EXPLOITING STRUCTURAL PROPERTIES IN THE ANALYSIS OF HIGH-DIMENSIONAL DYNAMICAL SYSTEMS

by
Hancheng Min

A dissertation submitted to Johns Hopkins University in conformity
with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
JULY, 2023

Abstract

The physical and cyber domains with which we interact are filled with high-dimensional dynamical systems. In machine learning, for instance, the evolution of overparametrized neural networks can be seen as a dynamical system. In networked systems, numerous agents or nodes dynamically interact with each other. A deep understanding of these systems can enable us to predict their behavior, identify potential pitfalls, and devise effective solutions for optimal outcomes. In this dissertation, we will discuss two classes of high-dimensional dynamical systems with specific structural properties that aid in understanding their dynamic behavior.

In the first scenario, we consider the training dynamics of multi-layer neural networks. The high dimensionality comes from overparametrization: a typical network has a large depth and hidden layer width. We are interested in the following question regarding convergence: Do network weights converge to an equilibrium point corresponding to a global minimum of our training loss, and how fast is the convergence rate? The key to those questions is the symmetry of the weights, a critical property induced by the multi-layer architecture. Such symmetry leads to a set of time-invariant quantities, called weight imbalance, that restrict the training trajectory to a low-dimensional manifold defined by the weight initialization. A tailored convergence analysis is developed over this low-dimensional manifold, showing improved rate bounds for several multi-layer network models studied in the literature, leading to novel characterizations of the effect of weight imbalance on the convergence rate.

In the second scenario, we consider large-scale networked systems with multiple weakly-connected groups. Such a multi-cluster structure leads to a time-scale separation between the fast intra-group interaction due to high intra-group connectivity, and the slow inter-group oscillation, due to the weak inter-group connection. We develop a novel frequency-domain network coherence analysis that captures both the coherent behavior within each group, and the dynamical interaction between groups, leading to a structure-preserving model-reduction methodology for large-scale dynamic networks with multiple clusters under general node dynamics assumptions.

Thesis Committee

Dr. Enrique Mallada (Primary Advisor)
Associate Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. René Vidal
Rachleff University Professor
Department of Electrical and Systems Engineering and Department of Radiology
University of Pennsylvania

Dr. Pablo Iglesias
Edward J. Schaefer Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. Mahyar Fazlyab
Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dedicated to My Family.

Acknowledgements

Throughout the course of my five-year Ph.D. study, I have been fortunate to receive invaluable assistance and unwavering support from numerous individuals.

I would like to first thank my advisor, Enrique Mallada, who is a wonderful advisor, collaborator, and also friend. What I appreciate the most from him as an advisor is the friendly and stress-free environment he creates between him and his advisee. Every week, I can bring up anything in the individual meeting with him. I talk about my progress and breakthroughs, also my concerns and struggles, and he is there to listen, guide, and help. As a researcher, he is one of the most diligent, intelligent, and passionate people I ever met. He is always open to discussion on research, full of creativity in problem-solving and excited about exploring new research avenues. I have learned from him a lot, from identifying and formulating research problems, to writing papers and preparing for presentations. Outside of academics and research, he is a great friend. He cares about our physical and mental health. He hosts group activities that bring lab members together like a family, and he cooks great Uruguayan Barbeque. I am fortunate to have him as my advisor and it makes my Ph.D. study even more enjoyable.

I would like to also thank my co-advisor, René Vidal, who is also among one of the most diligent, intelligent, and passionate researchers I ever met. The most important thing I learned from him is how to communicate as a researcher: how to express opinions in a research meeting, how to improve the clarity of a paper, how to deliver a poster or presentation efficiently, and more. He is the one who makes me realize the importance of these “non-math” parts of being a good researcher.

I thank Pablo Iglesias and Mahyar Fazlyab for serving on my thesis committee and for making valuable comments and suggestions on my dissertation. I also thank Jim Fill and Carey Priebe for serving on my graduate board oral exam committee.

I am fortunate to be able to collaborate with many people, including Fernando Paganini, Richard Pates, Juan Barzeque, Agustin Castellano, Salma Tarmoun, and Ziqing Xu.

I want to further extend my appreciation to my colleagues: Pengcheng You, Yan Jiang, Chengda Ji, Yue Shen, Tianqi Zheng, Rajni Kant Bansal, Eli Pivo, Mustafa Devrim Kaba, Haralampos Avraam, Gary Gao, Dhananjay Anand, Eliza Cohn, Jay Guthrie, and Roy Siegelmann from NetD Lab; Ambar Pal, Aditya Chattopadhyay, Kyle Poe, Liangzu Peng, Tianjiao Ding, Yutao Tang, Guilherme Franca, Paris Giampouras, Joshua Agterberg, Ryan Chan, Kaleab Kinfu, Darshan Thaker, and Carolina Pacheco from VisionLab.

I would like to end with my special thank to my family: my parents Yuanfu Min and Chunhong Tan, and my aunts Li Zhang and Yixing Fang. They encouraged me to embark on a journey of Ph.D. study, and I truly thank their endless love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The physical and cyber domains with which we interact are filled with high-dimensional dynamical systems. In machine learning, for instance, the evolution of overparametrized neural networks can be seen as a dynamical system. In networked systems, numerous agents or nodes dynamically interact with each other. A deep understanding of these systems can enable us to predict their behavior, identify potential pitfalls, and devise effective solutions for optimal outcomes. In this dissertation, we will discuss two classes of high-dimensional dynamical systems with specific structural properties that aid in understanding their dynamic behavior.

One is the *autonomous behavior*: starting from some initial state, how does the system evolve under no exogenous disturbance? In the long term, does it converge to some stable equilibrium point or exhibit some periodic behavior that corresponds to the desired operation of the system? The dynamical behavior of a low-dimensional system, if non-chaotic, can be generally inferred by first finding all of its stationary points, where time derivatives of all states are zero, then analyzing the stability around those points. However, when the number of states is large, there are generally infinitely many stationary points. One does not have an explicit characterization of those points, not to mention the difficulties in analyzing the system stability around those points. Without a generally applicable approach, understanding a given high-dimensional system heavily relies on its special properties, but what intrinsic properties could be useful in the analysis?

Another is the *input-output response*: upon reaching some stable equilibrium, which corresponds to a normal operation point for physical systems, the system is generally subject to exogenous disturbances. How does the system respond to those disturbances? and will the states leave the designated safe operation region? Those questions can be mostly answered by studying the linearized system around the stable equilibrium (normal operation point). However, for a high-dimensional dynamical system, its linearized system has the same number of states as the original one, and usually, there are as many sources of disturbance. Hence one needs to analyze a multi-input multi-output linear system with many internal states. Given a high-dimensional system, is there a way to approximate the linearized system by one with a much smaller scale? What properties of the system would allow such an approximation?

In my dissertation, I study two classes of high-dimensional dynamical systems with special structural properties that can be used to understand their dynamical behavior.

- The first scenario considers the training dynamics of multi-layer linear networks under some optimization algorithm. The system is high-dimensional due to the *overparametrization*: a typical neural network has a large depth and hidden layer width. Hence, there is an extremely large number of trainable weights evolving as the training iteration proceeds. We are interested in the following question regarding convergence: Do those network weights converge to an equilibrium point corresponding to a global minimum of our training loss, and if so, how fast is the convergence rate? The key to those questions about autonomous behavior is the *symmetry* of the weights, an important property induced by the multi-layer architecture: A class of transformations of the weights exists under which the corresponding input-output map remains the same. Such symmetry leads to a set of time-invariant quantities that depend on the differences between the weight matrices of adjacent layers, called *weight imbalance*. Such invariance restricts the training trajectory to a low-dimensional manifold defined by the weights initialization. A tailored convergence analysis is developed over this low-dimensional manifold, showing improved rate bounds for several multi-layer network models studied in the literature, leading to novel characterizations of the effect of weight imbalance on convergence.

- The second scenario considers large-scale networked systems with multiple weakly-connected groups. Upon reaching some equilibrium point, the network is often subject to exogenous but small disturbances at the individual node level. Understanding how some disturbance to a subset of nodes would affect all other nodes amounts to studying the transfer matrix of the entire network, making the analysis generally challenging due to the large scale of the system. However, the systems' multi-cluster structure leads to a time-scale separation between the fast intra-group interaction–due to high intra-group connectivity–and the slow inter-group coupling–due to the weak inter-group connection. As a result, the transfer matrix of the network has a low-rank structure. Building on my recent work on network coherence analysis in the Laplace domain, the input-output response can be well characterized by a structure-preserving reduced network model with the same size as the number of clusters.

1.1 Training Dynamics of Neural Networks

Training a neural network using a gradient descent algorithm with a small step size can be understood by studying a continuous-time gradient system $\dot{\theta} = -\nabla\mathcal{L}(\theta)$, called *gradient flow* (GF) dynamics, where θ contains all trainable weights and $\mathcal{L}$ is the empirical loss defined under certain training data. Due to overparametrization, there are generally infinitely many global minimums of $\mathcal{L}$ as well as local minimums, which are all stable equilibria of the GF. For what initializations do trajectories converge to global minimums? If so, at what rate?

1.1.1 Prior work

A vast body of work has tried to theoretically understand this phenomenon by analyzing either the loss landscape or the training dynamics of the network parameters from a specific initialization.

The *landscape-based analysis* is motivated by the empirical observation that deep neural networks used in practice often have a benign landscape [1], which can facilitate convergence. Existing theoretical analysis [2, 3, 4] shows that gradient descent converges when the loss function satisfies the following properties: 1) all of its local minimums are global minimums; and 2) every saddle point has a Hessian with at least one strict negative eigenvalue. Prior work suggests that the matrix factorization model [5], shallow networks [6], and certain positively homogeneous networks [7, 8] have such a landscape property, but unfortunately condition 2) does not hold for networks with multiple hidden layers [6]. Moreover, the landscape-based analysis generally fails to provide a good characterization of the convergence rate, except for a local rate around the equilibrium [2, 5]. In fact, during early stages of training, gradient descent could take exponential time to escape some saddle points if not initialized properly [9].

The *trajectory-based* analyses study the training dynamics of the weights given a specific initialization. For example, the case of small initialization has been studied for various models [10, 11, 12, 13, 14, 15, 16, 17]. Under this type of initialization, the trained model is implicitly biased towards low-rank [10, 11, 12, 13, 14, 17], or sparse [15] models. While the analysis for small initialization gives rich insights on the generalization of neural networks, the number of iterations required for gradient descent to find a good model often increases as the initialization scale decreases. Such dependence proves to be logarithmic on the reciprocal of the initialization scale for symmetric matrix factorization model [12, 13, 14], but for deep networks, existing analysis at best shows a polynomial dependency [15]. Therefore, the analysis for small initialization, while insightful in understanding the implicit bias of neural network training, is not suitable for understanding the training efficiency in practice since small initialization is rarely implemented due to its slow convergence. Another line of work studies the initialization in the kernel regime, where a randomly initialized sufficiently wide neural network can be well approximated by its linearization at initialization [18, 19, 20]. In this regime, gradient descent enjoys a linear rate of convergence toward the global minimum [21, 22, 23]. However, the width requirement in the analysis is often unrealistic, and empirical evidence has shown that practical neural networks generally do not operate in the kernel regime [19].

The study of non-small and non-kernel-regime initialization has been mostly centered around linear models. For matrix factorization models, spectral initialization [24, 11, 25] allows for decoupling the training dynamics into several scalar dynamics. For non-spectral initialization, the notion of weight *imbalance*, a quantity that depends on the differences between the weights matrices of adjacent layers, is crucial in most analyses. When the initialization is balanced, i.e., when the imbalance matrices are zero, it is sufficient for convergence when initial end-to-end linear model is close to its optimum [26, 27]. The effect of weight imbalance on the convergence has been only studied in the case where all imbalance matrices are positive semi-definite [28], which is often unrealistic in practice. Therefore, a convergence analysis that applies to deep linear networks under general initialization is still missing.

1.1.2 Thesis contribution

The contribution of this thesis to the understanding of the training dynamics of neural networks is twofold: convergence and implicit bias. Regarding **convergence**, we show that [29] the convergence of gradient flow for linear networks explicitly depends on two trajectory-specific quantities: 1) the *imbalance matrices*, which measure the difference between the weights of adjacent layers, and 2) a lower bound on the least singular values of *weight product* $W = W_1 W_2 \cdots W_L$. The former is time-invariant under gradient flow, thus determined at initialization, while the latter can be controlled by initializing the product sufficiently close to its optimum. With such observation, we provide two conditions on the initialization, *sufficient imbalance* and *sufficient margin*, with either of them being sufficient for guaranteeing convergence. Our results apply to various loss functions commonly used in regression tasks as well as those in classification tasks.

Specifically, for two-layer linear networks [30, 31], our convergence rate bound depends on two important quantities, *imbalance spectrum spread* and *spectral gap*, whose trade-off leads to a different rate characterization that applies to randomly initialized networks with varying width. Moreover, we provide a rate bound that applies to three-layer networks under general initialization. For deep networks, we study a broader class of initialization that covers most initialization schemes used in prior work [24, 25, 26, 27, 30, 28] for both multi-layer linear networks and diagonal linear networks while providing an improved rate bound. In addition, the analysis can be applied to late-stage training dynamics of two-layer ReLU networks under small initialization.

Regarding **implicit bias**, we study gradient flow for overparametrized two-layer linear networks [30, 31]. We show the existence of a subset of the parameter space defined by an orthogonality condition, which is invariant under gradient flow. All trajectories within this invariant set converge to a unique minimizer (w.r.t. the end-to-end function), which corresponds to the min-norm solution. As a result, initializing the network within this invariant set always yields the min-norm solution upon convergence. Next, we show that if we initialize each network weight as a sample from the distribution $\mathcal{N}(0, 1/h^{2\alpha})$ (where h is the hidden layer width and $1/4 < \alpha \leq 1/2$), then it holds with high probability that 1) the weight imbalance has sufficient spectral gap for exponential convergence; and 2) the aforementioned orthogonality condition is approximately satisfied throughout training. This results in a $\mathcal{O}(h^{2\alpha-\frac{1}{2}})$ upper bound on the operator norm distance between the trained network and the min-norm solution.

Finally, we provide a complete analysis [32] of the dynamics of gradient flow for the problem of training a two-layer ReLU network on well-separated data under the assumption of small initialization. Specifically, we show that if the initialization is sufficiently small, during the early phase of training the neurons in the first layer try to align with either the positive data or the negative data, depending on its corresponding weight on the second layer. Moreover, through a careful analysis of the neuron's directional dynamics we show that the time it takes for all neurons to achieve good alignment with the input data is upper bounded by $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$, where n is the number of data points and μ measures how well the data are separated. We also show that after the early alignment phase the loss converges to zero at a $\mathcal{O}(\frac{1}{t})$ rate and that the weight matrix on the first layer is approximately low-rank.

1.2 Network Coherence

Networked systems are formed by large group of agents/nodes dynamically interacting with each other through a communication network. Examples include social networks [33], power networks [34], and transportation networks [35]. The simplest but arguably the most important coordinated behavior in those networks is *consensus* [36, 37, 33]: each agent keeps updating its own state based on the neighboring state information until agreement among all the nodes is achieved. More interestingly, upon reaching a consensus, the agents have the ability to collectively respond to some exogenous disturbance while still maintaining some level of agreement on the state, which is often referred to as *coherence*.

1.2.1 Prior work

Classic slow coherence analyses [38, 39, 40, 41, 42] (with applications mostly to power networks) usually consider the second-order electro-mechanical model without damping: $\ddot{x} = -M^{-1}Lx$, where M is the diagonal matrix of machine inertias, and L is the Laplacian matrix whose elements are synchronizing coefficients between pair of machines. The coherency or synchrony [39] (a generalized notion of coherency) is identified by studying the first few slowest eigenmodes (eigenvectors with small eigenvalues) of $M^{-1}L$. The analysis can be carried over to the case of uniform [38] and non-uniform [40] damping. For a group of nodes that exhibit coherent behavior, one can construct dynamic equivalents [38, 39] that characterize the slow (coherent) behavior. Finding the dynamic equivalent, or an aggregate model, for interconnected power generators is a long-standing research subject in power system literature. Previously proposed aggregation model [43, 44, 45, 46, 47, 48, 40] mostly assume first- or second-order generator dynamics. As such, these state-space-based analyses are limited to very specific node dynamics and do not account for more complex dynamics or controllers that are usually present at a node level; e.g., in the power systems literature [49, 50, 51]. There is, therefore, the need for coherence identification and aggregation procedures that work for more general network systems. Moreover, it is widely known that such coherence is related to strong interconnection among the nodes, such relation is not formally justified in the aforementioned slow coherency analyses.

A vast body of work, triggered by the seminal paper [35], has quantitatively studied the role of the network topology in the emergence of coherence. Examples include, directed [52] and undirected [53] consensus networks, transportation networks [35], and power networks [48, 54, 55, 56]. The key technical approach amounts to quantify the level of coherence by computing the $\mathcal{H}_2$-norm of the system for appropriately defined nodal disturbance and performance signals. Broadly speaking, the analysis shows a reciprocal dependence between the performance metrics and the non-zero eigenvalues of the network graph Laplacian, validating the fact that strong network coherence (low $\mathcal{H}_2$-norm) results from the high connectivity of the network (large Laplacian eigenvalues). Unfortunately, the analysis strongly relies on a homogeneity [35, 52, 53, 54, 55, 56] or proportionality [48] assumption of the nodal transfer functions, and thus fails to characterize how individual heterogeneous node dynamics affect the overall coherent network response. Moreover, those analyses have not been generalized to multi-cluster networks. Therefore, a theoretical analysis that connects the network topology to the emergence of coherence in multi-cluster network systems with heterogeneous node dynamics is still missing.

1.2.2 Thesis contribution

In this thesis, we make the following contribution to the understanding of network coherence in large-scale systems:

We present a general framework in the frequency domain to analyze the coherence of heterogeneous networks [57, 58]. We show that network coherence emerges as a low-rank structure of the system transfer matrix as we increase the effective algebraic connectivity–a frequency-varying quantity that depends on the network coupling strength and dynamics. Unlike prior work [38, 39, 40, 41, 42], our analysis applies to networks with heterogeneous nodal dynamics, and further provides an explicit characterization in the frequency domain of the coherent response to disturbances as the harmonic mean of individual nodal dynamics. Thus, in this way, our results highlight the contribution of individual nodal dynamics to the network's coherent behavior. We also propose a balanced-truncation-based model reduction algorithm in order to reduce the complexity of obtained coherent dynamics.

We formally connect our frequency-domain results with explicit time-domain L_∞ bounds on the difference between individual nodal responses and the coherent dynamic response to certain classes of input signals, suggesting that network coherence is a frequency-dependent phenomenon. That is, the ability of nodes to respond coherently depends on the frequency composition of the input disturbance.

We extend our frequency-domain analysis to the case of multi-cluster network systems [59, 60]. We propose a structure-preserving model-reduction methodology for large-scale dynamic networks. Our analysis shows that networks with multiple coherent groups can be well approximated by a reduced network of the same size as the number of coherent groups, and we provide an upper bound on the approximation error when the network graph is randomly generated from a weight stochastic block model.

Chapter 2

Learning Dynamics of Overparametrized Neural Networks

In this chapter, we study the problem of training overparametrized neural networks. Consider a training dataset $\mathcal{D}$, a neural network parametrized by θ, and a loss function $\mathcal{L}(\theta; \mathcal{D})$[1], one seeks an optimal θ^* that solves

$$\min_{\theta} \mathcal{L}(\theta; \mathcal{D}) \,. \tag{2.1}$$

Generally, there is no closed-form solution to (2.1), and in practice, people use gradient-based algorithms to find an optimal θ^*. Arguably the simplest algorithm is the gradient descent (GD) algorithm:

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_\theta \mathcal{L}(\theta^{(k)}; \mathcal{D}),\ \theta^{(0)} = \theta_0 \,, \tag{2.2}$$

where $\nabla_\theta \mathcal{L}(\theta; \mathcal{D})$ is the gradient of $\mathcal{L}$ w.r.t. θ (for simplicity, here we assume $\mathcal{L}$ is differentiable). That is, the GD algorithm begins with some initialization θ_0, and at every iteration, moves the parameter θ along the direction of the negative gradient of $\mathcal{L}(\theta; \mathcal{D})$ with a step size η.

In typical deep learning applications, the neural networks are highly overparametrized: they consist of a large number of hidden layers, and there are many hidden neurons within each layer, resulting in millions to billions of trainable

[1] Exact definitions for $\mathcal{D}$, θ, and $\mathcal{L}$ are problem-dependent, which will be clear in later sections

weights. Such overparametrization makes the loss $\mathcal{L}$ highly non-convex with respect to the training parameters and leaves no theoretical guarantee that the GD algorithm would find a global minimum θ^*. However, in practice, it is generally the case that the GD algorithm converges to some θ^* that achieves zero loss.

One major part of the efforts in theoretically explaining such a phenomenon has been focused on studying the following *gradient flow* (GF) dynamics:

$$\dot{\theta} = -\nabla_\theta \mathcal{L}(\theta; \mathcal{D}),\ \theta(0) = \theta_0\,, \tag{2.3}$$

whose trajectory approximates those from GD with a small step size. Despite being easier to analyze than the discrete-time GD algorithm, the question of convergence largely remains the same for the GF dynamics: given some initialization θ_0, how do we know whether θ converges to a stationary point that corresponds to a global minimum of the loss? If so, is the convergence exponential? Additionally, there are many global minimums of the loss due to overparametrization, and they generally have different test errors, i.e., their performance varies in predicting a new data point that is different from those in the training data. How do we know if θ converges to a global minimum that achieves low test error?

The first question mostly concerns convergence, i.e., finding a good initialization that achieves zero loss asymptotically. The second question concerns the implicit bias of GF dynamics, i.e., understanding the special property of θ^* to which the training parameter θ converges. We will address these aforementioned questions mostly in the case of linear networks, with some discussion on nonlinear networks such as ReLU networks.

Chapter outline

This chapter is organized as follows: In Section 2.1, we study the problem of training two-layer linear networks under the l_2 loss, showing that the convergence of gradient flow for linear networks explicitly depends on two factors: 1) a weight imbalance matrix; and 2) the weight product matrix. With such observations, we provide two conditions on the initialization, namely *sufficient imbalance* and *sufficient margin*, either of which is sufficient to guarantee exponential convergence. Additionally, we demonstrate that an orthogonal condition on the initialization results in an implicit bias towards the min-norm solution for the underlying regression problem. In Section 2.2, we extend the convergence analysis of two-layer linear networks to multi-layer networks, showing that the convergence rate still depends on an imbalance matrix and the weight product. The general analysis considers a wide range of loss functions used for both regression and classification problems. Finally, in Section 2.3, we investigate the problem of training two-layer ReLU networks under small initialization. This problem involves two training phases: the first phase relies on our novel analysis of the directional dynamics of the neuron, while the second phase is related to training linear networks. Here, one can apply the convergence analysis of linear networks to demonstrate convergence and characterize the implicit bias of the weights.

Notation

For a matrix A, we denote its transpose as $A^\top$, its trace as $\mathrm{tr}(A)$, and its i-th eigenvalue and singular value as $\lambda_i(A)$ and $\sigma_i(A)$, respectively, in decreasing order (when adequate). For an $n \times m$ matrix A, we write $\sigma_{\min}(A) = \sigma_{\min\{n,m\}}(A)$, and we conventionally let $\lambda_i(A) = 0$ and $\sigma_i(A) = 0, \forall i > \min\{m, n\}$. Also, we let $[A]_{ij}$, $[A]_{i,:}$, and $[A]_{:,j}$ denote the (i, j)-th element, the i-th row and the j-th column of A, respectively, and we let $\|A\|_2$ and $\|A\|_F$ denote the spectral norm and the Frobenius norm of A, respectively. For a symmetric matrix A, we write $A \succ 0$, $A \succeq 0$, $A \prec 0$ or $A \preceq 0$ when A is positive definite, positive semi-definite, negative definite or negative semi-definite, respectively, and we write $A \succ B$, $A \succeq B$, $A \prec B$ and $A \preceq B$ to denote $A - B \succ 0$, $A - B \succeq 0$, $A - B \prec 0$ and $A - B \preceq 0$, respectively. For a scalar-valued or matrix-valued function of time, $F(t)$, we let $\dot{F} = \dot{F}(t) = \frac{d}{dt}F(t)$ denote its time derivative. Additionally, we let I_n denote the identity matrix of order n and $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2.

2.1 Two-layer linear networks

This section presents a novel analysis of the gradient flow dynamics of overparametrized two-layer linear networks, which provides a common set of conditions on initialization that lead to convergence and implicit bias. Specifically, this section makes the following contributions:

1. In Section 2.1.2, we show that the convergence of gradient flow for linear networks explicitly depends on: 1) a weight imbalance matrix; and 2) the weight product matrix. With such observation, we provide two conditions on the initialization, *sufficient imbalance* and *sufficient margin*, with either of them being sufficient for guaranteeing exponential convergence. Moreover, our convergence rate bound depends on two important quantities, *imbalance spectrum spread* and *spectral gap*, whose trade-off leads to a different rate characterization that applies to randomly initialized networks with varying width.

2. In Section 2.1.3 we study the implicit bias of gradient flow for overparametrized two-layer linear networks. We show the existence of a subset of the parameter space defined by an orthogonality condition, which is invariant under gradient flow. All trajectories within this invariant set converge to a unique minimizer (w.r.t. the end-to-end function), which corresponds to the min-norm solution. As a result, initializing the network within this invariant set always yields the min-norm solution upon convergence.

3. In Section 2.1.4, we show that if we initialize each network weight as a sample from the distribution $\mathcal{N}(0, 1/h^{2\alpha})$ (where h is the hidden layer width and $1/4 < \alpha \leq 1/2$), then it holds with high probability that 1) the weight imbalance has sufficient spectral gap for exponential convergence; and 2) the aforementioned orthogonality condition is approximately satisfied throughout training. This results in a $\mathcal{O}(h^{2\alpha-\frac{1}{2}})$ upper bound on the operator norm distance between the trained network and the min-norm solution.

Our analysis requires neither infinite width nor specific initializations such as spectral, balanced or random. Instead, we reveal general properties of initialization that facilitate the convergence and implicit bias, and show how prior work is related to these general properties in various ways. Moreover, our results provide new insights on how the network width and the level of overparametrization affect both the convergence and implicit bias. Hence, this analysis formally connects initialization, exponential convergence of the optimization task, overparametrization and implicit bias.

2.1.1 Problem setup

We study the dynamics of gradient flow for two-layer linear networks trained with the squared l_2-loss. More specifically, given N training samples $\{x^{(i)}, y^{(i)}\}_{i=1}^N$, where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathbb{R}^m$, we aim to solve the following linear regression problem

$$\min_{\Theta \in \mathbb{R}^{n\times m}} \mathcal{L} = \frac{1}{2}\sum_{i=1}^{N} \|y^{(i)} - \Theta^\top x^{(i)}\|_2^2 = \frac{1}{2}\|Y - X\Theta\|_F^2 \,, \tag{2.4}$$

where $Y = [y^{(1)}, \cdots, y^{(N)}]^\top \in \mathbb{R}^{N\times m}$ and $X = [x^{(1)}, \cdots, x^{(N)}]^\top \in \mathbb{R}^{N\times n}$. We do so by training a two-layer linear network $y = f(x; V, U) = VU^\top x$, where $V = \mathbb{R}^{m\times h}$, $U \in \mathbb{R}^{n\times h}$ and h is the hidden layer width, with gradient flow, i.e., gradient descent with "infinitesimal step size". In particular, we consider an *overparametrized* model [27, 25] such that $h \geq \min\{n, m\}$, i.e. there is no rank constraint on the linear model Θ obtained from the linear network $UV^\top$.

After rewriting the loss with respect to the parameters V and U,

$$\mathcal{L}(V, U) = \frac{1}{2} \sum_{i=1}^{N} \|y^{(i)} - VU^\top x^{(i)}\|_2^2 = \frac{1}{2} \|Y - XUV^\top\|_F^2 \,, \tag{2.5}$$

the gradient flow dynamics are given by

$$\dot{V}(t) = -\frac{\partial \mathcal{L}}{\partial V}(V(t), U(t)) = (Y - XU(t)V^\top(t))^\top XU(t) \,, \tag{2.6a}$$

$$\dot{U}(t) = -\frac{\partial \mathcal{L}}{\partial U}(V(t), U(t)) = X^\top (Y - XU(t)V^\top(t))V(t) \,. \tag{2.6b}$$

Remark 1. *For the remainder of this chapter, we drop the explicit dependence of scalar/matrix functions of time on the time parameter t whenever such dependence is clear from the context. For example, we will mostly write U and $\dot{U}$ instead of $U(t)$ and $\dot{U}(t)$, respectively.*

Our analysis requires reparametrization of the gradient flow dynamics. We let $r = \operatorname{rank}(X)$ and first consider the case $n > r$. The singular value decomposition (SVD) of X can be written as

$$X = W \begin{bmatrix} \Sigma_x^{1/2} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^\top \\ \Phi_2^\top \end{bmatrix} = W \Sigma_x^{1/2} \Phi_1^\top \,, \tag{2.7}$$

where $W \in \mathbb{R}^{N \times r}$, $\Phi_1 \in \mathbb{R}^{n \times r}$, and $\Phi_2 \in \mathbb{R}^{n \times (n-r)}$. Since $\Phi_1 \Phi_1^\top + \Phi_2 \Phi_2^\top = I_n$, we have

$$U = I_n U = (\Phi_1 \Phi_1^\top + \Phi_2 \Phi_2^\top) U = \Phi_1 \Phi_1^\top U + \Phi_2 \Phi_2^\top U \,, \tag{2.8}$$

and hence we can parametrize U as (U_1, U_2) using the bijection $U = \Phi_1 U_1 + \Phi_2 U_2$, with inverse $(U_1, U_2) = (\Phi_1^\top U, \Phi_2^\top U)$.

With this parametrization, we can rewrite the gradient flow in (2.6a)-(2.6b) explicitly as

$$\dot{V} = \left(Y - XUV^\top\right)^\top XU = E^\top \Sigma_x^{1/2} \Phi_1^\top U \,, \tag{2.9a}$$

$$\dot{U} = X^\top \left(Y - XUV^\top\right) V = \Phi_1 \Sigma_x^{1/2} E V \,, \tag{2.9b}$$

where

$$E = E(V, U_1) := W^\top(Y - XUV^\top) = W^\top Y - \Sigma_x^{1/2} U_1 V^\top \,, \tag{2.10}$$

is defined to be the *error*. Then from (2.9a)-(2.9b) we obtain the dynamics in the parameter space (V, U_1, U_2) as

$$\dot{V} = E^\top \Sigma_x^{1/2} U_1 \,,\; \dot{U}_1 = \Sigma_x^{1/2} E V \,,\; \dot{U}_2 = 0 \,. \tag{2.11}$$

Moreover, since W has orthonormal columns, we notice that

$$\begin{aligned}\mathcal{L}(V, U) =&\; \frac{1}{2}\|Y - XUV^\top\|_F^2 \\ =&\; \frac{1}{2}\|(I - WW^\top)(Y - XUV^\top) + WW^\top(Y - XUV^\top)\|_F^2 \\ =&\; \frac{1}{2}\|(I - WW^\top)Y + WE\|_F^2 = \frac{1}{2}\|WE\|_F^2 + \frac{1}{2}\|(I - WW^\top)Y\|_F^2 \\ =&\; \frac{1}{2}\|E\|_F^2 + \frac{1}{2}\|(I - WW^\top)Y\|_F^2 . \end{aligned} \tag{2.12}$$

Here the last term in (2.12) does not depend on V, U, and it is the residual

$$\mathcal{L}^* = \frac{1}{2}\|(I - WW^\top)Y\|_F^2 \,, \tag{2.13}$$

which is also the optimal value of (2.4). Therefore, for convergence, it suffices to analyze the convergence of the error E under the dynamics of (V, U_1) in (2.11). The remaining parameters U_2 are constant, i.e., $U_2(t) \equiv U_2(0)$ throughout the training trajectory, and the role of U_2 will be discussed when we study the implicit bias in Section 2.1.3.

Remark 2. *In the case of $n = r$, the SVD of X is $W\Sigma_x^{1/2}\Phi_1^\top$ with $\Phi_1 \in \mathbb{R}^{n\times n}$, i.e., Φ_2 is an empty matrix, and so is U_2. The convergence analysis follows exactly the same as those to be presented in Section 2.1.2. However, all global minimizer (U^*, V^*) of $\mathcal{L}(U, V)$ corresponds to the unique minimizer Φ^* of (2.4), and thus there is no need to study the implicit bias for this case.*

2.1.2 Convergence analysis

In this section, we study the convergence of gradient flow for the reparametrized dynamics

$$\dot{V} = E^\top \Sigma_x^{1/2} U_1 \,,\ \dot{U}_1 = \Sigma_x^{1/2} E V \,, \tag{2.14}$$

which is exactly the gradient flow dynamics of

$$\frac{1}{2}\|E\|_F^2 = \frac{1}{2}\|W^\top Y - \Sigma_x^{1/2} U_1 V^\top\|_F^2 \,. \tag{2.15}$$

In particular, when $\Sigma_x^{1/2} = I_r$, (2.15) reduces to $\frac{1}{2}\|W^\top Y - U_1 V^\top\|_F^2$, which is the loss function for a matrix factorization problem. To motivate our main result, we start with the simplest scalar version of this factorization problem.

Warm-up: scalar dynamics

Consider the gradient flow dynamics of the loss function $\mathcal{L}_s(u,v) = \frac{1}{2}|y - uv|^2$ given by

$$\dot{u} = (y - uv)v,\ \dot{v} = (y - uv)u \,. \tag{2.16}$$

One important feature of (2.16), is that the *imbalance* $d := u^2 - v^2$ is invariant under the gradient flow, namely

$$\dot{d} = 2u\dot{u} - 2v\dot{v} = 2(uv - vu)(y - uv) \equiv 0 \implies d(t) \equiv d(0). \tag{2.17}$$

One sufficient condition for exponential convergence of the loss $\mathcal{L}_s$ is a lower bound on the *instantaneous rate* $-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}$. To see this, notice that if there exist a constant $c > 0$ such that for all $t \geq 0$ we have $-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} \geq c > 0$, then

$$\int_0^t \frac{\dot{\mathcal{L}}_s(\tau)}{\mathcal{L}_s(\tau)} d\tau \leq \int_0^t -c d\tau \implies \log \frac{\mathcal{L}_s(t)}{\mathcal{L}_s(0)} \leq -ct \implies \mathcal{L}_s(t) \leq \exp(-ct)\mathcal{L}_s(0) \,. \tag{2.18}$$

Thus, a lower bound $c > 0$ on the instantaneous rate implies the loss converges to 0 exponentially at a rate at least c. Now under the scalar dynamics (2.16), one can verify that

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}=2(u^2+v^2)=2\sqrt{(u^2+v^2)^2}=\ 2\sqrt{(u^2-v^2)^2+4u^2v^2}\,. \tag{2.19}$$

Therefore, we have

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}=2(u^2+v^2)=2\sqrt{d^2+4(uv)^2}\,, \tag{2.20}$$

i.e. the instantaneous rate can be explicitly written as a function of the imbalance d and the product uv. More importantly, with proper initialization, we can control the value of d and uv throughout the entire trajectory to obtain the desired lower bound on (2.20). Specifically,

- Since the imbalance d is time-invariant, we have $d(t)=d(0)$. When $|d(0)|>0$, there is *sufficient imbalance* at initialization, and

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}=2\sqrt{d^2+4(uv)^2}\geq 2|d|=2|d(0)|\,. \tag{2.21}$$

- The product is tied to the loss function $\mathcal{L}_s=|y-uv|^2/2$, which is non-increasing since $\dot{\mathcal{L}}_s\leq 0$. This implies that $|y-uv|\leq|y-u(0)v(0)|$, from which it follows that $y-|y-u(0)v(0)|\leq uv\leq y+|y-u(0)v(0)|$, i.e. uv stays within a closed ball with radius $|y-u(0)v(0)|$ centered at y. Therefore, when $|y|-|y-u(0)v(0)|>0$, there is *sufficient margin* at initialization such that this ball is strictly bounded away from zero, and then so is $|uv|$:

$$|uv|\geq|y|-|y-uv|\geq|y|-|y-u(0)v(0)|\,, \tag{2.22}$$

  we have

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}=2\sqrt{d^2+4(uv)^2}\geq 4|uv|=4(|y|-|y-u(0)v(0)|)\,. \tag{2.23}$$

Combining the two observations above, we have

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}=2\sqrt{d^2+4(uv)^2}\geq 2\sqrt{d^2(0)+4(\max\{|y|-|y-u(0)v(0)|,0\})^2}\,. \tag{2.24}$$

That is, $\mathcal{L}_s$ converges to zero exponentially when either $|d(0)| > 0$ (sufficient imbalance) or $|y| - |y - u(0)v(0)| > 0$ (sufficient margin). One can carry out the analysis starting from any time epoch $t_0 > 0$: The convergence rate after time t_0 is lower bounded by

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} = 2\sqrt{d^2 + 4(uv)^2} \geq 2\sqrt{d^2(0) + 4(\max\{|y| - |y - u(t_0)v(t_0)|, 0\})^2}\,. \tag{2.25}$$

In particular, for an large time epoch t_0 (late-stage of the training), $u(t_0)v(t_0) \simeq y$ thus $-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} \simeq 2\sqrt{d^2 + 4y^2}$, and the margin term $\max\{|y| - |y - u(t_0)v(t_0)|, 0\} \simeq y$ well captures the effect of target y in determining the asymptotic rate. Moreover, similar computations can be done for the case where input data is considered, which corresponds to having a loss function $\tilde{\mathcal{L}}_s(u, v) = \frac{1}{2}|y - xuv|^2$. The instantaneous rate is given by $-\frac{d}{dt}\frac{\tilde{\mathcal{L}}_s}{\tilde{\mathcal{L}}_s} = 2x\sqrt{d^2 + 4(uv)^2}$, showing the effect of input data on the convergence rate.

Our main results in the next section show that such observation can be completely generalized to the matrix factorization problem, allowing us to derive exponential convergence guarantees for gradient flow on two-layer linear networks.

Main results

Now we turn to study the gradient dynamics in (2.11). Similar to the scalar dynamics, we define the *imbalance* of the two-layer linear network under input data X as

$$\textit{Imbalance}: \; D = U_1^\top U_1 - V^\top V \in \mathbb{R}^{h \times h}\,. \tag{2.26}$$

This imbalance matrix, as expected, is time-invariant under gradient flow dynamics (2.11). To see this, we compute the time derivative of $U_1^\top U_1$ and $V^\top V$ as

$$\frac{d}{dt}U_1^\top U_1 = \; \dot{U}_1^\top U_1 + U_1^\top \dot{U}_1 = V^\top E^\top \Sigma_x^{1/2} U_1 + U_1^\top \Sigma_x^{1/2} E V, \tag{2.27}$$

$$\frac{d}{dt}V^\top V = \; V^\top \dot{V} + \dot{V}^\top V = V^\top E^\top \Sigma_x^{1/2} U_1 + U_1^\top \Sigma_x^{1/2} E V\,. \tag{2.28}$$

Therefore, $\frac{d}{dt}U_1^\top U_1 \equiv \frac{d}{dt}V^\top V$, which implies that $\dot{D} = \frac{d}{dt}[U_1^\top U_1 - V^\top V] \equiv 0$. This time-invariant imbalance matrix has also been discussed in [26, 61].

Our first result is the lower bound on the instantaneous rate:

Proposition 2.1 (Bound on the instantaneous rate)**.** *Consider the continuous-time dynamics in* (2.11)*. Let* $\tilde{\mathcal{L}} := \mathcal{L} - \mathcal{L}^*$ *and* $D = U_1^\top U_1 - V^\top V$*, then we have*

$$-\frac{d}{dt}\frac{\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}} \geq \lambda_r(\Sigma_x)\Big(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_m^2(U_1V^\top)} \\ -\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_r^2(U_1V^\top)}\Big)\,, \tag{2.29}$$

where we define the following ***imbalance quantities****:*

$$\text{(Positive imbalance spectrum spread) } \Delta_+ = \max\{\lambda_1(D), 0\} - \max\{\lambda_r(D), 0\}\,, \tag{2.30}$$

$$\text{(Negative imbalance spectrum spread) } \Delta_- = \max\{\lambda_1(-D), 0\} - \max\{\lambda_m(-D), 0\}\,, \tag{2.31}$$

$$\text{(Imbalance spectral gap) } \underline{\Delta} = \max\{\lambda_r(D), 0\} + \max\{\lambda_m(-D), 0\}\,. \tag{2.32}$$

Imbalance quantities: First of all, the imbalance quantities Δ_+,Δ_- and $\underline{\Delta}$ are time-invariant because they are fully determined by the time-invariant imbalance matrix D. Therefore, we will always use Δ_+,Δ_- and $\underline{\Delta}$ to represent the values $\Delta_+(0)$,$\Delta_-(0)$ and $\underline{\Delta}(0)$ at initialization. Then, these imbalance quantities can be easily visualized when $h \geq r + m$. In this case, the imbalance matrix $D = U_1^\top U_1 - V^\top V$ has at least r non-negative eigenvalues and m non-positive eigenvalues, and the imbalance quantities are precisely the difference between some specific eigenvalues of the imbalance matrix, as shown in Figure 2-1. We will discuss how imbalance quantities affect convergence in the following remarks and in our numerical section.

Effect of imbalance and product: Our rate bound (2.29) reveals how weight imbalance D and weight product $U_1V^\top$ explicitly affect the convergence rate (the following assumes $\lambda_r(\Sigma_x) = 1$):

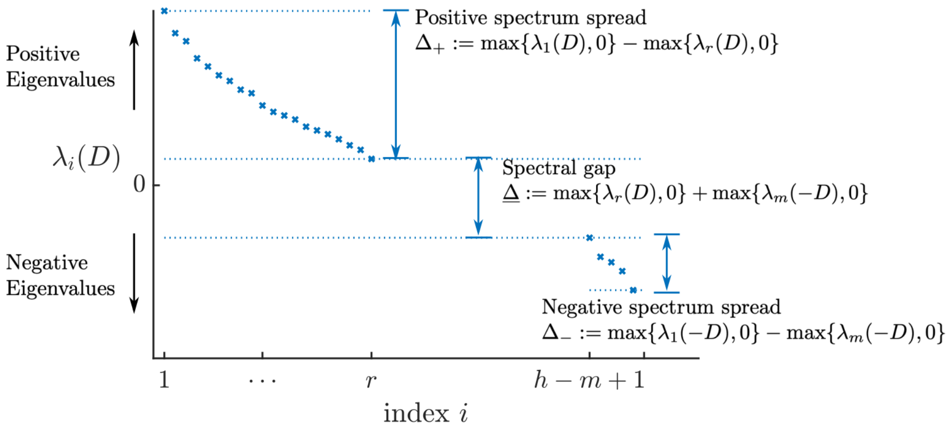
Positive spectrum spread
$\Delta_+ := \max\{\lambda_1(D), 0\} - \max\{\lambda_r(D), 0\}$
Positive Eigenvalues
$\lambda_i(D)$
0
Spectral gap
$\underline{\Delta} := \max\{\lambda_r(D), 0\} + \max\{\lambda_m(-D), 0\}$
Negative Eigenvalues
Negative spectrum spread
$\Delta_- := \max\{\lambda_1(-D), 0\} - \max\{\lambda_m(-D), 0\}$
1
$\cdots$
r
$h - m + 1$
index i

Figure 2-1. Illustration of imbalance quantities when $h \geq r + m$. The imbalance matrix D has rank at most $r + m$, we plot all its potentially non-zero eigenvalues.

1. (Effect of imbalance): Since

$$-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_m^2(U_1 V^\top)} \geq -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2} = \underline{\Delta}\,, \tag{2.33}$$

   it follows from (2.29) that $-\frac{\frac{d}{dt}\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}} \geq 2\underline{\Delta}$. Therefore, $2\underline{\Delta}$ is always a lower bound on the convergence rate. This means that, for initialization with an imbalance matrix bounded away from zero (characterized by $\underline{\Delta} > 0$), exponential convergence is guaranteed.

2. (Effect of product): The role of the product in (2.29) is more nuanced: Assume $n = m$ for simplicity so that $\sigma_n(U_1 V^\top) = \sigma_m(U_1 V^\top) = \sigma_{\min}(U_1 V^\top)$. We see that the non-negative quantities Δ_+ and Δ_- control how much the product affects the convergence. More precisely, the lower bound in (2.29) is a decreasing function of both Δ_+ and Δ_-. When $\Delta_+ = \Delta_- = 0$, the lower bound reduces to $\sqrt{\underline{\Delta}^2 + 4\sigma_{\min}^2(U_1 V^\top)}$, showing a joint contribution from both imbalance and product, which resembles (2.20) for the scalar case. However, as Δ_+ and Δ_- increase, the bound decreases towards $\underline{\Delta}$, which means that the effect of imbalance always exists, but the effect of the product diminishes for large Δ_+ and Δ_-.

The experiments in Section 2.1.5 show that, under random initialization, networks with large width fall into the first regime (Δ_+ and Δ_- are small), while networks with small width fall into the second regime (Δ_+ and Δ_- are large), and the loss trajectories behave differently in these two regimes.

Towards exponential convergence: As we illustrated with the scalar dynamics, the lower bound in Proposition 2.1, which depends explicitly on imbalance and product, is useful because one can control the two factors for the entire trajectory with proper initialization. This allows us to derive exponential convergence guarantees for the gradient flow, as stated in our main theorem next.

Theorem 2.1 (Exponential Convergence Guarantee)**.** *Consider the continuous dynamics in* (2.11)*. Let* $\tilde{Y} := W^\top Y$ *and define*

$$\begin{aligned} c(t) = & -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4(\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1(t) V(t)^\top\|_F, 0\})^2/\lambda_1(\Sigma_x)} \\ & - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4(\max\{\sigma_r(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1(t) V(t)^\top\|_F, 0\})^2/\lambda_1(\Sigma_x)}\,, \end{aligned} \tag{2.34}$$

where Δ_+, Δ_-*, and* $\underline{\Delta}$ *are defined as in* (2.30), (2.31) *and* (2.32)*, respectively.* $c(t) \geq 0, \forall t \geq 0$*, and we have*

$$(\mathcal{L}(t) - \mathcal{L}^*) \leq \exp\left(-\lambda_r(\Sigma_x) c(0) t\right) (\mathcal{L}(0) - \mathcal{L}^*), \forall t \geq 0\,. \tag{2.35}$$

That is, if $c(0) > 0$*, then the loss converges to its global minimum exponentially with a rate at least* $\lambda_r(\Sigma_x) c(0)$*.*

Theorem 2.1 unifies several previously discovered sufficient conditions for exponential convergence of the gradient flow on two-layer linear networks:

Corollary 2.1 (Sufficient imbalance [30])**.** *If at initialization,* $\underline{\Delta} > 0$*, then* $c(0) > 0$ *and the loss converges to zero exponentially with a rate at least* $\lambda_r(\Sigma_x) c(0) \geq 2\lambda_r(\Sigma_x)\underline{\Delta}$*.*

Proof. In (2.34), if we lower bound the term $(\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^\top\|_F, 0\})$ by 0, we have $c \geq 2\underline{\Delta} > 0$. □

Previous work [30] identifies the role of the *spectral gap* $\underline{\Delta}$ and proves the convergence result in Corollary 2.1. Our result generalizes it by showing the combined contribution of both the spectral gap and the margin to the convergence of the loss.

Corollary 2.2 (Sufficient margin). *If at initialization, $\sigma_{\min}(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^\top\|_F > 0$, then $c(0) > 0$ and the loss converges to zero exponentially with a rate at least $\lambda_r(\Sigma_x)c(0)$.*

Previous work [26] showed that when the initialization has a positive margin, i.e., $\sigma_{\min}(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^\top\|_F > 0$ and the imbalance has sufficiently small Frobenius norm (approximately balanced), then gradient flow converges exponentially. Corollary 2.2 improves upon it by showing that a positive margin is sufficient, regardless of the imbalance.

Corollary 2.3 (Characterizing local convergence rate). *If at some $t_0 > 0$, we have $c(t_0) > 0$, then*

$$(\mathcal{L}(t) - \mathcal{L}^*) \leq \exp\left(-\lambda_r(\Sigma_x)c(t_0)t\right)(\mathcal{L}(t_0) - \mathcal{L}^*), \forall t \geq t_0\,. \tag{2.36}$$

That is, after t_0, the loss converges to zero exponentially with a rate of at least $\lambda_r(\Sigma_x)c(t_0)$. Notably, given any trajectory that eventually converges to a global minimum for $\mathcal{L}$, for sufficiently large t_0, we have

$$\begin{aligned} c(t_0) \;\simeq\; & -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_m^2(\tilde{Y})/\lambda_1(\Sigma_x)} \\ & - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_r^2(\tilde{Y})/\lambda_1(\Sigma_x)}\,. \end{aligned} \tag{2.37}$$

For any trajectory that eventually converges, (2.37) is due to the fact that

$$\|\tilde{Y} - \Sigma_x^{1/2} U_1(t_0) V^\top(t_0)\|_F \simeq 0\,, \tag{2.38}$$

at sufficiently large t_0. This corollary suggests that the asymptotic convergence rate around the equilibrium depends on the imbalance D and on the training data (X, Y). Previous work [25] has shown that when $\Sigma_x = I_r$, $h = r = m$ and $D = \lambda I_h$ for some $\lambda \neq 0$, the asymptotic convergence rate of gradient flow is lower bounded by $2\sqrt{\lambda^2 + 4\sigma_{\min}^2(\tilde{Y})}$, and this can be exactly recovered from (2.37) with $\Delta_+ = \Delta_- = 0$ and $\underline{\Delta} = \lambda$. Our result has no additional assumption on the dimension nor on the imbalance structure.

The major limitation of previous analyses of convergence is the requirement that the initialization be exactly balanced [27] or homogeneously imbalanced [25]. These strong assumptions are made so that the dynamics of the product $U_1V^\top$ (the end-to-end function) can be solved for explicitly, from which the convergence results are derived. As illustrated in Figure 2-2, such analyses consider specific configurations in the parameter space and only allow for small variations [26]. Our analysis breaks such limitation by revealing fundamental relations between the convergence and the weight configuration (imbalance and product) , which provides convergence guarantees for a wide range of initializations.

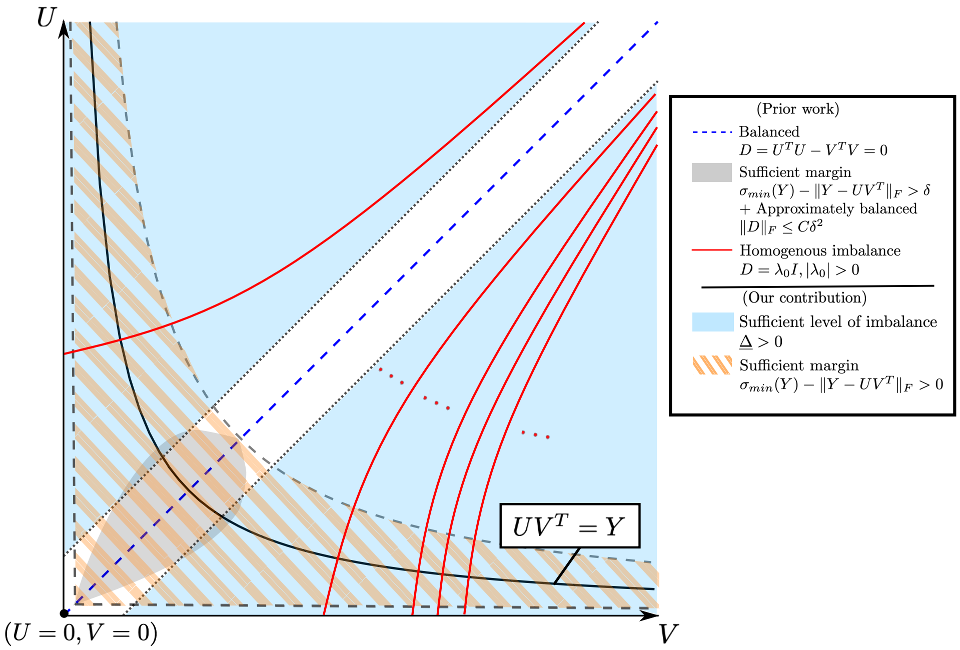

Figure 2-2. Illustration of non-spectral initialization studied for convergence of linear networks. Note: the conditions are presented for the gradient flow on $\frac{1}{2}\|Y - UV^\top\|^2$, which is the special case of ours when $X = I_n$.

2.1.3 Implicit bias

In the previous section, we studied two-layer linear networks trained with gradient flow and showed that the squared loss converges exponentially to the optimal loss with a rate that is determined by the imbalance and margin of the initialization. However, convergence of the loss does not necessarily imply convergence of the network weights. Moreover, since the end-to-end model $UV^\top$ does not uniquely determine the network weights U and V, what we actually care about is the convergence of $UV^\top$.

When $n = r = \mathrm{rank}(X)$ there is a unique end-to-end model $\Theta^* = X^\top(XX^\top)^{-1}Y$, thus we expect $UV^\top$ to converge to that model. However, when $n > r = \mathrm{rank}(X)$, the regression problem (2.4) has infinitely many solutions Θ^* that achieve the optimal loss. Therefore, it is important to understand which solution $UV^\top$ gradient flow converges to. Among all possible solutions, one that is of particular interest in high-dimensional linear regression is the *minimum norm solution* (min-norm solution)

$$\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{n\times m}}{\arg\min}\{\|\Theta\|_F : \|Y - X\Theta\|_F^2 = \min_{\Theta}\|Y - X\Theta\|_F^2\} = X^\top(XX^\top)^\dagger Y, \qquad (2.39)$$

because it has near-optimal generalization error for suitable data models [62, 63]. We are interested in conditions under which our trained network corresponds to an end-to-end model $UV^\top$ that is equal or close to the min-norm solution $\hat{\Theta}$.

In standard linear regression, where Θ follows the gradient flow on $\mathcal{L}(\Theta) = \frac{1}{2}\|Y - X\Theta\|_F^2$, it is well-known that one should decompose Θ into two parts: components $\Phi_1\Phi_1^\top\Theta$ within the subspace spanned by the data X and components $\Phi_2\Phi_2^\top\Theta$ orthogonal to the data subspace. Then $\Phi_1^\top\Theta$ converges to the min-norm solution $\hat{\Theta}$, and $\Phi_2\Phi_2^\top\Theta \equiv \Phi_2\Phi_2^\top\Theta(0)$ remains constant. Therefore, by decomposing the end-to-end model Θ into different components according to the data subspace and analyzing their dynamics separately, one derives the condition on $\Theta(0)$ for obtaining min-norm solution: $\Phi_2\Phi_2^\top\Theta(0) = 0$.

Although we are analyzing the dynamics of U, V instead of Θ, our approach is similar: We decompose the end-to-end model $UV^\top$ into $\Phi_1 U_1 V^\top$ and $\Phi_2 U_2 V^\top$ according to the SVD of data X, then showing that $\Phi_1 U_1 V^\top$ converges to $\hat{\Theta}$. The end-to-end model $UV^\top$ would converge to $\hat{\Theta}$ if $\Phi_2 U_2 V^\top \equiv 0$, for which we derive a sufficient condition that requires some orthogonality among $U_1(0), U_2(0), V(0)$ at initialization.

Decomposition of the end-to-end model: Notice that the end-to-end matrix $UV^\top \in \mathbb{R}^{n\times m}$ associated with the two-layer linear network can be decomposed according to the SVD of the data matrix X, (2.7), as

$$UV^\top = (\Phi_1\Phi_1^\top + \Phi_2\Phi_2^\top)UV^\top = \Phi_1 U_1 V^\top + \Phi_2 U_2 V^\top\,, \tag{2.40}$$

where Φ_1, Φ_2, U_1, and U_2 are defined in Section 2.1.2. The j-th column of $UV^\top$, $[UV^\top]_{:,j}$, is the linear predictor for the j-th output y_j, and is decomposed into two components within complementary subspaces $\mathrm{span}(\Phi_1)$ and $\mathrm{span}(\Phi_2)$. Moreover $[U_1V^\top]_{:,j}$ is the coordinate of $[UV^\top]_{:,j}$ w.r.t. the orthonormal basis consisting of the columns of Φ_1, and similarly $[U_2V^\top]_{:,j}$ is the coordinate w.r.t. basis Φ_2. Under gradient flow (2.11), the trajectory $U(t)V(t)^\top$, $t \geq 0$, is fully determined by the trajectories $U_1(t)V^\top(t)$ and $U_2(t)V^\top(t)$, $t \geq 0$.

Convergence of training parameters: First of all, we need to show that our end-to-end model $UV^\top$ converges to some $\hat{U}\hat{V}^\top$ before even analyzing how close $\hat{U}\hat{V}^\top$ is to $\hat{\Theta}$, which amounts to showing both $U_1V^\top$ and $U_2V^\top$ converges. We have shown in Section 2.1.2 that the error $E = W^\top Y - \Sigma_x^{1/2}U_1V^\top$ converges to zero (and the loss converges exponentially to $\mathcal{L}^*$) if $c(0) > 0$. This already shows $\lim_{t\to\infty} U_1V^\top = \Sigma_x^{-1/2}W^\top Y$. To show that $U_2V^\top$ converges to some $U_2(0)\hat{V}^T$ (U_2 part is time-invariant), we need to show that $\lim_{t\to\infty} V = \hat{V}$ for some $\hat{V}$. Yet, it is not immediate that from convergence of $U_1V^\top$ we know whether (U_1, V) would converge to some stationary point $(\hat{U}, \hat{V})$. Generally speaking, parameters in gradient flow dynamics either converge to some stationary point or diverge to infinity (their norms grow to infinity). Then to show (U_1, V) converges to some stationary point, one only needs to ensure the latter does not happen, as formally stated in the following proposition.

Proposition 2.2. *Consider the continuous dynamics in* (2.11). *If $c(0)$, defined in Theorem 2.1, is positive, then there exist some $\hat{V}$ and $\hat{U}_1$ such that $\lim_{t\to\infty} V(t) = \hat{V}$ and $\lim_{t\to\infty} U_1(t) = \hat{U}_1$. Moreover, $E(\hat{V}, \hat{U}_1) = W^\top Y - \Sigma_x^{1/2}\hat{U}_1\hat{V}^\top = 0$.*

In addition, notice that $\dot{U}_2(t) = 0$ in dynamics (2.11), hence $U_2(t) = U_2(0), \forall t > 0$. Therefore, when $c(0) > 0$, we know U_1 and V converge to $\hat{U}_1$ and $\hat{V}$, respectively, and $U_2 = U_2(0)$ remains constant. Having established the convergence of training parameters, together with the decomposition in (2.40), we know that the end-to-end model $U(t)V^\top(t)$ converges to

$$\hat{U}\hat{V}^\top = \Phi_1\hat{U}_1\hat{V}^\top + \Phi_2 U_2(0)\hat{V}^\top = \hat{\Theta} + \Phi_2 U_2(0)\hat{V}^\top\,, \tag{2.41}$$

where the second equality is from

$$\Phi_1\hat{U}_1\hat{V}^\top = \Phi_1\Sigma_x^{-1/2}W^\top Y = X^\top(XX^\top)^\dagger Y = \hat{\Theta}\,. \tag{2.42}$$

Constrained training via orthogonal initialization: Based on our analysis above, initializing $U_2(0)$ such that $U_2(0)\hat{V}^\top = 0$ in the limit, guarantees convergence to the min-norm solution via (2.41). However, this is not easily achievable, as one needs to know a priori $\hat{V}$. Instead, we can show that by choosing a proper initialization, one can constrain the trajectory of the matrix $U(t)V^\top(t)$ to lie identically in the set $\Phi_2^\top U(t)V^\top(t) \equiv 0$ for all $t \geq 0$, thus the min-norm solution is obtained upon convergence, as suggested by the following proposition.

Proposition 2.3. *Let $V(t)$, $U_1(t)$ and $U_2(t)$, $t > 0$, be the solution of* (2.11) *starting from some $V(0)$, $U_1(0)$ and $U_2(0)$. We assume $V(t)$ and $U_1(t)$, $t > 0$, converge to some $\hat{V}$ and*

$\hat{U}_1$ *with* $E(\hat{V}, \hat{U}_1) = 0$. *If the initialization satisfies*

$$V(0)U_2^\top(0) = 0,\ U_1(0)U_2^\top(0) = 0\,, \tag{2.43}$$

then we have

$$\hat{U}\hat{V}^\top = \hat{\Theta}\,. \tag{2.44}$$

Proof. From (2.11) we have

$$\frac{d}{dt}\begin{bmatrix} VU_2^\top \\ U_1U_2^\top \end{bmatrix} = \begin{bmatrix} 0 & E^\top\Sigma_x^{1/2} \\ \Sigma_x^{1/2}E & 0 \end{bmatrix}\begin{bmatrix} VU_2^\top \\ U_1U_2^\top \end{bmatrix}\,. \tag{2.45}$$

Since $VU_2^\top = 0,\ U_1U_2^\top = 0$ is an equilibrium point of (2.45), we have $V(t)U_2^\top(0) = 0, \forall t \geq 0$ under the initialization in (2.43), hence $\hat{V}U_2^\top(0) = 0$. From (2.41) we conclude that $\hat{U}\hat{V}^\top = \hat{\Theta}$. □

In the standard linear regression problem we described at the beginning of this subsection, where Θ follows the gradient flow on $\mathcal{L}(\Theta) = \frac{1}{2}\|Y - X\Theta\|_F^2$, it is well-known that if the columns of $\Theta(0)$ are initialized in $\mathrm{span}(\Phi_1)$, namely $\Theta^\top(0)\Phi_2 = 0$, then $\Theta(t)$ converges to $\hat{\Theta}$. Proposition 2.3 is the extension of such results to the overparameterized setting. It is worth-noting that initializing the columns of $U(0)V^\top(0)$ in $\mathrm{span}(\Phi_1)$, namely $V(0)U_2^\top(0) = 0$ is no longer sufficient for obtaining $\hat{\Theta}$ as the trained network, and additional condition $U_1(0)U_2^\top(0) = 0$ is required. Moreover, we note that while the zero initialization $\Theta(0) = 0$ works for the standard linear regression case, the initialization $V(0) = 0$ and $U(0) = 0$ is bad in the overparametrized case because it is an equilibrium point of the gradient flow, even though it satisfies the orthogonal condition $V(0)U_2^\top(0) = 0$ and $U_1(0)U_2^\top(0) = 0$.

Here the orthogonality constraints (2.43) defines an invariant subset of the parameter space $\{V, U : VU^\top\Phi_2 = 0, \Phi_1^\top UU^\top\Phi_2 = 0\}$ under the gradient flow. Proposition 2.3 shows that given an initialization within the invariant set, the trained network (after convergence) is exactly the min-norm solution, the only minimizer in the invariant set. While in practice we can initialize the weights exactly as in (2.43), or one can directly initialize $U_2(0) = 0$, such choices are data-dependent and require the SVD of the data matrix X. Nonetheless, we show in the next section that under (properly scaled) random initialization and sufficiently large hidden layer width h, the orthogonal condition in (2.43) is approximately satisfied, which is one of the key ingredients for studying the implicit bias of wide two-layer networks.

2.1.4 Wide two-layer linear networks

In the previous two sections, we provided deterministic conditions for convergence and minimum-norm implicit bias of two-layer linear networks. Specifically, we showed that (1) the loss converges exponentially to its optimal value if the initialization, $(U(0), V(0))$, satisfies $c(0) > 0$ (Theorem 2.1), and that the end-to-end model $UV^\top$ converges to the min-norm solution $\hat{\Theta}$ if the initialization satisfies the orthogonality condition (Proposition 2.3). Finding initialization that satisfies these conditions seems non-trivial. For example, one could achieve exponential convergence by having either sufficient imbalance $\underline{\Delta} > 0$ or sufficient margin $\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^\top\|_F > 0$, but both the spectral gap and the margin are defined based on the SVD of X. Similarly, the orthogonality conditions in (2.43) are stated for $U_1(0), U_2(0), V(0)$, which also depend on the SVD of X. On the other hand, practical training algorithms often succeed with random initialization, which is data-agnostic. Thus, we are still left with the mystery of why random initialization leads to exponential convergence of the loss and minimum-norm implicit bias on the end-to-end model.

In this section, we show that if the hidden layer network width h is sufficiently large and the network weights are initialized as i.i.d. zero-mean Gaussians, i.e.,

$$[U(0)]_{ij} \sim \mathcal{N}\left(0, \frac{1}{h^{2\alpha}}\right), \; 1 \leq i \leq n, 1 \leq j \leq h\,, \tag{2.46}$$

$$[V(0)]_{ij} \sim \mathcal{N}\left(0, \frac{1}{h^{2\alpha}}\right), \; 1 \leq i \leq m, 1 \leq j \leq h\,, \tag{2.47}$$

then both conditions on initialization for convergence (in Theorem 2.1) and for implicit bias towards the min-norm solution (in Proposition 2.3) are satisfied with high probability.

Concentration results at initialization: Recall form the last section, one can obtain exactly min-norm solution via proper initialization of the two-layer network. In particular, it requires 1) convergence of the error E to zero; and 2) the orthogonality conditions $V(0)U_2^\top(0) = 0$ and $U_1(0)U_2^\top(0) = 0$. Under random initialization and sufficiently large hidden layer width h, these two conditions are approximately satisfied. More specifically, the following lemma can be shown using basic random matrix theory.

Lemma 2.1. *Let $\frac{1}{4} < \alpha \leq \frac{1}{2}$. Given data matrix X whose SVD defines $\underline{\Delta}, U_1(0), U_2(0)$, $\forall \delta \in (0,1)$, $\forall h > h_0 = 16\left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, with probability at least $1-\delta$ over random initialization with $[U(0)]_{ij}, [V(0)]_{ij} \sim \mathcal{N}(0, h^{-2\alpha})$, the following conditions hold:*

1. *(Sufficient spectral gap $\underline{\Delta}$, defined in* (2.32)*)*

$$\underline{\Delta} > h^{1-2\alpha}\,, \tag{2.48}$$

2. *(Approximate orthogonality)*

$$\left\| \begin{bmatrix} V(0)U_2^\top(0) \\ U_1(0)U_2^\top(0) \end{bmatrix} \right\|_F \leq 2\sqrt{m+r}\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,, \tag{2.49}$$

$$\left\| U_1(0)V^\top(0) \right\|_F \leq 2\sqrt{m}\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,. \tag{2.50}$$

Maintain approximate orthogonality: From (2.49), we know that the parameters are initialized close to the invariant set of our interest, with the proximity measured by $P = \|VU_2^\top\|_F + \|U_1U_2^\top\|_F$. As long as the gradient flow dynamics keeps P small throughout the training trajectory, one can show that the end-to-end model converges to some $\hat{U}\hat{V}^\top$ that is close to the min-norm solution $\hat{\Theta}$, i.e.,

$$\|\hat{U}\hat{V}^\top - \hat{\Theta}\|_2 = \|U_2\hat{V}^\top\|_2 \leq \sup_{t\geq 0} P(t)\,. \tag{2.51}$$

However, As the training proceeds, the parameters may drift too much away from the invariant set so that P grows large, leaving us no guarantee of proximity to the min-norm solution upon convergence. Fortunately, The dynamics (2.45) quantify at time t how fast P can maximally increase given that its current value is non-zero:

$$\dot{P} \leq \lambda_1(\Sigma_x^{1/2})\|E\|P\,, \tag{2.52}$$

It is clear that the smaller norm the current error E has, the lower the rate at which this measure could increase, and the Grönwall bound gives

$$P(t) \leq \exp\left(\lambda_1(\Sigma_x^{1/2})\int_0^t \|E(\tau)\|d\tau\right) P(0)\,. \tag{2.53}$$

This suggests that as long as $\int_0^t \|E(\tau)\|d\tau$ is upper bounded by some constant, $\|VU_2^\top\|_F + \|U_1U_2^\top\|_F$ will not increase too much from its initial value, thus the approximate orthogonality is maintained throughout the trajectory. A constant upper bound on $\int_0^t \|E(\tau)\|d\tau$ is derived from the constant rate of exponential convergence of the error (given by (2.48)), and an initial error $E(0)$ that is bounded by some constant (derived from (2.50)).

Implicit bias of wide two-layer linear network: Knowing that, with high probability, the network weights converge to a global optimum of the loss (given by (2.48)) and they remain close to the invariant set of our interest(given by (2.53)), we expect the trained network to represent an end-to-end model that is close to the min-norm solution, as formalized in the following Theorem regarding the implicit bias of wide linear networks (We let $X^\dagger$ be the pseudoinverse of X).

Theorem 2.2. *Let $\frac{1}{4} < \alpha \leq \frac{1}{2}$. Let $V(t), U(t), t > 0$ be the trajectory of the continuous dynamics (2.11) starting from some $V(0), U(0)$. Then, $\forall \delta \in (0,1)$, and $\forall h > h_0^{1/(4\alpha-1)}$ with $h_0 = \max\left\{16, 4\frac{\lambda_1^2(\Sigma_x)}{\lambda_r^2(\Sigma_x)}\right\} m\left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, with probability $1-\delta$ over random initializations with $[U(0)]_{ij}, [V(0)]_{ij} \sim \mathcal{N}(0, h^{-2\alpha})$, we have the end-to-end model $UV^\top$ to converge to some $\hat{U}\hat{V}^\top$ with*

$$\|\hat{U}\hat{V}^\top - \hat{\Theta}\|_2 \leq 2C^{1/h^{1-2\alpha}}\sqrt{m+r}\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,. \tag{2.54}$$

Here $C = \exp\left(1 + \frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|X^\dagger Y\|_F\right)$, which depends on the data X, Y.

Previous works [20] show non-asymptotic results on bounding the difference of predictions between the trained network and the kernel predictor of the NTK over a finite number of testing point (non-global result) using more general network structure and activation functions. As we work with a simpler model, we are able to study it without going through non-asymptotic NTK analysis, which is considerably more complicated. We believe this theorem is a clear illustration of how overparametrization, in particular, in the hidden layer width, together with random initialization affects the convergence and implicit bias.

Notably, although our initialization is related to the NTK analysis [18, 20] and the kernel regime [19], we significantly simplify the non-asymptotic analysis with the exact charaterization of an invariant set tied to the regularized solution. Specifically, our analysis does not rely on approximating the training flow to one in the infinite width limit, or one from the linearized network at initialization. Instead, we have the exact characterization of the properties required to reach min-norm solution and show how such properties are approximately preserved during training.

2.1.5 Numerical experiments

In this section, we first illustrate how the imbalance quantities $\Delta_+, \Delta_-, \underline{\Delta}$ are obtained from the spectrum of the imbalance matrix, as well as the role of width in shaping the imbalance quantities under random initialization. Then we run gradient descent (with small step size) on linear regression problem to validate our lower bounds for the convergence rate. We also conduct numerical verification of our Theorem 2.2 on implicit bias of wide linear networks.

Imbalance Quantities

For simplicity, we consider the matrix factorization problem $\mathcal{L} = \frac{1}{2}\|Y - \frac{1}{\sqrt{mh}}UV^\top\|_F^2$, $U \in \mathbb{R}^{r\times h}, V \in \mathbb{R}^{m\times h}$ under random initialization [23]. The scaling factor $\frac{1}{\sqrt{mh}}$ ensures that at initialization, the product $UV^\top$ keeps the same scale as we vary the hidden layer width h. Our convergence results Proposition 2.1 and Theorem 2.1 apply to this case and the imbalance quantities $\Delta_+, \Delta_-, \underline{\Delta}$ are defined from the imbalance matrix $D = U^\top U - V^\top V$ at initialization.

When $h \geq n + m$, then with probability 1 under random initialization, the imbalance matrix D has $\mathrm{rank}(D) = n + m$ and it has n positive eigenvalues and m negative ones. Our experiment sets $n = 20, m = 5$ and consider the case of $h = 30$ (small width) and $h = 1000$ (large width). For initialization, we use $[U(0)]_{ij}, [V(0)]_{ij} \sim \mathcal{N}(0, 1)$.

Since the end-to-end model in $\mathcal{L}$ is scaled by $\frac{1}{\sqrt{mh}}$, the instantaneous rate is scaled by $\frac{1}{mh}$, hence we consider the scaled imbalance quantities. We plot in Figure 2-3 all the non-zero eigenvalues of imbalance D and the imbalance quantities, scaled by $\frac{1}{mh}$. As illustrated by the plot, the imbalance quantities can be understood as the gaps between certain eigenvalues. It is clear that, compare to small width $h = 50$, large width $h = 1000$ leads to a larger spectral gap and smaller spectrum spread.

Moreover, as the width varies, the loss curve behaves differently:

(Small width): When $h = 30$, spectrum spreads Δ_-, Δ_+ are larger compared to the spectral gap $\underline{\Delta}$. As we discussed in Section 2.1.2 after Proposition 2.1, the lower bound on the rate is approximately $2\underline{\Delta}$, which is not a good global bound for the

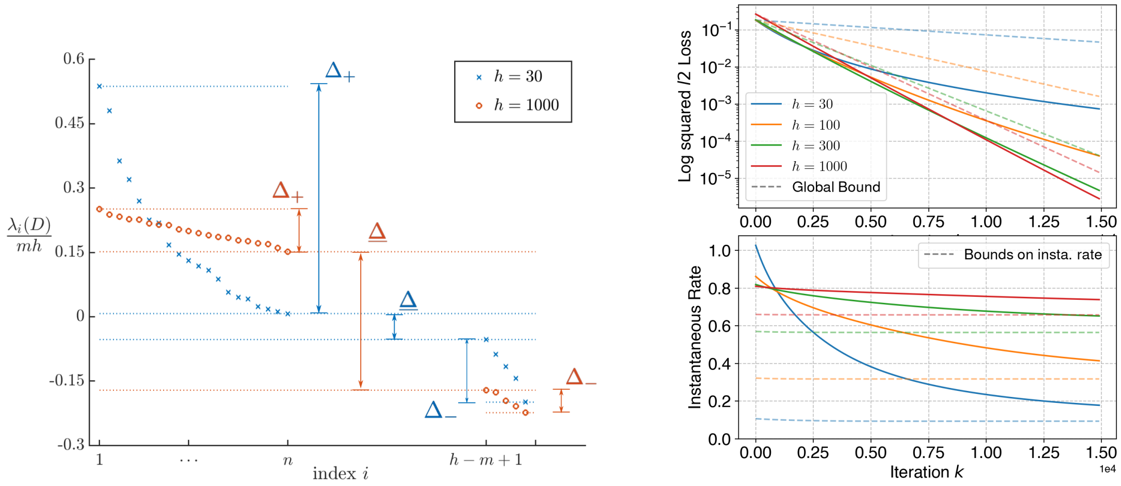

Figure 2-3. (Left): Scaled eigenvalues of the imbalance matrix D and the corresponding scaled imbalance quantities $\frac{1}{mh}\Delta_+$, $\frac{1}{mh}\Delta_-$, $\frac{1}{mh}\underline{\Delta}$ under random initialization, the scaling factor is omitted in the plot annotation for simplicity.
(Right): Gradient descent on $\mathcal{L} = \frac{1}{2}\left\|Y - \frac{1}{\sqrt{mh}}UV^\top\right\|_F^2$ for different network width. The dashed lines represent the bound provided by our results (Proposition 2.1 and Theorem 2.1).

convergence rate (see the top right plot in Figure 2-3). However, interestingly, the instantaneous rate (see the bottom right plot in Figure 2-3) starts off at large value and decreases as training proceeds. At the late stage of the training, our lower bound for the instantaneous rate is reasonably good.

(Large width): When $h = 1000$, the spectral gap $\underline{\Delta}$ is larger compared to spectrum spreads Δ_-, Δ_+. In this case $2\underline{\Delta}$ is a good global bound on the convergence rate (see the top right plot in Figure 2-3). As for the instantaneous rate, there is no significant variation in the rate and our bound Proposition 2.1 is reasonably good during training.

Convergence via imbalanced initialization

We train the linear network using gradient descent with a fixed small step size on the averaged loss $\mathcal{L}(U,V) = \|Y - XUV\|_F^2/n$. We use the initialization $U(0) = \sigma_U U_0, V(0) = \sigma_V V_0$ for some randomly sampled U_0, V_0 with i.i.d. standard normal entries, and scalars σ_U, σ_V. Under this setting, we can change the relative scales of σ_U, σ_V but keep their product fixed, so that we obtain initializations with different spectral gap $\underline{\Delta}$ while keeping the initial end-to-end matrix $U(0)V^\top(0)$ fixed. To eliminate the effect of ill-conditioned Σ_x on the convergence, we have $\Sigma_x = I_r$ in this experiment.

For comparison, we also consider the balanced initialization that corresponds to the same end-to-end matrix. For a given $\Theta(0) = U(0)V^\top(0)$, we choose an arbitrary $Q \in \mathbb{R}^{h\times m}$ with $Q^\top Q = I_m$, then a balanced initialization is given by

$$\begin{aligned} U_{\text{balanced}}(0) = \; & \Theta(0)\left[\Theta^\top(0)\Phi_1\Phi_1^\top\Theta(0)\right]^{-1/4}Q^\top, \\ V_{\text{balanced}}(0) = \; & \left[\Theta^\top(0)\Phi_1\Phi_1^\top\Theta(0)\right]^{1/4}Q\,. \end{aligned}$$

Such initialization ensures the imbalanced is the zero matrix while keeping the end-to-end matrix as $\Theta(0)$. We note here the choice of Q does not affect the error trajectory $E(t)$, hence the loss $\mathcal{L}(t)$.

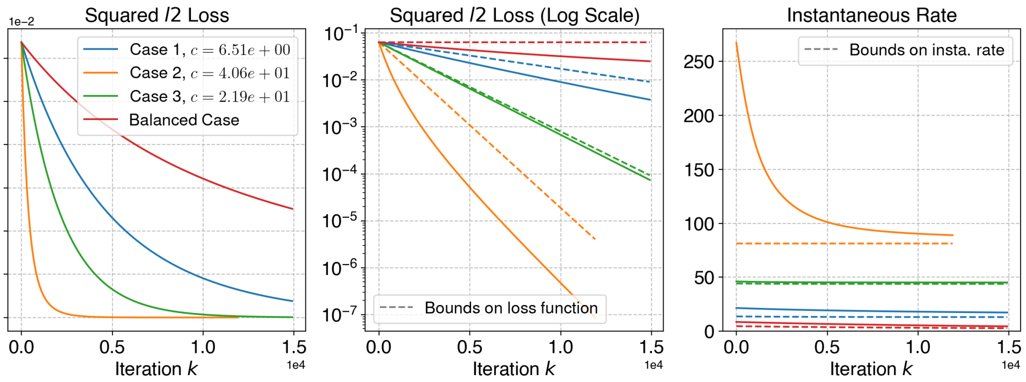

Figure 2-4. Convergence of gradient descent on linear networks with different initial imbalance matrices. We plot the loss function $\mathcal{L}$ on the left (regular scale) and the middle(log scale) figure. The instantaneous rate $-\dot{\mathcal{L}}/\mathcal{L}$ is shown on the right figure. The dashed line on the middle plot shows the bound on loss function by Theorem 2.1. Lastly, the dashed line on the right plot shows the lower bound by Proposition 2.1.

From Figure 2-4, we see that given fixed step size, the convergence rate is improved as we increase the level of the imbalance at initialization and the balanced initialization is the slowest among all cases. Notably, our lower bound on instantaneous rate is reasonably good for all cases except for case 2 at early training stage.

Moreover, the randomly initialized end-to-end function $\sigma_U \sigma_V U_0 V_0^\top$ has zero margin, as there is no bound provided for the balance case (Middle plot in Figure 2-4). Therefore, the margin-based convergence analysis [27] relies on carefully chosen initial end-to-end function and fail on the case of random initialization. On the contrary, random initialization almost surely yields a non-zero imbalance matrix, and our bound accounts for the effect of imbalance in convergence, resulting a much tighter bound on the rate.

Note that the goal of this experiment is to verify the improved convergence rate achieved by gradient flow initialized with a high spectral gap. To this end, we approximate the continuous dynamics using gradient descent with a fixed small step size. However, this does not imply that one can always accelerate gradient descent by increasing the spectral gap at initialization. This is because the step size for gradient descent is sometimes chosen to be close to the largest possible for convergence, but it is unknown how the spectral gap affects such choice. Analyzing the effect of large step size on convergence is subject of current research.

Implicit bias of wide linear networks

For the case of wide linear networks with random initialization considered in Section 2.1.4, when we set $\alpha = 1/2$, Theorem 2.2 suggests that $\|U(\infty)V^\top(\infty) - \hat{\Theta}\|_F \sim \mathcal{O}(h^{-1/2})$ We verify it by training linear networks with varying hidden layer width. We randomly initialize the network as in Section 2.1.4 and train it using gradient descent with a fixed small step size. The algorithm stops when the loss is below some fixed tolerance. We only vary the width h (from 500 to 10000) for different experiments and repeat 5 runs for each h.

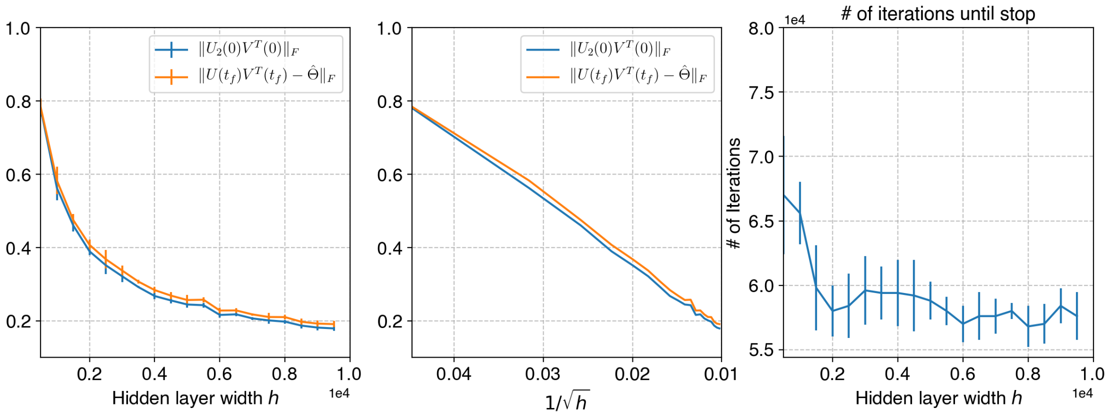

Figure 2-5. Implicit bias of wide two-layer linear network under random initialization. The line is plotting the average over 5 runs for each h, and the error bar shows the standard deviation. The gradient descent stops at iteration t_f.

Figure 2-5 clearly shows that the distance between the trained network and the min-norm solution, $\|U(t_f)V^\top(t_f) - \hat{\Theta}\|_F$, decreases as the width h increases and the middle plot verifies the asymptotic rate $\mathcal{O}(h^{-1/2})$. Besides, we also plot the initial distance in $\operatorname{span}(\Phi_2)$ between the network and the min-norm solution as

$$\|U_2(0)V^\top(0)\|_F = \|\Phi_2\Phi_2^\top(U(0)V^\top(0) - \hat{\Theta})\|_F\,.$$

A small $\|U_2V^\top\|_F$ is the exact property we want for a solution to be close to the min-norm solution. We see that the large width together with random initialization guarantees $\|U_2(0)V(0)\|_F \sim \mathcal{O}(h^{-1/2})$, and more importantly, since the initialization does not exactly fall into the invariant set defined by (2.43), $\|U_2V\|_F$ will deviate from its initial value. However, the deviation is well-controlled by the fast convergence of the error, i.e. as shown in the plot, $\|U_2(t_f)V^\top(t_f)\|_F \simeq \|U(t_f)V^\top(t_f) - \hat{\Theta}\|_F \sim \mathcal{O}(h^{-1/2})$.

Proofs of Proposition 2.1 and Theorem 2.1

The proof of our main results (Proposition 2.1 and Theorem 2.1) follows exactly the same procedure used for the scalar dynamics in our warm-up example, which is described in Section 2.1.2. First of all, we lower bound the instantaneous rate with singular values of U_1 and V as stated in the next lemma.

Lemma 2.2. *Consider the continuous dynamics in* (2.11)*. Let* $\tilde{\mathcal{L}} := \mathcal{L} - \mathcal{L}^*$*. Then,*

$$-\frac{d}{dt}\frac{\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}} \geq 2\lambda_r(\Sigma_x)(\lambda_r(U_1U_1^\top) + \lambda_m(VV^\top))\,. \tag{2.55}$$

Next, we derive lower bounds on $\lambda_r(U_1U_1^\top)$ and of $\lambda_m(VV^\top)$ by exploiting the fact that they satisfy a set of quadratic inequalities. The lower bounds are stated in the next lemma.

Lemma 2.3. *Suppose* $h \geq \min\{r, m\}$*. If* $A \in \mathbb{R}^{r\times h}$ *and* $B \in \mathbb{R}^{h\times m}$ *satisfy* $A^\top A - BB^\top = D$ *for some* $D \in \mathbb{R}^{h\times h}$*, then*

$$\lambda_m(B^\top B) \geq \frac{-\bar{\lambda} + \underline{\lambda} + \sqrt{(\bar{\lambda} + \underline{\lambda})^2 + 4\sigma_m^2(AB)}}{2}\,, \tag{2.56}$$

where $\bar{\lambda} = \max\{\lambda_1(D), 0\}$ *and* $\underline{\lambda} = \max\{\lambda_m(-D), 0\}$*.*

We defer the proofs of these lemmas to the end of this section. Combining Lemmas 2.2 and 2.3, we have the desired bound on the instantaneous rate:

Proof of Proposition 2.1. From Lemma 2.3, let $A = U_1, B = V^\top$, we have $A^\top A - BB^\top = D$, thus

$$\begin{aligned}\lambda_m(VV^\top) &\geq \frac{-\bar{\lambda}_+ + \underline{\lambda}_- + \sqrt{(\bar{\lambda}_+ + \underline{\lambda}_-)^2 + 4\sigma_m^2(U_1V^\top)}}{2}\,,\\ \bar{\lambda}_+ &= \max\{\lambda_1(D), 0\},\ \underline{\lambda}_- = \max\{\lambda_m(-D), 0\}\,,\end{aligned} \tag{2.57}$$

then let $A = V, B = U_1^\top$, we have $A^\top A - BB^\top = -D$, thus

$$\begin{aligned}\lambda_r(U_1U_1^\top) &\geq \frac{-\bar{\lambda}_- + \underline{\lambda}_+ + \sqrt{(\bar{\lambda}_- + \underline{\lambda}_+)^2 + 4\sigma_r^2(VU_1^\top)}}{2}\,.\\ \bar{\lambda}_- &= \max\{\lambda_1(-D), 0\},\ \underline{\lambda}_+ = \max\{\lambda_r(D), 0\}\end{aligned} \tag{2.58}$$

Now rewrite the lowerbounds (2.57)(2.58) in terms of

$$\Delta_+ := \bar{\lambda}_+ - \underline{\lambda}_+,\ \Delta_- := \bar{\lambda}_- - \underline{\lambda}_-,\ \underline{\Delta} := \underline{\lambda}_+ + \underline{\lambda}_-\,, \tag{2.59}$$

we have

$$\begin{aligned}\lambda_m(VV^\top) &\geq \frac{-\bar{\Delta}_+ + \underline{\lambda}_- - \underline{\lambda}_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_m^2(U_1V^\top)}}{2}\,,\\ \lambda_r(U_1U_1^\top) &\geq \frac{-\bar{\Delta}_- + \underline{\lambda}_+ - \underline{\lambda}_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_r^2(VU_1^\top)}}{2}\,.\end{aligned} \tag{2.60}$$

Then (2.29) follows immediately from Lemma 2.2. □

Similar to the warm-up example in Section 2.1.2, one can control the imbalance quantities and the singular value $\sigma_m^2(U_1V^\top)$ in the bound from Proposition 2.1 throughout the entire training trajectory: Δ_+, Δ_- and $\underline{\Delta}$ are time-invariant because the imbalance D is so, and the singular value $\sigma_m^2(U_1V^\top)$ can be controlled via a positive margin $\sigma_{\min}(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F$. This proves Theorem 2.1.

Proof of Theorem 2.1. When $m = r$, we have $\sigma_m(U_1V^\top) = \sigma_r(U_1V^\top) = \sigma_{\min}(U_1V^\top)$. When $m > r$, we only need to lower bound $\sigma_m(U_1V^\top)$ since $\sigma_r(U_1V^\top) = 0$, and vice versa when $r > m$.

Therefore, without loss of generality, we assume $m \leq r$ and derive the lower bound on $\sigma_m(U_1V^\top)$. By $\|A\|_F \geq \|A\|_2$ and Weyl's inequality [64, 7.3.P16], one has

$$\|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F + \sigma_m(\Sigma_x^{1/2}U_1V^\top) \geq \|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_2 + \sigma_m(\Sigma_x^{1/2}U_1V^\top) \geq \sigma_m(\tilde{Y})\,, \tag{2.61}$$

from which one obtain the lower bound

$$\sigma_m(U_1V^\top) \geq \sigma_m(\Sigma_x^{1/2}U_1V^\top)/\lambda_1^{1/2}(\Sigma_x) \geq (\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F)/\lambda_1^{1/2}(\Sigma_x)\,. \tag{2.62}$$

The lower bound is trivial when $\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F < 0$, thus we could write

$$\sigma_m(U_1V^\top) \geq \max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F, 0\}/\lambda_1^{1/2}(\Sigma_x)\,. \tag{2.63}$$

Now because $\|\tilde{Y} - \Sigma_x^{1/2}U_1V^\top\|_F = \sqrt{2\tilde{\mathcal{L}}}$ is non-decreasing under gradient flow, we have $\forall t \geq 0$,

$$\begin{aligned}\sigma_m^2(U_1(t)V^\top(t)) \geq&\ (\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1(t)V^\top(t)\|_F, 0\})^2/\lambda_1(\Sigma_x) \\ \geq&\ (\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2}U_1(0)V^\top(0)\|_F, 0\})^2/\lambda_1(\Sigma_x)\,.\end{aligned} \tag{2.64}$$

Finally using (2.64) to further lower bound (2.29) in Proposition 2.1, we have our desired lower bound on the instantaneous rate

$$-\frac{d}{dt}\frac{\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}} \geq \lambda_r(\Sigma_x)c(0)\,. \tag{2.65}$$

The result $\tilde{\mathcal{L}}(t) \leq \exp(-\lambda_r(\Sigma_x)c(0)t)\tilde{\mathcal{L}}(0)$ follows from Grönwall's inequality [65].

□

Proofs of Auxiliary Lemmas

Proof of Lemma 2.2. Under (2.11), the time derivative of error is given by

$$\dot{E} = -\Sigma_x^{1/2}U_1U_1^\top\Sigma_x^{1/2}E - \Sigma_x EVV^\top\,.$$

Consider the time derivative of $\|E\|_F^2$,

$$\frac{d}{dt}\|E\|_F^2 = \frac{d}{dt}\mathrm{tr}(E^\top E) = -2\mathrm{tr}\left(E^\top\Sigma_x^{1/2}U_1U_1^\top\Sigma_x^{1/2}E + E^\top\Sigma_x EVV^\top\right)\,. \tag{2.66}$$

Use the trace inequality [66, Lemma 1] to get the lower bound the trace of two matrices respectively as

$$\begin{aligned}\mathrm{tr}\left(E^\top\Sigma_x^{1/2}U_1U_1^\top\Sigma_x^{1/2}E\right) &= \mathrm{tr}\left(\Sigma_x^{1/2}EE^\top\Sigma_x^{1/2}U_1U_1^\top\right) \\ &\geq \lambda_r(U_1U_1^\top)\mathrm{tr}\left(\Sigma_x^{1/2}EE^\top\Sigma_x^{1/2}\right) \\ &= \lambda_r(U_1U_1^\top)\mathrm{tr}\left(\Sigma_x EE^\top\right) \\ &\geq \lambda_r(U_1U_1^\top)\lambda_r(\Sigma_x)\mathrm{tr}(EE^\top) \\ &= \lambda_r(U_1U_1^\top)\lambda_r(\Sigma_x)\|E\|_F^2\,,\end{aligned} \tag{2.67}$$

and

$$\begin{aligned}\operatorname{tr}\left(E^\top \Sigma_x E V V^\top\right) &\geq \lambda_m(VV^\top)\operatorname{tr}\left(E^\top \Sigma_x E\right)\\ &= \lambda_m(VV^\top)\operatorname{tr}\left(\Sigma_x E E^\top\right)\\ &\geq \lambda_m(VV^\top)\lambda_r(\Sigma_x)\operatorname{tr}(EE^\top)\\ &= \lambda_m(VV^\top)\lambda_r(\Sigma_x)\|E\|_F^2\,. \end{aligned} \tag{2.68}$$

Combine (2.66) with (2.67)(2.68), we have

$$\frac{d}{dt}\|E\|_F^2 \leq -2\lambda_r(\Sigma_x)\left(\lambda_r(U_1U_1^\top)+\lambda_m(VV^\top)\right)\|E\|_F^2 \tag{2.69}$$

Notice that $\frac{1}{2}\|E\|_F^2$ is exactly $\tilde{\mathcal{L}} = \mathcal{L} - \mathcal{L}^*$. It follows from (2.69) that

$$-\frac{d}{dt}\frac{\tilde{\mathcal{L}}}{\tilde{\mathcal{L}}} \geq 2\lambda_r(\Sigma_x)\left(\lambda_r(U_1U_1^\top)+\lambda_m(VV^\top)\right)\,.$$

□

Proof of Lemma 2.3. From the imbalance equation $A^\top A - BB^\top = D$, we have

$$(B^\top B)^2 = B^\top(BB^\top)B = B^\top(A^\top A - D)B = B^\top A^\top AB - B^\top DB\,.$$

Let $z_m \in \mathbb{S}^{m-1}$ be the eigenvector of $(B^\top B)^2$ (or $B^\top B$) associated with eigenvalue $\lambda_m^2(B^\top B)$ (or $\lambda_m(B^\top B)$). The one have

$$\begin{aligned}\lambda_m^2(B^\top B) = z_m^\top(B^\top B)^2 z_m = &\ z_m^\top B^\top A^\top AB z_m - z_m^\top B^\top DB z_m\\ \geq &\ \lambda_m(B^\top A^\top AB) - z_m^\top B^\top DB z_m\,,\\ = &\ \sigma_m^2(AB) - z_m^\top B^\top DB z_m \end{aligned} \tag{2.70}$$

and the rest of proof is to find a lower bound for $-z_m^\top B^\top DB z_m$.

First of all, we know that D has at most m negative eigenvalues: If D has more than m negative eigenvalues, then the subspace spanned by the all negative eigenvectors has dimension at least $m+1$, which must have non-trivial intersection with $\ker(B^\top)$, then there exists a nonzero vector $z \in \ker(B^\top)$ such that $z^\top Dz < 0$, which would imply $z^\top A^\top Az = z^\top Dz < 0$, a contradiction.

When D has less than m negative eigenvalues, then $\underline{\lambda} = 0$ and we simply lower bound $-z_m^\top B^\top DBz_m$ as

$$\begin{aligned}
\lambda_m^2(B^\top B) \geq &\ \sigma_m^2(AB) - z_m^\top B^\top DBz_m \\
\geq &\ \sigma_m^2(AB) - \bar{\lambda} z_m^\top B^\top Bz_m \\
= &\ \sigma_m^2(AB) - \bar{\lambda}\lambda_m(B^\top B)\,.
\end{aligned}$$

This quadratic inequality w.r.t. $\lambda_m(B^\top B)$ has nonnegative solutions

$$\lambda_m(B^\top B) \geq \frac{-\bar{\lambda} + \sqrt{\bar{\lambda}^2 + 4\sigma_m^2(AB)}}{2}\,,$$

which is exactly (2.56) when $\underline{\lambda} = 0$.

When D has exactly m negative eigenvalues, the easy case is one with $h = m$, i.e. all eigenvalues of D are negative. We simply lower bound $-z_m^\top B^\top DBz_m$ as

$$\begin{aligned}
\lambda_m^2(B^\top B) \geq &\ \sigma_m^2(AB) - z_m^\top B^\top DBz_m \\
\geq &\ \sigma_m^2(AB) - (-\underline{\lambda} z_m^\top B^\top Bz_m) \\
= &\ \sigma_m^2(AB) + \underline{\lambda}\lambda_m(B^\top B)\,.
\end{aligned}$$

This quadratic inequality w.r.t. $\lambda_m(B^\top B)$ has nonnegative solutions

$$\lambda_m(B^\top B) \geq \frac{\underline{\lambda} + \sqrt{\underline{\lambda}^2 + 4\sigma_m^2(AB)}}{2}\,,$$

which is exactly (2.56) when $\bar{\lambda} = 0$.

Now we only left to prove the bound for the case $h > m$. We first consider any orthogonal matrix $Q \in \mathcal{O}(h)$, we have $Q^\top A^\top AQ - Q^\top BB^\top Q = Q^\top DQ$, $AQQ^\top B = AB$, and $\lambda_m(B^\top Q^\top QB) = \lambda_m(B^\top B)$. Then it suffices to study the orthogonally transformed matrices $\tilde{A} = AQ, \tilde{B} = Q^\top B$, with $\tilde{A}^\top \tilde{A} - \tilde{B}\tilde{B}^\top = Q^\top DQ, \tilde{A}\tilde{B} = AB$

and find a lower bound on $\lambda_m(\tilde{B}^\top \tilde{B})$. We can pick Q that diagonalize D, thus *with out loss of generality, we assume D is diagonal and the eigenvalues are in decreasing order*.

Since $h > m$, we write the diagonal D as a block matrix $D = \begin{bmatrix} \Lambda_+ & 0 \\ 0 & -\Lambda_- \end{bmatrix}$, where

$$\begin{aligned}
&\Lambda_+ = \operatorname{diag}\{\lambda_1(D), \cdots, \lambda_{h-m}(D)\}\\
&\Lambda_- = \operatorname{diag}\{-\lambda_{h-m+1}(D), \cdots, -\lambda_h(D)\} = \operatorname{diag}\{\lambda_m(-D), \cdots, \lambda_1(-D)\}\,.
\end{aligned}$$

Here, notice that Λ_+ is positive semi-definite and Λ_- positive definite with

$$\Lambda_+ \preceq \bar{\lambda} I_{h-m},\ \Lambda_- \succeq \underline{\lambda} I_m\,. \tag{2.71}$$

Now we write A, B as block matrices as well

$$\begin{aligned}
&A = \begin{bmatrix} A_+ & A_- \end{bmatrix},\ B = \begin{bmatrix} B_+ \\ B_- \end{bmatrix},\\
&A_+ \in \mathbb{R}^{r\times(h-m)}, A_- \in \mathbb{R}^{r\times m}, B_+ \in \mathbb{R}^{(h-m)\times m}, B_- \in \mathbb{R}^{m\times m}\,,
\end{aligned}$$

from which we can rewrite equations $A^\top A - BB^\top = D$ as

$$\begin{bmatrix} A_+^\top \\ A_-^\top \end{bmatrix} \begin{bmatrix} A_+ & A_- \end{bmatrix} - \begin{bmatrix} B_+ \\ B_- \end{bmatrix} \begin{bmatrix} B_+^\top & B_-^\top \end{bmatrix} = \begin{bmatrix} \Lambda_+ & 0 \\ 0 & -\Lambda_- \end{bmatrix}\,.$$

By inspection, the equality for each block gives us

$$A_+^\top A_+ = B_+ B_+^\top + \Lambda_+\,, \tag{2.72}$$

$$A_-^\top A_- = B_- B_-^\top - \Lambda_-\,, \tag{2.73}$$

$$A_+^\top A_- = B_+ B_-^\top\,. \tag{2.74}$$

With these equalities, we know the following matrix is p.s.d., for any $\hat{\lambda} > \bar{\lambda} \geq 0$,

$$\begin{aligned}
\begin{bmatrix} B_+B_+^\top + \hat{\lambda} I_{h-m} & B_+B_-^\top \\ B_-B_+^\top & B_-B_-^\top - \underline{\lambda} I_m \end{bmatrix} &\overset{(2.71)}{\succeq} \begin{bmatrix} B_+B_+^\top + \Lambda_+ & B_+B_-^\top \\ B_-B_+^\top & B_-B_-^\top - \Lambda_- \end{bmatrix}\\
&= \begin{bmatrix} A_+^\top \\ A_-^\top \end{bmatrix} \begin{bmatrix} A_+ & A_- \end{bmatrix} \succeq 0\,.
\end{aligned} \tag{2.75}$$

Since $B_+B_+^\top + \hat{\lambda} I_{h-m} \succ 0$, positive semi-definiteness (2.75) is equivalent to

$$B_-B_-^\top - \underline{\lambda} I_m - B_-B_+^\top (B_+B_+^\top + \hat{\lambda} I_{h-m})^{-1} B_+B_-^\top \succeq 0\,. \tag{2.76}$$

Now we use Woodbury's Identity [64, 0.7.4], which says for matrices M, N, P with appropriate dimensions, we have

$$(M + P^\top N P)^{-1} = M^{-1} - M^{-1} P^\top (P M^{-1} P^\top + N^{-1})^{-1} P M^{-1} \,,$$

if all inverses exist. Let $M = I_m, N = \hat{\lambda}^{-1} I_{h-m}, P = B_+$, we have

$$(I_m + \hat{\lambda}^{-1} B_+^\top B_+)^{-1} = I_m - B_+^\top (\hat{\lambda} I_{h-m} + B_+ B_+^\top)^{-1} B_+ \,,$$

which leads to

$$B_-(I_m + \hat{\lambda}^{-1} B_+^\top B_+)^{-1} B_-^\top = B_- B_-^\top - B_- B_+^\top (\hat{\lambda} I_{h-m} + B_+ B_+^\top)^{-1} B_+ B_-^\top \,. \qquad (2.77)$$

Using (2.77), we can rewrite (2.76) as

$$\underline{\lambda} I_m - B_-(I_m + \hat{\lambda}^{-1} B_+^\top B_+)^{-1} B_-^\top \preceq 0 \,. \qquad (2.78)$$

Consider the following matrix congruence

$$\begin{aligned} & \begin{bmatrix} \underline{\lambda} I_m & B_- \\ B_-^\top & I_m + \hat{\lambda}^{-1} B_+^\top B_+ \end{bmatrix} \\ &= S_1 \begin{bmatrix} \underline{\lambda} I_m - B_-(I_m + \hat{\lambda}^{-1} B_+^\top B_+)^{-1} B_-^\top & 0 \\ 0 & I_m + \hat{\lambda}^{-1} B_+^\top B_+ \end{bmatrix} S_1^\top \qquad (2.79) \\ &= S_2 \begin{bmatrix} \underline{\lambda} I_m & 0 \\ 0 & I_m + \hat{\lambda}^{-1} B_+^\top B_+ - \underline{\lambda}^{-1} B_-^\top B_- \end{bmatrix} S_2^\top \qquad (2.80) \end{aligned}$$

where

$$S_1 = \begin{bmatrix} I_m & B_-(I_m + \hat{\lambda}^{-1} B_+^\top B_+)^{-1} \\ 0 & I_m \end{bmatrix} , \qquad S_2 = \begin{bmatrix} I_m & 0 \\ \underline{\lambda}^{-1} B_-^\top & I_m \end{bmatrix} ,$$

and S_1, S_2 are non-singular. By Sylvester's Intertia Theorem [64, Theorem 4.5.8], the block diagonal matrix shown in (2.79) has exactly the same number of positive eigenvalues as the one shown in (2.80), and the number of positive eigenvalues is m, according to (2.78). Then for the block diagonal matrix in (2.80), we must have

$$I_m + \hat{\lambda}^{-1} B_+^\top B_+ - \underline{\lambda}^{-1} B_-^\top B_- \preceq 0 \,,$$

hence

$$
\begin{aligned}
0 &\preceq -I_m - \hat{\lambda}^{-1} B_+^\top B_+ + \underline{\lambda}^{-1} B_-^\top B_- \\
0 &\preceq -\hat{\lambda}\underline{\lambda} I_m - \underline{\lambda} B_+^\top B_+ + \hat{\lambda} B_-^\top B_- \\
\hat{\lambda} B_+^\top B_+ - \underline{\lambda} B_-^\top B_- &\preceq -\hat{\lambda}\underline{\lambda} I_m - \underline{\lambda} B_+^\top B_+ + \hat{\lambda} B_-^\top B_- \\
&\qquad + \hat{\lambda} B_+^\top B_+ - \underline{\lambda} B_-^\top B_- \\
\hat{\lambda} B_+^\top B_+ - \underline{\lambda} B_-^\top B_- &\preceq -\hat{\lambda}\underline{\lambda} I_m + (\hat{\lambda} - \underline{\lambda})(B_+^\top B_+ + B_-^\top B_-) \\
\hat{\lambda} B_+^\top B_+ - \underline{\lambda} B_-^\top B_- &\preceq -\hat{\lambda}\underline{\lambda} I_m + (\hat{\lambda} - \underline{\lambda}) B^\top B \,, \qquad (2.81)
\end{aligned}
$$

where the last equivalence uses the fact $B^\top B = B_+^\top B_+ + B_-^\top B_-$. This suggests that

$$
\begin{aligned}
B^\top D B = B_+^\top \Lambda_+ B_+ - B_-^\top \Lambda_- B_- &\preceq \hat{\lambda} B_+^\top B_+ - \underline{\lambda} B_-^\top B_- \\
&\overset{(2.81)}{\preceq} -\hat{\lambda}\underline{\lambda} I_m + (\hat{\lambda} - \underline{\lambda}) B^\top B \qquad (2.82)
\end{aligned}
$$

Lastly, from (2.70) we have

$$
\begin{aligned}
\lambda_m^2(B^\top B) = z_m^\top (B^\top B)^2 z_m &\geq \sigma_m^2(AB) - z_m^\top B^\top D B z_m \\
&\overset{(2.82)}{\geq} \sigma_m^2(AB) + \hat{\lambda}\underline{\lambda} - (\hat{\lambda} - \underline{\lambda}) z_m^\top B^\top B z_m \\
&= \sigma_m^2(AB) + \hat{\lambda}\underline{\lambda} - (\hat{\lambda} - \underline{\lambda}) \lambda_m(B^\top B) \,.
\end{aligned}
$$

This quadratic inequality w.r.t. $\lambda_m(B^\top B)$ has nonnegative solutions

$$
\lambda_m(B^\top B) \geq \frac{\underline{\lambda} - \hat{\lambda} + \sqrt{(\underline{\lambda} - \hat{\lambda})^2 + 4\hat{\lambda}\underline{\lambda} + 4\sigma_m^2(AB)}}{2} = \frac{-\hat{\lambda} + \underline{\lambda} + \sqrt{(\hat{\lambda} + \underline{\lambda})^2 + 4\sigma_m^2(AB)}}{2} \,.
$$

Since we can choose any $\hat{\lambda} > \bar{\lambda} \geq 0$, we have

$$
\lambda_m(B^\top B) \geq \lim_{\hat{\lambda} \to \bar{\lambda}} \frac{-\hat{\lambda} + \underline{\lambda} + \sqrt{(\hat{\lambda} + \underline{\lambda})^2 + 4\sigma_m^2(AB)}}{2} = \frac{-\bar{\lambda} + \underline{\lambda} + \sqrt{(\bar{\lambda} + \underline{\lambda})^2 + 4\sigma_m^2(AB)}}{2} \,.
$$

This is exactly (2.56).

(Note that when $\bar{\lambda} > 0$, one can pick $\hat{\lambda} = \bar{\lambda}$ and obtain the desired bound directly. Taking the limit $\hat{\lambda} \to \bar{\lambda}$ is necessary only when $\bar{\lambda} = 0$). □

Proofs of Proposition 2.2, Lemma 2.1, and Theorem 2.2

Proof of Proposition 2.2. Since $c(0) > 0$, for the gradient system (2.11), the states (parameters) (U_1, V) converge either to an equilibrium point which minimizes the potential $\frac{1}{2}\|E\|_F^2 = \mathcal{L} - \mathcal{L}^*$ or have its l_2-norm grow to infinity [67].

Consider the following dynamics

$$\frac{d}{dt}\begin{bmatrix} V \\ U_1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & E^\top \Sigma_x^{1/2} \\ \Sigma_x^{1/2} E & 0 \end{bmatrix}}_{:=A_Z} \underbrace{\begin{bmatrix} V \\ U_1 \end{bmatrix}}_{:=Z}, \tag{2.83}$$

which can be viewed as a time-variant linear system. Notice that by [64, Theorem 7.3.3], we have $\|A_Z\|_2 = \|\Sigma_x^{1/2} E\|_2$.

From (2.83), we have

$$\begin{aligned} \frac{d}{dt}\|Z\|_F^2 &= 2\mathrm{tr}\left(Z^\top A_Z Z\right) \\ &= 2\mathrm{tr}\left(Z Z^\top A_Z\right) \\ &\leq 2\|A_Z\|_2 \mathrm{tr}\left(Z Z^\top\right) \\ &= 2\|\Sigma_x^{1/2} E\|_2 \|Z\|_F^2 \\ &\leq 2\lambda_1^{1/2}(\Sigma_x)\|E\|_2\|Z\|_F^2 \\ &\leq 2\lambda_1^{1/2}(\Sigma_x)\|E\|_F\|Z\|_F^2 . \end{aligned}$$

By Grönwall's inequality [65], we have

$$\|Z(t)\|_F^2 \leq \exp\left(\int_0^t 2\lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F^2 .$$

Finally, by Theorem 2.1, we have $\|E(t)\|_F \leq \exp\left(-\lambda_r(\Sigma_x)c(0)t/2\right)\|E(0)\|_F, \ \forall t > 0,$

since $\|E\|_F = \sqrt{2(\mathcal{L} - \mathcal{L}^*)}$, which leads to

$$\begin{aligned}
&\exp\left(\int_0^t 2\lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right) \\
\leq\ &\exp\left(2\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^t \exp\left(-\lambda_r(\Sigma_x)c(0)\tau/2\right)d\tau\right)\right) \\
\leq\ &\exp\left(2\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^\infty \exp\left(-\lambda_r(\Sigma_x)c(0)\tau/2\right)d\tau\right)\right) \\
=\ &\exp\left(\frac{4\lambda_1^{1/2}(\Sigma_x)}{c(0)\lambda_r(\Sigma_x)}\|E(0)\|_F\right).
\end{aligned}$$

Therefore we have

$$\|Z(t)\|_F^2 \leq \exp\left(\frac{4\lambda_1^{1/2}(\Sigma_x)}{c(0)\lambda_r(\Sigma_x)}\|E(0)\|_F\right)\|Z(0)\|_F^2\,,$$

which implies that the trajectory $V(t), U_1(t), t > 0$ is bounded, i.e. its l_2-norm can not grow to infinity, then it has to converge to some equilibrium point $(\hat{V}, \hat{U}_1)$ such that its potential is zero, i.e., $E(\hat{V}, \hat{U}_1) = 0$. □

Now we turn to prove Lemma 2.1 and Theorem 2.2. We need a basic result in random matrix theory

Lemma 2.4. *Given $m, n \in \mathbb{N}$ with $m \leq n$. Let A be an $n \times m$ random matrix with i.i.d. standard normal entries $A_{ij} \sim \mathcal{N}(0,1)$. For $\delta > 0$, with probability at least $1 - 2\exp(-\delta^2)$, we have*

$$\sqrt{n} - (\sqrt{m} + \delta) \leq \sigma_m(A) \leq \sigma_1(A) \leq \sqrt{n} + (\sqrt{m} + \delta)\,.$$

The proof can be found in [68, Theorem 2.13]. We also need the following inequality.

Lemma 2.5. *Let $A \in \mathbb{R}^{k\times n}, B \in \mathbb{R}^{n\times m}$. Suppose $n \leq m$, then*

$$\sigma_i(A)\sigma_n(B) \leq \sigma_i(AB)\,,$$

for $1 \leq i \leq \min\{k, n\}$.

Proof. We start with the case where $k = n$. When $\sigma_n(B^\top) = 0$, the result is trivial. When $\sigma_n(B^\top) \neq 0$, we have $BB^\dagger = I$, where $B^\dagger$ is the Moore–Penrose inverse of B. By Weyl's inequality [64, 7.3.P16], it follows that

$$\sigma_i(A) \leq \sigma_i(AB)\sigma_1(B^\dagger),\ \forall 1 \leq i \leq n\,.$$

Since $\sigma_1(B^\dagger) = \sigma_n^{-1}(B)$, we get the desired inequality.

When $k > n$, we have $\forall 1 \leq i \leq n$,

$$\sigma_i(A) = \sigma_i\left(\begin{bmatrix} A & 0_{k\times(k-n)} \end{bmatrix}\right) \leq \sigma_i\left(AB\right)\sigma_1(\begin{bmatrix} B^\dagger & 0_{m\times(k-n)} \end{bmatrix}) = \sigma_i(AB)\sigma_1(B^\dagger)\,,$$

which still leads to the desired result.

When $k < n$, consider replacing A with $\begin{bmatrix} A \\ 0_{(n-k)\times n} \end{bmatrix}$, we have $\forall 1 \leq i \leq k$,

$$\sigma_i(A)\sigma_n(B) = \sigma_i\left(\begin{bmatrix} A \\ 0_{(n-k)\times n} \end{bmatrix}\right)\sigma_n(B) \leq \sigma_i\left(\begin{bmatrix} AB \\ 0_{(n-k)\times m} \end{bmatrix}\right) = \sigma_i(AB)\,.$$

□

Now we are ready to prove Lemma 2.1.

Proof of Lemma 2.1. For readability we simply write $U(0), U_1(0), U_2(0), V(0), D(0)$ as U, U_1, U_2, V, D.

Consider the matrix $\begin{bmatrix} V^\top & U^\top \end{bmatrix}$ which is $h \times (m+n)$. Apply Lemma 2.4 to matrix $A = h^\alpha \begin{bmatrix} V^\top & U^\top \end{bmatrix}$, with probability at least $1-\delta$, we have

$$\sigma_{m+n}(h^\alpha \begin{bmatrix} V^\top & U^\top \end{bmatrix}) \geq \sqrt{h} - \left(\sqrt{m+n} + \delta\right)\,,$$

which leads to

$$\sigma_{m+n}(\begin{bmatrix} V^\top & U^\top \end{bmatrix}) \geq h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\,. \tag{2.84}$$

Regarding the first inequality, we write the imbalance as

$$U_1^\top U_1 - V^\top V = \begin{bmatrix} V^\top & U_1^\top \end{bmatrix}\begin{bmatrix} -V \\ U_1 \end{bmatrix} = \begin{bmatrix} V^\top & U^\top \end{bmatrix}\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix}\begin{bmatrix} V \\ U \end{bmatrix}\,.$$

For $h > \left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, assume event (2.84) happens, then

$$\sigma_{m+n}\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix}\right) \geq h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^\alpha} > 0\,,$$

hence we have

$$\begin{aligned}
\sigma_{r+m}(D) &= \sigma_{r+m}(U_1^\top U_1 - V^\top V) \\
&= \sigma_{r+m}\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix} \begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix}\right) \\
(\text{Lemma } 2.5) &\geq \sigma_{r+m}\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix} \begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix}\right) \sigma_{m+n}\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
&= \sigma_{r+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix}\right) \sigma_{m+n}\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
(\text{Lemma } 2.5) &\geq \sigma_{r+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix}\right) \sigma_{m+n}^2\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
&= \sigma_{r+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix}\right) \sigma_{m+n}^2\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix}\right) \\
&= \sigma_{m+n}^2\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix}\right)\,,
\end{aligned}$$

where the last equality is due to the fact that $\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^\top \end{bmatrix}$ has exactly $r+m$ non-zero singular value and all of them are 1.

We further assume $h > 16\left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, conditioned on event (2.84), with probability 1 we have

$$\begin{aligned}
\sigma_{r+m}(D) &\geq \sigma_{m+n}^2\left(\begin{bmatrix} V^\top & U^\top \end{bmatrix}\right) \\
&\geq \left(h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2 \\
&= h^{1-2\alpha} - 2\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}} + \left(\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2 \\
&> h^{1-2\alpha} - 2\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}} \geq \frac{1}{2}h^{1-2\alpha}\,. \qquad (2.85)
\end{aligned}$$

Lastly, due to the minimax property of symmetric matrix [64, Theorem 4.2.6], we have

$$
\begin{aligned}
\lambda_{r+1}(D) = & \min_{\dim(S)=h-r} \max_{0\neq x\in S} \frac{x^\top D x}{x^\top x} \\
(\dim(\ker(U_1)) \geq h-r) \leq & \min_{\substack{S\subseteq\ker(U_1)\\ \dim(S)=h-r}} \max_{0\neq x\in S} \frac{x^\top D x}{x^\top x} \\
= & \min_{\substack{S\subset\ker(U_1)\\ \dim(S)=r}} \max_{0\neq x\in S} \frac{x^\top (-V^\top V) x}{x^\top x} \leq 0\,,
\end{aligned}
$$

and

$$
\begin{aligned}
\lambda_{r}(D) = & \max_{\dim(S)=r} \min_{0\neq x\in S} \frac{x^\top D x}{x^\top x} \\
(\dim(\ker(V(0))) \geq h-m \geq r) \geq & \max_{\substack{S\subseteq\ker(V(0))\\ \dim(S)=r}} \min_{0\neq x\in S} \frac{x^\top D x}{x^\top x} \\
= & \max_{\substack{S\subset\ker(V(0))\\ \dim(S)=r}} \min_{0\neq x\in S} \frac{x^\top U_1^\top U_1 x}{x^\top x} \geq 0\,.
\end{aligned}
$$

Similarly, we have

$$
\lambda_{m+1}(-D) \leq \min_{\substack{S\subseteq\ker(V)\\ \dim(S)=h-m}} \max_{0\neq x\in S} \frac{x^\top (-U_1^\top U_1) x}{x^\top x} \leq 0\,,
$$

and

$$
\lambda_{m}(-D) \geq \max_{\substack{S\subseteq\ker(U_1(0))\\ \dim(S)=m}} \min_{0\neq x\in S} \frac{x^\top V^\top V x}{x^\top x} \geq 0\,.
$$

These inequalities together imply

$$
\min\{\lambda_r(D), \lambda_m(-D)\} = \sigma_{r+m}(D)\,.
$$

Here we also use the fact that D is symmetric. Now by (2.85), we immediately obtain that conditioned on event (2.84), with probability 1, the following holds,

$$
\underline{\lambda}_{+} + \underline{\lambda}_{-} = \lambda_r(D) + \lambda_m(-D) \geq \ 2\sigma_{r+m}(D) \geq h^{1-2\alpha}\,,
$$

which is exactly a lower bound on the spectral gap $\underline{\Delta}$.

Regarding the second and third inequality, using the fact that

$$\|A\|_F \leq \sqrt{\min\{n, m\}}\|A\|_2,\ \forall A \in \mathbb{R}^{n\times m}\,,$$

we have

$$\begin{aligned}
\frac{1}{\sqrt{m}}\left\|U_1V^\top\right\|_F \leq \left\|U_1V^\top\right\|_2 &= \left\|\begin{bmatrix}0 & \Phi_1^\top\end{bmatrix}\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix}\begin{bmatrix}I_m\\ 0\end{bmatrix}\right\|_2\\
&= \left\|\begin{bmatrix}0 & \Phi_1^\top\end{bmatrix}\left(\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right)\begin{bmatrix}I_m\\ 0\end{bmatrix}\right\|_2\\
&\leq \left\|\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right\|_2\text{, for any } \eta\in\mathbb{R}\,,
\end{aligned}$$

where the second equality is due to the fact that $\begin{bmatrix}0 & \Phi_1^\top\end{bmatrix}\begin{bmatrix}I_m\\ 0\end{bmatrix} = 0$. And

$$\begin{aligned}
\frac{1}{\sqrt{m+r}}\left\|\begin{bmatrix}VU_2^\top\\ U_1U_2^\top\end{bmatrix}\right\|_F \leq \left\|\begin{bmatrix}VU_2^\top\\ U_1U_2^\top\end{bmatrix}\right\|_2 &= \left\|\begin{bmatrix}I_m & 0\\ 0 & \Phi_1^\top\end{bmatrix}\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix}\begin{bmatrix}0\\ \Phi_2\end{bmatrix}\right\|_2\\
&= \left\|\begin{bmatrix}I_m & 0\\ 0 & \Phi_1^\top\end{bmatrix}\left(\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right)\begin{bmatrix}0\\ \Phi_2\end{bmatrix}\right\|_2\\
&\leq \left\|\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right\|_2\text{, for any } \eta\in\mathbb{R}\,,
\end{aligned}$$

where the second equality is due to the fact that $\begin{bmatrix}I_m & 0\\ 0 & \Phi_1^\top\end{bmatrix}\begin{bmatrix}0\\ \Phi_2\end{bmatrix} = 0$. Notice that

$$\left\|\begin{bmatrix}V\\ U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right\|_2 = \max_i\left|\sigma_i^2(\begin{bmatrix}V^\top & U^\top\end{bmatrix}) - \eta\right|\,.$$

Again we let $h > \left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$. When event (2.84) happens, $\sigma_i^2(\begin{bmatrix}V^\top & U^\top\end{bmatrix})$ are within the interval $\left[\left(h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2, \left(h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2\right]$. Since the choice of η is arbitrary, we pick

$$\eta = h^{1-2\alpha} + \left(\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2\,, \tag{2.86}$$

which is the mid-point of this interval, then we have

$$\begin{aligned}
&\max_i \left|\sigma_i^2(\begin{bmatrix}V^\top & U^\top\end{bmatrix}) - \eta\right| \\
\leq\ &\max\left\{\left|\left(h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2 - \eta\right|, \left|\left(h^{\frac{1}{2}-\alpha} + \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2 - \eta\right|\right\} \\
&(\eta \text{ is the mid-point}) \\
\leq\ &\left|\left(h^{\frac{1}{2}-\alpha} - \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2 - h^{1-2\alpha} - \left(\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^\alpha}\right)^2\right| \\
=\ &2\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,.
\end{aligned}$$

Therefore, when $h > \left(\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}\right)^2$, conditioned on event (2.84), with probability 1, we have

$$\begin{aligned}
\left\|U_1V^\top\right\|_F &\leq \sqrt{m}\left\|\begin{bmatrix}V\\U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right\|_2 \leq 2\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,, \\
\left\|\begin{bmatrix}VU_2^\top\\U_1U_2^\top\end{bmatrix}\right\|_F &\leq \sqrt{m+r}\left\|\begin{bmatrix}V\\U\end{bmatrix}\begin{bmatrix}V^\top & U^\top\end{bmatrix} - \eta I_{m+n}\right\|_2 \leq 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,,
\end{aligned} \tag{2.87}$$

where we choose η as in (2.86).

When $h > h_0 = 16\left(\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}\right)^2$ and conditioned on event (2.84), events (2.85) and (2.87) happen with probability 1, hence the probability that both (2.85) and (2.87) happen is at least the probability of event (2.84), which is at least $1-\delta$. □

With Lemma 2.1, we can prove Theorem 2.2.

Proof of Theorem 2.2. From Corollary 2.1 and Proposition 2.2, the stationary point $\hat{U}, \hat{V}$ satisfy

$$\hat{U}_1\hat{V}^\top = \Phi_1^\top\hat{\Theta}, \quad U_2(\infty) = U_2(0)\,,$$

provided that spectral gap $\underline{\lambda}_+ + \underline{\lambda}_-$ is non-zero, which is guaranteed with high probability by Lemma 2.1. Hence we have

$$\begin{aligned}
\|\hat{U}\hat{V}^\top - \hat{\Theta}\|_2 = &\ \|\Phi_1\hat{U}_1\hat{V}^\top + \Phi_2 U_2(\infty)\hat{V}^\top - \hat{\Theta}\|_2 \\
= &\ \|\Phi_1\Phi_1^\top\hat{\Theta} + \Phi_2 U_2(\infty)\hat{V}^\top - \hat{\Theta}\|_2 \\
= &\ \|\Phi_2 U_2(\infty)\hat{V}^\top\|_F \\
= &\ \|\Phi_2 U_2(0)\hat{V}^\top\|_F = \|U_2(0)\hat{V}^\top\|_2 \leq \|U_2(0)\hat{V}^\top\|_F\,.
\end{aligned}$$

Consider the following dynamics

$$\frac{d}{dt}\begin{bmatrix} VU_2^\top \\ U_1U_2^\top \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & E^\top\Sigma_x^{1/2} \\ \Sigma_x^{1/2}E & 0 \end{bmatrix}}_{:=A_Z}\underbrace{\begin{bmatrix} VU_2^\top \\ U_1U_2^\top \end{bmatrix}}_{:=Z}, \tag{2.88}$$

which can be viewed as a time-variant linear system, and in particular, by [64, Theorem 7.3.3], we have $\|A_Z\|_2 = \|\Sigma_x^{1/2}E\|_2$. Notice that here the Z is different from the one in the proof for Proposition 2.2.

From (2.88), we have

$$\begin{aligned}
\frac{d}{dt}\|Z\|_F^2 = &\ 2\mathrm{tr}\left(Z^\top A_Z Z\right) \\
= &\ 2\mathrm{tr}\left(ZZ^\top A_Z\right) \\
\leq &\, 2\|A_Z\|_2\mathrm{tr}\left(ZZ^\top\right) \\
= &\ 2\|\Sigma_x^{1/2}E\|_2\|Z\|_F^2 \\
\leq &\ 2\lambda_1^{1/2}(\Sigma_x)\|E\|_2\|Z\|_F^2 \leq 2\lambda_1^{1/2}(\Sigma_x)\|E\|_F\|Z\|_F^2\,.
\end{aligned}$$

By Grönwall's inequality [65], we have $\forall t \geq 0$,

$$\begin{aligned}
&\|Z(t)\|_F^2 \leq \exp\left(\int_0^t 2\lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F^2 \\
\Rightarrow\ &\|Z(t)\|_F \leq \exp\left(\int_0^t \lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F
\end{aligned} \tag{2.89}$$

Using Lemma 2.1, for $h > h_0' := 16\left(\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, with probability at least

$1-\delta$ we have all the following.

$$\begin{aligned} \underline{\Delta} >& \; h^{1-2\alpha}\,. \qquad (2.90)\\ \left\|U_1(0)V^\top(0)\right\| \leq& \; 2\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,, \qquad (2.91)\\ \|Z(0)\|_F = \left\|\begin{bmatrix}V(0)U_2^\top(0)\\U_1(0)U_2^\top(0)\end{bmatrix}\right\|_F \leq& \; 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}} \qquad (2.92)\end{aligned}$$

By Corollary 2.1, we have

$$\|E(t)\|_F^2 \leq \exp\left(-\lambda_r(\Sigma_x)\underline{\Delta}t\right)\|E(0)\|_F^2\,,$$

then by (2.90), we have

$$\begin{aligned} &\|E(t)\|_F^2 \leq \exp\left(-2h^{1-2\alpha}\lambda_r(\Sigma_x)t\right)\|E(0)\|_F^2\\ \Rightarrow &\|E(t)\|_F \leq \exp\left(-h^{1-2\alpha}\lambda_r(\Sigma_x)t\right)\|E(0)\|_F\,.\end{aligned}$$

Finally, from (2.89), we have

$$\begin{aligned}\|Z(t)\|_F \leq& \; \exp\left(\int_0^t \lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F\\ \leq& \; \exp\left(\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^t \exp\left(-h^{1-2\alpha}\lambda_r(\Sigma_x)\tau\right)d\tau\right)\right)\|Z(0)\|_F\\ \leq& \; \exp\left(\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^\infty \exp\left(-h^{1-2\alpha}\lambda_r(\Sigma_x)\tau\right)d\tau\right)\right)\|Z(0)\|_F\\ =& \; \exp\left(\frac{\lambda_1^{1/2}(\Sigma_x)}{h^{1-2\alpha}\lambda_r(\Sigma_x)}\|E(0)\|_F\right)\|Z(0)\|_F\,. \qquad (2.93)\end{aligned}$$

The initial error depends on the initialization but can be upper bounded as

$$\begin{aligned}\|E(0)\|_F =& \; \|W^\top Y - \Sigma_x^{1/2}U_1(0)V^\top(0)\|_F\\ =& \; \|\Sigma_x^{1/2}(\Sigma_x^{-1/2}W^\top Y - U_1(0)V^\top(0))\|_F\\ \leq& \; \lambda_1^{1/2}(\Sigma_x)\|\Phi_1^\top X^\dagger Y - U_1(0)V^\top(0)\|_F\\ \leq& \; \lambda_1^{1/2}(\Sigma_x)\|X^\dagger Y\|_F + \lambda_1^{1/2}(\Sigma_x)\|U_1(0)V^\top(0)\|_F\,,\end{aligned}$$

then we can write (2.93) as

$$\begin{aligned}
\|Z(t)\|_F \leq\; & \exp\left(\frac{\lambda_1(\Sigma_x)}{h^{1-2\alpha}\lambda_r(\Sigma_x)}\|X^\dagger Y\|_F\right)\exp\left(\frac{\lambda_1(\Sigma_x)}{h^{1-2\alpha}\lambda_r(\Sigma_x)}\|U_1(0)V^\top(0)\|_F\right)\|Z(0)\|_F \\
=\; & \left[\exp\left(\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|Y\|_F\right)\exp\left(\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|U_1(0)V^\top(0)\|_F\right)\right]^{1/h^{1-2\alpha}}\|Z(0)\|_F\,.
\end{aligned} \tag{2.94}$$

For the second exponential, we let $h_0 := \max\left\{h_0', 4\frac{\lambda_1^2(\Sigma_x)}{\lambda_r^2(\Sigma_x)}m\left(\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}\right)^2\right\}$, then $\forall h > h_0^{1/(4\alpha-1)}$, by (2.91) we have

$$\exp\left(\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|U_1(0)V^\top(0)\|_F\right) \leq \exp\left(2\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\right) \leq e\,. \tag{2.95}$$

Notice that $h > h_0^{1/(4\alpha-1)}$ also ensures $h > h_0^{1/(4\alpha-1)} \geq h_0 \geq h_0'$, hence the width condition for (2.90)(2.92)(2.91) to hold is satisfied.

Finally by (2.92)(2.95), we write (2.94) as

$$\begin{aligned}
\|Z(t)\|_F \leq\; & \left[\exp\left(1+\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|X^\dagger Y\|_F\right)\right]^{1/h^{1-2\alpha}}\|Z(0)\|_F \\
\leq\; & \underbrace{\left[\exp\left(1+\frac{\lambda_1(\Sigma_x)}{\lambda_r(\Sigma_x)}\|X^\dagger Y\|_F\right)\right]^{1/h^{1-2\alpha}}}_{:=C^{1/h^{1-2\alpha}}} 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}} \\
=\; & 2C^{1/h^{1-2\alpha}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,.
\end{aligned}$$

Therefore for some $C > 0$ that depends on the data (X, Y), given any $0 < \delta < 1$, when $h > h_0^{1/(4\alpha-1)}$ as defined above, with at least probability $1-\delta$, we have

$$\begin{aligned}
\|\hat{U}\hat{V}^\top - \hat{\Theta}\|_2 \leq\; & \|U_2(0)\hat{V}^\top\|_F \\
\leq\; & \sup_{t>0}\|U_2(0)V^\top(t)\|_F \\
\leq\; & \sup_{t>0}\|Z(t)\|_F \leq 2C^{1/h^{1-2\alpha}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}\,.
\end{aligned}$$

□

2.2 Multi-layer linear networks

This section aims to provide a general framework for analyzing the convergence of gradient flow on multi-layer linear models, that generalizes the convergence analysis for two-layer linear networks in Section 2.1. We consider a loss function of the form $\mathcal{L} = f(W_1 W_2 \cdots W_L)$, where f satisfies the gradient dominance property. Our analysis relies on a novel characterization of the gradient of the overparameterized loss as the composition of the non-overparametrized gradient with a time-varying (weight-dependent) linear operator whose smallest eigenvalue controls the convergence rate. The convergence analysis reduces to finding a uniform lower bound on the least eigenvalue of this time-varying linear operator over the entire training trajectory. However, finding such a uniform lower bound for general networks is extremely difficult even in the case of linear networks because the linear operator depends nontrivially on the weight matrix trajectories. As a consequence, in this work we focus on two- and three-layer neural networks as well as some classes of deep networks for which bounds are possible to obtain despite the complex dependency of the operator on the weight matrix trajectories. More specifically:

- Our analysis shows that the convergence rate depends on two trajectory-specific quantities: 1) the *imbalance matrices*, which measure the difference between the weights of adjacent layers, and 2) a lower bound on the least singular values of *weight product* $W = W_1 W_2 \cdots W_L$. The former is time-invariant under gradient flow, thus determined at initialization, while the latter can be controlled by initializing the product sufficiently close to its optimum.

- We provide a rate bound that applies to three-layer networks under general initialization. For deep networks, we study a broader class of initialization that covers most initialization schemes used in prior work [24, 25, 26, 27, 30, 28] for both multi-layer linear networks and diagonal linear networks while providing an improved rate bound.

- Our results directly apply to loss functions commonly used in regression tasks, and extend to loss functions used in classification tasks with an alternative assumption on f, under which we show $\mathcal{O}(1/t)$ convergence of the loss.

2.2.1 Problem setup

This section considers finding a matrix W that solves

$$\min_{W\in\mathbb{R}^{n\times m}} f(W)\,, \tag{2.96}$$

with the following assumption on f.

Assumption 2.1. *f is differentiable and satisfies*[2]*:*

***A1**: f satisfies the Polyak-Łojasiewicz (PL) condition, i.e. $\|\nabla f(W)\|_F^2 \geq \gamma(f(W) - f^*), \forall W$. This condition is also known as gradient dominance.*

***A2**: f is K-smooth and μ-strongly convex.*

While classic work [69] has shown that the gradient descent update on W with proper step size ensures a linear rate of convergence of $f(W)$ towards its optimal value f^*, the recent surge of research on the convergence and implicit bias of gradient-based methods for deep neural networks has led to a great amount of work on the *overparametrized* problem:

$$\min_{\{W_l\}_{l=1}^L} \mathcal{L}\left(\{W_l\}_{l=1}^L\right) = f(W_1W_2\cdots W_L)\,, \tag{2.97}$$

where $L \geq 2$, $W_l \in \mathbb{R}^{h_{l-1}\times h_l}, i = 1,\cdots,L$, with $h_0 = n, h_L = m$ and a width constraint $\min\{h_1,\cdots,h_{L-1}\} \geq \min\{n,m\}$. This assumption on $\min\{h_1,\cdots,h_{L-1}\}$

[2] Note that **A2** assumes μ-strong convexity, which implies **A1** with $\gamma = 2\mu$. However, we list **A1** and **A2** separately since they have different roles in our analysis.

is necessary to ensure that the optimal value of (2.97) is also f^*, and in this case, the *product* $\prod_{l=1}^{L} W_l$ can represent an *overparametrized linear network/model* [27, 25].

Convergence via gradient dominance

For problem (2.97), consider the gradient flow dynamics on the loss function $\mathcal{L}\left(\{W_l\}_{l=1}^{L}\right)$:

$$\dot{W}_l = -\frac{\partial}{\partial W_l}\mathcal{L}\left(\{W_l\}_{l=1}^{L}\right)\,, l = 1, \cdots, L\,. \tag{2.98}$$

The gradient flow dynamics can be viewed as gradient descent with "infinitesimal" step size and convergence results for gradient flow can be used to understand the corresponding gradient descent algorithm with sufficiently small step size [70]. We have the following result regarding the time-derivative of $\mathcal{L}$ under gradient flow.

Lemma 2.6. *Under continuous dynamics in* (2.98)*, we have*

$$\begin{aligned} \dot{\mathcal{L}} =& \; -\|\nabla\mathcal{L}\left(\{W_l\}_{l=1}^{L}\right)\|_F^2 \\ =& \; -\left\langle \mathcal{T}_{\{W_l\}_{l=1}^{L}}\nabla f(W), \nabla f(W)\right\rangle_F\,, \end{aligned} \tag{2.99}$$

where $W = \prod_{l=1}^{L} W_l$*, and* $\mathcal{T}_{\{W_l\}_{l=1}^{L}} = \sum_{l=1}^{L}\mathcal{T}_l$ *is a sum of* L *positive semi-definite linear operator on* $\mathbb{R}^{n\times m}$*:*

$$\mathcal{T}_l E = \left(\prod_{i=0}^{l-1} W_i\right)\left(\prod_{i=0}^{l-1} W_i\right)^\top E \left(\prod_{i=l+1}^{L+1} W_i\right)^\top\left(\prod_{i=l+1}^{L+1} W_i\right).$$

Proof. The gradient flow dynamics (2.98) satisfies

$$\frac{d}{dt}W_l = -\frac{\partial}{\partial W_l}\mathcal{L}\left(\{W_l\}_{l=1}^{L}\right) = -\left(\prod_{i=1}^{l-1} W_i\right)^\top \nabla f(W)\left(\prod_{i=l+1}^{L+1} W_i\right)^\top\,, \tag{2.100}$$

where $W = \prod_{l=1}^{L} W_i$ and $W_0 = I_n, W_{L+1} = I_m$.

Therefore

$$
\begin{aligned}
\dot{\mathcal{L}} =& \sum_{l=1}^{L} \left\langle \frac{\partial}{\partial W_l} \mathcal{L}\left(\{W_l\}_{l=1}^{L}\right), \frac{d}{dt} W_l \right\rangle_F \\
=& -\sum_{l=1}^{L} \left\| \frac{\partial}{\partial W_l} \mathcal{L}\left(\{W_l\}_{l=1}^{L}\right) \right\|_F^2 \\
=& -\sum_{l=1}^{L} \left\langle \left(\prod_{i=1}^{l-1} W_i \right)^\top \nabla f(W) \left(\prod_{i=l+1}^{L+1} W_i \right)^\top, \left(\prod_{i=1}^{l-1} W_i \right)^\top \nabla f(W) \left(\prod_{i=l+1}^{L+1} W_i \right)^\top \right\rangle_F \\
=& -\sum_{l=1}^{L} \left\langle \left(\prod_{i=1}^{l-1} W_i \right) \left(\prod_{i=1}^{l-1} W_i \right)^\top \nabla f(W) \left(\prod_{i=l+1}^{L+1} W_i \right)^\top \left(\prod_{i=l+1}^{L+1} W_i \right), \nabla f(W) \right\rangle_F \\
=& -\left\langle \sum_{l=1}^{L} \left(\prod_{i=1}^{l-1} W_i \right) \left(\prod_{i=1}^{l-1} W_i \right)^\top \nabla f(W) \left(\prod_{i=l+1}^{L+1} W_i \right)^\top \left(\prod_{i=l+1}^{L+1} W_i \right), \nabla f(W) \right\rangle_F \\
=& -\left\langle \mathcal{T}_{\{W_l\}_{l=1}^{L}} \nabla f(W), \nabla f(W) \right\rangle_F .
\end{aligned}
$$

□

With Lemma 2.6, our convergence analysis is as follows: For this overparameterized problem, the minimum $\mathcal{L}^*$ of (2.97) is f^*. Then from Lemma 2.6 and Assumption **A1**, we have

$$
\begin{aligned}
\dot{\mathcal{L}} =& -\left\langle \mathcal{T}_{\{W_l\}_{l=1}^{L}} \nabla f(W), \nabla f(W) \right\rangle_F \\
\leq& -\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^{L}}) \|\nabla f(W)\|_F^2 \qquad (2.101) \\
\overset{(\mathbf{A1})}{\leq}& -\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^{L}}) \gamma (f(W) - f^*) \qquad (2.102) \\
=& -\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^{L}}) \gamma (\mathcal{L} - \mathcal{L}^*).
\end{aligned}
$$

If we find an $\alpha > 0$ such that $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^{L}}) \geq \alpha, \forall t$, then the following inequality holds on the entire training trajectory $\frac{d}{dt}\left(\mathcal{L} - \mathcal{L}^*\right) \leq -\alpha\gamma\left(\mathcal{L} - \mathcal{L}^*\right)$. Therefore, by using Grönwall's inequality [65], we can show that the loss function $\mathcal{L}$ converges exponential to its minimum:

$$
\mathcal{L}(t) - \mathcal{L}^* \leq \exp\left(-\alpha\gamma t\right)\left(\mathcal{L}(0) - \mathcal{L}^*\right), \forall t \geq 0 . \qquad (2.103)
$$

Therefore, to show exponential convergence of the loss, we need to lower bound $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$.

Key challenge: Most existing work on the convergence of gradient flow/descent on linear networks implicitly provides a lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$ throughout the training trajectory, under particular assumptions on the initialization and network structure: For extremely wide networks under NTK initialization [23], the weights do not deviate too much from their initialization, from which one has $\mathcal{T}_{\{W_l(t)\}_{l=1}^L} \simeq \mathcal{T}_{\{W_l(0)\}_{l=1}^L}$, then the analysis reduces to finding eigenvalue bound for a fixed operator, rather than a time-varying one. Outside the kernel regime, one requires a uniform lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$ that accounts for the evolution of the weights. What has been facilitating the analysis are special initialization schemes that induce persistent structural properties on the weights, from which the operator can be simplified. For example, under balanced initialization [26], the linear operator would only depend on the product of the weights, instead of individual ones. To show convergence for general initialization without any structural property on the weights, one not only requires some analysis of the evolution of weights but, most importantly, also a careful eigenvalue analysis on $\mathcal{T}_{\{W_l(t)\}_{l=1}^L}$. However, the operator $\mathcal{T}_{\{W_l(t)\}_{l=1}^L}$ is a polynomial on the weight matrices whose degree depends on the network depth L, and *the higher the degree of* $\mathcal{T}_{\{W_l(t)\}_{l=1}^L}$*, the harder it is to bound its least eigenvalue.*

We first revisit our convergence analysis developed for two-layer networks from the last section, then we show that much of its ingredients hint at possible ways to lower bound $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$ for deep networks, then present our convergence results regarding deep networks.

Revisit two-layer linear networks

We revisit our convergence analysis developed for two-layer networks from the last section through the lens of our general convergence analysis. For $L = 2$, we have

$$\lambda_{\min}(\mathcal{T}_{\{W_1,W_2\}}) = \lambda_n(W_1 W_1^\top) + \lambda_m(W_2^\top W_2)\,. \tag{2.104}$$

In the proof of Proposition 2.1, there is already a lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_1,W_2\}})$ with the knowledge of the imbalance and the product.

Lemma 2.7. *When $L = 2$, given weights $\{W_1, W_2\}$ with imbalance matrix $D = W_1^\top W_1 - W_2 W_2^\top$ and product $W = W_1 W_2$, define Δ_+, Δ_-, and $\underline{\Delta}$ as in Proposition 2.1, then for the linear operator $\mathcal{T}_{\{W_1,W_2\}}$, we have*

$$\begin{aligned}2\lambda_{\min}\big(\mathcal{T}_{\{W_1,W_2\}}\big) \geq &- \Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_n^2\,(W)}\\ &- \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_m^2\,(W)}\,.\end{aligned} \tag{2.105}$$

Implication on convergence: Note that (2.105) is almost a lower bound for the eigenvalue $\lambda_{\min}\left(\mathcal{T}_{\{W_1(t),W_2(t)\}}\right), t \geq 0$, as the imbalance matrix D is time-invariant (so are $\Delta_+, \Delta_-, \underline{\Delta}$), except the right-hand side of (2.105) also depends on $\sigma_{\min}(W(t))$. If f satisfies **A2**, then f has a unique minimizer W^*. Moreover, one can show that given a initial product $W(0)$, $W(t)$ is constrained to lie within a closed ball

$$\left\{W : \|W - W^*\|_F \leq \sqrt{\frac{K}{\mu}}\|W(0) - W^*\|_F\right\},$$

i.e., $W(t)$ does not get too far away from W^* during training. We can use this to derive the following lower bound on $\sigma_{\min}(W(t))$:

$$\sigma_{\min}(W(t)) \geq \left[\sigma_{\min}(W^*) - \sqrt{\frac{K}{\mu}}\|W(0) - W^*\|_F\right]_+\,. \tag{2.106}$$

This margin term being positive guarantees that the closed ball excludes any W with $\sigma_{\min}(W) = 0$. With this observation, we find a lower bound $\lambda_{\min}\left(\mathcal{T}_{\{W_1(t),W_2(t)\}}\right), t \geq 0$ that depends on both the weight imbalance and margin, and the exponential convergence of loss $\mathcal{L}$ follows:

Theorem 2.3. *Let D be the imbalance matrix for $L = 2$. The continuous dynamics in* (2.98) *satisfy*

$$\mathcal{L}(t) - \mathcal{L}^* \leq \exp\left(-\alpha_2 \gamma t\right)\left(\mathcal{L}(0) - \mathcal{L}^*\right), \forall t \geq 0\,, \tag{2.107}$$

1. *If f satisfies only **A1**, then $\alpha_2 = \underline{\Delta}$;*

2. *If f satisfies both **A1** and **A2**, then*

$$\alpha_2 = \; -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\nu_n^2} - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\nu_m^2}\,, \tag{2.108}$$

where

$$\begin{aligned} \nu_n &= \left[\sigma_n\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+ \,, \\ \nu_m &= \left[\sigma_m\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+ \,, \end{aligned}$$

$W(0) = \prod_{l=1}^L W_l(0)$*, and W^* equal to the unique optimizer of f.*

Theorem 2.3 is new as it generalizes the result in Section 2.1.2, which is only for l_2 loss in linear regression. We consider a general loss function defined by f, including the losses for matrix factorization [26], linear regression [31], and matrix sensing [10]. Additionally, [26] first introduced the notion of margin for f in matrix factorization problems ($K = 1, \mu = 1$), and we extend it to any f that is smooth and strongly convex.

Towards deep Networks: So far, we revisited our results on two-layer networks, showing how $\lambda_{\min}(\mathcal{T}_{W_1,W_2})$ can be lower bounded by weight imbalance and product, from which the convergence result is derived. Can we generalize the analysis to deep networks? The challenge is that even computing $\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L})$ given the weights $\{W_l\}_{l=1}^L$ is complicated: For $L = 2$, $\lambda_{\min}(\mathcal{T}_{W_1,W_2}) = \lambda_n(W_1 W_1^\top) + \lambda_m(W_2^\top W_2)$, but such nice relation does not exist for $L > 3$, which makes the search for a tight lower bound potentially difficult. On the other hand, the findings in (2.105) shed light on what can be potentially shown for the deep layer case:

1. For two-layer networks, we always have the bound $\lambda_{\min}\left(\mathcal{T}_{\{W_1,W_2\}}\right) \geq \underline{\Delta}$, which depends only on the imbalance. *Can we find a lower bound on the convergence rate of a deep network that depends only on an imbalance quantity analogous to $\underline{\Delta}$? If yes, how does such a quantity depend on network depth?*

2. For two-layer networks, the bound reduces to $\sqrt{\underline{\Delta}^2 + 4\sigma_{\min}^2(W)}$ when the imbalance is "well-conditioned" (Δ_+, Δ_- are small). *For deep networks, can we characterize such joint contribution from the imbalance and product, given a similar assumption?*

We will answer these questions as we present our convergence results for deep networks.

2.2.2 Three-layer linear networks

To answer the first question of how weight imbalance effect convergence, we derive a novel rate bound for three-layer models showing the general effect of imbalance. For ease of presentation, we denote the two imbalance matrices for three-layer models, D_1 and D_2, as

$$-D_1 = W_2W_2^\top - W_1^\top W_1 := D_{21}\,, \tag{2.109}$$

$$D_2 = W_2^\top W_2 - W_3W_3^\top := D_{23}. \tag{2.110}$$

Our lower bound comes after a few definitions.

Definition 2.1. *Given two real symmetric matrices A, B of order n, we define a non-commutative binary operation $\wedge_r$ as $A \wedge_r B := \operatorname{diag}\{\min\{\lambda_i(A), \lambda_{i+1-r}(B)\}\}_{i=1}^n$, where $\lambda_j(\cdot) = +\infty, \forall j \leq 0$.*

Definition 2.2. *Given* $(D_{21}, D_{23}) \in \mathbb{R}^{h_1 \times h_1} \times \mathbb{R}^{h_2 \times h_2}$*, define*

$$\begin{aligned}
&\bar{D}_{h_1} = \mathrm{diag}\{\max\{\lambda_i(D_{21}), \lambda_i(D_{23}), 0\}\}_{i=1}^{h_1}, \bar{D}_{h_2} = \mathrm{diag}\{\max\{\lambda_i(D_{21}), \lambda_i(D_{23}), 0\}\}_{i=1}^{h_2}, \\
&\Delta_{21} = \mathrm{tr}(\bar{D}_{h_1}) - \mathrm{tr}(\bar{D}_{h_1} \wedge_n D_{21}), \qquad \Delta_{21}^{(2)} = \mathrm{tr}(\bar{D}_{h_1}^2) - \mathrm{tr}\left((\bar{D}_{h_1} \wedge_n D_{21})^2\right), \\
&\Delta_{23} = \mathrm{tr}(\bar{D}_{h_2}) - \mathrm{tr}(\bar{D}_{h_2} \wedge_m D_{23}), \qquad \Delta_{23}^{(2)} = \mathrm{tr}(\bar{D}_{h_2}^2) - \mathrm{tr}\left((\bar{D}_{h_2} \wedge_m D_{23})^2\right).
\end{aligned}$$

Theorem 2.4. *When* $L = 3$*, given weights* $\{W_1, W_2, W_3\}$ *with imbalance matrices* (D_{21}, D_{23}) *as defined in* (2.109)(2.110)*, then for the linear operator* $\mathcal{T}_{\{W_1,W_2,W_3\}}$*, we have*

$$\lambda_{\min}\left(\mathcal{T}_{\{W_1,W_2,W_3\}}\right) \geq \frac{1}{2}(\Delta_{21}^{(2)} + \Delta_{21}^2) + \Delta_{21}\Delta_{23} + \frac{1}{2}(\Delta_{23}^{(2)} + \Delta_{23}^2) := \Delta^*(D_{21}, D_{23}). \tag{2.111}$$

With the theorem, we have the following corollary.

Corollary 2.4. *When* $L = 3$*, given initialization with imbalance matrices* (D_{21}, D_{23}) *and* f *satisfying* ***A1****, the continuous dynamics in* (2.98) *satisfy*

$$\mathcal{L}(t) - \mathcal{L}^* \leq \exp(-\alpha_3 \gamma t)(\mathcal{L}(0) - \mathcal{L}^*), \forall t \geq 0, \tag{2.112}$$

where $\alpha_3 = \frac{1}{2}(\Delta_{21}^{(2)} + \Delta_{21}^2) + \Delta_{21}\Delta_{23} + \frac{1}{2}(\Delta_{23}^{(2)} + \Delta_{23}^2)$.

We make the following remarks regarding the contribution.

Optimal bound via imbalance: First of all, our bound should be considered as the best lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_1(t),W_2(t),W_3(t)\}})$ one can obtain given knowledge of the imbalance matrices only. More importantly, the bound works for ANY initialization and has the same role as $\underline{\Delta}$ does in two-layer networks, i.e., (2.111) quantifies the general effect imbalance on the convergence. Finding an improved bound that takes the effect of $\sigma_{\min}(W)$ into account is an interesting future research direction.

Implication on convergence: Corollary 2.4 suggests that the gradient flow starting at any initialization with positive $\Delta^*(D_{21}, D_{23})$ converges exponentially. However, due to its complicated expression, it is less clear under what initialization the bound is positive. We conjecture that most random initialization schemes would have a positive Δ^*, and through some numerical experiments in Section 2.2.5, we show that random initialization (outside NTK regime) is most likely to have a positive Δ^*, thus exponential convergence is guaranteed by our theorem.

Technical contribution: We highlight in Section 2.2.1 the challenge in bounding $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$ for deep networks. One needs to develop new mathematical tools for the eigenanalysis: The way we find the lower bound in (2.111) is by studying the generalized eigenvalue interlacing relation imposed by the imbalance constraints. Specifically, $W_2W_2^\top - W_1^\top W_1 = D_{21}$ suggests that $\lambda_{i+n}(W_2W_2^\top) \leq \lambda_i(D_{21}) \leq \lambda_i(W_2W_2^\top), \forall i$ because $W_2W_2^\top - D_{21}$ is a matrix of at most rank-n. We derive, from such interlacing relation, novel eigenvalue bounds (See Lemma 2.13) on $\lambda_n(W_1^\top W_1)$ and $\lambda_n(W_1W_2W_2^\top W_1)$ that depends on eigenvalues of both $W_2W_2^\top$ and D_{21}. Then the eigenvalues of $W_2W_2^\top$ can also be controlled by the fact that W_2 must satisfy both imbalance equations in (2.109)(2.110). Since imbalance equations like those in (2.109)(2.110) appear in deep networks and certain nonlinear networks [61, 71], we believe our mathematical results are potentially useful for understanding those networks.

Comparison with prior work: The convergence of multi-layer linear networks under balanced initialization ($D_l = 0, \forall l$) has been studied in [26, 27], and our result is complementary as we study the effect of non-zero imbalance on the convergence of three-layer networks. Some settings with imbalanced weights have been studied: [28] studies a special initialization scheme ($D_l \succeq 0, l = 1, \cdots, L-2$, and $D_{L-1} \succeq \lambda I_{h_{L-1}}$) that forces the partial ordering of the weights, and [72] uses similar initialization to study the linear residual networks. Our bound works for such initialization and also show such partial ordering is not necessary for convergence.

2.2.3 Deep linear networks

The lower bound we derived for three-layer networks applies to any initialization. However, the bound is a fairly complicated function of all the imbalance matrices that is hard to interpret. Searching for such a general bound is even more challenging for models with arbitrary depth ($L \geq 3$). Therefore, our results for deep networks will rely on extra assumptions on the weights that simplify the lower bound to facilitate interpretability. Specifically, we consider the following properties of the weights:

Definition 2.3. *A set of weights $\{W_l\}_{l=1}^L$ with imbalance matrices $\{D_l := W_l^\top W_l - W_{l+1}W_{l+1}^\top\}_{l=1}^{L-1}$ is said to be* ***unimodal with index*** *l^* if there exists $l^* \in [L]$ such that*

$$D_l \succeq 0, \quad \textit{for } \ l < l^* \qquad \textit{and} \qquad D_l \preceq 0, \quad \textit{for } \ l \geq l^* \,.$$

We define its ***cumulative imbalances*** *$\{\tilde{d}_{(i)}\}_{i=1}^{L-1}$ as*

$$\tilde{d}_{(i)} = \begin{cases} \sum_{l=l^*}^{i} \lambda_m(-D_l), & i \geq l^* \\ \sum_{l=i}^{l^*-1} \lambda_n(D_l), & i < l^* \end{cases} .$$

Furthermore, for weights with unimodality index l^*, if additionally, $D_l = d_l I_{h_l}, l = 1, \cdots, L-1$ for

$$d_l \geq 0, \quad \textit{for } \ l < l^* \qquad \textit{and} \qquad d_l \leq 0, \quad \textit{for } \ l \geq l^* \,,$$

those weights are said to have ***homogeneous imbalance****.*

The unimodality assumption enforces an ordering of the weights w.r.t. the positive semi-definite cone. This is more clear when considering scalar weights $\{w_l\}_{l=1}^L$, in which case unimodality requires w_l^2 to be descending until index l^* and ascending afterward. Under this unimodality assumption, we show that imbalance contributes to the convergence of the loss via a product of cumulative imbalanaces. Furthermore, we also show the combined effects of imbalance and weight product when the imbalance matrices are "well-conditioned" (in this case, homogeneous).

Table 2-I. Compare our rate bound with prior work on deep networks.

Assumptions	[26]	[28]	Ours
Unimodal weights	N/A	λ^{L-1}	$\prod_{l=1}^{L-1}\tilde{d}(i)$
Homogeneous imbalance	N/A	λ^{L-1}	$\sqrt{(\prod_{l=1}^{L-1}\tilde{d}(i))^2+(L\sigma_{\min}^{2-2/L}(W))^2}$
Balanced ($D_l=0,\forall l$)	$L\sigma_{\min}^{2-2/L}(W)$	N/A	

Theorem 2.5. *For weights $\{W_l\}_{l=1}^L$ with unimodality index l^* and product $W=\prod_{l=1}^L W_l$, we have*

$$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right)\geq\prod_{l=1}^{L-1}\tilde{d}_{(i)}\,. \tag{2.113}$$

Furthermore, if the weights have homogeneous imbalance,

$$\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L})\geq\sqrt{\left(\prod_{l=1}^{L-1}\tilde{d}_{(i)}\right)^2+\left(L\sigma_{\min}^{2-\frac{2}{L}}(W)\right)^2}\,, \tag{2.114}$$

We make the following remarks:

Connection to results for three-layer: For three-layer networks, we present an optimal bound given some imbalance. Interestingly, when comparing the three-layer bound (2.111) with our bound in (2.113), we have:

Claim 1. *When $L=3$, for weights $\{W_1,W_2,W_3\}$ with unimodality index l^*,*

1. *If $l^*=1$, then $\frac{1}{2}(\Delta_{23}^{(2)}+\Delta_{23}^2)=\prod_{l=1}^{L-1}\tilde{d}_{(i)}$ and $\frac{1}{2}(\Delta_{21}^{(2)}+\Delta_{21}^2)=\Delta_{21}\Delta_{23}=0$;*
2. *If $l^*=2$, then $\Delta_{21}\Delta_{23}=\prod_{l=1}^{L-1}\tilde{d}_{(i)}$ and $\frac{1}{2}(\Delta_{21}^{(2)}+\Delta_{21}^2)=\frac{1}{2}(\Delta_{23}^{(2)}+\Delta_{23}^2)=0$;*
3. *If $l^*=3$, then $\frac{1}{2}(\Delta_{21}^{(2)}+\Delta_{21}^2)=\prod_{l=1}^{L-1}\tilde{d}_{(i)}$ and $\frac{1}{2}(\Delta_{23}^{(2)}+\Delta_{23}^2)=\Delta_{21}\Delta_{23}=0$.*

The claim shows that the bound in (2.113) is optimal for three-layer unimodal weights as it coincides with the one in Theorem 2.4. We conjecture that (2.113) is also optimal for multi-layer unimodal weights and leave the proof for future research. Interestingly, while the bound for three-layer models is complicated, the three terms $\frac{1}{2}(\Delta_{23}^{(2)}+\Delta_{23}^2)$, $\Delta_{21}\Delta_{23}$, $\frac{1}{2}(\Delta_{21}^{(2)}+\Delta_{21}^2)$, seem to roughly capture how close the weights are to unimodality. This hints at potential generalization of Theorem 2.4 to the deep case where the bound should have L terms capturing how close the weights are to those with different unimodality ($l^*=1,\cdots,L$).

Effect of imbalance under unimodality: For simplicity, we assume unimodality index $l^* = L$. The bound $\prod_{i=1}^{L-1} \tilde{d}_{(i)}$, as a product of cumulative imbalances, generally grows exponentially with the depth L. Prior work [28] studies the case $D_l \succeq 0, l = 1, \cdots, L-2$, and $D_{L-1} \succeq \lambda I_{h_{L-1}}$, in which case $\prod_{i=1}^{L-1} \tilde{d}_{(i)} \geq \lambda^{L-1}$. Our bound $\prod_{i=1}^{L-1} \tilde{d}_{(i)}$ suggests the dependence on L could be super-exponential: When $\lambda_n(D_l) \geq \epsilon > 0$, for $l = 1, \cdots, L-1$, we have $\prod_{i=1}^{L-1} \tilde{d}_{(i)} = \prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(D_l) \geq \prod_{l=1}^{L-1} l\epsilon = \epsilon^{L-1}(L-1)!$, which grows faster in L than λ^{L-1} for any λ. Therefore, for gradient flow dynamics, the depth L could greatly improve convergence in the presence of weight imbalance. One should note, however, that such analysis can not be directly translated into fast convergence guarantees of gradient descent algorithm as one requires careful tuning of the step size for the discrete updates to follow the trajectory of the continuous dynamics [70].

Convergence under unimodality: Regarding exponential convergence, the following immediately comes from Theorem 2.5:

Corollary 2.5. *If the initialization weights $\{W_l(0)\}_{l=1}^L$ are unimodal, then the continuous dynamics in (2.98) satisfy*

$$\mathcal{L}(t) - \mathcal{L}^* \leq \exp\left(-\alpha_L \gamma t\right)(\mathcal{L}(0) - \mathcal{L}^*), \forall t \geq 0, \tag{2.115}$$

1. *If f satisfies **A1** only, then $\alpha_L = \Pi_{i=1}^{L-1} \tilde{d}_{(i)}$;*

2. *If f satisfies both **A1**, **A2**, and the weights additionally have homogeneous imbalance, then $\alpha_L = \sqrt{\left(\prod_{i=1}^{L-1} \tilde{d}_{(i)}\right)^2 + (L\nu_{\min})^2}$, where*

$$\nu_{\min} = \left[\sigma_{\min}(W^*) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+,$$

   $W(0) = \prod_{l=1}^L W_l(0)$ *and W^* equal to the unique optimizer of f.*

Spectral initialization under l_2 loss: Suppose $f = \frac{1}{2}\|Y - W\|_F^2$ and $W = \prod_{l=1}^L W_l$. We write the SVD of $Y \in \mathbb{R}^{n\times m}$ as $Y = P \begin{bmatrix} \Sigma_Y & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q \\ 0 \end{bmatrix} := P\tilde{\Sigma}_Y \tilde{Q}$, where $P \in$

$\mathcal{O}(n), Q \in \mathcal{O}(m)$. Consider the spectral initialization $W_1(0) = R\Sigma_1 V_1^\top$, $W_l(0) = V_{l-1}\Sigma_l V_l^\top, l = 2, \cdots, L-1$, $W_L(0) = V_{L-1}\Sigma_L \tilde{Q}$, where $\Sigma_l, l = 1, \cdots, L$ are diagonal matrices of our choice and $V_l \in \mathbb{R}^{n \times h_l}, l = 1, \cdots, L-1$ with $V_l^\top V_l = I_{h_l}$. It can be shown that [25]

$$W_1(t) = R\Sigma_1(t)V_1^\top, \;\; W_L(t) = V_{L-1}\Sigma_L(t)\tilde{Q},$$
$$W_l(t) = V_{l-1}\Sigma_l(t)V_l^\top, l = 2, \cdots, L-1.$$

Moreover, only the first m diagonal entries of Σ_l are changing. Let $\sigma_{i,l}, \sigma_{i,y}$ denote the i-th diagonal entry of Σ_l, and $\tilde{\Sigma}_Y$ respectively, then the dynamics of $\{\sigma_{i,l}\}_{l=1}^L$ follow the gradient flow on $\mathcal{L}_i(\{\sigma_{i,l}\}_{l=1}^L) = \frac{1}{2}|\sigma_{i,y} - \prod_{l=1}^L \sigma_{i,l}|^2$ for $i = 1, \cdots, m$, which is exactly a multi-layer model with scalar weights: $f(w) = |\sigma_{i,y} - w|^2/2, w = \prod_{l=1}^L w_l$. Therefore, spectral initialization under l_2 loss can be decomposed into m deep linear models with scalar weights, whose convergence is shown by Corollary 2.5. Note that networks with scalar weights are always unimodal, because the gradient flow dynamics remain the same under any reordering of the weights, and always have homogeneous imbalance, because the imbalances are scalars.

Diagonal linear networks: Consider f a function on $\mathbb{R}^n$ satisfying **A1** and $\mathcal{L} = f(w_1 \odot \cdots \odot w_L)$, where $w_l \in \mathbb{R}^n$ and $\odot$ denote the entrywise product. We can show that $\dot{\mathcal{L}} = -\|\nabla\mathcal{L}\|_F^2 \leq -(\min_{1\leq i\leq n} \lambda_{\min}(\mathcal{T}_{\{w_{l,i}\}_{l=1}^L}))\gamma(\mathcal{L} - \mathcal{L}^*)$, where $w_{l,i}$ is the i-th entry of w_l. Then Theorem 2.5 gives lower bound on each $\lambda_{\min}(\mathcal{T}_{\{w_{l,i}\}_{l=1}^L})$.

Comparison with prior work: Regarding unimodality, [28] studies the initialization scheme $D_l \succeq 0, l = 1, \cdots, L-2$ and $D_{L-1} \succeq \lambda I_{h_{L-1}}$, which is a special case ($l^* = L$) of ours. The homogeneous imbalance assumption was first shown in [25] for two-layer networks, and we generalize it to the deep case. We compare, in Table 2-I, our bound to existing work [26, 28] on convergence of deep linear networks outside the kernel regime. Note that [28] only studies a special case of unimodal weights ($l^* = L$ with $\tilde{d}_{(i)} \geq \lambda > 0, \forall i$). For homogeneous imbalance, [28] studied spectral initialization and diagonal linear networks, which necessarily have homogeneous imbalance, but the result does not generalize to the case of matrix weights. Our results for homogeneous imbalance works also for networks with matrix weights, and our rate also shown the effect of the product $L\sigma_{\min}^{2-2/L}(W)$, thus covers the balanced initialization [26] as well.

2.2.4 Convergence results for classification tasks

Note that the loss functions used in [73, 28] are classification losses, such as the exponential loss, which do not satisfy **A1**. However, we can show $\mathcal{O}\left(1/t\right)$ convergence with an alternative assumption.

Theorem 2.6. *Suppose f satisfies (**A1'**) $\|\nabla f(W)\|_F \geq \gamma(f(W) - f^*), \forall W \in \mathbb{R}^{n\times m}$. Given initialization $\{W_l(0)\}_{l=1}^L$ such that $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L}) \geq \alpha,\ \forall t$, then*

$$\mathcal{L}(t) - \mathcal{L}^* \leq \frac{\mathcal{L}(0) - \mathcal{L}^*}{(\mathcal{L}(0) - \mathcal{L}^*)\alpha\gamma^2 t + 1}\,. \tag{2.116}$$

The lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_l(t)\}_{l=1}^L})$ can be obtained for different networks by our results in previous sections.

2.2.5 Numerical experiments

In Section 2.2.2, we have shown a rate bound for three-layer networks under general initialization in Theorem 2.4. However, due to its complicated expression, it is less clear under what initialization the bound is positive. Through some numerical experiments, we show that our bound is very likely to be positive under various random initialization schemes. In Figure 2-6, we show a box plot of our bound $\Delta = \Delta^*(D_{21}, D_{23})$ in Theorem 2.4 under: NTK initialization [23], Xavier initialization [74], and Fan_{out} initialization. These initialization schemes all randomly sample the network weights with Gaussian distribution, but with different variances for each layer. Shown from the box plot, our bound is non-vacuous for random initialization:

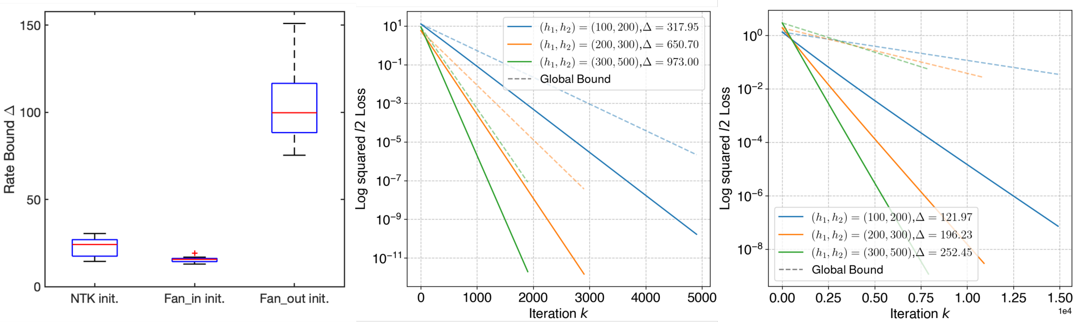

Figure 2-6. Three-layer network under random initialization most likely converges exponentially. Left: Box plot of our bound $\Delta = \Delta^*(D_{21}, D_{23})$ for different initialization schemes on a three-layer network with $n = 5, m = 1, h_1 = h_2 = 200$, each box is generated with 100 random samples of the weights; Middle: Gradient descent on three-layer network with $n = 1, m = 1$. Right: Gradient descent on three-layer network with $n = 5, m = 1$. For different network widths, we compare the actual loss with our theoretical bound.

All the sampled instances of random initialization, we have $\Delta^*(D_{21}, D_{23}) > 0$, i.e., exponential convergence is guaranteed for all cases, while no existing work provide exponential convergence guarantee for this experiment because the initialization has a non-zero imbalance ([26] requires balancedness), and the network has only a moderate width ([23] requires extremely large width).

Next, we run gradient descent on three-layer networks under Fan_{out} initialization with a loss function $\mathcal{L} = \|Y - W_1W_2W_3\|_F^2/2$, and compare our theoretical bound from Corollary 2.4 with the actual loss curve. We see that for certain cases $n = 1, m = 1$ (Middle plot in Figure 2-6), our bound provides a good characterization of the actual convergence rate, but appears less tight for problems with higher dimensions $n = 5, m = 1$ (Right plot in Figure 2-6). However, we note that even in the latter case, initialization with a large value of the bound Δ does converge faster, hence there exists some correlation between the bound Δ and the actual convergence rate, and formally justify such correlation is an interesting future research. Moreover, we view the fact that Δ fails to provide a tight bound for problems with larger scales as some evidence showing that imbalance constraint is relatively weaker in characterizing the eigenmodes of $\mathcal{T}_{\{W_l(t)\}_{l=1}^L}$ for deep networks, despite its usefulness in shallow networks [25, 30]. This suggests that we should be searching for new structural properties on the weights to fully understand the convergence of deep networks.

Proof of Theorem 2.3

The following Lemma will be used in the proof of Theorem 2.3.

Lemma 2.8. *If f satisfies **A2**, then the gradient flow dynamics (2.98) satisfies*

$$\sigma_{\min}\left(W(t)\right) \geq \sigma_{\min}\left(W^*\right) - \sqrt{\frac{K}{\mu}} \|W(0) - W^*\|_F \,, \forall t \geq 0$$

where $W(t) = \prod_{l=1}^{L} W_l(t)$ and W^* is the unique minimizer of f.

Proof. From [69], we know if f is μ-strongly convex, then it has unique minimizer W^* and

$$f(W) - f^* \geq \frac{\mu}{2} \|W - W^*\|_F^2 \,.$$

Additionally, if f is K-smooth, then

$$f(W) - f^* \leq \frac{K}{2} \|W - W^*\|_F^2 \,.$$

This suggests that for any $t \geq 0$,

$$\frac{K}{2} \|W(t) - W^*\|_F^2 \geq \mathcal{L}(t) - \mathcal{L}^* \geq \frac{\mu}{2} \|W - W^*\|_F^2 \,.$$

Therefore we have the following

$$\begin{aligned}
\sigma_{\min}\left(W(t)\right) =&\ \sigma_{\min}\left(W(t) - W^* + W^*\right) \\
\text{(Weyl's inequality [64, 7.3.P16])} \geq&\ \sigma_{\min}(W^*) - \|W(t) - W^*\|_2 \\
\geq&\ \sigma_{\min}(W^*) - \|W(t) - W^*\|_F \\
(f \text{ is } \mu\text{-strongly convex}) \geq&\ \sigma_{\min}(W^*) - \sqrt{\frac{2}{\mu}(\mathcal{L}(t) - \mathcal{L}^*)} \\
(\mathcal{L}(t) \text{ non-decreasing under (2.98)}) \geq&\ \sigma_{\min}(W^*) - \sqrt{\frac{2}{\mu}(\mathcal{L}(0) - \mathcal{L}^*)} \\
(f \text{ is } K\text{-smooth}) \geq&\ \sigma_{\min}(W^*) - \sqrt{\frac{K}{\mu}\|W(0) - W^*\|_F^2} \\
=&\ \sigma_{\min}\left(W^*\right) - \sqrt{\frac{K}{\mu}}\|W(0) - W^*\|_F \,.
\end{aligned}$$

□

Proof of Theorem 2.3. As shown in (2.101) in Section 2.2.1. We have

$$\frac{d}{dt}(\mathcal{L}(t) - \mathcal{L}^*) \leq -\lambda_{\min}\mathcal{T}_{\{W_1(t),W_2(t)\}}\gamma(\mathcal{L}(t) - \mathcal{L}^*)\,.$$

Consider any $\{W_1(t), W_2(t)\}$ on the trajectory, we have, by Lemma 2.7,

$$\begin{aligned}
\lambda_{\min}\mathcal{T}_{\{W_1(t),W_2(t)\}} \overset{\text{Lemma 2.7}}{\geq}&\ \frac{1}{2}\Big(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_n^2\left(W(t)\right)} \\
&\quad -\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_m^2\left(W(t)\right)}\Big) \\
\geq&\ \frac{1}{2}\left(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2} - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2}\right) = \underline{\Delta} := \alpha_2\,.
\end{aligned}$$

When f also satisfies A2: we need to prove

$$\sigma_n\left(W(t)\right) \geq \left[\sigma_n\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+\,, \tag{2.117}$$

$$\sigma_m\left(W(t)\right) \geq \left[\sigma_m\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+\,. \tag{2.118}$$

When $n = m$, both inequalities are equivalent to

$$\sigma_{\min}(W(t)) \geq \left[\sigma_{\min}(W^*) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+\,,$$

which is true by Lemma 2.8.

When $n \neq m$, one of the two inequalities become trivial. For example, if $n > m$, then (2.117) is trivially $0 \geq 0$, and (2.118) is equivalent to

$$\sigma_{\min}(W(t)) \geq \left[\sigma_{\min}(W^*) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+ ,$$

which is true by Lemma 2.8.

Overall, we have

$$\begin{aligned}
&\lambda_{min}\mathcal{T}_{\{W_1(t),W_2(t)\}} \\
&\overset{\text{Lemma 2.7}}{\geq} \frac{1}{2}\Big(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\sigma_n^2\left(W(t)\right)} \\
&\qquad\qquad -\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\sigma_m^2\left(W(t)\right)}\Big) \\
&\geq \frac{1}{2}\left(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\left(\left[\sigma_n\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+\right)^2}\right. \\
&\qquad \left. -\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\left(\left[\sigma_m\left(W^*\right) - \sqrt{K/\mu}\|W(0) - W^*\|_F\right]_+\right)^2}\right) \\
&:= \alpha_2 \,.
\end{aligned}$$

Either case, we have $\frac{d}{dt}(\mathcal{L}(t) - \mathcal{L}^*) \leq -\alpha_2\gamma(\mathcal{L}(t) - \mathcal{L}^*)$, and by Grönwall's inequality, we have

$$\mathcal{L}(t) - \mathcal{L}^* \leq \exp(-\alpha_2\gamma t)(\mathcal{L}(0) - \mathcal{L}^*)\,.$$

□

Proof of Theorem 2.4

Theorem 2.4 is the direct consequence of the following two results.

Lemma 2.9. *Given any set of weights* $\{W_1, W_2, W_3\} \in \mathbb{R}^{n\times h_1} \times \mathbb{R}^{h_1\times h_2} \times \mathbb{R}^{h_2\times m}$*, we have*

$$\lambda_{min}\mathcal{T}_{\{W_1,W_2,W_3\}} \geq \lambda_n(W_1W_2W_2^\top W_1^\top) + \lambda_n(W_1W_1^\top)\lambda_m(W_3^\top W_3) + \lambda_m(W_3^\top W_2^\top W_2W_3)\,.$$

Theorem 2.7. *Given imbalance matrices pair* $(D_{21}, D_{23}) \in \mathbb{R}^{h_1\times h_1} \times \mathbb{R}^{h_2\times h_2}$*, then the optimal value of*

$$\begin{aligned} \min_{W_1,W_2,W_3} \; & 2\left(\lambda_n(W_1W_2W_2^\top W_1^\top) + \lambda_n(W_1W_1^\top)\lambda_m(W_3^\top W_3) + \lambda_m(W_3^\top W_2^\top W_2W_3)\right) \\ s.t. \quad & W_2W_2^\top - W_1^\top W_1 = D_{21} \\ & W_2^\top W_2 - W_3W_3^\top = D_{23} \end{aligned}$$

is

$$\Delta^*(D_{21}, D_{23}) = \Delta_{21}^{(2)} + \Delta_{21}^2 + 2\Delta_{21}\Delta_{23} + \Delta_{23}^{(2)} + \Delta_{23}^2 \,.$$

Combining those two results gets $\lambda_{min}\mathcal{T}_{\{W_1,W_2,W_3\}} \geq \Delta^*(D_{21}, D_{23})/2$, as stated in Theorem 2.4. Lemma 2.9 is intuitive and easy to prove:

Proof of Lemma 2.9. Notice that $\mathcal{T}_{\{W_1,W_2,W_3\}}$ is the summation of three positive semi-definite linear operators on $\mathbb{R}^{n\times m}$, i.e.

$$\mathcal{T}_{\{W_1,W_2,W_3\}} = \mathcal{T}_{12} + \mathcal{T}_{13} + \mathcal{T}_{23} \,,$$

where

$$\mathcal{T}_{12}E = W_1W_2W_2^\top W_1^\top E, \;\mathcal{T}_{13}E = W_1W_1^\top EW_3^\top W_3, \;\mathcal{T}_{23}E = EW_3^\top W_2^\top W_2W_3 \,,$$

and $\lambda_{min}\mathcal{T}_{12} = \lambda_n(W_1W_2W_2^\top W_1^\top)$, $\lambda_{min}\mathcal{T}_{13} = \lambda_n(W_1W_1^\top)\lambda_m(W_3^\top W_3)$, $\lambda_{min}\mathcal{T}_{23} = \lambda_m(W_3^\top W_2^\top W_2W_3)$.Therefore, let E_{min} with $\|E_{min}\|_F = 1$ be the eigenmatrix associated with $\lambda_{min}\mathcal{T}_{\{W_1,W_2,W_3\}}$, we have

$$\begin{aligned} \lambda_{min}\mathcal{T}_{\{W_1,W_2,W_3\}} &= \left\langle \mathcal{T}_{\{W_1,W_2,W_3\}}, E_{min} \right\rangle_F \\ &= \langle \mathcal{T}_{12}, E_{min}\rangle_F + \langle \mathcal{T}_{13}, E_{min}\rangle_F + \langle \mathcal{T}_{23}, E_{min}\rangle_F \\ &\geq \lambda_{min}\mathcal{T}_{12} + \lambda_{min}\mathcal{T}_{13} + \lambda_{min}\mathcal{T}_{23} \,. \end{aligned}$$

□

The rest of this section is dedicated to proving Theorem 2.7

We will first state a few Lemmas that will be used in the proof, then show the proof for Theorem 2.7, and present the long proofs for the auxiliary Lemmas in the end.

Auxiliary lemmas

The main ingredient used in proving Theorem 2.7 is the notion of r-interlacing relation between the spectrum of two matrices, which is a natural generalization of the interlacing relation as seen in classical Cauchy Interlacing Theorem [64, Theorem 4.3.17].

Definition 2.4. *Given real symmetric matrices A, B of order n, write $A \succeq_r B$, if*

$$\lambda_{i+r}(A) \leq \lambda_i(B) \leq \lambda_i(A)\,, \forall i$$

where $\lambda_j(\cdot) = +\infty, j \leq 0$ and $\lambda_j(\cdot) = -\infty, j > n$. The case $r = 1$ gives the interlacing relation.

Claim 2. *We only need to check*

$$\lambda_{i+r}(A) \leq \lambda_i(B) \leq \lambda_i(A)\,, \forall i \in [n]\,,$$

for showing $A \succeq_r B$.

Proof. Any inequality regarding index outside $[n]$ is trivial. □

The following Lemma is a direct concequence of Weyl's inequality [64, Theorem 4.3.1], and stated as a special case of [64, Corollary 4.3.3]

Lemma 2.10. *Given real symmetric matrices A, B of order n, if $A - B$ is positive semi-definite and* $\operatorname{rank}(A - B) \leq r$*, then $A \succeq_r B$*

The converse is also true

Lemma 2.11. *Given real symmetric matrices A, B of order n, if $A \succeq_r B$, then there exists a positive semi-definite matrix $XX^\top$ with $\operatorname{rank}(XX^\top) \leq r$ and a real orthogonal matrix V such that $A - XX^\top = VBV^\top$.*

Proof. The case $r = 1$ is proved in [64, Theorem 4.3.26]. The case $r > 1$ is proved in [75, Theorem 1.3] by induction. □

Specifically for our problem, we also need the following ($\bar{D}_{h_1}$ and $\bar{D}_{h_2}$ are defined in Section 2.2.2)

Lemma 2.12. *Given imbalance matrices pair $(D_{21}, D_{23}) \in \mathbb{R}^{h_1 \times h_1} \times \mathbb{R}^{h_2 \times h_2}$, we have $\bar{D}_{h_1} \succeq_n D_{21}$ and $\bar{D}_{h_2} \succeq_m D_{23}$.*

In our analysis, the weights W_1, W_2, W_3 are "constrained" by the imbalance D_{21}, D_{23}, such constraints leads to some special eigenvalue bounds (The operation $\wedge_r$ was defined in Section 2.2.2):

Lemma 2.13. *Given an positive semi-definite matrix A of order n, and $Z \in \mathbb{R}^{r \times n}$ with $r \leq n$, when*

$$A - Z^\top Z = B\,,$$

we have

$$\lambda_r(ZZ^\top) \geq \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)\,,$$

and

$$2\lambda_r(ZAZ^\top) \geq \operatorname{tr}\left(A^2\right) - \operatorname{tr}\left((A \wedge_r B)^2\right) + \left(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)\right)^2$$

and this bound is actually tight

Lemma 2.14. *Given two real symmetric matrices A, B of order n, if $A \succeq_r B$ ($r \leq n$), then there exist $Z \in \mathbb{R}^{r \times n}$ and some orthogonal matrix $V \in \mathcal{O}(n)$, such that*

$$A - Z^\top Z = VBV^\top\,,$$

and

$$\begin{aligned}\lambda_r(ZZ^\top) &= \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)\,,\\ 2\lambda_r(ZAZ^\top) &= \operatorname{tr}\left(A^2\right) - \operatorname{tr}\left((A \wedge_r B)^2\right) + \left(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)\right)^2\,.\end{aligned}$$

Proof of Theorem 2.7

With these Lemmas, we are ready to prove Theorem 2.7.

Proof of Theorem 2.7. The proof is presented in two parts: First, we show $\Delta^*(D_{21}, D_{23})$ is a lower bound on the optimal value; Then we construct an optimal solution (W_1^*, W_2^*, W_3^*) that attains $\Delta^*(D_{21}, D_{23})$ as the objective value.

Showing $\Delta^*(D_{21}, D_{23})$ is a lower bound: Given any feasible triple (W_1, W_2, W_3), the imbalance equations

$$W_2W_2^\top - W_1^\top W_1 = D_{21}\,, \tag{2.119}$$

$$W_2^\top W_2 - W_3W_3^\top = D_{23}\,, \tag{2.120}$$

implies $W_2W_2^\top \succeq_n D_{21}$ and $W_2^\top W_2 \succeq_m D_{23}$ by Lemma 2.10. These interlacing relation shows

$$\lambda_i(W_2W_2^\top) \geq \lambda_i(D_{21}), \quad \lambda_i(W_2^\top W_2) \geq \lambda_i(D_{23}), \forall i\,,$$

which is

$$\lambda_i(W_2W_2^\top) = \lambda_i(W_2^\top W_2) \geq \max\{\lambda_i(D_{21}), \lambda_i(D_{21}), 0\} = \lambda_i(\bar{D}_{h_1}) \geq 0\,, \forall i \in [h_1] \tag{2.121}$$

Now by Lemma 2.13, imbalance equation (2.119) suggests

$$\lambda_n(W_1W_1^\top) \geq \operatorname{tr}(W_2W_2^\top) - \operatorname{tr}(W_2W_2^\top \wedge_n D_{21})\,,$$

and

$$\begin{aligned}&2\lambda_n(W_1W_2W_2^\top W_1^\top)\\ &\qquad \geq \operatorname{tr}\left((W_2W_2^\top)^2\right) - \operatorname{tr}\left((W_2W_2^\top \wedge_n D_{21})^2\right) + \left(\operatorname{tr}(W_2W_2^\top) - \operatorname{tr}(W_2W_2^\top \wedge_n D_{21})\right)^2\,.\end{aligned}$$

Notice that

$$\begin{aligned}
\lambda_r(W_1W_1^\top) \geq\ & \operatorname{tr}(W_2W_2^\top) - \operatorname{tr}(W_2W_2^\top \wedge_n D_{21}) \\
=\ & \sum_{i=1}^{h_1} \lambda_i(W_2W_2^\top) - \min\{\lambda_i(W_2W_2^\top), \lambda_{i+1-n}(D_{21})\} \\
=\ & \sum_{i=1}^{h_1} \max\{\lambda_i(W_2W_2^\top) - \lambda_{i+1-n}(D_{21}), 0\} \\
\geq\ & \sum_{i=1}^{h_1} \max\{\lambda_i(\bar{D}_{h_1}) - \lambda_{i+1-n}(D_{21}), 0\} \\
=\ & \operatorname{tr}(\bar{D}_{h_1}) - \operatorname{tr}(\bar{D}_{h_1} \wedge_n D_{21}) = \Delta_{21}\,, \qquad (2.122)
\end{aligned}$$

where the inequality holds because (2.121) and the fact that ReLU function $f(x) = \max\{x, 0\}$ is a monotonically non-decreasing function.

Since Δ_{21} can be viewed as summation of ReLU outputs, it has to be non-negative, then (2.122) also suggests

$$\left(\operatorname{tr}(W_2W_2^\top) - \operatorname{tr}(W_2W_2^\top \wedge_n D_{21})\right)^2 \geq \Delta_{21}^2\,. \qquad (2.123)$$

Next we have

$$\begin{aligned}
& 2\lambda_n(W_1W_2W_2^\top W_1^\top) \\
\geq\ & \operatorname{tr}\left((W_2W_2^\top)^2\right) - \operatorname{tr}\left((W_2W_2^\top \wedge_n D_{21})^2\right) + \left(\operatorname{tr}(W_2W_2^\top) - \operatorname{tr}(W_2W_2^\top \wedge_n D_{21})\right)^2 \\
\overset{(2.123)}{\geq}\ & \Delta_{21}^2 + \operatorname{tr}\left((W_2W_2^\top)^2\right) - \operatorname{tr}\left((W_2W_2^\top \wedge_n D_{21})^2\right) \\
=\ & \Delta_{21}^2 + \sum_{i=1}^{h_1} \lambda_i^2(W_2W_2^\top) - \left(\min\{\lambda_i(W_2W_2^\top), \lambda_{i+1-n}(D_{21})\}\right)^2 \\
\geq\ & \Delta_{21}^2 + \sum_{i=1}^{h_1} \lambda_i^2(\bar{D}_{h_1}) - \left(\min\{\lambda_i(\bar{D}_{h_1}), \lambda_{i+1-n}(D_{21})\}\right)^2 \\
=\ & \Delta_{21}^2 + \operatorname{tr}\left(\bar{D}_{h_1}^2\right) - \operatorname{tr}\left((\bar{D}_{h_1} \wedge_n D_{21})^2\right) = \Delta_{21}^2 + \Delta_{21}^{(2)}\,,
\end{aligned}$$

where the last inequality is because (2.121) and the fact that the function

$$g(x) = x^2 - (\min\{x, a\})^2 = \begin{cases} 0, & x \leq a \\ x^2 - a^2, & x > a \end{cases}\,,$$

is monotonically non-decreasing on $\mathbb{R}_{\geq 0}$ for any constant $a \in \mathbb{R}$.

At this point, we have shown

$$\lambda_n(W_1 W_1^\top) \geq \Delta_{21}\,, \qquad 2\lambda_n(W_1 W_2 W_2^\top W_1^\top) \geq \Delta_{21}^2 + \Delta_{21}^{(2)}\,. \tag{2.124}$$

We can repeat the proofs above with the following replacement

$$W_2 \to W_2^\top, W_1 \to W_3^\top, D_{21} \to D_{23}, \bar{D}_{h_1} \to \bar{D}_{h_2}\,,$$

and obtain

$$\lambda_m(W_3^\top W_3) \geq \Delta_{23}\,, \qquad 2\lambda_m(W_3^\top W_2^\top W_2 W_3) \geq \Delta_{23}^2 + \Delta_{23}^{(2)}\,. \tag{2.125}$$

These inequalities (2.124)(2.125) show that

$$\Delta^*(D_{21}, D_{23}) = \Delta_{21}^{(2)} + \Delta_{21}^2 + 2\Delta_{21}\Delta_{23} + \Delta_{23}^{(2)} + \Delta_{23}^2\,.$$

is a lower bound on the optimal value of our optimization problem. Now we proceed to show tightness.

Constructing optimal solution: By Lemma 2.12, we know $\bar{D}_{h_1} \succeq_n D_{21}$, and by Lemma 2.14, there exists $Z_1 \in \mathbb{R}^{n \times h_1}$ and orthogonal $V_1 \in \mathcal{O}(h_1)$ such that

$$\bar{D}_{h_1} - Z_1^\top Z_1 = V_1 D_{21} V_1^\top\,, \tag{2.126}$$

and most importantly,

$$\lambda_n(Z_1 Z_1^\top) = \Delta_{21}, \qquad 2\lambda_n(Z_1 \bar{D}_{h_1} Z_1^\top) = \Delta_{21}^{(2)} + \Delta_{21}^2\,. \tag{2.127}$$

Similarly, by Lemma Lemma 2.12, we know $\bar{D}_{h_2} \succeq_m D_{23}$, and by Lemma 2.14, there exists $Z_3 \in \mathbb{R}^{m \times h_2}$ and orthogonal $V_3 \in \mathcal{O}(h_2)$ such that

$$\bar{D}_{h_2} - Z_3^\top Z_3 = V_3 D_{23} V_3^\top\,, \tag{2.128}$$

and most importantly,

$$\lambda_m(Z_3 Z_3^\top) = \Delta_{23}, \qquad 2\lambda_m\left(Z_3 \bar{D}_{h_2} Z_3^\top\right) = \Delta_{23}^{(2)} + \Delta_{23}^2\,. \tag{2.129}$$

Let

$$W_2^* = \begin{cases} V_1^\top \begin{bmatrix} \bar{D}^{\frac{1}{2}} & \mathbf{0}_{h_1\times(h_2-h_1)} \end{bmatrix} V_3, & h_2 \geq h_1 \\ V_1^\top \begin{bmatrix} \bar{D}^{\frac{1}{2}} \\ \mathbf{0}_{(h_1-h_2)\times h_2} \end{bmatrix} V_3, & h_2 < h_1 \end{cases},$$

where $\bar{D} = \mathrm{diag}\{\max\{\lambda_i(D_{21}), \lambda_i(D_{21}), 0\}\}_{i=1}^{\min\{h_1,h_2\}}$, and

$$W_1^* = Z_1 V_1, \qquad W_3^* = V_3^\top Z_3^\top\,,$$

we have

$$\begin{aligned} W_2^*(W_2^*)^\top - (W_1^*)^\top W_1^* &= V_1^\top \bar{D}_{h_1} V_1 - V_1^\top Z_1^\top Z_1 V_1 = D_{21} \\ (W_2^*)^\top W_2^* - W_3^*(W_3^*)^\top &= V_3^\top \bar{D}_{h_2} V_3 - V_3^\top Z_3 Z_3^\top V_3 = D_{23}\,, \end{aligned}$$

and

$$\begin{aligned} &\lambda_r(W_1^*(W_1^*)^\top) = \lambda_r(Z_1 Z_1^\top) = \Delta_{21}\,, \\ &\lambda_m((W_3^*)^\top W_3^*) = \lambda_m(Z_3^\top Z_3) = \Delta_{23}\,, \\ &2\lambda_r(W_1^* W_2^* (W_2^*)^\top (W_1^*)^\top) = \lambda_r(Z_1 \bar{D}_{h_1} Z_1^\top) = \Delta_{21}^{(2)} + \Delta_{21}^2\,, \\ &2\lambda_m((W_3^*)^\top (W_2^*)^\top W_2^* W_3^*) = \lambda_m(Z_3^\top \bar{D}_{h_2} Z_3) = \Delta_{23}^{(2)} + \Delta_{23}^2\,, \end{aligned}$$

Therefore the lower bound $\Delta^*(D_{21}, D_{23})$ is tight. □

Proofs of auxiliary lemmas

We finish this section by providing the proofs of auxiliary lemmas we used in the last section.

Proof of Lemma 2.12. Since (D_{21}, D_{23}) is a pair of imbalance matrices, there exists $W_2 W_2^\top$, such that

$$W_2 W_2^\top \succeq_n D_{21}, W_2^\top W_2 \succeq_m D_{23}\,, \tag{2.130}$$

because at least our weight initialization $W_1(0), W_2(0), W_3(0)$ have to satisfy the imbalance constraints $W_2(0)W_2(0)^\top - W_1^\top(0)W_1(0) = D_{21}, W_2^\top(0)W_2(0) - W_3(0)W_3^\top(0) = D_{23}$.

Therefore, for $0 < i \leq h_1 - n$,

$$\lambda_{i+n}(\bar{D}_{h_1}) = \max\{\lambda_{i+n}(D_{21}), \lambda_{i+n}(D_{23}), 0\} \leq \lambda_{i+n}(W_2 W_2^\top) \leq \lambda_i(D_{21}) \leq \lambda_i(\bar{D}_{h_1}) \,,$$

where the first two inequalities uses (2.130) and the fact that $\lambda_{i+n}(W_2 W_2^\top) = \lambda_{i+n}(W_2^\top W_2)$. Also the last inequality is from the fact that

$$\lambda_i(\bar{D}_{h_1}) = \max\{\lambda_i(D_{21}), \lambda_i(D_{23}), 0\}, \forall i \in [h_1] \,.$$

For $h_1 \geq i > h_1 - n$, we still have

$$-\infty = \lambda_{i+n}(\bar{D}_{h_1}) \leq \lambda_i(D_{21}) \leq \lambda_i(\bar{D}_{h_1}) \,,$$

Overall, we have

$$\lambda_{i+n}(\bar{D}_{h_1}) \leq \lambda_i(D_{21}) \leq \lambda_i(\bar{D}_{h_1}) \,, \forall i \,,$$

which is exactly $\bar{D}_{h_1} \succeq_n D_{21}$.

Similarly, for $0 < i \leq h_2 - m$,

$$\lambda_{i+m}(\bar{D}_{h_2}) = \max\{\lambda_{i+m}(D_{21}), \lambda_{i+m}(D_{23}), 0\} \leq \lambda_{i+m}(W_2^\top W_2) \leq \lambda_i(D_{23}) \leq \lambda_i(\bar{D}_{h_2}) \,,$$

where the first two inequalities uses (2.130) and the fact that $\lambda_{i+m}(W_2 W_2^\top) = \lambda_{i+m}(W_2^\top W_2)$. Also the last inequality is from the fact that

$$\lambda_i(\bar{D}_{h_2}) = \max\{\lambda_i(D_{21}), \lambda_i(D_{23}), 0\}, \forall i \in [h_2] \,.$$

For $h_2 \geq i > h_2 - m$, we still have

$$-\infty = \lambda_{i+m}(\bar{D}_{h_2}) \leq \lambda_i(D_{23}) \leq \lambda_i(\bar{D}_{h_2}) \,,$$

Overall, we have

$$\lambda_{i+m}(\bar{D}_{h_2}) \leq \lambda_i(D_{23}) \leq \lambda_i(\bar{D}_{h_2}) \,, \forall i \,,$$

which is exactly $\bar{D}_{h_2} \succeq_m D_{23}$. □

Proof of Lemma 2.13. Notice that $\mathrm{rank}(Z^\top Z) \leq r$, hence we consider the eigendecomposition

$$Z^\top Z = \sum_{i=1}^{r} \lambda_i(Z^\top Z) v_i v_i^\top \,,$$

where v_i are unit eigenvectors of $Z^\top Z$. Then we can write

$$A - \lambda_r(Z^\top Z) v_i v_i^\top - \sum_{i=1}^{r-1} \lambda_i(Z^\top Z) v_i v_i^\top = B$$

We let $D = A - \lambda_r(Z^\top Z) v_i v_i^\top$, then by Lemma 2.10, we know $A \succeq_1 D$, and $D \succeq_{r-1} B$, which suggests that $\forall i$,

$$\lambda_{i+1}(A) \leq \lambda_i(D) \leq \lambda_i(A) \tag{2.131}$$

$$\lambda_{i+r-1}(D) \leq \lambda_i(B) \leq \lambda_i(D) \,. \tag{2.132}$$

In particular, we have $\lambda_i(D) \leq \lambda_i(A)$ from (2.131) and $\lambda_i(D) \leq \lambda_{i+1-r}(B)$ from (2.132), which suggests

$$\lambda_i(D) \leq \min\{\lambda_i(A), \lambda_{i+1-r}(B)\} = \lambda_i\left(A \wedge_r B\right), \forall i \,.$$

Hence

$$\mathrm{tr}(A \wedge_r B) \geq \mathrm{tr}(D) = \mathrm{tr}(A) - \lambda_r(Z^\top Z)\mathrm{tr}(v_i v_i^\top) = \mathrm{tr}(A) - \lambda_r(Z^\top Z) \,.$$

This proves the first inequality.

For the second the inequality, let x be the unit eigenvector associated with $\lambda_r(ZAZ^\top)$, then $\lambda_r(ZAZ^\top) = x^\top ZAZ^\top x$. Now write

$$A - Zxx^\top Z^\top - Z(I - xx^\top)Z^\top = B \,.$$

Let $\tilde{D} = A - Zxx^\top Z^\top$, then again by Lemma 2.10 we have $A \succeq_1 \tilde{D}$, and $\tilde{D} \succeq_{r-1} B$.

Notice that

$$\begin{aligned} \tilde{D}^2 &= (A - Zxx^\top Z^\top)^2 \\ &= A^2 + (Zxx^\top Z^\top)^2 - AZxx^\top Z^\top - Zxx^\top Z^\top A \,. \end{aligned}$$

Taking trace on both side of this equation and using the cyclic property of trace operation lead to

$$\mathrm{tr}(\tilde{D}^2) = \mathrm{tr}\left(A^2\right) + \|Zx\|^4 - 2\lambda_r(ZAZ^\top)\,. \tag{2.133}$$

We only need to lower bound $\|Zx\|^4 - \mathrm{tr}(\tilde{D}^2)$, for which we write the eigendecomposition $\tilde{D}$ using eigenpairs $\{(\lambda_i(\tilde{D}), u_i)\}_{i=1}^n$ as

$$\tilde{D} = \sum_{i=1}^{n} \lambda_i(\tilde{D}) u_i u_i^\top = \sum_{j=1}^{n-1} \lambda_i(\tilde{D}) u_i u_i^\top + \lambda_n(\tilde{D}) u_n u_n^\top\,.$$

Then we have

$$\begin{aligned}\|Zx\|^2 = \mathrm{tr}(Zxx^\top Z^\top) =&\ \mathrm{tr}(A) - \mathrm{tr}(\tilde{D}) \\ =&\ \mathrm{tr}(A) - \sum_{j=1}^{n-1} \lambda_j(\tilde{D}) - \lambda_n(\tilde{D}) \\ \geq&\ \mathrm{tr}(A) - \sum_{j=1}^{n-1} \lambda_j(A \wedge_r B) - \lambda_n(\tilde{D}) \\ =&\ \mathrm{tr}(A) - \mathrm{tr}(A \wedge_r B) + \lambda_n(A \wedge_r B) - \lambda_n(\tilde{D})\,,\end{aligned}$$

where the inequality follows similar argument in the previous part of the proof and uses

$$\lambda_i(\tilde{D}) \leq \min\{\lambda_i(A), \lambda_{i+1-r}(B)\} = \lambda_i\left(A \wedge_r B\right)\,, \tag{2.134}$$

from the fact that $A \succeq_1 \tilde{D}$, and $\tilde{D} \succeq_{r-1} B$.

Now examine the right-hand side carefully: The first component $\mathrm{tr}(A) - \mathrm{tr}(A \wedge_r B)$ is non-negative because $\lambda_i(A) \geq \lambda_i(A \wedge_r B), \forall i$. The second component $\lambda_n(A \wedge_r B) - \lambda_n(\tilde{D})$ is non-negative as well by (2.134). Therefore the right-hand side is non-negative and we can take square on both sides of the inequality, namely,

$$\|W_1 x\|^4 \geq \Big(\mathrm{tr}(A) - \mathrm{tr}(A \wedge_r B) + \lambda_n(A \wedge_r B) - \lambda_n(\tilde{D})\Big)^2\,. \tag{2.135}$$

We also have

$$\begin{aligned}
\operatorname{tr}(\tilde{D}^2) = & \sum_{i=1}^{n-1} \lambda_i^2(\tilde{D}) + \lambda_n^2(\tilde{D}) \\
\leq & \sum_{i=1}^{n-1} \lambda_i^2(A \wedge_r B) + \lambda_n^2(\tilde{D}) \\
= & \operatorname{tr}\left((A \wedge_r B)^2\right) - \lambda_n^2(A \wedge_r B) + \lambda_n^2(\tilde{D}) \,, \qquad (2.136)
\end{aligned}$$

The inequality holds because for $i = 1, \cdots, n-1$,

$$0 \leq \lambda_{i+1}(A) \leq \lambda_i(\tilde{D}) \leq \lambda_i(A \wedge_r B) \,,$$

where the inequality on the left is from $A \succeq_1 \tilde{D}$ and the inequality on the right is due to (2.134).

With those two inequalities (2.135)(2.136), we have (For simplicity, denote $\lambda_\wedge := \lambda_n(A \wedge_r B), \tilde{\lambda} := \lambda_n(\tilde{D})$)

$$\begin{aligned}
& \|W_1 x\|^4 - \operatorname{tr}(\tilde{D}^2) - \left[(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B))^2 - \operatorname{tr}\left((A \wedge_r B)^2\right)\right] \\
\geq \; & \lambda_\wedge^2 + \tilde{\lambda}^2 - 2\lambda_\wedge \tilde{\lambda} + 2(\lambda_\wedge - \tilde{\lambda})(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)) + \lambda_\wedge^2 - \tilde{\lambda}^2 \\
= \; & 2\lambda_\wedge^2 - 2\lambda_\wedge \tilde{\lambda} + 2(\lambda_\wedge - \tilde{\lambda})(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)) \\
= \; & 2(\lambda_\wedge - \tilde{\lambda})(\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B) + \lambda_\wedge) \geq 0 \,,
\end{aligned}$$

where the last inequality is due to the facts that $\lambda_\wedge \geq \tilde{\lambda}$ by (2.134) and

$$\begin{aligned}
& \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B) + \lambda_\wedge \\
= & \sum_{i=1}^{n-1} (\lambda_i(A) - \lambda_i(A \wedge_r B)) + \lambda_n(A) \geq 0 \,.
\end{aligned}$$

This shows

$$\|Zx\|^4 - \operatorname{tr}(\tilde{D}^2) \geq (\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B))^2 - \operatorname{tr}\left((A \wedge_r B)^2\right) \,.$$

Finally from (2.133) we have

$$\begin{aligned}
2\lambda_r(ZAZ^\top) = & \operatorname{tr}\left((A)^2\right) + \|Zx\|^4 - \operatorname{tr}(\tilde{D}^2) \\
\geq & \operatorname{tr}\left((A)^2\right) - \operatorname{tr}\left((A \wedge_r B)^2\right) + (\operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B))^2 \,.
\end{aligned}$$

□

To proof Lemma 2.14, we need one final lemma

Lemma 2.15. *Given two real symmetric matrices A, B of order n, for any $r \leq n$, if $A \succeq_r B$, then $A \succeq_1 (A \wedge_r B)$ and $(A \wedge_r B) \succeq_{r-1} B$.*

Proof. Denote $D := A \wedge_r B$, we show $A \succeq_1 D$ and $D \succeq_{r-1} B$. The following statements holds for any index $i \in [n]$.

First of all, we have

$$\lambda_i(D) = \min\{\lambda_i(A), \lambda_{i+1-r}(B)\} \leq \lambda_i(A)\,, \tag{2.137}$$

and

$$\lambda_{i+1}(A) \leq \min\{\lambda_i(A), \lambda_{i+1-r}(B)\} = \lambda_i(D)\,, \tag{2.138}$$

where $\lambda_{i+1}(A) \leq \lambda_{i+1-r}(B)$ is from $A \succeq_r B$. (2.137)(2.138) together show $A \succeq_1 D$.

Next, notice that

$$\lambda_i(B) \leq \min\{\lambda_i(A), \lambda_{i+1-r}(B)\} = \lambda_i(D)\,, \tag{2.139}$$

where $\lambda_i(B) \leq \lambda_i(A)$ is from $A \succeq_r B$, and

$$\lambda_{i+r-1}(D) = \min\{\lambda_{i+r-1}(A), \lambda_i(B)\} \leq \lambda_i(B) \tag{2.140}$$

(2.139)(2.140) together show $D \succeq_{r-1} B$. □

Then we are ready to prove Lemma 2.14

Proof of Lemma 2.14. Denote $D := A \wedge_r B$. We have shown in Lemma 2.15 that $A \succeq_1 D$ and $D \succeq_{r-1} B$.

With the two interlacing relations, we know there exist $x \in \mathbb{R}^{n\times 1}, X \in \mathbb{R}^{n\times(r-1)}$ and some orthogonal matrices $V_1, V_2 \in \mathcal{O}(n)$ such that

$$A - xx^\top = V_1 D V_1^\top, \qquad D - XX^\top = V_2 B V_2^\top\,, \tag{2.141}$$

then let $V := V_1 V_2$, we have

$$A - xx^\top - V_1 X X^\top V_1^\top = V_1 V_2 B V_2^\top V_1^\top = V B V^\top . \tag{2.142}$$

Notice that

$$xx^\top + V_1 X X^\top V_1^\top = \begin{bmatrix} x & V_1 X \end{bmatrix} \begin{bmatrix} x^\top \\ X^\top V_1^\top \end{bmatrix} ,$$

then with $Z^\top := \begin{bmatrix} x & V_1 X \end{bmatrix} \in \mathbb{R}^{n \times r}$, we can write

$$A - Z^\top Z = V_1 V_2 B V_2^\top V_1^\top = V B V^\top .$$

It remains to show $\lambda_r(ZZ^\top)$ and $2\lambda_r(ZAZ^\top)$ have the exact expressions as stated.

Notice that $A - xx^\top = V_1 D V_1^\top$, then we have

$$\|x\|^2 = \operatorname{tr}(xx^\top) = \operatorname{tr}(A - V_1 D V_1^\top) = \operatorname{tr}(A) - \operatorname{tr}(D) . \tag{2.143}$$

Moreover, taking trace on both sides of $(A - xx^\top)^2 = (V_1 D V_1^\top)^2$ yields

$$\operatorname{tr}\left((A)^2\right) - 2x^\top A x + \|x\|^4 = \operatorname{tr}(D^2) ,$$

from which we have

$$2x^\top A x = \operatorname{tr}(A) - \operatorname{tr}(D^2) + \|x\|^4 = \operatorname{tr}(A) - \operatorname{tr}(D^2) + \left(\operatorname{tr}(A) - \operatorname{tr}(D)\right)^2 . \tag{2.144}$$

Finally, notice that the first diagonal entry of

$$ZZ^\top = \begin{bmatrix} x^\top \\ X^\top V_1^\top \end{bmatrix} \begin{bmatrix} x & V_1 X \end{bmatrix} = \begin{bmatrix} \|x\|^2 & x^\top X \\ X^\top x & X^\top X \end{bmatrix}$$

is $\|x\|^2$, we have, by [64, Corollary 4.3.34],

$$\lambda_r(ZZ^\top) \leq \|x\|^2 = \operatorname{tr}(A) - \operatorname{tr}(D) = \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B) .$$

Since we have already shown in Lemma 2.13 that

$$\lambda_r(ZZ^\top) \geq \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B) ,$$

we must have the exact equality $\lambda_r(ZZ^\top) = \operatorname{tr}(A) - \operatorname{tr}(A \wedge_r B)$.

Similarly, the first diagonal entry of

$$ZAZ^\top = \begin{bmatrix} x^\top \\ X^\top V_1^\top \end{bmatrix} A \begin{bmatrix} x & V_1 X \end{bmatrix} = \begin{bmatrix} x^\top A x & x^\top A X \\ X^\top A x & X^\top A X \end{bmatrix}$$

is $x^\top Ax$, then we have, by [64, Corollary 4.3.34],

$$2\lambda_r(ZAZ^\top) \leq 2x^\top Ax = \mathrm{tr}\left(A^2\right) - \mathrm{tr}\left((A \wedge_r B)^2\right) + (\mathrm{tr}(A) - \mathrm{tr}(A \wedge_r B))^2 \; .$$

Again, Lemma 2.13 shows the inequality in the opposite direction, hence one must take the equality

$$2\lambda_r(ZAZ^\top) = x^\top Ax = \mathrm{tr}\left(A^2\right) - \mathrm{tr}\left((A \wedge_r B)^2\right) + (\mathrm{tr}(A) - \mathrm{tr}(A \wedge_r B))^2 \; .$$

□

Proof of Theorem 2.5

We prove Theorem 2.5 in two parts: First, we prove the lower bound under the unimodality assumption. Then we show the bound for the weights with homogeneous imbalance.

Lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L})$ under unimodality

We need the following two Lemmas:

Lemma 2.16. *Given $A \in \mathbb{R}^{n\times h}, B \in \mathbb{R}^{h\times m}$, and $D = A^\top A - BB^\top \in \mathbb{R}^{h\times h}$. If $\mathrm{rank}(A) \leq r$ and $D \succeq 0$, then*

1. $\mathrm{rank}(B) \leq r$, *and* $\mathrm{rank}(D) \leq r$.
2. *There exists $Q \in \mathbb{R}^{h\times r}$ with $Q^\top Q = I_r$, such that*

$$AQQ^\top B = AB,\; AQQ^\top A^\top = AA^\top,\; B^\top QQ^\top B = B^\top B\,,$$

and $\lambda_i(Q^\top DQ) = \lambda_i(D),\; i = 1, \cdots, r$.

Lemma 2.17. *For $W_1 \in \mathbb{R}^{n\times h_1}, W_2 \in \mathbb{R}^{h_1\times h_2} \cdots, W_{L-1} \in \mathbb{R}^{h_{L-2}\times h_{L-1}}$ and $W_L \in \mathbb{R}^{h_{L-1}\times h_L}$ such that*

$$D_l = W_l^\top W_l - W_{l+1}W_{l+1}^\top \succeq 0\,, \quad l = 1, \cdots, L-1$$

we have

$$\lambda_n(W_1W_2\cdots W_{L-1}W_{L-1}^\top \cdots W_2^\top W_1^\top) \geq \prod_{i=1}^{L-1}\sum_{l=i}^{L-1}\lambda_n(D_l)\,.$$

Then we can prove the following:

Theorem 2.8. *For weights $\{W_l\}_{l=1}^L$ with unimodality index l^*, we have*

$$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \prod_{l=1}^{L-1}\tilde{d}_{(i)}\,. \tag{2.145}$$

Proof. Recall that

$$\mathcal{T}_{\{W_l\}_{l=1}^L}E = \sum_{l=1}^{L}\left(\prod_{i=1}^{l-1}W_i\right)\left(\prod_{i=1}^{l-1}W_i\right)^\top E\left(\prod_{i=l+1}^{L+1}W_i\right)^\top\left(\prod_{i=l+1}^{L+1}W_i\right)\,.$$

For simplicity, define p.s.d. operators

$$\mathcal{T}_lE := \left(\prod_{i=1}^{l-1}W_i\right)\left(\prod_{i=1}^{l-1}W_i\right)^\top E\left(\prod_{i=l+1}^{L+1}W_i\right)^\top\left(\prod_{i=l+1}^{L+1}W_i\right)\,,\ l = 1, \cdots, L$$

Then $\mathcal{T}_{\{W_l\}_{l=1}^L} = \sum_{l=1}^L \mathcal{T}_l$.

When $l^* = L$, we have, by Lemma 2.17,

$$\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \lambda_{\min}(\mathcal{T}_L) = \lambda_n(W_1\cdots W_{L-1}W_{L-1}^\top\cdots W_1^\top) \geq \prod_{i=1}^{L-1}\sum_{l=i}^{L-1}\lambda_n(D_l) = \prod_{l=1}^{L-1}\tilde{d}_{(i)}\,.$$

When $l^* = 1$, we have, again by Lemma 2.17,

$$\begin{aligned}\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \lambda_{\min}(\mathcal{T}_1) = \lambda_m(W_L^\top\cdots W_2^\top W_2\cdots W_L) \geq\ & \prod_{i=1}^{L-1}\sum_{l=i}^{L-1}\lambda_m(-D_{L-l})\\ =\ & \prod_{i=1}^{L-1}\sum_{l=1}^{L-i}\lambda_m(-D_l)\\ =\ & \prod_{i=1}^{L-1}\sum_{l=1}^{i}\lambda_m(-D_l) = \prod_{l=1}^{L-1}\tilde{d}_{(i)}\,.\end{aligned}$$

(To see Lemma 2.17 applies to the case $l^* = 1$, consider the following

$$W_L^\top \to W_1, \cdots, W_{L-l+1}^\top \to W_l, \cdots, W_1^\top \to W_L\,,$$

and this naturally leads to $-D_{L-l} \to D_l$. The expressions on the right-hand side of the arrow are those appearing in Lemma 2.17.)

Now for unimodality index $1 < l^* < L$, we have

$$\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \lambda_{\min}(\mathcal{T}_{l^*}) = \lambda_n(W_1 \cdots W_{l^*-1} W_{l^*-1}^\top \cdots W_1)\lambda_m(W_L^\top \cdots W_{l^*+1}^\top W_{l^*+1} \cdots W_L)\,.$$

Now apply Lemma 2.17 to both $\{W_1, \cdots, W_{l^*-1}, W_{l^*}\}$ and $\{W_L^\top, \cdots, W_{l^*+1}^\top, W_{l^*}^\top\}$, we have

$$\lambda_n(W_1 \cdots W_{l^*-1} W_{l^*-1}^\top \cdots W_1) \geq \prod_{i=1}^{l^*-1} \sum_{l=i}^{l^*-1} \lambda_n(D_l) = \prod_{i=1}^{l^*-1} \tilde{d}_{(i)}\,, \tag{2.146}$$

and

$$\begin{aligned}\lambda_m(W_L^\top \cdots W_{l^*+1}^\top W_{l^*+1} \cdots W_L) &\geq \prod_{i=1}^{L-l^*} \sum_{l=i}^{L-l^*} \lambda_m(-D_{L-l}) \\ &= \prod_{i=1}^{L-l^*} \sum_{l=l^*}^{L-i} \lambda_m(-D_l) \\ &= \prod_{i=l^*}^{L-1} \sum_{l=l^*}^{i} \lambda_m(-D_l) = \prod_{i=l^*}^{L-1} \tilde{d}_{(i)}\,. \end{aligned} \tag{2.147}$$

Combining (2.146) and (2.147), we have

$$\lambda_n(W_1 \cdots W_{l^*-1} W_{l^*-1}^\top \cdots W_1)\lambda_m(W_L^\top \cdots W_{l^*+1}^\top W_{l^*+1} \cdots W_L) \geq \prod_{i=1}^{L-1} \tilde{d}_{(i)}\,, \tag{2.148}$$

which leads to $\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \prod_{i=1}^{L-1} \tilde{d}_{(i)}$. The proof is complete as we have shown $\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \prod_{i=1}^{L-1} \tilde{d}_{(i)}$ for any unimodality index $l^* \in [L]$. □

Lower bound on $\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L})$ under homogeneous imbalance

We need the following Lemma:

Lemma 2.18. *Given any set of scalars $\{w_l\}_{l=1}^L$ such that $d_{(i)} := w_i^2 - w_L^2 \geq 0, i = 1, \cdots, L-1$, we have*

$$\sum_{l=1}^L \prod_{i\neq l} w_i^2 = \sum_{l=1}^L \frac{w^2}{w_l^2} \geq \sqrt{\left(\prod_{i=1}^{L-1} d_{(i)}\right)^2 + \left(Lw^{2-2/L}\right)^2}\,, \tag{2.149}$$

where $w = \prod_{l=1}^L w_l$.

Then we can prove the following:

Theorem 2.9. *For weights $\{W_l\}_{l=1}^L$ with homogeneous imbalance, we have*

$$\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) \geq \sqrt{\left(\prod_{l=1}^{L-1} \tilde{d}_{(i)}\right)^2 + \left(L\sigma_{\min}^{2-2/L}(W)\right)^2}, \quad W = \prod_{l=1}^L W_l\,. \tag{2.150}$$

Proof. When all imbalance matrices are zero matrices, this is the balanced case [27] and $\lambda_{\min}\left(\mathcal{T}_{\{W_l\}_{l=1}^L}\right) = L\sigma_{\min}^{2-2/L}(W)$. Here we only prove the case when some $d_l \neq 0$.

Notice that given the homogeneous imbalance constraint

$$W_l^\top W_l - W_{l+1}W_{l+1}^\top = d_l I\,,$$

$W_l^\top W_l$ and $W_{l+1}W_{l+1}^\top$ must be co-diagonalizable: If we have $Q^\top Q = I$ such that $Q^\top W_l^\top W_l Q$ is diagonal, then $Q^\top W_{l+1}W_{l+1}^\top Q$ must be diagonal as well from the fact that $Q^\top W_l^\top W_l Q - Q^\top W_{l+1}W_{l+1}^\top Q = d_l I$.

Moreover, if the diagonal entries of $Q^\top W_l^\top W_l Q$ are in decreasing order, then so are those of $Q^\top W_{l+1}W_{l+1}^\top Q$ because the latter is the shifted version of the former by d_l.

This suggests that all $W_l, l = 1, \cdots, L$ have the same rank and one has the following decomposition of the weights:

$$W_l = Q_{l-1}\Sigma_l Q_l^\top\,, \tag{2.151}$$

Here, $\Sigma_l, l = 1, \cdots, L$ are diagonal matrix of size $k = \min\{n, m\}$ whose entries are in decreasing order. And $Q_l \in \mathbb{R}^{h_l \times \min\{n,m\}}$ with $Q_l^\top Q_l = I$. ($h_0 = n, h_L = m$). From such decomposition, we have

$$W = W_1 \cdots W_L = Q_0 \Sigma_1 Q_1^\top Q_1 \Sigma_2 Q_2^\top \cdots Q_{L-1} \Sigma_L Q_L^\top = Q_0 \left(\prod_{l=1}^L \Sigma_l \right) Q_L^\top \,, \tag{2.152}$$

thus

$$\sigma_{\min}(W) = \prod_{l=1}^L \lambda_{\min}(\Sigma_l) \,. \tag{2.153}$$

Regarding the imbalance, we have

$$Q_l^\top (W_l^\top W_l - W_{l+1} W_{l+1}^\top) Q_l = d_l I \quad \Rightarrow \quad \Sigma_l^2 - \Sigma_{l+1}^2 = d_l I \,, \tag{2.154}$$

which suggests that

$$\lambda_{\min}^2(\Sigma_l) - \lambda_{\min}^2(\Sigma_{l+1}) = d_l, l = 1, \cdots, L-1 \,. \tag{2.155}$$

Now consider the set of scalars $\{w_l\}_{l=1}^L$:

$$\begin{aligned} & w_l = \lambda_{\min}(\Sigma_l), l = 1, \cdots, l^* - 1 \\ & w_l = \lambda_{\min}(\Sigma_{l+1}), l = l^*, \cdots, L-1 \\ & w_L = \lambda_{\min}(\Sigma_{l^*}) \,. \end{aligned}$$

Then $\{w_l\}_{l=1}^L$ satisfy the assumption in Lemma 2.18:

$$w_i^2 - w_L^2 = \tilde{d}_{(i)} \geq 0, i = 1, \cdots, L-1 \,, \tag{2.156}$$

where $\tilde{d}_{(i)}$ is precisely the cumulative imbalance. Then Lemma 2.18 gives ((2.153) is also used here)

$$\sum_{l=1}^L \prod_{i \neq l} w_i^2 \geq \sqrt{\left(\prod_{i=1}^{L-1} \tilde{d}_{(i)} \right)^2 + \left(L \sigma_{\min}^{2-2/L}(W) \right)^2} \tag{2.157}$$

Recall that

$$\mathcal{T}_{\{W_l\}_{l=1}^L} E = \sum_{l=1}^L \left(\prod_{i=0}^{l-1} W_i \right) \left(\prod_{i=0}^{l-1} W_i \right)^\top E \left(\prod_{i=l+1}^{L+1} W_i \right)^\top \left(\prod_{i=l+1}^{L+1} W_i \right) .$$

For simplicity, define p.s.d. operators

$$\mathcal{T}_l E := \left(\prod_{i=0}^{l-1} W_i\right)\left(\prod_{i=0}^{l-1} W_i\right)^\top E \left(\prod_{i=l+1}^{L+1} W_i\right)^\top \left(\prod_{i=l+1}^{L+1} W_i\right)\,,\ l = 1, \cdots, L$$

Then $\mathcal{T}_{\{W_l\}_{l=1}^L} = \sum_{l=1}^L \mathcal{T}_l$.

Notice that the summand $\prod_{i\neq l} w_i^2$ exactly corresponds to one of $\lambda_{\min}(T_l)$. For example,

$$\lambda_{\min}(\mathcal{T}_1) = \lambda_{\min}(W_L^\top \cdots W_2^\top W_2 \cdots W_L) = \lambda_{\min}\left(Q_L^\top \left(\prod_{l=2}^L \Sigma_l^2\right) Q_L\right) = \prod_{i\neq 1} w_i^2\,. \tag{2.158}$$

More precisely, we have

$$\begin{aligned}
\lambda_{\min}(\mathcal{T}_l) &= \prod_{i\neq l} w_i^2, \quad l < l^* \\
\lambda_{\min}(\mathcal{T}_l) &= \prod_{i\neq l-1} w_i^2, \quad l > l^* \\
\lambda_{\min}(\mathcal{T}_l) &= \prod_{i\neq L} w_i^2, \quad l = l^*\,.
\end{aligned}$$

Therefore, we finally have

$$\lambda_{\min}(\mathcal{T}_{\{W_l\}_{l=1}^L}) \geq \sum_{l=1}^L \lambda_{\min}(\mathcal{T}_l) = \sum_{l=1}^L \prod_{i\neq l} w_i^2 \geq \sqrt{\left(\prod_{i=1}^{L-1} \tilde{d}_{(i)}\right)^2 + \left(L\sigma_{\min}^{2-2/L}(W)\right)^2}\,. \tag{2.159}$$

□

Proofs of auxiliary lemmas

Proof of Lemma 2.17. The proof is rather simple when $n = h_1 = h_2 = \cdots = h_{L-1}$: Notice that

$$\begin{aligned}
&\lambda_n(W_1 W_2 \cdots W_{L-1} W_{L-1}^\top \cdots W_2^\top W_1^\top) \\
\geq\ & \lambda_n(W_{L-1} W_{L-1}^\top) \cdot \lambda_n(W_1 W_2 \cdots W_{L-2} W_{L-2}^\top \cdots W_2^\top W_1^\top) \\
\geq\ & \lambda_n(W_{L-1} W_{L-1}^\top) \cdot \lambda_n(W_{L-2} W_{L-2}^\top) \cdot \lambda_n(W_1 W_2 \cdots W_{L-3} W_{L-3}^\top \cdots W_2^\top W_1^\top) \\
&\cdots \\
\geq\ & \prod_{i=1}^{L-1} \lambda_n(W_i W_i^\top)\,.
\end{aligned}$$

Then it remains to show that $\lambda_n(W_i W_i^\top) \geq \sum_{l=i}^{L-1} \lambda_n(D_l)$ for $i = 1, \cdots, L-1$.

Suppose $\lambda_n(W_k W_k^\top) \geq \sum_{l=k}^{L-1} \lambda_l(D)$ for some $k \in [L-1]$, then we have

$$\begin{aligned}
\lambda_n(W_{k-1} W_{k-1}^\top) &= \lambda_n(W_{k-1}^\top W_{k-1}) \\
&= \lambda_n(W_k W_k^\top + D_{k-1}) \\
&\geq \lambda_n(W_k W_k^\top) + \lambda_n(D_{k-1}) \\
&\geq \sum_{l=k}^{L-1} \lambda_n(D_l) + \lambda_n(D_{k-1}) = \sum_{l=k-1}^{L-1} \lambda_n(D_l)\,.
\end{aligned}$$

Therefore, we only need to show $\lambda_n(W_{L-1} W_{L-1}^\top) \geq \lambda_n(D_{L-1})$ then the rest follows by the induction above. Indeed

$$\lambda_n(W_{L-1} W_{L-1}^\top) = \lambda_n(W_{L-1}^\top W_{L-1}) = \lambda_n(W_L W_L^\top + D_{L-1}) \geq \lambda_n(D_{L-1})\,,$$

which finishes the proof for the case of $n = h_1 = h_2 = \cdots = h_{L-1}$.

When the above assumptions does not hold, Lemma 2.16 allows us to related the set of weights $\{W_l\}_{l=1}^L$ to the one $\{\tilde{W}_l\}_{l=1}^L$ that satisfy the equal dimension assumption. More specifically, apply Lemma 2.16 using each imbalance constraint

$$D_l = W_l^\top W_l - W_{l+1} W_{l+1}^\top \succeq 0\,, \quad l = 1, \cdots, L-1\,,$$

to obtain a $Q_l \in \mathbb{R}^{h_l \times n}$ that has all the property in Lemma (2.16). Use these $Q_l, l = 1, \cdots, L-1$ to define

$$\begin{aligned} \tilde{W}_l &= Q_{l-1}^\top W_l Q_l \,, l = 1, \cdots, L, \\ \tilde{D}_l &= \tilde{W}_l^\top \tilde{W}_l - \tilde{W}_{l+1}^\top \tilde{W}_{l+1} \,, l = 1, \cdots, L-1 \,, \end{aligned}$$

where $Q_0 = I, Q_L = I$. Now $\{\tilde{W}_l\}_{l=1}^L$ satisfies the assumption that $n = h_1 = \cdots = h_{L-1}$, then

$$\lambda_n(\tilde{W}_1 \tilde{W}_2 \cdots \tilde{W}_{L-1} \tilde{W}_{L-1}^\top \cdots \tilde{W}_2^\top \tilde{W}_1^\top) \geq \prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(\tilde{D}_l) \,. \tag{2.160}$$

Using the properties of $Q_l \in \mathbb{R}^{h_l \times n}, l = 1, \cdots, L-1$, we have

$$\begin{aligned} &\lambda_n(\tilde{W}_1 \tilde{W}_2 \cdots \tilde{W}_{L-1} \tilde{W}_{L-1}^\top \cdots \tilde{W}_2^\top \tilde{W}_1^\top) \\ =\, &\lambda_n(W_1 Q_1 Q_1^\top W_2 Q_2 \cdots Q_{L-2}^\top W_{L-1} Q_{L-1} Q_{L-1}^\top W_{L-1}^\top Q_{L-2}^\top \cdots Q_2^\top W_2^\top Q_1 Q_1^\top W_1^\top) \\ =\, &\lambda_n(W_1 W_2 \cdots W_{L-1} W_{L-1}^\top \cdots W_2^\top W_1^\top) \,, \end{aligned}$$

and

$$\prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(\tilde{D}_l) = \prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(Q_l^\top D_l Q_l) = \prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(D_l) \,.$$

Therefore, (2.160) is exactly

$$\lambda_n(W_1 W_2 \cdots W_{L-1} W_{L-1}^\top \cdots W_2^\top W_1^\top) \geq \prod_{i=1}^{L-1} \sum_{l=i}^{L-1} \lambda_n(D_l) \,. \tag{2.161}$$

□

Proof of Lemma 2.16. Since $\operatorname{rank}(A) \leq r$, A has a compact SVD $A = P \Sigma_A Q^\top$ such that $Q \in \mathbb{R}^{h \times r}$ and $Q^\top Q = I_r$.

This is exactly Q we are looking for. Let $Q_\perp Q_\perp^\top = I_h - QQ^\top$ be the projection onto the subspace orthogonal to the columns of Q. Then

$$\begin{aligned} D = A^\top A - BB^\top \;&\Rightarrow Q_\perp^\top D Q_\perp = Q_\perp^\top A^\top A Q_\perp - Q_\perp^\top B B^\top Q_\perp \\ &\Rightarrow Q_\perp^\top D Q_\perp + Q_\perp^\top B B^\top Q_\perp = 0 \,. \end{aligned}$$

$Q_\perp^\top D Q_\perp$ and $Q_\perp^\top BB^\top Q_\perp$ are two p.s.d. matrices whose sum is zero, which implies

$$Q_\perp^\top D Q_\perp = 0, \quad DQ_\perp = 0, \quad Q_\perp^\top BB^\top Q_\perp = 0, \quad B^\top Q_\perp = 0\,.$$

$Q_\perp^\top D Q_\perp = 0$ shows that the nullspace of D has at least dimension $h - r$, i.e., $\mathrm{rank}(D) \le r$.

Moreover

$$\begin{aligned} AQQ^\top B &= A(I_h - Q_\perp Q_\perp^\top)B = AB \\ AQQ^\top A^\top &= A(I_h - Q_\perp Q_\perp^\top)A^\top = AA^\top \\ B^\top QQ^\top B &= B^\top(I_h - Q_\perp Q_\perp^\top)B = B^\top B \end{aligned}$$

The last equality $B^\top B = B^\top QQ^\top B$ shows that $\mathrm{rank}(B) \le r$.

Lastly, we have, for $i = 1, \cdots, r$,

$$\lambda_i(Q^\top DQ) = \lambda_i(QQ^\top D) = \lambda_i((I_h - Q_\perp Q_\perp^\top)D) = \lambda_i(D)\,.$$

□

Before proving Lemma 2.18, we state a Lemma that will be used in the proof.

Lemma 2.19. *Given positive $x_i, i = 1, \cdots, n$, we have*

$$\sum_{i=1}^n x_i \ge n\left(\prod_{i=1}^n x_i\right)^{1/n}\,.$$

Proof. This is from the fact that arithmetic mean of $\{x_i\}_{i=1}^n$ is greater than the geometric mean of $\{x_i\}_{i=1}^n$. □

We are ready to prove Lemma 2.18.

Proof of Lemma 2.18. We denote

$$\tau_{\{w_l\}_{i=1}^L} := \sum_{l=1}^L \prod_{i\neq l} w_i^2 \tag{2.162}$$

Notice that $w_i^2 = w_L^2 + \sum_{j=i}^{L-1}(w_j^2 - w_{j+1}^2) = w_L^2 + d_{(i)}$. Let $d_{(L)} = 0$, we write the expression for τ as

$$\tau_{\{w_l\}_{i=1}^L} = \sum_{l=1}^L \prod_{i \neq l} w_i^2 = \sum_{l=1}^L \prod_{i \neq l} \left(w_L^2 + d_{(i)}\right) := \tau(w_L^2; \{d_{(i)}\}_{i=1}^{L-1}) \,.$$

Therefore, when fixing $\{d_{(i)}\}_{i=1}^{L-1}$, τ can be viewed as a function of w_L^2.

When $w = 0$: one of w_l must be zero, and because w_L^2 has the least value among all the weights, we know $w_L^2 = 0$. Then

$$\tau_{\{w_l\}_{i=1}^L} = \tau(0; \{d_{(i)}\}_{i=1}^{L-1}) = \prod_{i=1}^{L-1} d_{(i)} \,,$$

i.e. we actually have equality when $w = 0$.

When $w \neq 0$: then $w^2 \neq 0$ and we write

$$w^2 = \prod_{l=1}^L w_l^2 = w_L^2 \prod_{l=1}^{L-1} \left(w_L^2 + d_{(l)}\right) := p(w_L^2; \{d_{(i)}\}_{i=1}^{L-1}) \,,$$

which shows w^2 is a function of w_L^2 when $\{d_{(i)}\}_{i=1}^{L-1}$ are fixed. Here we use p to denote w^2 for simplicity. Moreover, function p: $\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ has differentiable inverse p^{-1} as long as $p > 0$, because

$$\frac{dp}{dw_L^2} = \sum_{l=1}^L \prod_{i \neq l} \left(w_L^2 + d_{(i)}\right) = \sum_{l=1}^L \prod_{i \neq l} w_i^2 \overset{\text{(Lemma 2.19)}}{\geq} L\left(p^{L-1}\right)^{1/L} > 0 \,,$$

and inverse function theorem [76] shows the existence of differentiable inverse. Whenever, p^{-1} exists, it derivative is

$$\frac{dw_L^2}{dp} = \left(\sum_{l=1}^L \prod_{i \neq l} \left(w_L^2 + d_{(i)}\right)\right)^{-1} = \tau^{-1} \,.$$

Now pick any $0 < p_0 \leq w^2$ we have, by Fundamental Theorem of Calculus,

$$\begin{aligned} \tau^2_{\{w_l\}_{l=1}^L} &= \tau^2(p^{-1}(w^2); \{d_{(i)}\}_{i=1}^{L-1}) \\ &= \tau^2(p^{-1}(p_0); \{d_{(i)}\}_{i=1}^{L-1}) + \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} \frac{d}{dw_L^2} \tau^2(w_L^2; \{d_{(i)}\}_{i=1}^{L-1}) dw_L^2 \end{aligned}$$

For the first part, we have

$$\begin{aligned}
&\tau^2(p^{-1}(p_0); \{d_{(i)}\}_{i=1}^{L-1}) \\
=& \left(\sum_{l=1}^{L} \prod_{i\neq l} \left(p^{-1}(p_0) + d_{(i)}\right) \right)^2 \geq \left(\prod_{i\neq L} \left(p^{-1}(p_0) + d_{(i)}\right) \right)^2 \geq \left(\prod_{i=1}^{L-1} d_{(i)} \right)^2 ,
\end{aligned}$$

and for the second part, we have

$$\begin{aligned}
& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} \frac{d}{dw_L^2} \tau^2 dw_L^2 \\
=& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau \frac{d}{dw_L^2} \tau dw_L^2 \\
=& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau \sum_{l=1}^{L} \sum_{i\neq l} \prod_{j\neq i, j\neq l} (w_L^2 + d_{(j)}) dw_L^2 \\
=& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau \sum_{l=1}^{L} \sum_{i\neq l} \frac{p}{w_i^2 w_l^2} dw_L^2 \\
\text{(Lemma 2.19)} \geq & \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau L(L-1) \left(\prod_{l=1}^{L} \prod_{i\neq l} \frac{p}{w_i^2 w_l^2} \right)^{\frac{1}{L(L-1)}} dw_L^2 \\
=& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau L(L-1) \left(\frac{p^{L(L-1)}}{p^{2L-2}} \right)^{\frac{1}{L(L-1)}} dw_L^2 \\
=& \int_{p^{-1}(p_0)}^{p^{-1}(w^2)} 2\tau L(L-1) p^{1-2/L} dw_L^2 \\
(dw_L^2 = \tau^{-1} dp) =& \int_{p_0}^{w^2} 2L(L-1) p^{1-2/L} dp = L^2 p^{2-2/L} \Big|_{p_0}^{w^2} = \left(L w^{2-2/L} \right)^2 - L^2 p_0^{2-2/L} .
\end{aligned}$$

Overall, for any $0 < p_0 \leq w^2$, we have

$$\tau^2_{\{w_l\}_{l=1}^L} \geq \left(\prod_{i=1}^{L-1} d_{(i)} \right)^2 + \left(L w^{2-2/L} \right)^2 - L^2 p_0^{2-2/L} .$$

Let $p_0 \to 0$, we have $\tau^2 \geq \left(\prod_{i=1}^{L-1} d_{(i)} \right)^2 + \left(L w^{2-2/L} \right)^2$, i.e.

$$\tau \geq \sqrt{ \left(\prod_{i=1}^{L-1} d_{(i)} \right)^2 + (L w^{2-2/L})^2 } .$$

□

2.3 Two-layer ReLU networks

In this section, we provide a complete analysis of the dynamics of gradient flow for the problem of training a two-layer ReLU network on well-separated data under the assumption of small initialization. Specifically, we show that if the initialization is sufficiently small, during the early phase of training the neurons in the first layer try to align with either the positive data or the negative data, depending on its corresponding weight on the second layer. Moreover, through a careful analysis of the neuron's directional dynamics we show that the time it takes for all neurons to achieve good alignment with the input data is upper bounded by $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$, where n is the number of data points and μ measures how well the data are separated. We also show that after the early alignment phase the loss converges to zero at a $\mathcal{O}(\frac{1}{t})$ rate and that the weight matrix on the first layer is approximately low-rank.

2.3.1 Preliminaries

We first discuss the problem setup. We then present some key ingredients for analyzing the training dynamics of ReLU networks under small initialization, and discuss some of the weaknesses/issues from prior work.

Problem setting

We are interested in a binary classification problem with dataset $[x_1, \cdots, x_n] \in \mathbb{R}^{D\times n}$ (input data) and $[y_1, \cdots, y_n]^\top \in \{-1, +1\}^n$ (labels). For the classifier, $f : \mathbb{R}^D \to \mathbb{R}$, we consider a two-layer ReLU network:

$$f(x; W, v) = v^\top \sigma(W^\top x) = \sum\nolimits_{j=1}^{h} v_j \sigma(w_j^\top x)\,, \tag{2.163}$$

parametrized by network weights $W := [w_1, \cdots, w_h] \in \mathbb{R}^{D\times h}, v := [v_1, \cdots, v_h]^\top \in \mathbb{R}^{h\times 1}$, where $\sigma(\cdot) = \max\{\cdot, 0\}$ is the ReLU activation function. We aim to find the network weights that minimize the training loss $\mathcal{L}(W, v) = \sum_{i=1}^{n} \ell(y_i, f(x_i; W, v))$, where $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$. The network is trained via the gradient flow (GF) dynamics

$$\dot{W} \in \partial_W \mathcal{L}(W, v),\ \dot{v} \in \partial_v \mathcal{L}(W, v), \tag{2.164}$$

where $\partial_W \mathcal{L}, \partial_v \mathcal{L}$ are Clark sub-differentials of $\mathcal{L}$. Therefore, (2.164) is a differential inclusion [77]. The work of [16] shows that there exist global solutions to (2.164) if one uses $\sigma'(x) = \mathbb{1}_{x>0}$ as the ReLU subgradient. Therefore, we follow this choice of subgradient for our analysis.

To initialize the weights, we consider the following initialization scheme. First, we start from a weight matrix $W_0 \in \mathbb{R}^{D \times h}$, and then and then initialize the weights as

$$W(0) = \epsilon W_0, \quad v_j(0) \in \{\|w_j(0)\|, -\|w_j(0)\|\}, \forall j \in [h]\,. \tag{2.165}$$

That is, the weight matrix W_0 determines the initial shape of the first-layer weights $W(0)$ and we use ϵ to control the initialization scale and we are interested in the regime where ϵ is sufficiently small. For the second layer weights $v(0)$, each $v_j(0)$ has magnitude $\|w_j(0)\|$ and we only need to decide its sign. Our results in later sections are stated for a deterministic choice of ϵ, W_0, and $v(0)$, then we comment on the case where W_0 is chosen randomly via some distribution.

The resulting weights in (2.165) are always "balanced", i.e., $v_j^2(0) - \|w_j(0)\|^2 = 0, \forall j \in [h]$, because $v_j(0)$ can only take two values: either $\|w_j(0)\|$ or $-\|w_j(0)\|$. More importantly, under GF (2.164), this balancedness is preserved [61]: $v_j^2(t) - \|w_j(t)\|^2 = 0, \forall t \geq 0, \forall j \in [h]$. In addition, it is shown in [16] that $\text{sign}(v_j(t)) = \text{sign}(v_j(0)), \forall t \geq 0, \forall j \in [h]$, and the dynamical behaviors of neurons will be divided into two types, depending on $\text{sign}(v_j(0))$.

Remark 3. *For our theoretical results, the balancedness condition is assumed for technical purposes: it simplifies the dynamics of GF and thus the analysis. It is a common assumption for many existing works on both linear [27] and nonlinear [78, 16] neural networks. For the* *experiments in Section 2.3.3, we use a standard Gaussian initialization with small variance, which is not balanced.*

Remark 4. *Without loss of generality, we consider the case where all columns of W_0 are nonzero, i.e., $\|w_j(0)\| > 0, \forall j \in [h]$. We make this assumption because whenever $w_j(0) = 0$, we also have $v_j(0) = 0$ from the balancedness, which together would imply $\dot{v}_j \equiv 0, \dot{w}_j \equiv 0$ under gradient flow. As a result, w_j and v_j would remain zero and thus they could be ignored in the convergence analysis.*

Remark 5. *Our main results will depend on both $\max_j \|w_j(0)\|$ and $\min_j \|w_j(0)\|$, as shown in our proofs in Appendices 2.3.3 and 2.3.3. Therefore, whenever we speak of small initialization, we will say that ϵ is small without worrying about the scale of W_0, which is already considered in our results.*

Neural alignment with small initialization: an overview

Prior work argues that the gradient flow dynamics (2.164) under small initialization (2.165), i.e., when ϵ is sufficiently small, can be roughly described as "align then fit" [79, 16]: During the early phase of training, every neuron $w_j, j \in [h]$ keeps a small norm $\|w_j\|^2 \ll 1$ while changing their directions $\frac{w_j}{\|w_j\|}$ significantly in order to locally maximize a "signed coverage" [79] of itself w.r.t. the training data. After the alignment phase, part of the neurons (potentially all neurons) start to grow their norms in order to fit the training data, and the loss decreases significantly. The analysis for the fitting phase generally depends on the resulting neuron directions at the end of the alignment phase [78, 16]. However, prior analysis of the alignment phase either is based on a vanishing initialization argument that can not be directly translated into the case finite but small initialization [79] or assumes some stringent assumption on the data [16]. In this section, we provide a brief overview of the existing analysis for neural alignment and then point out several weaknesses in prior work.

Prior analysis of the alignment phase: Since during the alignment phase all neurons have small norm, prior work mainly focuses on the directional dynamics, i.e., $\frac{d}{dt}\frac{w_j}{\|w_j\|}$, of the neurons. The analysis relies on the following approximation of the dynamics of every neuron $w_j, j \in [h]$:

$$\frac{d}{dt}\frac{w_j}{\|w_j\|} \simeq \operatorname{sign}(v_j(0))\mathcal{P}_{w_j(t)}x_a(w_j)\,, \quad (2.166)$$

where $\mathcal{P}_w = I - \frac{ww^\top}{\|w\|^2}$ is the projection onto the subspace orthogonal to w and

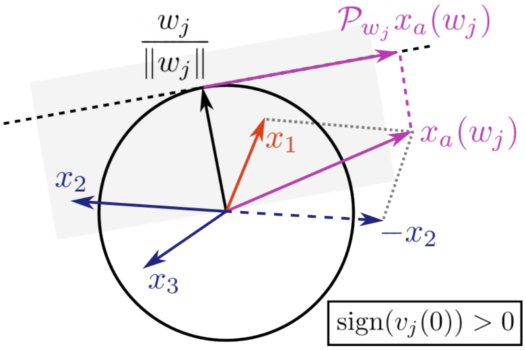
$\frac{w_j}{\|w_j\|}$
$\mathcal{P}_{w_j}x_a(w_j)$
x_1
$x_a(w_j)$
x_2
$-x_2$
x_3
$\operatorname{sign}(v_j(0)) > 0$

Figure 2-7. Illustration of $\frac{d}{dt}\frac{w_j}{\|w_j\|}$ during the early alignment phase. x_1 has $+1$ label, and x_2, x_3 have -1 labels, x_1, x_2 lie inside the halfspace $\langle x, w_j\rangle > 0$ (gray shaded), thus $x_a(w_j) = x_1 - x_2$. Since $\operatorname{sign}(v_j(0)) > 0$, GF pushes w_j towards $x_a(w_j)$.

$$x_a(w) := \sum\nolimits_{i:\langle x_i, w\rangle > 0} y_i x_i \quad (2.167)$$

denotes the signed combination of the data points activated by w.

First of all, (2.166) implies that the dynamics $\frac{w_j}{\|w_j\|}$ are approximately decoupled, and thus one can study each $\frac{w_j}{\|w_j\|}$ separately. Moreover, as illustrated in Figure 2-7, if $\operatorname{sign}(v_j(0)) > 0$, the flow (2.166) pushes w_j towards $x_a(w_j)$, since w_j is attracted by its currently activated positive data and repelled by its currently activated negative data. Intuitively, during the alignment phase, a neuron w_j with $\operatorname{sign}(v_j(0)) > 0$ would try to find a direction where it can activate as much positive data and as less negative data as possible. If $\operatorname{sign}(v_j(0)) < 0$, the opposite holds.

Indeed, [79] claims that the neuron w_j would be aligned with some "extreme vectors," defined as vector $w \in \mathbb{S}^{D-1}$ that locally maximizes $\sum_{i\in[n]} y_i\sigma(\langle x_i, w\rangle)$ (similarly, w_j with $\operatorname{sign}(v_j(0)) < 0$ would be aligned with the local minimizer), and there are only finitely many such vectors; thus the neurons are expected to converge to one of these extreme vectors in direction. The analysis is done by taking the limit $\epsilon \to 0$ on the initialization scale, under which the approximation in (2.166) is exact.

Weakness in prior analyses: Although [79] provides great insights into the dynamical behavior of the neurons in the alignment phase, the validity of the aforementioned approximation for finite but small ϵ remains in question. First, one needs to make sure that the error $\left\|\frac{d}{dt}\frac{w_j}{\|w_j\|} - \mathrm{sign}(v_j(0))\mathcal{P}_{w_j}x_a(w_j)\right\|$ is sufficiently small when ϵ is finite in order to justify (2.166) as a good approximation. Second, the error bound needs to hold for the entire alignment phase. [79] assumes $\epsilon \to 0$; hence there is no formal error bound. In addition, prior analyses on small initialization [13, 16] suggest the alignment phase only holds for $\Theta(\log\frac{1}{\epsilon})$ time. Thus, the claim in [79] would only hold if good alignment is achieved before the alignment phase ends. However, [79] provides no upper bound on the time it takes to achieve good alignment. Therefore, without a finite ϵ analysis, [79] fails to fully explain the training dynamics under small initialization. Understanding the alignment phase with finite ϵ requires additional analytical tools from dynamical systems theory. To the best of our knowledge, this has only been studied under a stringent assumption that all data points are orthogonal to each other [16].

Goal of this section: In this section, we want to address some of the aforementioned issues by developing a formal analysis for the early alignment phase with a finite but small initialization scale ϵ. We first discuss our main theorem that shows that a directional convergence can be achieved within bounded time under data assumptions that are less restrictive and have more practical relevance. Then, we discuss the error bound for justifying (2.166) in the proof sketch for the main theorem.

2.3.2 Convergence with small initialization

In this section, we present our main results, which require the following assumption on the training data (we will compare our assumption with those in prior work after the main theorem):

Assumption 2.2. *Any pair of input data with the same label are positively correlated, and any pair of inputs with different labels are negatively correlated, i.e.,*

$$\min_{i,j} \frac{\langle x_i y_i, x_j y_j \rangle}{\|x_i\|\|x_j\|} := \mu > 0. \tag{2.168}$$

Given a training dataset, we define $\mathcal{S}_+ := \{z \in \mathbb{R}^D : \mathbb{1}_{\langle x_i, z\rangle > 0} = \mathbb{1}_{y_i > 0}, \forall i\}$ to be the cone in $\mathbb{R}^n$ such that whenever neuron $w \in \mathcal{S}_+$, w is activated exclusively by every x_i with a positive label (see Figure 2-8). Similarly, for x_i with negative labels, we define $\mathcal{S}_- := \{z \in \mathbb{R}^D : \mathbb{1}_{\langle x_i, z\rangle > 0} = \mathbb{1}_{y_i < 0}, \forall i\}$. Finally, we define $\mathcal{S}_{\text{dead}} := \{z \in \mathbb{R}^D : \langle z, x_i \rangle \leq 0, \forall i\}$ to be the cone such that whenever $w \in \mathcal{S}_{\text{dead}}$, no data activates w. Given Assumption 2.2, it can be shown (see Appendix 2.3.3) that $\mathcal{S}_+$ ($\mathcal{S}_-$) is a non-empty, convex cone that contains all positive data $x_i, i \in \mathcal{I}_+$ (negative data $x_i, i \in \mathcal{I}_-$). $\mathcal{S}_{\text{dead}}$ is a convex cone as well, but not necessarily non-empty. We illustrate these cones in Figure 2-8 given some training data (red solid arrow denotes positive data and blue denotes negative ones).

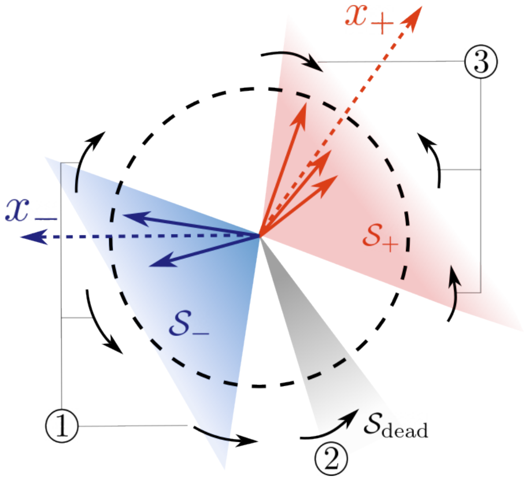

Figure 2-8. Neuron alignment under data that satisfies Assumption 2.2. For neurons in $\mathcal{V}_+$, ① if it lies inside $\mathcal{S}_-$, then it gets repelled by x_- and eventually escapes $\mathcal{S}_-$; Once it is outside $\mathcal{S}_-$, it may ② get continuously repelled by some negative data and eventually enters $\mathcal{S}_{\text{dead}}$. or ③ gain some activation on positive data and eventually enter $\mathcal{S}_+$, after which it gets constantly attracted by x_+.

Moreover, given some initialization from (2.165), we define $\mathcal{I}_+ := \{i \in [n] : y_i > 0\}$ to be the set of indices of positive data, and $\mathcal{I}_- := \{i \in [n] : y_i < 0\}$ for negative data. We also define $\mathcal{V}_+ := \{j \in [h] : \operatorname{sign}(v_j(t)) > 0\}$ to be the set of indices of neurons with positive second-layer entry and $\mathcal{V}_- := \{j \in [h] : \operatorname{sign}(v_j(t)) < 0\}$ for neurons with negative second-layer entry. Note that, as discussed in previous section, $\operatorname{sign}(v_j(t))$ does not change under balanced initialization, thus $\mathcal{V}_+, \mathcal{V}_-$ are time invariant. Further, as we discussed in Section 2.3.1 about the early alignment phase, we expect that every neuron in $\mathcal{V}_+$ will drift toward the region where positive data concentrate and thus eventually reach $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$, as visualized in Figure 2-8 (x_+, x_- shown in the figure are defined in Assumption 2.3). Similarly, all neurons in $\mathcal{V}_-$ would chase after negative data and thus reach $\mathcal{S}_-$ or $\mathcal{S}_{\text{dead}}$. Our theorem precisely characterizes this behavior.

Main results

Before we present our main theorem, we need the following assumption on the initialization, mostly for technical reasons.

Assumption 2.3. *The initialization from* (2.165) *satisfies that* $\max_{j\in\mathcal{V}_+}\langle\frac{w_j(0)}{\|w_j(0)\|},\frac{x_-}{\|x_-\|}\rangle < 1$, *and* $\max_{j\in\mathcal{V}_-}\langle\frac{w_j(0)}{\|w_j(0)\|},\frac{x_+}{\|x_+\|}\rangle < 1$, *where* $x_+ = \sum_{i\in\mathcal{I}_+} x_i$ *and* $x_- = \sum_{i\in\mathcal{I}_-} x_i$.

Assumption (2.3) essentially asks the neuron $w_j(0), j \in \mathcal{V}_+$ (or $w_j(0), j \in \mathcal{V}_-$, resp.) to not be completely aligned with x_+ (or x_-, resp.). We are now ready to present our main result (given Assumption 2.2 and Assumption 2.3):

Theorem 2.10. *Given some initialization from* (2.165), *if* $\epsilon = \mathcal{O}(\frac{1}{\sqrt{h}}\exp(-\frac{n}{\sqrt{\mu}}\log n))$, *then any solution to the gradient flow dynamics* (2.164) *satisfies*

1. *(Directional convergence in early alignment phase)* $\exists t_1 = \mathcal{O}(\frac{\log n}{\sqrt{\mu}})$, *such that*

   - $\forall j \in \mathcal{V}_+$, *either* $w_j(t_1) \in \mathcal{S}_+$ *or* $w_j(t_1) \in \mathcal{S}_{\textit{dead}}$. *Moreover, if* $\max_{i\in\mathcal{I}_+}\langle w_j(0), x_i\rangle > 0$, *then* $w_j(t_1) \in \mathcal{S}_+$.

   - $\forall j \in \mathcal{V}_-$, *either* $w_j(t_1) \in \mathcal{S}_-$ *or* $w_j(t_1) \in \mathcal{S}_{\textit{dead}}$. *Moreover, if* $\max_{i\in\mathcal{I}_-}\langle w_j(0), x_i\rangle > 0$, *then* $w_j(t_1) \in \mathcal{S}_-$.

2. *(Final convergence and low-rank bias)* $\forall t \geq t_1$ *and* $\forall j \in [h]$, *neuron* $w_j(t)$ *stays within* $\mathcal{S}_+$ *(*$\mathcal{S}_-$*, or* $\mathcal{S}_{\textit{dead}}$*) if* $w_j(t_1) \in \mathcal{S}_+$ *(*$\mathcal{S}_-$*, or* $\mathcal{S}_{\textit{dead}}$ *resp.). Moreover, if both* $\mathcal{S}_+$ *and* $\mathcal{S}_-$ *contains at least one neuron at time* t_1, *then*

- *$\exists\alpha > 0$ and $\exists t_2$ with $t_1 \leq t_2 = \Theta(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon})$, such that $\mathcal{L}(t) \leq \frac{\mathcal{L}(t_2)}{\mathcal{L}(t_2)\alpha(t-t_2)+1}$, $\forall t \geq t_2$.*
- *As $t \to \infty$, $\|W(t)\| \to \infty$ and $\|W(t)\|_F^2 \leq 2\|W(t)\|_2^2 + \mathcal{O}(\epsilon)$. Thus, the stable rank of $W(t)$ satisfies $\limsup_{t\to\infty} \|W(t)\|_F^2/\|W(t)\|_2^2 \leq 2$.*

We make the following remarks:

Early neuron alignment: The first part of the Theorem 2.10 describes the configuration of *all* neurons at the end of the alignment phase. Every neuron in $\mathcal{V}_+$ reaches either $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ by t_1, and stays there for the remainder of training. Obviously, we care about those neurons reaching $\mathcal{S}_+$ as any neuron in $\mathcal{S}_{\text{dead}}$ does not contribute to the final convergence at all. Luckily, Theorem 2.10 suggests that any neuron in $\mathcal{V}_+$ that starts with some activation on the positive data, i.e., it is initialized in the union of halfspaces $\cup_{i\in\mathcal{I}_+}\{w : \langle w, x_i\rangle > 0\}$, will eventually reach $\mathcal{S}_+$. A similar discussion holds for neurons in $\mathcal{V}_-$. We argue that randomly initializing W_0 ensures that with high probability, there will be at least a pair of neurons reaching $\mathcal{S}_+$ and $\mathcal{S}_-$ by time t_1 (please see the next remark). Lastly, we note that it is possible that $\mathcal{S}_{\text{dead}} = \emptyset$, in which case every neuron reaches either $\mathcal{S}_+$ or $\mathcal{S}_-$.

Merits of random initialization: Our theorem is stated for a deterministic initialization (2.165) given an initial shape W_0. In practice, one would use random initialization to find a W_0, for example, $[W_0]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/D)$. First, our Theorem 2.10 applies to this Gaussian initialization: Assumption 2.3 is satisfied with probability one because the events $\langle \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_-}{\|x_-\|}\rangle = 1$ and $\langle \frac{w_j(0)}{\|w_j(0)\|}, \frac{x_+}{\|x_+\|}\rangle = 1$ have probability zero. Moreover, any neuron in $\mathcal{V}_+$ has at least probability $1/2$ of being initialized within the union of halfspaces $\cup_{i\in\mathcal{I}_+}\{w : \langle w, x_i\rangle > 0\}$, which ensures that this neuron reaches $\mathcal{S}_+$. Thus when there are m neurons in $\mathcal{V}_+$, the probability that $\mathcal{S}_+$ has at least one neuron at time t_1 is lower bounded by $1 - 2^{-m}$ (same argument holds for $\mathcal{S}_-$), Therefore, with only very mild overparametrization on the network width h, one can make sure that with high probability there is at least one neuron in both

$\mathcal{S}_+$ and $\mathcal{S}_-$, leading to final convergence.

Duration of alignment phase: Our theorem shows that for sufficiently small ϵ, directional convergence, i.e., all neurons reaching either $\mathcal{S}_+, \mathcal{S}_-, \mathcal{S}_{\text{dead}}$, is achieved within $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ time (notably, independent of ϵ). Our bound quantitatively reveals the non-trivial dependency on the "data separation" μ for such directional convergence to occur. To the best of our knowledge, this is the first non-asymptotic bound on the time it takes for all neurons to achieve a desired configuration. [79] only shows such $t_1 > 0$ exists using an $\epsilon \to 0$ argument, without analyzing how large t_1 can be. [16] studies a different data assumption (we compare it with ours in later remarks) under which the alignment is studied only for neurons that has a specific activation pattern at initialization. Lastly, we note that $\mu \to 0$ leads to $t_1 \to \infty$, this is because when $\mu = 0$, there are more limiting directions to which neurons can converge, hence not all of them are "attracted" by $\mathcal{S}_+, \mathcal{S}_-, \mathcal{S}_{\text{dead}}$.

Refined alignment within $\mathcal{S}_+, \mathcal{S}_-$: Once a neuron in $\mathcal{V}_+$ reaches $\mathcal{S}_+$, it never leaves $\mathcal{S}_+$. Moreover, it always gets attracted by x_+. Therefore, every neuron gets well aligned with x_+, i.e., $\cos(w_j, x_+) \simeq 1, \forall w_j \in \mathcal{S}_+$. A similar argument shows neurons in $\mathcal{S}_-$ get attracted by x_-. We opt not to formally state it in Theorem 2.10 as the result would be similar to that in [16], and alignment with x_+, x_- is not necessary to guarantee convergence. Instead, we show this refined alignment through our numerical experiment in Section 2.3.3.

Final convergence and low-rank bias: The convergence analysis after t_1 is simple: All neurons in $\mathcal{S}_{\text{dead}}$ have small norm and do not move thus they can be ignored from the analysis. More interestingly, GF after t_1 can be viewed as fitting positive data $x_i, i \in \mathcal{I}_+$, with a subnetwork consisting of all neurons in $\mathcal{S}_+$, and fitting negative data with neurons in $\mathcal{S}_-$. By the fact that all neurons in $\mathcal{S}_+$ activate all $x_i, i \in \mathcal{I}_+$, the resulting subnetwork is linear, and so is the subnetwork for fitting $x_i, i \in \mathcal{I}_-$. The convergence analysis reduces to establishing $\mathcal{O}(1/t)$ convergence for two linear networks [26, 30, 28]. As for the stable rank, our result follows the analysis in [71], but in a simpler form since ours is for linear networks.

Comparison with [78]: Prior work [78] considers a similar data assumption to ours. However, [78] assumes that there exists a time t_1 such that at t_1, the neurons are in either $\mathcal{S}_+, \mathcal{S}_-$ or $\mathcal{S}_{\text{dead}}$ and their main contribution is the analysis of the implicit bias for the later stage of the training. [78] justifies their assumption by the analysis in [79], which does not necessarily apply to the case of finite ϵ, as we discussed in Section 2.3.1. Our work precisely establishes such directional convergence for finite but small ϵ, showing indeed the neurons achieve some good alignment with x_+, x_- within $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ time before they start to grow in norm. Moreover, [78] has no characterization on the convergence rate of the loss after the alignment phase, while we provide a $\mathcal{O}(1/t)$ bound on the loss. In addition, [78] considers the case where input data $x_i, i \in [n]$, spans the entire $\mathbb{R}^D$, which leads to $\mathcal{S}_{\text{dead}} = \emptyset$. This implicitly imposes the constraint that the number of data points n must be larger than the input dimension D. Our analysis allows for the case $\mathcal{S}_{\text{dead}} \neq \emptyset$ as we provide a sufficient condition for preventing a neuron from reaching $\mathcal{S}_{\text{dead}}$.

Comparison with [16]: In [16], the neuron alignment is carefully analyzed for the case all data points are orthogonal to each other, i.e., $\langle x_i, x_j \rangle = 0, \forall i \neq j \in [n]$. Such an assumption restricts the number of data points n to be smaller than the input dimension D and is often unrealistic. Our assumption does not restrict the size of the dataset and thus has more practical relevance (see our numerical experiments in Section 2.3.3).

Proof sketch for the alignment phase

We sketch the proof for our Theorem 2.10. First of all, it can be shown that $\mathcal{S}_+, \mathcal{S}_{\text{dead}}$ are trapping regions for all $w_j(t), j \in \mathcal{V}_+$, that is, whenever $w_j(t)$ gets inside $\mathcal{S}_+$ (or $\mathcal{S}_{\text{dead}}$), it never leaves $\mathcal{S}_+$ (or $\mathcal{S}_{\text{dead}}$). Similarly, $\mathcal{S}_-, \mathcal{S}_{\text{dead}}$ are trapping regions for all $w_j(t), j \in \mathcal{V}_-$. The alignment phase analysis concerns how long it takes for all neurons to reach one of the trapping regions, followed by the final convergence analysis on fitting data with $+1$ label by neurons in $\mathcal{S}_+$ and fitting data with -1 label by those in $\mathcal{S}_-$. We have discussed the final convergence analysis in the remark "Final convergence and low-rank bias", thus we focus on the proof sketch for the early alignment phase here, which is considered as our main technical contribution.

Approximating $\frac{d}{dt}\frac{w_j}{\|w_j\|}$: Our analysis for the neural alignment is rooted in the following Lemma:

Lemma 2.20. *Given some initialization from* (2.165)*, if $\epsilon = \mathcal{O}(\frac{1}{\sqrt{h}})$, then there exists $T = \Theta(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon})$ such that any solution to the gradient flow dynamics* (2.164) *satisfies that $\forall t \leq T$,*

$$\max_j \left\| \frac{d}{dt}\frac{w_j(t)}{\|w_j(t)\|} - \mathrm{sign}(v_j(0))\mathcal{P}_{w_j(t)}x_a(w_j(t)) \right\| = \mathcal{O}\left(\epsilon n\sqrt{h}\right) . \tag{2.169}$$

This Lemma shows that the error between $\frac{d}{dt}\frac{w_j(t)}{\|w_j(t)\|}$ and $\mathrm{sign}(v_j(0))\mathcal{P}_{w_j(t)}x_a(w_j(t))$ can be arbitrarily small with some appropriate choice of ϵ (to be determined later). This allows one to analyze the true directional dynamics $\frac{w_j(t)}{\|w_j(t)\|}$ using some property of $\mathcal{P}_{w_j(t)}x_a(w_j(t))$, which leads to a $t_1 = \mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ upper bound on the time it takes for the neuron direction to converge to the sets $\mathcal{S}_+$, $\mathcal{S}_-$, or $\mathcal{S}_{\text{dead}}$. Moreover, it also suggests ϵ can be made sufficiently small so that the error bound holds until the directional convergence is achieved, i.e. $T \geq t_1$. We will first illustrate the analysis for directional convergence, then close the proof sketch with the choice of a sufficiently small ϵ.

Activation pattern evolution: Given a sufficiently small ϵ, one can show that under Assumption 2.2, for every neuron w_j that is not in $\mathcal{S}_{\text{dead}}$ we have:

$$\frac{d}{dt}\left\langle \frac{w_j}{\|w_j\|}, \frac{x_i y_i}{\|x_i\|} \right\rangle \Bigg|_{\langle w_i, x_i\rangle = 0} > 0, \forall i \in [n], \text{if } j \in \mathcal{V}_+ , \tag{2.170}$$

$$\frac{d}{dt}\left\langle \frac{w_j}{\|w_j\|}, \frac{x_i y_i}{\|x_i\|} \right\rangle \Bigg|_{\langle w_i, x_i\rangle = 0} < 0, \forall i \in [n], \text{if } j \in \mathcal{V}_- . \tag{2.171}$$

This is because whenever a neuron satisfies $\langle x_i, w_j\rangle = 0$ for some i, and has activation on some other data, GF moves w_j towards $x_a(w_j) = \sum_{i:\langle x_i,w_j\rangle>0} x_i y_i$. Interestingly, Assumption 2.2 implies $\langle x_i y_i, x_a(w_j)\rangle > 0, \forall i \in [n]$, which makes $\frac{d}{dt}\frac{w_j}{\|w_j\|} \simeq \mathrm{sign}(v_j(0))\mathcal{P}_{w_j} x_a(w_j)$ point inward (or outward) the halfspace $\langle x_i y_i, w_j\rangle > 0$, if $\mathrm{sign}(v_j(0)) > 0$ (or $\mathrm{sign}(v_j(0)) < 0$, respectively). See Figure 2-9 for illustration.

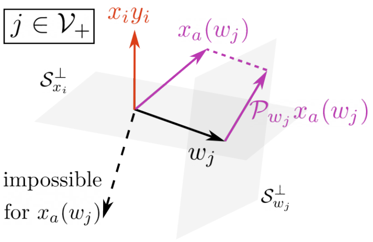
$j \in \mathcal{V}_+$
$x_i y_i$
$x_a(w_j)$
$\mathcal{S}^\perp_{x_i}$
$\mathcal{P}_{w_j} x_a(w_j)$
w_j
impossible
for $x_a(w_j)$
$\mathcal{S}^\perp_{w_j}$

Figure 2-9. For $j \in \mathcal{V}_+$, Assumption 2.2 enforces $\langle x_i y_i, x_a(w_j)\rangle > 0$, thus GF pushes w_j inward the half-space $\langle x_i y_i, w_j\rangle > 0$ at $\langle x_i, w_j\rangle = 0$ (i.e. towards gaining activation on x_i, if $y_i = +1$, or losing activation on x_i, if $y_i = -1$.). $\mathcal{S}^\perp_{x_i}$ and $\mathcal{S}^\perp_{w_j}$ denotes the subspace orthogonal to x_i and w_j, respectively.

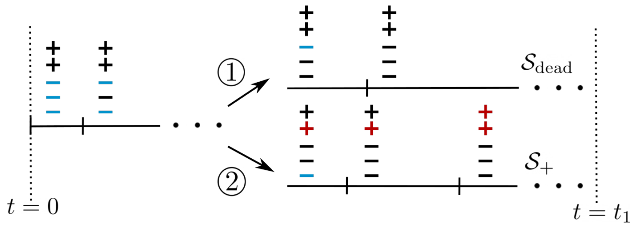

Figure 2-10. Illustration of the activation pattern evolution. The epochs on the time axis denote the time w_j changes its activation pattern by either losing one negative data (denoted by "$+$") or gaining one positive data (denoted by "$-$"). The markers are colored if it currently activates w_j. During the alignment phase $0 \le t \le t_1$, a neuron $w_j, j \in \mathcal{V}_+$ starts with activation on all negative data and no positive data, every $\mathcal{O}(1/n_a)$ time, it must change its activation, unless either ① it reaches $\mathcal{S}_{\text{dead}}$, or ② it activates some positive data at some epoch then eventually reaches $\mathcal{S}_+$.

As a consequence, a neuron can only change its activation pattern in a particular manner: a neuron in $\mathcal{V}_+$, whenever it is activated by some x_i with $y_i = +1$, never loses the activation on x_i thereafter, because (2.170) implies that GF pushes $\frac{w_j}{\|w_j\|}$ towards x_i at the boundary $\langle w_j, x_i\rangle = 0$. Moreover, (2.170) also shows that a neuron in $\mathcal{V}_+$ will never regain activation on a x_i with $y_i = -1$ once it loses the activation because GF pushes $\frac{w_j}{\|w_j\|}$ against x_i at the boundary $\langle w_i, x_i\rangle = 0$. Similarly, a neuron in $\mathcal{V}_-$ never loses activation on negative data and never gains activation on positive data.

Bound on activation transitions and duration: Equations (2.170) and (2.171) are key in the analysis of alignment because they limit how many times a neuron can change its activation pattern: a neuron in $\mathcal{V}_+$ can only gain activation on positive data and lose activation on negative data, thus at maximum, a neuron $w_j,\ j \in \mathcal{V}_+$, can start with full activation on all negative data and no activation on any positive one (which implies $w_j(0) \in \mathcal{S}_-$) then lose activation on every negative data and gain activation on every positive data as GF training proceeds (which implies $w_j(t_1) \in \mathcal{S}_+$), taking at most n changes on its activation pattern. See Figure 2-10 for an illustration. Then, since it is possible to show that a neuron w_j with $j \in \mathcal{V}_+$ that has $\cos(w_j, x_-) < 1$ (guaranteed by Assumption 2.3) and is not in $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$, must change its activation pattern after $\mathcal{O}(\frac{1}{n_a\sqrt{\mu}})$ time (that does not depend on ϵ), where n_a is the number of data that currently activates w_j, one can upper bound the time for w_j to reach $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ by some $t_1 = \mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ constant independent of ϵ.

Moreover, w_j must reach $\mathcal{S}_+$ if it initially has activation on at least one positive data, i.e., $\max_{i\in\mathcal{I}_+} \langle w_j(0), x_i\rangle > 0$ since it cannot lose this activation. A similar argument holds for $w_j, j \in \mathcal{V}_-$ that they reaches either $\mathcal{S}_-$ or $\mathcal{S}_{\text{dead}}$ before t_1.

Choice of ϵ: All the aforementioned analyses rely on the assumption that the approximation in equation (2.166) holds with some specific error bound. We show in Appendix 2.3.3 that the desired bound is $\left\|\frac{d}{dt}\frac{w_j(t)}{\|w_j(t)\|} - \operatorname{sign}(v_j(0))\mathcal{P}_{w_j(t)}x_a(w_j(t))\right\| \leq \mathcal{O}(\sqrt{\mu})$, which, by Lemma 2.20, can be achieved by a sufficiently small initialization scale $\epsilon_1 = \mathcal{O}(\frac{\sqrt{\mu}}{\sqrt{h}n})$. Moreover, the directional convergence (which takes $\mathcal{O}(\frac{\log n}{\sqrt{\mu}})$ time) should be achieved before the alignment phase ends, which happens at $T = \Theta(\frac{1}{n}\log\frac{1}{\sqrt{h}\epsilon})$. This is ensured by choosing another sufficiently small initialization scale $\epsilon_2 = \mathcal{O}(\frac{1}{\sqrt{h}}\exp(-\frac{n}{\sqrt{\mu}}\log n))$. Overall, the initialization scale should satisfy $\epsilon \leq \min\{\epsilon_1, \epsilon_2\}$. We opt to present ϵ_2 in our main theorem because ϵ_2 beats ϵ_1 when n is large.

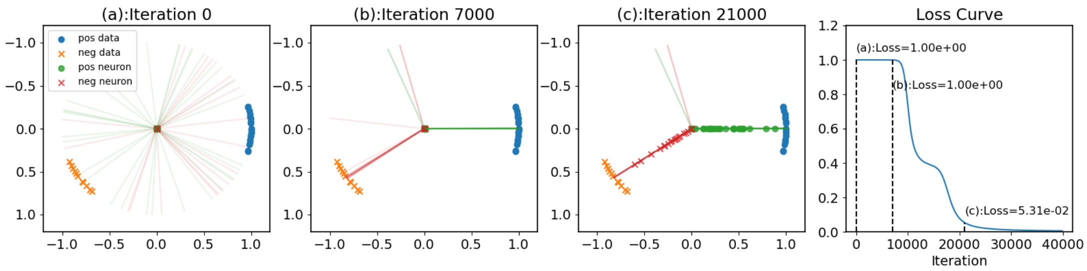

Figure 2-11. Illustration of gradient descent on two-layer ReLU network with small initialization. The marker represents either a data point or a neuron. Solid lines represent the directions of neurons. (a) at initialization, all neurons have small norm and are pointing in different directions; (b) around the end of the alignment phase, all neurons are in $\mathcal{S}_+, \mathcal{S}_-$, or $\mathcal{S}_{\text{dead}}$. Moreover, neurons in $\mathcal{S}_+$ ($\mathcal{S}_-$) are well aligned with x_+ (x_-); (c) With good alignment, neurons in $\mathcal{S}_-, \mathcal{S}_+$ start to grow in norm and the loss decreases. When the loss is close to zero, the resulting network has its first-layer weight approximately low-rank.

2.3.3 Numerical experiments

Illustrative example

We first illustrate our theorem using a toy example: we train a two-layer ReLU network with $h = 50$ neurons under a toy dataset in $\mathbb{R}^2$ (See Figure 2-11) that satisfies our Assumption 2.2, and initialize all entries of the weights as $[W]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha), v_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha), \forall i \in [n], j \in [h]$ with $\alpha = 10^{-6}$. Then we run gradient descent on both W and v with step size $\eta = 2 \times 10^{-3}$. Our theorem well predicts the dynamics of neurons at the early stage of the training: aside from neurons that ended up in $\mathcal{S}_{\text{dead}}$, neurons in $\mathcal{V}_+$ reach $\mathcal{S}_+$ and achieve good alignment with x_+, and neurons in $\mathcal{V}_-$ are well aligned with x_- in $\mathcal{S}_-$. Note that after alignment, the loss experiences two sharp decreases before it gets close to zero, which is studied and explained in [16].

Binary classification on two MNIST digits

Next, we consider a binary classification task for two MNIST digits. Such training data do not satisfy Assumption 2.2 since every data vector is a grayscale image with non-negative entries, making the inner product between any pair of data non-negative, regardless of their labels. However, we can preprocess the training data by centering: $x_i \leftarrow x_i - \bar{x}$, where $\bar{x} = \sum_{i\in[n]} x_i/n$. The preprocessed data, then, approximately satisfies our assumption (see the left-most plot in Figure 2-12): a pair of data points is very likely to have a positive correlation if they have the same label and to have a negative correlation if they have different labels. Thus we expect our theorem to make reasonable predictions on the training dynamics with preprocessed data. For the remaining part of this section, we use $x_i, i \in [n]$, to denote the preprocessed (centered) data and use $\bar{x}$ to denote the mean of the original data.

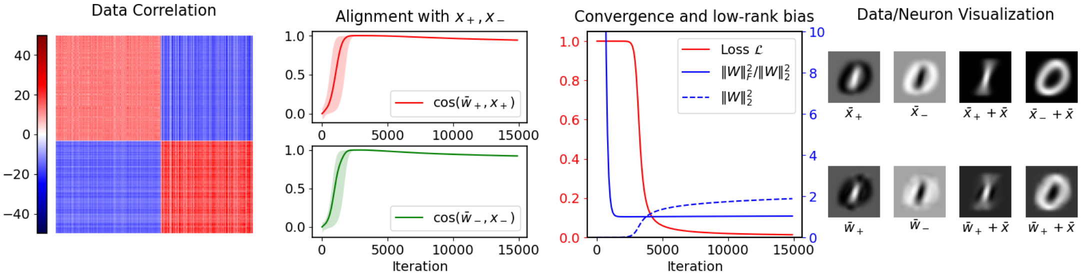

Figure 2-12. Training two-layer ReLU network under small initialization for binary classification on MNIST digits 0 and 1. The training data is preprocessed to be centered. (*First Plot*) Data correlation $[\langle x_i, x_j\rangle]_{ij}$ as a heatmap, where the data are reordered by their label (digit 1 first, then digit 0); (*Second Plot*) Alignment between neurons and the aggregate positive/negative data $x_+ = \sum_{i\in\mathcal{I}_+} x_i$, $x_- = \sum_{i\in\mathcal{I}_-} x_i$. In the top figure, the solid line shows $\cos(\bar{w}_+, x_+)$ during training, and the shaded region defines the range between $\min_{j\in\mathcal{V}_+}\cos(w_i, x_+)$ and $\max_{j\in\mathcal{V}_+}\cos(w_i, x_+)$. Similarly, in the bottom figure, the solid line shows $\cos(\bar{w}_-, x_-)$ during training, and the shaded region lies between $\min_{j\in\mathcal{V}_-}\cos(w_i, x_-)$ and $\max_{j\in\mathcal{V}_-}\cos(w_i, x_-)$; (*Third Plot*) The loss $\mathcal{L}$, the stable rank and the squared spectral norm of W during training; (*Fourth Plot*) Visualizing neuron centers $\bar{w}_+, \bar{w}_-$ and data centers $\bar{x}_+, \bar{x}_-$ (at iteration 15000) as grayscale images. $\bar{x}$ is the mean of the original training data, prior to preprocessing.

We build a two-layer ReLU network with $h = 50$ neurons and initialize all entries of the weights as $[W]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha), v_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \alpha), \forall i \in [n], j \in [h]$ with $\alpha = 10^{-5}$. Then we run gradient descent on both W and v with step size $\eta = 2 \times 10^{-3}$. Notice that here the weights are not initialized to be balanced as in (2.165). The numerical results are shown in Figure 2-12.

Alignment phase: Without balancedness, one no longer has $\mathrm{sign}(v_j(t)) = \mathrm{sign}(v_j(0))$. With a little abuse of notation, we denote $\mathcal{V}_+(t) = \{j \in [h] : \mathrm{sign}(v_j(t)) > 0\}$ and $\mathcal{V}_+(t) = \{j \in [h] : \mathrm{sign}(v_j(t)) > 0\}$, and we expect that at the end of the alignment phase, neurons in $\mathcal{V}_+$ are aligned with $x_+ = \sum_{i\in\mathcal{I}_+} x_i$, and neurons in $\mathcal{V}_-$ with $x_- = \sum_{i\in\mathcal{I}_-} x_i$. The second plot in Figure 2-12 shows such an alignment between neurons and x_+, x_-. In the top part, the red solid line shows $\cos(\bar{w}_+, x_+)$ during training, where $\bar{w}_+ = \sum_{j\in\mathcal{V}_+} w_j/|\mathcal{V}_+|$, and the shaded region defines the range between $\min_{j\in\mathcal{V}_+} \cos(w_i, x_+)$ and $\max_{j\in\mathcal{V}_+} \cos(w_i, x_+)$. Similarly, in the bottom part, the green solid line shows $\cos(\bar{w}_-, x_-)$ during training, where $\bar{w}_- = \sum_{j\in\mathcal{V}_-} w_j/|\mathcal{V}_-|$, and the shaded region shows the range between $\min_{j\in\mathcal{V}_-} \cos(w_i, x_-)$ and $\max_{j\in\mathcal{V}_-} \cos(w_i, x_-)$. Initially, every neuron is approximately orthogonal to x_+, x_- due to random initialization. Then all neurons in $\mathcal{V}_+$ ($\mathcal{V}_-$) start to move towards x_+ (x_-) and achieve good alignment after $\sim$2000 iterations. When the loss starts to decrease (after ~ 3000 iterations), the alignment drops a little. We conjecture that this is because the dataset does not exactly satisfy our Assumption 2.2, and the neurons in $\mathcal{V}_+$ have to fit some negative data, for which x_+ is not the best direction.

Final convergence: After ~ 3000 iterations, the norm $\|W\|_2^2$ starts to grow and the loss decreases, as shown in the third plot in Figure 2-12. Moreover, the stable rank $\|W\|_F^2/\|W\|_2^2$ decreases below 2. For this experiment, we almost have $\cos(x_+, x_-) \simeq -1$, thus the neurons in $\mathcal{V}_+$ (aligned with x_+) and those in $\mathcal{V}_-$ (aligned with x_-) are almost co-linear. Therefore, the stable rank $\|W\|_F^2/\|W\|_2^2$ is almost 1, as seen from the plot. Finally, at iteration 15000, we visualize the mean neuron $\bar{w}_+ = \sum_{j\in\mathcal{V}_+} w_j/|\mathcal{V}_+|$, $\bar{w}_- = \sum_{j\in\mathcal{V}_-} w_j/|\mathcal{V}_-|$ as grayscale images, and compare them with $\bar{x}_+ = x_+/|\mathcal{I}_+|, x_- = x_-/|\mathcal{I}_-|$, showing good alignment. We also show the images when the original data center $\bar{x}$ is added back.

Proof of Lemma 2.20

The following property of the exponential loss ℓ will be used throughout the Appendix for proofs of several results:

Lemma 2.21. *For exponential loss ℓ, we have*

$$|-\nabla_{\hat{y}}\ell(y,\hat{y})-y|\leq 2|\hat{y}|,\forall y\in\{+1,-1\},\quad \forall|\hat{y}|\leq 1\,. \tag{2.172}$$

Proof.

$$\begin{aligned}|-\nabla_{\hat{y}}\ell(y,\hat{y})-y| =&\ |y\exp(-y\hat{y})-y|\\ \leq&\ |y||\exp(-y\hat{y})-1|\\ \leq&\ |\exp(-y\hat{y})-1|\leq 2|\hat{y}|\,,\end{aligned}$$

where the last inequality is due to the fact that $2x\geq\max\{1-\exp(-x),\exp(x)-1\},\forall x\in[0,1]$. □

Formal statement

Denote: $X_{\max}=\max_i\|x_i\|, W_{\max}=\max_j\|[W_0]_{:,j}\|$. The formal statement of Lemma 2.20 is as follow:

Lemma 2.20. *Given some initialization from* (2.165)*, for any $\epsilon\leq\frac{1}{4\sqrt{h}X_{\max}W_{\max}^2}$, then any solution to the gradient flow dynamics* (2.164) *satisfies that $\forall t\leq T=\frac{1}{4nX_{\max}}\log\frac{1}{\sqrt{h}\epsilon}$,*

$$\max_j\left\|\frac{d}{dt}\frac{w_j(t)}{\|w_j(t)\|}-\operatorname{sign}(v_j(0))\mathcal{P}_{w_j(t)}x_a(w_j(t))\right\|\leq 4\epsilon n\sqrt{h}X_{\max}^2W_{\max}^2\,.$$

Lemma 2.20 is a direct result of the following two lemmas.

Lemma 2.22. *Given some initialization in* (2.165)*, then for any $\epsilon\leq\frac{1}{4\sqrt{h}X_{\max}W_{\max}^2}$, any solution to the gradient flow dynamics* (2.164) *satisfies*

$$\max_j\|w_j(t)\|^2\leq\frac{2\epsilon W_{\max}^2}{\sqrt{h}},\quad \max_i|f(x_i;W(t),v(t))|\leq 2\epsilon\sqrt{h}X_{\max}W_{\max}^2\,, \tag{2.173}$$

$\forall t\leq\frac{1}{4nX_{\max}}\log\frac{1}{\sqrt{h}\epsilon}$.

Lemma 2.23. *Consider any solution to the gradient flow dynamic (2.164) starting from initialization (2.165). Whenever* $\max_i |f(x_i; W, v)| \le 1$, *we have,* $\forall i \in [n]$,

$$\left\| \frac{d}{dt}\frac{w_j}{\|w_j\|} - \operatorname{sign}(v_j(0))\left(I - \frac{w_j w_j^\top}{\|w_j\|^2}\right)\left(\sum_{i:\langle x_i, w_j\rangle > 0} y_i x_i\right)\right\| \le 2nX_{\max}\max_i |f(x_i; W, v)| \,. \tag{2.174}$$

Proof of Lemma 2.22 and Lemma 2.23

Proof of Lemma 2.22. Under gradient flow, we have

$$\frac{d}{dt}w_j = -\sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \ge 0}\nabla_{\hat y}\ell(y_i, f(x_i; W, v))x_i v_j \,. \tag{2.175}$$

Balanced initialization enforces $v_j = \operatorname{sign}(v_j(0))\|w_j\|$, hence

$$\frac{d}{dt}w_j = -\sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \ge 0}\nabla_{\hat y}\ell(y_i, f(x_i; W, v))x_i \operatorname{sign}(v_j(0))\|w_j\| \,. \tag{2.176}$$

Let $T := \inf\{t :\ \max_i |f(x_i; W(t), v(t))| > 2\epsilon\sqrt{h}X_{\max}W_{\max}^2\}$, then $\forall t \le T, j \in [h]$, we have

$$\begin{aligned}
\frac{d}{dt}\|w_j\|^2 &= \left\langle w_j, \frac{d}{dt}w_j\right\rangle \\
&= -2\sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \ge 0}\nabla_{\hat y}\ell(y_i, f(x_i; W, v))\left\langle x_i, w_j\right\rangle \operatorname{sign}(v_j(0))\|w_j\| \\
&\le 2\sum_{i=1}^n |\nabla_{\hat y}\ell(y_i, f(x_i; W, v))|\,|\langle x_i, w_j\rangle|\,\|w_j\| \\
&\le 2\sum_{i=1}^n (|y_i| + 2|f(x_i; W, v)|)\,|\langle x_i, w_j\rangle|\,\|w_j\| && \text{(by Lemma 2.21)} \\
&\le 2\sum_{i=1}^n (1 + 4\epsilon\sqrt{h}X_{\max}W_{\max}^2)\,|\langle x_i, w_j\rangle|\,\|w_j\| && \text{(Since } t \le T\text{)} \\
&\le 2\sum_{i=1}^n (1 + 4\epsilon\sqrt{h}X_{\max}W_{\max}^2)\|x_i\|\|w_j\|^2 \\
&\le 2n(X_{\max} + 4\epsilon\sqrt{h}X_{\max}^2W_{\max}^2)\|w_j\|^2 \,.
\end{aligned} \tag{2.177}$$

Let $\tau_j := \inf\{t : \|w_j(t)\|^2 > \frac{2\epsilon W_{\max}^2}{\sqrt{h}}\}$, and let $j^* := \arg\min_j \tau_j$, then $\tau_{j^*} = \min_j \tau_j \leq T$ due to the fact that

$$|f(x_i; W, v)| = \left|\sum_{j\in[h]} \mathbb{1}_{\langle w_j, x_i\rangle > 0} v_j \langle w_j, x_i\rangle\right| \leq \sum_{j\in[h]} \|w_j\|^2 \|x_i\| \leq h X_{\max} \max_{j\in[h]} \|w_j\|^2 ,$$

which implies "$|f(x_i; W(t), v(t))| > 2\epsilon\sqrt{h} X_{\max} W_{\max}^2 \Rightarrow \exists j, s.t. \|w_j(t)\|^2 > \frac{2\epsilon W_{\max}^2}{\sqrt{h}}$".

Then for $t \leq \tau_{j^*}$, we have

$$\frac{d}{dt}\|w_{j^*}\|^2 \leq 2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)\|w_{j^*}\|^2 . \tag{2.178}$$

By Grönwall's inequality, we have $\forall t \leq \tau_{j^*}$

$$\begin{aligned}
\|w_{j^*}(t)\|^2 \leq &\ \exp\left(2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)t\right) \|w_{j^*}(0)\|^2 , \\
= &\ \exp\left(2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)t\right) \epsilon^2 \|[W_0]_{:,j^*}\|^2 \\
\leq &\ \exp\left(2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)t\right) \epsilon^2 W_{\max}^2 .
\end{aligned}$$

Suppose $\tau_{j^*} < \frac{1}{4nX_{\max}} \log\left(\frac{1}{\sqrt{h}\epsilon}\right)$, then by the continuity of $\|w_{j^*}(t)\|^2$, we have

$$\begin{aligned}
\frac{2\epsilon W_{\max}^2}{\sqrt{h}} \leq &\ \|w_{j^*}(\tau_{j^*})\|^2 \\
\leq &\ \exp\left(2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)\tau_{j^*}\right) \epsilon^2 W_{\max}^2 \\
\leq &\ \exp\left(2n(X_{\max} + 4\epsilon\sqrt{h} X_{\max}^2 W_{\max}^2)\frac{1}{4nX_{\max}} \log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right) \epsilon^2 W_{\max}^2 \\
\leq &\ \exp\left(\frac{1 + 4\epsilon\sqrt{h} X_{\max} W_{\max}^2}{2} \log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right) \epsilon^2 W_{\max}^2 \\
\leq &\ \exp\left(\log\left(\frac{1}{\sqrt{h}\epsilon}\right)\right) \epsilon^2 W_{\max}^2 = \frac{\epsilon W_{\max}^2}{\sqrt{h}} ,
\end{aligned}$$

which leads to a contradiction $2\epsilon \leq \epsilon$. Therefore, one must have $T \geq \tau_{j^*} \geq \frac{1}{4nX_{\max}} \log\left(\frac{1}{\sqrt{h}\epsilon}\right)$. This finishes the proof. □

Proof of Lemma 2.23. As we showed in the proof for Lemma 2.22, under balanced initialization,

$$\frac{d}{dt} w_j = -\sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \geq 0} \nabla_{\hat{y}} \ell(y_i, f(x_i; W, v)) x_i \text{sign}(v_j(0)) \|w_j\| . \tag{2.179}$$

Then for any $i \in [n]$,

$$\begin{aligned}
\frac{d}{dt}\frac{w_j}{\|w_j\|} &= -\mathrm{sign}(v_j(0))\sum_{i=1}^{n}\mathbb{1}_{\langle x_i,w_j\rangle>0}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\left(x_i-\frac{\langle x_i,w_j\rangle}{\|w_j\|^2}w_j\right)\\
&= -\mathrm{sign}(v_j(0))\sum_{i:\langle x_i,w_j\rangle>0}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\left(x_i-\frac{\langle x_i,w_j\rangle}{\|w_j\|^2}w_j\right)\\
&= -\mathrm{sign}(v_j(0))\left(I-\frac{w_jw_j^\top}{\|w_j\|^2}\right)\left(\sum_{i:\langle x_i,w_j\rangle>0}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))x_i\right).
\end{aligned}$$

Therefore, whenever $\max_i|f(x_i;W,v)|\leq 1$,

$$\begin{aligned}
&\left\|\frac{d}{dt}\frac{w_j}{\|w_j\|}-\mathrm{sign}(v_j(0))\left(I-\frac{w_jw_j^\top}{\|w_j\|^2}\right)\left(\sum_{i:\langle x_i,w_j\rangle>0}y_ix_i\right)\right\|\\
=&\left\|\mathrm{sign}(v_j(0))\left(\sum_{i:\langle x_i,w_j\rangle>0}\left(\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))+y_i\right)x_i\right)\right\|\\
\leq&\sum_{i=1}^{n}|\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))+y_i|\cdot\|x_i\|\\
\leq&\sum_{i=1}^{n}2|f(x_i;W,v)|\cdot\|x_i\|\leq 2nM_x\max_i|f(x_i;W,v)|\,. \qquad (2.180)
\end{aligned}$$

□

Proof for Theorem 2.10 (part one): early alignment phase

We break the proof of Theorem 2.10 into two parts: In Appendix 2.3.3 we prove the first part regarding directional convergence. Then in Appendix 2.3.3 we prove the remaining statement on final convergence and low-rank bias.

Auxiliary lemmas

The first several Lemmas concern mostly some conic geometry given the data assumption:

Consider the following conic hull

$$K=\mathcal{CH}(\{x_iy_i,i\in[n]\})=\left\{\sum_{i=1}^{n}a_ix_iy_i:a_i\geq 0,i\in[n]\right\}. \qquad (2.181)$$

It is clear that $x_iy_i \in K, \forall i$, and $x_a(w) \in K, \forall w$. The following lemma shows any pair of vectors in K is μ-coherent.

Lemma 2.24. $\cos(z_1, z_2) \geq \mu, \forall 0 \neq z_1, z_2 \in K$.

Proof. Since $z_1, z_2 \in K$, we let $z_1 = \sum_{i=1}^n x_iy_ia_{1i}$, and$z_2 = \sum_{j=1}^n x_jy_ja_{2j}$, where $a_{1i}, a_{2j} \geq 0$ but not all of them.

$$\begin{aligned}\cos(z_1, z_2) = \frac{1}{\|z_1\|\|z_2\|}\langle z_1, z_2\rangle = & \frac{1}{\|z_1\|\|z_2\|}\sum_{i,j\in[n]} a_{1i}a_{2j}\langle x_iy_i, x_jy_j\rangle \\ = & \frac{\sum_{i,j\in[n]}\|x_i\|\|x_j\|a_{1i}a_{2j}\mu}{\|z_1\|\|z_2\|} \geq \mu\,,\end{aligned}$$

where the last inequality is due to

$$\|z_1\|\|z_2\| \leq \left(\sum_{i=1}^n \|x_i\|a_{1i}\right)\left(\sum_{j=1}^n \|x_j\|a_{2j}\right) = \sum_{i,j\in[n]} \|x_i\|\|x_j\|a_{1i}a_{2j}\,.$$

□

The following lemma is some basic results regarding $\mathcal{S}_+$ and $\mathcal{S}_-$:

Lemma 2.25. *$\mathcal{S}_+$ and $\mathcal{S}_-$ are convex cones(excluding the origin).*

Proof. Since $\mathbb{1}_{\langle x_i, z\rangle} = \mathbb{1}_{\langle x_i, az\rangle}, \forall i \in [n], a > 0$, $\mathcal{S}_+, \mathcal{S}_-$ are cones. Moreover, $\langle x_i, z_1\rangle > 0$ and $\langle x_i, z_2\rangle > 0$ implies $\langle x_i, a_1z_1 + a_2z_2\rangle > 0, \forall a_1, a_2 > 0$, thus $\mathcal{S}_+, \mathcal{S}_-$ are convex cones. □

Now we consider the complete metric space $\mathbb{S}^{D-1}$ (w.r.t. $\arccos(\langle\cdot,\cdot\rangle)$) and we are interested in its subsets $K \cap \mathbb{S}^{D-1}$, $\mathcal{S}_+ \cap \mathbb{S}^{D-1}$, and $\mathcal{S}_- \cap \mathbb{S}^{D-1}$. First, we have (we use $\mathrm{Int}(S)$ to denote the interior of S)

Lemma 2.26. *$K \cap \mathbb{S}^{D-1} \subset \mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1})$, and $-K \cap \mathbb{S}^{D-1} \subset \mathrm{Int}(\mathcal{S}_- \cap \mathbb{S}^{D-1})$*

Proof. Consider any $x_c = \sum_{j=1}^n a_j x_j y_j \in K \cap \mathbb{S}^{D-1}$, For any $x_i, y_i, i \in [n]$, we have

$$\begin{aligned}\langle x_c, x_i \rangle &= \sum_{i=j}^{n} a_j \|x_j\| \left\langle \frac{x_j y_j}{\|x_j\|}, \frac{x_i y_i}{\|x_i\|} \right\rangle \frac{\|x_i\|}{y_i} \\ &\geq \mu y_i \|x_i\| \sum_{i=j}^{n} a_j \|x_j\| \begin{cases} \geq \mu X_{\min} > 0, & y_i > 0 \\ \leq -\mu X_{\min} < 0, & y_i < 0 \end{cases},\end{aligned}$$

where we use the fact that $1 = \|x_c\| = \|\sum_{j=1}^n a_j x_j y_j\| \leq \sum_{j=1}^n a_j \|x_j\|$. This already tells us $x_c \in \mathcal{S}_+ \cap \mathbb{S}^{D-1}$.

Since $f_i(z) = \langle z, x_i \rangle$ is a continuous function of $z \in \mathbb{S}^{D-1}$. There exists an open ball $\mathcal{B}(x_c, \delta_i)$ centered at x_c with some radius $\delta_i > 0$, such that $\forall z \in \mathcal{B}(x_c, \delta_i)$, one have $|f_i(z) - f_i(x_c)| \leq \frac{\mu X_{\min}}{2}$, which implies

$$\langle z, x_i \rangle \begin{cases} \geq \mu X_{\min}/2 > 0, & y_i > 0 \\ \leq -\mu X_{\min}/2 < 0, & y_i < 0 \end{cases}.$$

Hence $\cap_{i=1}^n \mathcal{B}\left(\frac{x_c}{\|x_c\|}, \delta_i\right) \in \mathcal{S}_+ \cap \mathbb{S}^{D-1}$. Therefore, $x_c \in \operatorname{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1})$. This suffices to show $K \cap \mathbb{S}^{D-1} \subset \operatorname{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1})$. The other statement $-K \cap \mathbb{S}^{D-1} \subset \operatorname{Int}(\mathcal{S}_- \cap \mathbb{S}^{D-1})$ is proved similarly. □

The following two lemmas are some direct results of Lemma 2.26.

Lemma 2.27. $\exists \zeta_1 > 0$ *such that*

$$\mathcal{S}_{x_+}^{\zeta_1} \subset \mathcal{S}_+, \qquad \mathcal{S}_{x_-}^{\zeta_1} \subset \mathcal{S}_-, \tag{2.182}$$

where $\mathcal{S}_x^\zeta := \{z \in \mathbb{R}^D : \cos(z, x) \geq \sqrt{1-\zeta}\}$.

Proof. By Lemma 2.26, $\frac{x_+}{\|x_+\|} \in K \subset \operatorname{Int}(S_+)$. Since $\mathbb{S}^{D-1}$ is a complete metric space (w.r.t $\arccos\langle \cdot, \cdot \rangle$), there exists a open ball centered at $\frac{x_+}{\|x_+\|}$ of some radius $\arccos(\sqrt{1-\zeta_1})$ that is a subset of $\mathcal{S}_+$, from which one can show $\mathcal{S}_{x_+}^{\zeta_1} \subset \mathcal{S}_+$. The other statement $\mathcal{S}_{x_-}^{\zeta_1} \subset \mathcal{S}_-$ simply comes from the fact that $x_+ = -x_-$ and $\operatorname{Int}(\mathcal{S}_+) = -\operatorname{Int}(\mathcal{S}_-)$. □

Lemma 2.28. $\exists \xi > 0$*, such that*

$$\sup_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathcal{S}_+ \cap \mathbb{S}^{D-1})^c \cap (\mathcal{S}_- \cap \mathbb{S}^{D-1})^c} |\cos(x_1, x_2)| \leq \sqrt{1-\xi}\,. \tag{2.183}$$

(S^c here is defined to be $\mathbb{S}^{D-1} - S$, the set complement w.r.t. complete space $\mathbb{S}^{D-1}$)

Proof. Notice that

$$\sup_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c} \langle x_1, x_2 \rangle = \inf_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c} \arccos \langle x_1, x_2 \rangle \ .$$

Since $\mathbb{S}^{D-1}$ is a complete metric space (w.r.t $\arccos \langle \cdot, \cdot \rangle$) and $K \cap \mathbb{S}^{D-1}$ and $x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c$ are two of its compact subsets. Suppose

$$\inf_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c} \arccos \langle x_1, x_2 \rangle = 0\,,$$

then $\exists x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c$ such that $\arccos \langle x_1, x_2 \rangle = 0$, i.e., $x_1 = x_2$, which contradicts the fact that $K \cap \mathbb{S}^{D-1} \subseteq \mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1})$ (Lemma 2.26). Therefore, we have the infimum strictly larger than zero, then

$$\sup_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (S_+ \cap \mathbb{S}^{D-1})^c} \langle x_1, x_2 \rangle \leq \sup_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathrm{Int}(\mathcal{S}_+ \cap \mathbb{S}^{D-1}))^c} \langle x_1, x_2 \rangle < 1\,. \tag{2.184}$$

Similarly, one can show that

$$\sup_{x_1 \in -K \cap \mathbb{S}^{D-1}, x_2 \in (\mathcal{S}_- \cap \mathbb{S}^{D-1})^c} \langle x_1, x_2 \rangle < 1\,. \tag{2.185}$$

Finally, find $\xi < 1$ such that

$$\max \left\{ \sup_{x_1 \in K \cap \mathbb{S}^{D-1}, x_2 \in (\mathcal{S}_+ \cap \mathbb{S}^{D-1})^c} \langle x_1, x_2 \rangle\,, \sup_{x_1 \in -K \cap \mathbb{S}^{D-1}, x_2 \in (\mathcal{S}_- \cap \mathbb{S}^{D-1})^c} \langle x_1, x_2 \rangle \right\} = \sqrt{1-\xi}\,,$$

then for any $x_1 \in K \cap \mathbb{S}^{D-1}$ and $x_2 \in (\mathcal{S}_+ \cap \mathbb{S}^{D-1})^c \cap (\mathcal{S}_- \cap \mathbb{S}^{D-1})^c$, we have

$$-\sqrt{1-\xi} \leq \langle x_1, x_2 \rangle \leq \sqrt{1-\xi}\,,$$

which is the desired result. □

The remaining two lemmas are technical but extensively used in the main proof.

Lemma 2.29. *Consider any solution to the gradient flow dynamic (2.164) starting from initialization (2.165). Let $x_r \in \mathbb{S}^{n-1}$ be some reference direction, we define*

$$\psi_{rj} = \left\langle x_r, \frac{w_j}{\|w_j\|} \right\rangle, \ \psi_{ra} = \left\langle x_r, \frac{x_a(w_j)}{\|x_a(w_j)\|} \right\rangle, \ \psi_{aj} = \left\langle \frac{w_j}{\|w_j\|}, \frac{x_a(w_j)}{\|x_a(w_j)\|} \right\rangle, \tag{2.186}$$

where $x_a(w_j) = \sum_{i:\langle x_i, w_j\rangle > 0} y_i x_i$.

Whenever $\max_i |f(x_i; W, v)| \leq 1$, we have

$$\left| \frac{d}{dt}\psi_{rj} - \operatorname{sign}(v_j(0)) \left(\psi_{ra} - \psi_{rj}\psi_{aj}\right) \|x_a(w_j)\| \right| \leq 2nX_{\max} \max_i |f(x_i; W, v)|. \tag{2.187}$$

Proof. A simple application of Lemma 2.23, together with Cauchy-Schwartz:

$$\begin{aligned}
&\left| \frac{d}{dt}\psi_{rj} - \operatorname{sign}(v_j(0)) \left(\psi_{ra} - \psi_{rj}\psi_{aj}\right) \|x_a(w_j)\| \right| \\
=& \left| x_r^\top \left(\frac{d}{dt}\frac{w_j}{\|w_j\|} - \operatorname{sign}(v_j(0)) \left(I - \frac{w_j w_j^\top}{\|w_j\|^2} \right) \left(\sum_{i:\langle x_i, w_j\rangle > 0} y_i x_i \right) \right) \right| \\
\leq& \ 2nX_{\max} \max_i |f(x_i; W, v)|.
\end{aligned}$$

□

Lemma 2.30.

$$\|x_a(w)\| \geq \sqrt{\mu} n_a(w) X_{\min}, \tag{2.188}$$

where $n_a(w) = |\{i \in [n] : \langle x_i, w\rangle > 0\}|$.

Proof. Let $\mathcal{I}_a(w)$ denote $\{i \in [n] : \langle x_i, w\rangle > 0\}$, then

$$\begin{aligned}
\|x_a(w)\| = \left\| \sum_{i:\langle x_i, w\rangle > 0} x_i y_i \right\| &= \sqrt{\sum_{i \in \mathcal{I}_a(w)} \|x_i\|^2 y_i^2 + \sum_{i,j \in \mathcal{I}_a(w), i<j} \|x_i\|\|x_j\| \left\langle \frac{x_i y_i}{\|x_i\|}, \frac{x_j y_j}{\|x_j\|} \right\rangle} \\
&\geq \sqrt{\sum_{i \in \mathcal{I}_a(w)} \|x_i\|^2 y_i^2 + \sum_{i,j \in \mathcal{I}_a(w), i<j} \|x_i\|\|x_j\||y_i||y_j|\mu} \\
&\geq \sqrt{n_a(w) X_{\min}^2 + \mu n_a(w) \left(n_a(w) - 1\right) X_{\min}^2} \\
&\geq \sqrt{n_a(w)(1 + \mu(n_a(w) - 1))} X_{\min} \\
&\geq \sqrt{\mu} n_a(w) X_{\min}.
\end{aligned}$$

□

Proof for early alignment phase

Proof of Theorem 2.10: First Part. Given some initialization in (2.165), by Assumption 2.3, $\exists \zeta_2 > 0$, such that

$$\max_{j\in\mathcal{V}_+} \cos(w_j(0), x_-) < \sqrt{1-\zeta_2}, \quad \max_{j\in\mathcal{V}_-} \cos(w_j(0), x_+) < \sqrt{1-\zeta_2}\,. \tag{2.189}$$

We define $\zeta := \max\{\zeta_1, \zeta_2\}$, where ζ_1 is from Lemma 2.27. In addition, by Lemma 2.28, $\exists \xi > 0$, such that

$$\sup_{x_1\in K\cap\mathbb{S}^{D-1}, x_2\in\mathcal{S}_-^c\cap\mathcal{S}_+^c\cap\mathbb{S}^{D-1}} |\cos(x_1, x_2)| \leq \sqrt{1-\xi}\,. \tag{2.190}$$

We pick a initialization scale ϵ that satisfies:

$$\begin{aligned}\epsilon &\leq \min\left\{\frac{\min\{\mu,\zeta,\xi\}\sqrt{\mu}X_{\min}}{4\sqrt{h}nX_{\max}^2W_{\max}^2}, \frac{1}{\sqrt{h}}\exp\left(-\frac{64nX_{\max}}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}\log n\right)\right\}\\ &\leq \frac{1}{4\sqrt{h}X_{\max}W_{\max}^2}\,.\end{aligned} \tag{2.191}$$

By Lemma 2.22, $\forall t \leq T = \frac{1}{4nX_{\max}}\log\frac{1}{\sqrt{h}\epsilon}$, we have

$$\max_i |f(x_i; W, v)| \leq \frac{\min\{\mu,\zeta,\xi\}\sqrt{\mu}X_{\min}}{4nX_{\max}}\,, \tag{2.192}$$

which is the key to analyzing the alignment phase. For the sake of simplicity, we only discuss the analysis of neurons in $\mathcal{V}_+$ here, the proof for neurons in $\mathcal{V}_-$ is almost identical.

Activation pattern evolution: Pick any w_j in $\mathcal{V}_+$ and pick $x_r = x_iy_i$ for some $i \in [n]$, and consider the case when $\langle w_j, x_i\rangle = 0$. From Lemma 2.29,we have

$$\left|\frac{d}{dt}\psi_{rj} - (\psi_{ra} - \psi_{rj}\psi_{aj})\,\|x_a(w_j)\|\right| \leq 2nX_{\max}\max_i |f(x_i; W, v)|\,.$$

$\langle w_j, x_i\rangle = 0$ implies $\psi_{rj} = \left\langle \frac{x_iy_i}{\|x_i\|}, \frac{w_j}{\|w_j\|}\right\rangle = 0$, thus we have

$$\left|\frac{d}{dt}\psi_{rj}|_{\langle w_j,x_i\rangle=0} - \psi_{ra}\|x_a(w_j)\|\right| \leq 2nX_{\max}\max_i |f(x_i; W, v)|\,.$$

Then whenever $w_j \notin \mathcal{S}_{\text{dead}}$, we have

$$\begin{aligned}\frac{d}{dt}\psi_{rj}|_{\langle w_j,x_i\rangle=0} &\geq \psi_{ra}\|x_a(w_j)\| - 2nX_{\max}\max_i|f(x_i;W,v)| \\ &\geq \mu\|x_a(w_j)\| - 2nX_{\max}\max_i|f(x_i;W,v)| && \text{(by Lemma 2.24)} \\ &\geq \mu^{3/2}X_{\min} - 2nX_{\max}\max_i|f(x_i;W,v)| && \text{(by Lemma 2.30)} \\ &\geq \mu^{3/2}X_{\min}/2 > 0\,. && \text{(by (2.192))}\end{aligned}$$

This is precisely (2.170) in our proof sketch.

Bound on activation transitions and duration: Next we show that if at time $t_0 < T$, $w_j(t_0) \notin \mathcal{S}_+ \cup \mathcal{S}_{\text{dead}}$, and the activation pattern of w_j is $\mathbb{1}_{\langle x_i, w_j(t_0)\rangle>0}$, then $\mathbb{1}_{\langle x_i, w_j(t_0+\Delta t))\rangle>0} \neq \mathbb{1}_{\langle x_i, w_j(t_0)\rangle>0}$, where $\Delta t = \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}n_a(w_j(t_0))}$ and $n_a(w_j(t_0))$ is defined in Lemma 2.30 as long as $t_0 + \Delta t < T$ as well. That is, during the alignment phase $[0, T]$, w_j must change its activation pattern within Δt time. There are two cases:

- The first case is when $w_j(t_0) \in \mathcal{S}_+^c \cap \mathcal{S}_-^c \cap \mathcal{S}_{\text{dead}}^c$. In this case, suppose that $\mathbb{1}_{\langle x_i, w_j(t_0+\tau))\rangle>0} = \mathbb{1}_{\langle x_i, w_j(t_0)\rangle>0}, \forall 0 \leq \tau \leq \Delta t$, i.e. w_j fixes its activation during $[t_0, t_0+\Delta t]$, then we have $x_a(w_j(t_0+\tau)) = x_a(w_j(t_0)), \forall 0 \leq \tau \leq \Delta t$. Let us pick $x_r = x_a(w_j(t_0))$, then Lemma 2.29 leads to

  $$\left|\frac{d}{dt}\cos(w_j, x_a(w_j)) - \left(1-\cos^2(w_j, x_a(w_j))\right)\|x_a(w_j)\|\right| \leq 2nX_{\max}\max_i|f(x_i;W,v)|\,.$$

  Since $x_a(w_j)$ is fixed, we have $\forall t \in [t_0, t_0+\Delta t]$,

  $$\begin{aligned}&\left|\frac{d}{dt}\cos(w_j, x_a(w_j(t_0))) - \left(1-\cos^2(w_j, x_a(w_j(t_0)))\right)\|x_a(w_j(t_0))\|\right| \\ &\leq 2nX_{\max}\max_i|f(x_i;W,v)|\,,\end{aligned}$$

$$\begin{aligned}
&\frac{d}{dt}\cos(w_j, x_a(w_j(t_0)))\\
\geq\;& \left(1-\cos^2(w_j, x_a(w_j(t_0)))\right)\|x_a(w_j(t_0))\|\\
&\qquad - 2nX_{\max}\max_i|f(x_i;W,v)|\\
\geq\;& \xi\|x_a(w_j(t_0))\| - 2nX_{\max}\max_i|f(x_i;W,v)| && \text{(by (2.190))}\\
\geq\;& \xi\sqrt{\mu}n_a(w_j(t_0))X_{\min} - 2nX_{\max}\max_i|f(x_i;W,v)| && \text{(by Lemma 2.30)}\\
\geq\;& \xi\sqrt{\mu}n_a(w_j(t_0))X_{\min}/2\,. && \text{(by (2.192))}\\
\geq\;& \min\{\xi,\zeta\}\sqrt{\mu}n_a(w_j(t_0))X_{\min}/2\,,
\end{aligned}$$

which implies that, by the Fundamental Theorem of Calculus,

$$\begin{aligned}
&\cos(w_j(t_0+\Delta t), x_a(w_j(t_0)))\\
=\;& \cos(w_j(t_0), x_a(w_j(t_0))) + \int_0^{\Delta t}\frac{d}{dt}\cos(w_j(t_0+\tau), x_a(w_j(t_0)))d\tau\\
\geq\;& \cos(w_j(t_0), x_a(w_j(t_0))) + \Delta t\cdot\min\{\xi,\zeta\}\sqrt{\mu}n_a(w_j(t_0))X_{\min}/2\\
=\;& \cos(w_j(t_0), x_a(w_j(t_0))) + 2 \geq 1\,,
\end{aligned}$$

which leads to $\cos(w_j(t_0+\Delta t), x_a(w_j(t_0))) = 1$. This would imply $w_j(t_0+\Delta t)\in\mathcal{S}_+$ because $x_a(w_j(t_0))\in\mathcal{S}_+$, which contradicts our original assumption that w_j fixes the activation pattern. Therefore, $\exists 0<\tau_0\leq\Delta t$ such that $\mathbb{1}_{\langle x_i, w_j(t_0+\tau_0))\rangle}\neq\mathbb{1}_{\langle x_i, w_j(t_0)\rangle>0}$, due to the restriction on how w_j can change its activation pattern, it cannot return to its previous activation pattern, then one must have $\mathbb{1}_{\langle x_i, w_j(t_0+\Delta t))\rangle}\neq\mathbb{1}_{\langle x_i, w_j(t_0)\rangle>0}$.

- The other case is when $w_j(t_0)\in\mathcal{S}_-$. For this case, we need first show that $w_j(t_0+\tau)\notin\mathcal{S}^\zeta_{x_-}, \forall 0\leq\tau\leq\Delta t$, or more generally, $\mathcal{S}^\zeta_{x_-}$ does not contain any w_j in $\mathcal{V}_+$ during $[0,T]$. To see this, let us pick $x_r = x_-$, then Lemma 2.29 suggests that

$$\left|\frac{d}{dt}\psi_{rj} - (\psi_{ra}-\psi_{rj}\psi_{aj})\|x_a(w_j)\|\right| \leq 2nX_{\max}\max_i|f(x_i;W,v)|\,.$$

Consider the case when $\cos(w_j, x_-)=\sqrt{1-\zeta}$, i.e. w_j is at the boundary of $\mathcal{S}^\zeta_{x_-}$.

We know that in this case, $w_j \in \mathcal{S}^{\zeta}_{x_-} \subseteq \mathcal{S}_-$ thus $x_a(w_j) = -x_-$, and

$$\left|\frac{d}{dt}\cos(w_j, x_-)\right|_{\cos(w_j,x_-)=\sqrt{1-\zeta}} + \left(1-\cos^2(w_j,x_-)\right)\|x_-\|\Bigg| \\ \leq 2nX_{\max}\max_i |f(x_i;W,v)|\,,$$

which is

$$\begin{aligned} \left|\frac{d}{dt}\cos(w_j, x_-)\right|_{\cos(w_j,x_-)=\sqrt{1-\zeta}} + \zeta\|x_-\|\Bigg| \leq & \quad 2nX_{\max}\max_i |f(x_i;W,v)| \\ \Rightarrow \frac{d}{dt}\cos(w_j,x_-)\Big|_{\cos(w_j,x_-)=\sqrt{1-\zeta}} & \\ \leq & -\zeta\|x_-\| + 2nX_{\max}\max_i |f(x_i;W,v)| \\ \leq & -\zeta\sqrt{\mu}X_{\min} + 2nX_{\max}\max_i |f(x_i;W,v)| && \text{(by Lemma 2.30)} \\ \leq & -\zeta\sqrt{\mu}X_{\min}/2 < 0\,. && \text{(by (2.192))} \end{aligned}$$

Therefore, during $[0,T]$, neuron w_j in $\mathcal{V}_+$ cannot enter $\mathcal{S}^{\zeta}_{x_-}$ if at initialization, $w_j(0) \notin \mathcal{S}^{\zeta}_{x_-}$, which is guaranteed by (2.189).

With the argument above, we know that $w_j(t_0+\tau) \notin \mathcal{S}^{\zeta}_{x_-}, \forall 0 \leq \tau \leq \Delta t$. Again we suppose that $w_j(t) \in \mathcal{S}_- - \mathcal{S}^{\zeta}_{x_-}, \forall t \in [t_0, t_0+\Delta t]$, i.e.,$w_j$ fixes its activation during $[t_0, t_0+\Delta t]$. Let us pick $x_r = x_-$, then Lemma 2.29 suggests that

$$\left|\frac{d}{dt}\cos(w_j,x_-) + \left(1-\cos^2(w_j,x_-)\right)\|x_-\|\right| \leq 2nX_{\max}\max_i |f(x_i;W,v)|\,,$$

which leads to $\forall t \in [t_0, t_0+\Delta t]$,

$$\begin{aligned} & \frac{d}{dt}\cos(w_j,x_-) \\ \leq & -\left(1-\cos^2(w_j,x_-)\right)\|x_-\| + 2nX_{\max}\max_i |f(x_i;W,v)| \\ \leq & -\zeta\|x_-\| + 2nX_{\max}\max_i |f(x_i;W,v)| && (w_j \notin \mathcal{S}^{\zeta}_{x_-}) \\ \leq & -\zeta\sqrt{\mu}n_a(w_j(t_0))X_{\min} + 2nX_{\max}\max_i |f(x_i;W,v)| && \text{(by Lemma 2.30)} \\ \leq & -\zeta\sqrt{\mu}n_a(w_j(t_0))X_{\min}/2\,. && \text{(by (2.192))} \\ \leq & -\min\{\xi,\zeta\}\sqrt{\mu}n_a(w_j(t_0))X_{\min}/2\,, \end{aligned}$$

Similarly, by FTC, we have

$$\cos(w_j(t_0 + \Delta t), x_-) \leq -1\,.$$

This would imply $w_j(t_0 + \Delta t) \in \mathcal{S}_+$ because $-x_- = x_a(w_j(t_0)) \in \mathcal{S}_+$, which contradicts our original assumption that w_j fixes its activation pattern. Therefore, one must have $\mathbb{1}_{\langle x_i, w_j(t_0+\Delta t)\rangle\rangle} \neq \mathbb{1}_{\langle x_i, w_j(t_0)\rangle\rangle > 0}$.

In summary, we have shown that, during $[0, T]$, a neuron in $\mathcal{V}_+$ can not keep a fixed activation pattern for a time longer than $\Delta t = \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}n_a}$, where n_a is the number of data points that activate w_j under the fixed activation pattern.

Bound on total travel time until directional convergence: As we have discussed in the proof sketch and also formally proved here, during alignment phase $[0, T]$, a neuron in $\mathcal{V}_+$ must change its activation pattern within $\Delta t = \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}n_a}$ time unless it is in either $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$. And the new activation it is transitioning into must contain no new activation on negative data points and must keep all existing activation on positive data points, together it shows that a neuron must reach either $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ within a fixed amount of time, which is the remaining thing we need to formally show here.

For simplicity of the argument, we first assume $T = \infty$, i.e., the alignment phase lasts indefinitely, and we show that a neuron in $\mathcal{V}_+$ must reach $\mathcal{S}_+$ or $\mathcal{S}_{\text{dead}}$ before $t_1 = \frac{16\log n}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}$. Lastly, such directional convergence can be achieved if $t_1 \leq T$, which is guaranteed by our choice of ϵ in (2.191).

- For a neuron in $\mathcal{V}_+$ that reaches $\mathcal{S}_{\text{dead}}$, the analysis is easy: It must start with no activation on positive data and then lose activation on negative data one by one until losing all of its activation. Therefore, it must reach $\mathcal{S}_{\text{dead}}$ before

$$\begin{aligned}\sum_{k=1}^{n_a(w_j(0))}\frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}k} &\leq \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}\left(\sum_{k=1}^{n}\frac{1}{k}\right)\\ &\leq \frac{16\log n}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}=t_1\,.\end{aligned}$$

- For a neuron in $\mathcal{V}_+$ that reaches $\mathcal{S}_+$, there is no difference conceptually, but it can switch its activation pattern in many ways before reaching $\mathcal{S}_+$, so it is not straightforward to see its travel time until $\mathcal{S}_+$ is upper bounded by t_1.

  To formally show the upper bound on the travel time, we need some definition of a path that keeps a record of the activation patterns of a neuron $w_j(t)$ before it reaches $\mathcal{S}_+$.

  Let $n_+=|\mathcal{I}_+|$, $n_-=|\mathcal{I}_-|$ be the number of positive, negative data respectively, then we call $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ a *path* of length-L, if

  1. $\forall 0\leq l\leq L$, we have $k^{(l)}=(k_+^{(l)},k_-^{(l)})\in\mathbb{N}\times\mathbb{N}$ with $0\leq k_+^{(l)}\leq n_+$, $0\leq k_-^{(l)}\leq n_-$;
  2. For $k^{(l_1)},k^{(l_2)}$ with $l_1<l_2$, we have either $k_+^{(l_1)}>k_+^{(l_2)}$ or $k_-^{(l_1)}<k_-^{(l_2)}$;
  3. $k^{(L)}=(n_+,0)$;
  4. $k^{(l)}\neq(0,0),\forall 0\leq l\leq L$.

  Given all our analysis on how a neuron $w_j(t)$ can switch its activation pattern in previous parts, we know that for any $w_j(t)$ that reaches $\mathcal{S}_+$, there is an associated $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ that keeps an ordered record of encountered values of

  $$\left(|\{i\in\mathcal{I}_+:\langle x_i,w_j(t)\rangle>0\}|,\ |\{i\in\mathcal{I}_-:\langle x_i,w_j(t)\rangle>0\}|\right)\,,$$

  before w_j reaches $\mathcal{S}_+$. That is, a neuron w_j starts with some activation pattern that activates $k_+(0)$ positive data and $k_-(0)$ negative data, then switch its activation pattern (by either losing negative data or gaining positive data) to one that activates $k_+(1)$ positive data and $k_-(1)$ negative data. By keep doing so, it reaches $\mathcal{S}_+$

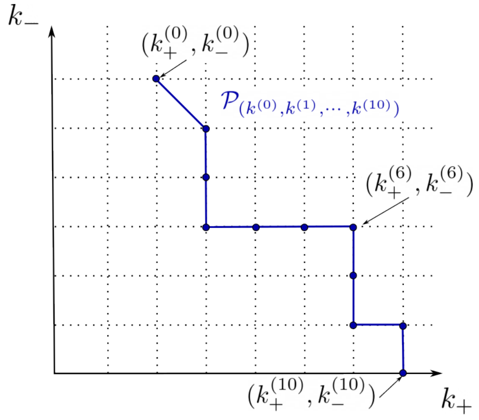
k_-
$(k_+^{(0)}, k_-^{(0)})$
$\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(10)})}$
$(k_+^{(6)}, k_-^{(6)})$
$(k_+^{(10)}, k_-^{(10)})$
k_+

Figure 2-13. Illustration of a path of length-10. Each dot on the grid represents one $k^{(l)}$.

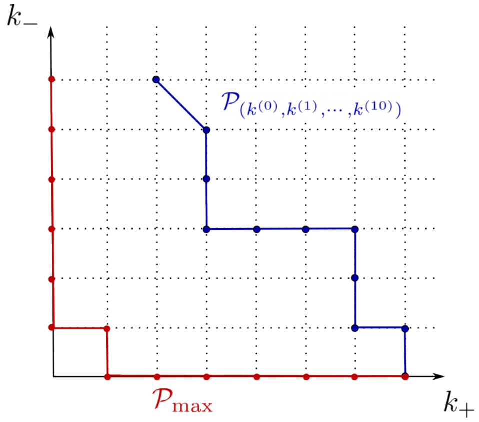

Figure 2-14. Illustration of a path and the maximal path

that activates $k_+(L) = n_+$ positive data and $k_-(L) = 0$ negative data. Please see Figure 2-13 for an illustration of a path.

Given a path $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ of neuron w_j, we define the *travel time* of this path as

$$T(\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}) = \sum_{l=0}^{L-1} \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}(k_+^{(l)} + k_-^{(l)})},$$

which is the traveling time from $k^{(0)}$ to $k^{(L)}$ if one spends $\frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}(k_+^{(l)}+k_-^{(l)})}$ on the edge between $k^{(l)}$ and $k^{(l+1)}$.

Our analysis shows that if w_j reaches $\mathcal{S}_+$, then

$$\inf\{t : w_j(t) \in \mathcal{S}_+\} \leq T(\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}).$$

Now we define the maximal path $\mathcal{P}_{\max}$ as a path that has the maximum length $n = n_+ + n_-$, which is uniquely determined by the following trajectory of $k^{(l)}$

$$(0, n_-), (0, n_- - 1), (0, n_- - 2), \cdots, (0, 1), (1, 1), (1, 0), \cdots, (n_+ - 1, 0), (n_+, 0).$$

Please see Figure 2-14 for an illustration.

The traveling time for $\mathcal{P}_{\max}$ is

$$\begin{aligned} T(\mathcal{P}_{\max}) &= \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}\left(\sum_{k=1}^{n_-}\frac{1}{k}+\frac{1}{2}+\sum_{k=1}^{n_+-1}\frac{1}{k}\right) \\ &\le \frac{4}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}}\left(2\sum_{k=1}^{n}\frac{1}{k}+\frac{1}{2}\right) \\ &\le \frac{16\log n}{\min\{\zeta,\xi\}\sqrt{\mu}X_{\min}} = t_1\,. \end{aligned}$$

The proof is complete by the fact that any path satisfies

$$T(\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}) \le T(\mathcal{P}_{\max})\,.$$

This is because there is a one-to-one correspondence between the edges $(k^{(l)},k^{(l+1)})$ in $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ and a subset of edges in $\mathcal{P}_{\max}$, and the travel time from of edge $(k^{(l)},k^{(l+1)})$ is shorter than the corresponding edge in $\mathcal{P}_{\max}$. Formally stating such correspondence is tedious and a visual illustration in Figure 2-15 and 2-16 is more effective (Putting all correspondence makes a clustered plot thus we split them into two figures):

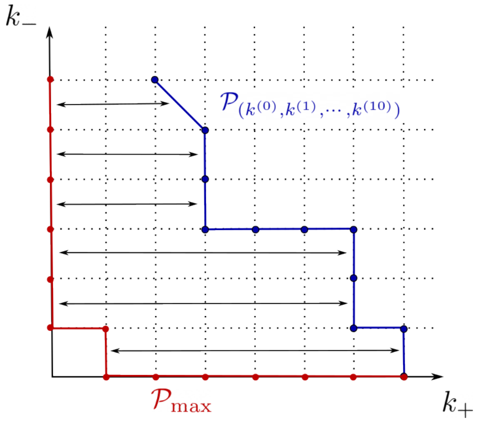

Figure 2-15. Correspondence between edges in $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ and $\mathcal{P}_{\max}$. (Part 1)

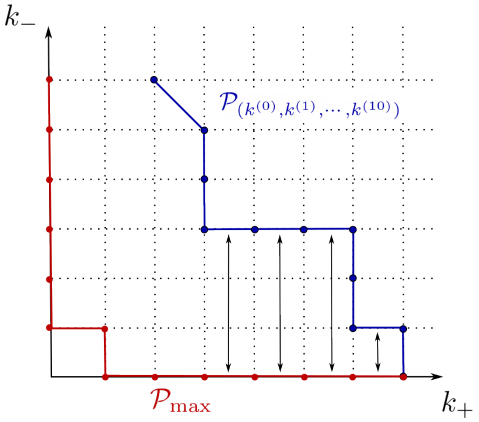

Figure 2-16. Correspondence between edges in $\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}$ and $\mathcal{P}_{\max}$. (Part 2)

Therefore, if w_j reaches $\mathcal{S}_+$, then it reaches $\mathcal{S}_+$ within t_1:

$$\inf\{t : w_j(t)\in\mathcal{S}_+\} \le T(\mathcal{P}_{(k^{(0)},k^{(1)},\cdots,k^{(L)})}) \le T(\mathcal{P}_{\max}) \le t_1\,.$$

So far we have shown when the alignment phase lasts long enough, i.e., T large enough, the directional convergence is achieved by t_1. We simply pick ϵ such that

$$T = \frac{1}{4nX_{\max}} \log \frac{1}{\sqrt{h}\epsilon} \geq t_1 = \frac{16 \log n}{\min\{\zeta, \xi\}\sqrt{\mu}X_{\min}}\,,$$

and (2.191) suffices. □

Proof for Theorem 2.10 (part two): final convergence

Since we have proved the first part of Theorem 2.10 in Section 2.3.3, we will use it as a fact, then prove the remaining part of Theorem 2.10.

Auxiliary lemmas

First, we show that $\mathcal{S}_+, \mathcal{S}_-, \mathcal{S}_{\text{dead}}$ are trapping regions.

Lemma 2.31. *Consider any solution to the gradient flow dynamic (2.164), we have the following:*

- *If at some time $t_1 \geq 0$, we have $w_j(t_1) \in \mathcal{S}_{dead}$, then $w_j(t_1 + \tau) \in \mathcal{S}_{dead},\ \forall \tau \geq 0$;*
- *If at some time $t_1 \geq 0$, we have $w_j(t_1) \in \mathcal{S}_+$ for some $j \in \mathcal{V}_+$, then $w_j(t_1+\tau) \in \mathcal{S}_+,\ \forall \tau \geq 0$;*
- *If at some time $t_1 \geq 0$, we have $w_j(t_1) \in \mathcal{S}_-$ for some $j \in \mathcal{V}_-$, then $w_j(t_1+\tau) \in \mathcal{S}_-,\ \forall \tau \geq 0$;*

Proof. The first statement is simple, if $w_j \in \mathcal{S}_{\text{dead}}$, then one have $\dot{w}_j = 0$, thus w_j remains in $\mathcal{S}_{\text{dead}}$.

For the second statement, we have, since $j \in \mathcal{V}_+$,

$$\frac{d}{dt} w_j = -\sum_{i=1}^{n} \mathbb{1}_{\langle x_i, w_j \rangle \geq 0} \nabla_{\hat{y}} \ell(y_i, f(x_i; W, v)) x_i \|w_j\|\,.$$

By the Fundamental Theorem of Calculus, one writes, $\forall \tau \geq 0$,

$$\begin{aligned} w_j(t_1+\tau) =&\ w_j(t_1) + \int_0^\tau \frac{d}{dt} w_j d\tau \\ =&\ w_j(t_1) + \int_0^\tau -\sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \geq 0} \nabla_{\hat{y}} \ell(y_i, f(x_i; W, v)) x_i \|w_j\| d\tau \\ =&\ w_j(t_1) + \int_0^\tau \sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \geq 0} y_i \exp(-y_i f(x_i; W, v)) x_i \|w_j\| d\tau \\ =&\ w_j(t_1) + \underbrace{\sum_{i\in\mathcal{I}_+} \left(\int_0^\tau \exp(-y_i f(x_i; W, v)) \|w_j\| d\tau\right) x_i}_{:=\tilde{x}_+} . \end{aligned}$$

Here $w_j(t_1) \in \mathcal{S}_+$ by our assumption, $\tilde{x}_+ \in K \subseteq \mathcal{S}_+$ because $\tilde{x}_+$ is a conical combination of $x_i, i \in \mathcal{I}_+$. Since $\mathcal{S}_+$ is a convex cone, we have $w_j(t_1+\tau) \in \mathcal{S}_+$ as well.

The proof of the third statement is almost identical: when $j \in \mathcal{V}_-$, we have

$$\frac{d}{dt} w_j = \sum_{i=1}^n \mathbb{1}_{\langle x_i, w_j\rangle \geq 0} \nabla_{\hat{y}} \ell(y_i, f(x_i; W, v)) x_i \|w_j\| ,$$

and

$$w_j(t_1+\tau) = w_j(t_1) + \underbrace{\sum_{i\in\mathcal{I}_-} \left(\int_0^\tau \exp(-y_i f(x_i; W, v)) \|w_j\| d\tau\right) x_i}_{:=\tilde{x}_-} .$$

Again, here $w_j(t_1) \in \mathcal{S}_-$ by our assumption, $\tilde{x}_- \in -K \subseteq \mathcal{S}_-$ because $\tilde{x}_-$ is a conical combination of $x_i, i \in \mathcal{I}_-$. Since $\mathcal{S}_-$ is a convex cone, we have $w_j(t_1+\tau) \in \mathcal{S}_+$ as well. □

Then the following Lemma provides a lower bound on neuron norms upon t_1.

Lemma 2.32. *Consider any solution to the gradient flow dynamic* (2.164) *starting from initialization* (2.165)*. Let t_1 be the time when directional convergence is achieved, as defined in Theorem 2.10, and we define $\tilde{\mathcal{V}}_+ : \{j : w_j(t_1) \in \mathcal{S}_+\}$ and $\tilde{\mathcal{V}}_- : \{j : w_j(t_1) \in \mathcal{S}_-\}$. If both $\tilde{\mathcal{V}}_+$ and $\tilde{\mathcal{V}}_-$ are non-empty, we have*

$$\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j(t_1)\|^2 \geq \exp(-4nX_{\max} t_1) \sum_{j\in\tilde{\mathcal{V}}_+} \|w_j(0)\|^2,$$

$$\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(t_1)\|^2 \geq \exp(-4nX_{\max}t_1)\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(0)\|^2,$$

Proof. We have shown that

$$\frac{d}{dt}\|w_j\|^2 = -2\sum_{i=1}^n \mathbb{1}_{\langle x_i,w_j\rangle\geq 0}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\,\langle x_i,w_j\rangle\,\mathrm{sign}(v_j(0))\|w_j\|\,.$$

Then before t_1, we have $\forall j\in[h]$

$$\begin{aligned}\frac{d}{dt}\|w_j\|^2 =&\ -2\sum_{i=1}^n \mathbb{1}_{\langle x_i,w_j\rangle\geq 0}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\,\langle x_i,w_j\rangle\,\mathrm{sign}(v_j(0))\|w_j\|\\ \geq&\ -2\sum_{i=1}^n(|y_i|+2\max_i|f(x_i;W,v)|)\|x_i\|\|w_j\|^2\\ \geq&\ -4\sum_{i=1}^n\|x_i\|\|w_j\|^2\geq -4nX_{\max}\|w_j\|^2\,,\end{aligned}$$

where the second last inequality is because $\max_i|f(x_i;W,v)|\leq\frac{1}{2}$ before t_1. Summing over $j\in\tilde{\mathcal{V}}_+$, we have

$$\frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\geq -4nX_{\max}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\,.$$

Therefore, we have the following bound:

$$\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_1)\|^2 \geq \exp(-4nX_{\max}t_1)\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(0)\|^2\,.$$

□

Moreover, after t_1, the neuron norms are non-decreasing, as suggested by

Lemma 2.33. *Consider any solution to the gradient flow dynamic (2.164) starting from initialization (2.165). Let t_1 be the time when directional convergence is achieved, as defined in Theorem 2.10, and we define $\tilde{\mathcal{V}}_+:\{j:w_j(t_1)\in\mathcal{S}_+\}$ and $\tilde{\mathcal{V}}_-:\{j:w_j(t_1)\in\mathcal{S}_-\}$. If both $\tilde{\mathcal{V}}_+$ and $\tilde{\mathcal{V}}_-$ are non-empty, we have $\forall\tau\geq 0$ and $t_2\geq t_1$,*

$$\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_2+\tau)\|^2\geq\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_2)\|,\qquad\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(t_2+\tau)\|^2\geq\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(t_2)\|\qquad(2.193)$$

Proof. It suffices to show that after t_1, the following derivatives:

$$\frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t)\|^2,\quad \frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(t)\|^2\,,$$

are non-negative.

For $j\in\tilde{\mathcal{V}}_+$, w_j stays in $\mathcal{S}_+$ by Lemma 2.31, and we have

$$\begin{aligned}\frac{d}{dt}\|w_j\|^2 =&\; -2\sum_{i\in\mathcal{I}_+}\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\,\langle x_i,w_j\rangle\,\|w_j\|\,.\\ =&\; 2\sum_{i\in\mathcal{I}_+}y_i\ell(y_i,f(x_i;W,v))\,\langle x_i,w_j\rangle\,\|w_j\|\geq 0\,.\end{aligned}$$

Summing over $j\in\tilde{\mathcal{V}}_+$, we have $\frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t)\|^2\geq 0$. Similarly, for neurons in $\tilde{\mathcal{V}}_-$, one has $\frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_-}\|w_j(t)\|^2\geq 0$. □

Finally, the following lemma is used for deriving the final convergence.

Lemma 2.34. *Consider the following loss function*

$$\mathcal{L}_{lin}(W,v)=\sum_{i=1}^n\ell\left(y_i,v^\top W^\top x_i)\right)\,,$$

if $\{x_i,y_i\},i\in[n]$ *are linearly separable, i.e.,* $\exists\gamma>0$ *and* $z\in\mathbb{S}^{D-1}$ *such that* $y_i\langle z,x_i\rangle\geq\gamma,\forall i\in[n]$*, then under the gradient flow on* $\mathcal{L}_{lin}(W,v)$*, we have*

$$\dot{\mathcal{L}}_{lin}\leq-\|v\|^2\mathcal{L}^2\gamma^2\,. \tag{2.194}$$

Proof.

$$\begin{aligned}\dot{\mathcal{L}}=-\|\nabla_W\mathcal{L}\|_F^2-\|\nabla_v\mathcal{L}\|_F^2\leq&\;-\|\nabla_W\mathcal{L}\|_F^2\\ =&\;-\left\|\sum_{i=1}^n y_i\ell(y_i,v^\top W^\top x_i)x_iv^\top\right\|_F^2\\ =&\;-\|v\|^2\left\|\sum_{i=1}^n y_i\ell(y_i,v^\top W^\top x_i)x_i\right\|^2\\ \leq&\;-\|v\|^2\left|\left\langle z,\sum_{i=1}^n y_i\ell(y_i,v^\top W^\top x_i)x_i\right\rangle\right|^2\\ \leq&\;-\|v\|^2\left|\sum_{i=1}^n\ell(y_i,v^\top W^\top x_i)\gamma\right|^2\leq-\|v\|^2\mathcal{L}^2\gamma^2\,.\end{aligned}$$

□

Proof of final convergence

Proof of Theorem 2.10: Second Part. By Lemma 2.31, we know that after t_1, neurons in $\mathcal{S}_+$ ($\mathcal{S}_-$) stays in $\mathcal{S}_+$ ($\mathcal{S}_-$). Thus the loss can be decomposed as

$$\mathcal{L} = \underbrace{\sum_{i\in\mathcal{I}_+} \ell\left(y_i, \sum_{j\in\tilde{\mathcal{V}}_+} v_j \langle w_j, x_i\rangle\right)}_{\mathcal{L}_+} + \underbrace{\sum_{i\in\mathcal{I}_-} \ell\left(y_i, \sum_{j\in\tilde{\mathcal{V}}_-} v_j \langle w_j, x_i\rangle\right)}_{\mathcal{L}_-}, \tag{2.195}$$

where $\tilde{\mathcal{V}}_+ : \{j : w_j(t_1) \in \mathcal{S}_+\}$ and $\tilde{\mathcal{V}}_- : \{j : w_j(t_1) \in \mathcal{S}_-\}$. Therefore, the training after t_1 is decoupled into 1) using neurons in $\tilde{\mathcal{V}}_+$ to fit positive data in $\mathcal{I}_+$ and 2) using neurons in $\tilde{\mathcal{V}}_-$ to fit positive data in $\mathcal{I}_-$.

Define $f_+(x_i; W, v) = \sum_{j\in\tilde{\mathcal{V}}_+} v_j \langle w_j, x_i\rangle$ and $t_2^+ = \inf\{t : \max_{i\in\mathcal{I}_+} |f_+(x_i; W, v)| > \frac{1}{4}\}$. Similarly, we also define $f_-(x_i; W, v) = \sum_{j\in\tilde{\mathcal{V}}_+} v_j \langle w_j, x_i\rangle$ and let $t_2^- = \inf\{t : \max_{i\in\mathcal{I}_-} |f_-(x_i; W, v)| > \frac{1}{4}\}$. Then $t_1 \leq \min\{t_2^+, t_2^-\}$, by Lemma 2.22.

$\mathcal{O}(1/t)$ **convergence after** t_2: We first show that when both t_2^+, t_2^- are finite, then it implies $\mathcal{O}(1/t)$ convergence on the loss. Then we show that they are indeed finite and $t_2 := \max\{t_2^+, t_2^-\} = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$.

At $t_2 = \max\{t_2^+, t_2^-\}$, by definition, $\exists i_+ \in \mathcal{I}_+$ such that

$$\frac{1}{4} \leq f_+(x_{i_+}; W, v) \leq \sum_{j\in\tilde{\mathcal{V}}_+} v_j \langle w_j, x_{i_+}\rangle \leq \sum_{j\in\tilde{\mathcal{V}}_+} \|w_j\|^2 \|x_{i_+}\|, \tag{2.196}$$

which implies, by Lemma 2.33, $\forall t \geq t_2$

$$\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j(t)\|^2 \geq \sum_{j\in\tilde{\mathcal{V}}_+} \|w_j(t_2)\|^2 \geq \frac{1}{4\|x_{i_+}\|} \geq \frac{1}{4X_{\max}}. \tag{2.197}$$

Similarly, we have $\forall t \geq t_2$,

$$\sum_{j\in\tilde{\mathcal{V}}_-} \|w_j(t)\|^2 \geq \frac{1}{4X_{\max}}. \tag{2.198}$$

Under the gradient flow dynamics (2.164), we apply Lemma 2.34 to the decomposed loss (2.195)

$$\dot{\mathcal{L}} \leq -\left(\sum_{j\in\tilde{\mathcal{V}}_+} v_j^2\right) \cdot \mathcal{L}_+^2 \cdot (\mu X_{\min})^2 - \left(\sum_{j\in\tilde{\mathcal{V}}_+} v_j^2\right) \cdot \mathcal{L}_-^2 \cdot (\mu X_{\min})^2.$$

Here, we can pick the same $\gamma = \mu X_{\min}$ for both $\mathcal{L}_+$ and $\mathcal{L}_-$ because $\{x_i, y_i\}, i \in \mathcal{I}_+$ is linearly separable with $z = \frac{y_1 x_1}{\|x_1\|}$: $\langle z, x_i y_i\rangle \geq \mu\|x_i\| \geq \mu X_{\min}$ by Assumption 2.2. And similarly, $\{x_i, y_i\}, i \in \mathcal{I}_-$ is linearly separable with $\langle z, x_i y_i\rangle \geq \mu\|x_i\| \geq \mu X_{\min}$. Replace v_i^2 by $\|w_j\|^2$ from balancedness, together with (2.197)(2.198), we have

$$\begin{aligned}\dot{\mathcal{L}} \leq & -\left(\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\right)\cdot\mathcal{L}_+^2\cdot(\mu X_{\min})^2 - \left(\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\right)\cdot\mathcal{L}_-^2\cdot(\mu X_{\min})^2 \\ \leq & -\frac{(\mu X_{\min})^2}{4X_{\max}}(\mathcal{L}_+^2+\mathcal{L}_-^2) \leq -\frac{(\mu X_{\min})^2}{8X_{\max}}(\mathcal{L}_+ + \mathcal{L}_-)^2 = -\frac{(\mu X_{\min})^2}{8X_{\max}}\mathcal{L}^2\,,\end{aligned}$$

which is

$$\frac{1}{\mathcal{L}^2}\dot{\mathcal{L}} \leq -\frac{(\mu X_{\min})^2}{8X_{\max}}\,.$$

Integrating both side from t_2 to any $t \geq t_2$, we have

$$\left.\frac{1}{\mathcal{L}}\right|_{t_2}^{\top} \leq -\frac{(\mu X_{\min})^2}{8X_{\max}}(t-t_2)\,,$$

which leads to

$$\mathcal{L}(t) \leq \frac{\mathcal{L}(t_2)}{\mathcal{L}(t_2)\alpha(t-t_2)+1}\,,\ \text{where } \alpha = \frac{(\mu X_{\min})^2}{8X_{\max}}\,.$$

Showing $t_2 = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$: The remaining thing is to show t_2 is $\mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$.

Since after t_1, the gradient dynamics are fully decoupled into two gradient flow dynamics (on $\mathcal{L}_+$ and on $\mathcal{L}_-$), it suffices to show $t_2^+ = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$ and $t_2^- = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$ separately, then combine them to show $t_2 = \max\{t_2^+, t_2^-\} = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$. The proof is almost identical for $\mathcal{L}_+$ and $\mathcal{L}_-$, thus we only prove $t_2^+ = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$ here.

Suppose

$$t_2 \geq t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}} + \frac{4}{\sqrt{\mu}n_+X_{\min}}\left(\log\frac{2}{\epsilon^2\sqrt{\mu}X_{\min}W_{\min}^2} + 4nX_{\max}t_1\right)\,, \tag{2.199}$$

where $n_+ = |\mathcal{I}_+|$. It takes two steps to show a contradiction: First, we show that for some $t_a \geq 0$, a refined alignment $\cos(w_j(t_1+t_a), x_+) \geq \frac{1}{4}, \forall j \in \tilde{\mathcal{V}}_+$ is achieved, and such refined alignment is maintained until at least t_2^+: $\cos(w_j(t), x_+) \geq \frac{1}{4}, \forall j \in \tilde{\mathcal{V}}_+$ for all $t_1 + t_a \leq t \leq t_2^+$. Then, keeping this refined alignment leads to a contradiction.

- For $j \in \tilde{\mathcal{V}}_+$, we have

$$\frac{d}{dt}\frac{w_j}{\|w_j\|} = \left(I - \frac{w_j w_j^\top}{\|w_j\|^2}\right)\underbrace{\left(\sum_{i\in\mathcal{I}_+} -\nabla_{\hat{y}}\ell(y_i, f_+(x_i; W, v))x_i\right)}_{:=\tilde{x}_a}.$$

  Then

$$\begin{aligned}\frac{d}{dt}\cos(x_+, w_j) =& \left(\cos(x_+, \tilde{x}_a) - \cos(x_+, w_j)\cos(\tilde{x}_a, w_j)\right)\|\tilde{x}_a\| \\ \geq& \left(\cos(x_+, \tilde{x}_a) - \cos(x_+, w_j)\right)\|\tilde{x}_a\|.\end{aligned}$$

  We can show that $\cos(x_+, \tilde{x}_a) \geq \frac{1}{3}$ and $\|\tilde{x}_a\| \geq \sqrt{\mu}n_+X_{\min}/2$ when $t_1 \leq t \leq t_2^+$ (we defer the proof to the end as it breaks the flow), thus within $[t_1, t_2^+]$, we have

$$\frac{d}{dt}\cos(x_+, w_j) \geq \left(\frac{1}{3} - \cos(x_+, w_j)\right)\sqrt{\mu}n_+X_{\min}/2\,. \tag{2.200}$$

  We use (2.200) in two ways: First, since

$$\left.\frac{d}{dt}\cos(x_+, w_j)\right|_{\cos(x_+, w_j)=\frac{1}{4}} \geq \frac{\sqrt{\mu}n_+X_{\min}}{24} > 0\,,$$

  $\cos(x_+, w_j) \geq \frac{1}{4}$ is a trapping region for w_j during $[t_1, t_2^+]$. Define $t_a := \inf\{t \geq t_1 : \min_{j\in\tilde{\mathcal{V}}_+}\cos(x_+, w_j(t)) \geq \frac{1}{4}\}$, then clearly, if $t_a \leq t_2^+$, then $\cos(w_j(t), x_+) \geq \frac{1}{4}, \forall j \in \tilde{\mathcal{V}}_+$ for all $t_1 + t_a \leq t \leq t_2^+$.

  Now we use (2.200) again to show that $t_a \leq t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}}$: Suppose that $t_a \geq t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}}$, then $\exists j^*$ such that $\cos(x_+, w_{j^*}(t)) < \frac{1}{4}, \forall t \in [t_1, t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}}]$, and we have

$$\frac{d}{dt}\cos(x_+, w_{j^*}) \geq \left(\frac{1}{3} - \cos(x_+, w_j)\right)\sqrt{\mu}n_+X_{\min}/2 \geq \frac{\sqrt{\mu}n_+X_{\min}}{24}\,. \tag{2.201}$$

  This shows

$$\cos(x_+, w_{j^*}(t_1+1)) \geq \cos(x_+, w_{j^*}(t_1)) + \frac{1}{4} \geq \frac{1}{4}\,,$$

  which contradicts that $\cos(x_+, w_{j^*}(t)) < \frac{1}{4}$. Hence we know $t_a \leq t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}}$.

  In summary, we have $\cos(w_j(t), x_+) \geq \frac{1}{4}, \forall j \in \tilde{\mathcal{V}}_+$ for all $t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}} \leq t \leq t_2^+$.

- Now we check the dynamics of $\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t)\|^2$ during $t_1+\frac{6}{\sqrt{\mu}n_+X_{\min}}\le t\le t_2^+$. For simplicity, we denote $t_1+\frac{6}{\sqrt{\mu}n_+X_{\min}}:=t_1'$.

  For $j\in\tilde{\mathcal{V}}_+$, we have, for $t_1'\le t\le t_2^+$,

$$\begin{aligned}
\frac{d}{dt}\|w_j\|^2 &= 2\sum_{i\in\mathcal{I}_+}-\nabla_{\hat{y}}\ell(y_i,f(x_i;W,v))\left\langle x_i,w_j\right\rangle\|w_j\| \\
&\ge \sum_{i\in\mathcal{I}_+}\left\langle x_i,w_j\right\rangle\|w_j\| && \text{(by (2.203))}\\
&= \left\langle x_+,w_j\right\rangle\|w_j\| \\
&= \|x_+\|\|w_j\|^2\cos(x_+,w_j) \\
&\ge \frac{1}{4}\|x_+\|\|w_j\|^2 && \text{(Since } t\ge t_1'\text{)}\\
&\ge \frac{\sqrt{\mu}n_+X_{\min}}{4}\|w_j\|^2\,, && \text{(by Lemma 2.30)}
\end{aligned}$$

  which leads to (summing over $j\in\tilde{\mathcal{V}}_+$)

$$\frac{d}{dt}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\ge\frac{\sqrt{\mu}n_+X_{\min}}{4}\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2\,.$$

  By Gronwall's inequality, we have

$$\begin{aligned}
&\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_2^+)\|^2\\
&\ge \exp\left(\frac{\sqrt{\mu}n_+X_{\min}}{4}(t_2^+-t_1')\right)\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_1')\|^2\\
&\ge \exp\left(\frac{\sqrt{\mu}n_+X_{\min}}{4}(t_2^+-t_1')\right)\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(t_1)\|^2 && \text{(By Lemma 2.33)}\\
&\ge \exp\left(\frac{\sqrt{\mu}n_+X_{\min}}{4}(t_2^+-t_1')\right)\exp\left(-4nX_{\max}t_1\right)\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j(0)\|^2 && \text{(By Lemma 2.32)}\\
&\ge \exp\left(\frac{\sqrt{\mu}n_+X_{\min}}{4}(t_2^+-t_1')\right)\exp\left(-4nX_{\max}t_1\right)\epsilon^2W_{\min}^2\,. && \text{(by (2.199))}\\
&\ge \frac{2}{\sqrt{\mu}X_{\min}}
\end{aligned}$$

However, at t_2^+, we have

$$\begin{aligned}
\frac{1}{4} \geq \frac{1}{n_+}\sum_{i\in\mathcal{I}_+} f_+(x_i;W,v) = &\ \frac{1}{n_+}\sum_{i\in\mathcal{I}_+}\sum_{j\in\tilde{\mathcal{V}}_+} v_j\left\langle w_j, x_i\right\rangle \\
= &\ \frac{1}{n_+}\sum_{j\in\tilde{\mathcal{V}}_+} v_j\left\langle w_j, x_+\right\rangle * \\
= &\ \frac{1}{n_+}\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j\|^2\cos(w_j,x_+)\|x_+\| \\
\geq &\ \frac{1}{4n_+}\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j\|^2\|x_+\| && \text{(Since } t\geq t_1'\text{)} \\
\geq &\ \frac{1}{4}\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j\|^2\sqrt{\mu}X_{\min}\,, && \text{(by Lemma 2.30)}
\end{aligned}$$

which suggests $\sum_{j\in\tilde{\mathcal{V}}_+}\|w_j\|^2 \leq \frac{1}{\sqrt{\mu}X_{\min}}$. A contradiction.

Therefore, we must have

$$t_2^+ \leq t_1 + \frac{6}{\sqrt{\mu}n_+X_{\min}} + \frac{4}{\sqrt{\mu}n_+X_{\min}}\left(\log\frac{2}{\epsilon^2\sqrt{\mu}X_{\min}W_{\min}^2} + 4nX_{\max}t_1\right)\,. \tag{2.202}$$

Since the dominant term here is $\frac{4}{\sqrt{\mu}n_+X_{\min}}\log\frac{2}{\epsilon^2\sqrt{\mu}X_{\min}W_{\min}^2}$, we have $t_2^+ = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$. A similar analysis shows $t_2^- = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$. Therefore $t_2 = \max\{t_2^+, t_2^-\} = \mathcal{O}(\frac{1}{n}\log\frac{1}{\epsilon})$

Complete the missing pieces: We have two claims remaining to be proved. The first is $\cos(x_+, \tilde{x}_a) \geq \frac{1}{2}$ when $t_1 \leq t \leq t_2^+$. Since $\tilde{x}_a = \sum_{i\in\mathcal{I}_+} -\nabla_{\hat{y}}\ell(y_i, f_+(x_i;W,v))x_i$ and $x_+ = \sum_{i\in\mathcal{I}_+} x_i$, we simply use the fact that before t_2^+, we have, by Lemma 2.21,

$$\frac{1}{2} \leq -\nabla_{\hat{y}}\ell(y_i, f_+(x_i;W,v)) =\leq \frac{3}{2}\,, \tag{2.203}$$

to show the following

$$\begin{aligned}
\cos(x_+,\tilde{x}_a) = & \frac{\langle x_+,\tilde{x}_a\rangle}{\|x_+\|\|\tilde{x}_a\|}\\
= & \frac{\sum_{i,j\in\mathcal{I}_+}(-\nabla_{\hat{y}}\ell(y_i,f_+(x_i;W,v)))\left\langle x_i,x_j\right\rangle}{\sqrt{\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}\sqrt{\sum_{i,j\in\mathcal{I}_+}(-\nabla_{\hat{y}}\ell(y_i,f_+(x_i;W,v)))^2\left\langle x_i,x_j\right\rangle}}\\
\geq & \frac{\frac{1}{2}\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}{\sqrt{\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}\sqrt{\sum_{i,j\in\mathcal{I}_+}(-\nabla_{\hat{y}}\ell(y_i,f_+(x_i;W,v)))^2\left\langle x_i,x_j\right\rangle}}\\
\geq & \frac{\frac{1}{2}\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}{\sqrt{\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}\sqrt{\sum_{i,j\in\mathcal{I}_+}(\frac{3}{2})^2\left\langle x_i,x_j\right\rangle}}\geq\frac{1}{3}\,,
\end{aligned}$$

since all $\left\langle x_i,x_j\right\rangle, i,j\in\mathcal{I}_+$ are non-negative.

The second claim is $\|\tilde{x}_a\|\geq\sqrt{\mu}n_+X_{\min}/2$ is due to that

$$\begin{aligned}
\|\tilde{x}_a\| &= \sqrt{\sum_{i,j\in\mathcal{I}_+}(-\nabla_{\hat{y}}\ell(y_i,f_+(x_i;W,v)))^2\left\langle x_i,x_j\right\rangle}\\
&\geq\frac{1}{2}\sqrt{\sum_{i,j\in\mathcal{I}_+}\left\langle x_i,x_j\right\rangle}=\frac{\|x_+\|}{2}\geq\frac{\sqrt{\mu}n_+X_{\min}}{2}\,,
\end{aligned}$$

where the last inequality is from Lemma 2.30. □

Proof of low-rank bias

So far we have proved the directional convergence at the early alignment phase and final $\mathcal{O}(1/t)$ convergence of the loss in the later stage. The only thing that remains to be shown is the low-rank bias. The proof is quite straightforward but we need some additional notations.

As we proved above, after t_1, neurons in $\mathcal{S}_+$ ($\mathcal{S}_-$) stays in $\mathcal{S}_+$ ($\mathcal{S}_-$). Thus the loss can be decomposed as

$$\mathcal{L}=\underbrace{\sum_{i\in\mathcal{I}_+}\ell\left(y_i,\sum_{j\in\tilde{\mathcal{V}}_+}v_j\left\langle w_j,x_i\right\rangle\right)}_{\mathcal{L}_+}+\underbrace{\sum_{i\in\mathcal{I}_-}\ell\left(y_i,\sum_{j\in\tilde{\mathcal{V}}_-}v_j\left\langle w_j,x_i\right\rangle\right)}_{\mathcal{L}_-}\,,$$

where $\tilde{\mathcal{V}}_+:\{j:w_j(t_1)\in\mathcal{S}_+\}$ and $\tilde{\mathcal{V}}_-:\{j:w_j(t_1)\in\mathcal{S}_-\}$. Therefore, the training after t_1 is decoupled into 1) using neurons in $\tilde{\mathcal{V}}_+$ to fit positive data in $\mathcal{I}_+$ and 2) using neurons in $\tilde{\mathcal{V}}_-$ to fit positive data in $\mathcal{I}_-$. We use

$$W_+ = [W]_{:,\tilde{\mathcal{V}}_+}, \quad W_- = [W]_{:,\tilde{\mathcal{V}}_-}$$

to denote submatrices of W by picking only columns in $\tilde{\mathcal{V}}_+$ and $\tilde{\mathcal{V}}_-$, respectively. Similarly, we define

$$v_+ = [v]_{\tilde{\mathcal{V}}_+}, \quad v_- = [v]_{\tilde{\mathcal{V}}_-}$$

for the second layer weight v. Lastly, we also define

$$W_{\text{dead}} = [W]_{:,\tilde{\mathcal{V}}_{\text{dead}}}, v_{\text{dead}} = [v]_{\tilde{\mathcal{V}}_{\text{dead}}}\,,$$

where $\tilde{\mathcal{V}}_{\text{dead}} := \{j : w_j(t_1) \in \mathcal{S}_{\text{dead}}\}$. Given these notations, after t_1 the loss is decomposed as

$$\mathcal{L} = \underbrace{\sum_{i\in\mathcal{I}_+} \ell\left(y_i, x_i^\top W_+ v_+\right)}_{\mathcal{L}_+} + \underbrace{\sum_{i\in\mathcal{I}_-} \ell\left(y_i, x_i^\top W_- v_-\right)}_{\mathcal{L}_-}\,,$$

and the GF on $\mathcal{L}$ is equivalent to GF on $\mathcal{L}_+$ and $\mathcal{L}_-$ separately. It suffices to study one of them. For GF on $\mathcal{L}_+$, we have the following important invariance [27] $\forall t \geq t_1$:

$$W_+^\top(t)W_+(t) - v_+(t)v_+^\top(t) = W_+^\top(t_1)W_+(t_1) - v_+(t_1)v_+^\top(t_1)\,,$$

from which one has

$$\begin{aligned}
\|W_+^\top(t)W_+(t) - v_+(t)v_+^\top(t)\|_2 =& \ \|W_+^\top(t_1)W_+(t_1) - v_+(t_1)v_+^\top(t_1)\|_2 \\
\leq& \ \|W_+^\top(t_1)W_+(t_1)\|_2 - \|v_+(t_1)v_+^\top(t_1)\|_2 \\
\leq& \ \operatorname{tr}(W_+^\top(t_1)W_+(t_1)) + \|v_+(t_1)\|^2 \\
=& \ 2\sum_{j\in\tilde{\mathcal{V}}_+} \|w_j(t_1)\|^2 \leq \frac{4\epsilon W_{\max}^2}{\sqrt{h}}|\tilde{\mathcal{V}}_+|\,,
\end{aligned}$$

where the last inequality is by Lemma 2.22. Then one can immediately get

$$\|v_+(t)v_+^\top(t)\|_2 - \|W_+^\top(t)W_+(t)\|_2 \leq \|W_+^\top(t)W_+(t) - v_+(t)v_+^\top(t)\|_2 \leq \frac{4\epsilon W_{\max}^2}{\sqrt{h}}|\tilde{\mathcal{V}}_+|\,,$$

which is precisely

$$\|W_+(t)\|_F^2 \leq \|W_+(t)\|_2^2 + \frac{4\epsilon W_{\max}^2}{\sqrt{h}}|\tilde{\mathcal{V}}_+|\,. \tag{2.204}$$

Similarly, we have

$$\|W_-(t)\|_F^2 \leq \|W_-(t)\|_2^2 + \frac{4\epsilon W_{\max}^2}{\sqrt{h}}|\tilde{\mathcal{V}}_-|\,. \tag{2.205}$$

Lastly, one has

$$\|W_{\text{dead}}\|_F^2 = \sum_{j\in\tilde{\mathcal{V}}_{\text{dead}}} \|w_j(t_1)\|^2 \leq \frac{4\epsilon W_{\max}^2}{\sqrt{h}}|\tilde{\mathcal{V}}_{\text{dead}}| \tag{2.206}$$

Adding (2.204)(2.205)(2.206) together, we have

$$\begin{aligned}\|W(t)\|_F^2 =& \ \|W_+(t)\|_F^2 + \|W_-(t)\|_F^2 + \|W_{\text{dead}}\|_F^2\\ \leq& \ \|W_+(t)\|_2^2 + \|W_-(t)\|_2^2 + \frac{4\sqrt{h}\epsilon W_{\max}^2}{\sqrt{h}} \leq 2\|W(t)\|_2^2 + 4\sqrt{h}\epsilon W_{\max}^2\,.\end{aligned}$$

Finally, since we have shown $\mathcal{L} \to 0$ as $t \to \infty$, then we have $\ell(y_i, f(x_i; W, v)) \to 0$, $\forall i \in [n]$. This implies

$$f(x_i; W, v) = -\frac{1}{y_i}\log \ell(y_i, f(x_i; W, v)) \to \infty\,.$$

Because we have shown that

$$f(x_i; W, v) \leq \sum_{j\in[h]} \|w_j\|^2\|x_i\| \leq \|W\|_F^2 X_{\max}\,,$$

$f(x_i; W, v) \to \infty$ enforces $\|W\|_F^2 \to \infty$ as $t \to \infty$, thus $\|W\|_2^2 \to \infty$ as well. This gets us

$$\limsup_{t\to\infty} \frac{\|W\|_F^2}{\|W\|_2^2} = 2\,.$$

2.4 Conclusion

In this chapter, we first study the explicit role of initialization on controlling the convergence and implicit bias of single-hidden-layer linear networks trained under gradient flow. We first provide a lower bound on the instantaneous rate based on the imbalance matrix and the product, from which convergence guarantees are derived based on sufficient imbalance or sufficient margin. We then show that proper initialization enforces the trajectory of network parameters to be exactly (or approximately) constrained in a low-dimensional invariant set, over which minimizing the loss yields the min-norm solution. Combining those results, we obtain a novel non-asymptotic bound regarding the implicit bias of wide linear networks under random initialization towards the min-norm solution. Our analysis, although on a simple overparametrized model, connects overparametrization, initialization, and optimization. Some concepts such as the imbalance extend to multi-layer linear networks, and eventually to neural networks with nonlinear activations, as shown in later sections. Next, we extend the convergence analysis to multi-layer linear models with a loss of general form $f(W_1W_2\cdots W_L)$. We show that with proper initialization, the loss converges to its global minimum exponentially. Our analysis applies to various types of multi-layer linear networks, and our assumptions on f are general.

Finally, we study the problem of training a binary classifier via gradient flow on two-layer ReLU networks under small initialization. We consider a training dataset with well-separated input vectors. A careful analysis of the neurons' directional dynamics allows us to provide an upper bound on the time it takes for all neurons to achieve good alignment with the input data. After the early alignment phase, the loss converges to zero at a $\mathcal{O}(\frac{1}{t})$ rate, and the weight matrix on the first layer is approximately low-rank. Lastly, our numerical experiment on classifying two digits from the MNIST dataset correlates with our theoretical findings.

Future directions include extending our gradient flow results to gradient descent algorithm [80] and to nonlinear networks. [61] shows the diagonal entries of the imbalance are preserved, and [71] shows a stronger version of such invariance given additional assumptions on the training trajectory. Therefore, the weight imbalance could be used to understand the training of nonlinear networks. Moreover, [81] shows that exploiting the symmetry that induces imbalance invariance could lead to an accelerated gradient descent algorithm, thus our general analysis could potentially also aid the algorithmic design.

Chapter 3

Coherence in Large-scale Networked Dynamical Systems

Network coherence generally refers to the emergence of simple aggregated dynamical behaviors, despite heterogeneity in the dynamics of the subsystems that constitute the network. Such a phenomenon usually results from a multi-cluster structure in large networks: Large-scale interconnected systems generally can be partitioned into multiple areas such that strong network coupling exists within each area while those between the areas are relatively weaker. This leads to a time-scale separation in the network responses to disturbances: the nodes in the same area get synchronized on a fast time scale through strong network interaction and move together, i.e., "coherently", in the long term. Then the slow dynamics, often referred to as *inter-area oscillation*, are characterized by the interaction between coherent areas through the weak connection. Inter-area oscillation potentially causes high-frequency fluctuation across the entire network, thus building an accurate and interpretable mathematical model for the inter-area oscillation is of paramount importance in understanding the system resilience of large-scale networks.

Network coherence has been a long-standing research problem and various analyses [38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48] have been developed over the last decades for identifying coherent areas, characterizing coherent dynamics, and modeling inter-area oscillation. Those analyses often assume simple first- or second-order node dynamics. However, new challenges arise as many practical network systems have changed their composition drastically. For example, in power networks, the increasing penetration of renewable energy sources introduces more inverted-based resources into the grid, and synchronous generators are being replaced. Compared to synchronous generators, based on which the classic coherence analyses are developed, inverted-based resources have very different dynamical characteristics and can hardly be well captured by only first- or second-order dynamics. Therefore, one requires coherence identification and aggregation procedures that work for more general network systems.

Chapter outline

The goal of the chapter is precisely to develop new analytical tools for understanding network coherence and inter-area oscillation in the face of networks with complex node dynamics. In Section 3.1, we focus on the case of a single coherent area and introduce our novel frequency-domain analysis for network coherence that works for networks with general node dynamics. This frequency domain analysis provides a deeper characterization of the role of both, network topology and node dynamics, on the coherent behavior of the network. In Section 3.2, we extend our analysis to the case of multiple coherent areas by proposing a structure-preserving network reduction model that captures the dominant inter-area oscillations among different areas. Lastly, we show that for power system applications, our reduction model still renders a high-order dynamical model, and propose algorithms to reduce the model complexity in Section 3.3.

Notation

For a vector x, $\|x\| = \sqrt{x^\top x}$ denotes the 2-norm of x, and for a matrix A, $\sigma_{\min}(A)$ denotes the minimum singular value of A, $\|A\|$ denotes the spectral norm of A. Particularly, if A is real symmetric, we let $\lambda_i(A)$ denote the ith smallest eigenvalue of A. We let $\mathrm{diag}\{x_i\}_{i=1}^n$ denote a $n \times n$ diagonal matrix with diagonal entries x_i. We let I_n denote the identity matrix of order n, $\mathbb{1}$ denote column vector $[1, \cdots, 1]^\top$, $[n]$ denote the set $\{1, 2, \cdots, n\}$ and $\mathbb{N}_+$ denote the set of positive integers. Also, we write complex numbers as $a + jb$, where $j = \sqrt{-1}$. We denote $\mathbb{C}$ the field of complex number, and define the following subsets $\mathcal{B}(s_0, \delta) := \{s \in \mathbb{C} : |s - s_0| \leq \delta\}$. For non-negative random variables $X(n), Y(n)$, we write $X(n) \sim \mathcal{O}_p(Y(n))$ if $\exists M > 0$, s.t. $\lim_{n\to\infty} \mathbb{P}\left(X(n) \leq MY(n)\right) = 1$. We write $X(n) \sim \Omega_p(Y(n))$ if $\exists M > 0$, s.t. $\lim_{n\to\infty} \mathbb{P}\left(X(n) \geq MY(n)\right) = 1$. *Notice that in this chapter, the eigenvalues $\lambda_i(A)$ of matrix A is in increasing order.*

3.1 Networks with One Coherent Cluster

In this section, we introduce our frequency domain analysis of network coherence and focus on the case when the entire network is coherent due to strong coupling among all nodes. Our analysis formalizes network coherence through a low-rank structure of the system transfer matrix that appears when the network feedback gain is high. This frequency domain analysis provides a deeper characterization of the role of both, network topology and node dynamics, on the coherent behavior of the network. In particular, our results make substantial contributions towards the understanding of coordinated and coherent behavior of network systems in many ways:

- We present a general framework in the frequency domain to analyze the coherence of heterogeneous networks. We show that network coherence emerges as a low-rank structure of the system transfer matrix as we increase the effective algebraic connectivity–a frequency-varying quantity that depends on the network coupling strength and dynamics.

- Our analysis applies to networks with heterogeneous nodal dynamics, and further provides an explicit characterization in the frequency domain of the coherent response to disturbances as the harmonic mean of individual nodal dynamics. Thus, in this way, our results highlight the contribution of individual nodal dynamics to the network's coherent behavior.
- We formally connect our frequency-domain results with explicit time-domain L_∞ bounds on the difference between individual nodal responses and the coherent dynamic response to certain classes of input signals, suggesting that network coherence is a frequency-dependent phenomenon. That is, the ability of nodes to respond coherently depends on the frequency composition of the input disturbance.
- By providing an exact characterization of the network's coherent dynamics, our analysis can be further applied in settings where only distributional information of the network composition is known. More precisely, we show that the coherent dynamics of tightly-connected networks with possibly random nodal dynamics are well approximated by a deterministic transfer function that only depends on the statistical distribution of node dynamics.

Notably, the problem of characterizing coherent dynamic response is unique to heterogeneous networks since the coherent dynamics for homogeneous networks are exactly equal to the common nodal dynamics. In real applications, however, such as power networks, such characterization is relevant to model reduction [44] and control design [50]. Our analysis provides, in the asymptotic sense, the exact characterization of coherent dynamics that can be used in control design for heterogeneous networks.

3.1.1 Problem setup

Consider a network consisting of n nodes ($n \geq 2$), indexed by $i \in [n]$ with the block diagram structure in Figure 3-1. L is the Laplacian matrix of the weighted graph that describes the network interconnection. We further use $f(s)$ to denote the transfer function representing the dynamics of network coupling, and $G(s) = \mathrm{diag}\{g_i(s)\}$ to denote the nodal dynamics, with $g_i(s),\ i \in [n]$, being an SISO transfer function representing the dynamics of node i. Throughout this chapter, we assume all $g_i(s),\ i = 1, \cdots, n$ and $f(s)$ are rational proper transfer functions, and the Laplacian matrix L is real symmetric.

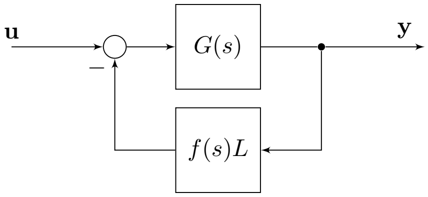

Figure 3-1. Block diagram of networked dynamical systems

Under this setting, we can compactly express the transfer matrix from the input signal vector $\mathbf{u}$ to the output signal vector $\mathbf{y}$ by

$$\begin{aligned} T(s) = &\ (I_n + G(s)f(s)L)^{-1}G(s) \\ = &\ (I_n + \mathrm{diag}\{g_i(s)\}f(s)L)^{-1}\mathrm{diag}\{g_i(s)\}\,. \end{aligned} \tag{3.1}$$

Many existing networks can be represented by this structure. For example, for the first-order consensus network [33, 82], $f(s) = 1$, and the node dynamics are given by $g_i(s) = \frac{1}{s}$. For power networks [55, 48], $f(s) = \frac{1}{s}$, $g_i(s)$ are the dynamics of the generators, and L is the Laplacian matrix representing the sensitivity of power injection w.r.t. bus phase angles. Finally, in transportation networks [37, 82], $g_i(s)$ represent the vehicle dynamics whereas $f(s)L$ describes local inter-vehicle information transfer.

Since L has an eigendecomposition $L = V\Lambda V^\top$ where $V = \left[\frac{\mathbb{1}}{\sqrt{n}}, V_\perp\right]$, $VV^\top = V^\top V = I_n$, and $\Lambda = \mathrm{diag}\{\lambda_i(L)\}$ with $0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_n(L)$, we can rewrite $T(s)$ as

$$\begin{aligned} T(s) =&\ (I_n + \mathrm{diag}\{g_i(s)\}f(s)L)^{-1}\mathrm{diag}\{g_i(s)\} \\ =&\ (\mathrm{diag}\{g_i^{-1}(s)\} + f(s)L)^{-1} \\ =&\ (\mathrm{diag}\{g_i^{-1}(s)\} + f(s)V\Lambda V^\top)^{-1} \\ =&\ V(V^\top \mathrm{diag}\{g_i^{-1}(s)\}V + f(s)\Lambda)^{-1}V^\top \,. \end{aligned} \tag{3.2}$$

As we mentioned in the introduction, we are interested in the regime where the closed-loop system $T(s)$ of (3.1) has a low-rank structure. To gain some insight, we first consider the following simplified example.

Motivating example: homogeneous network

Suppose $g_i(s)$ are homogeneous, i.e., $g_i(s) = g(s)$. Then using (3.2) one can decompose $T(s)$ as follows

$$T(s) = \frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top + V_\perp \mathrm{diag}\left\{\frac{1}{g^{-1}(s) + f(s)\lambda_i(L)}\right\}_{i=2}^{n} V_\perp^\top \,, \tag{3.3}$$

where the network dynamics decouple into two terms: 1) the dynamics $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$ that is independent of network topology and corresponds to the coherent behavior of the system; 2) the remaining dynamics that are dependent on the network structure via both, the eigenvalues $\lambda_i(L), i = 2, \cdots, n$ and the eigenvectors $V_\perp$. Notice that $|f(s)\lambda_2(L)| \leq |f(s)\lambda_i(L)|, i = 2, \ldots, n$, then $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$ is dominant in $T(s)$ as long as $|f(s)\lambda_2(L)|$ (later referred as *effective algbraic connectivity*), is large enough to make the norm of the second term in (3.3) sufficiently small. Following such observation, we can find two regimes where the coherent dynamics $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$ is dominant:

1. (*High network connectivity*) If a compact set $S \subset \mathbb{C}$ contains neither zeros nor poles of $g(s)$, then we have $\lim_{\lambda_2(L)\to\infty} \sup_{s\in S} \left\|T(s) - \frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top\right\| = 0\,.$

2. (*High gain in coupling dynamics*) If s_0 is a pole of $f(s)$, and the network is connected, i.e., $\lambda_2(L) > 0$, then we have $\lim_{s\to s_0} \left\|T(s) - \frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top\right\| = 0\,.$

Such convergence results suggest that if 1) the network has high algebraic connectivity, or 2) our point of interest in frequency domain is close to pole of $f(s)$, the response of the entire system is close to one of $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$. We refer $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$ as the coherent dynamics[1] in the sense that in such system, the inputs are aggregated, and all nodes have exactly the same response to the aggregate input. *Therefore, coherence of the network corresponds, in the frequency domain, to the property that the network's transfer matrix approximately having a particular rank-one structure.*

The aforementioned analysis can be extended to the case with proportionality assumption, i.e., $g_i(s) = p_i g(s)$ for some $g(s)$ and $p_i > 0, i = 1, \cdots, n$, where one can still obtain decoupled dynamics through proper coordinate transformation [48] and the coherent dynamics are again characterized by the common dynamics $g(s)$. However, it is challenging to analyze the transfer matrix $T(s)$ without the proportionality assumption: First, it is unclear whether low-rank structure would even emerge under high network connectivity or high gain in the coupling dynamics; Then most importantly, there is no obvious choice for coherent dynamics, hence characterizing the coherent dynamics is a non-trivial problem unique to heterogeneous networks, and no existing work has shown an explicit characterization.

Our work precisely aims at understanding the coherent dynamics of non-proportional heterogeneous networks. We would like to show that even when $g_i(s)$ are heterogeneous, similar results as in the motivating example still hold. More precisely, we show that, in Section 3.1.2, $T(s)$ converges to a rank-one transfer matrix of the form $\frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top$, as the effective algebraic connectivity $|f(s)\lambda_2(L)|$ increases. However, unlike the homogeneous node dynamics case where the coherent

[1] We also refer $g(s)$ as the coherent dynamics since transfer matrix of the form $\frac{1}{n}g(s)\mathbb{1}\mathbb{1}^\top$ is uniquely determined by its non-zero eigenvalue $g(s)$.

behavior is driven by $\bar{g}(s) = g(s)$, the coherent dynamics $\bar{g}(s)$ are given by the harmonic mean of $g_i(s), i = 1, \cdots, n$, i.e.,

$$\bar{g}(s) = \left(\frac{1}{n} \sum_{i=1}^{n} g_i^{-1}(s) \right)^{-1} . \tag{3.4}$$

The convergence results are presented in the aforementioned two regimes: high network connectivity and high gain in coupling dynamics. We then discuss in Section 3.1.3 their implications on network's time-domain response:

1. Network with high connectivity responds coherently to a wide class of input signals;
2. Network with coupling dynamics $f(s) = \frac{1}{s}$ is naturally coherent with respect to sufficiently low-frequency signals, regardless of its connectivity.

One additional feature of our analysis is that it can be further applied in settings where the composition of the network is unknown and only distributional information is present. More precisely, we, in Section 3.1.4, consider a network where node dynamics are given by random transfer functions. As the network size grows, the coherent dynamics $\bar{g}(s)$, the harmonic mean of all node dynamics, converges in probability to a deterministic transfer function. We term such a phenomenon, where a family of uncertain large-scale systems concentrates to a common deterministic system, *dynamics concentration*.

3.1.2 Coherence in frequency domain

In this section, we analyze the network coherence as the low-rank structure of the transfer matrix in the frequency domain. We start with an important lemma revealing how such coherence is related to the algebraic connectivity $\lambda_2(L)$ and the coupling dynamics $f(s)$.

Lemma 3.1. *Let $T(s)$ and $\bar{g}(s)$ be defined as in* (3.1) *and* (3.4)*, respectively. Suppose that for $s_0 \in \mathbb{C}$ that is not a pole of $f(s)$, we have*

$$|\bar{g}(s_0)| \leq M_1, \textit{and} \max_{1\leq i\leq n} |g_i^{-1}(s_0)| \leq M_2\,,$$

for some $M_1, M_2 > 0$. Then the following inequality holds:

$$\left\| T(s_0) - \frac{1}{n}\bar{g}(s_0)\mathbb{1}\mathbb{1}^\top \right\| \leq \frac{(M_1M_2+1)^2}{|f(s_0)|\lambda_2(L) - M_2 - M_1M_2^2}\,, \tag{3.5}$$

whenever $|f(s_0)|\lambda_2(L) \geq M_2 + M_1M_2^2$.

Lemma 3.1 provides an error bound for approximating $T(s)$ with a rank-one transfer matrix $\frac{1}{n}\bar{g}(s)$. It is a special version of Theorem 3.8 to be introduced in Section 3.2, which concerns approximating $T(s)$ with a rank-k transfer matrix.

Lemma 3.1 provides a non-asymptotic rate for our incoherence measure

$$\left\| T(s_0) - \frac{1}{n}\bar{g}(s_0)\mathbb{1}\mathbb{1}^\top \right\| \sim \mathcal{O}\left(\frac{M_1^2M_2^2}{|f(s_0)|\lambda_2(L)}\right)\,. \tag{3.6}$$

A large value of $|f(s_0)|\lambda_2(L)$ is sufficient to have the incoherence measure small, and we term this quantity as *effective algebraic connectivity*. We see that there are two possible ways to achieve such point-wise coherence: Either we increase the network algebraic connectivity $\lambda_2(L)$, by adding edges to the network and increasing edge weights, etc., or we move our point of interest s_0 to a pole of $f(s)$. This point-wise coherence via effective connectivity provides the basis of our subsequent analysis. As we mentioned above, we can achieve such coherence by increasing either $\lambda_2(L)$ or $|f(s_0)|$, provided that the other value is fixed and non-zero. Section 3.1.2 considers the former and Section 3.1.2 the latter.

Coherence under high network connectivity

It is intuitive that a network behaves coherently under high connectivity. A formal frequency domain characterization is stated as follow.

Theorem 3.1. *Let $T(s)$ and $\bar{g}(s)$ be defined as in (3.1) and (3.4), respectively. Given a compact set $S \subset \mathbb{C}$, if*

1. *S does not contain any zero or pole of $\bar{g}(s)$;*
2. $\inf_{s \in S} |f(s)| > 0\,,$

we have $\lim_{\lambda_2(L) \to +\infty} \sup_{s \in S} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| = 0\,.$

Proof. On the one hand, since S does not contain any pole of $\bar{g}(s)$, $\bar{g}(s)$ is continuous on the compact set S, and hence bounded [83, Theorem 4.15]. On the other hand, because S does not contain any zero of $\bar{g}(s)$, every $g_i^{-1}(s)$ must be continuous on S, and hence bounded as well. It follows that $\max_{1 \le i \le n} |g_i^{-1}(s)|$ is bounded on S, and the conditions of Lemma 3.1 are satisfied for all $s \in S$ with a uniform choice of M_1 and M_2. By (3.5), we have

$$\sup_{s \in S} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \le \frac{(M_1 M_2 + 1)^2}{F_l \lambda_2(L) - M_2 - M_1 M_2^2}\,,$$

where $F_l = \inf_{s \in S} |f(s)|$. We finish the proof by taking $\lambda_2(L) \to +\infty$ on both sides. □

Theorem 3.1 formally shows that high network connectivity leads to coherence. We emphasize that such coherence is frequency-dependent: the incoherence measure is defined over a compact set S. Roughly speaking, if we would like to see whether the network could have coherent response under certain input signal, then S should cover most of the frequency components of that signal, as well satisfies the assumptions in Theorem 3.1. We discuss the proper choice of S when we use Theorem 3.1 to infer the time-domain response in Section 3.1.3.

Coherence under high gain in coupling dynamics

However, high network connectivity is not necessary for coherence. A high gain in the coupling dynamics effectively amplifies the network connection, leading to the following frequency-domain coherence.

Theorem 3.2. *Let $T(s)$ and $\bar{g}(s)$ be defined as in* (3.1) *and* (3.4)*, respectively. Given a pole of $f(s)$, if*

1. *s_0 is neither a pole nor a zero of $\bar{g}(s)$;*

2. *$\lambda_2(L) > 0$,*

then $\lim_{s\to s_0} \left\|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\right\| = 0\,.$

Proof. Since s_0 is neither a zero nor a pole of $\bar{g}(s)$, $\exists \delta_1 > 0$ such that $\forall s \in \mathcal{B}(s_0, \delta_1)$, we have $|\bar{g}^{-1}(s)| \leq M_1$ and $\max_{1\leq i\leq n} |g_i^{-1}(s)| \leq M_2$ for some $M_1, M_2 > 0$.

Now notice that $\lim_{s\to s_0} |f(s)| = +\infty$, by the definition of the limit, we know that $\exists \delta_2 > 0$ such that $\forall s \in \mathcal{B}(s_0, \delta_2)$, we have $\frac{1}{2}|f(s)|\lambda_2(L) \geq M_2 + M_1 M_2^2$. By Lemma 3.1, let $\delta := \min\{\delta_1, \delta_2\}$, then $\forall s \in \mathcal{B}(s_0, \delta)$, the following holds

$$\begin{aligned}\left\|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\right\| &\leq \frac{(M_1 M_2 + 1)^2}{|f(s)|\lambda_2(L) - M_2 - M_1 M_2^2} \\ &\leq \frac{2\,(M_1 M_2 + 1)^2}{|f(s)|\lambda_2(L)}\,.\end{aligned}$$

Taking $s \to s_0$, the limit of right-hand side is 0. □

Theorem 3.2 suggests that for any connected network, some coupling dynamics causes coherent responses from the network under specific input signals. For example, when $f(s) = \frac{1}{s}$, the network $T(s)$ is naturally coherent around $s = 0$, which implies that such network behaves coherently under sufficiently low-frequency input signals. This is formally justified in Section 3.1.3, along with time-domain results for other choice of coupling dynamics.

3.1.3 Implications on time-domain response

In this section, we discuss how one can infer the network's time-domain response using the established frequency-domain coherence in Theorem 3.1 and 3.2. Provided that the network $T(s)$ and the coherent dynamics $\bar{g}(s)$ are BIBO stable, we let $\mathbf{y}(t) = [y_1(t), \cdots, y_i(t), \cdots, y_n(t)]^\top$ be the response of the network when the network input is $U(s)$, and let $\bar{y}(t)$ be the response of $\bar{g}(s)$ to $\frac{\mathbb{1}^\top}{n} U(s)$. The inverse Laplace transform [84] suggests that for all $i = 1, \cdots, n$, we have

$$|y_i(t) - \bar{y}(t)| = \left| \lim_{\omega \to \infty} \int_{\sigma - j\omega}^{\sigma + j\omega} e^{st} e_i^\top \left(T(s) - \frac{1}{n} \bar{g}(s) \mathbb{1}\mathbb{1}^\top \right) U(s) ds \right| , \tag{3.7}$$

with a proper choice of $\sigma > 0$. Here e_i is the i-th column of identity matrix I_n. This integral can be decomposed in two parts: one integral on the low-frequency band $(\sigma - j\omega_0, \sigma + j\omega_0)$; and another on the high-frequency band $(\sigma - j\infty, \sigma - j\omega_0) \cup (\sigma + j\omega_0, \sigma + j\infty)$, with some choice of ω_0. The former can be made small in absolute value by controlling the incoherence measure $\|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\|$ over the set $S : (\sigma - j\omega_0, \sigma + j\omega_0)$. In particular,

1. $\sup_{s \in S} \|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\|$ can be small under high network connectivity, as suggested by Theorem 3.1;

2. $\sup_{s \in S} \|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\|$ can be small when S is confined in a neighborhood around pole of coupling dynamics $f(s)$, suggested by Theorem 3.2. The case $f(s) = \frac{1}{s}$ is of the most interest.

Moreover, when $U(s)$ is a sufficiently low-frequency signal such that the high-frequency band $(\sigma - j\infty, \sigma - j\omega_0) \cup (\sigma + j\omega_0, \sigma + j\infty)$ does not include much of its frequency components, the latter integral can be made small. Given an upper bound on the integral in (3.7), we show that the time-domain response of every node in the network resembles the one from the coherent dynamics $\bar{g}(s)$. Similar to Section 3.1.2, we show such time-domain coherence in two regimes: high network connectivity or high gain in the coupling dynamics.

Remark 6. *In order to infer the time-domain response, it is necessary that both the transfer functions $T(s)$ and $\frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top$ are stable. Since our primary focus is on the interpretation of* *the frequency domain results, we are largely working under the tacit assumption that these transfer functions are stable whenever required. It should also be noted that there exist a range of scalable stability criteria in the literature that can be used to guarantee internal stability of the feedback setup in Figure 3-1. Perhaps the most well known is that if each $g_i(s)$ is strictly positive real, and $f(s)$ is positive real, then the transfer functions $\bar{g}(s)$ and*

$$\begin{bmatrix} G(s) \\ I \end{bmatrix} (I + f(s)LG(s))^{-1} \begin{bmatrix} f(s)L & I \end{bmatrix}$$

are stable (see e.g. [85]). Alternative approaches that can be easily adapted to our framework that give criteria that allow for different classes of transfer functions include [86, 87, 88].

Coherent response under high network connectivity

Our first result considers network with high connectivity.

Theorem 3.3. *Given a network with node dynamics $\{g_i(s)\}_{i=1}^n$ and coupling dynamics $f(s)$, assume that there exists $\gamma > 0$, such that $\|\bar{g}(s)\|_{\mathcal{H}_\infty} \leq \gamma$ and $\|T(s)\|_{\mathcal{H}_\infty} \leq \gamma$ for any symmetric Laplacian matrix L. Consider a network coupling $f(s)$ and a real input signal vector $\mathbf{u}(t)$ with its Laplace transform $U(s)$ such that for some $\sigma > 0$, we have*

1. $\inf_{\omega \in \mathbb{R}} |f(\sigma + i\omega)| > 0$;
2. $\sup_{Re(s) > \sigma} \|U(s)\|$ *is finite;*
3. $\lim_{\omega \to \infty} \int_{\sigma + j0}^{\sigma + j\omega} \|U(s)\| ds$ *is finite .*

Then for any $\epsilon > 0$, there exists a $\lambda > 0$, such that whenever $\lambda_2(L) \geq \lambda$, we have $\|\mathbf{y}(t) - \bar{y}(t)\mathbb{1}\|_{\mathcal{L}_\infty} \leq \epsilon$, i.e.,

$$\max_{i \in [n]} \sup_{t > 0} |y_i(t) - \bar{y}(t)| \leq \epsilon\,.$$

Theorem 3.3 provide a formal explanation of coherent behavior observed in practical networks and show its relation with network connectivity. That is, a stable network with high connectivity can respond coherently to a class of input signals. More importantly, the coherently response is well approximated by $\bar{g}(s)$, then it suffices to study $\bar{g}(s)$ for understanding the coherent behavior of a network with high connectivity.

While the theorem suggests that some level of coherence can be achieved by increasing the network connectivity, one should be cautious about the potential network instability caused by strong interconnection. Nonetheless, some simple passivity motivated criteria that ensure stability even as $\lambda_2(L)$ becomes arbitrarily large:

Theorem 3.4. *Suppose that all* $g_i(s), i = 1, \cdots, n$ *are* output strictly passive*:* $Re(g_i(s)) \geq \epsilon |g_i(s)|^2, \forall Re(s) > 0$ *, for some* $\epsilon > 0$*, and* $f(s)$ *is* positive real*:* $Re(f(s)) \geq 0, \forall Re(s) > 0$ *, then there exists* $\gamma > 0$*, such that given any positive semidefinite matrix* L*, we have*

$$\|\bar{g}(s)\|_{\mathcal{H}_\infty} \leq \gamma, \text{ and } \|T(s)\|_{\mathcal{H}_\infty} \leq \gamma \,.$$

Theorem 3.4, together with Theorem 3.3, shows that for certain passive networks, the coherence can be achieved over a class of input signals by increasing the network connectivity.

Remark 7. *Besides network stability as a prerequisite, a few assumptions are made: infimum on* $|f(s)|$ *ensures that the network coupling does not vanish over our domain of interest; supremum on* $\|U(s)\|$ *is needed for utilizing inverse Laplace transform; and the last assumption requires* $U(s)$ *to have light tail on the high-frequency range, a low-frequency signal with no abrupt change at* $t = 0$*, such as sinusoidal signal* $U(s) = \frac{\alpha}{s^2+\alpha^2}\mathbf{u}_0$*, or exponential approach signal* $U(s) = \frac{\alpha}{s(s+\alpha)}\mathbf{u}_0$ *of some shape* $\mathbf{u}_0 \in \mathbb{R}^n$*, satisfies the assumption.*

Coherent response under special coupling dynamics

As we discussed in Section 3.1.2, coherence is not all about network connectivity, and high gain in the coupling dynamics causes coherence as well. One simple and practically seen coupling dynamics are $f(s) = \frac{1}{s}$. Due to its high gain at $s = 0$, we expected the a coherent response under low-frequency signals, as formally shown below.

Theorem 3.5. *Given a network with node dynamics $\{g_i(s)\}_{i=1}^n$, coupling dynamics $f(s) = \frac{1}{s}$, and a fixed graph Laplacian L with $\lambda_2(L) > 0$, such that $\|\bar{g}(s)\|_{\mathcal{H}_\infty}$ and $\|T(s)\|_{\mathcal{H}_\infty}$ are finite, we let the network input be a sinusoidal signal $\mathbf{u}_\alpha(t) = \sin(\alpha t)\chi(t)\mathbf{u}_0$ in an arbitrary direction $\mathbf{u}_0 \in \mathbb{S}^{n-1}$. Then for any $\epsilon > 0$, there exists an $\alpha_0 > 0$ such that whenever $0 \leq \alpha \leq \alpha_0$, we have $\|\mathbf{y}(t) - \bar{y}(t)\mathbb{1}\|_{\mathcal{L}_\infty} \leq \epsilon$, i.e.,*

$$\max_{i\in[n]} \sup_{t>0} |y_i(t) - \bar{y}(t)| \leq \epsilon\,. \tag{3.8}$$

Theorem 3.5 shows that a stable network with $f(s) = \frac{1}{s}$ is naturally coherent subject to sufficiently low-frequency signals, regardless of its connectivity. Notably, the requirement on the node dynamics here is much weaker than one in Theorem 3.3 as we only need to establish stability for a given interconnection L, whereas Theorem 3.3 requires stability under any interconnection.

Comparison with different notions of coordination

Our Theorem 3.3 and 3.5 shows the coherent response of network in time domain. We compare our results to prior work that studies different forms of time-domain coordination in network systems.

The consensus [33] and synchronization [89, 90, 91] is arguably the simplest form of coordination in network systems, which can be viewed as a problem tracking some reference signal $\bar{y}(t)$ representing the final consensus or synchronization. However, one only requires $y_i(t) \to \bar{y}(t)$ when $t \to \infty$, i.e., that the node responses become close to $\bar{y}(t)$ in steady state. The coherent response considered here is different in that we have $y_i(t) \simeq \bar{y}(t), \forall t > 0$, i.e., $\bar{y}(t)$ is a good approximation for $y_i(t)$ for all time $t > 0$, hence our results can be also used for transient analysis.

The work on coherency and synchrony [92, 39, 93, 94] study a similar behavior as us, but characterized as pairwise coherence achieved under input signal of certain spatial shape: given a input signal vector $\mathbf{u}(t) = v(t)\mathbf{u}_0$, [92, 93] shows the condition on $\mathbf{u}_0$ such that the responses of some pair of nodes are similar (or generally, proportional [39]), i.e., $y_i(t) \simeq y_j(t)$ for some $i, j \in [n]$. Our results show that certain temporal shape $v(t)$ also causes coherence, and in a stronger form: our coherence does not depends on the shape u_0, and holds for all nodes.

3.1.4 Dynamics concentration in large-scale networks

In Section 3.1.2, we looked into convergence results of $T(s)$ for networks with fixed size n. However, one could easily see that such coherence depends mildly on the network size n: In Lemma 3.1, as long as the bounds regarding $g_i(s)$, i.e. M_1 and M_2 do not scale with respect to n, coherence can emerge as the network size increases. This is the topic of this section.

Coherence in large-scale networks

To start with, we revise the problem settings to account for variable network size: Let $\{g_i(s), i \in \mathbb{N}_+\}$ be a sequence of transfer functions, and $\{L_n, n \in \mathbb{N}_+\}$ be a sequence of real symmetric Laplacian matrices such that L_n is a square matrix of order n, particularly, let $L_1 = 0$. Then we define a sequence of transfer matrix $T_n(s)$ as

$$T_n(s) = \left(I_n + G_n(s)L_n\right)^{-1} G_n(s) \, , \tag{3.9}$$

where $G_n(s) = \mathrm{diag}\{g_1(s), \cdots, g_n(s)\}$. This is exactly the same transfer matrix shown in Figure 3-1 for a network of size n. We can then define the coherent dynamics for every $T_n(s)$ as $\bar{g}_n(s) = \left(\frac{1}{n}\sum_{i=1}^{n} g_i^{-1}(s)\right)^{-1}$.

For certain family $\{L_n, n \in \mathbb{N}_+\}$ of large-scale networks, the network algebraic connectivity $\lambda_2(L_n)$ increases as n grows. For example, when L_n is the Laplacian of a complete graph of size n with all edge weights being 1, we have $\lambda_2(L_n) = n$. As a result, network coherence naturally emerges as the network size grows. Recall that to prove the convergence of $T_n(s)$ to $\frac{1}{n}\bar{g}_n(s)\mathbb{1}\mathbb{1}^\top$ for fixed n, we essentially seek for $M_1, M_2 > 0$, such that $|\bar{g}_n(s)| \leq M_1$ and $\max_{1\leq i\leq n}|g_i^{-1}(s)| \leq M_2$ for s in a certain set. If it is possible to find a universal $M_1, M_2 > 0$ for all n, then the convergence results should be extended to arbitrarily large networks, provided that network connectivity increases as n grows. The results follows after we state the notion of uniform boundedness for a family of functions.

Definition 3.1. *Let $\{g_i(s), i \in I\}$ be a family of complex functions indexed by I. Given $S \subset \mathbb{C}$, $\{g_i(s), i \in I\}$ is uniformly bounded on S if*

$$\exists M > 0 \quad s.t. \quad |g_i(s)| \leq M, \quad \forall i \in I,\, \forall s \in S\,.$$

Theorem 3.6. *Suppose $\lambda_2(L_n) \to +\infty$ as $n \to \infty$. Given a compact set $S \subset \mathbb{C}$, if both $\{g_i^{-1}(s), i \in \mathbb{N}_+\}$ and $\{\bar{g}_n(s), n \in \mathbb{N}_+\}$ are uniformly bounded on a set $S \subset \mathbb{C}$, and $inf_{s\in S}|f(s)| > 0$, then we have*

$$\lim_{n\to\infty}\sup_{s\in S}\left\|T_n(s) - \frac{1}{n}\bar{g}_n(s)\mathbb{1}\mathbb{1}^\top\right\| = 0\,.$$

Proof. Since both $\{g_i^{-1}(s), i \in \mathbb{N}_+\}$ and $\{\bar{g}_n(s), n \in \mathbb{N}_+\}$ are uniformly bounded on S, $\exists M_1, M_2 > 0$ s.t. $|\bar{g}_n(s)| \leq M_1$ and $\max_{1\leq i\leq n}|g_i^{-1}(s)| \leq M_2$ for every $n \in \mathbb{N}_+$ and $s \in S$. By Lemma 3.1, $\forall n \in \mathbb{N}_+$,

$$\sup_{s\in S}\left\|T_n(s) - \frac{1}{n}\bar{g}_n(s)\mathbb{1}\mathbb{1}^\top\right\| \leq \frac{(M_1M_2+1)^2}{F_l\lambda_2(L_n) - M_2 - M_1M_2^2}\,, \tag{3.10}$$

where $F_l = \inf_{s\in S}|f(s)|$. We already assumed that $\lambda_2(L_n) \to +\infty$ as $n \to +\infty$, therefore the proof is finished by taking $n \to +\infty$ on both sides of (3.10). $\square$

Interestingly, in a stochastic setting where all $g_i(s)$ are unknown transfer functions independently drawn from some distribution, their harmonic mean $\bar{g}_n(s)$

eventually converges in probability to a deterministic transfer function as the network size increases. Consequently, a large-scale network consisting of random node dynamics (to be formally defined later) concentrates to deterministic a system. We term this phenomenon *dynamics concentration*.

Remark 8. *In this section, we only discuss the coherence due to connectivity, since the coherence from high gain in coupling dynamics shown in Theorem 3.2 can be applied to any connected network, regardless of its size.*

Dynamics concentration in large-scale networks

Now we consider the cases where the node dynamics are unknown (stochastic). For simplicity, we constraint our analysis to the setting where the node dynamics are independently sampled from the same random rational transfer function with all or part of the coefficients are random variables, i.e. the nodal transfer functions are of the form

$$g_i(s) \sim \frac{b_m s^m + \dots b_1 s + b_0}{a_l s^l + \dots a_1 s + a_0}\,, \tag{3.11}$$

for some $m, l > 0$, where $b_0, \cdots, b_m$, $a_0, \cdots, a_l$ are random variables.

To formalize the setting, we firstly define the random transfer function to be sampled. Let $\Omega = \mathbb{R}^d$ be the sample space, $\mathcal{F}$ the Borel σ-field of Ω, and $\mathbb{P}$ a probability measure on Ω. A sample $w \in \Omega$ thus represents a d-dimensional vector of coefficients. We then define a random rational transfer function $g(s, w)$ on $(\Omega, \mathcal{F}, \mathbb{P})$ such that all or part of the coefficients of $g(s, w)$ are random variables. Then for any $w_0 \in \Omega$, $g(s, w_0)$ is a rational transfer function.

Now consider the probability space $(\Omega^\infty, \mathcal{F}^\infty, \mathbb{P}^\infty)$. Every $\mathbf{w} \in \Omega^\infty$ give an instance of samples drawn from our random transfer function:

$$g_i(s, w_i) := g(s, w_i), i \in \mathbb{N}_+\,,$$

where w_i is the i-th element of $\mathbf{w}$. By construction, $g_i(s, w_i), i \in \mathbb{N}_+$ are i.i.d. random transfer functions. Moreover, for every $s_0 \in \mathbb{C}$, $g_i(s_0, w_i), i \in \mathbb{N}_+$ are i.i.d. random complex variables taking values in the extended complex plane (presumably taking value ∞).

Now given $\{L_n, n \in \mathbb{N}_+\}$ a sequence of $n \times n$ real symmetric Laplacian matrices, consider the random network of size n whose nodes are associated with the dynamics $g_i(s, w_i), i = 1, 2, \cdots, n$ and coupled through L_n. The transfer matrix of such a network is given by

$$T_n(s, \mathbf{w}) = (I_n + G_n(s, \mathbf{w})L_n)^{-1}G_n(s, \mathbf{w})\,, \tag{3.12}$$

where $G_n(s, \mathbf{w}) = \mathrm{diag}\{g_1(s, w_1), \cdots, g_n(s, w_n)\}$. Then under this setting, the coherent dynamics of the network is given by

$$\bar{g}(s, \mathbf{w}) = \left(\frac{1}{n}\sum_{i=1}^{n} g_i^{-1}(s, w_i)\right)^{-1} . \tag{3.13}$$

Now given a compact set $S \subset \mathbb{C}$ of interest, and assuming suitable conditions on the distribution of $g(s, w)$, we expect that the random coherent dynamics $\bar{g}(s, \mathbf{w})$ would converge uniformly in probability to its expectation

$$\hat{g}(s) = \left(\mathbb{E}g^{-1}(s, w))\right)^{-1} := \left(\int_{\Omega} g^{-1}(s, w)d\mathbb{P}(w)\right)^{-1} , \tag{3.14}$$

for all $s \in S$, as $n \to \infty$. The following Lemma provides a sufficient condition for this to hold.

Lemma 3.2. *Consider the probability space* $(\Omega^\infty, \mathcal{F}^\infty, \mathbb{P}^\infty)$. *Let* $\bar{g}_n(s, \mathbf{w})$ *and* $\hat{g}(s)$ *be defined as in* (3.13) *and* (3.14), *respectively, and given a compact set* $S \subset \mathbb{C}$, *let the following conditions hold:*

1. $g^{-1}(s, w)$ *is uniformly bounded on* $S \times \Omega$;

2. $\{\bar{g}_n(s, \mathbf{w}), n \in \mathbb{N}_+\}$ *are uniformly bounded on* $S \times \Omega^\infty$;

3. $\exists L > 0$ *s.t.* $|g_1^{-1}(s_1, w) - g_1^{-1}(s_2, w)| \leq L|s_1 - s_2|, \forall w \in \Omega, \forall s_1, s_2 \in S$;

4. *$\hat{g}(s)$ is uniformly continuous.*

Then, $\forall \epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\in S}\left\|\frac{1}{n}\bar{g}_n(s,\mathbf{w})\mathbb{1}\mathbb{1}^\top - \frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \geq \epsilon\right) = 0\,.$$

This lemma suggests that our coherent dynamics $\bar{g}_n(s,\mathbf{w})$, as n increases, converges uniformly on S to its expected version $\hat{g}(s)$. Then provided that the coherence is obtained as the network size grows, we would expect that the random transfer matrix $T_n(s,\mathbf{w})$ to concentrate to a deterministic one $\frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top$, as the following theorem shows.

Theorem 3.7. *Given probability space $(\Omega^\infty, \mathcal{F}^\infty, \mathbb{P}^\infty)$. Let $T_n(s,\mathbf{w})$ and $\hat{g}(s)$ be defined as in (3.12) and (3.14), respectively. Suppose $\lambda_2(L_n) \to +\infty$ as $n \to +\infty$. Given a compact set $S \subset \mathbb{C}$, if all the conditions in Lemma 3.2 hold, then $\forall \epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{s\in S}\left\|T_n(s,\mathbf{w}) - \frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \geq \epsilon\right) = 0\,.$$

Proof. Firstly, notice that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in S}\left\|T_n(s,\mathbf{w}) - \frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \geq \epsilon\right)\\
\leq\ &\mathbb{P}\left(\sup_{s\in S}\left\|T_n(s,\mathbf{w}) - \frac{1}{n}\bar{g}_n(s)\mathbb{1}\mathbb{1}^\top\right\| + \right.\\
&\qquad\qquad \left.\sup_{s\in S}\left\|\frac{1}{n}\bar{g}_n(s,\mathbf{w})\mathbb{1}\mathbb{1}^\top - \frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \geq \epsilon\right)\\
\leq\ &\mathbb{P}\left(\sup_{s\in S}\left\|T_n(s,\mathbf{w}) - \frac{1}{n}\bar{g}_n(s,\mathbf{w})\mathbb{1}\mathbb{1}^\top\right\| \geq \frac{\epsilon}{2}\right) + \\
&\qquad\qquad \mathbb{P}\left(\sup_{s\in S}\left\|\frac{1}{n}\bar{g}_n(s,\mathbf{w})\mathbb{1}\mathbb{1}^\top - \frac{1}{n}\hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \geq \frac{\epsilon}{2}\right)\,.
\end{aligned}$$

The second term converges to 0 as $n \to +\infty$ by Lemma 3.2. For the first term, we show below that it becomes exactly 0 for large enough n. Still, we assume $\{\bar{g}_n(s,\mathbf{w})\}$ and $\{g_i^{-1}(s,\mathbf{w})\}$ are uniformly bounded on S by $M_1, M_2 > 0$ respectively.

By Lemma 3.1, choosing large enough n s.t.

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in S}\left\|T_n(s,\mathbf{w})-\frac{1}{n}\bar{g}_n(s,\mathbf{w})\mathbb{1}\mathbb{1}^\top\right\|\geq\frac{\epsilon}{2}\right)\\
\leq\ &\mathbb{P}\left(\frac{(M_1M_2+1)^2}{F_l\lambda_2(L_n)-M_2-M_1M_2^2}\geq\frac{\epsilon}{2}\right)\ ,
\end{aligned}$$

then we can choose even larger n such that the probability on the right-hand side is 0 because $\lambda_2(L_n)\to+\infty$ as $n\to\infty$. □

In summary, because the coherent dynamics is given by the harmonic mean of all node dynamics $g_i(s)$, it concentrates to its harmonic expectation $\hat{g}(s)$ as the network size grows. As a result, in practice, the coherent behavior of large-scale networks depends on the empirical distribution of $g_i(s)$, i.e. a collective effect of all node dynamics rather than every individual node dynamics. For example, two different realizations of large-scale network with dynamics $T_n(s,\mathbf{w})$ exhibit similar coherent behavior with high probability, in spite of the possible substantial differences in individual node dynamics.

Remark 9. *With Theorem 3.7, one can adopt the analysis in Section 3.1.3 to derive a time-domain result similar to the one in Theorem 3.3. In this case, the network stability again relies on node passivity as required in Theorem 3.4. Nonetheless, for low-order rational transfer function, the condition of being passive is equivalent to its coefficients satisfying certain algebraic inequalities[95], hence there exists probability measure $\mathbb{P}$ on the coefficients such that the resulting transfer function is passive almost surely, under which the time-domain response of the network $T_n(s,\mathbf{w})$ can be inferred.*

3.1.5 Numerical Experiments

In this section, we apply our analysis to investigate coherence in power networks. For coherent generator groups, we find that $\frac{1}{n}\bar{g}(s)$ generalizes typical aggregate generator models which are often used for model reduction in power networks [96].

Moreover, we show that heterogeneity in generator dynamics usually leads to high-order aggregate dynamics, making it challenging to find a reasonably low-order approximation.

Consider the transfer matrix of power generator networks [48] linearized around its steady-state point, given by the following block diagram: This is exactly the block structure shown in Figure 3-1 with $f(s) = \frac{1}{s}$. Here, the network output, i.e., the frequency deviation of each generator, is denoted by ω. Generally, the $g_i(s)$ are modeled as strictly positive real transfer functions and we assume L is connected. Such interconnection is stable [85], regardless of the network connectivity.

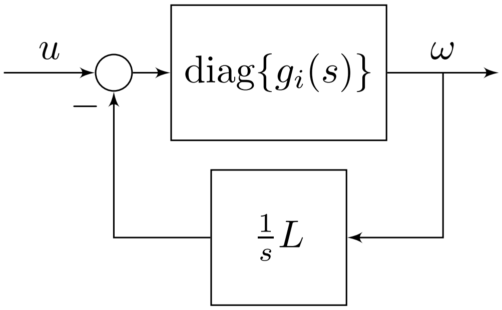

Figure 3-2. Block Diagram of Linearized Power Networks

We verify our theoretical results, Theorem 3.3 and Theorem 3.5, with numerical simulations on the Icelandic power grid [97] modeled as in Fig 3-2. We plot in Fig. 3-3 the frequency response of the power network model subject to various input disturbances. the network step response is more coherent, i.e. response of every single node (generator) is getting closer to the one of the coherent dynamics $\bar{g}(s)$, when the network connectivity is scaled up, as suggested by Theorem 3.3. In addition, the network responds more coherently when subject to lower-frequency signals (See the second and forth column in Fig 3-3), as suggested by Theorem 3.5. But most importantly, the coherent dynamics $\bar{g}(s)$ provides a good characterization of the coherent response. We also plot the Center-of-Inertia frequency of the grid $y_{\mathrm{COI}} = (\sum_{i=1}^{n} m_i y_i)/(\sum_{i=1}^{n} m_i)$, which is generally used for frequency response

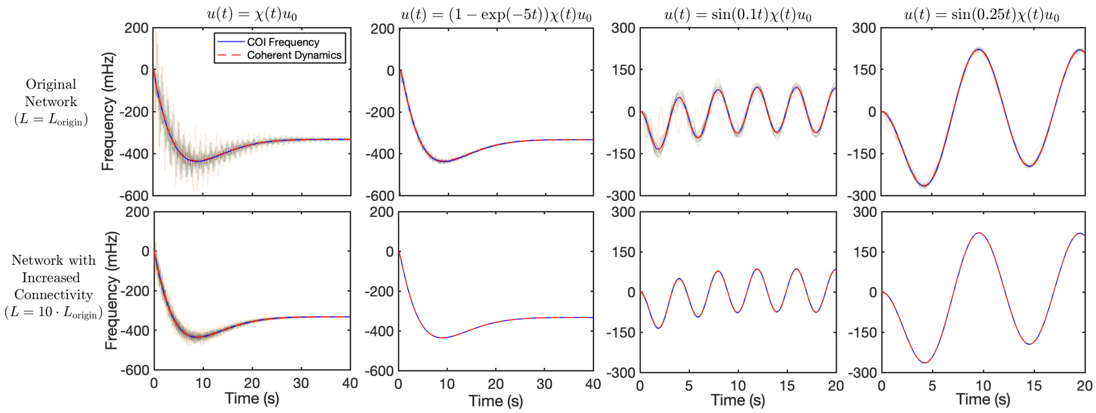

Figure 3-3. Coherent response of Icelandic Grid. Each column corresponds to a different input signal (from left to right: step, exponential approach, high-frequency sinusoidal, and low-frequency sinusoidal signal); The input signal has a shape $u_0 = -e_2$, i.e., only the second node is subject to disturbance. Top row shows the responses of original icelandic grid, and the bottom row shows the responses of network with increased connectivity. Red dashed line shows the response of $\bar{g}(s)$ subject to the averaged input $\bar{u}(t) = \mathbb{1}^\top u(t)/n$. Blue solid line shows the Center-of-Inertia frequency of the grid $y_{\text{COI}} = (\sum_{i=1}^n m_i y_i)/(\sum_{i=1}^n m_i)$.

assessment, and we see that it is well approximated by the response of $\bar{g}(s)$.

Proof of Theorem 3.3 and 3.5

We prove our time-domain results Theorem 3.3 and 3.5 here.

When the input to the network is $U(s)$, the output response of the i-th node is

$$Y_i(s) = e_i^\top T(s) U(s)\,,$$

where e_i is the i-th column of the identity matrix I_n.

Using Mellin's inverse formula [84, Theorem 3.20], we have

$$\begin{aligned}
|y_i(t) - \bar{y}(t)| &= \left| \frac{1}{2\pi j} \lim_{\omega\to\infty} \int_{\sigma - j\omega}^{\sigma + j\omega} e^{st} \left(Y_i(s) - e_i^\top \bar{g}(s) \mathbb{1} \frac{\mathbb{1}^\top}{n} U(s) \right) ds \right| \\
&\le \frac{e^\sigma}{2\pi} \lim_{\omega\to\infty} \int_{\sigma - j\omega}^{\sigma + j\omega} \left| e_i^\top T(s) U(s) - e_i^\top \bar{g}(s) \mathbb{1} \frac{\mathbb{1}^\top}{n} U(s) \right| ds \\
&\le \frac{e^\sigma}{2\pi} \lim_{\omega\to\infty} \int_{\sigma - j\omega}^{\sigma + j\omega} \left\| T(s) - \frac{1}{n} \bar{g}(s) \mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds \\
&= \frac{e^\sigma}{2\pi} \left((A) + (B) + (C) \right)\,,
\end{aligned}$$

where

$$(A) = \int_{\sigma - j\omega_0}^{\sigma + j\omega_0} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds\,,$$

$$(B) = \lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds\,,$$

$$(C) = \lim_{\omega\to\infty} \int_{\sigma - j\omega}^{\sigma - j\omega_0} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds\,.$$

Both proofs use such decomposition. By our assumption,

$$\begin{aligned}(B) = & \lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds \\ \leq & \lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \left(\|T(s)\| + \|\bar{g}(s)\|\right) \|U(s)\| ds \\ \leq & \; 2\gamma \lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \|U(s)\| ds\,,\end{aligned}$$

where the last inequality uses the fact that $\bar{g}(s)$ and $T(s)$ are stable:

$$\|\bar{g}(s)\|_{\mathcal{H}_\infty}, \|T(s)\|_{\mathcal{H}_\infty} \leq \gamma\,.$$

Because for the real input signals, we have $U(s^*) = U^*(s)$, hence $\int_{\sigma-j\omega}^{\sigma-j\omega_0} \|U(s)\| ds = \int_{\sigma+j\omega_0}^{\sigma+j\omega} \|U(s)\| ds$, which leads to

$$(C) \leq 2\gamma \lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \|U(s)\| ds\,.$$

Now we are ready to prove Theorem 3.3 and 3.5.

Proof of Theorem 3.3. First of all, Mellin's inverse formula requires that the vertical line $Re(s) = \sigma$ is on the right of all poles of the signal. This is the case from our assumption that $\sup_{Re(s)>\sigma} \|U(s)\| < +\infty$ and that $T(s), \bar{g}(s)$ being stable.

By the assumption that $\lim_{\omega\to\infty} \int_{\sigma+j0}^{\sigma+j\omega} \|U(s)\| ds$ is finite, one can pick an $\omega_0 > 0$, such that

$$\lim_{\omega\to\infty} \int_{\sigma + j\omega_0}^{\sigma + j\omega} \|U(s)\| ds \leq \frac{2\pi\epsilon}{6e^{\sigma}\gamma}\,,$$

which leads to

$$(B) \leq \ 2\gamma \lim_{\omega\to\infty} \int_{\sigma+j\omega_0}^{\sigma+j\omega} \|U(s)\| ds \leq \frac{2\pi\epsilon}{3e^{\sigma}} \,.$$

Similarly, we have $(C) \leq \frac{2\pi\epsilon}{3e^{\sigma}}$.

For the remaining term, we have

$$\begin{aligned}(A) = \ & \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \|U(s)\| ds \\ \leq \ & \sup_{w\in[-w_0,w_0]} \left\| T(\sigma+jw) - \frac{1}{n}\bar{g}(\sigma+jw)\mathbb{1}\mathbb{1}^\top \right\| \times \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \|U(s)\| ds\end{aligned}$$

Since $[\sigma - j\omega_0, \sigma + j\omega_0]$ is a compact set that satisfies the assumption in Theorem 3.1, we have

$$\lim_{\lambda_2(L)\to\infty} \sup_{w\in[-w_0,w_0]} \left\| T(\sigma+jw) - \frac{1}{n}\bar{g}(\sigma+jw)\mathbb{1}\mathbb{1}^\top \right\| = 0 \,.$$

Therefore, for sufficiently large $\lambda_2(L)$, we have $(A) \leq \frac{2\pi\epsilon}{3e^{\sigma}}$. Combining the upper-bounds for $(A), (B), (C)$, we have

$$|y_i(t) - \bar{y}(t)| \leq \epsilon \,.$$

Notice that the choice of $\lambda_2(L)$ does not depends on time t, hence this inequality holds for all $t > 0$. □

Proof of Theorem 3.5. Here, the input is a sinusoidal signal $U(s) = \frac{\alpha}{s^2+\alpha^2}u_0, u_0 \in \mathbb{S}^{n-1}$. Mellin's inverse formula requires that the vertical line $Re(s) = \sigma$ is on the right of all poles of the signal, which is satisfied under any choice $\sigma > 0$. For our purpose, we pick

$$\sigma = \alpha, \omega_0 = K\alpha \,,$$

for some $K > 0$ (to be determined later). By our assumption,

$$\begin{aligned}
(B) \leq\ & 2\gamma \lim_{\omega\to\infty} \int_{\sigma+j\omega_0}^{\sigma+j\omega} \left|\frac{\alpha}{s^2+\alpha^2}\right| \|u_0\| ds \\
=\ & 2\gamma \int_{\omega_0}^{+\infty} \frac{\alpha}{|(\sigma+j\omega)^2+\alpha^2|} d\omega \\
=\ & 2\gamma \int_{K\alpha}^{+\infty} \frac{\alpha}{|(\alpha+j\omega)^2+\alpha^2|} d\omega \\
=\ & 2\gamma \int_{K\alpha}^{+\infty} \frac{\alpha}{\sqrt{4\alpha^4+\omega^4}} d\omega \\
\leq\ & 2\sqrt{2}\gamma \int_{K\alpha}^{+\infty} \frac{\alpha}{2\alpha^2+\omega^2} d\omega \\
=\ & \gamma\left(\pi - 2\arctan\left(\frac{K}{\sqrt{2}}\right)\right) ,
\end{aligned} \tag{3.15}$$

where the last inequality use the fact that for $a, b > 0$, we have

$$\sqrt{a^2+b^2} \geq (a+b)/\sqrt{2}\,.$$

Similarly, we have

$$(C) \leq \gamma\left(\pi - 2\arctan\left(\frac{K}{\sqrt{2}}\right)\right) . \tag{3.16}$$

For the remaining term, we use the result in the proof of Theorem 3.2: $\exists\delta > 0$, such that $\forall s \in \mathcal{B}(0,\delta)$ such that

$$\left\|T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \leq \frac{2\left(M_1M_2+1\right)^2}{|f(s)|\lambda_2(L)}\,,$$

for some $M_1, M_2 > 0$. Then as long as we pick α, K appropriately such that $|\sigma +$

$j\omega_0| \leq \delta$, i.e., $\sqrt{1+K^2}\alpha \leq \delta$, we have

$$\begin{aligned}
(A) = & \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \left\| T(s) - \frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top \right\| \left| \frac{\alpha}{s^2+\alpha^2} \right| ds \\
\leq & \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \frac{2\left(M_1M_2+1\right)^2}{|f(s)|\lambda_2(L)} \left| \frac{\alpha}{s^2+\alpha^2} \right| ds \\
= & \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \frac{2\left(M_1M_2+1\right)^2}{\lambda_2(L)/|s|} \frac{\alpha}{|s^2+\alpha^2|} ds \\
= & \frac{2\left(M_1M_2+1\right)^2}{\lambda_2(L)} \int_{\sigma-j\omega_0}^{\sigma+j\omega_0} \frac{|s|\alpha}{|s^2+\alpha^2|} ds \\
= & \frac{4\left(M_1M_2+1\right)^2}{\lambda_2(L)} \int_{0}^{K\alpha} \frac{|\alpha+j\omega|\alpha}{|(\alpha+j\omega)^2+\alpha^2|} d\omega \\
= & \frac{4\left(M_1M_2+1\right)^2}{\lambda_2(L)} \int_{0}^{K\alpha} \frac{\sqrt{\alpha^2+\omega^2}\alpha}{\sqrt{4\alpha^4+\omega^4}} d\omega \\
\leq & \frac{2\sqrt{2}\left(M_1M_2+1\right)^2}{\lambda_2(L)} \int_{0}^{K\alpha} \frac{2(\alpha+\omega)\alpha}{2\alpha^2+\omega^2} d\omega\,,
\end{aligned}$$

where the last equality used the fact that for $a, b > 0$, we have

$$a+b \geq \sqrt{a^2+b^2} \geq (a+b)/\sqrt{2}\,,$$

to upper and lower bound the numerator and denominator respectively. Notice that

$$\begin{aligned}
& \int_{0}^{K\alpha} \frac{2(\alpha+\omega)\alpha}{2\alpha^2+\omega^2} d\omega \\
= & \ \alpha\left(\sqrt{2}\arctan\left(\frac{K}{\sqrt{2}}\right) + \log\left(1+\frac{K^2}{2}\right)\right) \\
\leq & \ 2\alpha\log\left(\frac{K^2}{2}\right)\,, \qquad (3.17)
\end{aligned}$$

for sufficiently large K. We have

$$(A) \leq \frac{4\sqrt{2}\left(M_1M_2+1\right)^2}{\lambda_2(L)}\alpha\log\left(\frac{K^2}{2}\right)\,. \qquad (3.18)$$

The last step is to find the right choice of α, K. Given $\epsilon > 0$, pick a $K > 0$, such that

$$2\gamma\left(\pi - 2\arctan\left(\frac{K}{\sqrt{2}}\right)\right) \leq \frac{\epsilon\pi}{2}\,.$$

Generally such a K is sufficient for (3.17) to hold. With this choice of K, let

$$\alpha_0 := \min\left\{\log 2, \frac{\epsilon\pi\lambda_2(L)}{8\sqrt{2}(M_1M_2+1)^2\log\left(\frac{K^2}{2}\right)}, \frac{\delta}{\sqrt{1+K^2}}\right\}.$$

Then, $\forall\alpha \leq \alpha_0$, combining (3.15)(3.16)(3.18), we have

$$\begin{aligned}|y_i(t)-\bar{y}(t)| &\leq \frac{e^{\sigma}}{2\pi}((A)+(B)+(C)) \\ &\leq \frac{e^{\alpha_0}}{2\pi}\left(2\gamma\left(\pi - 2\arctan\left(\frac{K}{\sqrt{2}}\right)\right) + \frac{4\sqrt{2}\,(M_1M_2+1)^2}{\lambda_2(L)}\alpha\log\left(\frac{K^2}{2}\right)\right) \\ &\leq \frac{1}{\pi}\left(\frac{\epsilon\pi}{2}+\frac{\epsilon\pi}{2}\right) = \epsilon\,.\end{aligned}$$

Notice that the choice of α_0, K does not depends on time t, nor the node index i, hence this inequality holds for all $t > 0$ and all $i \in [n]$. □

Proof of Theorem 3.4

Proof of Theorem 3.4. For each $g_i(s), i = 1, \cdots, n$, we have, by the OSP property,

$$Re(g_i(s)) \geq \epsilon|g_i(s)|^2, \forall Re(s) > 0\,.$$

That is,

$$Re(G(s)) \succeq \epsilon G^*(s)G(s)\,,$$

or equivalently, $\begin{bmatrix} G(s) \\ I \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} G(s) \\ I \end{bmatrix} \succeq 0\,.$ Since $g_i(s)$ are all OSP, then $g_i(s)$ is positive real [98]. A positive real function that is not zero function has no zero nor pole on the left half plane. Therefore $g_i(s)$ are invertible for all $Re(s) > 0$, which ensures that $G(s)$ is invertible for all $Re(s) > 0$. Then

$$(G^*(s))^{-1}\begin{bmatrix} G(s) \\ I \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} G(s) \\ I \end{bmatrix} G^{-1}(s) \succeq 0\,,$$

which is

$$\begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix} \succeq 0\,. \tag{3.19}$$

Notice that

$$T(s) = (I + G(s)f(s)L)^{-1}G(s) = (G^{-1}(s) + f(s)L)^{-1}\,,$$

then from (3.19) and the fact that $f(s)$ is PR, we have

$$\begin{aligned}
&\begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix} \\
= &\begin{bmatrix} I \\ G^{-1}(s) + f(s)L \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ G^{-1}(s) + f(s)L \end{bmatrix} \\
= &\begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix} + [f^*(s) + f(s)]L \\
\succeq &\begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ G^{-1}(s) \end{bmatrix} \succeq 0 \,.
\end{aligned}$$

Now for sufficiently large $\gamma > 0$, we have

$$\begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} + \begin{bmatrix} \frac{\epsilon}{2} I & 0 \\ 0 & -\gamma^2 \frac{\epsilon}{2} I \end{bmatrix} = \begin{bmatrix} -\frac{\epsilon}{2} I & I \\ I & -\gamma^2 \frac{\epsilon}{2} I \end{bmatrix} \preceq 0 \,,$$

since its Schur complement $(-\frac{\epsilon}{2} + \frac{2}{\epsilon\gamma^2})I \preceq 0$ for large γ. Therefore,

$$\begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\frac{\epsilon}{2} I & 0 \\ 0 & \gamma^2 \frac{\epsilon}{2} I \end{bmatrix} \begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix} \succeq \begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix}^* \begin{bmatrix} -\epsilon I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ T^{-1}(s) \end{bmatrix} \succeq 0 \,,$$

which is exactly, $\gamma^2 \frac{\epsilon}{2} (T^{-1}(s))^* (T^{-1}(s)) \succeq \frac{\epsilon}{2} I$. This shows that

$$\sigma_{min}^2(T^{-1}(s)) \geq \frac{1}{\gamma^2}, \forall Re(s) > 0 \,, \tag{3.20}$$

which is equivalent to $\|T(s)\|_2 \leq \gamma \,, \forall Re(s) > 0$. Moreover, (3.20) implies

$$|\bar{g}^{-1}(s)| = \left| \frac{\mathbb{1}^\top}{\sqrt{n}} T^{-1}(s) \frac{\mathbb{1}}{\sqrt{n}} \right|^2 \geq \frac{1}{\gamma^2}, \forall Re(s) > 0 \,,$$

which is equivalent to $\|\bar{g}(s)\|_2 \leq \gamma, \forall Re(s) > 0$. □

Proof of Lemma 3.2

Lemma 3.2. It suffices to show that $\forall \epsilon > 0$,

$$\lim_{n \to +\infty} \mathbb{P}\left(\sup_{s \in S} |\bar{g}_n(s, \mathbf{w}) - \hat{g}(s)| \geq \epsilon \right) = 0 \,, \tag{3.21}$$

since $|\bar{g}_n(s, \mathbf{w}) - \hat{g}(s)| = \left\| \frac{1}{n} \bar{g}_n(s, \mathbf{w}) \mathbb{1}\mathbb{1}^\top - \frac{1}{n} \hat{g}(s) \mathbb{1}\mathbb{1}^\top \right\|$.

By the assumptions, $\{\bar{g}_n(s, \mathbf{w}), n \in \mathbb{N}_+, \mathbf{w} \in \Omega^\infty\}$, and $\{g_i^{-1}(s, w), i \in \mathbb{N}_+, w \in \Omega\}$ are uniformly bounded by $M_1 > 0$ and $M_2 > 0$, respectively on S. Then, at any

$s \in S$, both $Re\left(g_i^{-1}(s,w)\right)$ and $Im\left(g_i^{-1}(s,w)\right)$ are random variables bounded within $[-M_2, M_2]$. We can simply bound their variances by

$$\text{Var}\left(Re\left(g_i^{-1}(s,w)\right)\right) \le (2M_2)^2 = 4M_2^2\,,\ \text{Var}\left(Im\left(g_i^{-1}(s,w)\right)\right) \le (2M_2)^2 = 4M_2^2\,.$$

Then it follows that

$$\text{Var}\left(Re\left(\bar{g}_n^{-1}(s,\mathbf{w})\right)\right) = \text{Var}\left(Re\left(n^{-1}\sum_{i=1}^{n} g_i^{-1}(s,w)\right)\right) \le 4M_2^2/n\,,$$

and

$$\text{Var}\left(Im\left(\bar{g}_n^{-1}(s,\mathbf{w})\right)\right) = \text{Var}\left(Im\left(n^{-1}\sum_{i=1}^{n} g_i^{-1}(s,w)\right)\right) \le 4M_2^2/n\,.$$

By definition of $\hat{g}(s)$ in (3.13), we have equalities $\mathbb{E}Re\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) = Re\left(\hat{g}(s)\right)$ and also $\mathbb{E}Im\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) = Im\left(\hat{g}(s)\right)$, then by Chebyshev's inequality, for $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\bar{g}_n^{-1}(s,\mathbf{w}) - \hat{g}^{-1}(s)\right| \ge \epsilon\right)\\
\le\ &\mathbb{P}\left(\left|Re\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) - Re\left(\hat{g}^{-1}(s)\right)\right| + \left|Im\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) - Im\left(\hat{g}^{-1}(s)\right)\right| \ge \epsilon\right)\\
\le\ &\mathbb{P}\left(\left|Re\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) - Re\left(\hat{g}^{-1}(s)\right)\right| \ge \epsilon/2\right)\\
&\qquad +\mathbb{P}\left(\left|Im\left(\bar{g}_n^{-1}(s,\mathbf{w})\right) - Im\left(\hat{g}^{-1}(s)\right)\right| \ge \epsilon/2\right) \qquad (3.22)\\
\le\ &\frac{4\text{Var}\left(Re\left(\bar{g}_n^{-1}(s,\mathbf{w})\right)\right)}{\epsilon^2} + \frac{4\text{Var}\left(Im\left(\bar{g}_n^{-1}(s,\mathbf{w})\right)\right)}{\epsilon^2}\\
\le\ &\frac{32M_2^2}{\epsilon^2 n}\,. \qquad (3.23)
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
|\bar{g}_n(s,\mathbf{w})| \le M_1 \Rightarrow\ &\left|\bar{g}_n^{-1}(s,\mathbf{w})\right| \ge \frac{1}{M_1}\\
\Rightarrow\ &\left|\bar{g}_n^{-1}(s,\mathbf{w}) - \hat{g}^{-1}(s) + \hat{g}^{-1}(s)\right| \ge \frac{1}{M_1}\\
\Rightarrow\ &\left|\hat{g}^{-1}(s)\right| \ge \frac{1}{M_1} - \left|\bar{g}_n^{-1}(s,\mathbf{w}) - \hat{g}^{-1}(s)\right|\,. \qquad (3.24)
\end{aligned}$$

Then given $\epsilon > 0, \forall n \in \mathbb{N}_+, \forall s \in S$, the following holds:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{g}(s)-\bar{g}_n(s,\mathbf{w})|\geq\epsilon\right)\\
&=\mathbb{P}\left(\left|\bar{g}_n(s,\mathbf{w})\hat{g}(s)\left(\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right)\right|\geq\epsilon\right)\\
&\leq\mathbb{P}\left(|\bar{g}_n(s,\mathbf{w})|\,|\hat{g}(s)|\,\left|\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right|\geq\epsilon\right)\\
&\leq\mathbb{P}\left(M_1\left|\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right|\geq\epsilon|\hat{g}^{-1}(s)|\right)\\
\text{(3.24)}\quad&\leq\mathbb{P}\left(M_1\left|\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right|\geq\frac{\epsilon}{M_1}-\epsilon\left|\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right|\right)\\
&=\mathbb{P}\left(\left|\bar{g}_n^{-1}(s,\mathbf{w})-\hat{g}^{-1}(s)\right|\geq\frac{\epsilon}{M_1(M_1+\epsilon)}\right)\\
\text{(3.23)}\quad&\leq\frac{32M_2^2M_1^2(M_1+\epsilon)^2}{\epsilon^2 n}.
\end{aligned}$$

By taking $n \to +\infty$ on both sides, we prove that $\bar{g}_n(s, \mathbf{w})$ converges point-wise to $\hat{g}(s)$ on S.

We now show that $\bar{g}_n(s, \mathbf{w})$ is also stochastic equicontinuous on S. For the definition of stochastic equicontinuity, please refer to [99]. We already assumed that $\bar{g}_n(s, \mathbf{w}) \leq M_1, \forall \mathbf{w} \in \Omega^\infty, s \in S$. Then $\forall \mathbf{w} \in \Omega^\infty, \forall s_1, s_2 \in S$, we have

$$\begin{aligned}
&|\bar{g}_n(s_1,\mathbf{w})-\bar{g}_n(s_2,\mathbf{w})|\\
\leq\ &\left|\bar{g}_n(s_1,\mathbf{w})||\bar{g}_n(s_2,\mathbf{w})||\bar{g}_n^{-1}(s_1,\mathbf{w})-\bar{g}_n^{-1}(s_2,\mathbf{w})\right|\\
\leq\ &M_1^2\left|\sum_{i=1}^{n}\left(g_i^{-1}(s_1,w_i)-g_i^{-1}(s_2,w_1)\right)\right|\\
\leq\ &M_1^2\sum_{i=1}^{n}\left|g_i^{-1}(s_1,w_i)-g_i^{-1}(s_2,w_i)\right|\leq nM_1^2L|s_1-s_2|\,,
\end{aligned}$$

where the last inequality is from our third assumption and also the fact that $g_i^{-1}(s, w) = g_1^{-1}(s, w)$ (identically distributed as random functions). By [99, Corollary 2.2], the inequality above is sufficient to establish stochatic equicontinuity of $\bar{g}_n(s, \mathbf{w})$ on S, and combining point-wise convergence and the fourth assumption that $\hat{g}(s)$ is uniform continuous, we get the uniform convergence of $\bar{g}_n(s, \mathbf{w})$ to $\hat{g}(s)$ on S, which gives (3.21). □

3.2 Networks with Multiple Coherent Clusters

We have shown, in the last section, that, under mild assumptions, the following holds[2] for almost any $s_0 \in \mathbb{C}$,

$$\lim_{\lambda_2(L)\to\infty} \|T(s_0) - \hat{g}(s_0)\mathbb{1}\mathbb{1}^\top\| = 0\,, \tag{3.25}$$

where

$$\hat{g}(s) = \left(\sum_{i=1}^{n} g_i^{-1}(s)\right)^{-1}\,. \tag{3.26}$$

That is, when the algebraic connectivity $\lambda_2(L)$ of the network is high, one can approximate $T(s)$ by a rank-one transfer matrix. Such a rank-one transfer matrix $\hat{g}(s_0)\mathbb{1}\mathbb{1}^\top$ precisely describes the coherent behavior of the network: The network takes the aggregated input $\hat{u} = \mathbb{1}^\top u = \sum_{i=1}^n u_i$, and responds coherently as $\hat{y}\mathbb{1}$, where $\hat{y} = \hat{g}(s)\hat{u}$. Therefore, it suffices to study $\hat{g}(s)$ to understand the coherent behavior in a tightly-connected network.

However, practical networks are not necessarily tightly-connected. Instead, they often contain multiple groups of nodes such that within each group, the nodes are tightly-connected while between groups, the nodes are weakly-connected. Then the network dynamics can be reduced to dynamic interactions among these groups. To approximate such interaction, it is natural first to identify *coherent groups*, or *coherent clusters*, in the network, then apply the aforementioned analysis to obtain the coherent dynamics $\hat{g}(s)$ for each group, and replace the entire coherent group by an aggregate node with $\hat{g}(s)$. Lastly, one needs to find a reduced network of the same size as the number of coherent groups, which characterize the interaction among these groups. The aggregate dynamics and the reduced network allow us to build a network model with exactly the same structure as the one in Figure 3-1 but with a much smaller size, for which we refer to such an approach as *structure-*

[2] In Section 3.1, the transfer matrix $\frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top$ appeared in the limit, where $\bar{g}(s) = \left(\frac{1}{n}\sum_{i=1}^n g_i^{-1}(s)\right)^{-1}$. It is easy to verify that $\frac{1}{n}\bar{g}(s)\mathbb{1}\mathbb{1}^\top = \hat{g}(s_0)\mathbb{1}\mathbb{1}^\top$

preserving model reduction and call the resulting reduction model *structure-preserving*. Figure 3-4 shows our proposed reduced model in the case of three coherent groups, for which the algorithm details are explained later.

Our algorithm

In this section, we propose a structure-preserving model reduction algorithm for networks with an arbitrary number of groups.

```
Algorithm 1: Structure-Preserving Network Reduction via Spectral Clustering
  Data: Network Model $(G(s) = \mathrm{diag}\{g_i(s)\}_{i=1}^n, L, f(s))$; Number of clusters $k$
   Do:
    1. $(\{\mathcal{I}_i\}_{i=1}^k, V_k, \Lambda_k) \leftarrow$ SpectralClustering$(L)$; // Spectral clustering
       Construct $P_{\{\mathcal{I}_i\}_{i=1}^k}$ as in (3.29);
    2. $\hat{g}_i(s) \leftarrow \left(\sum_{j\in\mathcal{I}_i} g_j^{-1}(s)\right)^{-1}$, $i = 1, \cdots, k$; // Aggregation
       $\hat{G}(s) = \mathrm{diag}\{\hat{g}_i(s)\}_{i=1}^k$;
    3. $S \leftarrow$(Solution to (3.34));
       $L_k = (S^{-1})^\top \Lambda_k S^{-1}$; // Construct reduced network
  Result: $\hat{T}_k(s) \leftarrow P_{\{\mathcal{I}_i\}_{i=1}^k}(I_k + \hat{G}_k(s)L_k f(s))^{-1}\hat{G}_k(s)P_{\{\mathcal{I}_i\}_{i=1}^k}^\top$
```

This algorithm, whose rationale will be explained in detail in Section 3.2.1, follows the same procedure as we discussed in the previous section: Firstly, we utilize some spectral clustering algorithm to obtain a k-way partition $\{\mathcal{I}_i\}_{i=1}^n$ of $[n]$ that encodes the clustering results. Notice that here any spectral clustering algorithm works. For subsequent steps, we also need to keep the first k smallest eigenvalues of L (in $\Lambda_k = \mathrm{diag}\{\lambda_i(L)\}_{i=1}^k$) and their associated eigenvectors (in $V_k = \begin{bmatrix} v_1(L) & v_2(L) & \cdots v_k(L) \end{bmatrix}$). Then the nodes in the same group $\mathcal{I}_i$ are aggregated into $\hat{g}_i(s)$. Lastly, the Laplacian matrix of the reduced network is constructed after solving an optimization problem (3.34) that can be viewed as a refinement process

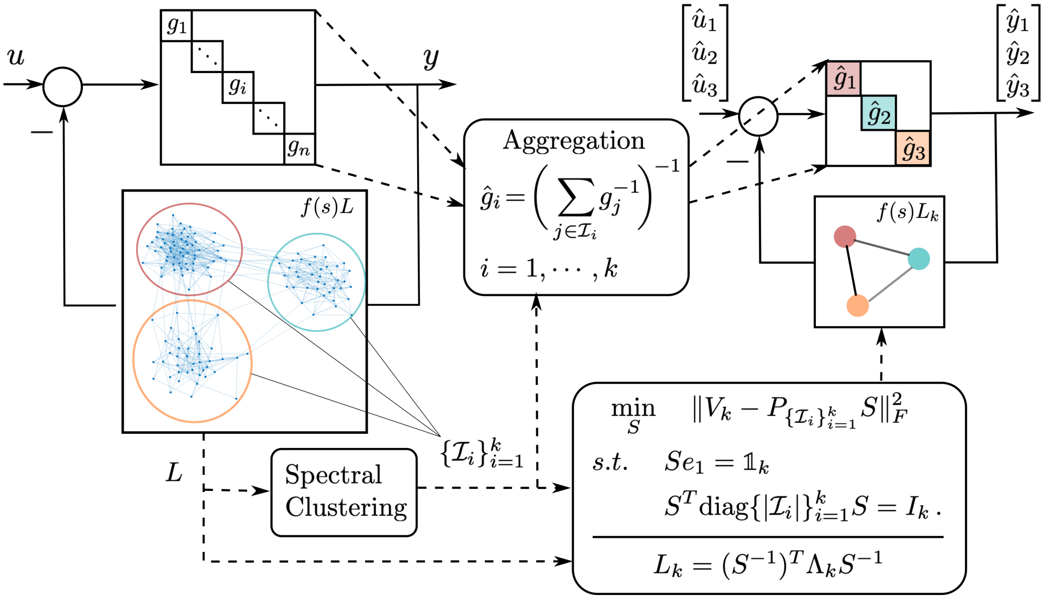
u
$-$
g_1
g_i
g_n
y
$f(s)L$
L
Spectral Clustering
$\{\mathcal{I}_i\}_{i=1}^k$
Aggregation
$\hat{g}_i = \left(\sum_{j\in\mathcal{I}_i} g_j^{-1}\right)^{-1}$
$i = 1, \cdots, k$
$\begin{bmatrix}\hat{u}_1\\ \hat{u}_2\\ \hat{u}_3\end{bmatrix}$
$-$
$\hat{g}_1$
$\hat{g}_2$
$\hat{g}_3$
$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\end{bmatrix}$
$f(s)L_k$
$\min_S \quad \|V_k - P_{\{\mathcal{I}_i\}_{i=1}^k} S\|_F^2$
$s.t. \quad Se_1 = \mathbb{1}_k$
$S^T \mathrm{diag}\{|\mathcal{I}_i|\}_{i=1}^k S = I_k\,.$
$L_k = (S^{-1})^T \Lambda_k S^{-1}$

Figure 3-4. Functional illustration of Algorithm 1.

on the Laplacian spectral embedding V_k. This algorithm will return a transfer matrix $\hat{T}_k(s)$ as an approximation model of the original transfer matrix $T(s)$. The algorithm is illustrated in Figure 3-4.

In the rest of the section, we first discuss how our algorithm is constructed based on the aforementioned coherence analysis [100] in Section 3.2.1, then show that our proposed approximation model is asymptotically accurate in a random graph setting where the network graph is sampled from a *weighted stochastic block model* [101] by showing an approximation error bound between the network $T(s)$ and the proposed reduced model $\hat{T}_k(s)$ (in Section 3.2.2). Lastly, we verify our theoretical findings through a numerical simulation in Section 3.2.3.

3.2.1 Structure-preserving network reduction via spectral clustering

Our algorithm roots in the analysis in Section 3.1 showing that the network transfer matrix $T(s)$ is approximately low rank for networks with Laplacian matrices satisfying some spectral property. Such a low-rank approximation is generally not structure-preserving, for which we use its closest structure-preserving approximation, obtained by spectral clustering on graph Laplacian L and a refinement process on its eigenvectors V_k, as our final reduction model for the original $T(s)$.

Low-rank approximation of network transfer matrix

Given the network Laplacian L and its first k smallest eigenvalues (in a diagonal matrix) $\Lambda_k = \mathrm{diag}\{\lambda_i(L)\}_{i=1}^k$ and the eigenvectors $V_k = \begin{bmatrix} v_1(L) & v_2(L) & \cdots v_k(L) \end{bmatrix}$ (we also refer it as *Laplacian spectral embedding*), we define the following rank-k transfer matrix

$$T_k(s) = V_k(V_k^\top G^{-1}(s)V_k + f(s)\Lambda_k)^{-1}V_k^\top \,, \tag{3.27}$$

and we have the following result:

Theorem 3.8. *For $s_0 \in \mathbb{C}$ that is not a pole of $f(s)$ and has these two quantities*

$$\|T_k(s_0)\| := M_1, \textit{and } \max_{1\leq i\leq n} |g_i^{-1}(s_0)| := M_2 \,,$$

finite. Then whenever $|f(s_0)|\lambda_{k+1}(L) > M_2 + M_1 M_2^2$, the following inequality holds:

$$\|T(s_0) - T_k(s_0)\| \leq \frac{(M_1 M_2 + 1)^2}{|f(s_0)|\lambda_{k+1}(L) - M_2 - M_1 M_2^2} \,. \tag{3.28}$$

Theorem 3.8 shows that in the large $\lambda_{k+1}(L)$ regime, one can somewhat approximate the original transfer matrix $T(s)$ by a low-rank one $T_k(s)$, but the approximation result in (3.28) is weaker than that the two transfer matrices $T(s)$ and $T_k(s)$ are close in the $\mathcal{H}_\infty$ sense. It heavily depends on the choice of s_0, the frequency of interest, as we should not expect $T(s)$ and $T_k(s)$ to behave similarly under input of any frequency. For the case of $k = 1$ (In Section 3.1), we have shown that if $\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\|$ is small for some $\eta > 0$, then one can show, provided that $T(s)$ and $T_k(s)$ are stable, the time domain responses of the two transfer matrices under low-frequency inputs (characterized by η) are close to each other.

Following such observation, we consider any $\hat{T}_k(s)$ with $\sup_{s\in(-j\eta,+j\eta)} \|T(s) - \hat{T}_k(s)\|$ being small for some $\eta > 0$ as a good approximation for the original network. Applying (3.28) uniformly over $\{s : s \in (-j\eta, +j\eta)\}$, one can show that $T_k(s)$ is such a good approximation when $\lambda_{k+1}(L)$ is large. However, $T_k(s)$ is, in general, not structure-preserving, and thus may not be interpreted as a reduced network of aggregate nodes. Therefore, we need to find a structure-preserving $\hat{T}_k(s)$ that is close to $T_k(s)$.

Structured low-rank approximation via spectral embedding refinement

We first discuss the case when $T_k(s)$ is structure-preserving. We show that a special property on the Laplacian spectral embedding V_k suffices. For some $\mathcal{I} \subseteq [n]$, we let $\mathbb{1}_{\mathcal{I}}$ be an $n \times 1$ vector such that $[\mathbb{1}_{\mathcal{I}}]_i = \begin{cases} 1, & i \in \mathcal{I} \\ 0, & i \notin \mathcal{I} \end{cases}$.

Definition 3.2. *A Laplacian matrix L is said to be* ***k-block-ideal*** *with respect to a k-way partition $\{\mathcal{I}_1, \cdots, \mathcal{I}_k\}$ of $[n]$, if there exists some invertible matrix $S \in \mathbb{R}^{k\times k}$ such that*

$$V_k := \begin{bmatrix} v_1(L) & v_2(L) & \cdots & v_k(L) \end{bmatrix} = \begin{bmatrix} \mathbb{1}_{\mathcal{I}_1} & \mathbb{1}_{\mathcal{I}_2} & \cdots & \mathbb{1}_{\mathcal{I}_k} \end{bmatrix} S\,.$$

We also say V_k is k-block-ideal in this case.

A k-block-ideal spectral embedding V_k, together with Λ_k containing the bottom k eigenvalues of L, would immediately lead to a reduced network: the k coherent groups are determined by the k-way partition $\{\mathcal{I}_i\}_{i=1}^k$, and the invertible matrix S, combined with Λ_k, characterize the interconnection in the reduced network, as show in the following theorem:

Theorem 3.9. *Given a k-block-ideal Laplacian L associated with a partition $\{\mathcal{I}_1, \cdots, \mathcal{I}_k\}$ and an invertible matrix S, and we define*

$$P_{\{\mathcal{I}_i\}_{i=1}^k} := \begin{bmatrix} \mathbb{1}_{\mathcal{I}_1} & \mathbb{1}_{\mathcal{I}_2} & \cdots & \mathbb{1}_{\mathcal{I}_k} \end{bmatrix}\,, \tag{3.29}$$

then

$$T_k(s) = P_{\{\mathcal{I}_i\}_{i=1}^k}(I_k + \hat{G}_k(s)L_kf(s))^{-1}\hat{G}_k(s)P_{\{\mathcal{I}_i\}_{i=1}^k}^\top\,, \tag{3.30}$$

where $\hat{G}(s) = \mathrm{diag}\{\hat{g}_i(s)\}_{i=1}^k\,,\ \hat{g}_i(s) = \left(\sum_{j\in\mathcal{I}_i} g_j^{-1}(s)\right)^{-1}$ *and* $L_k = (S^{-1})^\top \Lambda_k S^{-1}$.

Proof of Theorem 3.9. Since

$$T_k(s) = V_k(V_k^\top G^{-1}(s)V_k + f(s)\Lambda_k)^{-1}V_k^\top\,, \tag{3.31}$$

and

$$V_k = P_{\{\mathcal{I}_i\}_{i=1}^k}S\,, \tag{3.32}$$

we have

$$\begin{aligned}
T_k(s) =&\ P_{\{\mathcal{I}_i\}_{i=1}^k}S\left(S^\top P_{\{\mathcal{I}_i\}_{i=1}^k}^\top G^{-1}(s)P_{\{\mathcal{I}_i\}_{i=1}^k}S + f(s)\Lambda_k\right)^{-1}S^\top P_{\{\mathcal{I}_i\}_{i=1}^k}^\top \\
=&\ P_{\{\mathcal{I}_i\}_{i=1}^k}\left((S^\top)^{-1}\left(S^\top P_{\{\mathcal{I}_i\}_{i=1}^k}^\top G^{-1}(s)P_{\{\mathcal{I}_i\}_{i=1}^k}S + f(s)\Lambda_k\right)S^{-1}\right)^{-1}P_{\{\mathcal{I}_i\}_{i=1}^k}^\top \\
=&\ P_{\{\mathcal{I}_i\}_{i=1}^k}\left(P_{\{\mathcal{I}_i\}_{i=1}^k}^\top G^{-1}(s)P_{\{\mathcal{I}_i\}_{i=1}^k} + f(s)(S^\top)^{-1}\Lambda_k S^{-1}\right)^{-1}P_{\{\mathcal{I}_i\}_{i=1}^k}^\top \\
=&\ P_{\{\mathcal{I}_i\}_{i=1}^k}\left(\hat{G}_k^{-1}(s) + f(s)L_k\right)^{-1}P_{\{\mathcal{I}_i\}_{i=1}^k}^\top \\
=&\ P_{\{\mathcal{I}_i\}_{i=1}^k}\left(I + \hat{G}(s)L_kf(s)\right)^{-1}\hat{G}(s)P_{\{\mathcal{I}_i\}_{i=1}^k}^\top\,.
\end{aligned}$$

□

Theorem 3.9 shows that under k-block-ideal V_k, the dynamical behavior of T_k is structure-preserving since it is fully characterized by a reduced network with k nodes, with nodal dynamics $\hat{G}(s)$ and network coupling L_k. Each node $\hat{g}_i(s)$ represents the aggregate dynamics for nodes in $\mathcal{I}_i$. Any input u to $T_k(s)$ is aggregated into $\begin{bmatrix}\hat{u}_1 & \cdots & \hat{u}_k\end{bmatrix}^\top = P_{\{\mathcal{I}_i\}_{i=1}^k}^\top u$ as the input to the reduced network. Then the output $\begin{bmatrix}\hat{y}_1 & \cdots & \hat{y}_k\end{bmatrix}^\top$ is "broadcast" to the original nodes via $P_{\{\mathcal{I}_i\}_{i=1}^k}$ such that every node in the same $\mathcal{I}_i$ has the same response.

Notice that such structure-preserving property only depends on the Laplacian spectral embedding V_k. For V_k that is not k-block-ideal, we should be able to find a $\hat{V}_k$ close to V_k and is k-block-ideal. This gives rise to the following optimization problem:

$$\min_{S,\{\mathcal{I}_i\}_{i=1}^k} \|V_k - P_{\{\mathcal{I}_i\}_{i=1}^k} S\|_F^2, \quad s.t. \quad Se_1 = \mathbb{1}_k/\sqrt{n},\ S^\top \text{diag}\{|\mathcal{I}_i|\}_{i=1}^k S = I_k\,. \tag{3.33}$$

The resulting $\hat{V}_k = P_{\{\mathcal{I}_i\}_{i=1}^k} S$ is a refinement of V_k that is k-ideal, and the constraints in (3.33) ensures that the first column of $\hat{V}_k$ is $\mathbb{1}_n/\sqrt{n}$ and that $\hat{V}_k^\top \hat{V}_k = I_k$. Now

$$\hat{T}_k(s) = \hat{V}_k(\hat{V}_k^\top G^{-1}(s)\hat{V}_k + f(s)\Lambda_k)^{-1}\hat{V}_k^\top$$

is structure-preserving by Theorem 3.9. In the optimization problem (3.33), the need for identifying coherent groups is implicitly suggested by the fact that we are optimizing over all possible k-way partitions of n, and the reduced network interconnection is constructed by jointly optimizing over invertible S.

Generally, (3.33) is hard to solve. Notice, however, that given a fixed partition $\{\mathcal{I}_i\}_{i=1}^k$, one can find a closed-form solution (We show it at the end of this section) to the following optimization problem

$$\min_{S} \|V_k - P_{\{\mathcal{I}_i\}_{i=1}^k} S\|_F^2, \quad s.t. \quad Se_1 = \mathbb{1}_k/\sqrt{n},\ S^\top \text{diag}\{|\mathcal{I}_i|\}_{i=1}^k S = I_k\,. \tag{3.34}$$

This suggests that a computationally efficient way to find a sub-optimal solution to (3.33): First, we use any spectral clustering algorithm to find a good partition/clustering $\{\mathcal{I}_i\}_{i=1}^k$, then refine the spectral embedding V_k by optimizing (3.34) with the obtained partition, resulting in our Algorithm 1.

3.2.2 Performance analysis

In this section, we provide an error bound on $\sup_{s\in(-j\eta,j\eta)} \|T(s) - \hat{T}_k(s)\|$ for our proposed approximation model $\hat{T}_k(s)$ from Algorithm 1. As we discussed in Section 3.2.1, such error measure is related to how close the time-domain response of $\hat{T}_k(s)$ is to the one of $T(s)$ when subjected to low-frequency inputs. We consider a Laplacian sampled from a stochastic weighted block model.

Weighted stochastic block model

We first discuss how we sample our Laplacian matrix from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$. Here, $\{\mathcal{I}_i\}_{i=1}^k$ is a k-way partition of $[n]$, $Q \in [0,1]^{k\times k}$, and $W \in \mathbb{R}_{\geq 0}^{k\times k}$, where $Q_{ij} = Q_{ji}, W_{ij} = W_{ji}$. We let (j) denote the *block membership* of node j: when $j \in \mathcal{I}_i$, then $(j) = i$. The adjacency matrix A is sampled as follows:

$$A_{ij} = \begin{cases} W_{(i),(j)}, & \text{with probability } Q_{(i),(j)} \\ 0, & \text{with probability } 1 - Q_{(i),(j)} \end{cases}, \quad i \geq j, \qquad A_{ij} = A_{ji}, \quad i < j\,. \tag{3.35}$$

That is, each (undirected) edge i, j appears independently with probability $Q_{(i),(j)}$ that is determined by the block membership of node i, j, and has weight $W_{(i),(j)}$ if it appears. Then we have the Laplacian matrix L:

$$L = D_A - A, \quad D_A = \mathrm{diag}\{A\mathbb{1}\}\,. \tag{3.36}$$

Approximation error bound

Given the network model $(G(s), L, f(s))$ with L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$, we show that under certain assumptions, the error $\sup_{s\in(-j\eta,j\eta)} \|T(s) - \hat{T}_k(s)\|$ is small with high probability when the network size is sufficiently large. We start by stating our assumptions.

Assumption 3.1. *For our network model $(G(s), L, f(s))$ with L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$, we assume the following:*

1. *All $g_i(s), f(s)$ are rational. Moreover, node dynamics are **output strictly passive**: There exists $\gamma > 0$, such that for $i = 1, \cdots, n$, $Re(g_i(s)) \geq \frac{1}{\gamma}|g_i(s)|^2, \forall Re(s) > 0$, and network coupling $f(s)$ is **positive real**: $Re(f(s)) > 0, \forall Re(s) > 0$, and $Im(f(s)) = 0, \forall Re(s) = 0$*

2. *The node dynamics satisfies that for any $\eta > 0$, there exists $M(\eta)$ such that for $i = 1, \cdots, n$*

$$\sup_{s\in(-j\eta,+j\eta)} |g_i^{-1}(s)| \leq M(\eta)\,. \tag{3.37}$$

The network coupling $f(s)$ satisfies that $F_l(\eta) := \inf_{s\in(-j\eta,+j\eta)} |f(s)|$ is positive for all $\eta > 0$.

3. *The blocks are approximately balanced:*

$$\frac{n_{\max}}{n_{\min}} \leq \rho\,, \tag{3.38}$$

for some $\rho \geq 1$, where $n_{\max} := \max_{1\leq i\leq k} |\mathcal{I}_i|$ and $n_{\min} := \min_{1\leq i\leq k} |\mathcal{I}_i|$,

4. *The network has a stronger intra-block connection than the inter-block one:*

$$\min_i B_{ii} - 2\rho \max_i \sum_{j\neq i} B_{ij} \geq \Delta\,, \tag{3.39}$$

for some $\Delta > 0$, where $B = Q \odot W$. ($\odot$ is the Hadamard product)

The first assumption ensures the network $T(s)$ and our approximation model $\hat{T}_k(s)$ are stable. The second assumption ensures that our low-rank approximation $T_k(s)$ in Theorem 3.8 is valid on the interval of our interest $(-j\eta, +j\eta)$. The third assumption ensures our problem is non-degenerate: if the size of one block is too small, the network effectively has $k-1$ clusters. Such an assumption is standard in analyzing the consistency of spectral clustering algorithms on stochastic block models [102, 101]. Lastly, since we are interested in networks containing multiple groups of nodes such that within each group, the nodes are tightly-connected while between groups, the nodes are weakly-connected, the fourth assumption formally characterizes such a property.

In our algorithm, a spectral clustering algorithm is used to find a partition $\{\mathcal{I}_i\}_{i=1}^k$ that is used for aggregating node dynamics and constructing the reduced network. Ideally, we want some consistency property on the obtained partition.

Assumption 3.2. *Given L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$ satisfying Assumption 3.1, we have an asymptotically consistent spectral clustering algorithm in Algorithm 1: For any $\delta > 0$, there exists $\tilde{N}(\delta)$ such that for network with size $n > \tilde{N}(\delta)$, the spectral clustering algorithm on L returns the true $\{\mathcal{I}_i\}_{i=1}^k$ partition with probability at least $1 - \delta$.*

Formally justifying this assumption for some spectral clustering algorithms is an interesting future research topic. Nonetheless, such a consistency result has been studied for spectral clustering algorithms on the adjacency matrix from the stochastic block model [102] and the weighted stochastic block model [101].

With these assumptions, we have the following theorem regarding the error bound.

Theorem 3.10. *Consider the network model $(G(s), L, f(s))$ with L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$. If Assumption 3.1 and Assumption 3.2 hold, then Algorithm 1 returns a $\hat{T}_k(s)$ such that*

1. $\|T(s)\|_{\mathcal{H}_\infty} \leq \gamma$, $\|\hat{T}_k(s)\|_{\mathcal{H}_\infty} \leq \gamma$;

2. *For any $\eta, \epsilon > 0$, and $0 < \delta < 1$, there exists $N(\delta, \epsilon, \tilde{N}(\delta/2), \gamma, M(\eta), F_l(\eta), \rho, Q, W)$ such that for network with size $n \geq N$, with probability at least $1 - \delta$, we have*

$$\sup_{s \in (-j\eta, +j\eta)} \|T(s) - \hat{T}_k(s)\| \leq \epsilon\,. \tag{3.40}$$

Proof Sketch. For the stability of $T(s)$ and $\hat{T}_k(s)$, the proof is similar to the one in [100] and uses the assumption $g_i(s)$ are output strictly passive and $f(s)$ is positive real. The error bound relies on that the sampled Laplacian matrix L is close to one that is easy to analyze: Let A_{blk} be the expected value of the adjacency matrix A from the block model, and we can construct a Laplacian matrix $L_{\text{blk}} = D_{A_{\text{blk}}} - A_{\text{blk}},\ D_{A_{\text{blk}}} = \operatorname{diag}\{A_{\text{blk}}\mathbb{1}\}$. L_{blk} has (by (3.38)(3.39) in Assumption 3.1) all the desired properties:

1) $\lambda_{k+1}(L_{\text{blk}})$ grows linearly in network size n; 2) L_{blk} is k-block-ideal. [103] has shown that under the weighted stochastic block model, $\|L - L_{\text{blk}}\| \sim \mathcal{O}_p(\sqrt{n \log n})$, which is sufficient to show that 1) $\lambda_{k+1}(L) \sim \Omega_p(n)$ by Weyl's inequality [64]; 2) L is approximately k-block-ideal by Davis-Khan theorem [104]. The former shows that the error between $T(s)$ and $T_k(s)$ is small w.h.p. by Theorem 1 and the latter ensures the error between $T_k(s)$ and $\hat{T}_k(s)$ is small w.h.p.. □

Theorem 3.10 shows that our algorithms perform well for large networks with multiple coherent clusters, it also implies that the collective dynamic behavior of such networks can be modeled as a structured reduced network. This suggests a new avenue for data-driven system identification for such networks where only the reduced network model is learned from the data collected from the network.

3.2.3 Numerical experiments

The frequency response of synchronous generator (including grid-forming inverters) networks, linearized at its equilibrium point [34], can be modeled exactly as the network model in Fig 3-1 with $f(s) = \frac{1}{s}$ and second order node dynamics $g_i(s)$. We validate our algorithm with a synthetic test case, where the coefficients of generator dynamics are randomly sampled. The network adjacency matrix A is sampled from our weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$ with $k = 3$, and

$$\begin{bmatrix} |\mathcal{I}_1| & 0 & 0 \\ 0 & |\mathcal{I}_2| & 0 \\ 0 & 0 & |\mathcal{I}_3| \end{bmatrix} = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 40 & 0 \\ 0 & 0 & 20 \end{bmatrix}, Q = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, W = \begin{bmatrix} 20 & 0.4 & 0.8 \\ 0.4 & 20 & 0.7 \\ 0.8 & 0.7 & 20 \end{bmatrix}. \tag{3.41}$$

We use the spectral clustering algorithm proposed in [105]. Since the network size is not sufficiently large for the algorithm to return a true partition with high probability, when we run the experiments with multiple random seeds, we see a small fraction of the runs in which the algorithm fails to cover the true partition. For the case when the spectral clustering algorithm succeeds, we inject a step disturbance $u_2(t) = \chi(t)$

at the second node of the network and plot the step response of $T(s)$ in Fig 3-5, along with the response $\hat{y}$ of our approximate model $\hat{T}_3(s)$ from Algorithm 1. There is a clear difference between the dynamical response of generators from different groups, and the aggregate responses $\hat{y}$ capture such difference while providing a good approximation to the actual node responses. Here we only present the result of running Algorithm 1 on one instance of the randomly generated networks, but the results are consistent across multiple runs as long as the spectral clustering succeeds.

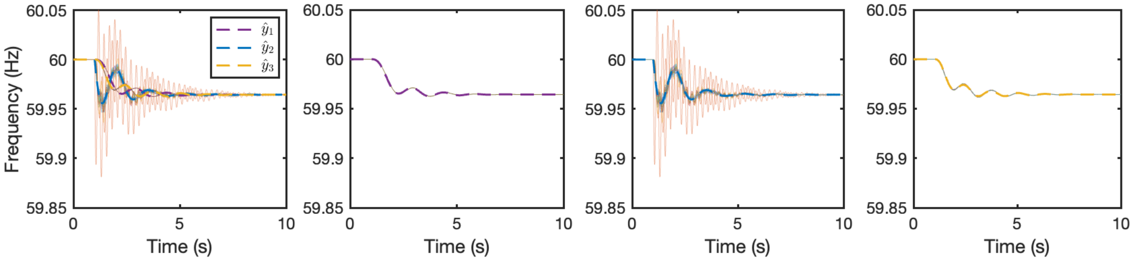

Figure 3-5. Left most plot shows the step response of $T(s)$ (solid lines) and $\hat{T}_k(s)$ (dashed lines) from algorithm 1. The three plots on the right show the response for each identified group $\mathcal{I}_i$. The node injected with step disturbance is in the group 2.

Solution to the Laplacian spectral embedding refinement problem in 3.34

In this section, we derive the analytical solution to (3.34):

$$\begin{aligned} \min_{S\in\mathbb{R}^{k\times k}} \quad & \|V_k - P_{\{\mathcal{I}_i\}_{i=1}^k} S\|_F^2 \\ s.t. \quad & Se_1 = \mathbb{1}_k/\sqrt{n} \\ & S^\top \mathrm{diag}\{|\mathcal{I}_i|\}_{i=1}^k S = I_k\,. \end{aligned}$$

First of all, there is nothing to optimize in the first column of S, since S must be of the form $S = \begin{bmatrix} \frac{\mathbb{1}_k}{\sqrt{n}} & \tilde{S} \end{bmatrix}$ for some $\tilde{S} \in \mathbb{R}^{k\times(k-1)}$. Since $V_k = \begin{bmatrix} \frac{\mathbb{1}_n}{\sqrt{n}} & \tilde{V}_k \end{bmatrix}$ with

$\tilde{V}_k = \begin{bmatrix} v_2(L) & \cdots & v_k(L) \end{bmatrix}$, solving (3.34) is equivalent to solving

$$\begin{aligned} \min_{\tilde{S} \in \mathbb{R}^{k \times (k-1)}} \quad & \|\tilde{V}_k - P_{\{\mathcal{I}_i\}_{i=1}^k} \tilde{S}\|_F^2 \\ s.t. \quad & \tilde{S}^\top \mathbb{1}_k = 0 \\ & \tilde{S}^\top \text{diag}\{|\mathcal{I}_i|\}_{i=1}^k \tilde{S} = I_k \,, \end{aligned} \tag{3.42}$$

where the first constraint in (3.34) is removed by excluding the first column of S, and the second constraint in (3.34) is rewritten as the two constraints in (3.42).

Let $\tilde{O} := \text{diag}\{\sqrt{|\mathcal{I}_i|}\}_{i=1}^k \tilde{S} \in \mathbb{R}^{k \times (k-1)}$ and $\tilde{P}_{\{\mathcal{I}_i\}_{i=1}^k} = P_{\{\mathcal{I}_i\}_{i=1}^k} \text{diag}\{(\sqrt{|\mathcal{I}_i|})^{-1}\}_{i=1}^k$, it is easy to see that (3.42) is equivalent to:

$$\begin{aligned} \min_{\tilde{O} \in \mathbb{R}^{k \times (k-1)}} \quad & \|\tilde{V}_k - \tilde{P}_{\{\mathcal{I}_i\}_{i=1}^k} \tilde{O}\|_F^2 \\ s.t. \quad & \tilde{O}^\top u_{\{\mathcal{I}_i\}_{i=1}^k} = 0 \\ & \tilde{O}^\top \tilde{O} = I_k \,, \end{aligned} \tag{3.43}$$

where $u_{\{\mathcal{I}_i\}_{i=1}^k} = \text{diag}\{(\sqrt{|\mathcal{I}_i|})^{-1}\}_{i=1}^k \mathbb{1}_k$. Now let $Q \in \mathbb{R}^{k \times (k-1)}$ be some matrix such that $Q^\top Q = I$ and $QQ^\top = I - u_{\{\mathcal{I}_i\}_{i=1}^k} u_{\{\mathcal{I}_i\}_{i=1}^k}^\top$, then $\{QO : \; O \in \mathbb{R}^{(k-1) \times (k-1)}, O^\top O = OO^\top = I_{k-1}$ are all the feasible solution to (3.43). Therefore (3.43) is equivalent to

$$\begin{aligned} \min_{O \in \mathbb{R}^{(k-1) \times (k-1)}} \quad & \|\tilde{V}_k - \tilde{P}_{\{\mathcal{I}_i\}_{i=1}^k} QO\|_F^2 \\ s.t. \quad & O^\top O = I_{k-1} \,. \end{aligned} \tag{3.44}$$

Given the SVD: $Q^\top \tilde{P}_{\{\mathcal{I}_i\}_{i=1}^k}^\top \tilde{V}_k = U \Sigma V^\top$, the optimal solution to (3.44) is $O^* = UV^\top$. Then the optimal solution to the original problem (3.34) is

$$S^* = \begin{bmatrix} \frac{\mathbb{1}_k}{\sqrt{n}} & \text{diag}\{(\sqrt{|\mathcal{I}_i|})^{-1}\}_{i=1}^k QO^* \end{bmatrix} \,. \tag{3.45}$$

Proof of Theorem 3.10

Before showing the proof of Theorem 3.10, we state a few lemmas that are used.

Auxiliary lemmas

Firstly, we need the following lemma concerning the stability of the original network $T(s)$ and our approximation model $T_k(s), \hat{T}_k(s)$.

Lemma 3.3. *Suppose all $g_i(s), f(s)$ satisfies Assumption 3.1, then for any V_k with $V_k^\top V_k = I$ and any $\Lambda_k \succeq 0$, the corresponding*

$$T(s) = V_k(V_k^\top \text{diag}\{g_i^{-1}(s)\}V_k + f(s)\Lambda_k)V_k^\top$$

has

$$\|T(s)\|_{\mathcal{H}_\infty} \leq \gamma\,.$$

This Lemma shows the stability of $T(s), T_k(s), \hat{T}_k(s)$ by choosing different V_k, Λ_k.

The following lemma concerns controlling the approximation error between $T_k(s)$ and $\hat{T}_k(s)$.

Lemma 3.4. *Suppose all $g_i(s), f(s)$ satisfies Assumption 3.1. Given two matrices $V_k, \hat{V}_k \in \mathbb{R}^{n\times k}$ with $V_k^\top V_k = \hat{V}_k^\top \hat{V}_k = I$ and some $\Lambda_k \succeq 0$. Define*

$$\begin{aligned} T_k(s) = &\ V_k(V_k^\top \text{diag}\{g_i^{-1}(s)\}V_k + f(s)\Lambda_k)^{-1}V_k^\top\,, \\ \hat{T}_k(s) = &\ \hat{V}_k(\hat{V}_k^\top \text{diag}\{g_i^{-1}(s)\}\hat{V}_k + f(s)\Lambda_k)^{-1}\hat{V}_k^\top\,. \end{aligned}$$

Given any $\eta > 0$, we have

$$\sup_{s\in(-j\eta,+j\eta)} \|T_k(s) - \hat{T}_k(s)\| \leq 2(\gamma + \gamma^2 M(\eta))\|V_k - \hat{V}_k\|_F\,,$$

where $M(\eta) = \sup_{s\in(-j\eta,+j\eta)} \max_i |g_i^{-1}(s)|$.

That is, since $\hat{T}_k(s)$ is obtained by replace V_k in $T_k(s)$ by $\hat{T}_k(s)$, the error can be controlled by the difference $\|V_k - \hat{V}_k\|_F$ between V_k and $\hat{V}_k$. Recall that Theorem 3.8 provides a bound on $\|T(s) - T_k(s)\|$, combing it with Lemma 3.4 allows us to control the error $\|T(s) - \hat{T}_k(s)\|$, as stated in the following lemma:

Lemma 3.5. *Consider the network model $(G(s), L, f(s))$ with L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, P, W)$. If Assumption 3.1 holds, Then given any $\eta > 0$, we have $\forall \epsilon > 0$,*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - \hat{T}_k(s)\| \geq \epsilon\right) \\
&\leq 2\mathbb{P}\left(\lambda_{k+1}(L) \leq \frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2 + M(\eta) + \gamma M^2(\eta)\right)\right) \\
&\qquad\qquad + \mathbb{P}\left(\|V_k - \hat{V}_k\| \geq \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))}\right),
\end{aligned}$$

where $\sup_{s\in(-j\eta,+j\eta)} \max_i |g_i^{-1}(s)| := M(\eta)$ *and* $F_l(\eta) := \inf_{s\in(-j\eta,+j\eta)} |f(s)|$.

That is, we need to lower bound $\lambda_{k+1}(L)$ and upper bound $\|V_k - \hat{V}_k\|$ for controlling the error. All of these are possible by studying the Laplacian matrix constructed from the expected adjacency matrix:

For a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$, we denote the expected value of adjacency matrix A as

$$A_{\text{blk}} = P_{\{\mathcal{I}_i\}_{i=1}^k} B P_{\{\mathcal{I}_i\}_{i=1}^k}^\top,\ B = Q \odot W\,, \tag{3.46}$$

and define

$$L_{\text{blk}} = \operatorname{diag}\{A_{\text{blk}}\mathbb{1}_n\} - A_{\text{blk}}\,, \tag{3.47}$$

and

$$V_k^{\text{blk}} = \begin{bmatrix}\frac{\mathbb{1}}{\sqrt{n}} & v_2(L_{\text{blk}}) & \cdots & v_k(L_{\text{blk}})\end{bmatrix}. \tag{3.48}$$

Firstly, if the spectral clustering algorithm returns the true block assignment, then it is sufficient to control the difference between V_k and V_k^{blk} for upper bounding $\|V_k - \hat{V}_k\|$:

Lemma 3.6. *Let $V_k^{blk} = \begin{bmatrix}\frac{\mathbb{1}}{\sqrt{n}} & v_2(L_{blk}) & \cdots & v_k(L_{blk})\end{bmatrix}$. Consider an L sampled from a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$, if the spectral clustering algorithm in Algorithm 1 returns the true block assignments $\{\mathcal{I}_i\}_{i=1}^k$, then optimizing (3.34) yields a $\hat{V}$ such that*

$$\|\hat{V}_k - V_k\| \leq \|\sin\Theta(V_k, V_k^{blk})\|_F\,. \tag{3.49}$$

The term $\|\sin\Theta(V_k, V_k^{\text{blk}})\|_F$ should be small given that L and L_{blk} are sufficiently close to each other with high probability(to be formalized later). Moreover, $\lambda_{k+1}(L)$ and $\lambda_{k+1}(L_{\text{blk}})$ should be close for the same reason. We discuss the spectrum of L_{blk} in detail in the next Appendix. The following Lemma is the direct consequence of Proposition 3.1 in the next Appendix:

Lemma 3.7. *Consider a weighted stochastic block model $(\{\mathcal{I}_i\}_{i=1}^k, Q, W)$ satisfying Assumption 3.1. Let $n_{\min} = \min_{1\le i\le k}|\mathcal{I}_i|$, and $b_{\min} := \min\{[B_k \mathbb{1}_k]_i : \ i = 1, \cdots, k\}$. We have*

1. *L_{blk} is k-block-ideal;*

2. *$\lambda_{k+1}(L_{blk}) \geq b_{\min} n_{\min}$;*

3. *$\lambda_{k+1}(L_{blk}) - \lambda_k(L_{blk}) \geq \Delta n_{\min}$.*

Now we are ready to proof our main theorem.

Proof of Theorem 3.10

Proof of Theorem 3.10. Define $\tilde{B} := P \odot W \odot W$, and $\tilde{b}_{\max} = \max_i \sum_j B_{ij}, \tilde{b}_{\min} = \min_j \sum_j B_{ij}$. We also define $W_{\max} =: \max_{ij}|W_{ij}|$. A direct application of Proposition 3 in [103] shows that for any $c > 0$, if

$$kn_{\min}\tilde{b}_{\max} \geq 16(c+1)\log n\,, \tag{3.50}$$

then for any $4n^{-c} \leq \frac{\delta}{6} < 1$, we have

$$\mathbb{P}\left(\|L - L_{\text{blk}}\| \geq 8\sqrt{kn_{\max}\tilde{b}_{\max}\log(24n/\delta)}\right) \leq \frac{\delta}{6} \tag{3.51}$$

If

$$b_{\min}n_{\min} - 8\sqrt{kn_{\max}\tilde{b}_{\max}\log(24n/\delta)} \geq \frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2 + M(\eta) + \gamma M^2(\eta)\right)\,, \tag{3.52}$$

then

$$\begin{aligned} \lambda_{k+1}(L) &\le \frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2 + M(\eta) + \gamma M^2(\eta)\right) \\ \Rightarrow \lambda_{k+1}(L) &\le b_{\min} n_{\min} - 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)} \\ (\text{Lemma 3.7}) \Rightarrow \lambda_{k+1}(L) &\le \lambda_{k+1}(L_{\text{blk}}) - 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)} \\ \Rightarrow 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)} &\le \lambda_{k+1}(L_{\text{blk}}) - \lambda_{k+1}(L) \\ (\text{Weyl's inequality [64]}) \Rightarrow 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)} &\le \|L - L_{\text{blk}}\| \,. \end{aligned}$$

That is, for a given $\delta \ge 4n^{-c}$, if (3.50)(3.52) hold, then

$$\begin{aligned} &\mathbb{P}\left(\lambda_{k+1}(L) \le \frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2 + M(\eta) + \gamma M^2(\eta)\right)\right) \\ &\qquad \le \mathbb{P}\left(\|L - L_{\text{blk}}\| \ge 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)}\right) \le \frac{\delta}{6}\,. \end{aligned}$$

Similarly, when

$$\frac{\epsilon}{8\sqrt{k}(\gamma + \gamma^2 M(\eta))}\Delta n_{\min} \ge 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)}\,, \tag{3.53}$$

and the spectral clustering (SC) returns the true $\{\mathcal{I}_i\}_{i=1}^k$, then

$$\begin{aligned} \|V_k - \hat{V}_k\| &\ge \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))} \\ (\text{Lemma 3.6}) \Rightarrow \|\sin\Theta(V_k, V_k^{\text{blk}})\|_F &\ge \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))} \\ (\text{Davis-Khan [104]}) \Rightarrow \frac{2\sqrt{k}\|L - L_{\text{blk}}\|}{\lambda_{k+1}(L_{\text{blk}}) - \lambda_k(L_{\text{blk}})} &\ge \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))} \\ (\text{Lemma 3.7}) \Rightarrow \|L - L_{\text{blk}}\| &\ge \frac{\epsilon}{8\sqrt{k}(\gamma + \gamma^2 M(\eta))}\Delta n_{\min} \\ \Rightarrow \|L - L_{\text{blk}}\| &\ge 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)}\,. \end{aligned}$$

That is, for a given $\delta \ge 4n^{-c}$, if (3.50)(3.53) hold, then

$$\begin{aligned} &\mathbb{P}\left(\|V_k - \hat{V}_k\| \ge \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))}, \text{“SC returns true } \{\mathcal{I}_i\}_{i=1}^k\text{”}\right) \\ &\qquad \le \mathbb{P}\left(\|L - L_{\text{blk}}\| \ge 8\sqrt{k n_{\max} \tilde{b}_{\max} \log(24n/\delta)}\right) \le \frac{\delta}{6}\,. \end{aligned}$$

Now by Lemma 3.5, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)}\|T(s)-\hat{T}_k(s)\|\geq\epsilon\right)\\
&\leq 2\mathbb{P}\left(\lambda_{k+1}(L)\leq\frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2+M(\eta)+\gamma M^2(\eta)\right)\right)\\
&\qquad\qquad\qquad\qquad+\mathbb{P}\left(\|V_k-\hat{V}_k\|\geq\frac{\epsilon}{4(\gamma+\gamma^2M(\eta))}\right)\\
&\leq 2\mathbb{P}\left(\lambda_{k+1}(L)\leq\frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2+M(\eta)+\gamma M^2(\eta)\right)\right)\\
&\qquad\quad+\mathbb{P}\left(\|V_k-\hat{V}_k\|\geq\frac{\epsilon}{4(\gamma+\gamma^2M(\eta))},\text{“SC returns true }\{\mathcal{I}_i\}_{i=1}^k\text{”}\right)\\
&\qquad\quad+\mathbb{P}\left(\|V_k-\hat{V}_k\|\geq\frac{\epsilon}{4(\gamma+\gamma^2M(\eta))},\text{“SC does not return true }\{\mathcal{I}_i\}_{i=1}^k\text{”}\right)\\
&\leq 2\mathbb{P}\left(\lambda_{k+1}(L)\leq\frac{1}{F_l(\eta)}\left(\frac{2}{\epsilon}(\gamma M(\eta)+1)^2+M(\eta)+\gamma M^2(\eta)\right)\right)\\
&\qquad\quad+\mathbb{P}\left(\|V_k-\hat{V}_k\|\geq\frac{\epsilon}{4(\gamma+\gamma^2M(\eta))},\text{“SC returns true }\{\mathcal{I}_i\}_{i=1}^k\text{”}\right)\\
&\qquad\quad+\mathbb{P}\left(\text{“SC does not return true }\{\mathcal{I}_i\}_{i=1}^k\text{”}\right)\\
&\leq 2\cdot\frac{\delta}{6}+\frac{\delta}{6}+\frac{\delta}{2}=\delta\,,
\end{aligned}
$$

In the last inequality, we upper bound the first and the second probability by picking a sufficiently large n such that (3.50)(3.52)(3.53) hold, and the last probability is upper bounded by $\frac{\delta}{2}$ if we pick $n\geq\tilde{N}\left(\frac{\delta}{2}\right)$ by our assumption 3.2. □

Proofs of auxiliary lemmas

Proof of Lemma 3.3. For each $g_i(s), i=1,\cdots,n$, we have, by the OSP property,

$$Re(g_i(s))\geq\frac{1}{\gamma}|g_i(s)|^2,\forall Re(s)>0\,.$$

we have, for the diagonal transfer matrix $G(s)=\mathrm{diag}\{g_i(s)\}_{i=1}^n$:

$$2Re(G(s))=G^*(s)+G(s)\succeq\frac{2}{\gamma}G^*(s)G(s)\,,\forall Re(s)>0\,. \tag{3.54}$$

Since $g_i(s)$ are all OSP, then all $g_i(s)$ are positive real [106]. A positive real function that is not a zero function has no zero nor pole on the left half plane. Therefore

$g_i(s)$ are invertible for all $Re(s) > 0$, which ensures that $G(s)$ is invertible for all $Re(s) > 0$. Multiply $(G^{-1})^*$ on the left and G^{-1} on the right of (3.54), we have

$$G^{-1}(s) + (G^{-1}(s))^* \succeq \frac{2}{\gamma}I, \forall Re(s) > 0\,. \tag{3.55}$$

Multiply $V_k^\top$ on the left and V_k on the right of (3.55), we have

$$V_k^\top G^{-1}(s)V_k + (V_k^\top G^{-1}(s)V_k)^* \succeq \frac{2}{\gamma}I, \forall Re(s) > 0\,,$$

using the fact that $f(s)\Lambda_k$ is PR, we have

$$V_k^\top G^{-1}(s)V_k + f(s)\Lambda_k + (V_k^\top G^{-1}(s)V_k + f(s)\Lambda_k)^* \succeq \frac{2}{\gamma}I, \forall Re(s) > 0\,, \tag{3.56}$$

Notice that we have defined $H_k(s) = V_k^\top G^{-1}(s)V_k + f(s)\Lambda_k$, then we conclude that $H_k(s) + (H_k(s))^* \succeq \frac{2}{\gamma}I$, or equivalently,

$$\begin{bmatrix} I \\ H_k(s) \end{bmatrix}^* \begin{bmatrix} -\frac{2}{\gamma}I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ H_k(s) \end{bmatrix} \succeq 0, \forall Re(s) > 0 \tag{3.57}$$

Moreover, we have

$$\begin{bmatrix} -\frac{2}{\gamma}I & I \\ I & 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{\gamma}I & 0 \\ 0 & -\gamma^2\frac{\epsilon}{2}I \end{bmatrix} = \begin{bmatrix} -\frac{1}{\gamma}I & I \\ I & -\gamma I \end{bmatrix} \preceq 0\,,$$

since its Schur complement is a zero matrix.

Therefore,

$$\begin{bmatrix} I \\ H_k(s) \end{bmatrix}^* \begin{bmatrix} -\frac{1}{\gamma}I & I \\ I & -\gamma I \end{bmatrix} \begin{bmatrix} I \\ H_k(s) \end{bmatrix} \succeq \begin{bmatrix} I \\ H_k(s) \end{bmatrix}^* \begin{bmatrix} -\frac{2}{\gamma}I & I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ H_k(s) \end{bmatrix} \succeq 0\,, \forall Re(s) > 0\,,$$

which is exactly,

$$\gamma(H_k(s))^*(H_k(s)) \succeq \frac{1}{\gamma}I\,, \forall Re(s) > 0\,.$$

This shows that

$$\sigma^2_{\min}(H_k(s)) \geq \frac{1}{\gamma^2}, \forall Re(s) > 0\,,$$

which leads to

$$\|T_k(s)\| = \|V_k H_k^{-1}(s) V_k^\top\| = \|H_k^{-1}(s)\| \leq \gamma\,, \forall Re(s) > 0\,.$$

This is exactly $\|T_k(s)\|_{\mathcal{H}_\infty} \leq \gamma$. □

Proof of Lemma 3.4. Denote $H_k(s) := V_k^\top \mathrm{diag}\{g_i^{-1}(s)\} V_k + f(s)\Lambda_k$ and

$$\hat{H}_k(s) := \hat{V}_k^\top \mathrm{diag}\{g_i^{-1}(s)\} \hat{V}_k + f(s)\Lambda_k \,,$$

then

$$\begin{aligned}
&\|T_k(s) - \hat{T}_k(s)\| \\
&= \|V_k H_k(s) V_k^\top - \hat{V}_k \hat{H}_k(s) \hat{V}_k^\top\| \\
&= \|\hat{V}_k H_k^{-1}(s)(V_k^\top - \hat{V}_k^\top) + (V_k - \hat{V}_k) H_k^{-1}(s) V_k^\top + \hat{V}_k (H_k^{-1}(s) - \hat{H}_k^{-1}(s)) \hat{V}_k^\top\| \\
&\leq \|\hat{V}_k H_k^{-1}(s)(V_k^\top - \hat{V}_k^\top)\| + \|(V_k - \hat{V}_k) H_k^{-1}(s) V_k^\top\| + \|\hat{V}_k (H_k^{-1}(s) - \hat{H}_k^{-1}(s)) \hat{V}_k^\top\|
\end{aligned}$$

Notice that for any $s \in (-j\eta, +j\eta)$,

$$\|\hat{V}_k H_k^{-1}(s)(V_k^\top - \hat{V}_k^\top)\| \leq \|H_k^{-1}(s)\| \|V_k - \hat{V}_k\| \leq \|H_k^{-1}(s)\|_{\mathcal{H}_\infty} \|V_k - \hat{V}_k\| \leq \gamma \|V_k - \hat{V}_k\| \,, \tag{3.58}$$

where the last inequality uses the intermediate result $\|H_k^{-1}(s)\|$ in the proof for Lemma 3.3. Similarly,

$$\|(V_k - \hat{V}_k) H_k^{-1}(s) V_k^\top\| \leq \gamma \|V_k - \hat{V}_k\| \,. \tag{3.59}$$

For the last term, we have for any $s \in (-j\eta, +j\eta)$,

$$\begin{aligned}
&\|\hat{V}_k (H_k^{-1}(s) - \hat{H}_k^{-1}(s)) \hat{V}_k^\top\| \\
&\leq \|H_k^{-1}(s) - \hat{H}_k^{-1}(s)\| \\
&= \|\hat{H}_k^{-1}(s)(\hat{H}_k(s) - H_k(s)) H_k^{-1}(s)\| \\
&\leq \|\hat{H}_k^{-1}(s)\| \|H_k^{-1}(s)\| \|\hat{H}_k(s) - H_k(s)\| \\
&\leq \gamma^2 \|V_k^\top \mathrm{diag}\{g_i^{-1}(s)\} V_k - \hat{V}_k^\top \mathrm{diag}\{g_i^{-1}(s)\} \hat{V}_k\| \\
&\leq \gamma^2 \|(V_k^\top - \hat{V}_k^\top) \mathrm{diag}\{g_i^{-1}(s)\} V_k + \hat{V}_k^\top \mathrm{diag}\{g_i^{-1}(s)\}(V_k - \hat{V}_k)\| \\
&\leq 2\gamma^2 \|\mathrm{diag}\{g_i^{-1}(s)\}\| \|V_k - \hat{V}_k\| \leq 2\gamma^2 M(\eta) \|V_k - \hat{V}_k\| \,.
\end{aligned} \tag{3.60}$$

Using the bounds in (3.58)(3.59)(3.60), we finally have

$$
\begin{aligned}
&\sup_{s\in(-j\eta,+j\eta)} \|T_k(s) - \hat{T}_k(s)\| \\
\leq\ &\sup_{s\in(-j\eta,+j\eta)} \Big(\|\hat{V}_k H_k^{-1}(s)(V_k^\top - \hat{V}_k^\top)\| + \|(V_k - \hat{V}_k) H_k^{-1}(s) V_k^\top\| \\
&\qquad\qquad\qquad + \|\hat{V}_k (H_k^{-1}(s) - \hat{H}_k^{-1}(s)) \hat{V}_k^\top\| \Big) \\
\leq\ & 2(\gamma + \gamma^2 M(\eta)) \|V_k - \hat{V}_k\| \leq 2(\gamma + \gamma^2 M(\eta)) \|V_k - \hat{V}_k\|_F\,.
\end{aligned}
$$

□

Proof of Lemma 3.5. We have defined $T_k(s) = V_k(V_k^\top \mathrm{diag}\{g_i^{-1}(s)\} V_k + f(s)\Lambda_k)^{-1} V_k^\top$, then

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - \hat{T}_k(s)\| \geq \epsilon \right) \\
\leq\ &\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| + \sup_{s\in(-j\eta,+j\eta)} \|T_k(s) - \hat{T}_k(s)\| \geq \epsilon \right) \\
\leq\ &\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \geq \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T_k(s) - \hat{T}_k(s)\| \geq \frac{\epsilon}{2} \right)\,.
\end{aligned}
\tag{3.61}
$$

For the first term, we have

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \geq \frac{\epsilon}{2} \right) \\
=\ &\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \geq \frac{\epsilon}{2},\ \lambda_{k+1}(L) \leq \frac{M(\eta) + \gamma M^2(\eta)}{F_l(\eta)} \right) \\
&\qquad + \mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \geq \frac{\epsilon}{2},\ \lambda_{k+1}(L) > \frac{M(\eta) + \gamma M^2(\eta)}{F_l(\eta)} \right) \\
\leq\ &\mathbb{P}\left(\lambda_{k+1}(L) \leq \frac{M(\eta) + \gamma M^2(\eta)}{F_l(\eta)} \right) \\
&\qquad + \mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \geq \frac{\epsilon}{2},\ \lambda_{k+1}(L) > \frac{M(\eta) + \gamma M^2(\eta)}{F_l(\eta)} \right) \\
\overset{(a)}{\leq}\ &\mathbb{P}\left(\lambda_{k+1}(L) \leq \frac{M(\eta) + \gamma M^2(\eta)}{F_l(\eta)} \right) + \mathbb{P}\left(\frac{(\gamma M(\eta) + 1)^2}{F_l(\eta)\lambda_{k+1}(L) - M(\eta) - \gamma M(\eta)} \geq \frac{\epsilon}{2} \right) \\
\overset{(b)}{\leq}\ & 2\mathbb{P}\left(\lambda_{k+1}(L) \leq \frac{1}{F_l(\eta)} \left(\frac{2}{\epsilon}(\gamma M(\eta) + 1)^2 + M(\eta) + \gamma M^2(\eta) \right) \right)\,,
\end{aligned}
\tag{3.62}
$$

where (a) is from the fact that when $\lambda_{k+1}(L) > \frac{M(\eta)+\gamma M^2(\eta)}{F_l(\eta)}$, we can apply Theorem 3.8 for any $s_0 \in (-j\eta, +j\eta)$, with a uniform bound

$$\|T(s_0) - T_k(s_0)\| \leq \frac{(\gamma M(\eta) + 1)^2}{F_l(\eta)\lambda_{k+1}(L) - M(\eta) - \gamma M(\eta)}\,,$$

then applying supremum gives us

$$\sup_{s\in(-j\eta,+j\eta)} \|T(s) - T_k(s)\| \leq \frac{(\gamma M(\eta) + 1)^2}{F_l(\eta)\lambda_{k+1}(L) - M(\eta) - \gamma M(\eta)}\,,$$

which implies the event $\frac{(\gamma M(\eta)+1)^2}{F_l(\eta)\lambda_{k+1}(L)-M(\eta)-\gamma M(\eta)} \geq \frac{\epsilon}{2}$. And the (b) is due to the fact that the second probability is always larger than the first one.

For the second term, we have, by Lemma 3.4

$$\begin{aligned}\mathbb{P}\left(\sup_{s\in(-j\eta,+j\eta)} \|T_k(s) - \hat{T}_k(s)\| \geq \frac{\epsilon}{2}\right) \leq\ & \mathbb{P}\left(2(\gamma + \gamma^2 M(\eta))\|V_k - \hat{V}_k\| \geq \frac{\epsilon}{2}\right) \\ =\ & \mathbb{P}\left(\|V_k - \hat{V}_k\| \geq \frac{\epsilon}{4(\gamma + \gamma^2 M(\eta))}\right)\,. \end{aligned} \tag{3.63}$$

Apply (3.62)(3.63) to (3.61) gives the desired bound. □

Proof of Lemma 3.6. Consider $V_k = \begin{bmatrix} \frac{\mathbb{1}}{\sqrt{n}} & v_2(L) & \cdots & v_k(L) \end{bmatrix}$ from the random Laplacian matrix, and $V_k^{\text{blk}} = \begin{bmatrix} \frac{\mathbb{1}}{\sqrt{n}} & v_2(L_{\text{blk}}) & \cdots & v_k(L_{\text{blk}}) \end{bmatrix}$ from L_{blk}. The singular values of $V_k^\top V_k^{\text{blk}}$ are exactly the cosine of the principal angles between the two subspace spanned by V_k and V_k^{blk}. Notice that

$$V_k^\top V_k^{\text{blk}} = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{V}_k^\top \tilde{V}_k^{\text{blk}} \end{bmatrix}\,, \tag{3.64}$$

where $\tilde{V}_k = \begin{bmatrix} v_2(L) & \cdots & v_k(L) \end{bmatrix}$ and $\tilde{V}_k^{\text{blk}} = \begin{bmatrix} v_2(L_{\text{blk}}) & \cdots & v_k(L_{\text{blk}}) \end{bmatrix}$. then there exists $O_1, O_2 \in \mathcal{O}^{(k-1)\times(k-1)}$ such that

$$\begin{bmatrix} 1 & 0 \\ 0 & O_1^\top \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \tilde{V}_k^\top \tilde{V}_k^{\text{blk}} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & O_2 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \cos\theta_2 & & \\ & & \ddots & \\ & & & \cos\theta_k \end{bmatrix}\,, \tag{3.65}$$

where the first diagonal term corresponds to the cosine of the first principal angle $\theta_1 = 0$. Now consider the orthogonal matrix $O = \begin{bmatrix} 1 & 0 \\ 0 & O_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & O_1^\top \end{bmatrix} \in \mathbb{R}^{k \times k}$, we have

$$\begin{aligned}
\|V_k^{\text{blk}} O - V_k\|_F^2 = &\ \operatorname{tr}\left((V_k^{\text{blk}} O - V_k)^\top (V_k^{\text{blk}} O - V_k)\right) \\
= &\ 2d - 2\operatorname{tr}(V_k^\top V_k^{\text{blk}} O) \\
= &\ 2d - 2\operatorname{tr}\left(\begin{bmatrix} 1 & 0 \\ 0 & O_1^\top \end{bmatrix} V_k^\top V_k^{\text{blk}} \begin{bmatrix} 1 & 0 \\ 0 & O_2 \end{bmatrix}\right) \\
= &\ 2d - \sum_{i=1}^k \cos\theta_i \\
\leq &\ 2d - \sum_{i=1}^k \cos^2\theta_i = \|\sin\Theta(V_k, V_k^{\text{blk}})\|_F^2 \,.
\end{aligned}$$

Now consider the $\hat{V}_k = P_{\{\mathcal{I}_i\}_{i=1}^k} S^*$, where $\{\mathcal{I}_i\}_{i=1}^k$ is the clustering result from the spectral clustering algorithm, and S^* from solving the optimization problem in (3.34). Notice that if the clustering result $\{\mathcal{I}_i\}_{i=1}^k$ is correct, i.e., L_{blk} is indeed k-block-ideal w.r.t. $P_{\{\mathcal{I}_i\}_{i=1}^k}$, then there exists some invertible matrix $\tilde{S} \in \mathbb{R}^{k \times k}$ such that

$$V_k^{\text{blk}} = P_{\{\mathcal{I}_i\}_{i=1}^k} \tilde{S} \,, \tag{3.66}$$

which implicitly requires $\tilde{S} e_1 = \mathbb{1}_k$ and $\tilde{S}^\top \operatorname{diag}\{n_i\}_{i=1}^k \tilde{S} = I_k$. It is easy to show that $S = \tilde{S} O$ is a feasible solution to (3.34):

$$S e_1 = \tilde{S} O e_1 = \tilde{S} e_1 = \mathbb{1}_k,\ S^\top \operatorname{diag}\{n_i\}_{i=1}^k S = O^\top \tilde{S}^\top \operatorname{diag}\{n_i\}_{i=1}^k \tilde{S} O = I_k \,. \tag{3.67}$$

Therefore

$$\|\hat{V}_k - V_k\|_F \leq \|P_{\{\mathcal{I}_i\}_{i=1}^k} S - V_k\|_F = \|V_k^{\text{blk}} O - V_k\|_F \leq \|\sin\Theta(V_k, V_k^{\text{blk}})\|_F \,. \tag{3.68}$$

□

Eigenvalues and eigenvectors of L_{blk}

Assume the network has the following adjacency matrix:

$$A_{\text{blk}} = \underbrace{\begin{bmatrix} \mathbb{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbb{1}_{n_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{1}_{n_k} \end{bmatrix}}_{:=P} B_k P^\top \,, \tag{3.69}$$

where $A_k \in \mathbb{R}^{k\times k}$ and

$$[B_k]_{ij} = \begin{cases} \alpha_{ii}, & i = j \\ \beta_{ij}, & i \le j \\ \beta_{ij}, & i > j \end{cases} . \tag{3.70}$$

The Laplacian

$$L_{\text{blk}} = \operatorname{diag}\{A_{\text{blk}}\mathbb{1}_n\} - A_{\text{blk}} \,. \tag{3.71}$$

Proposition 3.1. *Let* $n_{\min} = \min\{n_i \;:\; i = 1, \cdots, k\}$*, and* $n_{\max} = \max\{n_i \;:\; i = 1, \cdots, k\}$ *Suppose*

$$\min_i\{\alpha_{ii}\} - \frac{2n_{\max}}{n_{\min}} \max_i \sum_{j\neq i} \beta_{ij} := \Delta > 0 \,. \tag{3.72}$$

Define

$$\tilde{L}_k = \operatorname{diag}\{\tilde{B}_k\mathbb{1}_k\} - \tilde{A}_k, \quad \tilde{B}_k = B_k \cdot \operatorname{diag}\{n_i\}_{i=1}^k \,, \tag{3.73}$$

and let $v_i(\tilde{L}_k)$ *be the right eigenvector of* $\tilde{L}_k$ *associated with* $\lambda_i(\tilde{L}_k)$*. Then for* $i = 1, \cdots, k$*, we have*

1. *(L_{blk} is k-block-ideal)*

$$\lambda_i(L_{blk}) = \lambda_i(\tilde{L}_k), v_i(L_{blk}) = \begin{bmatrix} \mathbb{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbb{1}_{n_2} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{1}_{n_k} \end{bmatrix} v_i(\tilde{L}_k) \,. \tag{3.74}$$

Moreover, let $b_{\min} := \min\{[B_k\mathbb{1}_k]_i \,:\; i = 1, \cdots, k\}$ *be the minimum row sum of* B_k*, then we have*

2. *($\lambda_{k+1}(L_{blk})$ is large)*

$$\lambda_{k+1}(L_{blk}) \geq b_{\min} n_{\min}\,, \tag{3.75}$$

3. *(there is a sufficient spectral gap $\lambda_{k+1}(L_{blk}) - \lambda_k(L_{blk})$)*

$$\lambda_{k+1}(L_{blk}) - \lambda_k(L_{blk}) \geq \Delta n_{\min}\,. \tag{3.76}$$

Proof. The proof takes few steps: First we show that $\{(\lambda_i(\tilde{L}_k), v_i(\tilde{L}_k))\}_{i=1}^k$ are eigenpairs of L_{blk}, then we show that $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ are indeed the first k smallest eigenvalues of L_{blk}. Lastly we provide the lower bound on both $\lambda_{k+1}(L_{\text{blk}})$ and $\lambda_{k+1}(L_{\text{blk}}) - \lambda_k(L_{\text{blk}})\,.$

Show eigenpairs$\{(\lambda_i(\tilde{L}_k), v_i(\tilde{L}_k))\}_{i=1}^k$: Notice that

$$\begin{aligned}
L_{\text{blk}}P =& \;(\text{diag}\{A_{\text{blk}}\mathbb{1}_n\} - A_{\text{blk}})P \\
=& \;(\text{diag}\{PB_kP^\top\mathbb{1}_n\} - PB_kP^\top)P \\
=& \;\text{diag}\{PB_k\text{diag}\{n_i\}_{i=1}^k\mathbb{1}_k\} - PB_k\text{diag}\{n_i\}_{i=1}^k \\
=& \;P\text{diag}\{B_k\text{diag}\{n_i\}_{i=1}^k\mathbb{1}_k\} - PB_k\text{diag}\{n_i\}_{i=1}^k = P\tilde{L}_{\text{blk}}\,,
\end{aligned} \tag{3.77}$$

where we used an equality $P\text{diag}\{x\} = \text{diag}\{Px\}$ for any $x \in \mathbb{R}^k$ due to the special structure of P. We can obtain k eigenpairs through (3.77): Given any eigenpair $(\lambda_i(\tilde{L}_{\text{blk}}), v_i(\tilde{L}_k))$, we have

$$L_{\text{blk}}Pv_i(\tilde{L}_k) = P\tilde{L}_kv_i(\tilde{L}_k) = \lambda_i(\tilde{L}_{\text{blk}})Pv_i(\tilde{L}_k)\,, \tag{3.78}$$

which suggests $(\lambda_i(\tilde{L}_k), Pv_i(\tilde{L}_k))$ is an eigenpair of L_{blk}. This holds for every $i = 1, \cdots, k$.

Show that $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ are the first k smallest eigenvalues: The remaining eigenvalues of L_{blk} are easy to find:

- $\left(\lambda = n_1\alpha_{11} + \sum_{j\neq 1}\beta_{1j}n_j, v = \begin{bmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}\right)$ is an eigenpair for any $v_1 \in \mathbb{S}^{n_1-1}$ such that $\mathbb{1}_{n_1}^\top v_1 = 0$.

- $\left(\lambda = n_2\alpha_{22} + \sum_{j\neq 2}\beta_{2j}n_j, v = \begin{bmatrix} 0 \\ v_2 \\ \vdots \\ 0 \end{bmatrix}\right)$ is an eigenpair for any $v_2 \in \mathbb{S}^{n_2-1}$ such that $\mathbb{1}_{n_2}^\top v_2 = 0$.
- $\cdots$
- $\left(\lambda = n_k\alpha_{kk} + \sum_{j\neq k}\beta_{kj}n_j, v = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ v_k \end{bmatrix}\right)$ is an eigenpair for any $v_k \in \mathbb{S}^{n_k-1}$ such that $\mathbb{1}_{n_k}^\top v_k = 0$.

Any choice of such (λ, v) is an eigenpair because, for example, the eigenpair associated with some v_1 satisfies

$$
\begin{aligned}
& L_{\text{blk}} \begin{bmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \left(\begin{bmatrix} (n_1\alpha_{11} + \sum_{j\neq 1}\beta_{1j}n_j)I_{n_1} & & \\ & \ddots & \\ & & (n_k\alpha_{kk} + \sum_{j\neq k}\beta_{kj}n_j)I_{n_k} \end{bmatrix} - PB_kP^\top \right) \begin{bmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} (n_1\alpha_{11} + \sum_{j\neq 1}\beta_{1j}n_j)I_{n_1} & & \\ & \ddots & \\ & & (n_k\alpha_{kk} + \sum_{j\neq k}\beta_{kj}n_j)I_{n_k} \end{bmatrix} \begin{bmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \left(n_1\alpha_{11} + \sum_{j\neq 1}\beta_{1j}n_j \right) v_1 \,.
\end{aligned}
$$

A similar argument can be made for other pairs. This gives us all the rest of the eigenvalues: each eigenvalue $n_i\alpha_{ii} + \sum_{j\neq i}\beta_{ij}n_j$ has multiplicity $n_i - 1$. Together with the k eigenvalues $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ we have already found in previous derivation, we have all the eigenvalues of L_{blk}.

The claim that $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ are the first k smallest eigenvalues is shown by our assumption:

$$
\begin{aligned}
\min_i \left(n_i\alpha_{ii} + \sum_{j\neq i} \beta_{ij} n_j \right) &\geq \min_i n_i \alpha_{ii} \\
&\geq n_{\min} \min_i \alpha_{ii} \\
&\geq 2n_{\max} \max_i \sum_{j\neq i} \beta_{ij} + n_{\min}\Delta \\
&\geq \max_i (n_i + n_{\max}) \sum_{j\neq i} \beta_{ij} + n_{\min}\Delta \\
&\geq \max_i \left(n_i \sum_{j\neq i} \beta_{ij} + \sum_{j\neq i} \beta_{ij} n_j \right) + n_{\min}\Delta \\
&\geq \max_i \lambda_i(\tilde{L}_k) + n_{\min}\Delta \,, \qquad (3.79)
\end{aligned}
$$

where the last inequality is from the Gershgorin disk theorem [64] by noticing that for i-th column of $\tilde{L}_k$, the diagonal term is $\sum_{j\neq i} \beta_{ij} n_j$ and the sum of the absolute value of the off-diagonal terms is $n_i \sum_{j\neq i} \beta_{ij}$. (3.79) is more than enough to show $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ are the first k smallest eigenvalues of L_{blk}.

Bound on the eigenvalue and spectral gap

Knowing $\{\lambda_i(\tilde{L}_k)\}_{i=1}^k$ are the first k smallest eigenvalues, we have

$$
\lambda_{k+1}(L_{\text{blk}}) = \min_i \left(n_i\alpha_{ii} + \sum_{j\neq i} \beta_{ij} n_j \right) \geq \min_i \left(\alpha_{ii} + \sum_{j=i} \beta_{ij} \right) n_{\min} = b_{\min} n_{\min} \,,
$$

and (3.79) already shows

$$
\lambda_{k+1}(L_{\text{blk}}) = \min_i \left(n_i\alpha_{ii} + \sum_{j\neq i} \beta_{ij} n_j \right) \geq \max_i \lambda_i(\tilde{L}_k) + n_{\min}\Delta = \lambda_k(L_{\text{blk}}) + n_{\min}\Delta \,.
$$

□

3.3 Reducing Model Complexity

In this section, we apply our analysis to investigate coherence in power networks. For coherent generator groups, we find that $\frac{1}{n}\bar{g}(s) = \hat{g}(s)$ generalizes typical aggregate generator models which are often used for model reduction in power networks [96]. Moreover, we show that heterogeneity in generator dynamics usually leads to high-order aggregate dynamics. Therefore, although our previous analyses provide a structurally interpretable reduced model for networks, the resulting models are still potentially of high order due to aggregation. This asks for model reduction on the aggregate dynamics, in order to reduce the model complexity of our approximation model in Section 3.2. We will mostly discuss model reduction techniques for power networks, but the analyses could potentially be generalized to other networks.

We will resort to *frequency weighted balanced truncation* to develop a hierarchy of models of adjustable order and increasing accuracy. In particular, for aggregation of n second order generator models in power networks, we find that high accuracy can often be achieved by reducing the $(n+1)$-th order system to a 3rd order one. We further compare two alternatives: providing an aggregate model for a set of turbines, and subsequently closing the loop, versus performing the reduction directly on the closed loop $\hat{g}(s)$. The first is motivated by retaining the interpretation whereby the aggregate is represented by one or two equivalent turbines; nonetheless, we show how a similar interpretation may be available for the second, more accurate method.

The rest of the section is organized as follows. In Section 3.3.1, we provide the theoretical justification of the coherent dynamics $\hat{g}(s)$. In Section 3.3.2, we propose reduced-order models for $\hat{g}(s)$ by frequency weighted balanced truncation. We then show via numerical illustrations that the proposed models can achieve accurate approximation (Section 3.3.3).

3.3.1 Aggregate dynamics of coherent generators

Consider a group of n generators, indexed by $i = 1, \cdots, n$ and dynamically coupled through an AC network. Assuming the network is in steady-state, Figure 3-6 shows the block diagram of the linearized system around its operating point.

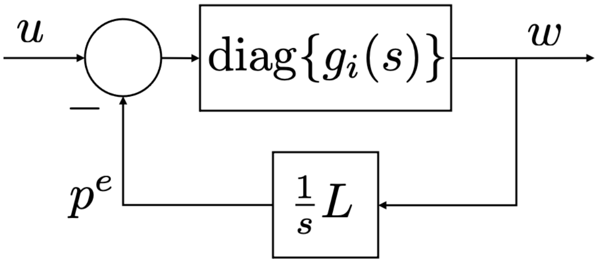

Figure 3-6. Block Diagram of a Linearized Power Network.

We refer to [34] for details on the linearization procedure. The signals $w = [w_1, \cdots, w_n]^\top, u = [u_1, \cdots, u_n]^\top, p^e = [p^e_1, \cdots, p^e_n]^\top$ are in vector form. For generator i, the transfer function $g_i(s)$ has input $(u_i - p^e_i)$, the net power deviation at its generator axis, resulting from disturbances u_i in mechanical power minus variations in electrical power p^e_i drawn from the network, relative to their equilibrium values. The ouptut w_i is the angular frequency deviation relative to equilibrium frequency.

The network power fluctuations p^e are given by a linearized (lossless) DC model $p^e(s) = \frac{1}{s}Lw(s)$ of the power flow equation. Here L is the Laplacian matrix of an undirected weighted graph, with its elements given by

$$L_{ij} = \frac{\partial}{\partial \theta_j} \sum_{k=1}^{n} |V_i||V_k| b_{ik} \sin(\theta_i - \theta_k) \Bigg|_{\theta=\theta_0} ,$$

where θ_0 are angle deviations at steady state, $|V_i|$ is the voltage magnitude at bus i and b_{ij} is the line susceptance. Without loss of generality, we assume the steady state angular difference $\theta_{0i} - \theta_{0j}$ across each line is smaller than $\frac{\pi}{2}$. Moreover, because L is a symmetric real Laplacian, its eigenvalues are given by $0 = \lambda_1(L) \le \lambda_2(L) \le \cdots \le \lambda_n(L)$. The overall linearized frequency dynamics of the generators is given by

$$w_i(s) = g_i(s)(u_i(s) - p^e_i(s)), \quad i = 1, \cdots, n, \tag{3.80a}$$

$$p^e(s) = \frac{1}{s}Lw(s). \tag{3.80b}$$

Generally, a group of generators coupled as in Figure 3-6 is considered *coherent* if their response in frequency is the same/similar under a disturbance u of any shape. We are interested in characterizing the dynamic response of coherent generators, which we term here the *coherent dynamics*. With this aim, we seek conditions on the network (3.80) under which the entire set of generators behave coherently. The same approach can be used on subgroups of generators.

To motivate our results, we start with summing over all equations in (3.80a) to get

$$\sum_{i=1}^{n} g_i^{-1}(s) w_i(s) = \sum_{i=1}^{n} u_i(s) - \sum_{i=1}^{n} p_i^e(s) = \sum_{i=1}^{n} u_i(s)\,. \tag{3.81}$$

Notice that the term $\sum_{i=1}^{n} p_i^e(s) = \mathbb{1}^\top \frac{L}{s} w(s) = 0$ since $\mathbb{1} = [1, \cdots, 1]^\top$ is a left eigenvector of $\lambda_1(L) = 0$.

A pragmatic approach to obtain a model of coherent behavior is to simply *impose* the equality $w_i(s) = \hat{w}(s)$ between the frequency outputs. Solving from (3.81) we obtain:

$$\hat{w}(s) = \left(\sum_{i=1}^{n} g_i^{-1}(s) \right)^{-1} \sum_{i=1}^{n} u_i(s) =: \hat{g}(s) \sum_{i=1}^{n} u_i(s); \tag{3.82}$$

the group of generators is aggregated into a single effective machine $\hat{g}(s)$, responding to the total disturbance.

Coherence in tightly connected networks

To properly justify the use of (3.82) as an accurate descriptor of the coherent dynamics, we state here a precise result. Our analysis will highlight the role of the algebraic connectivity $\lambda_2(L)$ of the network as a direct indicator of how coherent a group of generators is.

For the network shown in Figure 3-6, the transfer matrix from the disturbance u to the frequency deviation w is given by

$$T(s) = \left(I_n + \mathrm{diag}\{g_i(s)\} L/s\right)^{-1} \mathrm{diag}\{g_i(s)\}\,, \tag{3.83}$$

where I_n is the $n \times n$ identity matrix. We establish that the transfer matrix $T(s)$ converges, as the algebraic connectivity $\lambda_2(L)$ increases, to one where all entries are given by $\hat{g}(s)$.

We make several assumptions: 1) $T(s)$ is stable; 2) $\hat{g}(s)$ in (3.82) is stable 3) all $g_i(s)$ are minimum phase systems. All generator network models discussed here (Section 3.3.1, 3.3.1) satisfy these assumptions. In particular, the stability of $T(s)$ is guaranteed by passivity of the network [98]. We state the following result.

Theorem 3.11. *Given the assumptions above, the following holds for any $\eta_0 > 0$:*

$$\lim_{\lambda_2(L)\to+\infty} \sup_{\eta\in[-\eta_0,\eta_0]} \left\|T(j\eta) - \hat{g}(j\eta)\mathbb{1}\mathbb{1}^\top\right\| = 0\,,$$

where $j = \sqrt{-1}$ and $\mathbb{1} \in \mathbb{R}^n$ is the vector of all ones.

Proof. $\bar{g}(s)$ is stable because $\hat{g}(s)$ is stable, then $\bar{g}(s)$ is continuous on compact set $[-j\eta_0, j\eta_0]$. Then by [83, Theorem 4.15] there exists $M_1 > 0$, such that $\forall s \in [-j\eta_0, j\eta_0]$, we have $|\bar{g}(s)| \leq M_1$. Similarly, because all $g_i(s)$ are minimum-phase, all $g_i^{-1}(s)$ are stable hence continuous on $[-j\eta_0, j\eta_0]$. Again there exists $M_2 > 0$, such that $\forall s \in [-j\eta_0, j\eta_0]$, we have $\max_{1\leq i\leq n} |g_i^{-1}(s)| \leq M_2$.

Now we know that $\forall s \in [-j\eta_0, j\eta_0]$, we have $|\bar{g}(s)| \leq M_1, \max_{1\leq i\leq n} |g_i^{-1}(s)| \leq M_2$, i.e. the condition for Lemma 3.1 is satisfied for a common choice of $M_1, M_2 > 0$.

By Lemma 3.1, $\forall s \in [-j\eta_0, j\eta_0]$, we have:

$$\left\|T(s) - \hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \leq \frac{(M_1M_2+1)^2}{|\lambda_2(L)/s| - M_2 - M_1M_2^2}.$$

Taking $\sup_{s\in[-j\eta_0,j\eta_0]}$ on both sides gives:

$$\sup_{s\in[-j\eta_0,j\eta_0]} \left\|T(s) - \hat{g}(s)\mathbb{1}\mathbb{1}^\top\right\| \leq \frac{(M_1M_2+1)^2}{|\lambda_2(L)/\eta_0| - M_2 - M_1M_2^2}.$$

Lastly, take $\lambda_2(L) \to +\infty$ on both sides, the right-hand side gives 0 in the limit, which finishes the proof. □

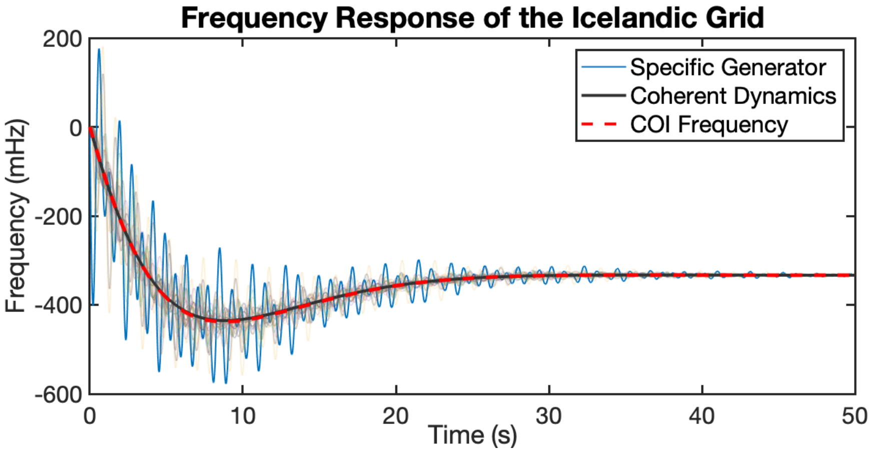

Figure 3-7. Step response of the Icelandic grid. Individual responses appear in light font in the background, with a specific one highlighted in blue. We also show the CoI frequency response, and step response of the coherent dynamics $\hat{g}(s)$. The Iceland network has Algebraic connectivity $\lambda_2(L) = 0.0915$.

The transfer matrix $\hat{g}(s)\mathbb{1}\mathbb{1}^\top$ has the property that for an arbitrary vector disturbance $u(s)$, the response is $w(s) = \hat{g}(s)\mathbb{1}\mathbb{1}^\top u(s) = (\hat{g}(s)\sum_{i=1}^n u_i(s))\,\mathbb{1}$; this says the vector of bus frequencies responds in unison, with all entries equal to the response $\hat{w}$ in (3.82). Theorem 3.11 states that in the limit of large connectivity, the true response $T(s)u(s)$ is approximated by the one in (3.82) for disturbances in the frequency band $[-\eta_0, \eta_0]$.

The limit of high connectivity analyzed in the theorem is a good assumption for many cases of tightly connected networks, but one may wonder about the relevance of $\hat{g}(s)$ in a less extreme case. We explore this through a numerical simulation on the Icelandic Power Grid [97], of moderate connectivty. As shown in Figure 3-7, the step response has incoherent oscillations from individual generators. Nevertheles, if one looks at the Center of Inertia (CoI) frequency $w_{\text{coi}} = (\sum_{i=1}^n m_i w_i)/(\sum_{i=1}^n m_i)$, a commonly used system-wide metric, we see it is very closely approximated by the coherent dynamics $\hat{g}(s)$. Thus we will proceed with this model of aggregate response. For certain generator models, however, the complexity of $\hat{g}(s)$ motivates the need for approximations.

Aggregate dynamics for different generator models

Having characterized how, the *coherent dynamics* given by $\hat{g}(s)$, represent the network's aggregate behavior, from now on we will use with no distinction the terms "aggregate" and "coherent" dynamics. Now we look into the explicit forms these dynamics take for different generator models.

Example 1. *Generators with 1st order model, of two types:*

1) For synchronous generators[48], $g_i(s) = \frac{1}{m_i s + d_i}$, where m_i, d_i are the inertia and damping of generator i, respectively. The coherent dynamics are $\hat{g}(s) = \frac{1}{\hat{m}s + \hat{d}}$, where $\hat{m} = \sum_{i=1}^{n} m_i$ and $\hat{d} = \sum_{i=1}^{n} d_i$.

2) For droop-controlled inverters[107], $g_i(s) = \frac{k_{P,i}}{\tau_{P,i}s+1}$, where $k_{P,i}$ and $\tau_{P,i}$ are the droop coefficient and the filter time constant of the active power measurement, respectively. The coherent dynamics are $\hat{g}(s) = \frac{\hat{k}_P}{\hat{\tau}_P s+1}$, where $\hat{k}_P = \left(\sum_{i=1}^{n} k_{P,i}^{-1}\right)^{-1}$, $\hat{\tau}_P = \hat{k}_P \left(\sum_{i=1}^{n} \tau_{P,i}/k_{P,i}\right)$.

Notice that both dynamics are of the same form; by suitable reparameterization, we may use the "swing" model $g_i(s) = \frac{1}{m_i s + d_i}$ to model both types of generators. In this case no order reduction is needed: the aggregate model given in Case 1 is consistent with the conventional approach of choosing inertia $\hat{m}$ and damping $\hat{d}$ as the respective sums over all generators. Theorem 3.11 explains why such a choice is indeed appropriate.

The aggregation is more complicated when considering generators with turbine droop control:

Example 2. *Synchronous generators given by the swing model with turbine droop[48]*

$$g_i(s) = \frac{1}{m_i s + d_i + \frac{r_i^{-1}}{\tau_i s+1}}, \tag{3.84}$$

where r_i^{-1} and τ_i are the droop coefficient and turbine time constant of generator i, respec-

tively. The coherent dynamics are given by

$$\hat{g}(s) = \frac{1}{\hat{m}s + \hat{d} + \sum_{i=1}^{n} \frac{r_i^{-1}}{\tau_i s+1}} \,. \tag{3.85}$$

When all generators have the same turbine time constant $\tau_i = \hat{\tau}$, then $\hat{g}(s)$ in (3.85) reduces to the typical effective machine model of the form (3.84) with parameters $(\hat{m}, \hat{d}, \hat{r}^{-1}, \hat{\tau})$, where $\hat{r}^{-1} = \sum_{i=1}^{n} r_i^{-1}$, i.e., the aggregation model is still obtained by choosing parameters as the respective sums of their individual values. However, if the τ_i are heterogeneous, then $\sum_{i=1}^{n} \frac{r_i^{-1}}{\tau_i s+1}$ is generally of high-order because the summands have distinct poles. As a result, the closed-loop dynamics $\hat{g}(s)$ is a high-order transfer function and cannot be accurately represented by a single generator model. The aggregation of generators thus requires a low-order approximation of $\hat{g}(s)$.

Aggregate dynamics for mixture of generators

We have shown the aggregate dynamics for generators of three different types. When a mixture of these different types is present[3], we adopt (3.84) as a general representation of the three types; in particular, the first order models can be regarded as (3.84) with $r_i^{-1} = 0$. Therefore, (3.85) provides a general representation of the aggregate dynamics resulting from a mixture of generators. Again, high-order coherent dynamics arise when heterogeneous turbines exist.

3.3.2 Reduced order model for coherent generators with heterogeneous turbines

As shown in the previous section, the coherent dynamics $\hat{g}(s)$ are of high-order if the coherent group has generators with different turbine time constants. This

[3] Generally, when considering a mixture of synchronous generators and grid-forming inverters, our network model is valid only when synchronous generators make up a significant portion of the composition.

suggests that substituting $\hat{g}(s)$ with an equivalent machine of the same order as each $g_i(s)$ may lead to a substantial approximation error. In this section we propose instead a hierarchy of reduced models with increasing order, based on balanced realization theory [108], such that eventually an accurate reduced model is obtained as the order of the reduction increases. An additional avenue of improvement is: instead of the standard approach [44, 45, 47] of reducing the aggregate of turbines, to apply the reduction methodology over the closed-loop coherent dynamics.

We use frequency weighted balanced truncation [109] to approximate $\hat{g}(s)$. Frequency weighted balanced truncation identifies the most significant dynamics with respect to particular LTI frequency weight by computing the weighted Hankel singular values, which decay fast in many cases, allowing us to accurately approximate high-order systems. Importantly, the reduction procedure favors approximation accuracy in certain frequency range specified by the weights. We defer the detailed procedure of frequency weighted balanced truncation at the end of this section. Given a SISO stable proper transfer function $G(s)$, and a stable frequency weight $W(s)$, the k-th order weighted balanced truncation returns

$$\tilde{G}_k(s) = \frac{b_{k-1}s^{k-1} + \cdots + b_1 s + b_0}{a_k s^k + \cdots + a_1 s + a_0}\,, \tag{3.86}$$

which is guaranteed to be stable [109], and such that the weighted error

$$\sup_{\eta\in\mathbb{R}} |W(j\eta)(G(j\eta) - \tilde{G}_k(j\eta))|$$

is upper bounded, with an upper bound decreasing to zero with the order k. For our purposes, $W(s)$ must have a high gain in the low frequency range, so that the DC gains of the original and the reduced dynamics are approximately matched, i.e., $G(0) \simeq \tilde{G}(0)$. Our two proposed model reduction approaches for high-order $\hat{g}(s)$ in (3.85) are both based on frequency weighted balanced truncation.

Model reduction on turbine dynamics

Our first model is based on applying balanced truncation to the turbine aggregate. Essentially, $\hat{g}(s)$ in (3.85) is of high order because it has high-order turbine dynamics $\sum_{i=1}^{n} \frac{r_i^{-1}}{\tau_i s+1}$; we seek to replace it with a reduced-order model. This is akin to the existing literature [44, 45] which replaces an aggregate of turbines in parallel by a first order turbine model with parameters obtained by minimizing certain error functions.

We denote the aggregate turbine dynamics as $\hat{g}_t(s) := \sum_{i=1}^{n} \frac{r_i^{-1}}{\tau_i s+1}$. We also denote the $(k-1)$-th reduction model of $\hat{g}_t(s)$ by frequency-weighted balanced truncation as $\tilde{g}_{t,k-1}(s)$. Then the k-th order reduction model of $\hat{g}(s)$ is given by

$$\tilde{g}_k^{tb}(s) = \frac{1}{\hat{m}s + \hat{d} + \tilde{g}_{t,k-1}(s)}, \tag{3.87}$$

with, again, $\hat{m} = \sum_{i=1}^{n} m_i, \hat{d} = \sum_{i=1}^{n} d_i$. We highlight two special instances of relevance for our numerical illustration.

2nd order reduction: When $k = 2$, the reduced model $\tilde{g}_{t,1}(s)$ can be interpreted as a first order turbine model

$$\tilde{g}_{t,1}(s) = \frac{b_0}{a_1 s + a_0} = \frac{b_0/a_0}{(a_1/a_0)s + 1} := \frac{\tilde{r}^{-1}}{\tilde{\tau}s + 1},$$

with parameters $(\tilde{r}^{-1}, \tilde{\tau})$ chosen by the weighted balanced truncation method. Then the overall reduced model $\tilde{g}_2^{tb}(s)$ is of second order, which is a single generator model.

Unlike [44, 45], there is a DC gain mismatch between $\tilde{g}_2^{tb}(s)$ and the original $\hat{g}(s)$ since $\tilde{r}^{-1} \neq \hat{r}^{-1} = \sum_{i=1}^{n} r_i^{-1}$. Later in the simulation section, by choosing a proper frequency weight $W(s)$, we effectively make the DC gain mismatch negligible. However, as we will see in the numerical section, $k = 2$ may not suffice to accurately approximate the coherent dynamics.

3rd order reduction: To obtain a more accurate reduced-order model, one may consider $k = 3$ as the next suitable option. In fact, as we see in the later numerical simulation, a 2nd order turbine model $\tilde{g}_{t,2}(s)$, i.e., $k = 3$, is sufficient to give an almost exact approximation of $\hat{g}_t(s)$.

We can also interpret $\tilde{g}_{t,2}(s)$, by means of partial fraction expansion, i.e.,

$$\tilde{g}_{t,2}(s) = \frac{b_1 s + b_0}{a_2 s^2 + a_1 s + a_0} = \frac{\tilde{r}_1^{-1}}{\tilde{\tau}_1 s + 1} + \frac{\tilde{r}_2^{-1}}{\tilde{\tau}_2 s + 1},$$

assuming the poles are real. Then the reduced dynamics $\tilde{g}_{t,2}(s)$ can be viewed as two first order turbines in parallel with parameters $(\tilde{r}_1^{-1}, \tilde{\tau}_1)$ and $(\tilde{r}_2^{-1}, \tilde{\tau}_2)$. In Section 3.3.3, we show such interpretation is valid for our numerical example.

Model reduction on closed-loop coherent dynamics

Our second proposal is to apply weighted balanced truncation directly on $\hat{g}(s)$, instead of reducing the turbine dynamics (3.87). Thus, we denote $\tilde{g}_k^{cl}(s)$ as the k-th order reduction model, via frequency weighted balanced truncation, of the coherent dynamics $\hat{g}(s)$. Again, DC gain mismatch can be made negligible by properly choosing $W(s)$.

As compared to the one in Section 3.3.2, this reduced model might not be easy to interpret. Nevertheless, the procedure described below often leads to a practical interpretation.

2nd order reduction: When $k = 2$, we wish to interpret $\tilde{g}_2^{cl}(s)$ in terms of a single generator with a first order turbine of the form in (3.84), with parameters $(\tilde{m}, \tilde{d}, \tilde{r}^{-1}, \tilde{\tau})$. Given

$$\tilde{g}_2^{cl}(s) = \frac{b_1 s + b_0}{a_2 s^2 + a_1 s + a_0} := \frac{N(s)}{D(s)},$$

obtained by the proposed method, we write the polynomial division $D(s) = Q(s)N(s) + R$, where $Q(s)$, R are quotient and remainder, respectively. This leads to the expression

$$\tilde{g}_2^{cl}(s) = \frac{N(s)}{Q(s)N(s)+R} = \frac{1}{Q(s)+\frac{R}{N(s)}} \, .$$

Here the first order polynomial $Q(s)$ can be matched to $\tilde{m}s+\tilde{d}$, and $\frac{R}{N(s)}$ to $\frac{\tilde{r}^{-1}}{\tilde{\tau}s+1}$. Provided the obtained constants $(\tilde{m}, \tilde{d}, \tilde{r}^{-1}, \tilde{\tau})$ are positive, the interpretation follows.

3rd order reduction: Similarly, when $k = 3$, the reduced model is $\tilde{g}_3^{cl}(s) = \frac{N(s)}{D(s)}$, with $N(s)$ of 2nd order and $D(s)$ of 3rd order. The polynomial division $D(s) = Q(s)N(s) + R(s)$, still gives a first order quotient $Q(s)$, which is interpreted as $\tilde{m}s + \tilde{d}$; the second order transfer function $\frac{R(s)}{N(s)}$ can be expressed, by partial fraction expansion, as two first order turbines in parallel, provided the obtained constants remain positive. We explore this in the examples studied below.

3.3.3 Numerical Simulations

We now evaluate the reduction methodologies proposed in the previous section, and compare their performance with the solutions proposed in [44, 45]. In our comparison, we consider 5 generators forming a coherent group[4]. All parameters are expressed in a common base of 100 MVA.

The test case: 5 generators, $\hat{m} = 0.0683(\text{s}^2/\text{rad})$, $\hat{d} = 0.0107$. The turbine and droop parameters of each generator are listed in Table 3-I. In all comparisons, a step change of -0.1 p.u. in disturbance power is used.

Table 3-I. Droop control parameters of generators in test case

Index / Parameter	1	2	3	4	5
droop r_i^{-1} (p.u.)	0.0218	0.0256	0.0236	0.0255	0.0192
time constant τ_i (s)	9.08	5.26	2.29	7.97	3.24

[4] More specifically, we assume sufficiently strong network coupling among these generators such that the frequency responses are coherent. The numerical simulation will only illustrate the approximation accuracy with respect to the coherent response rather than individual ones.

Remark 10. *In the test case, we only aggregate 5 generators and report all parameters explicitly in order to give insight on how the distribution of the time constant τ_i affects our approximations. It is worth noting that similar behavior is observed when reducing coherent groups with a much larger number of generators. In particular, the accuracy found below with 3rd order reduced models is also observed in these higher order problems.*

As mentioned in the previous section, one of the drawbacks of the balanced truncation method is the DC gain mismatch, which leads to a steady-state error. In our simulation, the DC gain mismatch is effectively cancelled by picking proper frequency weights for different reduced models.

Effect of reduction order k in accuracy

We now evaluate the effect of the reduction order on the accuracy. That is, we compare 2nd and 3rd order balanced truncation on the turbine dynamics, $\tilde{g}_2^{tb}(s)$ (BT2-tb), $\tilde{g}_3^{tb}(s)$ (BT3-tb), as well as balanced truncation on the closed-loop coherent dynamics $\tilde{g}_2^{cl}(s)$ (BT2-cl), $\tilde{g}_3^{cl}(s)$ (BT3-cl). The frequency weights are given by $W_{tb}(s) = \frac{s+3\cdot 10^{-2}}{s+10^{-4}}$ and $W_{cl}(s) = \frac{s+8\cdot 10^{-2}}{s+10^{-4}}$, respectively. The step response and step response error with respect to $\hat{g}(s)$ are shown in Figure 3-8.

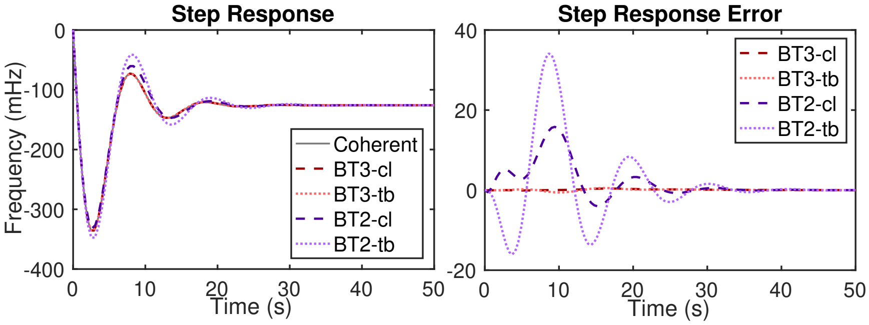

Figure 3-8. Comparison of all reduced-order models by balanced truncation

Compared to 2nd order models, 3rd order reduced models give a very accurate approximation of $\hat{g}(s)$. While it is expected that the approximation error goes down with the order, it is not trivial that a 3rd order model would provide this level of accuracy for an intrinsically high order system.

Moreover, when we examine the transfer function given by $\tilde{g}_3^{tb}(s)$ (from input u in p.u. to output w in rad/s), we find an interesting interpretation. That is, the turbine model for $\tilde{g}_3^{tb}(s)$ is given by

$$\tilde{g}_{t,2}(s) = \frac{0.0266s + 0.0057}{s^2 + 0.5046s + 0.0489} = \frac{0.0473}{2.68s + 1} + \frac{0.0684}{7.64s + 1},$$

where the latter is obtained by partial fraction expansion and can be viewed as two turbines (one fast turbine and one slow turbine) in parallel, and the choices of droop coefficients for these two turbines reflect the aggregate droop coefficients of fast turbines (generators 3 and 5) and slow turbines (generators 1,2, and 4), respectively, in $\hat{g}(s)$.

Reduction on turbines vs. closed-loop dynamics

Another observation from Figure 3-8 is that reduction on the closed-loop is more accurate than reduction on the turbine. For a more straightforward comparison, we list in Table 3-II the approximation errors of all 4 models in Fig 3-8 using the following metrics: 1) $\mathcal{L}_2$-norm of step response error[5] $e(t)$ (in $\mathrm{rad/s}^{1/2}$): $(\int_0^{+\infty} |e(t)|^2 dt)^{1/2}$; 2) $\mathcal{L}_\infty$-norm of $e(t)$ (in $\mathrm{rad/s}$): $\max_{t\geq 0} |e(t)|$; 3) $\mathcal{H}_\infty$-norm difference between reduced and original models (from input u in p.u. to output w in rad/s).

We observe from Table 3-II that for a given reduction order, balanced truncation on the closed-loop dynamics ($\tilde{g}_2^{cl}(s)$, $\tilde{g}_3^{cl}(s)$) has smaller approximation error than balanced truncation on turbine dynamics ($\tilde{g}_2^{tb}(s)$, $\tilde{g}_3^{tb}(s)$) *across all metrics*. Such observation seems to be true in general. For instance, Fig. 3-9 shows a similar trend by plotting the same configuration (metrics and models) of Table 3-II for different values of of the aggregate inertia $\hat{m}$, while keeping all other parameters the same.

[5] For reduced-order models obtained via frequency weighted balanced truncation, there exists an extremely small but non-zero DC gain mismatch that makes the $\mathcal{L}_2$-norm unbounded. We resolve this issue by simply scaling our reduced-order models to have exactly the same DC gain as $\hat{g}(s)$.

Table 3-II. Approximation errors of reduced order models

Metric Model	$\mathcal{L}_2$ diff. (rad/s$^{1/2}$)	$\mathcal{L}_\infty$ diff. (rad/s)	$\mathcal{H}_\infty$ diff.
Guggilam[45]	7.2956	3.8287	10.2748
Germond[44]	3.9594	1.9974	5.1431
BT2-tb	4.3737	2.1454	7.5879
BT2-cl	2.0376	0.9934	2.0381
BT3-tb	0.0967	0.0361	0.1315
BT3-cl	**0.0704**	**0.0249**	**0.0317**

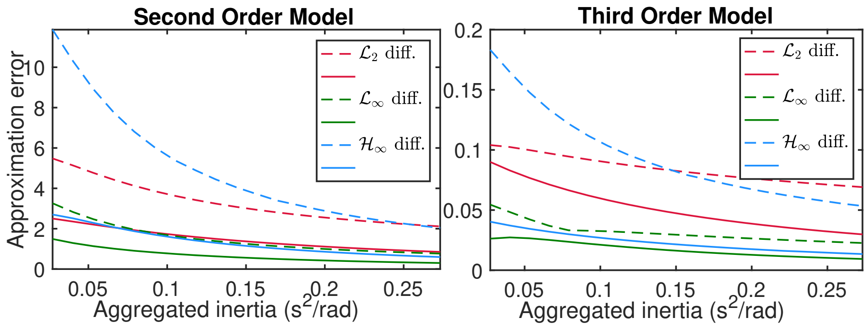

Figure 3-9. Approximation errors of second order models (left) and third order models (right) by balanced truncation in different metrics. Approximation errors of reduced-order models $\tilde{g}_2^{tb}(s)$, $\tilde{g}_3^{tb}(s)$ are shown in dashed lines; Approximation errors of reduced-order models $\tilde{g}_2^{cl}(s)$, $\tilde{g}_3^{cl}(s)$ are shown in solid lines. The approximation errors are in their respective units.

It can be seen from Fig. 3-9 that reduction on closed-loop dynamics improves the approximation in every metric, uniformly, for a wide range of aggregate inertia $\hat{m}$ values. The main reason is that, when applying reduction on the closed-loop dynamics, the algorithm has the flexibility to choose the corresponding values of inertia and damping to be different from the aggregate ones in order to better approximate the response. More precisely, from the reduced model we obtain

$$\tilde{g}_2^{cl}(s) = \frac{4.9733s + 1}{(0.06715s + 0.01464)(4.9733s + 1) + 0.1118},$$

from which we can get the equivalent swing and turbine models as:

$$\text{swing model: } \frac{1}{0.06715s + 0.01464}, \quad \text{turbine: } \frac{0.1118}{4.9733s + 1}.$$

The equivalent inertia and damping are $\tilde{m} = 0.06715$ and $\tilde{d} = 0.01464$, which are different from the aggregate values $\hat{m}, \hat{d}$. Therefore, when compared to reduction on turbine dynamics, reduction on closed-loop dynamics is less constrained on the parameter space, thus achieving smaller approximation errors.

Comparison with existing methods

Lastly, we compare reduced-order models via balanced truncation on the closed-loop dynamics, $\tilde{g}_2^{cl}(s)$, $\tilde{g}_3^{cl}(s)$, with the solutions proposed in [44, 45]. The step responses and the approximation errors are shown in Fig. 3-10 and Table. 3-II.

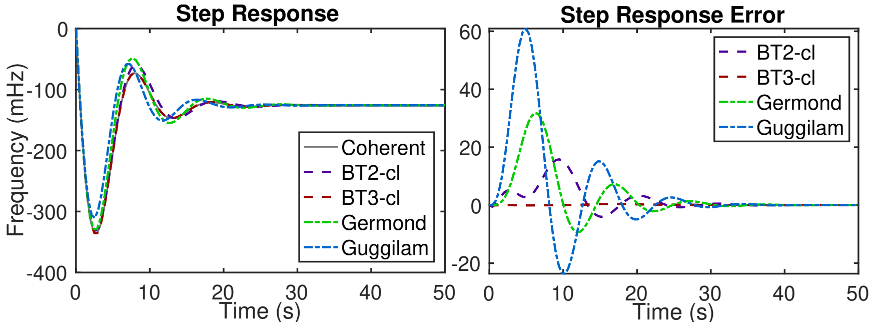

Figure 3-10. Comparison with existing reduced-order models

In the comparison, $\tilde{g}_3^{cl}(s)$ outperforms all other reduced-order models and is the most accurate reduced-order model of $\hat{g}(s)$. It is also worth noting that $\tilde{g}_2^{cl}(s)$ has the least approximation error among all 2nd order models. In general, our results suggest that to improve the accuracy of reduced-order models of the coherent dynamics of generators $\hat{g}(s)$, we should consider: 1) increasing the complexity (order) of the reduced model; 2) reduction on the closed-loop dynamics instead of the turbine dynamics.

Frequency Weighted balanced Truncation

Given a minimum realization of frequency weight $W(s)$ to be (A_W, B_W, C_W, D_W), the procedures of frequency weighted balanced truncation for a minimum, strictly proper and stable linear system (A, B, C) with order n are given as follow:

1. The extended system[6] is given by:

$$\left[\begin{array}{cc|c} A & \mathbb{0} & B \\ B_W C & A_W & \mathbb{0} \\ \hline D_W C & C_W & \mathbb{0} \end{array}\right] := \left[\begin{array}{c|c} \bar{A} & \bar{B} \\ \hline \bar{C} & \mathbb{0} \end{array}\right].$$

2. Compute the frequency weighted controllability and observability gramians X_c, Y_o from the gramians $\bar{X}_c, \bar{Y}_o$ of extended system:

$$\bar{X}_c = \int_0^\infty e^{\bar{A}t}\bar{B}\bar{B}^\top e^{\bar{A}^\top t}dt,\ \bar{Y}_o = \int_0^\infty e^{\bar{A}^\top t}\bar{C}^\top \bar{C}e^{\bar{A}t}dt$$

$$X_c = \begin{bmatrix} I_n & \mathbb{0} \end{bmatrix} \bar{X}_c \begin{bmatrix} I_n \\ \mathbb{0} \end{bmatrix},\ Y_c = \begin{bmatrix} I_n & \mathbb{0} \end{bmatrix} \bar{Y}_c \begin{bmatrix} I_n \\ \mathbb{0} \end{bmatrix}.$$

3. Perform the singular value decomposition of $X_c^{\frac{1}{2}} Y_o X_c^{\frac{1}{2}}$:

$$X_c^{\frac{1}{2}} Y_o X_c^{\frac{1}{2}} = U\Sigma U^*.$$

where U is unitary and Σ is diagonal, positive definite with its diagonal terms in decreasing order. Then compute the change of coordinates T given by:

$$T^{-1} = X_c^{\frac{1}{2}} U \Sigma^{-1}.$$

4. Apply change of coordinates T on (A, B, C) to get its balanced realization (TAT^{-1}, TB, CT^{-1}). Then the k-th order $(1 \le k \le n)$ reduction model (A_k, B_k, C_k) is given by truncating (TAT^{-1}, TB, CT^{-1}) as the following:

$$\begin{aligned} A_k &= \begin{bmatrix} I_k & \mathbb{0} \end{bmatrix} TAT^{-1} \begin{bmatrix} I_k \\ \mathbb{0} \end{bmatrix} \\ B_k &= \begin{bmatrix} I_k & \mathbb{0} \end{bmatrix} TB \\ C_k &= CT^{-1} \begin{bmatrix} I_k \\ \mathbb{0} \end{bmatrix}. \end{aligned}$$

Remark 11. *Balanced truncation only applies to systems in state space. For a transfer function, one should apply balanced truncation to its minimum realization, then obtain reduced order transfer function from the state-space reduction model.*

[6] When $W(s) = 1$, the extended system is exactly the same as original (A, B, C), then the procedures give unweighted standard balanced truncation.

3.4 Conclusion

In this chapter, we study network coherence as a low-rank property of the transfer matrix $T(s)$ in the frequency domain. The analysis leads to useful characterizations of coordinated behavior and justifies the relation between network coherence and network effective algebraic connectivity. Our results suggest that network coherence is a frequency-dependent phenomenon, which is numerically illustrated in power networks. Lastly, concentration results for large-scale networks are presented, revealing the exclusive role of the statistical distribution of node dynamics in determining the coherent dynamics of such networks. The network coherence analysis forms the basis for analyzing dominant dynamics in networks with multiple coherence clusters, discussed in the next section.

Next, we extend our frequency-domain analysis to the case of multi-cluster network systems. We propose a structure-preserving model-reduction methodology for large-scale dynamic networks. Our analysis shows that networks with multiple coherent groups can be well approximated by a reduced network of the same size as the number of coherent groups, and we provide an upper bound on the approximation error when the network graph is randomly generated from a weight stochastic block model.

Lastly, to address the high complexity of our proposed model for power network reduction. We seek tractable models for frequency dynamics in a power grid, starting with the characterization $\hat{g}(s) = \left(\sum_{i=1}^{n} g_i^{-1}(s)\right)^{-1}$ for the coherent response, which is shown to be asymptotically accurate as the coupling between generators (characterized via $\lambda_2(L)$) increases. Our characterization justifies existing aggregation approaches and also explains the difficulties of aggregating generators with heterogeneous turbine time constants. We leverage model reduction tools from control theory to find accurate reduced-order approximations to $\hat{g}(s)$. For $\{g_i(s)\}_{i=1}^{n}$

given by the 2nd order generator models, the numerical study shows that 3rd order models based on frequency weighted balanced truncation on closed-loop dynamics are sufficient to accurately represent $\hat{g}(s)$.

For future research, we believe our proposed model can be applied to power networks for studying the inter-area oscillation in the frequency response and allows new control designs based on the reduced network. Moreover, in the case of unknown node dynamics, one possible path is to learn coherent dynamics from output measurement data, then build a reduced network from the learned dynamics.

Chapter 4

Conclusions and General Discussion

In this thesis, we discuss two types of high-dimensional dynamical systems: training dynamics of neural networks and large-scale network systems. While classic dynamical systems and control tools do not scale with the dimensionality of the state-space, making it challenging to characterize and understand their dynamical behavior. One can exploit their structural properties to develop new analyses of these systems:

We first consider training multi-layer neural networks using gradient flow dynamics. The high dimensionality comes from overparametrization: a typical network has a large depth and hidden layer width. With the presence of millions and billions of training parameters, even the simplest questions such as whether the gradient flow converges to a global minimum of the loss become challenging and can only be answered for networks with certain architectures. For linear networks, the symmetry of the weights, a critical property induced by the multi-layer architecture, turns out to be the key to analyzing convergence. Such symmetry leads to a set of time-invariant quantities, called weight imbalance, that restrict the training trajectory to a low-dimensional manifold defined by the weight initialization. A tailored convergence analysis is developed over this low-dimensional manifold, showing improved rate bounds for several multi-layer network models studied in the literature, leading to novel characterizations of the effect of weight imbalance on the convergence rate. Moreover, our analysis is extended to the case of training two-layer ReLU networks under small initialization.

Then, we consider large-scale networked systems with multiple weakly-connected groups. Such a multi-cluster structure leads to a time-scale separation between the fast intra-group interaction due to high intra-group connectivity, and the slow inter-group oscillation, due to the weak inter-group connection. We develop novel frequency-domain network coherence analysis. Unlike prior work, our analysis applies to networks with heterogeneous nodal dynamics, and further provides an explicit characterization in the frequency domain of the coherent response to disturbances as the harmonic mean of individual nodal dynamics. The new frequency-domain analysis leads to a structure-preserving model-reduction methodology for large-scale dynamic networks with multiple clusters.

There are many related research topics worth exploring. Regarding the training dynamics of neural networks, there have been numerous developments in optimization algorithms and network architecture design that have led to the success of machine learning in many applications. How our neural network model benefits from these designs will continue to be one of the most important questions in unveiling the mystery of deep learning. In practice, the algorithms used for optimizing a neural network are more complicated than simple gradient descent algorithms. At every iteration, a mini-batch of data is fed to evaluate the gradient instead of the full data, and gradient descent updates are replaced by weight updates from accelerated methods such as momentum-based optimization algorithms. These optimization algorithms allow for efficient training of neural networks, and the trained networks are empirically shown to have superior generalization performance compared to those trained with vanilla gradient descent algorithms. Moreover, practical neural networks have also greatly benefited from architectural designs such as convolutional layers, residual connections, batch normalization, and self-attention mechanisms. Understanding the convergence of practical neural networks, such as CNNs and Transformers, under stochastic and accelerated optimization algorithms will provide theoretical guarantees for the training and inspire new algorithmic and architectural improvements.

Regarding large-scale network systems, the next step is the control design: with interpretable and structured reduction models for the networks, how can we design controllers for the network nodes, or improve the network connection to achieve desired network response to the disturbance so that the network can maintain safe operation? In the context of power networks, this is related to the problem regulating the frequency response of the entire network subject to the power imbalance in a certain area. The problem of control design in large networks is even more challenging when the node dynamics or network topology is unknown. Knowing that the network dynamics are governed by a few dominant modes that correspond to the coherent behaviors, how can we leverage such low-dimensional properties to design system identification algorithms for building a reduced model in a data-driven fashion? Studying these questions will eventually lead to a scalable, and interpretable control design for the safe and robust operation of large-scale networks.

Bibliography

[1] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[2] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *29th Annual Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, V. Feldman, A. Rakhlin, and O. Shamir, Eds., vol. 49. Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 1246–1257.

[3] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?" 2015.

[4] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1724–1732.

[5] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1233–1242.

[6] K. Kawaguchi, “Deep learning without poor local minima,” in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] B. D. Haeffele and R. Vidal, “Global optimality in tensor factorization, deep learning, and beyond,” *arXiv preprint arXiv:1506.07540*, 2015.

[8] ——, “Global optimality in neural network training,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7331–7339.

[9] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, B. Póczos, and A. Singh, “Gradient descent can take exponential time to escape saddle points,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1067–1077.

[10] S. Arora, N. Cohen, W. Hu, and Y. Luo, “Implicit regularization in deep matrix factorization,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] G. Gidel, F. Bach, and S. Lacoste-Julien, “Implicit regularization of discrete gradient dynamics in linear neural networks,” in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019, pp. 3202–3211.

[12] Y. Li, T. Ma, and H. Zhang, “Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations,” in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75. PMLR, 06–09 Jul 2018, pp. 2–47.

[13] D. Stöger and M. Soltanolkotabi, “Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction,” *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] Z. Li, Y. Luo, and K. Lyu, “Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning,” in *International Conference on Learning Representations*, 2021.

[15] J. Li, T. V. Nguyen, C. Hegde, and R. K. W. Wong, “Implicit sparse regularization: The impact of depth and early stopping,” in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[16] E. Boursier, L. Pullaud-Vivien, and N. Flammarion, “Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 20 105–20 118.

[17] N. Razin, A. Maman, and N. Cohen, “Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 422–18 462.

[18] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Advances in neural information processing systems*, 2018, pp. 8571–8580.

[19] L. Chizat, E. Oyallon, and F. Bach, “On lazy training in differentiable programming,” in *Advances in Neural Information Processing Systems*, 2019, pp. 2937–2947.

[20] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, “On exact computation with an infinitely wide neural net,” in *Advances in Neural Information Processing Systems*, 2019, pp. 8141–8150.

[21] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” in *International Conference on Learning Representations(ICLR), 2019*, 2019.

[22] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 242–252.

[23] S. Du and W. Hu, “Width provably matters in optimization for deep linear neural networks,” in *International Conference on Machine Learning*, 2019, pp. 1655–1664.

[24] A. M. Saxe, J. L. Mcclelland, and S. Ganguli, “Exact solutions to the non-linear dynamics of learning in deep linear neural network,” in *International Conference on Learning Representations*, 2014.

[25] S. Tarmoun, G. França, B. D. Haeffele, and R. Vidal, “Understanding the dynamics of gradient flow in overparameterized linear models,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 10 153–10 161.

[26] S. Arora, N. Cohen, N. Golowich, and W. Hu, “A convergence analysis of gradient descent for deep linear neural networks,” in *International Conference on Learning Representations*, 2018.

[27] S. Arora, N. Cohen, and E. Hazan, “On the optimization of deep networks: Implicit acceleration by overparameterization,” in *35th International Conference on Machine Learning*, 2018.

[28] C. Yun, S. Krishnan, and H. Mobahi, “A unifying view on implicit bias in training linear neural networks,” in *International Conference on Learning Representations*, 2020.

[29] H. Min, R. Vidal, and E. Mallada, “On the convergence of gradient flow on multi-layer linear models,” in *The 40th International Conference on Machine Learning (ICML)*, 2023, to appear.

[30] H. Min, S. Tarmoun, R. Vidal, and E. Mallada, “On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 18–24 Jul 2021, pp. 7760–7768.

[31] ——, “Convergence and implicit bias of gradient flow on overparametrized linear networks,” *arXiv preprint arXiv:2105.06351*, 2022.

[32] H. Min, R. Vidal, and E. Mallada, “Early neuron alignment in two-layer relu networks with small initialization,” *arXiv preprint arXiv:2307.12851*, 2023.

[33] R. Olfati-Saber and R. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1520–1533, 2004.

[34] C. Zhao, U. Topcu, N. Li, and S. Low, “Power system dynamics as primal-dual algorithm for optimal load control,” *arXiv preprint arXiv:1305.0585*, 2013.

[35] B. Bamieh, M. R. Jovanovic, P. Mitra, and S. Patterson, “Coherence in large-scale networks: Dimension-dependent limitations of local feedback,” *IEEE Trans. Automat. Contr.*, vol. 57, no. 9, pp. 2235–2249, 2012.

[36] M. H. DeGroot, “Reaching a consensus,” *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[37] A. Jadbabaie, J. Lin, and A. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” *IEEE Trans. Automat. Contr.*, vol. 48, no. 6, pp. 988–1001, 2003.

[38] J. H. Chow, *Time-scale modeling of dynamic networks with applications to power systems*. Springer, 1982.

[39] G. Ramaswamy, L. Rouco, O. Fillatre, G. Verghese, P. Panciatici, B. Lesieutre, and D. Peltier, “Synchronic modal equivalencing (sme) for structure-preserving dynamic equivalents,” *IEEE Transactions on Power Systems*, vol. 11, no. 1, pp. 19–29, 1996.

[40] D. Romeres, F. Dörfler, and F. Bullo, “Novel results on slow coherency in consensus and power networks,” in *2013 European Control Conference (ECC)*, 2013, pp. 742–747.

[41] I. Tyuryukanov, M. Popov, M. A. M. M. van der Meijden, and V. Terzija, “Slow coherency identification and power system dynamic model reduction by using orthogonal structure of electromechanical eigenvectors,” *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1482–1492, 2021.

[42] J. Fritzsch and P. Jacquod, “Long wavelength coherency in well connected electric power networks,” *IEEE Access*, vol. 10, pp. 19 986–19 996, 2022.

[43] P. M. Anderson and M. Mirheydar, “A low-order system frequency response model,” *IEEE Trans. Power Syst.*, vol. 5, no. 3, pp. 720–729, 1990.

[44] A. J. Germond and R. Podmore, “Dynamic aggregation of generating unit models,” *IEEE Trans. Power App. Syst.*, vol. PAS-97, no. 4, pp. 1060–1069, July 1978.

[45] S. S. Guggilam, C. Zhao, E. Dall’Anese, Y. C. Chen, and S. V. Dhople, “Optimizing DER participation in inertial and primary-frequency response,” *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 5194–5205, Sep. 2018.

[46] D. Apostolopoulou, P. W. Sauer, and A. D. Domínguez-García, “Balancing authority area model and its application to the design of adaptive AGC systems,” *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3756–3764, Sep. 2016.

[47] M. L. Ourari, L.-A. Dessaint, and V.-Q. Do, “Dynamic equivalent modeling of large power systems using structure preservation technique,” *IEEE Trans. Power Syst.*, vol. 21, no. 3, pp. 1284–1295, 2006.

[48] F. Paganini and E. Mallada, “Global analysis of synchronization performance for power systems: Bridging the theory-practice gap,” *IEEE Trans. Automat. Contr.*, vol. 65, no. 7, pp. 3007–3022, 2020.

[49] Y. Jiang, R. Pates, and E. Mallada, “Dynamic droop control in low inertia power systems,” *IEEE Transactions on Automatic Control*, vol. 66, no. 8, pp. 3518–3533, 8 2021.

[50] Y. Jiang, A. Bernstein, P. Vorobev, and E. Mallada, “Grid-forming frequency shaping control in low inertia power systems,” *IEEE Control Systems Letters (L-CSS)*, vol. 5, no. 6, pp. 1988–1993, 12 2021, also in ACC 2021.

[51] E. Ekomwenrenren, Z. Tang, J. W. Simpson-Porco, E. Farantatos, M. Patel, and H. Hooshyar, “Hierarchical coordinated fast frequency control using inverter-based resources,” *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 4992–5005, 2021.

[52] E. Tegling, B. Bamieh, and H. Sandberg, “Localized high-order consensus destabilizes large-scale networks,” in *2019 American Control Conference (ACC)*, July 2019, pp. 760–765.

[53] H. G. Oral, E. Mallada, and D. F. Gayme, “Performance of first and second order linear networked systems over digraphs,” in *IEEE 56th Annu. Conf. on Decision and Control*, Dec 2017, pp. 1688–1694.

[54] B. Bamieh and D. F. Gayme, “The price of synchrony: Resistive losses due to phase synchronization in power networks,” in *2013 American Control Conference*, 2013, pp. 5815–5820.

[55] M. Andreasson, E. Tegling, H. Sandberg, and K. H. Johansson, "Coherence in synchronizing power networks with distributed integral control," in *IEEE 56th Annu. Conf. on Decision and Control*, Dec 2017, pp. 6327–6333.

[56] M. Pirani, J. W. Simpson-Porco, and B. Fidan, "System-theoretic performance metrics for low-inertia stability of power networks," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 5106–5111.

[57] H. Min and E. Mallada, "Dynamics concentration of large-scale tightly-connected networks," in *IEEE 58th Conf. on Decision and Control*, 2019, pp. 758–763.

[58] H. Min, R. Pates, and E. Mallada, "A frequency domain analysis of slow coherency in networked systems," *arXiv preprint arXiv:2302.08438*.

[59] H. Min and E. Mallada, "Spectral clustering and model reduction for weakly-connected coherent network systems," in *American Control Conference (ACC)*, Jan 2023, pp. 1–7.

[60] ——, "Learning coherent clusters in weakly-connected network systems," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, vol. 211. PMLR, 6 2023, pp. 1167–1179.

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

[62] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, 2020.

[63] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and double descent curve," *arXiv preprint arXiv:1908.05355*, 2019.

[64] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[65] T. H. Grönwall, "Note on the derivatives with respect to a parameter of the solutions of a system of differential equations," *Annals of Mathematics*, vol. 20, no. 4, pp. 292–296, 1919.

[66] Sheng-De Wang, Te-Son Kuo, and Chen-Fa Hsu, "Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation," *IEEE Transactions on Automatic Control*, vol. 31, no. 7, pp. 654–656, 1986.

[67] M. W. Hirsch, R. L. Devaney, and S. Smale, *Differential equations, dynamical systems, and linear algebra*. Academic press, 1974, vol. 60.

[68] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and banach spaces," *Handbook of the geometry of Banach spaces*, vol. 1, no. 317-366, p. 131, 2001.

[69] B. T. Polyak, *Introduction to optimization*. New York: Optimization Software, Publications Division, 1987.

[70] O. Elkabetz and N. Cohen, "Continuous vs. discrete optimization of deep neural networks," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[71] T. Le and S. Jegelka, "Training invariances and the low-rank phenomenon: beyond linear networks," in *International Conference on Learning Representations*, 2022.

[72] L. Wu, Q. Wang, and C. Ma, "Global convergence of gradient descent for deep linear residual networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[73] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018.

[74] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256.

[75] Y. Wang and S. Zheng, "The converse of weyl's eigenvalue inequality," *Advances in Applied Mathematics*, vol. 109, pp. 65–73, 2019.

[76] W. Rudin, *Principles of mathematical analysis*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953.

[77] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet, "Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity," *Transactions of the American Mathematical Society*, vol. 362, no. 6, pp. 3319–3363, 2010.

[78] M. Phuong and C. H. Lampert, "The inductive bias of relu networks on orthogonally separable data," in *International Conference on Learning Representations*, 2021.

[79] H. Maennel, O. Bousquet, and S. Gelly, "Gradient descent quantizes relu network features," *arXiv preprint arXiv:1803.08367*, 2018.

[80] Z. Xu, H. Min, S. Tarmoun, E. Mallada, and R. Vidal, “Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 2262–2284.

[81] B. Zhao, N. Dehmamy, R. Walters, and R. Yu, “Symmetry teleportation for accelerated optimization,” *arXiv preprint arXiv:2205.10637*, 2022.

[82] R. Olfati-Saber, J. A. Fax, and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[83] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[84] G. E. Dullerud and F. Paganini, *A course in robust control theory: a convex approach*. Springer Science & Business Media, 2013, vol. 36.

[85] H. Marquez and C. Damaren, “Comments on "strictly positive real transfer functions revisited,” *IEEE Transactions on Automatic Control*, vol. 40, no. 3, pp. 478–479, 1995.

[86] I. Lestas and G. Vinnicombe, “Scalable decentralized robust stability certificates for networks of interconnected heterogeneous dynamical systems,” *IEEE Transactions on Automatic Control*, vol. 51, no. 10, pp. 1613–1625, 2006.

[87] U. T. Jönsson and C.-Y. Kao, “A scalable robust stability criterion for systems with heterogeneous lti components,” *IEEE Transactions on Automatic Control*, vol. 55, no. 10, pp. 2219–2234, 2010.

[88] R. Pates and E. Mallada, "Robust scale-free synthesis for frequency control in power systems," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1174–1184, 2019.

[89] R. E. Mirollo and S. H. Strogatz, "Synchronization of pulse-coupled biological oscillators," *SIAM Journal on Applied Mathematics*, vol. 50, no. 6, pp. 1645–1662, 1990.

[90] E. Mallada, X. Meng, M. Hack, L. Zhang, and A. Tang, "Skewless network clock synchronization without discontinuity: Convergence and performance," *IEEE/ACM Transactions on Networking (TON)*, vol. 23, no. 5, pp. 1619–1633, 10 2015.

[91] R. Sepulchre, D. Paley, and N. Leonard, "Stabilization of planar collective motion with limited communication," *IEEE Trans. Automat. Contr.*, vol. 53, no. 3, pp. 706–719, 2008.

[92] G. Ramaswamy, G. Verghese, L. Rouco, C. Vialas, and C. DeMarco, "Synchrony, aggregation, and multi-area eigenanalysis," *IEEE Transactions on Power Systems*, vol. 10, no. 4, pp. 1986–1993, 1995.

[93] F. Wu and N. Narasimhamurthi, "Coherency identification for power system dynamic equivalents," *IEEE Transactions on Circuits and Systems*, vol. 30, no. 3, pp. 140–147, 1983.

[94] S. Sastry and P. Varaiya, "Coherency for interconnected power systems," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 218–226, 1981.

[95] M. Z. Q. Chen and M. C. Smith, "A note on tests for positive-real functions," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 390–393, 2009.

[96] J. H. Chow, *Power system coherency and model reduction*. New York, NY, USA: Springer, 2013.

[97] U. of Edinburgh. Power systems test case archive. Mar. 2003.

[98] H. K. Khalil and J. W. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.

[99] W. K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, pp. 1161–1167, 1991.

[100] H. Min, R. Pates, and E. Mallada, "Coherence and concentration in tightly-connected networks," *arXiv preprint arXiv:2101.00981*, 2021.

[101] K. Ahn, K. Lee, and C. Suh, "Hypergraph spectral clustering in the weighted stochastic block model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 959–974, 2018.

[102] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, and C. E. Priebe, "Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding," *Electronic journal of statistics*, vol. 8, no. 2, pp. 2905–2922, 2014.

[103] H. Min and E. Mallada, "Spectral clustering and model reduction for weakly-connected coherent network systems," *arXiv preprint arXiv:2209.13701*, 2022.

[104] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the davis–kahan theorem for statisticians," 2014.

[105] F. R. Bach and M. I. Jordan, "Learning spectral clustering," in *Advances in Neural Information Processing Systems*, 2004, pp. 305–312.

[106] H. Khalil, "Nonlinear systems, printice-hall," *Upper Saddle River, NJ*, vol. 3, 1996.

[107] J. Schiffer, R. Ortega, A. Astolfi, J. Raisch, and T. Sezi, “Conditions for stability of droop-controlled inverter-based microgrids,” *Automatica*, vol. 50, no. 10, pp. 2457–2469, 2014.

[108] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.

[109] S. W. Kim, B. D. Anderson, and A. G. Madievski, “Error bound for transfer function order reduction using freqeuncy weighted balanced truncation,” *Systems & Control Letters*, vol. 24, no. 3, pp. 183 – 192, 1995.