ABSTRACT

Title of dissertation: STOCHASTIC PROCESSES ON GRAPHS: LEARNING REPRESENTATIONS AND APPLICATIONS

Addison W. Bohannon
Doctor of Philosophy, 2019

Dissertation directed by: Professor Radu V. Balan
Department of Mathematics

In this work, we are motivated by discriminating multivariate time-series with an underlying graph topology. Graph signal processing has developed various tools for the analysis of scalar signals on graphs. Here, we extend the existing techniques to design filters for multivariate time-series that have non-trivial spatiotemporal graph topologies. We show that such a filtering approach can discriminate signals that cannot otherwise be discriminated by competing approaches. Then, we consider how to identify spatiotemporal graph topology from signal observations. Specifically, we consider a generative model that yields a bilinear inverse problem with an observation-dependent left multiplication. We propose two algorithms for solving the inverse problem and provide probabilistic guarantees on recovery. We apply the technique to identify spatiotemporal graph components in electroencephalogram (EEG) recordings. The identified components are shown to discriminate between various cognitive task conditions in the data.

STOCHASTIC PROCESSES ON GRAPHS: LEARNING REPRESENTATIONS AND APPLICATIONS

by

Addison W. Bohannon

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Radu V. Balan, Chair/Advisor
Professor John J. Benedetto
Professor Ramalignan Chellappa
Professor Wojciech Czaja
Dr. Tom Goldstein

Dedication

To my wife, Gina.

Acknowledgments

I owe a debt of gratitude to many people who have supported, encouraged, and helped me along my path.

I would like to thank my advisor, Professor Radu Balan. It has been a pleasure to learn from him over the last four years, as he has modeled selfless mentorship, mathematical rigor, and scientific curiosity for me. It is a testament to him that I am here today. He is exceptionally generous with his time and great mathematical talent. He has patiently taught me nearly all I know of mathematics—how to learn it and how to practice it. As I encountered the innumerable setbacks of PhD research, our weekly meetings made that path less isolated and the setbacks less significant. I can only hope to credit him and his investment in me by striving to exhibit that same selflessness, rigor, and curiosity in all of my future endeavors.

I would like to thank Professors John Benedetto, Wojciech Czaja, and Rama Chellappa, and Dr. Tom Goldstein for agreeing to serve on my committee. I selected them because of the significant influence they had on my research through coursework, informal mentorship, and the community that they have established at the University of Maryland.

I would like to thank the many people from the AMSC program who have helped me along the way from applying for the program to graduating from the program: Professors Konstantina Trivisa, Howard Elman, and Dave Levermore, and Alverda McCoy, Jessica Sadler, and Christiana Sabett. I would also like to thank Cara Peters, Danielle Middlebrooks, David Russell, and Tengfei Su for their

friendship and support, and Chae Clark, Matt Guay, James Murphy, Dongmian Zou, and Weilin Li for advice and perspective.

I would like to thank the many people from ARL who are my friends, collaborators, and mentors. I would like to thank Brent Lance for seeing enough potential in me to support me finishing my PhD and also for helping me to understand the bigger picture of our research. I would like to thank Brian Sadler not only for mentorship, but also for inspiring me to pursue a career in science for the Army, connecting the work that we do to my own experience. I would like to thank Nick Waytowich and Vernon Lawhern first and foremost for their friendship, but also for teaching me principles of good scientific computing and how to solve problems on my own. I would like to thank Jean Vettel for her mentorship, both scientific and personal. I would like to thank Javier Garcia and Sean Fitzhugh for patiently teaching me the necessary science to apply mathematics to new problems. I would like to thank Jeremy Gaston, Arwen DeCostanza, Bill Evans, Amar Marathe, Piotr Franaszczuk, Kaleb McDowell, and Sandy Howard for supporting me in completing my degree. I would like to especially thank Marilynn Peterson and Ellen Wiley for the myriad of support that makes it possible to do research. I would also like to thank Heather Roy, Nina Lauharatanahirun, Derek Spangler, Steven Gutstein, Greg Lieberman, Amelia Solon, Ethan Stump, Jon Fink, and Alec Koppel for friendship, encouragement, and for always introducing me to new scientific ideas.

I would like to thank my family for teaching me the value of hard work and sacrifice, believing in me as I left my career to go back to school, and providing me with love and support throughout.

Finally, I would like to thank my wife, Gina Rotondo Bohannon, for loving, encouraging, and inspiring me. I could not imagine doing any of this without such a best friend and partner.

All experiments in this dissertation were conducted in Python using SciPy libraries [45]. All figures were generated using MatPlotLib [56].

Table of Contents

List of Figures

Notation

Object	Notation
vector	$\mathbf{x}$
matrix/operator	$\mathbf{A}$
matrix/operator conjugate transpose	$\mathbf{A}^*$
k singular value of matrix	$\sigma_k(\mathbf{A})$
k eigenvalue of a matrix	$\lambda_k(\mathbf{A})$
spectrum of a matrix/operator	$\Lambda(\mathbf{A})$
resolvent of a matrix/operator	$\mathcal{R}_{\mathbf{A}}(\cdot)$
set of bounded matrices/operators on $\mathcal{X}$	$\mathcal{B}(\mathcal{X})$
column of matrix	$\mathbf{A}_j$
entry of matrix	$A_{i,j}$
sequence	$(x[t])_{t\in\mathbb{Z}}$
function	$f(\cdot)$
norm	$\|\cdot\|_{\mathcal{X}}$
inner product	$\langle\cdot,\cdot\rangle_{\mathcal{H}}$
number class	$\mathbb{Z}$
set	$\mathcal{X}$
support	$\operatorname{supp}(f)$
Fourier transform	$\hat{f}$
inverse Fourier transform	$\check{f}$
asymptotic lower bound	$\Omega(\cdot)$
asymptotically equivalent	$\mathcal{O}(\cdot)$
asymptotically dominated	$o(\cdot)$
holomorphic functions on U	$\mathcal{H}(U)$
probability	$\mathbb{P}$
expectation	$\mathbb{E}$

Chapter 1: Introduction

We are interested in discriminative representations of multivariate time-series, $(x[t])_t$, where $x[t] \in \mathbb{R}^d$ (or $\mathbb{C}^d$). We are particularly motivated by applications of inferring functional connectivity in neuroimaging data and predicting teamwork processes in team psychology. In each of these applications, an underlying structure exists in the data. For network neuroscience, well-characterized networks of functional regions of the brain coordinate to achieve high-level cognitive processing. In teams, members with unique functional roles and relationships coordinate their activities to achieve a common goal. We want to leverage this underlying structure to make our inference task easier. Knowing how functional regions of the brain interact provides additional information about how to filter observed data. Similarly, knowing the personal relationships of team members or how distinct functional roles must interact to accomplish a task provides important information that we can use to filter observed data.

Building on the work of Mallat [75] and Bruna and Mallat [24], Bruna, *et al.* proposed a deep learning architecture that could encode domain-specific information in the form of a graph [25]. Deep learning, and specifically convolutional neural networks, hierarchically build rich representations of data by composing convolutional filters and element-wise nonlinearities [68]. The unrivaled success of deep

learning in speech, image, and video application domains can be attributed to a nonlinear filtering protocol which uniquely complements the relevant symmetries of these tasks [76]. How then do we generalize the phenomenon of deep learning's success to other application domains with different symmetries? It is this question to which Bruna, *et al.* propose graph convolutional neural networks [25]. Graph convolutional neural networks encode the statistical symmetries of functions with discrete domains by replacing traditional convolutional filters with functions of a given graph operator such as the weighted adjacency or graph Laplacian matrix. This has led to considerable follow-on work, *e.g.* [61, 62, 43, 38, 84]. For a recent review of deep learning on graphs, see Bronstein, *et al.* [22].

In parallel to the development of graph convolutional neural networks, signal processing researchers developed a conceptual framework for processing data on networks. This framework and collection of tools is collectively known as graph signal processing, and it combined with efforts toward generalizing convolutional neural networks to graphs to make a more complete body of data science techniques and theory. Graph signal processing began with the papers of Shuman, *et al.* [99] and Sandryhaila and Moura [90, 92]. These seminal works proposed a generalization of classical signal processing to data on graphs. The initial generalization included graph frequencies, graph Fourier transforms, graph filtering, and graph wavelets. Later work generalized a theory of graph sampling [32] and the uncertainty principle [109]. For a recent review of this rapidly growing field, see Ortega, *et al.* [81].

In the following chapters, we aim to contribute new theory and techniques for graph convolutional networks and graph signal processing of multivariate time-

series. In Chapter 2, we provide the relevant background for subsequent chapters. In Chapter 3, we propose a filtering framework which can encode statistical symmetries via an extended graph and associated graph operators. By using holomorphic functional calculus, we can realize a large class of linear filters with relatively few degrees of freedom. We show that our proposed approach provides a richer and more discriminable model than alternatives. In Chapter 4, we turn our attention to learning extended graphs and associated graph operators from observations. We specifically address learning graph operators in the presence of additive linear processes. This generative model leads to a linear mixture model, for which based on dictionary learning results, we propose an alternating minimization algorithm to solve it. We show that under suitable conditions, the algorithm converges linearly to the true solution.

1.1 Functional and effective connectivity in the human brain

The human brain displays both localized and global processing to execute high-level cognitive tasks. Historically, we have better understood the organizational principles that encourage localized processing in the brain, but modern neuroimaging techniques have precipitated greater understanding of how the brain executes global processing. Karl Friston refers to our understanding of these processes as functional segregation and functional integration respectively [46]. From case studies of patients with brain lesions and controlled animal studies, we have long theorized that our brain organizes information in functionally distinct regions. For example,

early visual processing takes place in the occipital lobe of the brain, whereas our sense of touch is processed in the parietal lobe. At this point, it is widely accepted that the human brain employs localized processing as an organizing principle. However, higher-order cognitive processing requires the contribution of many distinct functional units. It has been historically more difficult to identify the mechanisms by which the brain integrates discrete functional units to achieve higher-order processing. Modern neuroimaging modalities such as functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG) now provide neuroscientists with tools to observe brain activity non-invasively and in a controlled manner, facilitating hypothesis-driven experimentation.

Understanding how functionally segregated regions of the brain organize to yield high-order processing could manifest as either descriptions or generative models. Friston refers to the former as functional connectivity, *e.g.* reporting of observed dependencies of distinct regions of the brain during a cognitive task [46]. Functional connectivity can be measured via correlations in recorded data. If region $\mathcal{X}$ and $\mathcal{Y}$ must be functionally integrated for a cognitive task, then the observed correlation of their data will likely exceed some hypothesis-testing threshold. Such descriptive statistics of functional integration can be used to understand the global brain network via graph-theoretic techniques [26]. Alternatively, neuroscientists can propose a generative model for how functionally segregated regions of the brain integrate. The model can then be fit to observations of brain activity and explained variance indicates the likelihood that the generative model describes the underlying mechanism. Friston refers to this latter approach as effective connectivity [46]. Typical

techniques within this approach are dynamic causal modeling and autoregressive modeling.

fMRI and EEG modalities differ in their temporal and spatial resolution as well as the underlying physical phenomenon being measured. fMRI can reliably provide spatial resolution on the millimeter scale, while the methods for spatial localization of electrical activity in EEG yield far less reliability [26]. However, EEG provides far greater temporal resolution (1-100 Hz), while fMRI requires a long measurement cycle, resulting in sampling rates of 0.001-0.5 Hz [26]. Additionally, fMRI and EEG measure different signatures of neuronal activation. fMRI measures the blood oxygen level-dependent (BOLD) signal, essentially changes in magnetic susceptibility and tissue contrast [26]. EEG measures electrical activity at the scalp. Researchers choose the appropriate neuroimaging modality for their specific goals. Of note, graph signal processing has been applied to fMRI data in Medaglia, *et al.* [78] and Huang, *et al.* [55].

In either case, the observations of brain activity from fMRI and EEG comprises discretely sampled multivariate time-series, $(\mathbf{x}[t])_t$. The dimension d of the vector-valued samples $x[t] \in \mathbb{R}^d$ corresponds to either the number of sensors, sources, or aggregated regions based on an appropriate brain atlas. Representative tasks in which we are interested are predicting the cognitive state $y \in \mathcal{Y}$ from a complete observation $(\mathbf{x}[t])_t$, or predicting the following observation $\mathbf{x}[T+1]$ from a partial observation $(\mathbf{x}[t])_{t \leq T}$. Regardless of the task, the observations are very high-dimensional and require statistical techniques to account for this challenge [110]. Beyond the scientific goal of understanding the function of the human brain, the

identification of brain state via neuroimaging provides a mechanism to facilitate integration of humans and intelligent systems [67, 112, 16, 95]

1.2 Promoting teamwork with a systems approach

In a recent paper [37], DeCostanza, *et al.* propose a vision for adaptive technologies which promote effective teamwork in teams of humans and intelligent agents. By teams, we mean a collection of individuals working toward a common goal. It is well-understood that team performance depends on more than the individual capability of the constituent members [89, 34]. Mechanisms for dynamically organizing effort and competencies within the team underlie effective team performance [77, 88, 57, 27, 42, 87, 65]. Premised on this, DeCostanza, *et al.* argue that the science and technology exist today to design system-level approaches to target effective teamwork processes through the use of intelligent technology, *i.e.* artificial intelligence.

We can conceptualize this vision in a dynamical systems model. We consider the instantaneous team state $\mathbf{x}[t] \in \mathbb{R}^{d_1}$. The team state can represent dynamic properties of the team such as affect, effective communication, or shared understanding, which evolve as a function of context, history, and goals. As these states are not directly observable, our measurements of these states are filtered $\mathbf{y}[t] = g(\mathbf{x}[t]) \in \mathbb{R}^{d_2}$, where $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is a possibly nonlinear measurement operator. Adaptive technologies then serve as a control input to the dynamical system, *i.e.* we can design inputs $\mathbf{u}[t] \in \mathbb{R}^{d_3}$. This yields the following dynamical system

model [18]:

$$\begin{aligned} \mathbf{x}[t+1] &= f\left(\mathbf{x}[t], \ldots, \mathbf{x}[0], \mathbf{u}[t], \ldots, \mathbf{u}[0]\right) + \mathbf{n}_s[t] \\ \mathbf{y}[t] &= g(\mathbf{x}[t]) + \mathbf{n}_o[t] \end{aligned} \tag{1.1}$$

where f maps sequences of states and control inputs to a future state. Here, $\mathbf{n}_s[t] \in \mathbb{R}^{d_1}$ and $\mathbf{n}_o[t] \in \mathbb{R}^{d_2}$ are additive noise in the state and observation respectively. We can consider a linearization of this model, *i.e.* $\mathbf{A} = \nabla_{\mathbf{x}[t]} f$ and $\mathbf{B} = \nabla_{\mathbf{u}[t]} f$, which yields the following simplified state-space equation model:

$$\begin{aligned} \mathbf{x}[t+1] &= \mathbf{A}\mathbf{x}[t] + \mathbf{B}\mathbf{u}[t] + \mathbf{n}_s[t] \\ \mathbf{y}[t] &= g(\mathbf{x}[t]) + \mathbf{n}_o[t]. \end{aligned} \tag{1.2}$$

The proposed dynamical system model illuminates the scientific and technical challenges to realizing the vision of DeCostanza, *et al.* . We must understand how states evolve $\mathbf{A} : \mathbb{R}^{d_1} \to \mathbb{R}^{d_1}$, how control inputs manifest in the team state $\mathbf{B} : \mathbb{R}^{d_2} \to \mathbb{R}^{d_1}$, and how to infer states from observations $g^{-1} : \mathbb{R}^{d_3} \to \mathbb{R}^{d_1}$. This invites a host of modeling and inference problems for multivariate time-series, *e.g.* :

(1) Given state sequences $(\mathbf{x}[t])_{t \leq T}$, predict $\mathbf{x}[T+1]$;

(2) Given observations $(\mathbf{y}[t])_t$, infer states $(\mathbf{x}[t])_t$; and

(3) Given observations $(\mathbf{y}[t])_t$, design controls $(\mathbf{u}[t])_t$.

As in Sec. 1.1, teams may exhibit both functional segregation and integration in the performance of tasks. Implicit in the dynamical system model is the emergence of team states from individual states. This phenomena is perhaps most clear when we consider the observation function $g : \mathbb{R}^{d_1} \to \mathbb{R}^{d_3}$ and the control input function

$\mathbf{B} : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_1}$. DeCostanza, *et al.* propose wearable sensors as a viable technology for the non-obtrusive, continuous monitoring of teams [37]. This implies that d_3 scales with the size of the team. We can monitor and infer individual states, and use individual states to in turn predict team states. Similarly, candidate control inputs are envisioned at the individual team member level to facilitate personalization. This implies that d_2 also scales with the size of the team. Again, we can attempt to manipulate individual state in a deliberate way, and in turn shape the team state. Hence, we must understand the relationship between structures within the team to bridge the gap between individual states and emergent team states. Like functional connectivity in the brain, these structures can be modeled via graphical methods [60].

Chapter 2: Background

2.1 Mathematical preliminaries

We attempt to recount in this chapter the necessary mathematical background for the remainder of the chapters. We assume a basic understanding of analysis [101, 23], random variables [30, 52], and linear algebra [106, 48]. In the following, we will refresh some useful definitions and results specific to this work and quickly relate them to multivariate time-series. In any measure-theoretic statement, the measure is assumed to be the Lebesgue measure.

Definition 1. A *Banach space* $\mathcal{X}$ is a complete, normed vector space, $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$.

Definition 2. A *Hilbert space* $\mathcal{H}$ is a Banach space for which the norm arises from an inner product, $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, *i.e.* for any $h \in \mathcal{H}$,

$$\|h\|_{\mathcal{H}} = \sqrt{\langle h, h \rangle_{\mathcal{H}}}. \tag{2.1}$$

We will be primarily concerned with the following vector-valued sequence space,

$$\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right) = \left\{ f : \mathbb{Z} \to \mathbb{C}^d : \sum_{t \in \mathbb{Z}} \|f[t]\|^2 < \infty \right\}, \tag{2.2}$$

where the vector norm is the usual Euclidean norm on $\mathbb{C}^d$. As the Euclidean norm on $\mathbb{C}^d$ arises from an inner product, $\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$ is a Hilbert space with inner product

for any $\mathbf{x} = (\mathbf{x}[t])_{t\in\mathbb{Z}}, \mathbf{y} = (\mathbf{y}[t])_{t\in\mathbb{Z}} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{t\in\mathbb{Z}} \langle \mathbf{x}[t], \mathbf{y}[t]\rangle . \tag{2.3}$$

We will also use various vector-valued measurable spaces,

$$L^p\left([0,1];\mathbb{C}^d\right) = \left\{ f : [0,1] \to \mathbb{C}^d : \int_0^1 \|f(\omega)\|^p \, d\omega < \infty \right\}, \tag{2.4}$$

where $1 \leq p \leq \infty$ and again the norm arises from the Euclidean norm on $\mathbb{C}^d$.

Definition 3. The *Fourier transform* on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ is defined for any $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ by the map

$$\mathbf{x} \mapsto \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{x}[t]. \tag{2.5}$$

Definition 4. The *Fourier transform* on $L^2\left([0,1];\mathbb{C}^d\right)$ is defined for any $\mathbf{x} \in L^2\left([0,1];\mathbb{C}^d\right)$ by the map

$$\mathbf{x} \mapsto \left(\int_0^1 e^{-2\pi i\omega t}\mathbf{x}(\omega) d\omega \right)_{t\in\mathbb{Z}}. \tag{2.6}$$

Remark. The Fourier transform from $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right) \to L^2\left([0,1];\mathbb{C}^d\right)$ is bijective and unitary.

We will use the following notation to denote the respective Fourier transforms: $\hat{\mathbf{x}}(\omega) = \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{x}[t]$ and $\check{\mathbf{x}}[t] = \int_0^1 e^{-2\pi i\omega t}\mathbf{x}(\omega)d\omega$.

We use the following definition of a random variable.

Definition 5. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a *random variable* is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$.

This definition can be extended to random variables which take values in a Banach space $\mathcal{X}$. Now, we briefly introduce matrix norms and review the spectral theory of finite-dimensional linear operators.

Definition 6. The $p \to q$-*norm* of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is

$$\|\mathbf{A}\|_{p \to q} := \sup_{\mathbf{x} \in \mathbb{C}^n} \frac{\|\mathbf{A}\mathbf{x}\|_q}{\|\mathbf{x}\|_p}. \tag{2.7}$$

The usual spectral norm corresponds to the $2 \to 2$-norm. We omit the explicit notation $2 \to 2$ when it is clear from context. We also make use of mixed norms.

Definition 7. The p, q-*norm* of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$ is

$$\|\mathbf{A}\|_{p,q} := \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{m} |A_{j,i}|^p \right)^{q/p} \right)^{1/q}. \tag{2.8}$$

The Frobenius norm corresponds to the $2, 2$-norm, and we use the following notation: $\|\cdot\|_F = \|\cdot\|_{2,2}$.

Definition 8. $\mathbf{A} \in \mathbb{R}^{n \times m}$ satisfies a *s-restricted eigenvalue* condition if

$$\kappa_s := \min_{\substack{J \subset \{1,\ldots,m\} \\ |J| \leq s}} \min_{\substack{\mathbf{h} \neq \mathbf{0} \in \mathbb{R}^m \\ \|\mathbf{h}_{J^C}\|_1 \leq 3\|\mathbf{h}_J\|_1}} \frac{\|\mathbf{A}\mathbf{h}\|_2}{\sqrt{n}\,\|\mathbf{h}_J\|_2} > 0. \tag{2.9}$$

We make repeated use of the following result:

Theorem 2.1 (Jordan spectral representation)**.** *Let* $\mathbf{A} \in \mathcal{B}\left(\mathbb{C}^d\right)$*. Then, there exists* $m \leq d$ *distinct eigenvalues* $(\lambda_k \in \mathbb{C})_{k=1,\ldots,m}$*, projections* $\left(\mathbf{P}_k \in \mathcal{B}\left(\mathbb{C}^d\right)\right)_{k=1,\ldots,m}$*, and nilpotents* $\left(\mathbf{N}_k \in \mathcal{B}\left(\mathbb{C}^d\right)\right)_{k=1,\ldots,m}$ *such that*

$$\mathbf{A} = \sum_{k=1}^{m} \lambda_k \mathbf{P}_k + \mathbf{N}_k \tag{2.10}$$

with the following properties:

1. $\mathbf{P}_j \mathbf{P}_k = \mathbf{P}_k \mathbf{P}_j = \delta_{j,k} \mathbf{P}_k$*;*

2. $\mathbf{P}_k \mathbf{N}_k \mathbf{P}_k = \mathbf{N}_k$*;*

3. $(\mathbf{N}_k)^d = 0$; *and*

4. $\sum_{k=1}^{m} \mathbf{P}_k = \mathbf{I}$.

Properties 1–4 of Theorem 2.1 imply additionally that

$$\mathbf{P}_k \mathbf{N}_k = \mathbf{N}_k \mathbf{P}_k = \mathbf{N}_k \tag{2.11}$$

and

$$\mathbf{P}_j \mathbf{N}_k = \mathbf{N}_k \mathbf{P}_j = \delta_{j,k} \mathbf{N}_k. \tag{2.12}$$

For a proof of Thm. 2.1, see *e.g.* Kato [59].

2.2 Operator Theory

In this section, we provide a brief introduction to operator theory [86, 100]. Then, we introduce the primary two operator theory results that we will use throughout Chapter 3: the spectral theorem for Laurent operators and the holomorphic functional calculus. We begin with basic definitions.

Definition 9. A *bounded linear operator* $\mathbf{A} : \mathcal{X} \to \mathcal{Y}$ between two Banach spaces is a linear map for which the operator norm,

$$\|\mathbf{A}\|_{\mathcal{B}(\mathcal{X},\mathcal{Y})} = \sup_{\mathbf{x} \in \mathcal{X}} \frac{\|\mathbf{A}\mathbf{x}\|_{\mathcal{Y}}}{\|\mathbf{x}\|_{\mathcal{X}}}, \tag{2.13}$$

is finite.

Clearly, the operator norm generalizes the $p \to q$-norm of Sec. 2.1. The set of bounded linear operators, $\mathcal{B}(\mathcal{X}, \mathcal{Y})$, is itself a Banach space, and $\mathcal{B}(\mathcal{X}) = \mathcal{B}(\mathcal{X}, \mathcal{X})$

is a Banach algebra with identity. The following results relate to $\mathcal{B}(\mathcal{X})$ for any Banach space $\mathcal{X}$.

Definition 10. The *spectrum* of a bounded linear operator $\mathbf{A} \in \mathcal{B}(\mathcal{X})$ comprises all elements $\lambda \in \mathbb{C}$ for which $(\mathbf{A} - \lambda\mathbf{I})^{-1} \notin \mathcal{B}(\mathcal{X})$, and it is denoted $\Lambda(\mathbf{A})$.

The spectrum again generalizes its finite-dimensional counterpart, eigenvalues, *i.e.* $\lambda \in \mathbb{C}$ such that $(\lambda\mathbf{I} - \mathbf{A}) \notin \mathcal{B}\left(\mathbb{C}^d\right)$ for $\mathbf{A} \in \mathcal{B}\left(\mathbb{C}^d\right)$.

Definition 11. The *resolvent set* of a bounded linear operator $\mathbf{A} \in \mathcal{B}(\mathcal{X})$ is the complement of the spectrum, $\mathbb{C} \setminus \Lambda(\mathbf{A})$.

Definition 12. The *resolvent* of a bounded linear operator $\mathbf{A} \in \mathcal{B}(\mathcal{X})$ is the operator-valued function, $\mathbf{R}_{\mathbf{A}} : \mathbb{C} \setminus \Lambda(\mathbf{A}) \to \mathcal{B}(\mathcal{X})$,

$$z \mapsto (\mathbf{A} - \lambda\mathbf{I})^{-1}. \tag{2.14}$$

Remark. The spectrum of a bounded linear operator on a Banach space is nonempty, closed, and bounded [35].

Now, we recall a definition from complex analysis.

Definition 13. A complex-valued function $f : U \to \mathbb{C}$ defined on an open set $U \subset \mathbb{C}$ is said to be *holomorphic* if it has a well-defined derivative at each point in U.

We can also characterize a holomorphic function $f : U \to \mathbb{C}$ as having a convergent power series on an open disc with a positive radius, *i.e.* for every $z_0 \in U$, we can write $f(z) = \sum_{n=0}^{\infty} c_n (z - z_0)^n$ for some $(c_n)_{n \geq 0}$ and $|z - z_0| < r$, $r > 0$.

Theorem 2.2 (Holomorphic functional calculus)**.** *For a Banach space* $\mathcal{X}$*, let* $\mathbf{S} \in \mathcal{B}(\mathcal{X})$*,* $U \subset \mathbb{C}$ *be an open set such that* $\Lambda(\mathbf{S}) \subset U$*,* $\phi : U \to \mathbb{C}$ *be holomorphic, and* $\Gamma \subset int(U)$ *be a closed curve enclosing* $\Lambda(\mathbf{S})$*. Then, we define* $\phi(\mathbf{S}) \in \mathcal{B}\left(\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)\right)$*,*

$$\phi(\mathbf{S}) := \frac{1}{2\pi i} \oint_{\Gamma} \phi(z) \mathcal{R}_{\mathbf{S}}(z)\, dz. \tag{2.15}$$

Moreover, $\Lambda(\phi(\mathbf{S})) = \{\phi(\lambda) : \lambda \in \Lambda(\mathbf{S})\}$*,* $\phi \mapsto \phi(\mathbf{S})$ *is a continuous map from* $\sup_{\gamma \in \Gamma} |\phi(\gamma)|$ *to* $\|\cdot\|_{\mathcal{B}(\mathcal{X})}$*, and if* $\psi : \mathbb{C} \to \mathbb{C}$ *is holomorphic on* U*, then* $\phi(\mathbf{S})\psi(\mathbf{S}) = (\phi \cdot \psi)(\mathbf{S})$*.*

For a proof of Thm. 2.2, see *e.g.* Davies [35] or Simon [100]. We note the considerable enhancement of holomorphic functional calculus over the polynomial calculus. The holomorphic functional calculus includes all polynomials. In addition, we can consider functions which have poles outside of U. Even more powerful, if $\Lambda(\mathbf{S})$ can be separated, then we can define functions which manifest different behavior on the restriction to each separable component of the spectrum. This allows us to define projections onto connected components of the spectrum, and we use this property in Sec. 3.4 to define bandpass filters.

We can combine Thm. 2.1 with Thm. 2.2 for the following useful result. For $\mathbf{A} \in \mathcal{B}\left(\mathbb{C}^d\right)$ with Jordan spectral representation,

$$\mathbf{A} = \sum_{i=1}^{d'} \lambda_i \mathbf{P}_i + \mathbf{N}_i,$$

where $d' \leq d$, we can define an open set $U \subset \mathbb{C}$ such that $\Lambda(\mathbf{A}) \subset U$ and a holomorphic function $\phi : U \to \mathbb{C}$. Then, the holomorphic functional calculus has

the following spectral representation:

$$\phi(\mathbf{A}) = \sum_{i=1}^{d'} \phi(\lambda_i)\mathbf{P}_i + \phi'(\lambda_i)\mathbf{N}_i. \tag{2.16}$$

We will use Laurent operators extensively in Chapter 3, and we define them here.

Definition 14. A bounded operator $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ is said to be *Laurent* if there exists a matrix symbol $(\mathbf{K}[t])_{t\in\mathbb{Z}}$ such that for every $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$(\mathbf{A}\mathbf{x})[t] = \sum_{s\in\mathbb{Z}} \mathbf{K}[t-s]\mathbf{x}[s]. \tag{2.17}$$

The following spectral theorem for Laurent operators characterizes the admissible matrix symbols and spectrum of Laurent operators.

Theorem 2.3 (Spectral theorem for Laurent operators)**.** *Let* $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ *be Laurent with matrix symbol* $(\mathbf{K}[t])_{t\in\mathbb{Z}}$ *satisfying* $\sum_{t\in\mathbb{Z}} \|\mathbf{K}[t]\| < \infty$*. Then, for every* $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$\sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}(\mathbf{A}\mathbf{x})[t] = \hat{\mathbf{A}}(\omega)\cdot\hat{\mathbf{x}}(\omega), \tag{2.18}$$

where

$$\hat{\mathbf{A}}(\omega) = \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{K}[t]. \tag{2.19}$$

Additionally, $\Lambda(\mathbf{A}) = \cup_{\omega\in[0,1]}\Lambda\left(\hat{\mathbf{A}}(\omega)\right)$ *and*

$$\|\mathbf{A}\| = \max_{\omega\in[0,1]} \sigma_{\max}\left(\hat{\mathbf{A}}(\omega)\right). \tag{2.20}$$

For a proof of Thm. 2.3, see *e.g.* [21].

2.3 Matrix concentration inequalities

Scalar and vector concentration inequalities play an important role in learning theory and applied probability. Matrix concentration inequalities attempt to generalize the scalar counterparts for non-commutative algebras such as matrices. We briefly review two methods for concentrating the eigenvalues or singular values of a random matrix. One approach builds from the matrix Laplace transform method [6, 79]. The other approach leverages scalar concentration inequalities together with covering arguments, *e.g.* Vershynin [111]. We make extensive use of both in the technical arguments of Chapter 4.

2.3.1 Matrix Laplace transform methods

We follow the exposition of Tropp [108], in which proofs for all of the following results can be found.

Theorem 2.4 (Matrix Laplace transform method)**.** *Let* $\mathbf{Y} \in \mathcal{B}\left(\mathbb{C}^d\right)$ *be a random self-adjoint matrix. Then, for all* $t \in \mathbb{R}$*,*

$$\mathbb{P}\left(\{\lambda_{\max}\left(\mathbf{Y}\right) \geq t\}\right) \leq \inf_{\theta > 0} e^{-\theta \cdot t} \cdot \mathbb{E} tr \exp\left(\theta \mathbf{Y}\right). \tag{2.21}$$

This yields a bound for the largest eigenvalue of a random matrix in terms of the expectation of the matrix moment generating function. Extending this result to sums of random matrices proves impossible since the matrix exponential

does not commute in general. However, the matrix cumulant generating function is subadditive, and this yields the so-called master tail bound.

Theorem 2.5 (Master tail bound, Tropp [108], Theorem 3.6)**.** *Consider a finite sequence of independent, random, self-adjoint matrices* $(\mathbf{X}_k)_k$*. For all* $t \in \mathbb{R}$*,*

$$\mathbb{P}\left(\left\{\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \geq t\right\}\right) \leq \inf_{\theta>0} e^{-\theta\cdot t} \cdot tr\exp\left(\sum_k \log \mathbb{E}e^{\theta \mathbf{X}_k}\right). \tag{2.22}$$

We do not use the master tail bound directly. We instead use a corollary, the matrix Chernoff inequality.

Corollary 2.5.1 (Matrix Chernoff, Tropp [108], Corollary 5.2)**.** *Consider a finite sequence of independent, random, self-adjoint* d*-dimensional matrices* $(\mathbf{X}_k)_{k=1,\ldots,N}$ *that satisfy* $\mathbf{X}_k \succeq \mathbf{0}$ *and* $\lambda_{\max}(\mathbf{X}_k) \leq R$ *almost surely. Define*

$$\mu_{\min} := \lambda_{\min}\left(\sum_{k=1}^N \mathbb{E}\mathbf{X}_k\right) \quad \textit{and} \quad \mu_{\max} := \lambda_{\max}\left(\sum_{k=1}^N \mathbb{E}\mathbf{X}_k\right).$$

Then, for $\delta \in [0,1]$*,*

$$\mathbb{P}\left(\left\{\lambda_{\min}\left(\sum_{k=1}^N \mathbf{X}_k\right) \leq (1-\delta)\mu_{\min}\right\}\right) \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}, \tag{2.23}$$

and for $\delta > 0$*,*

$$\mathbb{P}\left(\left\{\lambda_{\max}\left(\sum_{k=1}^N \mathbf{X}_k\right) \geq (1+\delta)\mu_{\max}\right\}\right) \leq d \cdot \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/R}. \tag{2.24}$$

The matrix Chernoff inequality provides lower bounds on the smallest eigenvalue of a sum of random bounded matrices. This result will help us prove that a particular operator is full-rank with high probability in Chapter 4.

2.3.2 Covering argument methods

Covering argument methods have become the go-to method for bounding the singular values of random matrices for compressive sensing and machine learning applications. We follow the exposition of Vershynin [111], specifically that of random matrices with independent rows. The main result is given below.

Theorem 2.6 (Sub-Gaussian rows)**.** *Let $\mathbf{A} \in \mathbb{R}^{N \times n}$ be a random matrix whose rows are independent sub-Gaussian isotropic random vectors in $\mathbf{R}^n$. Then, for every $t \geq 0$, with probability at least $1 - 2e^{-ct^2}$, one has*

$$\sqrt{N} - C\sqrt{n} - t \leq \sigma_{\min}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{A}) \leq \sqrt{N} + C\sqrt{n} + t, \tag{2.25}$$

where C and c depend only on the maximum sub-Gaussian norm of any row of $\mathbf{A}$.

We do not use this result explicitly, but rather follow the structure of its proof. First, we convert the problem to bounding $\mathbf{A}^*\mathbf{A} - \mathbb{E}\mathbf{A}^*\mathbf{A}$. Then, we introduce a covering of the unit sphere, $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$. For every $\mathbf{x}$ in the net, we concentrate $|\langle \mathbf{x}, (\mathbf{A}^*\mathbf{A} - \mathbb{E}\mathbf{A}^*\mathbf{A})\,\mathbf{x}\rangle|$ using scalar concentration inequalities. Then, we take a union bound over all $\mathbf{x}$ in the net. The following lemmas are useful.

Lemma 2.7. *An ε-covering of the unit sphere in $\mathbf{R}^n$ has cardinality less than or equal to $(1 + 2/\varepsilon)^n$.*

Lemma 2.8. *Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix and consider an ε-covering of the unit sphere in $\mathbb{R}^n$. Then,*

$$\|\mathbf{A}\| \leq (1 - 2\varepsilon)^{-1} \max_{x \in \mathcal{N}_\varepsilon} |\langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle|. \tag{2.26}$$

Lastly, we make use of the Hanson-Wright inequality for a concentration argument in Chapter 4. We use the result of Rudelson and Vershynin [85].

Theorem 2.9 (Hanson-Wright inequality)**.** *Let $Z \in \mathbb{R}^n$ be a random vector with independent, centered components, and sub-Gaussian norm K. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(\{|Z^* \mathbf{A} Z - \mathbb{E} Z^* \mathbf{A} Z| > t\}\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4 \|\mathbf{A}\|_F^2}, \frac{t}{K^2 \|\mathbf{A}\|}\right)\right), \quad (2.27)$$

where c is a global constant.

2.4 Graph signal processing

2.4.1 Graph Theory

In this section, we recount now classical results in spectral graph theory [33, 103].

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with nodes $\mathcal{V} = \{1, \ldots, d\}$ such that $d = |\mathcal{V}| < \infty$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. A weight function assigns a relationship between any two nodes with an edge connecting them $w : \mathcal{E} \to \mathbb{R}$. The function, w, defines the entries of the adjacency, or weighted adjacency matrix, $\mathbf{S}$ ($[\mathbf{S}]_{j,k} = w(j,k)$). If w is symmetric (*i.e.* $w(j,k) = w(k,j)$), then the graph is undirected, otherwise it is directed. The degree of each node, $\sum_{j \in \mathcal{V}} w(i,j)$, and the degree matrix, $\mathbf{D} = \operatorname{diag}\left(\sum_{j \in \mathcal{V}} w(1,j), \ldots, \sum_{j \in \mathcal{V}} w(d,j)\right)$, follow from the definition of $\mathbf{S}$. The Laplacian of $\mathcal{G}$ is $\mathbf{L} = \mathbf{D} - \mathbf{S}$. If the weight matrix is undirected, then the Laplacian is symmetric and non-negative. Other matrices of interest are the normalized

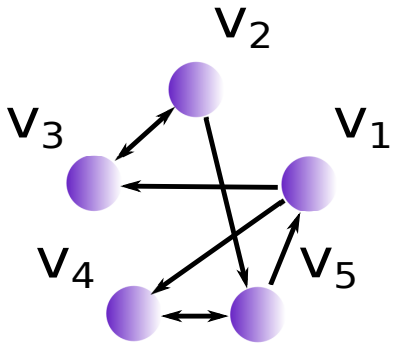


Figure 2.1: Example graph with five nodes. Here, the presence of an edge is depicted with an arrow. The graph has both directed and undirected edges. For example, edge $\mathcal{E}_{2,3}$ is undirected, and edge $\mathcal{E}_{2,5}$ is directed.

Laplacian, $\mathbf{L}_n = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$, and the random walk Laplacian, $\mathbf{L}_r = \mathbf{D}^{-1}\mathbf{L}$.

If $\mathcal{G}$ is undirected and connected, *i.e.* all nodes have at least one edge, then the minimum eigenvalue of $\mathbf{L}$ is 0 with multiplicity one, and it coincides with an eigenvector of $d^{-1/2}\mathbf{1}$. The second smallest eigenvalue is often known as the algebraic connectivity, or Fiedler value, and coincides with the Fiedler vector. The Fiedler vector can be used to solve minimum graph cut and other partitioning problems [98]. More generally, the eigenvectors of $\mathbf{L}$, $\mathbf{L} = \mathbf{U}\Lambda\mathbf{U}^*$ with $\mathbf{U}^*\mathbf{U} = \mathbf{U}\mathbf{U}^* = \mathbf{I}$, can be used for dimensionality reduction and manifold learning as in Laplacian eigenmaps [11].

Another interesting feature of the Laplacian of a graph $\mathcal{G}$ is that it can be used to impart physical intuition for the eigenvectors and eigenvalues. For any $\mathbf{x} \in \mathbb{R}^d$, we define an energy functional based on the weight function w:

$$E(\mathbf{x}) = \sum_{(i,j)\in\mathcal{E}} w_{i,j}x_i x_j = \langle \mathbf{x}, \mathbf{L}\mathbf{x} \rangle . \tag{2.28}$$

The variational characterization of the eigenvalues and eigenvectors of $\mathbf{L}$ coincides

with the energy function E. That is, the eigenvectors of $\mathbf{L}$ are the unique orthonormal set that minimize E, and the eigenvectors are the associated energy.

2.4.2 Graph Signals

We begin with a definition of graph signals. We attempt to present a version agnostic to the directedness of the graph.

Definition 15. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. A function taking values on $\mathcal{V}$, $x : \mathcal{V} \to \mathbb{C}$, is called a *graph signal*.

Graph signals are functions which have a domain with a topology given by a graph. That is, the discrete domain admits possibly nontrivial spatial relationships. We can contrast this with a finite sequence space such as $\ell^2(\mathbb{Z}/d\mathbb{Z})$, in which any two elements of the domain are related only by their distance apart on the number line. Graph signals on the other hand inherit a nontrivial spatial relationships from the edges of the graph, $\mathcal{E}$. The weight function, $w : \mathcal{E} \to \mathbb{R}$, defines a metric on the set $\mathcal{V}$, and in turn, induces a topology.

We can define a function space for graph signals.

Definition 16. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and $x : \mathcal{V} \to \mathbb{C}$, a graph signal. The *p-space of graph signals* is given by

$$\ell^p(\mathcal{V}) \left\{ x : \mathcal{V} \to \mathbb{C} : \sum_{i \in \mathcal{V}} |x_i|^p < \infty \right\}$$

for any $1 \leq p \leq \infty$.

As the underlying set is finite, the p-space of graph signals can be defined

equivalently with respect to any sequence p-norm. That is to say that, graph signals must take finite values on all nodes $\mathcal{V}$.

Remark. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = d$, the p-space of graph signals is isomorphic to $\left(\mathbb{C}^d, \|\cdot\|_p\right)$ for any $1 \leq p \leq \infty$.

Due to the equivalence between graph signals and a finite-dimensional vector, we will primarily identify graph signals with their vector-valued counterparts. Similarly, we can think of weight functions, $w : \mathcal{E} \to \mathbb{R}$, as defining bounded operators on $\mathbb{C}^d$, *i.e.* $d \times d$ matrices.

From this point forward, we will prefer to discuss general graph operators $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$, where $d = |\mathcal{V}|$. This is to allow us to speak about the adjacency matrix, weighted adjacency matrix, graph Laplacian, random walk Laplacian, and normalized Laplacian. When results or claims do not generalize to all, we will specify to which they do. The most important difference between the various graph operators stems from directedness and the symmetry that either does or does not follow from it. The various Laplacians yield symmetric operators on $\mathbb{C}^d$. In general, an adjacency matrix may or may not be defined to be symmetric.

2.4.3 Graph Fourier analysis

Fourier analysis provides a fundamental building block in classical signal processing, and we can generalize this analysis for graphs with self-adjoint graph operators. Other authors have proposed generalizations of graph Fourier analysis for non-self-adjoint graph operators, *e.g.* [92, 97], but here we present only the graph

Fourier analysis developed in Shuman, *et al.* [99]. We begin with a definition of a graph Fourier transform by analogy.

Definition 17. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with associated self-adjoint graph operator $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$. Let $\mathbf{S}$ have eigendecomposition $\mathbf{S} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^*$. Then, for any $\mathbf{x} \in \mathbb{C}^d$, we define the *graph Fourier transform* by the map

$$\mathbf{x} \mapsto \left(\langle \mathbf{x}, \mathbf{u}_i \rangle\right)_{i=1,\ldots,d}. \tag{2.29}$$

Accordingly, we define the *inverse graph Fourier transform* by the map

$$\left(\breve{x}_i\right)_{i=1,\ldots,d} \mapsto \sum_{i=1}^{d} \breve{x}_i \mathbf{u}_i. \tag{2.30}$$

The graph Fourier transform expands a signal in the invariant subspaces of a graph operator $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$. Since $\mathbf{S}$ is positive semi-definite, there exists an orthonormal set of eigenvectors $\left(\mathbf{u}_i\right)_{i=1,\ldots,d}$ that define the invariant subspaces. Moreover, the eigenvalues of $\mathbf{S}$, $\left(\lambda_i\right)_{i=1,\ldots,d}$ are real so that they have a natural ordering: $\lambda_1 \leq \cdots \leq \lambda_d$. We can use this ordering to impart a notion of frequency to the graph spectral domain, *i.e.* the image of the graph Fourier transform. The smallest eigenvalues of $\mathbf{S}$ correspond to "low frequency" eigenvectors. Then, the spectral representation of a graph signal, $\mathbf{x} \in \mathbb{C}^d$, $\left(\langle \mathbf{x}, \mathbf{u}_i \rangle\right)_{i=1,\ldots,d}$ decomposes a signal into components of increasing frequency. Graph signals with primarily "low frequency" spectral representations correspond to smooth, or regular, behaviors on the graph, whereas graph signals with primarily "high frequency" spectral representations correspond to nonsmooth, or irregular, behaviors.

2.4.4 Filtering graph signals

We can apply linear filters to graph signals as in classical signal processing [90, 99, 58, 96]. Here, we want to apply a linear transformation $\mathbb{C}^d \to \mathbb{C}^d$ that amplifies or attenuates desired properties of the graph signal. We will use the intuition from Sec. 2.4.3 to construct and interpret linear filters on graph signals.

In general, any bounded linear operator on $\mathbb{C}^d$ defines a linear filter. However, we restrict our attention to shift-invariant linear filters.

Definition 18. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with graph operator $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$. A linear operator $\mathbf{A} \in \mathcal{B}\left(\mathbb{C}^d\right)$ is *shift-invariant* if for all $\mathbf{x} \in \mathbb{C}^d$, $\mathbf{ASx} = \mathbf{SAx}$.

Note that we consider both self-adjoint and non-self-adjoint graph operators as opposed to Sec. 2.4.3, in which we only considered self-adjoint graph operators. This definition is inspired by linear time-invariant filters from classical signal processing. Time-invariant linear filters can be considered as commuting with the time-advancement operator. That is, let $x \in \ell^2\left(\mathbb{Z}/d\mathbb{Z}\right)$ be a time-series signal and $(\mathbf{T}x)[t] = x[(t-1) \bmod d]$. Then, a linear filter $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z}\right)\right)$ is time-invariant if $\mathbf{A}(\mathbf{T}x)[t] = (\mathbf{TA}x)[t] = \mathbf{A}x[(t-1) \bmod d]$.

Remark. The discrete time-advancement operator $T : \ell^2\left(\mathbb{Z}/d\mathbb{Z}\right)$ has a matrix rep-

resentation as a cyclic matrix:

$$\mathbf{T} = \begin{bmatrix} & & & 1 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}.$$

In this way, the property of time-invariance can be understood as commuting with $\mathbf{T}$.[1] This makes more clear the motivation for shift-invariance to general graph operators. The cyclic operator $\mathbf{T}$ defines a very simple graph structure, whereas graph filtering accommodates arbitrary graph operators.

Following Def. 18, we can characterize shift-invariant linear filters with the following theorem of Sandryhaila and Moura [90]:

Theorem 2.10 (Theorem 1, Sandryhaila and Moura)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with graph operator $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$. Assume that the characteristic and minimal polynomials of $\mathbf{S}$ are equal. Then, a graph filter $\mathbf{A} \in \mathcal{B}\left(\mathbb{C}^d\right)$ is shift-invariant if and only if there exists a d-order polynomial $h : \mathbb{C} \to \mathbb{C}$ over the complex field such that $\mathbf{A} = h(\mathbf{S})$.*

This result together with the spectral theorem for finite dimensional linear operators will help us understand filters from a spectral analysis. Polynomials on matrices act on the eigenvalues of the matrices, and so for $\mathbf{S} = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^*$ and a polynomial $h : \mathbb{C} \to \mathbb{C}$, we can understand the action of $\mathbf{A} = h(\mathbf{S})$ on a graph signal

[1] This observation was made in Sandryhaila and Moura [90].

$\mathbf{x} \in \mathbb{C}^d$ as

$$\mathbf{A}\mathbf{x} = \sum_{i=1}^{d} h(\lambda_i) \langle \mathbf{x}, \mathbf{v}_i \rangle \mathbf{u}_i. \tag{2.31}$$

From the definition of the graph Fourier transform (Def. 17), we can understand $(h(\lambda_i))_{i=1,\dots,d}$ as the transfer function of $\mathbf{A}$ as it acts multiplicatively on the spectral representation of $\mathbf{x}$, $(\langle \mathbf{x}, \mathbf{v}_i \rangle)_{i=1,\dots,d}$.

2.5 Autoregressive processes

In this section, we review some relevant results concerning autoregressive processes. In Chapters 3 and 4, we consider the matrix symbol associated with an autoregressive process as the graph operator for an extended graph. We will make these concepts clear in Sec. 3.1. Here, we consider autoregressive processes in the abstract. For further reading on autoregressive processes, see *e.g.* Lütkepohl [73] and Priestley [83].

An autoregressive process $\mathbf{x} = (\mathbf{x}[t])_{t \in \mathbb{Z}}$ is a random process generated by the recurrence relation:

$$\mathbf{x}[t] = \mathbf{K}[1]\mathbf{x}[t-1] + \cdots + \mathbf{K}[m]\mathbf{x}[t-m] + \mathbf{n}[t], \tag{2.32}$$

where, $m \in \mathbb{N}$ is known as the model order and $\mathbf{n} = (\mathbf{n}[t])_{t \in \mathbb{Z}}$ is a Gaussian process known as the innovation.[2] If $\mathbf{x}[t] \in \mathbb{R}^d$, then $\mathbf{K}[s] \in \mathcal{B}\left(\mathbb{R}^d\right)$ for all $s = 1, \dots, m$.

[2]There are some technical considerations for defining a valid Gaussian measure on $\ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)$ which exclude a true iid noise process. Some extension to this end include abstract Wiener spaces—pioneered by Wiener himself—but we gloss over this nuance and assume that the covariance of the noise process is bounded in Hilbert-Schmidt norm, decaying to zero safely far from where our analysis takes place. For further reading, see Bogachev [14].

Note that $(\mathbf{K}[s])_{s\in\mathbb{Z}}$ where $\mathbf{K}[s] = \mathbf{0}$ for all $s \neq 1, \ldots, m$ defines a Laurent operator.

Let $\mathbf{A} \in \ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)$ be the Laurent operator induced by the matrix symbol $(\mathbf{K}[s])_{s\in\mathbb{Z}}$. Then, (2.32) is given in moving average representation by

$$\mathbf{x}[t] = \left(\left(\mathbf{I} - \mathbf{A}\right)^{-1} \mathbf{n}\right)[t], \tag{2.33}$$

and the autoregressive process of (2.32) is then centered Gaussian with covariance function $\mathbf{Q} : \ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)^* \times \ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)^* \to \mathbb{C}$,

$$\{\mathbf{u}, \mathbf{v}\} \mapsto \left\langle \mathbf{u}, \left(\mathbf{I} - \mathbf{A}\right)^{-1} \Sigma \left(\mathbf{I} - \mathbf{A}\right)^{-*} \mathbf{v} \right\rangle_{\ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)}, \tag{2.34}$$

where $\Sigma : \ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)^* \times \ell^2\left(\mathbb{Z}; \mathbb{R}^d\right)^* \to \mathbb{C}$ is the covariance function of the Gaussian process $\mathbf{n}$. The matrix symbol of $\mathbf{Q}$, $\mathbf{R} = (\mathbf{R}[t])_{t\in\mathbb{Z}}$ where $\mathbf{R}[r-s] = \mathbb{E}\mathbf{x}_t\mathbf{x}^*_{t-(r-s)}$, is known as the autocovariance function. It is a Laurent operator which admits a spectral representation known as the spectral density function [83],

$$\hat{\mathbf{R}}(\omega) := \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{R}_t = \left(\mathbf{I} - \hat{\mathbf{A}}(\omega)\right)^{-1} \cdot \Sigma(\omega) \cdot \left(\mathbf{I} - \hat{\mathbf{A}}(\omega\right)^{-*}. \tag{2.35}$$

2.5.1 Analytical results

We review some terminology and associated results about autoregressive processes.

Definition 19. A Laurent operator is called *causal* (respectively *strictly causal*) if the symbol $a = (a[s])_{s\in\mathbb{Z}}$ has support on $\mathbb{N} \cup \{0\}$ (respectively $\mathbb{N} \setminus \{0\}$).

Definition 20. An autoregressive process is called *stable* if the spectral radius of the corresponding Laurent operator is strictly less than 1, *i.e.* $\mathrm{spr}\left(\mathbf{A}\right) < 1$.

Remark. An autoregressive process is causal by construction. The stability criterion can be understood from the moving average form (2.33). Note that $(\mathbf{I}-\mathbf{A})^{-1} = \phi(\mathbf{A})$ for $\phi(z) = (1-z)^{-1}$, a holomorphic function on the open unit disk. Therefore, if $\operatorname{spr}(\mathbf{A}) < 1$, $\phi(\mathbf{A})$ is well-defined and bounded. That is, stability is a sufficient condition for $\|\mathbf{x}\| < \infty$, *i.e.* that noise is not catastrophically amplified in the signal. A sufficient condition is given by $\|\mathbf{A}\| < 1$,

$$\begin{aligned}\|\mathbf{x}\| &= \left\|(\mathbf{I}-\mathbf{A})^{-1}\mathbf{n}\right\| \\ &\leq \left\|(\mathbf{I}-\mathbf{A})^{-1}\right\| \cdot \|\mathbf{n}\| \\ &= \frac{1}{\sigma_{\min}((\mathbf{I}-\mathbf{A}))} \cdot \|\mathbf{n}\| \\ &\leq \frac{1}{1-\|\mathbf{A}\|} \cdot \|\mathbf{n}\|.\end{aligned}$$

Theorem 2.11. *Let $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{R}^d\right)\right)$ be a Laurent operator with matrix symbol $(\mathbf{K}[t])_{t\in\mathbb{Z}}$. Let $\mathbf{x} = (\mathbf{x}[t])_{t\in\mathbb{Z}}$ be an autogregressive process defined by a Gaussian noise process $\mathbf{n} = (\mathbf{n}[t])_{t\in\mathbb{Z}}$ and $(\mathbf{K}[t])_{t\in\mathbb{Z}}$. Then, the spectrum of the autocovariance function is a subset of an annulus,*

$$\Lambda(\mathbf{R}) \subset \left\{z \in \mathbb{C} : \frac{\lambda_{\min}(\Sigma)}{(1+\|\mathbf{A}\|)^2} \leq |z| \leq \frac{\lambda_{\max}(\Sigma)}{(1-\|\mathbf{A}\|)^2}\right\}. \tag{2.36}$$

Proof. The covariance function $\mathbf{Q}$ of $\mathbf{x}$ is self-adjoint, which allows us to introduce the first set of inequalities. For all $\mathbf{v} \in \ell^2\left(\mathbb{Z};\mathbb{R}^d\right)$,

$$\begin{aligned}\lambda_{\min}(\Sigma) \cdot \left\|(\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}\right\|^2 &\leq \left\langle (\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}, \Sigma(\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}\right\rangle \\ &\leq \lambda_{\max}(\Sigma) \cdot \left\|(\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}\right\|^2.\end{aligned}$$

Now, we can further bound the inequalities in terms of the norm of $\mathbf{A}$,

$$\left\|(\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}\right\| \leq \frac{1}{\sigma_{\min}\left((\mathbf{I}-\mathbf{A})^{*}\right)}\cdot\|\mathbf{v}\| \leq \frac{1}{1-\|\mathbf{A}^{*}\|}\cdot\|\mathbf{v}\| = \frac{1}{1-\|\mathbf{A}\|}\cdot\|\mathbf{v}\|\,.$$

For a lower bound,

$$\begin{aligned}\left\|(\mathbf{I}-\mathbf{A})^{-*}\mathbf{v}\right\| &\geq \sigma_{\min}\left((\mathbf{I}-\mathbf{A})^{-*}\right)\cdot\|\mathbf{v}\| \\ &= \frac{1}{\|\mathbf{I}-\mathbf{A}^{*}\|}\cdot\|\mathbf{v}\| \\ &\geq \frac{1}{1+\|\mathbf{A}^{*}\|}\cdot\|\mathbf{v}\| \\ &\geq \frac{1}{1+\|\mathbf{A}\|}\cdot\|\mathbf{v}\|\,.\end{aligned}$$

□

2.5.2 Parameter estimation

We derive a simple maximum likelihood estimator (MLE) as a method to fit autoregressive models to data.[3] Maximum *a posteriori* (MAP) estimators can be derived by assuming prior beliefs on the distribution of matrix symbols, or graph operators.

Suppose that we observe a finite length subset $(\mathbf{x}[t])_{t=1,\ldots,T+m}$, generated by Eq. (2.32). We we want to find $(\mathbf{K}[s])_{s=1,\ldots,m}$. Since $\mathbf{x}^{*}[t] = \sum_{s=1}^{m} x^{*}[t-s]\mathbf{K}^{*}[s] + \mathbf{n}^{*}[t]$,

[3]The proposed estimator is more appropriately a quasi-MLE, as we make implicit assumptions about the initial observations for analytical expediency.

we can express the observed data in the following matrix-vector form:

$$\underbrace{\begin{bmatrix} \mathbf{x}^*[1+m] \\ \vdots \\ \mathbf{x}^*[T+m] \end{bmatrix}}_{\mathbf{Y}\in\mathbb{R}^{T\times d}} = \underbrace{\begin{bmatrix} \mathbf{x}^*[m] & \cdots & \mathbf{x}^*[1] \\ & \ddots & \vdots \\ \vdots & & \mathbf{x}^*[m] \\ & \ddots & \\ \mathbf{x}^*[T] & & \vdots \\ \vdots & \ddots & \\ \mathbf{x}^*[T+m-1] & \cdots & \mathbf{x}^*[T] \end{bmatrix}}_{\mathbf{X}\in\mathbb{R}^{T\times(d\cdot m)}} \underbrace{\begin{bmatrix} \mathbf{K}^*[1] \\ \vdots \\ \mathbf{K}^*[m] \end{bmatrix}}_{\mathbf{A}\in\mathbb{R}^{(d\cdot m)\times d}} + \underbrace{\begin{bmatrix} \mathbf{n}^*[1+m] \\ \vdots \\ \mathbf{n}^*[T+m] \end{bmatrix}}_{\mathbf{N}\in\mathbb{R}^{T\times d}}. \quad (2.37)$$

In this set-up, we can minimize the mean squared error,

$$\underset{\mathbf{A}\in\mathbb{R}^{(d\cdot m)\times d}}{\arg\min} \|\mathbf{Y}-\mathbf{X}\mathbf{A}\|_F^2, \quad (2.38)$$

with the ordinary least squares estimator:

$$\mathbf{A} = (\mathbf{X}^*\mathbf{X})^{-1}\mathbf{X}^*\mathbf{Y}. \quad (2.39)$$

This estimator coincides with the MLE for which we use the conditional distribution of $\mathbf{X}_t|\mathbf{X}_{t-1},\cdots,\mathbf{X}_{t-m}$ for $t = m+1,\ldots,T+m$ and assume a uniform prior over $\mathbf{X}_1,\ldots,\mathbf{X}_m$. We also note that this gives us a column-wise estimator: $\mathbf{A}_i = \left(\frac{1}{T}\mathbf{X}^*\mathbf{X}\right)^{-1}\mathbf{X}^*\mathbf{Y}_i$ for $i = 1,\ldots,d$.

2.6 Dictionary learning

Dictionary learning has a rich history in signal processing [80, 44, 4, 74, 113]. At its core is a generative model for observed signals, $(\mathbf{y}_i)_{i=1,\ldots,N}$,

$$\mathbf{y}_i = \sum_{j=1}^{r} c_{j,i}\mathbf{d}_j, \quad (2.40)$$

where $\mathbf{y}_i, \mathbf{d}_j \in \mathbb{R}^d$ and $c_{j,i} \in \mathbb{R}$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, r$. That is, $\mathbf{y}_i$ is a linear expansion in terms of a "dictionary" of vectors $(\mathbf{d}_j)_{j=1,\ldots,r}$. Importantly, it is assumed that $\mathbf{y}_i$ admits a sparse expansion, *i.e.* for all $i = 1, \ldots, N$, very few coefficients $(c_{j,i})_{j=1,\ldots,r}$ are nonzero. Let $s = |\{c_{j,i} \neq 0 : j = 1, \ldots, r\}|$, then very few means $s \ll r$. This sparsity condition lends it the alternative name "sparse coding." Dictionary learning is cast as the complement to basis pursuit [31]: if the "right" dictionary is not given, how do we learn it from observations?

This introduces the computational challenge of the dictionary learning problem. We observe only $(\mathbf{y}_i)_{i=1,\ldots,N}$, and we want to simultaneously learn the dictionary $(\mathbf{d}_j)_{j=1,\ldots,r}$ and coefficients of the expansion $(c_{j,i})_{i=1,\ldots,N,j=1,\ldots,r}$. Minimizing the mean squared error of the expansion yields the following optimization problem:

$$\underset{\mathbf{D}\in\mathbb{R}^{d\times r},\mathbf{C}\in\mathbb{R}^{r\times N}}{\arg\min} \|\mathbf{Y} - \mathbf{D}\mathbf{C}\|_F^2 . \tag{2.41}$$

Here, we have aligned the observations $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_N]$, coefficients $C_{i,j} = c_{j,i}$, and dictionary atoms $\mathbf{D} = [\mathbf{d}_1 \cdots \mathbf{d}_r]$ into matrices. This is a bi-convex problem in $\mathbf{C}$ and $\mathbf{D}$. It has no unique global minimizer. In fact, any solution $\mathbf{C}^\star, \mathbf{D}^\star$ is only unique up to any unitary transformation $\mathbf{U}$, $\mathbf{U}\mathbf{C}^\star, \mathbf{D}^\star\mathbf{U}^*$. To this point, we have not used the sparsity condition, which we can impose with a sparsity-inducing penalty, *e.g.*

$$\underset{\mathbf{D}\in\mathbb{R}^{d\times r},\mathbf{C}\in\mathbb{R}^{r\times N}}{\arg\min} \|\mathbf{Y} - \mathbf{D}\mathbf{C}\|_F^2 + \mu \|\mathbf{C}\|_{1,1} . \tag{2.42}$$

If we further fix $\|\mathbf{d}_j\| = 1$ for all $j = 1, \ldots, r$, then we can remove all but the sign and permutation ambiguity of a candidate solution. However, we have introduced a nonconvex feasible set along with the nonconvex objective.

Due to this sign and permutation ambiguity, we define the following pseudo-metric.

Definition 21. The *dictionary metric* for any dictionaries $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{d \times r}$ is given by

$$d(\mathbf{D}_1, \mathbf{D}_2) := \min_{\mathbf{P} \in \Pi, \mathbf{Z} \in \mathrm{diag}(\pm 1)} \|\mathbf{D}_1 - \mathbf{D}_2 \mathbf{P} \mathbf{Z}\|_F \,, \tag{2.43}$$

where, Π is the set of all r-dimensional permutation matrices, and $\mathrm{diag}\,(\pm 1)$ is the set of all r-dimensional diagonal matrices with entries ± 1.

For shorthand, we will at times use $d(\mathbf{D}_1, \mathbf{D}_2) = \|\mathbf{D}_1 - \mathbf{D}_2 \mathbf{P}\|_F$, where $\mathbf{P}$ is taken to be the signed permutation matrix that minimizes the norm.

Following the publication and empirical success of the K-SVD algorithm of Aharon, *et al.* for solving the dictionary learning problem [4], it took six years for publication of the first provably convergent dictionary learning algorithm in Spielman, *et al.* [104]. However, this result only worked for $\mathbf{D}$ nonsingular ($r = n$). Soon afterward, these results were extended to the overcomplete setting ($r > n$) in Arora, *et al.* [9] and Agarwal, *et al.* [2, 1]. All of these approaches follow a similar two-stage approach. The first stage estimates an initial dictionary $\mathbf{D}$, and the second stage refines that estimate by alternately estimating the coefficients $\mathbf{C}$ and dictionary $\mathbf{D}$.

In the first stage, we build a graph $\mathcal{G}$ with vertices corresponding to the observations, *i.e.* $|\mathcal{V}| = N$. Then, add edges based on the cross-correlation of the

observations,

$$\mathcal{E}_{i,j} = \begin{cases} 1 & |\langle \mathbf{y}_i, \mathbf{y}_j \rangle| > \tau \\ & \\ 0 & \text{o.w.} \end{cases}, \tag{2.44}$$

where $\tau > 0$ is a threshold depending on the generative model for $\mathbf{X}$. Given $\mathcal{G}$, then we can use an overlapping clustering algorithm in which we identify clusters which all share a single common atom. For example, if $c_{j,i} \neq 0$ for $i = 1, 2, 3$ and some $j = 1, \ldots, r$, then $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ should belong to a common cluster. From here, we can estimate the dictionary atom common to all cluster members using *e.g.* principal component analysis.

In the second stage, we iteratively refine the dictionary estimate and update the coefficient estimate. To update the coefficient estimate, we can fix $\mathbf{D}$ and solve Eq. (2.42) for $\mathbf{C}$. Then, we can fix $\mathbf{C}$ with this updated estimate and solve Eq. (2.42) for $\mathbf{D}$. This is rightly interpreted as an expectation-maximization scheme [39]. There is considerable flexibility in how to implement each of these respective steps. Variants of LASSO and basis pursuit, or their nonconvex corollaries can be used for the coefficient update provided that they can yield high-probability error bounds, *e.g.* [29, 40, 107]. The dictionary can be updated via least squares or a principal component analysis-type approach as in the K-SVD algorithm [5].

We recount the following representative result from which we will begin our analysis in Chapter 4. Let us first begin with a definition.

Definition 22. A random vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ has bounded ℓ-wise moments if the probability that X is nonzero in any subset $S \subset 1, \ldots, d$ such that

$|S| = \ell$ is at most $c^\ell \prod_{i \in S} \mathbb{P}(X_i \neq 0)$ for some $c \sim \mathcal{O}(1)$.

Theorem 2.12 (Theorem 4 (with noise), Arora, *et al.* [9]). *Suppose we observe the following sequence of observations* $(\mathbf{D}^\star \mathbf{c}_i + \mathbf{n}_i)_{i=1,\ldots,N}$. *Assume*

(1) The dictionary $\mathbf{D}^\star \in \mathbb{R}^{d \times r}$ *has columns of unit norm and satisfies* $|\langle \mathbf{d}_i, \mathbf{d}_j \rangle| < \mu/\sqrt{d}$ *for all* $\{i \neq j : i, j = 1, \ldots, r\}$ *and some* $\mu \sim \mathcal{O}(\log d)$*;*

(2) $\mathbf{c}_i \in \mathbb{R}^r$ *are iid random vectors with random support of size* s*; the nonzero coefficients are drawn independently from a centered distribution with support* $[-C, 1] \cup [1, C]$ *for some* $C \sim \mathcal{O}(1)$*; the distribution of* $\mathbf{c}_i$ *has bounded 3-wise moments;*

(3) $s \leq c \cdot \min\left(r^{2/5}, \sqrt{d}/(\mu \log d)\right)$ *for some* $c > 0$*; and*

(4) $\mathbf{n}_i \sim_{iid} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ *with* $\sigma \sim o\left(\sqrt{d}\right)$.

Then, if $N \sim \Omega\left(\sigma^2/\varepsilon^2 \cdot [(r^2/s^2) \log r + r s^2 \log r + r \log r \log(1/\varepsilon)]\right)$*, there exists a polynomial time algorithm such that with high probability, the algorithm will return a dictionary estimate* $\mathbf{D}_0$ *that satisfies* $d(\mathbf{D}^\star, \mathbf{D}_0) \leq \varepsilon$.

Chapter 3: Filtering stochastic processes on graphs

3.1 Stochastic processes on graphs

In Section 2.4.2, we defined graph signals and related them to complex vectors. In this section, we extend scalar graph signals to time-indexed graph signals. Then, we relate these stochastic processes on graphs to vector-valued sequences.

Before we can define time-varying graph signals, we must introduce the notion of an extended graph. Ultimately, we want to define graph signals that are indexed in time, and in so doing, we want to accommodate more general graph topologies. Consider a set of nodes $\mathcal{V}$ on which we observe a graph signal over time $t \in \mathbb{Z}$. We could of course interpret this observation as different graph signals on the same graph. Another interpretation is that we observe a single graph signal taking values over a much larger graph. That is, the graph extends in time to support the signal. We define an extended graph according to this latter perspective.

Definition 23. Let $\mathcal{V}$ be a set of nodes $|\mathcal{V}| < \infty$, and let time be indexed by $\mathbb{Z}$. Then, an extended graph $\mathcal{G}$ is defined by a node set $\mathbb{Z} \times \mathcal{V}$ and edge set $\mathcal{E} \subseteq (\mathbb{Z} \times \mathcal{V}) \times (\mathbb{Z} \times \mathcal{V})$.

In the definition of the extended graph, note that we can accommodate edges

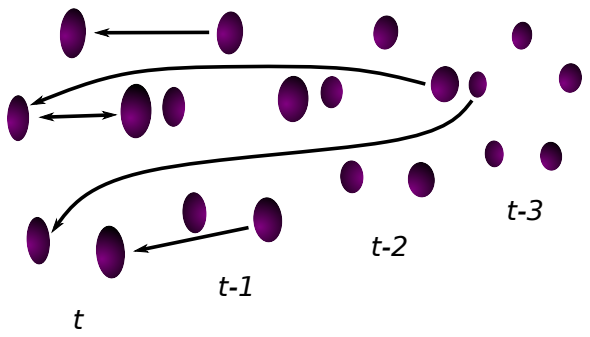


Figure 3.1: Example extended graph. Nodes are indexed by both time and space. The extended graph allows edges between any two nodes in time and space.

that span time so that nodes can interact at multiple time scales. See Fig. 3.1 for an example.

We also introduce a notion of stationarity for extended graphs.

Definition 24. An extended graph is said to be *stationary* if the existence of an edge for $v_1, v_2 \in \mathcal{V}$ and $t_1, t_2 \in \mathbb{Z}$ implies the existence of an edge for v_1, v_2 at $t_1 + m, t_2 + m$ for all $m \in \mathbb{Z}$.

Stationarity of an extended graph means that the relationship of nodes is a function of their distance apart in time and not an arbitrary function of time, similar to stationarity in a random process. See Fig. 3.2 for an example.

Now, we define time-varying graph signals.

Definition 25. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an extended graph with time indexed by $\mathbb{Z}$. Then, a *time-varying graph signal* is a function $x : \mathbb{Z} \times \mathcal{V} \to \mathbb{C}$.

Remark. A random time-varying graph signal is called a *stochastic process on a graph*.

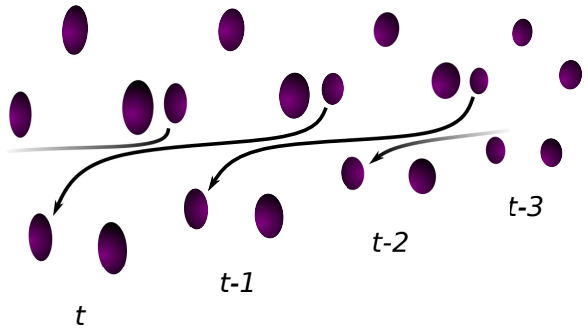


Figure 3.2: Example stationary graph. The edges present between nodes are a function of distance in time.

As in Sec. 2.4.2, we can define an appropriate function space for time-varying graph signals.

Definition 26. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an extended graph with time indexed by $\mathbb{Z}$ and $x : \mathbb{Z} \times \mathcal{V} \to \mathbb{C}$ be a time-varying graph signal. The *p, q-space of time-varying graph signals* is given by

$$\ell^{p,q}(\mathbb{Z}; \mathcal{V}) = \left\{ x : \mathbb{Z} \times \mathcal{V} \to \mathbb{C} : \sum_{t \in \mathbb{Z}} \left(\sum_{i \in \mathcal{V}} |x_{t,i}|^p \right)^{q/p} < \infty \right\} \tag{3.1}$$

for any $1 \leq p, q \leq \infty$.

As in the graph signal case, the p, q-space of time-varying graph signals is isomorphic to a Euclidean space, $\ell^q\left(\mathbb{Z}; \mathbb{C}^d\right)$. For the remainder of the chapter, we concern ourselves with $\ell^{2,2}(\mathbb{Z}; \mathcal{V})$, which is isomorphic to $\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$, a Hilbert space. We also restrict our attention to stationary extended graphs. This excludes general time-varying graphs.

We can again understand weight functions of extended graphs $w : \mathcal{E} \to \mathbb{C}$ as defining bounded linear operators on $\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$. We require that graph operators

exhibit the same stationarity as the underlying extended graph. That is, if $\mathcal{G}$ is a stationary graph, then only Laurent operators are admissible graph operators.

3.2 Linear, Time-invariant Filtering

In this section, we use time-invariant filtering on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ to preview our approach to filtering stochastic processes on graphs. We define an appropriate notion of time-invariance that maintains consistency with classical linear time-invariant signal processing, characterize filters which exhibit time-invariance, and then propose constructive approaches to realize time-invariant filters.

Let time-evolution of a discrete signal $(x[t])_{t\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$ be associated with the time-shift operator, $\mathbf{T} \in \mathcal{B}(\ell^2(\mathbb{Z}))$,

$$(\mathbf{T}x)[t] = x[t-1]. \tag{3.2}$$

An operator $\mathbf{A} \in \mathcal{B}(\ell^2(\mathbb{Z}))$ is said time-invariant if for any $(x[t])_{t\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$,

$$(\mathbf{TA}x)[t] = (\mathbf{AT}x)[t]. \tag{3.3}$$

The time-shift operator $\mathbf{T}$ is a Laurent operator with scalar symbol $(\delta[t+1])_{t\in\mathbb{Z}}$. We can generalize this definition to a time-shift operator defined by the matrix symbol $(\delta[t+1]\mathbf{I})_{t\in\mathbb{Z}}$, where $\mathbf{I} \in \mathcal{B}\left(\mathbb{C}^d\right)$. Thus, the time-shift operator advances a vector-valued signal $(\mathbf{x}[t])_{t\in\mathbb{Z}} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$, in whole, one step in time:

$$(\mathbf{Tx})[t] = \mathbf{x}[t-1]. \tag{3.4}$$

The definition of time-invariance extends immediately, and we formalize it in the following definition:

Definition 27. $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ is called *time-invariant* if $\mathbf{A}$ is in the commutant of $\mathcal{T} = \langle \mathbf{T} \rangle$, $\left\{\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right) : \mathbf{A}\mathcal{T} = \mathcal{T}\mathbf{A}\right\}$.

Next, recalling the definition of a Laurent operator (Def. 14) we prove an important relationship about Laurent operators and time-invariance.

Theorem 3.1. $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ *is time-invariant if and only if* $\mathbf{A}$ *is Laurent.*

Proof. $\Leftarrow$ Let $\mathbf{A}$ be Laurent with matrix symbol $(\mathbf{K}[t])_{t\in\mathbb{Z}}$. We want to show that $\mathbf{A}\mathbf{T}^m = \mathbf{T}^m\mathbf{A}$ for any $m \in \mathbb{Z}$.

$$\begin{aligned}
\left(\mathbf{A}\mathbf{T}^m\mathbf{x}\right)[t] &= \lim_{N\to\infty} \sum_{s=-N}^{N} \mathbf{K}[t-s]\left(\mathbf{T}^m\mathbf{x}\right)[s] \\
&= \lim_{N\to\infty} \sum_{s=-N}^{N} \mathbf{K}[t-s]\mathbf{x}[s-m] \\
&= \lim_{N\to\infty} \sum_{s'=-N}^{N} \mathbf{K}[t-(s'+m)]\mathbf{x}[s'] \\
&= \mathbf{T}^m \left(\lim_{N\to\infty} \sum_{s'=-N}^{N} \mathbf{K}[t-s']\mathbf{x}[s'] \right) \\
&= \left(\mathbf{T}^m\mathbf{A}\mathbf{x}\right)[t]
\end{aligned}$$

$\Rightarrow$ Let $\mathbf{A}$ be time-invariant. Since $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$, there exists a kernel function $\mathbf{K} : \mathbb{Z}\times\mathbb{Z} \to \mathcal{B}\left(\mathbb{C}^d\right)$ such that for every $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$\left(\mathbf{A}\mathbf{x}\right)[t] = \sum_{s\in\mathbb{Z}} \mathbf{K}[t,s]\mathbf{x}[s].$$

By Def. 27, $\mathbf{A}$ commutes with $\mathcal{T}$. Without loss of generality, choose $\mathbf{T}^m \in \mathcal{T}$ for

some $m \in \mathbb{Z}$. It is necessary to show that $\mathbf{K}[t,s] = \mathbf{K}[t+m, s+m]$.

$$
\begin{aligned}
\left(\mathbf{A}\mathbf{T}^m\mathbf{x}\right)[t+m] &= \left(\mathbf{T}^m\mathbf{A}\mathbf{x}\right)[t+m] \\
\lim_{N\to\infty}\sum_{s=-N}^{N}\mathbf{K}[t+m,s]\left(\mathbf{T}^m\mathbf{x}\right)[s] &= \mathbf{T}^m\left(\lim_{N\to\infty}\sum_{s=-N}^{N}\mathbf{K}[t+m,s]\mathbf{x}[s]\right) \\
\lim_{N\to\infty}\sum_{s=-N}^{N}\mathbf{K}[t+m,s]\mathbf{x}[s-m] &= \lim_{N\to\infty}\sum_{s=-N}^{N}\mathbf{K}[t,s]\mathbf{x}[s] \\
\lim_{N\to\infty}\sum_{s'=-N}^{N}\mathbf{K}[t+m,s'+m]\mathbf{x}[s'] &= \lim_{N\to\infty}\sum_{s=-N}^{N}\mathbf{K}[t,s]\mathbf{x}[s]
\end{aligned}
$$

By the uniqueness of $\mathbf{K}$, $\mathbf{A}$ is Laurent. □

By Thm. 2.3, we know that Laurent operators on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ are defined by essentially bounded matrix symbols, $\sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{K}[t] \in L^\infty\left([0,1];\mathcal{B}\left(\mathbb{C}^d\right)\right)$. We can extend these results to the further claim that linear, time-invariant filters on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ can be realized by the independent choice of $d\times d$ essentially bounded functions, $\left\{\hat{k}_{i,j} \in L^\infty\left([0,1]\right) : i,j = 1,\ldots,d\right\}$.

Corollary 3.1.1. *Let* $\left\{\hat{k}_{i,j} \in L^\infty\left([0,1]\right) : i,j = 1,\ldots,d\right\}$ *define a Laurent operator* $\mathbf{A}$. *That is, let* $\mathbf{A}$ *have matrix symbol* $\left(\mathbf{K}[t]\right)_{t\in\mathbb{Z}}$,

$$K_{i,j}[t] = \int_0^1 e^{-2\pi i\omega t}k_{i,j}(\omega)d\omega$$

for all $i,j = 1,\ldots,d$. *Then,* $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ *is time-invariant. Moreover,* $\|\mathbf{A}\| = \operatorname{ess\,sup}_{\omega\in[0,1]}\left\|\hat{\mathbf{A}}(\omega)\right\|$.

Proof. Consider a fiber $\omega \in [0,1]$ of $\hat{\mathbf{A}}(\omega)$. $\hat{\mathbf{A}}(\omega) \in \mathbb{C}^{d\times d}$ comprises $d\times d$ finite entries. Because of the equivalence of all matrix norms in finite dimensions, $\hat{\mathbf{A}}(\omega) \in$

$\mathcal{B}\left(\mathbb{C}^d\right)$. This is simultaneously true of all fibers $\omega \in [0,1]$, and by Thm. 2.3, $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$.[1] □

The significance of Corollary 3.1.1 is that it provides a constructive means of designing time-invariant filters on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$. There are $\mathcal{O}\left(d^2\right)$ degrees of freedom, the d^2 essentially bounded functions $\left\{\hat{k}_{i,j} : [0,1] \to \mathbb{C}\right\}$

We can visualize the action of a time-invariant operator in the spectral domain and time domain. Let $\mathbf{A} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ be defined as in Corollary 3.1.1 by $\left\{\hat{k}_{i,j} \in L^\infty\left([0,1]\right) : i,j = 1,\ldots,d\right\}$. Then, for any $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$\sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\left(\mathbf{A}\mathbf{x}\right)[t] = \begin{bmatrix} \hat{k}_{1,1}(\omega) & \cdots & \hat{k}_{1,d}(\omega) \\ \vdots & \ddots & \vdots \\ \hat{k}_{d,1}(\omega) & \cdots & \hat{k}_{d,d}(\omega) \end{bmatrix} \cdot \begin{bmatrix} \hat{x}_1(\omega) \\ \vdots \\ \hat{x}_d(\omega) \end{bmatrix}. \tag{3.5}$$

Here, the matrix symbol is in fact the transfer function of the filter. Likewise,

$$\begin{aligned} &\left(\mathbf{A}\mathbf{x}\right)[t] \\ &= \sum_{s\in\mathbb{Z}} \begin{bmatrix} \int_0^1 e^{-2\pi i\omega s}\hat{k}_{1,1}(\omega)d\omega & \cdots & \int_0^1 e^{-2\pi i\omega s}\hat{k}_{1,d}(\omega)d\omega \\ \vdots & \ddots & \vdots \\ \int_0^1 e^{-2\pi i\omega s}\hat{k}_{d,1}(\omega)d\omega & \cdots & \int_0^1 e^{-2\pi i\omega s}\hat{k}_{d,d}(\omega)d\omega \end{bmatrix} \cdot \begin{bmatrix} x_1[t-s] \\ \vdots \\ x_d[t-s] \end{bmatrix}, \end{aligned} \tag{3.6}$$

where the matrix symbol is the impulse-response function of the filter.

[1] To be rigorous, we consider a fiber $\omega \in [0,1]\backslash\mathcal{U}_0$, where $\mathcal{U}_0 \subset [0,1]$ is a set of Lebesque measure zero where $\hat{k}_{i,j}$ can be unbounded. Hence, the essential supremum is unaffected by the partition.

3.3 Linear, Shift-invariant Filtering

Our ultimate objective is to leverage the graph structure in our filtering approach. Therefore, we want to define a class of filters which respect the graph structure as Laurent operators respect time. This work follows the developments of Bohannon, *et al.* [15, 19] Motivated by Def. 27, we define shift-invariance.

Definition 28. $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ is called *shift-invariant* to an extended graph $\mathcal{G}$ with an associated graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ if $\mathbf{A}$ is in the commutant of $\mathcal{S} = \langle \mathbf{S} \rangle$,

$$\left\{\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right) : \mathbf{A}\mathcal{S} = \mathcal{S}\mathbf{A}\right\}.$$

This definition leads to our first characterization of shift-invariant filters, which specializes a more general result about commuting operators. Recall that stationary graphs admit only Laurent graph operators so that if $\mathcal{G}$ is stationary, then $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ must be Laurent by Def. 28.

Theorem 3.2. *Let $\mathcal{G}$ be a stationary extended graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$, and consider a Laurent operator $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(Z;\mathbb{C}^d\right)\right)$. Suppose $\mathbf{S}$ and $\mathbf{A}$ admit Jordan spectral representations given pointwise for $\omega \in [0,1]$ by*

$$\hat{\mathbf{S}}(\omega) = \sum_{k=1}^{m(\omega)} \lambda_k(\omega)\mathbf{P}_k(\omega) + \mathbf{N}_k(\omega) \tag{3.7}$$

and

$$\hat{\mathbf{A}}(\omega) = \sum_{j=1}^{p(\omega)} \nu_j(\omega)\mathbf{Q}_j(\omega) + \mathbf{M}_j(\omega) \tag{3.8}$$

respectively. Then, $\mathbf{A}$ *is shift-invariant to* $\mathcal{G}$ *if and only if*

$$\mathbf{P}_k(\omega)\mathbf{Q}_j(\omega) = \mathbf{Q}_j(\omega)\mathbf{P}_k(\omega) \tag{3.9}$$

$$\mathbf{P}_k(\omega)\mathbf{M}_j(\omega) = \mathbf{M}_j(\omega)\mathbf{P}_k(\omega) \tag{3.10}$$

$$\mathbf{N}_k(\omega)\mathbf{Q}_j(\omega) = \mathbf{Q}_j(\omega)\mathbf{N}_k(\omega) \tag{3.11}$$

$$\mathbf{N}_k(\omega)\mathbf{M}_j(\omega) = \mathbf{M}_j(\omega)\mathbf{N}_k(\omega) \tag{3.12}$$

for all $k \in \{1, \ldots, m(\omega)\}$, $j \in \{1, \ldots, p(\omega)\}$, *and* $\omega \in [0, 1]$ *a.e.*

Before proving Thm. 3.2, it will help to establish an intermediate result.

Lemma 3.3. *Let* $\mathbf{A}, \mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)\right)$ *be time-invariant. Then* $\mathbf{A}$ *commutes with* $\mathcal{S}$ *if and only if*

$$\hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{S}}(\omega) = \hat{\mathbf{S}}(\omega) \cdot \hat{\mathbf{A}}(\omega) \tag{3.13}$$

almost everywhere for $\omega \in [0, 1]$.

Proof. $\Leftarrow$ Assume that Eq. (3.13) is true for $\omega \in [0, 1]$ a.e. Then,

$$\hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{S}}(\omega) \cdot \hat{\mathbf{x}}(\omega) = \hat{\mathbf{S}}(\omega) \cdot \hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{x}}(\omega)$$

$$\int_0^1 e^{-2\pi i\omega t} \hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{S}}(\omega) \cdot \hat{\mathbf{x}}(\omega) d\omega = \int_0^1 \hat{\mathbf{S}}(\omega) \cdot \hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{x}}(\omega) d\omega$$

$$(\mathbf{ASx})[t] = (\mathbf{SAx})[t].$$

A similar argument holds for $\hat{\mathbf{A}}(\omega) \cdot \mathbf{S}^m(\omega) = \mathbf{S}^m(\omega) \cdot \hat{\mathbf{A}}(\omega)$ for any $m \in \mathbb{Z}$.

$\Rightarrow$ Assume that $\mathbf{A}$ commutes with any $\mathbf{S}^m \in \mathcal{S}$.

$$(\mathbf{AS}^m\mathbf{x})[t] = (\mathbf{S}^m\mathbf{Ax})[t]$$

$$\sum_{t\in\mathbb{Z}} e^{2\pi i\omega t} (\mathbf{AS}^m\mathbf{x})[t] = \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t} (\mathbf{S}^m\mathbf{Ax})[t]$$

$$\hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{S}}^m(\omega) \cdot \hat{\mathbf{x}}(\omega) = \hat{\mathbf{S}}^m(\omega) \cdot \hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{x}}(\omega)$$

□

With this result, we prove Thm. 3.2.

Proof of Thm. 3.2. We want to show that $\hat{\mathbf{A}}(\omega)$ commutes with $\hat{\mathbf{S}}(\omega)$ almost everywhere on $\omega \in [0,1]$ if and only if Eqs. (3.9), (3.10), (3.11), and (3.12) hold for all $k \in \{1,\dots,m(\omega)\}$, $j \in \{1,\dots,p(\omega)\}$, and $\omega \in [0,1]$ a.e. Then, we can use Lemma 3.3 to complete the argument.

$\Leftarrow$ Assume that the respective projections and nilpotents commute. To make it more readable, the dependence on ω is dropped, but it is to be understood that this condition must hold pointwise for $\omega \in [0,1]$ a.e.

$$\begin{aligned}
\hat{\mathbf{A}} \cdot \hat{\mathbf{S}} &= \left(\sum_{j=1}^{p} \nu_j \mathbf{Q}_j + \mathbf{M}_j\right) \cdot \left(\sum_{k=1}^{m} \lambda_k \mathbf{P}_k + \mathbf{N}_k\right) \\
&= \sum_{j=1}^{p}\sum_{k=1}^{m} \nu_j\lambda_k \mathbf{Q}_j\mathbf{P}_k + \nu_j \mathbf{Q}_j\mathbf{N}_k + \lambda_k \mathbf{M}_j\mathbf{P}_k + \mathbf{M}_j\mathbf{N}_k \\
&= \sum_{j=1}^{p}\sum_{k=1}^{m} \nu_j\lambda_k \mathbf{P}_k\mathbf{Q}_j + \nu_j \mathbf{N}_k\mathbf{Q}_j + \lambda_k \mathbf{P}_k\mathbf{M}_j + \mathbf{N}_k\mathbf{M}_j \\
&= \left(\sum_{k=1}^{m} \lambda_k \mathbf{P}_k + \mathbf{N}_k\right) \cdot \left(\sum_{j=1}^{p} \nu_j \mathbf{Q}_j + \mathbf{M}_j\right) \\
&= \hat{\mathbf{S}} \cdot \hat{\mathbf{A}}
\end{aligned}$$

We have used that the projections and nilpotents commute in the third equality.

$\Rightarrow$

Now, assume that $\hat{\mathbf{A}}(\omega)$ and $\hat{\mathbf{S}}(\omega)$ commute almost everywhere on $\omega \in [0,1]$. We will first show that this implies that the resolvents commute. Let $z_1 \in \mathbb{C} \setminus$

$\Lambda\left(\hat{\mathbf{A}}(\omega)\right)$ and $z_2 \in \mathbb{C} \setminus \Lambda\left(\hat{\mathbf{S}}(\omega)\right)$.

$$\begin{aligned}\left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right) \cdot \left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right) &= \hat{\mathbf{A}}(\omega) \cdot \hat{\mathbf{S}}(\omega) - z_2\hat{\mathbf{A}}(\omega) - z_1\hat{\mathbf{S}}(\omega) + z_1 z_2\mathbf{I} \\ &= \hat{\mathbf{S}}(\omega)\hat{\mathbf{A}}(\omega) - z_2\hat{\mathbf{A}}(\omega) - z_1\hat{\mathbf{S}}(\omega) + z_1 z_2\mathbf{I} \\ &= \left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right) \cdot \left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right)\end{aligned}$$

Now, by taking the inverse of both sides, the intermediate claim follows.

$$\begin{aligned}\left[\left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right) \cdot \left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right)\right]^{-1} &= \left[\left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right) \cdot \left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right)\right]^{-1} \\ \left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right)^{-1} \cdot \left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right)^{-1} &= \left(\hat{\mathbf{A}}(\omega) - z_1\mathbf{I}\right)^{-1} \cdot \left(\hat{\mathbf{S}}(\omega) - z_2\mathbf{I}\right)^{-1} \\ \mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega) \cdot \mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega) &= \mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega) \cdot \mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega)\end{aligned}$$

Let $\gamma_1 \subset \mathbb{C} \setminus \Lambda\left(\hat{\mathbf{A}}(\omega)\right)$ be a closed curve that encloses only $\nu_j(\omega)$ and $\gamma_2 \subset \mathbb{C} \setminus \Lambda\left(\hat{\mathbf{S}}(\omega)\right)$ be a closed curve that encloses only $\lambda_k(\omega)$. Let $f_1(z_1)$ and $f_2(z_2)$ be two holomorphic functions on an open set which includes the curves γ_1 and γ_2 respectively. Then

$$\begin{aligned}&\left(-\frac{1}{2\pi i}\right)^2 \oint_{\gamma_1}\oint_{\gamma_2} f_1(z_1) f_2(z_2) \mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega) \cdot \mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega)\, dz_1 dz_2 \\ &\quad = \left(-\frac{1}{2\pi i}\right)^2 \oint_{\gamma_1}\oint_{\gamma_2} f_1(z_1) f_2(z_2) \mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega) \cdot \mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega)\, dz_1 dz_2\end{aligned}$$

The order of integration can be interchanged by Fubini's theorem since the resolvent is an analytic function on the resolvent set [41], which means that the integrals are bounded on γ_1 and γ_2. This allows the integrals to factor,

$$\begin{aligned}&\left(-\frac{1}{2\pi i}\oint_{\gamma_2} f_2(z_2)\mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega)\, dz_2\right) \cdot \left(-\frac{1}{2\pi i}\oint_{\gamma_1} f_1(z_1)\mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega)\, dz_1\right) \\ &\quad = \left(-\frac{1}{2\pi i}\oint_{\gamma_1} f_1(z_1)\mathcal{R}_{\hat{\mathbf{A}}}(z_1, \omega)\, dz_1\right) \cdot \left(-\frac{1}{2\pi i}\oint_{\gamma_2} f_2(z_2)\mathcal{R}_{\hat{\mathbf{S}}}(z_2, \omega)\, dz_2\right).\end{aligned} \tag{3.14}$$

We use the functional definition of the projection associated with the eigenvalue $\lambda_k(\omega)$,

$$\mathbf{P}_k(\omega) = -\frac{1}{2\pi i} \oint_{\gamma_k} \mathcal{R}_{\mathbf{S}}(z,\omega)\, dz, \tag{3.15}$$

and of the nilpotent,

$$\mathbf{N}_k(\omega) = -\frac{1}{2\pi i} \oint_{\gamma_k} (z - \lambda_k(\omega))\, \mathcal{R}_{\mathbf{S}}(z,\omega)\, dz, \tag{3.16}$$

where γ_k is a closed curve around $\lambda_k(\omega)$ [59].

For $f_1 = f_2 = 1$, Eq. (3.14) produces $\mathbf{P}_k \cdot \mathbf{Q}_j = \mathbf{Q}_j \cdot \mathbf{P}_k$, *i.e.* Eq. (3.9). For $f_1(z_1) = z_1 - \nu_j(\omega)$, $f_2 = 1$, Eq. (3.14) yields $\mathbf{P}_k \cdot \mathbf{M}_j = \mathbf{M}_j \cdot \mathbf{P}_k$, *i.e.* (3.10). Similarly, the choice $f_1 = 1$ and $f_2(z_2) = z_2 - \lambda_k(\omega)$ turns Eq. (3.14) into Eq. (3.11), whereas the choice $f_1(z_1) = z_1 - \nu_j(\omega)$ and $f_2(z_2) = z_2 - \lambda_k(\omega)$ turns Eq. (3.14) into Eq. (3.12). □

Theorem 3.2 is neither constructive nor particularly illuminating, and so, the following corollary provides a more intuitive characterization of shift-invariant filters.

Corollary 3.3.1. *Let $\mathcal{G}$ be a stationary extended graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$. Let $\mathbf{S}$ have a Jordan spectral representation given pointwise almost everywhere for $\omega \in [0,1]$ by*

$$\hat{\mathbf{S}}(\omega) = \sum_{k=1}^{m(\omega)} \lambda_k(\omega)\mathbf{P}_k(\omega) + \mathbf{N}_k(\omega). \tag{3.17}$$

Then, for any $\cup_{\omega\in[0,1]} \{\nu_k(\omega) < \infty : k = 0, \ldots, m(\omega)\}$, $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ defined by its corresponding spectral representation,

$$\hat{\mathbf{A}}(\omega) = \sum_{k=1}^{m(\omega)} \nu_k(\omega)\mathbf{P}_k(\omega) + \mathbf{N}_k(\omega), \tag{3.18}$$

is shift-invariant to $\mathcal{G}$.

Proof. The result follows immediately from Thm. 2.1 by noting that $\mathbf{P}_k$ and $\mathbf{N}_k$ satisfy Eqs. (3.9), (3.10), (3.11), and (3.12) by the definition of the Jordan spectral decomposition. We show that $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ using first Plancherel's theorem and then the inclusion of $L^\infty\left([0,1];\mathbb{C}^d\right) \subset L^2\left([0,1];\mathbb{C}^d\right)$.

$$\|\mathbf{A}\| = \left\|\hat{\mathbf{A}}\right\|_{L^2\left([0,1];\mathbb{C}^d\right)} \leq \left\|\hat{\mathbf{A}}\right\|_{L^\infty\left([0,1];\mathbb{C}^d\right)} < \infty.$$

The last inequality follows from the essential boundedness of ν_k for all $k = 1, \ldots, m$.

□

Corollary 3.3.1 provides a constructive means to design shift-invariant filters on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$. There are $\mathcal{O}(d)$ degrees of freedom, $\cup_{\omega\in[0,1]}\{\nu_k(\omega) < \infty : k = 0, \ldots, m(\omega)\}$. As opposed to Corollary 3.1.1, the filters are constructed pointwise instead of by essentially bounded functions. This alludes to a more significant problem in the construction of shift-invariant filters: the pointwise nature of the spectral theory for Laurent operators. In general, the spectrum, projections, and nilpotents are not continuous and thus cannot be defined in any unique way as functions of ω. The rank can abruptly change as a function of ω. A more thorough discussion of this phenomenon can be found in Kato [59]. See Fig. 3.3 for an example. This motivates consideration of graph operators with matrix symbols that decay exponentially fast. For such operators, the spectral theory admits a holomorphic parameterization of the spectrum, projections, and nilpotents.

Theorem 3.4. *Let $\mathcal{G}$ be a graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$. Let $\mathbf{S}$ have matrix symbol $(\mathbf{K}[t])_{t\in\mathbb{Z}}$ such that for $\varepsilon > 0$ there*

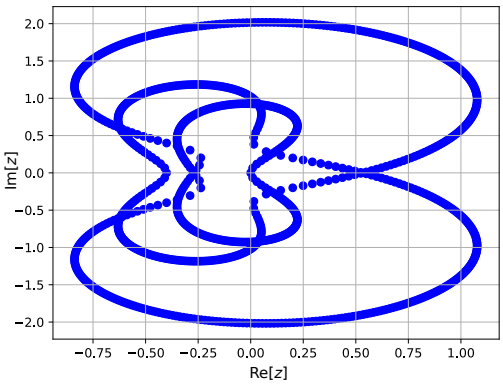


Figure 3.3: Pointwise spectrum of a Laurent operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^2\right)\right)$. The figure was created by uniformly sampling $\omega \in [0,1]$ and numerically computing an eigendecomposition. That is, each sample yields two points of the scatter plot. Even for this relatively well-behaved operator, unique components cannot be identified.

exist constants $c_1, c_2 > 0$ *such that for all* $t > 0$,

$$\|\mathbf{K}[t]\| \leq \frac{c_1}{(1+\varepsilon)^t}, \tag{3.19}$$

and for all $t < 0$,

$$\|\mathbf{K}[t]\| \leq c_2(1-\varepsilon)^t. \tag{3.20}$$

Then, the analytic continuation of $\hat{\mathbf{S}}$,

$$\hat{\hat{\mathbf{S}}}(z) = \sum_{t\in\mathbb{Z}} z^t \mathbf{K}[t], \tag{3.21}$$

is a holomorphic matrix-valued function on $U = \{z \in \mathbb{C} : 1-\varepsilon < |z| < 1+\varepsilon\}$. *Moreover, there are* d *holomorphic functions* $\{\lambda_k : U \to \mathbb{C} : k = 1,\ldots,d\}$ *with at most algebraic singularities such that*

$$det\left(\hat{\hat{\mathbf{S}}}(z) - \lambda_k(z)\mathbf{I}\right) = 0 \tag{3.22}$$

for all $z \in U$.

Proof. That the analytic continuation of $\hat{\mathbf{S}}$ converges to a holomorphic function on an annulus $z \in \{z \in \mathbb{C} : 1 - \varepsilon < |z| < 1 + \varepsilon\}$ follows from the exponential norm decay of the matrix symbol [23]. The holomorphicity of the spectrum and nature of the singularities follows from Eq. (3.22). This is an algebraic equation for which the solutions vary analytically as a function of the elements of $\hat{\hat{\mathbf{S}}}$ [63], which are holomorphic on account of the equivalence of all norms on $\mathbb{C}^{d\times d}$ [48]. □

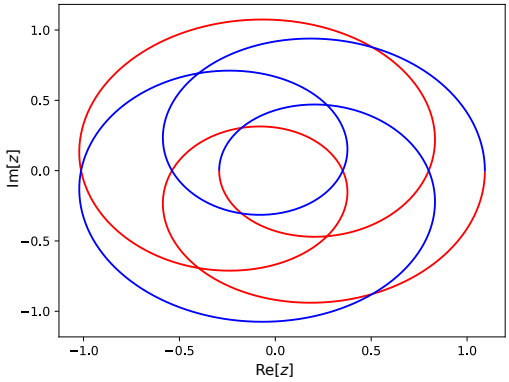


Figure 3.4: λ-group. The connected components of the spectrum compose a group, called a λ-group in Kato [59]. In this example, the connected components correspond to periodic functions with period greater than one.

That is to say that $\hat{\mathbf{S}}(\omega) = \hat{\hat{\mathbf{S}}}\left(e^{2\pi i\omega}\right)$ is a holomorphic function of $\omega \in [0, 1]$. For analytic perturbations of finite-dimensional linear operators, the eigenvalue functions form groups, the multi-valued complex functions in the spectrum. See Fig. 3.4 for an example. Each group, along with any other group it intersects in the complex plane, has an associated total projection. The total projection follows from

the functional definition of a projection with the curve drawn so as to include the entire group and any other intersecting group. The total projection is bounded and holomorphic on the annulus of holomorphy to include exceptional points. These results can be found in [59]. A simple but useful corollary follows immediately.

Corollary 3.4.1. *Let $\mathcal{G}$ be a stationary extended graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$. Let $\mathbf{S}$ have matrix symbol $\left(\mathbf{K}[t]\right)_{t\in\mathbb{Z}}$ such that for some $0 < N < \infty$, $\|\mathbf{K}[t]\| = 0$ for all $|t| > N$. Then, the analytic continuation of $\hat{\mathbf{S}}$,*

$$\hat{\hat{\mathbf{S}}}(z) = \sum_{t\in\mathbb{Z}} z^t \mathbf{K}[t], \tag{3.23}$$

is a holomorphic matrix-valued function on $U = \mathbb{C}\setminus\{0\}$. Moreover, there are d holomorphic functions $\{\lambda_k : U \to \mathbb{C} : k = 1,\ldots,d\}$ with at most algebraic singularities such that

$$det\left(\hat{\hat{\mathbf{S}}}(z) - \lambda_k(z)\mathbf{I}\right) = 0 \tag{3.24}$$

for all $z \in U$.

That is, for finitely supported matrix symbols, the spectral theory admits an almost everywhere holomorphic decomposition. This is significant because in applications, we will likely model extended graphs as supporting edges only over finite distances. Then, the graph operators associated with those extended graphs will have spectra and projections composed of smooth functions.

As in the time-invariant case, we can visualize the transfer function and impulse-response function of a shift-invariant filter. Let $\mathbf{A} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ be defined as in Corollary 3.3.1 by $\cup_{\omega\in[0,1]}\{\nu_k(\omega) < \infty : k = 1,\ldots,m(\omega)\}$ to be shift-invariant to a stationary extended graph $\mathcal{G}$ with Laurent graph operator $\mathbf{S}$. We can

also suppose that $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ admits an analytic extension on the torus in the spectral domain, which implies that $m(\omega) = m$ a.e. Then, for any $\mathbf{x} \in \ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$,

$$\sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\left(\mathbf{A}\mathbf{x}\right)[t] = \sum_{k=1}^{m} \nu_k(\omega)\cdot\mathbf{P}_k(\omega)\cdot\hat{\mathbf{x}}(\omega) + \mathbf{N}_k(\omega)\cdot\hat{\mathbf{x}}(\omega) \tag{3.25}$$

and

$$\left(\mathbf{A}\mathbf{x}\right)[t] = \sum_{k=1}^{m} \left(\left(\left(\check{\nu}_k * \check{\mathbf{P}}_k\right) + \check{\mathbf{N}}_k\right) * \mathbf{x}\right)[t]. \tag{3.26}$$

3.3.0.1 Filtering with holomorphic functional calculus

This brings us to the pinnacle of our filtering approach. Designing filters through the use of holomorphic functional calculus as in Thm. 2.2 offers a clever means to avoid the challenges of disambiguating pointwise spectral decompositions.

Theorem 3.5. *Let $\mathcal{G}$ be a stationary extended graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$. Define an open set, $U \subset \mathbb{C}$, such that $\Lambda\left(\mathbf{S}\right) \subset U$. Then, for any $\phi : U \to \mathbb{C}$, a holomorphic function, $\mathbf{A} = \phi(\mathbf{S}) \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ is shift-invariant to $\mathcal{G}$.*

Proof. Shift-invariance follows from the algebra homomorphism property. Note that $\mathbf{S} = \mathbf{1}(\mathbf{S})$ where $\mathbf{1}(z) = z$ is the identity mapping. Then,

$$\mathbf{A}\mathbf{S} = \phi\left(\mathbf{S}\right)\mathbf{1}\left(\mathbf{S}\right) = \left(\phi\cdot\mathbf{1}\right)\left(\mathbf{S}\right) = \left(\mathbf{1}\cdot\phi\right)\left(\mathbf{S}\right) = \mathbf{1}\left(\mathbf{S}\right)\phi\left(\mathbf{S}\right) = \mathbf{S}\mathbf{A}.$$

□

Theorem 3.5 says that we can specify a single degree of freedom $\phi : U \to \mathbb{C}$ to design a shift-invariant filter. That encompasses a large class of functions without

concerns about the projections and nilpotents of $\mathbf{S}$. The holomorphic functional calculus encompasses more than entire functions, since it only requires that ϕ admit a power series representation at each point in U. Moreover, as discussed in Sec. 2.1, U can be the union of disjoint sets. This means that ϕ can be different holomorphic functions on each disjoint set enclosing a connected component of the spectrum. We will employ this technique in Sec. 3.4 to design ideal bandpass filters.

We can give a spectral representation to the action of a linear filter defined via holomorphic functional calculus. That is, we can characterize the transfer function.

Corollary 3.5.1. *Let $\mathcal{G}$ be a stationary extended graph with associated Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ with $\Lambda\left(\mathbf{S}\right) \subset U$ for an open set, $U \subset \mathbb{C}$. Further, let $\mathbf{S}$ have Jordan spectral representation given pointwise almost everywhere for $\omega \in [0,1]$ by*

$$\hat{\mathbf{S}}(\omega) = \sum_{k=1}^{m(\omega)} \lambda_k(\omega)\mathbf{P}_k(\omega) + \mathbf{N}_k(\omega). \tag{3.27}$$

Then, for any holomorphic function $\phi : U \to \mathbb{C}$, the action of $\mathbf{A} = \phi(\mathbf{S}) \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ is given by

$$\left(\mathbf{A}\mathbf{x}\right)[t] = \int_0^1 e^{-2\pi i\omega t}\left[\sum_{k=1}^{m(\omega)} \phi\left(\lambda_k(\omega)\right)\mathbf{P}_k(\omega) + \phi'\left(\lambda_k(\omega)\right)\mathbf{N}_k(\omega)\right] \cdot \hat{\mathbf{x}}(\omega)d\omega \tag{3.28}$$

Proof. To begin the proof, we note that $\phi\left(\hat{\mathbf{S}}(\omega)\right)$ is given by

$$\sum_{k=1}^{m(\omega)} \phi\left(\lambda_k(\omega)\right)\mathbf{P}_k(\omega) + \phi'\left(\lambda_k(\omega)\right)\mathbf{N}_k(\omega).$$

This follows from a remark in Sec. 2.1 relating the holomorphic functional calculus with the Jordan spectral representation. Therefore, we want to show that

$$\left(\mathbf{A}\mathbf{x}\right)[t] = \int_0^1 e^{-2\pi i\omega t}\phi\left(\hat{\mathbf{S}}(\omega)\right) \cdot \hat{\mathbf{x}}(\omega)d\omega.$$

Using the definition of the holomorphic functional calculus, we have

$$(\mathbf{Ax})[t] = \int_0^1 e^{-2\pi i\omega t}\left[\frac{1}{2\pi i}\oint_\gamma \phi(z)\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,dz\right]\cdot\hat{\mathbf{x}}(\omega)d\omega.$$

for $\Lambda(\mathbf{S}) \subset \gamma \subset \operatorname{int}(U)$. By Fubini's theorem,

$$\begin{aligned}(\mathbf{Ax})[t] &= \frac{1}{2\pi i}\oint_\gamma \phi(z)\left[\int_0^1 e^{-2\pi i\omega t}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\cdot\hat{\mathbf{x}}(\omega)d\omega\right]dz \\ &= \frac{1}{2\pi i}\oint_\gamma \phi(z)\left[\sum_{s\in\mathbb{Z}}\left(\int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega\right)\mathbf{x}[s]\right]dz.\end{aligned}$$

By Fubini's theorem again,

$$(\mathbf{Ax})[t] = \sum_{s\in\mathbb{Z}}\left[\frac{1}{2\pi i}\oint_\gamma \phi(z)\left(\int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega\right)dz\right]\mathbf{x}[s].$$

If we can show that $\mathcal{R}_{\mathbf{S}}(z) = \int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega$, then, we have completed the argument since

$$\phi(\mathbf{S}) = \frac{1}{2\pi i}\oint_\gamma \phi(z)\mathcal{R}_{\mathbf{S}}(z)\,dz,$$

and we know that $\phi(\mathbf{S})$ is Laurent so that it has a matrix symbol $(\mathbf{K}[t])_{t\in\mathbb{Z}}$. To prove that $\mathcal{R}_{\mathbf{S}}(z) = \int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega$, note that $z\in\gamma$ satisfies $z > \|\mathbf{S}\|$, and so we can use the Neumann series

$$\mathcal{R}_{\mathbf{S}}(z) = z^{-1}\sum_{n=1}^{\infty} z^{-n}\mathbf{S}^n.$$

Since $\mathbf{S}$ is Laurent, $\mathbf{S}^n = \underbrace{\mathbf{K}*\cdots*\mathbf{K}}_{n}$. With this in mind, we return to $\int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega$:

$$\begin{aligned}\int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}(z,w)\,d\omega &= \int_0^1 e^{-2\pi i\omega(t-s)}\left[z^{-1}\sum_{n=1}^{\infty} z^{-n}\left(\hat{\mathbf{S}}(\omega)\right)^n\right]d\omega \\ &= z^{-1}\sum_{n=1}^{\infty} z^{-n}\left[\int_0^1 e^{-2\pi i\omega(t-s)}\left(\hat{\mathbf{S}}(\omega)\right)^n d\omega\right].\end{aligned}$$

As $\hat{\mathbf{S}}(\omega) = \sum_{t\in\mathbb{Z}} e^{2\pi i\omega t}\mathbf{K}[t]$, we can eventually show that

$$\int_0^1 e^{-2\pi i\omega(t-s)}\mathcal{R}_{\hat{\mathbf{S}}}\,(z,w)\,d\omega = z^{-1}\sum_{n=1}^{\infty} z^{-n}\underbrace{\mathbf{K}*\cdots*\mathbf{K}}_{n},$$

completing the argument. □

3.4 Applications of shift-invariant filtering

Linear filtering is an important tool in signal processing and accordingly an important extension to graph signal processing. Here, we use our developed filtering techniques to first design ideal bandpass filters for the purpose of discriminating components of a signal and second to build discriminative representations.

3.4.1 Bandpass filtering

Our proposed filtering technique is distinguishable from existing literature in two ways: it accommodates non-self-adjoint graph operators, and it supports edges across time. We attempt to highlight what is gained by these two features by composing an example signal as the combination of two pure frequency components and designing a bandpass filter to discriminate the two components. Existing proposals for filtering time-varying graph signals consider product graph models instead of the extended graph, *e.g.* [91, 72, 49]. These models lack the ability to model more general time-dependent structure between nodes. This manifests as trivial structure in the spectrum of product graph operators and an inability to discriminate the components of our example signal. Although there are considerable computational and analytical advantages to using self-adjoint graph operators, we show that

a symmetrized version of the graph operator cannot discriminate the pure frequency components of the example signal.

3.4.1.1 Bandpass filtering with the proposed approach

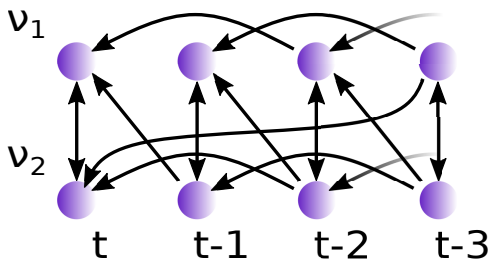


Figure 3.5: Extended graph of $\mathcal{G}$. Note the presence of edges both within time and across time. This allows nodes to interact in time and space, which leads to defining a richer class of graph operators.

Consider an extended graph $\mathcal{G}$, depicted in Fig. 3.5, with Laurent graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ and matrix symbol

$$\begin{aligned} \mathbf{K}[0] &= \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \quad \mathbf{K}[1] = \begin{bmatrix} 0 & -\frac{4}{5} \\ 0 & 0 \end{bmatrix} \\ \mathbf{K}[2] &= \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{2}{5} \end{bmatrix} \quad \mathbf{K}[3] = \begin{bmatrix} 0 & 0 \\ \frac{3}{5} & 0 \end{bmatrix} \end{aligned} \tag{3.29}$$

and $\mathbf{K}[t] = \mathbf{0}$ otherwise. $\mathbf{S}$ has the following spectral representation:

$$\hat{\mathbf{S}}(\omega) = \begin{bmatrix} \frac{1}{5}e^{4\pi i\omega} & -1-\frac{4}{5}e^{2\pi i\omega} \\ 1+\frac{3}{5}e^{6\pi i\omega} & \frac{2}{5}e^{4\pi i\omega} \end{bmatrix}. \tag{3.30}$$

In accordance with Corollary 3.4.1, the spectral representation is a holomorphic matrix-valued function of $\omega \in [0,1]$, and we can compute a pointwise eigendecom-

position with eigenvalues,

$$\lambda_{\pm}(\omega) = \frac{1}{10}\left(3e^{4\pi i\omega} \pm \sqrt{-100 - 80e^{2\pi i\omega} - 60e^{6\pi i\omega} + 47e^{8\pi i\omega}}\right), \tag{3.31}$$

and right (unnormalized) eigenvectors,

$$\mathbf{u}_{\pm}(\omega) = \begin{bmatrix} e^{4\pi i\omega} \pm \sqrt{-100 + 80e^{2\pi i\omega} - 60e^{6\pi i\omega} + 47e^{8\pi i\omega}} \\ \\ 10 + 6e^{6\pi i\omega} \end{bmatrix}. \tag{3.32}$$

The spectrum, $\Lambda(\mathbf{S}) = \cup_{\omega\in[0,1]}\{\lambda_{\pm}(\omega)\}$ is plotted in Fig. 3.6.

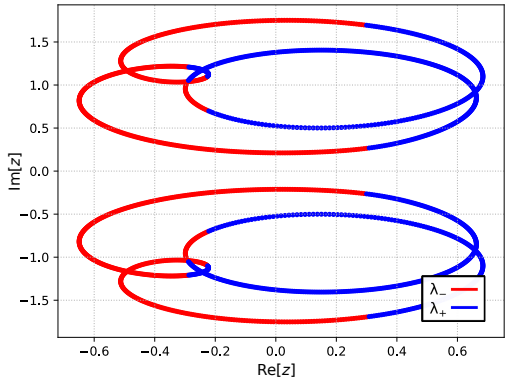


Figure 3.6: Spectrum of $\mathbf{S}$. The spectrum comprises smooth curves due to the holomorphicity of $\hat{\mathbf{S}}$. However, due to the square root, we observe coherent λ-groups, not coherence of $\cup_{\omega\in[0,1]}\lambda_{-}(\omega)$ or $\cup_{\omega\in[0,1]}\lambda_{+}(\omega)$. Note that the spectrum admits separation by disjoint open sets.

Suppose that we observe a time-varying graph signal,

$$\mathbf{x}[t] = e^{-2\pi i\omega_1 t}\mathbf{u}_{-}(\omega_1) + e^{-2\pi i\omega_2 t}\mathbf{u}_{+}(\omega_2),$$

where $\omega_1, \omega_2 \in [0,1]$. We want to define a shift-invariant filter via holomorphic functional calculus as in Thm. 3.5 to discriminate the pure frequency components

of the signal. We define an open set $U = U_1 \cup U_2$ and Jordan curve $\Gamma = \gamma_1 \cup \gamma_2$ according to Fig. 3.7. Then, provided that $\lambda_-(\omega_1) \in U_1$ and $\lambda_+(\omega_2) \in U_2$, we can define a holomorphic function $\phi : U \to \mathbb{C}$ such that $\phi(z) = 1\big|_{z\in U_1}$ and $\phi(z) = 0\big|_{z\in U_2}$ so that by Cor. 3.5.1:

$$\begin{aligned}
(\mathbf{Ax})[t] &= \int_0^1 e^{-2\pi i\omega t}\phi\left(\hat{\mathbf{S}}(\omega)\right)\cdot\hat{\mathbf{x}}(\omega)d\omega \\
&= \int_0^1 e^{-2\pi i\omega t}\left[\phi(\lambda_-(\omega))\mathbf{P}_-(\omega) + \phi(\lambda_+(\omega))\mathbf{P}_+(\omega)\right] \\
&\quad \cdot\left[\delta(\omega-\omega_1)\mathbf{u}_-(\omega_1) + \delta(\omega-\omega_2)\mathbf{u}_+(\omega_1)\right]d\omega \\
&= e^{-2\pi i\omega_1}\phi(\lambda_-(\omega_1))\mathbf{P}_-(\omega_1)\mathbf{u}_-(\omega_1) + e^{-2\pi i\omega_1}\phi(\lambda_+(\omega_1))\mathbf{P}_+(\omega_1)\mathbf{u}_-(\omega_1) \\
&\quad + e^{-2\pi i\omega_2}\phi(\lambda_-(\omega_2))\mathbf{P}_-(\omega_2)\mathbf{u}_+(\omega_2) + e^{-2\pi i\omega_2}\phi(\lambda_+(\omega_2))\mathbf{P}_+(\omega_2)\mathbf{u}_+(\omega_2).
\end{aligned}$$

The cross-products will disappear since $\mathbf{P}_-(\omega)\mathbf{u}_+(\omega) = 0$ and $\mathbf{P}_-(\omega)\mathbf{u}_+(\omega)$ for all $\omega \in [0,1]$. Then, we need only make sure that $\lambda_-(\omega_1) \in U_1$ and $\lambda_+(\omega_2) \in U_2$ to yield

$$(\mathbf{Ax})[t] = e^{-2\pi i\omega_1}\mathbf{u}_-(\omega_1). \tag{3.33}$$

In summary, it is the separability of the spectrum that allows us to design a filter that discriminates the pure frequency components of the signal.

3.4.1.2 Bandpass filtering with product graph model

Now, we restrict our model to using only product graphs. That is, we want to express the extended graph $\mathcal{G}$ in terms of product graphs as is done in Sandryhaila and Moura [91]. In this formulation, two graphs $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ with respective graph operators $\mathbf{S}_1$ and $\mathbf{S}_2$ are combined, *i.e.* $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$, to yield

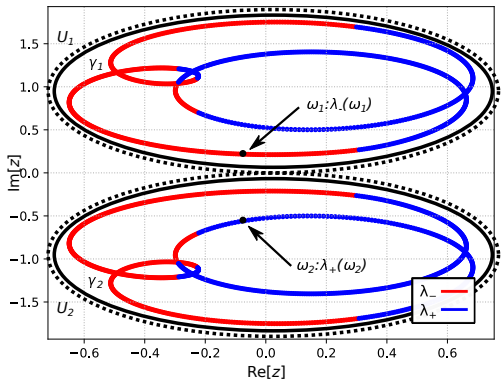


Figure 3.7: Open sets containing the spectrum of $\mathbf{S}$. We can define $U = U_1 \cup U_2$ such that $\Lambda(\mathbf{S}) \subset U$ and $U_1 \cap U_2 = \emptyset$. For defining $\phi(\mathbf{S})$, we illustrate γ_1 and γ_2, Jordan curves enclosing the connected components of the spectrum. We also identify $\lambda_-(\omega - 1)$ and $\lambda_+(\omega_2)$.

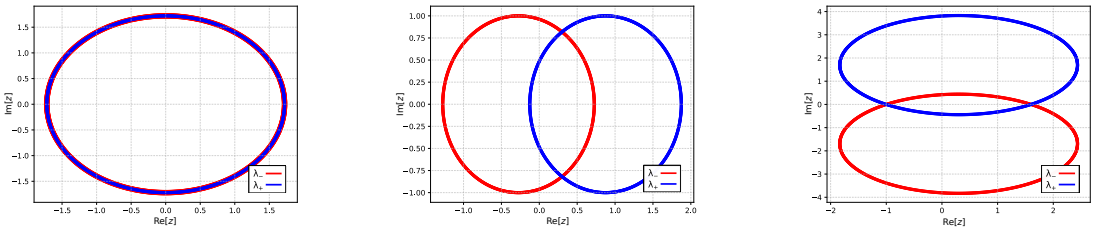
a resultant graph operator according to one of three product rules:

(1) Kronecker product: $\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{S}_2$;

(2) Cartesian product: $\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{S}_2$; or

(3) strong product: $\mathbf{S} = \mathbf{S}_1 \otimes \mathbf{S}_2 + \mathbf{S}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{S}_2$.

In Sandryhaila and Moura [91], the authors propose to filter time-varying graph signals using a product graph of $\mathbf{S}$ and $\mathbf{T}$, where $\mathbf{S} \in \mathcal{B}\left(\mathbb{C}^d\right)$ is the graph operator for the d-node graph and $\mathbf{T} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ is the scalar time-advancement operator (3.2). To implement the factor graph approach, we interpret $\mathbf{K}[0], \ldots, \mathbf{K}[3]$ of Eq. (3.29) as each a graph operator on a d-node graph. For each product graph, the spectrum is not separable, and so we cannot implement a bandpass filter to discriminate the pure frequency components (see Fig. 3.8).



(a) Kronecker product (b) Cartesian product (c) Strong product

Figure 3.8: Spectrum of the product graph operators. In all cases, the spectrum does not admit a separable projection. Note that the spectrum overlaps completely in the Kronecker product graph.

(1) Kronecker product graph For the extended graph with graph operator given by Eq. (3.29), the Kronecker product graph is given by

$$\mathbf{S}_K = \sum_{t=0}^{3} \mathbf{T} \otimes \mathbf{K}[t] = \mathbf{T} \otimes \left(\sum_{t=0}^{3} \mathbf{K}[t] \right). \tag{3.34}$$

It has a spectral representation,

$$\hat{\mathbf{S}}_K(\omega) = e^{2\pi i \omega} \begin{bmatrix} 1/5 & -9/5 \\ 8/5 & 2/5 \end{bmatrix}, \tag{3.35}$$

and spectrum,

$$\Lambda\left(\mathbf{S}_K\right) = \bigcup_{\omega \in [0,1]} \left\{ \frac{3 \pm i\sqrt{287}}{10} e^{2\pi i \omega} \right\}. \tag{3.36}$$

Surprisingly, there is almost no spectral information shared between the Kronecker product graph operator and the graph operator of the proposed approach. The spectrum is plotted in Fig. 3.8a, in which we observe that it is not separated. Any filter defined by a holomorphic function ϕ will have to apply uniformly to the entire spectrum, and so we cannot bandpass the signal for a generic signal. For

special cases, we may be able to define a holomorphic function $\phi : U \to \mathbb{C}$ such that $\phi\left(\lambda_{-}(\omega_1)\right) = 1$ and $\phi\left(\lambda_{-}(\omega_2)\right) = 0$, but this will be the exception and not the rule.

(2) Cartesian product graph The Cartesian product graph for the graph operator given by Eq. (3.29) is given by

$$\mathbf{S}_C = \sum_{t=0}^{3} \mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{K}[t] = \mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \left(\sum_{t=0}^{3} \mathbf{K}[t] \right). \tag{3.37}$$

It has a spectral representation

$$\hat{\mathbf{S}}_C(\omega) = e^{2\pi i \omega} \mathbf{I} + \begin{bmatrix} 1/5 & -9/5 \\ 8/5 & 2/5 \end{bmatrix}, \tag{3.38}$$

and spectrum

$$\Lambda\left(\mathbf{S}\right) = \bigcup_{\omega \in [0,1]} \left\{ \frac{1}{10} \left(3 \pm \sqrt{33} + 10 e^{2\pi i \omega} \right) \right\}. \tag{3.39}$$

We find a similar result in which the spectral information of the two graph operators share almost no information. The spectrum is plotted in Fig. 3.8b. Although the spectrum is not completely overlapping as in the Kronecker product graph, the spectrum is not separable. For the same reasons as in the Kronecker product graph, the Cartesian product graph does not allow us to implement a general bandpass filter.

(3) Strong product graph The strong product graph for our problem is given by

$$\begin{aligned} \mathbf{S}_S &= \mathbf{T} \otimes \left(\sum_{t=0}^{3} \mathbf{K}[t] \right) + \mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \left(\sum_{t=0}^{3} \mathbf{K}[t] \right) \\ &= \mathbf{T} \otimes \left(\mathbf{I} + \sum_{t=0}^{3} \mathbf{K}[t] \right) + \mathbf{I} \otimes \left(\sum_{t=0}^{3} \mathbf{K}[t] \right) \end{aligned} \tag{3.40}$$

It has a spectral representation

$$\hat{\mathbf{S}}_S(\omega) = e^{2\pi i\omega}\begin{bmatrix} 6/5 & -4/5 \\ -8/5 & 7/5 \end{bmatrix} + \begin{bmatrix} 1/5 & -9/5 \\ 8/5 & 2/5 \end{bmatrix}, \tag{3.41}$$

and spectrum

$$\Lambda\left(\mathbf{S}_C\right) = \bigcup_{\omega\in[0,1]} \left\{ \frac{1}{10}\left(3 + 13e^{2\pi i\omega} \pm i\sqrt{287}\left(1 + e^{2\pi i\omega}\right)\right)\right\}. \tag{3.42}$$

Again the spectrum is not separable, and we cannot implement the desired bandpass filter using the strong product graph either. See Fig. 3.8c for the plot.

3.4.1.3 Bandpass filtering with a self-adjoint graph operator

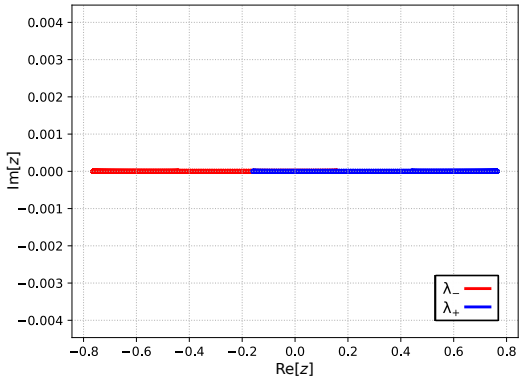


Figure 3.9: Spectrum of self-adjoint graph operator. In this case, the spectrum is restricted to the real line and overlaps.

Here, we use a symmetrization of $\mathbf{S}$, *i.e.* $\mathbf{S}_{\text{sym}} = \frac{1}{2}\left(\mathbf{S} + \mathbf{S}^*\right)$. $\mathbf{S}_{\text{sym}}$ has a spectral representation

$$\hat{\mathbf{S}}(\omega) = \begin{bmatrix} \frac{1}{10}\cos 4\pi\omega & -\frac{4}{10}e^{2\pi i\omega} + \frac{3}{10}e^{-6\pi i\omega} \\ -\frac{4}{10}e^{-2\pi i\omega} + \frac{3}{10}e^{6\pi i\omega} & \frac{2}{10}\cos 4\pi\omega, \end{bmatrix} \tag{3.43}$$

and a pointwise eigendecomposition

$$\lambda_{\pm}(\omega) = \frac{1}{20}\left(6\cos 4\pi\omega \pm \sqrt{102 - 94\cos 8\pi\omega}\right) \tag{3.44}$$

and

$$\mathbf{u}_{\pm}(\omega) = \begin{bmatrix} 2\cos 4\pi\omega \pm \sqrt{102 - 94\cos 8\pi\omega} \\ 8e^{-2\pi i\omega} - 6e^{6\pi i\omega} \end{bmatrix}. \tag{3.45}$$

The spectrum is plotted in Fig. 3.9. Note that the spectrum $\cup_{\omega\in[0,1]}\{\lambda_{\pm}(\omega)\}$ is the real projection of $\Lambda(\mathbf{S})$.

Clearly, $\Lambda(\mathbf{S}_{\text{sym}})$ is not separable, but that does not definitively mean that we cannot discriminate the pure frequency components of the candidate signal. Self-adjoint operators accommodate an ever larger class of functions via functional calculus, namely all Borel measurable functions on $\Lambda(\mathbf{S}_{\text{sym}})$ (see *e.g.* [86]). Accordingly, we could define a bump function $h : U \to \mathbb{C}$, the indicator of any Borel subset $B \subset \Lambda(\mathbf{S}_{\text{sym}})$,

$$h(z) = \begin{cases} 1 & z \in B \\ 0 & \text{o.w.} \end{cases}. \tag{3.46}$$

This means that it may still be possible to discriminate the pure frequency components of the signal $\mathbf{x}$.

However, we can choose a signal so as to make it impossible for a filter defined via the Borel functional calculus on $\mathbf{S}_{\text{sym}}$ to discriminate the pure frequency components of $\mathbf{x}$. To do this, we must find ω_1, ω_2 such that $\lambda_-(\omega_1) = \lambda_+(\omega_2)$. Then, for any $h : \mathbb{R} \to \{0, 1\}$ such that $h(\lambda_-(\omega_1)) = 1$, $h(\lambda_+(\omega_2)) = 1$ as well. Consider h

such that $h(\lambda_+(\omega_1)) = h(\lambda_-(\omega_1)) = 1$ and $h(\lambda_-(\omega_2))$, then we have

$$\begin{aligned}
(h(\mathbf{S}_{\text{sym}})\mathbf{x})\,[t] &= \int_0^1 e^{-2\pi i\omega t}\,[h(\lambda_-(\omega))\mathbf{P}_-(\omega) + h(\lambda_+(\omega))\mathbf{P}_+(\omega)] \\
&\quad \cdot [\delta(\omega-\omega_1)\mathbf{u}_-(\omega_1) + \delta(\omega-\omega_2)\mathbf{u}_+(\omega_1)]\,d\omega \\
&= e^{-2\pi i\omega_1} h(\lambda_-(\omega_1))\mathbf{P}_-(\omega_1)\mathbf{u}_-(\omega_1) + e^{-2\pi i\omega_1} h(\lambda_+(\omega_1))\mathbf{P}_+(\omega_1)\mathbf{u}_-(\omega_1) \\
&\quad + e^{-2\pi i\omega_2} h(\lambda_-(\omega_2))\mathbf{P}_-(\omega_2)\mathbf{u}_+(\omega_2) + e^{-2\pi i\omega_2} h(\lambda_+(\omega_2))\mathbf{P}_+(\omega_2)\mathbf{u}_+(\omega_2) \\
&= e^{-2\pi i\omega_1}\mathbf{u}_-(\omega_1) + e^{-2\pi i\omega_2}\mathbf{P}_+(\omega_2)\mathbf{u}_+(\omega_2).
\end{aligned}$$

As the $\lambda_-(\omega) = \lambda_+(\omega)$ on a measurable subset of $\Lambda(\mathbf{S}_{\text{sym}})$ (see Fig. 3.9), we can find such a $\omega_1, \omega_2 \in [0,1]$ as is the case for ω_1, ω_2 in Fig. 3.7.

3.4.2 Discriminative representations

In this section, we apply shift-invariant filtering for the purpose of building invariant representations. This application is loosely inspired by the scattering transform of Mallat [75] and recently extended to the graph domain by Zou and Lerman [114] and Gama, *et al.* [47]. We begin with a few definitions.

Definition 29. Random variables X, Y taking values in a Banach space $\mathcal{X}$ are called *separable* if there exists a continuous function $f : \mathcal{X} \to [0,1]$ such that $f(X) \neq f(Y)$ almost surely. They are called *strongly separable* if there exists a Lipschitz continuous function $f : \mathcal{X} \to [0,1]$ such that $f(X) \neq f(Y)$ almost surely. They are called *linearly separable* if there exists a $f \in \mathcal{X}^*$ such that $f(X) \neq f(Y)$ almost surely.

Definition 30. Let $f : \mathcal{X} \to \mathcal{Y}$ be a function between two Banach spaces, and let

$\mathcal{G} = \{g_i : \mathcal{X} \to \mathcal{X}\}$ be a collection of functions on $\mathcal{X}$. $\mathcal{G}$ is said to be a *symmetry* of f if for every $x \in \mathcal{X}$ and every $g_i \in \mathcal{G}$, $(f \circ g_i)(x) = f(x)$.

Definition 31. Let $\mathcal{X}$ be a Banach space and $\mathcal{G} = \{g_i : \mathcal{X} \to \mathcal{X}\}$ be a collection of functions on $\mathcal{X}$. A map $h : \mathcal{X} \to \mathcal{Y}$ to a Banach space $\mathcal{Y}$ is said to *linearize* $\mathcal{G}$ if there exists a $C > 0$ such that for all $x \in \mathcal{X}$ and $g_i \in \mathcal{G}$,

$$\|(h \circ g_i)(x) - h(x)\|_{\mathcal{Y}} \leq C \cdot |g_i|_{\mathcal{G}} \cdot \|x\|_{\mathcal{X}}, \tag{3.47}$$

where $|g_i|_{\mathcal{G}}$ is a metric measuring the difference of g_i from $\mathbf{I}$.

Finally, we introduce new notation: For an open set $U \subset \mathbb{C}$, we denote the set of holomorphic functions $f : U \to \mathbb{C}$, $\mathcal{H}(U)$ and the set of holomorphic functions with a multiplicative inverse, $(f \cdot f^{-1})(z) = z$ for all $z \in U$, $\mathcal{L}(U) \subset \mathcal{H}(U)$. Now, we can state our main result.

Theorem 3.6. *Let $W_1, W_2 \in \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$ be random variables strongly separated by a function $f : \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right) \to [0,1]$ and $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)\right)$ such that $\Lambda(\mathbf{S}) \subset U$ for an open set $U \subset \mathbb{C}$. Then, for every $\varepsilon > 0$, there exists a $\delta > 0$ such that for every finite subset $\left\{\psi_i \in \mathcal{L}(U) : |\psi_i|_{\mathcal{L}(U)} < \delta, i \in \mathcal{I}\right\} \subset \mathcal{L}(U)$, there exists a Lipschitz continuous function $\Phi : \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right) \to \ell^2(\mathcal{I})$ and a positive constant $C > 0$ such that*

$$\begin{aligned} \inf_{\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}} \mathbb{E} \frac{\|\Phi(\psi_j(\mathbf{S}) W_1) - \Phi(\psi_k(\mathbf{S}) W_2)\|}{\frac{1}{2}\sum_{\ell=1,2} \|\Phi(\psi_j(\mathbf{S}) W_\ell) - \Phi(\psi_k(\mathbf{S}) W'_\ell)\|} \\ \geq C \cdot \mathbb{E} \frac{|f(W_1) - f(W_2)|}{\frac{1}{2}\sum_{\ell=1,2} \|W_\ell - W'_\ell\|} - \varepsilon. \end{aligned} \tag{3.48}$$

Our result is motivated by the following problem set-up. Suppose that there exists a classification function $c : \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right) \to \{0,1\}$ and strongly separable random variables $W_1, W_2 \in \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$. The classification function classifies W_1 and

W_2, *i.e.* $c(W_1) = 0$ and $c(W_2) = 1$ almost surely. Further, there exists a collection of bounded linear operators, $\{\psi_i(\mathbf{S}) : \psi_i \in \mathcal{L}(U), i \in \mathcal{I}\}$, that are a symmetry of c, *i.e.* $c(\psi_i(\mathbf{S})W_1) = 0$ and $c(\psi_j(\mathbf{S})W_2) = 1$ almost surely for all $i, j \in \mathcal{I}$. Can we approximate c? Figure 3.10 depicts the challenge of the problem. In the result, we show that Φ approximately recovers the discriminability of W_1, W_2, which are strongly separable prior to arbitrary linear transformation from the group, $\{\psi_i(\mathbf{S}) : \psi_i \in \mathcal{L}(U), i \in \mathcal{I}\}$.

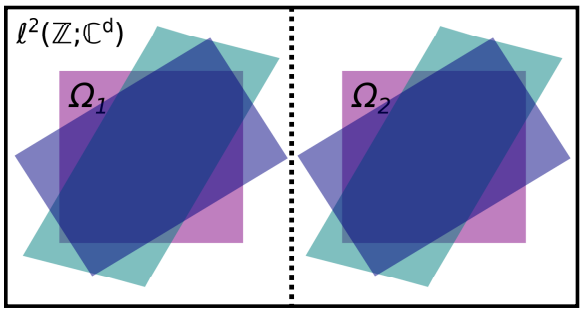


Figure 3.10: Depiction of Thm. 3.6. Let $\Omega_1 = \text{supp}(W_1)$ and $\Omega_2 = \text{supp}(W_2)$ Then, $\psi_j(\mathbf{S})W_1$ and $\psi_k(\mathbf{S})W_2$ stretch and rotate the support. Thus, the function which originally separated W_1 and W_2 may no longer.

Before continuing with the proof of Thm. 3.6, we will propose a candidate Φ and also a metric on $\mathcal{L}(U)$. Since W_1 and W_2 are strongly separable, there exists a Lipschitz continuous function $f : \ell^2(\mathbb{Z};\mathbb{C}^d) \to [0,1]$ such that $f(W_1) \neq f(W_2)$ almost surely. Let us define $\Phi : \ell^2(\mathbb{Z};\mathbb{C}^d) \to \ell^2(\mathcal{I})$ by the map

$$\mathbf{w} \mapsto (f(\phi_i(\mathbf{S})\,\mathbf{w}))_{i\in\mathcal{I}}, \tag{3.49}$$

where $\phi_i : U \to \mathbb{C}$ come from the set $\{\phi_i \in \mathcal{H}(U) : (\phi_i \cdot \psi_i)(z) = z, \forall z \in U\}$. See

Fig. 3.11 for a visualization of Φ. We define the following metric on $(\mathcal{L}(U), \mathbf{S})$:

$$|\psi|_{\mathcal{L}(U)} := \|\psi(\mathbf{S}) - \mathbf{I}\|. \tag{3.50}$$

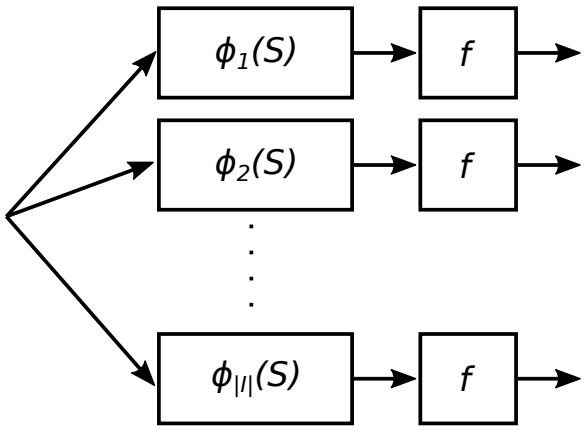


Figure 3.11: Visualization of $\Phi : \ell^2(\mathbb{Z}; \mathbb{C}^d) \to \ell^2(\mathcal{I})$. From this visualization, we see that Φ is a parallel array of linear filters, $(\phi_i(\mathbf{S}))_{i=1,\ldots,|\mathcal{I}|}$. Then, we apply $f : \ell^2(\mathbb{Z}; \mathbb{C}^d) \to [0,1]$ to all paths of the filtered signal.

We will also make use of the following intermediate results.

Lemma 3.7 (Lipschitz)**.** *Let $\mathbf{S} \in \mathcal{B}(\ell^2(\mathbb{Z}; \mathbb{C}^d))$ such that $\Lambda(\mathbf{S}) \subset U$ for an open set $U \subset \mathbb{C}$. For a Lipschitz continuous function $f : \ell^2(\mathbb{Z}; \mathbb{C}^d) \to [0,1]$ and any finite set of holomorphic functions $\{\phi_i : U \to \mathbb{C} : i \in \mathcal{I}\}$, the map $\Phi : \ell^2(\mathbb{Z}; \mathbb{C}^d) \to \ell^2(\mathcal{I})$,*

$$\mathbf{w} \mapsto (f(\phi_i(\mathbf{S})\mathbf{w}))_{i \in \mathcal{I}},$$

is Lipschitz continuous.

Proof of Lem. 3.7. To show that Φ is Lipschitz, note that for any $\mathbf{w}, \mathbf{w}' \in \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$,

$$\begin{aligned}\|\Phi(\mathbf{w}) - \Phi(\mathbf{w}')\| &= \left(\sum_{i\in\mathcal{I}} \left|f\left(\phi_i\left(\mathbf{S}\right)\mathbf{w}\right) - f\left(\phi_i\left(\mathbf{S}\right)\mathbf{w}'\right)\right|^2\right)^{1/2} \\ &\leq \|f\|_L \cdot \left(\sum_{i\in\mathcal{I}} \left\|\phi_i\left(\mathbf{S}\right)\mathbf{w} - \phi_i\left(\mathbf{S}\right)\mathbf{w}'\right\|^2\right)^{1/2} \\ &\leq \|f\|_L \cdot \left(\sum_{i\in\mathcal{I}} \left\|\phi_i\left(\mathbf{S}\right)\right\|^2\right)^{1/2} \cdot \left\|\mathbf{w} - \mathbf{w}'\right\|.\end{aligned}$$

The first inequality follows from the Lipschitz continuity of f. The second inequality follows from $\phi_i(\mathbf{S})$ defining a bounded linear operator. By the spectral mapping theorem, $\|\phi_i(\mathbf{S})\| = \max\left|\phi_i(\Lambda\left(\mathbf{S}\right))\right|$. Since $|\mathcal{I}| < \infty$, $\left(\sum_{i\in\mathcal{I}} \left\|\phi_i\left(\mathbf{S}\right)\right\|^2\right)^{1/2} < \infty$. Therefore,

$$\|\Phi\|_L = \|f\|_L \cdot \left(\sum_{i\in\mathcal{I}} \left\|\phi_i\left(\mathbf{S}\right)\right\|^2\right)^{1/2}. \tag{3.51}$$

□

Lemma 3.8 (Linearization)**.** *Let* $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)\right)$ *such that* $\Lambda\left(\mathbf{S}\right) \subset U$ *for an open set* $U \subset \mathbb{C}$*. Consider a finite subset* $\Psi = \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\} \subset \mathcal{L}(U)$*. Then,* Ψ *is linearized by* $\Phi : \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right) \to \ell^2\left(\mathcal{I}\right)$ *as defined in Eq. (3.49).*

Proof of Lem. 3.8. Fix a $\psi_j \in \Psi$. For any $\mathbf{w} \in \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$,

$$\left\|\Phi\left(\psi_j\left(\mathbf{S}\right)\mathbf{w}\right) - \Phi\left(\mathbf{w}\right)\right\| \leq \|\Phi\|_L \cdot \|\psi_j(\mathbf{S})\mathbf{w} - \mathbf{w}\| \leq \|\Phi\|_L \cdot \left\|\psi_j\left(\mathbf{S}\right) - \mathbf{I}\right\| \cdot \|\mathbf{w}\|.$$

The first inequality follows from Lemma 3.7, and the second inequality follows from $\psi_j\left(\mathbf{S}\right) - \mathbf{I}$ defining a bounded linear operator. We can simplify this notation by using the metric defined in Eq. (3.50),

$$\left\|\Phi\left(\psi_j\left(\mathbf{S}\right)\mathbf{w}\right) - \Phi\left(\mathbf{w}\right)\right\| \leq \left|\psi_j\right|_{\mathcal{L}(U)} \cdot \|\Phi\|_L \cdot \|\mathbf{w}\|.$$

As the choice of ψ_j was arbitrary, the result holds for all $\psi_j \in \Psi$. □

Lemma 3.9. *For $a, b > 0$,*

$$\frac{a-x}{b+x} \geq \frac{a}{b} - x \cdot \frac{a+b}{b^2},$$

for all $x > -b$.

Proof. To prove the statement, note that $\frac{a-x}{b+x}\Big|_{x>-b}$ is a convex function. To get the desired result, we take a tangent line at $x = 0$. □

Now, we proceed with the proof of Thm. 3.6.

Proof of Thm. 3.6. Since we want to find a lower bound, we can lower bound the numerator and upper bound the denominator. We begin with the denominator.

(Denominator) Fix $\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}$. We can upper bound the denominator with a telescoping argument:

$$\begin{aligned}
&\frac{1}{2}\sum_{\ell=1,2} \|\Phi\left(\psi_j\left(\mathbf{S}\right) W_\ell\right) - \Phi\left(\psi_k\left(\mathbf{S}\right) W'_\ell\right)\| \\
&= \frac{1}{2}\sum_{\ell=1,2} \|\Phi\left(\psi_j\left(\mathbf{S}\right) W_\ell\right) - \Phi\left(\psi_j\left(\mathbf{S}\right) W'_\ell\right) + \Phi\left(\psi_j\left(\mathbf{S}\right) W'_\ell\right) - \Phi\left(\psi_k\left(\mathbf{S}\right) W'_\ell\right)\| \\
&\leq \underbrace{\frac{1}{2}\sum_{\ell=1,2} \|\Phi\left(\psi_j\left(\mathbf{S}\right) W_\ell\right) - \Phi\left(\psi_j\left(\mathbf{S}\right) W'_\ell\right)\|}_{\text{(a)}} \\
&\quad + \underbrace{\frac{1}{2}\sum_{\ell=1,2} \|\Phi\left(\psi_j\left(\mathbf{S}\right) W'_\ell\right) - \Phi\left(\psi_k\left(\mathbf{S}\right) W'_\ell\right)\|}_{\text{(b)}}.
\end{aligned}$$

We have used the triangle inequality for the inequality. Now, we separately upper bound terms (a) and (b).

For term (a), we have

$$\frac{1}{2}\sum_{\ell=1,2}\left\|\Phi\left(\psi_j\left(\mathbf{S}\right)W_\ell\right)-\Phi\left(\psi_j\left(\mathbf{S}\right)W'_\ell\right)\right\| \leq \left\|\Phi\right\|_L \cdot \frac{1}{2}\sum_{\ell=1,2}\left\|\psi_j\left(\mathbf{S}\right)W_\ell-\psi_j\left(\mathbf{S}\right)W'_\ell\right\|$$
$$\leq \left\|\Phi\right\|_L \cdot \left\|\psi_j(\mathbf{S})\right\| \cdot \frac{1}{2}\sum_{\ell=1,2}\left\|W_\ell-W'_\ell\right\|.$$

The first equality follows from Lemma 3.7, and the second inequality follows from $\psi_j\left(\mathbf{S}\right)$ defining a bounded linear operator on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$. Note that we have our desired denominator multiplied by a constant, $C_1 = \left\|\Phi\right\|_L \cdot \left\|\psi_j(\mathbf{S})\right\|$.

Now, for term (b), we have

$$\frac{1}{2}\sum_{\ell=1,2}\left\|\Phi\left(\psi_j\left(\mathbf{S}\right)W'_\ell\right)-\Phi\left(\psi_k\left(\mathbf{S}\right)W'_\ell\right)\right\| \leq \left\|\Phi\right\|_L \cdot \frac{1}{2}\sum_{\ell=1,2}\left\|\psi_j\left(\mathbf{S}\right)W'_\ell-\psi_k\left(\mathbf{S}\right)W'_\ell\right\|$$
$$\leq \left\|\Phi\right\|_L \cdot \left\|\psi_j\left(\mathbf{S}\right)-\psi_k\left(\mathbf{S}\right)\right\| \cdot \frac{1}{2}\sum_{\ell=1,2}\left\|W'_\ell\right\|.$$

The first inequlity follows from Lemma 3.7. The second equality follows from the algebra homomorphism property of the holomorphic functional calculus so that $\left(\psi_j-\psi_k\right)\left(\mathbf{S}\right)$ defines a bounded linear operator on $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$. We can further bound this using the following observation:

$$\begin{aligned}\left\|\left(\psi_j-\psi_k\right)\left(\mathbf{S}\right)\right\| &= \left\|\left(\psi_j-1+1-\psi_k\right)\left(\mathbf{S}\right)\right\| \\ &\leq \left\|\psi_j\left(\mathbf{S}\right)-\mathbf{I}\right\|+\left\|\psi_k\left(\mathbf{S}\right)-\mathbf{I}\right\| \\ &\leq 2\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)},\left|\psi_k\right|_{\mathcal{L}(U)}\right).\end{aligned}$$

Putting this together, we have

$$\frac{1}{2}\sum_{\ell=1,2}\left\|\Phi\left(\psi_j\left(\mathbf{S}\right)W_\ell\right)-\Phi\left(\psi_k\left(\mathbf{S}\right)W'_\ell\right)\right\|$$
$$\leq C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\left\|W_\ell-W'_\ell\right\|+2\cdot\left\|\Phi\right\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)},\left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\left\|W'_\ell\right\|.$$

(Numerator) With the same $\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}$, we consider the numerator. We can derive a lower bound with a telescoping argument:

$$\begin{aligned}
&\|\Phi(\psi_j(\mathbf{S})W_1) - \Phi(\psi_k(\mathbf{S})W_2)\| \\
&= \|\Phi(\psi_j(\mathbf{S})W_1) - \Phi(\psi_j(\mathbf{S})W_2) + \Phi(\psi_j(\mathbf{S})W_2) - \Phi(\psi_k(\mathbf{S})W_2)\| \\
&\geq \underbrace{\|\Phi(\psi_j(\mathbf{S})W_1) - \Phi(\psi_j(\mathbf{S})W_2)\|}_{\text{(c)}} - \underbrace{\|\Phi(\psi_j(\mathbf{S})W_2) - \Phi(\psi_k(\mathbf{S})W_2)\|}_{\text{(d)}}.
\end{aligned}$$

We can now proceed with bounding terms (c) and (d) independently.

For term (c), using the finite-dimensional norm relationship $\|\cdot\|_2 \geq \|\cdot\|_\infty$, we have

$$\begin{aligned}
&\|\Phi(\psi_j(\mathbf{S})W_1) - \Phi(\psi_j(\mathbf{S})W_2)\| \\
&= \left(\sum_{i\in\mathcal{I}} |f(\phi_i(\mathbf{S})\psi_j(\mathbf{S})W_1) - f(\phi_i(\mathbf{S})\psi_j(\mathbf{S})W_2)|^2\right)^{1/2} \\
&= \left(\sum_{i\in\mathcal{I}} |f((\phi_i\cdot\psi_j)(\mathbf{S})W_1) - f((\phi_i\cdot\psi_j)(\mathbf{S})W_2)|^2\right)^{1/2} \\
&\geq \max_{i\in\mathcal{I}} |f((\phi_i\cdot\psi_j)(\mathbf{S})W_1) - f((\phi_i\cdot\psi_j)(\mathbf{S})W_2)|.
\end{aligned}$$

Since $\psi_j \in \mathcal{L}(U)$, by the construction of Φ, there is a ϕ_i, $i \in \mathcal{I}$, such that $(\phi_i \cdot \psi_j)(z) = z$. Therefore,

$$\max_{i\in\mathcal{I}} |f((\phi_i\cdot\psi_j)(\mathbf{S})W_1) - f((\phi_i\cdot\psi_j)(\mathbf{S})W_2)| \geq |f(W_1) - f(W_2)|.$$

For term (d), we have a similar result as from term (b) in the denominator:

$$\|\Phi(\psi_j(\mathbf{S})W_2) - \Phi(\psi_k(\mathbf{S})W_2)\| \leq 2\cdot\|\Phi\|_L \cdot \max\left(|\psi_j|_{\mathcal{L}(U)}, |\psi_k|_{\mathcal{L}(U)}\right)\cdot\|W_2\|.$$

We note that the choice to bound W_2 in term (d) was arbitrary, so we can choose to bound $\arg\min(\|W_1\|, \|W_2\|)$. Then, using $\min(\|W_1\|, \|W_2\|) \leq \frac{1}{2}\sum_{\ell=1,2}\|W_\ell\|$,

we have that

$$\begin{aligned}
&\|\Phi\left(\psi_j\left(\mathbf{S}\right)W_1\right) - \Phi\left(\psi_k\left(\mathbf{S}\right)W_2\right)\| \\
&\geq \left|f\left(W_1\right) - f\left(W_2\right)\right| - 2\cdot\|\Phi\|_L \cdot \max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|.
\end{aligned}$$

(Combined) Putting together the respective results from the numerator and denominator, we have for a fixed $\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}$,

$$\begin{aligned}
&\inf_{\psi_j,\psi_k\in\{\psi_i\in\mathcal{L}(U):i\in\mathcal{I}\}} \mathbb{E}\frac{\|\Phi\left(\psi_j\left(\mathbf{S}\right)W_1\right) - \Phi\left(\psi_k\left(\mathbf{S}\right)W_2\right)\|}{\frac{1}{2}\sum_{\ell=1,2}\|\Phi\left(\psi_j\left(\mathbf{S}\right)W_\ell\right) - \Phi\left(\psi_k\left(\mathbf{S}\right)W_\ell'\right)\|} \\
&\geq \mathbb{E}\frac{\left|f\left(W_1\right) - f\left(W_2\right)\right| - 2\cdot\|\Phi\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|}{C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell - W_\ell'\| + 2\cdot\|\Phi\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|}.
\end{aligned}$$

Note that the second term in the numerator and denominator are equal and linear in $\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)$. Using Lemma 3.9, we have

$$\begin{aligned}
&\mathbb{E}\frac{\left|f\left(W_1\right) - f\left(W_2\right)\right| - 2\cdot\|\Phi\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|}{C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell - W_\ell'\| + 2\cdot\|\Phi\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|} \\
&\quad\geq \mathbb{E}\frac{\left|f\left(W_1\right) - f\left(W_2\right)\right|}{C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell - W_\ell'\|} \\
&\qquad - 2\cdot\|\Phi\|_L\cdot\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right)\cdot\mathbb{E}\left(\frac{1}{2}\sum_{\ell=1,2}\|W_\ell'\|\right) \\
&\qquad\cdot\frac{\left|f\left(W_1\right) - f\left(W_2\right)\right| + C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell - W_\ell'\|}{\left(C_1\cdot\frac{1}{2}\sum_{\ell=1,2}\|W_\ell - W_\ell'\|\right)^2}.
\end{aligned}$$

By the statement of the lemma, $\max\left(\left|\psi_j\right|_{\mathcal{L}(U)}, \left|\psi_k\right|_{\mathcal{L}(U)}\right) < \delta$, and so we can choose $\delta > 0$ small enough to make the second term less than ε.

To finish the argument, we note that the choice of $\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}$ was arbitrary, so the argument holds for all $\psi_j, \psi_k \in \{\psi_i \in \mathcal{L}(U) : i \in \mathcal{I}\}$. □

We have shown that we can use shift-invariant filtering to construct representations that retain the underlying discriminability of the signal for a large class of linear symmetries.

Chapter 4: Learning the graph structure of stochastic processes

4.1 Introduction

In Chapter 3, we developed filters for time-varying graph signals in $\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)$ where the graph operator $\mathbf{S} \in \mathcal{B}\left(\ell^2\left(\mathbb{Z};\mathbb{C}^d\right)\right)$ was known *a priori* . In practice, we may need to estimate the underlying extended graph $\mathcal{G}$ and graph operator $\mathbf{S}$ from observations. For this reason, we consider finite observations, for which we will want to estimate graph operators. Estimating the graph operator can be posed as a parameter estimation problem as in the following example.

Example 1 (Instantaneous covariance)**.** Suppose that we observe a vector-valued sequence $\mathbf{x} = (\mathbf{x}[t])_{t=1,\dots,T}$ with $\mathbf{x}[t] \in \mathbb{R}^d$ for all $t = 1,\dots,T$, in which observations have no temporal dependence. Therefore, the time-series can be treated as independent and identically distributed observations. If $\mathbf{x}[t] \sim_{\text{iid}} \mathcal{N}(\mathbf{0},\mathbf{K})$, where $\mathbf{K} \in \mathbb{R}^{d\times d}$ is unknown, then we can estimate $\mathbf{K}$ from the observations. $\mathbf{K}$ captures the spatial dependence of the observations and is thus a viable graph operator on $\mathcal{V} = \{1,\dots,d\}$. We can infer the underlying graph by assigning an edge between nodes i and j if $|K_{i,j}| > \tau$ for some threshold $\tau > 0$. The MLE for this model is

given by

$$\arg\min_{\mathbf{K}\in\mathbb{R}^{d\times d}} -\frac{1}{2}\log\det\left(\mathbf{K}\right) + \frac{1}{2T}\sum_{t=1}^{T}\left\langle \mathbf{x}[t], \mathbf{K}^{-1}\mathbf{x}[t]\right\rangle, \tag{4.1}$$

and the minimizer is $\mathbf{K} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}[t]\mathbf{x}^{*}[t]$. We could then extend $\mathbf{K}$ to a graph operator on a time-varying graph signal using a product graph formulation as we did in Sec. 3.4.

We can employ a similar method for estimating autoregressive graph operators using the estimator in Sec. 2.5. However, the graph estimation problem can be complicated by confounding processes. We use another example to illustrate this point.[1]

Example 2 (Confounding autoregressive processes)**.** Suppose that we observe a time-series $(\mathbf{x}[t])_{t=1,\ldots,T}$, $\mathbf{x}[t]\in\mathbb{R}^{d}$. If we partition $\{1,\ldots,T\}$ into r sets, $\{1,\ldots,T_1\}$, $\{T_1+1,\ldots,T_2\}$, $\ldots$, $\{T_{r-1}+1,\ldots,T\}$, where $T_1 < T_2 < \ldots < T_{r-1} < T$, then $\mathbf{x}[t]$ for $t\in\{T_{j-1}+1,\ldots,T_j\}$ is generated according to the recurrence relation,

$$\mathbf{x}[t] = \left(\mathbf{K}_0[1] + \mathbf{K}_j[1]\right)\mathbf{x}[t-1] + \cdots + \left(\mathbf{K}_0[m] + \mathbf{K}_j[m]\right)\mathbf{x}[t-m] + \mathbf{n}[t],$$

with $\mathbf{n}[t] \sim_{\text{iid}} \mathcal{N}\left(\mathbf{0},\mathbf{I}\right)$ for each $j=1,\ldots,r$ with $\mathbf{K}_j[s]\in\mathcal{B}\left(\mathbb{R}^{d}\right)$ for all $j=0,\ldots,r$ and $s=1,\ldots,m$. That is to say that there is one persistent autoregressive process given by the matrix symbol $(\mathbf{K}_0[t])_{t=1,\ldots,m}$ and a different confounding process $(\mathbf{K}_j[t])_{t=1,\ldots,m}$ present in each observation window $\{T_{j-1}+1,\ldots,T_j\}$.

Estimating the parameters of the model in Example 2 could be a difficult computational problem without considerable *a priori* knowledge of

[1] This work grew out of an observation made in a workshop presentation about learning graph operators [17].

$\left\{(\mathbf{K}_j[t])_t : j = 0, \ldots, r\right\}$. More practically, these confounding processes may be incidental so that we want to estimate $(\mathbf{K}_0[t])_t$ in a robust way. This is a viable model for observed network activity in the brain and team interaction as presented in Secs. 1.1 and 1.2. Multiple causal processes may be operating in parallel. Two regions of the brain may be interacting for a visual processing task, while different regions of the brain are coordinating movement, and the resultant activity would reflect both processes. In teams, this may correspond to a sub-team planning future actions and another sub-team rehearsing tasks, and the observed activity may appear to be a single coordinated action, when it really reflects incidental processes.

If we want to detect the presence of a particularly important causal process, then the confounding processes should be a symmetry for the detection function. To make this more precise, let d be a detection function, $(\mathbf{x}[t])_t \mapsto \{0, 1\}$, where $d((\mathbf{x}[t])_t) = 1$ if $(\mathbf{x}[t])_t$ is causally generated by $(\mathbf{K}_0[t])_t$ and 0 otherwise. By a symmetry, we mean $d((\mathbf{x}[t])_t) = 1$ if $(\mathbf{x}[t])_t$ is causally generated by $(\mathbf{K}_0[t] + \mathbf{K}_j[t])_t$ for any $j = 1, \ldots, r$ (and 0 otherwise). This has implications for how to design filters in accordance with Chapter 3, but here we address how we should estimate the graph operators in the presence of confounding processes.

Let us modify the generative model of Example 2. Let $\beta_1, \ldots, \beta_r \in \mathbb{R}$ be random variables and $(\mathbf{x}[t])_t$ with $\mathbf{x}[t] \in \mathbb{R}^d$ be causally generated by

$$\mathbf{x}[t] = \left(\sum_{j=0}^{r} \beta_j \mathbf{K}_j[1]\right) \mathbf{x}[t-1] + \cdots + \left(\sum_{j=0}^{r} \beta_j \mathbf{K}_j[m]\right) \mathbf{x}[t-m] + \mathbf{n}[t], \tag{4.2}$$

with $\mathbf{n}[t] \sim_{\text{iid}} \mathcal{N}(\mathbf{0}, \mathbf{I})$. If we use the matrix-vector form of Eq. (2.37), then we have

$$\mathbf{Y} = \sum_{j=0}^{r} \beta_j \mathbf{X} \begin{bmatrix} \mathbf{K}_j^*[1] \\ \vdots \\ \mathbf{K}_j^*[m] \end{bmatrix} + \mathbf{N}. \tag{4.3}$$

Equation (4.3) begins to look like a sparse coding problem, especially if we suppose that $\left|\text{supp}\left((\beta_j)_{j=1,\ldots,r}\right)\right| = s \ll r$. It is this observation that inspires our work.

We propose to learn atomic autoregressive processes from multiple independent observations. We consider two methods for estimating the atomic autoregressive components. One is based on a two-stage process in which we estimate the autoregressive coefficients of each independent observation using Eq. (2.39) and then use existing dictionary learning results recounted in Sec. 2.6 to disambiguate the atomic components. The second attempts to directly solve for the atomic components using an alternating minimization algorithm, an approach similar in spirit to the dictionary refinement of Arora, *et al.* [9] and Agarwal, *et al.* [2].

This problem shares common elements with any model that depends on a signal decomposition with a common decomposition across subjects or modalities. This problem has instigated considerable research in neuroimaging due to a reliance on analyses of independent components [28], where the goal is to find common independent components across subjects or imaging modalities. Joint signal representation also arises in multiview clustering in which multiple images of an object are encoded in a hopefully common subspace via simultaneous nonnegative matrix factorization [8, 66, 71]. Notably, the problem shares a similar time-based bilinear structure with compressve sensing of video [93, 94], in which the linear dynamics and sequential

states of video can be estimated from compressive measurements. It differs from recent efforts to solve bi-convex problems via lifting, *e.g.* [7, 69, 70], as we attempt to solve the problem in its natural domain, as in Aghasi, *et al.* [3] and Sun, *et al.* [105].

4.1.1 Problem

Recall the m-order autoregressive model of Section 2.5 for a sequence $(\mathbf{x}[t])_{t=1,\dots,T+m}$ in matrix-vector form:

$$\underbrace{\begin{bmatrix} \mathbf{x}^*[1+m] \\ \vdots \\ \mathbf{x}^*[T+m] \end{bmatrix}}_{\mathbf{Y}\in\mathbb{R}^{T\times d}} = \underbrace{\begin{bmatrix} \mathbf{x}^*[m] & \cdots & \mathbf{x}^*[1] \\ & \ddots & \vdots \\ \vdots & & \mathbf{x}^*[m] \\ & \ddots & \\ \mathbf{x}^*[T] & & \vdots \\ \vdots & \ddots & \\ \mathbf{x}^*[T+m-1] & \cdots & \mathbf{x}^*[T] \end{bmatrix}}_{\mathbf{X}\in\mathbb{R}^{T\times(m\cdot d)}} \underbrace{\begin{bmatrix} \mathbf{K}^*[1] \\ \vdots \\ \mathbf{K}^*[m] \end{bmatrix}}_{\mathbf{A}\in\mathbb{R}^{(m\cdot d)\times d}} + \underbrace{\begin{bmatrix} \mathbf{n}^*[1+m] \\ \vdots \\ \mathbf{n}^*[T+m] \end{bmatrix}}_{\mathbf{N}\in\mathbb{R}^{T\times d}}. \quad (4.4)$$

We want to find a sparse encoding of the columns $\{\mathbf{A}_i \in \mathbb{R}^{(m\cdot d)} : i = 1,\dots,d\}$. That is,

$$\mathbf{A} = \underbrace{\begin{bmatrix} \mathbf{d}_1 & \cdots & \mathbf{d}_r \end{bmatrix}}_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}} \underbrace{\begin{bmatrix} c_{1,1} & \cdots & c_{1,d} \\ \vdots & \ddots & \vdots \\ c_{r,1} & \cdots & c_{r,d} \end{bmatrix}}_{\mathbf{C}\in\mathbb{R}^{r\times d}}. \quad (4.5)$$

Here, $\mathbf{d}_j \in \mathbb{R}^{(m\cdot d)}$ for $j = 1,\dots,r$ are the dictionary atoms.

A single time-series is not sufficient to learn a sparse encoding, and so we

consider repeated independent observations of time-series,

$$\mathbf{Y}_{(k)} = \mathbf{X}_{(k)}\mathbf{D}^{\star}\mathbf{C}^{\star}_{(k)} + \mathbf{N}_{(k)}, \tag{4.6}$$

$k = 1, \ldots, N$, for which we want to estimate $\mathbf{D}^{\star}$ and $\mathbf{C}^{\star}_{(k)}$.

One might ask why we choose to model the columns of $\mathbf{A}$ as independent. From Sec. 2.6, we know that provable guarantees for dictionary learning require $\mathcal{O}(r \log r)$ observations, where r scales like the dimension of the vector space for the observed signals. In our case, the autoregressive coefficients would be in a $m \cdot d^2$-dimensional vector space. By modeling the columns as independent, we can reduce the sampling burden since $r \sim \mathcal{O}(m \cdot d)$.

4.2 Two-stage approach

In this section, we will analyze a two-stage procedure in which we first estimate the autoregressive coefficients of each observation independently. This yields a noisy dictionary learning problem, for which we can apply the theoretical results of Thm. 2.12. This will yield a probabilistic statement on the recovery of autoregressive atoms. Formally, we analyze Algorithm 1. The functions *OverlappingCluster*, *OverlappingSVD*, and *IterativeAverage* are described in general terms in Sec. 2.6, and details for these sub-routines can be found in Arora, *et al.* [9]. Before proceeding with our analysis, we give the generative model for our observations.

Model 1. We observe N independent and identical observations of vector-valued time-series $\left(\mathbf{x}_{(k)}[t]\right)_{t=1,\ldots,T+m}$, $k = 1, \ldots, N$, generated as follows:

1. Dictionary, $\mathbf{D}^{\star}$:

(a) The columns of $\mathbf{D}^\star \in \mathbb{R}^{(m\cdot d)\times r}$ are unit norm ($\|\cdot\|_2$)

(b) $\left\langle \mathbf{D}_i^\star, \mathbf{D}_j^\star \right\rangle \leq \mu/\sqrt{m\cdot d}$ for all $i \neq j$ and for some $\mu \sim \mathcal{O}\left(\log(m\cdot d)\right)$

2. Coefficients, $\left(\mathbf{C}_{(k)}^\star\right)_{k=1,\ldots,N}$:

(a) Columns are drawn independently and identically

(b) The support for a column is chosen uniformly at random from

$\{U \subset \{1,\ldots,r\} : |U| = s\}$

(c) The non-zero coefficients are drawn independently and identically from a centered distribution with support on $[-C,-c]\cup[c,C]$

3. Innovation, $\left(\mathbf{N}_{(k)}\right)_{k=1,\ldots,N}$:

(a) Entries are drawn independently and identically from $\mathcal{N}\left(0,\nu^2\right)$ with $\nu \sim \mathcal{O}\left(\sqrt{d}\right)$

4.2.1 Estimation of the autoregressive coefficients

In this section, we analyze the error of the estimator of Eq. (2.39),

$$\mathbf{A} = \left(\mathbf{X}^*\mathbf{X}\right)^{-1}\mathbf{X}^*\mathbf{Y}. \tag{4.7}$$

This estimator seeds the dictionary learning algorithm. Lemma 4.1 provides a probabilistic bound on the error of the estimator. For large N, and T larger, the estimator is well-behaved. The result yields an asymptotic characterization of the estimator, *i.e.* as $T, N \to \infty$, for $T \sim \Omega\left(N^p\right)$ for $p > 1$, the estimator is consistent.

Algorithm 1: Two-stage approach

Data: $\left(\mathbf{Y}_{(k)}\right)_{k=1,\dots,N}$, $\left(\mathbf{X}_{(k)}\right)_{k=1,\dots,N}$

Result: $\hat{\mathbf{D}} \in \mathbb{R}^{(m\cdot d)\times r}$

for $k = 1, \dots, N$ **do**

$$\hat{\mathbf{A}}_{(k)} = \underset{\mathbf{A}\in\mathbb{R}^{(m\cdot d)\times d}}{\arg\min} \frac{1}{2(T\cdot d)} \left\|\mathbf{Y}_{(k)} - \mathbf{X}_{(k)}\mathbf{A}\right\|_F^2;$$

end

$$(\mathcal{C}_j)_{j=1,\dots,r} = \text{OverlappingCluster}\left(\left(\hat{\mathbf{A}}_{(k)}\right)_{k=1,\dots,N}\right);$$

$$\hat{\mathbf{D}} = \text{OverlappingSVD}\left((\mathcal{C}_j)_{j=1,\dots,r}\right);$$

while *not converged* **do**

$$\hat{\mathbf{D}} = \text{IterativeAverage}\left(\left(\mathbf{Y}_{(k)}\right)_{k=1,\dots,N}, \hat{\mathbf{D}}\right);$$

end

Lemma 4.1. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 1. Then for any* $\varepsilon > 0$ *and* $\delta_1 \in (0,1)$, *with probability at least* $1 - \frac{N\cdot m\cdot d}{T\cdot(1-\delta_1)^2\cdot\varepsilon^2}\cdot\frac{\left(1+\sqrt{d}\cdot s\cdot C\right)^4}{\left(1-\sqrt{d}\cdot s\cdot C\right)^2} - C_1\cdot N\cdot\exp\left(-c_1\cdot T/m\cdot\min\left(\left[\frac{\delta_1\cdot\nu^4}{2K_1\cdot(1-d\cdot s^2\cdot C^2)^2}\right]^2, \frac{\delta_1\cdot\nu^4}{2K_1\cdot(1-d\cdot s^2\cdot C^2)^2}\right)\right)$, *the estimates of Eq. (2.39)* $\left(\mathbf{A}_{(k)}\right)_{k=1,\dots,N}$ *will satisfy*

$$\left\|\mathbf{D}^\star\mathbf{C}^\star_{(k)} - \mathbf{A}_{(k)}\right\|_{2,\infty} \leq \varepsilon \tag{4.8}$$

for all $k = 1, \dots, N$, *where* $C_1 = 2\cdot m\cdot 9^{(m\cdot d)}$, $c_1 > 0$ *is a global constant, and* $K_1 \sim \mathcal{O}(1)$.

Proof. We will prove the statement in two parts: (1) for a fixed k and appropriate assumptions, we will bound the error of the estimate; and (2) we will compute the probability that the assumptions hold and extend it for all $k = 1, \dots, N$ using a

union bound.

(1) To begin, we fix $k \in \{1, \ldots, N\}$ and assume

(i) $\left\| \frac{1}{T}\mathbf{X}^*\mathbf{X} - \mathbf{R} \right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s \cdot C} \right)^2$ and

(ii) $\left\| \frac{1}{T}\mathbf{X}^*\mathbf{N} \right\|_{2,\infty} \leq (1-\delta_1) \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s \cdot C} \right)^2 \cdot \varepsilon$.

Since $\mathbf{A}$ is a minimizer with respect to each column $i = 1, \ldots, d$,

$$\|\mathbf{Y}_i - \mathbf{X}\mathbf{A}_i\|^2 \leq \|\mathbf{Y}_i - \mathbf{X}\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i\|^2 .$$

Substituting $\mathbf{Y} = \mathbf{X}\mathbf{D}^\star\mathbf{C}^\star + \mathbf{N}$ yields

$$\|\mathbf{X}\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i + \mathbf{N}_i - \mathbf{X}\mathbf{A}_i\|^2 \leq \|\mathbf{X}\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i + \mathbf{N}_i - \mathbf{X}\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i\|^2 = \|\mathbf{N}_i\|^2 .$$

Expanding the square on the left-hand side gives us

$$\|\mathbf{X}\left[\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\right]\|^2 + \langle \mathbf{X}\left[\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\right], \mathbf{N}_i \rangle + \|\mathbf{N}_i\|^2 \leq \|\mathbf{N}_i\|^2 .$$

Some straightforward manipulations to include moving $\mathbf{X}$ to the other side of the inner product and applying the Cauchy-Schwarz inequality yield

$$\|\mathbf{X}\left[\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\right]\|^2 \leq \|\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\| \cdot \|\mathbf{X}^*\mathbf{N}_i\| .$$

Multiplying both sides by $1/T$, we can lower bound the left-hand side as

$$\lambda_{\min}\left(\frac{1}{T}\mathbf{X}^*\mathbf{X}\right) \cdot \|\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\|^2 \leq \|\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\| \cdot \left\| \frac{1}{T}\mathbf{X}^*\mathbf{N}_i \right\| .$$

This yields the following error bound:

$$\|\left(\mathbf{D}^\star\mathbf{C}^\star\right)_i - \mathbf{A}_i\| \leq \left[\lambda_{\min}\left(\frac{1}{T}\mathbf{X}^*\mathbf{X}\right)\right]^{-1} \cdot \left\| \frac{1}{T}\mathbf{X}^*\mathbf{N}_i \right\| .$$

By our assumptions, we have a bound on $\left\|\frac{1}{T}\mathbf{X}^*\mathbf{N}_i\right\|$ uniform for $i = 1, \ldots, d$, and

$$\lambda_{\min}\left(\frac{1}{T}\mathbf{X}^*\mathbf{X}\right) \geq \lambda_{\min}\left(\mathbf{R}\right) - \left\|\frac{1}{T}\mathbf{X}^*\mathbf{X} - \mathbf{R}\right\| \geq (1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2,$$

where we have used Thm. 2.11 to get

$$\begin{aligned}\lambda_{\min}\left(\mathbf{R}\right) \geq \left(\frac{\nu}{1+\|\mathbf{D}^\star\mathbf{C}^\star\|}\right)^2 \geq \left(\frac{\nu}{1+\|\mathbf{D}^\star\mathbf{C}^\star\|_F}\right)^2 \geq \left(\frac{\nu}{1+\|\mathbf{C}^\star\|_{1,2}}\right)^2 \\ \geq \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2.\end{aligned}$$

Therefore, $\|(\mathbf{D}^\star\mathbf{C}^\star)_i - \mathbf{A}_i\| \leq \varepsilon$ for all $i = 1, \ldots, d$, and we note that the bound is independent of the realization of $\mathbf{C}^\star$.

(2) We can compute the probabilities of $\left\{\left\|\frac{1}{T}\mathbf{X}^*\mathbf{X} - \mathbf{R}\right\| \leq \delta_1\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\}$ and $\left\|\frac{1}{T}\mathbf{X}^*\mathbf{N}\right\|_{2,\infty} \leq (1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\cdot\varepsilon$ using Lemmas 4.10 and 4.11 respectively. Note that these sets do not depend on the realization of $\mathbf{C}^\star$. Thus, we can use a simple union bound to compute

$$\begin{aligned}&\mathbb{P}\left(\bigcap_{k=1}^{N}\left\{\left\|\mathbf{D}^\star\mathbf{C}^\star_{(k)} - \mathbf{A}_{(k)}\right\|_{2,\infty} \leq \varepsilon\right\}\right) \\ &\quad\geq 1 - \sum_{k=1}^{N}\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta_1\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\}\right) \\ &\qquad - \sum_{k=1}^{N}\mathbb{P}\left(\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} > (1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\cdot\varepsilon\right).\end{aligned}$$

□

Lemma 4.1 tells us that for large T, we will have observations $k = 1, \ldots, N$,

$$\mathbf{A}_{(k)} = \mathbf{D}^\star\mathbf{C}^\star_{(k)} + \mathbf{W}_{(k)}, \tag{4.9}$$

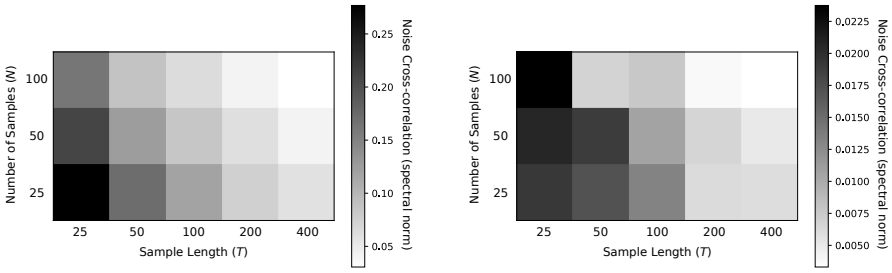


Figure 4.1: Sample mean (left) and standard deviation (right) of $\left\| \left(\frac{1}{T} \mathbf{X}^*_{(k)} \mathbf{X}_{(k)} \right)^{-1} \frac{1}{T} \mathbf{X}^*_{(k)} \mathbf{N}_{(k)} \right\|$. Simulations under Model 1 reveal the expected decay of cross-correlation.

where $\mathbf{W}_{(k)}$ is random but uniformly bounded, *i.e.* $\left\| \mathbf{W}_{(k)} \right\|_{2,\infty} \leq \delta$. We would like to say more, *e.g.* the columns of $\mathbf{W}_{(k)}$ are normally distributed. We have

$$\left(\mathbf{X}^* \mathbf{X} \right)^{-1} \mathbf{X}^* \mathbf{Y} = \left(\mathbf{X}^* \mathbf{X} \right)^{-1} \mathbf{X}^* \left(\mathbf{X} \mathbf{D}^\star \mathbf{C}^\star + \mathbf{N} \right) = \mathbf{D}^\star \mathbf{C}^\star + \left(\frac{1}{T} \mathbf{X}^* \mathbf{X} \right)^{-1} \frac{1}{T} \mathbf{X}^* \mathbf{N}.$$

Conditioning on the coefficient matrix $\mathbf{C}^\star$, we have $\left(\frac{1}{T} \mathbf{X}^* \mathbf{X} \right)^{-1} \Big| \mathbf{C}^\star \xrightarrow{p} \mathbf{R}$ and $\frac{1}{T} \mathbf{X}^* \mathbf{N} \big| \mathbf{C}^\star \xrightarrow{p} \mathbf{0}$. By Slutsky's theorem and the continuous mapping theorem [52], we get

$$\left(\mathbf{X}^* \mathbf{X} \right)^{-1} \mathbf{X}^* \mathbf{N} \xrightarrow{p} \mathbf{0}.$$

This gives us asymptotic consistency, and numerical experiments bear this out as shown in Fig. 4.1.

4.2.2 Finding the autoregressive dictionary atoms

Given this estimate of the autoregressive coefficients of each observation

$$\mathbf{A}_{(k)} = \mathbf{D}^\star \mathbf{C}^\star_{(k)} + \mathbf{W}_{(k)}, \tag{4.10}$$

we can use Thm. 2.12 to find $\mathbf{D}^\star$. This result is summarized in the below theorem.

Theorem 4.2. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\ldots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\ldots,N}$ *be generated according to Model 1. Then, if* $N \sim \Omega\left((r^2/s^2)\log r + rs^2 \log r + r \log r \log(1/\varepsilon)\right)$ *and* $T \sim \Omega\left(N \cdot m \cdot d/\varepsilon^3 + d \cdot m^2 \log m \log N/(\nu^4 \cdot \log \varepsilon)\right)$, *with high probability, Alg. 1 will return a dictionary estimate* $\mathbf{D}_0$ *that satisfies* $d(\mathbf{D}^\star, \mathbf{D}_0) \leq \varepsilon$, *where* $d : \mathbb{R}^{(m\cdot d)\times r} \times \mathbb{R}^{(m\cdot d)\times r}$ *is the dictionary metric defined in Eq. (2.43).*

Proof. This is a rather straightforward application of Thm. 2.12 to the result of Lemma 4.1. Note that for $T \sim \Omega\left(N \cdot m \cdot d/\varepsilon^3 + d \cdot m^2 \log m \log N/(\nu^4 \cdot \log \varepsilon)\right)$, Lemma 4.1 provides a high probability bound on the coefficient error by ε, *i.e.* $1 - \mathcal{O}(\varepsilon)$. This means that the additive noise for our observations of $\mathbf{D}^\star \mathbf{C}^\star_{(k)}$ has variance bounded by ε^2. That $\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)^{-1} \frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)} - \mathbf{D}^\star\mathbf{C}^\star_{(k)}$ is not spherical Gaussian does not change any of the $\Omega(\cdot)$ factors in the statement of the theorem (see p. 10 of Arora, *et al.* [9]). □

4.2.3 Simulations

In this section, we report results from applying Algorithm 1 on simulated data. The simulated data was generated according to Model 1 using $d = 4$, $m = 2$, and $s = 2$. We evaluated accuracy using the dictionary metric (2.43) scaled by the number of dictionary atoms. For each condition, we ran twenty simulations and reported statistics over those twenty experiments. The algorithm runs in polynomial time $\mathcal{O}(rN^2)$, the inner loops of the algorithm entail non-trivial computations polynomial in d, m, r, and T, and the run-time grows rapidly with increasing problem size. This

makes reaching $N \sim \mathcal{O}\left(r \log r \log(1/\varepsilon)\right)$ for small ε difficult. As shown in Fig. 4.2, even for small m and d, it was difficult to simulate problem sizes that revealed the finite sample behavior of Thm. 4.2. However, we can more clearly see the effect of increasing redundancy of the dictionary in Fig. 4.3.

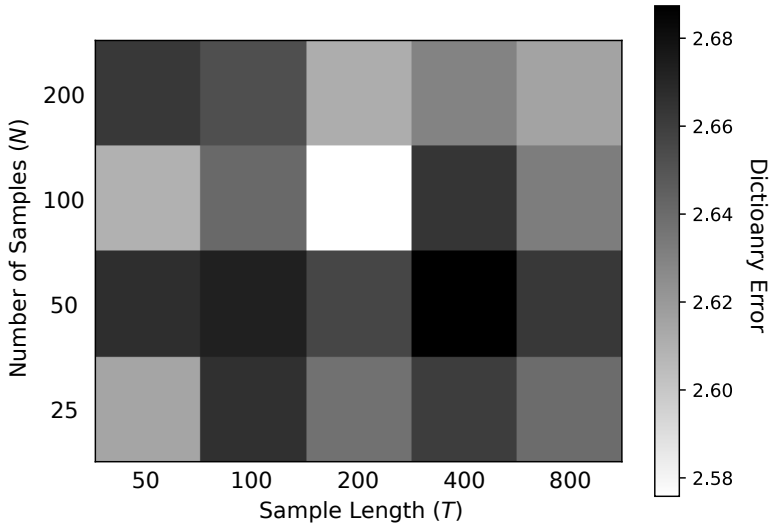


Figure 4.2: $r^{-1/2} d\left(\mathbf{D}^\star, \hat{\mathbf{D}}\right)$. We cannot discern the expected behavior of the error as a function of N and T. Here, we use $r = 9$ so that we require $N \sim \mathcal{O}\left(20/\varepsilon\right)$ and $T \sim \mathcal{O}\left(N/\varepsilon^3\right)$ to expect ε accuracy. Likely, we cannot simulate sufficient size problems to overcome the finite sample factors in the result.

4.3 Direct Approach

In Section 4.2, we considered an algorithm for recovering $\mathbf{D}^\star$ and $\mathbf{C}^\star_{(k)}$ from observations

$$\mathbf{Y}_{(k)} = \mathbf{X}_{(k)} \mathbf{D}^\star \mathbf{C}^\star_{(k)} + \mathbf{N}_{(k)}. \tag{4.11}$$

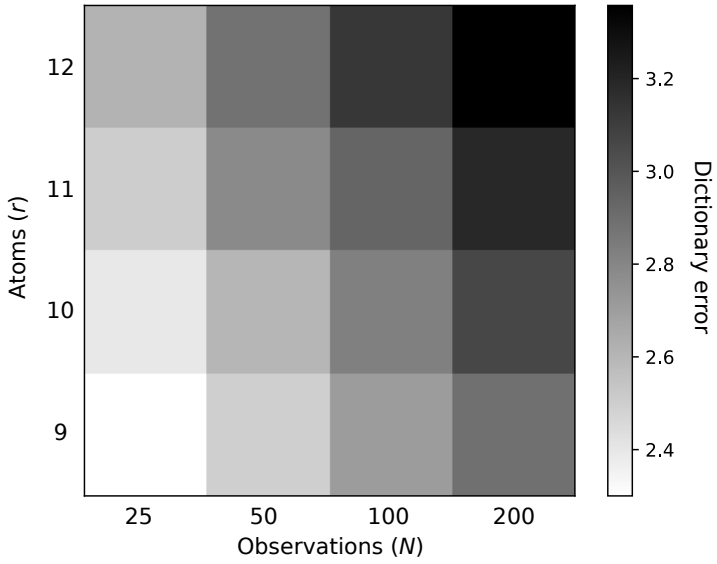


Figure 4.3: $r^{-1/2}d\left(\mathbf{D}^\star, \hat{\mathbf{D}}\right)$. We see linear increase in error with increases in dictionary redundancy. The mixed effects are less discernible due to the implicit dependence on observation length.

We implemented a two-stage approach, finding $\mathbf{A}_{(k)} \approx \mathbf{D}^\star \mathbf{C}^\star_{(k)}$, and then applying existing results from theoretical dictionary learning. In this section, we explore a direct approach to recovering $\mathbf{D}^\star$ and $\mathbf{C}^\star_{(k)}$ via an alternating minimization algorithm. It is not a direct substitute for the two-stage approach since for provable recovery it would need to be paired with an initialization procedure such as *OverlappingCluster* and *OverlappingSVD*, but empirical results and recent theoretical results (see Sun, *et al.* [105]) indicate that the loss landscape of dictionary learning contains many equivalent solutions, and so a fixed point of an alternating minimization algorithm is likely to be equivalent to the global minimum. The approach we now analyze follows more closely that of Agarwal, *et al.* [2].

Our analysis leads us to a largely negative result. It illustrates the challenges of directly solving this problem without stronger assumptions than those we use. We can give conditions under which a direct approach works with positive probability, but those conditions are impractical even in asymptotic analysis.

Before proceeding with our analysis, we give the generative model for our observations.

Model 2. For a fixed $s \in \{1, \ldots, r\}$, we define the following generative model:

1. Dictionary, $\mathbf{D}^\star$:
   (a) The columns of $\mathbf{D}^\star \in \mathbb{R}^{(m \cdot d) \times r}$ are unit norm ($\|\cdot\|_2$)
   (b) $\mathbf{D}^\star$ satisfies a s-restricted eigenvalue condition with $\kappa_s^\star > 4\varepsilon^{(0)}\sqrt{\frac{s}{m \cdot d}}$
2. Coefficients, $\left(\mathbf{C}_{(k)}^\star\right)_{k=1,\ldots,N}$:
   (a) Columns are drawn independently and identically
   (b) The support for a column is chosen uniformly at random from $\{U \subset \{1, \ldots, r\} : |U| = s\}$
   (c) The non-zero coefficients are drawn independently and identically from a centered distribution with support on $[-C, C]$ and variance c^2
3. Innovation ($\left(\mathbf{N}_{(k)}\right)_{k=1,\ldots,N}$):
   (a) Entries are drawn independently and identically from $\mathcal{N}(0, \nu^2)$ with $\nu \sim \mathcal{O}\left(\sqrt{d}\right)$

Thus far, the primary different between Models 2 and 1 are the dictionary condition and distribution of non-zero coefficients. However, we require an additional assumption on the growth of the moments of the autocorrelation function of the autoregressive process:

$$\left\| \mathbb{E}\left[\left(\frac{1}{T} \mathbf{X}_{(k)}^{*} \mathbf{X}_{(k)} \right)^{2} \middle| \mathbf{C}_{(k)}^{\star} \right] \right\| \sim \mathcal{O}\left(\left\| \mathbb{E}\left[\left(\frac{1}{T} \mathbf{X}_{(k)}^{*} \mathbf{X}_{(k)} \right) \middle| \mathbf{C}_{(k)}^{\star} \right] \right\|^{2} \right). \tag{4.12}$$

We anticipate that such a condition imposes additional constraints on the higher order moments of the non-zero coefficients $\mathbf{C}_{(k)}^{\star}$.

Our analysis will consider Algorithm 2. The algorithm features a straightforward application of alternating minimization. Our results refer to recovery of $\mathbf{D}^{\star}$ in Model 2; however, the alternating minimization algorithm derives from minimizing the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}, \left\{ \mathbf{C}_k \in \mathbb{R}^{r \times d} \right\}}{\arg\min} \frac{1}{N \cdot T \cdot m \cdot d} \left\| \mathbf{Y}_{(k)} - \mathbf{X}_{(k)} \mathbf{D} \mathbf{C}_{(k)} \right\| + \mu \left\| \mathbf{C}_{(k)} \right\|_{1,1} \\ &\text{s.t. } \|\mathbf{D}_j\| = 1, \ j = 1, \ldots, r \end{aligned} \tag{4.13}$$

for some $\mu > 0$ which we are allowed to adapt at each iteration. The two steps of the algorithm follow from fixing the dictionary with the current estimate $\mathbf{D}^{(\ell)}$ while updating the estimate of the coefficients $\left(\mathbf{C}_{(k)}^{(\ell)} \right)_{k=1,\ldots,N}$. Then, we fix the coefficients with the updated estimate and update our estimate of the dictionary. Accordingly, we refer throughout this section to the coefficient and dictionary update steps respectively. The coefficient update encompasses the assignments to $\mathbf{C}_{k}^{(\ell)}$ for all $k = 1, \ldots, N$, and the dictionary update encompasses the assignment to $\mathbf{D}^{(\ell)}$.

Algorithm 2: Direct Approach

Data: $\left(\mathbf{Y}_{(k)}\right)_{k=1,\dots,N}$, $\left(\mathbf{X}_{(k)}\right)_{k=1,\dots,N}$, $\left(\mu^{(\ell)}\right)_{\ell\geq 1}$

Result: $\hat{\mathbf{D}} \in \mathbb{R}^{(d\cdot m)\times r}$

Initialize $\mathbf{D}^{(0)}$;

while *not converged* **do**

for $k = 1, \dots, N$ **do**

$$\mathbf{C}_k^{(\ell)} = \underset{\mathbf{C}\in\mathbb{R}^{r\times d}}{\arg\min} \frac{1}{T\cdot m\cdot d} \left\|\mathbf{Y}_k - \mathbf{X}_k\mathbf{D}^{(\ell-1)}\mathbf{C}\right\|_F^2 + 2\mu^{(\ell)} \left\|\mathbf{C}\right\|_{1,1};$$

end

$$\mathbf{D}^{(\ell)} = \underset{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}{\arg\min} \frac{1}{N\cdot T\cdot d} \sum_{k=1}^{N} \left\|\mathbf{Y}_k - \mathbf{X}_k\mathbf{D}\mathbf{C}_k^{(\ell)}\right\|_F^2 \text{ s.t. } \left\|\mathbf{D}_j^{(\ell)}\right\| = 1,$$

$j = 1, \dots, r;$

end

In the present work, we make no claim as to whether the fixed point of the alternating minimization algorithm coincides with the minimizer of a corresponding optimization problem. Such questions in dictionary learning are the subject of *e.g.* Gribonval and Schnass [51] and Gribonval, *et al.* [50].

The dictionary update of Algorithm 2 requires solving a nonconvex problem. In our analysis, we assume that we can find a global minimum to such a problem. This is clearly a strong assumption at its face. However, the nonconvex constraint decouples with respect to the dictionary atoms $\{\mathbf{D}_j\}$, which allows us to implement atom-wise proximal algorithms such as that of Bolte, *et al.* [20]. This algorithm guarantees convergence to a fixed point of the problem. Given that we are initializing the dictionary ε-close to the true dictionary, it is perhaps not unreasonable to expect

that our current estimate and the global minimizer lie within the same convex region of the objective. If not, it is possible to solve the unconstrained convex problem and then project onto the unit sphere. The analysis then requires less elegant techniques for addressing how large this projection can be. This is done in [2], but it is not particularly illuminating to understanding the mechanism by which the alternating minimization algorithm—which enjoys widespread practical application and success—works. Therefore, our analysis assumes that we can solve the nonconvex problem

In the following, we present our main result which characterizes the convergence of Algorithm 2. In the following sections, we provide the key supporting results.

4.3.1 Main result

Our main result for the direct approach has a markedly different character than that of the two-stage approach. Here, we require an accurate initialization of the dictionary estimate. Then, we characterize the probability of recovering the true dictionary to arbitrary accuracy. As mentioned in Section 4.3, the result is negative. In order to have convergence with nonzero probability, we require impractical conditions on the model, namely vanishing variance of the non-zero coefficients and innovation.

Theorem 4.3. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 2. Assume that* $d\left(\mathbf{D}^\star, \mathbf{D}^{(0)}\right) \leq (\kappa_s^\star/8) \cdot \sqrt{m \cdot d/s}$. *Then, for any* $\varepsilon \in$

$\left(0, (\kappa_s^\star/8) \cdot \sqrt{m \cdot d/s}\right)$, *there exists a finite* $n_\varepsilon \sim \mathcal{O}(\log \varepsilon)$ *and a sequence* $\left(\mu^{(\ell)}\right)_{\ell=1,\ldots,n_\varepsilon}$ *such that for*

- $T \sim \Omega\left(N \cdot r^{1/2} \cdot \nu^4/(m \cdot d \cdot \varepsilon^3) + d \cdot m^2 \log m \cdot \log N/(\nu^4 \cdot \log \varepsilon)\right)$,
- $N \sim \Omega\left(s \cdot r \cdot \log(m \cdot d \cdot r)/\log \varepsilon\right)$, *and*
- $c^2 \cdot \nu^4 \sim \mathcal{O}\left(m^{3/2} \cdot d^{1/2} \cdot r \cdot \varepsilon/(N \cdot s^{3/2} \cdot \exp(m \cdot d \cdot r)\right)$,

with high probability, the dictionary estimate of Algorithm 2 will satisfy $d\left(\mathbf{D}^\star, \mathbf{D}^{(n_\varepsilon)}\right) \leq \varepsilon$.

Proof. We will begin by constructing an ε-sequence $\left(\varepsilon^{(\ell)}\right)_{\ell=1,\ldots,n_\varepsilon}$ such that $\varepsilon^{(n_\varepsilon)} \leq \varepsilon$. Let $\varepsilon^{(0)} = (\kappa_s^\star/8) \cdot \sqrt{m \cdot d/s}$. Let us choose $\varepsilon^{(\ell)} = \alpha \cdot \varepsilon^{(\ell-1)}$ for $\alpha \in (0,1)$ so that $n_\varepsilon = \lceil \log_\alpha \varepsilon/\varepsilon^{(0)} \rceil$. Recall that $d(\mathbf{D}^\star, \mathbf{D}^{(0)}) < \varepsilon^{(0)}$ implies that there exists a signed permutation matrix $\mathbf{P}$ such that $\left\|\mathbf{D}^\star - \mathbf{D}^{(0)}\mathbf{P}\right\|_F < \varepsilon^{(0)}$. By Lemma 4.4, we know that for $\ell = 1$, if conditions (2) and (3) are satisfied, then there exists a $\mu^{(1)}$ such that if conditions (4) and (5) are satisfied for $\mu^{(1)}$, then $\left\|\mathbf{D}^\star - \mathbf{D}^{(1)}\mathbf{P}\right\|_F \leq \varepsilon^{(1)}$, satisfying condition (1) for $\ell = 2$. We can then repeat the reasoning steps for all $\ell = 2, \ldots, n_\varepsilon$.

The remainder of the proof will comprise assembling the probabilities of conditions (2), (3), (4), and (5) from Lemma 4.4. Conditions (2) and (3) are global conditions that we can establish independent of the μ- and ε-sequences. Conditions (4) and (5) will require us to construct a nested sequence of events from which we can compute a global probability.

Let us define events $\mathcal{A}_2$, $\mathcal{A}_3$, $\mathcal{A}_4$, and $\mathcal{A}_5$ corresponding to conditions (2), (3),

(4), and (5) as follows:

$$\mathcal{A}_2 = \bigcap_{k=1}^{N} \left\{ \left\| \frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)} \right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s \cdot C} \right)^2 \right\};$$

$$\mathcal{A}_3 = \left\{ \lambda_{\min} \left(\frac{1}{N}\sum_{k=1}^{N} \left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star} \left(\mathbf{P}\mathbf{C}_{(k)}^{\star} \right)^{*} \right) \otimes \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} \right) \right) \right.$$
$$\left. \geq (1-\delta_2) \cdot \frac{s \cdot c^2}{r} \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s \cdot C} \right)^2 \right\};$$

$$\mathcal{A}_4 = \bigcap_{k=1}^{N} \bigcap_{\ell=1}^{n_\varepsilon} \left\{ \left\| \frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)} \right\|_{2,\infty} \leq \frac{m \cdot d \cdot \mu^{(\ell)}}{4} \right\}; \text{ and}$$

$$\mathcal{A}_5 = \bigcap_{k=1}^{N} \bigcap_{\ell=1}^{n_\varepsilon} \left\{ \max_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\| \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} \right) \mathbf{D}\mathbf{C}_{(k)}^{\star} \right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq \frac{m \cdot d \cdot \mu^{(\ell)}}{4 \cdot \varepsilon^{(\ell-1)}} \right\}.$$

Let us also denote the desired event $\mathcal{A}_0 = \left\{ \left\| \mathbf{D}^{\star} - \mathbf{D}^{(n_\varepsilon)}\mathbf{P} \right\|_F \leq \varepsilon \right\}$. Then,

$$\mathbb{P}\left(\mathcal{A}_0\right) = \mathbb{P}\left(\mathcal{A}_3 \middle| \mathcal{A}_2\right) \cdot \mathbb{P}\left(\mathcal{A}_2\right) \cdot \mathbb{P}\left(\mathcal{A}_3\right) \cdot \mathbb{P}\left(\mathcal{A}_5\right)$$
$$\geq 1 - \mathbb{P}\left(\mathcal{A}_3^C \middle| \mathcal{A}_2\right) - \mathbb{P}\left(\mathcal{A}_2^C\right) - \mathbb{P}\left(\mathcal{A}_4^C\right) - \mathbb{P}\left(\mathcal{A}_5^C\right).$$

We can calculate $\mathbb{P}\left(\mathcal{A}_2^C\right)$ using Lemma 4.10 and a union bound over $k = 1, \ldots, N$. This event will happen with vanishing probability, *i.e.* $\mathcal{O}(\varepsilon)$, if $T \sim \Omega\left(d \cdot m^2 \log m \cdot \log N / (\nu^4 \cdot \log \varepsilon)\right)$.

We can compute $\mathbb{P}\left(\mathcal{A}_3^C \middle| \mathcal{A}_2\right)$ using Lemma 4.14. This event will happen with vanishing probability if $N \sim \Omega\left(s \cdot r \cdot \log(m \cdot d \cdot r) / \log \varepsilon\right)$.

Computing $\mathbb{P}\left(\mathcal{A}_4^C\right)$ requires some additional consideration. Note that in proving Lemma 4.4, we showed that $\left(\mu^{(\ell)}\right)_{\ell=1,\ldots,n_\varepsilon}$ is a monotonically decreasing sequence so that

$$\left\{ \left\| \frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)} \right\|_{2,\infty} \leq \frac{m \cdot d \cdot \mu^{(n_\varepsilon)}}{4} \right\} \subseteq \cdots \subseteq \left\{ \left\| \frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)} \right\|_{2,\infty} \leq \frac{m \cdot d \cdot \mu^{(1)}}{4} \right\}.$$

Therefore, $\mathcal{A}_4 = \bigcap_{k=1}^{N} \left\{ \left\| \frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{N}_{(k)} \right\|_{2,\infty} \leq \frac{m \cdot d \cdot \mu^{(n_\varepsilon)}}{4} \right\}$. Moreover, in Lemma 4.4, we see that for small $\varepsilon^{(\ell-1)}$, $\mu^{(\ell)} \sim \mathcal{O}\left(r^{-1/2} \cdot \varepsilon^{(\ell-1)}\right)$. Using Lemma 4.11 and a union bound over $k = 1, \ldots, N$, we can compute a bound on $\mathbb{P}\left(\mathcal{A}_4^C\right)$. This event will occur with vanishing probability for $T \sim \Omega\left(N \cdot r^{1/2} \cdot \nu^4 / (m \cdot d \cdot \varepsilon^3\right)$.

Finally, we consider $\mathcal{A}_5$. We note that $\mu^{(\ell)}/\varepsilon^{(\ell-1)}$ is a monotonically increasing sequence in ℓ. Thus,

$$\mathcal{A}_5 = \left\{ \max_{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}} \frac{\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq \frac{m \cdot d \cdot \mu^{(1)}}{4 \cdot \varepsilon^{(0)}} \right\}.$$

For $\varepsilon^{(0)} \in (\varepsilon, \kappa_s^\star / 4 \cdot \sqrt{m \cdot d / s})$, we have that $\mu^{(1)}/\varepsilon^{(0)} \sim \mathcal{O}\left(r^{-1/2} \cdot \left(\varepsilon^{(0)}\right)^{-1/2}\right)$. This requires us to bound

$$\left\{ \max_{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}} \frac{\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty}}{\|\mathbf{D}\|_F} > C_1 \cdot \sqrt{m \cdot d \cdot s} \right\}$$

for some $C_1 > 0$, which we can do using Lemma 4.13 and a union bound over all $k = 1, \ldots, N$. From this, we see that there is vanishing probability only if $c^2 \cdot \nu^4 \sim \mathcal{O}\left(m^{3/2} \cdot d^{1/2} \cdot r^{3/2} \cdot \varepsilon / (N \cdot s^{3/2} \cdot \exp(m \cdot d \cdot r))\right)$. □

4.3.2 Results for a single iteration of the algorithm

The following result forms the crux of the convergence result for Algorithm 2. It tells us that for an arbitrary iteration of the alternating minimization algorithm, given that we are ε-close to recovering the true dictionary, there exists a penalty parameter that will improve our estimate of the dictionary. This result depends on several probabilistic events captured in the assumptions of Lemma 4.4.

The approach is standard (*c.f.* [2]). From the dictionary update step, we conclude that our estimate minimizes the objective at least as well as any other candidate solution while simultaneously satisfying the atom-wise constraints. As discussed previously in Sec. 4.3, we are assuming that we have access to a solver which provides a global minimum to this nonconvex problem. From here, the argument is largely an exercise in book-keeping. The set of $\mu^{(\ell)}$ which satisfy our per-iteration goal come from a quadratic inequality, and yield a function logarithmic in $\varepsilon^{(\ell-1)}$, *i.e.* $\mu^{(\ell)} \sim \mathcal{O}\left(\sqrt{\varepsilon^{(\ell)}}\right)$.[2] In practice, we should use larger steps in μ-space (decrease $\mu^{(\ell)}$ more aggressively) as we achieve better accuracy in recovery ($\varepsilon^{(\ell-1)}$ gets closer to zero).

It would be tempting to conclude from this result, that we can conclude our main result by an induction argument. However, assumptions (4) and (5) depend on the penalty parameter $\mu^{(\ell)}$ and dictionary error $\varepsilon^{(\ell-1)}$. It is these events that we need to show have positive probability for some sequence of dictionary estimates in Theorem 4.3.

Lemma 4.4. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 2. Assume the following conditions for some* $\delta_1, \delta_2 \in (0,1)$*:*

(1) $\left\|\mathbf{D}^\star - \mathbf{D}^{(\ell-1)}\mathbf{P}\right\|_F \leq \varepsilon^{(\ell-1)}$;

(2) $\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$ *for all* $k = 1,\dots,N$*; and*

(3) $\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{P}\mathbf{C}^\star_{(k)}\left(\mathbf{P}\mathbf{C}^\star_{(k)}\right)^*\right)\otimes\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\right) \geq (1-\delta_2)\cdot\frac{s\cdot c^2}{r}\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$.

[2] This statement requires that we assume that we are in an error regime where $\varepsilon^{(\ell-1)} \ll \kappa^\star_s\sqrt{m\cdot d/s}/4$.

Then, for any $\alpha \in (0,1)$ *independent of* ℓ*, there exists a* $\mu^{(\ell)} > 0$ *such that if the following conditions hold:*

(4) $\left\| \frac{1}{T} \mathbf{X}_{(k)}^{*} \mathbf{N}_{(k)} \right\|_{2,\infty} \leq \frac{m \cdot d \cdot \mu^{(\ell)}}{4}$ *for all* $k = 1, \ldots, N$ *and*

(5) $\max_{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}} \frac{\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^{*} \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^{\star} \right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq \frac{m \cdot d \cdot \mu^{(\ell)}}{4 \cdot \varepsilon^{(\ell-1)}}$ *for all* $k = 1, \ldots, N$.

Then, $d\left(\mathbf{D}^{\star}, \mathbf{D}^{(\ell)}\right) \leq \alpha \cdot \varepsilon^{(\ell-1)}$.

Proof. We will prove the statement in multiple parts. The first part will comprise algebraic manipulations to isolate the dictionary error, *i.e.*

$$\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}_{(k)}^{(\ell)} \right\|_F^2 \leq \text{r.h.s.}$$

The second part of the argument will comprise upper bounding the right-hand side (r.h.s.) in terms of computable quantities so that it is linear in the dictionary error, $\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F$. The third part of the argument will lower bound the left-hand side (l.h.s.) with respect to a quadratic term of the dictionary error, *i.e.* $\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F^2$. The fourth part will complete the argument, showing that there exists a $\mu^{(\ell)} > 0$ such that $\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F \leq \alpha \cdot \varepsilon^{(\ell-1)}$.

(1) As $\mathbf{D}^{(\ell)}$ is a global minimizer, we have

$$\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{Y}_{(k)} - \mathbf{X}_{(k)} \mathbf{D}^{(\ell)} \mathbf{C}_{(k)}^{(\ell)} \right\|_F^2 \leq \frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{Y}_{(k)} - \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{C}_{(k)}^{(\ell)} \right\|_F^2.$$

We substitute $\mathbf{Y}_{(k)} = \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}_{(k)}^{\star} + \mathbf{N}_{(k)}$ to yield:

$$\begin{aligned} &\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}_{(k)}^{\star} - \mathbf{D}^{(\ell)} \mathbf{C}_{(k)}^{(\ell)} \right) + \mathbf{N}_{(k)} \right\|_F^2 \\ &\leq \frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)} \right) + \mathbf{N}_{(k)} \right\|_F^2. \end{aligned}$$

Expanding the squares and combining like terms yields

$$\begin{aligned}
&\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left[\left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{D}^{(\ell)} \mathbf{C}^{(\ell)}_{(k)} \right) \right\|_F^2 \right. \\
&\qquad \left. + 2 \left\langle \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{D}^{(\ell)} \mathbf{C}^{(\ell)}_{(k)} \right), \mathbf{N}_{(k)} \right\rangle_F \right] \\
&\leq \frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left[\left\| \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)} \right) \right\|_F^2 \right. \\
&\qquad \left. + 2 \left\langle \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)} \right), \mathbf{N}_{(k)} \right\rangle_F \right].
\end{aligned}$$

We can telescope $\mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{D}^{(\ell)} \mathbf{C}^{(\ell)}_{(k)}$ to yield

$$\begin{aligned}
&\mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{D}^{(\ell)} \mathbf{C}^{(\ell)}_{(k)} \\
&= \mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{C}^{(\ell)}_{(k)} + \mathbf{D}^{\star} \mathbf{P}^{*} \mathbf{C}^{(\ell)}_{(k)} - \mathbf{D}^{(\ell)} \mathbf{C}^{(\ell)}_{(k)} \\
&= \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)} \right) + \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}^{(\ell)}_{(k)}.
\end{aligned}$$

Substitution, expanding the square, and combining common terms yields

$$\begin{aligned}
&\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}^{(\ell)}_{(k)} \right\|_F^2 \\
&\leq -\frac{2}{N \cdot T \cdot d} \sum_{k=1}^{N} \left[\left\langle \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}^{(\ell)}_{(k)}, \mathbf{N}_{(k)} \right\rangle_F \right. \\
&\qquad \left. + \left\langle \mathbf{X}_{(k)} \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)} \right), \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}^{(\ell)}_{(k)} \right\rangle_F \right].
\end{aligned}$$

Rearranging, the right-hand side yields two terms linear in the dictionary error:

$$\begin{aligned}
&\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}^{(\ell)}_{(k)} \right\|_F^2 \\
&\leq \underbrace{- \left\langle \frac{2}{N} \sum_{k=1}^{N} \left(\frac{1}{T} \mathbf{X}^{*}_{(k)} \mathbf{N}_{(k)} \right) \left(\frac{1}{d} \mathbf{C}^{(\ell)}_{(k)} \right)^{*}, \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\rangle_F}_{\text{(a)}} \\
&\quad \underbrace{- \left\langle \frac{2}{N} \sum_{k=1}^{N} \left(\frac{1}{T} \mathbf{X}^{*}_{(k)} \mathbf{X}_{(k)} \right) \mathbf{D}^{\star} \mathbf{P}^{*} \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)} \right) \left(\frac{1}{d} \mathbf{C}^{(\ell)}_{(k)} \right)^{*}, \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\rangle_F}_{\text{(b)}}.
\end{aligned}$$

(2) From the previous equation, we will telescope $\frac{1}{d}\mathbf{C}^{(\ell)}_{(k)} = \frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)} - \frac{1}{d}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)}\right)$ to yield four total terms (a1), (a2), (b1), and (b2), where 1 refers to the $\mathbf{P}\mathbf{C}^{\star}_{(k)}$ term and 2 refers to the $\mathbf{P}\mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)}$ term. Our goal is to bound each term with respect to computable quantities and so that they are linear in the dictionary error, $\left\|\mathbf{D}^{\star}\mathbf{P}^{*} - \mathbf{D}^{(\ell)}\right\|_F$.

Before computing upper bounds for (a1), (a2), (b1), and (b2), we derive some quantities that will be used repeatedly in the following bounds. That these are uniform in $k = 1, \ldots, N$ will be important. First, using condition (2) of the lemma, we can upper bound $\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right\|$ uniformly for all $k = 1, \ldots, N$ with a deviation:

$$\begin{aligned}\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right\| &\leq \left\|\mathbf{R}_{(k)}\right\| + \left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \\ &\leq \left(\frac{\nu}{1 - \sqrt{d}\cdot s\cdot C}\right)^2 + \delta_1 \cdot \left(\frac{\nu}{1 + \sqrt{d}\cdot s\cdot C}\right)^2.\end{aligned}$$

Also using conditions (1), (2), (4), and (5), we can use the result of Lemma 4.5 to bound the coefficient error uniformly for all $k = 1, \ldots, N$:

$$\left\|\left(\mathbf{P}\mathbf{C}^{\star}_{(k)} - \mathbf{C}^{(\ell)}_{(k)}\right)\right\|_{1,\infty} \leq \left(\frac{1 + \sqrt{d}\cdot s\cdot C}{\nu}\right)^2 \cdot \frac{48\cdot s\cdot \mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa^{\star}_s - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}.$$

Lastly, $\left\|\mathbf{P}\mathbf{C}^{\star}_{(k)}\right\|_{1,2} \leq \sqrt{d}\cdot s\cdot C$ uniformly for all $k = 1, \ldots, N$ by Lemma 4.7. Now, we proceed with computing upper bounds to the right-hand side.

(a1) Although we expect this term to be small,

$$\begin{aligned}&\mathbb{E}\left[\frac{2}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{N}_{(k)}\right)\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right] \\ &= 2\cdot\mathbb{E}\left[\mathbb{E}\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{N}_{(k)}\right)\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\middle|\mathbf{C}^{\star}_{(k)}\right] = \mathbf{0},\end{aligned}$$

we accept a worst-case bound due to possible correlation with the dictionary error. Applying Cauchy-Schwarz, the triangle inequality, and Hölder's inequality yields the following upper bound:

$$\begin{aligned}
&\left|\left\langle \frac{2}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)}\right)\left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*},\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\rangle_{F}\right| \\
&\quad\leq \frac{2}{N}\sum_{k=1}^{N}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)}\right)\right\|_{2,\infty}\cdot\left\|\left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*}\right\|_{1,1}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F} \\
&\quad\leq \frac{C\cdot s\cdot m\cdot d\cdot\mu^{(\ell)}}{2}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F}.
\end{aligned}$$

(a2) In this term, we cannot avoid correlation between the noise cross-correlation term and coefficient error since they exhibit covariance. Again, a worst case bound follows from Cauchy-Schwarz, the triangle inequality, and Hölder's inequality:

$$\begin{aligned}
&\left|\left\langle \frac{2}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)}\right)\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)^{*},\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\rangle_{F}\right| \\
&\quad\leq \frac{2}{N}\sum_{k=1}^{N}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)}\right)\right\|_{2,\infty}\cdot\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,1}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F} \\
&\quad\leq \left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^{2}\cdot\frac{24\cdot s\cdot m\cdot d\cdot\left(\mu^{(\ell)}\right)^{2}}{(1-\delta_{1})\cdot\left(\kappa_{s}^{\star}-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^{2}}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F}.
\end{aligned}$$

(b1) We follow a familiar strategy of Cauchy-Schwarz, the triangle inequality, and Hölder's inequality to yield

$$\begin{aligned}
&\left|\left\langle \frac{2}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}^{\star}\mathbf{P}^{*}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*},\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\rangle_{F}\right| \\
&\leq \frac{2}{N}\sum_{k=1}^{N}\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right\|\cdot\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,2}\cdot\left\|\mathbf{P}\mathbf{C}_{(k)}^{\star}\right\|_{1,2}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F} \\
&\leq \left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^{2}+\delta_{1}\right]\cdot\frac{96\cdot C\cdot s^{2}\cdot\mu^{(\ell)}}{(1-\delta_{1})\cdot\left(\kappa_{s}^{\star}-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^{2}} \\
&\quad\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F}.
\end{aligned}$$

(b2) Again, the same sequence results in the following bound:

$$\begin{aligned}
&\left|\left\langle \frac{2}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}^{\star}\mathbf{P}^{*}\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)^{*},\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\rangle_{F}\right| \\
&\leq \frac{2}{N}\sum_{k=1}^{N}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\right\|\cdot\frac{1}{d}\cdot\left\|\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,2}^{2}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F} \\
&\leq \left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^{2}+\delta_{1}\right]\cdot\left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^{2} \\
&\quad\cdot\left[\frac{48\cdot s\cdot\mu^{(\ell)}}{(1-\delta_{1})\cdot\left(\kappa_{s}^{\star}-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^{2}}\right]^{2}\cdot\left\|\mathbf{D}^{\star}\mathbf{P}^{*}-\mathbf{D}^{(\ell)}\right\|_{F}.
\end{aligned}$$

(3) A lower bound for the left-hand side follows from the conditions of the statement and Lemma 4.6,

$$\begin{aligned}
&\frac{1}{N \cdot T \cdot d} \sum_{k=1}^{N} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right) \mathbf{C}_{(k)}^{(\ell)} \right\|_F^2 \\
&\geq \lambda_{\min} \left(\frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{d} \mathbf{C}_{(k)}^{(\ell)} \left(\mathbf{C}_{(k)}^{(\ell)} \right)^{*} \right) \otimes \left(\frac{1}{T} \mathbf{X}_{(k)}^{*} \mathbf{X}_{(k)} \right) \right) \cdot \left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F^2 \\
&\geq \left\{ \frac{(1 - \delta_2) \cdot s \cdot c^2}{r} \cdot \left(\frac{\nu}{1 + \sqrt{d} \cdot s \cdot C} \right)^2 \right. \\
&\quad \left. - \left(\frac{1 + \sqrt{d} \cdot s \cdot C}{\nu} \right)^2 \cdot \frac{24 \cdot s \cdot m \cdot d \cdot \left(\mu^{(\ell)} \right)^2}{(1 - \delta_1) \cdot \varepsilon^{(\ell-1)} \cdot \left(\kappa_s^{\star} - 4 \cdot \sqrt{\frac{s}{m \cdot d}} \cdot \varepsilon^{(\ell-1)} \right)^2} \right\} \\
&\quad \cdot \left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F^2 .
\end{aligned}$$

Dividing through by $\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F$ will yield an upper bound on the dictionary error in terms of computable quantities. We will make sense of the result in the next part.

(4) At a high-level, we have for some positive constants C_1, C_2, C_3, and C_4:

$$\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F \leq \frac{C_1 \cdot \mu^{(\ell)} + C_2 \cdot \left(\mu^{(\ell)} \right)^2}{C_3 - C_4 \cdot \left(\mu^{(\ell)} \right)^2}.$$

This is good news because this is positive for $\mu^{(\ell)} \in (0, C_3/C_4)$. Moreover, it achieves the correct asymptotic behavior:

$$\lim_{\mu^{(\ell)} \to 0^{+}} \frac{C_1 \cdot \mu^{(\ell)} + C_2 \cdot \left(\mu^{(\ell)} \right)^2}{C_3 - C_4 \cdot \left(\mu^{(\ell)} \right)^2} = 0.$$

Through this bound, we can drive the error to zero with $\mu^{(\ell)}$.

We would like to achieve linear convergence, *i.e.* $\left\| \mathbf{D}^{\star} \mathbf{P}^{*} - \mathbf{D}^{(\ell)} \right\|_F \leq \alpha \cdot \varepsilon^{(\ell-1)}$, where $\alpha \in (0, 1)$ is independent of ℓ. This leads to the following quadratic equation

in $\mu^{(\ell)}$:

$$\begin{aligned}
&\frac{C\cdot s\cdot m\cdot d\cdot\mu^{(\ell)}}{2}+\left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2\cdot\frac{24\cdot s\cdot m\cdot d\cdot\left(\mu^{(\ell)}\right)^2}{(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}\\
&+\left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^2+\delta_1\right]\cdot\frac{96\cdot C\cdot s^2\cdot\mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}\\
&+\left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^2+\delta_1\right]\cdot\left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2\\
&\quad\cdot\left[\frac{48\cdot s\cdot\mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}\right]^2\\
&\leq\alpha\cdot\varepsilon^{(\ell-1)}\cdot\left\{\frac{(1-\delta_2)\cdot s\cdot c^2}{r}\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right.\\
&\left.-\left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2\cdot\frac{24\cdot s\cdot m\cdot d\cdot\left(\mu^{(\ell)}\right)^2}{(1-\delta_1)\cdot\varepsilon^{(\ell-1)}\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}\right\}.
\end{aligned}$$

The solution will yield one positive real root, $\mu_+\left(\varepsilon^{(\ell-1)}\right)$, and one negative real root, $\mu_-\left(\varepsilon^{(\ell-1)}\right)$, functions of $\varepsilon^{(\ell-1)}$. Since we are interested in only positive solutions, any $\mu^{(\ell)}\in\left(0,\mu_+\left(\varepsilon^{(\ell-1)}\right)\right)$ will satisfy the desired result. We have

$$\begin{aligned}
&\mu_+\left(\varepsilon^{(\ell-1)}\right)\\
&=\frac{(1-\delta_1)\cdot B\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}{96\cdot\sqrt{r}\cdot\left(m\cdot d\cdot(1+\alpha)\cdot(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2+96\cdot B\cdot s\right)}\\
&\left[-C\cdot m\cdot d\cdot\sqrt{r}\cdot(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2\right.\\
&+\left(C^2\cdot m^2\cdot d^2\cdot r\cdot(1-\delta_1)^2\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^4\right.\\
&+384\cdot\alpha\cdot\varepsilon^{(\ell-1)}\cdot(1-\delta_2)\cdot c^2\\
&\left.\left.\cdot\left(m\cdot d\cdot(1-\delta_1)\cdot\left(\kappa_s^\star-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2\cdot(1+\alpha)+96\cdot B\cdot s\right)\right)^{1/2}\right],
\end{aligned}$$

where $B=\left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^2+\delta_1\right]$. □

4.3.3 Bounding the error of the coefficient estimation

In this section, we bound the error of the coefficient estimate of Algorithm 2. The proof technique in principle follows that of Bickel, *et al.* [12], to which we owe the restricted eigenvalue analysis technique. The adaptation to autoregressive models is not novel, as it has been done previously in Basu and Michalidis [10] and Kock and Callot [64]. In fact, there is a considerable history of using ℓ^1 penalization for autoregressive modeling [110, 102, 54, 53, 36]. Our problem introduces new challenges due to the matrix factorization of the autoregressive parameters; however, we retain the linear dependence on the sparsity s and penalty parameter $\mu^{(\ell)}$ as in Basu and Michalidis [10] and Kock and Callot [64]. This is important as the linear dependence on $\mu^{(\ell)}$ provides us a coercive mechanism to drive the estimate toward zero with successive iterations.

Assumption (4) of Lemma 4.5 is unique to our problem. It can be understood as a control on the second moment of the cross correlation of the autocovariance and coefficients of the autoregressive process. Recall $\mathbb{E}\left[\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}|\mathbf{C}_{(k)}\right] = \mathbf{R}\left(\mathbf{C}_{(k)}^{\star}\right)$. In this way, we are trying to bound a map

$$\mathbf{D} \mapsto \left\langle \mathbf{D}, \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)^{2} \mathbf{D}\mathbf{C}_{(k)}^{\star}\left(\mathbf{C}_{(k)}^{\star}\right)^{*} \right\rangle_{F}$$

This condition represents the most difficult event to characterize analytically, and also achieve probabilistically. The other assumptions are standard (*c.f.* [10, 64]).

Lemma 4.5. *Let* $\mathbf{D}^{\star}$, $\left(\mathbf{C}_{(k)}^{\star}\right)_{k=1,\ldots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\ldots,N}$ *be generated according to Model 2. Assume the following conditions:*

(1) $\left\|\mathbf{D}^{\star} - \mathbf{D}^{(\ell-1)}\mathbf{P}\right\|_F \leq \varepsilon^{(\ell-1)}$;

(2) $\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$ *for all* $k = 1, \ldots, N$;

(3) $\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{N}_{(k)}\right\|_{2,\infty} \leq \frac{m\cdot d\cdot \mu^{(\ell)}}{4}$ *for all* $k = 1, \ldots, N$; *and*

(4) $\max_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\mathbf{C}_{(k)}^{\star}\right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq \frac{m\cdot d\cdot \mu^{(\ell)}}{4\cdot \varepsilon^{(\ell-1)}}$ *for all* $k = 1, \ldots, N$.

Then, the coefficient update of Algorithm 2 will satisfy

$$\left\|\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right\|_{1,\infty} \leq \left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2 \cdot \frac{48\cdot s\cdot \mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa_s^{\star} - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}. \tag{4.14}$$

Proof. First, note that the optimization problem decouples along the columns of the coefficient matrix,

$$\begin{aligned}
&\arg\min_{\mathbf{C}\in\mathbb{R}^{r\times d}} \frac{1}{T\cdot m\cdot d}\left\|\mathbf{Y}_{(k)} - \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{C}\right\|_F^2 + 2\mu^{(\ell)}\left\|\mathbf{C}\right\|_{1,1} \\
&\quad = \left(\arg\min_{\mathbf{c}\in\mathbb{R}^{r}} \frac{1}{T\cdot m\cdot d}\left\|\left(\mathbf{Y}_{(k)}\right)_i - \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{c}\right\|^2 + 2\mu^{(\ell)}\left\|\mathbf{c}\right\|_1\right)_{i=1,\ldots,d}.
\end{aligned}$$

We will restrict our attention to a single column at this time. As $\left(\mathbf{C}_{(k)}^{(\ell)}\right)_i$ is a minimizer, we have

$$\begin{aligned}
&\frac{1}{T\cdot m\cdot d}\left\|\left(\mathbf{Y}_{(k)}\right)_i - \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)_i\right\|^2 + 2\mu^{(\ell)}\left\|\left(\mathbf{C}_{(k)}^{(\ell)}\right)_i\right\|_1 \\
&\quad \leq \frac{1}{T\cdot m\cdot d}\left\|\left(\mathbf{Y}_{(k)}\right)_i - \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)_i\right\|^2 + 2\mu^{(\ell)}\left\|\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)_i\right\|_1.
\end{aligned}$$

where $\mathbf{C}_{(k)}^{\star}$ is the true coefficient matrix. We substitute $\left(\mathbf{Y}_{(k)}\right)_i = \mathbf{X}_{(k)}\mathbf{D}^{\star}\left(\mathbf{C}_{(k)}^{\star}\right)_i + \left(\mathbf{N}_{(k)}\right)_i$ which yields

$$\begin{aligned}
&\frac{1}{T\cdot m\cdot d}\left\|\mathbf{X}_{(k)}\left(\mathbf{D}^{\star}\left(\mathbf{C}_{(k)}^{\star}\right)_i - \mathbf{D}^{(\ell-1)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)_i\right) + \left(\mathbf{N}_{(k)}\right)_i\right\|^2 + 2\mu^{(\ell)}\left\|\left(\mathbf{C}_{(k)}^{(\ell)}\right)_i\right\|_1 \\
&\quad \leq \frac{1}{T\cdot m\cdot d}\left\|\mathbf{X}_{(k)}\left(\mathbf{D}^{\star} - \mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}_{(k)}^{\star}\right)_i + \left(\mathbf{N}_{(k)}\right)_i\right\|^2 + 2\mu^{(\ell)}\left\|\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)_i\right\|_1.
\end{aligned}$$

Expanding the norms, combining common terms, and rearranging yields

$$\begin{aligned}
&\frac{1}{T \cdot m \cdot d} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} \left(\mathbf{C}^{\star}_{(k)} \right)_i - \mathbf{D}^{(\ell-1)} \left(\mathbf{C}^{(\ell)}_{(k)} \right)_i \right) \right\|^2 \\
&\quad \leq \frac{1}{T \cdot m \cdot d} \left\| \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} - \mathbf{D}^{(\ell-1)} \mathbf{P} \right) \left(\mathbf{C}^{\star}_{(k)} \right)_i \right\|^2 \\
&\qquad + \frac{2}{T \cdot m \cdot d} \left\langle \mathbf{X}_{(k)} \mathbf{D}^{(\ell-1)} \left(\left(\mathbf{C}^{(\ell)}_{(k)} \right)_i - \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right), \left(\mathbf{N}_{(k)} \right)_i \right\rangle \\
&\qquad + 2\mu^{(\ell)} \left(\left\| \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right\|_1 - \left\| \left(\mathbf{C}^{(\ell)}_{(k)} \right)_i \right\|_1 \right).
\end{aligned}$$

We then substitute

$$\begin{aligned}
&\mathbf{D}^{\star} \left(\mathbf{C}^{\star}_{(k)} \right)_i - \mathbf{D}^{(\ell-1)} \left(\mathbf{C}^{(\ell)}_{(k)} \right)_i \\
&\quad = \left(\mathbf{D}^{\star} - \mathbf{D}^{(\ell-1)} \mathbf{P} \right) \left(\mathbf{C}^{\star}_{(k)} \right)_i - \mathbf{D}^{(\ell-1)} \left(\left(\mathbf{C}^{(\ell)}_{(k)} \right)_i - \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right),
\end{aligned}$$

expand the norm, combine common terms, and rearrange to yield

$$\begin{aligned}
&\frac{1}{T \cdot m \cdot d} \left\| \mathbf{X}_{(k)} \mathbf{D}^{(\ell-1)} \left(\left(\mathbf{C}^{(\ell)}_{(k)} \right)_i - \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right) \right\|^2 \\
&\quad \leq \frac{2}{T \cdot m \cdot d} \left\langle \mathbf{X}_{(k)} \left(\mathbf{D}^{\star} - \mathbf{D}^{(\ell-1)} \mathbf{P} \right) \left(\mathbf{C}^{\star}_{(k)} \right)_i, \mathbf{X}_{(k)} \mathbf{D}^{(\ell-1)} \left(\left(\mathbf{C}^{(\ell)}_{(k)} \right)_i - \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right) \right\rangle \\
&\qquad + \frac{2}{T \cdot m \cdot d} \left\langle \mathbf{X}_{(k)} \mathbf{D}^{(\ell-1)} \left(\left(\mathbf{C}^{(\ell)}_{(k)} \right)_i - \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right), \left(\mathbf{N}_{(k)} \right)_i \right\rangle \\
&\qquad + 2\mu^{(\ell)} \left(\left\| \left(\mathbf{P} \mathbf{C}^{\star}_{(k)} \right)_i \right\|_1 - \left\| \left(\mathbf{C}^{(\ell)}_{(k)} \right)_i \right\|_1 \right).
\end{aligned}$$

We would like to upper bound the first and second terms on the right hand

side with respect to $\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1$ and $\mu^{(\ell)}$. Consider the first term:

$$\begin{aligned}
&\frac{2}{T\cdot m\cdot d}\left\langle \mathbf{X}_{(k)}\left(\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}^{\star}_{(k)}\right)_i, \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\left(\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right)\right\rangle \\
&= \frac{2}{m\cdot d}\left\langle \left(\mathbf{D}^{(\ell-1)}\right)^*\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\left(\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}^{\star}_{(k)}\right)_i, \left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\rangle \\
&\le \frac{2}{m\cdot d}\left\|\left(\mathbf{D}^{(\ell-1)}\right)^*\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\left(\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}^{\star}_{(k)}\right)_i\right\|_\infty \\
&\qquad\cdot\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 \\
&\le \frac{2}{m\cdot d}\left\|\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\left(\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}^{\star}_{(k)}\right)_i\right\|\cdot\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1.
\end{aligned}$$

Now, using conditions (1) and (4) of the lemma, we have:

$$\begin{aligned}
&\frac{2}{m\cdot d}\cdot\left\|\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\left(\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right)\left(\mathbf{C}^{\star}_{(k)}\right)_i\right\| \\
&\quad\le \frac{2}{m\cdot d}\cdot\frac{m\cdot d\cdot\mu^{(\ell)}}{4\cdot\varepsilon^{(\ell-1)}}\cdot\left\|\mathbf{D}^{\star}-\mathbf{D}^{(\ell-1)}\mathbf{P}\right\|_F \\
&\quad\le \frac{\mu^{(\ell)}}{2}.
\end{aligned}$$

Now, consider the second term:

$$\begin{aligned}
&\frac{2}{T\cdot m\cdot d}\cdot\left\langle \mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\left(\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right), \left(\mathbf{N}_{(k)}\right)_i\right\rangle \\
&= \frac{2}{m\cdot d}\cdot\left\langle \left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i, \frac{1}{T}\left(\mathbf{D}^{(\ell-1)}\right)^*\mathbf{X}^*_{(k)}\left(\mathbf{N}_{(k)}\right)_i\right\rangle \\
&\le \frac{2}{m\cdot d}\cdot\left\|\frac{1}{T}\left(\mathbf{D}^{(\ell-1)}\right)^*\mathbf{X}^*_{(k)}\left(\mathbf{N}_{(k)}\right)_i\right\|_\infty\cdot\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 \\
&\le \frac{2}{m\cdot d}\cdot\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\left(\mathbf{N}_{(k)}\right)_i\right\|\cdot\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1.
\end{aligned}$$

By the conditions of the lemma, $2/(m\cdot d)\cdot\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\left(\mathbf{N}_{(k)}\right)_i\right\| \le \mu^{(\ell)}/2$.

Now, we have the following upper for the left-hand side:

$$\begin{aligned}
&\frac{1}{T\cdot m\cdot d}\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\left(\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right)\right\|^2 \\
&\quad\le \mu^{(\ell)}\left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 + 2\mu^{(\ell)}\left(\left\|\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 - \left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i\right\|_1\right).
\end{aligned}$$

Let $J = \operatorname{supp}\left(\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right)$ and define $\mathbf{h} = \left(\mathbf{C}^{(\ell)}_{(k)}\right)_i - \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i$. We can upper bound $1/(T\cdot m\cdot d)\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2$ using $\left\|\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 - \left\|\left(\mathbf{C}^{(\ell)}_{(k)}\right)_i\right\|_1 \leq \|\mathbf{h}\|_1$,

$$\frac{1}{T\cdot m\cdot d}\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2 \leq 3\cdot\mu^{(\ell)}\cdot\|\mathbf{h}\|_1 .$$

Before proceeding, we note from $1/(T\cdot m\cdot d)\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2 \geq 0$ that

$$\begin{aligned}0 &\leq \mu^{(\ell)}\|\mathbf{h}_J\|_1 + \mu^{(\ell)}\|\mathbf{h}_{J^C}\|_1 + 2\mu^{(\ell)}\left(\left\|\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 - \left\|\mathbf{h}_J + \left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)_i\right\|_1 - \|\mathbf{h}_{J^C}\|_1\right)\\ &\leq 3\mu^{(\ell)}\|\mathbf{h}_J\|_1 - \mu^{(\ell)}\|\mathbf{h}_{J^C}\|_1 .\end{aligned}$$

This leads to the conclusion $\|\mathbf{h}_{J^C}\|_1 \leq 3\|\mathbf{h}_J\|_1$, and eventually $\|\mathbf{h}\|_1^2 \leq 16\cdot s\cdot\|\mathbf{h}_J\|^2$. We will require these relationships to complete the proof.

Now, we begin lower bounding $1/(T\cdot m\cdot d)\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|_F^2$. By the assumptions of the proposition,

$$\begin{aligned}\sigma_{\min}\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right) &\geq \sigma_{\min}\left(\mathbf{R}_{(k)}\right) - \left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\|\\ &\geq (1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2 .\end{aligned}$$

Therefore,

$$\frac{1}{T\cdot m\cdot d}\left\|\mathbf{X}_{(k)}\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2 \geq (1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\cdot\frac{1}{m\cdot d}\left\|\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2 .$$

By Lemma 4.12,

$$\frac{1}{m\cdot d}\left\|\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|^2 = \left(\frac{\left\|\mathbf{D}^{(\ell-1)}\mathbf{h}\right\|}{\sqrt{m\cdot d}}\right)^2 \geq \left(\kappa^{\star}_s - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2\cdot\|\mathbf{h}_J\|^2 .$$

We now have

$$(1-\delta_1)\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\cdot\left(\kappa^{\star}_s - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2\cdot\|\mathbf{h}_J\|^2 \leq 3\mu^{(\ell)}\|\mathbf{h}\|_1 .$$

Using $\|\mathbf{h}\|_1^2 \leq 16 \cdot s \cdot \|\mathbf{h}_J\|^2$ yields

$$\|\mathbf{h}\|_1 \leq \left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2 \cdot \frac{48\cdot s\cdot \mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa_s^\star - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}.$$

Although we have only calculated a bound for the coefficient error of a single column, the bound is uniform across the columns $i = 1, \ldots, d$ by the assumptions of the Lemma. This completes the argument. □

4.3.4 Deriving a lower bound to isolate the dictionary error

This result is unique to this atomic decomposition of autoregressive processes. In Lemma 4.4, we end-up with an expression,

$$\frac{1}{N\cdot T\cdot d}\sum_{k=1}^{N}\left\|\mathbf{X}_k\left(\mathbf{D}^\star\mathbf{P}^* - \mathbf{D}^{(\ell)}\right)\mathbf{C}_k^{(\ell)}\right\|_F^2 \leq \text{r.h.s.}$$

We want to lower bound this quantity with respect to the dictionary error, and so we require a bound on the smallest eigenvalue of the non-negative map

$$\mathbf{D} \mapsto \left\langle \mathbf{D}, \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)}\right)\mathbf{D}\left(\frac{1}{d}\mathbf{C}_{(k)}^{(\ell)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)^*\right)\right\rangle_F.$$

We prove Lemma 4.6 using a straightforward deviation to separate the purely random part of this operator from that which depends on Algorithm 2. This allows us to bound the smallest eigenvalue of the purely random operator,

$$,\mathbf{D} \mapsto \left\langle \mathbf{D}, \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)}\right)\mathbf{D}\left(\frac{1}{d}\mathbf{C}_{(k)}^{\star}\left(\mathbf{C}_{(k)}^{\star}\right)^*\right)\right\rangle_F$$

using Corollary 2.5.1, a matrix concentration inequality of Tropp [108]. This result is captured in condition (4) of Lemma 4.6.

Lemma 4.6. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 2. Assume the following conditions:*

(1) $\left\|\mathbf{D}^\star - \mathbf{D}^{(\ell-1)}\mathbf{P}\right\|_F \leq \varepsilon^{(\ell-1)}$;

(2) $\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$ *for all* $k = 1,\dots,N$;

(3) $\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} \leq \frac{m\cdot d\cdot \mu^{(\ell)}}{4}$ *for all* $k = 1,\dots,N$;

(4) $\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{P}\mathbf{C}^\star_{(k)}\left(\mathbf{P}\mathbf{C}^\star_{(k)}\right)^*\right)\otimes\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\right) \geq (1-\delta_2)\cdot\frac{s\cdot c^2}{r}\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$; *and*

(5) $\max_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}\frac{\left\|\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\mathbf{D}\mathbf{C}^\star_{(k)}\right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq \frac{m\cdot d\cdot\mu^{(\ell)}}{4\cdot\varepsilon^{(\ell-1)}}$ *for all* $k = 1,\dots,N$.

Then, the coefficient estimates of Algorithm 2 will satisfy

$$\begin{aligned}\lambda_{\min}&\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{C}^{(\ell)}_{(k)}\left(\mathbf{C}^{(\ell)}_{(k)}\right)^*\right)\otimes\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\right)\\ &\geq \frac{(1-\delta_2)\cdot s\cdot c^2\cdot\nu^2}{r\cdot\left(1+\sqrt{d}\cdot s\cdot C\right)^2} - \left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^2\\ &\quad\cdot\frac{24\cdot s\cdot m\cdot d\cdot\left(\mu^{(\ell)}\right)^2}{(1-\delta_1)\cdot\varepsilon^{(\ell-1)}\cdot\left(\kappa^\star_s - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^2}.\end{aligned} \tag{4.15}$$

Proof. We begin with telescoping

$$
\begin{aligned}
&\frac{1}{d}\mathbf{C}_{(k)}^{(\ell)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)^{*} \\
&= \frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*} - \left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*} - \frac{1}{d}\mathbf{C}_{(k)}^{(\ell)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)^{*}\right) \\
&= \frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*} - \frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)^{*} \\
&\quad - \frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)\left(\mathbf{C}_{(k)}^{(\ell)}\right)^{*} \\
&= \frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*} - \frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)^{*} \\
&\quad - \frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*} + \frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)^{*}.
\end{aligned}
$$

From this relationship, we can derive the following using Weyl's inequality:

$$
\begin{aligned}
&\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{C}_{(k)}^{(\ell)}\left(\mathbf{C}_{(k)}^{(\ell)}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\right) \\
&\geq \underbrace{\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{P}\mathbf{C}_{(k)}^{\star}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\right)}_{\text{(a)}} \\
&\quad - 2\cdot\underbrace{\left\|\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star} - \mathbf{C}_{(k)}^{(\ell)}\right)\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\right\|}_{\text{(b)}}.
\end{aligned}
$$

We have a lower bound for term (a) from condition (4) of the lemma, so we want to upper bound term (b). After applying the triangle inequality, we can decouple the norms of the respective matrices in the Kronecker product. Then, we will use

the variational definition of a singular value:

$$\begin{aligned}
&\max_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}\right)\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)^{*}\right)\right\|_{F}}{\|\mathbf{D}\|_{F}} \\
&\leq \left(\max_{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\mathbf{P}\mathbf{C}_{(k)}^{\star}\right\|_{2,\infty}}{\|\mathbf{D}\|_{F}}\right)\cdot\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,1} \\
&= \left(\max_{\tilde{\mathbf{D}}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\tilde{\mathbf{D}}\mathbf{C}_{(k)}^{\star}\right\|_{2,\infty}}{\left\|\tilde{\mathbf{D}}\mathbf{P}^{-1}\right\|_{F}}\right)\cdot\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,1} \\
&= \left(\max_{\tilde{\mathbf{D}}\in\mathbb{R}^{(m\cdot d)\times r}} \frac{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\tilde{\mathbf{D}}\mathbf{C}_{(k)}^{\star}\right\|_{2,\infty}}{\left\|\tilde{\mathbf{D}}\right\|_{F}}\right)\cdot\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,1}.
\end{aligned}$$

We can upper bound the first term for all $k = 1, \ldots, N$ using condition (5) of the lemma, and using conditions (1)-(3), and (5), we can use Lemma 4.5 to bound the second term for all $k = 1, \ldots, N$:

$$\left\|\frac{1}{d}\left(\mathbf{P}\mathbf{C}_{(k)}^{\star}-\mathbf{C}_{(k)}^{(\ell)}\right)\right\|_{1,1} \leq \left(\frac{1+\sqrt{d}\cdot s\cdot C}{\nu}\right)^{2}\cdot\frac{48\cdot s\cdot\mu^{(\ell)}}{(1-\delta_1)\cdot\left(\kappa_s^{\star}-4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon^{(\ell-1)}\right)^{2}}.$$

□

4.3.5 Simulations

In this section, we report results from applying Algorithm 2 to simulated data. The simulated data was generated according to Model 2 using $d = 4$, $m = 2$, and $s = 2$. We evaluated accuracy using the dictionary metric scaled by the number of dictionary atoms. For each condition, we ran twenty simulations and reported statistics over those experiments. The inner loop of the algorithm includes an iterative estimator for the coefficient estimate of every observation.

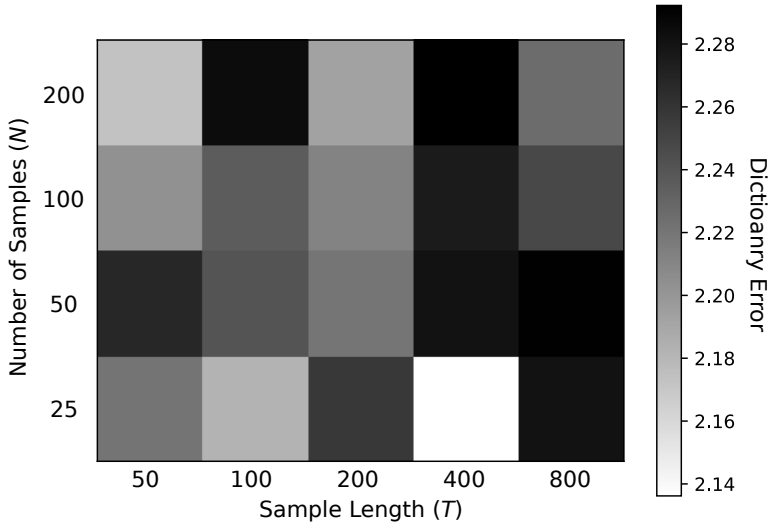


Figure 4.4: $r^{-1/2}d\left(\mathbf{D}^{\star}, \hat{\mathbf{D}}\right)$. We cannot discern the expected behavior of the error as a function of N and T. Here, we use $r = 9$. Likely, we cannot simulate sufficient size problems to overcome the finite sample factors in the result.

4.4 Lemmata

In this section, we prove the majority of technical lemmas required for results in Sec. 4.2 and 4.3.

4.4.1 Consequences of Models 1 and 2

Lemma 4.7. *Let* $\left(\mathbf{C}^{\star}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 1 or 2. Then, for all* $k = 1, \dots, N$,

$$\left\|\mathbf{C}^{\star}_{(\mathbf{k})}\right\| \leq \left\|\mathbf{C}^{\star}_{(k)}\right\|_{1,2} \leq \sqrt{d} \cdot s \cdot C.$$

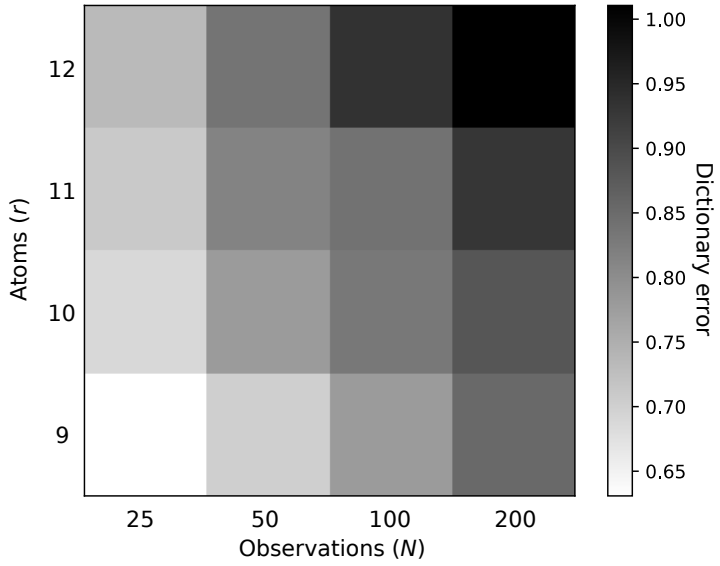


Figure 4.5: $r^{-1/2}d\left(\mathbf{D}^\star, \hat{\mathbf{D}}\right)$. We see linear increase in error with increases in dictionary redundancy. The mixed effects are less discernible due to the implicit dependence on observation length.

Proof. This is shown through Hölder's inequality:

$$\begin{aligned}
\left\|\mathbf{C}^\star_{(k)}\right\| &= \sup_{\mathbf{v}\in\mathbf{S}^{r-1}} \left\|\mathbf{C}^\star_{(k)}\mathbf{v}\right\| \\
&= \sup_{\mathbf{v}\in\mathbf{S}^{r-1}} \left(\sum_{j=1}^{d} \left|\left\langle \left(\mathbf{C}^\star_{(k)}\right)_j, \mathbf{v}\right\rangle\right|^2\right)^{1/2} \\
&\leq \sup_{\mathbf{v}\in\mathbf{S}^{r-1}} \left(\sum_{j=1}^{d} \left\|\left(\mathbf{C}^\star_{(k)}\right)_j\right\|_1^2 \cdot \|\mathbf{v}\|_\infty^2\right)^{1/2} \\
&= \left\|\mathbf{C}^\star_{(k)}\right\|_{1,2} \\
&\leq \sqrt{d}\cdot s \cdot C.
\end{aligned}$$

The final inequality follows from Model 1 and 2, in which the column support is fixed of size s and entries bounded by C. □

Lemma 4.8. *Let* $\mathbf{C}^\star_{(k)}$, $\mathbf{D}^\star$, *and* $\mathbf{N}_{(k)}$ *satisfy the conditions of Model 1 or 2. Then, for all* $k = 1, \ldots, N$

$$\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2 \leq \sigma_{min}\left(\mathbf{R}_{(k)}\right) \leq \sigma_{max}\left(\mathbf{R}_{(k)}\right) \leq \left(\frac{\nu}{1-\sqrt{d}\cdot s\cdot C}\right)^2 \tag{4.16}$$

almost surely.

Proof. This is a simple consequence of Thm. 2.11 and Lemma 4.7 after noting

$$\begin{aligned}\left\|\mathbf{D}^\star\mathbf{C}^\star_{(k)}\right\| &\leq \left\|\mathbf{D}^\star\mathbf{C}^\star_{(k)}\right\|_F \\ &\leq \left\|\mathbf{D}^\star\right\|_{2,\infty}\cdot\left\|\mathbf{C}^\star_{(k)}\right\|_{1,2} \\ &= \left\|\mathbf{C}^\star_{(k)}\right\|_{1,2}.\end{aligned}$$

Here, we have used a variant of Hölder's inequality for the Frobenius norm and that $\mathbf{D}^\star$ has unit norm columns. □

Lemma 4.9. *Let* $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\ldots,N}$ *be generated according to Model 2. Then,*

$$\mathbb{E}\frac{1}{d}\mathbf{C}^\star_{(k)}\left(\mathbf{C}^\star_{(k)}\right)^* = \frac{s\cdot c^2}{r}\mathbf{I}$$

for all $k = 1, \ldots, N$.

Proof. Note that $\frac{1}{d}\mathbf{C}^\star_{(k)}\left(\mathbf{C}^\star_{(k)}\right)^*$ is the sample covariance:

$$\frac{1}{d}\mathbf{C}^\star_{(k)}\left(\mathbf{C}^\star_{(k)}\right)^* = \frac{1}{d}\sum_{i=1}^{d}\left(\mathbf{C}^\star_{(k)}\right)_i\left(\mathbf{C}^\star_{(k)}\right)_i^*.$$

Therefore, we first recognize that this is equivalent to the column-wise covariance. There are two independent random experiments that determine the coefficients of a column: the choice of support and the value of non-zero coefficients. The support must be of size s, and it is chosen uniformly at random. Thus, the probability of

an element being in the support is s/r, and that two elements are in the support $(s/r)^2$. However, the non-zero coefficients are then chosen independently from a centered Gaussian with variance c^2, which eliminates all cross-terms. This provides the desired result. □

4.4.2 Concentration of $\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\|$

Lemma 4.10. *Let $\mathbf{D}^{\star}$, $\mathbf{C}_{(k)}^{\star}$, and $\mathbf{N}_{(k)}$ be consistent with Model 2 or Model 1. Then, for any $k = 1, \ldots, N$ and every $\delta_1 \in (0,1)$,*

$$\mathbb{P}\left(\left\{\left\|\mathbf{R}_{(k)} - \frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right\| > \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\}\right) \leq 2\cdot m\cdot 9^{(m\cdot d)} \cdot \exp\left(-c_1\cdot(T/m)\cdot\min\left(\left[\frac{\delta_1\cdot\nu^4}{2K_1\cdot(1-d\cdot s^2\cdot C^2)^2}\right]^2, \frac{\delta_1\cdot\nu^4}{2K_1\cdot(1-d\cdot s^2\cdot C^2)^2}\right)\right), \tag{4.17}$$

where $c_1 > 0$ is a global constant.

Proof. We will prove the statement in two parts. First, we will fix a $\mathbf{C}_{(k)}^{\star}$ and concentrate $\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\|$. Then, we will use this result to prove the global bound. This proof approach is adapted from that of Thm. 2.6 of Vershynin [111].

(Fixed $\mathbf{C}_{(k)}^{\star}$) For a fixed $\mathbf{C}_{(k)}^{\star}$ and $\delta > 0$, we want to upper bound

$$\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta\right\}\middle|\mathbf{C}_{(k)}^{\star}\right)$$

Fix a 1/4-net of $\mathbf{S}^{(m\cdot d)-1}$, $\mathcal{K}_{1/4}$, so that by Lemma 2.8,

$$\left\|\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \leq 2\cdot\max_{\mathbf{v}\in\mathcal{K}_{1/4}}\left|\left\langle\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right)\mathbf{v},\mathbf{v}\right\rangle\right|.$$

Now, we want to show that $\left|\frac{1}{T}\left\|\mathbf{X}\mathbf{v}\right\|_2^2 - \left\langle\mathbf{R}_{(k)}\mathbf{v},\mathbf{v}\right\rangle_2\right| \leq \delta/2$ for all $\mathbf{v}\in\mathcal{K}_{1/4}$.

For a fixed $\mathbf{v} \in \mathcal{K}_{1/4}$, we want to concentrate $\left|\frac{1}{T}\left\|\mathbf{X}_{(k)}\mathbf{v}\right\|^2 - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle\right|$. To do so, we want to express $\frac{1}{T}\left\|\mathbf{X}_{(k)}\mathbf{v}\right\|_2^2$ as a quadratic function of a Gaussian random vector. Let $\mathbf{v}^* = [\mathbf{v}_1^* \cdots \mathbf{v}_m^*]$, then,

$$\mathbf{X}_{(k)}\mathbf{v} = \left(\sum_{j=1}^{m} \langle \mathbf{x}_{t+j}, \mathbf{v}_j\rangle\right)_{t=0,\ldots,T-1}.$$

Recall that $\mathbf{x}_t = \left[(\mathbf{I}-\mathbf{A})^{-1}\mathbf{n}\right]_t = \left[(\mathbf{I}-\mathbf{A})^{-1}(\nu\mathbf{z})\right]_t$, and note $\sum_{j=1}^{m}\langle \mathbf{x}_{t+j}, \mathbf{v}_j\rangle = \left\langle \mathbf{R}_{(k)}^{1/2}\mathbf{z}_t, \mathbf{v}\right\rangle$ where $\mathbf{R}_{(k)}^{1/2}$ is the finite dimensional truncation of $\mathbf{z} \mapsto \left(\mathbf{I}-\mathbf{A}_{(k)}\right)^{-*}(\nu\mathbf{z})$ and $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Now, consider $\left\langle \mathbf{v}, \frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)}\mathbf{v}\right\rangle$,

$$\left\langle \mathbf{v}, \frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)}\mathbf{v}\right\rangle = \frac{1}{T}\sum_{t=1}^{T}\left(\left\langle \mathbf{R}_{(k)}^{1/2}\mathbf{z}_t, \mathbf{v}\right\rangle\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left\langle \mathbf{z}_t, \mathbf{R}_{(k)}^{1/2}\mathbf{v}\mathbf{v}^*\mathbf{R}_{(k)}^{1/2}\mathbf{z}_t\right\rangle.$$

Since $\mathbf{z}_t^* = \left[\mathbf{z}_{t,1}^* \cdots \mathbf{z}_{t,m}^*\right]$ is highly correlated with $\mathbf{z}_{t+1}, \ldots, \mathbf{z}_{t+m-1}$, we note the following (without loss of generality, assume $T(\mathrm{mod} m) = 0$):

$$\left\langle \mathbf{v}, \frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)}\mathbf{v}\right\rangle = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{(T/m)}\sum_{t=1}^{T/m}\left\langle \mathbf{z}_{t\cdot m+j}, \mathbf{R}_{(k)}^{1/2}\mathbf{v}\mathbf{v}^*\mathbf{R}_{(k)}^{1/2}\mathbf{z}_{t\cdot m+j}\right\rangle.$$

From this, we can derive the upper bound

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\frac{1}{T}\|\mathbf{X}\mathbf{v}\|_2^2 - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle_2\right| \le \delta/2\right\} \middle| \mathbf{C}_{(k)}^{\star}\right) \\
&\quad \le \sum_{j=1}^{m}\mathbb{P}\left(\left\{\left|\frac{1}{T/m}\sum_{t=1}^{T/m}\left\langle \mathbf{z}_{t\cdot m+j}, \mathbf{R}_{(k)}^{1/2}\mathbf{v}\mathbf{v}^*\mathbf{R}_{(k)}^{1/2}\mathbf{z}_{t\cdot m+j}\right\rangle - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle\right| \le \delta/2\right\}\right).
\end{aligned}$$

Note that we now have a sum of centered random variables which will allow us to use a Bernstein type inequality to bound the sum.

Theorem 2.9 (Hanson-Wright inequality) tells us that the summands are at

least sub-exponential:

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\left\langle \mathbf{z}_{t\cdot m+j}, \mathbf{R}_{(k)}^{1/2}\mathbf{v}\mathbf{v}^*\mathbf{R}_{(k)}^{1/2}\mathbf{z}_{t\cdot m+j}\right\rangle - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle\right| > c_t\right\}\right) \\
&\quad \leq 2\cdot\exp\left(-c\cdot\min\left(\left[\frac{c_t}{2K^2\cdot\left\langle \mathbf{v}, \mathbf{R}_{(k)}\mathbf{v}\right\rangle}\right]^2, \frac{c_t}{2K^2\cdot\left\langle \mathbf{v}, \mathbf{R}_{(k)}\mathbf{v}\right\rangle}\right)\right) \\
&\quad \leq 2\cdot\exp\left(-c\cdot\min\left(\left[\frac{c_t}{2K^2\cdot\left\|\mathbf{R}_{(k)}\right\|}\right]^2, \frac{c_t}{2K^2\cdot\left\|\mathbf{R}_{(k)}\right\|}\right)\right),
\end{aligned}$$

for any $c_t > 0$, c a global constant, and K the sub-Gaussian norm of $\mathbf{z}$. The second inequality follows from the fact that $\left\langle \mathbf{v}, \mathbf{R}_{(k)}\mathbf{v}\right\rangle \leq \left\|\mathbf{R}_{(k)}\right\|$ for all $\mathbf{v} \in \mathcal{K}_{1/4}$. The sub-exponential norm is $K_1/\left\|\mathbf{R}_{(k)}\right\|$ for some $K_1 > 0$ depending on K and c above. Thus, by Prop. 5.16 of Vershynin [111], we have

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\frac{1}{T/m}\sum_{t=1}^{T/m}\left\langle \mathbf{z}_{t\cdot m+j}, \mathbf{R}_{(k)}^{1/2}\mathbf{v}\mathbf{v}^*\mathbf{R}_{(k)}^{1/2}\mathbf{z}_{t\cdot m+j}\right\rangle - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle\right| \leq \delta/2\right\}\right) \\
&\quad \leq 2\cdot\exp\left(-c_1\cdot(T/m)\cdot\min\left(\left[\frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right]^2, \frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right)\right),
\end{aligned}$$

for $c_1 > 0$ a global constant. As this is independent of $j = 1, \ldots, m$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left|\frac{1}{T}\left\|\mathbf{X}_{(k)}\mathbf{v}\right\|_2^2 - \left\langle \mathbf{R}_{(k)}\mathbf{v}, \mathbf{v}\right\rangle_2\right| \leq \delta/2\right\}\middle|\mathbf{C}_{(k)}^{\star}\right) \\
&\quad \leq 2\cdot m\cdot\exp\left(-c_1\cdot(T/m)\cdot\min\left(\left[\frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right]^2, \frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right)\right).
\end{aligned}$$

Finally, by a union bound over all $\mathbf{v} \in \mathcal{K}_{1/4}$, with $\left|\mathcal{K}_{1/4}\right| \leq 9^{(m\cdot d)}$ by Lemma 2.7, we have the following upper bound:

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}_{(k)}^*\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta\right\}\middle|\mathbf{C}_{(k)}^{\star}\right) \\
&\quad \leq 2\cdot m\cdot 9^{(m\cdot d)}\cdot\exp\left(-c_1\cdot(T/m)\cdot\min\left(\left[\frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right]^2, \frac{\delta\cdot\left\|\mathbf{R}_{(k)}\right\|}{2\cdot K_1}\right)\right).
\end{aligned}$$

(Global bound) The above result concentrates $\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta\right\}$ in terms of $\left\|\mathbf{R}_{(k)}\right\|$, a function of $\mathbf{C}^\star_{(k)}$. Therefore, we let δ be defined as in the statement of the lemma and marginalize over $\mathbf{C}^\star_{(k)}$:

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\}\right) \\
&\quad = \int \mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| > \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\} \middle| \mathbf{C}^\star_{(k)}\right) \mathbb{P}\left(\mathbf{C}^\star_{(k)}\right).
\end{aligned}$$

We can lower bound the integrand with a global bound from Lemma 4.8, $\left\|\mathbf{R}_{(k)}\right\| \leq \left(\frac{\nu}{1-\sqrt{d}\cdot s\cdot C}\right)^2$, to yield the desired result. □

4.4.3 Concentration of the cross-correlation of the observations and noise

Lemma 4.11. *Let $\mathbf{D}^\star$, $\mathbf{C}^\star_{(k)}$, and $\mathbf{N}_{(k)}$ be consistent with Model 2 or Model 1. Then, for any $k = 1, \ldots, N$ and every $\delta > 0$,*

$$\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} > \delta\right\}\right) \leq \frac{m\cdot d\cdot \nu^4}{T\cdot\left(1-\sqrt{d}\cdot s\cdot C\right)^2\cdot\delta^2}. \tag{4.18}$$

Proof. We will prove the statement in two parts, first concentrating $\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} \leq \delta\right\}$ for a fixed $\mathbf{C}^\star_{(k)}$ and then marginalizing over $\mathbf{C}^\star_{(k)}$.

(Fixed $\mathbf{C}^\star_{(k)}$) We want to show that

$$\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} > \delta\right\} \middle| \mathbf{C}^\star_{(k)}\right) \leq \frac{m\cdot d\cdot \nu^4}{T\cdot\left(1-\sqrt{d}\cdot s\cdot C\right)^2\cdot\delta^2}$$

for every $\delta > 0$. For the remainder of this section, we will drop the k subscript for readability. Let us first visualize $\frac{1}{T}\mathbf{X}^*\mathbf{N}$:

$$\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)} = \underbrace{\begin{bmatrix} \frac{1}{T}\sum_{t=m}^{T+m} \mathbf{x}[t]\mathbf{n}^*[t+1] \\ \vdots \\ \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}[t]\mathbf{n}^*[t+m] \end{bmatrix}}_{\mathbb{R}^{(m\cdot d)\times d}}.$$

$\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)}$ contains the sample cross-correlations of $(\mathbf{x}[t])_{t=1,\ldots,T+m}$ and $(\mathbf{n}[t])_{t=m,\ldots,T+m+1}$ at lags of $1,\ldots,m$. As the observed process is causally dependent on the noise process, $\mathbf{x}[t] = \left[(\mathbf{I}-\mathbf{A})^{-1}\mathbf{n}\right][t]$, in expectation, the blocks of $\frac{1}{T}\mathbf{X}^*\mathbf{N}$ are $\mathbf{0}$. Any observed correlation is spurious.

Now, we consider the quantity that we must bound, the maximum column norm of $\frac{1}{T}\mathbf{X}^*\mathbf{N}$. Without loss of generality, assume that $T(\mathrm{mod}m) = 0$, and let $\tilde{\mathbf{x}}[t] = [\mathbf{x}[t]^* \cdots \mathbf{x}[t-m+1]^*]^*$. Note that $(\tilde{\mathbf{x}}[t])_t$ are identically distributed with covariance $\mathbf{R}_{(k)}$.

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^*\mathbf{N}\right\|_{2,\infty} > \delta\right\}\middle|\mathbf{C}^\star_{(k)}\right) \\
&= \mathbb{P}\left(\left\{\max_{i=1,\ldots,d}\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{n}_i[t+m]\tilde{\mathbf{x}}[t+m-1]\right\| > \delta\right\}\middle|\mathbf{C}^\star_{(k)}\right) \\
&= \mathbb{P}\left(\left\{\max_{i=1,\ldots,d}\left\|\frac{1}{m}\sum_{j=1}^{m}\frac{1}{T/m}\sum_{t=1}^{T/m}\mathbf{n}_i[t\cdot m+m]\tilde{\mathbf{x}}[t\cdot m+j-1]\right\| > \delta\right\}\middle|\mathbf{C}^\star_{(k)}\right) \\
&\leq \sum_{i=1}^{d}\sum_{j=1}^{m}\mathbb{P}\left(\left\{\left\|\frac{1}{T/m}\sum_{t=1}^{T/m}\mathbf{n}_i[t\cdot m+j]\tilde{\mathbf{x}}[t\cdot m+j-1]\right\| > \delta\right\}\middle|\mathbf{C}^\star_{(k)}\right)
\end{aligned}$$

We have attempted to separate the sequence into nearly iid terms, $(\mathbf{x}[t\cdot m+j-1])_{t=1,\ldots,T/m}$ for each $j = 1,\ldots,m$. We can bound the probability

using a Markov inequality:

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{T/m}\sum_{t=1}^{T/m}\mathbf{n}_i[t\cdot m+j]\tilde{\mathbf{x}}[t\cdot m+j-1]\right\|>\delta\right\}\middle|\mathbf{C}^{\star}_{(k)}\right)\\
&\leq\frac{1}{\delta^2}\cdot\mathbb{E}\left\|\frac{1}{T/m}\sum_{t=1}^{T/m}\mathbf{n}_i[t\cdot m+j]\tilde{\mathbf{x}}[t\cdot m+j-1]\right\|^2\\
&\leq\frac{1}{\delta^2}\cdot\mathbb{E}\left(\frac{1}{T/m}\right)^2\sum_{s,t=1}^{T/m}\mathbf{n}_i[t\cdot m+j]\mathbf{n}_i[s\cdot m+j]\left\langle\tilde{\mathbf{x}}[t\cdot m+j-1],\tilde{\mathbf{x}}[s\cdot m+j-1]\right\rangle
\end{aligned}$$

What we would like to do at this point, is eliminate all but the diagonal terms, $s=t$ due to the independence of $\mathbf{n}_i[t\cdot m+j]$ and $\mathbf{n}_i[s\cdot m+j]$ for any $s\neq t$ and all $j=1,\ldots,m$. We will observe dependence between $\tilde{\mathbf{x}}[t\cdot m+j-1]$ and $\mathbf{n}_i[s\cdot m+j]$ for all $s=1,\ldots,T/m$ that satisfy $s=t-1$. However, when this happens, $\mathbf{n}_i[t\cdot m+j]$ is independent of the rest of the terms, and $\mathbb{E}\mathbf{n}_i[t\cdot m+j]=0$. Thus, we can eliminate all cross terms. Moreover, the noise terms are independent of the observations by causality. This leads to the following conclusion:

$$\begin{aligned}
&\frac{1}{\delta^2}\cdot\mathbb{E}\left(\frac{1}{T/m}\right)^2\sum_{s,t=1}^{T/m}\mathbf{n}_i[t\cdot m+j]\mathbf{n}_i[s\cdot m+j]\left\langle\tilde{\mathbf{x}}[t\cdot m+j-1],\tilde{\mathbf{x}}[s\cdot m+j-1]\right\rangle\\
&\leq\frac{1}{\delta^2}\cdot\left(\frac{1}{T/m}\right)\cdot\left(\mathbb{E}\mathbf{n}_i^2\right)\cdot\mathbb{E}\left(\frac{1}{T/m}\sum_{t=1}^{T/m}\left\|\tilde{\mathbf{x}}[j\cdot m+t-1]\right\|^2\right)\\
&\leq\frac{1}{\delta^2}\cdot\left(\frac{1}{T/m}\right)\cdot\left(\mathbb{E}\mathbf{n}_i^2\right)\cdot\left(\mathbb{E}\left\|\tilde{\mathbf{x}}\right\|^2\right)\\
&\leq\frac{\nu^2\cdot\mathrm{tr}\left(\mathbf{R}_{(k)}\right)}{T/m\cdot\delta^2}.
\end{aligned}$$

(Global bound) Now, we marginalize over $\mathbf{C}^{\star}_{(k)}$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} > \delta\right\}\right) \\
&\quad = \int \mathbb{P}\left(\left\{\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{N}_{(k)}\right\|_{2,\infty} > \delta\right\} \middle| \mathbf{C}^{\star}_{(k)}\right) \mathbb{P}\left(\mathbf{C}^{\star}_{(k)}\right) \\
&\quad \leq \frac{\nu^2}{T/m \cdot \delta^2} \cdot \int \operatorname{tr}\left(\mathbf{R}_{(k)}\right) \mathbb{P}\left(\mathbf{C}^{\star}_{(k)}\right).
\end{aligned}$$

To complete the proof, we note that $\operatorname{tr}\left(\mathbf{R}_{(k)}\right) \leq d \cdot \left\|\mathbf{R}_{(k)}\right\| \leq d \cdot \left(\frac{\nu}{1-\sqrt{d}\cdot s\cdot C}\right)^2$ almost surely from Lemma 4.8. □

4.4.4 Restricted eigenvalue condition of the dictionary estimate

Lemma 4.12. *Let $\mathbf{D}^{\star}$ satisfy the s-restricted eigenvalue condition with $\kappa^{\star}_s > 4 \cdot \sqrt{\frac{s}{m\cdot d}} \cdot \varepsilon$ for some $\varepsilon > 0$. Then, for any $\mathbf{D}$ which satisfies*

$$d\left(\mathbf{D}, \mathbf{D}^{\star}\right) \leq \varepsilon, \tag{4.19}$$

$\mathbf{D}$ satisfies a s-restricted eigenvalue condition with $\kappa_s > 0$.

Proof. We will control the s-restricted eigenvalue condition of $\mathbf{D}$ by considering the magnitude of the perturbation from $\mathbf{D}^{\star}\mathbf{P}^{*}$, where $\mathbf{P}$ is the signed permutation matrix of the dictionary metric (eq. (2.43)). This is linear in ε. Let $\mathbf{h} \neq 0 \in \mathbb{R}^m$,

then

$$\begin{aligned}
\|\mathbf{Dh}\|_2 &= \|\mathbf{D}^\star\mathbf{P}^* - (\mathbf{D}^\star\mathbf{P}^* - \mathbf{DP})\,\mathbf{h}\|_2 \\
&\geq \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 - \|(\mathbf{D}^\star\mathbf{P}^* - \mathbf{D})\,\mathbf{h}\|_2 \\
&\geq \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 - \left\|\sum_{i=1}^{m} (\mathbf{P}^*\mathbf{h})_i\,(\mathbf{D}^\star - \mathbf{DP})_i\right\|_2 \\
&\geq \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 - \sum_{i=1}^{m} |(\mathbf{P}^*\mathbf{h})_i| \cdot \|(\mathbf{D}^\star - \mathbf{DP})_i\|_2 \\
&\geq \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 - \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 \cdot \|\mathbf{D}^\star - \mathbf{DP}\|_F \\
&\geq \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 - \varepsilon \cdot \|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2 .
\end{aligned}$$

Substituting this relationship into the definition of s-restricted eigenvalue condition gives

$$\begin{aligned}
&\min_{\substack{J\subset\{1,\ldots,r\}\\ |J|\leq s}} \min_{\substack{\mathbf{h}\neq 0\in\mathbb{R}^r\\ \|\mathbf{h}_{J^C}\|_1 \leq 3\|\mathbf{h}_J\|_1}} \frac{\|\mathbf{Dh}\|_2}{\sqrt{m\cdot d}\,\|\mathbf{h}_J\|_2} \\
&\geq \min_{\substack{J\subset\{1,\ldots,r\}\\ |J|\leq s}} \min_{\substack{\mathbf{h}\neq 0\in\mathbb{R}^r\\ \|\mathbf{h}_{J^C}\|_1 \leq 3\|\mathbf{h}_J\|_1}} \frac{\|\mathbf{D}^\star\mathbf{P}^*\mathbf{h}\|_2}{\sqrt{n}\,\|(\mathbf{P}^*\mathbf{h})_J\|_2} - \frac{\varepsilon\cdot\|(\mathbf{P}^*\mathbf{h})\|_2}{\sqrt{m\cdot d}\,\|(\mathbf{P}^*\mathbf{h})_J\|_2}.
\end{aligned}$$

Note that $\tilde{\mathbf{h}} = \mathbf{P}^*\mathbf{h}$ satisfies $\left\|\tilde{\mathbf{h}}_{J^C}\right\|_1 \leq 3\left\|\tilde{\mathbf{h}}_J\right\|_1$, and for such vectors, $\left\|\tilde{\mathbf{h}}\right\|_2 \leq \left\|\tilde{\mathbf{h}}\right\|_1 \leq 4\left\|\tilde{\mathbf{h}}_J\right\|_1 \leq 4\sqrt{s}\left\|\tilde{\mathbf{h}}_J\right\|_2$. Ultimately, this yields

$$\min_{\substack{J\subset\{1,\ldots,r\}\\ |J|\leq s}} \min_{\substack{\mathbf{h}\neq 0\in\mathbb{R}^r\\ \|\mathbf{h}_{J^C}\|_1 \leq 3\|\mathbf{h}_J\|_1}} \frac{\|\mathbf{Dh}\|_2}{\sqrt{m\cdot d}\,\|\mathbf{h}_J\|_2} \geq \kappa_s^\star - 4\cdot\sqrt{\frac{s}{m\cdot d}}\cdot\varepsilon.$$

This quantity is positive by the statement of the lemma. □

4.4.5 Concentration of $\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty}$

Lemma 4.13. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}_{(k)}^\star \right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)} \right)_{k=1,\dots,N}$ *be generated according to Model 2. Then, for any* $\delta > 0$ *and all* $k = 1, \dots, N$,

$$\begin{aligned} &\mathbb{P}\left(\left\{ \max_{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}} \frac{\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty}}{\|\mathbf{D}\|_F} > \delta \right\} \right) \\ &\quad \leq C_1 \cdot 5^{m \cdot d \cdot r} \cdot \frac{d \cdot s \cdot c^2}{r \cdot \delta^2} \cdot \left(\frac{\nu}{1 + \sqrt{d} \cdot s \cdot C} \right)^4, \end{aligned} \tag{4.20}$$

for some positive constant C_1.

Proof. We will make use of a covering argument. Let $\mathcal{K}_{1/2}$ be a 1/2-net covering of the unit $\|\cdot\|_F$-sphere in $\mathbb{R}^{(m \cdot d) \times r}$. By Lemma 2.8, we have

$$\max_{\mathbf{D} \in \mathbb{R}^{(m \cdot d) \times r}} \frac{\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty}}{\|\mathbf{D}\|_F} \leq 2 \cdot \max_{\mathbf{D} \in \mathcal{K}_{1/2}} \left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty},$$

and so we want to show that for all $\mathbf{D} \in \mathcal{K}_{1/2}$,

$$\left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty} \leq \delta/2.$$

Now, we want establish the probability of the complement of this event. As the columns of $\mathbf{C}_{(k)}^\star$ are chosen independently,

$$\begin{aligned} &\mathbb{P}\left(\left\{ \left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \mathbf{C}_{(k)}^\star \right\|_{2,\infty} > \delta/2 \right\} \right) \\ &= \sum_{i=1}^{d} \mathbb{P}\left(\left\{ \left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \left(\mathbf{C}_{(k)}^\star \right)_i \right\| > \delta/2 \right\} \right) \\ &= \sum_{i=1}^{d} \mathbb{P}\left(\left\{ \left\| \left(\frac{1}{T} \mathbf{X}_{(k)}^* \mathbf{X}_{(k)} \right) \mathbf{D} \left(\mathbf{C}_{(k)}^\star \right)_i \right\|^2 > \delta^2/4 \right\} \right). \end{aligned}$$

By Markov's inequality,

$$\mathbb{P}\left(\left\{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\|^{2} > \delta^{2}/4\right\}\right)$$
$$\leq \frac{4}{\delta^{2}} \cdot \mathbb{E}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\|^{2}.$$

We consider the term under expectation:

$$\begin{aligned}
&\mathbb{E}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\|^{2} \\
&= \mathbb{E}\left\langle\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}, \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\rangle \\
&= \mathbb{E}\left\langle\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}, \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)^{2}\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\rangle \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\langle\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}, \left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)^{2}\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\rangle\middle|\mathbf{C}_{(k)}^{\star}\right]\right].
\end{aligned}$$

Note that by the assumptions of Model 2 and Lemma 4.8, we have a bound independent of the realization $\mathbf{C}_{(k)}^{\star}$,

$$\left\|\mathbb{E}\left[\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)^{2}\middle|\mathbf{C}_{(k)}^{\star}\right]\right\| \leq C_{0} \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^{4}.$$

for some $\mathbf{C}_{0}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\right\|^{2} \\
&\leq C_{0} \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^{4} \cdot \left\langle\mathbf{D}\left(\mathbb{E}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}\left(\mathbf{C}_{(k)}^{\star}\right)_{i}^{*}\right), \mathbf{D}\right\rangle \\
&\leq C_{0} \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^{4} \cdot \frac{s\cdot c^{2}}{r}.
\end{aligned}$$

Putting this together, we have

$$\mathbb{P}\left(\left\{\left\|\left(\frac{1}{T}\mathbf{X}_{(k)}^{*}\mathbf{X}_{(k)}\right)\mathbf{D}\mathbf{C}_{(k)}^{\star}\right\|_{2,\infty} > \delta/2\right\}\right) \leq C_{1} \cdot \frac{d\cdot s\cdot c^{2}}{r\cdot\delta^{2}} \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^{4}$$

for some positive constant C_1. This provides the probability for a fixed $\mathbf{D} \in \mathcal{K}_{1/2}$. The total probability is given by a union bound. $\left|\mathbf{K}_{1/2}\right| = 5^{m \cdot d \cdot r}$ is given by Lemma 2.7. This completes the argument. □

4.4.6 Smallest eigenvalue of left-hand side operator

Lemma 4.14. *Let* $\mathbf{D}^\star$, $\left(\mathbf{C}^\star_{(k)}\right)_{k=1,\dots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\dots,N}$ *be generated according to Model 2. Assume that* $\left\|\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} - \mathbf{R}_{(k)}\right\| \leq \delta_1 \cdot \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2$. *Then, for any* $\delta_2 \in (0,1)$,

$$\begin{aligned}
&\mathbb{P}\left(\left\{\lambda_{\min}\left(\frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{d}\mathbf{P}\mathbf{C}^\star_{(k)}\left(\mathbf{P}\mathbf{C}^\star_{(k)}\right)^*\right)\otimes\left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)}\right)\right)\right.\\
&\qquad\left.\leq (1-\delta_2)\cdot\frac{s\cdot c^2}{r}\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2\right\}\Bigg) \qquad (4.21)\\
&\qquad\leq m\cdot d\cdot r\cdot\left[\frac{e^{-\delta_2}}{(1-\delta_2)^{1-\delta_2}}\right]^{N\cdot\left(s\cdot r\cdot\left[\left(\frac{1+\sqrt{d}\cdot s\cdot C}{1-\sqrt{d}\cdot s\cdot C}\right)^2+\delta_1\right]\right)^{-1}}.
\end{aligned}$$

Proof. We will apply Cor. 2.5.1. By Lemma 4.15,

$$\mu_{\min} = N\cdot\frac{s\cdot c^2}{r}\cdot\left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2.$$

Therefore, we need only calculate an absolute bound R,

$$\begin{aligned}
&\lambda_{\max}\left(\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\right)\\
&=\sup_{\substack{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}\\ \|\mathbf{D}\|_F=1}}\left\langle\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right),\mathbf{D}\right\rangle_F\\
&=\sup_{\substack{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}\\ \|\mathbf{D}\|_F=1}}\left\langle\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right),\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)\right\rangle_F\\
&\leq\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right\|\cdot\sup_{\substack{\mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}\\ \|\mathbf{D}\|_F=1}}\left\langle\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right),\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)\right\rangle_F\\
&\leq\left\|\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right\|\cdot\left\|\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right\|^2
\end{aligned}$$

By the assumption of the lemma and Model 2, we have

$$\begin{aligned}
&\lambda_{\max}\left(\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\right)\\
&\leq s^2\cdot C^2\cdot\left[\left(\frac{\nu}{1-\sqrt{d}\cdot s\cdot C}\right)^2+\delta_1\cdot\left(\frac{\nu}{1-\sqrt{d}\cdot s\cdot C}\right)^2\right].
\end{aligned}$$

The result follows from apply Cor. 2.5.1. □

Lemma 4.15. *Let* $\mathbf{D}^{\star}$, $\left(\mathbf{C}^{\star}_{(k)}\right)_{k=1,\ldots,N}$, *and* $\left(\mathbf{N}_{(k)}\right)_{k=1,\ldots,N}$ *be generated according to Model 2. Then,*

$$\lambda_{min}\left(\mathbb{E}\left[\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\right]\right)\geq\frac{s\cdot c^2\cdot\nu^2}{r\cdot\left(1+\sqrt{d}\cdot s\cdot C\right)^2}. \quad (4.22)$$

Proof. We first use iterated expectation to isolate the randomness resulting from the coefficients,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\right]\\
&=\mathbb{E}\left[\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\mathbb{E}\left(\frac{1}{T}\mathbf{X}^{*}_{(k)}\mathbf{X}_{(k)}\right)\Big|\mathbf{C}^{\star}_{(k)}\right]\\
&=\mathbb{E}\left[\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^{*}\right)\otimes\mathbf{R}_{(k)}\right].
\end{aligned}$$

Next, we use the variational definition of a minimum eigenvalue for linear operators on $\left(\mathbb{R}^{(m\cdot d)\times r}, \|\cdot\|_F\right)$,

$$\begin{aligned}
&\lambda_{min}\left(\mathbb{E}\left[\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^*\right)\otimes \mathbf{R}_{(k)}\right]\right)\\
&= \min_{\substack{\|\mathbf{D}\|_F=1\\ \mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}} \mathbb{E}\left\langle \mathbf{D}, \mathbf{R}_{(k)}\mathbf{D}\left(\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^*\right)\right\rangle_F\\
&= \min_{\substack{\|\mathbf{D}\|_F=1\\ \mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}} \mathbb{E}\left\langle \mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right), \mathbf{R}_{(k)}\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)\right\rangle_F,
\end{aligned}$$

where the second line follows from the properties of $\langle\cdot,\cdot\rangle_F$. Now, we can introduce a lower bound by using the uniform lower bound on the smallest eigenvalue of $\mathbf{R}_{(k)}$,

$$\begin{aligned}
&\min_{\substack{\|\mathbf{D}\|_F=1\\ \mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}} \mathbb{E}\left\langle \mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right), \mathbf{R}_{(k)}\mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)\right\rangle_F\\
&\quad\geq \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2 \cdot \min_{\substack{\|\mathbf{D}\|_F=1\\ \mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}} \mathbb{E}\left\langle \mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right), \mathbf{D}\left(\frac{1}{\sqrt{d}}\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)\right\rangle_F\\
&\quad= \left(\frac{\nu}{1+\sqrt{d}\cdot s\cdot C}\right)^2 \cdot \min_{\substack{\|\mathbf{D}\|_F=1\\ \mathbf{D}\in\mathbb{R}^{(m\cdot d)\times r}}} \left\langle \mathbf{D}, \mathbf{D}\left(\mathbb{E}\frac{1}{d}\mathbf{P}\mathbf{C}^{\star}_{(k)}\left(\mathbf{P}\mathbf{C}^{\star}_{(k)}\right)^*\right)\right\rangle_F
\end{aligned}$$

The second equality follows again from the properties of $\langle\cdot,\cdot\rangle_F$. Finally, we use Lemma 4.9 to complete the argument. □

4.5 Application to EEG data

In this section, we apply Alg. 1 to neuroimaging data. Subjects completed multiple trials of various cognitive tasks while being monitored with EEG. We learn a dictionary of autoregressive components from the recorded data for each individual. Then, we show that these autoregressive components carry discriminative information of the underlying cognitive task.

Each subject enrolled in the study reported to the laboratory every two weeks for four months. Each of the visits corresponds to one session, and subjects returned for eight sessions. While at the laboratory, the subject would perform three cognitive tasks: a dot-probe task (dot), a dynamic attention task (dyn), and a visual working memory task (vwm). Each task implicates a different network, emotion, attention, and memory respectively. The session comprised repeated trials of the dynamic attention task during a single session and only one trial of the dot-probe and visual working memory tasks. For our analysis, we excluded all subjects with fewer than 50 trials.

fMRI and EEG were simultaneously recorded using commercially available EEG hardware from Brain Products (GmbH, Germany). Trials varied from 5-12 minutes, and EEG recordings were digitally sampled at 640 Hz on a 61-channel headset (see Fig. 4.6). Brain Products (GmbH, Germany) software was used to remove the two major sources of MRI-related artifacts: the gradient artifacts and the cardioballistic artifacts. Both artifacts were removed via standard procedures: an average artifact subtraction method where a "template" EEG response to the onset of both the gradient and the heart beat (as measured with electrocardiogram) is subtracted from each gradient pulse or heart beat. After this standard procedure, any residual artifact was removed with standard procedures in Matlab (Mathworks, Inc.) as suggested by Bigdely-Shamlo, *et al.* [13] and included an independent component analysis and artifact subspace removal ($SD = 3$). Finally, each trial was de-meaned and bandpass filtered at 1-30 Hz using a sixth order Butterworth bandpass filter implemented on cascaded second order sections before downsampling

to 80 Hz. This processing was performed in SciPy using the signal processing toolbox [45].

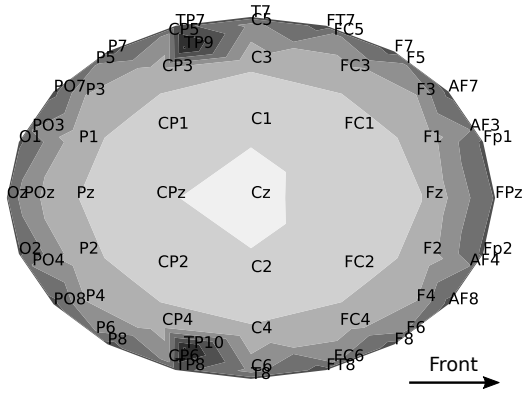


Figure 4.6: 61-channel electrode layout for EEG recording.

In order to implement Alg. 1 on significantly higher dimensional data ($d = 61$, $N \sim 50$), we made several modifications to the algorithm. The most significant obstacle was the computational burden of finding an overlapping clustering, for which the run time is dominated in the high-dimensional case by a $\mathcal{O}\left(d \log^2 d \cdot N^2\right)$ loop. In order to reduce this complexity to $\mathcal{O}\left(d \log^2 d \cdot N \log N\right)$, we further randomized the algorithm by sampling. The algorithm randomly samples pairs and looks for potential triples over all N elements. We instead look for potential triples over a random subset of $\log N$ elements. In addition, the convergence analysis of the algorithm depends on a very specific coefficient distribution given by Model 1. In the OverlappingCluster algorithm, we connected nodes if the inner product exceeded two standard deviations above the mean instead of a fixed $1/2$ as used for the convergence analysis in Arora, *et al.* [9]. Instead of prescribing the number of atoms, we implemented a pruning routine after OverlappingSVD that iteratively identified the

most coherent pair of atoms and removed one until the coherence of the dictionary was below 0.85.[3] When fitting the autoregressive model, we used an ordinary least squares estimator with a singular value (relative) threshold of 1×10^{-3} in order to provide greater numerical stability.

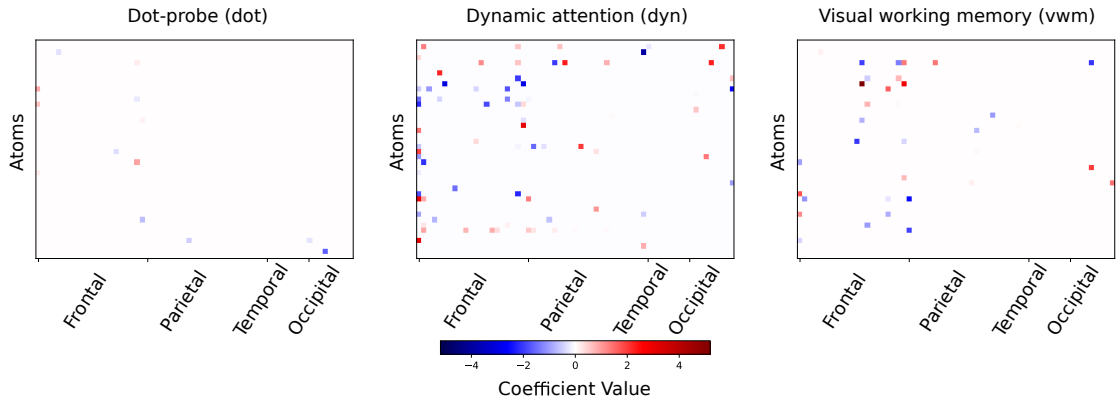


Figure 4.7: $\mathbf{W}^{\text{dot}}, \mathbf{W}^{\text{dyn}}, \mathbf{W}^{\text{vwm}}$ for an example subject. Note that the dynamic attention task has most of the coefficients concentrated in the frontal and parietal lobes, while the visual working memory task has most of the coefficients concentrated in the frontal and occipital lobes.

We implemented Alg. 1 on each subject individually with $m = 2$ and $s = 4$. These parameters were selected from a four-fold cross-validation that maximized explained variance. This provided us with a dictionary estimate $\mathbf{D}$ and coefficient estimate $\mathbf{C}_{(k)}$ for each such trial. Ultimately, we wanted to determine if the autoregressive atoms were discriminative. To test this, we fit a logistic regression model to each of the cognitive tasks, *i.e.* dot-probe, dynamic attention, and visual working

[3] Algorithm 1 does not explicitly enforce the desired sparsity s or desired number of dictionary atoms r.

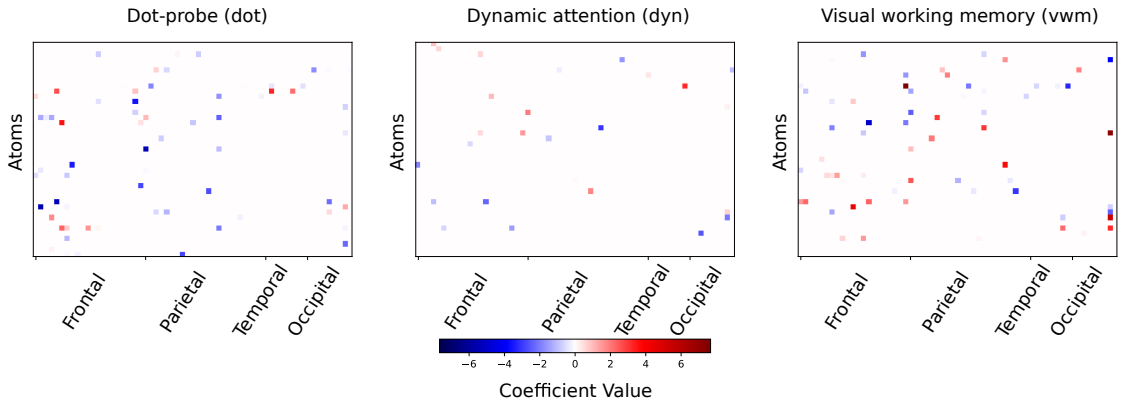


Figure 4.8: $\mathbf{W}^{\text{dot}}, \mathbf{W}^{\text{dyn}}, \mathbf{W}^{\text{vwm}}$ for another example subject. Again, the coefficients are heavily concentrated in the frontal and parietal lobes for the dynamic attention task and the frontal, parietal, and occipital lobes in the visual working memory task.

memory. For each trial, we modeled whether or not the task was *e.g.* dot-probe as a Bernoulli random variable $\beta_{(k)}^{\text{dot}}$:

$$\beta_{(k)}^{\text{dot}} | \mathbf{C}_{(k)} = \left(1 + \exp\left(-\left\langle \mathbf{W}^{\text{dot}}, \mathbf{C}_{(k)} \right\rangle_F + w_0^{\text{dot}}\right)\right)^{-1}, \tag{4.23}$$

where $\mathbf{W}^{\text{dot}} \in \mathbb{R}^{r \times d}$ and $w_0^{\text{dot}} \in \mathbb{R}$ are the parameters of the dot-probe model. We fit such a model using an ℓ^1 penalty for each cognitive task, *i.e.* dot-probe, dynamic attention, and visual working memory.[4] To implement the logistic regression, we used Scikit Learn [82]. We report our prediction accuracy on a 25% hold-out test set within each subject in Table 4.1. For two of the three subjects, we predicted the cognitive task better than chance for the multiclass problem. Due to a class imbalance from the repeated trials of the the dynamic attention task, results are reported in balanced accuracy.

[4]The penalty parameter was selected with a four-fold cross-validation which maximized balanced accuracy.

Subject	Observations (N)	Atoms (r)	Score
1	61	42	0.42
2	63	41	0.40
3	54	67	0.33

Table 4.1: Results of logistic regression in terms of balanced accuracy (chance: 1/3).

We can also use the coefficients of the logistic regression, *e.g.* $\mathbf{W}_{\text{dot}}$, to assess the discriminative information in the dictionary atoms. By visualizing the coefficients and their magnitude, we can observe whether the same atoms are predictive of all tasks, or if instead different atoms are important for predicting each cognitive task. Example results are given in Figs. 4.7 and 4.8. Here, we show the learned coefficients of Eq. (4.23) for two different subjects. The plots identify which relationships between past activity across the whole brain and current activity localized at a single channel are predictive of the cognitive task. We grouped electrodes into the four primary lobes: frontal, parietal, temporal, and occipital. The patterns reveal a significant difference in predictive components between the cognitive task conditions. Then, in Fig. 4.9, we visualize some example autoregressive atoms for a subject.

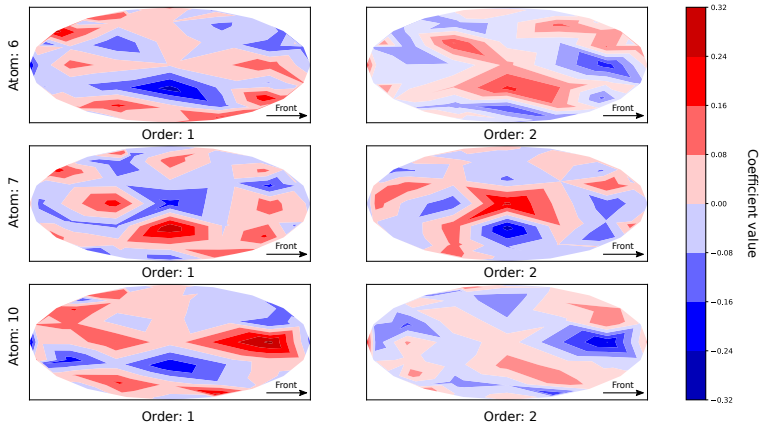


Figure 4.9: Representative atoms for each of the cognitive tasks for one subject. Front is to the right as in Fig. 4.6. Each row corresponds to an atom, representing essentially different network activity.

Chapter 5: Conclusion

We briefly conclude with some final thoughts and ideas for future work.

Is it possible to use the filters of Chapter 3 in a learning framework? Deep learning offers not only a powerful statistical model and computational graph, but also a learning algorithm in the form of stochastic gradient descent [68]. Considering an architecture like that of Φ in Sec. 3.4.2, is it possible to learn the appropriate holomorphic functions $(\phi_i)_{i\in\mathcal{I}}$? That is, can we minimize some functional $J : \bigoplus_{i\in\mathcal{I}} \mathcal{H}(U) \to \mathbb{R}$? Solving this variational problem requires a generalized derivative on $\mathcal{H}(U)$. We introduce first a definition and then state a minor claim.

Definition 32. Let $U \subset \mathcal{X}$ be an open subset of a locally convex space $\mathcal{X}$. The *Gâteaux derivative* of a function $f : U \to \mathbb{R}$ at an element $u \in U$ is

$$df(u; x) := \lim_{t\to 0} \frac{f(u + tx) - f(u)}{t}, \tag{5.1}$$

defined for all $x \in \mathcal{X}$.

Claim 5.1. *Let $U \subset \mathbb{C}$ be an open set. The set of holomorphic functions $f : U \to \mathbb{C}$ is a locally convex topological vector space when equipped with the family of seminorms*

$$\left(\sup_{z\in\mathcal{A}_i} |f(z)| \right)_{i\in\mathcal{I}}, \tag{5.2}$$

where $\{\mathcal{A}_i : i \in \mathcal{I}\}$ *is the collection of compact subsets of* U*.*

Claim 5.1, together with Def. 32, seems to offer a path towards a learning framework. We could perhaps update $(\phi_i)_{i\in\mathcal{I}}$ iteratively using a Gâteaux derivative of J.

Can we improve the results of Thm. 4.3? The finite sample factors and assumptions of the direct approach in Sec. 4.3 are unsatisfying. These results can be improved along two dual paths. First, the growth of observation length $T \sim \mathcal{O}(\varepsilon^{-3})$ could be improved by using a Bernstein-type concentration inequality to bound $\left\| \frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{N}_{(k)} \right\|_{2,\infty}$. Second, $\mathcal{A}_5$ could be eliminated from the proof of Thm. 4.3 if we could show the following:

$$\left\| \left(\frac{1}{T}\mathbf{X}^*_{(k)}\mathbf{X}_{(k)} \right) \left(\mathbf{D}^\star - \mathbf{D}^{(\ell-1)}\mathbf{P} \right) \left(\mathbf{C}_{(k)} \right)_i \right\| \leq \frac{\mu^{(\ell)}}{2}$$

for all $i = 1, \ldots, d$ and $\ell = 1, \ldots, n_\varepsilon$. This would eliminate condition (5) from both Lemma 4.5 and Lemma 4.6. It would be sufficient that

$$\frac{\mu^{(1)}}{\varepsilon^{(0)}} \geq \frac{4 \cdot \sqrt{s} \cdot C}{m \cdot d} \cdot \left[\left(\frac{\nu}{1 - \sqrt{d} \cdot s \cdot C} \right)^2 + \delta_1 \left(\frac{\nu}{1 + \sqrt{d} \cdot s \cdot C} \right)^2 \right]$$

for some $\delta_1 > 0$, a deterministic condition.

Component analysis of autoregressive processes? Chapter 4 introduces a linear mixture model of autoregressive processes. We focus on the problem of finding atomic components of autoregressive processes. Dictionary learning can be seen as an alternative to various other blind source separation models such as independent component analysis and principal component analysis. In this light, we can ask:

what would it mean to consider a principal component analysis of autoregressive processes? Let $\mathbf{n} \in \ell^2\left(\mathbb{Z}; \mathbb{C}^d\right)$ be a Gaussian process and $\mathbf{A} = \left(\mathbf{A}_{(k)}[t]\right)_{t \in \mathbb{Z}_+}$ be the causal matrix symbol of an autoregressive process. It seems that a reasonable path for defining the principal component $\mathbf{U} = (\mathbf{U}[t])_{t \in \mathbb{Z}_+}$ would be something of the form:

$$\mathbf{U}_1 = \arg\min_{\mathbf{U}} \mathbb{E} \frac{1}{2} \left\| \left[(\mathbf{I} - \mathbf{U})^{-1} - (\mathbf{I} - \mathbf{A})^{-1} \right] \mathbf{n} \right\| \text{ s.t. } \operatorname{rank}(\mathbf{U}) = 1. \tag{5.3}$$

Then, we could define subsequent components likewise,

$$\begin{aligned} \mathbf{U}_j = \arg\min_{\mathbf{U}} & \mathbb{E} \frac{1}{2} \left\| \left[(\mathbf{I} - \mathbf{U})^{-1} - (\mathbf{I} - \mathbf{A})^{-1} \right] \mathbf{n} \right\| \\ & \text{s.t. } \operatorname{rank}(\mathbf{U}) = 1 \text{ and} \\ & \left(\sum_{s \in \mathbb{Z}} \langle \mathbf{U}[s], \mathbf{U}_\ell[t-s] \rangle_F \right)_{t \in \mathbb{Z}} = \mathbf{0}, \ \ell = 1, \ldots, j-1. \end{aligned} \tag{5.4}$$

Although these are not rigorous definitions, it offers a path towards defining an "orthogonal" component analysis of autoregressive processes.

Further applications in neuroimaging. Section 4.5 showed promising results for using Alg. 1 to discern autoregressive components in EEG data. These initial results offer proof-of-concept for the method. Future work will entail collaboration with neuroscientists to validate the results and apply the method toward understanding the functional integration of segregated regions of the brain.

Bibliography

[1] A. Agarwal, A. Anandkumar, and P. Netrapalli. "A Clustering Approach to Learning Sparsely Used Overcomplete Dictionaries". In: *IEEE Transactions on Information Theory* 63.1 (Jan. 2017), pp. 575–592.

[2] A. Agarwal et al. "Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization". In: *SIAM Journal on Optimization* 26.4 (Jan. 1, 2016), pp. 2775–2799.

[3] Alireza Aghasi et al. "A convex program for bilinear inversion of sparse vectors". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 8557–8567.

[4] M. Aharon, M. Elad, and A. Bruckstein. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation". In: *IEEE Transactions on Signal Processing* 54.11 (Nov. 2006), pp. 4311–4322.

[5] Michal Aharon, Michael Elad, and Alfred M. Bruckstein. "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them". In: *Linear Algebra and its Applications*. Special Issue devoted to the Haifa 2005 conference on matrix theory 416.1 (July 1, 2006), pp. 48–67.

[6] R. Ahlswede and A. Winter. "Strong converse for identification via quantum channels". In: *IEEE Transactions on Information Theory* 48.3 (Mar. 2002), pp. 569–579.

[7] A. Ahmed, B. Recht, and J. Romberg. "Blind Deconvolution Using Convex Programming". In: *IEEE Transactions on Information Theory* 60.3 (Mar. 2014), pp. 1711–1732.

[8] Zeynep Akata, Christian Thurau, and Christian Bauckhage. "Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction". In: 16th Computer Vision Winter Workshop. Feb. 2, 2011.

[9] Sanjeev Arora, Rong Ge, and Ankur Moitra. "New Algorithms for Learning Incoherent and Overcomplete Dictionaries". In: *Conference on Learning Theory*. Conference on Learning Theory. May 29, 2014, pp. 779–806.

[10] Sumanta Basu and George Michailidis. "Regularized estimation in sparse high-dimensional time series models". In: *The Annals of Statistics* 43.4 (Aug. 2015), pp. 1535–1567.

[11] Mikhail Belkin and Partha Niyogi. “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation”. In: *Neural Computation* 15.6 (June 1, 2003), pp. 1373–1396.

[12] Peter J. Bickel, Ya’acov Ritov, and Alexandre B. Tsybakov. “Simultaneous analysis of Lasso and Dantzig selector”. In: *The Annals of Statistics* 37.4 (Aug. 2009), pp. 1705–1732.

[13] Nima Bigdely-Shamlo et al. “The PREP pipeline: standardized preprocessing for large-scale EEG analysis”. In: *Frontiers in Neuroinformatics* 9 (2015).

[14] Vladimir Igorevich Bogachev. *Gaussian Measures*. Vol. 62. Mathematical Surveys and Monographs. American Mathematical Soc., 1998. 449 pp.

[15] A. W. Bohannon, B. M. Sadler, and R. V. Balan. “LEARNING FLEXIBLE REPRESENTATIONS OF STOCHASTIC PROCESSES ON GRAPHS”. In: *2018 IEEE Data Science Workshop (DSW)*. 2018 IEEE Data Science Workshop (DSW). June 2018, pp. 61–65.

[16] A. W. Bohannon et al. “Collaborative image triage with humans and computer vision”. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Oct. 2016, pp. 004046–004051.

[17] Addison W. Bohannon. “Estimating Functional Connectivity from fMRI Data Using a Frequency-Sparse Multivariate Autoregressive Model”. Poster. 2018 Graph Signal Processing Workshop. Laussanne, Switzerland, 2018.

[18] Addison W. Bohannon, Sean M. Fitzhugh, and Arwen H. DeCostanza. “A framework for enhancing human-agent teamwork through adaptive individualized technologies”. In: Proceedings of the SPIE Defense and Commercial Sensing Conference on AI/ML for Multidomain Operations. (accepted). SPIE, 2019.

[19] Addison W. Bohannon, Brian M. Sadler, and Radu V. Balan. “A Filtering Framework for Time-Varying Graph Signals”. In: *Vertex-Frequency Analysis of Graph Signals*. Ed. by Ljubiša Stanković and Ervin Sejdić. Signals and Communication Technology. Cham: Springer International Publishing, 2019, pp. 341–376.

[20] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. “Proximal alternating linearized minimization for nonconvex and nonsmooth problems”. In: *Mathematical Programming* 146.1 (Aug. 1, 2014), pp. 459–494.

[21] Albrecht Böttcher and Bernd Silbermann. *Analysis of Toeplitz Operators*. 2nd. Monographs in Mathematics. Springer, 2006. 511 pp.

[22] M. M. Bronstein et al. “Geometric Deep Learning: Going beyond Euclidean data”. In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 18–42.

[23] James Ward Brown and Ruel V. Churchill. *Complex Variables and Applications*. 7th. McGraw-Hill, 2004.

[24] J. Bruna and S. Mallat. "Invariant Scattering Convolution Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (Aug. 2013), pp. 1872–1886.

[25] Joan Bruna et al. "Spectral Networks and Locally Connected Networks on Graphs". In: *Proceedings of the International Conference on Learning Representations 2014*. Banff, Canada, Apr. 16, 2014. arXiv: 1312.6203.

[26] Ed Bullmore and Olaf Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". In: *Nature Reviews Neuroscience* 10.3 (Mar. 2009), pp. 186–198.

[27] C. Shawn Burke et al. "Understanding team adaptation: A conceptual analysis and model". In: *Journal of Applied Psychology* 91.6 (2006), pp. 1189–1207.

[28] Vince D. Calhoun, Jingyu Liu, and Tülay Adalı. "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data". In: *NeuroImage*. Mathematics in Brain Imaging 45.1 (Mar. 1, 2009), S163–S172.

[29] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223.

[30] George Casella and Roger Berger. *Statistical inference*. Thomson Learning, June 18, 2002.

[31] S. Chen, D. Donoho, and M. Saunders. "Atomic Decomposition by Basis Pursuit". In: *SIAM Review* 43.1 (Jan. 1, 2001), pp. 129–159.

[32] S. Chen et al. "Discrete Signal Processing on Graphs: Sampling Theory". In: *IEEE Transactions on Signal Processing* 63.24 (Dec. 2015), pp. 6510–6523.

[33] Fan R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics 92. American Mathematical Soc., 1997. 228 pp.

[34] Matthew A. Cronin, Laurie R. Weingart, and Gergana Todorova. "Dynamics in Groups: Are We There Yet?" In: *The Academy of Management Annals* 5.1 (June 1, 2011), pp. 571–612.

[35] E. Brian Davies. *Linear Operators and their Spectra*. Cambridge Studies in Advanced Mathematics 106. Cambridge University Press, Apr. 26, 2007. 436 pp.

[36] Richard A. Davis, Pengfei Zang, and Tian Zheng. "Sparse Vector Autoregressive Modeling". In: *Journal of Computational and Graphical Statistics* 25.4 (Oct. 1, 2016), pp. 1077–1096.

[37] Arwen H. DeCostanza et al. *Enhancing HumanAgent Teaming with Individualized, Adaptive Technologies: A Discussion of Critical Scientific Questions*. ARL-TR-8359. US Army Research Laboratory Aberdeen Proving Ground United States, May 4, 2018.

[38] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 3844–3852.

[39] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[40] D. L. Donoho, M. Elad, and V. N. Temlyakov. "Stable recovery of sparse overcomplete representations in the presence of noise". In: *IEEE Transactions on Information Theory* 52.1 (Jan. 2006), pp. 6–18.

[41] Nelson Dunford and Jacob T. Schwartz. *Linear operators part I: general theory*. Vol. VII. Pure and Applied Mathematics. Interscience Publishers, 1958.

[42] Francis T. Durso. *Handbook of Applied Cognition*. John Wiley & Sons, Feb. 6, 2007. 918 pp.

[43] David K Duvenaud et al. "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 2224–2232.

[44] K. Engan, S. O. Aase, and J. Hakon Husoy. "Method of optimal directions for frame design". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258). Vol. 5. Mar. 1999, 2443–2446 vol.5.

[45] Eric Jones, Travis Oliphant, and Pearu Peterson. *SciPy: Open source scientific tools for Python*. 2001. URL: `http://www.scipy.org/`.

[46] Karl J. Friston. "Functional and Effective Connectivity: A Review". In: *Brain Connectivity* 1.1 (Jan. 1, 2011), pp. 13–36.

[47] Fernando Gama, Alejandro Ribeiro, and Joan Bruna. "Diffusion Scattering Transforms on Graphs". In: *arXiv:1806.08829 [cs, stat]* (June 22, 2018). arXiv: `1806.08829`.

[48] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. JHU Press, Dec. 27, 2012. 738 pp.

[49] F. Grassi et al. "A Time-Vertex Signal Processing Framework: Scalable Processing and Meaningful Representations for Time-Series on Graphs". In: *IEEE Transactions on Signal Processing* 66.3 (Feb. 2018), pp. 817–829.

[50] R. Gribonval, R. Jenatton, and F. Bach. "Sparse and Spurious: Dictionary Learning With Noise and Outliers". In: *IEEE Transactions on Information Theory* 61.11 (Nov. 2015), pp. 6298–6319.

[51] R. Gribonval and K. Schnass. "Dictionary Identification—Sparse Matrix-Factorization viaℓ_1-Minimization". In: *IEEE Transactions on Information Theory* 56.7 (July 2010), pp. 3523–3539.

[52] Geoffrey Grimmett et al. *Probability and Random Processes*. Oxford University Press, May 31, 2001. 612 pp.

[53] Fang Han and Han Liu. "Transition Matrix Estimation in High Dimensional Time Series". In: *International Conference on Machine Learning*. International Conference on Machine Learning. Feb. 13, 2013, pp. 172–180.

[54] Fang Han, Huanran Lu, and Han Liu. "A Direct Estimation of High Dimensional Stationary Vector Autoregressions". In: *arXiv:1307.0293 [stat]* (July 1, 2013). arXiv: 1307.0293.

[55] W. Huang et al. "Graph Signal Processing of Human Brain Imaging Data". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Apr. 2018, pp. 980–984.

[56] John D. Hunter. "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3 (May 1, 2007), pp. 90–95.

[57] Daniel R. Ilgen et al. "Teams in Organizations: From Input-Process-Output Models to IMOI Models". In: *Annual Review of Psychology* 56.1 (2005), pp. 517–543.

[58] E. Isufi et al. "Autoregressive Moving Average Graph Filtering". In: *IEEE Transactions on Signal Processing* 65.2 (Jan. 2017), pp. 274–288.

[59] Tosio Kato. *Perturbation theory for linear operators*. 2nd. Springer, 1995. 61 pp.

[60] Nancy Katz et al. "Network Theory and Small Groups". In: *Small Group Research* 35.3 (June 1, 2004), pp. 307–332.

[61] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *arXiv:1609.02907 [cs, stat]* (Sept. 9, 2016). arXiv: 1609.02907.

[62] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *Proceedings of the International Conference on Learning Representations 2017*. Toulon, France, Apr. 23, 2017. arXiv: 1609.02907.

[63] Konrad Knopp. *Theory of Functions*. Dover Publications, Inc., 1996. 340 pp.

[64] Anders Bredahl Kock and Laurent Callot. "Oracle inequalities for high dimensional vector autoregressions". In: *Journal of Econometrics*. High Dimensional Problems in Econometrics 186.2 (June 1, 2015), pp. 325–344.

[65] Steve W. J. Kozlowski et al. "Teams, teamwork, and team effectiveness: Implications for human systems integration". In: *APA handbook of human systems integration*. APA handbooks in psychology. Washington, DC, US: American Psychological Association, 2015, pp. 555–571.

[66] Abhishek Kumar, Piyush Rai, and Hal Daume. "Co-regularized Multi-view Spectral Clustering". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., 2011, pp. 1413–1421.

[67] B. J. Lance et al. "Brain–Computer Interface Technologies in the Coming Decades". In: *Proceedings of the IEEE* 100 (Special Centennial Issue May 2012), pp. 1585–1599.

[68] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (May 2015), pp. 436–444.

[69] Y. Li, K. Lee, and Y. Bresler. "Identifiability in Blind Deconvolution With Subspace or Sparsity Constraints". In: *IEEE Transactions on Information Theory* 62.7 (July 2016), pp. 4266–4275.

[70] Shuyang Ling and Thomas Strohmer. "Self-calibration and biconvex compressive sensing". In: *Inverse Problems* 31.11 (2015), p. 115002.

[71] J. Liu et al. "Multi-View Clustering via Joint Nonnegative Matrix Factorization". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. 0 vols. Proceedings. Society for Industrial and Applied Mathematics, May 2, 2013, pp. 252–260.

[72] A. Loukas and D. Foucard. "Frequency analysis of time-varying graph signals". In: *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Dec. 2016, pp. 346–350.

[73] Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Dec. 6, 2005. 765 pp.

[74] Julien Mairal et al. "Online Dictionary Learning for Sparse Coding". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. event-place: Montreal, Quebec, Canada. New York, NY, USA: ACM, 2009, pp. 689–696.

[75] Stéphane Mallat. "Group Invariant Scattering". In: *Communications on Pure and Applied Mathematics* 65.10 (Oct. 1, 2012), pp. 1331–1398.

[76] Stéphane Mallat. "Understanding deep convolutional networks". In: *Phil. Trans. R. Soc. A* 374.2065 (Apr. 13, 2016), p. 20150203.

[77] Michelle A. Marks, John E. Mathieu, and Stephen J. Zaccaro. "A Temporally Based Framework and Taxonomy of Team Processes". In: *The Academy of Management Review* 26.3 (2001), pp. 356–376.

[78] John D. Medaglia et al. "Functional Alignment with Anatomical Networks is Associated with Cognitive Flexibility". In: *Nature human behaviour* 2.2 (2018), pp. 156–164.

[79] Roberto Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *Electronic Communications in Probability* 15 (2010), pp. 203–212.

[80] Bruno A. Olshausen and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* 37.23 (Dec. 1, 1997), pp. 3311–3325.

[81] Antonio Ortega et al. "Graph Signal Processing". In: *arXiv:1712.00468 [eess]* (Dec. 1, 2017). arXiv: 1712.00468.

[82] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (Oct 2011), pp. 2825–2830.

[83] Maurice B. Priestley. *Spectral analysis and time series*. Academic Press, 1981.

[84] Charles R. Qi et al. "PointNet : Deep Learning on Point Sets for 3D Classification and Segmentation". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 652–660.

[85] Mark Rudelson and Roman Vershynin. "Hanson-Wright inequality and subgaussian concentration". In: *Electronic Communications in Probability* 18 (2013).

[86] Walter Rudin. *Functional Analysis*. 2nd. International Series in Pure and Applied Mathematics. McGraw-Hill, Jan. 1991. 456 pp.

[87] Eduardo Salas, Gerald F. Goodwin, and C. Shawn Burke. *Team Effectiveness In Complex Organizations: Cross-Disciplinary Perspectives and Approaches*. Routledge, Nov. 20, 2008. 624 pp.

[88] Eduardo Salas, Dana E. Sims, and C. Shawn Burke. "Is there a "Big Five" in Teamwork?" In: *Small Group Research* 36.5 (Oct. 1, 2005), pp. 555–599.

[89] Eduardo Salas et al. "The wisdom of collectives in organizations: An update of the teamwork competencies". In: *Team Effectiveness In Complex Organizations: Cross-Disciplinary Perspectives and Approaches*. Routledge, Nov. 20, 2008, pp. 39–79.

[90] A. Sandryhaila and J. M. F. Moura. "Discrete Signal Processing on Graphs". In: *IEEE Transactions on Signal Processing* 61.7 (Apr. 2013), pp. 1644–1656.

[91] A. Sandryhaila and J. M. F. Moura. "Big Data Analysis with Signal Processing on Graphs: Representation and processing of massive data sets with irregular structure". In: *IEEE Signal Processing Magazine* 31.5 (Sept. 2014), pp. 80–90.

[92] A. Sandryhaila and J. M. F. Moura. "Discrete Signal Processing on Graphs: Frequency Analysis". In: *IEEE Transactions on Signal Processing* 62.12 (June 2014), pp. 3042–3054.

[93] Aswin C. Sankaranarayanan et al. "Compressive Acquisition of Dynamic Scenes". In: *Computer Vision – ECCV 2010*. Ed. by Kostas Daniilidis, Petros Maragos, and Nikos Paragios. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 129–142.

[94] A. Sankaranarayanan et al. "Compressive Acquisition of Linear Dynamical Systems". In: *SIAM Journal on Imaging Sciences* 6.4 (Jan. 1, 2013), pp. 2109–2133.

[95] S. Saproo et al. “Cortically Coupled Computing: A New Paradigm for Synergistic Human-Machine Interaction”. In: *Computer* 49.9 (Sept. 2016), pp. 60–68.

[96] S. Segarra, A. G. Marques, and A. Ribeiro. “Optimal Graph-Filter Design and Applications to Distributed Linear Network Operators”. In: *IEEE Transactions on Signal Processing* 65.15 (Aug. 2017), pp. 4117–4131.

[97] Harry Sevi, Gabriel Rilling, and Pierre Borgnat. “Harmonic analysis on directed graphs and applications: from Fourier analysis to wavelets”. In: *arXiv:1811.11636 [math, stat]* (Nov. 28, 2018). arXiv: 1811.11636.

[98] Jianbo Shi and J. Malik. “Normalized cuts and image segmentation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (Aug. 2000), pp. 888–905.

[99] D. I. Shuman et al. “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains”. In: *IEEE Signal Processing Magazine* 30.3 (May 2013), pp. 83–98.

[100] Barry Simon. *Operator Theory*. A comprehensive course in analysis 4. American Mathematical Soc., Dec. 4, 2015. 769 pp.

[101] Barry Simon. *Real Analysis*. A comprehensive course in analysis 1. American Mathematical Soc., Nov. 2, 2015. 811 pp.

[102] Song Song and Peter J. Bickel. “Large Vector Auto Regressions”. In: *arXiv:1106.3915 [q-fin, stat]* (June 20, 2011). arXiv: 1106.3915.

[103] Daniel Spielman. “Spectral Graph Theory”. In: *Combinatorial Scientific Computing*. Ed. by Uwe Naumann and Olaf Schenk. Jan. 25, 2012.

[104] Daniel A. Spielman, Huan Wang, and John Wright. “Exact Recovery of Sparsely-Used Dictionaries”. In: *Conference on Learning Theory*. Conference on Learning Theory. June 16, 2012, pp. 37.1–37.18.

[105] J. Sun, Q. Qu, and J. Wright. “Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust-Region Method”. In: *IEEE Transactions on Information Theory* 63.2 (Feb. 2017), pp. 885–914.

[106] Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, June 1, 1997. 356 pp.

[107] Joel A. Tropp. “Just relax: convex programming methods for identifying sparse signals in noise”. In: *IEEE Transactions on Information Theory* 52.3 (Mar. 2006), pp. 1030–1051.

[108] Joel A. Tropp. “User-Friendly Tail Bounds for Sums of Random Matrices”. In: *Foundations of Computational Mathematics* 12.4 (Aug. 1, 2012), pp. 389–434.

[109] M. Tsitsvero, S. Barbarossa, and P. Di Lorenzo. “Signals on Graphs: Uncertainty Principle and Sampling”. In: *IEEE Transactions on Signal Processing* 64.18 (Sept. 2016), pp. 4845–4860.

[110] Pedro A. Valdés-Sosa et al. "Estimating Brain Functional Connectivity with Sparse Multivariate Autoregression". In: *Philosophical Transactions: Biological Sciences* 360.1457 (2005), pp. 969–981.

[111] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed Sensing, Theory and Applications*. Ed. by Gitta Kutyniok and Yonina C. Eldar. Cambridge University Press, 2012. arXiv: 1011.3027.

[112] Nicholas R. Waytowich et al. "Spectral Transfer Learning Using Information Geometry for a User-Independent Brain-Computer Interface". In: *Frontiers in Neuroscience* 10 (2016).

[113] Q. Zhang and B. Li. "Discriminative K-SVD for dictionary learning in face recognition". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. June 2010, pp. 2691–2698.

[114] Dongmian Zou and Gilad Lerman. "Graph Convolutional Neural Networks via Scattering". In: *arXiv:1804.00099 [cs, eess, math]* (Mar. 30, 2018). arXiv: 1804.00099.