UNIVERSITY OF CALIFORNIA SAN DIEGO

Networked Dynamical Systems: Privacy, Control, and Cognition

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Engineering Sciences (Mechanical Engineering) and Cognitive Science

by

Erfan Nozari

Committee in charge:

Professor Jorge Cortés, Chair
Professor Miroslav Krstić
Professor Sonia Martinez
Professor Eran Mukamel
Professor Nuno Vasconcelos

2019

The dissertation of Erfan Nozari is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California San Diego

2019

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation is the outcome of five years of continuous work and progress, five years of growth – academic, and personal. This growth was more profound than I could have ever imagined, and could only occur with a tremendous amount of support; one that I received. I cannot be more humble and grateful for it.

My sincere thanks and appreciation go to my Ph.D. advisor, Prof. Jorge Cortés, who was the leader of my unforgettable journey during these years. When applying for graduate school, I thought, like most, that I know how important the role of one's advisor is. It was far more momentous. And I learned, year after year, how fortunate I was. His guidance, dedication, supportiveness, insightfulness, courage, and, last but not the least, friendliness and cheerfulness elegantly shaped my experience. He raised me from a fresh undergraduate to one who will raise others. In particular, I would like to thank him for allowing me to bend course towards my passion almost in the middle of my Ph.D. It was a highly uncommon and insightful decision that will have an ever-lasting effect on my life.

I would like to further thank my dissertation committee, Prof. Miroslav Krstić, Prof. Sonia Martinez, Prof. Eran Mukamel, and Prof. Nuno Vasconcelos for taking out time and providing me with suggestions and criticisms.

My special thanks go to Prof. Gary Cottrell, the director, and to the members of the UCSD Cognitive Science Interdisciplinary Ph.D. Program (IDP). The existence and the incredibly open doors of this program allowed me to bridge a telescopic gap and integrate an old and personal passion of mine, that for mind and cognition, into my research on network control theory. I am also highly indebted to the UCSD Graduate Council for their unprecedented approval of this interdisci-

plinary degree and my participation in IDP.

At UCSD, I was fortunate to learn from many of the best teachers I have ever had. I would like to express my gratitude to them all, Professors Jorge Cortés, Sonia Martinez, Miroslav Kirstic, Eran Mukamel, Terry Sejnowsky, Douglas Nitz, Robert Bitmead, Raymond de Callafon, Mauricio de Oliveira, William Mceneaney, Bradley Voytek, Lara Rangel, Angela Yu, Cory Miller, John Serences, David Kleinfeld, and Adam Aron. What I learned from them has laid the foundation of this dissertation and my research in the years to come.

I was also fortunate to enjoy the collaboration of wonderful researchers throughout my Ph.D. My special thanks go to my collaborators and coauthors Prof. Pavankumar Tallapragada, Prof. Fabio Pasqualetti, and Dr. Yingbo Zhao. I am also greatly thankful to Dr. Erik Peterson, Dr. John Iversen, and Eric Leonardis for the many valuable discussions that we had and inspirations that they gave me.

I am further particularly grateful to my friend Ashish Cherukuri and my friend, mentor, and brother-in-law, Afshin Nikzad. Throughout these years, they have been that guide walking a few steps ahead and advising me whenever I doubted my decisions. They have done that so wisely that I have not regretted any steps I have taken following their advice, and I am indebted to them for all their help and support.

Looking back at my life, I cannot help but see the roots of my each and every accomplishment in the fabulous atmosphere in which I grew up, and I am deeply indebted for that to my parents and my sister, Azadeh. From my first breath up until now, they have sacrificed so much to provide me with more. They taught me, without me even realizing, the paramount importance of education, of fulfilling my potentials and responding to my inner calls, of valuing and finding joy in what I have, of trusting myself and standing on my own feet, and of having balance in life. These were the

pillars and foundations of my life and work throughout my Ph.D and I owe my deepest gratitude to them. Special thanks also goes to my parents-in-law. Their profound appreciation of education and intellectual growth was the perfect reinforcement that my wife and I needed to embrace the challenges throughout these years and see the ultimate opportunities behind. I truly appreciate their continuous support.

At last comes the greatest. My heartfelt appreciation goes to my wife, Parisa, who carried no less than half the burden of this Ph.D. The woman who never thought twice before prioritizing my accomplishments over hers; who could not demand less and support more. Throughout these years, her incredible strength and courage overturned my hesitations into confidence and the peace and positivity of her heart dissolved my otherwise debilitating fears and frustrations. I cannot be more grateful for having her on my side during each and every step and excited about the journey that we have ahead in the years to come.

The research reported in this dissertation was supported in part by the National Science Foundation Awards CNS-1329619, CNS-1446891, FA9550-15-1-0108, CMMI-1826065.

Chapter 3 is taken, in part, from the work published as “Differentially private average consensus: obstructions, trade-offs, and optimal algorithm design" by E. Nozari, P. Tallapragada, and J. Cortés in *Automatica*, vol. 81, pp. 221–231, 2017. The dissertation author was the primary investigator and author of this paper.

Chapter 4 is taken, in part, from the work published as “Differentially private distributed convex optimization via functional perturbation" by E. Nozari, P. Tallapragada, and J. Cortés in *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 5 is taken, in part, from the work which has been submitted for publication as

“Event-triggered stabilization of nonlinear systems with time-varying sensing and actuation delay" by E. Nozari, P. Tallapragada, and J. Cortés in *Automatica*. The dissertation author was the primary investigator and author of this paper.

Chapter 6 is taken, in part, from the work which is to appear as “Heterogeneity of central nodes explains the benefits of time-varying control scheduling in complex dynamical networks" by E. Nozari, F. Pasqualetti, and J. Cortés in *Journal of Complex Networks*. The dissertation author was the primary investigator and author of this paper.

Chapter 7 is taken, in part, from the work published as “Network identification with latent nodes via auto-regressive models" by E. Nozari, Y. Zhao, and J. Cortés in *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 722–736, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter 8 is taken, in part, from the work which has been submitted for publication as “Hierarchical selective recruitment in linear-threshold brain networks. Part I: Intra-layer dynamics and selective inhibition" by E. Nozari and J. Cortés in *IEEE Transactions on Automatic Control*, as well as the work which has been submitted for publication as “Hierarchical selective recruitment in linear-threshold brain networks. Part II: Inter-layer dynamics and top-down recruitment" by E. Nozari and J. Cortés in *IEEE Transactions on Automatic Control*. The dissertation author was the primary investigator and author of these papers.

Chapter 9 is taken, in part, from the work which is to appear as “Oscillations and coupling in interconnections of two-dimensional brain networks" by E. Nozari and J. Cortés in *American Control Conference*, Philadelphia, PA, July 2019. The dissertation author was the primary investigator and author of this paper.

VITA

2013	Bachelor of Science in Electrical Engineering – Control, Isfahan University of Technology
2015	Master of Science in Engineering Sciences (Mechancial Engineering), University of California San Diego
2019	Doctor of Philosophy in Engineering Science (Mechanical Engineering) and Cognitive Science, University of California San Diego

PUBLICATIONS

Journal publications:

[1] E. Nozari and J. Cortés, "Hierarchical selective recruitment in linear-threshold brain networks. Part I: Intra-layer dynamics and selective inhibition," *IEEE Transactions on Automatic Control*, 2018, submitted.

[2] E. Nozari and J. Cortés, "Hierarchical selective recruitment in linear-threshold brain networks. Part II: Inter-layer dynamics and top-down recruitment," *IEEE Transactions on Automatic Control*, 2018, submitted.

[3] E. Nozari, P. Tallapragada, and J. Cortés, "Event-triggered stabilization of nonlinear systems with time-varying sensing and actuation delay," *Automatica*, 2018, submitted.

[4] E. Nozari, F. Pasqualetti, and J. Cortés, "Heterogeneity of central nodes explains the benefits of time-varying control scheduling in complex dynamical networks," *Journal of Complex Networks*, 2019, to appear.

[5] E. Nozari, Y. Zhao, and J. Cortés, "Network identification with latent nodes via auto-regressive models," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 722–736, 2018.

[6] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private distributed convex optimization via functional perturbation," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, 2018.

[7] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private average consensus: obstructions, trade-offs, and optimal algorithm design," *Automatica*, vol. 81, pp. 221–231, 2017.

Conference proceedings:

[8] P. V. Chanekar, E. Nozari, and J. Cortés, "Network modification using a novel Gramian-based edge centrality," in *IEEE Conf. on Decision and Control*, Nice, France, Dec. 2019, submitted.

[9] E. Nozari and J. Cortés, "Oscillations and coupling in interconnections of two-dimensional brain networks," in *American Control Conference*, Philadelphia, PA, July 2019, to appear.

[10] E. Nozari and J. Cortés, "Selective recruitment in hierarchical complex dynamical networks with linear-threshold rate dynamics," in *IEEE Conf. on Decision and Control*, Miami Beach, FL, Dec. 2018, pp. 5227–5232.

[11] E. Nozari and J. Cortés, "Stability analysis of complex networks with linear-threshold rate dynamics," in *American Control Conference*, Milwaukee, WI, May 2018, pp. 191–196.

[12] E. Nozari, F. Pasqualetti, and J. Cortés, "Time-invariant versus time-varying actuator scheduling in complex networks," in *American Control Conference*, Seattle, WA, May 2017, pp. 4995–5000.

[13] E. Nozari, P. Tallapragada, and J. Cortés, "Event-triggered control for nonlinear systems with time-varying input delay," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016, pp. 495–500.

[14] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private distributed convex optimization via objective perturbation," in *American Control Conference*, Boston, MA, July 2016, pp. 2061–2066.

[15] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private average consensus with optimal noise selection," *IFAC-PapersOnLine*, vol. 48, no. 22, pp. 203–208, 2015, *IFAC Workshop on Distributed Estimation and Control in Networked Systems*, Philadelphia, PA.

ABSTRACT OF THE DISSERTATION

Networked Dynamical Systems: Privacy, Control, and Cognition

by

Erfan Nozari

Doctor of Philosophy in Engineering Sciences (Mechanical Engineering) and Cognitive Science

University of California San Diego, 2019

Professor Jorge Cortés, Chair

Many natural and man-made systems, ranging from the nervous system to power and transportation grids to societies, exhibit dynamic behaviors that evolve over a sparse and complex network. This networked aspect raises significant challenges and opportunities for the identification, analysis, and control of such dynamic behaviors. While some of these challenges emanate from the networked aspect *per se* (such as the sparsity of connections between system components and the interplay between nodal *communication* and network dynamics), various challenges arise from the specific application areas (such as privacy concerns in cyber-physical systems or the need for *scalable* algorithm designs due to the large size of various biological and engineered networks).

On the other hand, networked systems provide significant opportunities and allow for performance and robustness levels that are far beyond reach for centralized systems, with examples ranging from the Internet (of Things) to the smart grid and the brain. This dissertation aims to address several of these challenges and harness these opportunities.

The dissertation is divided into three parts. In the first part, we study privacy concerns whose resolution is vital for the utility of networked cyber-physical systems. We study the problems of average consensus and convex optimization as two principal distributed computations occurring over networks and design algorithm with rigorous privacy guarantees that provide a *best achievable* tradeoff between network utility and privacy. In the second part, we analyze networks with resource constraints. More specifically, we study three problems of stabilization under communication (bandwidth and latency) limitations in sensing and actuation, optimal time-varying control scheduling problem under limited number of actuators and control energy, and the structure identification problem of under-sensed networks (i.e., networks with latent nodes). Finally in the last part, we focus on the intersection of networked dynamical systems and neuroscience and draw connections between brain network dynamics and two extensively studied but yet not fully understood neuro-cognitive phenomena: goal-driven selective attention and neural oscillations. Using a novel axiomatic approach, we establish these connections in the form of necessary and/or sufficient conditions on the network structure that match the network output trajectories with experimentally observed brain activity.

Chapter 1

Introduction

Consider any natural or man-made system of your choice. Most likely, this system is composed of smaller subsystems (or components) and its properties and function can be described, at one level or another, in terms of the properties and function of its subsystems. By the mere fact that these components constitute a larger system, they need to be interacting with each other. This ubiquitous pattern of interacting components (or better, the abstraction of this pattern) is what we call a *network*.

My favorite example is the brain. The idea of the brain as a network is more than a century old and perhaps sounds already familiar, but the precise process of abstracting the brain into a network is anything but a matter of consensus. The reason is that the definitions of “components" and “interactions", i.e., the “nodes" and the “edges", are not unique and depend on the particular properties and function of the brain that one wants to study. We will discuss this in more detail later, but the same ambiguity and non-uniqueness of network abstractions is present in almost all real-world systems, from genetic processes to social phenomena to infrastructure systems such as the power and transportation grids and the internet of things (IoT), to name a few.

A further source of complexity is time. It is often the case that the nodes of the network (and therefore the network as a whole) change with time, and this change is the result of the interactions between the nodes and with the environment. In this case, one is often interested in the time evolution of one (scalar or vector) variable of each node, and thus defines that as the *state* variable of that node. This is in fact one of the most important reasons for abstracting a system as a network as it allows us to break the complexity of its overall dynamics into simpler node-to-node and node-to-environment interactions. Such systems will be hereafter referred to as *networked dynamical systems (NDS)*.

A few examples can ground the discussion. Consider again the example of the brain. The trivial network abstraction defines individual neurons as nodes and the axonal connections between them as edges. In this case, the state variable of each neuron can be its activity level (voltage across its membrane), either as a continuous (analogue) or discrete (digital) value. In the case of a transportation network, one network abstraction considers each terminal as a node with state variable equal to its passenger population and each route as an edge. Another example is that of a social network whereby each person may constitute a node, his/her opinion be his/her state, and the friendships or physical contacts between them form the edges. Finally, in the setting of distributed computation, the dynamical network often coincides with the physical computer network and the state variable of each node is often the estimate of the final output of the desired distributed computation that it updates after each round of communication with its neighbors.

As helpful as dynamical network abstractions are, their analysis poses significant challenges. These include large network size (number of nodes), poor knowledge of its structure, node heterogeneity, nonlinearity, the effects of noise and disturbances, communication and/or computation resource constraints, privacy and/or security, under-sensing and/or under-actuation, heterogeneity

of timescales, and human interactions, to name a few, and arise as one seeks to *identify, analyze, and control* NDS. Addressing these challenges is the central theme of this dissertation.

The dissertation is divided into three parts that are focused on different (sets of) challenges. In Part I (Chapters 3 and 4), we address the privacy challenges that arise when a distributed computation is carried over a dynamical network. Given the various existing forms of distributed computations, we here focus on perhaps the two most fundamental ones, consensus (Chapter 3) and optimization (Chapter 4), and use the elegant notion of differential privacy from the computer science literature to analyze and control the privacy of these networks. Part II (Chapters 5, 6, and 7) is motivated by another critical challenge pertaining the limitation of resources in NDS. We particularly address the limitations of communication (Chapter 5), actuation (Chapter 6), and sensing (Chapter 7) resources in identification and control of NDS. Finally, Part III (Chapters 8 and 9) seeks to make headway into one of the greatest challenges of our time, i.e., understanding the relationship between the NDS of the brain and cognition. We start with the analysis of goal-driven selective attention as one of the most fundamental processes in the brain (Chapter 8), which in turn motivates the analysis of oscillations brain dynamics (Chapter 9). Chapter 10 will conclude the dissertation and discuss a few directions for future research.

1.1 Literature Review

The study of networked systems has a long and vast history [28–30] and has roots in various domains such as computer networks [1–3], neuroscience [4–7], systems and synthetic biology [8–11], sociology [12–15], information privacy [16–19], robotics [20–23], and energy systems [24–27], to name a few. Interestingly, despite the major differences in these domains, there

are various domain-free properties that network abstractions are designed to capture. These include, in particular, the *complexity* (large size, poorly known structure, nonlinearity, etc.) [31, 32] and the existence of dynamics that evolve over these networks [33–35], both playing central roles throughout this dissertation.

A central dichotomy in the study of networked systems is between node-oriented (a.k.a. microscopic) and network-oriented (a.k.a. macroscopic) analysis [36–40]. In the former, the focus is on the properties, objectives, and limitations of individual nodes (at the same time that their interactions lead to a certain collective behavior) while the latter is more interested in the emerging collective behavior and may treat individual nodes as passive and cooperative building blocks or even dissolve them into densities altogether. A particular scenario where this distinction becomes critical is the study of privacy in network dynamical systems.

The study of privacy becomes critical when individual entities (nodes) carry private information and thus have self-prioritizing concerns which may be in tension with the global performance and utility of the networked system [41, 42]. Examples include online user databases (and the pioneering de-anonymization of the Netflix Prize dataset [43]), the smart grid [44], and smart transportation management [45, 46], to name a few, and are rapidly increasing with the ever expansion of the private and public data acquisition systems of our age. This contrast often leads to some form of tradeoff between what the group can achieve (and at what computational and other costs) and how much information each node agrees to disclose. In turn, various notions of privacy and corresponding privacy-ensuring algorithms have been proposed, including k-anonimity [47] and its extensions [48], information-theoretic privacy [49], conditions based on observability [50–52], and differential privacy [53, 54]. The latter has gained significant popularity due to its desirable characteristics such as mathematically elegant formulation, independence from side information,

independence to the adversarial algorithms or capabilities, and immunity to post-processing, and thus will be the basis of our analysis and design of privacy-aware distributed network computations in Part I.

Another barrier to the global utility of networked dynamical systems even in the absence of privacy concerns is the resources available for their analysis and control. Resource limitations are important in almost any engineering system, but become increasingly critical as the size of the system grows, as is the case in most applications of networked systems. One of the most widely studied constraints is the sparsity of communication links between nodes in distributed computations [55, 56]. Of further significant interest have been the limitations on the number of sensors and actuators in networked control systems, motivating the design of optimal sensor [57–62] and actuator [62–65] scheduling algorithms, respectively. An independent line of research has also focused on the communication channels between the nodes and the constraints that realistic channels impose, including time delays [66–73], bandwidth limitations [74–81], and quantization [82–87]. These limitations motivate Part II of this dissertation.

In addition to raising privacy concerns, the dichotomy between microscopic and macroscopic phenomena also complicates network analysis and control. One elegant approach in breaking this complexity has been the study of networks at the mesoscopic or subnetwork level as a middle ground between the two extremes. This can be either implicit in the very definition of the nodes (i.e., by (re-)defining each node to encompass several microscopic entities and assigning to it an aggregate state variable) or explicit in the analysis and control of the network in terms of its smaller subnetworks [88–92], and has been employed in the study of neuronal [7, 93–95], transportation [96–98], social [99–102], epidemic [103–105], and ecological [106–108] networks. In Part III, we combine the implicit and the explicit approach to break the complexity of one the most

complex networked dynamical systems –the brain.

1.2 Statement of Contributions

Differentially Private Average Consensus: In Chapter 3, we study the average consensus problem where a group of agents seek to compute and agree on the average of their local variables while seeking to keep them differentially private against an adversary with potential access to all group communications. This privacy requirement also applies to the case where each agent wants to keep its initial state private against the rest of the group (e.g., due to the possibility of communication leakages). The main contributions of this work are the characterization and optimization of the fundamental trade-offs between differential privacy and average consensus.

Our first contribution is the formulation and formal proof of a general impossibility result. We show that as long as a coordination algorithm is differentially private, it is impossible to guarantee the convergence of agents' states to the average of their initial values, even in distribution. This result automatically implies the same impossibility result for stronger notions of convergence. Motivated by it, our second contribution is the design of a linear Laplacian-based consensus algorithm that achieves average consensus in expectation —the most that one can expect. We prove the almost sure convergence and differential privacy of our algorithm and characterize its accuracy and convergence rate.

Our final contribution is the computation of the optimal values of the design parameters to achieve the most accurate consensus possible. Letting the agents fix a (local) desired value of the privacy requirement, we minimize the variance of the algorithm convergence point as a function of the noise-to-state gain and the amplitude and decay rate of the noise. We show that the mini-

mum variance is achieved by the one-shot perturbation of the initial states by Laplace noise. This result reveals the optimality of one-shot perturbation for static average consensus, previously (but implicitly) shown in the sense of information-theoretic entropy. Various simulations are presented to illustrate our results.

Differentially Private Distributed Optimization: In Chapter 4, we consider a group of agents that seek to minimize the sum of their individual objective functions over a communication network in a differentially private manner. Our first contribution is to show that coordination algorithms which rely on perturbing the agents' messages with noise cannot satisfy the requirements of differential privacy if the underlying noiseless dynamics are locally asymptotically stable. The presence of noise necessary to ensure differential privacy is known to affect the algorithm accuracy in solving the distributed convex optimization problem. However, this result explains why message-perturbing strategies incur additional inaccuracies that are present even if no noise is added.

Our second contribution is motivated by the goal of guaranteeing that the algorithm accuracy is only affected by the presence of noise. We propose a general framework for functional differential privacy over Hilbert spaces and introduce a novel definition of adjacency using adjacency spaces. The latter notion is quite flexible and includes, as a special case, the conventional bounded-difference notion of adjacency. We carefully specify these adjacency spaces within the L_2 space such that the requirement of differential privacy can be satisfied with bounded perturbations.

Our third contribution builds on these results on functional perturbation to design a class of distributed, differentially private coordination algorithms. We let each agent perturb its own objective function based on its desired level of privacy, and then the group uses any provably correct distributed coordination algorithm to optimize the sum of the individual perturbed functions.

Two challenges arise to successfully apply this strategy: the fact that the perturbed functions might lose the smoothness and convexity properties of the original functions and the need to characterize the effect of the added noise on the minimizer of the resulting problem. We address the first challenge using a cascade of smoothening and projection steps that maintain the differential privacy of the functional perturbation step. We address the second challenge by explicitly bounding the absolute expected deviation from the original optimizer using a novel Lipschitz characterization of the arg min map. By construction, the resulting coordination algorithms satisfy the requirement of recovering perfect accuracy in the absence of noise. This chapter also includes various simulations that illustrate our results.

Event-Triggered Stabilization of Delayed Systems: In Chapter 5, we turn to a different challenge in the analysis and control of NDS, i.e., the existence of imperfect communication channels between the network nodes. Our contributions are threefold. First, we design an event-triggered controller for stabilization of nonlinear systems with arbitrarily large sensing and actuation delays. We employ the method of predictor feedback to compensate for the delay in both and then co-design the control law and triggering strategy to guarantee the monotonic decay of a Lyapunov-Krasovskii functional.

Our second contribution involves the closed-loop analysis of the event-triggered law, proving that the closed-loop system is globally asymptotically stable and the inter-event times are uniformly lower bounded (and thus no Zeno behavior may exist). Due to the importance of linear systems in numerous applications, we briefly discuss the simplifications of the design and analysis in this case.

Our final contribution pertains to the trade-off between convergence rate and sampling. Our analysis in this part is limited to linear systems, where closed-form solutions are derivable for

(exponential) convergence rate and minimum inter-event times. We provide a quantitative account of the well-known trade-off between sampling and convergence in event-triggered designs and show how this trade-off can be biased in either direction by tuning a design parameter. Finally, we present simulations to illustrate the effectiveness of our design and address its numerical implementation.

Time-Varying Control Scheduling (TVCS): In Chapter 6, we address the limitation of actuation resources in the control of networked dynamical systems. First, we show that $2k$-*communicability*, a new notion of nodal centrality that we introduce, plays a fundamental role in TVCS. This notion measures the centrality of each node in the network at different spatial scales. Throughout this work, the *spatial scale* (or simply *scale*) of any notion of centrality is defined as the maximum topological distance between pairs of nodes that allows them to affect the centrality of each other, where topological distance between a pair of nodes refers to the minimum number of edges in the graph of the network that should be traversed to go from one to the other. In particular, the spatial scale of degree centrality is 1, while the spatial scale of eigenvector centrality is ∞. Based on the distinction between local and global nodal centralities (i.e., centralities with small and large spatial scales, respectively), we show that the optimal control node at every time instance is the node with the largest centrality at the appropriate scale (i.e., the node with the largest $2k$-communicability at an appropriate k). Accordingly, our main conclusion is that the benefit of TVCS is directly related to the *scale-heterogeneity of central nodes* in the network: the most benefit is gained in networks where the highest centrality is attained by various nodes at different spatial scales, while this benefit starts to decay as fewer nodes dominate the network at all scales (i.e., scale-homogeneity).

Moreover, we provide an extensive discussion of how the dynamical adjacency matrix of a network can (and should) be extracted from its static connectivity, a vital step that is often ignored

in the literature. Indeed, our simulation results show that this step has a significant effect on the benefit of TVCS, with *transmission* networks (networks with states that represent physical quantities transmitted over the network) benefiting significantly more than *induction* networks (those with non-physical states that induce state dynamics over the network) from TVCS.

Structure Identification with Latent Nodes: In Chapter 7, we consider a scenario where one can only directly actuate and measure a subset of the nodes, termed manifest, of a large linear time-invariant network whose total number of nodes and interaction topology are unknown. The objective is to identify the manifest transfer function, which is the submatrix corresponding to the manifest nodes of the transfer function matrix of the entire network. To achieve this, we study the transfer functions provided by linear autoregressive (AR) models. Our discussion shows how AR models can be used to effectively distinguish direct interactions between manifest nodes from indirect interactions mediated by latent nodes. Our first contribution shows that, if no inputs act on the latent nodes, then there exists a class of AR models whose transfer functions converge exponentially in the H_∞ norm to the manifest transfer function as the model order increases. We also show that, if the latent subnetwork is acyclic, then this approximation is exact above a specific model order.

Our second contribution characterizes the properties of using least-squares auto-regressive estimation to construct the AR model from measured data. We establish that the least-squares matrix estimate converges in probability to the optimal matrix sequence identified in our first contribution, enabling us to determine whether two manifest nodes interact directly or indirectly through latent nodes. We also show that the least-squares auto-regressive method guarantees an arbitrarily small H_∞-norm error as the length of data and the model order grow. In fact, once the order of the AR model candidates exceeds a certain threshold, the H_∞-norm error decays exponentially.

Finally, we show that, when the latent subnetwork is acyclic, the method achieves perfect

identification of the manifest transfer function. Throughout a series of remarks, we also discuss how our results can be extended to the identification of linear network models of arbitrary order. Simulations on a directed ring network, Erdős–Rényi random graphs, and real EEG data illustrate our results.

Hierarchical Selective Recruitment: In Chapter 8, we continue our study of the relationship between brain network dynamics and cognition, and focus on the particular phenomenon of goal-driven selective attention (GDSA) as one of the most fundamental processes in the brain that enable cognition. This chapter contains six main contributions. First, we analyze the internal dynamics of a single-layer linear-threshold network as a basis for our study of hierarchical structures. Our results are a combination of previously known results (for which we give simpler proofs) and novel ones, providing a comprehensive characterization of the dynamical properties of linear-threshold networks. Specifically, we show that existence and uniqueness of equilibria, asymptotic stability, and boundedness of trajectories can be characterized using simple algebraic conditions on the network structure in terms of the class of P-matrices (matrices with positive principal minors), totally-Hurwitz matrices (those with Hurwitz principal submatrices, shown to be a sub-class of P-matrices), and Schur-stable matrices, respectively. In addition to forming the basis of HSR, these results solve some of the long-standing open problems in the characterization of linear-threshold networks and are of independent interest.

Our second contribution pertains the problem of selective inhibition in a bilayer network composed of two subnetworks. Motivated by the mechanisms of inhibition in the brain, we study feedforward and feedback inhibition mechanisms. We provide necessary and sufficient conditions on the network structure that guarantee selective inhibition of task-irrelevant nodes at the lower-level while simultaneously guaranteeing various dynamical properties of the resulting (partly in-

hibited, partly active) subnetwork, including existence and uniqueness of equilibria and asymptotic stability. Interestingly, under both mechanisms, these conditions require that (i) there exist at least as many independent inhibitory control inputs as the number of nodes to be inhibited, and (ii) the (not-inhibited) task-relevant part of the lower-level subnetwork intrinsically satisfies the desired dynamical properties. Therefore, when sufficiently many inhibitory control inputs exist, the intrinsic dynamical properties of the task-relevant part are the sole determiner of the dynamical properties achievable under feedforward and feedback inhibitory control. This is particularly important for selective inhibition as asymptotic stability underlies it. These results unveil the important role of task-relevant nodes in constraining the dynamical properties achievable under selective inhibition and have further implications for the number and centrality of nodes that need to be inhibited in order for an unstable-in-isolation subnetwork to gain stability through selective inhibition. For subnetworks that are not stable as a whole, these results provide conditions on the task-relevant/irrelevant partitioning of the nodes that allow for stabilization using inhibitory control.

Third, we use the timescale separation in hierarchical brain networks and the theory of singular perturbations to provide an analytic account of top-down recruitment in terms of conditions on the network structure. These conditions guarantee the stability of the task-relevant part of a (fast) linear-threshold subnetwork towards a reference trajectory set by a slower subnetwork. This, in particular, subsumes the standard "modulation" (enhancement) of the activity of task-relevant nodes (as the most widely observed phenomena in GDSA) but is significantly more general, and can account for recent, complex observations such as shifts in neuronal receptive fields under GDSA. We further combine these results with the results of Part I to allow for simultaneous selective inhibition and top-down recruitment, as observed in GDSA.

Fourth, we extend this combination to hierarchical structures with an arbitrary number of

layers, as observed in nature, to yield a fully developed HSR framework. Here, we derive an extension of the stability results in Part I that guarantees GES of a multi-layer multiple timescale linear-threshold network. This, together with a recursive application of singular perturbation theory, guarantees top-down recruitment of the task-relevant nodes in each layer towards the desired trajectory set by the layer above.

Fifth, to validate the proposed HSR framework, we provide a detailed case study of GDSA in real brain networks. Using single-unit recordings from two brain regions of rodents performing a selective listening task, we provide an in-depth analysis of appropriate choices of neuronal populations in each brain region as well as the timescales of their dynamics. We propose a novel hierarchical structure for these populations, tune the parameters of the resulting network to match its state trajectories with their measured estimates using a novel objective function, and show that the resulting structure conforms to the theoretical results and requirements of HSR while explaining more than 90% of variability in the data.

As part of our technical approach, our sixth and final contribution is a novel converse Lyapunov theorem that extends the state of the art on GES for state-dependent switched affine systems. This result only requires continuity of the vector field and guarantees the existence of an infinitely smooth quadratically-growing Lyapunov function if the dynamics is GES. Because of independent interest, we formulate and prove the result for general state-dependent switched affine systems.

Oscillations and Coupling in Brain Networks: Finally Chapter 9 is motivated by the extension of the results of Chapter 8 (which is based on encoding of sensory information in equilibrium attractors) to more complex attractors. As a first step, this chapter tackles the long-standing problem of characterizing oscillatory activity in terms of brain network structure. Our contributions here are threefold. First, we obtain an exact characterization of existence of limit cycles for

two-dimensional excitatory-inhibitory network motifs described by bounded linear-threshold dynamics (*E-I pairs*). These two-dimensional motifs serve as models of small brain regions that can then be connected to model large-scale brain dynamics.

Accordingly, our second contribution is the study of such networks of oscillators with arbitrary size and connectivity where each oscillator is itself an E-I pair. We derive exact conditions for the lack of stable equilibria and show, through extensive simulations, that this condition is indeed a tight proxy for oscillatory behavior. Finally, using this condition, we study synchronization and PAC as the two most prominent forms of oscillatory coupling in the brain. We show numerically that increasing the inter-oscillator connectivity strength has the same (enhancing) effect on both synchronization and PAC, while increasing frequency mismatch between the oscillators has an opposing effect on them (decreasing synchronization, increasing PAC). Together, these analytical and numerical results provide great insight into the nature of brain oscillations and its relation to the structure of the underlying networks.

Chapter Bibliography

[1] D. E. Comer, *The Internet book: everything you need to know about computer networking and how the Internet works*. Chapman and Hall/CRC, 2018.

[2] L. L. Peterson and B. S. Davie, *Computer networks: a systems approach*. Elsevier, 2007.

[3] D. E. Comer and R. E. Droms, *Computer networks and internets*. Prentice Hall, 2003.

[4] O. Sporns, *Networks of the Brain*, 1st ed. The MIT Press, 2010.

[5] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neuroscience*, vol. 20, no. 3, p. 353, 2017.

[6] A. Fornito, A. Zalesky, and E. Bullmore, *Fundamentals of brain network analysis*. Academic Press, 2016.

[7] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, ser. Computational Neuroscience. Cambridge, MA: MIT Press, 2001.

[8] G. Sanguinetti and V. A. Huynh-Thu, Eds., *Gene Regulatory Networks: Methods and Protocols*. Humana Press, New York, NY, 2019.

[9] H. Kitano, *Foundations of systems biology*. MIT Press, 2001.

[10] P. A. Iglesias and B. P. Ingalls, *Control theory and systems biology*. MIT Press, 2010.

[11] D. D. Vecchio and R. M. Murray, *Biomolecular Feedback Systems*. Princeton University Press, 2015.

[12] S. Wasserman, *Advances in social network analysis: Research in the social and behavioral sciences*. Sage, 1994.

[13] J. Scott and P. J. Carrington, *The SAGE handbook of social network analysis*. SAGE, 2011.

[14] O. Serrat, "Social network analysis," in *Knowledge solutions*. Springer, 2017, pp. 39–43.

[15] D. S. Grewal, *Network power: The social dynamics of globalization*. Yale University Press, 2008.

[16] R. Song and L. Korba, "Review of network-based approaches for privacy," in *Proceedings of the 14th Annual Canadian Information Technology Security Symposium*. Citeseer, 2002, pp. 13–17.

[17] N. Li, N. Zhang, S. K. Das, and B. Thuraisingham, "Privacy preservation in wireless sensor networks: A state-of-the-art survey," *Ad Hoc Networks*, vol. 7, no. 8, pp. 1501–1514, 2009.

[18] S. Zhong, H. Zhong, X. Huang, P. Yang, J. Shi, L. Xie, and K. Wang, "Networking cyber-physical systems: System fundamentals of security and privacy for next-generation wireless networks," in *Security and Privacy for Next-Generation Wireless Networks*. Springer, 2019, pp. 1–32.

[19] X. Chen and S. Shi, "A literature review of privacy research on social network sites," in *2009 International Conference on Multimedia Information Networking and Security*, vol. 1. IEEE, 2009, pp. 93–97.

[20] A. Gautam and S. Mohan, "A review of research in multi-robot systems," in *2012 IEEE 7th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2012, pp. 1–5.

[21] J. C. Barca and Y. A. Sekercioglu, "Swarm robotics reviewed," *Robotica*, vol. 31, no. 3, pp. 345–359, 2013.

[22] Y. Mohan and S. G. Ponnambalam, "An extensive review of research in swarm robotics," in *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*. IEEE, 2009, pp. 140–145.

[23] E. Bahceci, O. Soysal, and E. Sahin, "A review: Pattern formation and adaptation in multi-robot systems," *Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-03-43*, 2003.

[24] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control*. McGraw-hill New York, 1994, vol. 7.

[25] J. A. Momoh, *Smart grid: fundamentals of design and analysis*. Wiley, 2012, vol. 63.

[26] D. P. Kaundinya, P. Balachandra, and N. H. Ravindranath, "Grid-connected versus stand-alone energy systems for decentralized power – a review of literature," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 8, pp. 2041–2050, 2009.

[27] Y. Xiao, *Communication and networking in smart grids*. CRC press, 2012.

[28] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[29] T. G. Lewis, *Network science: Theory and applications*. Wiley, 2011.

[30] A. L. Barabási, *Network science*. Cambridge University Press, 2016.

[31] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[32] R. van der Hofstad, *Random Graphs and Complex Networks*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016, no. v. 1.

[33] G. Chen, X. Wang, and X. Li, *Fundamentals of complex networks: models, structures and dynamics*. Wiley, 2014.

[34] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[35] M. Newman, A. L. Barabasi, and D. J. Watts, *The structure and dynamics of networks*. Princeton University Press, 2011, vol. 12.

[36] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 462–470.

[37] B. Tadić, S. Thurner, and G. J. Rodgers, "Traffic on complex networks: Towards understanding global statistical properties from microscopic density fluctuations," *Physical Review E*, vol. 69, no. 3, p. 036102, 2004.

[38] G. Robins, P. Pattison, and J. Woolcock, "Small and other worlds: Global network structures from local processes," *American Journal of Sociology*, vol. 110, no. 4, pp. 894–936, 2005.

[39] V. A. Maksimenko, A. Lüttjohann, V. V. Makarov, M. V. Goremyko, A. A. Koronovskii, V. Nedaivozov, A. E. Runnova, G. van Luijtelaar, A. E. Hramov, and W. Boccaletti, "Macroscopic and microscopic spectral properties of brain networks during local and global synchronization," *Physical Review E*, vol. 96, no. 1, p. 012316, 2017.

[40] M. Doroud, P. Bhattacharyya, S. F. Wu, and D. Felmlee, "The evolution of ego-centric triads: A microscopic approach toward predicting macroscopic network properties," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 172–179.

[41] R. H. Weber, "Internet of things–new security and privacy challenges," *Computer law & security review*, vol. 26, no. 1, pp. 23–30, 2010.

[42] J. Cortés, G. E. Dullerud, S. Han, J. L. Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016, pp. 4252–4272.

[43] A. Narayanan and V. Shmatikov, "How to break anonymity of the Netflix Prize dataset," 2006, arXiv:cs/0610105.

[44] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[45] B. Hoh, T. Iwuchukwu, Q. Jacobson, D. Work, A. M. Bayen, R. Herring, J. C. Herrera, M. Gruteser, M. Annavaram, and J. Ban, "Enhancing privacy and accuracy in probe vehicle-based traffic monitoring via virtual trip lines," *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, pp. 849–864, 2011.

[46] W. Xin, J. Chang, S. Muthuswamy, and M. Talas, ""midtown in motion": A new active traffic management methodology and its implementation in New York City," Tech. Rep., 2013.

[47] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[48] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[49] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[50] M. Xue, W. Wang, and S. Roy, "Security concepts for the dynamics of autonomous vehicle networks," *Automatica*, vol. 50, no. 3, pp. 852–857, 2014.

[51] N. E. Manitara and C. N. Hadjicostis, "Privacy-preserving asymptotic average consensus," in *European Control Conference*, Zurich, Switzerland, 2013, pp. 760–765.

[52] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 110–127, 2015.

[53] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Theory of Cryptography Conference*, New York, NY, Mar. 2006, pp. 265–284.

[54] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, Venice, Italy, July 2006, pp. 1–12.

[55] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, electronically available at http://coordinationbook.info.

[56] G. Tel, *Introduction to distributed algorithms*. Cambridge University Press, 2000.

[57] V. Gupta, T. H. Chung, B. Hassibi, and R. M. Murray, "On a stochastic sensor selection algorithm with applications in sensor scheduling and sensor coverage," *Automatica*, vol. 42, no. 2, pp. 251–260, 2006.

[58] L. Zhao, W. Zhang, J. Hu, A. Abate, and C. J. Tomlin, "On the optimal solutions of the infinite-horizon linear sensor scheduling problem," *IEEE Transactions on Automatic Control*, vol. 59, no. 10, pp. 2825–2830, 2014.

[59] S. T. Jawaid and S. L. Smith, "Submodularity and greedy algorithms in sensor scheduling for linear dynamical systems," *Automatica*, vol. 61, pp. 282–288, 2015.

[60] D. Han, J. Wu, H. Zhang, and L. Shi, "Optimal sensor scheduling for multiple linear dynamical systems," *Automatica*, vol. 75, pp. 260–270, 2017.

[61] J. Lin, W. Xiao, F. L. Lewis, and L. Xie, "Energy-efficient distributed adaptive multisensor scheduling for target tracking in wireless sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 6, pp. 1886–1896, 2008.

[62] T. H. Summers and J. Lygeros, "Optimal sensor and actuator placement in complex dynamical networks," in *IFAC World Congress*, Cape Town, South Africa, 2014, pp. 3784–3789.

[63] A. Olshevsky, "Minimal controllability problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 249–258, 2014.

[64] Y. Zhao, F. Pasqualetti, and J. Cortés, "Scheduling of control nodes for improved network controllability," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016, pp. 1859–1864.

[65] V. Tzoumas, M. A. Rahimian, G. J. Pappas, and A. Jadbabaie, "Minimal actuator placement with bounds on control effort," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 67–78, 2016.

[66] O. J. M. Smith, "A controller to overcome dead time," *ISA Transactions*, vol. 6, no. 2, pp. 28–33, 1959.

[67] D. Q. Mayne, "Control of linear systems with time delay," *Electronics Letters*, vol. 4, no. 20, pp. 439–440, October 1968.

[68] A. Manitius and A. Olbrot, "Finite spectrum assignment problem for systems with delays," *IEEE Transactions on Automatic Control*, vol. 24, no. 4, pp. 541–552, Aug 1979.

[69] M. T. Nihtila, "Finite pole assignment for systems with time-varying input delays," in *Decision and Control, Proceedings of the 30th IEEE Conference on*, vol. 1, Dec 1991, pp. 927–928.

[70] M. Krstic, *Delay Compensation for Nonlinear, Adaptive, and PDE Systems*, 1st ed., ser. Systems & Control: Foundations & Applications. Birkhäuser, 2009.

[71] I. Karafyllis and M. Krstic, "Nonlinear stabilization under sampled and delayed measurements, and with inputs subject to delay and zero-order hold," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1141–1154, May 2012.

[72] W. Zhang, M. S. Branicky, and S. M. Phillips, "Stability of networked control systems," *IEEE Control Systems*, vol. 21, no. 1, pp. 84–99, 2001.

[73] L. Zhang, Y. Shi, T. Chen, and B. Huang, "A new method for stabilization of networked control systems with random delays," *IEEE Transactions on Automatic Control*, vol. 50, no. 8, pp. 1177–1181, 2005.

[74] H. Kopetz, *Operating Systems of the 90s and Beyond: International Workshop Proceedings*. Springer Berlin Heidelberg, 1991, ch. Event-triggered versus time-triggered real-time systems, pp. 86–101.

[75] K. J. Åström and B. M. Bernhardsson., "Comparison of Riemann and Lebesgue sampling for first-order stochastic systems," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, Dec. 2002, pp. 2011–2016.

[76] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1680–1685, 2007.

[77] X. Wang and M. D. Lemmon, "Event-triggering in distributed networked control systems," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 586–601, 2011.

[78] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *IEEE Conf. on Decision and Control*, Maui, HI, 2012, pp. 3270–3285.

[79] M. Abdelrahim, R. Postoyan, J. Daafouz, and D. Nešić, "Robust event-triggered output feedback controllers for nonlinear systems," *Automatica*, vol. 75, pp. 96–108, 2017.

[80] M. Velasco, J. M. Fuertes, C. Lin, P. Marti, and S. Brandt, "A control approach to bandwidth management in networked control systems," in *Annual Conference of IEEE Industrial Electronics Society*, vol. 3. IEEE, 2004, pp. 2343–2348.

[81] H. Li, G. Chen, T. Huang, and Z. Dong, "High-performance consensus control in networked systems with limited bandwidth communication and time-varying directed topologies," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 5, pp. 1043–1054, 2016.

[82] L. Keyong and J. Baillieul, "Robust quantization for digital finite communication bandwidth (dfcb) control," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, pp. 1573–1584, 2004.

[83] ——, "Robust and efficient quantization and coding for control of multidimensional linear systems under data rate constraints," *International Journal on Robust and Nonlinear Control*, vol. 17, pp. 898–920, 2007.

[84] P. Tallapragada and N. Chopra, "On co-design of event trigger and quantizer for emulation based control," in *American Control Conference*, Montreal, Canada, June 2012, pp. 3772–3777.

[85] E. Garcia and P. J. Antsaklis, "Model-based event-triggered control for systems with quantization and time-varying network delays," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 422–434, 2013.

[86] D. Lehmann and J. Lunze, "Event-based control using quantized state information," in *IFAC Workshop on Distributed Estimation and Control in Networked Systems*, Annecy, France, Sept. 2010, pp. 1–6.

[87] L. Li, X. Wang, and M. D. Lemmon, "Stabilizing bit-rates in quantized event triggered control systems," in *International Conference on Hybrid Systems: Computation and Control*, Beijing, China, Apr. 2012, pp. 245–254.

[88] X. Wang, P. Cui, J. Wang, J. Pei, W. Zhu, and S. Yang, "Community preserving network embedding," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[89] T. Menara, G. Baggio, D. S. Bassett, and F. Pasqualetti, "Stability conditions for cluster synchronization in networks of heterogeneous Kuramoto oscillators," *IEEE Transactions on Control of Network Systems*, 2019, in press.

[90] D. Y. Kenett, M. Perc, and S. Boccaletti, "Networks of networks–an introduction," *Chaos, Solitons & Fractals*, vol. 80, pp. 1–6, 2015.

[91] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of complex networks*, vol. 2, no. 3, pp. 203–271, 2014.

[92] S. Boccaletti, G. Bianconi, R. Criado, C. I. D. Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.

[93] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, vol. 12, no. 1, pp. 1–24, 1972.

[94] H. Liljenström, "Mesoscopic brain dynamics," *Scholarpedia*, vol. 7, no. 9, p. 4601, 2012.

[95] D. Malagarriga, A. E. P. Villa, J. Garcia-Ojalvo, and A. J. Pons, "Mesoscopic segregation of excitation and inhibition in a brain network model," *PLOS Computational Biology*, vol. 11, no. 2, p. e1004007, 2015.

[96] J. Barceló, Ed., *Fundamentals of traffic simulation*. Springer, 2010, vol. 145.

[97] M. Marinov and J. Viegas, "A mesoscopic simulation modeling methodology for analyzing and evaluating freight train operations in a rail network," *Simulation Modeling Practice and Theory*, vol. 19, no. 1, pp. 516–539, 2011.

[98] T. Hou, H. S. Mahmassani, R. M. Alfelor, J. Kim, and M. Saberi, "Calibration of traffic flow models under adverse weather and application in mesoscopic network simulation," *Transportation Research Record*, vol. 2391, no. 1, pp. 92–104, 2013.

[99] G. Tibély, L. Kovanen, M. Karsai, K. Kaski, J. Kertész, and J. Saramäki, "Communities and beyond: mesoscopic analysis of a large social network with complementary methods," *Physical Review E*, vol. 83, no. 5, p. 056125, 2011.

[100] S. Lozano, A. Arenas, and A. Sánchez, "Mesoscopic structure conditions the emergence of cooperation on social networks," *PLOS One*, vol. 3, no. 4, p. e1892, 2008.

[101] J. P. Bagrow and Y. R. Lin, "Mesoscopic structure and social aspects of human mobility," *PLOS One*, vol. 7, no. 5, p. e37676, 2012.

[102] P. Esmailian, S. E. Abtahi, and M. Jalili, "Mesoscopic analysis of online social networks: The role of negative ties," *Physical Review E*, vol. 90, no. 4, p. 042817, 2014.

[103] L. Gauvin, A. Panisson, A. Barrat, and C. Cattuto, "Revealing latent factors of temporal networks for mesoscale intervention in epidemic spread," *arXiv preprint arXiv:1501.02758*, 2015.

[104] H. Kashisaz and A. H. Darooneh, "The influence of society's mesoscopic structure on the rate of epidemic spreading," *Chaos*, vol. 26, no. 6, p. 063114, 2016.

[105] T. Kobayashi and N. Masuda, "Immunizing networks by targeting collective influencers at a mesoscopic level," *arXiv preprint arXiv:1605.03694*, 2016.

[106] S. Pilosof, M. A. Porter, M. Pascual, and S. Kéfi, "The multilayer nature of ecological networks," *Nature Ecology & Evolution*, vol. 1, no. 4, p. 0101, 2017.

[107] J. Bascompte, P. Jordano, C. J. Melián, and J. M. Olesen, "The nested assembly of plant-animal mutualistic networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9383–9387, 2003.

[108] L. J. Gilarranz, M. Sabatino, M. A. Aizen, and J. Bascompte, "Hot spots of mutualistic networks," *Journal of Animal Ecology*, vol. 84, no. 2, pp. 407–413, 2015.

Chapter 2

Preliminaries

Here, we introduce notational conventions and review basic concepts on graph theory, matrix analysis, probability theory, Hilbert spaces and orthonormal bases, input-to-state stability of dynamical systems, and dynamical rate models of brain networks. The reader familiar with these topics may safely skip this chapter.

2.1 Notation

2.1.1 Vectors and Matrices

We use $\mathbb{R}$, $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{\leq 0}$, $\mathbb{R}_{>0}$, $\mathbb{N}$, and $\mathbb{Z}_{\geq 0}$ to denote the set of reals, nonnegative reals, nonpositive reals, positive reals, positive integers, and nonnegative integers, respectively. We let $\mathbb{R}^{\mathbb{N}}$ and $(\mathbb{R}^n)^{\mathbb{N}}$ denote the space of scalar- and n-vector-valued sequences, respectively. We use bold-faced letters for vectors and matrices. For a sequence $\{x(k)\}_{k=0}^{\infty} \subset \mathbb{R}^{\mathbb{N}}$, we use the shorthand notation $\mathring{x} = \{x(k)\}_{k=0}^{\infty}$, $\mathring{x}_k = \{x(j)\}_{j=0}^{k}$, and $\{\mathring{x}\}_{k_1}^{k_2} = \{x(j)\}_{j=k_1}^{k_2}$, with bold-faced counterparts for vector-valued sequences. If the index of $\mathring{x}$ starts at $k = 1$, with a slight abuse of notation we also denote

$\{x(j)\}_{j=1}^{k}$ by $\mathring{x}_k$ (the starting index will be clear from the context). $\mathbf{1}_n$, $\mathbf{0}_n$, $\mathbf{0}_{m\times n}$, and $\mathbf{I}_n$ stand for the n-vector of all ones, the n-vector of all zeros, the m-by-n zero matrix, and the identity n-by-n matrix (we omit the subscripts when clear from the context). We let $\mathbf{\Pi}_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. Note that $\mathbf{\Pi}_n$ is diagonalizable, has one eigenvalue equal to 1 with eigenspace

$$\mathbb{R}\mathbf{1}_n \triangleq \{a\mathbf{1}_n \mid a \in \mathbb{R}\},$$

and all other eigenvalues equal 0.

Given a vector $\mathbf{x} \in \mathbb{R}^n$, x_i and $(\mathbf{x})_i$ refer to its ith component. Given $\mathbf{A} \in \mathbb{R}^{n\times m}$, a_{ij} and $\mathbf{A}_i$ refer to the (i,j)th entry and ith column, respectively. For block-partitioned $\mathbf{x}$ and $\mathbf{A}$, $\mathbf{x}_i$, $\mathbf{A}_i$, and $\mathbf{A}_{ij}$ refer to the ith block of $\mathbf{x}$, ith block (e.g., row) of $\mathbf{A}$, and (i,j)th block of $\mathbf{A}$, respectively. In block representation of matrices, $\star$ denotes arbitrary blocks whose value is immaterial to the discussion. If $\mathbf{x}$ and $\mathbf{y}$ are vectors, $\mathbf{x} \leq \mathbf{y}$ denotes $x_i \leq y_i$ for all i. For $\mathbf{x} \in \mathbb{R}^n$, $\mathrm{Ave}(\mathbf{x}) = \frac{1}{n}\mathbf{1}_n^T\mathbf{x}$ denotes the average of its components.

For a vector $\boldsymbol{\sigma} \in \{0,1\}^n$, we make the convention that $\mathbf{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix with the elements of $\boldsymbol{\sigma}$ on its diagonal. Likewise, for two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathrm{diag}(\mathbf{A},\mathbf{B})$ denotes the block-diagonal matrix with $\mathbf{A}$ and $\mathbf{B}$ on its diagonal. We use $\|\cdot\|_p$ for the p-norm in both finite and infinite-dimensional normed vector spaces and drop the index for $p = 2$. For matrices, $\|\mathbf{A}\|_{\max} = \max_{i,j} |a_{ij}|$.

For a matrix $\mathbf{A}$, its trace, element-wise absolute value, determinant, spectral radius, and eigenvalue with smallest magnitude are denoted by $\mathrm{tr}(\mathbf{A})$, $|\mathbf{A}|$, $\det(\mathbf{A})$, $\rho(\mathbf{A})$, and $\lambda_{\min}(\mathbf{A})$, respectively, while $\mathrm{range}(\mathbf{A})$ denotes the subspace of $\mathbb{R}^m$ spanned by the columns of $\mathbf{A}$. For symmetric $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{A} > \mathbf{0}$ ($\mathbf{A} < \mathbf{0}$) denotes that $\mathbf{A}$ is positive (negative) definite. A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called

stable if all its eigenvalues have magnitude strictly less than 1. The singular values of $\mathbf{A} \in \mathbb{R}^{m\times n}$ are denoted by $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_{\min\{m,n\}}(\mathbf{A}) \geq 0$.

2.1.2 Sets

For a set S, S^c, $|S|$, S^o, and $\bar{S}$ denotes its complement, cardinality, interior, and closure, respectively. For a function $f : X \to Y$ and sets $A \subseteq X$ and $B \subseteq Y$, we use $f(A) = \{f(\mathbf{x}) \in Y | \mathbf{x} \in A\}$ and $f^{-1}(B) = \{\mathbf{x} \in X | f(\mathbf{x}) \in B\}$. In general, $f(f^{-1}(B)) \subseteq B$. For any topological space X, we denote by $\mathcal{B}(X)$ the set of Borel subsets of X. We let $B(\mathbf{c}, r)$ denote the closed ball with center $\mathbf{c}$ and radius r in Euclidean space and given $\mathbf{m} \in \mathbb{R}^n_{>0}$, $[\mathbf{0}, \mathbf{m}] = [0, m_1] \times \cdots \times [0, m_n]$. Throughout the dissertation, measure-theoretic statements refer to the Lebesgue measure which we denote by $\mu_L(\cdot)$. We denote by $\ell_2 \subset \mathbb{R}^{\mathbb{N}}$ the space of square-summable infinite sequences and by $L_2(S)$ and $C^2(S)$ the set of square-integrable measurable functions and the set of twice continuously differentiable functions over a set S, respectively. If $\{E_k\}_{k=1}^{\infty}$ is a sequence of subsets of Ω such that $E_k \subseteq E_{k+1}$ and $E = \bigcup_k E_k$, then we write $E_k \uparrow E$ as $k \to \infty$. We say $E_k \downarrow E$ as $k \to \infty$ if $E_k^c \uparrow E^c$ as $k \to \infty$.

In $\mathbb{R}^n$, a *hyper-plane* with normal vector $\mathbf{n} \in \mathbb{R}^n$ passing through $\mathbf{x} \in \mathbb{R}^n$ is the $(n-1)$-dimensional space $\{\mathbf{y} \mid \mathbf{n}^T(\mathbf{x}-\mathbf{y}) = 0\}$. A set of n hyperplanes is *degenerate* [1] if their intersection is a point or, equivalently, the matrix composed of their normal vectors is nonsingular. A set $S \subseteq \mathbb{R}^n$ is called

- a *polytope* if it has the form $S = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ for some $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, m \in \mathbb{N}$,
- a *cone* if $c\mathbf{x} \in S$ for any $\mathbf{x} \in S$ and $c \in \mathbb{R}_{\geq 0}$,
- a *translated cone apexed at* $\mathbf{y}$ if $\{\mathbf{x} \,|\, \mathbf{x} + \mathbf{y} \in S\}$ is a cone,

- *convex* if $(1-\theta)\mathbf{x}+\theta\mathbf{y} \in S$ for any $\mathbf{x}, \mathbf{y} \in S, \theta \in [0,1]$,
- *solid* if it has a non-empty interior.

Given a subspace W of a vector space V, $W^{\perp}$ denotes the orthogonal complement of W in V. Given any closed and convex subset $S \subseteq \mathcal{H}$ of a Hilbert space, we denote by proj_S the orthogonal projection operator onto S.

2.1.3 Functions

A function $\alpha : [0,\infty) \to [0,\infty)$ belongs to class $\mathcal{K}$ if it is continuous and strictly increasing and $\alpha(0) = 0$. A function α belongs to $\mathcal{K}_\infty$ if $\alpha \in \mathcal{K}$ and $\lim_{r\to\infty} \alpha(r) = \infty$. Similarly, a function $\beta : [0,\infty) \times [0,\infty) \to [0,\infty)$ belongs to class $\mathcal{KL}$ if $\beta(\cdot, s)$ belongs to class $\mathcal{K}$ for any $s \in [0,\infty)$ and $\beta(r,\cdot)$ is decreasing and $\lim_{s\to\infty} \beta(r,s) = 0$ for any $r \in [0,\infty)$. A map $M : X \to Y$ between two normed vector spaces is $\mathcal{K}$-Lipschitz if there exists $\kappa \in \mathcal{K}_\infty$ such that $\|M(\mathbf{x}_1) - M(\mathbf{x}_2)\| \leq \kappa(\|\mathbf{x}_1 - \mathbf{x}_2\|)$ for all $\mathbf{x}_1, \mathbf{x}_2 \in X$. We use the notation $\mathcal{L}_f S = \nabla S \cdot f$ for the Lie derivative of a function $S : \mathbb{R}^n \to \mathbb{R}$ along the trajectories of a vector field f taking values in $\mathbb{R}^n$. The H_∞-norm of a discrete transfer function $\mathbf{T}$ is $\|\mathbf{T}\|_\infty \triangleq \sup_{-\pi\leq\omega\leq\pi} |\mathbf{T}(\omega)|$.

For $x \in \mathbb{R}$, $[x]^+ = \max\{0, x\}$ and $[x]_0^m = \min\{\max\{x, 0\}, m\}$, which are extended entry-wise to $[\mathbf{x}]^+$ and $[\mathbf{x}]_{\mathbf{0}}^{\mathbf{m}}$, respectively. Given $t \in \mathbb{R}$ and a function f on $\mathbb{R}$, $f(t^+) \triangleq \lim_{s\to t^+} f(s)$ and $f(t^-) \triangleq \lim_{s\to t^-} f(s)$. For $q \in (0,1)$, the Euler function is given by $\varphi(q) = \prod_{k=1}^{\infty}(1-q^k) > 0$. Note that

$$\lim_{k\to\infty} \prod_{j=k}^{\infty}(1-q^j) = \lim_{k\to\infty} \frac{\varphi(q)}{\prod_{j=1}^{k-1}(1-q^j)} = 1.$$

2.2 Graph Theory

We present some useful notions on algebraic graph theory following [2]. Let $\mathcal{G} = (V, E, \mathbf{A})$ denote a weighted undirected graph with vertex set V of cardinality n, edge set $E \subset V \times V$, and symmetric adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}_{\geq 0}$. A path from i to j is a sequence of vertices starting from i and ending in j such that any pair of consecutive vertices is an edge. For $k \geq 1$, $(\mathbf{A}^k)_{ij}$ gives the (weighted) number of paths of length k between nodes i and j. The set of neighbors $\mathcal{N}_i$ of i is the set of nodes j such that $(i, j) \in E$. A graph is connected if for each node there exists a path to any other node. A regular graph of degree k is a graph where all the vertices have k neighbors. A strongly regular graph with parameters (n, k, λ, μ) is a regular graph of n nodes with degree k where any two adjacent vertices have λ common neighbors and any pair of non-adjacent vertices have μ neighbors in common. A cone on $\mathcal{G}$ is a network with $n + 1$ nodes where the last one is connected to all others.

The weighted degree matrix is the diagonal matrix $\mathbf{D} \in \mathbb{R}^{n\times n}$ with diagonal $\mathbf{A}\mathbf{1}_n$. The Laplacian is $\mathbf{L} = \mathbf{D} - \mathbf{A}$ and has the following properties:

- $\mathbf{L}$ is symmetric and positive semi-definite;
- $\mathbf{L}\mathbf{1}_n = \mathbf{0}$ and $\mathbf{1}_n^T\mathbf{L} = \mathbf{0}$, i.e., 0 is an eigenvalue of $\mathbf{L}$ corresponding to the eigenspace $\mathbb{R}\mathbf{1}_n$;
- $\mathcal{G}$ is connected if and only if $\operatorname{rank}(\mathbf{L}) = n - 1$, so 0 is a simple eigenvalue of $\mathbf{L}$;
- All eigenvalues of $\mathbf{L}$ belong to $[0, 2d_{\max}]$, where $d_{\max}$ is the largest element of $\mathbf{D}$.

For convenience, we define $\mathbf{L}_{\text{cpt}} = \mathbf{I}_n - \mathbf{\Pi}_n$.

2.2.1 Network centrality

We briefly review here three centrality measures with spectral characterizations. Consider a network of size n represented by the adjacency matrix $\mathbf{A}$.

Eigenvector centrality [3, 4]: Let $v_i \in \mathbb{R}_{\geq 0}$ denote the centrality value of node $i \in \mathcal{N}$. Eigenvector centrality is based on the idea that the influential nodes are the ones that are connected to other influential nodes. In other words, $v_i \propto \sum_{j=0}^{n} a_{ij} v_j$ for all i. This requires the existence of a constant $\lambda > 0$ such that $\lambda v_i = \sum_{j=0}^{n} a_{ij} v_j$ for all i. In matrix notation, $\mathbf{v} = [v_1 \ \cdots \ v_n]^T$, this becomes $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$, which is an eigenvalue problem. Since $\mathbf{A}$ is non-negative, by the Perron-Frobenius Theorem [5, Fact 4.11.4], there always exists a pair $(\lambda, \mathbf{v}) \in \mathbb{R}_{>0} \times \mathbb{R}^n_{\geq 0}$ such that $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. This vector $\mathbf{v}$ is thus defined as the vector of (right) eigenvector centralities. The same argument can be repeated by reversing the direction of influence flow in the network, leading to the vector of left eigenvector centralities (i.e., a positive vector $\mathbf{u}$ such that $\mathbf{u}^T \mathbf{A} = \lambda \mathbf{u}^T$).

Exponential and resolvent communicability [6, 7]: The communicability of a node measures its ability to communicate with the rest of the network. Different notions of communicability have been proposed for complex networks. For a given node i, these include exponential communicability $(e^{\beta \mathbf{A}})_{ii}$ and the resolvent communicability $((I - \beta \mathbf{A})^{-1})_{ii}$, respectively, where $\beta > 0$. From the power series expansion of $e^{\beta \mathbf{A}}$ and $(I - \beta \mathbf{A})^{-1}$, it follows that the exponential and resolvent communicabilities count the total number of cycles that pass through node i, weighting the "importance" of cycles of length k by $\beta^k / k!$ and β^k, respectively. Thus, the role of β is to determine how local/global these measures are: increasing β increases the weights of longer cycles. One can show [7] that in the extreme cases of $\beta \to \infty$ in the exponential case and $\beta \to \frac{1}{\lambda_{\max}(\mathbf{A})}$ in the resolvent

case, both notions result in the same rankings of nodes as eigenvector centrality.

Degree centrality: The degree centrality of node i is the sum of the i-th row (or column) of $\mathbf{A}$ and provides a measure of the immediate influence of node i on its neighbors.

2.3 Matrix Analysis

Here, we define and characterize several matrix classes of interest and their inclusion relationships. These matrices play a key role in Chapter 8.

Definition 2.3.1. ***(Matrix classes).*** A matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ (not necessarily symmetric) is

(i) *absolutely Schur stable* if $\rho(|\mathbf{A}|) < 1$;

(ii) *totally $\mathcal{L}$-stable*, denoted $\mathbf{A} \in \mathcal{L}$, if there exists $\mathbf{P} = \mathbf{P}^T > \mathbf{0}$ such that $(-\mathbf{I} + \mathbf{A}^T\mathbf{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \mathbf{\Sigma}\mathbf{A}) < \mathbf{0}$ for all $\mathbf{\Sigma} = \text{diag}(\boldsymbol{\sigma})$ and $\boldsymbol{\sigma} \in \{0, 1\}^n$;

(iii) *totally Hurwitz*, denoted $\mathbf{A} \in \mathcal{H}$, if all the principal submatrices of $\mathbf{A}$ are Hurwitz;

(iv) *a P-matrix*, denoted $\mathbf{A} \in \mathcal{P}$, if all the principal minors of $\mathbf{A}$ are positive. □

In working with P-matrices, the principal pivot transform of a matrix plays an important role. Given

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

with nonsingular $\mathbf{A}_{22}$, its principal pivot transform is the matrix

$$\pi(\mathbf{A}) \triangleq \begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Note that $\pi(\pi(\mathbf{A})) = \mathbf{A}$. The next result formalizes several equivalent characterizations of P-matrices.

Lemma 2.3.2. ***(Properties of P-matrices [8, 9]).*** $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is a P-matrix if and only if any of the following holds:*

(i) $\mathbf{A}^{-1}$ *is a P-matrix;*

(ii) all real eigenvalues of all the principal submatrices of $\mathbf{A}$ *are positive;*

(iii) for any $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ *there is* k *such that* $x_k(\mathbf{A}\mathbf{x})_k > 0$*;*

(iv) the principal pivot transform of $\mathbf{A}$ *is a P-matrix.*

The next result states inclusion relationships among the matrix classes in Definition 2.3.1 that will be used in our ensuing discussion.

Lemma 2.3.3. ***(Inclusions among matrix classes).*** *For* $\mathbf{A}, \mathbf{W} \in \mathbb{R}^{n\times n}$*, we have*

(i) $\rho(|\mathbf{W}|) < 1 \Rightarrow -\mathbf{I} + \mathbf{W} \in \mathcal{H}$*;*

(ii) $\|\mathbf{W}\| < 1 \Rightarrow \mathbf{W} \in \mathcal{L}$*;*

(iii) $\mathbf{W} \in \mathcal{L} \Rightarrow -\mathbf{I} + \mathbf{W} \in \mathcal{H}$*;*

(iv) $\mathbf{A} \in \mathcal{H} \Rightarrow -\mathbf{A} \in \mathcal{P}$.

Proof. (i). From [5, Fact 4.11.19], we have that $\rho(|\mathbf{W}_\sigma|) < 1$ for any principal submatrix $\mathbf{W}_\sigma$ of $\mathbf{W}$, which in turn implies $\rho(\mathbf{W}_\sigma) < 1$ by [5, Fact 4.11.17], implying the result.

(ii) It is straightforward to check that $\mathbf{P} = \mathbf{I}_n$ satisfies $(-\mathbf{I} + \mathbf{W}^T\boldsymbol{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}) < \mathbf{0}$ for all $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma}), \boldsymbol{\sigma} \in \{0, 1\}^n$.

(iii) Pick an arbitrary $\boldsymbol{\sigma} \in \{0, 1\}^n$ and let the permutation $\boldsymbol{\Pi} \in \mathbb{R}^{n\times n}$ be such that

$$\boldsymbol{\Pi}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Pi}^T = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{W}}_{21} & \hat{\mathbf{W}}_{22} \end{bmatrix},$$

where $\hat{\mathbf{W}}_{22}$ is the principal submatrix of $\mathbf{W}$ corresponding to $\boldsymbol{\sigma}$. Then

$$\begin{aligned} \mathbf{P}(-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}) &= \mathbf{P}\boldsymbol{\Pi}^T \begin{bmatrix} -\mathbf{I} & \mathbf{0} \\ \hat{\mathbf{W}}_{21} & -\mathbf{I} + \hat{\mathbf{W}}_{22} \end{bmatrix} \boldsymbol{\Pi} \\ &= \boldsymbol{\Pi}^T \Big(\underbrace{\boldsymbol{\Pi}\mathbf{P}\boldsymbol{\Pi}^T}_{\hat{\mathbf{P}}} \begin{bmatrix} -\mathbf{I} & \mathbf{0} \\ \hat{\mathbf{W}}_{21} & -\mathbf{I} + \hat{\mathbf{W}}_{22} \end{bmatrix} \Big) \boldsymbol{\Pi} \\ &= \boldsymbol{\Pi}^T \begin{bmatrix} \star & \star \\ \star & \hat{\mathbf{P}}_{22}(-\mathbf{I} + \hat{\mathbf{W}}_{22}) \end{bmatrix} \boldsymbol{\Pi}, \end{aligned}$$

where $\hat{\mathbf{P}} = \begin{bmatrix} \hat{\mathbf{P}}_{11} & \hat{\mathbf{P}}_{12} \\ \hat{\mathbf{P}}_{21} & \hat{\mathbf{P}}_{22} \end{bmatrix} = \hat{\mathbf{P}}^T > \mathbf{0}$. Thus, by assumption,

$$\begin{aligned} &\mathbf{\Pi}^T \begin{bmatrix} \star & \star \\ \star & (-\mathbf{I} + \hat{\mathbf{W}}_{22}^T)\hat{\mathbf{P}}_{22} + \hat{\mathbf{P}}_{22}(-\mathbf{I} + \hat{\mathbf{W}}_{22}) \end{bmatrix} \mathbf{\Pi} < \mathbf{0} \\ \Rightarrow &\begin{bmatrix} \star & \star \\ \star & (-\mathbf{I} + \hat{\mathbf{W}}_{22}^T)\hat{\mathbf{P}}_{22} + \hat{\mathbf{P}}_{22}(-\mathbf{I} + \hat{\mathbf{W}}_{22}) \end{bmatrix} < \mathbf{0} \\ \Rightarrow &(-\mathbf{I} + \hat{\mathbf{W}}_{22}^T)\hat{\mathbf{P}}_{22} + \hat{\mathbf{P}}_{22}(-\mathbf{I} + \hat{\mathbf{W}}_{22}) < \mathbf{0}, \end{aligned}$$

proving that $-\mathbf{I} + \hat{\mathbf{W}}_{22}$ is Hurwitz. Since $\boldsymbol{\sigma}$ is arbitrary, $-\mathbf{I} + \mathbf{W}$ is totally Hurwitz.

(iv) The result follows from Lemma 2.3.2(ii). □

For a general matrix $\mathbf{W}$, neither of $\rho(|\mathbf{W}|)$ and $\|\mathbf{W}\|$ is bounded by the other. However, if $\mathbf{W}$ satisfies the *Dale's law* (as biological neuronal networks do), i.e., each column is either nonnegative or nonpositive, then $\mathbf{W} = |\mathbf{W}|\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix such that $|\mathbf{D}| = \mathbf{I}$. Then,

$$\begin{aligned} \|\mathbf{W}\| &= \sqrt{\rho(\mathbf{W}\mathbf{W}^T)} = \sqrt{\rho(|\mathbf{W}|\mathbf{D}\mathbf{D}|\mathbf{W}|^T)} \\ &= \sqrt{\rho(|\mathbf{W}||\mathbf{W}|^T)} = \||\mathbf{W}|\| \geq \rho(|\mathbf{W}|), \end{aligned}$$

showing that, in this case, $\rho(|\mathbf{W}|) < 1$ is a less restrictive condition. Figure 2.1 depicts a Venn diagram of the various matrix classes of interest to help visualize their relationships.

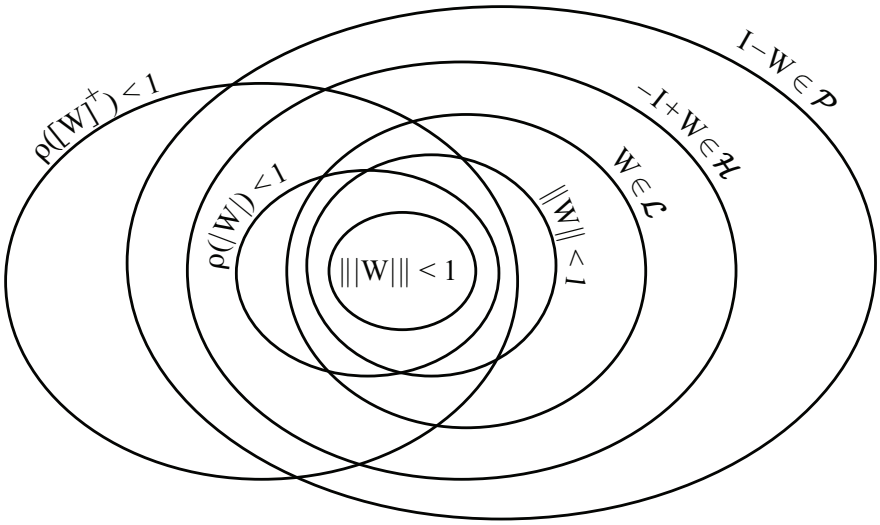


Figure 2.1: Inclusion relationships between the matrix classes introduced in Definition 2.3.1 (cf. Lemma 2.3.3).

2.4 Probability Theory

Here we briefly review basic notions on probability following [10,11]. Consider a probability space $(\Omega, \Sigma, \mathbb{P})$. If $E, F \in \Sigma$ are two events with $E \subseteq F$, then $\mathbb{P}\{E\} \leq \mathbb{P}\{F\}$. For simplicity, we may sometimes denote events of the type $E_p = \{\omega \in \Omega \mid p(\omega)\}$ by $\{p\}$, where p is a logical statement on the elements of Ω. Clearly, for two statements p and q,

$$(p \Rightarrow q) \Rightarrow (\mathbb{P}\{p\} \leq \mathbb{P}\{q\})\,. \tag{2.1}$$

A random variable is a measurable function $X : \Omega \to \mathbb{R}$. For any $N \in \mathbb{R}_{>0}$ and any random variable X with finite expected value μ and finite nonzero variance σ^2, Chebyshev's inequality states that

$$\mathbb{P}\{|X - \mu| \geq N\sigma\} \leq \frac{1}{N^2}. \tag{2.2}$$

For a random variable X, let $\mathbb{E}[X]$ and F_X denote its expectation and cumulative distribution function, respectively. A sequence of random variables $\{X_k\}_{k\in\mathbb{Z}_{\geq 0}}$ converges to a random variable X

- almost surely (a.s.) if $\mathbb{P}\{\lim_{k\to\infty} X_k = X\} = 1$;
- in mean square if $\mathbb{E}[X_k^2], \mathbb{E}[X^2] < \infty$ for all $k \in \mathbb{Z}_{\geq 0}$ and $\lim_{k\to\infty} \mathbb{E}[(X_k - X)^2] = 0$;
- in probability, denoted $\operatorname{plim}_{k\to\infty} X_k = X$, if $\lim_{k\to\infty} \mathbb{P}\{|X_k - X| < \upsilon\} = 1$ for any $\upsilon > 0$;
- in distribution or weakly if $\lim_{k\to\infty} F_{X_k}(x) = F_X(x)$ for any $x \in \mathbb{R}$ at which F_X is continuous.

These definitions are extended to vectors and matrices in an element-wise manner. Almost sure convergence and convergence in mean square imply convergence in probability, which itself implies convergence in distribution. Moreover, if $\mathbb{P}\{|X_k| \leq \bar{X}\} = 1$ for all $k \in \mathbb{Z}_{\geq 0}$ and some fixed random variable $\bar{X}$ with $\mathbb{E}[\bar{X}^2] < \infty$, then convergence in probability implies mean square convergence, and if X is a constant, then convergence in distribution implies convergence in probability.

For $a, b \in \mathbb{R}$, $\mathcal{U}(a, b)$ denotes the uniform distribution over $[a, b]$. The (zero-mean) Laplace distribution with scale $b \in \mathbb{R}_{>0}$ is a continuous distribution with probability density function

$$\mathcal{L}(x; b) = \frac{1}{2b} e^{-\frac{|x|}{b}}.$$

It is clear that $\frac{\mathcal{L}(x;b)}{\mathcal{L}(y;b)} \leq e^{\frac{|x-y|}{b}}$. We use $\eta \sim \text{Lap}(b)$ to denote a random variable η with Laplace distribution. It is easy to see that if $\eta \sim \text{Lap}(b)$, $|\eta|$ has an exponential distribution with rate $\lambda = \frac{1}{b}$. Similarly, we use the notation $\eta \sim \mathcal{N}(\mu, \sigma^2)$ when η is normally distributed with mean μ and variance σ^2.

The error function $\mathrm{erf} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\mathrm{erf}(x) \triangleq \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt \geq 1 - e^{-x^2}.$$

Therefore, $\mathbb{P}\{|\eta| \leq r\} = \mathrm{erf}(r/\sqrt{2}\sigma)$ if $\eta \sim \mathcal{N}(0, \sigma^2)$. Given any random variable η and any convex function ϕ, Jensen's inequality states that $\mathbb{E}[\phi(\eta)] \geq \phi(\mathbb{E}[\eta])$. The opposite inequality holds if ϕ is concave.

2.5 Hilbert Spaces and Orthonormal Bases

We review some basic facts on Hilbert spaces and refer the reader to [12] for a comprehensive treatment. A Hilbert space $\mathcal{H}$ is a complete inner-product space. A set $\{e_k\}_{k \in I} \subset \mathcal{H}$ is an orthonormal system if $\langle e_k, e_j \rangle = 0$ for $k \neq j$ and $\langle e_k, e_k \rangle = \|e_k\|^2 = 1$ for all $k \in I$. If, in addition, the set of linear combinations of elements of $\{e_k\}_{k \in I}$ is dense in $\mathcal{H}$, then $\{e_k\}_{k \in I}$ is an orthonormal basis. Here, I might be uncountable: however, if $\mathcal{H}$ is separable (i.e., it has a countable dense subset), then any orthonormal basis is countable. In this case, for any $h \in \mathcal{H}$,

$$h = \sum_{k=1}^{\infty} \langle h, e_k \rangle e_k.$$

We define the coefficient sequence $\mathring{\theta} \in \mathbb{R}^{\mathbb{N}}$ by $\theta_k = \langle h, e_k \rangle$ for $k \in \mathbb{N}$. Then, $\mathring{\theta} \in \ell_2$ and, by Parseval's identity, $\|h\| = \|\mathring{\theta}\|$. For ease of notation, we define $\Phi : \ell_2 \to \mathcal{H}$ to be the linear bijection that maps the coefficient sequence $\mathring{\theta}$ to h. For an arbitrary $D \subseteq \mathbb{R}^d$, $L_p(D)$ is a Hilbert space if and only if $p = 2$, and the inner product is the integral of the product of functions. Moreover, $L_2(D)$ is separable. In the following (particularly in Chapter 4), we assume is an orthonormal basis

for $L_2(D)$ is chosen and $\Phi : \ell_2 \to L_2(D)$ is the corresponding linear bijection between coefficient sequences and functions.

2.6 Input-to-State Stability of Dynamical Systems

2.6.1 Discrete-Time Systems

This section briefly describes notions of robustness for discrete-time systems following [13, 14]. Consider a discrete-time dynamical system of the form

$$\mathbf{x}(k+1) = f(\mathbf{x}(k), \mathbf{u}(k)), \tag{2.3}$$

where $\mathbf{u} \in \mathbb{R}^m$ is a disturbance input and $\mathbf{x} \in \mathbb{R}^n$ is the state. Given an equilibrium point $\mathbf{x}^* \in \mathbb{R}^n$ of the unforced system (i.e., $\mathbf{x}^* = f(\mathbf{x}^*, \mathbf{0})$), we say that (2.3) is

(i) *0-input locally asymptotically stable (0-LAS) relative to* $\mathbf{x}^*$ if by setting $\mathbf{u}(k) = \mathbf{0}, \forall k$, there exists $\rho > 0$ and $\gamma \in \mathcal{KL}$ such that, for every initial condition $\mathbf{x}(0) \in B(\mathbf{x}^*, \rho)$, we have for all $k \in \mathbb{Z}_{\geq 0}$,

$$\|\mathbf{x}(k) - \mathbf{x}^*\| \leq \gamma(\|\mathbf{x}(0) - \mathbf{x}^*\|, k);$$

(ii) *locally input-to-state stable (LISS) relative to* $\mathbf{x}^*$ if there exist $\rho > 0$, $\gamma \in \mathcal{KL}$, and $\kappa \in \mathcal{K}$ such that, for every initial condition $\mathbf{x}(0) \in B(\mathbf{x}^*, \rho)$ and every input satisfying $\|\mathring{\mathbf{u}}\|_\infty \leq \rho$,

we have

$$\|\mathbf{x}(k) - \mathbf{x}^*\| \leq \max\{\gamma(\|\mathbf{x}(0) - \mathbf{x}^*\|, k), \kappa(\|\mathring{\mathbf{u}}_{k-1}\|_\infty)\}, \tag{2.4}$$

for all $k \in \mathbb{N}$. In this case, we refer to ρ as the *robust stability radius* of (2.3) relative to $\mathbf{x}^*$;

(iii) *globally input-to-state stable (ISS) relative to* $\mathbf{x}^*$ if there exists $\gamma \in \mathcal{KL}$ and $\kappa \in \mathcal{K}$ such that, for any bounded input $\mathring{\mathbf{u}}$, any initial condition $\mathbf{x}(0) \in \mathbb{R}^n$, and all $k \in \mathbb{Z}_{\geq 0}$,

$$\|\mathbf{x}(k) - \mathbf{x}^*\| \leq \max\{\gamma(\|\mathbf{x}(0) - \mathbf{x}^*\|, k), \kappa(\|\mathring{\mathbf{u}}_{k-1}\|_\infty),$$

where $\|\mathring{\mathbf{u}}_{k-1}\|_\infty = \sup\{\|\mathbf{u}(j)\| \mid j = 0, \dots, k-1\}$.

By definition, if the system (2.3) is LISS, then it is also 0-LAS, but the converse is also true, cf. [14, Theorem 1].

The system (2.3) is said to have a $\mathcal{K}$-asymptotic gain if there exists a $\kappa_a \in \mathcal{K}$ such that, for any initial condition $\mathbf{x}(0) \in \mathbb{R}^n$,

$$\limsup_{k\to\infty} \|\mathbf{x}(k) - \mathbf{x}^*\| \leq \kappa_a\Big(\limsup_{k\to\infty} \|\mathbf{u}(k)\|\Big).$$

If a system is ISS, then it has a $\mathcal{K}$-asymptotic gain [13, Lemma 3.8]. The following is a local version of this result.

Proposition 2.6.1. ***(Asymptotic gain of LISS systems).*** *Assume system* (2.3) *is LISS relative to* $\mathbf{x}^*$ *with associated robust stability radius* ρ. *If* $\mathbf{x}(0) \in B(\mathbf{x}^*, \rho)$ *and* $\|\mathring{\mathbf{u}}\|_\infty \leq \min\{\kappa^{-1}(\rho), \rho\}$ *(where*

$\kappa^{-1}(\rho) = \infty$ *if* ρ *is not in the range of* κ*), then*

$$\limsup_{k\to\infty} \|\mathbf{x}(k) - \mathbf{x}^*\| \leq \kappa(\limsup_{k\to\infty} \|\mathbf{u}(k)\|).$$

In particular, $\mathbf{x}(k) \to \mathbf{x}^*$ *if* $\mathbf{u}(k) \to \mathbf{0}$ *as* $k \to \infty$.

Proof. From (2.4), we have

$$\begin{aligned} \|\mathbf{x}(k) - \mathbf{x}^*\| &\leq \max\{\gamma(\|\mathbf{x}(0) - \mathbf{x}^*\|, k), \kappa(\|\mathring{\mathbf{u}}\|_\infty)\} \\ &\leq \max\{\gamma(\rho, k), \kappa(\|\mathring{\mathbf{u}}\|_\infty)\}, \end{aligned} \tag{2.5}$$

where we have used $\mathbf{x}(0) \in B(\mathbf{x}^*, \rho)$. Now, for each $k \in \mathbb{N}$, let $\mathring{\mathbf{u}}_{[k]} \in (\mathbb{R}^n)^{\mathbb{N}}$ be defined by $\mathbf{u}_{[k]}(\ell) = \mathbf{u}(k + \ell)$ for all $\ell \in \mathbb{Z}_{\geq 0}$. If there exists k_0 such that $\|\mathring{\mathbf{u}}_{[k_0]}\|_\infty = 0$, then we need to show that $\lim_{k\to\infty} \|\mathbf{x}(k) - \mathbf{x}^*\| = 0$. Since $\gamma \in \mathcal{KL}$, there exists $K \in \mathbb{Z}_{\geq 0}$ such that $\gamma(\rho, k) \leq \rho$ for all $k \geq K$, and since $\kappa(\|\mathring{\mathbf{u}}\|_\infty) \leq \rho$ as well, it follows from (2.5) that $\mathbf{x}(k) \in B(\mathbf{x}^*, \rho)$ for all $k \geq K$. Let $\overline{k} = \max\{k_0, K\}$. Using (2.4), we get

$$\|\mathbf{x}(k) - \mathbf{x}^*\| \leq \gamma(\|\mathbf{x}(\overline{k}) - \mathbf{x}^*\|, k - \overline{k}), \quad \forall k > \overline{k},$$

and the result follows. Assume then that no k_0 exists such that $\|\mathring{\mathbf{u}}_{[k_0]}\|_\infty = 0$. Let $K_0 = 0$ and, for each $j \in \mathbb{N}$, let K_j be such that $\gamma(\rho, k - K_{j-1}) \leq \kappa(\|\mathring{\mathbf{u}}_{[K_{j-1}]}\|_\infty)$ for all $k \geq K_j$ (this sequence is well-defined because $\gamma \in \mathcal{KL}$). Since $\kappa(\|\mathring{\mathbf{u}}_{[K_{j-1}]}\|_\infty) \leq \kappa(\|\mathring{\mathbf{u}}\|_\infty) \leq \rho$, (2.4) holds if we set the

"initial" state to $\mathbf{x}(K_{j-1})$ which implies that $\|\mathbf{x}(k)-\mathbf{x}^*\| \le \kappa(\|\mathring{\mathbf{u}}_{[K_{j-1}]}\|_\infty)$ for all $k \ge K_j$. Therefore,

$$\limsup_{k\to\infty} \|\mathbf{x}(k)-\mathbf{x}^*\| \le \kappa(\|\mathring{\mathbf{u}}_{[K_j]}\|_\infty), \quad \forall j \in \mathbb{Z}_{\ge 0}.$$

The result follows by taking limit of both sides as $j \to \infty$. □

Finally, any LTI system $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k)$ is ISS if $\mathbf{A}$ is stable.

2.6.2 Continuous-Time Systems

Here, we follow [15] to review the definition of input-to-state stability for continuous-time systems. Consider a nonlinear system of the form

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t)), \qquad t \ge 0, \text{ a.e.}, \tag{2.6}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuously differentiable, $f(\mathbf{0},\mathbf{0}) = \mathbf{0}$, and "a.e." (almost everywhere) denotes the fact that $\mathbf{x}$ may not be differentiable on a set of Lebesgue measure zero.

System (2.6) is (globally) input-to-state stable (ISS) if there exist $\alpha \in \mathcal{K}$ and $\beta \in \mathcal{KL}$ such that for any measurable locally essentially bounded input $\mathbf{u} : \mathbb{R}_{\ge 0} \to \mathbb{R}^m$ and any initial condition $\mathbf{x}(0) \in \mathbb{R}^n$, its solution satisfies

$$\|\mathbf{x}(t)\| \le \beta(\|\mathbf{x}(0)\|, t) + \alpha\big(\operatorname{ess\,sup}_{t\ge 0} \|\mathbf{u}(t)\|\big),$$

for all $t \ge 0$. For this system, a continuously differentiable function $S : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ is called an

ISS-Lyapunov function if there exist $\alpha_1, \alpha_2, \gamma, \rho \in \mathcal{K}_\infty$ such that

$$\alpha_1(\|\mathbf{x}\|) \leq S(\mathbf{x}) \leq \alpha_2(\|\mathbf{x}\|), \qquad \forall \mathbf{x} \in \mathbb{R}^n \tag{2.7a}$$

$$\mathcal{L}_f S(\mathbf{x}, \mathbf{u}) \leq -\gamma(\|\mathbf{x}\|) + \rho(\|\mathbf{u}\|) \qquad \forall (\mathbf{x}, \mathbf{u}) \in \mathbb{R}^{n+m}. \tag{2.7b}$$

According to [15, Theorem 1], the system (2.6) is ISS if and only if it admits an ISS-Lyapunov function.

2.7 Dynamical Rate Models of Brain Networks

Here we briefly review, following [16, §7], the fundamental concepts and assumptions that underlie the linear-threshold network model used in Part III. In a lumped model, neural circuits are composed of neurons, each receiving an electrical signal at its *dendrites* from other neurons and generating an electrical response to other neurons at its *axon*. The transmission of activity from one neuron to another takes place at a *synapse*, thus the terms *pre-synaptic* and *post-synaptic* for the two neurons, respectively. Both the input and output signals mainly consist of a sequence of spikes (action-potentials), as shown in Figure 2.2 (top panel), which are modeled as impulse trains of the form

$$\rho(t) = \sum_k \delta(t - t_k),$$

where $\delta(\cdot)$ denotes the Dirac delta function. In many brain areas the exact timing $\{t_k\}$ of $\rho(t)$ seems essentially random, with the information mainly encoded in its firing rate (number of spikes per second). Thus, $\rho(t)$ is modeled as the sample path of an inhomogeneous Poisson point process with

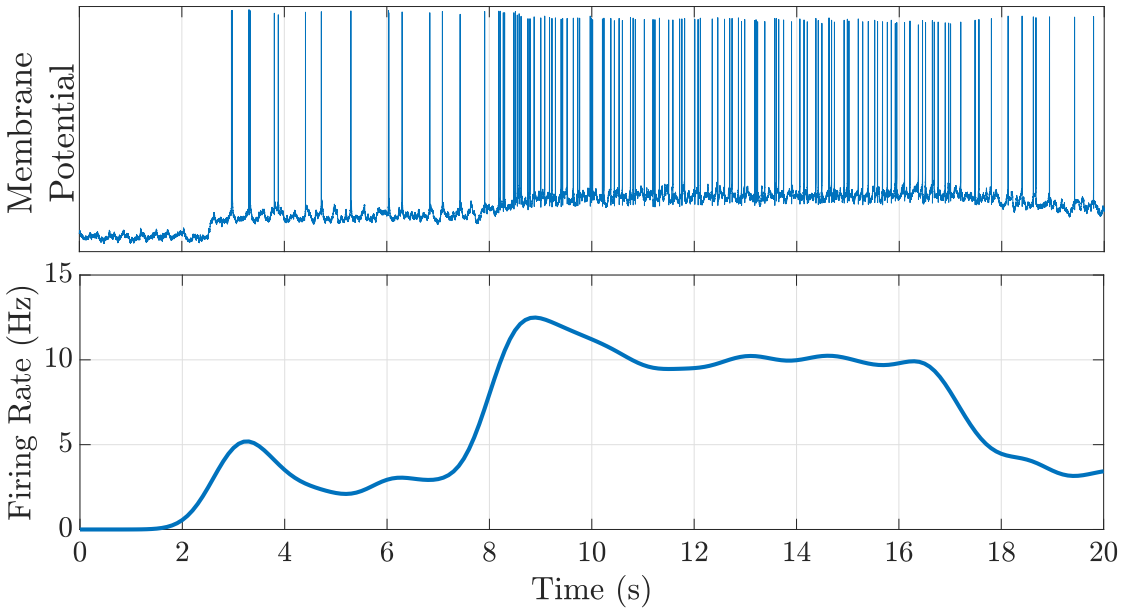


Figure 2.2: A sample intracellular recording illustrating the spike train used for neuronal communication (top panel, measured intracellularly [17, 18]) and the corresponding (estimate of) firing rate (bottom panel).

rate, say, $x(t)$ (cf. Figure 2.2, bottom panel).

Now, consider a pair of pre- and post-synaptic neurons with rates $x_{\text{pre}}(t)$ and $x_{\text{post}}(t)$, respectively. As a result of $x_{\text{pre}}(t)$, an electrical current $I_{\text{post}}(t)$ forms in the post-synaptic neuron's dendrites and soma (body). Assuming fast synaptic dynamics, $I_{\text{post}}(t) \propto x_{\text{pre}}(t)$. Let $w_{\text{post,pre}}$ be the proportionality constant, so $I_{\text{post}}(t) = w_{\text{post,pre}} x_{\text{pre}}(t)$. The pre-synaptic neuron is called excitatory if $w_{\text{post,pre}} > 0$ and inhibitory if $w_{\text{post,pre}} < 0$. In other words, excitatory neurons increase the activity of their out-neighbors while inhibitory neurons decrease it. Notice that this is a property of neurons, not synapses, so a neuron either excites all its out-neighbors or inhibits them.

If the post-synaptic neuron receives input from multiple neurons, $I_{\text{post}}(t)$ follows a superposition law,

$$I_{\text{post}}(t) = \sum_j w_{\text{post},j} x_j(t), \tag{2.8}$$

where the sum is taken over its in-neighbors. If I_{post} is constant, the post-synaptic rate follows $x_{\text{post}} = F(I_{\text{post}})$, where F is a nonlinear "response function". Among the two widely used response functions, namely, sigmoidal and linear-threshold, we use the latter ($F(\cdot) = [\cdot]^+$ or $F(\cdot) = [\cdot]_0^m$) due to its analytical tractability. Finally, if $I_{\text{post}}(t)$ is time-varying, $x_{\text{post}}(t)$ "lags" $F(I_{\text{post}}(t))$ with a time constant τ, i.e.,

$$\tau \dot{x}_{\text{post}}(t) = -x_{\text{post}}(t) + [I_{\text{post}}(t)]^+. \tag{2.9}$$

Equations (2.8)-(2.9) will be the basis for our network model in Chapters 8 and 9.

Chapter Bibliography

[1] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, ser. Comprehensive Studies in Mathematics. New York: Springer, 1998, vol. 317.

[2] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, electronically available at http://coordinationbook.info.

[3] P. Bonacich, "Some unique properties of eigenvector centrality," *Social Networks*, vol. 29, no. 4, pp. 555–564, 2007.

[4] ——, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[5] D. S. Bernstein, *Matrix Mathematics*, 2nd ed. Princeton University Press, 2009.

[6] E. Estrada and N. Hatano, "Communicability in complex networks," *Physical Review E*, vol. 77, p. 036111, 2008.

[7] C. Klymko, "Centrality and communicability measures in complex networks: Analysis and algorithms," Ph.D. dissertation, Emory University, 2013.

[8] M. Fiedler and V. Ptak, "On matrices with non-positive off-diagonal elements and positive principal minors," *Czechoslovak Mathematical Journal*, vol. 12, no. 3, pp. 382–400, 1962.

[9] O. Slyusareva and M. Tsatsomeros, "Mapping and preserver properties of the principal pivot transform," *Linear and Multilinear Algebra*, vol. 56, no. 3, pp. 279–292, 2008.

[10] A. Papoulis and S. U. Pillai, Eds., *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[11] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[12] E. Kreyszig, *Introductory Functional Analysis with Applications*. John Wiley & Sons, 1989.

[13] Z.-P. Jiang and Y. Wang, "Input-to-state stability for discrete-time nonlinear systems," *Automatica*, vol. 37, no. 6, pp. 857–869, 2001.

[14] C. Cai and A. R. Teel, "Results on input-to-state stability for hybrid systems," in *44th IEEE Conference on Decision and Control and European Control Conference*. Seville, Spain: IEEE, Dec. 2005, pp. 5403–5408.

[15] E. D. Sontag and Y. Wang, "On characterizations of the input-to-state stability property," *Systems & Control Letters*, vol. 24, no. 5, pp. 351–359, 1995.

[16] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, ser. Computational Neuroscience. Cambridge, MA: MIT Press, 2001.

[17] D. A. Henze, Z. Borhegyi, J. Csicsvari, M. A. Mamiya, K. D. Harris, and G. Buzsaki, "Intracellular features predicted by extracellular recordings in the hippocampus in vivo," *Journal of Neurophysiology*, vol. 84, no. 1, pp. 390–400, 2000.

[18] D. A. Henze, K. D. Harris, Z. Borhegyi, J. Csicsvari, A. Mamiya, H. Hirase, A. Sirota, and G. Buzsaki, "Simultaneous intracellular and extracellular recordings from hippocampus region ca1 of anesthetized rats," CRCNS.org, 2009. [Online]. Available: http://dx.doi.org/10.6080/K02Z13FP

Part I

Privacy-Aware Dynamic Network Computation

Chapter 3

Differentially Private Average Consensus

The social adoption of new technologies in networked cyberphysical systems relies heavily on the privacy preservation of individual information. Social networking, the power grid, and smart transportation are only but a few examples of domains in need of privacy-aware design of control and coordination strategies. In these scenarios, the ability of a networked system to fuse information, compute common estimates of unknown quantities, and agree on a common view of the world is critical. Motivated by these observations, in this chapter we begin our analysis of differentially private distributed algorithms with the problem of distributed average consensus. We consider a group of agents who seek to agree on the average of their individual values by only interchanging information with their neighbors, a problem that has numerous applications in synchronization, network management, and distributed control/computation/optimization.

In the context of privacy preservation, the notion of differential privacy has gained significant popularity due to its rigorous formulation and proven security properties, including resilience to post-processing and side information, and independence from the model of the adversary. Roughly speaking, a strategy is differentially private if the information of an agent has no

significant effect on the aggregate output of the algorithm, and hence its data cannot be inferred by an adversary from its execution. This chapter is a contribution to the emerging body of research that studies privacy preservation in cooperative network systems, specifically focused on gaining insight into the achievable trade-offs between privacy and performance in multi-agent average consensus.

We first establish that a differentially private consensus algorithm cannot guarantee convergence of the agents' states to the exact average in distribution, which in turn implies the same impossibility for other stronger notions of convergence. This result motivates our design of a novel differentially private Laplacian consensus algorithm in which agents linearly perturb their state-transition and message-generating functions with exponentially decaying Laplace noise. We prove that our algorithm converges almost surely to an unbiased estimate of the average of agents' initial states, compute the exponential mean-square rate of convergence, and formally characterize its differential privacy properties. We show that the optimal choice of our design parameters (with respect to the variance of the convergence point around the exact average) corresponds to a one-shot perturbation of initial states and compare our design with various counterparts from the literature. We end the chapter by numerical simulations that illustrate our results.

3.1 Prior Work

The problem of multi-agent average consensus has been a subject of extensive research in networked systems and it is impossible to survey here the vast amount of results in the literature. We provide [1–4] and the references therein as a starting point for the interested reader.

In cyberphysical systems, privacy at the physical layer provides protection beyond the use of higher-level encryption-based techniques. Information-theoretic approaches to privacy at the

physical layer have been actively pursued [5,6]. Recently, these ideas have also been utilized in the context of control [7]. The paper [6] also surveys the more recent game-theoretic approach to the topic. In computer science, the notion of differential privacy, first introduced in [8,9], and the design of differentially private mechanisms have been widely studied in the context of privacy preservation of databases. The work [10] provides a recent comprehensive treatment. A well-known advantage of differential privacy over other notions of privacy is its immunity to post-processing and side information, which makes it particularly well-suited for multi-agent scenarios where agents do not fully trust each other and/or the communication channels are not fully secure. While secure multi-party computation also deals with scenarios where no trust exists among agents, the maximum number of agents that can collude (without the privacy of others being breached) is bounded, whereas using differential privacy provides immunity against arbitrary collusions [11,12]. As a result, differential privacy has been adopted by recent works in a number of areas pertaining to networked systems, such as control [13–15], estimation [16], and optimization [17–19].

Of relevance to our present work, the paper [13] studies the average consensus problem with differentially privacy guarantees and proposes an adjacency-based distributed algorithm with decaying Laplace noise and mean-square convergence. The algorithm preserves the differential privacy of the agents' initial states but the expected value of its convergence point depends on the network topology and may not be the exact average, even in expectation. By contrast, the algorithm proposed in this work enjoys almost sure convergence, asymptotic unbiasedness, and an explicit characterization of its convergence rate. Our results also allow individual agents to independently choose their level of privacy.

The problem of privacy-preserving average consensus has also been studied using other notions of privacy. The work [20] builds on [21] to let agents have the option to add to their first

set of transmitted messages a zero-sum noise sequence with finite random length, which in turn allows the coordination algorithm to converge to the exact average of their initial states. As long as an adversary cannot listen to the transmitted messages of an agent as well as all its neighbors, the privacy of that agent is preserved, in the sense that different initial conditions may produce the same transmitted messages. This idea is further developed in [22,23], where agents add an infinitely-long exponentially-decaying zero-sum sequence of Gaussian noise to their transmitted messages. The algorithm has guaranteed mean-square convergence to the average of the agents' initial states and preserves the privacy of the nodes whose messages and those of their neighbors are not listened to by the malicious nodes, in the sense that the maximum-likelihood estimate of their initial states has nonzero variance. Finally, [24] considers the problem of privacy preserving maximum consensus.

3.2 Problem statement

Consider a group of n agents whose interaction topology is described by an undirected connected graph $\mathcal{G}$. The group objective is to compute the average of the agents' initial states while preserving the privacy of these values against potential adversaries eavesdropping on all the network communications. Note that this privacy requirement is the same as the case where each agent wants to keep its initial state private against the rest of the group due to the possibility of communication leakages. We next generalize the exposition in [13] to provide a formal presentation of this problem. The state of each agent $i \in \{1, \dots, n\}$ is represented by $\theta_i \in \mathbb{R}$. The message that agent i shares with its neighbors about its current state is denoted by $x_i \in \mathbb{R}$. For convenience, the aggregated network state and the vector of transmitted messages are denoted by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ and $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, respectively. Agents update their states in discrete time according to some

rule,

$$\boldsymbol{\theta}(k+1) = f(\boldsymbol{\theta}(k), \mathbf{x}(k)), \quad k \in \mathbb{Z}_{\geq 0}, \tag{3.1}$$

with initial states $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$, where the state-transition function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is such that its ith element depends only on θ_i and $\{x_j\}_{j \in \mathcal{N}_i \cup \{i\}}$. The messages are calculated as

$$\mathbf{x}(k) = h(\boldsymbol{\theta}(k), \boldsymbol{\eta}(k)), \quad k \in \mathbb{Z}_{\geq 0}, \tag{3.2}$$

where $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is such that its ith element depends only on θ_i and η_i. For simplicity, we assume that f and h are continuous. $\boldsymbol{\eta}(k) \in \mathbb{R}^n$ is a vector random variable, with $\eta_i(k)$ being the noise generated by agent i at time k from an arbitrary distribution. Consequently, $\mathring{\boldsymbol{\theta}}$ and $\mathring{\mathbf{x}}$ are sequences of vector random variables on the total sample space $\boldsymbol{\Omega} = (\mathbb{R}^n)^{\mathbb{N}}$ whose elements are noise sequences $\mathring{\boldsymbol{\eta}}$. Although one could choose h to only depend on $\boldsymbol{\theta}$, corrupting the messages by noise is necessary to preserve privacy. Given an initial state $\boldsymbol{\theta}_0$, $\mathring{\mathbf{x}}$ is uniquely determined by $\mathring{\boldsymbol{\eta}}$ according to (3.1)-(3.2). Therefore, the function $X_{\boldsymbol{\theta}_0} : (\mathbb{R}^n)^{\mathbb{N}} \to (\mathbb{R}^n)^{\mathbb{N}}$ such that

$$X_{\boldsymbol{\theta}_0}(\mathring{\boldsymbol{\eta}}) = \mathring{\mathbf{x}}$$

is well defined.

Definition 3.2.1. ***(Differential privacy).*** Given $\delta \in \mathbb{R}_{>0}$, the initial network states $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ are

δ-adjacent if, for some $i_0 \in \{1, \dots, n\}$,

$$|\theta_{0,i}^{(2)} - \theta_{0,i}^{(1)}| \leq \begin{cases} \delta & \text{if } i = i_0, \\ 0 & \text{if } i \neq i_0, \end{cases} \tag{3.3}$$

for $i \in \{1, \dots, n\}$. Given $\delta, \epsilon \in \mathbb{R}_{\geq 0}$, the dynamics (3.1)-(3.2) is ϵ-differentially private if, for any pair $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ of δ-adjacent initial states and any set $\mathcal{O} \in \mathcal{B}\big((\mathbb{R}^n)^{\mathbb{N}}\big)$,

$$\mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \,|\, X_{\boldsymbol{\theta}_0^{(1)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\} \leq e^{\epsilon} \mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \mid X_{\boldsymbol{\theta}_0^{(2)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\}. \qquad \bullet$$

A final aspect to consider is that, because of the presence of noise, the agents' states under (3.1) might not converge exactly to their initial average $\mathrm{Ave}(\boldsymbol{\theta}_0)$, but to a neighborhood of it. This is captured by the notion of accuracy.

Definition 3.2.2. ***(Accuracy).*** For $p \in [0, 1]$ and $r \in \mathbb{R}_{\geq 0}$, the dynamics (3.1)-(3.2) is (p, r)-accurate if, from any initial state $\boldsymbol{\theta}_0$, the network state $\boldsymbol{\theta}(k)$ converges to $\boldsymbol{\theta}_\infty \in \mathbb{R}^n$ as $k \to \infty$, with $\mathbb{E}[\boldsymbol{\theta}_\infty] = \mathrm{Ave}(\boldsymbol{\theta}_0)\mathbf{1}_n$ and $\mathbb{P}\{\|\boldsymbol{\theta}_\infty - \mathrm{Ave}(\boldsymbol{\theta}_0)\mathbf{1}_n\| \leq r\} \geq 1 - p$. $\square$

In Definition 3.2.2, the type of convergence of $\boldsymbol{\theta}(k)$ to $\boldsymbol{\theta}_\infty$ can be any of the four classes described in Section 2.4. Furthermore, for each notion of convergence, $(0, 0)$-accuracy is equivalent to the convergence of $\boldsymbol{\theta}(k)$ to $\mathrm{Ave}(\boldsymbol{\theta}_0)\mathbf{1}_n$. We are finally ready to formally state our problem.

Problem 1. ***(Differentially private average consensus).*** *Design the dynamics* (3.1)*, the inter-agent messages* (3.2)*, and the distribution of noise sequences $\mathring{\boldsymbol{\eta}}$ such that asymptotic average consensus is achieved with (p, r)-accuracy while guaranteeing ϵ-differential privacy for (finite) ϵ, r, and $p \in \mathbb{R}_{\geq 0}$ as small as possible.* $\square$

3.3 Obstructions to Exact Differentially Private Average Consensus

In this section we establish the impossibility of solving Problem 1 with $(0, 0)$-accuracy, even if considering the weakest notion of convergence.

Proposition 3.3.1 (Impossibility Result)**.** *Consider a group of agents executing a distributed algorithm of the form* (3.1) *with messages generated according to* (3.2)*. Then, for any $\delta, \epsilon > 0$, agents cannot simultaneously converge to the average of their initial states in distribution and preserve ϵ-differential privacy of their initial states.*

Proof. We reason by contradiction. Assume there exists an algorithm that achieves convergence in distribution to the exact average of the network initial state and preserves ϵ-differential privacy of it. Since the algorithm must preserve the privacy of *any* pair of δ-adjacent initial conditions, consider a specific pair satisfying

$$\theta^{(2)}_{0,i_0} = \theta^{(1)}_{0,i_0} + \delta,$$

for some $i_0 \in \{1, \dots, n\}$ and $\theta^{(2)}_{0,i} = \theta^{(1)}_{0,i}$ for all $i \neq i_0$. Since $\mathrm{Ave}(\boldsymbol{\theta}_0)$ is fixed (i.e., deterministic), the convergence of $\theta_i(k)$, $i \in \{1, \dots, n\}$ to $\mathrm{Ave}(\boldsymbol{\theta}_0)$ is also in probability. Thus, for any $i \in \{1, \dots, n\}$ and any $\upsilon > 0$, we have $\lim_{k\to\infty} \mathbb{P}\{|\theta^{(\ell)}_i(k) - \mathrm{Ave}(\boldsymbol{\theta}^{(\ell)}_0)| < \upsilon\} = 1$, for $\ell = 1, 2$. Therefore, for any $\upsilon' > 0$, there exists $k \in \mathbb{Z}_{\geq 0}$ such that for all $i \in \{1, \dots, n\}$,

$$\mathbb{P}\{|\theta^{(\ell)}_i(k) - \mathrm{Ave}(\boldsymbol{\theta}^{(\ell)}_0)| < \upsilon\} > 1 - \upsilon', \quad \ell = 1, 2. \tag{3.4}$$

Now, considering (3.1)-(3.2), it is clear that, for any fixed initial state $\boldsymbol{\theta}_0$ and any $k \in \mathbb{Z}_{\geq 0}$, $\mathring{\mathbf{x}}_k$ is uniquely determined by $\mathring{\boldsymbol{\eta}}_k$ and $\mathring{\boldsymbol{\theta}}_k$ is uniquely determined by $\mathring{\mathbf{x}}_k$. Therefore, the functions $X_{k,\boldsymbol{\theta}_0}, \Theta_{k,\boldsymbol{\theta}_0} : \mathbb{R}^{n(k+1)} \to \mathbb{R}^{n(k+1)}$ such that

$$X_{k,\boldsymbol{\theta}_0}(\mathring{\boldsymbol{\eta}}_k) = \mathring{\mathbf{x}}_k, \quad \Theta_{k,\boldsymbol{\theta}_0}(\mathring{\mathbf{x}}_k) = \mathring{\boldsymbol{\theta}}_k, \tag{3.5}$$

are well defined and continuous (due to continuity of f and g). Next, for $\ell = 1, 2$, define $\boldsymbol{R}_k^{(\ell)} = X^{-1}_{k,\boldsymbol{\theta}_0^{(\ell)}}\left(\Theta^{-1}_{k,\boldsymbol{\theta}_0^{(\ell)}}\left(\mathcal{N}_k^{(\ell)}\right)\right)$, where $\mathcal{N}_k^{(\ell)} \triangleq \mathbb{R}^{nk} \times \left(\mathcal{I}^{(\ell)}\right)^n$ and $\mathcal{I}^{(\ell)} \subset \mathbb{R}$ is the υ-neighborhood of $\mathrm{Ave}(\boldsymbol{\theta}_0^{(\ell)})$. By (3.4), we have

$$\mathbb{P}(\boldsymbol{R}_k^{(\ell)}) > 1 - \upsilon', \quad \ell = 1, 2. \tag{3.6}$$

Note that $\boldsymbol{R}_k^{(1)}$ is open as it is the continuous pre-image of an open set, so $\mathcal{O}_k \triangleq X_{k,\boldsymbol{\theta}_0^{(1)}}\left(\boldsymbol{R}_k^{(1)}\right)$ is Borel. To reach a contradiction, we define $\boldsymbol{R}_k^{\prime(2)} = X^{-1}_{k,\boldsymbol{\theta}_0^{(2)}}(\mathcal{O}_k)$ and show that $\mathbb{P}(\boldsymbol{R}_k^{\prime(2)})$ can be made arbitrarily small by showing that $\boldsymbol{R}_k^{\prime(2)} \cap \boldsymbol{R}_k^{(2)} = \varnothing$. To do this, note that by the definitions of $\boldsymbol{R}_k^{\prime(2)}$, $\mathcal{O}_k$ and $\boldsymbol{R}_k^{(1)}$, we have

$$\Theta_{k,\boldsymbol{\theta}_0^{(1)}}\left(X_{k,\boldsymbol{\theta}_0^{(2)}}\left(\boldsymbol{R}_k^{\prime(2)}\right)\right) \subseteq \mathcal{N}_k^{(1)}. \tag{3.7}$$

Recall that in (3.1), f is such that the next state of each agent only depends on its current state and the messages it receives. Hence, since for all $i \neq i_0$, $\theta_{0,i}^{(2)} = \theta_{0,i}^{(1)}$, we have from (3.7) that

$$\Theta_{k,\boldsymbol{\theta}_0^{(2)}}\left(X_{k,\boldsymbol{\theta}_0^{(2)}}\left(\boldsymbol{R}_k^{\prime(2)}\right)\right) \subseteq \overline{\mathcal{N}}_k^{(1)},$$

where $\overline{\mathcal{N}}_k^{(1)} \triangleq \mathbb{R}^{nk} \times (\mathcal{I}^{(1)})^{i_0-1} \times \mathbb{R} \times (\mathcal{I}^{(1)})^{n-i_0}$ is the same as $\mathcal{N}_k^{(1)}$ except that the requirement on $\theta_{i_0}(k)$ (to be close to Ave($\boldsymbol{\theta}_0^{(1)}$)) is relaxed. Now, since $\Theta_{k,\boldsymbol{\theta}_0^{(2)}}\left(X_{k,\boldsymbol{\theta}_0^{(2)}}\left(R_k^{(2)}\right)\right) \subseteq \mathcal{N}_k^{(2)}$ and, by choosing $\upsilon < \frac{\delta}{2n}$, we get $\overline{\mathcal{N}}_k^{(1)} \cap \mathcal{N}_k^{(2)} = \varnothing$, we conclude that $\Theta_{k,\boldsymbol{\theta}_0^{(2)}}\left(X_{k,\boldsymbol{\theta}_0^{(2)}}\left(R_k^{(2)}\right)\right) \cap \Theta_{k,\boldsymbol{\theta}_0^{(2)}}\left(X_{k,\boldsymbol{\theta}_0^{(2)}}\left(R_k^{\prime(2)}\right)\right) = \varnothing$, which implies $R_k^{(2)} \cap R_k^{\prime(2)} = \varnothing$, so we get

$$\mathbb{P}(R_k^{(2)}) < \upsilon', \tag{3.8}$$

as desired. Now, let $\mathcal{O} = \mathcal{O}_k \times (\mathbb{R}^n)^{\mathbb{N}} \in \mathcal{B}\left((\mathbb{R}^n)^{\mathbb{N}}\right)$. For any initial condition $\boldsymbol{\theta}_0$,

$$\mathbb{P}\{\mathring{\boldsymbol{\eta}} | X_{\boldsymbol{\theta}_0}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\} = \mathbb{P}\{\mathring{\boldsymbol{\eta}}_k | X_{k,\boldsymbol{\theta}_0}(\mathring{\boldsymbol{\eta}}_k) \in \mathcal{O}_k\}.$$

Hence, since the algorithm is ϵ-differentially private,

$$\mathbb{P}(R_k^{(1)}) = \mathbb{P}\{\mathring{\boldsymbol{\eta}}_k | X_{k,\boldsymbol{\theta}_0^{(1)}}(\mathring{\boldsymbol{\eta}}_k) \in \mathcal{O}_k\} \leq e^{\epsilon}\mathbb{P}\{\mathring{\boldsymbol{\eta}}_k | X_{k,\boldsymbol{\theta}_0^{(2)}}(\mathring{\boldsymbol{\eta}}_k) \in \mathcal{O}_k\} = e^{\epsilon}\mathbb{P}(R_k^{\prime(2)}).$$

Thus, using (3.6) and (3.8), we have for all $\upsilon' > 0$,

$$1 - \upsilon' < e^{\epsilon}\upsilon' \Rightarrow \frac{1}{1+e^{\epsilon}} < \upsilon',$$

which is clearly a contradiction because ϵ is a finite number, completing the proof. □

Since convergence in distribution is the weakest notion of convergence, Proposition 3.3.1 implies that a differentially private algorithm cannot guarantee any type of convergence to the exact average. Therefore, in our forthcoming discussion, we relax the exact convergence requirement and allow for convergence to a random variable that is at least unbiased (i.e., centered at the true

average).

3.4 Differentially Private Average Consensus Algorithm

Here, we develop a solution to Problem 1. Consider the following linear distributed dynamics,

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) - h\mathbf{L}\mathbf{x}(k) + \mathbf{S}\boldsymbol{\eta}(k), \tag{3.9}$$

for $k \in \mathbb{Z}_{\geq 0}$, where $h < (d_{\max})^{-1}$ is the step size, $\mathbf{S}$ is a diagonal matrix with diagonal $(s_1, \dots, s_n)$ and $s_i \in (0, 2)$ for each $i \in \{1, \dots, n\}$, and the messages are generated as

$$\mathbf{x}(k) = \boldsymbol{\theta}(k) + \boldsymbol{\eta}(k), \tag{3.10}$$

where the ith component of the noise vector $\boldsymbol{\eta}(k)$ has the Laplace distribution $\eta_i(k) \sim \text{Lap}(b_i(k))$ at any time $k \in \mathbb{Z}_{\geq 0}$ with

$$b_i(k) = c_i q_i^k, \quad c_i \in \mathbb{R}_{>0},\ q_i \in (|s_i - 1|, 1). \tag{3.11}$$

Note that (3.9) is a special case of (3.1) (since $\boldsymbol{\eta}(k) = \mathbf{x}(k) - \boldsymbol{\theta}(k)$) and (3.10) a special case of (3.2). Also note that without the term $\mathbf{S}\boldsymbol{\eta}(k)$, the average of the agents' initial states would be preserved throughout the evolution.

Remark 3.4.1. ***(Comparison with the literature).*** The proposed algorithm (3.9)-(3.11) has similarities and differences with the algorithm proposed in [13] which can be expressed (with a slight

change of notation in using s_i instead of σ_i) as

$$\begin{aligned}\theta(k+1) &= (\mathbf{I}_n - \mathbf{S})\theta(k) + \mathbf{S}\mathbf{D}^{-1}\mathbf{A}\mathbf{x}(k) \\ &= \left[\mathbf{I}_n - \mathbf{S}\mathbf{D}^{-1}\mathbf{L}\right]\theta(k) + \left[\mathbf{S} - \mathbf{S}\mathbf{D}^{-1}\mathbf{L}\right]\boldsymbol{\eta}(k).\end{aligned}$$

If each agent selects $s_i = d_i h < 1$, then we recover (3.9)-(3.11). As we show later, this particular choice results in an unbiased convergence point, while in general the expected value of the convergence point of the algorithm in [13] depends on the graph structure and may not be the true average. Furthermore, this algorithm is only shown to exhibit mean square convergence of $\boldsymbol{\theta}(k)$ for $s_i \in (0, 1)$, while here we provide an explicit expression for the convergence point and establish convergence in the stronger a.s. sense for larger range of $s_i \in (0, 2)$. As we show later, the inclusion of $s_i = 1$ is critical, as it leads to identifying the optimal algorithm performance. On a different note, the algorithms in [15] and [22, 23] add a noise sequence to the messages which is correlated over time – the latter using a different notion of privacy. [15] generate a single noise at time $k = 0$ and add a scaled version of it to the messages at every time $k \geq 1$, leading to an effectively "one-shot"-type of perturbation. We show in Section 3.4.3 that the one-shot approach is optimal for static average consensus while sequential perturbation is necessary for dynamic scenarios. □

3.4.1 Convergence Analysis

This section analyzes the asymptotic correctness of the algorithm (3.9)-(3.11) and characterizes its rate of convergence. We start by establishing convergence.

Proposition 3.4.2. ***(Asymptotic convergence).*** *Consider a network of n agents executing the dis-*

tributed dynamics (3.9)-(3.11). *Define the random variable* θ_∞ *as*

$$\theta_\infty \triangleq \mathrm{Ave}(\boldsymbol{\theta}_0) + \sum_{i=1}^{n} \frac{s_i}{n} \sum_{j=0}^{\infty} \eta_i(j). \tag{3.12}$$

Then, θ_∞ *is well-defined a.s., and the states of all agents converge to* θ_∞ *almost surely.*

Proof. Note that $s_i \in (0, 2)$ ensures that $(|s_i - 1|, 1)$ is not empty. Substituting (3.10) into (3.9), the system dynamics is

$$\boldsymbol{\theta}(k+1) = \mathbf{A}\boldsymbol{\theta}(k) + \mathbf{B}\boldsymbol{\eta}(k), \tag{3.13}$$

with $\mathbf{A} = \mathbf{I}_n - h\mathbf{L}$ and $\mathbf{B} = \mathbf{S} - h\mathbf{L}$. For any $\boldsymbol{\theta} \in \mathbb{R}^n$, let

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathrm{Ave}(\boldsymbol{\theta})\mathbf{1}_n = \mathbf{L}_{\mathrm{cpt}}\boldsymbol{\theta} \in (\mathbb{R}\mathbf{1}_n)^\perp. \tag{3.14}$$

Multiplying both sides of (3.13) by $\mathbf{L}_{\mathrm{cpt}}$ on the left and using the fact that $\mathbf{L}_{\mathrm{cpt}}$ and $\mathbf{L}$ commute, the dynamics of $\tilde{\boldsymbol{\theta}}$ can be expressed as

$$\tilde{\boldsymbol{\theta}}(k+1) = (\mathbf{I}_n - h\mathbf{L})\tilde{\boldsymbol{\theta}}(k) + \mathbf{L}_{\mathrm{cpt}}(\mathbf{S} - h\mathbf{L})\boldsymbol{\eta}(k). \tag{3.15}$$

Notice that $(\mathbb{R}\mathbf{1}_n)^\perp$ is forward invariant under (3.15). Therefore, considering $(\mathbb{R}\mathbf{1}_n)^\perp$ as the state space for (3.15) and noting that $\mathbf{I}_n - h\mathbf{L}$ is stable on it, we deduce that (3.15) is ISS. Consequently, this dynamics has a $\mathcal{K}$-asymptotic gain (c.f. Section 2.6), i.e., there exists $\gamma_a \in \mathcal{K}$ such that

$$\limsup_{k\to\infty} \|\tilde{\boldsymbol{\theta}}(k)\| \leq \gamma_a\Big(\limsup_{k\to\infty} \|\boldsymbol{\eta}(k)\|\Big).$$

Therefore, $\lim_{k\to\infty} \tilde{\boldsymbol{\theta}}(k) \neq 0$ implies $\lim_{k\to\infty} \|\boldsymbol{\eta}(k)\| \neq 0$. By definition, the latter means that there is $\upsilon > 0$ such that for all $K \in \mathbb{N}$ there exists $k \geq K$ with $\|\boldsymbol{\eta}(k)\| > \upsilon$. In other words, there exists a subsequence $\left\{\boldsymbol{\eta}(k_\ell)\right\}_{\ell\in\mathbb{N}}$ such that $\|\boldsymbol{\eta}(k_\ell)\| > \upsilon$ for all $\ell \in \mathbb{N}$. This, in turn, implies that for all $\ell \in \mathbb{N}$, $\|\boldsymbol{\eta}(k_\ell)\|_\infty > \upsilon/\sqrt{n}$, i.e.,

$$\exists i_\ell \in \{1, \dots, n\} \text{ with } |\eta_{i_\ell}(k_\ell)| > \frac{\upsilon}{\sqrt{n}}.$$

According to (2.1), this chain of implications gives

$$\begin{aligned}\mathbb{P}\{\lim_{k\to\infty} \tilde{\boldsymbol{\theta}}(k) \neq 0\} &\leq \mathbb{P}\{\forall \ell \in \mathbb{N},\ \exists i_\ell \text{ s.t. } |\eta_{i_\ell}(k_\ell)| > \frac{\upsilon}{\sqrt{n}}\} \\ &= \prod_{\ell=1}^{\infty} e^{-\frac{\upsilon}{\sqrt{n} b_{i_\ell}(k_\ell)}} = 0.\end{aligned}$$

The last equality holds because $\lim_{\ell\to\infty} b_{i_\ell}(k_\ell) = \lim_{\ell\to\infty} c_{i_\ell} q_{i_\ell}^{k_\ell} = 0$. Therefore, we conclude

$$\mathbb{P}\{\lim_{k\to\infty} \tilde{\boldsymbol{\theta}}(k) = 0\} = 1. \tag{3.16}$$

From (3.14), we see that a.s. convergence of $\boldsymbol{\theta}$ requires a.s. convergence of $\text{Ave}(\boldsymbol{\theta})$ as well. Left multiplying (3.9) by $\mathbf{1}_n^T$, we obtain for all $k \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned}\frac{1}{n}\mathbf{1}_n^T\boldsymbol{\theta}(k+1) &= \frac{1}{n}\mathbf{1}_n^T\boldsymbol{\theta}(k) + \frac{1}{n}\mathbf{1}_n^T\mathbf{S}\boldsymbol{\eta}(k) \\ &= \frac{1}{n}\mathbf{1}_n^T\boldsymbol{\theta}_0 + \frac{1}{n}\sum_{j=0}^{k}\sum_{i=1}^{n} s_i\eta_i(j),\end{aligned}$$

which in turn implies

$$\mathrm{Ave}(\boldsymbol{\theta}(k)) = \mathrm{Ave}(\boldsymbol{\theta}_0) + \sum_{i=1}^{n} \frac{s_i}{n} \sum_{j=0}^{k-1} \eta_i(j). \tag{3.17}$$

We next prove that $\mathrm{Ave}(\boldsymbol{\theta}(k))$ converges almost surely to θ_∞. For the latter to be well-defined over Ω, we simply set $\theta_\infty \triangleq \mathrm{Ave}(\boldsymbol{\theta}_0)$ when the series does not converge. Clearly, for any $\mathring{\boldsymbol{\eta}} \in \Omega$ such that $\sum_{j=0}^{\infty} \eta_i(j)$ converges for all $i \in \{1, \dots, n\}$, we have $\lim_{k\to\infty} \mathrm{Ave}(\boldsymbol{\theta}(k)) = \theta_\infty$. Hence, using (2.1),

$$\mathbb{P}\{\lim_{k\to\infty} \mathrm{Ave}(\boldsymbol{\theta}(k)) = \theta_\infty\} \geq \prod_{i=1}^{n} \mathbb{P}\Big\{ \sum_{j=0}^{\infty} \eta_i(j) \text{ converges} \Big\}.$$

Note that, for each $i \in \{1, \dots, n\}$ and any $\ell \in \mathbb{N}$, if $|\eta_i(j)| \leq \frac{1}{j^2}$ for all $j \geq \ell$, then $\sum_{j=0}^{\infty} \eta_i(j)$ converges. Hence, using (2.1) and the definition of Laplace distribution, we get for all $\ell \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\{\lim_{k\to\infty} \mathrm{Ave}(\boldsymbol{\theta}(k)) = \theta_\infty\} &\geq \prod_{i=1}^{n} \prod_{j=\ell}^{\infty} \mathbb{P}\Big\{ |\eta_i(j)| \leq \frac{1}{j^2} \Big\} \\ &= \prod_{i=1}^{n} \prod_{j=\ell}^{\infty} \Big(1 - e^{-\frac{1}{c_i q_i^j j^2}} \Big).\end{aligned}$$

For each $i \in \{1, \dots, n\}$, because $0 < q_i < 1$, there exists β_i such that $\frac{1}{c_i q_i^j j^2} \geq \beta_i j$ for $j \geq 1$. Therefore, using the Euler function φ,

$$\mathbb{P}\{\lim_{k\to\infty} \mathrm{Ave}(\boldsymbol{\theta}(k)) = \theta_\infty\} \geq \prod_{i=1}^{n} \frac{\varphi(e^{-\beta_i})}{\prod_{j=1}^{\ell-1}(1 - e^{-\beta_i j})},$$

for all $\ell \in \mathbb{N}$, and hence,

$$\mathbb{P}\{\lim_{k\to\infty} \mathrm{Ave}(\boldsymbol{\theta}(k)) = \theta_\infty\} \geq \lim_{\ell\to\infty} \prod_{i=1}^{n} \frac{\varphi(e^{-\beta_i})}{\prod_{j=1}^{\ell-1}(1-e^{-\beta_i j})} = 1.$$

This, together with (3.14) and (3.16), implies that $\mathbb{P}\{\lim_{k\to\infty} \boldsymbol{\theta}(k) = \theta_\infty \mathbf{1}_n\} = 1$, which completes the proof. □

Remark 3.4.3 (Mean-Square Convergence)**.** From (3.13) and the fact that $\|\mathbf{A}\| = 1$, we have

$$\begin{aligned}\|\boldsymbol{\theta}(k)\| &\leq \|\boldsymbol{\theta}_0\| + \|\mathbf{B}\| \sum_{j=0}^{k-1} \|\boldsymbol{\eta}(j)\| \\ &\leq \|\boldsymbol{\theta}_0\| + \|\mathbf{B}\| \sum_{j=0}^{\infty} \|\boldsymbol{\eta}(j)\| \triangleq Z,\end{aligned}$$

for all $k \in \mathbb{Z}_{\geq 0}$. It is straightforward to show $\mathbb{E}[Z^2] < \infty$, so, using Proposition 3.4.2, $\boldsymbol{\theta}(k)$ also converges to $\theta_\infty \mathbf{1}_n$ in mean square. □

Our next aim is to characterize the convergence rate of the distributed dynamics (3.9)-(3.11). Given the result in Proposition 3.4.2, we define the exponential mean-square convergence rate of the dynamics (3.9)-(3.11) as

$$\mu = \lim_{k\to\infty} \left(\sup_{\boldsymbol{\theta}(0)\in\mathbb{R}^n} \frac{\mathbb{E}\left[(\boldsymbol{\theta}(k) - \theta_\infty \mathbf{1}_n)^T (\boldsymbol{\theta}(k) - \theta_\infty \mathbf{1}_n)\right]}{\mathbb{E}\left[(\boldsymbol{\theta}(0) - \theta_\infty \mathbf{1}_n)^T (\boldsymbol{\theta}(0) - \theta_\infty \mathbf{1}_n)\right]} \right)^{\frac{1}{2k}}.$$

In the absence of noise ($\mathring{\boldsymbol{\eta}} = 0$), this definition coincides with the conventional exponential convergence rate of autonomous linear systems, see e.g., [1].

Proposition 3.4.4. ***(Convergence rate).*** *Under the hypotheses of Proposition 3.4.2, the exponential*

mean-square convergence rate of the distributed dynamics (3.9)-(3.11) *is*

$$\mu = \max\{\overline{q}, \overline{\lambda}\} \in (0, 1), \tag{3.18}$$

where $\overline{q} = \max_{1 \le i \le n} q_i$ *and* $\overline{\lambda} < 1$ *is the spectral radius of* $\mathbf{I}_n - h\mathbf{L} - \mathbf{\Pi}_n$.

Proof. For convenience, we let $\hat{\boldsymbol{\theta}}(k) = \boldsymbol{\theta}(k) - \theta_\infty \mathbf{1}_n$ denote the convergence error at $k \in \mathbb{Z}_{\ge 0}$ and $\hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}(0)$. Our first goal is to obtain an expression for $\mathbb{E}\left[\hat{\boldsymbol{\theta}}(k)^T \hat{\boldsymbol{\theta}}(k)\right]$. From (3.12) and the proof of Proposition 3.4.2, we have

$$\theta_\infty = \frac{1}{n}\mathbf{1}_n^T \boldsymbol{\theta}_0 + \frac{1}{n}\mathbf{1}_n^T \mathbf{S} \sum_{j=0}^{\infty} \boldsymbol{\eta}(j),$$

almost surely. Then, from (3.13), we have almost surely for all $k \in \mathbb{Z}_{\ge 0}$,

$$\hat{\boldsymbol{\theta}}(k) = \mathbf{A}^k \boldsymbol{\theta}_0 + \sum_{j=0}^{k-1} \mathbf{A}^{k-1-j} \mathbf{B} \boldsymbol{\eta}(j) - \mathbf{\Pi}_n \boldsymbol{\theta}_0 - \mathbf{\Pi}_n \mathbf{B} \sum_{j=0}^{\infty} \boldsymbol{\eta}(j),$$

where we have used the fact that $\mathbf{\Pi}_n \mathbf{S} = \mathbf{\Pi}_n \mathbf{B}$. Next, note that for all $k \in \mathbb{N}$,

$$\begin{aligned} (\mathbf{A} - \mathbf{\Pi}_n)^k &= \sum_{j=0}^{k} \binom{k}{j} (-\mathbf{\Pi}_n)^{k-j} \mathbf{A}^j \\ &= \mathbf{A}^k + \sum_{j=0}^{k-1} \binom{k}{j} (-1)^{k-j} \mathbf{\Pi}_n = \mathbf{A}^k - \mathbf{\Pi}_n, \end{aligned} \tag{3.19}$$

where we have used the facts that $\mathbf{\Pi}_n$ is idempotent and $\mathbf{\Pi}_n \mathbf{A}^j = \mathbf{\Pi}_n$ for any $j \in \mathbb{Z}_{\ge 0}$. Let $\boldsymbol{\mathcal{A}} = \mathbf{A} - \mathbf{\Pi}_n$. Notice that $\boldsymbol{\mathcal{A}}$ has spectral radius $\overline{\lambda} < 1$ and the same eigenvectors as $\mathbf{L}$. Then, using (3.19)

twice, we have almost surely for all $k \in \mathbb{N}$,

$$\hat{\theta}(k) = \mathcal{A}^k \theta_0 + \sum_{j=0}^{k-2} \mathcal{A}^{k-1-j} \mathbf{B}\boldsymbol{\eta}(j) + \mathbf{L}_{\text{cpt}} \mathbf{B}\boldsymbol{\eta}(k-1) - \sum_{j=k}^{\infty} \boldsymbol{\Pi}_n \mathbf{S}\boldsymbol{\eta}(j).$$

By the independence of $\{\boldsymbol{\eta}(j)\}_{j=0}^{\infty}$ over time, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}(k)^T \hat{\theta}(k)\right] = \theta_0^T \mathcal{A}^{2k} \theta_0 + \sum_{j=0}^{k-2} \mathbb{E}[\boldsymbol{\eta}(j)^T \mathbf{B} \mathcal{A}^{2k-2-2j} \mathbf{B}\boldsymbol{\eta}(j)] + \mathbb{E}[\boldsymbol{\eta}(k-1)^T \mathbf{B}\mathbf{L}_{\text{cpt}}^2 \mathbf{B}\boldsymbol{\eta}(k-1)] \\ + \sum_{j=k}^{\infty} \mathbb{E}[\boldsymbol{\eta}(j)^T \mathbf{S}\boldsymbol{\Pi}_n^2 \mathbf{S}\boldsymbol{\eta}(j)], \end{aligned} \tag{3.20}$$

for all $k \in \mathbb{N}$. Next, we upper bound the exponential mean-square convergence rate μ. Let $\overline{c} = \max_{1 \le i \le n} c_i$ and note that for any $\mathbf{N} \in \mathbb{R}^{n \times n}$ and any $j \in \mathbb{Z}_{\ge 0}$,

$$\begin{aligned}\mathbb{E}[\boldsymbol{\eta}(j)^T \mathbf{N}^T \mathbf{N}\boldsymbol{\eta}(j)] &= \sum_{i=1}^{n} 2b_i^2(j)(\mathbf{N}^T\mathbf{N})_{ii} \\ &\le 2\overline{c}^2 \overline{q}^{2j} \operatorname{tr}(\mathbf{N}^T\mathbf{N}) = 2\overline{c}^2 \overline{q}^{2j} \|\mathbf{N}\|_F^2,\end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Therefore,

$$\begin{aligned}\mathbb{E}\left[\hat{\theta}(k)^T \hat{\theta}(k)\right] \le \theta_0^T \mathcal{A}^{2k} \theta_0 + 2\overline{c}^2 \sum_{j=0}^{k-2} \overline{q}^{2j} \|\mathcal{A}^{k-1-j} \mathbf{B}\|_F^2 + 2\overline{c}^2 \overline{q}^{2(k-1)} \|\mathbf{L}_{\text{cpt}} \mathbf{B}\|_F^2 \\ + 2\overline{c}^2 \sum_{j=k}^{\infty} \overline{q}^{2j} \|\boldsymbol{\Pi}_n \mathbf{S}\|_F^2.\end{aligned}$$

Since the Frobenius norm is submultiplicative, $\|\mathbf{N}\|_F^2 \le n\|\mathbf{N}\|^2$ for any matrix $\mathbf{N}$, and $\|\mathcal{A}\| = \overline{\lambda}$,

we have

$$\mathbb{E}\left[\hat{\theta}(k)^T\hat{\theta}(k)\right] \leq \theta_0^T\mathcal{A}^{2k}\theta_0 + C_1\sum_{j=0}^{k-2}\overline{q}^{2j}\overline{\lambda}^{2k-4-2j} + C_2\overline{q}^{2k},$$

where $C_1 = 2n\overline{c}^2\|\mathbf{B}\|_F^2\overline{\lambda}^2$ and $C_2 = 2\overline{c}^2(\|\mathbf{L}_{\text{cpt}}\mathbf{B}\|_F^2/\overline{q}^2 + \|\mathbf{\Pi}_n\mathbf{S}\|_F^2/(1-\overline{q}^2))$ are constants. Note that for any $0 \leq j \leq k-2$, we have $\overline{q}^{2j}\overline{\lambda}^{2k-4-2j} \leq \max\{\overline{q},\overline{\lambda}\}^{2k-4}$. Therefore, using the fact that the supremum of a sum is less than or equal to the sum of suprema, we have

$$\sup_{\theta_0\in\mathbb{R}^n}\frac{\mathbb{E}\left[\hat{\theta}(k)^T\hat{\theta}(k)\right]}{\mathbb{E}\left[\hat{\theta}_0^T\hat{\theta}_0\right]} \leq \sup_{\theta_0\in\mathbb{R}^n}\frac{\theta_0^T\mathcal{A}^{2k}\theta_0}{\mathbb{E}\left[\hat{\theta}_0^T\hat{\theta}_0\right]} + \sup_{\theta_0\in\mathbb{R}^n}\frac{C_3(k-1)\max\{\overline{q},\overline{\lambda}\}^{2k} + C_2\overline{q}^{2k}}{\mathbb{E}\left[\hat{\theta}_0^T\hat{\theta}_0\right]},$$

where $C_3 = C_1\max\{\overline{q},\overline{\lambda}\}^{-4}$. Let $\tilde{\theta}_0 = \mathbf{L}_{\text{cpt}}\theta_0$ be the initial disagreement vector. It is straightforward to verify that $\theta_0^T\mathcal{A}^{2k}\theta_0 = \tilde{\theta}_0^T\mathcal{A}^{2k}\tilde{\theta}_0$ and

$$\mathbb{E}\left[\hat{\theta}_0^T\hat{\theta}_0\right] = \tilde{\theta}_0^T\tilde{\theta}_0 + \frac{1}{n}\sum_{i=1}^{n}\frac{2c_i^2s_i^2}{1-q_i^2} \triangleq \tilde{\theta}_0^T\tilde{\theta}_0 + C_4.$$

Therefore,

$$\begin{aligned}\sup_{\theta_0\in\mathbb{R}^n}\frac{\mathbb{E}\left[\hat{\theta}(k)^T\hat{\theta}(k)\right]}{\mathbb{E}\left[\hat{\theta}_0^T\hat{\theta}_0\right]} &\leq \sup_{\tilde{\theta}_0\in(\mathbb{R}\mathbf{1}_n)^\perp}\frac{\tilde{\theta}_0^T\mathcal{A}^{2k}\tilde{\theta}_0}{\tilde{\theta}_0^T\tilde{\theta}_0 + C_4} + \frac{C_3(k-1)\max\{\overline{q},\overline{\lambda}\}^{2k} + C_2\overline{q}^{2k}}{\inf_{\tilde{\theta}_0\in(\mathbb{R}\mathbf{1}_n)^\perp}(\tilde{\theta}_0^T\tilde{\theta}_0 + C_4)} \\ &= \overline{\lambda}^{2k} + C_3C_4^{-1}(k-1)\max\{\overline{q},\overline{\lambda}\}^{2k} + C_2C_4^{-1}\overline{q}^{2k}.\end{aligned}$$

By raising the right hand side of the above expression to the power $1/2k$ and taking the limit as $k \to \infty$, the constant/polynomial factors converge to 1 and the terms containing $\max\{\overline{q},\overline{\lambda}\}$ dominate the sum, proving that $\mu \leq \max\{\overline{q},\overline{\lambda}\}$. Similarly, we can lower bound μ as follows.

From (3.20), we have for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\hat{\theta}(k)^T \hat{\theta}(k)\right] \geq \theta_0^T \mathcal{A}^{2k} \theta_0 \Rightarrow \mu \geq \lim_{k\to\infty} \left(\sup_{\tilde{\theta}_0 \in (\mathbb{R}\mathbf{1}_n)^\perp} \frac{\tilde{\theta}_0^T \mathcal{A}^{2k} \tilde{\theta}_0}{\tilde{\theta}_0^T \tilde{\theta}_0 + C_4} \right)^{1/2k} = \overline{\lambda},$$

and

$$\begin{aligned} \mathbb{E}\left[\hat{\theta}(k)^T \hat{\theta}(k)\right] &\geq \mathbb{E}[\boldsymbol{\eta}(k)^T \mathbf{S} \boldsymbol{\Pi}_n^2 \mathbf{S} \boldsymbol{\eta}(k)] = \sum_{i=1}^{n} C_{5i} q_i^{2k} \\ \Rightarrow \mu &\geq \lim_{k\to\infty} \left(\sup_{\tilde{\theta}_0 \in (\mathbb{R}\mathbf{1}_n)^\perp} \frac{\sum_{i=1}^{n} C_{5i} q_i^{2k}}{\tilde{\theta}_0^T \tilde{\theta}_0 + C_4} \right)^{1/2k} = \overline{q}, \end{aligned}$$

where $C_{5i} = 2c_i^2 (\mathbf{S}\boldsymbol{\Pi}_n^2 \mathbf{S})_{ii}$ for all $i \in \{1, \ldots, n\}$. Therefore, $\mu \geq \max\{\overline{q}, \overline{\lambda}\}$, completing the proof.

□

Note that $\overline{\lambda}$ is the convergence rate of the noise-free (and non-private) Laplacian-based average consensus algorithm, while $\overline{q}$ is the worst-case decay rate of the noise sequence among the agents. From (3.18), the convergence rate μ is the larger of these two values, confirming our intuition that the slower rate among them is the bottleneck for convergence speed. Also, note that $\overline{\lambda}$ depends on the network topology $\mathcal{G}$ while $\overline{q}$ is independent of it.

3.4.2 Accuracy and Differential Privacy

Having established the convergence properties of the algorithm (3.9), here we characterize the extent to which our design solves Problem 1 by providing guarantees on its accuracy and differential privacy. The next result elaborates on the statistical properties of the agreement value.

Corollary 3.4.5 (Accuracy)**.** *Under the hypotheses of Proposition 3.4.2, the convergence point* θ_∞

is an unbiased estimate of $\mathrm{Ave}(\boldsymbol{\theta}_0)$ *with bounded dispersion,*

$$var\left\{\theta_\infty\right\} = \frac{2}{n^2}\sum_{i=1}^{n}\frac{s_i^2 c_i^2}{1-q_i^2}. \tag{3.21}$$

As a result, the algorithm (3.9)-(3.11) *is* $\left(p, \frac{1}{n}\sqrt{\frac{2}{p}\sum_{i=1}^{n}\frac{s_i^2 c_i^2}{1-q_i^2}}\right)$*-accurate for any* $p \in (0,1)$.

Proof. Since noises are independent over time and among agents, we deduce from (3.17) that for any $k \in \mathbb{Z}_{\geq 0}$, $E\{\mathrm{Ave}(\boldsymbol{\theta}(k))\} = \mathrm{Ave}(\boldsymbol{\theta}_0)$ and

$$\mathrm{var}\{\mathrm{Ave}(\boldsymbol{\theta}(k))\} = \frac{2}{n^2}\sum_{j=0}^{k}\sum_{i=1}^{n} s_i^2 c_i^2 q_i^{2j},$$

which establishes unbiasedness and bounded dispersion for any time. As $k \to \infty$, we get $E\{\theta_\infty\} = \mathrm{Ave}(\boldsymbol{\theta}_0)$ and

$$\mathrm{var}\{\theta_\infty\} = \frac{2}{n^2}\sum_{i=1}^{n}\frac{s_i^2 c_i^2}{1-q_i^2}.$$

The (p,r)-accuracy follows directly by applying Chebyshev's inequality (2.2) for $N = 1/\sqrt{p}$. □

Remark 3.4.6. ***(Comparison with the literature – cont'd).*** Proposition 3.4.2 and Corollary 3.4.5 establish almost sure convergence, with the expected value of convergence being the average of the agents' initial states. In contrast, the results in [13] establish convergence in mean square, and the expected value of convergence depends on the network topology. In both cases, the accuracy radius r decreases with the number of agents as $O(1/\sqrt{n})$. □

The expression for (p,r)-accuracy in Corollary 3.4.5 shows that one cannot obtain the ideal case of $(0,0)$-accuracy, and that r is a decreasing function of p, with $r \to \infty$ as $p \to 0$. This is an

(undesirable) consequence of the lack of preservation of the average under (3.9) due to the term $\mathbf{S}\boldsymbol{\eta}$. In turn, the presence of this expression helps establish the differential privacy of the algorithm with bounded, asymptotically vanishing noise, as we show next.

Proposition 3.4.7 (Differential Privacy)**.** *Under the hypotheses of Proposition 3.4.2, let*

$$\epsilon_i = \delta \frac{q_i}{c_i(q_i - |s_i - 1|)}, \tag{3.22}$$

for each $i \in \{1, \dots, n\}$, where δ is the adjacency bound in (3.3)*. Then, the algorithm preserves the ϵ_i-differential privacy of agent i's initial state for all $i \in \{1, \dots, n\}$. Consequently, the algorithm is ϵ-differential private with $\epsilon = \max_i \epsilon_i$.*

Proof. Consider any pair of δ-adjacent initial conditions $\boldsymbol{\theta}_0^{(1)}$ and $\boldsymbol{\theta}_0^{(2)}$ and an arbitrary set $\mathcal{O} \subset (\mathbb{R}^n)^{\mathbb{N}}$. For any $k \in \mathbb{Z}_{\geq 0}$, let

$$R_k^{(\ell)} = \{\mathring{\boldsymbol{\eta}}_k \in \Omega_k \mid X_{k,\boldsymbol{\theta}_0^{(\ell)}}(\mathring{\boldsymbol{\eta}}_k) \in \mathcal{O}_k\},\ \ell = 1, 2, \tag{3.23}$$

where $\Omega_k = \mathbb{R}^{n(k+1)}$ is the sample space up to time k, $X_{k,\boldsymbol{\theta}_0}$ is given in (3.5), and $\mathcal{O}_k \subseteq \mathbb{R}^{n(k+1)}$ is the set composed by truncating the elements of $\mathcal{O}$ to finite subsequences of length $k + 1$. Then, by the continuity of probability [25, Theorem 1.1.1.iv],

$$\mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \mid X_{\boldsymbol{\theta}_0^{(\ell)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\} = \lim_{k \to \infty} \int_{R_k^{(\ell)}} f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(\ell)}) d\mathring{\boldsymbol{\eta}}_k^{(\ell)}, \tag{3.24}$$

for $\ell = 1, 2$, where $f_{n(k+1)}$ is the $n(k+1)$-dimensional joint Laplace pdf given by

$$f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k) = \prod_{i=1}^{n} \prod_{j=0}^{k} \mathcal{L}(\eta_i(j); b_i(j)). \tag{3.25}$$

Next, we define a bijection between $R_k^{(1)}$ and $R_k^{(2)}$. Without loss of generality, assume $\theta_{0,i_0}^{(2)} = \theta_{0,i_0}^{(1)} + \delta_1$ for some $i_0 \in \{1, \ldots, n\}$, where $0 \leq \delta_1 \leq \delta$ and $\theta_{0,i}^{(2)} = \theta_{0,i}^{(1)}$ for all $i \neq i_0$. Then, for any $\mathring{\boldsymbol{\eta}}_k^{(1)} \in R_k^{(1)}$, define $\mathring{\boldsymbol{\eta}}_k^{(2)}$ by

$$\eta_i^{(2)}(j) = \begin{cases} \eta_i^{(1)}(j) - (1 - s_i)^j \delta_1, & \text{if } i = i_0, \\ \eta_i^{(1)}(j), & \text{if } i \neq i_0, \end{cases}$$

for $j \in \{0, \ldots, k\}$. It is not difficult to see that $X_{k,\boldsymbol{\theta}_0^{(1)}}(\mathring{\boldsymbol{\eta}}_k^{(1)}) = X_{k,\boldsymbol{\theta}_0^{(2)}}(\mathring{\boldsymbol{\eta}}_k^{(2)})$, so $\mathring{\boldsymbol{\eta}}_k^{(2)} \in R_k^{(2)}$. Since the converse argument is also true, the above defines a bijection. Therefore, for any $\mathring{\boldsymbol{\eta}}_k^{(2)} \in R_k^{(2)}$ there exists a unique $(\mathring{\boldsymbol{\eta}}_k^{(1)}, \Delta\mathring{\boldsymbol{\eta}}_k) \in R_k^{(1)} \times \mathbb{R}^{n(k+1)}$ such that

$$\mathring{\boldsymbol{\eta}}_k^{(2)} = \mathring{\boldsymbol{\eta}}_k^{(1)} + \Delta\mathring{\boldsymbol{\eta}}_k.$$

Note that $\Delta\mathring{\boldsymbol{\eta}}_k$ is fixed and does not depend on $\mathring{\boldsymbol{\eta}}_k^{(2)}$. Thus, we can use a change of variables to get

$$\mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \mid X_{\boldsymbol{\theta}_0^{(2)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\} = \lim_{k\to\infty} \int_{R_k^{(1)}} f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(1)} + \Delta\mathring{\boldsymbol{\eta}}_k) d\mathring{\boldsymbol{\eta}}_k^{(1)}. \tag{3.26}$$

Comparing (3.24) for $\ell = 1$ with (3.26), we see that both integrals are over $R_k^{(1)}$ with different

integrands. Dividing the integrands for any $\mathring{\boldsymbol{\eta}}_k^{(1)} \in R_k^{(1)}$ yields,

$$\begin{aligned}
\frac{f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(1)})}{f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(1)} + \Delta\mathring{\boldsymbol{\eta}}_k)} &= \frac{\prod_{i=1}^n \prod_{j=0}^k \mathcal{L}(\eta_i^{(1)}(j); b_i(j))}{\prod_{i=1}^n \prod_{j=0}^k \mathcal{L}(\eta_i^{(1)}(j) + \Delta\eta_i(j); b_i(j))} \\
&= \frac{\prod_{j=0}^k \mathcal{L}(\eta_{i_0}^{(1)}(j); b_{i_0}(j))}{\prod_{j=0}^k \mathcal{L}(\eta_{i_0}^{(1)}(j) + \Delta\eta_{i_0}(j); b_{i_0}(j))} \leq \prod_{j=0}^k e^{\frac{\left|\Delta\eta_{i_0}(j)\right|}{b_{i_0}(j)}} \leq e^{\sum_{j=0}^k \frac{|1-s_{i_0}|^j \delta}{c_{i_0} q_{i_0}^j}} \\
\Rightarrow f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(1)}) &\leq e^{\frac{\delta}{c_{i_0}} \sum_{j=0}^k \left(\frac{|1-s_{i_0}|}{q_{i_0}}\right)^j} f_{n(k+1)}(\mathring{\boldsymbol{\eta}}_k^{(1)} + \Delta\mathring{\boldsymbol{\eta}}_k).
\end{aligned}$$

Due to (3.11), the geometric series in the exponent of the multiplicative term is convergent. Therefore, integrating both sides over $R_k^{(1)}$ and letting $k \to \infty$, we have

$$\mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \mid X_{\theta_0^{(1)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\} \leq e^{\delta \frac{q_{i_0}}{c_{i_0}(q_{i_0} - |1-s_{i_0}|)}} \mathbb{P}\{\mathring{\boldsymbol{\eta}} \in \Omega \mid X_{\theta_0^{(2)}}(\mathring{\boldsymbol{\eta}}) \in \mathcal{O}\},$$

which establishes the ϵ_{i_0}-differential privacy for agent i_0. The fact the i_0 can be any agent establishes (3.22), while the last statement follows from Definition 3.2.1. □

Since the algorithm (3.9)-(3.11) converges almost surely (cf. Proposition 3.4.2) and is differentially private (cf. Proposition 3.4.7), Proposition 3.3.1 implies that it cannot achieve $(0, 0)$-accuracy, as noted above when discussing Corollary 3.4.5. The explicit privacy-accuracy trade-off is given by the relation between $\text{var}\{\theta_\infty\}$ and $\{\epsilon_i\}_{i=1}^n$, i.e., (c.f. (3.21), (3.22))

$$\text{var}\left\{\theta_\infty\right\} = \frac{2\delta^2}{n^2} \sum_{i=1}^n \frac{s_i^2 q_i^2}{\epsilon_i^2 (q_i - |s_i - 1|)^2 (1 - q_i^2)}, \tag{3.27}$$

so $\text{var}\left\{\theta_\infty\right\}$ increases as any ϵ_i is decreased and vice versa. We optimize this trade-off over $\{s_i, q_i\}_{i=1}^n$ in Section 3.4.3 and depict the optimal trade-off curve for a test network in Section 3.5.

Remark 3.4.8 (Laplacian Noise Distribution)**.** Even though the choice of Laplacian noise in (3.11) is not the only one that can be made to achieve differential privacy, it is predominant in the literature [8, 9]. The work [15] shows that Laplacian noise is optimal (among all possible distributions) in the sense that it minimizes the entropy of the transmitted messages while preserving differential privacy. □

Remark 3.4.9. ***(Comparison with the literature – cont'd).*** Proposition 3.4.7 guarantees the ϵ_i-differential privacy of agent i's initial state independently of the noise levels chosen by other agents. Therefore, each agent can choose its own level of privacy, and even opt not to add any noise to its messages, without affecting the privacy of other agents. In contrast, in [13], agents need to agree on the level of privacy before executing the algorithm. In both cases, privacy is achieved against an adversary that can hear everything, independently of how it processes the information. In contrast, the algorithm in [22, 23] assumes the adversary uses maximum likelihood estimation and only preserves the privacy of those agents who are sufficiently "far" from it in the graph (an agent is sufficiently far if the adversary cannot listen to it and all of its neighbors). The latter work uses a different notion of privacy based on the covariance of the maximum likelihood estimate which allows for guaranteed exact convergence, in the mean-square sense, to the true average. □

3.4.3 Optimal Noise Selection

In this section, we discuss the effect on the algorithm's performance of the free parameters present in our design. Given the trade-off between accuracy and privacy, cf. (3.27), we fix the privacy levels $\{\epsilon_i\}_{i=1}^n$ constant and study the best achievable accuracy of the algorithm as a function of the remaining free parameters. Each agent $i \in \{1, \dots, n\}$ gets to select the parameters s_i, c_i, q_i

determining the amount of noise introduced in the dynamics, with the constraint that $(s_i, c_i, q_i) \in \mathcal{P}$, where

$$\mathcal{P} = \{(s, c, q) \mid s \in (0, 2), c > 0, q \in (|s - 1|, 1)\}.$$

Given the characterization of accuracy in Corollary 3.4.5, we consider as cost function the variance of the agents' convergence point, i.e., θ_∞, around $\text{Ave}(\boldsymbol{\theta}_0)$, giving

$$J(\{s_i, c_i, q_i\}_{i=1}^n) = \frac{2}{n^2} \sum_{i=1}^n \frac{s_i^2 c_i^2}{1 - q_i^2}. \tag{3.28}$$

The next result characterizes its global minimization.

Proposition 3.4.10. ***(Optimal parameters for variance minimization).*** *For the adjacency bound $\delta > 0$ and privacy levels $\{\epsilon_i\}_{i=1}^n$ fixed, the optimal value of the variance of the agents' convergence point is*

$$J^* = \inf_{\{s_i, c_i, q_i\}_{i=1}^n \in \mathcal{P}^n} J(\{s_i, c_i, q_i\}_{i=1}^n) = \frac{2\delta^2}{n^2} \sum_{i=1}^n \frac{1}{\epsilon_i^2}.$$

The infimum is not attained over $\mathcal{P}^n$ but approached as

$$c_i = \delta \frac{q_i}{\epsilon_i (q_i - |s_i - 1|)} \quad , \quad s_i = 1, \tag{3.29}$$

and $q_i \to 0$ for all $i \in \{1, \dots, n\}$.

Proof. For each $i \in \{1, \dots, n\}$, with the privacy level fixed, the expression (3.29) follows directly

from (3.22). For convenience, we re-parameterize the noise decaying ratio q_i as

$$\alpha_i = \frac{q_i - |s_i - 1|}{1 - |s_i - 1|} \in (0, 1). \tag{3.30}$$

Note that $q_i = \alpha_i + (1 - \alpha_i)|s_i - 1|$. Substituting (3.29) and (3.30) into (3.28), we obtain (with a slight abuse of notation, we also use J to denote the resulting function),

$$\begin{aligned} J(\{s_i, \alpha_i\}_{i=1}^n) &= \frac{2}{n^2} \sum_{i=1}^n \frac{\delta^2}{\epsilon_i^2} \phi(\alpha_i, s_i), \\ \phi(\alpha, s) &= \frac{s^2(\alpha + (1-\alpha)|s-1|)^2}{\alpha^2(1 - |s-1|)^2\left[1 - (\alpha + (1-\alpha)|s-1|)^2\right]}. \end{aligned}$$

Therefore, to minimize J, each agent has to independently minimize the same function ϕ of its local parameters (α_i, s_i) over $D = (0, 1) \times (0, 2)$. Figure 3.1 illustrates the graph of this function over D.

Since D is not compact, the infimum might not be attained, and in fact, this is the case. It is easy verify that $\lim_{\alpha \to 0} \phi(\alpha, 1) = 1$. Now, for all $(\alpha, s) \in D$, $1 - (\alpha + (1 - \alpha)|s - 1|)^2 < 1$ so

$$\phi(\alpha, s) > \phi_1^2(\alpha, s), \ \phi_1(\alpha, s) = \frac{(\alpha + (1-\alpha)|s-1|)s}{\alpha(1 - |s-1|)}.$$

If $s \leq 1$, then $\phi_1(\alpha, s) = s + \frac{1-s}{\alpha} > 1$. If $s > 1$, then $\phi_1(\alpha, s) > 1 + \frac{s-1}{\alpha(2-s)} > 1$. Therefore, for all $(\alpha, s) \in D$, $\phi(\alpha, s) > 1$, which completes the proof. □

Given that differential privacy is resilient to post-processing, an alternative design strategy to preserve the differential privacy of agents' initial states is to inject noise only at the initial time, $k = 0$. From (3.11), the introduction of a one-shot noise by agent i corresponds to $q_i = 0$ which is

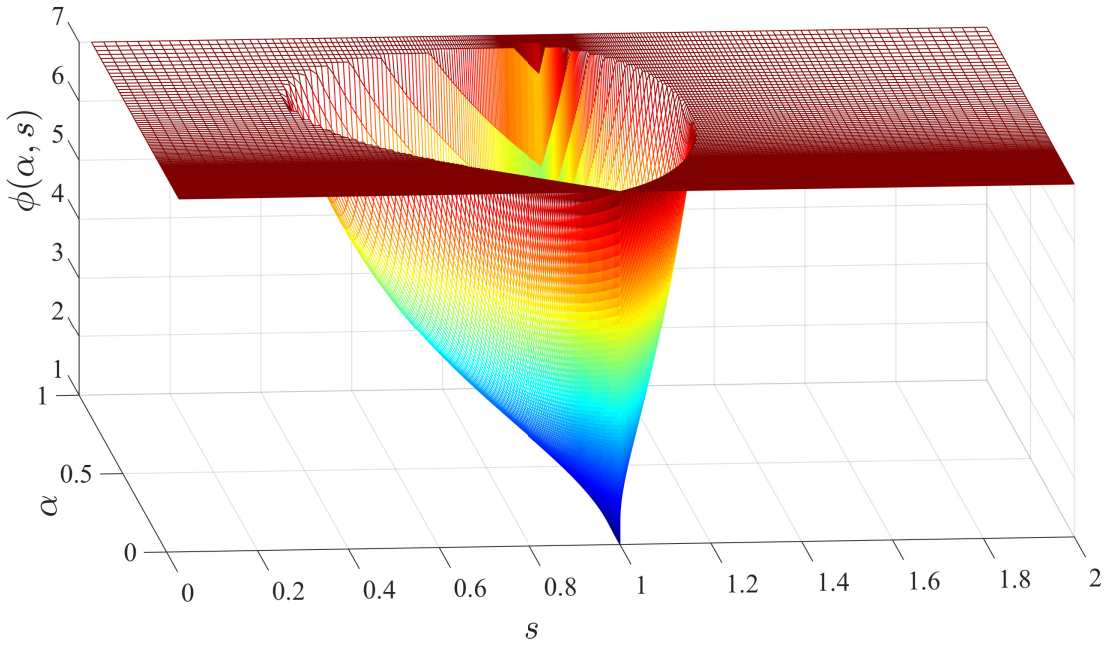


Figure 3.1: Local objective function ϕ of each agent as a function of its parameters. s is the noise-to-state gain and α is related to the noise decaying ratio. We cap the function values at 7 for visualization purposes. The function approaches its infimum as $\alpha \to 0$ while $s = 1$.

not feasible if $s_i \neq 1$. This can also be seen by rewriting (3.9) as

$$\theta(k+1) = (\mathbf{I}_n - \mathbf{S})\theta(k) + (\mathbf{S} - h\mathbf{L})\mathbf{x}(k),$$

so if $s_i \neq 1$ for any i, $\theta_i(k)$ directly (not only through $x_i(0)$) depend on $\theta_i(0)$. However, if $s_i = 1$, one can verify using a simplified version of the proof of Proposition 3.4.7 that $q_i = 0$ also preserves ϵ_i-differential privacy of $\theta_i(0)$ with $\epsilon_i = \frac{\delta}{c_i}$. This results in a cost of

$$J = \frac{2}{n^2}\sum_{i=1}^{n} c_i^2 = \frac{2\delta^2}{n^2}\sum_{i=1}^{n}\frac{1}{\epsilon_i^2} = J^*,$$

showing that the optimal accuracy is also achieved by one-shot perturbation of the initial state at time $k = 0$ and injection of no noise thereafter. A similar conclusion (that one-shot Laplace

perturbation minimizes the output entropy) can be drawn from [15], albeit this is not explicitly mentioned therein.

Remark 3.4.11. ***(Dynamic average consensus).*** In dynamic average consensus [26–28], agents seek to compute the average of individual exogenous, time-varying signals (the "static" average consensus considered here would be a special case corresponding to the exogenous signals being constant). In such scenarios, it is straightforward to show, using an argument similar to Proposition 3.3.1, that one-shot perturbation would no longer preserve the differential privacy of time-varying input signals. The reason is that in this case, there is a recurrent flow of information at each node whose privacy can no longer be preserved with one-shot perturbation. Sequential perturbation as in (3.10)-(3.11) is then necessary and the variance of the noise sequence has to dynamically depend on the rate of information flow to each node. Although the detailed design of such algorithms is beyond the scope of this work, such an algorithm can be designed following the idea of the sequential perturbation design of this work and the proof of its privacy in Proposition 3.4.7. To see this, note that (for $\mathbf{S} \neq \mathbf{I}_n$) we "tune" the amount of noise injection $\eta_i(k)$ so that the privacy of $(1-s_i)^k\theta_{0,i}$ is preserved at each round $k \geq 1$, but $(1-s_i)^k\theta_{0,i}$ is the amount of "retained information" of $\theta_{0,i}$ at round k and plays the same role as $u(k)$ in the dynamic average consensus problem. □

3.5 Simulations

In this section, we report simulation results of the distributed dynamics (3.9)-(3.11) on a network of $n = 50$ agents. Figure 3.2 shows the random graph used throughout the section, where edge weights are i.i.d. and each one equals a sum of two i.i.d. Bernoulli random variables with $p = 0.1$. The agents' initial states are also i.i.d. with distribution $\mathcal{N}(50, 100)$. As can

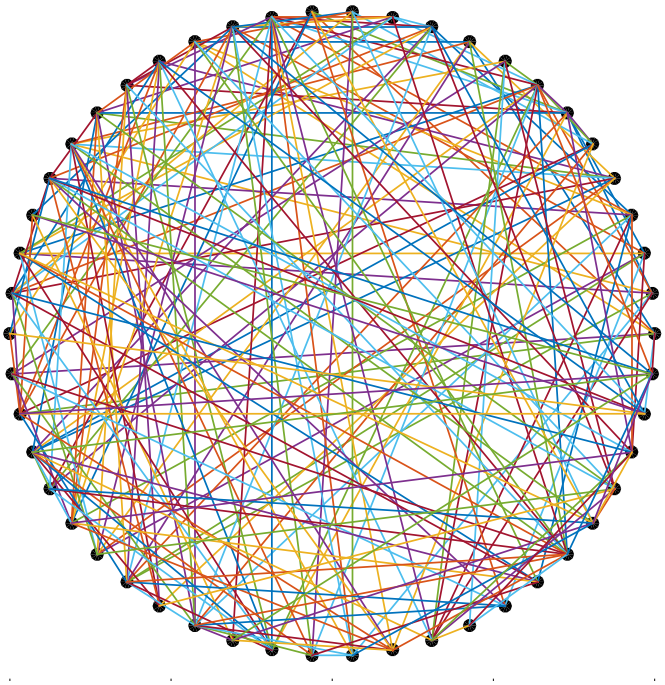
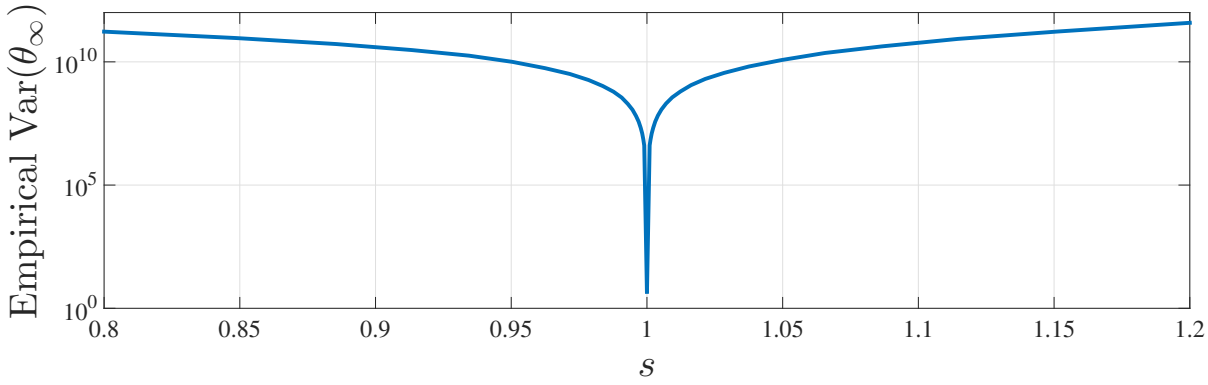


Figure 3.2: Random graph used for simulation.

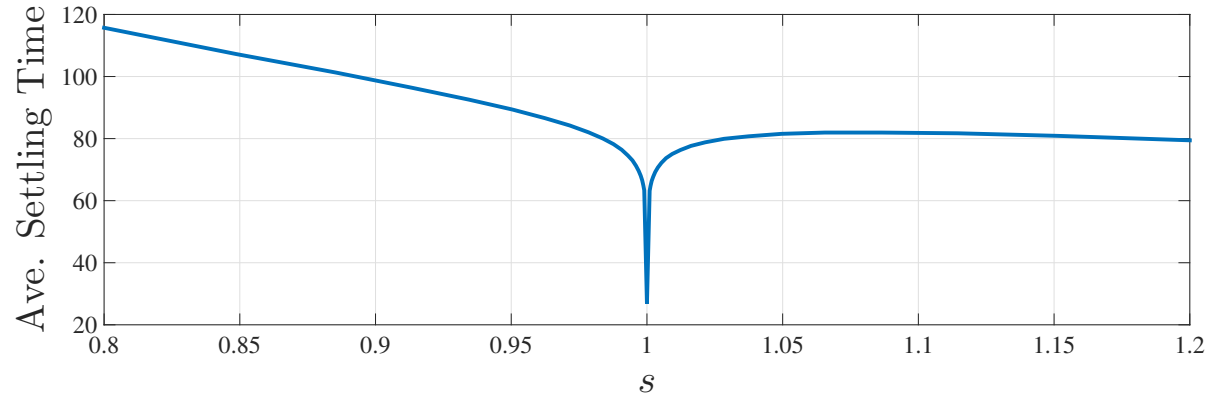
be seen from (3.21) and (3.22), neither accuracy nor privacy depend on the initial values or the communication topology (albeit according to (3.18) the convergence rate depends on the latter). In all the simulations, $\delta = 1$ and $c_i = \delta q_i / \epsilon_i (q_i - |s_i - 1|)$ for all $i \in \{1, \dots, n\}$.

Figure 3.3 depicts simulations with $\boldsymbol{\epsilon} = 0.1 \cdot \mathbf{1}_n$ and $\mathbf{S} = s\mathbf{I}_n$ while sweeping s over $[0.8, 1.2]$ with logarithmic step size. For each value of s, we set $q_i = \alpha_i + (1 - \alpha_i)|s - 1|$ with $\alpha_i = 10^{-6}$ for each $i \in \{1, \dots, n\}$ and repeat the simulation 10^4 times. For each run, to capture the statistical properties of the convergence point, the graph topology and initial conditions are the same and only noise realizations change. Figure 3.3(a) shows the empirical (sample) standard deviation of the convergence point as a function of s, verifying the optimality of one-shot perturbation. In particular, notice the sensitivity of the accuracy to s close to $s = 1$. Figure 3.3(b) shows the 'settling time', defined as the number of rounds until convergence (measured by a tolerance of 10^{-2}), as a function of s. The fastest convergence is achieved for $s = 1$, showing that one-shot noise is also optimal in

the sense of convergence speed. We have observed the same trends as in Figure 3.3 for different random choices of initial conditions and network topologies. Note that the settling time depends on both the convergence rate and the initial distance from the convergence point $\|\boldsymbol{\theta}(0) - \theta_\infty \mathbf{1}_n\|$. The former is constant at $\mu = \overline{\lambda} = 0.84$ for $s \in [0.8, 1.2]$. The latter depends on $\{c_i\}_{i=1}^n$, which in turn depend on s by (3.29). This explains the trend observed in Figure 3.3(b).



(a)

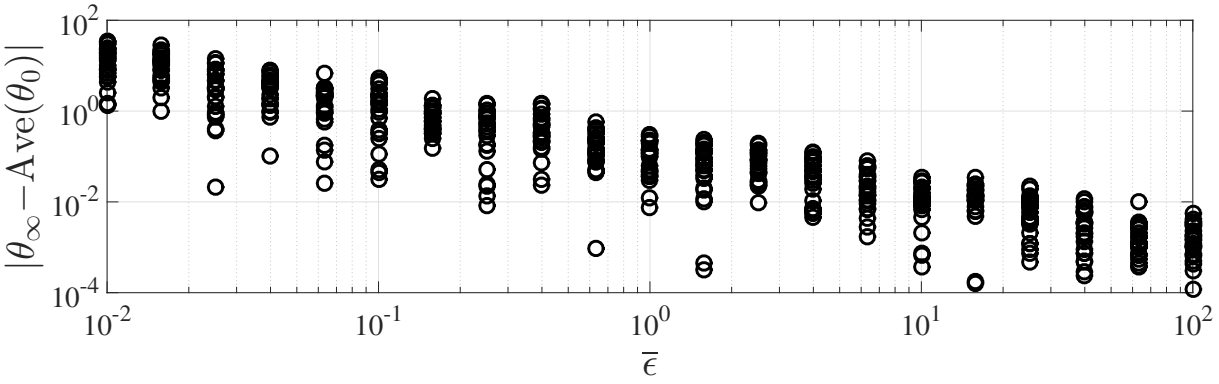


(b)

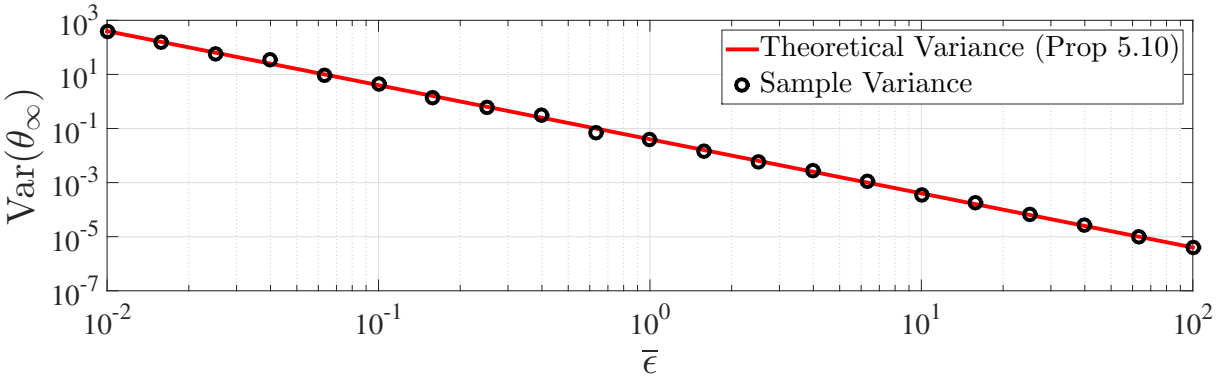
Figure 3.3: Executions of the algorithm (3.9)-(3.11) for random topology and initial conditions. (a) shows the empirical (i.e., sample) variance of the convergence point and (b) shows the settling time. The trend in (a) validates Proposition 3.4.10 while (b) shows the optimality of one-shot perturbation for convergence speed.

Figure 3.4 depicts the privacy-accuracy trade-off for the proposed algorithm. We have set $\mathbf{S} = \mathbf{I}_n$, $\mathbf{q} = \mathbf{0}_n$, and $\boldsymbol{\epsilon} = \overline{\epsilon}\mathbf{1}_n$ and then swept $\overline{\epsilon}$ logarithmically over $[10^{-2}, 10^2]$. In Figure 3.4(a), the algorithm is run 25 times for each value of the $\overline{\epsilon}$ and the error $|\theta_\infty - \text{Ave}(\boldsymbol{\theta}_0)|$ for each run is

plotted as a circle. In Figure 3.4(b), the sample variance of the convergence point θ_∞ is shown as a function of $\bar{\epsilon}$ together with the theoretical value given in Proposition 3.4.10. In both plots, we see an inversely-proportional relationship between accuracy and privacy, as expected.



(a)



(b)

Figure 3.4: The privacy-accuracy trade-off for the proposed algorithm (3.9)-(3.11) for random topology and initial conditions. (a) shows the norm of the error for 25 different realizations of the noise and (b) shows the sample variance over 100 noise realizations as well as the theoretical value provided by Proposition 3.4.10. The trend in both figures conforms with the theoretical characterization of θ_∞ given in Corollary 3.4.5.

Figure 3.5 shows the histogram of convergence points for 10^6 runs of the algorithm with $\epsilon = 0.1 \cdot \mathbf{1}_n$, $\mathbf{S} = \mathbf{I}_n$ and $\mathbf{q} = \mathbf{0}_n$ (optimal accuracy). The distribution of the convergence point is a bell-shaped curve with mean exactly at the true average, in accordance with Corollary 3.4.5.

Although the distribution of θ_∞ is provably non-Gaussian, the central limit theorem, see e.g., [25], implies that it is very close to Gaussian since the number of agents is large.

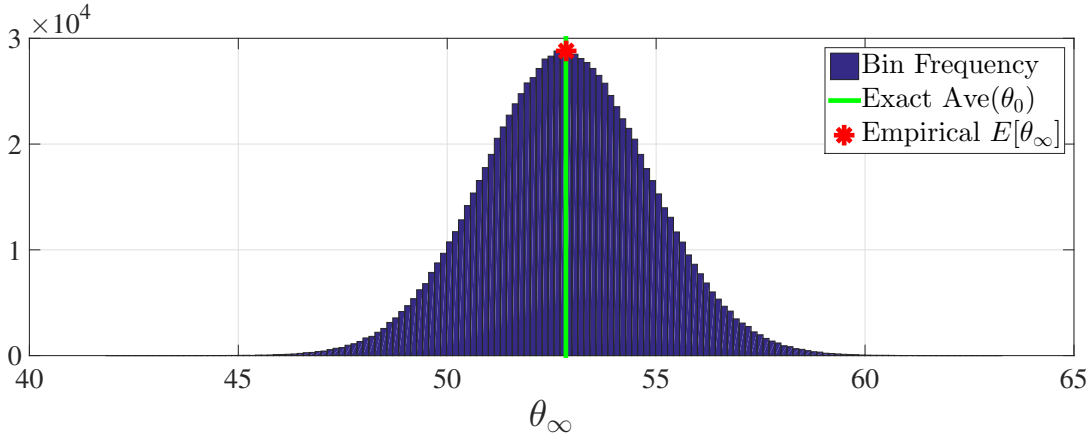


Figure 3.5: Statistical distribution of the convergence point. The sample mean (starred) matches the true average (green vertical line).

Finally, Figure 3.6 illustrates the convergence rate of the algorithm. Here, for $\boldsymbol{\epsilon} = 0.1 \cdot \mathbf{1}_n$, $\mathbf{S} = 0.9\mathbf{I}_n$, $\mathbf{q} = 0.2 \cdot \mathbf{1}_n$, and the same topology as in the previous plots, the initial agents states are randomly selected and the whole algorithm is run 100 times with different noise realizations $\mathring{\boldsymbol{\eta}}$, each time until 100 iterations. For each value of initial states and each $k \in \{1, \dots, 100\}$, we empirically approximate the quantity

$$\left(\frac{\mathbb{E}\left[(\boldsymbol{\theta}(k) - \theta_\infty \mathbf{1}_n)^T (\boldsymbol{\theta}(k) - \theta_\infty \mathbf{1}_n) \right]}{\mathbb{E}\left[(\boldsymbol{\theta}(0) - \theta_\infty \mathbf{1}_n)^T (\boldsymbol{\theta}(0) - \theta_\infty \mathbf{1}_n) \right]} \right)^{1/2k}$$

by taking the sample mean instead of the expectation in the numerator and denominator. We repeat this whole process 50 times for different random initial conditions and plot the result, together with the theoretical value of μ (which in this case equals $\overline{\lambda}$) given by Proposition 3.4.4. As Figure 3.6 shows, the supremum of the resulting curves converges to μ as $k \to \infty$, as expected.

Acknowledgements: This chapter is taken, in part, from the work published as "Differentially private average consensus: obstructions, trade-offs, and optimal algorithm design" by E. Nozari, P. Tallapragada, and J. Cortés in *Automatica*, vol. 81, pp. 221–231, 2017. The dis-

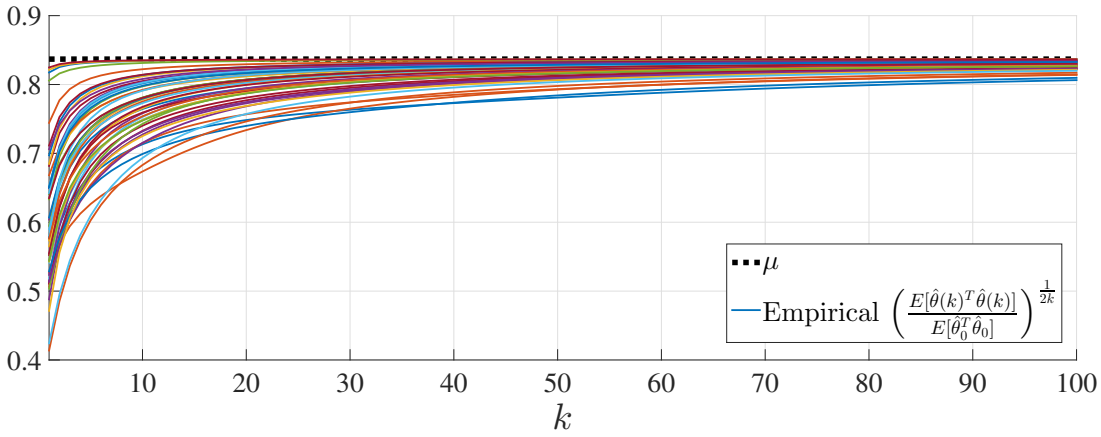


Figure 3.6: Illustration of the convergence rate of the algorithm (3.9)-(3.11). The limit of the supremum of the solid lines converges to the theoretical value of the exponential mean-square convergence rate μ given by Proposition 3.4.4. The curves with higher values correspond to initial states $\boldsymbol{\theta}_0$ that are closer to the eigenvector of $\mathbf{I}_n - h\mathbf{L} - \mathbf{\Pi}_n$ associated with $\overline{\lambda}$.

sertation author was the primary investigator and author of this paper.

Chapter Bibliography

[1] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, electronically available at http://coordinationbook.info.

[2] W. Ren and R. W. Beard, *Distributed Consensus in Multi-Vehicle Cooperative Control*, ser. Communications and Control Engineering. Springer, 2008.

[3] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*, ser. Applied Mathematics Series. Princeton University Press, 2010.

[4] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[5] D. Gündüz, E. Erkip, and H. Poor, "Source coding under secrecy constraints," in *Securing Wireless Communications at the Physical Layer*. Boston, MA: Springer US, 2010, pp. 173–199.

[6] A. Mukherjee, S. Fakoorian, J. Huang, and A. Swindlehurst, "Principles of physical layer security in multiuser wireless networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1550–1573, 2014.

[7] T. Tanaka and H. Sandberg, "SDP-based joint sensor and controller design for information-regularized optimal LQG control," in *IEEE Conf. on Decision and Control*, Osaka, 2015, pp. 4486–4491.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Theory of Cryptography Conference*, New York, NY, Mar. 2006, pp. 265–284.

[9] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, Venice, Italy, July 2006, pp. 1–12.

[10] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, Aug. 2014.

[11] P. Kairouz, S. Oh, and P. Viswanath, "Secure multi-party differential privacy," in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 2008–2016.

[12] M. Pettai and P. Laud, "Combining differential privacy and secure multiparty computation," in *Proceedings of the 31st Annual Computer Security Applications Conference*, ser. ACSAC 2015. ACM, 2015, pp. 421–430.

[13] Z. Huang, S. Mitra, and G. Dullerud, "Differentially private iterative synchronous consensus," in *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, New York, NY, 2012, pp. 81–90.

[14] Z. Huang, Y. Wang, S. Mitra, and G. E. Dullerud, "On the cost of differential privacy in distributed control systems," in *Proceedings of the 3rd International Conference on High Confidence Networked Systems (HiCoNS)*, Berlin, Germany, Apr. 2014, pp. 105–114.

[15] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, "Entropy-minimizing mechanism for differential privacy of discrete-time linear feedback systems," in *IEEE Conf. on Decision and Control*, Los Angeles, CA, Dec. 2014, pp. 2130–2135.

[16] J. L. Ny and G. J. Pappas, "Differentially private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2014.

[17] S. Han, U. Topcu, and G. J. Pappas, "Differentially private convex optimization with piecewise affine objectives," in *IEEE Conf. on Decision and Control*, Los Angeles, CA, Dec. 2014, pp. 2160–2166.

[18] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, Pilani, India, Jan. 2015.

[19] E. Nozari, P. Tallapragada, and J. Cortés, "Differentially private distributed convex optimization via functional perturbation," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, 2018.

[20] N. E. Manitara and C. N. Hadjicostis, "Privacy-preserving asymptotic average consensus," in *European Control Conference*, Zurich, Switzerland, 2013, pp. 760–765.

[21] M. Kefayati, M. S. Talebi, B. H. Khalaj, and H. R. Rabiee, "Secure consensus averaging in sensor networks using random offsets," in *IEEE Intern. Conf. on Telec., and Malaysia Intern. Conf. on Communications*, Penang, May 2007, pp. 556–560.

[22] Y. Mo and R. M. Murray, "Privacy preserving average consensus," in *IEEE Conf. on Decision and Control*, Los Angeles, CA, Dec. 2014, pp. 2154–2159.

[23] ——, "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, Feb 2017, submitted, available at http://yilinmo.github.io/public/papers/tac2014privacy.pdf.

[24] X. Duan, J. He, P. Cheng, Y. Mo, and J. Chen, "Privacy preserving maximum consensus," in *IEEE Conf. on Decision and Control*, Osaka, 2015, pp. 4517–4522.

[25] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[26] H. Bai, R. A. Freeman, and K. M. Lynch, "Robust dynamic average consensus of time-varying inputs," in *IEEE Conf. on Decision and Control*, Atlanta, GA, Dec. 2010, pp. 3104–3109.

[27] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2010.

[28] S. S. Kia, J. Cortés, and S. Martinez, "Dynamic average consensus under limited control authority and privacy requirements," *International Journal on Robust and Nonlinear Control*, vol. 25, no. 13, pp. 1941–1966, 2015.

Chapter 4

Differentially Private Distributed Optimization

In this chapter, we continue our design and analysis of privacy-aware distributed algorithms. In various application areas of networked dynamical systems, e.g., power networks, manufacturing systems, and smart transportation, the problem of optimizing the operation of a group of networked resources is a common and important task, where the individual objective functions associated to the entities, the estimates of the optimizer, or even the constraints on the optimization might reveal sensitive information. Our work here is motivated by the goal of synthesizing distributed coordination algorithms that accurately solve networked optimization problems with privacy guarantees.

We consider a class of distributed convex constrained optimization problems where a group of agents aim to minimize the sum of individual objective functions while each desires that any information about its objective function is kept private. We prove the impossibility of achieving differential privacy using strategies based on perturbing the inter-agent messages with noise when the underlying noise-free dynamics are asymptotically stable. This justifies our algorithmic solu-

tion based on the perturbation of individual functions with Laplace noise. To this end, we establish a general framework for differentially private handling of functional data. We further design post-processing steps that ensure the perturbed functions regain the smoothness and convexity properties of the original functions while preserving the differentially private guarantees of the functional perturbation step. This methodology allows us to use any distributed coordination algorithm to solve the optimization problem on the noisy functions. Finally, we explicitly bound the magnitude of the expected distance between the perturbed and true optimizers which leads to an upper bound on the privacy-accuracy trade-off curve. We end the chapter with numerical simulations that illustrate our results.

4.1 Prior Work

Our work builds upon the existing literature of distributed convex optimization and differential privacy. In the area of networked systems, an increasing body of research, e.g., [1–6] and references therein, designs and analyzes algorithms for distributed convex optimization both in discrete and continuous time as well as in deterministic and stochastic scenarios. While these works consider an ambitious suite of topics related to convergence and performance under various constraints imposed by real-world applications, privacy is an aspect generally absent in their treatment. The concept of differential privacy [7, 8] was originally proposed for databases of individual records subject to public queries and has been extended to several areas thereafter. The recent work [9] provides a comprehensive recent account of this area.

In machine learning, the problem of differentially private optimization has received attention, see e.g. [10–16], as an intermediate, usually centralized, step for solving other learning or

statistical tasks. The common paradigm is having the sensitive information correspond to the entries of a finite database of records or training data that usually constitute the parameters of an additive objective function. Threat models are varied, including releasing to the adversary the whole sequence of internal states of the optimization process or only the algorithm's final output. The work [10] designs a differentially private classifier by perturbing the objective function with a linear finite-dimensional function (hyper-plane). The work [11] shows that this method works also in the presence of constraints and non-differentiable regularizers. Although this is sufficient to preserve the privacy of the underlying finite-dimensional parameter set (learning samples), it cannot keep the whole objective functions private. The work [12] designs a sensitivity-based differentially private algorithm for regression analysis which, instead of perturbing the optimal weight vector, perturbs the regression cost function by injecting noise into the coefficients of the quadratic truncation of its Taylor expansion. This truncation limits the functional space to the (finite-dimensional) space of quadratic functions. The work [13] proposes the addition of a sample path of a Gaussian random process to the objective function, but does not explore the generalization to arbitrary dimensions or ensures the smoothness and convexity of the resulting function. In general, the proposed algorithms are not distributed and neither designed for nor capable of preserving the privacy of infinite-dimensional objective functions. Furthermore, the work in this area does not rigorously study the effect of added noise on the global optimizer or on the smoothness and convexity properties of the objective functions. In addition to addressing these issues, the present treatment is applicable to scenarios where the sensitive information consists of objective functions coming from the (infinite-dimensional) space of L_2 functions.

Of more relevance to our work are recent papers [17–19] that study differentially private distributed optimization problems for multi-agent systems. These papers consider as private infor-

mation, respectively, the objective functions, the optimization constraints, and the agents' states. The underlying commonality is the algorithm design approach based on the idea of message perturbation. This idea consists of adopting a standard distributed optimization algorithm and modifying it by having agents perturb the messages to their neighbors or a central coordinator with Laplace or Gaussian noise. This approach has the advantage of working with the original objective functions and thus is easy to implement. However, for fixed design parameters, the algorithm's output does not correspond to the true optimizer in the absence of noise, suggesting the presence of a steady-state accuracy error. The work [18] addresses this problem by terminating the algorithm after a finite number of steps, and optimizing this number offline as a function of the desired level of privacy. Nevertheless, for any fixed level of privacy, there exists an amount of bias in the algorithm's output which is not due to the added noise but to the lack of asymptotic stability of the underlying noiseless dynamics. To address this issue, our approach explores the use of functional perturbation to achieve differential privacy. The concept of functional differential privacy combines the benefits of metrics and adjacency relations. The work [20] also employs metrics instead of binary adjacency relations in the context of differential privacy. This approach has the advantage that the difference between the probabilities of events corresponding to any pair of data sets is bounded by a function of the distance between the data sets, eliminating the need for the computation of conservative sensitivity bounds.

4.2 Problem Statement

Consider a group of n agents whose communication topology is described by a digraph $\mathcal{G}$. Each agent $i \in \{1, \dots, n\}$ has a local objective function $f_i : D \to \mathbb{R}$, where $D \subset \mathbb{R}^d$ is convex and

compact and has nonempty interior. We assume that each $f_i, i \in \{1, \dots, n\}$ is convex and twice continuously differentiable, and use the shorthand notation $F = \{f_i\}_{i=1}^n$. Consider the following convex optimization problem

$$\begin{aligned} \min_{\mathbf{x} \in D} \quad & f(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(\mathbf{x}) \\ \text{s.t.} \quad & G(\mathbf{x}) \leq \mathbf{0}, \\ & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

where the component functions of $G : D \to \mathbb{R}^m$ are convex, $\mathbf{A} \in \mathbb{R}^{s \times d}$, and $\mathbf{b} \in \mathbb{R}^s$. Denote by $X \subseteq D$ the feasibility set. The optimization problem can be equivalently written as,

$$\min_{\mathbf{x} \in X} \ f(\mathbf{x}). \tag{4.1}$$

We assume that X is a global piece of information known to all agents.

The group objective is to solve the convex optimization problem (4.1) in a distributed and private way. By distributed, we mean that each agent can only interact with its neighbors in the graph $\mathcal{G}$. Regarding privacy, we consider the case where the function f_i (or some of its attributes) constitute the local and sensitive information known to agent $i \in \{1, \dots, n\}$ that has to be kept confidential. Each agent assumes that the adversary has access to all the "external" information (including all the network communications and all other objective functions). This setting is sometimes called local (differential) privacy in the literature, see e.g., [21]. In order to properly define privacy, let us first introduce the notion of adjacency. Given any normed vector space $(\mathcal{V}, \|\cdot\|_{\mathcal{V}})$ with $\mathcal{V} \subseteq L_2(D)$, two sets of functions $F, F' \subset L_2(D)$ are $\mathcal{V}$-adjacent if there exists $i_0 \in \{1, \dots, n\}$

such that

$$f_i = f'_i,\ i \neq i_0 \quad \text{and} \quad f_{i_0} - f'_{i_0} \in \mathcal{V}.$$

The set $\mathcal{V}$ is a design choice that we specify later in Section 4.4.2. Moreover, this definition can be readily extended to the case where $\mathcal{V}$ is any subset of another normed vector space $\mathcal{W} \subseteq L_2(D)$. With this generalization, the conventional bounded-difference notion of adjacency (also used in Chapter 3) becomes a special case of the definition above, where $\mathcal{V}$ is a closed ball around the origin. We provide next a more general definition of differential privacy for a map.

Definition 4.2.1. ***(Differential privacy).*** Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and consider a random map

$$\mathcal{M} : L_2(D)^n \times \Omega \to \mathcal{X}$$

from the function space $L_2(D)^n$ to an arbitrary set $\mathcal{X}$. Given $\boldsymbol{\epsilon} \in \mathbb{R}^n_{>0}$, the map $\mathcal{M}$ is $\boldsymbol{\epsilon}$-differentially private if, for any two $\mathcal{V}$-adjacent sets of functions F and F' that (at most) differ in their i_0'th element and any set $\mathcal{O} \subseteq \mathcal{X}$, one has

$$\mathbb{P}\{\omega \in \Omega \mid \mathcal{M}(F', \omega) \in \mathcal{O}\} \leq e^{\epsilon_{i_0} \| f_{i_0} - f'_{i_0} \|_{\mathcal{V}}} \mathbb{P}\{\omega \in \Omega \mid \mathcal{M}(F, \omega) \in \mathcal{O}\}. \tag{4.2}$$

□

Essentially, this notion requires the statistics of the output of $\mathcal{M}$ to change only (relatively) slightly if the objective function of one agent changes (and the change is in $\mathcal{V}$), making it hard to

an "adversary" that observes the output of $\mathcal{M}$ to determine this change. In the case of an iterative asymptotic distributed optimization algorithm, one should think of $\mathcal{M}$ as representing the action (observed by the adversary) of the algorithm on the set of local functions F. In other words, $\mathcal{M}$ is the map (parameterized by the initial network condition) that assigns to F the whole sequence of messages transmitted over the network. In this case, (4.2) has to hold for all allowable values of the initial conditions. We are ready to formally state the network objective.

Problem 2. ***(Differentially private distributed optimization).*** *Design a distributed and differentially private optimization algorithm whose guarantee on accuracy improves as the level of privacy decreases, leading to the exact optimizer of the aggregate objective function in the absence of privacy.* □

The reason for the requirement of recovering the exact optimizer in the absence of privacy in Problem 2 is the following. It is well-known in the literature of differential privacy that there always exists a cost for an algorithm to be differentially private, i.e., the algorithm inevitably suffers a performance loss that increases as the level of privacy increases. This phenomenon is a result of the noise added in the map $\mathcal{M}$, whose variance increases as ϵ decreases. With this requirement on the noise-free behavior of the algorithm, we aim to make sure that the cause of this performance loss is *only* due to the added noise and not to any other factor.

Example 4.2.2. ***(Linear classification with logistic loss function).*** We introduce here a supervised classification problem that will serve to illustrate the discussion along the chapter. Consider a database of training records composed by the labeled samples $\{(\mathbf{a}_i, b_i)\}_{i=1}^N$, where each $\mathbf{a}_i \in \mathbb{R}^d$ (containing the features of a corresponding object) may belong to one of two possible classes and $b_i \in \{-1, 1\}$ determines to which class it belongs. The goal is to train a classifier with the samples

so that it can automatically classify future unlabeled samples. For simplicity, we let $d = 2$ and assume $\mathbf{a}_i \in [0,1]^2$ and $b_i \in \{-1, 1\}$ are independently and uniformly randomly selected. The aim is to find the best hyperplane $\mathbf{x}^T\mathbf{a}$ that can separate the two classes. The parameters $\mathbf{x}$ defining the hyperplane can be found by solving the convex problem,

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in X} \sum_{i=1}^{N} \left(\ell(\mathbf{x}; \mathbf{a}_i, b_i) + \frac{\lambda}{2}\|\mathbf{x}\|^2 \right), \tag{4.3}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}_{>0}$ is the loss function and $(\lambda/2)\|\mathbf{x}\|^2$ is the regularizing term. Since the objective function is strongly convex, we choose X large enough so that $\mathbf{x}^*$ is the same as the unique unconstrained minimizer. Popular choices of ℓ are the logistic loss $\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \ln(1 + e^{-b_i \mathbf{a}_i^T \mathbf{x}})$ and the hinge loss $\ell(\mathbf{x}; \mathbf{a}_i, b_i) = \max\{0, 1 - b_i \mathbf{a}_i^T \mathbf{x}\}$. We focus on the logistic loss here due to its smoothness.

Consider a group of n agents, each one owning a portion $N_d = N/n$ of the training samples, who seek to collectively solve (4.3) in a distributed fashion, i.e., only through communication with their neighbors (without a central aggregator). Various iterative algorithms have been proposed in the literature, cf. [2–6], to address this problem formulation. As an example, [2] proposes to have each agent $i \in \{1, \dots, n\}$ start with an initial estimate $\mathbf{x}_i(0)$ of $\mathbf{x}^*$ and, at each iteration k, update its estimate as

$$\mathbf{x}_i(k+1) = \text{proj}_X(\mathbf{z}_i(k) - \alpha_k \nabla f_i(\mathbf{z}_i(k))), \tag{4.4a}$$

$$\mathbf{z}_i(k) = \sum_{j=1}^{n} a_{ij} \mathbf{x}_j(k), \tag{4.4b}$$

where $\{a_{ij}\}_{j=1}^n$ are the edge weights of the communication graph at node i and α_k is the stepsize.

From (4.4b), one can see that agents only need to share their estimates with their neighbors to run the algorithm. Under reasonable connectivity assumptions, one can show [2] that $\mathbf{x}_i(k)$ converges to $\mathbf{x}^*$ asymptotically if the sequence of stepsizes is square-summable ($\sum_k \alpha_k^2 < \infty$) but not summable ($\sum_k \alpha_k = \infty$). We are interested in endowing distributed coordination algorithms such as this with privacy guarantees so that their execution does not reveal information about the local objective functions to the adversary. □

4.3 Rationale for Design Strategy

In this section, we discuss two algorithm design strategies to solve Problem 2 based on the perturbation of either inter-agent messages or the local objective functions. We point out an important limitation of the former, and this provides justification for the ensuing design of our objective-perturbing algorithm based on functional differential privacy.

4.3.1 Limitations of Message-Perturbing Strategies

We use the term *message-perturbing strategy* to refer to the result of modifying any of the distributed optimization algorithms available in the literature by adding (Gaussian or Laplace) noise to the messages agents send to either neighbors or a central aggregator in order to preserve privacy. A generic message-perturbing distributed algorithm takes the form

$$\begin{aligned} \mathbf{x}(k+1) &= a_{\mathcal{I}}(\mathbf{x}(k), \boldsymbol{\xi}(k)), \\ \boldsymbol{\xi}(k) &= \mathbf{x}(k) + \boldsymbol{\eta}(k), \end{aligned} \tag{4.5}$$

where $\mathring{\boldsymbol{\xi}}, \mathring{\boldsymbol{\eta}} : \mathbb{Z}_{\geq 0} \to \mathbb{R}^n$ represent the sequences of messages and perturbations, respectively, and $a_{\mathcal{I}} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ depends on the agents' sensitive information set $\mathcal{I}$ with associated optimizer $\mathbf{x}^*_{\mathcal{I}}$. This formulation is quite general and can also encode algorithmic solutions for optimization problems other than the one in Section 4.2, such as the ones studied in [18, 19]). In the problem of interest here, $\mathcal{I} = F = \{f_i\}_{i=1}^n$.

The following result provides conditions on the noise variance that ensure that the noise vanishes asymptotically almost surely and remains bounded with nonzero probability.

Lemma 4.3.1. ***(Convergence and boundedness of Laplace and normal random sequences with decaying variance).*** *Let $\mathring{\eta}$ be a sequence of independent random variables defined over the sample space $\Omega = \mathbb{R}^{\mathbb{N}}$, with $\eta(k) \sim Lap(b(k))$ or $\eta(k) \sim \mathcal{N}(0, b(k))$ for all $k \in \mathbb{N}$. Given $r > 0$, consider the events*

$$E = \{\mathring{\eta} \in \Omega \mid \lim_{k\to\infty} \eta(k) = 0\},$$

$$F_r = \{\mathring{\eta} \in \Omega \mid \forall k \in \mathbb{N} \quad |\eta(k)| \leq r\}.$$

If $b(k)$ is $O(\frac{1}{k^p})$ for some $p > 0$, then $\mathbb{P}(E) = 1$ and $\mathbb{P}(F_r) = \mathbb{P}(F_r \cap E) > 0$ for all $r > 0$.

Proof. First, consider the case where $\eta(k) \sim \text{Lap}(b(k))$. By the independence of the random variables and the fact that $|\eta(k)|$ is exponentially distributed with rate $\frac{1}{b(k)}$,

$$\mathbb{P}(F_r) = \prod_{k=1}^{\infty} \left(1 - e^{-\frac{r}{b(k)}}\right).$$

By assumption, $b(k) \leq \frac{c}{k^p}$ for all $k \in \mathbb{N}$ and some $p, c > 0$. Therefore,

$$\mathbb{P}(F_r) \geq \prod_{k=1}^{\infty} \left(1 - e^{-\frac{r}{c}k^p}\right) > 0,$$

because the series $\sum_{k=1}^{\infty} e^{-\frac{r}{c}k^p}$ converges [22, §1.14]. Next, let $E_{\ell,K} = \{\mathring{\eta} \in \Omega \mid \forall k \geq K \quad |\eta(k)| < \upsilon_\ell\}$ where $\{\upsilon_\ell\}_{\ell=1}^{\infty}$ is a monotonically decreasing sequence that converges to zero as $\ell \to \infty$ (e.g., $\upsilon_\ell = \frac{1}{\ell}$). Note that

$$E = \bigcap_{\ell=1}^{\infty} \bigcup_{K=1}^{\infty} E_{\ell,K}.$$

$E_{\ell,K} \uparrow \bigcup_{K=1}^{\infty} E_{\ell,K}$ for all $\ell \in \mathbb{N}$ as $K \to \infty$, and $\bigcup_{K=1}^{\infty} E_{\ell,K} \downarrow E$ as $\ell \to \infty$. Therefore,

$$\begin{aligned}\mathbb{P}(E) &= \lim_{\ell\to\infty} \lim_{K\to\infty} \mathbb{P}(E_{\ell,K}) = \lim_{\ell\to\infty} \lim_{K\to\infty} \prod_{k=K}^{\infty} \left(1 - e^{-\frac{\upsilon_\ell}{b(k)}}\right) \\ &\geq \lim_{\ell\to\infty} \lim_{K\to\infty} \prod_{k=K}^{\infty} \left(1 - e^{-\frac{\upsilon_\ell}{c}k^p}\right) = 1.\end{aligned}$$

Then, $\mathbb{P}(F_r \cap E) = \mathbb{P}(F_r) - \mathbb{P}(F_r \cap E^c) = \mathbb{P}(F_r) > 0$. For the case of normal distribution of random variables,

$$\mathbb{P}\{|\eta(k)| \leq r\} = \operatorname{erf}\left(\frac{r}{\sqrt{2b(k)}}\right) \geq 1 - e^{-\frac{r^2}{2b(k)}},$$

and the results follows from the arguments above. □

Note that Lemma 4.3.1 also ensures that the probability that the noise simultaneously converges to zero and remains bounded is nonzero. One might expect that Lemma 4.3.1 would hold if

$b(k) \to 0$ at any rate. However, this is not true. For instance, if $b(k) = \frac{1}{\log k}$, one can show that the probability that $\eta(k)$ eventually remains bounded is zero for any bound $r \leq 1$, so the probability that $\eta(k) \to 0$ is zero as well.

The following result shows that a message-perturbing algorithm of the form (4.5) cannot achieve differential privacy if the underlying (noise-free) dynamics are asymptotically stable. For convenience, we employ the short-hand notation $\tilde{a}_{\mathcal{I}}(\mathbf{x}(k), \boldsymbol{\eta}(k)) = a_{\mathcal{I}}(\mathbf{x}(k), \mathbf{x}(k) + \boldsymbol{\eta}(k))$ to refer to (4.5).

Proposition 4.3.2. ***(Impossibility result for 0-LAS message-perturbing algorithms).*** *Consider the dynamics* (4.5) *with either* $\eta_i(k) \sim Lap(b_i(k))$ *or* $\eta_i(k) \sim \mathcal{N}(0, b_i(k))$. *If* $\tilde{a}_{\mathcal{I}}$ *is LISS relative to* $\mathbf{x}^*_{\mathcal{I}}$ *for two information sets* $\mathcal{I}$ *and* $\mathcal{I}'$ *with different optimizers* $\mathbf{x}^*_{\mathcal{I}} \neq \mathbf{x}^*_{\mathcal{I}'}$ *and associated robust stability radii* ρ *and* ρ', *respectively,* $b_i(k)$ *is* $O(\frac{1}{k^p})$ *for all* $i \in \{1, \dots, n\}$ *and some* $p > 0$, *and at least one of the following holds,*

(i) $\mathbf{x}^*_{\mathcal{I}}$ *is not an equilibrium point of* $\mathbf{x}(k+1) = \tilde{a}_{\mathcal{I}'}(\mathbf{x}(k), \mathbf{0})$ *and* $\tilde{a}_{\mathcal{I}'}$ *is continuous,*

(ii) $\mathbf{x}^*_{\mathcal{I}}$ *belongs to the interior of* $B(\mathbf{x}^*_{\mathcal{I}'}, \rho')$,

then no algorithm of the form (4.5) *can preserve the* ϵ*-differentially privacy of the information set* $\mathcal{I}$ *for any* $\epsilon > \mathbf{0}$.

Proof. Our proof strategy consists of establishing that, if the initial state is close to the equilibrium of the system for one information set, the state trajectory converges to that equilibrium with positive probability but to the equilibrium of the system with the other information set with probability zero. We then use this fact to rule out differential privacy. For any fixed initial state $\mathbf{x}_0$, if either of $\mathring{\boldsymbol{\xi}}$ or $\mathring{\boldsymbol{\eta}}$ is known, the other one can be uniquely determined from (4.5). Therefore, the mapping

$\Xi_{\mathcal{I},\mathbf{x}_0} : (\mathbb{R}^n)^{\mathbb{N}} \to (\mathbb{R}^n)^{\mathbb{N}}$ such that

$$\Xi_{\mathcal{I},\mathbf{x}_0}(\mathring{\boldsymbol{\eta}}) = \mathring{\boldsymbol{\xi}}$$

is well-defined and bijective. Let $\kappa, \kappa' \in \mathcal{K}$ be as in (2.4) corresponding to $\tilde{a}_{\mathcal{I}}$ and $\tilde{a}_{\mathcal{I}'}$, respectively. Consider as initial condition $\mathbf{x}_0 = \mathbf{x}^*_{\mathcal{I}}$ and define

$$R = \left\{ \mathring{\boldsymbol{\eta}} \in \Omega \mid \forall i \in \{1, \dots, n\},\ \lim_{k\to\infty} \eta_i(k) = 0 \text{ and } |\eta_i(k)| \leq \min\left\{\kappa^{-1}(\rho), \rho\right\},\ \forall k \in \mathbb{N} \right\}.$$

By Lemma 4.3.1, we have $\mathbb{P}(R) > 0$. By Proposition 2.6.1, since $\|\mathbf{x}_0 - \mathbf{x}^*_{\mathcal{I}}\| = 0 \leq \rho$ and $\|\mathring{\boldsymbol{\eta}}\|_\infty \leq \min\left\{\kappa^{-1}(\rho), \rho\right\}$ for all $\mathring{\boldsymbol{\eta}} \in R$, the sequence $\Xi_{\mathcal{I},\mathbf{x}_0}(\mathring{\boldsymbol{\eta}})$ converges to $\mathbf{x}^*_{\mathcal{I}}$. Let $\mathcal{O} = \Xi_{\mathcal{I},\mathbf{x}_0}(R)$ and $R' = \Xi^{-1}_{\mathcal{I}',\mathbf{x}_0}(\mathcal{O})$ (where we are using the forward and reverse images of sets, respectively). Next, we show that no $\mathring{\boldsymbol{\eta}}' \in R'$ converges to $\mathbf{0}$ under either hypothesis (i) or (ii) of the statement. Under (i), there exists a neighborhood of $(\mathbf{x}^*_{\mathcal{I}}, \mathbf{0}) \in \mathbb{R}^{2n}$ in which the infimum of the absolute value of at least one of the components of $\tilde{a}_{\mathcal{I}'}(\mathbf{x}, \boldsymbol{\eta})$ is positive, so whenever $(\mathbf{x}, \boldsymbol{\eta})$ enters this neighborhood, it exits it in finite time. Therefore, given that any $\mathring{\mathbf{x}} \in \mathcal{O}$ converges to $\mathbf{x}^*_{\mathcal{I}}$, no $\mathring{\boldsymbol{\eta}}' \in R'$ can converge to zero. Under (ii), there exists a neighborhood of $\mathbf{x}^*_{\mathcal{I}}$ included in $B(\mathbf{x}^*_{\mathcal{I}'}, \rho')$. Since $\Xi_{\mathcal{I}',\mathbf{x}_0}(\mathring{\boldsymbol{\eta}}') \to \mathbf{x}^*_{\mathcal{I}}$, there exists $K \in \mathbb{N}$ such that $\Xi_{\mathcal{I}',\mathbf{x}_0}(\mathring{\boldsymbol{\eta}}')(k)$ belongs to $B(\mathbf{x}^*_{\mathcal{I}'}, \rho')$ for all $k \geq K$. Therefore, if $\|\boldsymbol{\eta}'(k)\| \leq \min\left\{(\kappa')^{-1}(\rho'), \rho'\right\}$ indefinitely after any point of time, $\Xi_{\mathcal{I}',\mathbf{x}_0}(\mathring{\boldsymbol{\eta}}') \to \mathbf{x}^*_{\mathcal{I}'}$ by Proposition 2.6.1 which is a contradiction, so $\mathring{\boldsymbol{\eta}}'$ cannot converge to zero. In both cases, by Lemma 4.3.1,

$$\mathbb{P}(R') = 0,$$

which, together with $\mathbb{P}(R) > 0$ and the definition of ϵ-differential privacy, cf. (4.2), implies the

result. □

Note that the hypotheses of Proposition 4.3.2 are mild and easily satisfied in most cases. In particular, the result holds if the dynamics are continuous and globally asymptotically stable relative to $\mathbf{x}_I^*$ for two information sets. The main take-away message of this result is that a globally asymptotically stable distributed optimization algorithm cannot be made differentially private by perturbing the inter-agent messages with asymptotically vanishing noise. This observation is at the core of the design choices made in the literature regarding the use of stepsizes with finite sum to make the zero-input dynamics not asymptotically stable, thereby causing a steady-state error in accuracy which is present independently of the amount of noise injected for privacy. For instance, the algorithmic solution proposed in [17] replaces (4.4b) by $\mathbf{z}_i(k) = \sum_{j=1}^n a_{ij}\boldsymbol{\xi}_j(k)$, where $\boldsymbol{\xi}_j(k) = \mathbf{x}_j(k) + \eta_j(k)$ is the perturbed message received from agent j, and chooses a finite-sum sequence of stepsizes $\{\alpha_k\}$ in the computation (4.4a), leading to a dynamical system which is not 0-GAS, see Figure 4.1. Similar observations can be made in the scenario considered in [18], where the agents' local constraints are the sensitive information (instead of the objective function). This algorithmic solution uses a constant-variance noise, which would make the dynamics unstable if executed over an infinite time horizon. This problem is circumvented by having the algorithm terminate after a finite number of steps, and optimizing this number offline as a function of the desired level of privacy $\boldsymbol{\epsilon}$.

4.3.2 Algorithm Design via Objective Perturbation

To overcome the limitations of message-perturbing strategies, here we outline an alternative design strategy to solve Problem 2 based on the perturbation of the agents' objective functions. The

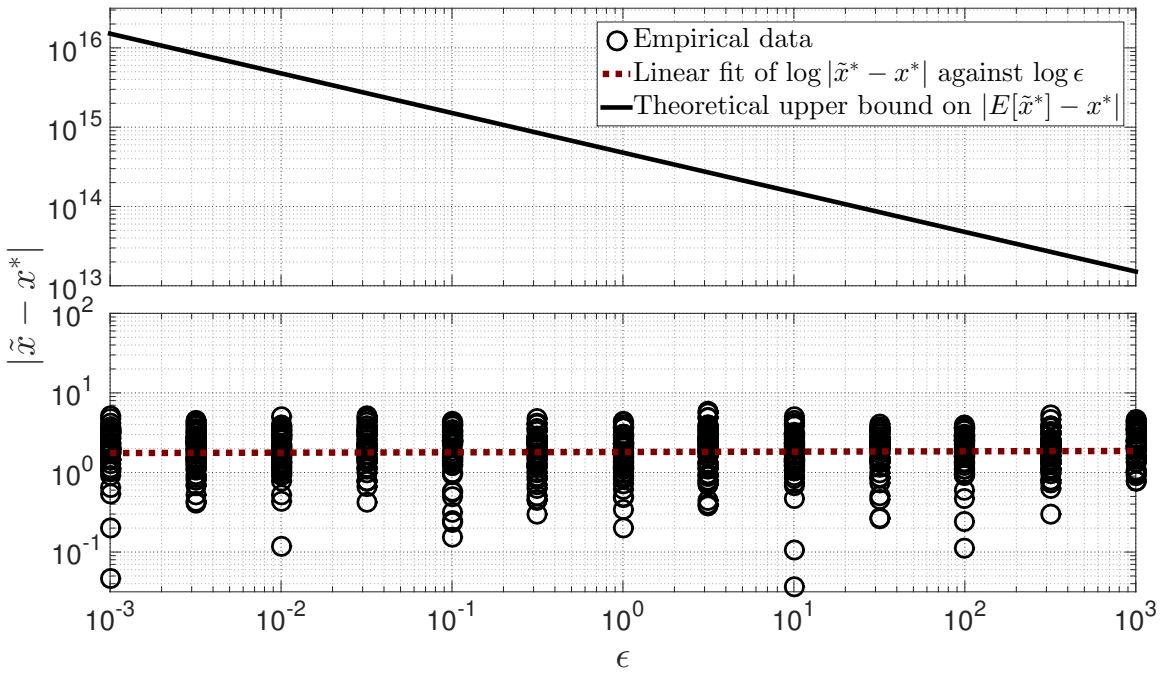


Figure 4.1: Privacy-accuracy trade-off for the algorithm proposed in [17] applied to Example 4.2.2 with $D = X = [-5, 5]^2$, $n = 10$, $N_d = 100$, and $\lambda = 0.01$. With that paper's notation, the algorithm's parameters are $q = 0.1$, $p = 0.11$, $c = 0.5$. The stepsize $\alpha_k = cq^{k-1}$ has finite sum. The circles, dotted line, and solid line illustrate simulation results for 50 executions, their best linear fit in logarithmic scale, and the upper bound on accuracy provided in [17], respectively. We have broken the vertical axis to better display the scale of the algorithm output.

basic idea is to have agents independently perturb their objective functions in a differentially private way and then have them participate in a distributed optimization algorithm with the perturbed objective functions instead of their original ones. In the context of Example 4.2.2, this would correspond to leave (4.4b) and the sequence of stepsizes unchanged, and instead use perturbed functions in the computation (4.4a). The latter in turn automatically adds noise to the estimates shared with neighbors. The following result, which is a special case of [23, Theorem 1], ensures that the combination with the distributed optimization algorithm does not affect the differential privacy at the functional level.

Proposition 4.3.3. ***(Resilience to post-processing).*** *Let* $\mathcal{M} : L_2(D)^n \times \Omega \to L_2(D)^n$ *be* ϵ-

differentially private (cf. Definition 4.2.1) and $\mathcal{F} : L_2(D)^n \to \mathcal{X}$*, where* $(\mathcal{X}, \Sigma_{\mathcal{X}})$ *is an arbitrary measurable space. Then,* $\mathcal{F} \circ \mathcal{M} : L_2(D)^n \times \Omega \to \mathcal{X}$ *is* ϵ*-differentially private.*

Proof. Consider the σ-algebra $\mathcal{P}(L_2(D)^n)$ on $L_2(D)^n$ where $\mathcal{P}$ denotes the power set. With the notation from [23, Theorem 1], the map $M_2 = \mathcal{F} \circ \mathcal{M}$ is a deterministic function of the output of the map $M_1 = \mathcal{M}$. Then, it is easy to verify that, for any $S \in \Sigma_{\mathcal{X}}$,

$$\mathbb{P}(M_2(F) \in S \mid M_1(F)) = \chi_S(\mathcal{F}(M_1(F))),$$

(with $\chi_\cdot$ the indicator function) is measurable as a function of $M_1(F)$ (because $\mathcal{F}$ and S are trivially measurable) and defines a probability measure on $(\mathcal{X}, \Sigma_{\mathcal{X}})$ (associated to a singleton), so it is a probability kernel. Hence, the conditions of [23, Theorem 1] hold and $\mathcal{F} \circ \mathcal{M}$ is ϵ-differentially private in the sense of Definition 4.2.1. □

Our design strategy based on the perturbation of the individual objective functions requires solving the following challenges:

(i) establishing a differentially private procedure to perturb the individual objective functions;

(ii) ensuring that the resulting perturbed functions enjoy the smoothness and regularity properties required by distributed optimization algorithms to converge;

(iii) with (i) and (ii) in place, characterizing the accuracy of the resulting differentially private, distributed coordination algorithm.

Section 4.4 addresses (i) and Section 4.5 deals with (ii) and (iii).

4.4 Functional Differential Privacy

We explore here the concept of functional differential privacy to address the challenge (i) laid out in Section 4.3.2. The generality of this notion makes it amenable for problems where the sensitive information is a function or some of its attributes (e.g., sample points, optimizers, derivatives and integrals). For simplicity of exposition and without loss of generality, we limit our discussion to the privacy of a single function.

4.4.1 Functional Perturbation via Laplace Noise

Let $f \in L_2(D)$ be a function whose differential privacy has to be preserved. With the notation of Section 2, we decompose f into its coefficients $\Phi^{-1}(f)$ and perturb this sequence by adding noise to all of its elements. Specifically, we set

$$\mathcal{M}(f, \mathring{\eta}) = \Phi\left(\Phi^{-1}(f) + \mathring{\eta}\right) = f + \Phi(\mathring{\eta}), \tag{4.6}$$

where

$$\eta_k \sim \text{Lap}(b_k), \tag{4.7}$$

for all $k \in \mathbb{N}$. Clearly, for $\mathring{\eta}$ to belong to ℓ_2 and for the series $\Phi(\mathring{\eta})$ to converge, the scales $\{b_k\}_{k=1}^{\infty}$ cannot be arbitrary. The next result addresses this issue.

Lemma 4.4.1. ***(Sufficient condition for boundedness of perturbed functions).*** *If there exists $K \in$*

$\mathbb{N}$ *such that, for some* $p > \frac{1}{2}$ *and* $s > 1$,

$$b_k \leq \frac{1}{k^p \log k^s}, \quad \forall k \geq K, \tag{4.8}$$

then $\mathring{\eta}$ *defined by* (4.7) *belongs to* ℓ_2 *with probability one. In particular, if for some* $p > \frac{1}{2}$ *and* $\gamma > 0$,

$$b_k \leq \frac{\gamma}{k^p}, \quad \forall k \in \mathbb{N}, \tag{4.9}$$

then $\mathring{\eta}$ *defined by* (4.7) *belongs to* ℓ_2 *with probability one.*

Proof. Equation (4.8) can be equivalently written as $e^{-\frac{1}{k^p b_k}} \leq \frac{1}{k^s}$, for $k \geq K$. In particular, this implies that $\sum_{k=1}^{\infty} e^{-\frac{1}{k^p b_k}}$ is convergent. Therefore [22, §1.14], $\prod_{k=1}^{\infty} \left(1 - e^{-\frac{1}{k^p b_k}}\right)$ converges (i.e., the limit exists and is nonzero), so

$$1 = \lim_{K \to \infty} \prod_{k=K}^{\infty} \left(1 - e^{-\frac{1}{k^p b_k}}\right) = \lim_{K \to \infty} \mathbb{P}(E_K),$$

where $E_K = \left\{\mathring{\eta} \in \mathbb{R}^{\mathbb{N}} \mid \forall k \geq K,\ |\eta_k| \leq \frac{1}{k^p}\right\}$ and we have used the fact that $|\eta_k|$ is exponentially distributed with rate $\frac{1}{b_k}$. Since $E_K \uparrow \bigcup_{K=1}^{\infty} E_K$ as $K \to \infty$, we have

$$\begin{aligned} 1 &= \mathbb{P}\Big(\bigcup_{K=1}^{\infty} E_K\Big) \\ &= \mathbb{P}\left\{\mathring{\eta} \in \mathbb{R}^{\mathbb{N}} \mid \exists K \in \mathbb{N} \text{ s.t. } \forall k \geq K,\ |\eta_k| \leq \frac{1}{k^p}\right\} \\ &\leq \mathbb{P}\{\mathring{\eta} \in \ell_2\}, \end{aligned}$$

as stated. If equation (4.9) holds, we define $\bar{p} = \frac{1}{2}(p + \frac{1}{2})$ and equivalently write (4.9) as

$$b_k \leq \frac{1}{k^{\bar{p}}} \frac{\gamma}{k^{p-\bar{p}}}, \quad \forall k \in \mathbb{N}.$$

Since $p - \bar{p} > 0$, for any $s > 1$ there exists $K \in \mathbb{N}$ such that $k^{p-\bar{p}} \geq \gamma \log k^s$ for all $k \geq K$, and the result follows. □

Having established conditions on the noise variance under which the map (4.6) is well defined, we next turn our attention to establish its differentially private character.

4.4.2 Differential Privacy of Functional Perturbation

Here we establish the differential privacy of the map (4.6). In order to do so, we first specify our choice of adjacency space. Given $q > 1$, consider the weight sequence $\{k^q\}_{k=1}^{\infty}$ and define the adjacency vector space to be the image of the resulting weighted ℓ_2 space under Φ, i.e.,

$$\mathcal{V}_q = \Phi\Big(\Big\{\mathring{\delta} \in \mathbb{R}^{\mathbb{N}} \mid \sum_{k=1}^{\infty} (k^q \delta_k)^2 < \infty\Big\}\Big). \tag{4.10}$$

It is not difficult to see that $\mathcal{V}_q$ is a vector space. Moreover,

$$\|f\|_{\mathcal{V}_q} \triangleq \Big(\sum_{k=1}^{\infty} (k^q \delta_k)^2\Big)^{\frac{1}{2}}, \quad \text{with } \mathring{\delta} = \Phi^{-1}(f),$$

is a norm on $\mathcal{V}_q$. The next result establishes the differential privacy of the map (4.6) for an appropriately chosen noise scale sequence $\mathring{b}$.

Theorem 4.4.2. ***(Differential privacy of functional perturbation).*** *Given $q > 1$, $\gamma > 0$ and $p \in$*

$\left(\frac{1}{2}, q - \frac{1}{2}\right)$, *let*

$$b_k = \frac{\gamma}{k^p}, \quad k \in \mathbb{N}. \tag{4.11}$$

Then, the map (4.6) *is* ϵ*-differentially private with*

$$\epsilon = \frac{1}{\gamma}\sqrt{\zeta(2(q-p))}, \tag{4.12}$$

where ζ *is the Riemann zeta function.*

Proof. Note that the map $\mathcal{M}$ defined by (4.6) is well defined because (4.11) ensures, by Lemma 4.4.1, that $\mathring{\eta}$ belongs to ℓ_2 almost surely. Our proof consists of showing that $\mathcal{M}$ satisfies the definition of differential privacy, cf. Definition 4.2.1. To this effect, consider two functions f and f', with $f - f' \in \mathcal{V}_q$, and an arbitrary set $\mathcal{O} \subseteq L_2(D)$. Let $\Phi_K^{-1} : L_2(D) \to \mathbb{R}^K$ be the map that returns the first K coefficients of $\Phi^{-1}(\cdot)$ and

$$\mathcal{L}^K(\mathring{\eta}_K; \mathring{b}_K) \triangleq \prod_{k=1}^{K} \mathcal{L}(\eta_k; b_k).$$

We have

$$\begin{aligned}\mathbb{P}\{f + \Phi(\mathring{\eta}) \in \mathcal{O}\} &= \mathbb{P}\{\mathring{\eta} \in \Phi^{-1}(\mathcal{O} - f)\} \\ &= \lim_{K\to\infty} \int_{\Phi_K^{-1}(\mathcal{O}-f)} \mathcal{L}^K(\mathring{\eta}_K; \mathring{b}_K) d\mathring{\eta}_K,\end{aligned}$$

where $\Phi_K^{-1}(\mathcal{O} - f)$ denotes the inverse image of the set $\mathcal{O} - f = \{g \in L_2(D) \mid g + f \in \mathcal{O}\}$ and the second equality follows from the continuity of probability [24, Theorem 1.1.1.iv] (since

$\Phi_K^{-1}(\mathcal{O}-f)\times\mathbb{R}^{\mathbb{N}} \downarrow \Phi^{-1}(\mathcal{O}-f)$ as $K\to\infty$). Similarly,

$$\mathbb{P}\{f'+\Phi(\mathring{\eta}')\in\mathcal{O}\}=\lim_{K\to\infty}\int_{\Phi_K^{-1}(\mathcal{O}-f')}\mathcal{L}^K(\mathring{\eta}'_K;\mathring{b}_K)d\mathring{\eta}'_K.$$

By linearity of Φ_K, we have

$$\Phi_K^{-1}(\mathcal{O}-f')=\Phi_K^{-1}(\mathcal{O}-f)+\mathring{\delta}_K,$$

where $\mathring{\delta}=\Phi^{-1}(f-f')$. Therefore,

$$\mathbb{P}\{f'+\Phi(\mathring{\eta}')\in\mathcal{O}\}=\lim_{K\to\infty}\int_{\Phi_K^{-1}(\mathcal{O}-f)}\mathcal{L}^K(\mathring{\eta}_K+\mathring{\delta}_K;\mathring{b}_K)d\mathring{\eta}_K.$$

Note that

$$\frac{\mathcal{L}^K(\mathring{\eta}_K+\mathring{\delta}_K;\mathring{b}_K)}{\mathcal{L}^K(\mathring{\eta}_K;\mathring{b}_K)}=\prod_{k=1}^{K}\frac{\mathcal{L}(\eta_k+\delta_k;b_k)}{\mathcal{L}(\eta_k;b_k)}\le e^{\sum_{k=1}^{K}\frac{|\delta_k|}{b_k}}.$$

After multiplying both sides by $\mathcal{L}^K(\mathring{\eta}_K;\mathring{b}_K)$, integrating over $\Phi_K^{-1}(\mathcal{O}-f)$ and letting $K\to\infty$, we have

$$\mathbb{P}\{f'+\Phi(\mathring{\eta}')\in\mathcal{O}\}\le e^{\sum_{k=1}^{\infty}\frac{|\delta_k|}{b_k}}\mathbb{P}\{f+\Phi(\mathring{\eta})\in\mathcal{O}\}.$$

Finally, the coefficient of the exponential can be upper bounded using Holder's inequality with $p = q = 2$ as

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{|\delta_k|}{b_k} = \sum_{k=1}^{\infty} \frac{k^q |\delta_k|}{k^q b_k} &\leq \left(\sum_{k=1}^{\infty} \frac{1}{(k^q b_k)^2} \right)^{\frac{1}{2}} \left(\sum_{k=1}^{\infty} (k^q \delta_k)^2 \right)^{\frac{1}{2}} \\ &= \left(\sum_{k=1}^{\infty} \frac{1}{(\gamma k^{q-p})^2} \right)^{\frac{1}{2}} \left\| f - f' \right\|_{\mathcal{V}_q} \\ &= \frac{1}{\gamma} \sqrt{\zeta(2(q-p))} \left\| f - f' \right\|_{\mathcal{V}_q},\end{aligned}$$

which completes the proof. □

Remark 4.4.3. ***(Choice of q).*** The choice of parameter q affects the trade-off between the size of the adjacency space $\mathcal{V}_q$ and the noise required to preserve privacy. From (4.10), one can see that decreasing q makes $\mathcal{V}_q$ larger, which allows for the privacy preservation of a larger collection of functions. However, as expected, preserving privacy in a larger space requires more noise. From (12), we see that for a fixed ϵ, γ will be larger (since p cannot be decreased by the same amount as q and ζ is monotonically decreasing), resulting in larger b_k and larger noise. We show later in Theorem VI.2 that the guaranteed upper bound on the expected minimizer deviation also increases as $\{q_i\}_{i=1}^n$ decrease. □

4.5 Differentially Private Distributed Optimization

In this section, we employ functional differential privacy to solve the differentially private distributed optimization problem formulated in Section 4.2 for a group of $n \in \mathbb{N}$ agents. For convenience, we introduce the shorthand notation $S_0 = C^2(D) \subset L_2(D)$ and, for given $\overline{u} > 0$,

$0 < \alpha < \beta$,

$$\mathcal{S} = \{h \in \mathcal{S}_0 \mid \|\nabla h(\mathbf{x})\| \leq \overline{u},\ \forall \mathbf{x} \in D \text{ and } \alpha \mathbf{I}_d \leq \nabla^2 h(\mathbf{x}) \leq \beta \mathbf{I}_d,\ \forall \mathbf{x} \in D^o\},$$

for twice continuously differentiable functions with bounded gradients and Hessians. In the rest of the chapter, we assume that the agents' local objective functions $f_1, \ldots, f_n$ belong to $\mathcal{S}$.

4.5.1 Smoothness and Regularity of the Perturbed Functions

We address here the challenge (ii) laid out in Section 4.3.2. To exploit the framework of functional differential privacy for optimization, we need to ensure that the perturbed functions have the smoothness and regularity properties required by the distributed coordination algorithm. In general, the output (4.6) might neither be smooth nor convex. We detail next how to address these problems by defining appropriate maps that, when composed with $\mathcal{M}$ in (4.6), yield functions with the desired properties. Proposition 4.3.3 ensures that differential privacy is retained throughout this procedure.

Ensuring Smoothness

To ensure smoothness, we rely on the fact that $\mathcal{S}_0$ is dense in $L_2(D)$ and, therefore, given any function g in $L_2(D)$, there exists a smooth function arbitrarily close to it, i.e.,

$$\forall \varepsilon > 0,\ \exists \hat{g}^s \in \mathcal{S}_0 \quad \text{such that} \quad \|g - \hat{g}^s\| < \varepsilon.$$

Here, ε is a design parameter and can be chosen sufficiently small (later, we show how to do this so that the accuracy of the coordination algorithm is not affected).

Remark 4.5.1. ***(Smoothening and truncation).*** A natural choice for the smoothening step, if the basis functions are smooth (i.e., $\{e_k\}_{k=1}^{\infty} \subset \mathcal{S}_0$), is truncating the infinite expansion of g. Such truncation is also inevitable in practical implementations due to the impossibility of handling infinite series. The appropriate truncation order depends on the specific function, the basis set, and the noise decay rate (p in (4.11)). □

Ensuring Strong Convexity and Bounded Hessian

The next result ensures that the orthogonal projection from $\mathcal{S}_0$ onto $\mathcal{S}$ is well defined, and can therefore be used to ensure strong convexity and bounded Hessian of the perturbed functions.

Proposition 4.5.2. ***(Convexity of $\mathcal{S}$ and closedness relative to $\mathcal{S}_0$).*** *The set $\mathcal{S}$ is convex and closed as a subset of $\mathcal{S}_0$ under the 2-norm.*

Proof. The set $\mathcal{S}$ is clearly convex because, if $h_1, h_2 \in \mathcal{S}$ and $\lambda \in [0, 1]$, then for all $\mathbf{x} \in D^o$,

$$\begin{aligned}\nabla^2((1-\lambda)h_1(\mathbf{x}) + \lambda h_2(\mathbf{x})) &= (1-\lambda)\nabla^2 h_1(\mathbf{x}) + \lambda\nabla^2 h_2(\mathbf{x}) \\ &\geq (1-\lambda)\alpha\mathbf{I}_d + \lambda\alpha\mathbf{I}_d = \alpha\mathbf{I}_d.\end{aligned}$$

Similarly, $\nabla^2((1-\lambda)h_1(\mathbf{x}) + \lambda h_2(\mathbf{x})) \leq \beta\mathbf{I}_d$. Also,

$$\begin{aligned}\|\nabla((1-\lambda)h_1(\mathbf{x}) + \lambda h_2(\mathbf{x}))\| &\leq (1-\lambda)\|\nabla h_1(\mathbf{x})\| + \lambda\|\nabla h_2(\mathbf{x})\| \\ &\leq (1-\lambda)\overline{u} + \lambda\overline{u} \leq \overline{u},\end{aligned}$$

for all $\mathbf{x} \in D$. To establish closedness, let

$$\mathcal{S}_1 = \{h \in \mathcal{S}_0 \mid \alpha \mathbf{I}_d \leq \nabla^2 h(\mathbf{x}) \leq \beta \mathbf{I}_d, \ \forall \mathbf{x} \in D^o\},$$

$$\mathcal{S}_2 = \{h \in \mathcal{S}_0 \mid \|\nabla h(\mathbf{x})\| \leq \overline{u}, \ \forall \mathbf{x} \in D\}.$$

Since $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$, it is enough to show that $\mathcal{S}_1$ and $\mathcal{S}_2$ are both closed subsets of $\mathcal{S}_0$.

To show that $\mathcal{S}_1$ is closed, let $\{h_k\}_{k=1}^{\infty}$ be a sequence of functions in $\mathcal{S}_1$ such that $h_k \xrightarrow{\|\cdot\|_2} h \in \mathcal{S}_0$. We show that $h \in \mathcal{S}$. Since $h_k - \frac{\alpha}{2}\|\mathbf{x}\|^2 \xrightarrow{\|\cdot\|_2} h - \frac{\alpha}{2}\|\mathbf{x}\|^2$ and L_2 convergence implies pointwise convergence of a subsequence almost everywhere, there exists $\{h_{k_\ell}\}_{\ell=1}^{\infty}$ and $Y \subset D$ such that $D \setminus Y$ has zero (Lebesgue) measure and $h_{k_\ell}(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}\|^2 \to h(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}\|^2$ for all $\mathbf{x} \in Y$. It is straightforward to verify that Y is dense in D and therefore $Y \cap D^o$ is dense in D^o. Then, by [25, Theorem 10.8], $h - \frac{\alpha}{2}\|\mathbf{x}\|^2$ is convex on D^o, so $\alpha \mathbf{I}_d \leq \nabla^2 h(\mathbf{x})$ for all $\mathbf{x} \in D^o$. Similarly, one can show that $\nabla^2 h(\mathbf{x}) \leq \beta \mathbf{I}_d$ for all $\mathbf{x} \in D^o$. Therefore, $h \in \mathcal{S}_1$.

Next, we prove the closedness of $\mathcal{S}_2$ by contradiction. Assume that $\{h_k\}_{k=1}^{\infty}$ is a sequence of functions in $\mathcal{S}_2$ such that $h_k \xrightarrow{\|\cdot\|_2} h \in \mathcal{S}_0$ but $h \notin \mathcal{S}_2$. Therefore, there exist $\mathbf{x}_0 \in D^o$ such that $\|\nabla h(\mathbf{x}_0)\| > \overline{u}$ and, by continuity of ∇h, $\delta_0 > 0$ and $\upsilon_0 > 0$ such that

$$\|\nabla h(\mathbf{x})\| \geq \overline{u} + \upsilon_0, \quad \forall \mathbf{x} \in B(\mathbf{x}_0, \delta_0) \subseteq D.$$

Let $\mathbf{u}_0 = \frac{\nabla h(\mathbf{x}_0)}{\|\nabla h(\mathbf{x}_0)\|}$. By continuity of ∇h, for all $\upsilon_1 > 0$ there exists $\delta_1 \in (0, \delta_0]$ such that

$$\nabla h(\mathbf{x}) \cdot \mathbf{u}_0 \geq (1 - \upsilon_1)\|\nabla h(\mathbf{x})\|, \quad \forall \mathbf{x} \in B(\mathbf{x}_0, \delta_1).$$

As mentioned above, L_2 convergence implies pointwise convergence of a subsequence $\{h_{k_\ell}\}_{\ell=1}^{\infty}$ almost everywhere. In turn, this subsequence converges to h almost uniformly, i.e., for all $\upsilon_2 > 0$ and all $\upsilon_3 > 0$, there exist $E \subset D$ and $L \in \mathbb{N}$ such that $\mu_L(E) < \upsilon_2$ and

$$|h_{k_\ell}(\mathbf{x}) - h(\mathbf{x})| < \upsilon_3, \quad \forall \mathbf{x} \in D \setminus E \text{ and } \ell \geq L. \tag{4.13}$$

For ease of notation, let $\delta_2 = \delta_1/2$. Using the fundamental theorem of line integrals [26], for all $\mathbf{x} \in B(\mathbf{x}_0, \delta_2) \setminus E$,

$$\begin{aligned} h(\mathbf{x} + \delta_2 \mathbf{u}_0) - h(\mathbf{x}) &= \int_{\mathbf{x}}^{\mathbf{x}+\delta_2\mathbf{u}_0} \nabla h \cdot d\mathbf{r} = \int_{\mathbf{x}}^{\mathbf{x}+\delta_2\mathbf{u}_0} \nabla h \cdot \mathbf{u}_0 \|d\mathbf{r}\| \\ &\geq \int_{\mathbf{x}}^{\mathbf{x}+\delta_2\mathbf{u}_0} (1-\upsilon_1)\|\nabla h\|\|d\mathbf{r}\| \geq (1-\upsilon_1)(\overline{u}+\upsilon_0)\delta_2. \end{aligned} \tag{4.14}$$

Similarly, for all $\mathbf{x} \in B(\mathbf{x}_0, \delta_2) \setminus E$ and all $\ell \in \mathbb{N}$,

$$\begin{aligned} h_{k_\ell}(\mathbf{x} + \delta_2 \mathbf{u}_0) - h_{k_\ell}(\mathbf{x}) &= \int_{\mathbf{x}}^{\mathbf{x}+\delta_2\mathbf{u}_0} \nabla h_{k_\ell} \cdot d\mathbf{r} \\ &\leq \int_{\mathbf{x}}^{\mathbf{x}+\delta_2\mathbf{u}_0} \|\nabla h_{k_\ell}\|\|d\mathbf{r}\| \leq \overline{u}\delta_2. \end{aligned} \tag{4.15}$$

Putting (4.14), (4.15), and (4.13) together and choosing $\upsilon_3 = \upsilon_1\delta_2\overline{u}$, we have for all $\mathbf{x} \in B(\mathbf{x}_0, \delta_2)\setminus E$ and all $\ell \geq L$,

$$\begin{aligned} h(\mathbf{x} + \delta_2 \mathbf{u}_0) - h_{k_\ell}(\mathbf{x} + \delta_2 \mathbf{u}_0) &\geq h(\mathbf{x}) - h_{k_\ell}(\mathbf{x}) + \delta_2(1-\upsilon_1)(\overline{u}+\upsilon_0) - \delta_2\overline{u} \\ &\geq \delta_2(1-\upsilon_1)(\overline{u}+\upsilon_0) - \delta_2(1+\upsilon_1)\overline{u} \triangleq \upsilon_4. \end{aligned} \tag{4.16}$$

The quantity υ_4 can be made strictly positive choosing $\upsilon_1 = \frac{\upsilon_0}{4\vec{u}+3\upsilon_0} > 0$. Let $E^+ = E + \delta_2 \mathbf{u}_0$ and $\mathbf{x}_1 = \mathbf{x}_0 + \delta_2 \mathbf{u}_0$. Then, (4.16) can be rewritten as

$$h(\mathbf{x}) - h_{k_\ell}(\mathbf{x}) \geq \upsilon_4, \quad \forall \mathbf{x} \in \mathcal{N}_{\delta_2}(\mathbf{x}_1) \setminus E^+ \text{ and } \ell \geq L,$$

which, by choosing $\upsilon_2 = \frac{1}{2}\mu_L(B(\mathbf{x}_1, \delta_2))$, implies

$$\int_{\mathcal{N}_{\delta_2}(\mathbf{x}_1)\setminus E^+} |h(\mathbf{x}) - h_{k_\ell}(\mathbf{x})|^2 d\mathbf{x} \geq \upsilon_4^2 \cdot \mu_L(B(\mathbf{x}_1, \delta_2) \setminus E^+)$$

$$\Rightarrow \|h - h_{k_\ell}\| \geq \upsilon_4 \sqrt{\mu_L(B(\mathbf{x}_1, \delta_2))/2} > 0,$$

contradicting $h_{k_\ell} \xrightarrow{\|\cdot\|_2} h$, so $\mathcal{S}_2$ must be closed. □

Given the result in Proposition 4.5.2, the best approximation in $\mathcal{S}$ of a function $h \in \mathcal{S}_0$ is its unique projection onto $\mathcal{S}$, i.e.,

$$\tilde{h} = \text{proj}_{\mathcal{S}}(h).$$

By definition, the projected function has bounded gradient and Hessian.

4.5.2 Algorithm Design and Analysis

We address here the challenge (iii) laid out in Section 4.3.2 and put together the discussion above to propose a class of differentially private, distributed optimization algorithms that solve Problem 2. Unlike message-perturbing distributed coordination algorithms, that have agents use the original objective functions in the computations and rely on perturbing the inter-agent messages

with appropriately chosen noise, here we propose that agents locally perturb their objective functions and use them in their computations, without adding any additional noise to the inter-agent messages. Therefore, we require each agent $i \in \{1, \dots, n\}$ to first compute

$$\hat{f}_i = \mathcal{M}(f_i, \mathring{\eta}_i) = f_i + \Phi(\mathring{\eta}_i), \tag{4.17a}$$

where $\mathring{\eta}_i$ is a sequence of Laplace noise generated by i according to (4.7) with the choice (4.11), then select $\hat{f}_i^s \in \mathcal{S}_0$ such that

$$\|\hat{f}_i - \hat{f}_i^s\| < \varepsilon_i, \tag{4.17b}$$

and finally compute

$$\tilde{f}_i = \text{proj}_{\mathcal{S}}(\hat{f}_i^s). \tag{4.17c}$$

After this process, agents participate in *any* distributed optimization algorithm with the modified objective functions $\{\tilde{f}_i\}_{i=1}^n$. Let

$$\tilde{\mathbf{x}}^* = \arg\min_{\mathbf{x} \in X} \sum_{i=1}^n \tilde{f}_i \quad \text{and} \quad \mathbf{x}^* = \arg\min_{\mathbf{x} \in X} \sum_{i=1}^n f_i,$$

denote, respectively, the output of the distributed algorithm and the optimizer for the original optimization problem (with objective functions $\{f_i\}_{i=1}^n$). The following result establishes the connection between the algorithm's accuracy and the design parameters.

Theorem 4.5.3. ***(Accuracy of a class of distributed, differentially private coordination algo-***

rithm). *Consider a group of n agents which perturb their local objective functions according to* (4.17) *with Laplace noise* (4.7) *of variance* (4.11)*, where* $q_i > 1$*,* $\gamma_i > 0$*, and* $p_i \in \left(\frac{1}{2}, q_i - \frac{1}{2}\right)$ *for all* $i \in \{1, \dots, n\}$*. Let the agents participate in any distributed coordination algorithm that asymptotically converges to the optimizer* $\tilde{\mathbf{x}}^*$ *of the perturbed aggregate objective function. Then,* ϵ_i*-differential privacy of each agent i's original objective function is preserved with* $\epsilon_i = \sqrt{\zeta(2(q_i - p_i))}/\gamma_i$ *and*

$$\|\mathbb{E}[\tilde{\mathbf{x}}^*] - \mathbf{x}^*\| \leq \sum_{i=1}^{n} \kappa_n\left(\gamma_i \sqrt{\zeta(2p_i)}\right) + \kappa_n(\varepsilon_i),$$

where the function $\kappa_n \equiv \kappa_{n\alpha,n\beta}$ *is defined in Proposition 4.A.2.*

Proof. Since the distributed algorithm is a post-processing step on the perturbed functions, privacy preservation of the objective functions follows from Theorem 4.4.2 and Proposition 4.3.3. For convenience, let $\Delta = \|\mathbb{E}[\tilde{\mathbf{x}}^* - \mathbf{x}^*]\|$. Note that

$$\Delta \leq \mathbb{E}\,\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\| = \mathbb{E}\left\| \arg\min_{\mathbf{x}\in X} \sum_{i=1}^{n} \tilde{f}_i - \arg\min_{\mathbf{x}\in X} \sum_{i=1}^{n} f_i \right\|.$$

Since $\mu_{n\alpha,n\beta}$ is convex and belongs to class $\mathcal{K}_\infty$ (so is monotonically increasing), κ_n is concave and belongs to class $\mathcal{K}_\infty$ and so is subadditive. Therefore, using Proposition 4.A.2,

$$\begin{aligned}\Delta &\leq \mathbb{E}\left[\kappa_n\left(\left\| \sum_{i=1}^{n} \tilde{f}_i - \sum_{i=1}^{n} f_i \right\|\right)\right] \\ &\leq \mathbb{E}\left[\kappa_n\left(\sum_{i=1}^{n} \|\tilde{f}_i - f_i\| \right)\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[\kappa_n\left(\|\tilde{f}_i - f_i\|\right)\right].\end{aligned}$$

Then, by the non-expansiveness of projection, we have

$$\begin{aligned}\Delta &\leq \sum_{i=1}^{n} \mathbb{E}\big[\kappa_n\big(\|\hat{f}_i^s - f_i\|\big)\big] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\big[\kappa_n\big(\|\hat{f}_i^s - \hat{f}_i\|\big) + \kappa_n\big(\|\hat{f}_i - f_i\|\big)\big] \\ &\leq \sum_{i=1}^{n} \big(\kappa_n(\varepsilon_i) + \mathbb{E}\big[\kappa_n\big(\|\mathring{\eta}_i\|\big)\big]\big). \end{aligned} \tag{4.18}$$

By invoking Jensen's inequality twice, for all $i \in \{1, \ldots, n\}$,

$$\begin{aligned}\mathbb{E}\big[\kappa_n(\|\mathring{\eta}_i\|)\big] &\leq \kappa_n\big(\mathbb{E}\big[\|\mathring{\eta}_i\|\big]\big) = \kappa_n\big(\mathbb{E}\big[\sqrt{\|\mathring{\eta}_i\|^2}\big]\big) \\ &\leq \kappa_n\Big(\sqrt{\mathbb{E}\big[\|\mathring{\eta}_i\|^2\big]}\Big) = \kappa_n\Big(\sqrt{\textstyle\sum_{k=1}^{\infty} b_{i,k}^2}\Big) \\ &= \kappa_n\Big(\gamma_i\sqrt{\zeta(2p_i)}\Big). \end{aligned} \tag{4.19}$$

The result follows from (4.18) and (4.19). □

The following result describes the trade-off between accuracy and privacy. The proof follows by direct substitution.

Corollary 4.5.4. ***(Privacy-accuracy trade-off).*** *Under the hypotheses of Theorem 4.5.3, if* $p_i = \frac{q_i}{2}$ *in* (4.11) *for all* i*, then*

$$\|\mathbb{E}[\tilde{\mathbf{x}}^*] - \mathbf{x}^*\| \leq \sum_{i=1}^{n} \kappa_n\left(\frac{\zeta(q_i)}{\epsilon_i}\right) + \kappa_n(\varepsilon_i). \tag{4.20}$$

In Corollary 4.5.4, q_i and ϵ_i are chosen independently, which in turn determines the value of γ_i according to (4.12). Also, it is clear from (4.20) that in order for the accuracy of the coordination

algorithm not to be affected by the smoothening step, each agent $i \in \{1, \dots, n\}$ has to take the value of ϵ_i sufficiently small so that it is negligible relative to $\zeta(2p_i)/\epsilon_i$. In particular, this procedure can be executed for any arbitrarily large value of ϵ_i, so that in case of no privacy requirements at all, perfect accuracy is recovered, as specified in Problem 2.

Remark 4.5.5. ***(Accuracy bound for sufficiently large domains).*** One can obtain a less conservative bound than (4.20) on the accuracy of the proposed class of algorithms if the minimizers of all the agents' objective functions are sufficiently far from the boundary of X. This can be made precise via Corollary 4.A.3. If the aggregate objective function satisfies (4.25) and the amount of noise is also sufficiently small so that the minimizer of the sum of the perturbed objective functions satisfies this condition, then invoking Corollary 4.A.3, one can obtain

$$\begin{aligned} \|\mathbb{E}[\tilde{\mathbf{x}}^*] - \mathbf{x}^*\| &\leq \frac{L}{n^2} \sum_{i=1}^{n} \left(\gamma_i^{\frac{2}{d+4}} \zeta(2p_i)^{\frac{1}{d+4}} + \epsilon_i^{\frac{2}{d+4}} \right) \\ &= \frac{L}{n^2} \sum_{i=1}^{n} \left[\left(\frac{\zeta(q_i)}{\epsilon_i} \right)^{\frac{2}{d+4}} + \epsilon_i^{\frac{2}{d+4}} \right], \end{aligned}$$

where the equality holds under the assumption that $p_i = \frac{q_i}{2}$ in (4.11) for all $i \in \{1, \dots, n\}$. □

4.6 Simulations

In this section, we report simulation results for our algorithm design for Example 4.2.2 with $D = X = [-5, 5]^2$, $n = 10$, $N_d = 100$, and $\lambda = 0.01$. The orthonormal basis of $L_2(D)$ is constructed from the Gram-Schmidt orthogonalization of the Taylor functions and the series is truncated to the second, sixth, and fourteenth orders, resulting in 15, 28, and 120-dimensional coefficient spaces, respectively. This truncation also acts as the smoothening step described in Sec-

tion 4.5.1, where higher truncation orders result in smaller ϵ. We evaluate the projection operator in (4.17c) by numerically solving the convex optimization problem $\min_{\tilde{f}_i \in S} \|\tilde{f}_i - \hat{f}_i^s\|$, where $\hat{f}_i^s$ is the result of the truncation. The parameters of S are given by $\alpha = N_d\lambda$, $\beta = N_d\lambda + N_d r_D\sqrt{2} + e^{2r_D}$, and $\bar{u} = \sqrt{2}N_d(\lambda r_D + e^{2r_D})$ where $r_D = 5$. Rather than implementing any specific distributed coordination algorithm, we use an iterative interior-point algorithm on $\tilde{f}$ and f to find the perturbed $\tilde{\mathbf{x}}^*$ and original $\mathbf{x}^*$ optimizers, respectively (these points correspond to the asymptotic behavior of any provably correct distributed optimization algorithm with the perturbed and original functions, respectively).

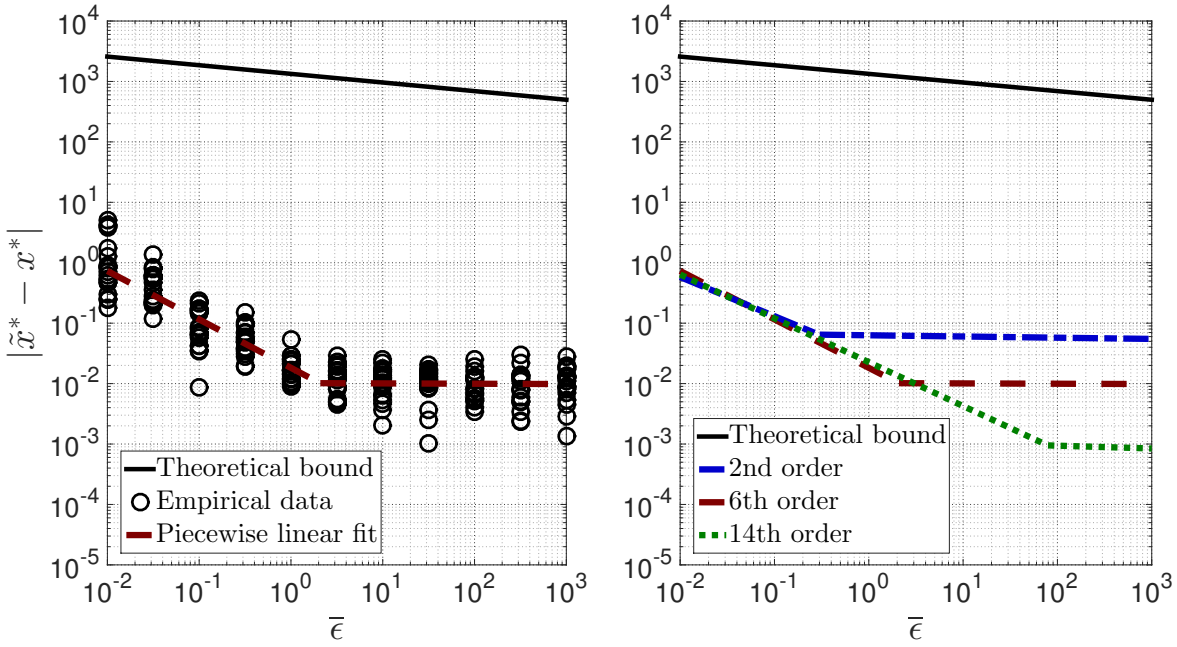


Figure 4.2: Privacy-accuracy trade-off curve of the proposed class of distributed, differentially private algorithms in Section 4.5.2 for Example 4.2.2 (with the same data as Figure 4.1) and different truncation orders. Left: empirical data and its best piecewise linear fit for 6th-order truncation of the function expansions, together with the theoretical upper bound of Corollary 4.5.4. Right: piecewise linear fit of empirical data for 2nd, 6th, and 14th order truncations as well as the theoretical upper bound. Accuracy improves with the truncation order.

The privacy levels are chosen the same for all agents, i.e., $\boldsymbol{\epsilon} = \bar{\epsilon}\mathbf{1}_n$, and $\bar{\epsilon}$ is swept logarithmically over $[10^{-2}, 10^3]$. For each $i \in \{1, \ldots, n\}$, we set $q_i = 2p_i = 1.1$ and $\gamma_i = \sqrt{\zeta(2(q_i - p_i))}/\bar{\epsilon}$. For

each value of $\overline{\epsilon}$ and truncation order, the simulations are repeated 20 times to capture the stochasticity of the solutions. Figure 4.2 illustrates the error $\|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|$ as a function of $\overline{\epsilon}$ for different truncation orders, together with the best linear fit of $\log \|\tilde{\mathbf{x}}^* - \mathbf{x}^*\|$ against $\log \overline{\epsilon}$, and the upper bound obtained in Corollary 4.5.4. The conservative nature of this upper bound can be explained by noting the approximations leading to the computation of L in Proposition 4.A.2, suggesting there is room for refining this bound. Figure 4.2 shows that accuracy keeps improving as the privacy requirement is relaxed until the ϵ-term (resulting from the smoothening/truncation) dominates the error. This saturation value can be decreased by increasing the truncation order (which comes at the expense of more computational complexity), in contrast with the behavior of message-perturbing algorithms, cf. Figure 4.1. It is important to mention that the respective error values for a fixed ϵ cannot be compared between Figures 4.1 and 4.2 because, in [17], ϵ is defined as the total exponent in (4.2), i.e., $\epsilon_{i_0}\|f_{i_0} - f'_{i_0}\|_{\mathcal{V}}$. However, it can be seen that the accuracy in Figure 4.1 is almost indifferent to the value of ϵ and is in the same order as $r_D = 5$. This is explained by the impossibility result of Proposition IV.2: since the noise-free algorithm of [17] is not asymptotically stable, depending on the specific application, its accuracy may not be desirable regardless of the value of ϵ. In contrast, the accuracy in Figure 4.2 keeps improving as ϵ is increased (with an appropriate choice of truncation order).

Appendix

4.A $\mathcal{K}$-Lipschitz Property of the arg min Map

Here we establish the Lipschitzness of the arg min map under suitable assumptions. This is a strong result of independent interest given that arg min is not even continuous for arbitrary C^2 functions. Our accuracy analysis for the proposed class of distributed, differentially private coordination algorithms in Section 4.5.2 relies on this result. We begin with an auxiliary result stating a geometric property of balls contained in convex, compact domains.

Lemma 4.A.1. ***(Minimum portion of balls contained in convex compact domains).*** *Assume $D \subset \mathbb{R}^d$ is convex, compact, and has nonempty interior and let $r_D > 0$ denote its inradius. Then, there exists $\lambda_D \in (0, 1)$ such that,*

$$\mu_L(B(\mathbf{x}, r) \cap D) \geq \lambda_D \mu_L(B(\mathbf{x}, r)),$$

for any $\mathbf{x} \in D$ and $r \leq r_D$.

Proof. Let $B(\mathbf{c}_D, r_D)$ be the inball of D, i.e., the largest ball contained in D. If this ball is not unique, we pick one arbitrarily. Since $D^o \neq \emptyset$, $r_D > 0$. Let R_D be the radius of the largest ball centered at $\mathbf{c}_D$ that contains D. Since D is compact, $R_D < \infty$. For any $\mathbf{x} \in D$ that is on or outside of $B(\mathbf{c}_D, r_D)$, let Σ be the intersection of $B(\mathbf{c}_D, r_D)$ and the hyperplane passing through $\mathbf{c}_D$ and perpendicular to $\mathbf{c}_D - \mathbf{x}$. Consider the cone $C = \text{conv}(\Sigma \cup \{\mathbf{x}\})$ where conv denotes convex hull. Since D is convex, $C \subseteq D$. Note that C has half angle $\theta_x = \tan^{-1} \frac{r_D}{\|\mathbf{x}-\mathbf{c}_D\|}$ so the solid angle at its apex is

$$\Omega_{\theta_x} = \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_0^{\theta_x} \sin^{d-2}(\phi) d\phi. \tag{4.21}$$

Therefore, for any $r \leq r_D$, the proportion $\frac{\Omega_{\theta_x}}{\Omega_d}$ of $B(\mathbf{x}, r)$ is contained in D where Ω_d is the total d-dimensional solid angle given by

$$\Omega_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}.$$

For any $\mathbf{x}$ inside $B(\mathbf{c}_D, r_D)$, the same argument holds with

$$\theta_x = \max_{\|\mathbf{x}-\mathbf{c}_D\| \geq r_D} \tan^{-1} \frac{r_D}{\|\mathbf{x} - \mathbf{c}_D\|} = \frac{\pi}{4}.$$

Therefore, for arbitrary $\mathbf{x} \in D$, the statement holds with

$$\lambda_D = \min_{\mathbf{x} \in D} \frac{\Omega_{\theta_x}}{\Omega_d} = \frac{1}{\Omega_d} \Omega_{\tan^{-1}(r_D/R_D)}.$$

□

We are now ready to establish the $\mathcal{K}$-Lipschitzness of the arg min map.

Proposition 4.A.2. ***($\mathcal{K}$-Lipschitzness of* arg min*).*** *For any two functions $f, g \in \mathcal{S}$,*

$$\left\| \arg\min_{\mathbf{x} \in X} f - \arg\min_{\mathbf{x} \in X} g \right\| \leq \kappa_{\alpha,\beta}(\|f - g\|), \tag{4.22}$$

where $\kappa_{\alpha,\beta} \in \mathcal{K}_\infty$ is given by

$$\kappa^{-1}_{\alpha,\beta}(r) = \frac{\alpha^2 \pi^{\frac{d}{2}}}{d 2^{d+3} \Gamma(\frac{d}{2})} \lambda_D \left(\frac{r_D}{d_D} \right)^d r^4 \mu^d_{\alpha,\beta}(r), \quad \forall r \in [0, \infty),$$

r_D and λ_D are as in Lemma 4.A.1, d_D is the diameter of D, and $\mu_{\alpha,\beta} \in \mathcal{K}_\infty$ is defined for all

$r \in [0, \infty)$ *by*

$$\mu_{\alpha,\beta}(r) = \frac{\alpha r^2}{2\sqrt{\alpha\beta r^2 + 2(\beta + \alpha)\bar{u}r + 4\bar{u}^2}}.$$

Proof. We consider the case where $\mathbf{a} = \arg\min_{\mathbf{x}\in X} f(\mathbf{x}) \neq \arg\min_{\mathbf{x}\in X} g(\mathbf{x}) = \mathbf{b}$ since the statement is trivial otherwise. Let $m_{\mathbf{a}} = f(\mathbf{a})$, $m_{\mathbf{b}} = g(\mathbf{b})$, $m = m_{\mathbf{a}} - m_{\mathbf{b}}$, $\mathbf{u}_{\mathbf{a}} = \nabla f(\mathbf{a})$, and $\mathbf{u}_{\mathbf{b}} = \nabla g(\mathbf{b})$. Without loss of generality, assume $m \geq 0$. Define,

$$f_l(\mathbf{x}) = \frac{\alpha}{2}\|\mathbf{x} - \mathbf{a}\|^2 + \mathbf{u}_{\mathbf{a}}^T(\mathbf{x} - \mathbf{a}) + m_{\mathbf{a}},$$
$$g_u(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \mathbf{u}_{\mathbf{b}}^T(\mathbf{x} - \mathbf{b}) + m_{\mathbf{b}},$$

for all $\mathbf{x} \in D$. Since $f, g \in S$, we can integrate $\nabla^2 f \geq \alpha \mathbf{I}_d$ and $\nabla^2 g \leq \beta \mathbf{I}_d$ twice to get,

$$\forall \mathbf{x} \in D \quad f_l(\mathbf{x}) \leq f(\mathbf{x}) \text{ and } g(\mathbf{x}) \leq g_u(\mathbf{x}). \tag{4.23}$$

It follows that, for all $\mathbf{x} \in D$,

$$|f(\mathbf{x}) - g(\mathbf{x})| \geq [f_l(\mathbf{x}) - g_u(\mathbf{x})]^+ \geq [f_l(\mathbf{x}) - g_u(\mathbf{x}) - m]^+,$$

where $[z]^+ = \max\{z, 0\}$ for any $z \in \mathbb{R}$. After some computations, one can get

$$f_l(\mathbf{x}) - g_u(\mathbf{x}) - m = -\frac{\beta - \alpha}{2}\Big(\|\mathbf{x} - \mathbf{c}\|^2 - r^2 \Big),$$

where

$$\mathbf{c} = \frac{\beta\mathbf{b} - \alpha\mathbf{a} + \mathbf{u_a} - \mathbf{u_b}}{\beta - \alpha},$$
$$r^2 = \frac{\alpha\beta\|\mathbf{a} - \mathbf{b}\|^2}{(\beta - \alpha)^2} + \frac{\|\mathbf{u_a} - \mathbf{u_b}\|^2}{(\beta - \alpha)^2} + \frac{2(\beta\mathbf{u_a} - \alpha\mathbf{u_b})^T(\mathbf{b} - \mathbf{a})}{(\beta - \alpha)^2}.$$

Therefore, the region where $f_l - g_u - m \geq 0$ is $B(\mathbf{c}, r)$. Next, we seek to identify a subset inside this ball where we can determine a strictly positive lower bound of $f_l - g_u$ that depends on the difference $\|\mathbf{a} - \mathbf{b}\|$. To this effect, note that $\mathbf{b} \in B(\mathbf{c}, r)$, since

$$r^2 - \|\mathbf{c} - \mathbf{b}\|^2 = \frac{\alpha}{\beta - \alpha}\|\mathbf{a} - \mathbf{b}\|^2 + \frac{2}{\beta - \alpha}\mathbf{u_a}^T(\mathbf{b} - \mathbf{a}),$$

and, by the convexity of the problem, $\mathbf{u_a}^T(\mathbf{b} - \mathbf{a}) \geq 0$. Let $\underline{r} = r - \|\mathbf{c} - \mathbf{b}\| > 0$ be the radius of the largest ball centered at $\mathbf{b}$ and contained in $B(\mathbf{c}, r)$. We have,

$$r^2 - \|\mathbf{c} - \mathbf{b}\|^2 = (r - \|\mathbf{c} - \mathbf{b}\|)(r + \|\mathbf{c} - \mathbf{b}\|) \geq \frac{\alpha}{\beta - \alpha}\|\mathbf{a} - \mathbf{b}\|^2$$
$$\Rightarrow \underline{r} \geq \frac{\frac{\alpha}{\beta-\alpha}\|\mathbf{a} - \mathbf{b}\|^2}{r + \|\mathbf{c} - \mathbf{b}\|} \geq \frac{\frac{\alpha}{\beta-\alpha}\|\mathbf{a} - \mathbf{b}\|^2}{2r} \geq \mu_{\alpha,\beta}(\|\mathbf{a} - \mathbf{b}\|),$$

where in the last inequality we have used $\|\mathbf{u_a}\|, \|\mathbf{u_b}\| \leq \overline{u}$. Next, note that for all $\mathbf{x} \in B(\mathbf{c}, \frac{r+\|\mathbf{c}-\mathbf{b}\|}{2})$,

$$f_l(\mathbf{x}) - g_u(\mathbf{x}) - m \geq -\frac{\beta - \alpha}{2}\Big(\frac{r^2 + \|\mathbf{c} - \mathbf{b}\|^2 + 2r\|\mathbf{c} - \mathbf{b}\|}{4} - r^2\Big). \tag{4.24}$$

Using the bound $2r\|\mathbf{c}-\mathbf{b}\| \leq r^2 + \|\mathbf{c}-\mathbf{b}\|^2$, we get after some simplifications,

$$(f_l - g_u)(\mathbf{x}) - m \geq \frac{\alpha}{4}\|\mathbf{a}-\mathbf{b}\|^2 + \frac{1}{2}\mathbf{u}_{\mathbf{a}}^T(\mathbf{b}-\mathbf{a}) \geq \frac{\alpha}{4}\|\mathbf{a}-\mathbf{b}\|^2,$$

for all $\mathbf{x} \in B(\mathbf{b}, \frac{r}{2}) \subset B(\mathbf{c}, \frac{r+\|\mathbf{c}-\mathbf{b}\|}{2})$. Therefore,

$$\begin{aligned}
\|f-g\|^2 &= \int_D |f(\mathbf{x}) - g(\mathbf{x})|^2 d\mathbf{x} \\
&\geq \int_D ([f_l(\mathbf{x}) - g_u(\mathbf{x}) - m]^+)^2 d\mathbf{x} \\
&\geq \int_{B(\mathbf{b},\frac{r}{2})\cap D} (f_l(\mathbf{x}) - g_u(\mathbf{x}) - m)^2 d\mathbf{x} \\
&\geq \frac{\alpha^2}{16}\|\mathbf{a}-\mathbf{b}\|^4 m\left(B(\mathbf{b}, \tfrac{r}{2}) \cap D\right) \\
&\geq \frac{\alpha^2}{16}\|\mathbf{a}-\mathbf{b}\|^4 m\left(B(\mathbf{b}, \tfrac{\mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|)}{2}) \cap D\right).
\end{aligned}$$

Now, we invoke Lemma 4.A.1 to lower bound the last term. Note that $\mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|) \leq \|\mathbf{a}-\mathbf{b}\| \leq d_D$ for all $\mathbf{a}, \mathbf{b} \in D$. Therefore, $\frac{r_D}{d_D}\frac{\mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|)}{2} \leq \min\{r_D, \mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|)/2\}$, so by Lemma 4.A.1,

$$\begin{aligned}
\|f-g\|^2 &\geq \frac{\alpha^2}{16}\|\mathbf{a}-\mathbf{b}\|^4 m\left(B\left(\mathbf{b}, \frac{r_D\mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|}{2d_D}\right) \cap D\right) \\
&\geq \frac{\alpha^2}{16}\|\mathbf{a}-\mathbf{b}\|^4 \lambda_D \frac{2\pi^{\frac{d}{2}}}{d\Gamma(\frac{d}{2})}\frac{r_D^d}{2^d d_D^d}(\mu_{\alpha,\beta}(\|\mathbf{a}-\mathbf{b}\|))^d,
\end{aligned}$$

which yields (4.22). □

The next result shows that if the minimizers of f and g are sufficiently far from the boundary of D, then their gradients need not be uniformly bounded and yet one can obtain a less conservative characterization of the $\mathcal{K}$-Lipschitz property of the arg min map.

Corollary 4.A.3. ***($\mathcal{K}$-Lipschitzness of* $\arg\min$ *for sufficiently large domains).*** *If f and g belong to $\mathcal{S}_1 = \{h \in \mathcal{S}_0 \mid \alpha \mathbf{I}_d \leq \nabla^2 h(\mathbf{x}) \leq \beta \mathbf{I}_d,\ \forall \mathbf{x} \in D^o\}$ and*

$$\arg\min_{\mathbf{x}\in X} f(\mathbf{x}),\ \arg\min_{\mathbf{x}\in X} g(\mathbf{x}) \in B(\mathbf{c}_D, \underline{r}_D) \cap X^o, \tag{4.25}$$

where $\underline{r}_D = \frac{\beta-\alpha}{\alpha+\beta+2\sqrt{\alpha\beta}} r_D$ and $B(\mathbf{c}_D, r_D) \subset D$, then

$$\left\| \arg\min_{\mathbf{x}\in X} f - \arg\min_{\mathbf{x}\in X} g \right\| \leq L \|f - g\|^{\frac{2}{d+4}},$$

where

$$L = \frac{d(d+2)(d+4)(\beta-\alpha)^{d+2}\Gamma(d/2)}{4(\alpha\beta)^{d/2+2}\pi^{d/2}}.$$

Proof. The proof follows the same lines as the proof of Proposition 4.A.2 (and we use here the same notation). Since the minimizers of f and g lie in the interior of X, $\mathbf{u_a} = \mathbf{u_b} = \mathbf{0}$. The main difference here is that due to (4.25), we have for all $\mathbf{x} \in B(\mathbf{c}, r)$ that

$$\begin{aligned} \|\mathbf{x} - \mathbf{c}_D\| &\leq \|\mathbf{x} - \mathbf{c}\| + \|\mathbf{c} - \mathbf{b}\| + \|\mathbf{b} - \mathbf{c}_D\| \\ &\leq \frac{\alpha + \sqrt{\alpha\beta}}{\beta - \alpha} 2\underline{r}_D + \underline{r}_D = r_D, \end{aligned}$$

so $B(\mathbf{c}, r) \subset D$. Therefore, one can integrate $(f_l - g_u - m)^2$ on the whole $B(\mathbf{c}, r)$ instead of its lower bound (4.24) on the smaller ball $B(\mathbf{c}, \frac{r+\|\mathbf{c}-\mathbf{b}\|}{2})$. To explicitly calculate the value of the resulting integral, one can use the change of variables $x_i = c_i + r y_i^{1/2}, i \in \{1, \dots, d\}$ and then use the

formula

$$\forall a_i > -1,\ \int_{S_d} y_1^{a_1} \cdots y_d^{a_d}\, dy = \frac{\Gamma(a_1+1)\cdots\Gamma(a_d+1)}{\Gamma(a_1+\cdots+a_d+d+1)},$$

where $S_d = \{\mathbf{y} \in \mathbb{R}^d \mid \sum_{i=1}^d y_i = 1 \text{ and } y_i \geq 0,\ i \in \{1,\dots,d\}\}$. The result then follows from straightforward simplifications of the integral. □

Acknowledgements: This chapter is taken, in part, from the work published as "Differentially private distributed convex optimization via functional perturbation" by E. Nozari, P. Tallapragada, and J. Cortés in *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 395–408, 2018. The dissertation author was the primary investigator and author of this paper.

Chapter Bibliography

[1] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.

[2] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[3] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Control and Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[4] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.

[5] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 781–786, 2014.

[6] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Theory of Cryptography Conference*, New York, NY, Mar. 2006, pp. 265–284.

[8] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, Venice, Italy, July 2006, pp. 1–12.

[9] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, Aug. 2014.

[10] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *The Journal of Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.

[11] D. Kifer, A. Smith, and A. Thakurta, "Private convex empirical risk minimization and high-dimensional regression," in *25th Annual Conference on Learning Theory*, vol. 23, 2012, pp. 25.1–25.40.

[12] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional mechanism: Regression analysis under differential privacy," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1364–1375, July 2012.

[13] R. Hall, A. Rinaldo, and L. Wasserman, "Differential privacy for functions and functional data," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 703–727, Jan. 2013.

[14] A. Rajkumar and S. Agarwal, "A differentially private stochastic gradient descent algorithm for multiparty classification," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, vol. 22, 2012, pp. 933–941.

[15] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *Proceedings of the Global Conference on Signal and Information Processing*. IEEE, Dec. 2013, pp. 245–248.

[16] K. Chaudhuri, D. J. Hsu, and S. Song, "The large margin mechanism for differentially private maximization," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 1287–1295.

[17] Z. Huang, S. Mitra, and N. Vaidya, "Differentially private distributed optimization," in *Proceedings of the 2015 International Conference on Distributed Computing and Networking*, Pilani, India, Jan. 2015.

[18] S. Han, U. Topcu, and G. J. Pappas, "Differentially private distributed constrained optimization," *IEEE Transactions on Automatic Control*, 2016, to appear.

[19] M. T. Hale and M. Egerstedt, "Differentially private cloud-based multi-agent optimization with constraints," in *American Control Conference*, Chicago, IL, July 2015, pp. 1235–1240.

[20] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 7981, pp. 82–102.

[21] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, Oct 2013, pp. 429–438.

[22] H. Jeffreys and B. S. Jeffreys, *Methods of Mathematical Physics*, 3rd ed. Cambridge University Press, 1999.

[23] J. L. Ny and G. J. Pappas, "Differentially private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2014.

[24] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[25] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[26] R. E. Williamson and H. F. Trotter, *Multivariable Mathematics*, 4th ed. Pearson Education, Inc., June 2003.

Part II

Network Control Under Resource Constraints

Chapter 5

Event-Triggered Stabilization of Delayed Network Systems

In this chapter, we turn to a different challenge in the analysis and control of network dynamical systems, the existence of imperfect communication channels between the network nodes. We consider a network control system composed of two nodes, a (possibly) nonlinear plant and a controller, and study the problem of stabilization of the plant using event-triggered control where time-varying delays can exist in both sensing and actuation.

Event-triggered and self-triggered approaches have recently gained popularity for controlling cyberphysical systems. The basic premise is that of abandoning the assumption of continuous or periodic updating of the control signal and instead adopt an opportunistic perspective that leads to deliberate, aperiodic updates. The challenge resides in determining precisely when control signals should be updated to improve efficiency while still guaranteeing convergence. We here expand the state-of-the-art in opportunistic state-triggered control by designing predictor-based event-triggered control strategies that stabilize nonlinear systems with *known* delays in both sensing and actuation

that can be *arbitrarily large* and *time-varying*.

Our proposed strategy seeks to opportunistically minimize the number of control updates while guaranteeing stabilization and builds on predictor feedback to compensate for arbitrarily large known time-varying delays. We establish, using a Lyapunov approach, the global asymptotic stability of the closed-loop system as long as the open-loop system is globally input-to-state stabilizable in the absence of time delays and event-triggering. We further prove that the proposed event-triggered law has inter-event times that are uniformly lower bounded and hence does not exhibit Zeno behavior. For the particular case of a stabilizable linear system, we show global exponential stability of the closed-loop system and analyze the trade-off between the rate of exponential convergence and average sampling frequency. We illustrate these results in simulation and also examine the properties of the proposed event-triggered strategy beyond the class of systems for which stabilization can be guaranteed.

5.1 Prior Work

There exists a vast literature on both event-triggered control and the control of time-delay systems. Here, we review the works most closely related to our treatment. Originating from event-based and discrete-event systems [1, 2], the concept of event-triggered control (i.e., updating the control signal in an opportunistic fashion) was proposed in [3, 4] and has found its way into the efficient use of sensing, computing, actuation, and communication resources in networked control systems, see e.g., [5–8] and references therein.

On the other hand, the notion of predictor feedback is a powerful method in dealing with controlled systems subject to time delay [9–14]. In essence, a predictor feedback controller antic-

ipates the future evolution of the plant using its forward model and sends the control signal early enough to compensate for the delay. Here, we pursue a Lyapunov-based analysis of predictor feedback following [15]. Given that the numerical implementations of predictor feedback controllers are particularly challenging [16, 17], we further discuss several methods for the numerical implementation of our proposed controller and show that a carefully designed "closed-loop" method is numerically stable and robust to errors in delay compensation.

The joint treatment of time delay and event-triggering is particularly challenging. By its opportunistic nature, an event-triggered controller keeps the control value unchanged until the plant is close to instability and then updates the control value according to the current state. Now, if time delays exist, the controller only has access to some past state of the plant (delayed sensing) and it takes some time for an updated control action to reach the plant (delayed actuation), jointly increasing the possibility of the updated control value being already obsolete when it is implemented in the plant, resulting in instability. Therefore, the controller needs to be sufficiently proactive and update the control value sufficiently ahead of time to maintain closed-loop stability. This makes the design problem challenging. Delays in actuation and sensing may be due to communication delays between controller-actuator and controller-sensor pairs, and in that sense, previous work on the event-triggered control literature that specifically considers delays in the communication channel deals with a similar problem setup as the one considered here. Several event-triggered designs consider scenarios where the system dynamics are linear, see, e.g. [18–23]. The inclusion of nonlinearity, however, makes the problem more challenging. When digital controllers are used and the delay is smaller than the sampling time, [24, 25] design event-triggered controllers for the resulting delay-free discretized system. Robust event-triggered stabilizing controllers are also designed for nonlinear systems with sensing delays in [26] and with both sensing and actuation

delays in [5, 27].

In all these works, however, a key assumption is that the (maximum) delay is smaller than the (minimum) inter-transmission time. This assumption (also called the small-delay case) allows for the *treatment of delay as a disturbance* and, by construction, can tolerate unknown delays. In reality, however, (minimum) inter-transmission times can be very small, making this assumption restrictive. Similar to our preliminary work [28], we take a different perspective here and consider arbitrarily large delays, with the expected tradeoff in our treatment that the delay can no longer be unknown. The technical approach is based on using predictors that capture the effect of the delay on the system to compensate for it. We rigorously analyze the case when the delay is accurately known and show in simulation that our design is indeed robust to small variations when the delay is only approximately known. Unlike [28], here we consider imperfect signal transmission with event-triggering and time-varying delay both in the sensing and actuation. Further, given the well-known difficulties in the computation of predictor-feedback controllers, we here provide a detailed discussion of the numerical challenges that arise in the implementation of predictor feedback and effective solutions to resolve them. Finally, this chapter provides a complete and thorough technical treatment, including the proofs of all results, which are not available in [28].

5.2 Problem Statement

Consider the nonlinear system ("plant") with dynamics

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}_p(t)), \qquad t \geq 0, \text{ a.e.}, \tag{5.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. Our goal is to provide a state-feedback controller ensuring global asymptotic stability under the following challenges:

(i) **Actuation delay:** Let $\mathbf{u}(t)$ be the control signal generated by the controller. Actuation delay is modeled as

$$\mathbf{u}_p(t) = \mathbf{u}(\phi(t)), \quad t \geq 0, \tag{5.2}$$

where $t - \phi(t) > 0$ is the amount of time that it takes for a control action generated at time $\phi(t)$ to reach the plant/actuator. For instance, In the case of a constant actuation delay D, we have $\phi(t) = t - D$. This delay further requires an initial value $\{\mathbf{u}(t) \mid \phi(0) \leq t < 0\}$ on the control input for (5.1) to be well-defined.

(ii) **Sensing delay:** We allow the existence of a delay between the sensor and the controller such that at any time t, the controller may have access to $\mathbf{x}(s), s \leq \psi(t)$ (alternatively, $\mathbf{x}(t)$ takes $\psi^{-1}(t) - t$ seconds to reach the controller) for some delay function $\psi(t) \leq t$.

(iii) **Actuation event-triggering:** We aim to design opportunistic event-triggered controllers that do not require continuous updating of the control signals. This is motivated by practical concerns about the implementability of the controller in real time, and also by considerations about the efficient use of the available resources (to prevent, for instance, wear-and-tear of the actuator, or to accommodate bandwidth limitations when sensor, controller, and actuator are not co-located). We seek to design a controller that updates $\mathbf{u}(t)$ only at a sequence of discrete times $\{t_k\}_{k=0}^{\infty}$,

$$\mathbf{u}(t) = \mathbf{u}(t_k), \quad t \in [t_k, t_{k+1}), \quad k \geq 0. \tag{5.3}$$

(iv) **Sensing event-triggering:** We further allow for the possibility that the sensor can only send an event-triggered sequence of states $\{\mathbf{x}(\tau_\ell)\}_{\ell=0}^{\infty}$ to the controller. In this case, we let for simplicity that $\tau_0 = 0$, $t_0 = \psi^{-1}(0)$, and $\mathbf{u}(t)$ be arbitrarily set in $[0, t_0)$ as the controller has not received any state information yet.

In the sequel, we impose the following assumptions on the system dynamics.

Assumption 5.2.1. ***(Standing assumptions).***

(i) f is continuously differentiable, $f(0,0) = 0$, and (5.1) is forward complete (does not exhibit finite escape time) for all initial conditions and bounded inputs;

(ii) the initial control $\{\mathbf{u}(t) \mid \phi(0) \leq t < 0\}$ is given and continuously differentiable;

(iii) the delay function ϕ is continuously differentiable;

(iv) the delay functions ϕ and ψ are monotonically increasing so the argument of $\mathbf{u}(\phi(t))$ and $\mathbf{x}(\psi(t))$ do not go back in time;

(v) the origin of (5.1) is robustly globally asymptotically stabilizable in the absence of delays and with continuous sensing and actuation. Formally, there exists a globally Lipschitz feedback law $K : \mathbb{R}^n \to \mathbb{R}^m$, $K(\mathbf{0}) = \mathbf{0}$, that makes

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), K(\mathbf{x}(t)) + \mathbf{w}(t)), \tag{5.4}$$

ISS with respect to the additive input disturbance $\mathbf{w}$;

(vi) the ISS gain function ρ in (2.7b) satisfies $\int_0^1 \frac{\rho(r)}{r} < \infty$;

(vii) the delay function ϕ is known to the controller; on the other hand, ψ need not be known a priori or in full, but only a posteriori and at times when state is measured;

(viii) the delay function ϕ and its derivative are bounded, i.e., there exist $M_0 > 0$, $M_1 \geq 1$, and $0 < m_2 \leq 1$ such that

$$t - \phi(t) \leq M_0 \text{ and } m_2 \leq \dot{\phi}(t) \leq M_1, \ \forall t \geq 0; \tag{5.5}$$

(ix) the sensing triggering times $\{\tau_\ell\}_{\ell=0}^{\infty}$ are given (determined by the sensor independently of our design). In particular, the sensor ensures that $\{t_\ell\}_{\ell \geq 0} \cap [a, b]$ is finite for any $a, b < \infty$ (lack of Zeno behavior). □

Assumption 5.2.1(i)-(iv) are standard in predictor-based control of delay systems. Together with the piecewise-constant form of $\mathbf{u}_p$, Assumption 5.2.1(i) also ensures existence and uniqueness of solutions for (5.1). Assumption 5.2.1(v) on the availability of a feedback law in the absence of event-triggering and delays is also standard in event-triggered control. This allows us to focus on the challenges that arise by time delays and event-triggered control. Assumption (vi) is satisfied by $\rho(r) = r^c$ for any $c > 0$ and thus it is also satisfied by any $\rho \in \mathcal{K}$ such that $\rho(r) \leq r^c, r \in [0, \bar{r}]$ for some $c, \bar{r} > 0$ (e.g., $\rho(r) = \tan^{-1}(r)$, $\rho(r) = e^r - 1$, or $\rho(r) = \log(1 + r)$). Further, the a priori knowledge of ϕ in Assumption 5.2.1(vii) is most realistic in applications where the same control task is repeatedly executed and thus a data-driven estimate of future ϕ can be computed using its history. Moreover, note that Assumption 5.2.1(viii) is trivially satisfied for a constant delay ($\phi(t) = t - D$) with $M_0 = D$ and $M_1 = m_2 = 1$. Finally, Assumption 5.2.1(ix) is imposed for simplicity and to let us focus on the design of the actuation triggering times. In fact, the values of

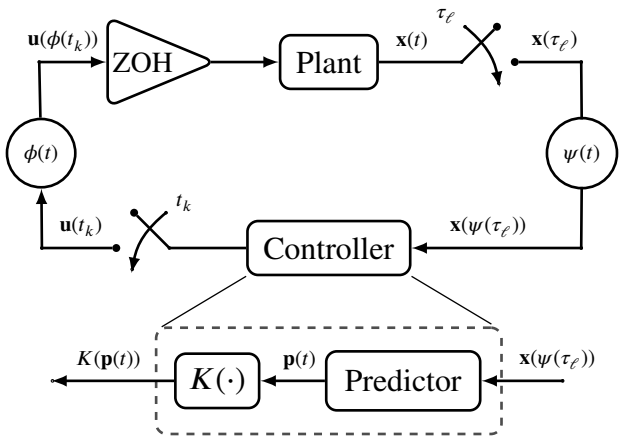


Figure 5.1: The considered networked control scheme with sensing and actuation delays and event-triggering (top) and the proposed predictor-based controller (bottom).

$\{\tau_\ell\}$ other than τ_0 are irrelevant theoretically but practically critical for stability, a point we discuss in detail in Sections 5.3.4 and 5.5.

The resulting networked control scheme is illustrated in Figure 5.1. Our considered problem is then as follows.

Problem 3. ***(Event-triggered stabilization under sensing and actuation delay).*** *Design the sequence of actuation triggering times*[1] $\{t_k\}_{k=1}^{\infty}$ *and the corresponding control values* $\{\mathbf{u}(t_k)\}_{k=0}^{\infty}$ *such that* $\{t_k\}_{k\geq 0} \cap [a,b]$ *is finite for any* $0 \leq a \leq b < \infty$ *and the closed-loop system* (5.1) *is globally asymptotically stable using the piecewise constant control* (5.3) *and the delayed information* $\{\mathbf{x}(\tau_\ell)\}_{\ell=0}^{\infty}$ *received, resp., at* $\{\psi^{-1}(\tau_\ell)\}_{\ell=0}^{\infty}$.[2] □

The requirement that $\{t_k\}_{k\geq 0} \cap [a,b]$ be finite for any $0 \leq a \leq b < \infty$ ensures the resulting design is implementable by avoiding finite accumulation points, i.e., Zeno behavior. We propose a solution to Problem 3 in the next section.

[1] Recall that $t_0 = \psi^{-1}(0)$ is fixed.

[2] We require that the control law is causal, i.e., t_k and $\mathbf{u}(t_k)$ depend only on the states $\{\mathbf{x}(\tau_\ell)\}$ that have reached the controller by the time t_k. While sampling may be modeled as a specific type of delay, we capture it with the prediction error $\mathbf{e}(t)$ (defined later). The values $\phi(t)$ and $\psi(t)$ only capture the delays in actuation and sensing, resp.

5.3 Event-Triggered Design and Analysis

In this section, we propose an event-triggered control policy to solve Problem 3. We start our analysis with the simpler case where the controller receives state feedback continuously (i.e., $\{\mathbf{x}(t)\}_{t=0}^{\infty}$ instead of $\{\mathbf{x}(\tau_\ell)\}_{\ell=0}^{\infty}$) without delays (i.e., $\psi(t) = t$), and later extend it to the general case.

5.3.1 Predictor Feedback Control for Time-Delay Systems

Here we review the continuous-time stabilization of the dynamics (5.1) by means of a predictor-based feedback control [15]. For convenience, we denote the inverse of ϕ by $\sigma(t) = \phi^{-1}(t)$, for all $t \geq 0$. The inverse exists since ϕ is strictly monotonically increasing. From (5.5), for all $t \geq \phi(0)$,

$$\dot{\sigma}(t) \leq M_2 \triangleq m_2^{-1}.$$

To compensate for the delay, at any time $t \geq \phi(0)$, the controller makes the following prediction of the future state of the plant,

$$\mathbf{p}(t) = \mathbf{x}(\sigma(t)) = \mathbf{x}([t]^+) + \int_{\phi([t]^+)}^{t} \dot{\sigma}(s) f(\mathbf{p}(s), \mathbf{u}(s)) ds. \tag{5.6}$$

This integral is computable by the controller since it only requires knowledge of the initial or current state of the plant (gathered from the sensors) and the history of $\mathbf{u}(t)$ and $\mathbf{p}(t)$, both of which are available to the controller. Nevertheless, for general nonlinear vector fields f, (5.6) may not have a closed-form solution and it has to be computed using numerical integration methods, cf. Re-

mark 5.5.1 below. As shown in Figure 5.1, the controller applies the control law K on the prediction $\mathbf{p}$ in order to compensate for the delay, i.e.,

$$\mathbf{u}(t) = K(\mathbf{p}(t)), \qquad t \geq 0. \tag{5.7}$$

The next result shows convergence for the closed-loop system.

Proposition 5.3.1. ***(Asymptotic stabilization by predictor feedback [15]).*** *Under Assumption 5.2.1, the closed-loop system* (5.1) *under the controller* (5.7) *is globally asymptotically stable, i.e., there exists $\beta \in \mathcal{KL}$ such that for any $\mathbf{x}(0) \in \mathbb{R}^n$ and bounded $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$, for all $t \geq 0$,*

$$\|\mathbf{x}(t)\| + \sup_{\phi(t)\leq\tau\leq t} \|\mathbf{u}(\tau)\| \leq \beta\Big(\|\mathbf{x}(0)\| + \sup_{\phi(0)\leq\tau\leq 0} \|\mathbf{u}(\tau)\|, t \Big).$$

5.3.2 Design of Event-triggered Control Law

Following Section 5.3.1, we let the controller make the prediction $\mathbf{p}(t)$ according to (5.6) for all $t \geq \phi(0)$. Since the controller can only update $\mathbf{u}(t)$ at discrete times $\{t_k\}_{k=0}^{\infty}$, it uses the piecewise-constant control (5.3) and assigns the control

$$\mathbf{u}(t_k) = K(\mathbf{p}(t_k)), \tag{5.8}$$

for all $k \geq 0$. In order to design the triggering times $\{t_k\}_{k=1}^{\infty}$, we use Lyapunov stability tools to determine when the controller has to update $\mathbf{u}(t)$ to prevent instability. We define the triggering

error for all $t \geq \phi(0)$ as

$$\mathbf{e}(t) = \begin{cases} \mathbf{p}(t_k) - \mathbf{p}(t) & \text{if } t \in [t_k, t_{k+1}) \text{ for } k \geq 0, \\ 0 & \text{if } t \in [\phi(0), t_0), \end{cases} \tag{5.9}$$

so that $\mathbf{u}(t) = K(\mathbf{p}(t) + \mathbf{e}(t))$, for $t \geq t_0$. Let

$$\mathbf{w}(t) = \mathbf{u}(t) - K(\mathbf{p}(t) + \mathbf{e}(t)), \qquad t \geq \phi(0), \tag{5.10}$$

where $\mathbf{w}(t) = 0$ for $t \geq t_0$ but $\mathbf{w}(t)$ is in general nonzero for $t \in [\phi(0), t_0)$. Computing $\mathbf{u}(\phi(t))$ from (5.10) and substituting it in (5.1), the closed-loop system can be written as

$$\dot{\mathbf{x}}(t) = f\left(\mathbf{x}(t), K\left(\mathbf{x}(t) + \mathbf{e}(\phi(t))\right) + \mathbf{w}(\phi(t))\right), \tag{5.11}$$

for all $t \geq 0$. Notice that (5.11) simplifies to [5, Eq. (3)] in the absence of delay ($\phi(t) = t$). Let $g(\mathbf{x}, \mathbf{w}) = f(\mathbf{x}, K(\mathbf{x}) + \mathbf{w})$ for all $\mathbf{x}, \mathbf{w}$. By Assumption 5.2.1(v), there exists a continuously differentiable function $S : \mathbb{R}^n \to \mathbb{R}$ and $\alpha_1, \alpha_2, \gamma, \rho \in \mathcal{K}_\infty$ such that

$$\alpha_1(\|\mathbf{x}(t)\|) \leq S(\mathbf{x}(t)) \leq \alpha_2(\|\mathbf{x}(t)\|), \tag{5.12}$$

and $(\mathcal{L}_g S)(\mathbf{x}, \mathbf{w}) \leq -\gamma(\|\mathbf{x}\|) + \rho(\|\mathbf{w}\|)$. Therefore, we have

$$\begin{aligned}(\mathcal{L}_f S)\big(\mathbf{x}(t), K\big(\mathbf{x}(t) + \mathbf{e}(\phi(t))\big) + \mathbf{w}(\phi(t))\big) \\ = (\mathcal{L}_g S)\big(\mathbf{x}(t), K\big(\mathbf{x}(t)+\mathbf{e}(\phi(t))\big)+\mathbf{w}(\phi(t))-K(\mathbf{x}(t))\big) \\ \leq -\gamma(\|\mathbf{x}(t)\|) + \rho\Big(\Big\|K\big(\mathbf{x}(t)+\mathbf{e}(\phi(t))\big)+\mathbf{w}(\phi(t))-K(\mathbf{x}(t))\Big\|\Big).\end{aligned} \tag{5.13}$$

Then, given Assumption 5.2.1(vi), define

$$V(t) = S(\mathbf{x}(t)) + \frac{2}{b}\int_0^{2L(t)} \frac{\rho(r)}{r} dr, \tag{5.14a}$$

$$L(t) = \sup_{t \leq \tau \leq \sigma(t)} e^{b(\tau - t)} \|\mathbf{w}(\phi(\tau))\|, \tag{5.14b}$$

and $b > 0$ is a design parameter. Note that the second term in (5.14a) may only be nonzero for $\phi(t) \in [\phi(0), t_0)$ since the system is open-loop over this interval (cf. (5.9),(5.10)). The next result establishes an upper bound on the time derivative of V.

Proposition 5.3.2. ***(Upper-bounding $\dot{V}(t)$).*** *For the system* (5.1) *under the control defined by* (5.3) *and* (5.8) *and the predictor* (5.6)*, we have*

$$\dot{V}(t) \leq -\gamma(\|\mathbf{x}(t)\|) - \rho(2L(t)) + \rho(2L_K\|\mathbf{e}(\phi(t))\|), \tag{5.15}$$

for all $t \neq \bar{t}$ and $V(\bar{t}^-) \geq V(\bar{t}^+)$, where L_K is the Lipschitz constant of K and $\bar{t} \in [0, \sigma(t_0)]$ is the smallest time such that $\mathbf{w}(\phi(t)) = \mathbf{0}$ for all $t > \bar{t}$.

Proof. Using (5.13), we have

$$\begin{aligned}\mathcal{L}_f S(\mathbf{x}(t)) &\leq -\gamma(\|\mathbf{x}(t)\|) + \rho\big(\|\mathbf{w}(\phi(t))\| + \|K(\mathbf{x}(t)+\mathbf{e}(\phi(t))) - K(\mathbf{x}(t))\|\big) \\ &\leq -\gamma(\|\mathbf{x}(t)\|) + \rho\big(\|\mathbf{w}(\phi(t))\| + L_K\|\mathbf{e}(\phi(t))\|\big) \\ &\leq -\gamma(\|\mathbf{x}(t)\|) + \rho(2\|\mathbf{w}(\phi(t))\|) + \rho(2L_K\|\mathbf{e}(\phi(t))\|). \end{aligned} \tag{5.16}$$

Since $e^{-b(t-\tau)}\|\mathbf{w}(\phi(\tau))\|$ is bounded for $\tau \in [t, \sigma(t)]$ and any $t \geq 0$ and $[t, \sigma(t)]$ has finite measure, the sup-norm in (5.14b) equals the limit of the corresponding p-norm as $p \to \infty$, i.e.,

$$L(t) = \lim_{n\to\infty} \left[\int_t^{\sigma(t)} e^{2nb(\tau-t)} \|\mathbf{w}(\phi(\tau))\|^{2n} d\tau \right]^{\frac{1}{2n}} \triangleq \lim_{n\to\infty} L_n(t).$$

In fact, it can be shown that this convergence is uniform over $[0, \underline{t}]$ for any $\underline{t} < \bar{t}$. Therefore, since $\dot{L}_n(t) = -bL_n(t) - \frac{L_n}{2n}\left(\frac{\|\mathbf{w}(\phi(t))\|}{L_n}\right)^{2n}$, $\frac{\|\mathbf{w}(\phi(t))\|}{L_n} < 1$ for $t \in [0, \underline{t}]$ and sufficiently large n and b, and $\underline{t} \in [0, \bar{t})$ is arbitrary, it follows from [29, Thm 7.17] that $\dot{L}(t) = -bL(t)$ for $t \in (0, \infty) \setminus \{\bar{t}\}$. Combining this and (5.16), we get

$$\begin{aligned}\dot{V}(t) &\leq -\gamma(\|\mathbf{x}(t)\|) + \rho(2\|\mathbf{w}(\phi(t))\|) + \rho(2L_K\|\mathbf{e}(\phi(t))\|) + \frac{2}{b}2\dot{L}(t)\frac{\rho(2L(t))}{2L(t)} \\ &\leq -\gamma(\|\mathbf{x}(t)\|) + \rho(2\|\mathbf{w}(\phi(t))\|) + \rho(2L_K\|\mathbf{e}(\phi(t))\|) - 2\rho(2L(t)).\end{aligned}$$

for $t \in (0, \infty) \setminus \{\bar{t}\}$. Equation (5.15) thus follows since $\|\mathbf{w}(\phi(t))\| \leq L(t)$ (c.f. (5.14b)) and the fact that ρ is strictly increasing. Finally, since $S(\mathbf{x}(t))$ is continuous, $L(\bar{t}^-) \geq 0$, and $L(\bar{t}^+) = 0$, we get $V(\bar{t}^-) \geq V(\bar{t}^+)$. □

Proposition 5.3.2 is the basis for our event-trigger design. Formally, we select $\theta \in (0, 1)$

and require

$$\rho(2L_K\|\mathbf{e}(\phi(t))\|) \leq \theta\gamma(\|\mathbf{x}(t)\|), \qquad t \geq 0,$$

which can be equivalently written as

$$\|\mathbf{e}(t)\| \leq \frac{\rho^{-1}(\theta\gamma(\|\mathbf{p}(t)\|))}{2L_K}, \qquad t \geq \phi(0). \tag{5.17}$$

Notice from (5.9) and the fact $t = 0$ that (5.17) holds on $[\phi(0), t_0]$. Equation (5.17) fully specifies the sequence of times $\{t_k\}_{k=1}^{\infty}$ and its dependence on the actuation delay. For each k, after each time t_k, the controller keeps evaluating (5.17) until it reaches equality. At this time, labeled t_{k+1}, the controller triggers the next event that sets $\mathbf{e}(t_{k+1}) = 0$ and maintains (5.17). Notice that "larger" γ and "smaller" ρ (corresponding to "stronger" input-to-state stability in (2.7)) are then more desirable, as they are intuitively expected to let the controller update $\mathbf{u}$ less often. Our ensuing analysis shows global asymptotic stability of the closed-loop system and the existence of a uniform lower bound on the inter-event times.

5.3.3 Convergence Analysis under Event-triggered Law

In this section we show that our event triggered law (5.17) solves Problem 3 by showing, in the following result, that the inter-event times are uniformly lower bounded (so, in particular, there is no finite accumulation point in time) and the closed-loop system achieves global asymptotic stability.

Theorem 5.3.3. ***(Uniform lower bound for the inter-event times and global asymptotic stability).***

Suppose that the class $\mathcal{K}_\infty$ function $\mathcal{G} : r \mapsto \gamma^{-1}(\rho(r)/\theta)$ is (locally) Lipschitz. For the system (5.1) *under the control* (5.8) *and the triggering condition* (5.17)*, the following hold:*

(i) there exists $\delta = \delta(\mathbf{x}(0), \{\mathbf{u}(t)\}_{t=\phi(0)}^{0}) > 0$ such that $t_{k+1} - t_k \geq \delta$ for all $k \geq 1$,

(ii) there exists $\beta \in \mathcal{KL}$ such that for any $\mathbf{x}(0) \in \mathbb{R}^n$ and bounded $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$, we have for all $t \geq 0$,

$$\|\mathbf{x}(t)\| + \sup_{\phi(t)\leq\tau\leq t} \|\mathbf{u}(\tau)\| \leq \beta\Big(\|\mathbf{x}(0)\| + \sup_{\phi(0)\leq\tau\leq 0} \|\mathbf{u}(\tau)\|, t \Big). \tag{5.18}$$

Proof. Let $[0, t_{\max})$ be the maximal interval of existence of the solutions of the closed-loop system. The proof involves three steps. First, we prove that (ii) holds for $t < t_{\max}$. Then, we show that (i) holds until $t_{\max}$, and finally that $t_{\max} = \infty$.

Step 1: From Proposition 5.3.2 and (5.17), we have

$$\begin{aligned} \dot{V}(t) &\leq -(1-\theta)\gamma(\|\mathbf{x}(t)\|) - \rho(2L(t)) \\ &\leq -\gamma_{\min}(\|\mathbf{x}(t)\| + L(t)), \qquad\qquad t \in [0, t_{\max}) \setminus \{\bar{t}\}, \end{aligned}$$

where $\gamma_{\min}(r) = \min\{(1-\theta)\gamma(r), \rho(2r)$ for all $r \geq 0$, so $\gamma_{\min} \in \mathcal{K}$. Also, note that

$$V(t) \leq \alpha_2(\|\mathbf{x}(t)\|) + \alpha_0(L(t)) \leq 2\alpha_{\max}(\|\mathbf{x}(t)\| + L(t)),$$

where $\alpha_{\max}(r) = \max\{\alpha_2(r), \alpha_0(r)\}$ and $\alpha_0(r) = \frac{2}{b}\int_0^{2r} \frac{\rho(s)}{s} ds$ for all $r \geq 0$. Since $\alpha_0, \alpha_2 \in \mathcal{K}_\infty$, we

have $\alpha_{\max} \in \mathcal{K}_\infty$, so $\alpha_{\max}^{-1} \in \mathcal{K}$. Hence,

$$\dot{V}(t) \leq -\alpha_{\min}(\alpha_{\max}^{-1}(V(t)/2)) \triangleq \overline{\alpha}(V(t)), \qquad t \in [0, t_{\max}) \setminus \{\bar{t}\},$$

where $\overline{\alpha} \in \mathcal{K}$. Therefore, using the Comparison Principle [30, Lemma 3.4], [30, Lemma 4.4], and $V(\bar{t}^-) \geq V(\bar{t}^+)$, there exists $\beta_1 \in \mathcal{KL}$ such that $V(t) \leq \beta_1(V(0), t)$, $t < t_{\max}$. Therefore,

$$\|\mathbf{x}(t)\| + L(t) \leq \beta_2(\|\mathbf{x}(0)\| + L(0), t), \qquad t < t_{\max},$$

where $\beta_2(r, s) = \alpha_{\min}^{-1}(\overline{\beta}(2\alpha_{\max}(r), s))$ for any $r, s \geq 0$. Note that $\beta_2 \in \mathcal{KL}$. Since we have

$$\sup_{\phi(t) \leq \tau \leq t} \|\mathbf{w}(\tau)\| \leq L(t) \leq e^{bM_0} \sup_{\phi(t) \leq \tau \leq t} \|\mathbf{w}(\tau)\|,$$

it then follows that

$$\|\mathbf{x}(t)\| + \sup_{\phi(t) \leq \tau \leq t} \|\mathbf{w}(\tau)\| \leq \beta_3\Big(\|\mathbf{x}(0)\| + \sup_{\phi(0) \leq \tau \leq 0} \|\mathbf{w}(\tau)\|, t\Big),$$

for all $t < t_{\max}$, where $\beta_3(r, s) = \beta_2(e^{bM_0} r, s)$. This inequality leads to (5.18) using the same steps as in [15, Lemmas 8.10, 8.11] (the only difference being the multiplicity of inputs).

Step 2: Equation (5.17) can be rewritten as

$$\|\mathbf{p}(t)\| \geq \gamma^{-1}\Big(\frac{\rho(2L_K\|\mathbf{e}(t)\|)}{\theta}\Big).$$

From step 1, the prediction $\mathbf{p}(t) = \mathbf{x}(\sigma(t))$ and its error $\mathbf{e}(t) = \mathbf{p}(t_k) - \mathbf{p}(t)$ are bounded. Therefore,

there exists $L_{\gamma^{-1}\rho/\theta} > 0$ such that for all $t \geq 0$,

$$\gamma^{-1}\Big(\frac{\rho(2L_K\|\mathbf{e}(t)\|)}{\theta}\Big) \leq 2L_{\gamma^{-1}\rho/\theta}L_K\|\mathbf{e}(t)\|.$$

where $L_{\gamma^{-1}\rho/\theta}$ is the Lipschitz constant of $\mathcal{G}$ on the compact set that contains $\{\mathbf{e}(t)\}_{t=0}^{t_{\max}}$. Hence, a sufficient (stronger) condition for (5.17) is

$$\|\mathbf{p}(t)\| \geq 2L_{\gamma^{-1}\rho/\theta}L_K\|\mathbf{e}(t)\|. \tag{5.19}$$

Note that (5.19) is only for the purpose of analysis and is *not* executed in place of (5.17). Clearly, if the inter-event times of (5.19) are lower bounded, so are the inter-event times of (5.17). Let $r(t) = \frac{\|\mathbf{e}(t)\|}{\|\mathbf{p}(t)\|}$ for any $t \geq 0$ (with $r(t) = 0$ if $\mathbf{p}(t) = 0$). For any $k \geq 0$, we have $r(t_k) = 0$ and $t_{k+1} - t_k$ is greater than or equal to the time that it takes for $r(t)$ to go from 0 to $\frac{1}{2L_{\gamma^{-1}\rho/\theta}L_K}$. Note that for any $t \geq 0$,

$$\begin{aligned}
\dot{r} &= \frac{d}{dt}\frac{\|\mathbf{e}\|}{\|\mathbf{p}\|} = \frac{d}{dt}\frac{(\mathbf{e}^T\mathbf{e})^{1/2}}{(\mathbf{p}^T\mathbf{p})^{1/2}} \\
&= \frac{(\mathbf{e}^T\mathbf{e})^{-1/2}\mathbf{e}^T\dot{\mathbf{e}}(\mathbf{p}^T\mathbf{p})^{1/2} - (\mathbf{p}^T\mathbf{p})^{-1/2}\mathbf{p}^T\dot{\mathbf{p}}(\mathbf{e}^T\mathbf{e})^{1/2}}{\mathbf{p}^T\mathbf{p}} \\
&= -\frac{\mathbf{e}^T\dot{\mathbf{p}}}{\|\mathbf{e}\|\|\mathbf{p}\|} - \frac{\|\mathbf{e}\|\mathbf{p}^T\dot{\mathbf{p}}}{\|\mathbf{p}\|^3} \leq \frac{\|\dot{\mathbf{p}}\|}{\|\mathbf{p}\|} + \frac{\|\mathbf{e}\|\|\dot{\mathbf{p}}\|}{\|\mathbf{p}\|^2} = (1+r)\frac{\|\dot{\mathbf{p}}\|}{\|\mathbf{p}\|},
\end{aligned}$$

where the time arguments are dropped for better readability. To upper bound the ratio $\|\dot{\mathbf{p}}(t)\|/\|\mathbf{p}(t)\|$, we have from (5.6) that $\dot{\mathbf{p}}(t) = \dot{\sigma}(t)f(\mathbf{p}(t), \mathbf{u}(t))$ for all $t \geq \phi(0)$. By continuous differentiability of f (which implies Lipschitz continuity on compacts) and global asymptotic sta-

bility of the closed loop system, there exists $L_f > 0$ such that

$$\begin{aligned}
\|\dot{\mathbf{p}}(t)\| &= \|\dot{\sigma}(t) f(\mathbf{p}(t), \mathbf{u}(t))\| \leq M_2 \|f(\mathbf{p}(t), K(\mathbf{p}(t) + \mathbf{e}(t)))\| \\
&\leq M_2 L_f \|(\mathbf{p}(t), K(\mathbf{p}(t) + \mathbf{e}(t)))\| \\
&\leq M_2 L_f (\|\mathbf{p}(t)\| + \|K(\mathbf{p}(t) + \mathbf{e}(t))\|) \\
&\leq M_2 L_f (\|\mathbf{p}(t)\| + L_K \|\mathbf{p}(t) + \mathbf{e}(t)\|) \\
&\leq M_2 L_f (1 + L_K) \|\mathbf{p}(t)\| + M_2 L_f L_K \|\mathbf{e}(t)\| \\
\Rightarrow \dot{r}(t) &\leq M_2 (1 + r(t))(L_f (1 + L_K) + L_f L_K |r(t)|).
\end{aligned}$$

Thus, using the Comparison Principle [30, Lemma 3.4], we have $t_{k+1} - t_k \geq \delta, k \geq 0$ where δ is the time that it takes for the solution of

$$\dot{r} = M_2 (1 + r)(L_f (1 + L_K) + L_f L_K r), \tag{5.20}$$

to go from 0 to $\frac{1}{2L_{\gamma^{-1}\rho/\theta} L_K}$.

Step 3: Since all system trajectories are bounded and $t_k \xrightarrow{k \to \infty} \infty$, we have $t_{\max} = \infty$, completing the proof. □

A particular corollary of Theorem 5.3.3 is that the proposed event-triggered law does not suffer from Zeno behavior, i.e., t_k accumulating to a finite point $t_{\max}$. Also, note that the lower bound δ in general depends on the initial conditions $\mathbf{x}(0)$ and $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$ through the Lipschitz constant $L_{\gamma^{-1}\rho/\theta}$.

5.3.4 Delayed and Event-Triggered Sensing

So far, we have not considered any delays in the availability of the sensing information about the plant state, which we consider next. Our treatment here shows that the above event-triggered controller with *the same triggering condition* (5.17)*, and with slight adjustments in the employed control and predictor signals,* globally asymptotically stabilizes the plant while maintaining the same lower bound on the inter-event times.

To address the general scenario in Problem 3, let

$$\bar{\ell} = \bar{\ell}(t) = \max\{\ell \geq 0 \mid \tau_\ell \leq \psi(t)\},$$

be the index of the last plant state available at the controller at time t. Then, the best estimate of $\mathbf{x}(\sigma(t))$ available to the controller, namely,

$$\mathbf{p}(t) = \mathbf{x}(\tau_{\bar{\ell}}) + \int_{\phi(\tau_{\bar{\ell}})}^{t} \dot{\sigma}(s) f(\mathbf{p}(s), \mathbf{u}(s)) ds, \ t \geq \psi^{-1}(0), \tag{5.21}$$

is used as the prediction signal in place of (5.6)[3]. Since $\mathbf{p}(t)$ is not available before $\psi^{-1}(0)$, the control signal (5.3), (5.8) is updated as

$$\mathbf{u}(t) = \begin{cases} K(\mathbf{p}(t_k)) & \text{if} \quad t \in [t_k, t_{k+1}), \ k \geq 0, \\ 0 & \text{if} \quad t \in [0, t_0), \end{cases} \tag{5.22}$$

where the first event time is now $t_0 = \psi^{-1}(0)$. We next provide the same guarantees as Theo-

[3]This only requires the controller to know $\psi(\tau_\ell)$ for every received state (not the full function ψ), which is realized by having a time-stamp for $\mathbf{x}(\tau_\ell)$.

rem 5.3.3.

Theorem 5.3.4. *Consider the plant dynamics* (5.1) *driven by the predictor-based event-triggered controller* (5.22) *with the predictor* (5.21) *and triggering condition* (5.17)*. Under Assumption 5.2.1, the closed-loop system is globally asymptotically stable, namely, there exists* $\beta \in \mathcal{KL}$ *such that* (5.18) *holds for all* $\mathbf{x}(0) \in \mathbb{R}^n$*, continuously differentiable* $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$*, and* $t \geq 0$*. Furthermore, there exists* $\delta = \delta(\mathbf{x}(0), \{\mathbf{u}(t)\}_{t=\phi(0)}^{0}) > 0$ *such that* $t_{k+1} - t_k \geq \delta$ *for all* $k \geq 0$*.*

Proof. For simplicity, let $U(t) = \sup_{\phi(t)\leq\tau\leq t} \|\mathbf{u}(t)\|$. Since the open-loop system exhibits no finite escape time behavior, the state remains bounded during the initial period $[0, t_0]$. Hence, for any $\mathbf{x}(0)$ and any $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$ there exists $\Xi > 0$ such that $\|\mathbf{x}(t)\| \leq \Xi$ for $t \in [0, t_0]$. Without loss of generality, Ξ can be chosen to be a class $\mathcal{K}$ function of $\|\mathbf{x}(0)\| + U(0)$. Thus,

$$\begin{aligned} \|\mathbf{x}(t)\| + U(t) &\leq \Xi(\|\mathbf{x}(0)\| + U(0)) + U(0) \\ &\leq \left[\Xi(\|\mathbf{x}(0)\| + U(0)) + U(0)\right] e^{-(t-t_0)}, \quad t \in [0, t_0]. \end{aligned} \tag{5.23}$$

As soon as the controller receives $\mathbf{x}(0)$ at t_0, it can estimate the state $\mathbf{x}(t)$ by simulating the dynamics (5.1), i.e.,

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t f(\mathbf{x}(s), \mathbf{u}(\phi(s)))ds. \tag{5.24}$$

This estimation is updated whenever a new state $\mathbf{x}(\tau_\ell)$ arrives and used to compute the predictor (5.6), which combined with (5.24) takes the form (5.21). Since the controller now has access to the same prediction signal $\mathbf{p}(t)$ as before, the same Lyapunov analysis as above holds for $[t_0, \infty)$.

Therefore, let $\hat{\beta} \in \mathcal{KL}$ be such that (5.18) holds for $t \geq t_0$. By (5.23),

$$\|\mathbf{x}(t)\| + U(t) \leq \hat{\beta}\big(\Xi(\|\mathbf{x}(0)\| + U(0)) + U(0), t - t_0\big) \quad t \geq t_0.$$

Therefore, (5.18) holds by choosing $\beta(r,t) = \max\left\{\hat{\beta}\big(\Xi(r) + r, t - t_0\big), \big[\Xi(r) + r\big]e^{-(t-t_0)}\right\}$. Finally, since the triggering condition (5.17) has not changed, $t_{k+1} - t_k \geq \delta, k \geq 0$ for the same $\delta > 0$ as in Theorem 5.3.3. □

Remark 5.3.5. ***(Separation of sensing and actuation delays).*** It is a standard practice in the literature to combine the sensing and actuation delays into a single quantity, i.e., "networked induced delays". This is in fact the basis of the predictor design in equation (23). However, in our treatment, it is beneficial to keep the two delays distinct since their sources are often physically distinct and the assumptions on the sensing delay ψ are significantly weaker than on the actuator delay ϕ (cf. Assumption 5.2.1). □

Remark 5.3.6. ***(Practical importance of feedback).*** While the controller can theoretically discard $\{\mathbf{x}(\tau_\ell)\}_{\ell=1}^{\infty}$ and rely on $\mathbf{x}(0)$ for estimating the state at all future times, closing the loop using the most recent state value $\mathbf{x}(\tau_{\bar{\ell}})$ is in practice critical for preventing the estimator (5.24) from drifting due to noise and un-modeled dynamics, even when the system dynamics are perfectly known. This is apparent, for instance, in Example 5.5.2 shown later, where facing the errors caused by the numerical approximation of the prediction signal. □

5.4 The Linear Case

In this section, we show how the general treatment of Section 5.3 is specialized and simplified if the dynamics (5.1) is linear, i.e, when we have

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(\phi(t)), \qquad t \geq 0, \tag{5.25}$$

subject to initial conditions $\mathbf{x}(0) \in \mathbb{R}^n$ and bounded $\{\mathbf{u}(t)\}_{t=\phi(0)}^{0}$. For simplicity, we restrict our attention to the perfect sensing case, as the generalization to sensing channels with time delay does not change the controller or stability guarantees (cf. Theorem 5.3.4). Assuming that the pair $(\mathbf{A}, \mathbf{B})$ is stabilizable, we can use pole placement to find a linear feedback law $\mathbf{K}$ that satisfies Assumption 5.2.1(v). Moreover, $\mathbf{p}(t)$ can be explicitly solved from (5.6) to obtain

$$\mathbf{p}(t) = e^{\mathbf{A}(\sigma(t)-[t]^+)}\mathbf{x}([t]^+) + \int_{\phi([t]^+)}^{t} \dot{\sigma}(s) e^{\mathbf{A}(\sigma(t)-\sigma(s))}\mathbf{B}\mathbf{u}(s)\,ds, \tag{5.26}$$

for all $t \geq \phi(0)$ and the closed-loop system takes the form

$$\dot{\mathbf{x}}(t) = (\mathbf{A} + \mathbf{B}\mathbf{K})\mathbf{x}(t) + \mathbf{B}\mathbf{w}(\phi(t)) + \mathbf{B}\mathbf{K}\mathbf{e}(\phi(t)).$$

Furthermore, given an arbitrary $\mathbf{Q} = \mathbf{Q}^T > \mathbf{0}$, the continuously differentiable function $S : \mathbb{R}^n \to \mathbb{R}$ is $S(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x}$, where $\mathbf{P} = \mathbf{P}^T > \mathbf{0}$ is the unique solution to the Lyapunov equation $(\mathbf{A}+\mathbf{B}\mathbf{K})^T\mathbf{P} + \mathbf{P}(\mathbf{A} + \mathbf{B}\mathbf{K}) = -\mathbf{Q}$. Clearly, (5.12) holds with $\alpha_1(r) = \lambda_{\min}(\mathbf{P})r^2$ and $\alpha_2(r) = \lambda_{\max}(\mathbf{P})r^2$. To

show (5.13), notice that using Young's inequality [31],

$$\mathcal{L}_f S(\mathbf{x}(t)) = -\mathbf{x}(t)^T \mathbf{Q}\mathbf{x}(t) + 2\mathbf{x}(t)^T \mathbf{P}\mathbf{B}(\mathbf{w}(\phi(t)) + \mathbf{K}\mathbf{e}(\phi(t))),$$

so (5.13) holds with $\gamma(r) = \frac{1}{2}\lambda_{\min}(\mathbf{Q})r^2$ and $\rho(r) = \frac{2\|\mathbf{P}\mathbf{B}\|^2}{\lambda_{\min}(\mathbf{Q})}r^2$. In this case, the trigger (5.17) takes the simpler form

$$\|\mathbf{e}(t)\| \leq \frac{\lambda_{\min}(\mathbf{Q})\sqrt{\theta}}{4\|\mathbf{P}\mathbf{B}\|\|\mathbf{K}\|}\|\mathbf{p}(t)\|. \tag{5.27}$$

In addition to the simplifications, we show next that the closed-loop system is globally exponentially stable in the linear case.

5.4.1 Exponential Stabilization under Event-triggered Control

We next show that, in the linear case, we obtain the stronger feature of global exponential stability, though this requires a slightly different Lyapunov-Krasovskii functional.

Theorem 5.4.1. ***(Exponential stability of the linear case).*** *The system* (5.25) *subject to the piecewise-constant closed-loop control* $\mathbf{u}(t) = \mathbf{K}\mathbf{p}(t_k)$, $t \in [t_k, t_{k+1})$, *with* $\mathbf{p}(t)$ *given in* (5.26) *and* $\{t_k\}_{k=1}^{\infty}$ *determined according to* (5.27) *satisfies*

$$\|\mathbf{x}(t)\|^2 + \int_{\phi(t)}^{t} \|\mathbf{u}(\tau)\|^2 d\tau \leq Ce^{-\mu t}\Big(\|\mathbf{x}(0)\|^2 + \int_{\phi(0)}^{0} \|\mathbf{u}(\tau)\|^2 d\tau\Big),$$

for some $C > 0$, $\mu = \frac{(2-\theta)\lambda_{min}(\mathbf{Q})}{4\lambda_{max}(\mathbf{P})}$, *and all* $t \geq 0$.

Proof. For $t \geq 0$, let $L(t) = \int_t^{\sigma(t)} e^{b(\tau-t)}\mathbf{w}(\phi(\tau))^2 d\tau$. One can see that $\dot{L}(t) = -\mathbf{w}(\phi(t))^2 - bL(t)$,

$t \geq 0$. Define $V(t) = \mathbf{x}(t)^T \mathbf{P}\mathbf{x}(t) + \frac{4\|\mathbf{PB}\|^2}{\lambda_{\min}(\mathbf{Q})} L(t)$. Therefore, using (5.27),

$$\begin{aligned}
\dot{V}(t) &= -\mathbf{x}(t)^T \mathbf{Q}\mathbf{x}(t) + 2\mathbf{x}(t)^T \mathbf{PBw}(\phi(t)) - \frac{4\|\mathbf{PB}\|^2 b}{\lambda_{\min}(\mathbf{Q})} L(t) \\
&\quad + 2\mathbf{x}(t)^T \mathbf{PBKe}(\phi(t)) - \frac{4\|\mathbf{PB}\|^2}{\lambda_{\min}(\mathbf{Q})} \mathbf{w}(\phi(t))^2 \\
&\leq -\frac{2-\theta}{4} \lambda_{\min}(\mathbf{Q}) \|\mathbf{x}(t)\|^2 - \frac{4\|\mathbf{PB}\|^2 b}{\lambda_{\min}(\mathbf{Q})} L(t) \leq -\mu V(t),
\end{aligned}$$

where $\mu = \min\left\{\frac{(2-\theta)\lambda_{\min}(\mathbf{Q})}{4\lambda_{\max}(\mathbf{P})}, b\right\} = \frac{(2-\theta)\lambda_{\min}(\mathbf{Q})}{4\lambda_{\max}(\mathbf{P})}$ if b is chosen sufficiently large. Hence, by the Comparison Principle [30, Lemma 3.4], we have $V(t) \leq e^{-\mu t} V(0)$, $t \geq 0$. Let $W(t) = \|\mathbf{x}(t)\|^2 + \int_{\phi(t)}^{t} \|\mathbf{u}(\tau)\|^2 d\tau$. From [15, Eq.(6-70),(6-88)], $c_1 W(t) \leq V(t) \leq c_2 W(t)$, for some $c_1, c_2 > 0$ and all $t \geq 0$. Hence, the result follows with $C = c_2/c_1$. □

From Theorem 5.4.1, the convergence rate μ depends both on the ratio $\frac{\lambda_{\min}(\mathbf{Q})}{\lambda_{\max}(\mathbf{P})}$ and the parameter θ. The former can be increased by placing the eigenvalues of $\mathbf{A} + \mathbf{BK}$ at larger negative values, though large eigenvalues result in noise amplification. Decreasing θ, however, comes at the cost of faster control updates, a trade-off we study next.

5.4.2 Optimizing the Sampling-Convergence Trade-off

In this section, we analyze the trade-off between sampling and convergence speed in our proposed event-triggered scheme. In general, it is clear from the Lyapunov analysis of Section 5.3 that more updates (smaller θ) hasten the decay of $V(t)$ and help the convergence. This trade-off becomes clearer in the linear case since explicit expressions are derivable for convergence rate and minimum inter-event times. To this end, we define two objective functions and formulate the trade-off as a multi-objective optimization. Let δ be the time that it takes for the solution of (5.20) to go

from 0 to $\frac{1}{2L_{\gamma^{-1}\rho/\theta}L_K}$. As shown in Section 5.3.3, the inter-event times are lower bounded by δ, so it can be used to roughly measure the cost of implementing the control scheme. Let

$$a = M_2 L_f L_K, \quad c = M_2 L_f (1 + L_K), \quad R = \frac{1}{2L_{\gamma^{-1}\rho/\theta}L_K},$$

where $L_f = \sqrt{2}(\|\mathbf{A}\| + \|\mathbf{B}\|)$, $L_K = \|\mathbf{K}\|$, and $L_{\gamma^{-1}\rho/\theta} = \frac{2\|\mathbf{PB}\|}{\lambda_{\min}(\mathbf{Q})\sqrt{\theta}}$. Then, the solution of (5.20) with initial condition $r(0) = 0$ is given by $r(t) = \frac{ce^{at} - ce^{ct}}{ae^{ct} - ce^{at}}$. Solving $r(\delta) = R$ for δ gives $\delta = \frac{\ln \frac{c+Ra}{c+Rc}}{a-c}$. The objective is to maximize δ and μ by tuning the optimization variables θ and $\mathbf{Q}$. For simplicity, let $\theta = \nu^2$ and $\mathbf{Q} = q\mathbf{I}_n$ where $\nu, q > 0$. Then,

$$\delta(\nu) = \frac{1}{a-c} \ln \frac{c + \frac{\nu}{\|\mathbf{P}_1\mathbf{B}\|\|\mathbf{K}\|}a}{c + \frac{\nu}{\|\mathbf{P}_1\mathbf{B}\|\|\mathbf{K}\|}c}, \quad \mu(\nu) = \frac{2 - \nu^2}{4\lambda_{\max}(\mathbf{P}_1)},$$

where $\mathbf{P}_1 = q^{-1}\mathbf{P}$ is the solution of the Lyapunov equation $(\mathbf{A} + \mathbf{BK})^T\mathbf{P}_1 + \mathbf{P}_1(\mathbf{A} + \mathbf{BK}) = -\mathbf{I}_n$. Figure 5.2(a) depicts δ and μ as functions of ν and illustrates the sampling-convergence trade-off.

To balance these two objectives, we define the aggregate objective function as a convex combination of δ and μ, i.e.,

$$J(\nu) = \lambda\delta(\nu) + (1 - \lambda)\mu(\nu),$$

where $\lambda \in [0, 1]$ determines the (subjective) relative importance of convergence rate and sampling. Notice that due to the difference between the (physical) units of δ and μ, one might multiply either one by a unifying constant, but we are not doing this as it leads to an equivalent optimization problem with a different λ. It is straightforward to verify that J is strongly convex and its unique maximizer is given by the positive real solution of $c_3\nu^3 + c_2\nu^2 + c_1\nu + c_0 = 0$ where $c_3 = a(1 - \lambda)$,

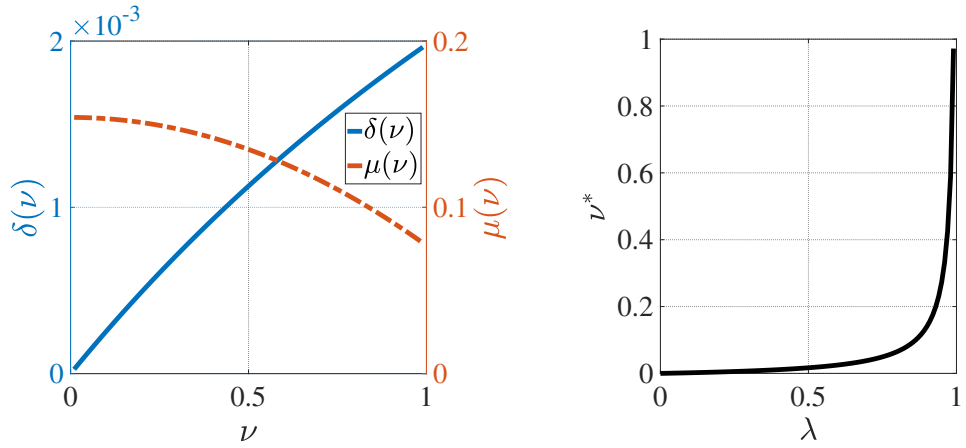


Figure 5.2: Sampling-convergence trade-off for event-triggered control of linear systems. **(Left)**, values of the lower bound of the inter-event times (δ) and exponential rate of convergence (μ) for different values of the optimization parameter ν for a third-order unstable linear system with $M_2 = 1$. **(Right)**, the unique maximizer ν^* of the aggregate objective function $J(\nu)$ for different values of the weighting factor λ. As λ goes from 0 to 1, more weight is given to the maximization of δ, which increases ν^*.

$c_2 = (a+c)\|\mathbf{P}_1\mathbf{B}\|\|\mathbf{K}\|(1-\lambda)$, $c_1 = c\|\mathbf{P}_1\mathbf{B}\|^2\|\mathbf{K}\|^2(1-\lambda)$, and $c_0 = -2\lambda_{\max}(\mathbf{P}_1)\|\mathbf{P}_1\mathbf{B}\|\|\mathbf{K}\|\lambda$.

Figure 5.2(b) illustrates the optimizer of the aggregate objective function $J(\nu)$ for different values of the weighting factor λ.

5.5 Simulations

Here we illustrate the performance of our event-triggered predictor-based design. Example 5.5.2 is a two-dimensional nonlinear system that satisfies all the hypotheses required to ensure global asymptotic convergence of the closed-loop system. Example 5.5.3 is a different two-dimensional nonlinear system which instead does not, but for which we observe convergence in simulation. We start by discussing some numerical challenges that arise because of the particular hybrid nature of our design, along with our approach to tackle them.

Remark 5.5.1. ***(Numerical implementation of event-triggered control law).*** The main challenge

in the numerical simulation of the proposed event-trigger law is the computation of the prediction signal $\mathbf{p}(t) = \mathbf{x}(\sigma(t))$. To this end, at least three methods can be used, as follows:line(i) *Open-loop*: One can solve $\dot{\mathbf{p}}(t) = \dot{\sigma}(t) f(\mathbf{p}(t), \mathbf{u}(t))$ directly starting from $\mathbf{p}(\phi(0)) = \mathbf{x}(0)$. The closed-loop system takes the form of a hybrid system (see, e.g., [32] for an introduction to hybrid systems) with flow map

$$\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(\phi(t))), \qquad t \geq 0, \tag{5.28a}$$

$$\dot{\mathbf{p}}(t) = \dot{\sigma}(t) f(\mathbf{p}(t), \mathbf{u}(t)), \qquad t \geq \phi(0), \tag{5.28b}$$

$$\dot{\mathbf{p}}_{tk}(t) = \mathbf{0}, \qquad t \geq t_0, \tag{5.28c}$$

$$\mathbf{u}(t) = K(\mathbf{p}_{tk}(t)), \qquad t \geq t_0, \tag{5.28d}$$

jump map $\mathbf{p}_{tk}([t_k]^+) = \mathbf{p}([t_k]^+)$, jump set $D = \left\{(\mathbf{x}, \mathbf{p}, \mathbf{p}_{tk}) \mid \|\mathbf{p}_{tk} - \mathbf{p}\| = \frac{\rho^{-1}(\theta\gamma(\|\mathbf{p}\|))}{2L_K}\right\}$, and flow set $C = \mathbb{R}^{3n} \setminus D$. This formulation is computationally efficient but, if the original system is unstable, it is prone to numerical instabilities. The reason, suggesting the name "open-loop", is that the $(\mathbf{p}, \mathbf{p}_{tk})$-subsystem is completely decoupled from the $\mathbf{x}$-subsystem. Therefore, as stated in Remark 5.3.6, if any mismatch occurs between $\mathbf{x}(t)$ and $\mathbf{p}(\phi(t))$ due to numerical errors, the $\mathbf{x}$-subsystem tends to become unstable, and this is not "seen" by the $(\mathbf{p}, \mathbf{p}_{tk})$-subsystem. (ii) *Semi-closed-loop*: One can add a feedback path from the $\mathbf{x}$-subsystem to the $(\mathbf{p}, \mathbf{p}_{tk})$ subsystem by computing $\mathbf{p}$ directly from (5.21) every time a new state value arrives (i.e., at every $\psi^{-1}(\tau_\ell)$). This requires a numerical integration of $f(\mathbf{p}(s), \mathbf{u}(s))$ over the "history" of $(\mathbf{p}, \mathbf{u})$ from $\phi(\tau_{\bar{\ell}})$ to t. This method is more computationally expensive but improves the numerical robustness. However, since we are still integrating over the history of $\mathbf{p}$, any mismatch in the prediction takes more time to die out, which may not be tolerable for an unstable system. (iii) *Closed-loop*: To further increase robustness, one

can solve the differential form in (5.28b) rather than the integral form in (5.21) every time a new state value arrives (i.e., at every $\psi^{-1}(\tau_\ell)$) from $\phi(\tau_{\bar{\ell}})$ to t with "initial" condition $\mathbf{p}(\phi(\tau_{\bar{\ell}})) = \mathbf{x}(\tau_{\bar{\ell}})$. This method is as computationally expensive as (ii) but is considerably more robust. This is therefore the recommended method for the numerical implementation of the proposed predictor-based controller and used below in Examples 5.5.2 and 5.5.3. □

Example 5.5.2. ***(Compliant nonlinear system).*** Consider the 2-dimensional system given by

$$f(\mathbf{x}, u) = \begin{bmatrix} x_1 + x_2 \\ \tanh(x_1) + x_2 + u \end{bmatrix}, \quad \phi(t) = t - \frac{(t-5)^2 + 2}{2(t-5)^2 + 2},$$

$$\tau_\ell = \ell \Delta_\tau, \quad \ell \geq 0, \qquad\qquad \psi(t) = t - D_\psi,$$

where Δ_τ and D_ψ are constants. This system satisfies Assumption 5.2.1 with the feedback law $K(\mathbf{x}) = -6x_1 - 5x_2 - \tanh(x_1)$, $S(\mathbf{x}) = \mathbf{x}^T\mathbf{P}\mathbf{x}$, and

$$L_f = \frac{\sqrt{2\sqrt{17} + 10}}{2}, \quad L_K = \sqrt{74}, \quad (M_1, m_2) = 1 \pm \frac{3\sqrt{3}}{16},$$
$$M_0 = 1, \quad \gamma(r) = \frac{\lambda_{\min}(\mathbf{Q})}{2} r^2, \quad \rho(r) = \frac{2\|\mathbf{P}\mathbf{B}\|^2}{\lambda_{\min}(\mathbf{Q})} r^2,$$

where $\mathbf{P} = \mathbf{P}^T > 0$ is the solution of $(\mathbf{A} + \mathbf{B}\mathbf{k})^T\mathbf{P} + \mathbf{P}(\mathbf{A} + \mathbf{B}\mathbf{k}) = -\mathbf{Q}$ for $\mathbf{A} = [1\ 1;\ 0\ 1]$, $\mathbf{B} = [0;\ 1]$, $\mathbf{k} = [-6\ \ -5]$, and arbitrary $\mathbf{Q} = \mathbf{Q}^T > 0$ (we use $\mathbf{Q} = \mathbf{I}$). A sample simulation result of this system is depicted in Figure 5.3(a). It is to be noted that for this example, (5.17) simplifies to $\|\mathbf{e}(t)\| \leq \overline{\rho}\|\mathbf{p}(t)\|$ with $\overline{\rho} = 0.022$, but the closed-loop system remains stable when increasing $\overline{\rho}$ about until 0.8 (Figure 5.3(b)).

While Theorem 5.3.3 guarantees the global asymptotic stability of the continuous-time sys-

tem, discretization accuracy/error plays an important role in its digital implementation. It is with this in mind that one should interpret Figure 5.3(c), where depending on the discretization scheme and the stepsize employed, the numerical approximation errors in computing the prediction signal, cf. Remark 5.5.1, make the evolution of the Lyapunov function V not monotonically decreasing (whereas we know from Theorem 5.3.3 that it is monotonically decreasing for the continuous-time system). We see that, at least for this example, the effect on the evolution of V is sensitive to both the order of discretization and the stepsize (h), and benefits more from decreasing the latter.

Stability is also critically dependent on the sensing sampling rate $1/\Delta_\tau$, as noted in Remark 5.3.6. We can also see from Figure 5.3(c) that the decay of V clearly deteriorates for large Δ_τ (insufficient sampling) due to (in this example only discretization) noise but can be made monotonic for sufficiently small Δ_τ. To visualize this effect on stability more systematically, we varied Δ_τ and D_ψ and computed $\|\mathbf{x}(25)\|$ as a measure of asymptotic stability. The average result is depicted in Figure 5.3(d) for 10 random initial conditions, showing that unlike our theoretical expectation, large Δ_τ and/or D_ψ result in instability even in the absence of noise because of the numerical error that degrades the estimation (5.24) over time (c.f. Remark 5.5.1). Nevertheless, taking the delays and sampling into account while designing the controller using the predictor-based scheme (5.8) significantly increases the robustness of the closed-loop system relative to a design that is oblivious to delays and sampling. As shown in [33], the asymptotic stability of the latter can only be guaranteed for this example *without actuation delays and event-triggering* if $\Delta_\tau + D_\psi \leq 7.1 \times 10^{-3}$ (given that, using the notation therein, we have $c_1 = 25, c_2 = 29/9, c_3 = 772$), which is more than two orders of magnitude more conservative than the empirical bound shown in Figure 5.3(d).

Finally, we have investigated the robustness of the closed-loop system to external disturbances (which are not theoretically included in our analysis but inevitably exist in practice). In

an event-triggered system, disturbances may lead to instability and/or Zeno behavior, cf. [27]. However, as shown in Figure 5.3(e-f), neither instability nor Zeno behavior occurs when adding (any strength of) the disturbance here, highlighting the practical relevance of the proposed event-triggered scheme. □

Example 5.5.3. ***(Non-compliant nonlinear system).*** Here, we consider an example that violates several of our assumptions. Let

$$f(\mathbf{x}, u) = (\mathbf{A} + \Delta\mathbf{A})\mathbf{x} + \mathbf{B}u + \mathbf{E}x_1^3, \quad \mathbf{E} = [0 \;\; 1]^T,$$
$$t - \phi(t) = D + a\sin(t),\ \tau_\ell = \ell\Delta_\tau,\ \psi(t) = t - \frac{1 - e^{-t}}{2},$$

where $\mathbf{A}$ and $\mathbf{B}$ are as in Example 5.5.2. The nominal delay D and nominal coefficient matrix $\mathbf{A}$ are known but their perturbations $a\sin(t)$ and $\Delta\mathbf{A}$ are not (the controller *assumes* $\phi(t) = t - D$ and $f(\mathbf{x}, u) = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{E}x_1^3$). We generate the elements of $\Delta\mathbf{A}$ independently from $\mathcal{N}(0, \sigma_A^2)$. Furthermore, in our simulation, the actual time that it takes for a sensor message $\mathbf{x}(\tau_\ell)$ to reach the controller is *not* the nominal delay $\psi^{-1}(\tau_\ell) - \tau_\ell$ but a random variable D_ℓ^ψ, where

$$E[D_\ell^\psi] = \psi^{-1}(\tau_\ell) - \tau_\ell, \quad \mathrm{Var}(D_\ell^\psi) = \sigma_\psi > 0.$$

This serves to illustrate how the delay function ψ (and similarly ϕ), though being continuous and deterministic in our treatment, can be used to compensate for (in addition to physical sensor lag) computation and communication delays that are discrete and stochastic in nature[4].

Moreover, $K(\mathbf{x}) = -6x_1 - 5x_2 - x_1^3$ makes the closed-loop system ISS but is not globally

[4]Since the triggering times τ_ℓ are themselves random and vary from execution to execution, the function ψ is defined for all t even though only the discrete sequence $\{\psi^{-1}(\tau_\ell)\}$ is relevant for each execution.

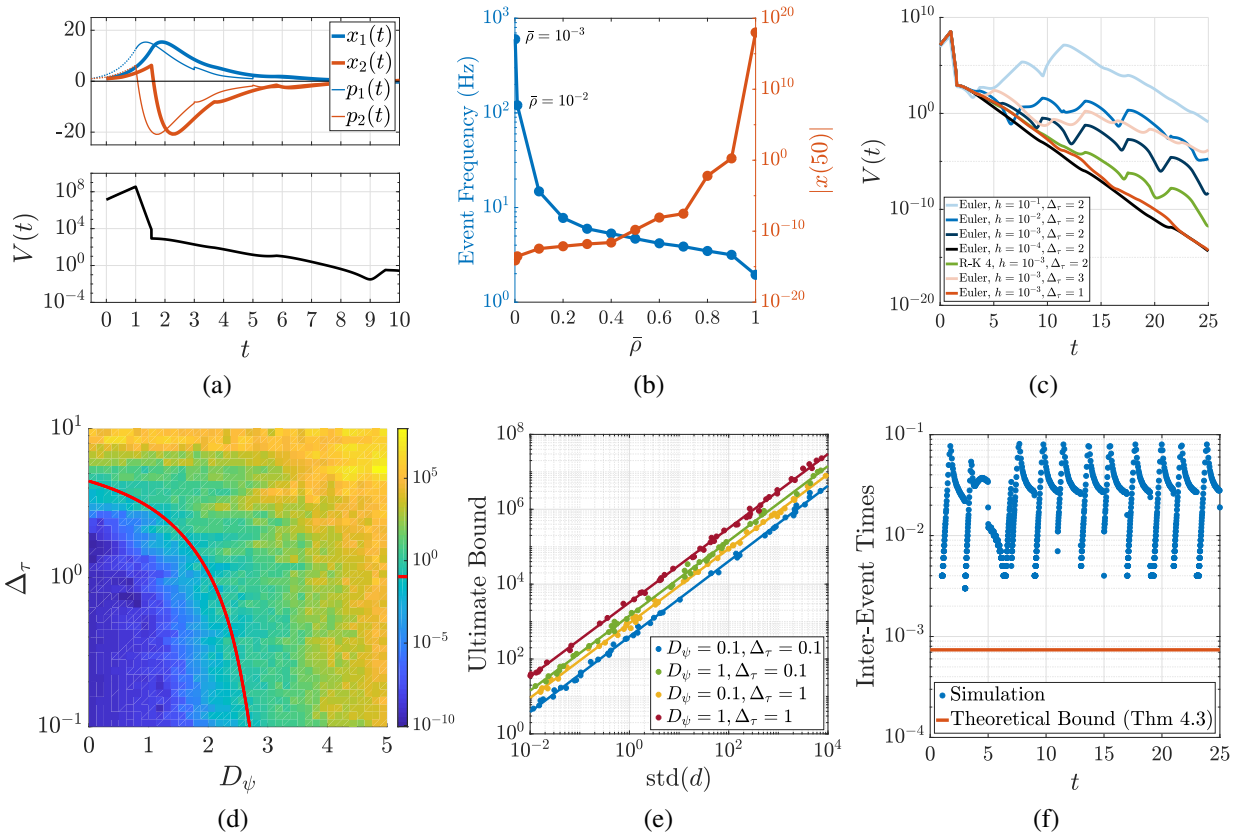


Figure 5.3: Simulation results of the compliant system in Example 5.5.2. Unless otherwise stated, we use $\mathbf{x}(0) = (1, 1)$, $\theta = 0.5$, $b = 10$, $\Delta_\tau = 2$, $D_\psi = 1$, and Euler discretization with $h = 10^{-3}$. **(a)** Sample trajectories. The dotted portion of $\mathbf{p}(t)$ corresponds to the times $[\phi(0), \psi^{-1}(0))$ and is plotted only for illustration purposes (not used by the controller). **(b)** The event-frequency and average of $\|\mathbf{x}(50)\|$ over 100 random initial conditions as a function of $\bar{\rho}$. **(c)** The effect of discretization and state sampling on stability. While stepsize h and sampling rate $1/\Delta_\tau$ have a strong impact on stability (blue and red curves, respectively), the effect of discretization order is less significant (green curve, 4th order Runge-Kutta). **(d)** Heat map of the average of $\|\mathbf{x}(25)\|$ over 10 random initial conditions drawn from standard normal distribution. The red line shows an approximate border of stability. **(e-f)** Numerical verification of the robustness of the proposed event-triggered controller to additive disturbances for the system of Example 5.5.2. Here, we augment (5.1) such that $\dot{\mathbf{x}} = f(\mathbf{x}, u_p) + \mathbf{d}$ where $\mathbf{d}$ is zero-mean, white, and Gaussian. **(e)** The estimate of the ultimate bound of state ($\max_{i=1,2} \limsup_{t\to\infty} \|x_i(t)\|$) for varying standard deviation of d_1 and d_2 (which are equal and denoted by std($\mathbf{d}$)) . The value of the ultimate bound depends on sampling delay and frequency, but the state always remained bounded for bounded disturbances and the best linear fit always has a slope $\simeq 1$, a behavior akin to globally input-to-state stable linear systems. **(f)** The inter-event times $\{t_{k+1} - t_k\}_{k\geq 0}$ for std($\mathbf{d}$) = 1. We highlight that unlike [34], the minimum inter-event time is lower bounded by δ given in Theorem 5.3.3 irrespective of the existence or strength of disturbance (as long as $\Delta_\tau > \delta$). This is due to the fact that sensing only occurs at discrete time instances $\{\tau_\ell\}$, making the controller oblivious to disturbance over each Δ_τ period. This may in principle lead to instability ($\|\mathbf{x}\| \to \infty$) but we see from (e) that, at least here, this is not the case.

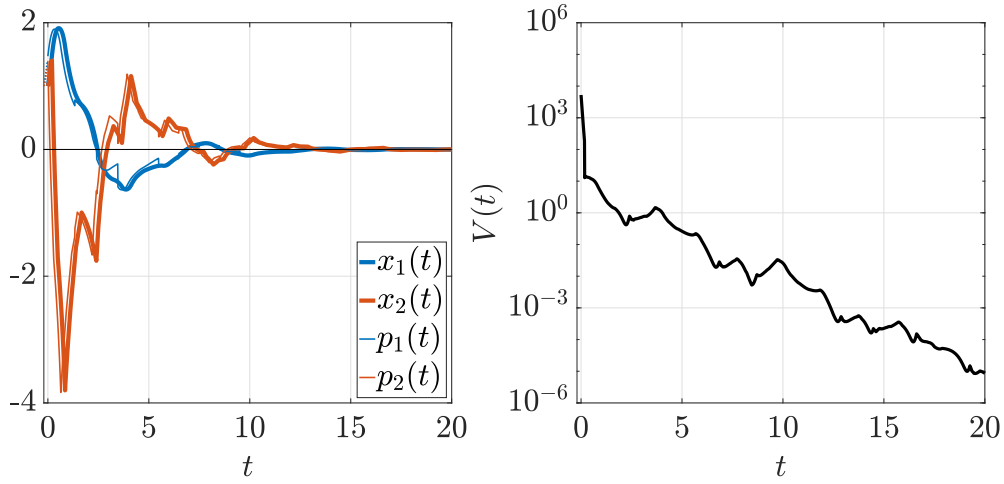


Figure 5.4: Simulation of the non-compliant system in Example 5.5.3. We have used $\mathbf{x}(0) = (1, 1)$, $\theta = 0.5$, $b = 10$, $a = 0.01$, $D = 0.2$, $\Delta_\tau = 1$, $\mu_\psi = 0.1$, $\sigma_\psi = \sigma_A = 0.02$, triggering condition $\|\mathbf{e}(t)\| \leq 0.5\|\mathbf{p}(t)\|$, and Euler discretization of the continuous-time dynamics with $h = 10^{-2}$.

Lipschitz, and the zero-input system exhibits finite escape time. The simulation results of this example are illustrated in Figure 5.4. It can be seen that although V is significantly non-monotonic, the event-triggered controller is able to stabilize the system. While a thorough investigation of the stability of the resulting stochastic dynamical system reaches far beyond our theoretical guarantees, this example suggests that the proposed controller is robust to small violations of its assumptions and is thus applicable to a wider class of systems than those satisfying Assumption 5.2.1. □

Acknowledgements: This chapter is taken, in part, from the work which has been submitted for publication as "Event-triggered stabilization of nonlinear systems with time-varying sensing and actuation delay" by E. Nozari, P. Tallapragada, and J. Cortés in *Automatica*. The dissertation author was the primary investigator and author of this paper.

Chapter Bibliography

[1] C. G. Cassandras and S. Lafortune, *Introduction to Discrete-Event Systems*, 2nd ed. Springer, 2007.

[2] L. Zou, Z. D. Wang, and D. H. Zhou, "Event-based control and filtering of networked systems: A survey," *International Journal of Automation and Computing*, vol. 14, no. 3, pp. 239–253, 2017.

[3] H. Kopetz, *Operating Systems of the 90s and Beyond: International Workshop Proceedings*. Springer Berlin Heidelberg, 1991, ch. Event-triggered versus time-triggered real-time systems, pp. 86–101.

[4] K. J. Åström and B. M. Bernhardsson., "Comparison of Riemann and Lebesgue sampling for first-order stochastic systems," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, Dec. 2002, pp. 2011–2016.

[5] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1680–1685, 2007.

[6] X. Wang and M. D. Lemmon, "Event-triggering in distributed networked control systems," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 586–601, 2011.

[7] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *IEEE Conf. on Decision and Control*, Maui, HI, 2012, pp. 3270–3285.

[8] M. Abdelrahim, R. Postoyan, J. Daafouz, and D. Nešić, "Robust event-triggered output feedback controllers for nonlinear systems," *Automatica*, vol. 75, pp. 96–108, 2017.

[9] O. J. M. Smith, "A controller to overcome dead time," *ISA Transactions*, vol. 6, no. 2, pp. 28–33, 1959.

[10] D. Q. Mayne, "Control of linear systems with time delay," *Electronics Letters*, vol. 4, no. 20, pp. 439–440, October 1968.

[11] A. Manitius and A. Olbrot, "Finite spectrum assignment problem for systems with delays," *IEEE Transactions on Automatic Control*, vol. 24, no. 4, pp. 541–552, Aug 1979.

[12] M. T. Nihtila, "Finite pole assignment for systems with time-varying input delays," in *Decision and Control, Proceedings of the 30th IEEE Conference on*, vol. 1, Dec 1991, pp. 927–928.

[13] M. Krstic, *Delay Compensation for Nonlinear, Adaptive, and PDE Systems*, 1st ed., ser. Systems & Control: Foundations & Applications. Birkhäuser, 2009.

[14] I. Karafyllis and M. Krstic, "Nonlinear stabilization under sampled and delayed measurements, and with inputs subject to delay and zero-order hold," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1141–1154, May 2012.

[15] N. Bekiaris-Liberis and M. Krstic, *Nonlinear Control Under Nonconstant Delays*, ser. Advances in Design and Control. SIAM, 2013.

[16] L. Mirkin, "On the approximation of distributed-delay control laws," *Systems & Control Letters*, vol. 51, no. 5, pp. 331–342, 2004.

[17] Q. C. Zhong, "On distributed delay in linear control laws-part i: discrete-delay implementations," *IEEE Transactions on Automatic Control*, vol. 49, no. 11, pp. 2074–2080, Nov 2004.

[18] X. M. Zhang, Q. L. Han, and B. L. Zhang, "An overview and deep investigation on sampled-data-based event-triggered control and filtering for networked systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 4–16, 2017.

[19] J. Chen, S. Meng, and J. Sun, "Stability analysis of networked control systems with aperiodic sampling and time-varying delay," *IEEE Transactions on Cybernetics*, vol. 47, no. 8, pp. 2312–2320, 2017.

[20] A. Selivanov and E. Fridman, "Predictor-based networked control under uncertain transmission delays," *Automatica*, vol. 70, pp. 101–108, 2016.

[21] ——, "Observer-based input-to-state stabilization of networked control systems with large uncertain delays," *Automatica*, vol. 74, pp. 63–70, 2016.

[22] X. Ge and Q. L. Han, "Distributed event-triggered H_∞ filtering over sensor networks with communication delays," *Information Sciences*, vol. 291, no. Supplement C, pp. 128–142, 2015.

[23] E. Garcia and P. J. Antsaklis, "Model-based event-triggered control for systems with quantization and time-varying network delays," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 422–434, 2013.

[24] L. Hetel, J. Daafouz, and C. Iung, "Stabilization of arbitrary switched linear systems with unknown time-varying delays," *IEEE Transactions on Automatic Control*, vol. 51, no. 10, pp. 1668–1674, Oct 2006.

[25] W. Wu, S. Reimann, D. Görges, and S. Liu, "Suboptimal event-triggered control for time-delayed linear systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1386–1391, May 2015.

[26] L. Li, X. Wang, and M. D. Lemmon, "Stabilizing bit-rate of disturbed event triggered control systems," in *Proceedings of the 4th IFAC Conference on Analysis and Design of Hybrid Systems*, Eindhoven, Netherlands, June 2012, pp. 70–75.

[27] V. S. Dolk, D. P. Borgers, and W. P. M. H. Heemels, "Output-based and decentralized dynamic event-triggered control with guaranteed $\mathcal{L}_p$-gain performance and Zeno-freeness," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 34–49, 2017.

[28] E. Nozari, P. Tallapragada, and J. Cortés, "Event-triggered control for nonlinear systems with time-varying input delay," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016, pp. 495–500.

[29] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976.

[30] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.

[31] W. H. Young, "On classes of summable functions and their Fourier series," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 87, no. 594, pp. 225–229, 1912.

[32] R. Goebel, R. G. Sanfelice, and A. R. Teel, *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press, 2012.

[33] F. Mazenc, M. Malisoff, and T. N. Dinh, "Robustness of nonlinear systems with respect to delay and sampling of the controls," *Automatica*, vol. 49, no. 6, pp. 1925–1931, 2013.

[34] D. P. Borgers and W. P. M. H. Heemels, "Event-separation properties of event-triggered control systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 10, pp. 2644–2656, 2014.

Chapter 6

Time-Varying Control Scheduling in Complex Dynamical Networks

The ability to control network dynamics in various application domains is not only a theoretically challenging problem but also a barrier to fundamental breakthroughs across science and engineering. In this chapter, we focus our attention to the limitation of actuation resources in the control of networked dynamical systems (the dual analysis would apply to limitations of sensing resources). While multiple studies have addressed various aspects of this problem, several fundamental questions remain unanswered, including to what extent the capability of controlling a different set of nodes over time can improve network controllability. Most of the existing studies and practical control methods limit their focus to time-invariant control schedules (TICS). This is both due to their simplicity and the fact that the benefits of time-varying control schedules (TVCS) have remained largely uncharacterized.

We consider networks with linear and discrete-time dynamics and analyze the role of network structure in TVCS. First, we show that TVCS can significantly enhance network controllability

over TICS both in small and large networks. Through the analysis of a scale-dependent notion of nodal centrality, we then show that optimal TVCS involves the actuation of the most central nodes at appropriate spatial scales at all times. Consequently, it is the scale-heterogeneity of the central nodes in a network that determine whether, and to what extent, TVCS outperforms conventional policies based on TICS. Here, scale-heterogeneity of a network refers to how diverse the central nodes of the network are at different spatial (local vs. global) scales. Several analytical results and case studies support and illustrate this relationship.

6.1 Prior Work

Controllability of a dynamical network (i.e., a network that supports the temporal evolution of a well-defined set of nodal *states*) is classically defined as the possibility of steering its state arbitrarily around the state space through the application of external inputs to (i.e., *actuation* of) certain *control nodes* [1]. This raises a fundamental question: how does the choice of control nodes affect network controllability? Hereafter, we refer to this as the *control scheduling problem* [2–4]. Notice that in this classical setting, attention is only paid to the *possibility* of arbitrarily steering the network state, but not to the *difficulty and energy cost* of doing so. This has motivated the introduction of several controllability metrics to quantify the required effort in the control scheduling problem [5–9]. While a comprehensive solution has remained elusive, these works have collectively revealed the role of several factors in the control scheduling problem such as the network size and structure [6], nodal dynamics [3] and centralities [2, 7], the number of control nodes [6], and the choice of controllability metric [8]. This problem has also close connections with the optimal sensor scheduling problem, see, e.g. [10–13] and the references therein.

The majority of the above literature, however, implicitly relies on the assumption of time-invariant control schedules (TICS), namely, that the control node(s) is fixed over time. Depending on the specific network structure, this assumption may come at the expense of a significant limitation on its controllability, especially for large-scale systems where distant nodes inevitably exist relative to any control node. Intuitively, the possibility of time-varying control schedules (TVCS), namely, the ability to control different nodes at different times, allows for targeted interventions at different network locations and can ultimately decrease the control effort to accomplish a desired task. On the other hand, from a practical standpoint, the implementation of TVCS requires the ability to geographically relocate actuators or the presence of actuation mechanisms at different, ideally all, network nodes, and more sophisticated control policies. This leads to a critical trade-off between the benefits of TVCS and its implementation costs which has not received enough, if any, attention in the literature.

The significant potential of time-varying schedules for control (and also sensing, which has a dual interpretation to control) has led to the design of (sub)optimal sensor [14, 15] and control [16,17] scheduling algorithms in recent years. While constituting a notable leap forward and the benchmark for the methods developed in this chapter, these works are oblivious to the fundamental question of whether, and to what extent, TVCS provides an improvement in network controllability compared to TICS. Our previous work [18] has studied the former question (i.e., whether TVCS provides *any* improvement over TICS) in the case of undirected networks, but did not consider directed networks or, more importantly, addressed the latter question of how large the relative improvement in network controllability is. Given the trade-off between benefits and costs of TVCS, a clear answer to this question is vital for the practical application of TVCS in real-world complex networks.

In this chapter, we address these two questions in the context of discrete-time linear dynamics evolving over directed networks. Since the implementation costs of TVCS are greatly domain-specific and do not follow any common pattern of dependence on the control schedule, we here provide an in-depth analysis of the benefits of TVCS. This provides the necessary information for comparison with the costs of implementing TVCS in any specific application in order to decide between TICS and TVCS.

6.2 Problem Statement

We consider a network of n nodes that communicate over a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E}, \mathbf{A})$ that is in general weighted and directed (see Appendix 6.A for methods of obtaining $\mathbf{A}$ from network connectivity structure). Each node i has a *state* value $x_i \in \mathbb{R}$ that evolves over time through the interaction of node i with its neighbors in $\mathcal{G}$ and an external control u. Assuming that these interactions are linear and time-invariant, we have

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}(k)u(k), \quad k \in \{0, \dots, K-1\}, \tag{6.1}$$

where $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ is the network state, $u(k) \in \mathbb{R}$ is the control input, $\mathbf{b}(k) \in \mathbb{R}^n$ is the *time-varying* input vector, and K is the time horizon. For simplicity of exposition, we consider only one control input at a time, but the discussion is generalizable to multi-input networks (cf. Appendix 6.E). Define

$$\iota_k \in \mathcal{N}, \tag{6.2}$$

to be the index of the node to which the control signal $u(k)$ is applied at time k. Then, $\mathbf{b}(k)$ is equal to the ι_k'th column of the identity matrix. For the sake of simplicity, we here assume that all the network nodes are actuatable, so $\iota_k \in \mathcal{N}$. If a subset of nodes are *latent*, (i.e., not actuatable), further challenges arise and thus we postpone the analysis of this case to Section 6.3.4.

The dynamical network (6.1) is *controllable* if its state can be steered from arbitrary $\mathbf{x}(0) = \mathbf{x}_0$ to arbitrary $\mathbf{x}(K) = \mathbf{x}_f$ using the control input $\{u(k)\}_{k=0}^{K-1}$ or, equivalently, if the *controllability Gramian*

$$\mathcal{W}_K = \sum_{k=0}^{K-1} \mathbf{A}^k \mathbf{b}(K-1-k)\mathbf{b}(K-1-k)^T (\mathbf{A}^T)^k, \tag{6.3}$$

is nonsingular [19]. In general, the eigenvalues of $\mathcal{W}_K$ determine how large the *unit-energy reachability set* (the set of states $\mathbf{x}_f$ that can be reached from the origin $\mathbf{x}_0 = \mathbf{0}$ using controls with unit energy) is (cf. Appendix 6.B for derivation). Therefore, various measures of controllability based on the eigenvalues of $\mathcal{W}_K$ have been proposed, most notably $\mathrm{tr}(\mathcal{W}_K)$, $\mathrm{tr}(\mathcal{W}_K^{-1})^{-1}$, $\det(\mathcal{W}_K)$, $\lambda_{\min}(\mathcal{W}_K)$. Each metric has its own benefits and limitations, on which we elaborate more in the following.

Assume, for now, that $f(\mathcal{W}_K) \geq 0$ is any of the aforementioned controllability measures. In *optimal control scheduling*, we seek to choose the control nodes $\{\iota_k\}_{k=0}^{K-1}$ (or, equivalently, $\{\mathbf{b}(k)\}_{k=0}^{K-1}$) optimally. The conventional approach in the literature [2–9] is to assume a constant

control node, thus called the *time-invariant control scheduling (TICS) problem*:

$$\text{TICS:} \qquad \max_{\iota_0,\ldots,\iota_{K-1}\in\mathcal{N}} \quad f(\mathcal{W}_K) \tag{6.4a}$$

$$\text{s.t.} \qquad \iota_0 = \cdots = \iota_{K-1} \tag{6.4b}$$

The main advantage of TICS is its simplicity, from theoretical, computational, and implementation perspectives. However, this simplicity comes at a possibly significant cost in terms of network controllability, compared to the case where the control nodes $\{\iota_k\}_{k=0}^{K-1}$ are independently chosen, namely,

$$\text{TVCS:} \qquad \max_{\iota_0,\ldots,\iota_{K-1}\in\mathcal{N}} \quad f(\mathcal{W}_K). \tag{6.5}$$

This approach, namely, time-varying control scheduling (TVCS), is at least as good as TICS, but has the potential to improve network controllability significantly. Figure 6.1(a-b) illustrates a small network of $n = 5$ nodes together with the optimal values of equations (6.4) and (6.5) and the relative advantage of TVCS over TICS, defined as

$$\chi = \frac{f_{\max}^{\text{TV}} - f_{\max}^{\text{TI}}}{f_{\max}^{\text{TI}}}. \tag{6.6}$$

Three observations are worth highlighting. First, the value of χ is extremely dependent on the choice of controllability measure f, and different choices lead to orders of magnitude change in χ. Second, the relative advantage of TVCS over TICS is significant for all choices of the controllability measure, with the minimum improvement of $\chi = 35\%$ for the choice of $f(\cdot) = \text{tr}(\cdot)$. The fact that

$f(\cdot) = \mathrm{tr}(\cdot)$ results in the smallest value of χ relative to other measures is consistently observed in synthetic and real-world networks, and stems from the fact that $\mathrm{tr}(\mathcal{W}_K)$ has the smallest sensitivity (greatest robustness) to the choice of control schedule. Finally, even with optimal TVCS, $\lambda_{\min}(\mathcal{W}_K)$ is orders of magnitude less than 1, indicating the inevitable existence of very hard-to-reach directions in the state space. This shows that efficient controllability cannot be maintained in all directions in the state space even using TVCS and even in very small networks with control over $1/5 = 20\%$ of the nodes. Except for $\mathrm{tr}(\mathcal{W}_K)$, all the measures rely heavily on this least-controllable direction, while $\mathrm{tr}(\mathcal{W}_K)$ trades this off for improved controllability in the most efficient directions in the state space. See Appendix 6.B for further discussion of this tradeoff.

Despite the significant increase in size and complexity, the same core principles outlined above apply to controllability of real-world networks. The large size of these networks, however, imposes new constraints on the choice of the controllability measure f that make the use of $f(\cdot) = \lambda_{\min}(\cdot)$, $\mathrm{tr}((\cdot)^{-1})^{-1}$, and $\det(\cdot)$ numerically infeasible and theoretically over-conservative, as discussed in detail in Appendix 6.B. As a result, we resort to the particular choice of controllability measure

$$f(\mathcal{W}_K) = \mathrm{tr}(\mathcal{W}_K), \tag{6.7}$$

for networks beyond $n \simeq 15$. Since this measure has the smallest sensitivity to the choice of $\{\imath_k\}_{k=0}^{K-1}$ (Figure 6.1(b)), we expect any network that benefits from TVCS using the choice of equation (6.7) to also benefit from it using other Gramian-based measures (while the converse is not necessarily true, i.e., there are networks that significantly benefit from TVCS using other measures but show no benefit in terms of $\mathrm{tr}(\mathcal{W}_K)$). Figure 6.1(c) illustrates an air transportation network among the

busiest airports in the United States, comprising of $n = 500$ nodes. Using (6.7), we see $\chi \simeq 20\%$ improvement in controllability, verifying our expectation about the benefits of TVCS.

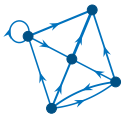
In spite of this potential benefit, TVCS has usually higher computational and implementation costs. These include the higher computational cost of computing the optimal TVCS, and that of installing an actuator at several (ideally all) nodes of the network. Further, not all networks benefit from TVCS alike. A simple directed chain network with the same size as that of Figure 6.1(a) gains absolutely no benefit from TVCS, independently of the choice of f (Figure 6.1(d-e)). Similarly, $\chi = 0$ is also observed in larger, complex networks, indicating that the optimal TVCS and the optimal TICS are the same (Figure 6.1(f)).

These observations collectively raise a fundamental question that constitutes the main problem studied in this chapter. Before formally stating the problem, we need a definition for ease of reference.

Definition 6.2.1. ***(Class $\mathcal{V}$ and $\mathcal{I}$ networks).*** Consider a dynamical network described by (6.1) and the measure χ introduced in (6.6). We say that the network belongs to class $\mathcal{V}$ if it has $\chi > 0$ and we say it belongs to class $\mathcal{I}$ otherwise ($\chi = 0$). □

In words, class $\mathcal{V}$ networks are those that benefit from TVCS and class $\mathcal{I}$ networks are those that do not. Our main problem of interest is then as follows.

Problem 4. *Given the set of all dynamical networks described by dynamics of the form* (6.1)*, characterize the sets $\mathcal{V}$ and $\mathcal{I}$ in terms of the network structure* $\mathbf{A}$ *and develop efficient and easy-to-interpret methods for distinguishing between them.* □



(a)

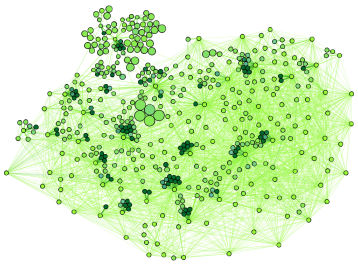
$f(\cdot)$	$f^{\text{TI}}_{\max}$	$f^{\text{TV}}_{\max}$	χ
$\text{tr}(\cdot)$	2.00	2.70	0.35
$\text{tr}((\cdot)^{-1})^{-1}$	1.26×10^{-7}	8.22×10^{-4}	6.5×10^{3}
$\det(\cdot)$	9.90×10^{-11}	7.42×10^{-10}	6.49
$\lambda_{\min}(\cdot)$	1.27×10^{-7}	1.10×10^{-4}	8.7×10^{2}

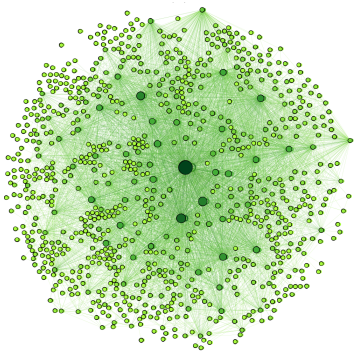
(b)

(d)

$f(\cdot)$	$f^{\text{TI}}_{\max}$	$f^{\text{TV}}_{\max}$	χ
$\text{tr}(\cdot)$	5	5	0
$\text{tr}((\cdot)^{-1})^{-1}$	0.2	0.2	0
$\det(\cdot)$	1	1	0
$\lambda_{\min}(\cdot)$	1	1	0

(e)





(f)

Figure 6.1: Advantage of TVCS in dynamic networks. **(a)** A small example network of $n = 5$ nodes. The thickness of each edge (i, j) illustrates its weight a_{ij}. **(b)** The optimal values of TICS and TVCS (equations (6.4) and (6.5), respectively) and the relative TVCS advantage (equation (6.6)) for the network in (a). **(c)** An air transportation network among the busiest airports in the United States (see 'air500' in Table 6.1 for details). The network is undirected, and the dynamical adjacency matrix $\mathbf{A}$ is computed from static connectivity using the transmission method (cf. Appendix 6.A). This is an example of a network that significantly benefits from TVCS with $\chi \simeq 20\%$. **(d)** A small example network of the same size as (a) but with no benefit from TVCS. **(e)** The optimal values of TICS and TVCS (equations (6.4) and (6.5), respectively) and the relative TVCS advantage (equation (6.6)) for the network in (d). We see that the network does not benefit from TVCS independently of the choice of controllability metric. **(f)** A social network of students at the University of California, Irvine (see 'UCI Forum' in Table 6.1 for details). Similar to (c), the network is undirected and the adjacency matrix is computed using the transmission method. This network, however, does not benefit from TVCS ($\chi = 0$). In (c) and (f), the controllability measure of equation (6.7) is used due to the large size of the network. In both cases, the color intensity and size of nodes represent their values of $R_i(1)$ and $R_i(K-1)$, respectively ($K = 10$). While there is a close correlation between nodal size and color intensity in (f) (i.e., the darkest nodes are also the largest), this is not the case in (c), which is the root cause for the difference in their χ-values. The interested reader can find comprehensive discussions of the network control problem for air transportation in [20–23], social opinion in [24–28], and social epidemic dynamics in [29–34] and references therein.

In the following, we restrict our attention to the choice of controllability measure in equation (6.7) due to its applicability to all network sizes and carry a thorough analysis of its properties in order to address Problem 4.

6.3 Main Results

In this section, we present our main results regarding Problem 4. First, we introduce a new notion of communicability that is pivotal to the solution of Problem 4. Then, we present our results regarding the characterization of class $\mathcal{V}$ and $\mathcal{I}$ networks and, finally, study the case of networks with latent nodes declared earlier.

6.3.1 $2k$-Communicability and Scale-Heterogeneity

Consider the TVCS problem in equation (6.5) with $f(\cdot) = \mathrm{tr}(\cdot)$. Using the definition of the controllability Gramian in (6.3) and the invariance property of trace under cyclic permutations, we can write

$$\mathrm{tr}(\mathcal{W}_K) = \sum_{k=0}^{K-1} \mathbf{b}(K-1-k)^T (\mathbf{A}^k)^T \mathbf{A}^k \mathbf{b}(K-1-k).$$

Therefore,

$$\max_{\iota_0,\ldots,\iota_{K-1}} \mathrm{tr}(\mathcal{W}_K) = \sum_{k=0}^{K-1} \max_{\iota_{K-1-k}} \mathbf{b}(K-1-k)^T (\mathbf{A}^k)^T \mathbf{A}^k \mathbf{b}(K-1-k),$$

where each term $\mathbf{b}(K-1-k)^T (\mathbf{A}^k)^T \mathbf{A}^k \mathbf{b}(K-1-k)$ is the ι_{K-1-k}'th diagonal entry of $(\mathbf{A}^k)^T \mathbf{A}^k$ (cf. equation (6.2)). Therefore, the optimization in (6.5) boils down to finding the largest diagonal

element of $(\mathbf{A}^k)^T\mathbf{A}^k$ and applying $u(K-1-k)$ to this node. On the other hand, for the TICS problem in (6.4) we have

$$\mathrm{tr}(\mathcal{W}_K) = \mathbf{b}^T\left(\sum_{k=0}^{K-1}(\mathbf{A}^k)^T\mathbf{A}^k\right)\mathbf{b},$$

so one has to instead find the largest diagonal entry of $\sum_{k=0}^{K-1}(\mathbf{A}^k)^T\mathbf{A}^k$ and apply all the control inputs $u(0), \ldots, u(K-1)$ to this same node, which is clearly sub-optimal with respect to TVCS. This discussion motivates the following definition.

Definition 6.3.1. *($2k$-communicability).* Given the network dynamics (6.1), the $2k$-communicability of a node $i \in \mathcal{N}$ is defined as

$$R_i(k) = ((\mathbf{A}^k)^T\mathbf{A}^k)_{ii}, \qquad i \in \mathcal{N}, \quad k \geq 0. \tag{6.8}$$

□

Figure 6.2(a-b) illustrates the evolution of $R_i(k)$ as a function of k for all $i \in \mathcal{N}$ for a sample network of $n = 20$ nodes.

Perhaps the most salient property of $2k$-communicability is the extent to which it relies on the local interactions among the nodes. Recall, cf. [35], that for any k, the (i, j) entry of $\mathbf{A}^k$ equals the total number of paths of length k from node i to j (if the graph is weighted, each path counts as its weight, equal to the product of the weights of its edges). From equation (6.8), we see that $R_i(k)$ equals the sum of the squares of the total (weighted) number of paths of length k ending in node i. In other words, $R_i(k)$ only depends on connections of node i with its k-hop out-neighbors, and is independent of the rest of the network. Therefore, $R_i(k)$ is a local notion of centrality for small k

and it incorporates more global information as k grows. In particular, as shown in Appendix 6.C, $R_i(k)$ is closely related to

- the out-degree centrality of node i for $k = 1$;
- the left eigenvector centrality of node i for $k \to \infty$.

This scaling property of $2k$-communicability is illustrated in Figure 6.2(a-d) for an example network of $n = 100$ nodes. Accordingly, we take the left eigenvector centrality squared as the definition of $R_i(\infty)$ in the sequel.

The scaling property of $2k$-communicability also plays an important role in Problem 4. For ease of reference, let

$$r(k) \in \mathcal{N}$$

denote the index of the node that has the largest $R_i(k)$. Then, according to the discussion above,

$$\iota_k^* = r(K-1-k), \tag{6.9}$$

which forms the core connection between $2k$-communicability and TVCS. From this, we see that the optimal TVCS involves the application of $u(0)$ to the node $r(K-1)$ with the highest global centrality and gradually moving the control node until we apply $u(K-2)$ to the node $r(1)$ with the highest local centrality (the control node at time $K-1$ is arbitrary as $R_i(0) = 1$ for all i). The intuition behind this procedure is simple. At $k = 0$, the control input has enough time to propagate through the network, which is why the highest globally-central node should be controlled. As we reach the control horizon K, the control input has only a few time steps to disseminate through

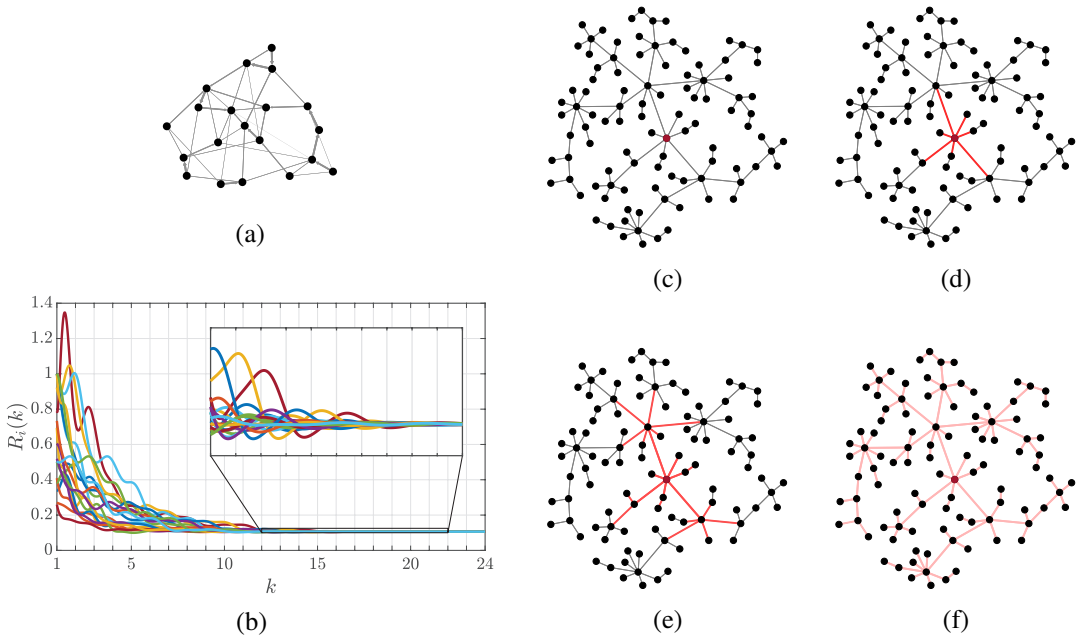


Figure 6.2: $2k$-communicability of dynamical networks. **(a)** An example network of $n = 20$ nodes for illustration of the dependence on k of nodal $2k$-communicabilities. The thickness of the edges is proportional to their weights. **(b)** The evolution of the functions $\{R_i(k)\}_{i=1}^{n}$. Although these functions are originally only defined over integer values of k, we have extended their domain to real numbers for better illustration of their crossings and oscillatory behavior. Oscillatory behavior only arises when $\mathbf{A}$ has complex-valued eigenvalues (otherwise, $R_i(k)$ is strictly convex). **(c)** An example network of $n = 100$ nodes for illustration of the scaling property of $2k$-communicability. The node whose $2k$-communicabilities are to be computed (i.e., "node i") is depicted in red. **(d-f)** The 2-, 4-, and 14-communicability of the node depicted in red, as determined by its 1-, 2-, and 7-hop incoming paths. We see that $R_i(1)$ only depends on the immediate (out-)neighbors of i, but as k grows, $R_i(k)$ encodes more global information.

the network, hence the optimality of locally-central nodes. This further motivates our definition of *scale-heterogeneity*, as follows.

Definition 6.3.2. ***(Scale-heterogeneity of dynamical networks).*** Consider the network dynamics (6.1) subject to the TVCS problem (6.5) with $2k$-communicability as defined in Definition 6.3.1. The network is called *scale-homogeneous* if $r(1) = r(2) = \cdots = r(\infty)$ and scale-heterogeneous otherwise. Accordingly, the more varied $\{r(k)\}_{k=1}^{\infty}$ and $\{R_{r(k)}(k)\}_{k=1}^{\infty}$ are, the more scale-heterogeneous

the network is. □

Based on this definition, we see that the scale-heterogeneity is the main factor in the benefit of TVCS over TICS. In fact, scale-homogeneous and scale-heterogenous networks are the same as class $\mathcal{I}$ and $\mathcal{V}$ networks, respectively, due to (6.9). Further, note that the degree of scale-heterogeneity provides a *geometric and qualitative* characterization of the amount of benefit TVCS has over TICS and distinguishes between networks in $\mathcal{V}$ that only marginally benefit from TVCS and those the benefit significantly (while $2k$-communicability is a more *quantitative* notion used for computational assignment of networks to class $\mathcal{V}$ or $\mathcal{I}$).

It follows immediately from Definition 6.3.2 that determining the scale-heterogeneity of a network requires computation of all $\{r(k)\}_{k=1}^{\infty}$ which is infeasible. Next, we seek simple and computationally efficient conditions to be used as a proxy for scale-heterogeneity.

6.3.2 Identifying Class $\mathcal{V}$ Networks

In this section we discuss a sufficient condition for scale-heterogeneity that, when satisfied, ensures that a network belongs to class $\mathcal{V}$. This condition, given next, relies on the fact that $r(1)$ and $r(\infty)$ are particularly important elements of $\{r(k)\}_{k=1}^{\infty}$ in determining scale-heterogeneity. The proof of this theorem is given in Appendix 6.G.

Theorem 6.3.3. ***(Class $\mathcal{V}$ networks).*** *Consider the TVCS problem* (6.5) *for the network dynamics* (6.1)*. Assume that the adjacency matrix* $\mathbf{A}$ *is irreducible, aperiodic, and diagonalizable. If*

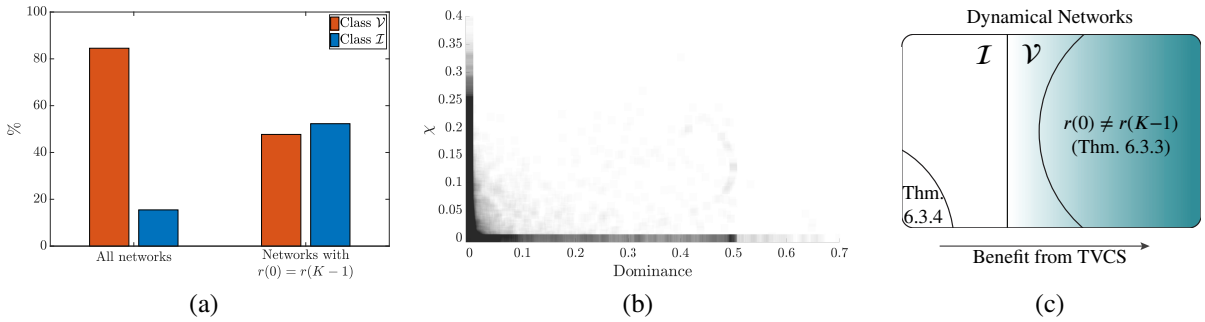
$$\arg\max_{i\in\mathcal{N}} R_i(1) \cap \arg\max_{i\in\mathcal{N}} R_i(\infty) = \emptyset,$$

then the network belongs to class $\mathcal{V}$ for sufficiently large K. □

The condition of **A** being irreducible is equivalent to the network being strongly connected, and thus not restrictive. Likewise, **A** being aperiodic is not restrictive as it requires that there exists no integer number greater than 1 that divides the length of every cycle in the network (satisfied, in particular, if any self-loops exist). Finally, **A** is almost always diagonalizable in the Lebesgue sense, i.e., the set of non-diagonalizable **A** has Lebesgue measure zero.

Consider again the networks of Figure 6.1(c and f). Here, the color intensity of each node indicates its value $R_i(1)$ while its size corresponds to its value $R_i(K-1)$. Clearly, the first few largest and darkest nodes are distinct in Figure 6.1(c), while there is a close correlation between nodal size and darkness in Figure 6.1(f), illustrating the root cause of their difference in benefiting from TVCS.

If a network has $r(0) = r(K-1)$, it is still possible that the network belongs to class $\mathcal{V}$. In fact, about half of the networks with $r(0) = r(K-1)$ still belong to $\mathcal{V}$ (Figure 6.3(a)). However, these networks have a value of χ of no more than 3% on average, and in turn this value quickly decreases with the dominance of the node $r(0)$ over the rest of the network nodes (Figure 6.3(b)). This is a strong indication that, for most practical purposes, the test based on $2k$-communicability is a valid indicator of whether a network benefits from TVCS. Furthermore, in the case of undirected networks, it is possible to analytically prove that a network belongs to class $\mathcal{I}$ ($\chi = 0$) if certain conditions based on the eigen-decomposition of the adjacency matrix **A** are satisfied, as shown next.



(a) (b) (c)

Figure 6.3: The role of $2k$-communicability in distinguishing between networks of class $\mathcal{V}$ ($\chi > 0$) and $\mathcal{I}$ ($\chi = 0$). **(a)** The proportion of random networks in $\mathcal{V}$ and $\mathcal{I}$. A total of 10^5 random connectivity matrices were generated with logarithmically-uniform n in $[10^1, 10^3]$, uniform sparsity p in $[0, 1]$, and uniform pairwise connectivity weight in $[0, 1]$, and then transformed to adjacency matrices $\mathbf{A}$ using the transmission method (cf. Appendix 6.A). A time-horizon of $K = 10$ is used for all networks. While more than 80% of all networks belong to class $\mathcal{V}$, this number drops to less than 50% among networks with $r(1) = r(K-1)$ (i.e., networks where the same node has the greatest local and global centralities). **(b)** The χ-value of the same networks as in (a) that have $r(1) = r(K-1)$ as a function of the dominance of the node $r(0)$. For the node $r(0)$, its *dominance* (over the rest of the network) is a measure of how distinctly $R_{r(0)}(1)$ and $R_{r(0)}(K-1)$ are larger than $R_i(1)$ and $R_i(K-1)$, respectively, for $i \neq r(0)$ (cf. Appendix 6.D). Each gray square represents one randomly generated network, so the darkness of each area represents the probability of observing random networks with that value of (dominance, χ). A rapid decay of χ with dominance is clear, such that networks with positive dominance have very low probability of having $\chi > 0$. **(c)** A Venn diagram illustrating the decomposition of dynamical networks based on the extent to which they benefit from TVCS. The color gradient is a depiction of this extent, as measured by χ (equation (6.6)), where darker areas correspond to higher χ. As shown in (a) and (b), the class of networks for which $r(0) \neq r(K-1)$ is only a subset of $\mathcal{V}$ but provides a good approximation for it.

6.3.3 Identifying Class $\mathcal{I}$ Networks

Complementary to Section 6.3.2, here we discuss some necessary conditions for scale-heterogeneity based on the eigen-structure of the network that characterize subsets of $\mathcal{I}$. Let $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ be the eigen-decomposition of $\mathbf{A}$, where $\mathbf{V} = [v_{ij}]_{n\times n}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix of eigenvalues with $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Further, let $\mathbf{W} = [w_{ij}]_{n\times n}$ be the doubly stochastic matrix such that $w_{ij} = v_{ij}^2$ for all $i, j \in \{1, \dots, n\}$. The next result, proven in Appendix 6.G, characterizes three undirected sub-classes of $\mathcal{I}$.

Theorem 6.3.4. ***(Class 1 networks).*** *Consider the TVCS problem* (6.5) *for the network dynamics* (6.1)*. Assume that the network is undirected (i.e.,* $\mathbf{A} = \mathbf{A}^T$*) and that, without loss of generality, the node with the largest eigenvector centrality is labeled as node* 1*. If any of the following conditions holds:*

(i) $\frac{1-w_{11}}{w_{11}} \leq \frac{|\lambda_1|-|\lambda_2|}{|\lambda_1|-|\lambda_n|}$,

(ii) $w_{11} + w_{12} = 1$,

(iii) the network has three or fewer nonzero eigenvalues with different absolute values and $1 \in \arg\max_i R_i(1)$,

then,

$$1 \in \arg\max_{1 \leq i \leq n} R_i(k), \quad \forall k \in \{0, \ldots, K-1\}, \tag{6.10}$$

i.e., selecting the node with the largest eigenvector centrality at every time step is the solution to (6.5)*.* □

The conditions in Theorem 6.3.4 are based on the eigen-decomposition of the network adjacency matrix $\mathbf{A}$ and thus abstract. However, these conditions can be interpreted as follows:

(i) Condition *(i)* holds for networks where there is a sufficiently distinct central node, in the sense of eigenvector centrality, and the network dynamics is dominated by the largest eigenvalue. An extreme case of such networks is a totally disconnected network where $\mathbf{W} = \mathbf{I}$ and the highest authority is the node with the largest self-loop.

(ii) Condition *(ii)* holds for networks where the eigenvector centrality of all nodes is determined by the weight of the link to the most eigenvector-central node. To see this, note that we have $w_{1j} = 0$ for $j \geq 3$, implying $v_{1j} = 0, j \geq 3$. Since the rows of $\mathbf{V}$ are orthogonal, we deduce $v_{i2} = \alpha v_{i1}$ for all $i \geq 2$, where $\alpha = -v_{11}/v_{12}$ is constant. Using $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, we have

$$a_{1i} = \lambda_1 v_{11} v_{i1} + \lambda_2 v_{12} v_{i2} = (v_{11}\lambda_1 + \alpha v_{12}\lambda_2) v_{i1},$$

so $v_{i1} \propto a_{1i}$ for all $i \geq 2$. Examples of such networks are star networks with no (or small-weight) self-loops (cf. Proposition 6.F.3).

(iii) Regarding condition *(iii)*, the most well-known families of networks with three distinct eigenvalues are the complete bipartite networks and connected strongly regular networks. Moreover, cones on (n, k, λ, μ)-strongly regular graphs satisfying $\lambda_{\min}(\mathbf{A})(\lambda_{\min}(\mathbf{A}) - k) = n$ are also known to have three distinct eigenvalues [36]. The other condition $1 \in \arg\max_i R_i(1)$ holds when the node with the largest eigenvector centrality (i.e., $r(\infty)$) has also the largest 2-communicability. The simplest example of a network with these properties is the star network (with no or equal self-loops).

The general abstraction from these cases is that a network belongs to class $\mathcal{I}$ if it contains a sufficiently distinct central node, which reinforces our main conclusion that $\mathcal{V}$ is the class of networks with multiple scale-heterogeneous central nodes. The inclusion relationships between the various classes of networks introduced in this section are summarized in Figure 6.3(c).

While Theorem 6.3.4 is only applicable to undirected networks, it has a straightforward extension to normal networks (i.e., directed networks with normal $\mathbf{A}$). Using the same proof technique

as in Theorem 6.3.4, it can be shown that the exact same results hold if one replaces the eigenvalues and eigenvectors with singular values and singular vectors of $\mathbf{A}$. Interestingly in this case, $R_i(\infty)$ coincides with HITS hub/authority centrality of node i squared [37].

6.3.4 Networks with Latent Nodes

As mentioned in Section 6.2, in many real-world applications of TVCS not all the nodes are available/accessible for control. In this case, we call a node *manifest* if it can be actuated and *latent* if it cannot. The natural solution would then be to choose the control nodes optimally among the manifest nodes. If the adjacency matrix $\mathbf{A}$ of the network is fixed and given, this is the best solution. However, there are cases where $\mathbf{A}$ itself can be changed, at least among the manifest nodes. We call such a change of structure an *(edge) manipulation*. Edge manipulations are primarily possible in man-made (power, transportation, etc.) networks, since the edges are originally engineered, but are also becoming increasingly feasible in biological networks due to advances in bioengineering, see, e.g., [38, 39] for brain and [40, 41] for gene networks. When manipulation is possible in a network with latent nodes, another solution to TVCS is to manipulate the network among the manifest nodes such that the optimal control nodes (when computed *without* any restrictions on control scheduling) lie among the manifest nodes for all time. The following result provides a guarantee that this is always possible, provided that the manipulation is sufficiently strong and not acyclic.

Theorem 6.3.5. ***(Network manipulation and TVCS in networks with latent nodes).*** *Consider the optimal node selection problem* (6.5) *over a time horizon K. Given a network of n nodes with adjacency matrix $\mathbf{A}_0 \in \mathbb{R}^{n\times n}$, let $\mathbf{E} \in \mathbb{R}^{n\times n}$ be a nonnegative matrix of the form*

$$\mathbf{E} = \begin{bmatrix} \overbrace{\star}^{n_1} & \overbrace{\mathbf{0}}^{n-n_1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{matrix} \} n_1 \\ \} n-n_1 \end{matrix},$$

corresponding to the manifest subnetwork involving the first $n_1 < n$ nodes (this is without loss of generality, since nodes can be renumbered) and consider the dynamic network described by (6.1) *with adjacency matrix $\mathbf{A} = \mathbf{A}_0 + \alpha\mathbf{E}$, where $\alpha > 0$. Then, if $\mathbf{E}$ is not acyclic, there exists $\overline{\alpha} > 0$ such that for $\alpha > \overline{\alpha}$,*

$$r(k) \in \{1, \dots, n_1\}, \tag{6.11}$$

for all $k \in \{0, \dots, K-1\}$. Furthermore, if $\mathbf{A}_0$ and $\mathbf{E}$ are symmetric (the corresponding networks are undirected), $\bar{\alpha}$ can be found in closed form and (6.11) *holds for all $k \geq 1$.* □

Both requirements of Theorem 6.3.5 (that $\alpha\mathbf{E}$ is sufficiently strong and acyclic) have clear interpretations. First, depending on how large the size of the manifest subnetwork is and how central its nodes already are (pre-manipulation), larger manipulation may be necessary to turn them into central nodes at various scales (i.e., $r(k)$ for $k = \{0, \dots, K-1\}$). Second, for the manifest nodes to become central at arbitrarily global scales (i.e., $r(k)$ for $k \sim K \to \infty$), the manipulation must contain paths of arbitrarily long lengths, which are absent in acyclic networks.

According to Theorem 6.3.5, manipulation of the manifest subnetwork is effective even when the manifest nodes are among the least central nodes of the network (before the manipulation). In this case, as we increase α from 0, the manifest nodes usually first turn into the most locally-

central nodes ($\alpha \not\geq \bar{\alpha}$ yet), and then also into globally-central nodes ($\alpha > \bar{\alpha}$). The following example illustrates this phenomenon in a simple star network where the center node is latent and the peripheral nodes are manifest.

Example 6.3.6. ***(Undirected star networks with varying self-loop weights).*** Consider an undirected uniform star network given by

$$\mathbf{A}_0 = \begin{bmatrix} l_p \mathbf{I}_{n-1} & a_{cp} \mathbf{1}_{n-1} \\ a_{cp} \mathbf{1}_{n-1}^T & l_c \end{bmatrix},$$

where $\mathbf{1}_{n-1}$ denotes the $(n-1)$-dimensional vector of all ones and the positive constants l_c, l_p, and a_{cp} are the central self-loop weight, peripheral self-loop weight, and the link weight between the center node and any peripheral node, respectively. The $2k$-communicabilities of this network are computed analytically in Proposition 6.F.3 (nodes are re-labeled here for conformity with Theorem 6.3.5). It follows from (6.25) that for any $i \in \{1, \dots, n-1\}$,

$$R_n(1) - R_i(1) = l_c^2 - l_p^2 + (n-2)a_{cp}^2. \tag{6.12}$$

Therefore, if $l_p \leq l_c$, then $R_n(k) > R_i(k)$ for all $k \geq 1$, i.e., the center node is the optimal control node at all times. However, when $l_c < l_p$, the network can exhibit different behaviors. From (6.25), we can also see that

$$\lim_{k\to\infty} R_n(k) > \lim_{k\to\infty} R_i(k) \Leftrightarrow \lambda_1 - l_p > a_{cp}. \tag{6.13}$$

Define $\underline{l}_p = \sqrt{l_c^2 + (n-2)a_{cp}^2}$ and $\bar{l}_p = l_c + (n-2)a_{cp}$. Using (6.12)-(6.13) and after some compu-

tations, one can see that

$$\begin{aligned} &r(k) = n \text{ for all } k, && \text{if } l_p \leq \underline{l}_p, \\ &r(1) = \{1, \dots, n-1\} \text{ but } r(k) = n \text{ for large enough } k, && \text{if } \underline{l}_p < l_p < \bar{l}_p, \\ &r(k) = \{1, \dots, n-1\} \text{ for all } k, && \text{if } l_p \geq \bar{l}_p. \end{aligned}$$

In other words, when the manipulation is weak, the (latent) center node is the optimal control node at all times. As the manipulation gains strength, scale-heterogeneity emerges, making the (manifest) peripheral nodes the optimal control node at local scales while the center node remains still the optimal control node at global scales. Finally, when the manipulation is strong enough, scale-heterogeneity vanishes, leaving the (manifest) peripheral nodes as the optimal control nodes at all scales. Notice that with the terminology of Theorem 6.3.5,

$$\mathbf{E} = \begin{bmatrix} \mathbf{I}_{n-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad n_1 = n-1, \quad \alpha = l_p, \quad \text{and} \quad \overline{\alpha} = \bar{l}_p. \qquad \bullet$$

A fair concern, however, exists regarding the minimum size of the manipulation needed to make the TVCS all-manifest. If this is excessively high, the prescribed approach may be infeasible in practice. Nevertheless, among networks of various size and structure, random manipulations with norm of about 10% of the norm of $\mathbf{A}$ are on average sufficient (Figure 6.4). Here, we see that the largest manipulations are needed for manifest subnetworks of about 10% the total size of the network. This is because when the size of the manifest subnetwork is extremely small, manipulations are focused on this small subset of nodes and thus more efficient, while with extremely large manifest subnetworks, the majority of the nodes are accessible for control and there is little restriction on the TVCS.

Finally, Figure 6.4 also shows the comparison, in terms of controllability, of the manipulation-based approach against the alternative approach of selecting an optimal TVCS with

the additional constraint that control nodes must be manifest (without any manipulation of the dynamics), which results in a sub-optimal all-manifest TVCS. For the comparison to be fair, we normalize each network by its spectral radius (largest magnitude of its eigenvalues), and then compare the optimal value of their TVCS (equation (6.5)). We see that the amount of relative advantage produced by manifest subnetwork manipulation is comparable to the relative size of the manipulation, except for medium-sized manifest subnetworks ($5 \sim 20\%$ of nodes), where the manipulation advantage is about two times its size.

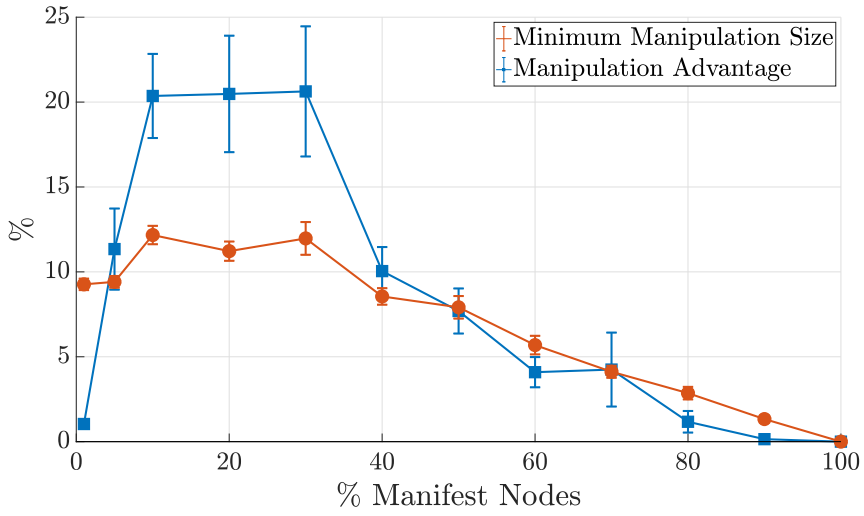


Figure 6.4: Manipulation of manifest subnetworks in order to obtain an all-manifest optimal TVCS. The horizontal axis represents the percentage of manifest nodes in the network. In red, we show the minimum size of manipulation needed for the optimal TVCS to only include manifest nodes, relative to the size of the initial adjacency matrix (both measured by induced matrix 2-norm). In blue, we depict the optimal (i.e., maximal) value of $\mathrm{tr}(\mathcal{W}_K)$ for the case where the minimal manifest manipulation is applied, relative to the maximal value of $\mathrm{tr}(\mathcal{W}_K)$ subject to the constraint that all the control nodes are manifest (the former is with manipulation and without constraints on the control nodes, while the latter has no manipulation but control node constraints). Results are for 10^3 random networks of logarithmically-uniform sizes in $[10^1, 10^3]$ but otherwise similar to Figure 6.3. Markers (circles/squares) represent average values and error bars represent standard error of the mean (s.e.m). In both cases, the overall adjacency matrix is normalized by its spectral radius for fairness of comparison. We see that medium-sized manifest subnetworks ($5 \sim 20\%$) are the hardest yet most fruitful to manipulate.

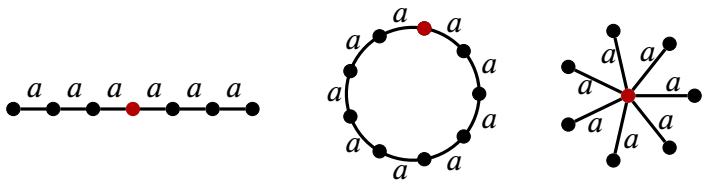


Figure 6.5: Simple networks with closed-form $2k$-communicabilities. **(a)** A line network, **(b)** a ring network, and **(c)** a star network. All networks are undirected and have homogeneous edge weights a. The $2k$-communicabilities of these networks are analytically computed (cf. Appendix 6.F), concluding that all networks belong to class $\mathcal{I}$, with the optimal control node depicted in red in each case (the optimal control node is arbitrary in a ring network due to its symmetry).

6.4 Case Study: TVCS in Synthetic and Real Networks

Here, we discuss the benefits of TVCS and its relation to network structure for several examples of synthetic and real networks. We start with the classical deterministic examples of undirected line, ring, and star networks (Figure 6.5). Due to their simple structure, the $2k$-communicabilities of these networks can be analytically computed in closed form (cf. Appendix 6.F). Using these results, it follows that for the line and star networks, the optimal control node is always the center node (or any of the two center nodes if a line has even number of nodes), while the optimal control node is arbitrary in a ring network. Notice that in all cases, it is the *homogeneity* of these networks that results in a single node having the greatest centrality at all scales (cf. Example 6.3.6 for non-homogeneous star networks that have scale-heterogeneous central nodes and thus belong to class $\mathcal{V}$).

Next, we analyze the role of TVCS in three classes of probabilistic complex networks that are widely used to capture the behavior of various dynamical networks. These include the Erdös-Rényi (ER) random networks, Barabási-Albert (BA) scale-free networks, and Watts-Strogatz (WS) small-world networks. Each network has its own characteristic properties, and these properties

lead to different behaviors under TVCS. The average χ-values of these networks are computed for various values of n and network parameters (Figure 6.6). For ER networks, χ is in general small, and decays with n. This is because ER networks, especially when n is large, are extremely homogeneous. This homogeneity is further increased during the transmission method, leading to a network matrix $\mathbf{A}$ that is extremely insensitive to the choice of control nodes.

The connectivity structure of BA networks, in contrast, is extremely inhomogeneous, with one (sometimes 2) highly central nodes and a hierarchy down to peripheral leafs. As one would expect, this implies a small χ-value since the center node has the highest centrality at all scales (Figure 6.8). However, when the connectivity matrix is transformed to $\mathbf{A}$ using the transmission method, the incoming links to all nodes are made uniform (adding up to 1). This in turns make the centrality levels of all the nodes comparable, leading to high χ-values observed (notice that the underlying connectivity structures are still highly inhomogeneous, distinguishing them from the homogeneous ER networks). Notice that as the growth rate m_a is increased, smaller networks tend towards complete graphs and high χ values *shift* to larger n.

As our last class of probabilistic networks, WS networks have the broadest range of size-parameter values with significant χ. As one would expect, χ is low near $\beta = 0, 1$, corresponding to regular ring lattice and ER networks, respectively. For $\beta \sim 0.2$, there is a sufficiently high probability of having multiple nodes that are close to many rewired links (increasing their centrality), yet there is a low probability that these nodes, and the nodes close to them, are rewired all alike, resulting in heterogeneous central nodes and high χ-values. This heterogeneity is increased with n as larger networks have more possibilities of rewiring every edge.

Finally, we used the tools and concepts introduced so far to analyze TVCS in several real-world dynamical networks (Table 6.1). These networks are chosen from a wide range of application

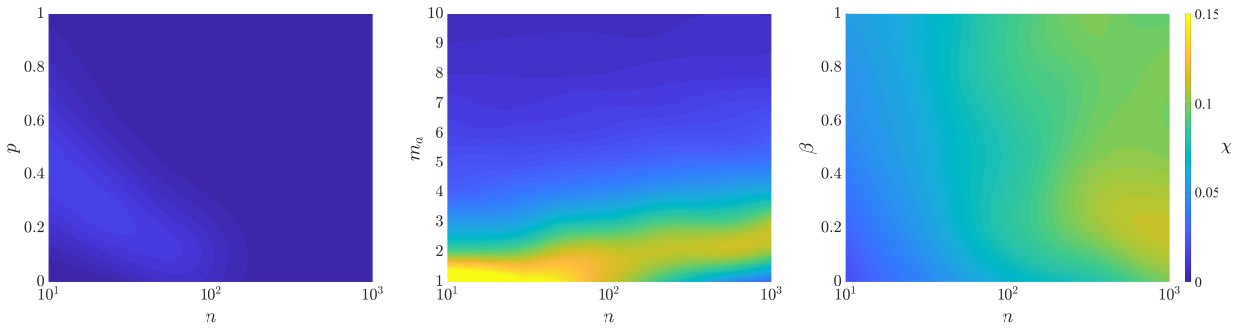


Figure 6.6: The average χ-value for **(a)** ER, **(b)** BA, and **(c)** WS probabilistic networks. The horizontal axis determines the size of the network n in all cases, while the vertical axis determines the values of the corresponding parameters for each network: edge probability p for ER, growth (link attachment) rate m_a for BA, and rewiring probability β for WS. After constructing the unweighted connectivity according to each algorithm (ER, BA, or WS), standard uniformly random weights are assigned to each edge, which is then converted to $\mathbf{A}$ using transmission method (cf. Appendix 6.A). For each value of n and network parameter over a coarse mesh ($\sim$ 100 points), 100 networks are generated and the average of their χ-value is computed, which is then smoothly interpolated over a fine mesh (MATLAB `csaps`).

domains, from neuronal networks to transportation and social networks. According to the type of dynamics evolving over each network, we have used either the transmission or induction method to obtain its dynamical adjacency matrix from its static connectivity (the "$\mathbf{C} \rightarrow \mathbf{A}$" column, cf. Appendix 6.A).

We have computed the χ-value for each network using a variable time horizon $K \leq 50$, with the results ranging from 0 to more than 30% for different networks. These large variations even within each category signify both the potential benefits of TVCS and the possibility of its redundancy, a contrast that has been pivotal to our discussion. In particular, four facts about these results worth highlighting. (i) As measured by $\mathrm{tr}(\mathcal{W}_K)$, the majority of networks tested do not benefit from TVCS, but a few do so significantly. (ii) Despite coming from various domains, the networks that do significantly benefit from TVCS share *scale-heterogeneity* as their common qualitative property (cf. Section 6.3.1). (iii) Networks with inductive $\mathbf{C} \rightarrow \mathbf{A}$ transformation benefit

significantly less from TVCS than those with transmission $\mathbf{C} \rightarrow \mathbf{A}$ transformation. (iv) Significantly higher values of χ are expected for all networks if using $\lambda_{\min}(\mathcal{W}_K)$ or similar measures for controllability, cf. Appendix 6.B.

In the last column, we have also indicated whether the most local and most global central nodes coincide in each network. Recall that this is a sufficient but not necessary condition for a network to be in class $\mathcal{V}$ (Theorem 6.3.3 and Figure 6.3). Though only sufficient, this simple metric can correctly classify class members of $\mathcal{V}$ from $\mathcal{I}$ among these networks, except for the WesternUS power network, for which $r(0) = r(K-1)$ only marginally holds (the dominance of $r(0)$ is 0) (cf. Figure 6.3(b)).

6.5 Discussion

A striking finding that defied our expectations is the effect of network dynamics, beyond its raw connectivity structure, on TVCS. Here, we differentiated between the raw connectivity structure of a network (obtained using specific field knowledge and measure the *relative* strength of nodal connections) and its dynamical adjacency matrix which determines the evolution of network state over time. Depending on the nature of network state, we proposed two methods, transmission and induction, for obtaining the dynamical adjacency matrix from static connectivity. The effects of these methods, however, is noteworthy on the benefits of TVCS, even though the underlying network connectivity is the same (Table 6.1 and Figure 6.8). While the transmission method significantly enhances the merit of TVCS, the induction method depresses it (both compared to raw connectivity). We believe the reason for the former is the additional *homogeneity* that the transmission method introduces among the nodes, while the latter is due to the conversion from continuous to

Table 6.1: Characteristics of the real-world networks studied. For each network, we have reported the number of nodes n, number of edges $|\mathcal{E}|$ (with each bidirectional edge counted twice), whether the network is directed, the method used for obtaining dynamical adjacency matrix $\mathbf{A}$ from static connectivity $\mathbf{C}$ ($\mathbf{A} \to \mathbf{C}$), the χ value (equation (6.6)), and whether the most local and global central nodes coincide ($r(0) = r(K-1)$). Since the value of χ is a function of K, we have chosen the value of $K \leq 50$ that has the largest χ for each network. Detailed descriptions of these datasets are provided in Appendix 6.H.

Category	Name	n	$\lvert E\rvert$	Directed	$\mathbf{C} \to \mathbf{A}$	$\chi(\%)$	$r(0) = r(K-1)$	Dominance of $r(0)$ ($\times 10^{-3}$)	ref.
Neuronal	BCTNet fMRI	638	37250	N	T	1.8	N	N/A	[42]
	Cocomac	58	1078	Y	T	5.5	N	N/A	[43]
	BCTNet Cat	95	2126	Y	T	1.9	N	N/A	[42]
	C. elegans	306	2345	Y	T	0	Y	0	[44]
Transportation	air500	500	5960	N	T	22.4	N	N/A	[45]
	airUS	1858	28236	Y	T	0	Y	0	[46]
	airGlobal	7976	30501	Y	T	0	Y	0	[46]
	Chicago	1467	2596	N	T	0	Y	0	[47,48]
Gene Regulatory	E. coli	4053	127544	N	T	0	Y	0	[49]
PPI	Yeast	2361	13828	N	T	0	Y	0	[50]
	Stelzl	1706	6207	Y	T	0	Y	0	[51]
	Figeys	2239	6452	Y	T	0	Y	0	[52]
	Vidal	3133	12875	N	T	0	Y	0	[53]
Power	WesternUS	4941	13188	N	T	33.7	Y	0	[44]
Food	Florida	128	2106	Y	T	34.6	N	N/A	[54]
	LRL	183	2494	Y	T	27.3	N	N/A	[55]
Social	Facebook group	4039	176468	N	I	0.4	N	N/A	[56]
	E-mail	1005	25571	Y	I	0	Y	40.5	[57,58]
	Southern Women	18	278	N	I	0	Y	1.6	[59]
	UCI P2P	1899	20296	Y	I	0	Y	5.5	[60]
	UCI Forum	899	142760	N	I	0	Y	2.8	[61]
	Freeman's EIES	48	830	Y	I	0	Y	1.4	[62]
	Dolphins	62	318	N	I	0	Y	0.7	[63]
Trust	Physicians	241	1098	Y	I	8.8	N	N/A	[64]
	Org. Consult Advice	46	879	Y	I	0	Y	0.1	[65]
	Org. Consult Value	46	858	Y	I	0	Y	1.2	[65]
	Org. R&D Advice	77	2228	Y	I	6×10^{-3}	N	N/A	[65]
	Org. R&D Aware	77	2326	Y	I	0	Y	0.3	[65]

discrete-time dynamics, which enables long-distance connections even over small sampling times (due to the fact that interactions occur over infinitesimal intervals in continuous time) (cf. Section 6.A and Figure 6.7). These results suggest that controllability of network dynamics is not only a function of its structural connectivity, but also greatly relies on the type of dynamics evolving over the network, an aspect that has received little attention in the existing literature and warrants future research.

Our discussion so far applies to networks with and without self-loops alike. However, it follows from the results in Section 6.3 that self-loops play an important role in TVCS. This is because (i) the self-loop of each node directly adds to its $2k$-communicability for all k, and (ii) the self-loop of each node also contributes indirectly to the $2k$-communicability of its neighbors less than $k-1$ hops away. As a result, the self-loop of any node has the largest effect on its own $2k$-communicability for all k, but also a lesser effect on the $2k$ communicability of all other nodes in the network. This latter effect becomes smaller and limited to higher k for more distant nodes. A clear demonstration of the effects of self-loops can be seen in Example 6.3.6, where as the self-loops of the peripheral nodes get stronger, they gradually become the central nodes in the network, first at local scales (small k) and eventually at all scales.

Further, the focus of our discussion has so far been on single input networks where one node is controlled at a time, in order to enhance the simplicity and clarity of concepts. Nevertheless, our results have straightforward generalizations to multiple-input networks (cf. Appendix 6.E). If m denotes the number of control inputs, the optimal TVCS involves applying these control inputs to the m nodes with the highest centralities at the appropriate scale at every time instance (i.e., the m nodes with the largest $R_i(K-1-k)$ have to be controlled at every time instance k). It is clear that the additional flexibility due to the additional inputs makes $\mathcal{V}$ larger, i.e., more networks have $\chi > 0$.

Nevertheless, this additional flexibility also makes TICS significantly more efficient. Therefore, it is not immediately clear whether this enlargement of $\mathcal{V}$ also entails larger χ for networks with the same size and sparsity. In fact, increasing m reduces average χ for all the classes of ER, BA, and WS networks (Figure 6.9), suggesting that the additional flexibility is more advantageous for TICS than TVCS.

Regardless of the number of inputs (1 or more), an important implicit assumption of TVCS is that this number is limited, i.e., no more than m nodes can be controlled at every time instance. This may at first seem over-conservative since TVCS requires, by its essence, the installation of actuators at all (or many) nodes of the network. Therefore, one might wonder why limit the control to only m nodes at every time instance when all the nodes are ready for actuation. The answer lies within the practical limitations of actuators. For ideal actuators, distributing the control energy over as many nodes as possible is indeed optimal. However, this is not possible in many scenarios, including when (i) actuators exhibit nonlinear *dead-zone* behaviors, so that each one requires a sizable activation energy. In many applications ranging from distributed industrial processes to opinion dynamics in social networks, nodes cannot be actuated with arbitrarily small amounts of control energy. If $E_{\min}$ is the minimum activation energy of any actuator, at least $mE_{\min}$ is required for actuation of m nodes at a time. Thus, when $E_{\min}$ is sizable and n is large, simultaneous actuation of all nodes ($m = n$) requires a significant amount of control energy which is often infeasible (notice that the dead-zone behavior of actuators does not violate the linearity assumption in (6.1) as one can replace u with $v = \phi(u)$, where ϕ denotes dead-zone nonlinearity); (ii) actuators are geographically disperse so that precise coordination becomes difficult or time-consuming. A familiar example of this is the social opinion dynamics in pre-election times during political campaigns, where rallies and speeches by candidates act as control inputs to the network. Even though all nodes may be

actuatable, at most one node can be actuated at every time; (iii) simultaneous control of proximal nodes results in actuator interference. This is the case in many biological networks. In neuronal networks, for instance, common control technologies such as TMS do not allow for simultaneous actuation of all cortical areas due, in part, to electromagnetic interference between multiple sources of actuation (note that TVCS is still possible by installation and sequential activation of multiple coils at different locations); and when (iv) actuators are controlled via communication channels with limited capacity, so that only a small number of devices can be simultaneously operated. This may be the case in industrial applications where large numbers of geographically distributed actuators are remotely (and centrally) controlled over shared communication channels with limited bandwidth. In all these scenarios, TVCS has the potential to significantly enhance network controllability, conditioned on the scale-heterogeneity of the central nodes in the network.

Although the dynamics of all real networks have some degrees of nonlinearity, the analysis of linear(ized) dynamics is a standard first step in analysis of dynamical properties of complex networks [2–9, 14–17]. This is mainly due to the fact that stability and controllability of linearized dynamics of a nonlinear network implies the same properties *locally* for the original nonlinear dynamics, making linear dynamics a powerful tool in analyzing many dynamical properties that are in general intractable for nonlinear dynamics. The local validity of linearization, however, is a main limitation of this work, particularly in networks where the change of state is significant relative to the size of the domain over which the linearization is valid. For these networks, whether the nonlinearity enhances or decreases the benefits TVCS with respect to its linearization is in general dependent on the type of nonlinearity. However, for saturation nonlinearities, being perhaps the most widespread, we expect TVCS to be more beneficial than linear counterparts. This is because in TICS all the control input is injected through a fixed node, requiring the state of that node to

potentially undergo large over- and undershoots in order to convey sufficient input to the rest of the network. Saturation clearly prevents this from happening, further limiting the scope of TICS. The generalization of this work to nonlinear dynamics with saturation and linear *time-varying* dynamics (namely, $\mathbf{A}(k)$ instead of $\mathbf{A}$ in equation (6.1)) is a warranted next step for future exploration of the role of TVCS in general nonlinear networks.

Appendix

6.A Obtaining Dynamical Adjacency Matrix from Static Connectivity

A standard starting point for the analysis of network dynamics of the form (6.1) is the assumption that the network adjacency matrix $\mathbf{A}$ is known. While this is a valid assumption (as the construction of $\mathbf{A}$ is itself the subject of vast research in network identification and corresponding field sciences), care should be taken in how one interprets raw network connectivity matrices. Usually, the network structure is described not by its dynamic adjacency matrix $\mathbf{A}$ (which determines the evolution of network *state* according to (6.1)) but rather by its static connectivity matrix $\mathbf{C}$ (our implicit assumption is that each node has a well-defined state that evolves over time through network dynamics, so our discussion is not applicable to completely static networks). While for any $i, j \in \mathcal{N}$, a_{ij} describes the impact of x_j on x_i over one time step (relative to x_j), c_{ij} often describes the strength of the link (i, j) in arbitrary units (e.g., number of synapses between two neurons, capacity of high-voltage lines between two generators, or number of seats on a flight). In particular, multiplying $\mathbf{C}$ by a positive constant results in an equivalent description of the network

structure, yet multiplying $\mathbf{A}$ by a constant significantly alters network dynamics. Here, we outline two methods for obtaining $\mathbf{A}$ from $\mathbf{C}$, and describe example domains where each method seems more relevant. Consider an arbitrary link $(i, j) \in \mathcal{E}$.

- **Transmission:** This method applies to dynamical networks where at each time step, the value of the state of node i is itself affected (reduced) as a result of interaction with neighbor node j. Here, the state of each node corresponds to a physical quantity that is *transmitted* to its neighbors in order to affect their states. Neuronal, transportation, food, gene regulatory, protein-protein interaction, and power networks are all examples of this type of interaction. If the sampling time is chosen long enough such that "current" state of a node is completely diffused through the network until the next time step, we can obtain $\mathbf{A}$ from $\mathbf{C}$ using

$$\mathbf{A} = \mathbf{CDC}, \mathrm{in}^{-1},$$

  where $\mathbf{DC}, \mathrm{in}$ is the *augmented in-degree matrix* of $\mathbf{C}$ (a diagonal matrix with the sum of the columns of $\mathbf{C}$ on its diagonal, except where the sum of a column of $\mathbf{C}$ is zero, in which case the corresponding diagonal element of $\mathbf{DC}, \mathrm{in}$ is 1). This means that over each time step, x_i is transmitted to the in-neighbors of node i proportionally to their connectivity strength, if i has any in-neighbors, and preserved otherwise.

- **Induction:** This method is appropriate for networks in which nodal states are not physical quantities and thus do not reduce as a result of network interactions. Opinion or epidemic dynamics evolving over social and/or trust networks have such properties. Here, in order to compute $\mathbf{A}$ from $\mathbf{C}$, we start from the underlying continuous-time dynamics $\dot{\mathbf{x}} = (-\alpha\mathbf{I} +$

$\mathbf{C})\mathbf{x}$ where $\alpha > 0$ is chosen such that $-\alpha\mathbf{I} + \mathbf{C}$ is stable (Hurwitz), and then discretize it to obtain (6.1), where

$$\mathbf{A} = e^{(-\alpha\mathbf{I}+\mathbf{C})T_s},$$

and T_s is the sampling time [19, eq. (4.17)]. From the expansion of matrix exponential ($e^{\mathbf{M}} = \mathbf{I} + \mathbf{M} + \frac{\mathbf{M}^2}{2} + \frac{\mathbf{M}^3}{3!} + \cdots$), we see that $\mathbf{A}$ does not inherit the sparsity pattern of $\mathbf{C}$ (and G) since nodes interact in continuous time. However, if $\|(-\alpha\mathbf{I} + \mathbf{C})^2 T_s^2/2\| \ll \|(-\alpha\mathbf{I} + \mathbf{C})T_s\|$, then the sparsity pattern of $\mathbf{C}$ is almost preserved in $\mathbf{A}$. Therefore, in this work we use $T_s = \gamma_{\text{ind}}/\|\alpha\mathbf{I} + \mathbf{C}\|$ for the induction method with $\gamma_{\text{ind}} = 0.2$ unless otherwise stated. Further, Figure 6.7 shows the effect of γ_{ind} on the value of χ when using the induction method. As expected, the larger γ_{ind}, the larger T_s, the closer $\mathbf{A}$ gets to $\lim_{k\to\infty} \mathbf{A}^k$, the more similar $2k$-communicabilities for different k become, and the smaller χ becomes.

Unless otherwise stated, we use the transmission method in this work. Nevertheless, it is to be noted that the method used for obtaining $\mathbf{A}$ from $\mathbf{C}$ can have profound effects on network controllability and should thus be chosen carefully. Figure 6.8 illustrates this concept by showing the mean χ-value of ER, BA, and WS networks for a number of different choices for this transformation.

6.B Comparison Between Gramian-based Measures of Controllability

In this section, we first derive and elaborate on the relationship between the eigenvalues of the Gramian and control energy. Then, we discuss the different Gramian-based measures of

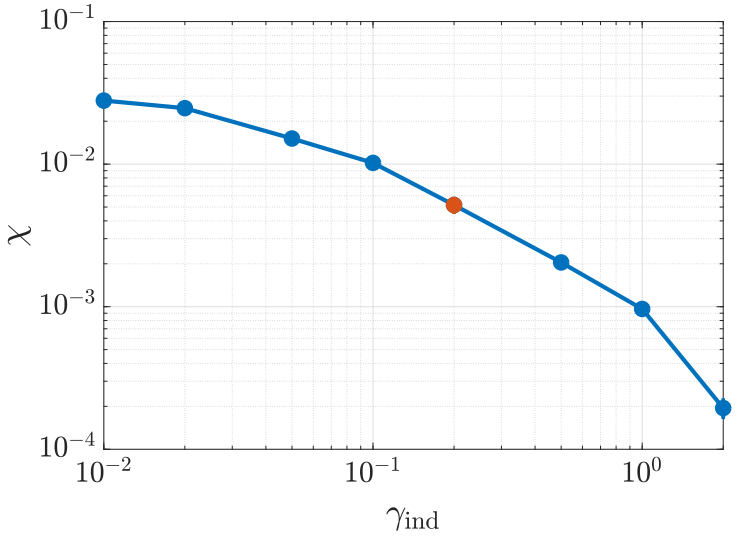


Figure 6.7: The average value of χ for the induction method and varying values of γ_{ind} (corresponding to varying discretization step sizes T_s). Each point represents the average value of χ for 50 realizations of ER networks with $n = 100$ and $p = 0.2$ and vertical bars (when visible) show one standard error of the mean (s.e.m.). For each network, the value of $K \leq 10^3$ that gives the largest value of χ is chosen. The average value of χ drops with γ_{ind}, showing the effect of discretization on χ and the merit of TVCS. The red point corresponds to $\gamma_{\text{ind}} = 0.2$ used throughout this work.

controllability and their respective properties.

Assume that $\mathcal{W}_K$ is invertible (the network dynamics (6.1) is controllable). Then, for any $\mathbf{x}_f \in \mathbb{R}^n$, among the (usually infinitely many) choices of $\{u(k)\}_{k=0}^{K-1}$ that take the network from $\mathbf{x}(0) = \mathbf{0}$ to $\mathbf{x}(K) = \mathbf{x}_f$, the one that has the smallest energy is given by [19, Thm 6.1]

$$u^*(k) = \mathbf{b}(k)^T (\mathbf{A}^T)^{K-1-k} \mathcal{W}_K^{-1} \mathbf{x}_f, \quad k \in \{0, \dots, K-1\}.$$

Similar expression holds for arbitrary $\mathbf{x}_0$, but it is customary to evaluate control energy starting from the network's unforced equilibrium $\mathbf{x} = \mathbf{0}$. It is immediate to verify that this gives the minimal energy $\sum_{k=0}^{K-1} u^{*2}(k) = \mathbf{x}_f^T \mathcal{W}_K^{-1} \mathbf{x}_f$. Therefore, the unit-energy reachability set is given by

$$\{\mathbf{x}_f \in \mathbb{R}^n \mid \mathbf{x}_f^T \mathcal{W}_K^{-1} \mathbf{x}_f \leq 1\}.$$

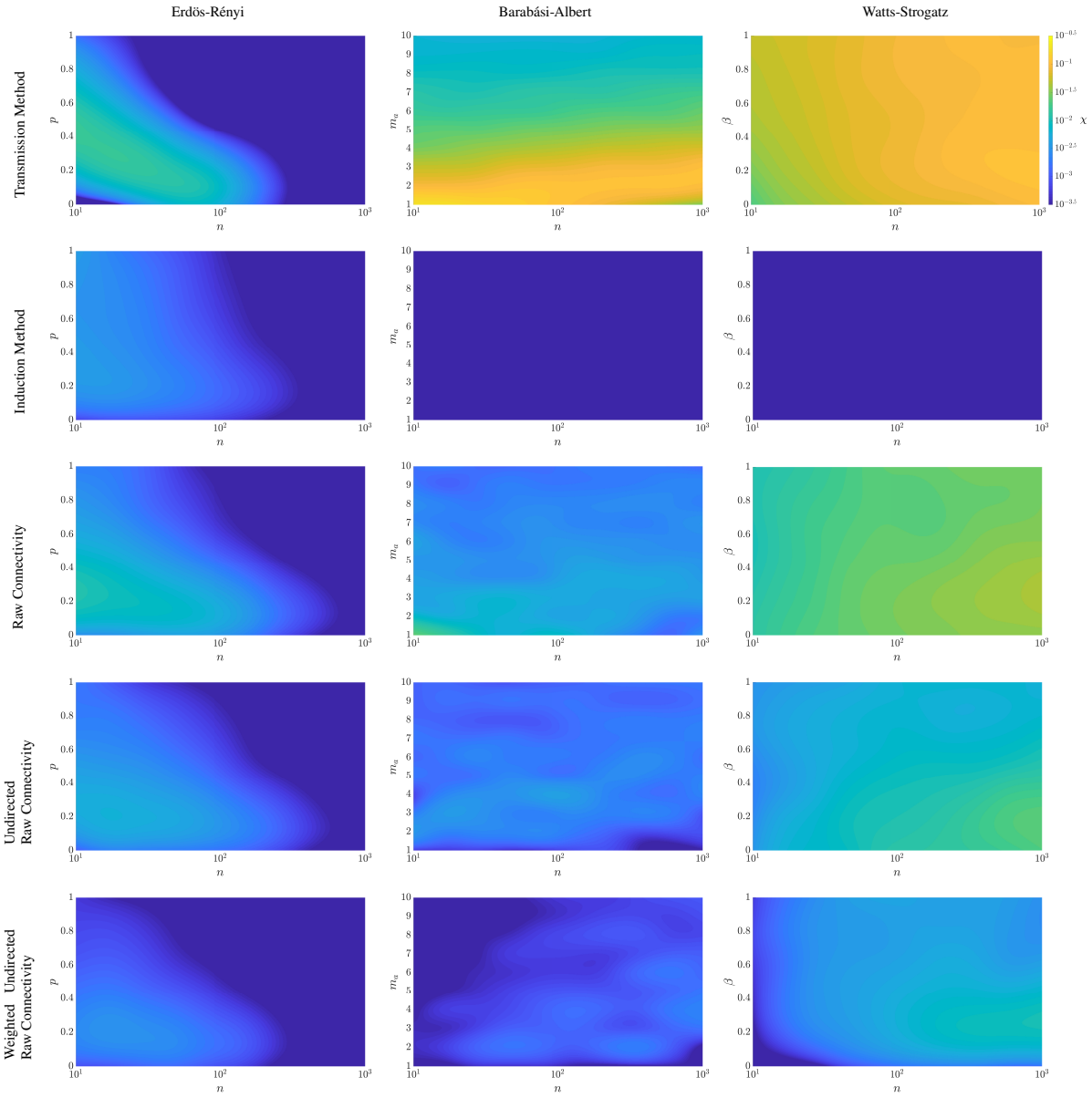


Figure 6.8: Average value of χ for different methods of obtaining dynamical adjacency matrix $\mathbf{A}$ from static connectivity $\mathbf{C}$. The plots show the effect of these methods on TVCS. The details on how to obtain the plots are similar to Figure 6 in the main text. All matrices are normalized by their spectral radius for uniformity and comparison. The plots show a sizable enhancement (respectively depression) of χ by the transmission (respectively induction) method compared to raw connectivity, except for Erdös-Rényi networks whose χ maintains a robust pattern irrespective of the method of obtaining the dynamic adjacency matrix $\mathbf{A}$ from the raw static connectivity $\mathbf{C}$.

Since $\mathcal{W}_K^{-1}$ is positive definite, this is a hyper-ellipsoid in $\mathbb{R}^n$, with axes aligned with the eigenvectors of $\mathcal{W}_K$. Let $(\lambda_i, \mathbf{v}_i)$ be an eigen-pair of $\mathcal{W}_K$ and $\mathbf{x}_f = c\mathbf{v}_i$. Then,

$$\mathbf{x}_f^T \mathcal{W}_K^{-1} \mathbf{x}_f \leq 1 \Leftrightarrow c^2 \lambda_i^{-1} \leq 1 \Leftrightarrow |c| \leq \lambda_i^{1/2},$$

showing that the axis lengths of this hyper-ellipsoid are given by the square roots of the eigenvalues of $\mathcal{W}_K$. Intuitively, the "larger" the reachability hyper-ellipsoid, the "more controllable" the network dynamics (equation (6.1)) are. To quantify how large the hyper-ellipsoid is, several measures based on the eigenvalues of $\mathcal{W}_K$ have been proposed in the literature [6, 8, 66]. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ denote the eigenvalues of $\mathcal{W}_K$. The most widely used Gramian-based measures are

- $\text{tr}(\mathcal{W}_K) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$,
- $\text{tr}(\mathcal{W}_K^{-1})^{-1} = (\lambda_1^{-1} + \lambda_2^{-1} + \cdots + \lambda_n^{-1})^{-1}$,
- $\det(\mathcal{W}_K) = \lambda_1 \lambda_2 \cdots \lambda_n$,
- $\lambda_{\min}(\mathcal{W}_K) = \lambda_n$.

It is clear from these relationships that all these measures, except for $\text{tr}(\mathcal{W}_K)$, approach 0 if $\lambda_n \to 0$. This property, i.e., the behavior of a measure as $\lambda_n \to 0$, is the most critical difference between $\text{tr}(\mathcal{W}_K)$ and the other three measures. For the rest of this discussion, let $f_c(\cdot)$ be any of $\text{tr}((\cdot)^{-1})^{-1}$, $\det(\cdot)$, or $\lambda_{\min}(\cdot)$. Since the network is (Kalman-) controllable if and only if $\lambda_n > 0$, having $f_c(\mathcal{W}_K) > 0$ guarantees network controllability while $\text{tr}(\mathcal{W}_K) > 0$ does not. This is a major disadvantage of $\text{tr}(\mathcal{W}_K)$ for small networks, where controllability in all directions in state space is

both achievable and desirable. As the size of the network grows, however, λ_n typically decays exponentially fast to zero [6], irrespective of network structure. This exponential decay of worst-case controllability is even evident in the example network of Figure 6.1(a) comprising of only $n = 5$ nodes.

Computationally, this means that λ_n (and in turn $f_c(\mathcal{W}_K)$) can quickly drop below machine precision as n grows. In fact, for $K = 10$ and double-precision arithmetics, this happens for $n \sim 15$, making the TVCS (equation (6.5)) with $f = f_c$ numerically infeasible (as it involves the comparison of $f_c(\mathcal{W}_K)$ for different $\{b_k\}_{k=0}^{K-1}$, which may be zero up to machine accuracy). Further, notice that the computational complexity of TVCS for $f = f_c$ grows as n^K due to the NP-hardness of TVCS, enforcing the use of sub-optimal greedy algorithms even if machine precision was not a concern (see [16] and the references therein for details).

In addition to the computational aspects of TVCS, the exponential decay of λ_n also has theoretical implications for the choice of f. When using $f = f_c$, TVCS seeks to assign the control nodes $\{\imath_k\}_{k=0}^{K-1}$ such that controllability is maintained in all directions in the state space, with special emphasis on the hardest-to-reach directions. The use of $\mathrm{tr}(\mathcal{W}_K)$, on the other hand, involves maximizing the average of Gramian eigenvalues, which usually strengthens the largest eigenvalues and spares the few smallest ones. In large networks, the latter is in general more realistic as controllability is hardly needed in all n directions of the state space. As discussed in detail in [67], this seems to be the case in the resting-state structural brain networks: this paper shows that $\mathrm{tr}(\mathcal{W}_K)$ is maximized by controlling specific brain regions that have long been identified as the structural "core" or "hubs" of the cerebral cortex, while the Gramian is itself close to singular.

Further, due to the same strong dependence of $f_c(\mathcal{W}_K)$ but not $\mathrm{tr}(\mathcal{W}_K)$ on λ_n, we often ob-

serve that $\mathrm{tr}(\mathcal{W}_K)$ is significantly less sensitive to the choice of the control nodes $\{\iota_k\}_{k=0}^{K-1}$, leading to orders of magnitude smaller χ than that of $f_c(\mathcal{W}_K)$ (Figure 6.1(b)). This means that $\mathcal{V}$ is only a small subclass of networks that benefit from TVCS measured by f_c. This also has a clear interpretation, since maintaining controllability in all directions in the state space requires a broader distribution of the control nodes that facilitates the reach of the control action $\{u(k)\}_{k=0}^{K-1}$ to all the nodes in the network.

Finally, we highlight the need for development and analysis of measures that are neither strongly reliant on the least controllable directions (such as $f_c(\mathcal{W}_K)$) nor mainly ignore them (such as $\mathrm{tr}(\mathcal{W}_K)$). Two such candidates are:

- $\mathrm{tr}(\mathbf{C}^T\mathcal{W}_K\mathbf{C})$ where $\mathbf{C}$ is a matrix (or vector) with columns that point towards some particular directions of interest in the state space. This measure is a modular set function similar to $\mathrm{tr}(\mathcal{W}_K)$ [8], but the extensions of the notion of $2k$-communicability and the relationship between class $\mathcal{I}/\mathcal{V}$ networks and scale-heterogeneity are unclear;

- appropriate approximations of $\log(f_c(\mathcal{W}_K))$. While computing the exact value of $\log(f_c(\mathcal{W}_K))$ is subject to the same issues as $f_c(\mathcal{W}_K)$ itself, approximations can be used that provide a *mitigation* of the effects of the smallest eigenvalues of $\mathcal{W}_K$. In the case of $f_c(\cdot) = \det(\cdot)$, e.g., various algorithms have been proposed to approximate $\log\det$ of large matrices, see, e.g. [68–74]. These algorithms, however, are predominantly designed with the aim of reducing the computational complexity of determinant calculation and not mitigation of the effects of its high condition number, and often rely on assumptions (such as sparsity or knowledge of lower and upper bounds on matrix eigenvalues) that do not apply to $\mathcal{W}_K$. Thus, development of *appropriate* approximations of $\log(f_c(\mathcal{W}_K))$ constitutes a warranted

direction for future research.

6.C Relationships Between $2k$-Communicability, Degree, and Eigenvector Centrality

The notion of $2k$-communicability introduced in this article has close connections with the degree and eigenvector centrality in the limit cases of $k = 1$ and $k \to \infty$, respectively. Recall that the out-degree centrality and 2-communicability of a node $i \in \mathcal{N}$ are defined as, respectively,

$$d_i^{\text{out}} = \sum_{j=1}^{n} a_{ji},$$
$$R_i(1) = \sum_{j=1}^{n} a_{ji}^2.$$

Therefore, if the network is unweighted (i.e., all the edges have the same weight), then $R_i(1) \propto d_i^{\text{out}}$, so 2-communicability and out-degree centrality result in the same ranking of the nodes (in particular, $r(1)$ is the node with the largest out-degree). As edge weights become more heterogenous, these two rankings become less correlated, with 2-communicability putting more emphasis on stronger weights.

A similar relation exists between ∞-communicability and left eigenvector centrality, as we show next. Let $\mathbf{v}_1, \mathbf{u}_1 \in \mathbb{R}^n$ be the right and left Perron-Frobenius eigenvectors of $\mathbf{A}$, respectively, normalized such that $\mathbf{v}_1^T \mathbf{v}_1 = \mathbf{u}_1^T \mathbf{v}_1 = 1$ (notice that $\mathbf{u}_1$ has unit inner product with $\mathbf{v}_1$ but does *not* in general have unit length). Since the network is by assumption strongly connected and aperiodic,

we have

$$\lim_{k\to\infty}\left(\frac{1}{\rho(\mathbf{A})}\mathbf{A}\right)^k = \mathbf{v}_1\mathbf{u}_1^T. \tag{6.14}$$

Thus for any $i \in \mathcal{N}$,

$$\lim_{k\to\infty}\left(\frac{1}{\rho(\mathbf{A})}\right)^{2k} R_i(k) = \lim_{k\to\infty}\left(\frac{1}{\rho(\mathbf{A})}\right)^{2k}\left((\mathbf{A}^k)^T\mathbf{A}^k\right)_{ii} = (\mathbf{u}_1\mathbf{v}_1^T\mathbf{v}_1\mathbf{u}_1^T)_{ii} = u_{1,i}^2.$$

Given that dividing $R_i(k)$ by $\rho(\mathbf{A})^{2k}$ for all i does not change the ranking of nodes, we define $R_i(\infty) = u_{1,i}^2$ for all i. Since squaring non-negative numbers preserves their order, nodal rankings based on ∞-communicability and left eigenvector centrality are identical.

6.D Nodal Dominance

Among the networks where the nodes with the greatest $R_i(1)$ and $R_i(\infty)$ coincide (i.e., $r(0) = r(\infty)$), there is a higher chance (than in general) that any network belongs to class $\mathcal{I}$. However, about half of these networks still belong to class $\mathcal{V}$, meaning that there exists $1 < k < \infty$ such that $r(k) \neq r(0)$. To assess the importance of this time-variation of optimal control nodes, we define the *dominance* of the node $r(0)$ (over the rest of the network) as follows. Let $r'(0)$ be the index of the node with the second largest $R_i(1)$ (largest after removing $r(0)$). Similarly, let $r'(\infty)$ be the index of the second largest $R_i(\infty)$. We define

$$\text{Dominance of } r(0) = \min\left\{\frac{R_{r(0)}(0) - R_{r'(0)}(0)}{R_{r(0)}(0)}, \frac{R_{r(0)}(\infty) - R_{r'(\infty)}(\infty)}{R_{r(0)}(\infty)}\right\}.$$

A small dominance indicates that another node has very similar value $R_i(0)$ *or* $R_i(\infty)$ to $r(0)$, while a large dominance is an indication of a large gap between $R_{r(0)}(k)$ and the next largest $R_i(k)$ for both $k = 0$ *and* $k \to \infty$.

6.E Networks with Multiple Inputs

Consider a multiple-input network, namely, a network in which $m \geq 1$ nodes are controlled at every time step. Let $\imath_k^1, \dots, \imath_k^m \in \mathcal{N}$ denote the indices of the control nodes at every time k, and $\boldsymbol{\imath}_k = \{\imath_k^1, \dots, \imath_k^m\}$. Then, the corresponding TICS and TVCS are defined as

$$\max_{\boldsymbol{\imath}_0,\dots,\boldsymbol{\imath}_{K-1} \in \mathcal{N}} \quad f(\boldsymbol{\mathcal{W}}_K) \tag{6.15a}$$

$$\text{s.t.} \qquad \boldsymbol{\imath}_0 = \cdots = \boldsymbol{\imath}_{K-1} \tag{6.15b}$$

and

$$\max_{\boldsymbol{\imath}_0,\dots,\boldsymbol{\imath}_{K-1} \in \mathcal{N}} \quad f(\boldsymbol{\mathcal{W}}_K), \tag{6.16}$$

respectively. Accordingly, a multiple-input network is said to belong to class $\mathcal{I}$ if the solution of (6.16) satisfies (6.15b), and to class $\mathcal{V}$ otherwise.

Clearly, for a multiple-input network to belong to class $\mathcal{V}$, any of the first m largest of $\{R_i(k)\}_{i=1}^n$ should change over time, which is often implied by (but does not imply) a change in $r(k)$. Therefore, the condition of Theorem 6.3.3 can still be used as a tight proxy for networks in $\mathcal{V}$, but is too conservative and can be relaxed as follows: assume that $\mathbf{A}$ is irreducible, aperiodic, and di-

agonalizable. Let $\{r_j^d \in \mathbb{R}^n \mid j \in \mathcal{J}^d\}$ be the set of nodes with the m highest 2-communicabilities, where the index set $\mathcal{J}^d$ accounts for different choices of rankings if there are nodes with equal 2-communicabilities. Similarly, let $\{r_j^c \in \mathbb{R}^n \mid j \in \mathcal{J}^c\}$ be the set of nodes with the m highest centralities. Then, if $r_{j_1}^d \neq r_{j_2}^c$ for all $(j_1, j_2) \in \mathcal{J}^d \times \mathcal{J}^c$, the network belongs to class $\mathcal{V}$ when K is sufficiently large. The proof of this statement is a straightforward generalization of the proof of Theorem 6.3.3 and thus omitted.

Similarly, the three conditions in Theorem 6.3.4 can be generalized to undirected multiple-input networks as follows (with similar proofs as the proof of Theorem 6.3.4):

(i) For all $i \in \{1, \dots, m\}$,

$$\frac{1 - w_{i1}}{\sum_{\ell \leq m+1, \ell \neq i+1} w_{\ell 1}} \leq \frac{|\lambda_1| - |\lambda_2|}{|\lambda_1| - |\lambda_n|}.$$

This condition can be simplified, at the expense of being more conservative, to $\frac{1-w_{i1}}{i w_{i1}} \leq \frac{|\lambda_1|-|\lambda_2|}{|\lambda_1|-|\lambda_n|}$, for all $i \in \{1, \dots, m\}$,

(ii) for all $i \in \{1, \dots, m\}$, $w_{i2} = 1 - w_{i1}$,

(iii) the network has three or fewer nonzero eigenvalues with different absolute values and

$$R_1(1) \geq R_2(1) \geq \cdots \geq R_m(1) \geq R_i(1),$$

for all $i \in \{m + 1, \dots, n\}$.

Finally, Figure 6.9 illustrates the effect of m on χ-values of ER, BA, and WS networks discussed in the main text.

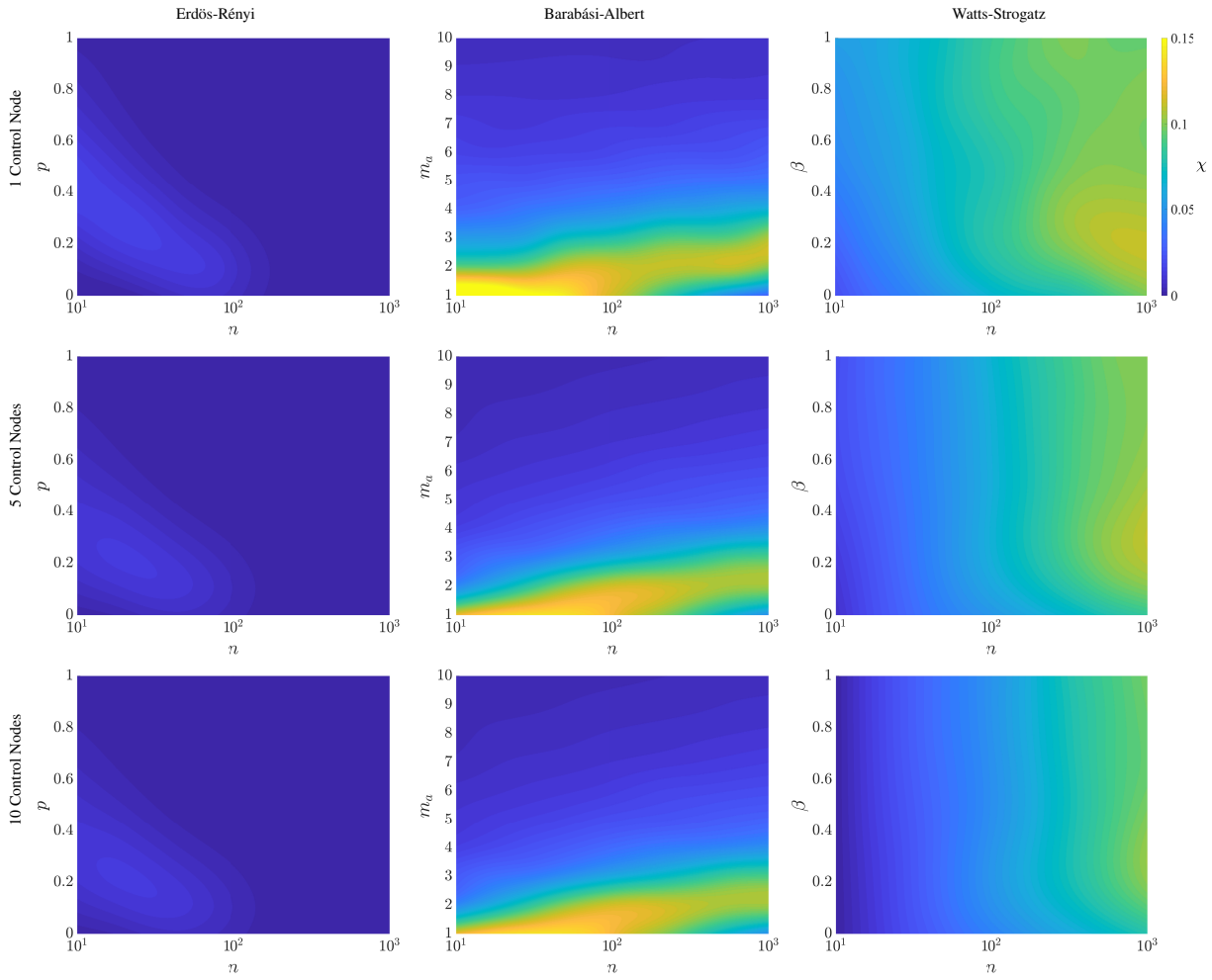


Figure 6.9: Average value of χ for networks with increasing number of inputs. The plots show the effect of multiple inputs on TVCS. The details on how to obtain the plots are similar to Figure 6 in the main text. The dynamic adjacency matrix $\mathbf{A}$ is obtained from the raw static connectivity $\mathbf{C}$ using the transmission method in all cases. These plots show a slight depression in the benefit of TVCS as the number of control nodes grows, despite the fact that networks with more control nodes have a higher probability of belonging to $\mathcal{V}$ (namely, having $\chi > 0$).

6.F $2k$-Communicabilities of Simple Networks

As mentioned in the main text, cf. Figure 5, the simple structure of homogeneous undirected line, ring, and star networks allows us to compute their $2k$-communicabilities analytically in closed form, as derived in the following. Throughout, $\mathbb{Z}$ denotes the set of integers and for $a, b \in \mathbb{Z}$, $a \,|\, b$ denotes that a divides b.

Proposition 6.F.1. *($2k$-**communicabilities of line networks**). Consider a line network of n nodes with uniform link weights a (and no self-loops). Then, for $i \in \mathcal{N}$ and $k \in \mathbb{N}$,*

$$R_i(k) = a^{2k} \sum_{p \in \mathcal{I}} \left[\binom{2k}{k+p(n+1)} - \binom{2k}{k+p(n+1)-i} \right], \tag{6.17}$$

where $\mathcal{I} = \{-\lceil \frac{k}{n+1} \rceil, \dots, \lceil \frac{k}{n+1} \rceil\}$ and $\binom{n}{k} \triangleq 0$ if $k \notin \{0, \dots, n\}$. In particular, if $i \leq \lceil \frac{n}{2} \rceil$ and $k \leq \lceil \frac{n}{2} \rceil - 1$,

$$R_i(k) = a^{2k} \left[\binom{2k}{k} - \binom{2k}{k-i} \right]. \tag{6.18}$$

Proof. From [75, Lemma 1.77], we have

$$\lambda_j = 2a \cos \frac{j\pi}{n+1} \quad \text{and} \quad w_{ij} \propto \sin^2 \frac{ij\pi}{n+1},$$

for $i, j \in \{1, \dots, n\}$ where $\propto$ accounts for normalization. In order to normalize the eigenvectors, we use the identities $\sin^2 \alpha = \frac{1}{2}(1 - \cos 2\alpha)$ and

$$\sum_{j=1}^{n} \cos \frac{2sj\pi}{n+1} = -1 \quad \text{for all } s \nmid n+1, \tag{6.19}$$

to get $w_{ij} = \frac{2}{n+1} \sin^2 \frac{ij\pi}{n+1}$ for all $i, j \in \{1, \dots, n\}$ (one can show (6.19) by multiplying and dividing the LHS by $\sin \frac{s\pi}{n+1}$ and using the identity $2 \sin \alpha \cos \beta = \sin(\alpha + \beta) + \sin(\alpha - \beta)$ for each term). Thus, by substitution, we have $R_i(k) = \frac{2a^{2k}}{n+1} \sum_{j=1}^{n} \tau_{ijk}^2$ where

$$\tau_{ijk} = 2^k \sin \frac{ij\pi}{n+1} \cos^k \frac{j\pi}{n+1}.$$

By using the identity $2 \sin \alpha \cos \beta = \sin(\alpha + \beta) + \sin(\alpha - \beta)$, k times and collecting terms, we get

$$\tau_{ijk} = \sum_{\ell=0}^{k} \binom{k}{\ell} \sin \frac{(i + k - 2\ell)j\pi}{n+1}.$$

Hence, by squaring τ_{ijk} and substituting it in $R_i(k)$, and using the identity $2 \sin \alpha \sin \beta = \cos(\alpha - \beta) - \cos(\alpha + \beta)$, we get

$$R_i(k) = \frac{a^{2k}}{n+1} \sum_{\ell,r=0}^{k} \binom{k}{\ell} \binom{k}{r} \Big[\sum_{j=1}^{n} \cos \frac{2(\ell - r)j\pi}{n+1} - \sum_{j=1}^{n} \cos \frac{2(i + k - \ell - r)j\pi}{n+1} \Big]. \tag{6.20}$$

However, by (6.19), the two sums in (6.20) cancel each other unless $\ell - r \,|\, n+1$ or $i+k-\ell-r \,|\, n+1$ (the cases where both of these happen need not be excluded since they automatically cancel). Thus,

$$R_i(k) = a^{2k} \left[\sum_{\mathcal{I}_1} \binom{k}{\ell} \binom{k}{r} - \sum_{\mathcal{I}_2} \binom{k}{\ell} \binom{k}{r} \right], \tag{6.21}$$

where

$$\mathcal{I}_1 = \{(\ell, r) \in \{0, \dots, k\}^2 \mid \ell - r \,|\, n+1\},$$

$$\mathcal{I}_2 = \{(\ell, r) \in \{0, \dots, k\}^2 \mid i + k - \ell - r \,|\, n+1\}.$$

Defining $p = \frac{n+1}{\ell - r}$ in the first and $p = \frac{i+k-\ell-r}{n+1}$ in the second sum in (6.21), we get

$$R_i(k) = a^{2k} \sum_{p \in \mathcal{I}} \left[\sum_{\ell=0}^{k} \binom{k}{\ell} \binom{k}{\ell - p(n+1)} - \sum_{\ell=0}^{k} \binom{k}{\ell} \binom{k}{\ell + p(n+1) - i} \right], \tag{6.22}$$

where we have used the identity $\binom{k}{s} = \binom{k}{k-s}$. Equation (6.17) then follows by applying the formula $\sum_{\ell=0}^{k} \binom{k}{\ell} \binom{k}{\ell \pm s} = \binom{2k}{k \pm s}$ [76, Eq. 6.69-70] to each of the two sums in (6.22). To get (6.18), note that if $i \leq \lceil \frac{n}{2} \rceil$ and $k \leq \lceil \frac{n}{2} \rceil - 1$, then the only nonzero term in (6.17) is the one corresponding to $p = 0$. □

According to this result, in the case of no self-loops, the value of $R_i(k)$ increases with i until $i = \lceil \frac{n}{2} \rceil$ (i.e., the middle node) for $k \leq \lceil \frac{n}{2} \rceil - 1$ (this can be observed from the expression (6.18)). For general k, it can be shown that the value of the sum in (6.17) for $R_i(k)$ is strongly dominated by the summand corresponding to the index $p = 0$, which increases with i until $i = \lceil \frac{n}{2} \rceil$ and decreases afterwards. Thus, **the optimal control node corresponds always to (one of) the center node(s)**, i.e., $\mathbf{b}^*(k) = \mathbf{e}_{\lceil \frac{n}{2} \rceil}$ for all k. If nodes have uniform self-loops (i.e., self-loops all with the same weight), $R_i(k)$ can no longer be computed analytically but simulations show the exact same behavior;

Proposition 6.F.2. *(2k-communicabilities of ring networks). Consider a ring network of n nodes and uniform link weights a (with no self-loops). Then, for $i \in \mathcal{N}$ and $k \in \mathbb{N}$,*

$$R_i(k) = \frac{(2a)^{2k}}{n} \left[1 + 2 \sum_{j=1}^{\lceil \frac{n}{2} \rceil - 1} \cos^{2k} \left(\frac{2j\pi}{n} \right) + \delta_n^E \right], \tag{6.23}$$

where δ_n^E equals one if n is even and zero otherwise.

Proof. From [75, Lemma 1.77], we have $\lambda_j = 2a\cos\frac{2j\pi}{n}$ and (after normalization of eigenvectors),

$$w_{ij} = \begin{cases} \frac{2}{n}\cos^2\frac{2(i-1)j\pi}{n} & \text{if } 1 \le j < \frac{n}{2}, \\ \frac{2}{n}\sin^2\frac{2(i-1)(n-j)\pi}{n} & \text{if } \frac{n}{2} < j < n, \\ \frac{1}{n} & \text{if } j = n, \text{ or } n \in \mathbb{Z} \text{ even and } j = \frac{n}{2}, \end{cases}$$

for $i, j \in \{1, \dots, n\}$. Note that to normalize the eigenvectors, we follow a similar procedure to the one described in the proof of Lemma 6.F.1 (setting $s = 2i$ and substituting n by $n-1$ in (6.19)). The result then follows by substituting these expressions in $R_i(k)$. □

We can infer from the preceding result that without self-loops, the value of $R_i(k)$ is independent of i (as shown by (6.23)) for a uniform ring network, so **the optimal control node is arbitrary for all** k. Similar result can be proved analytically if the nodes have uniform self-loops.

Proposition 6.F.3. ***(2k-communicabilities of star networks).*** *Consider a star network given by*

$$\mathbf{A} = \begin{bmatrix} l_c & \mathbf{a}^T \\ \mathbf{a} & l_p\mathbf{I}_{n-1} \end{bmatrix}, \tag{6.24}$$

where $\mathbf{a} \in \mathbb{R}^{n-1}$ *contains the link weights between the center node and peripheral nodes. Then*

$$\begin{aligned} R_1(k) &= \frac{(\lambda_1 - l_p)^2}{(\lambda_1 - l_p)^2 + \|\mathbf{a}\|^2}\lambda_1^{2k} + \frac{(l_p - \lambda_2)^2}{(l_p - \lambda_2)^2 + \|\mathbf{a}\|^2}\lambda_2^{2k}, \\ R_i(k) &= \frac{a_{i-1}^2}{(\lambda_1 - l_p)^2 + \|\mathbf{a}\|^2}\lambda_1^{2k} + \frac{a_{i-1}^2}{(l_p - \lambda_2)^2 + \|\mathbf{a}\|^2}\lambda_2^{2k} + \frac{\|\mathbf{a}\|^2 - a_{i-1}^2}{\|\mathbf{a}\|^2}l_p^{2k}, \end{aligned} \tag{6.25}$$

for all $k \in \mathbb{N} \cup \{0\}$ *and* $i \in \{2, \dots, n\}$*, where*

$$\lambda_{1,2} = \frac{l_c + l_p \pm \sqrt{(l_c - l_p)^2 + 4\|\mathbf{a}\|^2}}{2}. \tag{6.26}$$

Proof. Using the formula

$$\det\left(\begin{bmatrix} P & \mathbf{Q} \\ \mathbf{R} & \mathbf{S} \end{bmatrix}\right) = (P-1)\det(\mathbf{S}) + \det(\mathbf{S} - \mathbf{RQ}),$$

for scalar P, row vector $\mathbf{Q}$, column vector $\mathbf{R}$, and square matrix $\mathbf{S}$, and some algebra, we get $|s\mathbf{I}_n - \mathbf{A}| = \left(s^2 - (l_c + l_p)s + l_c l_p - \|\mathbf{a}\|^2\right)(s - l_p)^{n-2}$, so the eigenvalues of $\mathbf{A}$ are given by

$$\lambda_{3,\dots,n} = l_p, \tag{6.27}$$

and $\lambda_{1,2}$ in (6.26). Note that we may or may not have $|\lambda_1| \geq \cdots \geq |\lambda_n|$ as the order depends on the values of the parameter. By solving $(\mathbf{A} - \lambda_j \mathbf{I}_n)\mathbf{v}_j = 0$ for $j = 1, 2$, and then using the orthogonality of eigenvectors, we get

$$\mathbf{v}_{1,2} \propto \begin{bmatrix} \lambda_{1,2} - l_p \\ \mathbf{a} \end{bmatrix}, \quad (\mathbf{v}_j)_1 = 0 \quad \forall j \in \{3, \dots, n\}, \tag{6.28}$$

where $\propto$ accounts for normalization. The result then follows by substituting (6.26)-(6.28) into $R_i(1) = \sum_j v_{ij}^2 \lambda_j^2$ separately for $i = 1$ and $i \geq 2$, and simplifying. □

Using this result, if all self-loop weights are the same ($l_c = l_p$ in (6.24)), then $R_1(1) > R_i(1)$ for all $i \geq 2$ from (6.12). Therefore Theorem 6.3.4(iii) implies that **the center node is the optimal**

control node at all times.

6.G Additional Lemmas and Proofs

In this section, we formulate and prove a number of lemmas that underlie the main results of this chapter and also provide the proofs of the main results presented in the main text. Throughout, $\mathbb{C}$ denotes the set of complex numbers and for $\mathbf{M} \in \mathbb{C}^{n\times n}$, $\overline{\mathbf{M}}$ and $\mathbf{M}^*$ denote its complex conjugate and complex conjugate transpose, respectively, and $\mathbf{M}^{-*} = (\mathbf{M}^*)^{-1}$. Further, for $\lambda \in \mathbb{R}^n$ and $\ell \in \mathbb{N} \cup \{0\}$, $\lambda^\ell \triangleq [\lambda_1^\ell \ \cdots \ \lambda_n^\ell]^T$ and $|\lambda| \triangleq [|\lambda_1| \ \cdots \ |\lambda_n|]^T$.

Proof of Theorem 6.3.3. Define

$$\mathbf{U} = \mathbf{V}^{-\star}.$$

Notice that the columns of $\mathbf{U}$ are the left eigenvectors of $\mathbf{A}$, with the same order as in $\mathbf{\Lambda}$ and $\mathbf{V}$. Since for any k,

$$(\mathbf{A}^k)^T\mathbf{A}^k = (\mathbf{A}^k)^*\mathbf{A}^k = (\mathbf{V}\mathbf{\Lambda}^k\mathbf{U}^*)^*\mathbf{V}\mathbf{\Lambda}^k\mathbf{U}^*,$$

it follows that for any i and k,

$$R_i(k) = \left((\mathbf{A}^k)^T\mathbf{A}^k\right)_{ii} = (\mathbf{V}\mathbf{\Lambda}^k\mathbf{U}_{i,:}^*)^*\mathbf{V}\mathbf{\Lambda}^k\mathbf{U}_{i,:}^* = \|\mathbf{V}\mathbf{\Lambda}^k\mathbf{U}_{i,:}^*\|_2^2,$$

where $\mathbf{U}_{i,:}$ denotes the ith row of $\mathbf{U}$. For simplicity, define $\mathbf{c}^{(i,k)} = \mathbf{V}\mathbf{\Lambda}^k\mathbf{U}_{i,:}^* \in \mathbb{C}^n$. It is straightfor-

ward to check that

$$c_\ell^{(i,k)} = \sum_{j=1}^{n} v_{\ell j}\overline{u}_{ij}\lambda_j^k, \qquad \ell \in \{1,\dots,n\}$$

so

$$\begin{aligned} R_i(k) = \sum_{\ell=1}^{n}\left|c_\ell^{(i,k)}\right|^2 = \sum_{\ell=1}^{n} c_\ell^{(i,k)}\overline{c_\ell^{(i,k)}} &= \sum_{\ell=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n} v_{\ell j}\overline{v}_{\ell m}\overline{u}_{ij}u_{im}\lambda_j^k\overline{\lambda}_m^k \\ &= \sum_{j,m=1}^{n}\underbrace{\left(\sum_{\ell=1}^{n} v_{\ell j}\overline{v}_{\ell m}\overline{u}_{ij}u_{im}\right)}_{\beta_i^{(j,m)}}\lambda_j^k\overline{\lambda}_m^k. \end{aligned}$$

Dividing both sides by λ_1^{2k} and taking the limits as $k \to \infty$, we see that for all i, $\beta_i^{(1,1)} = u_{i1}^2 = R_i(\infty)$ (notice that $u_{i1} \in \mathbb{R}$ for all i since $\lambda_1 \in \mathbb{R}_{>0}$ according to the Perron-Frobenius Theorem [77, Fact 4.11.4]). Choose $r(1) \in \arg\max_i R_i(1)$ and $r(\infty) \in \arg\max_i R_i(\infty)$. The network belongs to class $\mathcal{V}$ if for some $k > 1$,

$$\begin{aligned} R_{r(\infty)}(k) > R_{r(1)}(k) &\Leftrightarrow R_{r(\infty)}(\infty)\lambda_1^{2k} + \sum_{(j,m)\neq(1,1)}\beta_{r(\infty)}^{(j,m)}\lambda_j^k\overline{\lambda}_m^k > R_{r(1)}(\infty)\lambda_1^{2k} + \sum_{(j,m)\neq(1,1)}\beta_{r(1)}^{(j,m)}\lambda_j^k\overline{\lambda}_m^k \\ &\Leftrightarrow \left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]\lambda_1^{2k} > \sum_{(j,m)\neq(1,1)}\left[\beta_{r(1)}^{(j,m)} - \beta_{r(\infty)}^{(j,m)}\right]\lambda_j^k\overline{\lambda}_m^k \\ &\overset{\text{(a)}}{\Leftarrow} \left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]\lambda_1^{2k} > \lambda_1^k|\lambda_2|^k\left|\sum_{(j,m)\neq(1,1)}\beta_{r(1)}^{(j,m)} - \beta_{r(\infty)}^{(j,m)}\right| \\ &\Leftrightarrow \left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]\lambda_1^{2k} > \lambda_1^k|\lambda_2|^k\sum_{(j,m)\neq(1,1)}\left|\beta_{r(1)}^{(j,m)}\right| + \left|\beta_{r(\infty)}^{(j,m)}\right| \\ &\Leftarrow \left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]\lambda_1^{k} > |\lambda_2|^k \cdot 2\max_{i\in\{1,\dots,n\}}\sum_{j,m=1}^{n}\left|\beta_i^{(j,m)}\right|, \end{aligned}$$

where in (a) we have used the fact that $|\lambda_j\overline{\lambda}_m| \leq \lambda_1|\lambda_2|$ for any $(j,m) \neq (1,1)$. Now, using the

definition of $\beta_i^{(j,m)}$,

$$\begin{aligned}\sum_{j,m=1}^{n}\left|\beta_i^{(j,m)}\right| &\leq \sum_{j,m=1}^{n}\sum_{\ell=1}^{n}|v_{\ell j}||v_{\ell m}||u_{ij}||u_{im}| \\ &= \sum_{j,m=1}^{n}|u_{ij}||u_{im}|\left(\sum_{\ell=1}^{n}|v_{\ell j}||v_{\ell m}|\right) \\ &\overset{\text{(b)}}{\leq} \sum_{j,m=1}^{n}|u_{ij}||u_{im}|\underbrace{\|\mathbf{V}_{:,j}\|_2}_{1}\underbrace{\|\mathbf{V}_{:,m}\|_2}_{1} = \|\mathbf{U}_{i,:}\|_1^2 \leq \|\mathbf{U}\|_\infty^2,\end{aligned}$$

where (b) follows from the Cauchy-Schwarz inequality. Thus,

$$\begin{aligned}R_{r(\infty)}(k) > R_{r(1)}(k) &\Leftarrow \left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]\lambda_1^k > |\lambda_2|^k \cdot 2\|\mathbf{U}\|_\infty^2 \\ &\Leftrightarrow k > \frac{\log 2\|\mathbf{U}\|_\infty^2 - \log\left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]}{\log\lambda_1 - \log|\lambda_2|}.\end{aligned}$$

Therefore, the result follows by choosing $K > \bar{K}$, where

$$\bar{K} = \lceil\frac{\log 2\|\mathbf{U}\|_\infty^2 - \log\left[R_{r(\infty)}(\infty) - R_{r(1)}(\infty)\right]}{\log\lambda_1 - \log|\lambda_2|}\rceil.$$

□

The following lemma will be useful in the proof of Theorem 6.3.4.

Lemma 6.G.1. *Let* $\mathbf{W} \in \mathbb{R}^{n\times n}$ *be a doubly-stochastic matrix and* $\boldsymbol{\gamma} \in \mathbb{R}^n_{\geq 0}$ *be such that* $\gamma_1 \geq \cdots \geq \gamma_n$. *If* $\frac{1-w_{11}}{w_{11}} \leq \frac{\gamma_1-\gamma_2}{\gamma_1-\gamma_n}$, *then* $1 \in \arg\max_{1\leq i\leq n}(\mathbf{W}\gamma)_i$.

Proof. Note that we have

$$\begin{aligned}\frac{1-w_{11}}{w_{11}} \leq \frac{\gamma_1-\gamma_2}{\gamma_1-\gamma_n} &\Leftrightarrow (\gamma_1-\gamma_2)w_{11} \geq (\gamma_1-\gamma_n)(1-w_{11}) \\ &\Leftrightarrow \gamma_n + w_{11}(\gamma_1-\gamma_n) \geq \gamma_2 + (1-w_{11})(\gamma_1-\gamma_2) \\ &\Rightarrow \forall i \geq 2 \quad \gamma_n + w_{11}(\gamma_1-\gamma_n) \geq \gamma_2 + w_{i1}(\gamma_1-\gamma_2),\end{aligned}$$

where the last implication is because $w_{i1} \leq 1 - w_{11}$ for all $i \in \{1, \dots, n\}$. The last inequality can be equivalently expressed, for any $i \in \{2, \dots, n\}$, as

$$w_{11}\gamma_1 + (1-w_{11})\gamma_n \geq w_{i1}\gamma_1 + (1-w_{i1})\gamma_2,$$

which, given that $\gamma_n \leq \gamma_j \leq \gamma_2$ for all $j \in \{2, \dots, n\}$, implies

$$w_{11}\gamma_1 + \sum_{j=2}^{n} w_{1j}\gamma_j \geq w_{i1}\gamma_1 + \sum_{j=2}^{n} w_{ij}\gamma_j,$$

for any $i \in \{2, \dots, n\}$. This can be equivalently written as

$$\sum_{j=1}^{n} w_{1j}\gamma_j \geq \sum_{j=1}^{n} w_{ij}\gamma_j \Leftrightarrow (\mathbf{W}\boldsymbol{\gamma})_1 \geq (\mathbf{W}\boldsymbol{\gamma})_i,$$

completing the proof. □

Proof of Theorem 6.3.4. For convenience, let $\boldsymbol{\lambda} = [\lambda_1 \ \cdots \ \lambda_n]^T$. After some algebraic manipula-

tions, one can show that

$$R_i(k) = (\mathbf{A}^{2k})_{ii} = \sum_{j=1}^{n} v_{ij}^2 \lambda_j^{2k} = (\mathbf{W}\boldsymbol{\lambda}^{2k})_i. \tag{6.29}$$

The assumption that node 1 has the largest eigenvector centrality is equivalent to the largest element of the first column of $\mathbf{W}$ being w_{11}, i.e.,

$$w_{11} = \max_{1 \le i \le n} w_{i1}, \tag{6.30}$$

or, also equivalently, $r(\infty) = 1$. This can always be realized by a permutation of the rows of $\mathbf{W}$ achieved by relabeling the node with the largest centrality as node 1 (note that relabeling the nodes only permutes the rows of $\mathbf{W}$ and not its columns. The order of its columns is arbitrary and corresponds to the order of the diagonal elements of $\mathbf{\Lambda}$).

The claim of the theorem is trivial in all cases for $k = 0$. Under condition *(i)*, for $k = 1$, we have

$$\frac{\lambda_1^2 - \lambda_2^2}{\lambda_1^2 - \lambda_n^2} = \frac{|\lambda_1| - |\lambda_2|}{|\lambda_1| - |\lambda_n|} \frac{|\lambda_1| + |\lambda_2|}{|\lambda_1| + |\lambda_n|} \ge \frac{|\lambda_1| - |\lambda_2|}{|\lambda_1| - |\lambda_n|} \ge \frac{1 - w_{11}}{w_{11}}.$$

For $k \ge 2$, using the above inequality, we have

$$\frac{\lambda_1^{2k} - \lambda_2^{2k}}{\lambda_1^{2k} - \lambda_n^{2k}} = \frac{\lambda_1^2 - \lambda_2^2}{\lambda_1^2 - \lambda_n^2} \frac{\lambda_1^{2k-2} + \cdots + \lambda_2^{2k-2}}{\lambda_1^{2k-2} + \cdots + \lambda_n^{2k-2}} \ge \frac{1 - w_{11}}{w_{11}}.$$

Thus, the result follows from Lemma 6.G.1.

Under condition *(ii)*, for any $k \geq 1$,

$$1 \in \arg\max_{1\leq i\leq n} R_i(k) \Leftrightarrow \sum_{j=1}^{n} w_{1j}\lambda_j^{2k} \geq \sum_{j=1}^{n} w_{ij}\lambda_j^{2k} \Leftrightarrow w_{11}\lambda_1^{2k} + (1-w_{11})\lambda_2^{2k} \geq \sum_{j=1}^{n} w_{ij}\lambda_j^{2k}$$

$$\Leftarrow w_{11}\lambda_1^{2k} + (1-w_{11})\lambda_2^{2k} \geq w_{i1}\lambda_1^{2k} + (1-w_{i1})\lambda_2^{2k}$$

$$\Leftrightarrow (w_{11}-w_{i1})(\lambda_1^{2k} - \lambda_2^{2k}) \geq 0,$$

where the last inequality is always true (cf. equation (6.30)).

Finally, under condition *(iii)*, first consider the case when $|\lambda_1| > |\lambda_2|$. By contradiction, assume $R_i(k) > R_1(k)$ for some $i \in \{2, \dots, n\}$ and $k \geq 2$. Since $|\lambda_1| > |\lambda_i|$ for all $i \in \{2, \dots, n\}$, there exists a sufficiently large $\overline{k}$ where $R_1(\overline{k}) > R_i(\overline{k})$ (recall our node labeling convention in (6.30)). Note that it is not necessary for $\overline{k}$ to be less than K. Thus, R_1 and R_i swap orders at least 2 times. However, since $\mathbf{A}$ has (at most) three distinct nonzero eigenvalues, [78, Theorem 1] implies that R_1 and R_i can swap orders at most once, which is a contradiction. On the other hand, if $|\lambda_1| = |\lambda_2|$, then each R_i is essentially the sum of at most two distinct exponential functions and thus, using [78, Theorem 1] again, the order of all R_i's remains unchanged for all k, yielding the result. □

Proof of Theorem 6.3.5. We first prove the first part of the theorem for general (not necessarily symmetric) $\mathbf{A}_0$ and $\mathbf{E}$. Recall that for $k \in \{0, \dots, K-1\}$

$$r(k) = \arg\max_{i\in\mathcal{N}} R_i(k) = \arg\max_{i\in\mathcal{N}} \left(((\mathbf{A}+\alpha\mathbf{E})^k)^T(\mathbf{A}+\alpha\mathbf{E})^k\right)_{ii}$$

$$\overset{(a)}{=} \arg\max_{i\in\mathcal{N}} \left(((\alpha^{-1}\mathbf{A}+\mathbf{E})^k)^T(\alpha^{-1}\mathbf{A}+\mathbf{E})^k\right)_{ii},$$

where (a) holds because the maximizer of a set is invariant to the scaling of all the elements of the set by a constant. Using $\lim_{\alpha\to\infty} \alpha^{-1}\mathbf{A} + \mathbf{E} = \mathbf{E}$ and the continuity of polynomials, we get

$$\lim_{\alpha\to\infty} R_i(k) = \tilde{R}_i(k),$$

where $\tilde{R}_i$ denotes the $2k$-communicabilities of a node i in the additive network $\mathbf{E}$. Since $\mathbf{E}$ is not acyclic, powers of $\mathbf{E}$ never vanish, and thus

$$\forall k \in \{0, \dots, K-1\}\ \exists i \in \{1, \dots, n_1\} \qquad \tilde{R}_i(k) > 0,$$

while $\tilde{R}_i(k) = 0$ for $i \in \{n_1 + 1, \dots, n\}$ and all k. Therefore, for any $k \in \{0, \dots, K-1\}$, there exists $\bar{\alpha}_k > 0$ such that

$$r(k) \in \{1, \dots, n_1\},$$

for $\alpha > \bar{\alpha}_k$. The claim follows by taking $\bar{\alpha} = \max_{k\in\{0,\dots,K-1\}} \bar{\alpha}_k$.

Now, assume $\mathbf{A}_0$ and $\mathbf{E}$ are symmetric. As before, let $\boldsymbol{\lambda} = [\lambda_1 \ \cdots \ \lambda_n]^T \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ be the vector of eigenvalues (with $|\lambda_1| \geq \cdots \geq |\lambda_n|$) and the matrix of eigenvectors of $\mathbf{A}$, respectively, and $\mathbf{W}$ be the element-wise square of $\mathbf{V}$. Recall that this gives

$$R_i(k) = (\mathbf{A}^{2k})_{ii} = \sum_{j=1}^{n} v_{ij}^2 \lambda_j^{2k} = (\mathbf{W}\boldsymbol{\lambda}^{2k})_i.$$

Let $i^* \in \{1, \dots, n_1\}$ be the node with the greatest eigenvector centrality in $\mathbf{E}$ and $\boldsymbol{\gamma} \in \mathbb{R}^n$ be any vector such that $\gamma_1 \geq \cdots \geq \gamma_n \geq 0$. Fix $i \in \{n_1 + 1, \dots, n\}$ arbitrarily and let $r \leq n_1$ be the rank of

E. Using the inequalities

$$\sum_{j=1}^{n} w_{i^*j}\gamma_j \geq w_{i^*1}\gamma_1,$$
$$\sum_{j=1}^{r} w_{ij}\gamma_j \leq \gamma_1 \sum_{j=1}^{r} w_{ij},$$
$$\sum_{j=r+1}^{n} w_{ij}\gamma_j \leq \gamma_{r+1},$$

it follows that $(\mathbf{W}\gamma)_{i^*} > (\mathbf{W}\gamma)_i$ if

$$w_{i^*1}\gamma_1 > \gamma_1 \sum_{j=1}^{r} w_{ij} + \gamma_{r+1}. \tag{6.31}$$

Note that if (6.31) holds for $\boldsymbol{\gamma} = |\lambda|$, then it holds for $\boldsymbol{\gamma} = \lambda^{2k}$ for any $k \geq 1$. This is because

$$w_{i^*1}\lambda_1^{2k} = |\lambda_1|^{2k-1} \cdot w_{i^*1}|\lambda_1| > |\lambda_1|^{2k-1}\Big(|\lambda_1| \sum_{j=1}^{r} w_{ij} + |\lambda|_{r+1}\Big) > \lambda_1^{2k} \sum_{j=1}^{r} w_{ij} + \lambda_{r+1}^{2k}.$$

Therefore, our proof strategy is to find $\overline{\alpha}$ such that (6.31) holds for $\boldsymbol{\gamma} = |\lambda|$ if $\alpha > \overline{\alpha}$. To this end, let $\tilde{\lambda} = [\tilde{\lambda}_1 \ \cdots \ \tilde{\lambda}_n]^T \in \mathbb{R}^n$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{n\times n}$ be the vector of eigenvalues (with $|\tilde{\lambda}_1| \geq \cdots \geq |\tilde{\lambda}_n|$) and the matrix of eigenvectors of $\mathbf{E}$, respectively, and $\tilde{\mathbf{W}}$ be the element-wise square of $\tilde{\mathbf{V}}$. Note that $\tilde{\mathbf{W}}$ has the structure

$$\tilde{\mathbf{W}} = \begin{bmatrix} \overbrace{\star}^{n_1} & \overbrace{\mathbf{0}}^{n-n_1} \\ \mathbf{0} & \star \end{bmatrix} \begin{matrix} \}n_1 \\ \}n-n_1 \end{matrix}. \tag{6.32}$$

In the following, we bound λ and $\mathbf{V}$ using perturbation theory of eigenvalues and eigenvectors. For

simplicity of exposition, we only deal with the case where the r nonzero eigenvalues of $\mathbf{E}$ are all distinct (the proof for the general case proceeds along the same lines but is more involved).

To bound the eigenvalues in λ, let $\pi_{\mathbf{A}} : \{1, \dots, n\} \to \{1, \dots, n\}$ be a permutation that reorders the eigenvalues of $\mathbf{A}$ based on their *signed* value, i.e., $\lambda_{\pi_{\mathbf{A}}(1)} \geq \lambda_{\pi_{\mathbf{A}}(2)} \geq \cdots \geq \lambda_{\pi_{\mathbf{A}}(n)}$. Define $\pi_{\mathbf{E}}$ similarly for $\mathbf{E}$ (i.e., such that $\tilde{\lambda}_{\pi_{\mathbf{E}}(1)} \geq \tilde{\lambda}_{\pi_{\mathbf{E}}(2)} \geq \cdots \geq \tilde{\lambda}_{\pi_{\mathbf{E}}(n)}$). By Weyl's Theorem [79, Thm 4.3.1],

$$|\lambda_{\pi_{\mathbf{A}}(j)} - \alpha\tilde{\lambda}_{\pi_{\mathbf{E}}(j)}| \leq \rho(\mathbf{A}_0), \tag{6.33}$$

for all $j \in \{1, \dots, n\}$. We know from the Perron-Frobenius theorem [77, Fact 4.11.4] for nonnegative matrices that $\pi_{\mathbf{A}}(1) = \pi_{\mathbf{E}}(1) = 1$. Therefore, (6.33) implies that

$$\alpha\rho(\mathbf{E}) - \rho(\mathbf{A}_0) \leq \lambda_1 \leq \alpha\rho(\mathbf{E}) + \rho(\mathbf{A}_0). \tag{6.34a}$$

Moreover, since $\mathbf{E}$ has $n - r$ zero eigenvalues, (6.33) implies that $\mathbf{A}$ has *at least* $n - r$ eigenvalues with absolute value less than or equal to $\rho(\mathbf{A}_0)$, i.e.,

$$|\lambda_{r+1}| \leq \rho(\mathbf{A}_0). \tag{6.34b}$$

Next, we bound the eigenvectors in $\mathbf{V}$. Define

$$\delta_E = \min\{\tilde{\lambda}_{\pi_{\mathbf{E}}(j)} - \tilde{\lambda}_{\pi_{\mathbf{E}}(j+1)} \mid \tilde{\lambda}_{\pi_{\mathbf{E}}(j)} - \tilde{\lambda}_{\pi_{\mathbf{E}}(j+1)} > 0,\ j \in \{1, \dots, n-1\}\}.$$

Using [80, Cor. 1], we have

$$\|\mathbf{v}_{\pi_{\mathbf{A}}(j)} - \tilde{\mathbf{v}}_{\pi_{\mathbf{E}}(j)}\| \leq \frac{2^{3/2}\|\mathbf{A}_0\|}{\alpha\delta_E}, \tag{6.35}$$

for $j \in \pi_{\mathbf{E}}^{-1}(\{1,\dots,r\})$. To see this, set $\mathbf{\Sigma} = \alpha\mathbf{E}$ and $\hat{\mathbf{\Sigma}} = \mathbf{A}_0$ in [80, Cor. 1]. This is the only place where we need the nonzero eigenvalues of $\mathbf{E}$ to be distinct. If $\mathbf{E}$ has a repeated nonzero eigenvalue, then the corresponding eigenvectors are no longer unique, i.e., one has to study the perturbation of eigenspaces rather than eigenvectors. Therefore, one can no longer use the simplified variant [80, Cor. 1] of the Davis-Kahan Theorem but the original result itself, which provides essentially the same result but is more technically involved.

Using $\pi_{\mathbf{A}}(1) = \pi_{\mathbf{E}}(1) = 1$ and (6.35), we get

$$\begin{aligned} |w_{i^*1} - \tilde{w}_{i^*1}| = |v_{i^*1}^2 - \tilde{v}_{i^*1}^2| &\leq 2||v_{i^*1}| - |\tilde{v}_{i^*1}|| \\ &\leq 2|v_{i^*1} - \tilde{v}_{i^*1}| \leq 2\|\mathbf{v}_1 - \tilde{\mathbf{v}}_1\| \leq \frac{2^{5/2}\|\mathbf{A}_0\|}{\alpha\delta_E}, \end{aligned} \tag{6.36}$$

which together with $\tilde{w}_{i^*1} \geq \frac{1}{n_1}$ gives

$$w_{i^*1} \geq \frac{1}{n_1} - \frac{2^{5/2}\|\mathbf{A}_0\|}{\alpha\delta_E}. \tag{6.37a}$$

To derive similar bounds on $w_{ij}, j \in \{1,\dots,r\}$ (recall that we fixed $i \in \{n_1+1,\dots,n\}$ arbitrarily at the beginning of the proof), we need to choose $\alpha > \frac{2\rho(\mathbf{A}_0)}{|\tilde{\lambda}_r|}$. This choice of α guarantees that $\pi_{\mathbf{A}}(j) \in \{1,\dots,r\}$ for all $j \in \pi_{\mathbf{E}}^{-1}(\{1,\dots,r\})$. Therefore, using (6.35) and (6.32) and following the

same steps as in (6.36), we get

$$w_{ij} \leq \frac{2^{5/2}\|\mathbf{A}_0\|}{\alpha\delta_E}, \quad j \in \{1,\dots,r\}. \tag{6.37b}$$

Now, using (6.34) and (6.37), (6.31) holds with $\boldsymbol{\gamma} = |\boldsymbol{\lambda}|$ if

$$\left(\frac{1}{n_1} - \frac{2^{5/2}\|\mathbf{A}_0\|}{\alpha\delta_E}\right)\left(\alpha\tilde{\lambda}_1 - \rho(\mathbf{A}_0)\right) > \left(\alpha\tilde{\lambda}_1 + \rho(\mathbf{A}_0)\right)\frac{r2^{5/2}\|\mathbf{A}_0\|}{\alpha\delta_E} + \rho(\mathbf{A}_0),$$

which itself holds if $\alpha > \overline{\alpha}$, where

$$\overline{\alpha} \triangleq \max\Big\{1, \frac{2\rho(\mathbf{A}_0)}{\tilde{\lambda}_r}, \frac{8\|\mathbf{A}_0\|}{\delta_E}\Big(1+\frac{\rho(\mathbf{A}_0)}{\rho(\mathbf{E})}\Big)n_1^2 + 2\frac{\rho(\mathbf{A}_0)}{\rho(\mathbf{E})}n_1\Big\},$$

completing the proof. □

6.H Description of the Analyzed Real Networks

The real networks studied in this work have been acquired from a multitude of sources, which we list here for easier reproduction of our results. All the databases are freely and publicly available.

- **BCTNet fMRI [42]:** This is a human whole-brain functional network. Nodes represent brain areas and edges represent fMRI co-activations. The dataset is available online at `https://sites.google.com/site/bctnet/datasets`.
- **Cocomac [43]:** This is a macaque whole-brain structural network based on the Felleman and Van Essen atlas. Nodes represent brain areas and edges represent axonal projections

(nerve tracts) between them. The dataset is retrieved from http://cocomac.g-node.org/services/axonal_projections.php by entering the specifications detailed in http://cocomac.g-node.org/main/faq.php#connectivitymatrix.

- **BCTNet Cat [42]:** This represents the cat structural thalamocortical network. Nodes represent thalamocortical areas and edges represent nerve tracts between them. The dataset is available online at https://sites.google.com/site/bctnet/datasets.

- **C. elegans [44]:** This dataset contains the neural network of Caenorhabditis elegans worm (C. elegans). Nodes represent individual neurons and edges represent the total number of synapses and gap junctions between any pair of neurons. The dataset of available online at https://toreopsahl.com/datasets/#celegans.

- **air500 [45]:** This is the network of the 500 busiest commercial airports in the United States in 2002. Nodes represent airports and edges represent flights between them. The dataset is available online at https://toreopsahl.com/datasets/#usairports.

- **airUS [46]:** This is the complete US airport network in 2010. Nodes and edges represent airports and flights between them, respectively. The dataset is available online at https://toreopsahl.com/datasets/#usairports.

- **airGlobal [46]:** This dataset contains the global airport network according to OpenFlights.org. Nodes and edges represent airports and flights between them, respectively. The dataset is available online at https://toreopsahl.com/datasets/#usairports.

- **Chicago [47, 48]:** This dataset represents the road transportation network of the Chicago region, USA. Nodes are transport nodes while edges represent connections between them.

The dataset is available online as [81].

- **E. coli [49]:** This is the probabilistic functional gene network of E. coli. Nodes represent genes and edges represent interactions between them. The dataset is available online at `http://www.inetbio.org/ecolinet/downloadnetwork.php` (The integrated network).

- **Yeast [50]:** This network represents the protein-protein interaction (PPI) network in the budding yeast. Nodes and edges represent proteins and the interactions among them, respectively. The dataset is available online at `http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm`.

- **Stelzl [51]:** This is a protein-protein interaction network in humans. Nodes and edges represent proteins and the interactions among them, respectively. The dataset is available online as [82].

- **Figeys [52]:** Similar to above, this is a protein-protein interaction network in humans where nodes and edges represent proteins and the interactions among them, respectively. The dataset is available online as [83].

- **Vidal [53]:** Similar to above, this is a protein-protein interaction network in humans where nodes and edges represent proteins and the interactions among them, respectively. The dataset is available online as [84].

- **westernUS [44]:** This dataset describes the high voltage power grid in the Western States of the US. Nodes represent transformers, substations, and generators, and the edges represent high-voltage transmission lines. The dataset is available online at `https://toreopsahl.com/datasets/#uspowergrid`.

- **Florida [54]:** This network describes the food web in the cypress wetlands of South Florida during the wet season. Nodes represent taxa and an an edge denotes that a taxon uses another taxon as food. The dataset is available online as [85].

- **LRL [55]:** The networks describes the food web of Little Rock Lake, Wisconsin, USA. Nodes represent autotrophs, herbivores, carnivores and decomposers while links represent food sources. The dataset is available online as [86].

- **Facebook group [56]:** This dataset describes the social interactions among a group of Facebook users. Nodes and edges represent profiles and the connections between them, respectively. The dataset is available online at `http://snap.stanford.edu/data/egonets-Facebook.html`.

- **E-main [57, 58]:** This datasets contains E-main communications in a research institution. Nodes represent institution members and edges exist between any ordered pair of members if one has sent at least one E-main to the other. The dataset is available online at `http://snap.stanford.edu/data/email-Eu-core.html`.

- **Southern Women [59]:** This is a social network of 18 Southern women. Nodes are individuals and edges represent mutual attendance at one of the 14 events recorded. The dataset is available online at `https://toreopsahl.com/datasets/#southernwomen`.

- **UCI P2P [60]:** This dataset describes an online community among the students of the University of California, Irvine. Nodes represent individuals and edges represent at least one message sent between any pair of them. The dataset is available online at `https://toreopsahl.com/datasets/#online_social_network`.

- **UCI Forum [61]:** This network is based on the same online community as in UCR P2P, but an edge exists between two individuals if they posted on the same topic in a forum. This dataset is also available online at `https://toreopsahl.com/datasets/#online_social_network`.

- **Freeman's EIES [62]:** This is a network of researchers working on social network analysis. Nodes represent researchers and edges represent personal relationships between them. The dataset is available online at `https://toreopsahl.com/datasets/#FreemansEIES` (the second dataset in the list).

- **Dolphins [63]:** This is a social network of bottlenose dolphins observed between 1994 and 2001. The nodes are the bottlenose dolphins and edges indicate a frequent association between them. The dataset is available online as [87].

- **Physicians [64]:** This network captures innovation spread among 246 physicians in four towns in Illinois, USA. A node represents a physician and an edge represents that one physician recognizes the other as theor friend or that they turn to them if they need advice or are interested in a discussion. The dataset is available online as [88].

- **Org. Consult Advice & Value [65]:** These are intra-organizational networks between employees of a consulting company. The nodes are individuals, and the edges represent frequency of information or advice requests (Org. Consult Advice) and the value placed on the information or advice received (Org. Consult Value). The datasets are available online at `https://toreopsahl.com/datasets/#Cross_Parker`.

- **Org. R&D Advice & Aware [65]:** Similar to the networks above, these describe intra-

organizational interactions among the members of a research team in a manufacturing company. Nodes represent individuals, and edges represent the extent to which individuals received advice from their peers to accomplish their work (Org. R&D Advice) and employees' awareness of each others' knowledge and skills (Org. R&D Aware). The datasets are available online at `https://toreopsahl.com/datasets/#Cross_Parker`.

Acknowledgements: This chapter is taken, in part, from the work which is to appear as "Heterogeneity of central nodes explains the benefits of time-varying control scheduling in complex dynamical networks" by E. Nozari, F. Pasqualetti, and J. Cortés in *Journal of Complex Networks*. The dissertation author was the primary investigator and author of this paper.

Chapter Bibliography

[1] R. E. Kalman, "Mathematical description of linear dynamical systems," *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, vol. 1, no. 2, pp. 152–192, 1963.

[2] Y. Y. Liu, J. J. Slotine, and A. L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167–173, 2011.

[3] N. J. Cowan, E. J. Chastain, D. A. Vilhena, J. S. Freudenberg, and C. T. Bergstrom, "Nodal dynamics, not degree distributions, determine the structural controllability of complex networks," *PLOS One*, vol. 7, no. 6, pp. 1–5, 06 2012.

[4] A. Olshevsky, "Minimal controllability problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 249–258, 2014.

[5] G. Yan, J. Ren, Y. Lai, C. Lai, and B. Li, "Controlling complex networks: How much energy is needed?" *Physical Review Letters*, vol. 108, no. 21, p. 218703, 2012.

[6] F. Pasqualetti, S. Zampieri, and F. Bullo, "Controllability metrics, limitations and algorithms for complex networks," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 40–52, 2014.

[7] T. H. Summers and J. Lygeros, "Optimal sensor and actuator placement in complex dynamical networks," in *IFAC World Congress*, Cape Town, South Africa, 2014, pp. 3784–3789.

[8] T. H. Summers, F. L. Cortesi, and J. Lygeros, "On submodularity and controllability in complex dynamical networks," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 91–101, 2016.

[9] V. Tzoumas, M. A. Rahimian, G. J. Pappas, and A. Jadbabaie, "Minimal actuator placement with bounds on control effort," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 67–78, 2016.

[10] S. Pequito, P. Bogdan, and G. J. Pappas, "Minimum number of probes for brain dynamics observability," in *IEEE Conf. on Decision and Control*. IEEE, 2015, pp. 306–311.

[11] M. A. Belabbas, "Geometric methods for optimal sensor design," *Proceedings of the Royal Society of London Series A*, vol. 472, p. 20150312, 2016.

[12] H. Zhang, R. Ayoub, and S. Sundaram, "Sensor selection for Kalman filtering of linear dynamical systems: Complexity, limitations and greedy algorithms," *Automatica*, vol. 78, pp. 202–210, 2017.

[13] V. Tzoumas, Y. Xue, S. Pequito, P. Bogdan, and G. J. Pappas, "Selecting sensors in biological fractional-order systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 709–721, June 2018.

[14] L. Zhao, W. Zhang, J. Hu, A. Abate, and C. J. Tomlin, "On the optimal solutions of the infinite-horizon linear sensor scheduling problem," *IEEE Transactions on Automatic Control*, vol. 59, no. 10, pp. 2825–2830, 2014.

[15] S. T. Jawaid and S. L. Smith, "Submodularity and greedy algorithms in sensor scheduling for linear dynamical systems," *Automatica*, vol. 61, pp. 282–288, 2015.

[16] Y. Zhao, F. Pasqualetti, and J. Cortés, "Scheduling of control nodes for improved network controllability," in *IEEE Conf. on Decision and Control*, Las Vegas, NV, 2016, pp. 1859–1864.

[17] D. Han, J. Wu, H. Zhang, and L. Shi, "Optimal sensor scheduling for multiple linear dynamical systems," *Automatica*, vol. 75, pp. 260–270, 2017.

[18] E. Nozari, F. Pasqualetti, and J. Cortés, "Time-invariant versus time-varying actuator scheduling in complex networks," in *American Control Conference*, Seattle, WA, May 2017, pp. 4995–5000.

[19] C. T. Chen, *Linear System Theory and Design*, 3rd ed. New York, NY, USA: Oxford University Press, Inc., 1998.

[20] H. Balakrishnan, "Control and optimization algorithms for air transportation systems," *Annual Reviews in Control*, vol. 41, pp. 39–46, 2016.

[21] B. Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 485–497, 2010.

[22] W. Zhang, M. Kamgarpour, D. Sun, and C. J. Tomlin, "A hierarchical flight planning framework for air traffic management," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 179–194, 2012.

[23] D. Teodorovic and K. Vukadinovic, *Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach*, ser. International Series in Intelligent Technologies. Springer, 2012.

[24] G. Albi, L. Pareschi, and M. Zanella, "Boltzmann-type control of opinion consensus through leaders," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 372, no. 2028, 2014.

[25] C. Qian, J. Cao, J. Lu, and J. Kurths, "Adaptive bridge control strategy for opinion evolution on social networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 21, no. 2, p. 025116, 2011.

[26] R. J. Bursik, "The informal control of crime through neighborhood networks," *Sociological Focus*, vol. 32, no. 1, pp. 85–97, 1999.

[27] A. V. Proskurnikov, A. S. Matveev, and M. Cao, "Opinion dynamics in social networks with hostile camps: Consensus vs. polarization," *IEEE Transactions on Automatic Control*, vol. 61, no. 6, pp. 1524–1536, June 2016.

[28] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, p. 888, 2010.

[29] C. Nowzari, V. M. Preciado, and G. J. Pappas, "Analysis and control of epidemics: A survey of spreading processes on complex networks," *IEEE Control Systems*, vol. 36, no. 1, pp. 26–46, Feb 2016.

[30] M. Salathe and J. H. Jones, "Dynamics and control of diseases in networks with community structure," *PLOS Computational Biology*, vol. 6, no. 4, 04 2010.

[31] L. B. Shaw and I. B. Schwartz, "Enhanced vaccine control of epidemics in adaptive networks," *Physical Review E*, vol. 81, p. 046120, 2010.

[32] L. Hufnagel, D. Brockmann, and T. Geisel, "Forecast and control of epidemics in a globalized world," *Proceedings of the National Academy of Sciences*, vol. 101, no. 42, pp. 15 124–15 129, 2004.

[33] R. Huerta and L. S. Tsimring, "Contact tracing and epidemics control in social networks," *Physical Review E*, vol. 66, p. 056115, 2002.

[34] Y. Chen, G. Paul, S. Havlin, F. Liljeros, and H. E. Stanley, "Finding a better immunization strategy," *Physical Review Letters*, vol. 101, p. 058701, 2008.

[35] C. D. Godsil and G. F. Royle, *Algebraic Graph Theory*, ser. Graduate Texts in Mathematics. Springer, 2001, vol. 207.

[36] E. R. van Daam, "Graphs with few eigenvalues: An interplay between combinatorics and algebra," Ph.D. dissertation, Tilburg University, Oct. 1996.

[37] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[38] J. Rebesco, I. Stevenson, K. Koerding, S. Solla, and L. Miller, "Rewiring neural interactions by micro-stimulation," *Frontiers in Systems Neuroscience*, vol. 4, p. 39, 2010.

[39] D. J. Guggenmos, M. Azin, S. Barbay, J. D. Mahnken, C. Dunham, P. Mohseni, and R. J. Nudo, "Restoration of function after brain damage using a neural prosthesis," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 21 177–21 182, 2013.

[40] T. K. Lu, A. S. Khalil, and J. J. Collins, "Next-generation synthetic gene networks," *Nature biotechnology*, vol. 27, no. 12, p. 1139, 2009.

[41] D. D. Vecchio and R. M. Murray, *Biomolecular Feedback Systems*. Princeton University Press, 2015.

[42] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[43] R. Bakker, T. Wachtler, and M. Diesmann, "CoCoMac 2.0 and the future of tract-tracing databases," *Frontiers in neuroinformatics*, vol. 6, 2012.

[44] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[45] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics*, vol. 3, pp. 276–282, 2007.

[46] T. Opsahl, "Why anchorage is not (that) important: Binary ties and sample selection," 2011, available at http://wp.me/poFcY-Vw.

[47] R. W. Eash, K. S. Chon, Y. J. Lee, and D. E. Boyce, "Equilibrium traffic assignment on an aggregated highway network for sketch planning," *Transportation Research Record*, vol. 994, pp. 30–37, 1983.

[48] D. E. Boyce, K. S. Chon, M. E. Ferris, Y. J. Lee, K.-T. Lin, and R. W. Eash, "Implementation and evaluation of combined models of urban travel and location on a sketch planning network," *Chicago Area Transportation Study*, pp. xii+169, 1985.

[49] H. Kim, J. E. Shim, J. Shin, and I. Lee, "Ecolinet: a database of cofunctional gene network for Escherichia coli," *Database*, vol. 2015, 2015.

[50] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, "Topological structure analysis of the protein–protein interaction network in budding yeast," *Nucleic acids research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[51] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toks?z, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker, "A human protein-protein interaction network: A resource for annotating the proteome," *Cell*, vol. 122, pp. 957–968, 2005.

[52] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J. P. P. Lambert, H. S. Duewel, I. I. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S. L. L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou, and D. Figeys, "Large-scale mapping of human protein-protein interactions by mass spectrometry," *Molecular Systems Biology*, vol. 3, 2007.

[53] J. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, and N. Ayivi-Guedehoussou, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, no. 7062, pp. 1173–1178, 2005.

[54] R. E. Ulanowicz, J. J. Heymans, and M. S. Egnotovich, "Network analysis of trophic dynamics in South Florida ecosystems, FY 99: The graminoid ecosystem," *Annual Report to the United States Geological Service Biological Resources Division Ref. No.[UMCES] CBL 00-0176, Chesapeake Biological Laboratory, University of Maryland*, 2000.

[55] N. D. Martinez, J. J. Magnuson, T. Kratz, and M. Sierszen, "Artifacts or attributes? effects of resolution on the Little Rock Lake food web," *Ecological Monographs*, vol. 61, pp. 367–392, 1991.

[56] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 539–547.

[57] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 555–564.

[58] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.

[59] A. Davis, B. B. Gardner, and M. R. Gardner, *Deep South*. Chicago, IL: University of Chicago Press, 1941.

[60] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, 2009.

[61] T. Opsahl, "Triadic closure in two-mode networks: Redefining the global and local clustering coefficients," *Social Networks*, vol. 35, no. 2, pp. 159–167, 2013.

[62] S. C. Freeman and L. C. Freeman, *The Networkers Network: A Study of the Impact of a New Communications Medium on Sociometric Structure*, ser. Social sciences research reports. School of Social Sciences University of Calif., 1979.

[63] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, pp. 396–405, 2003.

[64] J. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, pp. 253–270, 1957.

[65] R. L. Cross and A. Parker, *The Hidden Power of Social Networks: Understanding how Work Really Gets Done in Organizations*. Harvard Business School Press, 2004.

[66] P. C. Müller and H. I. Weber, "Analysis and optimization of certain qualities of controllability and observability for linear dynamical systems," *Automatica*, vol. 8, no. 3, pp. 237–246, 1972.

[67] S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, A. B. Yu, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton, and D. S. Bassett, "Controllability of structural brain networks," *Nature Communications*, vol. 6, pp. 8414 EP–, 10 2015.

[68] Z. Bai and G. H. Golub, "Bounds for the trace of the inverse and the determinant of symmetric positive definite matrices," *Annals of Numerical Mathematics*, vol. 4, pp. 29–38, 1996.

[69] R. P. Barry and R. K. Pace, "Monte Carlo estimates of the log determinant of large sparse matrices," *Linear Algebra and its Applications*, vol. 289, no. 1-3, pp. 41–54, 1999.

[70] A. Reusken, "Approximation of the determinant of large sparse symmetric positive definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 3, pp. 799–818, 2002.

[71] I. C. F. Ipsen and D. J. Lee, "Determinant approximations," *arXiv preprint arXiv:1105.0437*, 2011.

[72] I. Han, D. Malioutov, and J. Shin, "Large-scale log-determinant computation through stochastic chebyshev expansions," in *International Conference on Machine Learning*, 2015, pp. 908–917.

[73] C. Boutsidis, P. Drineas, P. Kambadur, E. Kontopoulou, and A. Zouzias, "A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix," *Linear Algebra and its Applications*, vol. 533, pp. 95–117, 2017.

[74] J. Fitzsimons, K. Cutajar, M. Osborne, S. Roberts, and M. Filippone, "Bayesian inference of log determinants," *arXiv preprint arXiv:1704.01445*, 2017.

[75] F. Bullo, J. Cortés, and S. Martinez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, electronically available at http://coordinationbook.info.

[76] H. W. Gould, "Combinatorial identities: Table I: Intermediate techniques for summing finite series," 2010, from the seven unpublished manuscripts of H. W. Gould. Edited and compiled by J. Quaintance.

[77] D. S. Bernstein, *Matrix Mathematics*, 2nd ed. Princeton University Press, 2009.

[78] T. Tossavainen, "On the zeros of finite sums of exponential functions," *Gazette of the Australian Mathematical Society*, vol. 33, no. 1, pp. 47–50, 2006.

[79] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[80] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis-Kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015.

[81] "Chicago network dataset – KONECT," Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/tntp-ChicagoRegional

[82] “Human protein (stelzl) network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-Stelzl

[83] “Human protein (figeys) network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-figeys

[84] “Human protein (vidal) network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-vidal

[85] “Florida ecosystem wet network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/foodweb-baywet

[86] “Little rock lake network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/maayan-foodweb

[87] “Dolphins network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/dolphins

[88] “Physicians network dataset – KONECT,” Apr. 2017. [Online]. Available: http://konect.uni-koblenz.de/networks/moreno_innovation

Chapter 7

Network Identification with Latent Nodes

In this chapter, we continue our treatment of networked dynamical systems under resource constraints and focus on the problem of network system identification. This problem has numerous applications in science and engineering. In neuroscience, for instance, researchers seek to understand how different regions of the brain cooperate with each other by having subjects perform certain goal-directed tasks while measuring their brain activity via multi-channel recordings such as electroencephalograms (EEG) [1–5]. In systems biology, genetic network identification uses data from RNA micro-array experiments to identify the interaction pattern between genes in a regulatory network [6, 7]. Similar examples exist in other areas including finance, social networks, and physics.

Roughly speaking, the objective in network identification is to determine causal relationships among the nodes in the network that model the direction and strength of the interactions between them. While network control and coordination has made much progress on problems where the interaction topology is either given or the design objective itself, not so much attention has been devoted to develop techniques to address the identification of unknown topologies from measured

data. In many applications of complex network systems, only a *manifest* subset of the nodes can be directly actuated and measured while the state of the remaining *latent* nodes and their number are unknown. Our goal is to identify the transfer function of the manifest subnetwork and determine whether interactions between manifest nodes are direct or mediated by latent nodes.

We show that, if there are no inputs to the latent nodes, the manifest transfer function can be approximated arbitrarily well in the H_∞-norm sense by the transfer function of an auto-regressive model and present a least-squares estimation method to construct the auto-regressive model from measured data. We show that the least-squares auto-regressive method guarantees an arbitrarily small H_∞-norm error in the approximation of the manifest transfer function, exponentially decaying once the model order exceeds a certain threshold. Finally, we show that when the latent subnetwork is acyclic, the proposed method achieves perfect identification of the manifest transfer function above a specific model order as the length of the data increases. We end the chapter with various examples that illustrate our results.

7.1 Prior Work

An increasing number of works study topology identification problems to better understand the interactions in large-scale networks and their role in determining the network behavior. A complex network is commonly represented as a directed graph, and the interactions among neighboring nodes are represented by directed edges whose weights reflect the interaction strength. In this sense, topology identification aims at identifying the adjacency matrix of the network graph [8] or its Boolean structure [9]. The work [10] studies the complete characterization of the interaction topology of consensus-type networks using a series of node-knockout experiments, where nodes

are sequentially forced to broadcast a zero state without being removed from the network. The work [11] also uses node-knockout experiments to identify the topology of directed linear time-invariant networks relying on the cross-power spectral densities of the network response to wide-sense stationary noise. The work [12] presents a method to infer the topology of a network of coupled phase oscillators from its stable response dynamics, assuming that one can manipulate every individual node and perform large number of experiments. In general, without such assumption, it is difficult or impossible, depending on the additional structural information available, to accurately identify the topology of a general network. As a result, a main focus has been on particular network realizations that explain the measured data, such as the sparsest realization, sometimes with a design parameter to manage the trade-off between model accuracy and sparsity, see e.g., [7, 13]. Along these lines, the work [14] considers the identification of networked linear systems with tree topologies.

The above-referenced works rely on knowledge of the number of nodes in the network. However, it is often impossible to sample the state of all nodes, or even know the existence of some of them. The work [15] studies the problem of learning latent tree graphical models where samples are available only from a subset of the nodes, and proposes computationally efficient algorithms for learning trees without any redundant hidden nodes. The work [16] proposes a method to identify the latent nodes and consistently reconstruct the topology under the assumptions that the network is a polytree and the degree of each latent node is at least three, with out-degree of at least two. Unlike the topology identification algorithms proposed in [14, 16], our approach here allows for the possibility of cycles in the network topology. Using the notion of the dynamical structure function of a network with latent nodes [17], the work [18] proposes a convex optimization-based approach to find the best Boolean structure among manifest nodes which consists of computing and comparing

the distance between an estimated transfer matrix or data to all possible Boolean structures. The problem of minimal state-space realization of a given dynamical structure function was further studied in [19]. In the present work, however, we use a least-square autoregressive identification approach to identify not only whether a pair of manifest nodes are dynamically connected, but also whether this connection is direct or indirect (latent-mediated) and, in the latter case, the length of the shortest path between the two.

Recent work has employed sparse plus low-rank (S+L) decomposition to identify general graphical models (with the possibility of cycles) with latent variables for static [20] and dynamic [21] models. The present work has two main differences with respect to this paper. First, the S+L decomposition assumes that the subnetwork among manifest nodes is sparse and the number of latent nodes is (considerably) smaller than the number of manifest ones, while our method is applicable to arbitrary networks. Second, although the identification procedure of [21] also leads to an auto-regressive (AR) model, it is based on the so-called maximum-entropy covariance extension. This method, with origins in seismic vibrations and human voice analysis, seeks to *maximize* the prediction error [22] (while our approach seeks to *minimize* it), leading to very different models.

Finally, our work is inspired by the wide use in neuroscience of AR models to analyze brain data via Granger causality and its variants and the study of effective connectivity among different areas of the brain, see e.g., [2,3,23]. The Granger causality measure is a mainly descriptive tool that captures influence and interconnection among time series. A popular variant of Granger causality, direct directed transfer function (dDTF) [5, 24] distinguishes between direct and indirect interconnections between two nodes by multiplying the directed transfer function (DTF, the normalized transfer function between the two nodes) by the partial coherence between them in the frequency domain. We are motivated here by understanding to what extent the reconstruction results ob-

tained via methods that build on Granger causality are sensitive to the presence of latent nodes. Furthermore, we propose a method using (multivariate) AR models for networks with latent nodes that distinguishes between direct and indirect (i.e., latent-mediated) interconnections between two nodes in the time domain based on the order of the interconnection between them.

7.2 Problem Statement

We consider a discrete-time, linear time-invariant (LTI) network dynamics with state-space representation

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{u}(k), \\ \mathbf{y}(k) &= \mathbf{C}\mathbf{x}(k), \end{aligned} \tag{7.1}$$

where $k \in \mathbb{Z}_{\geq 0}$ is the time index, $\mathbf{x}(k) \in \mathbb{R}^n$ is the network state (with $x_i(k)$ representing the state of node $i \in \{1, \dots, n\}$), $\mathbf{u}(k) \in \mathbb{R}^n$ is the control input (with $u_i(k)$ acting on node i), and $\mathbf{y}(k) \in \mathbb{R}^m$ is the network output. Here, $\mathbf{A} \in \mathbb{R}^{n \times n}$ is the adjacency matrix of the network, characterizing the interactions among neighboring nodes, and $\mathbf{C} \in \mathbb{R}^{m \times n}$ is the output matrix. Since natural system are usually driven by noise, the input, state, and output sequences are in general stochastic processes over the sample space of noise realizations. For simplicity, the dynamical description (7.1) assumes that all nodes are of order 1, that is, $\mathbf{x}(k+1)$ depends directly only on $\mathbf{x}(k)$ and is conditionally independent of $\mathring{\mathbf{x}}_{k-1}$ given $\mathbf{x}(k)$. Nevertheless, as we discuss later (see e.g., Remark 7.3.4), all of the subsequent results are generalizable to systems whose dynamics (in the original "physical" variables) are described by difference equations of order higher than 1.

Even though there is a control input at every node in the network dynamics (7.1), we do not assume that all the control inputs are user-specified. In fact, in a large-scale network, it is common that one can actuate only a small subset of the nodes due to computational constraints, physical limitations, or cost. A similar observation can be made regarding the number of nodes whose state can be directly measured. For these reasons, here we assume that the nodes of the network are divided into $n_m \leq n$ *manifest* nodes, which can be directly actuated and measured by the user, and $n - n_m$ *latent* nodes, which can neither be directly actuated nor measured by the user. With this distinction, and using a permutation of the indices in $(1, 2, \dots, n)$ if necessary, we can decompose the network and input state as $\mathbf{x} = [\mathbf{x}_m^T, \mathbf{x}_l^T]^T$ and $\mathbf{u} = [\mathbf{u}_m^T, \mathbf{u}_l^T]^T$, respectively, where the subindex 'm' corresponds to manifest nodes and the subindex 'l' corresponds to latent nodes. With this convention, the output matrix takes the form $\mathbf{C} = [\mathbf{I}_{n_m}, \mathbf{0}_{n_m \times (n-n_m)}]$. With the decomposition of the nodes into manifest and latent, the state-space representation (7.1) becomes

$$\begin{bmatrix} \mathbf{x}_m(k+1) \\ \mathbf{x}_l(k+1) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_m(k) \\ \mathbf{x}_l(k) \end{bmatrix} + \begin{bmatrix} \mathbf{u}_m(k) \\ \mathbf{u}_l(k) \end{bmatrix},$$

$$\mathbf{y}(k) = \mathbf{x}_m(k). \tag{7.2}$$

In the remainder of this chapter, we consider the network in the relabeled form (7.2). Fig. 7.1 illustrates this relabeling procedure (corresponding to a linear transformation) in a ring.

Since the focus of this work is on network identification and not stabilization, we make the following standard assumption.

Assumption 7.2.1. The adjacency matrix of the complete network as well as the latent subnetwork are Schur stable, i.e., $\rho(\mathbf{A}) < 1$ and $\rho(\mathbf{A}_{22}) < 1$. □

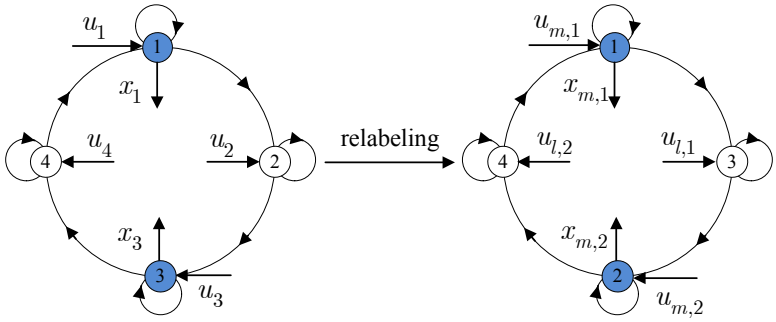


Figure 7.1: Node relabeling in a directed ring with 4 nodes. Nodes 1 and 3 are manifest, nodes 2 and 4 are latent. The permutation $(1, 2, 3, 4) \rightarrow (1, 3, 2, 4)$ makes manifest and latent nodes have consecutive indices, as in (7.2).

Remark 7.2.2. ***(Direct versus latent interactions).*** The interaction graph of the manifest subnetwork is characterized by $\mathbf{A}_{11}$. In particular, the state of node p affects the state of node q *directly* if and only if the entry on the q-th row and the p-th column, denoted by $\mathbf{A}_{11}(q, p)$, is nonzero. However, even if $\mathbf{A}_{11}(q, p) = 0$, it is still possible that node p affects node q *indirectly* through some latent nodes. The distinction between direct and indirect connections is an important point to which we come back later in our discussion. □

We refer to a latent node as *passive* if its corresponding input is zero. Throughout the chapter, we only deal with passive latent nodes, so that $\mathring{\mathbf{u}}_l \equiv 0$. We make the following assumption on the input to the manifest nodes.

Assumption 7.2.3. The input $\mathring{\mathbf{u}}_m$ to the manifest subnetwork is a zero-mean stochastic process with independent and identically distributed (i.i.d.) absolutely continuous[1] random vectors $\mathbf{u}_m(k)$, with covariance $\mathbf{I}_{n_m}$. □

Assumption 7.2.3 guarantees that $\mathring{\mathbf{u}}_m$ is persistently exciting of arbitrary order and its power spectral density does not vanish at any frequency. Similar assumptions are common in system iden-

[1]Recall that an absolutely continuous random variable/vector is one that has a probability density function (e.g., Gaussian).

tification, see e.g., [11, 25]. The zero-mean assumption can be relaxed by assuming a nonzero but known $\mathbb{E}[\mathbf{u}_m(k)]$ corresponding to the scenario where the designer injects a deterministic stimulating signal into every manifest node, which itself is subject to the disturbance of a zero-mean white noise. Without loss of generality and for simplicity, we assume $\mathbb{E}[\mathbf{u}_m(k)] \equiv \mathbf{0}_{n_m}$.

Given the setup above, our objective is to identify the transfer function $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}(\omega)$ of the manifest subnetwork, that is, the transfer function from $\mathbf{u}_m$ to $\mathbf{x}_m$, absent any knowledge of the latent nodes.

Problem 5. ***(Identification of the manifest transfer function).*** *Given the measured data $\mathring{\mathbf{y}}_N$, find a linear auto-regressive model of order τ, with $N \gg \tau$, of the form*

$$\tilde{\mathbf{x}}_m(k+1) = \sum_{i=0}^{\tau-1} \tilde{\mathbf{A}}_i \tilde{\mathbf{x}}_m(k-i) + \mathbf{u}_m(k), \tag{7.3}$$

such that the associated transfer function $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}$ from $\mathbf{u}_m$ to $\tilde{\mathbf{x}}_m$ and the transfer function $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ from $\mathbf{u}_m$ to $\mathbf{x}_m$ in (7.1) *are close in the H_∞-norm, i.e., $\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m} - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty$ is small.* □

There are alternative methods to identify the transfer function matrix $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ besides the AR method in (7.3). Our adoption here of AR model candidates is motivated by their widespread use in neuroscience to determine causality and interconnections in human brain connectivity models, see e.g., [2–4][2]. Equipped with time series data obtained during the performance of a cognitive task, the conventional procedure consists of first estimating an AR model, then computing its associated transfer function matrix, and finally evaluating the Granger causality connectivity measure, or generalizations of it, in the frequency domain. We are particularly motivated by the prospect of

[2]In general, the main advantage of AR models over more general models such as ARMA or BJ is their simplicity, only capturing the internal dynamics and assuming negligible input *noise correlation* (though putting no restriction on input *signal correlation*, which is significant in brain dynamics). As a result, prediction error minimization has a closed-form solution for an AR model while it is non-convex in the ARMA or BJ cases.

understanding the sensitivity of these approaches to the presence of latent nodes corresponding to brain regions that are active during the cognitive task but are not directly measured.

7.3 Asymptotically Exact Identification of the Manifest Transfer Function

In this section we establish that, given an arbitrary precision, there exists an AR model solving Problem 5. More precisely, we show that there exists a sequence of AR models of the form (7.3) with increasing order whose transfer functions converge to $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ exponentially in the H_∞ sense. We later show that, if the latent subnetwork is acyclic, then this approximation can be made exact.

We start our discussion with a useful auxiliary result.

Lemma 7.3.1. ***(Upper bound on*** $\|\mathbf{A}_{22}^i\|$***).*** *For any Schur stable* $\mathbf{A}_{22} \in \mathbb{R}^{n_l \times n_l}$ *and any* $\bar{\rho} \in (\rho(\mathbf{A}_{22}), 1)$*, there exists* $\kappa \in \mathbb{R}_{>0}$ *such that* $\|\mathbf{A}_{22}^i\| \leq \kappa \cdot \bar{\rho}^i$*, for all* $i \in \mathbb{Z}_{\geq 0}$*.*

Proof. The result is an immediate consequence of the spectral radius formula $\lim_{i\to\infty} \|\mathbf{A}_{22}^i\|^{1/i} = \rho(\mathbf{A}_{22})$. □

We are now ready to state the main result of this section.

Theorem 7.3.2. ***(AR model whose transfer function converges to the manifest transfer function).*** *Consider the LTI network described by* (7.2) *where all the latent nodes are passive. For any* $\bar{\rho} \in (\rho(\mathbf{A}_{22}), 1)$*, there exists* $\bar{\gamma} \in \mathbb{R}_{>0}$ *such that for all* $\tau \in \mathbb{Z}_{\geq 0}$*, the AR model* (7.3) *with*

$$\tilde{\mathbf{A}}_0^* = \mathbf{A}_{11}, \quad \tilde{\mathbf{A}}_i^* = \mathbf{A}_{12}\mathbf{A}_{22}^{i-1}\mathbf{A}_{21},\ i \in \{1, \dots, \tau - 1\}, \tag{7.4}$$

guarantees

$$\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau) - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty \leq \bar{\gamma}\cdot\bar{\rho}^\tau. \tag{7.5}$$

Proof. We obtain from (7.2) that

$$\begin{aligned}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}(\omega) &= (z\mathbf{I}_{n_m} - \mathbf{A}_{11} - \mathbf{A}_{12}(z\mathbf{I}_{n_l} - \mathbf{A}_{22})^{-1}\mathbf{A}_{21})^{-1} \\ &\overset{(a)}{=} (z\mathbf{I}_{n_m} - \mathbf{A}_{11} - \sum_{i=1}^{\infty} z^{-i}\mathbf{A}_{12}\mathbf{A}_{22}^{i-1}\mathbf{A}_{21})^{-1},\end{aligned} \tag{7.6}$$

where $z = e^{j\omega}$ and (a) follows by using the relation $(z\mathbf{I}_{n_l} - \mathbf{A}_{22})^{-1} = \sum_{i=1}^{\infty} z^{-i}\mathbf{A}_{22}^{i-1}$. Similarly, from (7.3) we obtain

$$\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\omega,\tau) = (z\mathbf{I}_{n_m} - \sum_{i=0}^{\tau-1} z^{-i}\tilde{\mathbf{A}}_i^*)^{-1}. \tag{7.7}$$

Here we write the transfer function as $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\omega,\tau)$ to emphasize its dependence on τ. It then follows directly that

$$\begin{aligned}\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau) - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty &= \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}(\mathbf{T}^{-1}_{\mathbf{x}_m\mathbf{u}_m} - \mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau))\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty \\ &\overset{(a)}{\leq} \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty\|\mathbf{T}^{-1}_{\mathbf{x}_m\mathbf{u}_m} - \mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty \\ &\overset{(b)}{\leq} \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty \sum_{i=\tau}^{\infty}\|z^{-i}\mathbf{A}_{12}\mathbf{A}_{22}^{i-1}\mathbf{A}_{21}\|_\infty \\ &\overset{(c)}{\leq} \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty\|\mathbf{A}_{12}\|\|\mathbf{A}_{21}\|\sum_{i=\tau}^{\infty}\|\mathbf{A}_{22}^{i-1}\| \\ &\overset{(d)}{\leq} \gamma(\tau)\cdot\bar{\rho}^\tau,\end{aligned}$$

where

$$\gamma(\tau) \triangleq \frac{\kappa \|\mathbf{T}_{\mathbf{x}_m \mathbf{u}_m}\|_\infty \|\mathbf{A}_{12}\| \|\mathbf{A}_{21}\|}{\bar{\rho} - \bar{\rho}^2} \|\mathbf{T}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\cdot, \tau)\|_\infty. \tag{7.8}$$

Here, (a) follows from the sub-multiplicativity of induced norms, (b) follows by the sub-additivity of norms, (c) follows by the definition of the H_∞-norm and also the sub-multiplicativity of induced norms, and (d) follows from Lemma 7.3.1. The remainder of the proof is devoted to showing the existence of a uniform upper bound $\bar{\gamma}$ for $\gamma(\tau)$. By the definition of the H_∞-norm,

$$\begin{aligned} \|\mathbf{T}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\cdot, \tau)\|_\infty &= \sup_{-\pi \le \omega \le \pi} \sigma_{\max}\big(\mathbf{T}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\omega, \tau)\big) \\ &\overset{(a)}{=} \Big(\inf_{-\pi \le \omega \le \pi} \sigma_{\min}\big(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\omega, \tau)\big) \Big)^{-1}, \end{aligned} \tag{7.9}$$

where (a) holds due to the fact that $\sigma_{\max}(\mathbf{M}) = \sigma^{-1}_{\min}(\mathbf{M}^{-1})$ for any invertible matrix $\mathbf{M}$. To complete the proof, we only need to show that

$$\vartheta \triangleq \inf_{\tau \in \mathbb{Z}_{\ge 0}} \inf_{-\pi \le \omega \le \pi} \sigma_{\min}\big(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\omega, \tau)\big) > 0. \tag{7.10}$$

We show this in two steps.

(i) It follows from (7.6) and (7.7) that

$$\lim_{\tau \to \infty} \mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\omega, \tau) = \mathbf{T}^{-1}_{\mathbf{x}_m \mathbf{u}_m}(\omega), \quad \forall \omega \in [-\pi, \pi].$$

It is straightforward to show, using the exponential decay of $\mathbf{A}^\tau_{22}$ and definition of uniform

convergence, that each entry of $\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)$ converges uniformly to the corresponding entry of $\mathbf{T}^{-1}_{\mathbf{x}_m\mathbf{u}_m}$. Hence, given the uniform continuity of matrix eigenvalues as a function of matrix entries [26, Thm 7.8c], $\sigma_{\min}\big(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\big)$ converges uniformly to $\sigma_{\min}\big(\mathbf{T}^{-1}_{\mathbf{x}_m\mathbf{u}_m}\big)$. Thus, since $\inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\mathbf{T}^{-1}_{\mathbf{x}_m\mathbf{u}_m}(\omega)\big) = \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty > 0$ (which itself holds by Assumption 7.2.1), there exists $\tau_0 \in \mathbb{Z}_{\ge 0}$ such that

$$\inf_{\tau\ge\tau_0}\inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\omega,\tau)\big) > 0.$$

(ii) For any finite τ, we show that $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)$ is BIBO stable and thus has no poles on the unit circle (which in turn guarantees $\inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\omega,\tau)\big) > 0$). For any bounded input $\mathbf{u}_m$, let the corresponding outputs of $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)$ and $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ be denoted by $\tilde{\mathbf{x}}_m$ and $\mathbf{x}_m$, resp. (with initial states set to zero). We then have

$$\mathbf{x}_m(k) - \tilde{\mathbf{x}}_m(k) = \mathbf{A}_{12}\mathbf{A}_{22}^{\tau-1}\mathbf{x}_l(k-\tau),$$

where $\mathbf{x}_l$ is the (internal) state of the latent nodes in $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$. By Assumption 7.2.1, both $\mathbf{x}_m(k)$ and $\mathbf{A}_{12}\mathbf{A}_{22}^{\tau-1}\mathbf{x}_l(k-\tau)$ are bounded, proving the BIBO stability of $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)$.

Hence, (7.10) follows by combining (i) and (ii) and the fact that the decomposition $\mathbb{Z}_{\ge 0} = \{0\}\cup\{1\}\cup\cdots\cup\{\tau_0-1\}\cup\{\tau_0,\tau_0+1,\dots\}$ is finite. Equivalently, there exists $U > 0$ such that $\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty < U$ for all $\tau \in \mathbb{Z}_{\ge 0}$, so (7.5) holds with $\bar{\gamma} = \kappa U\|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty\|\mathbf{A}_{12}\|\|\mathbf{A}_{21}\|/(\bar{\rho}-\bar{\rho}^2)$. □

Theorem 7.3.2 shows that the presence of latent nodes in the network, as long as they do not receive any external input, does not affect the achievable accuracy of the identification via AR modeling of the manifest transfer function.

Remark 7.3.3. ***(Direct versus latent interactions – cont'd).*** It follows from the network dynamics (7.2) that

$$\mathbf{x}_m(k+1) = \sum_{i=0}^{k} \tilde{\mathbf{A}}_i^* \mathbf{x}_m(k-i) + \mathbf{A}_{12}\mathbf{A}_{22}^k \mathbf{x}_l(0) + \mathbf{u}_m(k). \tag{7.11}$$

By virtue of (7.11), we can distinguish whether two manifest nodes interact directly or indirectly through latent nodes by looking at the matrix sequence $\{\tilde{\mathbf{A}}_i^*\}$. First, the state of manifest node p affects the state of manifest node q directly if and only if $\tilde{\mathbf{A}}_0^*(q,p) = \mathbf{A}_{11}(q,p) \neq 0$. Similarly, the state of manifest node p affects the state of manifest node q indirectly through latent nodes if and only if $\tilde{\mathbf{A}}_i^*(q,p) \neq 0$ for some $i \geq 1$. In particular, from the relation $\tilde{\mathbf{A}}_i^* = -\mathbf{A}_{12}\mathbf{A}_{22}^{i-1}\mathbf{A}_{21}$, one can see that the state of p first affects some latent nodes (that correspond to the nonzero entries in the p-th column of $\mathbf{A}_{21}$) through $\mathbf{A}_{21}$, then propagates through the latent subnetwork, reflected by $\mathbf{A}_{22}^{i-1}$, and finally affects q through $\mathbf{A}_{12}$. Furthermore, if the latent subnetwork is acyclic, then $\tilde{\mathbf{A}}_i^*(q,p) \neq 0$ implies that there are exactly i latent nodes in a path connecting p to q. □

Remark 7.3.4. ***(Systems described by higher-order difference equations).*** Unlike the system description in (7.1), the dynamic behavior of many real-world complex systems such as the brain cortical networks is described by difference equations of orders significantly greater than 1, i.e.,

$$\mathbf{x}(k+1) = \mathbf{A}^{(0)}\mathbf{x}(k) + \mathbf{A}^{(1)}\mathbf{x}(k-1) + \cdots + \mathbf{A}^{(\nu-1)}\mathbf{x}(k-\nu+1) + \mathbf{u}(k),\ \nu \gg 1 \tag{7.12}$$

where $x_1, \ldots, x_{n_m}$ still denote the manifest (sensed and actuated) nodes and $x_{n_m+1}, \ldots, x_n$ are the latent ones. In this description, the vector $\mathbf{x}$ corresponds to some relevant physical variables. Defining the state vector $\boldsymbol{\xi}(k) = [\mathbf{x}(k)^T\ \mathbf{x}(k-1)^T\ \cdots\ \mathbf{x}(k-\nu+1)^T]^T$, one can rewrite (7.12) in order-1 form

as

$$\begin{bmatrix} \boldsymbol{\xi}_m(k+1) \\ \boldsymbol{\xi}_l(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mathcal{A}}_{11} & \boldsymbol{\mathcal{A}}_{12} \\ \boldsymbol{\mathcal{A}}_{21} & \boldsymbol{\mathcal{A}}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_m(k) \\ \boldsymbol{\xi}_l(k) \end{bmatrix} + \begin{bmatrix} \mathbf{u}_m(k) \\ \mathbf{0} \end{bmatrix}, \tag{7.13}$$

where $\boldsymbol{\xi}_m(k) = \mathbf{x}_m(k)$, $\boldsymbol{\xi}_l(k) = [\mathbf{x}_l(k)^T\ \mathbf{x}_m(k-1)^T\ \mathbf{x}_l(k-1)^T\ \cdots\ \mathbf{x}_m(k-\nu+1)^T\ \mathbf{x}_l(k-\nu+1)^T]^T$, $\boldsymbol{\mathcal{A}}_{11} = \mathbf{A}_{11}^{(0)}$, and

$$\boldsymbol{\mathcal{A}}_{12} = \begin{bmatrix} \mathbf{A}_{12}^{(0)} & \mathbf{A}_{11}^{(1)} & \mathbf{A}_{12}^{(1)} & \cdots & \mathbf{A}_{11}^{(\tau-1)} & \mathbf{A}_{12}^{(\tau-1)} \end{bmatrix},$$

$$\boldsymbol{\mathcal{A}}_{21} = \begin{bmatrix} (\mathbf{A}_{21}^{(0)})^T & \mathbf{I}_{n_m} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}^T,$$

$$\boldsymbol{\mathcal{A}}_{22} = \begin{bmatrix} \mathbf{A}_{22}^{(0)} & \mathbf{A}_{21}^{(1)} & \mathbf{A}_{22}^{(1)} & \cdots & \mathbf{A}_{21}^{(\tau-2)} & \mathbf{A}_{22}^{(\tau-2)} & \mathbf{A}_{21}^{(\tau-1)} & \mathbf{A}_{22}^{(\tau-1)} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{n_l} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{n_m} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_{n_l} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

In this description, we view $\boldsymbol{\xi}_m$ as the actual "manifest state" of the system while the whole vector $\boldsymbol{\xi}_l$ is the "latent state". The reason for this interpretation is that, at any time k, only $\mathbf{x}_m(k)$ is directly sensed/actuated while $\mathbf{x}(k-1), \ldots, \mathbf{x}(k-\nu+1)$ are quantities stored in the system. Interestingly, for the order-1 description (7.1), this observation brings up the possibility of some of the latent variables $\mathbf{x}_l$ simply being a relayed version of manifest variables. Note that, under this interpretation,

the matrices $\mathbf{A}_{11}^{(1)}, \ldots, \mathbf{A}_{11}^{(\nu-1)}$ represent manifest-latent (rather than manifest-manifest) interactions. From (7.13), it is clear that all the treatment for (7.1) is readily applicable. Nevertheless, as ν increases, larger τ is necessary in order for (7.3) to represent the system accurately. This is both intuitive and clear from (7.5) and (7.8), where increasing ν results in larger $\|\mathbf{A}_{12}\|$ and $\|\mathbf{A}_{21}\|$ as well as (usually) $\|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|$ and $\rho(\mathbf{A}_{22})$. This, in turn, may result in numerical difficulties when one constructs the AR model from recorded input-output data (which is the subject of the next section).

□

Next, we show that there exists an AR model (7.3) whose transfer function coincides with the manifest transfer function if the latent subnetwork is acyclic.

Corollary 7.3.5. ***(Exact manifest transfer function identification for acyclic latent subnetworks).*** *Under the assumptions of Theorem 7.3.2, further assume that the latent subnetwork is acyclic, i.e., there exists* $\tau_{22} \in \mathbb{Z}_{\geq 1}$ *such that* $\mathbf{A}_{22}^{\tau_{22}} = \mathbf{0}_{n_l \times n_l}$. *Then, the matrix sequence* $\tilde{\mathbf{A}}_0^*, \cdots, \tilde{A}_{\tau_{22}}^*$ *in* (7.4) *ensures* $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m} = \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$.

The proof of the result follows by comparing (7.6) and (7.7), and using the assumption that the latent subnetwork is acyclic. Theorem 7.3.2 and Corollary 7.3.5 show that it is possible to identify the transfer function of the manifest subnetwork without any knowledge of the passive latent nodes. However, (7.4) cannot be directly applied to determine the auto-regressive model because its evaluation requires knowledge of the adjacency matrix A of the whole network, which is unknown. This problem can be circumvented by employing the measured data sequence $\mathring{\mathbf{y}}_N \subset \mathbb{R}^{n_m}$, as explained in the next section.

7.4 Identification via Least-Squares Estimation

In this section we employ least-squares estimation to compute from data the sequence of matrices defining the auto-regressive model. We show that the estimates resulting from this method asymptotically converge in probability, as the data length N and model order τ increase, to the optimal matrix sequence identified in Theorem 7.3.2. Finally, we particularize our discussion to the case of acyclic latent subnetworks.

7.4.1 Least-Squares Auto-Regressive Estimation

Given a vector sequence $\mathring{\mathbf{y}}_N \subset \mathbb{R}^{n_m}$, the problem of least-squares auto-regressive (LSAR) model estimation with order $\tau \in \mathbb{Z}_{\geq 1}$ is to find a matrix sequence $\mathring{\hat{\mathbf{A}}}_{\tau-1} = \{\hat{\mathbf{A}}_0, \ldots, \hat{\mathbf{A}}_{\tau-1}\} \subset \mathbb{R}^{n_m \times n_m}$ that minimizes the 2-norm of the residual sequence $\mathring{\mathbf{e}}_\tau^{N-1} \subset \mathbb{R}^{n_m}$ defined by

$$\mathbf{e}(k) = \mathbf{y}(k+1) - \sum_{i=0}^{\tau-1} \hat{\mathbf{A}}_i \mathbf{y}(k-i), \tag{7.14}$$

for $k \in \{\tau, \ldots, N-1\}$. Equation (7.14) can be written in compact vector form as

$$\mathring{\mathbf{y}}_{\tau+1}^{N} = \mathring{\hat{\mathbf{A}}}_{\tau-1} \mathbf{\Phi}_N + \mathring{\mathbf{e}}_\tau^{N-1}, \tag{7.15}$$

where

$$\mathring{\mathbf{y}}_{\tau+1}^{N} = \begin{bmatrix} \mathbf{y}(\tau+1) & \mathbf{y}(\tau+2) & \cdots & \mathbf{y}(N) \end{bmatrix} \in \mathbb{R}^{n_m \times (N-\tau)},$$
$$\mathring{\mathbf{e}}_{\tau}^{N-1} = \begin{bmatrix} \mathbf{e}(\tau) & \mathbf{e}(\tau+1) & \cdots & \mathbf{e}(N-1) \end{bmatrix} \in \mathbb{R}^{n_m \times (N-\tau)},$$
$$\mathring{\hat{\mathbf{A}}}_{\tau-1} = \begin{bmatrix} \hat{\mathbf{A}}_0 & \hat{\mathbf{A}}_1 & \cdots & \hat{\mathbf{A}}_{\tau-1} \end{bmatrix} \in \mathbb{R}^{n_m \times n_m \tau},$$
$$\mathbf{\Phi}_N = \begin{bmatrix} \mathbf{y}(\tau) & \mathbf{y}(\tau+1) & \cdots & \mathbf{y}(N-1) \\ \mathbf{y}(\tau-1) & \mathbf{y}(\tau) & \cdots & \mathbf{y}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}(1) & \mathbf{y}(2) & \cdots & \mathbf{y}(N-\tau) \end{bmatrix}.$$

Using the square of the prediction error [25]

$$\operatorname{tr}(\mathring{\mathbf{e}}_{\tau}^{N-1}(\mathring{\mathbf{e}}_{\tau}^{N-1})^T) = \operatorname{tr}\left((\mathring{\mathbf{y}}_{\tau+1}^{N} - \mathring{\hat{\mathbf{A}}}_{\tau-1}\mathbf{\Phi}_N)(\mathring{\mathbf{y}}_{\tau+1}^{N} - \mathring{\hat{\mathbf{A}}}_{\tau-1}\mathbf{\Phi}_N)^T\right)$$

as the cost function, we compute its gradient

$$\frac{\partial \operatorname{tr}(\mathring{\mathbf{e}}_{\tau}^{N-1}(\mathring{\mathbf{e}}_{\tau}^{N-1})^T)}{\partial \mathring{\hat{\mathbf{A}}}_{\tau-1}} = (\mathring{\mathbf{y}}_{\tau+1}^{N} - \mathring{\hat{\mathbf{A}}}_{\tau-1}\mathbf{\Phi}_N)(-\mathbf{\Phi}_N^T) = \mathring{\hat{\mathbf{A}}}_{\tau-1}\mathbf{\Phi}_N\mathbf{\Phi}_N^T - \mathring{\mathbf{y}}_{\tau+1}^{N}\mathbf{\Phi}_N^T.$$

Setting this to zero, we get a system of linear equations for which a solution is guaranteed to exist (since the rows of $\mathring{\mathbf{y}}_{\tau+1}^{N}\mathbf{\Phi}_N^T$ belong to the row space of $\mathbf{\Phi}_N\mathbf{\Phi}_N^T$, which is the same as the row space of $\mathbf{\Phi}_N^T$). By Assumption 7.2.3, $\det(\mathbf{\Phi}_N\mathbf{\Phi}_N^T) \neq 0$ and this solution is unique with probability one.[3]

[3]This is because (each element of) $\mathring{\mathbf{y}}_{N-1}$ is an affine function of $\mathring{\mathbf{u}}_{N-2}$, and $\det(\mathbf{\Phi}_N\mathbf{\Phi}_N^T)$ is a polynomial function of $\mathring{\mathbf{y}}_{N-1}$, so $\det(\mathbf{\Phi}_N\mathbf{\Phi}_N^T)$ is a polynomial function of $\mathring{\mathbf{u}}_{N-2}$. Therefore, the level set $\mathcal{N} = \{\mathring{\mathbf{u}}_{N-2} \mid \det(\mathbf{\Phi}_N\mathbf{\Phi}_N^T) = 0\}$ has Lebesgue measure zero. Thus, by Assumption 7.2.3, $\Pr(\mathcal{N}) = 0$.

If $\det(\mathbf{\Phi}_N \mathbf{\Phi}_N^T) = 0$, the minimum-norm solution can be found as

$$\mathring{\hat{\mathbf{A}}}_{\tau-1} = \mathring{\mathbf{y}}_{\tau+1}^N \mathbf{\Phi}_N^T (\mathbf{\Phi}_N \mathbf{\Phi}_N^T)^{-1} = \mathring{\mathbf{y}}_{\tau+1}^N \mathbf{\Phi}_N^+, \tag{7.16}$$

where $(\cdot)^+$ denotes the Moore-Penrose pseudo-inverse. Since (7.16) is also valid for the nonsingular case, it is taken as the solution to the LSAR estimation problem. In order to indicate the dependency of the solution upon the measured data sequence, we sometimes use the notation $\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N)$.

7.4.2 Convergence in Probability to Manifest Transfer Function

Here we study the transfer function resulting from the LSAR estimation method and characterize its convergence properties, as the data length and the model order increase, with respect to the transfer function of the manifest subnetwork. Our first result establishes that the LSAR matrix estimate (7.16) converges in probability to the optimal matrix sequence identified in Theorem 7.3.2.

Proposition 7.4.1. ***(The LSAR estimate converges in probability to optimal matrix sequence).*** *Consider the LTI network described by* (7.2) *where all latent nodes are passive. Given the measured data sequence* $\mathring{\mathbf{y}}_N$ *generated from the dynamics* (7.2) *stimulated by the white noise input* $\mathring{\mathbf{u}}_m$ *according to Assumption 7.2.3 and any* $\bar{\rho} \in (\rho(\mathbf{A}_{22}), 1)$*, there exists* $\beta \in \mathbb{R}_{>0}$ *(depending only on the adjacency matrix A) such that the LSAR estimate* $\hat{\mathbf{A}}_\tau(\mathring{\mathbf{y}}_N)$ *in* (7.16) *satisfies*

$$\|\operatorname*{plim}_{N\to\infty} \mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) - \mathring{\tilde{\mathbf{A}}}^*_{\tau-1}\|_{\max} \leq \beta \tau \bar{\rho}^\tau, \tag{7.17}$$

where $\mathring{\tilde{\mathbf{A}}}^*_{\tau-1} = \begin{bmatrix} \tilde{\mathbf{A}}^*_0 & \tilde{\mathbf{A}}^*_1 & \dots & \tilde{\mathbf{A}}^*_{\tau-1} \end{bmatrix} \in \mathbb{R}^{n_m \times n_m \tau}$ *is the optimal matrix sequence given by* (7.4).

Proof. For any quasi-stationary signal[4] $\mathring{\mathbf{s}}$, let

$$\mathbf{R}_{\mathbf{s}}(j) \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[\mathbf{s}(i)\mathbf{s}(i-j)^T].$$

Using the Birkhoff's Ergodic Theorem [27, Thm 7.2.1] (see also [27, Thm 7.1.3]) and the fact that $\mathring{\mathbf{y}}$ is the output of a stable system (and thus the effects of initial conditions asymptotically vanish), we can show that

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}(i)\mathbf{y}(i-j)^T = \mathbf{R}_{\mathbf{y}}(j).$$

As a result, $\frac{1}{N}\boldsymbol{\Phi}_N\boldsymbol{\Phi}_N^T \in \mathbb{R}^{n_m\tau\times n_m\tau}$ also converges in probability and

$$\mathbf{R}_{\boldsymbol{\Phi}} \triangleq \operatorname*{plim}_{N\to\infty} \frac{1}{N}\boldsymbol{\Phi}_N\boldsymbol{\Phi}_N^T = \begin{bmatrix} \mathbf{R}_{\mathbf{y}}(0) & \mathbf{R}_{\mathbf{y}}(1) & \cdots & \mathbf{R}_{\mathbf{y}}(\tau-1) \\ \mathbf{R}_{\mathbf{y}}^T(1) & \mathbf{R}_{\mathbf{y}}(0) & \cdots & \mathbf{R}_{\mathbf{y}}(\tau-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{\mathbf{y}}^T(\tau-1) & \mathbf{R}_{\mathbf{y}}^T(\tau-2) & \cdots & \mathbf{R}_{\mathbf{y}}(0) \end{bmatrix}.$$

Define

$$\boldsymbol{\nu}(k) \triangleq \mathbf{y}(k+1) - \sum_{i=0}^{\tau-1} \tilde{\mathbf{A}}_i^* \mathbf{y}(k-i), \tag{7.18}$$

[4] Basically, a signal is quasi-stationary if it has a well-defined covariance function. See [25, Def 2.1] for a formal definition.

and note that the transfer function from $\mathbf{u}_m$ to $\boldsymbol{\nu}$ is $\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$, where $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ and $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}$ are given by (7.6) and (7.7), respectively. Equation (7.18) can be written in compact vector form as

$$\mathring{\mathbf{y}}^N_{\tau+1} = \mathring{\hat{\mathbf{A}}}^*_{\tau-1}\boldsymbol{\Phi}_N + \mathring{\boldsymbol{\nu}}^{N-1}_{\tau}, \tag{7.19}$$

with $\mathring{\boldsymbol{\nu}}^{N-1}_{\tau} \triangleq [\boldsymbol{\nu}(\tau)\ \ \boldsymbol{\nu}(\tau+1)\ \ \cdots\ \ \boldsymbol{\nu}(N-1)] \in \mathbb{R}^{n_m\times(N-\tau)}$. From (7.16) and (7.19), it follows that

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) &= \operatorname*{plim}_{N\to\infty}\frac{1}{N}\mathring{\mathbf{y}}^N_{\tau+1}\boldsymbol{\Phi}^T_N(\frac{1}{N}\boldsymbol{\Phi}_N\boldsymbol{\Phi}^T_N)^{-1}\\ &= \mathring{\tilde{\mathbf{A}}}^*_{\tau-1} + \operatorname*{plim}_{N\to\infty}\frac{1}{N}\mathring{\boldsymbol{\nu}}^{N-1}_{\tau}\boldsymbol{\Phi}^T_N(\frac{1}{N}\boldsymbol{\Phi}_N\boldsymbol{\Phi}^T_N)^{-1}.\end{aligned} \tag{7.20}$$

Moreover, Assumption 7.2.3 renders $\mathbf{u}_m(k)$ independent of $\mathring{\mathbf{y}}_k$, which further implies that $\operatorname{plim}_{N\to\infty}\frac{1}{N}\mathring{\mathbf{u}}^{N-1}_{m,\tau}\boldsymbol{\Phi}^T_N = \mathbf{0}_{n_m\times n_m\tau}$, where $\mathring{\mathbf{u}}^{N-1}_{m,\tau} \triangleq [\mathbf{u}_m(\tau)\ \mathbf{u}_m(\tau+1)\ \cdots\ \mathbf{u}_m(N-1) \in \mathbb{R}^{n_m\times(N-\tau)}$. Therefore,

$$\operatorname*{plim}_{N\to\infty}\frac{1}{N}\mathring{\boldsymbol{\nu}}^{N-1}_{\tau}\boldsymbol{\Phi}^T_N = \operatorname*{plim}_{N\to\infty}\frac{1}{N}(\mathring{\boldsymbol{\nu}}^{N-1}_{\tau} - \mathring{\mathbf{u}}^{N-1}_{m,\tau})\boldsymbol{\Phi}^T_N = \boldsymbol{\Psi}, \tag{7.21}$$

where $\boldsymbol{\Psi} \triangleq \begin{bmatrix}\boldsymbol{\Psi}_1 & \boldsymbol{\Psi}_2 & \ldots & \boldsymbol{\Psi}_\tau\end{bmatrix} \in \mathbb{R}^{n_m\times n_m\tau}$, with

$$\boldsymbol{\Psi}_j \triangleq \operatorname*{plim}_{N\to\infty}\frac{1}{N}\sum_{i=\tau}^{N-1}(\boldsymbol{\nu}(i) - \mathbf{u}_m(i))\mathbf{y}^T(i-j+1) \in \mathbb{R}^{n_m\times n_m}.$$

Thus, using $\operatorname{plim}_{N\to\infty}(\frac{1}{N}\boldsymbol{\Phi}_N\boldsymbol{\Phi}^T_N)^{-1} = \mathbf{R}^{-1}_{\boldsymbol{\Phi}}$, we have

$$\operatorname*{plim}_{N\to\infty}\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) - \mathring{\tilde{\mathbf{A}}}^*_{\tau-1} = \boldsymbol{\Psi}\mathbf{R}^{-1}_{\boldsymbol{\Phi}}.$$

By the sub-additivity of the max norm, it holds for any $j \in \{1, \dots, \tau\}$ that

$$\begin{aligned}
\|\mathbf{\Psi}_j\|_{\max} &\leq \operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum_{i=\tau}^{N-1} \|(\boldsymbol{\nu}(i) - \mathbf{u}_m(i))\mathbf{y}^T(i-j+1)\|_{\max} \\
&\overset{(a)}{\leq} \operatorname*{plim}_{N\to\infty} \frac{\bar{\rho}^{-\tau}}{N} \sum_{i=\tau}^{N-1} (\boldsymbol{\nu}(i)-\mathbf{u}_m(i))^T(\boldsymbol{\nu}(i)-\mathbf{u}_m(i)) + \operatorname*{plim}_{N\to\infty} \frac{\bar{\rho}^{\tau}}{N} \sum_{i=\tau}^{N-1} \mathbf{y}^T(i-j+1)\mathbf{y}(i-j+1) \\
&= \bar{\rho}^{-\tau}\mathrm{tr}(\mathbf{R}_{\boldsymbol{\nu}-\mathbf{u}_m}(0)) + \bar{\rho}^{\tau}\mathrm{tr}(\mathbf{R}_{\mathbf{y}}(0)),
\end{aligned} \tag{7.22}$$

where (a) follows from Lemma 7.A.1 in the appendix with the positive scalar M chosen as $\bar{\rho}^{\tau}$. Using the fact that the transfer function from $\mathbf{u}_m$ to $\boldsymbol{\nu} - \mathbf{u}_m$ is $\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m} - \mathbf{I}_{n_m}$, we obtain

$$\begin{aligned}
\mathbf{R}_{\boldsymbol{\nu}-\mathbf{u}_m}(0) &\triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mathbb{E}[(\boldsymbol{\nu} - \mathbf{u}_m)(i)(\boldsymbol{\nu} - \mathbf{u}_m)^T(i)] \\
&\overset{(a)}{=} \frac{1}{2\pi} \int_{-\pi}^{\pi} (\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}(\omega) - \mathbf{I}_{n_m})(\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}(\omega) - \mathbf{I}_{n_m})^* d\omega \\
&\overset{(b)}{\leq} \|\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m} - \mathbf{I}_{n_m}\|^2_{\infty}\mathbf{I}_{n_m} \overset{(c)}{\leq} \|\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m} - \mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\|^2_{\infty}\|\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\|^2_{\infty}\mathbf{I}_{n_m} \\
&\overset{(d)}{\leq} \hat{\gamma}\bar{\rho}^{2\tau}\mathbf{I}_{n_m},
\end{aligned} \tag{7.23}$$

where $\hat{\gamma} \triangleq \bar{\gamma}^2\big(1 + \|\mathbf{A}_{11}\| + \|\mathbf{A}_{12}\|\|\mathbf{A}_{21}\|\kappa(1-\bar{\rho})^{-1}\big)^2$ is constant, (a) follows from [28, eq. (9-193)], (b) follows by the definition of H_∞-norm, (c) follows by the sub-multiplicativity of induced norms, and (d) holds because of Theorem 7.3.2 and the observation that, by Lemma 7.3.1,

$$\|\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\|_{\infty} \leq 1 + \|\mathbf{A}_{11}\| + \|\mathbf{A}_{12}\|\|\mathbf{A}_{21}\|\kappa(1-\bar{\rho})^{-1}.$$

We obtain from (7.22) and (7.23),

$$\|\mathbf{\Psi}_j\|_{\max} \leq \bar{\rho}^{\tau}(\hat{\gamma} n_m + \mathrm{tr}(\mathbf{R}_{\mathbf{y}}(0))),$$

and from (7.20) and (7.21),

$$\begin{aligned} \|\operatorname*{plim}_{N\to\infty} \mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) - \mathring{\tilde{\mathbf{A}}}^*_{\tau-1}\|_{\max} &= \|\mathbf{\Psi}\mathbf{R}_{\mathbf{\Phi}}^{-1}\|_{\max} \leq n_m \tau \|\mathbf{R}_{\mathbf{\Phi}}^{-1}\|_{\max} \|\mathbf{\Psi}\|_{\max} \\ &= n_m \tau \|\mathbf{R}_{\mathbf{\Phi}}^{-1}\|_{\max} \max_j \|\mathbf{\Psi}_j\|_{\max} \leq \beta \tau \bar{\rho}^{\tau}, \end{aligned}$$

where $\beta = (\hat{\gamma} n_m^2 + \mathrm{tr}(\mathbf{R}_{\mathbf{y}}(0)) n_m) \|\mathbf{R}_{\mathbf{\Phi}}^{-1}\|_{\max}$, as claimed. □

When it is clear from context, we refer to $\operatorname{plim}_{N\to\infty} \hat{\mathbf{A}}_i(\mathring{\mathbf{y}}_N)$ simply as $\hat{\mathbf{A}}_i$.

Remark 7.4.2. ***(Direct versus latent interactions – cont'd).*** Proposition 7.4.1 shows that $\hat{\mathbf{A}}_i$ converges in probability to $\tilde{\mathbf{A}}_i^*$ exponentially as the model order τ increases. Therefore, within a margin of error that can be tuned as desired, we deduce from the discussion in Remark 7.3.3 that the LSAR estimate $\hat{\mathbf{A}}_0$ allows us to determine whether two manifest nodes interact directly and the LSAR estimates $\{\hat{\mathbf{A}}_i\}_{i\geq 1}$ allow us to determine whether two manifest nodes interact indirectly through latent nodes with high probability as the length of measurement data grows. □

Given the result in Proposition 7.4.1, we next turn our attention to the transfer function from $\mathbf{e}$ to $\mathbf{y}$ resulting from the LSAR estimation (7.14), which we denote by $\mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N, \tau)$. The next result shows that the H_∞-norm of this transfer function is uniformly upper bounded with respect to the model order τ.

Lemma 7.4.3. ***(H_∞-norm of $\mathbf{T}_{\mathbf{ye}}$ is uniformly upper bounded).*** *Under the assumptions of Propo-*

sition 7.4.1, there exist positive scalars τ_0 and $U^{\infty}_{\mathbf{T}_{\mathbf{ye}}}$ such that, for $\tau \geq \tau_0$,

$$\|\operatorname*{plim}_{N\to\infty} \mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N, \tau)\|_{\infty} \leq U^{\infty}_{\mathbf{T}_{\mathbf{ye}}}. \tag{7.24}$$

Proof. By definition of H_{∞}-norm, we have

$$\begin{aligned}\|\operatorname*{plim}_{N\to\infty} \mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N, \tau)\|_{\infty} &= \sup_{-\pi\leq\omega\leq\pi} \sigma_{\max}\big(\operatorname*{plim}_{N\to\infty} \mathbf{T}_{\mathbf{ye}}(\omega, \tau)\big) \\ &= \Big(\inf_{-\pi\leq\omega\leq\pi} \sigma_{\min}\big(\operatorname*{plim}_{N\to\infty} \mathbf{T}^{-1}_{\mathbf{ye}}(\omega, \tau)\big)\Big)^{-1}. \end{aligned} \tag{7.25}$$

Note that, for every $\omega \in [-\pi, \pi]$ and $\tau \in \mathbb{Z}_{\geq 0}$,

$$\operatorname*{plim}_{N\to\infty} \mathbf{T}^{-1}_{\mathbf{ye}}(\omega, \tau) = z\mathbf{I}_{n_m} - \sum_{i=0}^{\tau-1} z^{-i}\hat{\mathbf{A}}_i = \mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m \mathbf{u}_m}(\omega, \tau) - \sum_{i=0}^{\tau-1} z^{-i}(\hat{\mathbf{A}}_i - \tilde{\mathbf{A}}^*_i), \tag{7.26}$$

where $z = e^{j\omega}$. However, for every $\omega \in [-\pi, \pi]$ and $\tau \in \mathbb{Z}_{\geq 0}$,

$$\begin{aligned}\Big\| \sum_{i=0}^{\tau-1} z^{-i}(\hat{\mathbf{A}}_i - \tilde{\mathbf{A}}^*_i)\Big\| \leq \sum_{i=0}^{\tau-1} \|\hat{\mathbf{A}}_i - \tilde{\mathbf{A}}^*_i\| &\overset{(a)}{\leq} n_m \sum_{i=0}^{\tau-1} \|\hat{\mathbf{A}}_i - \tilde{\mathbf{A}}^*_i\|_{\max} \\ &\overset{(b)}{\leq} n_m \tau \max_i \|\hat{\mathbf{A}}_i - \tilde{\mathbf{A}}^*_i\|_{\max} \leq n_m \beta \tau^2 \bar{\rho}^{\tau},\end{aligned}$$

where (a) follows from the fact that $\|\mathbf{A}\| \leq n_m \|\mathbf{A}\|_{\max}$ for any matrix $\mathbf{A} \in \mathbb{R}^{n_m \times n_m}$ and (b) follows from Proposition 7.4.1. Therefore, using Weyl's theorem for the perturbation of singular values [29]

in (7.26) and taking $\inf_{-\pi\le\omega\le\pi}$ of both sides, we get

$$\begin{aligned}\inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\operatorname*{plim}_{N\to\infty}\mathbf{T}_{\mathbf{ye}}^{-1}(\omega,\tau)\big) &\ge \inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}^{-1}(\omega,\tau)\big) - \Big\|\sum_{i=0}^{\tau-1} z^{-i}(\hat{\mathbf{A}}_i-\tilde{\mathbf{A}}_i^*)\Big\| \\ &\ge \inf_{-\pi\le\omega\le\pi}\sigma_{\min}\big(\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}^{-1}(\omega,\tau)\big) - n_m\beta\tau^2\bar{\rho}^{\tau}.\end{aligned}$$

In view of (7.10), let τ_0 be such that

$$n_m\beta\tau^2\bar{\rho}^{\tau} \le \frac{\vartheta}{2}, \qquad \forall\tau\ge\tau_0. \tag{7.27}$$

Then, the result follows from (7.25) with $U_{\mathbf{T}_{\mathbf{ye}}}^{\infty} = \frac{2}{\vartheta}$. □

We are finally ready to show that the transfer function $\mathbf{T}_{\mathbf{ye}}$ obtained from the LSAR method converges in probability to the transfer function $\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$ of the manifest subnetwork.

Theorem 7.4.4. ***(The LSAR method consistently estimates the manifest transfer function).*** *Under the assumptions of Proposition 7.4.1, for any $\bar{\rho}\in(\rho(\mathbf{A}_{22}),1)$, there exist positive scalars $\bar{\beta}$, $\bar{\gamma}$ and τ_0 such that, for $\tau\ge\tau_0$,*

$$\|\operatorname*{plim}_{N\to\infty}\mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N,\tau) - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty \le (\bar{\beta}\tau^2+\bar{\gamma})\bar{\rho}^{\tau}. \tag{7.28}$$

Consequently, $\operatorname{plim}_{N\to\infty,\tau\to\infty}\mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N,\tau) = \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$.

Proof. We only need to prove (7.28) as it directly implies the last equation in the statement. By the

sub-additivity and sub-multiplicity of induced norms,

$$\begin{aligned}\|\mathbf{T}_{\mathbf{ye}}(\cdot,\tau)-\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty &\le \|\mathbf{T}_{\mathbf{ye}}(\cdot,\tau)-\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty+\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)-\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty\\ &\le \|\mathbf{T}_{\mathbf{ye}}(\cdot,\tau)\|_\infty\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)\|_\infty\|\mathbf{T}_{\mathbf{ye}}^{-1}(\cdot,\tau)-\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}^{-1}(\cdot,\tau)\|_\infty\\ &\quad+\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)-\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty. \end{aligned}\tag{7.29}$$

Next, by (7.9), Lemma 7.4.3, and Theorem 7.3.2, there exist positive scalars τ_0, $U_{\mathbf{T}_{\mathbf{ye}}}^{\infty}$ and ϑ such that for $\tau \ge \tau_0$,

$$\|\operatorname*{plim}_{N\to\infty}\mathbf{T}_{\mathbf{ye}}(\cdot,\tau)-\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty \le U_{\mathbf{T}_{\mathbf{ye}}}^{\infty}\vartheta^{-1}\|\operatorname*{plim}_{N\to\infty}\mathbf{T}_{\mathbf{ye}}^{-1}(\cdot,\tau)-\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}^{-1}(\cdot,\tau)\|_\infty+\bar{\gamma}\bar{\rho}^{\tau}.\tag{7.30}$$

Finally, according to the definition of $\mathbf{T}_{\mathbf{ye}}(\cdot,\tau)$ in (7.14) and $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\cdot,\tau)$ in (7.7), it follows that

$$\begin{aligned}\|\operatorname*{plim}_{N\to\infty}\mathbf{T}_{\mathbf{ye}}^{-1}(\cdot,\tau)-\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}^{-1}(\cdot,\tau)\|_\infty &= \|\sum_{i=0}^{\tau-1}z^{-i}(\operatorname*{plim}_{N\to\infty}\hat{\mathbf{A}}_i-\tilde{\mathbf{A}}_i^*)\|_\infty\\ &\overset{(a)}{\le}\sum_{i=0}^{\tau-1}\|\operatorname*{plim}_{N\to\infty}\hat{\mathbf{A}}_i-\tilde{\mathbf{A}}_i^*\|\overset{(b)}{\le} n_m\beta\tau^2\bar{\rho}^{\tau},\end{aligned}\tag{7.31}$$

where (a) holds by the sub-additivity and sub-multiplicity of $\|\cdot\|$ and (b) follows by Proposition 7.4.1 and the fact that $\|\mathbf{A}\| \le n_m\|\mathbf{A}\|_{\max}$ for any matrix $\mathbf{A}\in\mathbb{R}^{n_m\times n_m}$. Thus, we obtain (7.28) for $\tau \ge \tau_0$, where $\bar{\beta} \triangleq U_{\mathbf{T}_{\mathbf{ye}}}^{\infty}\vartheta^{-1}n_m\beta$ is a constant. □

According to Theorem 7.4.4, when the length N of the measurement data is sufficiently large and the model order τ exceeds a certain threshold, the error $\|\mathbf{T}_{\mathbf{ye}}(\tau)-\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty$ obtained by the LSAR method decreases exponentially with τ.

Remark 7.4.5. ***(Identification of manifest transfer function requires higher-order models as sta-***

bility margin of latent subnetwork decreases). Even though an explicit expression of the threshold τ_0 in Theorem 7.4.4 as a function of the network is difficult to obtain, we can still make some useful observations. From inequality (7.27) in the proof of Lemma 7.4.3, one can see that τ_0 is an increasing function of $\bar{\rho}$. Hence, as the latent subnetwork becomes less stable ($\rho(\mathbf{A}_{22})$ gets closer to 1), the corresponding τ_0 becomes larger, requiring the order of the AR model to be higher to ensure exponential convergence. □

Remark 7.4.6. ***(Systems described by higher-order difference equations – cont'd).*** As explained in Remark 7.3.4, the AR representation of systems with order $\nu > 1$ is identical to the $\nu = 1$ case, although they require larger AR order τ. For large-scale systems ($n \gg 1$), increasing τ rapidly raises the number of parameters in (7.15), which leads to over-parametrization of the LSAR identification. Our simulations in Section 7.5 show how this can be overcome both by increasing N (which is computationally costly) and exponential regularization. Also, note that when $\nu > 1$, the only member of the sequence of matrices $\mathbf{A}_{11}^{(0)}, \ldots, \mathbf{A}_{11}^{(\nu-1)}$ (denoting all current and past interactions among manifest *nodes*) that is identifiable by the LSAR method is $\mathbf{A}_{11}^{(0)}$ (representing direct interactions among manifest *states*) while the others are only identifiable in the aggregate form (7.5). □

7.4.3 Exact Identification for Acyclic Latent Subnetworks

Here we show that the transfer function of the manifest subnetwork can be perfectly identified using the LSAR method with a finite model order if the latent subnetwork is acyclic. We start by refining the result in Proposition 7.4.1 and showing how, in this case, the convergence of the LSAR matrix estimate (7.16) to the optimal matrix sequence identified in Theorem 7.3.2 holds in

the mean-square sense.

Proposition 7.4.7. ***(The LSAR estimate converges in mean square to optimal matrix sequence for acyclic latent subnetworks).*** *Consider the LTI network described by* (7.2) *where all latent nodes are passive. Further assume that the latent subnetwork is acyclic, i.e., there exists* $\tau_{22} \in \mathbb{Z}_{\geq 1}$ *such that* $\mathbf{A}_{22}^{\tau_{22}} = \mathbf{0}_{n_l \times n_l}$. *Given the measured data sequence* $\mathring{\mathbf{y}}_N$ *generated from the dynamics* (7.2) *stimulated by the white noise input* $\mathring{\mathbf{u}}_m$ *according to Assumption 7.2.3, the LSAR estimate* $\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N)$ *in* (7.16) *satisfies, for any* $\tau \geq \tau_{22} + 1$,

$$\lim_{N\to\infty} \mathbb{E}[(\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) - \mathring{\hat{\mathbf{A}}}^*_{\tau-1})^T(\mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) - \mathring{\hat{\mathbf{A}}}^*_{\tau-1})] = \mathbf{0}_{n_m\tau \times n_m\tau}.$$

Proof. If $\mathbf{A}_{22}$ is nilpotent, using Corollary 7.3.5, we deduce that the transfer function from $\mathbf{u}_m$ to $\boldsymbol{\nu}$ defined in (7.18) is $\mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}\mathbf{T}_{\mathbf{x}_m\mathbf{u}_m} = \mathbf{I}_{n_m}$. Consequently, the random vectors $\boldsymbol{\nu}(k)$'s are i.i.d. with zero mean and finite second moment $\mathbb{E}[\boldsymbol{\nu}(k)\boldsymbol{\nu}^T(k)] = \mathbf{I}_{n_m}$. Define

$$\mathbf{Z}_N \triangleq \frac{1}{N}(\mathring{\hat{\mathbf{A}}}_{\tau-1} - \mathring{\hat{\mathbf{A}}}^*_{\tau-1})\boldsymbol{\Phi}_N\boldsymbol{\Phi}_N^T \overset{(a)}{=} \frac{1}{N}(\mathring{\boldsymbol{\nu}}_\tau^{N-1} - \mathring{\mathbf{e}}_\tau^{N-1})\boldsymbol{\Phi}_N^T \overset{(b)}{=} \frac{1}{N}\mathring{\boldsymbol{\nu}}_\tau^{N-1}\boldsymbol{\Phi}_N^T,$$

where (a) follows from (7.15) and (7.19) and (b) follows from the fact that the least-squares estimate $\mathring{\hat{\mathbf{A}}}_{\tau-1}$ in (7.16) renders $\mathring{\mathbf{e}}_\tau^{N-1}\boldsymbol{\Phi}_N^T = \mathbf{0}_{n_m\times n_m\tau}$. Combining the fact that the $\boldsymbol{\nu}(k)$'s are i.i.d. and the fact that $\mathring{\mathbf{y}}_k$ is a function of $\mathring{\boldsymbol{\nu}}_{k-1}$, we deduce that $\boldsymbol{\nu}(k)$ are independent of $\mathring{\mathbf{y}}_k$. This further implies that $\mathbb{E}[\mathbf{Z}_N] = \mathbf{0}_{n_m\times n_m\tau}$. Furthermore,

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}[\mathbf{Z}_N^T\mathbf{Z}_N] &= \lim_{N\to\infty} \frac{1}{N^2}\mathbb{E}[\boldsymbol{\Phi}_N\mathring{(}\boldsymbol{\nu}_\tau^{N-1})^T\mathring{\boldsymbol{\nu}}_\tau^{N-1}\boldsymbol{\Phi}_N^T] \\ &= \lim_{N\to\infty} \frac{1}{N}\mathbf{R}_{\boldsymbol{\Phi}} = \mathbf{0}_{n_m\tau\times n_m\tau}.\end{aligned}$$

Therefore,

$$\lim_{N\to\infty} \mathbb{E}[\mathring{\hat{\mathbf{A}}}_{\tau-1} - \mathring{\tilde{\mathbf{A}}}^{*}_{\tau-1}] = \lim_{N\to\infty} \mathbb{E}[\mathbf{Z}_N]\mathbf{R}_{\mathbf{\Phi}}^{-1} = \mathbf{0}_{n_m \times n_m \tau},$$

and

$$\lim_{N\to\infty} \mathbb{E}[(\mathring{\hat{\mathbf{A}}}_{\tau-1} - \tilde{\mathbf{A}}^{*}_{\tau})^T(\mathring{\hat{\mathbf{A}}}_{\tau-1} - \mathring{\tilde{\mathbf{A}}}^{*}_{\tau-1})] = \mathbf{R}_{\mathbf{\Phi}}^{-1} \lim_{N\to\infty} \mathbb{E}[\mathbf{Z}_N^T\mathbf{Z}_N]\mathbf{R}_{\mathbf{\Phi}}^{-1} = \mathbf{0}_{n_m\tau \times n_m\tau},$$

as claimed. □

We build on this result to show that the manifest transfer function can be perfectly identified using the LSAR method with a finite model order if the latent subnetwork is acyclic.

Theorem 7.4.8. ***(Exact manifest transfer function identification for acyclic latent subnetworks).*** *Under the assumptions of Proposition 7.4.7, for any* $\tau \geq \tau_{22} + 1$*,*

$$\operatorname*{plim}_{N\to\infty} \mathbf{T}_{\mathbf{ye}}(\mathring{\mathbf{y}}_N, \tau) = \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}.$$

Proof. We have $\operatorname{plim}_{N\to\infty} \mathring{\hat{\mathbf{A}}}_{\tau-1}(\mathring{\mathbf{y}}_N) = \mathring{\tilde{\mathbf{A}}}^{*}_{\tau-1}$ from Proposition 7.4.7, which combined with (7.31) implies

$$\operatorname*{plim}_{N\to\infty} \mathbf{T}_{\mathbf{ye}}^{-1}(\tau) = \mathbf{T}^{-1}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\tau).$$

Moreover, from Corollary 7.3.5, we have $\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m}(\tau) = \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}$. The statement then follows from (7.29) and Lemma 7.4.3. □

7.5 Simulations

In this section, we illustrate the performance of least-squares auto-regressive estimation in identifying the manifest transfer function in two examples, a deterministic directed ring network and a group of Erdős–Rényi random networks. We pay particular attention to the behavior displayed as the length of measured data and the model order change. In both examples, the input signal is a white Gaussian process with unit variance.

Example 7.5.1. ***(Directed ring network).*** Consider a directed ring network of 40 nodes with self-loops and all edge weights equal to $\alpha = 0.25$. The nodes with indices $\{5, 23, 33, 34, 36\}$ are manifest and the remaining 35 nodes are passive latent. Fig. 7.2.(a) shows a 3D plot of the identification error $\|\mathbf{T}_{\mathbf{ye}} - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty$ of the LSAR method, with axes corresponding to length of measured data and model order, respectively. We note that, when the measured data length N is small, increasing the AR model order τ does not provide better estimation of the manifest transfer function. Similarly, when the model order τ is too low, increasing the data length N is not helpful either. Instead, when N and τ increase simultaneously, the LSAR method provides good estimation of the manifest transfer function without any knowledge of the latent nodes, as predicted by Theorem 7.4.4. In Fig. 7.2.(b), we fix $N = 10^6$ and show that the error of the model obtained by the LSAR method is quite similar to the error $\|\mathbf{T}_{\tilde{\mathbf{x}}_m\mathbf{u}_m} - \mathbf{T}_{\mathbf{x}_m\mathbf{u}_m}\|_\infty$ of the ideal AR model from Theorem 7.3.2. Note that the latter requires knowledge of the true adjacency matrix A, and we use it here merely for comparison purposes. □

Example 7.5.2. ***(Erdős–Rényi random network).*** Here we consider a group of 10 Erdős–Rényi random networks [30]. Each network in the group is of type $G(10, 0.35)$, with 5 manifest nodes chosen randomly and the remaining 5 nodes are latent. Each pair of edges $(i, j), (j, i)$, $1 \leq i <$

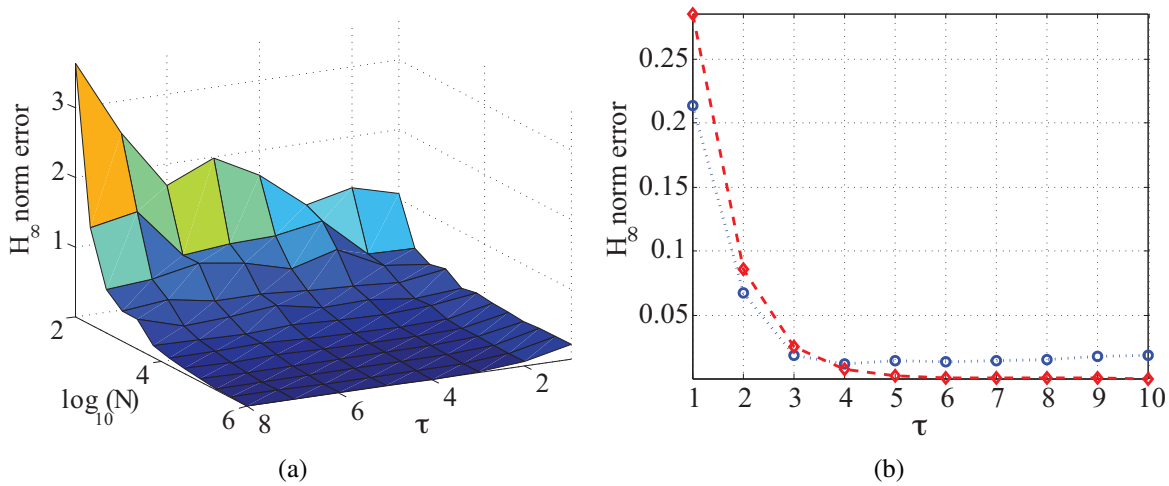


Figure 7.2: H_∞-norm errors for the directed ring network of Example 7.5.1. (a) The H_∞-norm error of the LSAR method as a function of data length N and model order τ. Performance improves as N and τ increase. (b) Comparison of the H_∞-norm errors of the LSAR method (blue dotted lines) and the optimal AR model from Theorem 7.3.2 (red dashed lines) for $N = 10^6$.

$j \leq 10$ has nonzero weights with probability 0.35 (we choose edges in pairs so that, when plotting the graph, the edge direction can be omitted). The weight of each edge has a uniform distribution in $\{x \in \mathbb{R} \mid 0.1 < x < 0.35\}$ (note that (i, j) and (j, i) can have different weights). Because of rounding errors in the numerical computation, the estimated coefficient matrices (7.16) of the AR model are usually full matrices. The lower bound on the edge weights allows us to discard entries in $\hat{\mathbf{A}}_0$ that are smaller than 0.1. We consider a fixed length $N = 10^6$ of measured data and analyze the effect of varying model order. Fig. 7.3 shows a 3D plot of the error in the identification of the manifest transfer function by the LSAR estimation, with axes corresponding to network index and model order, respectively. One can see the improvement in performance as the model order increases for all 10 networks. Fig. 7.4 compares the identification error of the LSAR method for the networks with indices $1, 6, 8, 10$ in Fig. 7.3 against the error of the optimal AR model from Theorem 7.3.2. The latent subnetwork of network 6 is acyclic (with $\mathbf{A}_{22} = \mathbf{0}_{5\times5}$), and the estimation error goes to 0 when the AR model has order higher than $\tau_{22} = 1$, as predicted by Theorem 7.4.8.

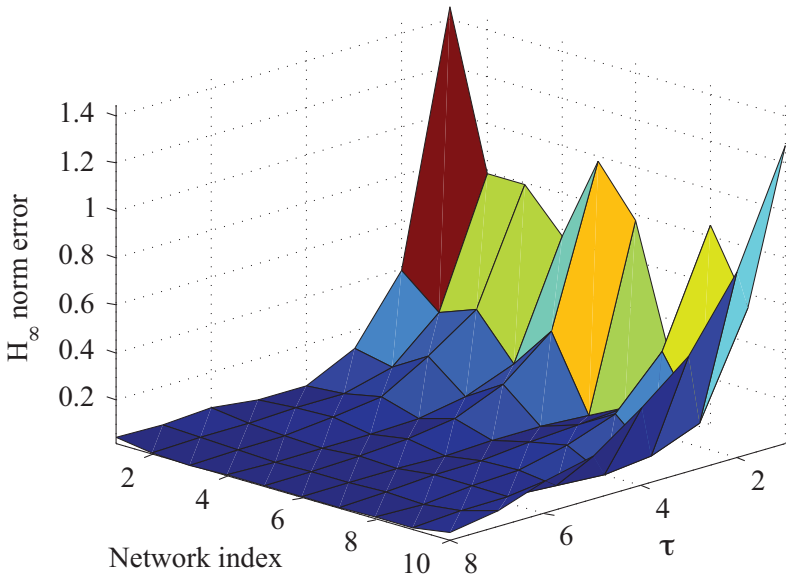


Figure 7.3: Illustration of the H_∞-norm error of the LSAR with respect to the model order τ for the group of $G(10, 0.35)$ Erdős–Rényi random networks of Example 7.5.2. Performance improves as the model order τ increases for all 10 networks. The length of measured data is $N = 10^6$.

To illustrate our observations in Remark 7.4.2 regarding the identification of manifest and latent interactions, Fig. 7.5 shows on the left the networks with indices $1, 6, 8, 10$ of Fig. 7.3 and on the right the corresponding reconstructions obtained with the LSAR method. The indirect interactions represented by dashed edges in these plots imply the presence of latent nodes. For comparison, we have also used the brain connectivity estimator technique called direct directed transfer function (dDTF) measure [5, 24] from neuroscience to identify direct connections between nodes. This technique is a refinement of the directed transfer function (DTF) approach, which instead cannot distinguish between direct and indirect connections. We have employed the dynamical modeling method within the Source Information Flow Toolbox (SIFT) [31, 32] in EEGLAB [33], which is a widely used open-source toolbox for EEG analysis. Fig. 7.6 shows the interaction topology among the 5 manifest nodes in network 10 identified by SIFT using the dDTF measure. The dDTF measure is in the frequency domain and can also be a function of time (e.g., for time-varying networks).

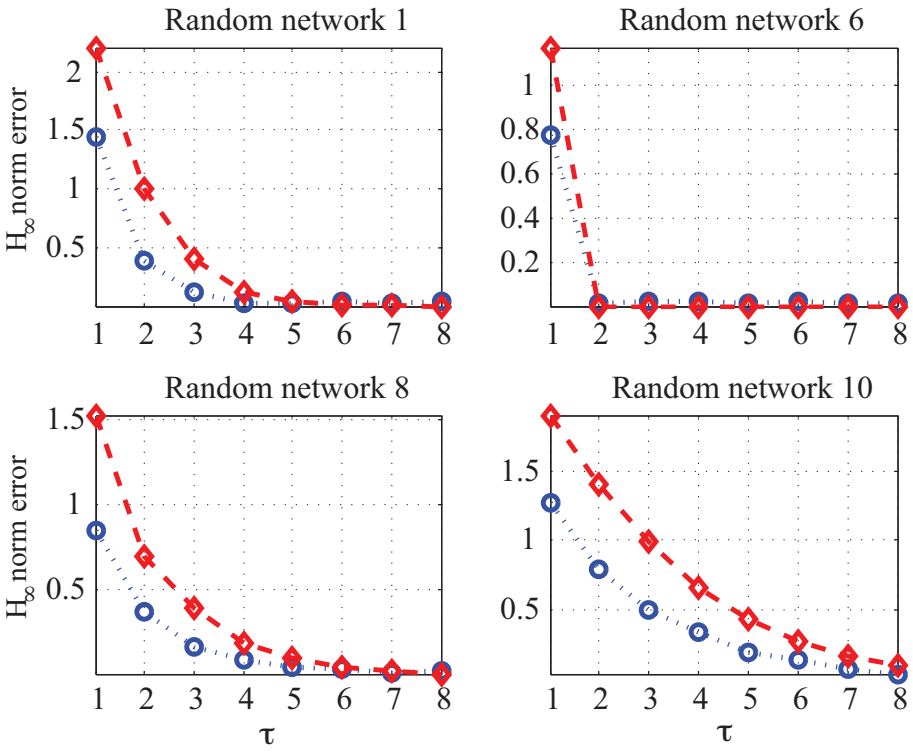


Figure 7.4: Comparison of the H_∞-norm errors of the LSAR method (red dashed lines) and the optimal AR model from Theorem 7.3.2 (blue dotted lines) for the Erdős–Rényi random networks with indices $1, 6, 8, 10$ in Fig. 7.3. The estimation error for network 6 becomes 0 when the AR model has order higher than 1 because the latent subnetwork is acyclic with $\tau_{22} = 1$. The length of measured data is $N = 10^6$.

Since our networks are time-invariant, the time axis can be ignored. The plot shows that the dDTF identifies roughly equally strong connections for $(2, 4)$ (which is in reality mediated by latent nodes) and $(4, 5)$ (which is a true direct connection). This is in contrast with the identification made with the LSAR method presented in Fig. 7.5(d). □

Example 7.5.3. ***(Cortical brain network identification from EEG data).*** In this example, we apply our method to a multi-channel electroencephalogram (EEG) time-series recorded from a human scalp during a selective visual attention experiment in order to identify the manifest and latent-mediated connections among the channels. The EEG data is taken from the sample dataset available in the EEGLAB MATLAB toolbox [33]. This dataset contains recordings from 32 channels

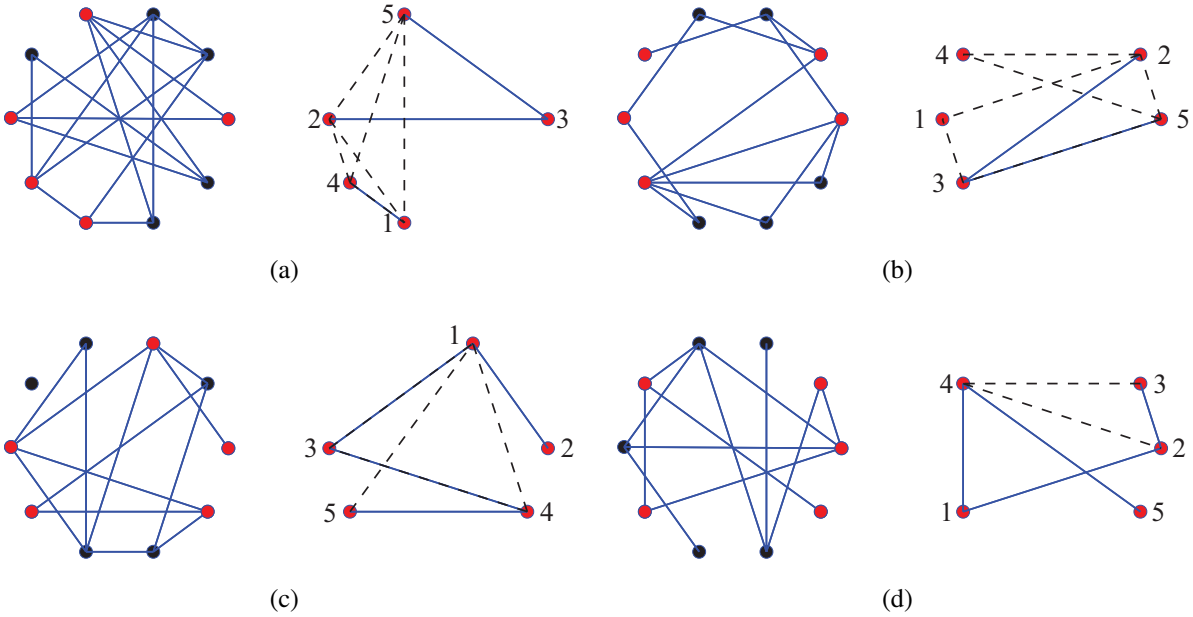


Figure 7.5: Left: Erdős–Rényi random networks corresponding to the networks with indices 1 (a), 6 (b), 8 (c), 10 (d) in Fig. 7.3, where red circles represent manifest nodes and black circles represent latent nodes. Right: reconstructed interaction graphs of the manifest subnetworks using the LSAR method. The numbers next to these nodes indicate their indices. A blue solid edge represents direct interaction and a black dashed edge represents indirect interaction through latent nodes. Note that the latent subnetwork of network 6 is acyclic.

for more than 3 seconds with $T_s = 7.8$ ms sampling time (128 Hz sampling frequency). Channel locations are shown in Fig. 7.10(a) on a top (axial) view of the skull. During the experiment, the subject is asked to perform specific motor actions in response to certain visual stimuli, requiring coordination among several cortices. We take the first 13 EEG channels corresponding to the fronto-temporal cortical areas (shown as blue squares in Fig. 7.10(a)) as the manifest nodes and the remaining channels as well as the truly hidden brain regions (the ones not probed in the test) as the latent nodes. In the following, we present the results of identifying the direct and indirect connections among the manifest nodes using the LSAR method as well as the dDTF algorithm [5, 24] and the S+L algorithm of [21]. For each method, we only keep the edges whose identified weights are above a certain threshold θ (which we choose as a proportional constant $\alpha \in (0, 1)$ times the largest

edge weight in the network).

In neuroscience, the brain dynamics generating the EEG data are usually approximated by a high-order AR model of the form (7.12) ($\nu \gtrsim 10$). As mentioned in Remark 7.4.6, larger τ and thus larger number of parameters are then required, which may lead to over-parametrization. To prevent this, we use an exponentially-regularized version of (7.16) by minimizing

$$\mathrm{tr}(\mathring{\mathbf{e}}_{\tau}^{N-1}(\mathring{\mathbf{e}}_{\tau}^{N-1})^T + \gamma \mathring{\hat{\mathbf{A}}}_{\tau-1}\mathbf{P}\mathbf{P}^T\mathring{\hat{\mathbf{A}}}_{\tau-1}^T), \tag{7.32}$$

where $\mathbf{P} = \mathrm{diag}(1, \rho_0^{-1}, \ldots, \rho_0^{-(\tau-1)}) \otimes \mathbf{I}_{n_m}$ and, ideally, $\rho_0 = \rho(\mathbf{A}_{22})$ (in practice, it is found by trial and error). The role of the exponential regularizer is to encourage the higher-order AR terms to decay exponentially, as $\tilde{\mathbf{A}}_i^*$ do. In the simulations that follow, we have used $\gamma = 10$ and $\rho_0 = 0.9$.

Fig. 7.7 shows the reconstructed manifest subnetwork with direct and indirect connections using the LSAR method for $\tau = 15$ and different values of α. One can observe that the sensitivity of the network structure to the threshold ratio α is significant, showing that the majority of network links are relatively weak with respect to the largest link (which is usually a self-loop). This sensitivity, however, is smaller for the indirect connections. Note that increasing α is a way of enforcing sparsity among the manifest nodes similar (but not equivalent) to [21]. Also, note that unlike [21], the manifest subnetwork estimated by our method is directed (though directions are not shown in Fig. 7.7 for simplicity).

For comparison, Fig. 7.8 shows the reconstructed manifest subnetwork with direct and indirect connections using the S+L method of [21] for $n = 5$[5]. Although the use of a threshold value is

[5] n represents the model order in [21]. While the role of the model order is not discussed in the reference, the use of higher-order models significantly increases the computational cost of the algorithm. Also, note that there is no one-to-one correspondence between the subfigures of Figs. 7.7-7.9.

not prescribed in [21], we have used a fixed value of $\alpha = 0.01$ for all values of (λ, γ), since the absence of a threshold ($\alpha = 0$) results in all nodes being estimated to be (both directly and indirectly) connected. This lack of sparsity occurs for all values of (λ, γ) (no matter how large they are chosen), unless extremely large values are employed, which results in a fully disconnected network. From various plots, we see that even with the use of a threshold value all the nodes are estimated to be indirectly connected, with the sparsity of direct connections and the estimated number of latent nodes being determined by (λ, γ). This abundance of indirect connections and parameter-based tuning of direct connectivity is similar to our results in Fig. 7.7, even though the details of the reconstructed networks do not exactly match.

Fig. 7.9 shows the result of applying both the Directed Transfer Function (DTF) [34] and direct Directed Transfer Function (dDTF) methods to the EEG channel data to estimate the indirect and direct connections between the manifest nodes, respectively, for different frequency bands. Both methods are applied to the data using the EEGLAB SIFT plugin for $\tau = 15$ (selected based on SIFT Model Order Selection). In all cases, a constant threshold ratio $\alpha = 0.1$ is used and the value of the threshold is computed with respect to the largest *off-diagonal* link weight in the same frequency. As can be seen, the connectivity pattern is considerably different between lower and higher frequencies, where several pairs are not even indirectly connected over the δ-θ band. This is in contrast to the reconstructed networks of Fig. 7.7 in which most pairs are at least indirectly connected, even for threshold values as large as $\alpha = 0.15$. Nevertheless, a common feature of all the reconstructed networks in Figs. 7.7-7.9 is that the density of direct connections is higher in the fronto-central (FC) areas and lower in central (C) areas and midline frontal pole (FPz). The independence of this sparsity pattern from the employed reconstruction method and parameter value suggests that it is a robust feature of the actual brain connectivity among these areas.

Since the *true* network structure is unknown for this example (and hence the methods are not directly comparable), we validate our LSAR estimated connectivity based on its ability to predict *future* (i.e., unseen) channel activity. Thus, we used the first 80% of data for LSAR estimation and the last 20% for evaluation, which is based on

$$R^2 = 1 - \frac{\sum_{k=N+1}^{N'} \|\mathbf{e}(k)\|^2}{\sum_{k=N+1}^{N'} \|\mathbf{y}(k)\|^2}, \tag{7.33}$$

denoting the percentage of the future channel activity that is correctly predicted by the model [25, §16.4], where $\mathring{\mathbf{y}}_{N+1}^{N'}$ is the latter data sequence not used for estimation. The blue curve in Fig. 7.10(c) shows the value of $R \times 100\%$ for the LSAR method as a function of model order for the same selection of nodes as above (i.e., anterior)[6]. This shows that the method is capable of predicting more than 96.5% of unseen data with model orders $\tau = 15 \sim 20$ (which is relatively low given the large number of latent nodes and the high order of the underlying brain dynamics). It should be noted that the R-value is not a suitable measure for comparison among the networks obtained by the LSAR, S+L, and dDTF methods. On the one hand, the AR model underlying the dDTF method is almost identical to the LSAR model used here, resulting in almost identical R values, while the reconstructed networks are considerably different (c.f. Figs. 7.7 and 7.9) due to different interpretations of on the model implications for network connectivity. On the other hand, the R value is not well-defined for the S+L method since the right-hand side of (7.33) is negative, i.e., the reconstructed AR model has extremely poor *prediction* performance. This is not surprising as the S+L method is aimed at maximizing the entropy (and thus minimizing predictability).

Next, we analyzed the effect of the choice of manifest nodes on the reconstructed network.

[6]Edge values are not thresholded ($\alpha = 0$) for computing R values.

In addition to selecting the 13 most anterior cortical nodes as above, we performed other runs where we selected the 13 most posterior nodes and 13 random nodes to reconstruct the manifest network using the LSAR method. We show in Fig. 7.10 these node choices (a), the reconstructed network for the posterior (b) and random (d) selections ($\alpha = 0.12$), and (c) the R values for all three cases. Interestingly, the density of direct connections is significantly higher among the posterior nodes. Also, the LSAR prediction performance is significantly lower in this case, suggesting less conformity of the occipito-parietal cortex to the simplifying assumptions of our AR model (linearity and passivity of latent nodes). Consistently, the network density and R value of the random case interpolates between the anterior and posterior cases, as expected.

Finally, an interesting observation in Fig. 7.10(c) is that, even an AR model with $\tau = 2$ can predict about 95% of unseen data in all cases. This, at first glance, questions the need for any higher-order models as far as prediction is concerned. Nevertheless, notice that even an AR model with $\tau = 1$, corresponding to an *isolated* manifest subnetwork, can predict 90% of unseen data, while the visual discrimination task performed by the subject heavily relies on coordination between posterior (visual) and anterior (motor planning and execution) areas. The reason why this model can predict unseen data so well is in the strong dominance of first-order local dynamics of every area (the diagonal of $\tilde{A}_0$) over the rest of network dynamics.[7] Thus, the prediction performance of a first-order model serves as a *baseline* for higher orders, capturing the contribution of local interactions to the overall brain dynamics. This enlightens why the $\sim 1\%$ improvement in prediction performance as we go from $\tau = 2$ to $\tau = 15 \sim 20$ is significant. □

[7] This can be easily seen by inspecting the AR coefficients $\tilde{A}_i$ estimated from data, and is physiologically justified as each area is composed of millions of neurons that are locally densely connected and serve specific purposes but only (relatively) sparsely connected with remote areas.

Appendix

7.A Auxiliary Result

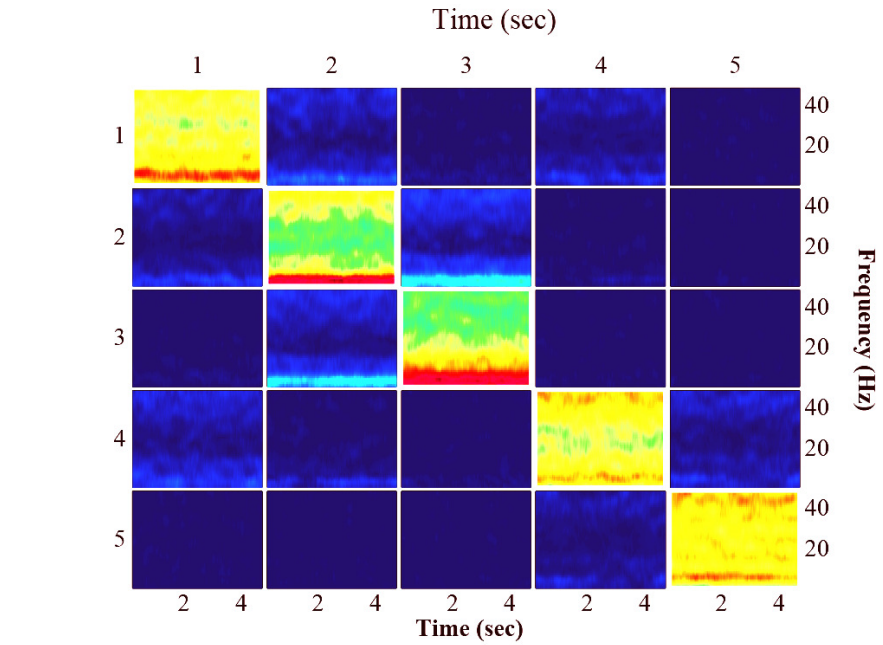
Lemma 7.A.1. *Given two vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$*, it holds for any* $M \in \mathbb{R}_{>0}$ *that* $\|\mathbf{a}\mathbf{b}^T\|_{\max} \leq M^{-1}\mathbf{a}^T\mathbf{a} + M\mathbf{b}^T\mathbf{b}$.

Proof. By definition of the max norm,

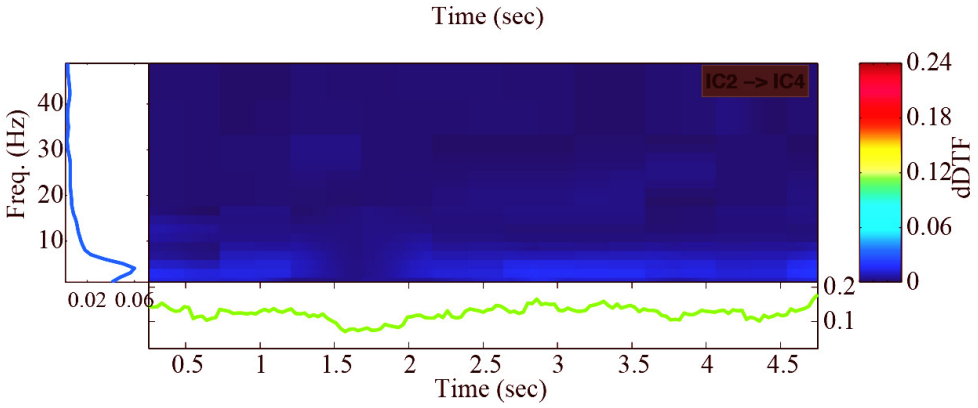
$$\begin{aligned}\|\mathbf{a}\mathbf{b}^T\|_{\max} &= \max_{1\leq i,j\leq n} \left|a_i b_j\right| \leq \sum_{i=1}^{n} (M^{-1}\left|a_i\right|^2 + M\left|b_i\right|^2) \\ &= M^{-1}\mathbf{a}^T\mathbf{a} + M\mathbf{b}^T\mathbf{b}.\end{aligned}$$

□

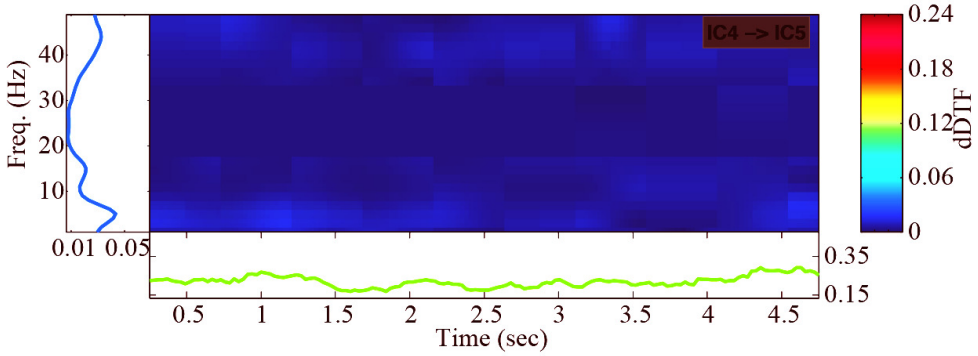
Acknowledgements: This chapter is taken, in part, from the work published as "Network identification with latent nodes via auto-regressive models" by E. Nozari, Y. Zhao, and J. Cortés in *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 722–736, 2018. The dissertation author was the primary investigator and author of this paper.



(a)

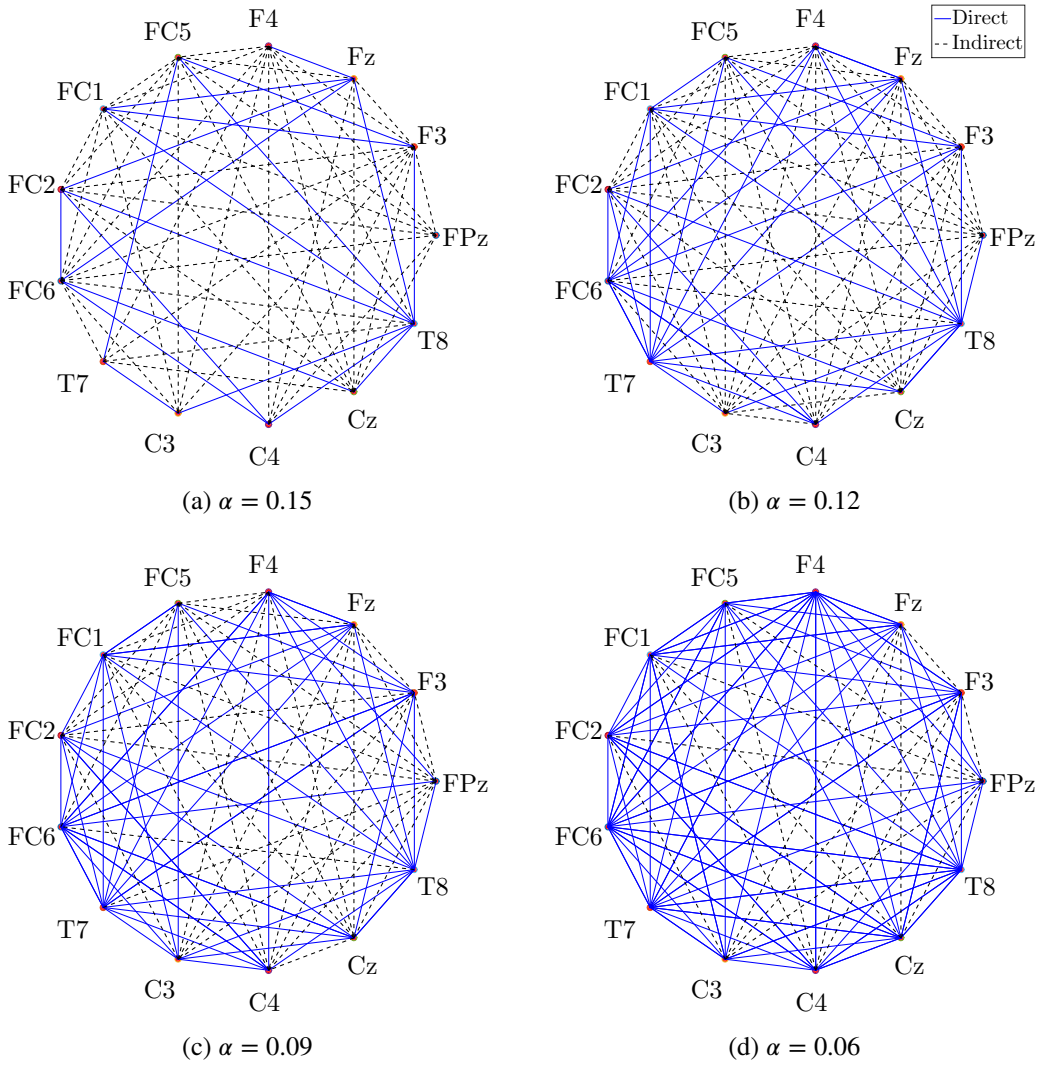


(b)



(c)

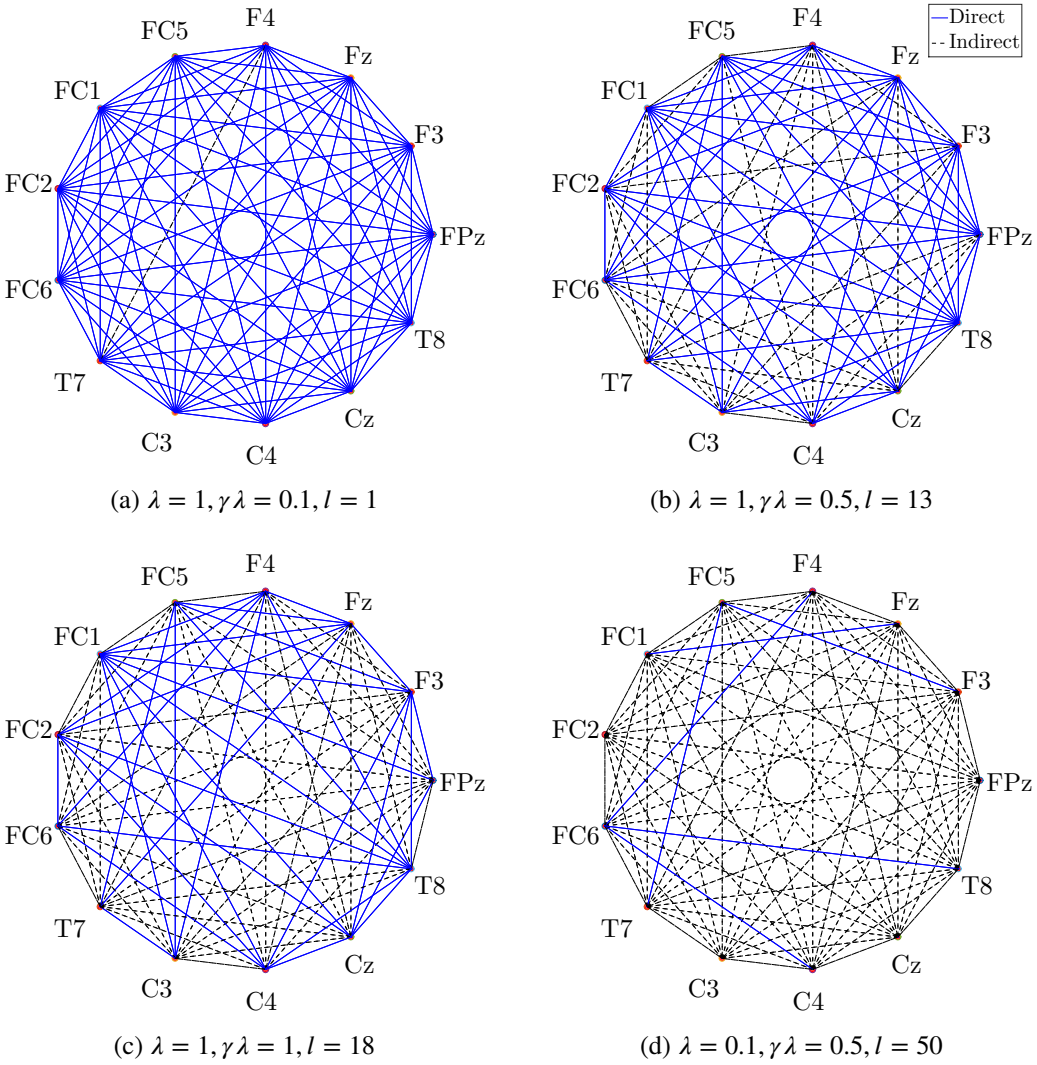
Figure 7.6: (a) The interaction topology identified by the dDTF method for the Erdős–Rényi network with index 10. (b and c) A zoom-in of the (indirect) connection $(2, 4)$ and the (direct) connection $(4, 5)$, resp.



(a) $\alpha = 0.15$ (b) $\alpha = 0.12$

(c) $\alpha = 0.09$ (d) $\alpha = 0.06$

Figure 7.7: Reconstructed manifest subnetwork for the EEG data in Example 7.5.3 using our proposed method with the exponentially-regularized objective function (7.32) and $\gamma = 10$, $\rho_0 = 0.9$, and $\tau = 15$. The direct (solid blue) and indirect (dashed black) connections are depicted for different values of threshold ratio α. For each value of α, the connections whose weights are smaller than α times the largest network weight are removed.



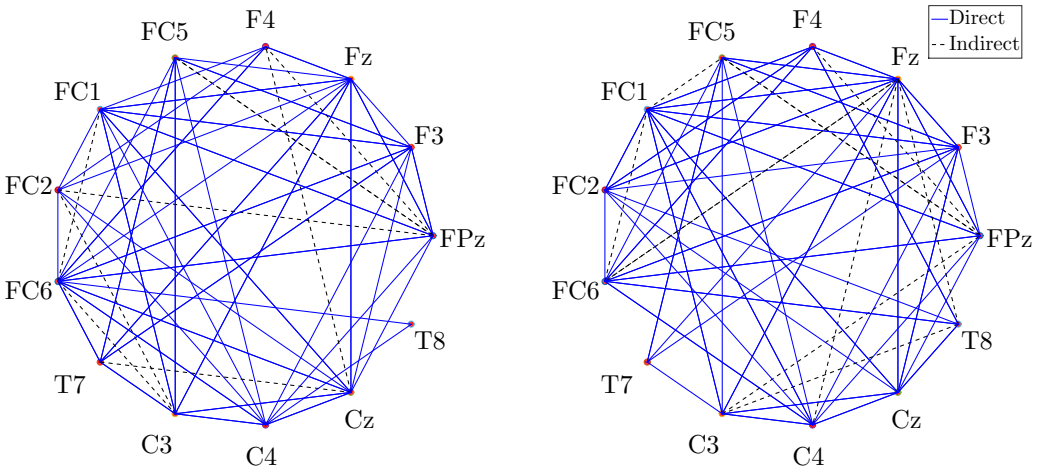
(a) $\lambda = 1, \gamma\lambda = 0.1, l = 1$

(b) $\lambda = 1, \gamma\lambda = 0.5, l = 13$

(c) $\lambda = 1, \gamma\lambda = 1, l = 18$

(d) $\lambda = 0.1, \gamma\lambda = 0.5, l = 50$

Figure 7.8: Reconstructed manifest subnetwork for Example 7.5.3 using the S+L method in [21]. The direct (solid blue) and indirect (dashed black) connections are depicted for different values of weight parameters (λ, γ) and fixed threshold ratio $\alpha = 0.01$. l represents the estimated number of latent nodes.



(a) $f = 1^{Hz}$ (δ band) (b) $f = 5^{Hz}$ (θ band)

(c) $f = 20^{Hz}$ (β band) (d) $f = 50^{Hz}$ (γ band)

Figure 7.9: Reconstructed manifest subnetwork for Example 7.5.3 using the combination of DTF and dDTF estimation methods. The direct (solid blue) and indirect (dashed black) connections are illustrated for different frequency values and fixed threshold ratio $\alpha = 0.1$.

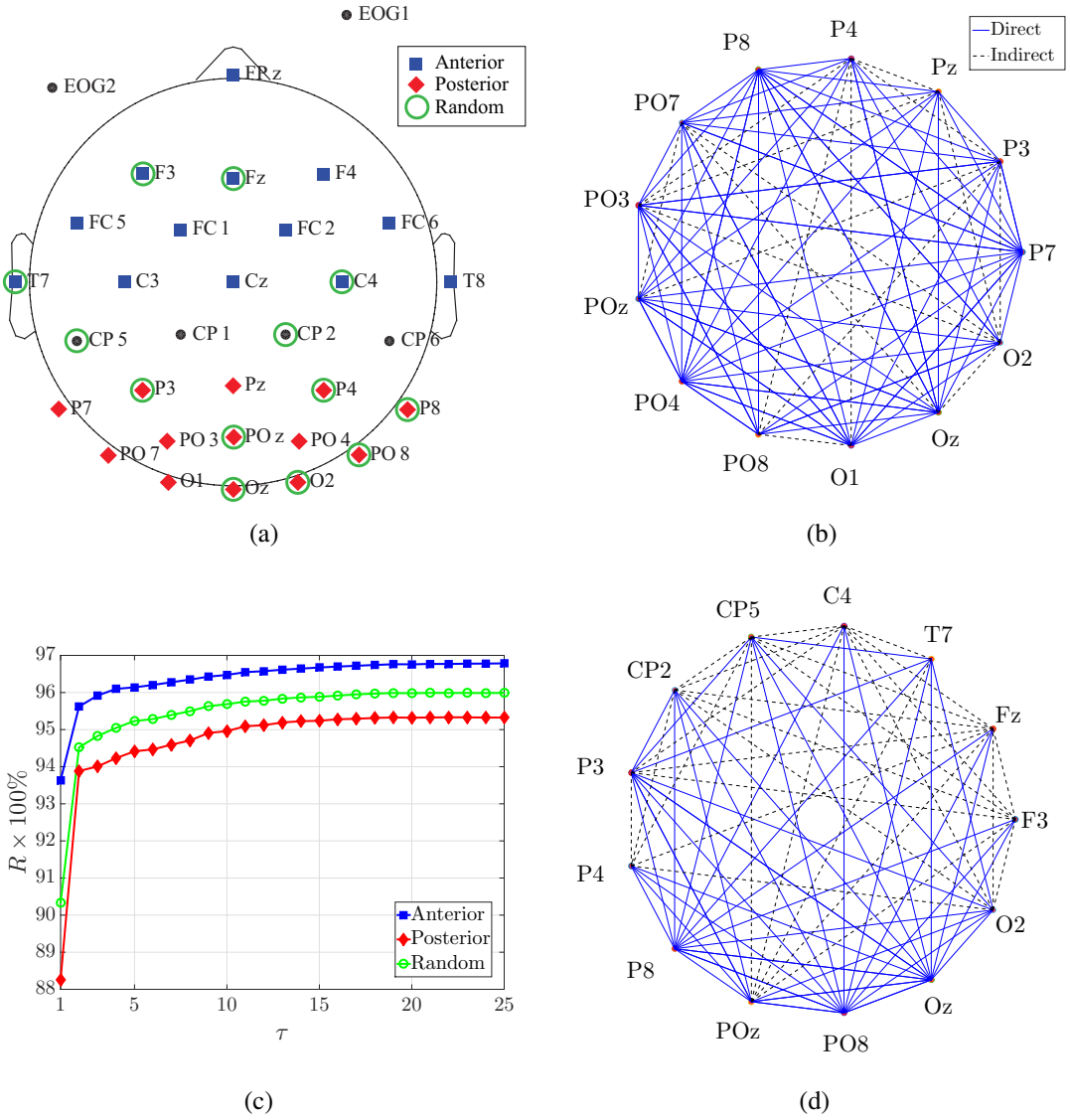


Figure 7.10: Comparison between different selections of manifest nodes in Example 7.5.3: (a) Electrode locations. (b and d) The reconstructed network for the 13 posterior nodes and 13 random nodes, resp. ($\alpha = 0.12$). (c) Prediction performance R for the three different choices of manifest nodes (reconstructed network for anterior selection is given in Fig. 7.7(b)).

Chapter Bibliography

[1] V. Sakkalis, "Review of advanced techniques for the estimation of brain connectivity measured with eeg/meg," *Computers in Biology and Medicine*, vol. 41, no. 12, pp. 1110–1117, 2011.

[2] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: a well established methodology," *NeuroImage*, vol. 58, no. 2, pp. 323–329, 2011.

[3] J. R. Iversen, A. Ojeda, T. Mullen, M. Plank, J. Snider, G. Cauwenberghs, and H. Poizner, "Causal analysis of cortical networks involved in reaching to spatial targets," in *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, 2014, pp. 4399–4402.

[4] A. Korzeniewska, C. M. Crainiceanu, R. Kuś, P. J. Franaszczuk, and N. E. Crone, "Dynamics of event-related causality in brain electrical activity," *Human Brain Mapping*, vol. 29, no. 10, pp. 1170–1192, 2008.

[5] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological Cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[6] Y. X. R. Wang and H. Huang, "Review on statistical methods for gene network reconstruction using expression data," *Journal of Theoretical Biology*, vol. 362, pp. 53–61, 2014.

[7] A. Julius, M. Zavlanos, S. Boyd, and G. J. Pappas, "Genetic network identification using convex programming," *IET Systems Biology*, vol. 3, no. 3, pp. 155–166, 2009.

[8] C. D. Godsil and G. F. Royle, *Algebraic Graph Theory*, ser. Graduate Texts in Mathematics. Springer, 2001, vol. 207.

[9] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[10] M. Nabi-Abdolyousefi and M. Mesbahi, "Network identification via node knockout," *IEEE Transactions on Automatic Control*, vol. 57, no. 12, pp. 3214–3219, 2012.

[11] S. Shahrampour and V. M. Preciado, "Topology identification of directed dynamical networks via power spectral analysis," *IEEE Transactions on Automatic Control*, vol. 60, no. 8, pp. 2260–2265, 2015.

[12] M. Timme, "Revealing network connectivity from response dynamics," *Physical Review Letters*, vol. 98, no. 22, p. 224101, 2007.

[13] D. Materassi, G. Innocenti, L. Giarré, and M. V. Salapaka, "Model identification of a network as compressing sensing," *Systems & Control Letters*, vol. 62, no. 8, pp. 664–672, 2013.

[14] D. Materassi and G. Innocenti, "Topological identification in networks of dynamical systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1860–1871, 2010.

[15] M. J. Choi, V. Y. Tan, A. Anandkumar, and A. S. Willsky, "Learning latent tree graphical models," *Journal of Machine Learning Research*, vol. 12, pp. 1771–1812, 2011.

[16] D. Materassi and M. V. Salapaka, "Network reconstruction of dynamical polytrees with unobserved nodes," in *IEEE Conf. on Decision and Control*, Maui, Hawaii, USA, 2012, pp. 4629–4634.

[17] J. Gonçalves and S. Warnick, "Necessary and sufficient conditions for dynamical structure reconstruction of LTI networks," *IEEE Transactions on Automatic Control*, vol. 53, no. 7, pp. 1670–1674, 2008.

[18] Y. Yuan, G. B. Stan, S. Warnick, and J. Goncalves, "Robust dynamical network structure reconstruction," *Automatica*, vol. 47, no. 6, pp. 1230–1235, 2011, special Issue on Systems Biology.

[19] Y. Yuan, K. Glover, and J. Goncalves, "On minimal realisations of dynamical structure functions," *Automatica*, vol. 55, pp. 159–164, 2015.

[20] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, "Latent variable graphical model selection via convex optimization," *The Annals of Statistics*, vol. 40, no. 4, pp. 2005–2013, 2012.

[21] M. Zorzi and R. Sepulchre, "AR identification of latent-variable graphical models," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2327–2340, 2016.

[22] C. I. Byrnes, S. V. Gusev, and A. Lindquist, "From finite covariance windows to modeling filters: A convex optimization approach," *SIAM Review*, vol. 43, no. 4, pp. 645–675, 2001.

[23] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.

[24] A. Korzeniewska, M. Mańczak, M. Kamiński, K. J. Blinowska, and S. Kasicki, "Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method," *Journal of Neuroscience Methods*, vol. 125, no. 1, pp. 195–207, 2003.

[25] L. Ljung, *System Identification: Theory for the User*, ser. Prentice Hall information and system sciences series. Prentice Hall, 1999.

[26] P. Henrici, *Applied and Computational Complex Analysis, Volume 1: Power Series Integration Conformal Mapping Location of Zero*. Wiley, 1988.

[27] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[28] A. Papoulis and S. U. Pillai, Eds., *Probability, Random Variables and Stochastic Processes*. McGraw-Hill, 2002.

[29] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, vol. 71, pp. 441–479, 1912.

[30] B. Bollobás, *Random Graphs*, 2nd ed. Cambridge University Press, 2001.

[31] T. Mullen, A. Delorme, C. Kothe, and S. Makeig, "An electrophysiological information flow toolbox for EEGLAB," *Biological Cybernetics*, vol. 83, pp. 35–45, 2010.

[32] A. Delorme, T. Mullen, C. Kothe, Z. A. Acar, N. Bigdely-Shamlo, A. Vankov, and S. Makeig, "EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing," *Computational Intelligence and Neuroscience*, vol. 2011, p. 10, 2011.

[33] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[34] M. J. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological Cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.

Part III

Network Dynamics and Cognition

Chapter 8

Hierarchical Selective Recruitment

In this chapter, we being our analysis of the relationship between the network dynamics of the brain and cognition, which serves as one of the most important and concrete application areas of network dynamical systems. The human brain is constantly under the influx of sensory inputs and is responsible for integrating and interpreting them to generate appropriate decisions and actions. This influx contains not only the pieces of information relevant to the present task(s), but also a myriad of distractions. Goal-driven selective attention (GDSA) refers to the active prioritization of the processing of task-relevant information over task-irrelevant distractions according to one's goals and desires and is vital for our ability to construct a dynamic yet coherent perception of the world.[1] Examples of GDSA range from selective audition in a crowded place to selective vision in cluttered environments to selective taste/smell in food. As a result, a long standing question in neuroscience involves understanding the brain's complex mechanisms underlying selective attention [1–6]. Despite major advances, a fundamental understanding of GDSA and, in particular, how

[1] Note the distinction of this with stimulus-driven selective attention (the reactive shift of focus based on saliency of stimuli) which is not the focus of this work.

it emerges from the dynamics of the underlying neuronal networks, is still missing. The aim of this chapter is to reduce this gap by bringing tools and insights from systems and control theory into these questions from neuroscience.

In this chapter, we propose the novel theoretical framework of Hierarchical Selective Recruitment (HSR) for GDSA. This framework is inspired by the extensive experimental research [1–16] that has discovered some of the fundamental aspects of the neuronal mechanisms underlying GDSA. These include (i) the brain's hierarchical organization, so that (cognitively-)higher areas (e.g., prefrontal cortex) control the activity of lower level ones (e.g., primary sensory and motor cortices), (ii) the separation of timescales between subsequent hierarchical layers, (iii) its sparse coding, so that task-relevant and task-irrelevant information is represented and processed by different and mostly distinct neuronal populations (particularly for sufficiently distinct stimuli), (iv) the distributed and graded nature of GDSA, so that selective attention happens at multiple layers of the hierarchy, and (v) the concurrence of the suppression and enhancement of task-irrelevant and task-relevant activity, respectively (formulated as *selective inhibition* and *top-down recruitment* in HSR, respectively).

We begin our development of HSR by analyzing the internal dynamics of each layer of the hierarchy described as a network with linear-threshold dynamics and deriving conditions on its structure to guarantee existence and uniqueness of equilibria, asymptotic stability, and boundedness of trajectories. We also provide mechanisms that enforce selective inhibition using the biologically-inspired schemes of feedforward and feedback inhibition. Despite their differences, both schemes lead to the same conclusion: the intrinsic dynamical properties of the (not-inhibited) task-relevant subnetworks are the sole determiner of the dynamical properties that are achievable under selective inhibition. Based on these results, we derive conditions on the joint structure of the hierarchical sub-

networks that guarantee top-down recruitment of the task-relevant part of each subnetwork by the subnetwork at the layer immediately above, while inhibiting the activity of task-irrelevant subnetworks at all the hierarchical layers. To further verify the merit and applicability of this framework, we carry out a comprehensive case study of selective listening in rodents and show that a small network with HSR-based structure and minimal size can explain the data with remarkable accuracy while satisfying the theoretical requirements of HSR. Our technical approach relies on the theory of switched systems and provides a novel converse Lyapunov theorem for state-dependent switched affine systems that is of independent interest.

8.1 Prior Work

In this work we use dynamical networks with linear-threshold nonlinearities (also called rectified linear units, ReLU, in machine learning) to model the activity of neuronal populations. Linear-threshold models allow for a unique combination between the tractability of linear systems and the dynamical versatility of nonlinear systems, and thus have been widely used in computational neuroscience. They were first proposed as a model for the lateral eye of the horseshoe crab in [17] and their dynamical behavior has been studied at least as early as [18]. A detailed stability analysis of symmetric (undirected) linear-threshold networks has been carried out in continuous [19] and discrete [20] time: however, this has limited relevance for biological neuronal networks, which are fundamentally asymmetric (due to the presence of excitatory and inhibitory neurons). An initial summary of the properties of general (possibly asymmetric) networks, including the existence and uniqueness of equilibria and asymptotic stability was given in [21], with limited rigorous justification provided later in [22]. Lyapunov-based methods were used in a number of later studies for

discrete-time linear-threshold networks [23–25], but the extension of these results to continuous-time dynamics, which has more relevance to biological neuronal networks, is not clear. In fact, the use of Lyapunov-based techniques in continuous-time networks has remained limited to planar dynamics [26] and restrictive conditions for boundedness of trajectories [26, 27]. Recently, [28] presents interesting properties of competitive (i.e., fully inhibitory) linear-threshold networks, particularly regarding the emergence of limit cycles. However, the majority of neurons in biological neuronal networks are excitatory, making the implications of these results limited. Moreover, all the preceding works are limited to networks with constant exogenous inputs whereas time-varying inputs are essential for modeling inter-layer connections in HSR.

A critical property of linear-threshold networks is that their nonlinearity, while enriching their behavior beyond that of linear systems, is piecewise linear. Accordingly, almost all the theoretical analysis of these networks builds upon the formulation of them as switched affine systems. There exists a vast literature on the analysis of general switched linear/affine systems, see, e.g., [29–31]. Nevertheless, we have found that the conditions obtained by applying these results to linear-threshold dynamics are more conservative than the ones we obtain using direct analysis of the system dynamics. This is mainly due to the fact that such results, by the essence of their generality, are oblivious to the particular structure of linear-threshold dynamics that can be leveraged in direct analysis.

A critical component of GDSA is selective inhibition. Selective inhibition has been the subject of extensive research in neuroscience. A number of early studies [3,9,10] provided evidence for a mechanism of selective visual attention based on a biased competition between the subnetwork of task-relevant nodes and the subnetwork of task-irrelevant ones. In this model, nodes belonging to these subnetworks compete at each layer by mutually suppressing the activity of each other, and

this competition is biased towards task-relevant nodes by the layer immediately above. Later studies [11, 12] further supported this theory using functional magnetic resonance imaging (fMRI) and showed [32], in particular, the suppression of activity of task-irrelevant nodes as a result of GDSA. This suppression of activity is further shown to occur in multiple layers along the hierarchy [33], grow with increasing attention [34, 35], and be inversely related to the power of the task-irrelevant nodes' state trajectories in the alpha frequency band ($\sim$ 8-14$^{\text{Hz}}$) [14]. Here, we use insights from this body of work in developing a theoretical framework for selective inhibition.

Also critical for GDSA is the hierarchical organization of the brain which has been recognized for decades [36–38] and applies to multiple aspects of brain structure and function. These aspects include (i) network topology [38–41] (where nodes are assigned to layers based on their position on bottom-up and top-down pathways), (ii) encoding properties [42,43] (where nodes that have larger response fields and/or encode more abstract stimulus properties constitute higher layers), (iii) dynamical timescale [39,41,44–55] (where nodes are grouped into layers according to the timescale of their dynamics), (iv) nodal clustering [56–59] (where nodes only constitute the leafs of a clustering tree), and (v) oscillatory activity [60] (where layers correspond to nested oscillatory frequency bands). Note that while hierarchical layers are composed of brain regions in (i)-(iii), this is not the case for (iv) and (v). The hierarchies (i)-(iii) are remarkably similar, and here we particularly focus on (iii) (the timescale separation between hierarchical layers) as it plays a pivotal role in HSR.

Studies of timescale separation between cortical regions are more recent. Several experimental works have demonstrated a clear increase in intrinsic timescales as one moves up the hierarchy using indirect measures such as the length of stimulus presentation that elicits a response [44, 45], resonance frequency [46], the length of the largest time window over which the

responses to successive stimuli interfere [47], and how quickly the activation level of any brain region can track changes in sensory stimuli [48]. Direct evidence for this hierarchical separation of timescales was indeed provided in [49] using the decay rate of spike-count autocorrelation functions. This was shown even more comprehensively in [50] using linear-threshold rate models and the concept of *continuous hierarchies* [39, 41] (whereby the layer of each node can vary continuously according to its intrinsic timescale, therefore removing the rigidity and arbitrariness of node assignment in classical hierarchical structures). Interestingly, recent studies show that this timescale variability may have roots not only in synaptic dynamics of individual neurons [51], but also in sub-neuronal genetic factors [52] as well as supra-neuronal network structures [53]. In terms of applications, computational models of motor control were perhaps the first to exploit this cortical hierarchy of timescales [54, 55]. Despite the vastness of the literature on its roots and applications, we are not aware of any theoretical analysis of the effects of this separation of timescales on the hierarchical dynamics of neuronal networks.

Finally, we use tools and concepts from singular perturbation theory to rigorously leverage this separation of timescales. The classical result on singularly perturbed differential equations goes back to Tikhonov [61], [62, Thm 11.1] and has since inspired an extensive literature, see, e.g. [63–66]. Tikhonov's result, however, requires smoothness of the vector fields, which is not satisfied by linear-threshold dynamics. Fortunately, several works have sought extensions to non-smooth differential equations and even differential inclusions [67–70], culminating in the work [71] which we use here. Similar to Tikhonov's original work, [71] only applies to finite intervals. Extensions to infinite intervals exist [72, 73] but, as expected, they require asymptotic stability of the reduced-order model which we do not in general have.

8.2 Problem Statement

Consider a network of neurons evolving according to (2.8)-(2.9). Since the number of neurons in a brain region is very large, it is common to consider a population of neurons with similar activation patterns as a single *node*. The "firing rate" of such a node is then defined as the average of the individual firing rates. This convention also has the advantage of getting more consistent rates, as the firing pattern of individual neurons may be at times sparse. Accordingly, we use "node" and "population" interchangeably.[2] Combining the nodal rates in a vector $\mathbf{x} \in \mathbb{R}^n$ and synaptic weights in a matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$, we obtain, according to (2.8)-(2.9), the *linear-threshold network dynamics*

$$\tau \dot{\mathbf{x}}(t) = -\mathbf{x}(t) + [\mathbf{W}\mathbf{x}(t) + \mathbf{p}(t)]^+, \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{8.1}$$

The extra term $\mathbf{p}(t) \in \mathbb{R}^n$ captures the *external inputs* to the network including un-modeled background activity and possibly nonzero thresholds (i.e., if a node becomes active when its net input exceeds a threshold other than zero). Note that the right-hand side of (8.2) is a continuous (though not smooth) vector field in $\mathbf{x}$ and thus solutions, in the classical sense, are well defined.

Consistent with the vision for hierarchical selective recruitment (HSR) outlined above, we consider a hierarchical neuronal network $\mathcal{N}$ of the form (8.2), as depicted in Figure 8.1, whereby the nodes in each layer $\mathcal{N}_i$ are further decomposed into a task-irrelevant part $\mathcal{N}_i^{\circ}$ and a task-relevant part $\mathcal{N}_i^{1}$. We make the convention that $\mathcal{N}_i$ is higher in the hierarchy than $\mathcal{N}_j$ if $i < j$.

The state evolution of each layer $\mathcal{N}_i$ is modeled with linear-threshold network dynamics of

[2] Our discussion is nevertheless valid irrespective of whether network nodes represent individual neurons or groups of them.

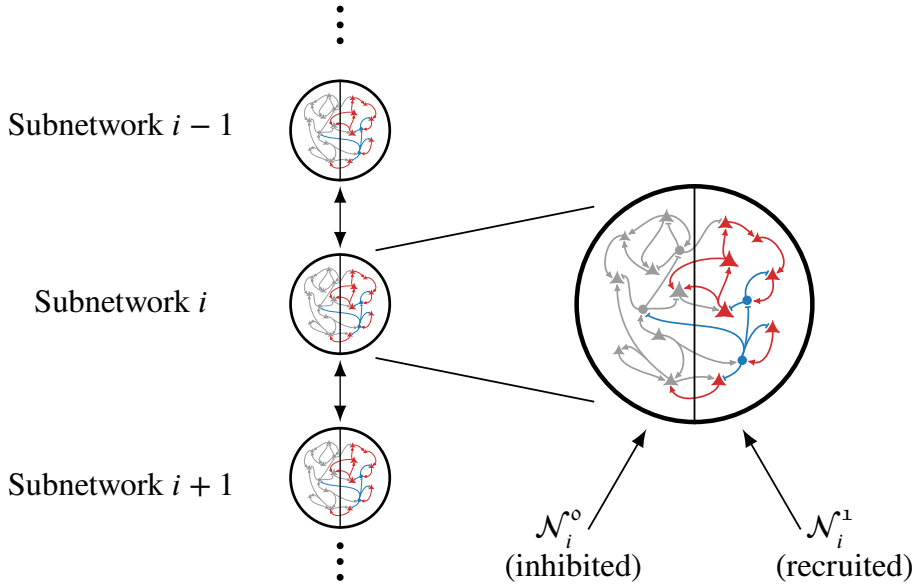


Figure 8.1: The hierarchical network structure considered in this work. Each layer is only directly connected to the layers below and above it. Longer-range connections between non-successive layers do exist in thalamocortical hierarchies but are weaker than those between successive layers and are not considered in this work for simplicity.

the form (8.1), i.e.,

$$\tau_i \dot{\mathbf{x}}_i(t) = -\mathbf{x}_i(t) + [\mathbf{W}_{i,i}\mathbf{x}_i(t) + \mathbf{p}_i(t)]^+, \tag{8.2}$$

where $\mathbf{x}_i \in \mathbb{R}^{n_i}$, $\mathbf{W}_{i,i} \in \mathbb{R}^{n_i \times n_i}$, and $\mathbf{p}_i \in \mathbb{R}^{n_i}$ denote the state, internal synaptic connectivity, and external inputs of $\mathcal{N}_i$, respectively. In this model, the $\mathbf{W}_{i,i}\mathbf{x}_i$ term represents the internal subnetwork connectivity and the $\mathbf{p}_i$ term incorporates, in addition to background activity and nonzero thresholds, the incoming signals from other subnetworks $j \neq i$.

It is worth noticing that each layer of the network of networks exhibits rich dynamical behavior when considered in isolation. Indeed, simulations of the dynamics (8.1) with random instances of $\mathbf{W}$ and constant $\mathbf{p}$ reveal that

– locally, the dynamics may have zero, one, or many equilibrium points, where each equilibrium may be stable or unstable independently of others,

– globally, the dynamics is capable of exhibiting different nonlinear phenomena such as limit cycles, multi-stability, and chaos,

– the state trajectories grow unbounded (in reality until saturation) if the excitatory subnetwork $[\mathbf{W}]^+$ is sufficiently strong.

This richness of behavior can only increase if the layer is subject to a time-varying external input $\mathbf{p}(t)$, and in particular when interconnected with other layers in the network of networks. Motivated by these observations, our ultimate goal in this work is to characterize the dynamics of complex networks composed of hierarchically-connected linear-threshold subnetworks and the conditions under which their collective dynamics can give rise to HSR. Specifically, we tackle the following problems:

(i) the analysis of the relationship between structure ($\mathbf{W}_{i,i}$) and dynamical behavior for each subnetwork when operating in isolation from the rest of the network ($\mathbf{p}_i(t) \equiv \mathbf{p}_i$);

(ii) the analysis of the conditions on the joint structure of each two successive layers $\mathcal{N}_i$ and $\mathcal{N}_{i+1}$ that allows for selective inhibition of $\mathcal{N}_{i+1}^{\circ}$ by its input from $\mathcal{N}_i$, being equivalent to the stabilization of $\mathcal{N}_{i+1}^{\circ}$ to the origin (inactivity);

(iii) the analysis of the conditions on the joint structure of each two successive layers $\mathcal{N}_i$ and $\mathcal{N}_{i+1}$ that allows for top-down recruitment of $\mathcal{N}_{i+1}^{1}$ by its input from $\mathcal{N}_i$, being equivalent to the stabilization of $\mathcal{N}_{i+1}^{1}$ toward a desired trajectory set by $\mathcal{N}_i$ (activity);

(iv) the combination of (ii) and (iii) in a unified framework and its extension to the complete N-layer network of networks.

We let

$$\mathbf{p}_i(t) = \mathbf{B}_i \mathbf{u}_i(t) + \tilde{\mathbf{p}}_i(t),$$

where $\mathbf{u}_i \in \mathbb{R}^{m_i}_{\geq 0}$ is the top-down control used for inhibition of $\mathcal{N}_i^{\circ}$. While in Part I we assume for simplicity that $\tilde{\mathbf{p}}_i(t)$ is given and constant, we here consider its complete form

$$\tilde{\mathbf{p}}_i(t) = \mathbf{W}_{i,i-1}\mathbf{x}_{i-1}(t) + \mathbf{W}_{i,i+1}\mathbf{x}_{i+1}(t) + \mathbf{c}_i,$$

where the inter-layer connectivity matrices $\mathbf{W}_{i,i-1}$ and $\mathbf{W}_{i,i+1}$ have appropriate dimensions and $\mathbf{c}_i \in \mathbb{R}^{n_i}$ captures un-modeled background activity and possibly nonzero activation thresholds. Substituting these into (8.2), the dynamics of each layer $\mathcal{N}_i$ is given by

$$\tau_i \dot{\mathbf{x}}_i(t) = -\mathbf{x}_i(t) + [\mathbf{W}_{i,i}\mathbf{x}_i(t) + \mathbf{W}_{i,i-1}\mathbf{x}_{i-1}(t) + \mathbf{W}_{i,i+1}\mathbf{x}_{i+1}(t) + \mathbf{B}_i\mathbf{u}_i(t) + \mathbf{c}_i]^+. \tag{8.3}$$

Also following Part I, we partition $\mathbf{x}_i$, $\mathbf{W}_{i,j}$, $\mathbf{B}_i$, and $\mathbf{c}_i$ as[3]

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^{\circ} \\ \mathbf{x}_i^{1} \end{bmatrix}, \quad \mathbf{W}_{i,j} = \begin{bmatrix} \mathbf{W}_{i,j}^{\circ\circ} & \mathbf{W}_{i,j}^{\circ 1} \\ \mathbf{W}_{i,j}^{1\circ} & \mathbf{W}_{i,j}^{11} \end{bmatrix}, \quad \mathbf{B}_i = \begin{bmatrix} \mathbf{B}_i^{\circ} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{c}_i = \begin{bmatrix} \mathbf{c}_i^{\circ} \\ \mathbf{c}_i^{1} \end{bmatrix}, \tag{8.4}$$

where $\mathbf{x}_i^{\circ} \in \mathbb{R}^{r_i}$ for all $i, j \in \{1, \ldots, N\}$. By convention, $\mathbf{W}_{1,0} = \mathbf{0}$, $\mathbf{W}_{N,N+1} = \mathbf{0}$, and $r_1 = 0$ (so

[3]This sparsity pattern can always be achieved by (re-)labeling the nodes.

$\mathbf{B}_1 = \mathbf{0}$ and the first subnetwork has no inhibited part). Throughout, we assume that the hierarchical layers have sufficient timescale separation, i.e.,

$$\tau_1 \gg \tau_2 \gg \cdots \gg \tau_N.$$

Finally, define

$$\epsilon = (\epsilon_1, \ldots, \epsilon_{N-1}), \quad \epsilon_i = \frac{\tau_{i+1}}{\tau_i}, \qquad i = \{1, \ldots, N-1\}.$$

In the following, we first analyze the internal dynamics of each layer of the hierarchy and then formulate and analyze selective inhibition and selective recruitment, respectively. A comprehensive case study is presented at the end to illustrate and verify the developed framework using.

8.3 Internal Dynamics of Single-Layer Networks

In this section, we provide an in-depth study of the basic dynamical properties of the network dynamics (8.1) in isolation. In such case, the external input $\mathbf{p}(t)$ boils down to background activity and possibly nonzero thresholds, which are constant relative to the timescale τ. The dynamics (8.1) thus simplify to

$$\tau \dot{\mathbf{x}}(t) = -\mathbf{x}(t) + [\mathbf{W}\mathbf{x}(t) + \mathbf{p}]^+, \quad t \geq 0. \tag{8.5}$$

In the following, we derive conditions in terms of the network structure for EUE, local/global asymptotic stability, and boundedness of trajectories.

8.3.1 Dynamics as Switched Affine System

The nonlinear dynamics (8.5) is a switched affine system with 2^n modes. Each mode of this system corresponds to a switching index $\boldsymbol{\sigma} \in \{0,1\}^n$, where for each $i \in \{1,\dots,n\}$, $\sigma_i = 1$ if the node is active (i.e., $(\mathbf{W}\mathbf{x}(t)+\mathbf{p})_i > 0$) and $\sigma_i = 0$ if the node is inactive (i.e., $(\mathbf{W}\mathbf{x}(t)+\mathbf{p})_i \leq 0$). Clearly, the mode of the system varies with time and within the one corresponding to $\boldsymbol{\sigma} \in \{0,1\}^n$, we have

$$[\mathbf{W}\mathbf{x}(t)+\mathbf{p}]^+ = \boldsymbol{\Sigma}(\mathbf{W}\mathbf{x}(t)+\mathbf{p}),$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma})$. This switched representation of the dynamics motivates the following assumptions on the weight matrix $\mathbf{W}$.

Assumption 8.3.1. Assume

(i) $\det(\mathbf{W}) \neq 0$;

(ii) $\det(\mathbf{I}-\boldsymbol{\Sigma}\mathbf{W}) \neq 0$ for all the 2^n matrices $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\sigma}), \boldsymbol{\sigma} \in \{0,1\}^n$. □

Assumption 8.3.1 is not a restriction in practice since the set of matrices for which it is not satisfied can be expressed as a finite union of measure-zero sets, and hence has measure zero. By Assumption 8.3.1(i), the system of equations $\mathbf{W}\mathbf{x}+\mathbf{p} = \mathbf{0}$ defines a non-degenerate set of n hyperplanes partitioning $\mathbb{R}^n$ into 2^n solid convex polytopic translated cones apexed at $-\mathbf{W}^{-1}\mathbf{p}$. For each $\boldsymbol{\sigma} \in \{0,1\}^n$, let $\Omega_{\boldsymbol{\sigma}}$ be the associated switching region,

$$\Omega_{\boldsymbol{\sigma}} = \{\mathbf{x} \in \mathbb{R}_{\geq 0} \mid (2\boldsymbol{\Sigma}-\mathbf{I})(\mathbf{W}\mathbf{x}+\mathbf{p}) \geq \mathbf{0}\}.$$

The piecewise-affine dynamics (8.5) can be written in the equivalent form

$$\tau\dot{\mathbf{x}} = (-\mathbf{I} + \mathbf{\Sigma W})\mathbf{x} + \mathbf{\Sigma p}, \quad \forall \mathbf{x} \in \Omega_{\boldsymbol{\sigma}},\ \boldsymbol{\sigma} \in \{0,1\}^n. \tag{8.6}$$

Unlike linear systems, the existence of equilibria is not guaranteed for this system. In fact, for each $\boldsymbol{\sigma} \in \{0,1\}^n$, according to (8.6), the point

$$\mathbf{x}^*_{\boldsymbol{\sigma}} = \mathbf{x}^*_{\boldsymbol{\sigma}}(\mathbf{p}) = (\mathbf{I} - \mathbf{\Sigma W})^{-1}\mathbf{\Sigma p}, \tag{8.7}$$

is the corresponding *equilibrium candidate*. This equilibrium candidate is indeed an equilibrium if it belongs to the switching region $\Omega_{\boldsymbol{\sigma}}$ where the description (8.6) is valid. We next identify conditions for this to be true.

8.3.2 Existence and Uniqueness of Equilibria

Here we characterize the EUE for the dynamics (8.5). Given $\mathbf{W} \in \mathbb{R}^{n\times n}$, define the *equilibria set-valued map* $h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n_{\geq 0}$ by

$$h(\mathbf{p}) = h_{\mathbf{W}}(\mathbf{p}) \triangleq \{\mathbf{x} \in \mathbb{R}^n_{\geq 0} \mid \mathbf{x} = [\mathbf{W}\mathbf{x} + \mathbf{p}]_+\}. \tag{8.8}$$

The map h can, in particular, take empty values. EUE then precisely corresponds to h being single-valued on $\mathbb{R}^n$. If so, with a slight abuse of notation, we take $h : \mathbb{R}^n \to \mathbb{R}^n_{\geq 0}$ to be an ordinary function.

From the definition (8.7) of equilibrium candidate, note that $\mathbf{x}^*_{\boldsymbol{\sigma}} \in h(\mathbf{p})$ if and only if $\mathbf{x}^*_{\boldsymbol{\sigma}} \in$

$\Omega_{\boldsymbol{\sigma}}$. Then, using Assumption 8.3.1, and after some manipulations, we have

$$\begin{aligned}\mathbf{W}\mathbf{x}_{\boldsymbol{\sigma}}^{*}+\mathbf{p} &= \mathbf{W}(\mathbf{I}-\boldsymbol{\Sigma}\mathbf{W})^{-1}\boldsymbol{\Sigma}\mathbf{p}+\mathbf{p} \\ &= (\mathbf{W}^{-1}-\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}\mathbf{p}+\mathbf{p} \\ &= (\mathbf{I}-\mathbf{W}\boldsymbol{\Sigma})^{-1}\mathbf{W}\boldsymbol{\Sigma}\mathbf{p}+\mathbf{p} \\ &= \left[(\mathbf{I}-\mathbf{W}\boldsymbol{\Sigma})^{-1}\mathbf{W}\boldsymbol{\Sigma}+\mathbf{I}\right]\mathbf{p} = (\mathbf{I}-\mathbf{W}\boldsymbol{\Sigma})^{-1}\mathbf{p}.\end{aligned} \tag{8.9}$$

Therefore,

$$\begin{aligned}&\mathbf{x}_{\boldsymbol{\sigma}}^{*}\in h(\mathbf{p}) \Leftrightarrow \mathbf{x}_{\boldsymbol{\sigma}}^{*}\in\Omega_{\boldsymbol{\sigma}} \\ &\qquad\Leftrightarrow \underbrace{(2\boldsymbol{\Sigma}-\mathbf{I})(\mathbf{I}-\mathbf{W}\boldsymbol{\Sigma})^{-1}}_{\triangleq\mathbf{M}_{\boldsymbol{\sigma}}}\mathbf{p}\geq\mathbf{0}.\end{aligned} \tag{8.10}$$

Accordingly, for $\boldsymbol{\sigma}\in\{0,1\}^{n}$, let

$$\Delta_{\boldsymbol{\sigma}}\triangleq\{\mathbf{p}\in\mathbb{R}^{n}\mid\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}\geq\mathbf{0}\},$$

be the set of external inputs $\mathbf{p}$ such that (8.5) has an equilibrium in $\Omega_{\boldsymbol{\sigma}}$, which is a closed convex polytopic cone. Also, note that h can be equivalently written in the piecewise-linear form

$$h(\mathbf{p})=\{(\mathbf{I}-\boldsymbol{\Sigma}\mathbf{W})^{-1}\boldsymbol{\Sigma}\mathbf{p}\mid\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}\geq 0,\boldsymbol{\Sigma}=\operatorname{diag}(\boldsymbol{\sigma}),\ \boldsymbol{\sigma}\in\{0,1\}^{n}\}. \tag{8.11}$$

Note that if $\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}\geq\mathbf{0}$ for exactly one $\boldsymbol{\sigma}\in\{0,1\}^{n}$, then a unique equilibrium exists according to (8.10). However, when $\mathbf{M}_{\boldsymbol{\sigma}_{\ell}}\mathbf{p}\geq\mathbf{0}$ for multiple $\boldsymbol{\sigma}_{\ell}\in\{0,1\}^{n},\ell\in\{1,\ldots,\bar{\ell}\}$, the network may

have either multiple equilibria or a unique one $\mathbf{x}^*_{\boldsymbol{\sigma}_1} = \cdots = \mathbf{x}^*_{\boldsymbol{\sigma}_{\bar{\ell}}}$ lying on the boundary between $\{\Omega_{\sigma_\ell}\}_{\ell=1}^{\bar{\ell}}$. The next result shows that the quantities $\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}$ can be used to distinguish between these two latter cases.

Lemma 8.3.2. ***(Existence of multiple equilibria).*** *Assume* $\mathbf{W}$ *satisfies Assumption 8.3.1,* $\mathbf{p} \in \mathbb{R}^n$ *is arbitrary, and* $\mathbf{M}_{\boldsymbol{\sigma}}$ *is defined as in* (8.10) *for* $\boldsymbol{\sigma} \in \{0,1\}^n$*. If there exist* $\boldsymbol{\sigma}_1 \neq \boldsymbol{\sigma}_2$ *such that* $\mathbf{p} \in \Delta_{\boldsymbol{\sigma}_1} \cap \Delta_{\boldsymbol{\sigma}_2}$*, then* $\mathbf{x}^*_{\boldsymbol{\sigma}_1} = \mathbf{x}^*_{\boldsymbol{\sigma}_2}$ *if and only if* $\mathbf{M}_{\boldsymbol{\sigma}_1}\mathbf{p} = \mathbf{M}_{\boldsymbol{\sigma}_2}\mathbf{p}$*.*

Proof. Clearly,

$$\begin{aligned} \mathbf{x}^*_{\boldsymbol{\sigma}_1} = \mathbf{x}^*_{\boldsymbol{\sigma}_2} &\Leftrightarrow \mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}_1} + \mathbf{p} = \mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}_2} + \mathbf{p} \\ &\Leftrightarrow (\mathbf{I} - \mathbf{W}\boldsymbol{\Sigma}_1)^{-1}\mathbf{p} = (\mathbf{I} - \mathbf{W}\boldsymbol{\Sigma}_2)^{-1}\mathbf{p}, \end{aligned} \tag{8.12}$$

where we have used (8.9). Since both $\mathbf{M}_{\boldsymbol{\sigma}_1}\mathbf{p}$ and $\mathbf{M}_{\boldsymbol{\sigma}_2}\mathbf{p}$ are nonnegative, (8.12) holds if and only if $\left((\mathbf{I} - \mathbf{W}\boldsymbol{\Sigma}_1)^{-1}\mathbf{p}\right)_i = \left((\mathbf{I} - \mathbf{W}\boldsymbol{\Sigma}_2)^{-1}\mathbf{p}\right)_i = 0$ for any i such that $\sigma_{1,i} \neq \sigma_{2,i}$, which is equivalent to $\mathbf{M}_{\boldsymbol{\sigma}_1}\mathbf{p} = \mathbf{M}_{\boldsymbol{\sigma}_2}\mathbf{p}$. □

Our next result provides an optimization-based condition for EUE that is both necessary and sufficient.

Proposition 8.3.3. ***(Optimization-based condition for EUE).*** *Let* $\mathbf{W}$ *satisfy Assumption 8.3.1 and* $\mathbf{M}_{\boldsymbol{\sigma}}$ *be as defined in* (8.10) *for* $\boldsymbol{\sigma} \in \{0,1\}^n$*. For* $\mathbf{p} \in \mathbb{R}^n$*, define* $\mu_1(\mathbf{p})$ *and* $\mu_2(\mathbf{p})$ *to be the largest and second largest elements of the set*

$$\left\{ \min_{i=1,\ldots,n} (\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p})_i \mid \boldsymbol{\sigma} \in \{0,1\}^n \right\},$$

respectively. Then, (8.5) *has a unique equilibrium for each* $\mathbf{p} \in \mathbb{R}^n$ *if and only if*

$$\max_{\|\mathbf{p}\|=1} \mu_1(\mathbf{p})\mu_2(\mathbf{p}) < 0. \tag{8.13}$$

Proof. First, note that $\mathbf{p} = \mathbf{0}$ is a degenerate case where the origin is the unique equilibrium belonging to all $\Omega_{\boldsymbol{\sigma}}$. For any $\mathbf{p} \neq \mathbf{0}$ and $\boldsymbol{\sigma} \in \{0,1\}^n$, $\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p} \geq \mathbf{0}$ if and only if $\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}/\|\mathbf{p}\| \geq \mathbf{0}$. Thus, EUE for all $\mathbf{p} \in \mathbb{R}^n$ and all $\|\mathbf{p}\| = 1$ are equivalent. The result then follows from Lemma 8.3.2 and the fact that, for any $\mathbf{p}$,

$$\begin{aligned} \mu_1(\mathbf{p})\mu_2(\mathbf{p}) < 0 &\Leftrightarrow \mu_1(\mathbf{p}) > 0 \text{ and } \mu_2(\mathbf{p}) < 0 \\ &\Leftrightarrow \exists \text{ unique } \boldsymbol{\sigma} \in \{0,1\}^n \quad \mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p} \geq \mathbf{0}. \end{aligned}$$

□

The optimization involved in (8.13) is usually highly non-convex. However, since the search space $\|\mathbf{p}\| = 1$ is compact, global search methods can be used to verify (8.13) numerically. Next, we give our main result regarding the EUE that not only is verifiable analytically but also provides insight into the class of matrices $\mathbf{W}$ that satisfy EUE.

Theorem 8.3.4. ***(Existence and uniqueness of equilibria).*** *Consider the network dynamics* (8.5) *and assume the weight matrix* $\mathbf{W}$ *satisfies Assumption 8.3.1. Then,* (8.5) *has a unique equilibrium for each* $\mathbf{p} \in \mathbb{R}^n$ *if* $\mathbf{I} - \mathbf{W} \in \mathcal{P}$.

Proof. The uniqueness can be shown as a corollary to both [74, Thm 5.3] and [75, Thm 2.2]. However, we provide in the following a novel proof technique based on (8.10) that establishes both existence and uniqueness. From Lemma 8.3.2, we know that, for any $\mathbf{p}$, (8.5) has a unique

equilibrium if and only if exactly one element of $\{\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p} \mid \boldsymbol{\sigma} \in \{0,1\}^n\}$ is nonnegative. To check this, we need to check whether

$$\exists \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \in \{0,1\}^n \quad \text{s.t.} \quad \mathbf{M}_{\boldsymbol{\sigma}_1}\mathbf{p} \geq \mathbf{0} \ \text{ and } \ \mathbf{M}_{\boldsymbol{\sigma}_2}\mathbf{p} \geq \mathbf{0}.$$

Note that this is equivalent to saying that there exist $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n_{\geq 0}$ such that $\mathbf{M}_{\boldsymbol{\sigma}_1}^{-1}\mathbf{x} = \mathbf{p} = \mathbf{M}_{\boldsymbol{\sigma}_2}^{-1}\mathbf{y}$. A more general question, which is relevant to our discussion, is whether

$$\mathbf{M}_{\boldsymbol{\sigma}_1}^{-1}\mathbf{x} = \mathbf{M}_{\boldsymbol{\sigma}_2}^{-1}\mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathcal{O}_n \setminus \{\mathbf{0}\}, \tag{8.14}$$

for *any* orthant $\mathcal{O}_n$ of $\mathbb{R}^n$ (including $\mathcal{O}_n = \mathbb{R}^n_{\geq 0}$ as a special case). Note that $\mathbf{x} = \mathbf{y} = \mathbf{0}$ is a trivial case and thus excluded. This is the question we analyze in the following. Notice that for any $\boldsymbol{\sigma} \in \{0,1\}^n$,

$$\mathbf{M}_{\boldsymbol{\sigma}}^{-1} = (\mathbf{I} - \mathbf{W}\boldsymbol{\Sigma})(2\boldsymbol{\Sigma} - \mathbf{I}) = (2\boldsymbol{\Sigma} - \mathbf{I}) - \mathbf{W}\boldsymbol{\Sigma}. \tag{8.15}$$

Since nodes can be relabeled arbitrarily, we can assume without loss of generality that $\boldsymbol{\sigma}_1 = [\mathbf{1}_{n_1}^T \ \mathbf{1}_{n_2}^T \ \mathbf{0}_{n_3}^T \ \mathbf{0}_{n_4}^T]^T$ and $\boldsymbol{\sigma}_2 = [\mathbf{1}_{n_1}^T \ \mathbf{0}_{n_2}^T \ \mathbf{1}_{n_3}^T \ \mathbf{0}_{n_4}^T]^T$ where $n_1, \ldots, n_4 \geq 0, \sum_{i=1}^4 n_i = n$. Then, it

follows from (8.15) that

$$\mathbf{M}_{\sigma_1}^{-1} = \left[\begin{array}{cc:cc} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} & \mathbf{0} & \mathbf{0} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} & \mathbf{0} & \mathbf{0} \\ \hdashline -\mathbf{W}_{31} & -\mathbf{W}_{32} & -\mathbf{I}_{n_3} & \mathbf{0} \\ -\mathbf{W}_{41} & -\mathbf{W}_{42} & \mathbf{0} & -\mathbf{I}_{n_4} \end{array}\right],$$

$$\mathbf{M}_{\sigma_2}^{-1} = \left[\begin{array}{cccc} \mathbf{I}_{n_1} - \mathbf{W}_{11} & \mathbf{0} & -\mathbf{W}_{13} & \mathbf{0} \\ -\mathbf{W}_{21} & -\mathbf{I}_{n_2} & -\mathbf{W}_{23} & \mathbf{0} \\ -\mathbf{W}_{31} & \mathbf{0} & \mathbf{I}_{n_3} - \mathbf{W}_{33} & \mathbf{0} \\ -\mathbf{W}_{41} & \mathbf{0} & -\mathbf{W}_{43} & -\mathbf{I}_{n_4} \end{array}\right],$$

where $\mathbf{W}_{ij}$'s are submatrices of $\mathbf{W}$ with appropriate dimensions. Taking the inverse of $\mathbf{M}_{\sigma_1}^{-1}$ as a 2-by-2 block-triangular matrix [76, Prop 2.8.7] (with the indicated blocks), we get

$$\mathbf{M}_{\sigma_1} = \left[\begin{array}{cc} \left[\begin{array}{cc} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} \end{array}\right]^{-1} & \mathbf{0} \\ -\left[\begin{array}{cc} \mathbf{W}_{31} & \mathbf{W}_{32} \\ \mathbf{W}_{41} & \mathbf{W}_{42} \end{array}\right]\left[\begin{array}{cc} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} \end{array}\right]^{-1} & -\mathbf{I}_{n_3+n_4} \end{array}\right],$$

so direct multiplication gives $\mathbf{M}_{\sigma_1}\mathbf{M}_{\sigma_2}^{-1} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{B}_3 & \mathbf{B}_4 \end{bmatrix}$, with

$$\mathbf{B}_1 = \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{W}_{11} & \mathbf{0} \\ -\mathbf{W}_{21} & -\mathbf{I}_{n_2} \end{bmatrix},$$

$$\mathbf{B}_2 = -\begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{W}_{13} & \mathbf{0} \\ \mathbf{W}_{23} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{B}_3 = -\begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \\ \mathbf{W}_{41} & \mathbf{W}_{42} \end{bmatrix} \mathbf{B}_1 + \begin{bmatrix} \mathbf{W}_{31} & \mathbf{0} \\ \mathbf{W}_{41} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{B}_4 = -\begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \\ \mathbf{W}_{41} & \mathbf{W}_{42} \end{bmatrix} \mathbf{B}_2 - \begin{bmatrix} \mathbf{I}_{n_3} - \mathbf{W}_{33} & \mathbf{0} \\ -\mathbf{W}_{43} & -\mathbf{I}_{n_4} \end{bmatrix}.$$

With this, after some computations one can show that

$$\mathbf{M}_{\sigma_1}\mathbf{M}_{\sigma_2}^{-1} = \begin{bmatrix} \mathbf{I}_{n_1} & \star & \star & \mathbf{0} \\ \mathbf{0} & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} \star & \star \\ \star & \star \end{bmatrix}$}} & \mathbf{0} \\ \mathbf{0} & & & \mathbf{0} \\ \mathbf{0} & \star & \star & \mathbf{I}_{n_4} \end{bmatrix}. \tag{8.16}$$

Let $\mathbf{\Gamma} \in \mathbb{R}^{(n_2+n_3)\times(n_2+n_3)}$ be the bracketed block in $\mathbf{M}_{\sigma_1}\mathbf{M}_{\sigma_2}^{-1}$. Similarly, let $\mathbf{x}_{23}, \mathbf{y}_{23} \in \mathbb{R}^{n_2+n_3}$ denote the corresponding $n_2 + n_3$-dimensional sub-vectors in the decomposition of $\mathbf{x}$ and $\mathbf{y}$, respectively.

Then, according to Lemma 8.A.1, (8.14) holds if and only if $\mathbf{y}_{23} = \mathbf{0}$ or

$$\mathbf{y}_{23}, \mathbf{\Gamma}\mathbf{y}_{23} \in \mathcal{O}_{n_2+n_3} \setminus \{\mathbf{0}\}, \tag{8.17}$$

for some orthant $\mathcal{O}_{n_2+n_3} \subset \mathbb{R}^{n_2+n_3}$. To check (8.17), define

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} \end{bmatrix}^{-1},$$

$$\mathbf{R} \triangleq \mathbf{I}_{n_3} - \mathbf{W}_{33} - \begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \end{bmatrix} \mathbf{Q} \begin{bmatrix} \mathbf{W}_{13} \\ \mathbf{W}_{23} \end{bmatrix}.$$

It can be shown that

$$\mathbf{\Gamma} = -\begin{bmatrix} \mathbf{Q}_{22} & \begin{bmatrix} \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{13} \\ \mathbf{W}_{23} \end{bmatrix} \\ -\begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{12} \\ \mathbf{Q}_{22} \end{bmatrix} & \mathbf{R} \end{bmatrix}.$$

Inverting the left-hand-side matrix as a 2-by-2 block matrix [76, Prop 2.8.7] (the first block is $\mathbf{Q}^{-1}$) and applying the matrix inversion lemma [76, Cor 2.8.8] to the first block of the result, we obtain

$$\left[\begin{array}{cc:c} \mathbf{I}_{n_1} - \mathbf{W}_{11} & -\mathbf{W}_{12} & -\mathbf{W}_{13} \\ -\mathbf{W}_{21} & \mathbf{I}_{n_2} - \mathbf{W}_{22} & -\mathbf{W}_{23} \\ \hdashline -\mathbf{W}_{31} & -\mathbf{W}_{32} & \mathbf{I}_{n_3} - \mathbf{W}_{33} \end{array}\right]^{-1} = \begin{bmatrix} \star & \star & \star \\ \star & \hat{\mathbf{B}}_1 & \hat{\mathbf{B}}_2 \\ \star & \hat{\mathbf{B}}_3 & \hat{\mathbf{B}}_4 \end{bmatrix}$$

where

$$\hat{\mathbf{B}}_1 = \mathbf{Q}_{22} + \begin{bmatrix} \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{13} \\ \mathbf{W}_{23} \end{bmatrix} \mathbf{R}^{-1} \begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{12} \\ \mathbf{Q}_{22} \end{bmatrix},$$

$$\hat{\mathbf{B}}_2 = \begin{bmatrix} \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{W}_{13} \\ \mathbf{W}_{23} \end{bmatrix} \mathbf{R}^{-1},$$

$$\hat{\mathbf{B}}_3 = \mathbf{R}^{-1} \begin{bmatrix} \mathbf{W}_{31} & \mathbf{W}_{32} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_{12} \\ \mathbf{Q}_{22} \end{bmatrix}, \qquad \hat{\mathbf{B}}_4 = \mathbf{R}^{-1}.$$

Therefore, $-\mathbf{\Gamma}$ is the principal pivot transform of $\begin{bmatrix} \hat{\mathbf{B}}_1 & \hat{\mathbf{B}}_2 \\ \hat{\mathbf{B}}_3 & \hat{\mathbf{B}}_4 \end{bmatrix}$. Since $\mathbf{I} - \mathbf{W} \in \mathcal{P}$, Lemma 2.3.2(v) guarantees that $-\mathbf{\Gamma} \in \mathcal{P}$, which by Lemma 2.3.2(iv) implies that for every nonzero $\mathbf{y}_{23}$ there exists an index $k \in \{1, \ldots, n_2 + n_3\}$ such that $(\mathbf{y}_{23})_k (\mathbf{\Gamma} \mathbf{y}_{23})_k < 0$, i.e., there does not exist any $\mathbf{y}_{23} \neq \mathbf{0}$ such that (8.17) holds, in which case, by (8.16), $\mathbf{M}_{\sigma_1} \mathbf{M}_{\sigma_2}^{-1} \mathbf{y} = \mathbf{y}$. Therefore, we have shown that, for any $\mathbf{y} \in \mathbb{R}^n$,

$$\mathbf{y}, \mathbf{M}_{\sigma_1} \mathbf{M}_{\sigma_2}^{-1} \mathbf{y} \in \mathcal{O}_n \text{ for some orthant } \mathcal{O}_n \Leftrightarrow \mathbf{y} = \mathbf{M}_{\sigma_1} \mathbf{M}_{\sigma_2}^{-1} \mathbf{y} \Leftrightarrow \mathbf{y}_{23} = \mathbf{0}.$$

Recalling that $\mathbf{y} = \mathbf{M}_{\sigma_2} \mathbf{p}$, we have

$$\begin{aligned} \mathbf{M}_{\sigma_1} \mathbf{p}, \mathbf{M}_{\sigma_2} \mathbf{p} \in \mathcal{O}_n \text{ for some orthant } \mathcal{O}_n &\Leftrightarrow \mathbf{M}_{\sigma_1} \mathbf{p} = \mathbf{M}_{\sigma_2} \mathbf{p} \\ &\Leftrightarrow \Big((\mathbf{M}_{\sigma_2} \mathbf{p})_i = 0 \quad \forall i \text{ s.t. } \sigma_{1,i} \neq \sigma_{2,i} \Big). \end{aligned} \tag{8.18}$$

Using Lemma 8.3.2, the first equivalence in (8.18) for the particular case of $\mathcal{O}_n = \mathbb{R}^n_{\geq 0}$ shows the

uniqueness of equilibrium. To prove existence, consider the following assignment procedure. For any $\boldsymbol{\sigma}_0 \in \{0,1\}^n$, let $\mathcal{O}_n$ be the orthant that contains $\mathbf{M}_{\boldsymbol{\sigma}_0}\mathbf{p}$ (if there are more than one such orthants, we pick one arbitrarily). Define

$$\mathcal{S}_{\mathcal{O}_n} = \{\boldsymbol{\sigma} \in \{0,1\}^n \mid \mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p} \in \mathcal{O}_n\}.$$

If $\mathcal{S}_{\mathcal{O}_n} = \{\boldsymbol{\sigma}_0\}$, we assign $\boldsymbol{\sigma}_0$ to $\mathcal{O}_n$. If $|\mathcal{S}_{\mathcal{O}_n}| > 1$, it follows from (8.18) that

(i) $\mathbf{M}_{\boldsymbol{\sigma}}\mathbf{p}$ is the same for all $\boldsymbol{\sigma} \in \mathcal{S}_{\mathcal{O}_n}$; let $\mathbf{x}^*_{\mathcal{O}_n}$ be this shared value,

(ii) $|\mathcal{S}_{\mathcal{O}_n}| = 2^k$ for some $k \geq 1$,

(iii) there exists $\mathcal{I}_{\mathcal{O}_n} \subseteq \{1,\ldots,n\}$ with $|\mathcal{I}_{\mathcal{O}_n}| = k$ such that for any $i \notin \mathcal{I}_{\mathcal{O}_n}$, σ_i is the same for all $\boldsymbol{\sigma} \in \mathcal{S}_{\mathcal{O}_n}$,

(iv) $(\mathbf{x}^*_{\mathcal{O}_n})_i = 0$ for all $i \in \mathcal{I}_{\mathcal{O}_n}$.

Due to (iv), $\mathbf{x}^*_{\mathcal{O}_n}$ belongs to the intersection of 2^k orthants ($\mathcal{O}_n$ and $2^k - 1$ other orthants that differ from $\mathcal{O}_n$ along the dimensions in $\mathcal{I}_{\mathcal{O}_n}$). Therefore, we can assign each of the 2^k elements of $\mathcal{S}_{\mathcal{O}_n}$ to its corresponding orthant from the 2^k orthants that contain $\mathbf{x}^*_{\mathcal{O}_n}$. By repeating this procedure for all unassigned $\boldsymbol{\sigma}_0 \in \{0,1\}^n$, we can uniquely assign every $\boldsymbol{\sigma}_0 \in \{0,1\}^n$ to an orthant in $\mathbb{R}^n$ such that no two are assigned to the same orthant. Therefore, one $\boldsymbol{\sigma} \in \{0,1\}^n$ must be assigned to the positive orthant $\mathbb{R}^n_{\geq 0}$, showing the existence of an equilibrium and completing the proof. □

The fact that the condition in Lemma 2.3.2(iv) is an *equivalent* characterization of P-matrices suggests that the sufficient condition of Theorem 8.3.4 is tight. Indeed, extensive simulations with random matrices did not reveal any instance of $\mathbf{W}$ that is not a P-matrix but for which (8.5)

has a unique equilibrium for all $\mathbf{p} \in \mathbb{R}^n$ (where we checked the latter using Proposition 8.3.3). This leads us to the following conjecture, which was also made in [21] as a claim without proof.

Conjecture 8.3.5. ***(Necessity of $\mathbf{I} - \mathbf{W} \in \mathcal{P}$ for EUE).*** *Assume the weight matrix* $\mathbf{W}$ *satisfies Assumption 8.3.1. Then,* (8.5) *has a unique equilibrium for all* $\mathbf{p} \in \mathbb{R}^n$ *if* and only if $\mathbf{I} - \mathbf{W} \in \mathcal{P}$.□

Remark 8.3.6. ***(Computational complexity of verifying $\mathbf{I} - \mathbf{W} \in \mathcal{P}$).*** Although the problem of determining whether a matrix is in $\mathcal{P}$ is straightforward for small n, it is known to be co-NP-complete [77], and thus expensive for large networks. Indeed, [78] shows that all the 2^n principal minors of $\mathbf{A}$ have to be checked to prove $\mathbf{A} \in \mathcal{P}$ (though disproving $\mathbf{A} \in \mathcal{P}$ is usually much easier). In these cases, one may need to rely on more conservative sufficient conditions such as $\rho(|\mathbf{W}|) < 1$ or $\|\mathbf{W}\| < 1$ (cf. Lemma 2.3.3) to establish $\mathbf{I} - \mathbf{W} \in \mathcal{P}$. These conditions, moreover, have the added benefit of providing intuitive connections between the distribution of synaptic weights, network size, and stability. We elaborate more on this point in Section 8.4.3. □

Example 8.3.7. ***(Uniform excitatory-inhibitory networks).*** Consider a network of n nodes in which $\alpha n, \alpha \in (0, 1)$ are excitatory, $(1 - \alpha)n$ are inhibitory, and the synaptic weight between any pair of nodes only depends on their type (the synaptic weight of any inhibitory-to-excitatory connection is $w_{ei} < 0$, and similarly for $w_{ee} > 0, w_{ie} > 0, w_{ii} < 0$). Also, assume common external inputs $p_e, p_i \in \mathbb{R}$ for all excitatory and inhibitory nodes, respectively. Let $x_e(t)$ and $x_i(t)$ be the average firing rates of excitatory and inhibitory nodes, respectively. Then,

$$\tau \begin{bmatrix} \dot{x}_e \\ \dot{x}_i \end{bmatrix} = -\begin{bmatrix} x_e \\ x_i \end{bmatrix} + \left[\begin{bmatrix} \alpha n w_{ee} & (1-\alpha) n w_{ei} \\ \alpha n w_{ie} & (1-\alpha) n w_{ii} \end{bmatrix} \begin{bmatrix} x_e \\ x_i \end{bmatrix} + \begin{bmatrix} p_e \\ p_i \end{bmatrix}\right]^+ .$$

This simplification of n-dimensional networks to planar dynamics is commonly known as the

Wilson-Cowan model [79] and is a widely used model in computational neuroscience, see e.g., [80, 81]. Let $\mathbf{W}_{EI} \in \mathbb{R}^{2\times 2}$ be the corresponding weight matrix above. One can check that

$$\mathbf{I} - \mathbf{W}_{EI} \in \mathcal{P} \Leftrightarrow \alpha n w_{ee} < 1,$$

and

$$\rho(|\mathbf{W}_{EI}|) < 1 \Leftrightarrow \alpha n w_{ee} < 1,\ (1-\alpha) n |w_{ii}| < 1, \text{ and}$$
$$\alpha(1-\alpha) n^2 w_{ie} |w_{ei}| < (1 - \alpha n w_{ee})(1 - (1-\alpha) n |w_{ii}|).$$

Thus, according to Theorem 8.3.4, EUE only requires the excitatory dynamics to be stable (note that w_{ee} has to be smaller as n grows), while the more conservative $\rho(|\mathbf{W}_{EI}|) < 1$ requires two extra conditions: the stability of inhibitory dynamics and a weak interconnection between excitatory and inhibitory subnetworks. □

8.3.3 Asymptotic Stability

The EUE, as discussed above, is an *opportunity* to shape the network state, provided the equilibrium corresponds to a desired state (e.g., a memory, the encoding of a spatial location, or eye position) *and* it attracts network trajectories [82–86]. Here we investigate when the latter holds, i.e., the network equilibrium is asymptotically stable. Our main result on asymptotic stability is the following.

Theorem 8.3.8. ***(Asymptotic stability).*** *Consider the network dynamics* (8.5) *and assume* $\mathbf{W}$ *satisfies Assumption 8.3.1.*

(i) [Sufficient condition] If $\mathbf{W} \in \mathcal{L}$*, then for all* $\mathbf{p} \in \mathbb{R}^n$*, the network is globally exponentially stable (GES) relative to a unique equilibrium* $\mathbf{x}^*$*;*

(ii) [Necessary condition] If for all $\mathbf{p} \in \mathbb{R}^n$ *the network is locally asymptotically stable relative to a unique equilibrium* $\mathbf{x}^*$*, then* $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$.

Proof. (i) The EUE follows from Lemma 2.3.3(iii)&(iv) and Theorem 8.3.4. GES follows from [87, Thm 1], but we give a simpler proof here for completeness. Consider an arbitrary trajectory $\mathbf{x}(t)$ of (8.5) and define

$$\boldsymbol{\xi}(t) = \mathbf{x}(t) - \mathbf{x}^*.$$

After some manipulations, one can show that

$$\tau\dot{\boldsymbol{\xi}}(t) = (-\mathbf{I} + \mathbf{M}(t)\mathbf{W})\boldsymbol{\xi}(t), \tag{8.19}$$

where $\mathbf{M}(t)$ is a diagonal matrix with diagonal entries

$$m_{ii}(t) \triangleq \begin{cases} \frac{[\mathbf{W}_i\mathbf{x}(t)+p_i]^+ - [\mathbf{W}_i\mathbf{x}^*+p_i]^+}{\mathbf{W}_i\boldsymbol{\xi}(t)} & \text{if } \mathbf{W}_i\boldsymbol{\xi}(t) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Since the function $[\cdot]^+$ is monotonically increasing and Lipschitz with constant 1, $m_{ii}(t) \in [0, 1]$ for all $i \in \{1, \ldots, n\}$. Thus, $\mathbf{M}(t)$ belongs to the convex hull of $\{\boldsymbol{\Sigma}\}_{\sigma\in\{0,1\}^n}$ for all $t \geq 0$. Let

$(\alpha_{\boldsymbol{\sigma}}(t))_{\boldsymbol{\sigma}\in\{0,1\}^n}$ be a convex combination such that

$$\mathbf{M}(t) = \sum_{\boldsymbol{\sigma}\in\{0,1\}^n} \alpha_{\boldsymbol{\sigma}}(t)\boldsymbol{\Sigma}, \quad t \geq 0.$$

By assumption, there exists $\mathbf{P} = \mathbf{P}^T > \mathbf{0}$ and $\lambda > 0$ such that

$$(-\mathbf{I} + \mathbf{W}^T\boldsymbol{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}) \leq -\lambda\mathbf{I}, \quad \boldsymbol{\sigma} \in \{0,1\}^n.$$

Therefore, the evolution of the Lyapunov function $V(\boldsymbol{\xi}) = \boldsymbol{\xi}^T\mathbf{P}\boldsymbol{\xi}$ along (8.19) satisfies

$$\begin{aligned}
\tau\frac{dV(\boldsymbol{\xi}(t))}{dt} &= \boldsymbol{\xi}^T[(-\mathbf{I} + \mathbf{W}^T\mathbf{M}(t))\mathbf{P} + \mathbf{P}(-\mathbf{I} + \mathbf{M}(t)\mathbf{W})]\boldsymbol{\xi} \\
&= \boldsymbol{\xi}^T\Big[\sum_{\boldsymbol{\sigma}\in\{0,1\}^n}(-\mathbf{I} + \alpha_{\boldsymbol{\sigma}}(t)\mathbf{W}^T\boldsymbol{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \alpha_{\boldsymbol{\sigma}}(t)\boldsymbol{\Sigma}\mathbf{W})\Big]\boldsymbol{\xi} \\
&= \boldsymbol{\xi}^T\Big[\sum_{\boldsymbol{\sigma}\in\{0,1\}^n}\alpha_{\boldsymbol{\sigma}}(t)[(-\mathbf{I} + \mathbf{W}^T\boldsymbol{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W})]\Big]\boldsymbol{\xi} \\
&\leq -\lambda\|\boldsymbol{\xi}\|^2 \leq -\frac{\lambda}{\rho(\mathbf{P})}V(\boldsymbol{\xi}(t)),
\end{aligned}$$

proving GES.

(ii) Assume, by contradiction, that $-\mathbf{I} + \mathbf{W} \notin \mathcal{H}$, which means that there exists $\boldsymbol{\sigma} \in \{0,1\}^n$ such that $-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}$ is not Hurwitz. Then, consider the choice

$$\mathbf{p} = (2\mathbf{I} - \mathbf{W})\boldsymbol{\sigma} - \mathbf{1}_n.$$

It is straightforward to show that $\mathbf{x}^* = \boldsymbol{\sigma}$ is an equilibrium point for (8.5) lying in the interior of $\Omega_{\boldsymbol{\sigma}}$. By assumption, $\mathbf{x}^*$ is (unique and) locally asymptotically stable, which contradicts $-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}$

not being Hurwitz. This completes the proof. □

Similar to the problem of verifying whether a matrix is a P-matrix, cf. Remark 8.3.6, verifying total-Hurwitzness becomes computationally expensive for large n. The next result gives a usually more conservative but computationally inexpensive alternative.

Proposition 8.3.9. ***(Computationally feasible sufficient conditions for GES).*** *Consider the network dynamics* (8.5) *and assume the weight matrix* $\mathbf{W}$ *satisfies Assumption 8.3.1. If* $\rho(|\mathbf{W}|) < 1$ *or* $\|\mathbf{W}\| < 1$*, then for all* $\mathbf{p} \in \mathbb{R}^n$*, the network has a unique equilibrium* $\mathbf{x}^*$ *and it is GES relative to* $\mathbf{x}^*$.

Proof. If $\|\mathbf{W}\| < 1$, the result follows from Lemma 2.3.3(ii) and Theorem 8.3.8. For the case $\rho(|\mathbf{W}|) < 1$, however, the claim follows from the argument in [22, Prop. 3], which we bring here for completeness. For simplicity, assume that $\mathbf{W}$ is irreducible, i.e., the network topology contains a path from any node to any other (this assumption is without loss of generality since, if $\mathbf{W}$ is not irreducible, the continuity of the spectral radius guarantees the existence of $\epsilon > 0$ such that $\rho(|\mathbf{W}| + \epsilon \mathbf{1}_n \mathbf{1}_n^T) < 1$. The same argument can then be employed for $|\mathbf{W}| + \epsilon \mathbf{1}_n \mathbf{1}_n^T$). By the Perron-Frobenius Theorem [76, Fact 4.11.4], there exists $\boldsymbol{\alpha} \in \mathbb{R}^n_{>0}$ such that $\boldsymbol{\alpha}^T |\mathbf{W}| = \rho(|\mathbf{W}|)\boldsymbol{\alpha}^T$. The map $\|\cdot\|_{\boldsymbol{\alpha}} : \mathbf{v} \to \|\mathbf{v}\|_{\boldsymbol{\alpha}} \triangleq \boldsymbol{\alpha}^T |\mathbf{v}|$ is indeed a norm on $\mathbb{R}^n$. To show EUE, note that for any $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$,

$$\begin{aligned}
\|[\mathbf{W}\mathbf{y} + \mathbf{p}]^+ - [\mathbf{W}\mathbf{z} + \mathbf{p}]^+\|_{\boldsymbol{\alpha}} &= \boldsymbol{\alpha}^T |[\mathbf{W}\mathbf{y} + \mathbf{p}]^+ - [\mathbf{W}\mathbf{z} + \mathbf{p}]^+| \\
&\leq \boldsymbol{\alpha}^T |\mathbf{W}(\mathbf{y} - \mathbf{z})| \leq \boldsymbol{\alpha}^T |\mathbf{W}||\mathbf{y} - \mathbf{z}| = \rho(|\mathbf{W}|)\boldsymbol{\alpha}^T |\mathbf{y} - \mathbf{z}| \\
&= \rho(|\mathbf{W}|)\|\mathbf{y} - \mathbf{z}\|_{\boldsymbol{\alpha}},
\end{aligned}$$

so $\mathbf{x} \mapsto [\mathbf{W}\mathbf{x} + \mathbf{p}]^+$ is a contraction on $\mathbb{R}^n_{\geq 0}$ and has a unique fixed point by the Banach Fixed-Point

Theorem, denoted $\mathbf{x}^*$.

To show GES, let $t \mapsto \mathbf{x}(t)$ be an arbitrary trajectory and consider $\boldsymbol{\xi}(t) \triangleq (\mathbf{x}(t) - \mathbf{x}^*)e^t$. We have

$$\tau\dot{\boldsymbol{\xi}}(t) = \mathbf{M}(t)\mathbf{W}\boldsymbol{\xi}(t), \tag{8.20}$$

where $\mathbf{M}(t)$ is the same as in (8.19). Then, by using [21, Lemma] (which is essentially a careful application of Gronwall-Bellman's Inequality [62, Lemma A.1] to (8.20)), we get

$$\|\boldsymbol{\xi}(t)\|_\alpha \leq \|\boldsymbol{\xi}(0)\|_\alpha e^{\rho(|\mathbf{W}|)t} \Rightarrow \|\mathbf{x}(t) - \mathbf{x}^*\|_\alpha \leq \|\mathbf{x}(0) - \mathbf{x}^*\|_\alpha e^{-(1-\rho(|\mathbf{W}|))t},$$

which gives GES by the equivalence of norms on $\mathbb{R}^n$. □

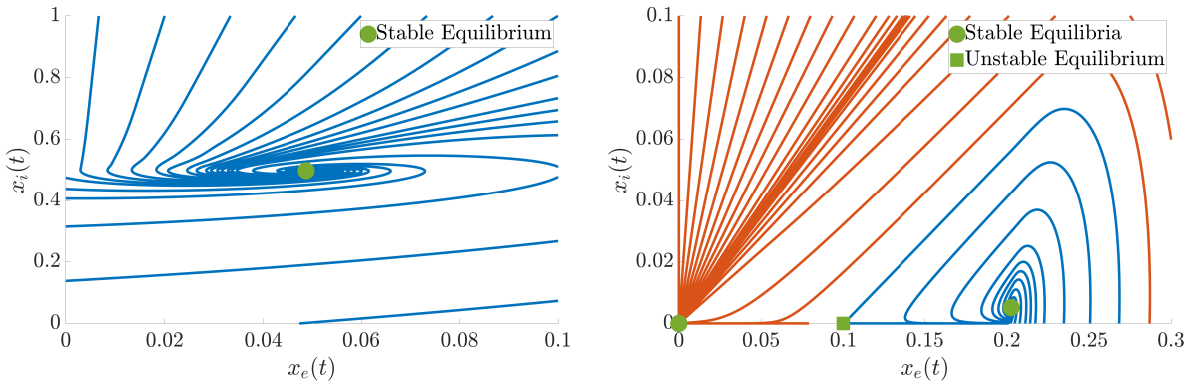
From Lemma 2.3.3(iii), the conditions of Theorem 8.3.8 and Proposition 8.3.9 are not conclusive when $\mathbf{W}$ satisfies $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$ but neither $\mathbf{W} \in \mathcal{L}$ nor $\rho(|\mathbf{W}|) < 1$. However,

(i) If a unique equilibrium $\mathbf{x}^*$ lies in the interior of a switching region Ω_σ (a condition that can be shown to hold for Lebesgue-almost all $\mathbf{p}$), then $\mathbf{x}^*$ is at least locally exponentially stable.

(ii) In our extensive simulations with random $(\mathbf{W}, \mathbf{p})$, any system satisfying $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$ was GES for all $\mathbf{p}$.

These observations lead us to the following conjecture, whose analytic characterization remains an open problem.

Conjecture 8.3.10. ***(Sufficiency of total-Hurwitzness for GES).*** *Consider the dynamics* (8.5) *and assume* $\mathbf{W}$ *satisfies Assumption 8.3.1. The network has a unique GES equilibrium for all* $\mathbf{p} \in \mathbb{R}^n$ *If*



(a) (b)

Figure 8.2: Network trajectories for the excitatory-inhibitory network of Example 8.3.11. a) When $\mathbf{W}_{EI} = [0.9, -2; 5, -1.5]$, $\mathbf{p}_{EI} = [1; 1]$, network has a unique GES equilibrium. b) However, for $\mathbf{W}_{EI} = [1.1, -2; 5, -1.5]$, $\mathbf{p}_{EI} = [-0.01; -1]$, the network exhibits bi-stable behavior. The colors of the trajectories correspond to the equilibria to which they converge. Note that although $\alpha n w_{ee} > 1$, the network is GES for most values of $\mathbf{p}_{EI}$, so we used Proposition 8.3.3 for finding a $\mathbf{p}_{EI}$ that leads to multi-stability.

and only if $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$. □

We next study the GES of the uniform excitatory-inhibitory networks of Example 8.3.7.

Example 8.3.11. ***(Uniform excitatory-inhibitory networks, cont'd).*** Consider again the excitatory-inhibitory network of Example 8.3.7. One can verify that

$$-\mathbf{I} + \mathbf{W}_{EI} \in \mathcal{H} \Leftrightarrow \alpha n w_{ee} < 1. \tag{8.21}$$

Thus, the (sufficient) condition for EUE and (necessary) condition for GES coincide, and they interestingly only restrict w_{ee} while w_{ei}, w_{ie}, and w_{ii} are completely free. Figure 8.2 shows sample phase portraits for the cases $\alpha n w_{ee} < 1$ and $\alpha n w_{ee} > 1$. □

8.3.4 Boundedness of Solutions

Here we study the boundedness of solutions under the network dynamics (8.2). While our discussion so far has been about the dynamics (8.5) (with constant $\mathbf{p}$), we switch for the remainder of this section to (8.2) for the sake of generality, as the same results are applicable without major modifications. Note that in reality, the firing rate of any neuron is bounded by a maximum rate dictated by its refractory period (the minimum inter-spike duration). Unboundedness of solutions in the model corresponds in practice to the so-called "run-away" excitations where the firing of neurons grow beyond sustainable rates for prolonged periods of time, which is neither desirable nor safe [88]. Since GES implies boundedness of solutions, any condition that is sufficient for GES is also sufficient for boundedness. However, boundedness of solutions can be guaranteed under less restrictive conditions. The next result shows that inhibition, overall, preserves boundedness.

Lemma 8.3.12. ***(Inhibition preserves boundedness).*** *Let $t \mapsto \mathbf{x}(t)$ be the solution of* (8.2) *starting from initial state $\mathbf{x}(0) = \mathbf{x}_0$. Consider the system*

$$\tau \dot{\bar{\mathbf{x}}}(t) = -\bar{\mathbf{x}}(t) + [[\mathbf{W}]^+ \bar{\mathbf{x}}(t) + \mathbf{p}(t)]^+, \quad \bar{\mathbf{x}}(0) = \mathbf{x}_0. \tag{8.22}$$

Then, $\mathbf{x}(t) \leq \bar{\mathbf{x}}(t)$ for all $t \geq 0$.

Proof. Since $\mathbf{x}(t) \geq \mathbf{0}$ for all t, we can write (8.2) as

$$\tau \dot{\mathbf{x}}(t) = -\mathbf{x}(t) + [[\mathbf{W}]^+ \mathbf{x}(t) + \mathbf{p}(t) + \boldsymbol{\delta}(t)]^+, \tag{8.23}$$

where $\boldsymbol{\delta}(t) \triangleq (\mathbf{W} - [\mathbf{W}]^+)\mathbf{x}(t) \leq \mathbf{0}$. Since the vector field $(\mathbf{x}, t) \mapsto -\mathbf{x} + [[\mathbf{W}]^+ \mathbf{x} + \mathbf{p}(t)]^+$ is

quasi-monotone nondecreasing[4], the result follows by using the monotonicity of the function $[\cdot]^+$ and applying the vector-valued extension of the Comparison Principle given in [89, Lemma 3.4] to (8.22) and (8.23). □

While the result about preservation of boundedness under inhibition in Lemma 8.3.12 is intuitive, one must interpret it carefully: it is *not* in general true that adding inhibition to any dynamics (8.2) can only decrease $\mathbf{x}(t)$. This is only true if the network vector field is quasi-monotone nondecreasing, as is the case with the excitatory-only dynamics (8.22). Intuitively, this is because, if the network has inhibitory nodes, adding inhibition to their input can in turn "disinhibit" and increase the activity of the rest of the network.

The next result identifies a condition on the excitatory part of the dynamics to determine if trajectories are bounded.

Theorem 8.3.13. ***(Boundedness).*** *Consider the network dynamics* (8.2)*. If the corresponding excitatory-only dynamics* (8.22) *has bounded trajectories, the trajectories of* (8.2) *are also bounded by the same bound as those of* (8.22)*.*

The proof of this result follows from Lemma 8.3.12 and is therefore omitted. The following result, similar to Proposition 8.3.9, provides a more conservative but computationally feasible test for boundedness.

Corollary 8.3.14. ***(Boundedness).*** *Consider the network dynamics* (8.2) *and assume that* $\mathbf{p}(t)$ *is bounded, i.e., there exists* $\bar{\mathbf{p}} \in \mathbb{R}^n_{>0}$ *such that* $\mathbf{p}(t) \leq \bar{\mathbf{p}}, t \geq 0$*. If* $\rho([\mathbf{W}]^+) < 1$*, then the network trajectories remain bounded for all* $t \geq 0$*.*

[4]A vector field $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is *quasi-monotone nondecreasing* [89] if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any $i \in \{1, \ldots, n\}$,

$$\left(x_i = y_i \text{ and } x_j \leq y_j \text{ for all } j \neq i\right) \Rightarrow f(\mathbf{x}, t) \leq f(\mathbf{y}, t).$$

Proof. If $\mathbf{p}(t)$ is constant, the result follows from Theorem 8.3.13 and Proposition 8.3.9. If $\mathbf{p}(t)$ is not constant, the same argument proves boundedness of trajectories for the dynamics

$$\tau\dot{\bar{\mathbf{x}}}(t) = -\bar{\mathbf{x}}(t) + [[\mathbf{W}]^+\bar{\mathbf{x}}(t) + \bar{\mathbf{p}}]^+, \quad \bar{\mathbf{x}}(0) = \mathbf{x}_0. \tag{8.24}$$

The result then follows from the quasi-monotonicity of $(\mathbf{x}, t) \mapsto -\mathbf{x} + [\mathbf{W}^+\mathbf{x} + \bar{\mathbf{p}}]_+$, similar to Lemma 8.3.12. □

Example 8.3.15. ***(Uniform excitatory-inhibitory networks, cont'd).*** Consider again the excitatory-inhibitory network of Example 8.3.7. Clearly, the corresponding excitatory-only dynamics have bounded trajectories if and only if

$$\rho([\mathbf{W}_{EI}]^+) < 1 \Leftrightarrow \alpha n w_{ee} < 1, \tag{8.25}$$

which is the same condition as (8.21). However, an exhaustive inspection of the switching regions $\{\Omega_\sigma\}_\sigma$ and the eigenvalues of $\{-\mathbf{I} + \mathbf{\Sigma W}\}_\sigma$ reveals that (8.25) can be relaxed to

$$-\mathbf{I} + \mathbf{W} \text{ be Hurwitz} \Leftrightarrow \begin{cases} (1 - \alpha n w_{ee}) + (1 - (1 - \alpha) n w_{ii}) > 0, \text{ and} \\ (1 - \alpha n w_{ee})(1 - (1 - \alpha) n w_{ii}) > \alpha(1 - \alpha) n^2 w_{ie} w_{ei}, \end{cases}$$

showing that there is room for sharpening Theorem 8.3.13. □

Remark 8.3.16. ***(Comparison with the literature).*** In this section, we have provided a comprehensive list of conditions that both extend and simplify the state of the art on stability of dynamically isolated linear-threshold networks. Regarding network equilibria, we have extended [74, Thm 5.3]

to guarantee both existence and uniqueness when $\mathbf{I} - \mathbf{W} \in \mathcal{P}$ (Theorem 8.3.4) and provided a novel optimization-based if and only if condition for EUE (Proposition 8.3.3). On exponential stability, Theorem 8.3.8 gives a simpler proof than [87, Thm 1] for the sufficiency of $\mathbf{W} \in \mathcal{L}$ and a novel proof for the necessity of $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$. Finally, our result on boundedness of trajectories (Theorem 8.3.13) extends Corollary 8.3.14 (also available in [27, Thm 1]) to a much wider class of networks by exploiting the quasi-monotonicity of excitatory-only dynamics. □

Building on our understanding of single-layer dynamics, we next analyze multilayer dynamics beginning with mechanisms of selective inhibition.

8.4 Selective Inhibition in Bilayer Networks

Here, we study selective inhibition in bilayer networks as a building block towards the understanding of hierarchical selective recruitment in multilayer networks. With respect to the model described in Section 8.2, we consider two layers ($N = 2$), where the dynamics of the lower layer $\mathcal{N}_2$ is described by (8.2) and the dynamics of the upper layer $\mathcal{N}_1$ is temporarily arbitrary (for generality). Our goal is to study the selective inhibition of $\mathcal{N}_2^o$ via the input that it receives from $\mathcal{N}_1$.

As pointed out in Section 8.2, when a group of neurons are inhibited, their activity is substantially decreased, ideally such that their net input (their respective component of $\mathbf{W}\mathbf{x}(t) + \mathbf{p}(t)$) becomes negative and their firing rate decays exponentially to zero. Therefore, the problem of selective inhibition is equivalent to the exponential stabilization of the nodes $\mathcal{N}_2^o$ to the origin. To this end, we decompose $\mathbf{p}(t)$ as

$$\mathbf{p}(t) = \mathbf{B}\mathbf{u}(t) + \tilde{\mathbf{p}}. \tag{8.26}$$

The role of $\mathbf{u}(t) \in \mathbb{R}^m_{\geq 0}$ is to stabilize $\mathcal{N}_2^{\circ}$ to the origin while the role of $\tilde{\mathbf{p}} \in \mathbb{R}^n$ is to shape the activity of $\mathcal{N}_2^{\shortmid}$ by determining its equilibrium For the purpose of this section, we assume $\tilde{\mathbf{p}}$ is given and constant.

Let $r \leq n$ be the size of $\mathcal{N}_2^{\circ}$. We partition $\mathbf{x}$, $\mathbf{W}$, $\mathbf{B}$, and $\tilde{\mathbf{p}}$ similar to (8.4), i.e.,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{\circ} \\ \mathbf{x}^{\shortmid} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}^{\circ\circ} & \mathbf{W}^{\circ\shortmid} \\ \mathbf{W}^{\shortmid\circ} & \mathbf{W}^{\shortmid\shortmid} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}^{\circ} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{p}} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{p}}^{\shortmid} \end{bmatrix}, \tag{8.27}$$

where $\mathbf{W}^{\circ\circ} \in \mathbb{R}^{r\times r}, \mathbf{B}^{\circ} \in \mathbb{R}^{r\times m}_{\leq 0}$ is nonpositive to deliver inhibition, and $\tilde{\mathbf{p}}^{\shortmid} \in \mathbb{R}^{n-r}$. The first r rows of $\mathbf{B}$ are nonzero to allow for the inhibition of $\mathcal{N}_2^{\circ}$ while the remaining $n-r$ rows are zero to make this inhibition selective to $\mathcal{N}_2^{\circ}$. The sparsity of the entries of $\tilde{\mathbf{p}}$ is opposite to the rows of $\mathbf{B}$ due to the complementary roles of $\mathbf{Bu}(t)$ and $\tilde{\mathbf{p}}$.

The mechanisms of inhibition in the brain are broadly divided [90] into two categories, feedforward and feedback, based on how the signal $\mathbf{u}(t)$ is determined. In the following, we separately study each scenario, analyzing the interplay between the corresponding mechanism and network structure. We will later combine both mechanisms when we discuss the complete HSR framework in [91], as natural selective inhibition is not purely feedback or feedforward.

8.4.1 Feedforward Selective Inhibition

Feedforward inhibition [90] refers to the scenario where $\mathcal{N}_1$ provides an inhibition based on its own "desired" activity/inactivity pattern for $\mathcal{N}_2$ and irrespective of the current state of $\mathcal{N}_2$. This is indeed possible if the inhibition is sufficiently strong, as excessive inhibition has no effect on nodal dynamics due to the thresholding in $[\cdot]^+$. However, this independence from the activity

level of $\mathcal{N}_2$ requires some form of guaranteed boundedness, as defined next.

Definition 8.4.1. ***(Monotone boundedness).*** The dynamics (8.2) is *monotonically bounded* if for any $\bar{\mathbf{p}} \in \mathbb{R}^n$ there exists $\boldsymbol{\nu}(\bar{\mathbf{p}}) \in \mathbb{R}^n_{\geq 0}$ such that $\mathbf{x}(t) \leq \boldsymbol{\nu}(\bar{\mathbf{p}}), t \geq 0$ for any $\mathbf{p}(t) \leq \bar{\mathbf{p}}, t \geq 0$. □

From Lemma 8.3.12 and Proposition 8.3.9, (8.2) is monotonically bounded if $\rho([\mathbf{W}]^+) < 1$ and the initial condition $\mathbf{x}_0$ is restricted to a bounded domain. Also in reality, the state of any biological neuronal network is uniformly bounded due to the refractory period of its neurons, implying monotone boundedness. The next result shows that the GES of $\mathcal{N}_2^{\text{1}}$ is both necessary and sufficient for feedforward selective inhibition.

Theorem 8.4.2. ***(Feedforward selective inhibition).*** *Consider the dynamics* (8.2)*, where the external input is given by* (8.26)-(8.27) *with a constant feedforward control*

$$\mathbf{u}(t) \equiv \mathbf{u} \geq \mathbf{0}.$$

Assume that (8.2) *is monotonically bounded and*

$$\textit{range}([\mathbf{W}^{\text{oo}} \ \mathbf{W}^{\text{o1}}]) \subseteq \textit{range}(\mathbf{B}^{\text{o}}). \tag{8.28}$$

Then, for any $\tilde{\mathbf{p}}^{\text{1}} \in \mathbb{R}^{n-r}$*, there exists* $\bar{\mathbf{u}} \in \mathbb{R}^m_{\geq 0}$ *such that for all* $\mathbf{u} \geq \bar{\mathbf{u}}$*,* $\mathcal{N}_2$ *is GES relative to a unique equilibrium of the form* $\mathbf{x}_* = [\mathbf{0}_r^T \ (\mathbf{x}_*^{\text{1}})^T]^T$ *if and only if* $\mathbf{W}^{\text{11}}$ *is such that the internal* $\mathcal{N}_2^{\text{1}}$ *dynamics*

$$\tau \dot{\mathbf{x}}^{\text{1}} = -\mathbf{x}^{\text{1}} + [\mathbf{W}^{\text{11}}\mathbf{x}^{\text{1}} + \tilde{\mathbf{p}}^{\text{1}}]^+, \tag{8.29}$$

is GES relative to a unique equilibrium.

Proof. ($\Leftarrow$) Define $\mathbf{u}_s$ to be a solution of

$$\mathbf{B}^{\circ}\mathbf{u}_s = -[[\mathbf{W}^{\circ\circ}\ \mathbf{W}^{\circ 1}]]^{+}\boldsymbol{\nu}(\tilde{\mathbf{p}}). \tag{8.30}$$

This solution exists by assumption (8.28). Let $\bar{\mathbf{u}} = [\mathbf{u}_s]^{+}$ and note that $\mathbf{B}^{\circ}\mathbf{u} \leq \mathbf{B}^{\circ}\bar{\mathbf{u}} \leq \mathbf{B}^{\circ}\mathbf{u}_s$. By construction, (8.2), (8.26), (8.27), (8.30) simplify to

$$\begin{aligned}
\tau\dot{\mathbf{x}}^{\circ} &= -\mathbf{x}^{\circ}, \\
\tau\dot{\mathbf{x}}^{1} &= -\mathbf{x}^{1} + [\mathbf{W}^{1\circ}\mathbf{x}^{\circ} + \mathbf{W}^{11}\mathbf{x}^{1} + \tilde{\mathbf{p}}^{1}]^{+},
\end{aligned} \tag{8.31}$$

whose GES follows from Lemma 8.A.2.

($\Rightarrow$) By monotone boundedness and nonpositivity of $\mathbf{B}^{\circ}$, $\mathbf{x}(t) \leq \boldsymbol{\nu}(\tilde{\mathbf{p}})$ for all $t \geq 0$ and any $\mathbf{u} \geq \bar{\mathbf{u}}$. Let $\mathbf{u} = \bar{\mathbf{u}} + [\mathbf{u}_s]^{+}$ where $\mathbf{u}_s$ is a solution to (8.30). Similar to above, this simplifies (8.2), (8.26), (8.27), (8.30) to (8.31), which is GES by assumption. However, for any initial condition of the form $\mathbf{x}(0) = [\mathbf{0}_r^T\ \mathbf{x}^{1}(0)^T]^T$, the trajectories of (8.31) are the same as (8.29), and the result follows. □

As we show next, the condition (8.28) on the ability to influence the dynamics of the task-irrelevant nodes through control also plays a key role in feedback selective inhibition. We defer the discussion about the interpretation of this condition to Section 8.4.3 below.

8.4.2 Feedback Selective Inhibition

The core idea of feedback inhibition [90], as found throughout the brain, is the dependence of the amount of inhibition on the activity level of the nodes that are to be inhibited. This dependence is in particular relevant to GDSA, as the stronger and more salient a source of distraction, the harder one must try to suppress its effects on perception. The next result provides a novel characterization of several equivalences between the dynamical properties of $\mathcal{N}_2$ under linear full-state feedback inhibition and those of $\mathcal{N}_2^{\mathbf{1}}$.

Theorem 8.4.3. ***(Feedback selective inhibition).*** *Consider the dynamics* (8.2)*, where the external input is given by* (8.26)-(8.27) *with a linear state feedback* $\mathbf{u}$

$$\mathbf{u}(t) = \mathbf{K}\mathbf{x}(t), \tag{8.32}$$

and $\mathbf{K} \in \mathbb{R}^{m\times n}$ *is a constant control gain. Assume that* (8.28) *holds. Then, there exists* $\mathbf{K} \in \mathbb{R}^{m\times n}$ *such that*

(i) $\mathbf{I} - (\mathbf{W} + \mathbf{B}\mathbf{K}) \in \mathcal{P}$ *if and only if* $\mathbf{I} - \mathbf{W}^{\mathbf{11}} \in \mathcal{P}$*;*

(ii) $-\mathbf{I} + (\mathbf{W} + \mathbf{B}\mathbf{K}) \in \mathcal{H}$ *if and only if* $-\mathbf{I} + \mathbf{W}^{\mathbf{11}} \in \mathcal{H}$*;*

(iii) $\mathbf{W} + \mathbf{B}\mathbf{K} \in \mathcal{L}$ *if and only if* $\mathbf{W}^{\mathbf{11}} \in \mathcal{L}$*;*

(iv) $\rho(|\mathbf{W} + \mathbf{B}\mathbf{K}|) < 1$ *if and only if* $\rho(|\mathbf{W}^{\mathbf{11}}|) < 1$*;*

(v) $\|\mathbf{W} + \mathbf{B}\mathbf{K}\| < 1$ *if and only if* $\|[\mathbf{W}^{\mathbf{10}}\ \mathbf{W}^{\mathbf{11}}]\| < 1$.

Proof. *(i)* $\Rightarrow$) For any $\mathbf{K} = [\mathbf{K}^{\circ}\ \mathbf{K}^{1}] \in \mathbb{R}^{m\times n}$,

$$\mathbf{W} + \mathbf{B}\mathbf{K} = \begin{bmatrix} \mathbf{W}^{\circ\circ} + \mathbf{B}^{\circ}\mathbf{K}^{\circ} & \mathbf{W}^{\circ 1} + \mathbf{B}^{\circ}\mathbf{K}^{1} \\ \mathbf{W}^{1\circ} & \mathbf{W}^{11} \end{bmatrix}. \tag{8.33}$$

Thus, since any principal submatrix of a P-matrix is a P-matrix, $\mathbf{I} - \mathbf{W}^{11} \in \mathcal{P}$.

$\Leftarrow$) Since $m \geq R$, there Lebesgue-almost always exists $\bar{\mathbf{K}} \in \mathbb{R}^{m\times n}$ such that

$$-\begin{bmatrix} \mathbf{W}^{\circ\circ} & \mathbf{W}^{\circ 1} \end{bmatrix} = \mathbf{B}^{\circ}\bar{\mathbf{K}}. \tag{8.34}$$

Using the fact that the determinant of any block-triangular matrix is the product of the determinants of the blocks on its diagonal [76, Prop 2.8.1], it follows that $\mathbf{I} - (\mathbf{W} + \mathbf{B}\bar{\mathbf{K}}) \in \mathcal{P}$.

(ii) $\Rightarrow$) This follows from (8.33) and the fact that a principal submatrix of a totally-Hurwitz matrix is totally-Hurwitz.

$\Leftarrow$) Using the matrix $\bar{\mathbf{K}}$ in (8.34), the result follows from the fact that the eigenvalues of a block-triangular matrix are the eigenvalues of its diagonal blocks.

(iii) $\Rightarrow$) Let $\mathbf{P} = \mathbf{P}^T > \mathbf{0}$ be such that

$$(-\mathbf{I} + (\mathbf{W} + \mathbf{B}\mathbf{K})^T\boldsymbol{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \boldsymbol{\Sigma}(\mathbf{W} + \mathbf{B}\mathbf{K})) < \mathbf{0} \tag{8.35}$$

for all $\boldsymbol{\sigma} \in \{0,1\}^n$. Consider, in particular, $\boldsymbol{\sigma} = [\mathbf{0}_r^T\ (\boldsymbol{\sigma}^1)^T]^T$ where $\boldsymbol{\sigma}^1 \in \{0,1\}^{n-r}$ is arbitrary. Let

$\mathbf{\Sigma}^{1} = \mathrm{diag}(\boldsymbol{\sigma}^{1})$ and partition $\mathbf{P}$ in 2-by-2 block form similarly to $\mathbf{W}$. Since

$$(-\mathbf{I} + (\mathbf{W} + \mathbf{B}\mathbf{K})^T\mathbf{\Sigma})\mathbf{P} + \mathbf{P}(-\mathbf{I} + \mathbf{\Sigma}(\mathbf{W} + \mathbf{B}\mathbf{K})) = \begin{bmatrix} \star & \star \\ \star & (-\mathbf{I} + \mathbf{\Sigma}^{1}\mathbf{W}^{11})^T\mathbf{P}^{11} + \mathbf{P}^{11}(-\mathbf{I} + \mathbf{\Sigma}^{1}\mathbf{W}^{11}) \end{bmatrix},$$

and any principal submatrix of a negative definite matrix is negative definite, we deduce $\mathbf{W}^{11} \in \mathcal{L}$.

$\Leftarrow$) Let $\mathbf{P}^{11} \in \mathbb{R}^{(n-r)\times(n-r)}$ be such that

$$(-\mathbf{I} + (\mathbf{W}^{11})^T\mathbf{\Sigma}^{1})\mathbf{P}^{11} + \mathbf{P}^{11}(-\mathbf{I} + \mathbf{\Sigma}^{1}\mathbf{W}^{11}) < \mathbf{0},$$

for all $\boldsymbol{\sigma}^{1} \in \{0, 1\}^{n-r}$ and $\bar{\mathbf{K}}$ be as in (8.34). For any $\boldsymbol{\sigma} = [(\boldsymbol{\sigma}^{\circ})^T\ (\boldsymbol{\sigma}^{1})^T]^T$, (8.34) gives

$$-\mathbf{I} + \mathbf{\Sigma}(\mathbf{W} + \mathbf{B}\bar{\mathbf{K}}) = \begin{bmatrix} -\mathbf{I} & \mathbf{0} \\ \star & -\mathbf{I} + \mathbf{\Sigma}^{1}\mathbf{W}^{11} \end{bmatrix}.$$

Thus, the dynamics

$$\tau\dot{\mathbf{x}} = \big(-\mathbf{I} + \mathbf{\Sigma}(\mathbf{W} + \mathbf{B}\bar{\mathbf{K}})\big)\mathbf{x},$$

is a cascade of $\tau\dot{\mathbf{x}}^{\circ} = -\mathbf{x}^{\circ}$ and

$$\tau\dot{\mathbf{x}}^{1} = (-\mathbf{I} + \mathbf{\Sigma}^{1}\mathbf{W}^{11})\mathbf{x}^{1} + \star \cdot \mathbf{x}^{\circ},$$

where the latter has the ISS Lyapunov function $V^{1}(\mathbf{x}^{1}) = (\mathbf{x}^{1})^T\mathbf{P}^{11}\mathbf{x}^{1}$. Using [92, Thm 3], (8.35)

holds for $\mathbf{K} = \bar{\mathbf{K}}$, $\mathbf{P} = \mathrm{diag}(\mathbf{I}, \mathbf{P}^{11})$, and any $\boldsymbol{\sigma} \in \{0,1\}^n$, giving $\mathbf{W} + \mathbf{B}\bar{\mathbf{K}} \in \mathcal{L}$.

(iv) $\Rightarrow$) This follows from (8.33) and [76, Fact 4.11.19].

$\Leftarrow$) Consider the matrix $\bar{\mathbf{K}}$ in (8.34). The result then follows from the fact that the eigenvalues of a block-triangular matrix are the eigenvalues of its diagonal blocks.

(v) $\Rightarrow$) Note that for any $\mathbf{K} \in \mathbb{R}^{m \times n}$,

$$\begin{aligned}\|\mathbf{W} + \mathbf{B}\mathbf{K}\|^2 &= \rho\left(\begin{bmatrix} \star & \star \\ \star & \mathbf{W}^{10}(\mathbf{W}^{10})^T + \mathbf{W}^{11}(\mathbf{W}^{11})^T \end{bmatrix}\right) \\ &\geq \rho(\mathbf{W}^{10}(\mathbf{W}^{10})^T + \mathbf{W}^{11}(\mathbf{W}^{11})^T) = \left\| \begin{bmatrix} \mathbf{W}^{10} & \mathbf{W}^{11} \end{bmatrix} \right\|^2,\end{aligned}$$

where the inequality follows from the well-known interlacing property of eigenvalues of principal submatrices (c.f. [93]).

$\Leftarrow$) Consider the matrix $\bar{\mathbf{K}}$ in (8.34) and note that

$$\begin{aligned}\|\mathbf{W} + \mathbf{B}\bar{\mathbf{K}}\|^2 &= \rho\left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^{10}(\mathbf{W}^{10})^T + \mathbf{W}^{11}(\mathbf{W}^{11})^T \end{bmatrix}\right) \\ &= \rho(\mathbf{W}^{10}(\mathbf{W}^{10})^T + \mathbf{W}^{11}(\mathbf{W}^{11})^T) = \left\| \begin{bmatrix} \mathbf{W}^{10} & \mathbf{W}^{11} \end{bmatrix} \right\|^2 < 1,\end{aligned}$$

completing the proof. □

Remark 8.4.4. ***(Feedback inhibition with nonnegative*** $\mathbf{u}(t)$***).*** Even though Theorem 8.4.3 is motivated by feedback inhibition in the brain, the result illustrates some fundamental properties of linear-threshold dynamics and the corresponding matrix classes that is of independent interest, which motivates the generality of its formulation. The particular application to brain networks

requires nonnegative inputs, which we discuss next. The core principle of Theorem 8.4.3 is the cancellation of local input $[\mathbf{W}^{\circ\circ}\ \mathbf{W}^{\circ 1}]\mathbf{x}$ to $\mathcal{N}_2^\circ$ with the top-down feedback input $\mathbf{B}^\circ\bar{\mathbf{K}}\mathbf{x}$, simplifying the dynamics of $\mathcal{N}_2^\circ$ to $\tau\dot{\mathbf{x}}^\circ = -\mathbf{x}^\circ$ that guarantee its inhibition. However, the resulting input signal $\mathbf{u} = \bar{\mathbf{K}}\mathbf{x}$ (being the firing rate of some neuronal population) may not remain nonnegative at all times. This can be easily addressed as follows. Similar to the proof of Theorem 8.4.2, we let

$$\mathbf{u}(t) = [\bar{\mathbf{K}}\mathbf{x}(t)]^+.$$

This makes $\mathbf{u}(t)$ nonnegative without affecting the selective inhibition of $\mathcal{N}_2^\circ$ in (8.2) due to the nonpositivity of $\mathbf{B}^\circ$ and the thresholding in $[\cdot]^+$. □

8.4.3 Network Size, Weight Distribution, and Stabilization

Underlying the discussion above is the requirement that $\mathcal{N}_2$ can be asymptotically stabilized towards an equilibrium which has some components equal to zero and the remaining components determined by $\tilde{\mathbf{p}}$. Here, it is important to distinguish between the stability of $\mathcal{N}_2$ in the absence and presence of selective inhibition. In reality, the large size of biological neuronal networks often leads to highly unstable dynamics if all the nodes in a layer, say $\mathcal{N}_2$, are active. Therefore, the selective inhibition of $\mathcal{N}_2^\circ$ is not only responsible for the suppression of the task-irrelevant activity of $\mathcal{N}_2^\circ$, but also for the overall stabilization of $\mathcal{N}_2$ that allows for top-down recruitment of $\mathcal{N}_2^1$. This poses limitations on the size and structure of the subnetworks $\mathcal{N}_2^\circ$ and $\mathcal{N}_2^1$. It is in this context that one can analyze the condition (8.28) assumed in both Theorems 8.4.2 and 8.4.3. This condition requires, essentially, that there are sufficiently many "independent" external controls $\mathbf{u}$ to enforce inhibition on $\mathcal{N}_2^\circ$. The following result formalizes this statement.

Lemma 8.4.5. ***(Equivalent characterization of*** (8.28)***).*** *Let the matrices* $\mathbf{W}^\circ$ *and* $\mathbf{B}^\circ$ *have dimensions* $r \times n$ *and* $r \times m$*, respectively. Then,* range$(\mathbf{W}^\circ) \subseteq$ range$(\mathbf{B}^\circ)$ *for Lebesgue-almost all* $(\mathbf{W}^\circ, \mathbf{B}^\circ) \in \mathbb{R}^{r\times n} \times \mathbb{R}^{r\times m}$ *if and only if* $m \geq r$*.*

Proof. $\Rightarrow$) Assume, by contradiction, that $m < r$, so range$(\mathbf{B}^\circ) \subsetneq \mathbb{R}^r$ for any $\mathbf{B}^\circ$. Let $\mathbf{Q} = \mathbf{Q}(\mathbf{B}^\circ)$ be a matrix whose columns form a basis for range$(\mathbf{B}^\circ)^\perp$. Then, range$(\mathbf{W}^\circ) \subseteq$ range$(\mathbf{B}^\circ)$ if and only if $\mathbf{Q}(\mathbf{B}^\circ)^T\mathbf{W}^\circ = \mathbf{0}$. By Fubini's theorem [94, Ch. 20],

$$\begin{aligned}\int_{\mathbb{R}^{r\times n}\times\mathbb{R}^{r\times m}} \mathbb{I}_{\{\mathbf{Q}(\mathbf{B}^\circ)^T\mathbf{W}^\circ=\mathbf{0}\}}(\mathbf{W}^\circ, \mathbf{B}^\circ) d(\mathbf{W}^\circ, \mathbf{B}^\circ) &= \int_{\mathbb{R}^{r\times m}} d\mathbf{B}^\circ \int_{\mathbb{R}^{r\times n}} \mathbb{I}_{\{\mathbf{Q}(\mathbf{B}^\circ)^T\mathbf{W}^\circ=\mathbf{0}\}}(\mathbf{W}^\circ, \mathbf{B}^\circ) d\mathbf{W}^\circ \\ &= \int_{\mathbb{R}^{r\times m}} 0 \, d\mathbf{B}^\circ = 0,\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function. This contradiction proves $m \geq r$.

$\Leftarrow$) Let $\mathbf{B}^\circ = [\mathbf{B}_1^\circ \; \mathbf{B}_2^\circ]$ where $\mathbf{B}_1^\circ \in \mathbb{R}^{r\times r}$. It is straightforward to show that

$$\{(\mathbf{W}^\circ, \mathbf{B}^\circ) \mid \text{range}(\mathbf{W}^\circ) \nsubseteq \text{range}(\mathbf{B}^\circ)\} \subseteq \mathbb{R}^{r\times n} \times A,$$

where $A = \{\mathbf{B}^\circ \mid \det(\mathbf{B}_1^\circ) = 0\}$. Since A has measure zero, the result follows from a similar argument as above invoking Fubini's theorem. □

Based on intuitions from linear systems theory, it may be tempting to seek a relaxation of (8.28) for the case where $m < r$. This is due to the fact that for a *linear* system $\tau\dot{\mathbf{x}} = \mathbf{W}\mathbf{x} + \mathbf{B}\mathbf{u}$, it is known [95, eq (4.5) and Thm 3.5] that the set of all reachable states from the origin is given by

$$\text{range}\left(\begin{bmatrix} \mathbf{B}^\circ & \mathbf{W}^{\circ\circ}\mathbf{B}^\circ & \cdots & (\mathbf{W}^{n-1})^{\circ\circ}\mathbf{B}^\circ \\ \mathbf{0} & \mathbf{W}^{1\circ}\mathbf{B}^\circ & \cdots & (\mathbf{W}^{n-1})^{1\circ}\mathbf{B}^\circ \end{bmatrix}\right),$$

which is usually much larger than range($\mathbf{B}$). Therefore, it is reasonable to expect that (8.28) could be relaxed to

$$\text{range}([\mathbf{W}^{\circ\circ}\ \mathbf{W}^{\circ 1}]) \subseteq \text{range}([\mathbf{B}^{\circ}\ \mathbf{W}^{\circ\circ}\mathbf{B}^{\circ}\ \cdots\ (\mathbf{W}^{n-1})^{\circ\circ}\mathbf{B}^{\circ}]). \tag{8.36}$$

However, it turns out that this relaxation is not possible, the reason being the (apparently simple, yet intricate) nonlinearity in (8.2). We show this by means of an example.

Example 8.4.6. ***(Tightness of*** (8.28)***).*** Consider the feedback dynamics (8.2), (8.26), (8.32), where $n = 3$, $m = 1$, $r = 2$, and

$$\mathbf{W} = \left[\begin{array}{cc|c} 2\alpha & 0 & 0 \\ 0 & 3\alpha & 0 \\ \hline 0 & 0 & \alpha \end{array}\right], \quad \mathbf{B} = \left[\begin{array}{c} 1 \\ 1 \\ \hline 0 \end{array}\right], \quad \alpha \in (0.5, 1).$$

Clearly, (8.28) does not hold (so Theorem 8.4.3(iv) does not apply), but range($[\mathbf{W}^{\circ\circ}\ \mathbf{W}^{\circ 1}]$) $\subseteq$ range($[\mathbf{B}^{\circ}\ \mathbf{W}^{\circ\circ}\mathbf{B}^{\circ}]$). One can show that for all $\mathbf{K} \in \mathbb{R}^{1\times 3}$,

$$\rho(|\mathbf{W} + \mathbf{B}\mathbf{K}|) \geq 2\alpha > 1,$$

while $\rho(\mathbf{W}^{11}) = \alpha < 1$, verifying that (8.28) is necessary and cannot be relaxed to (8.36). □

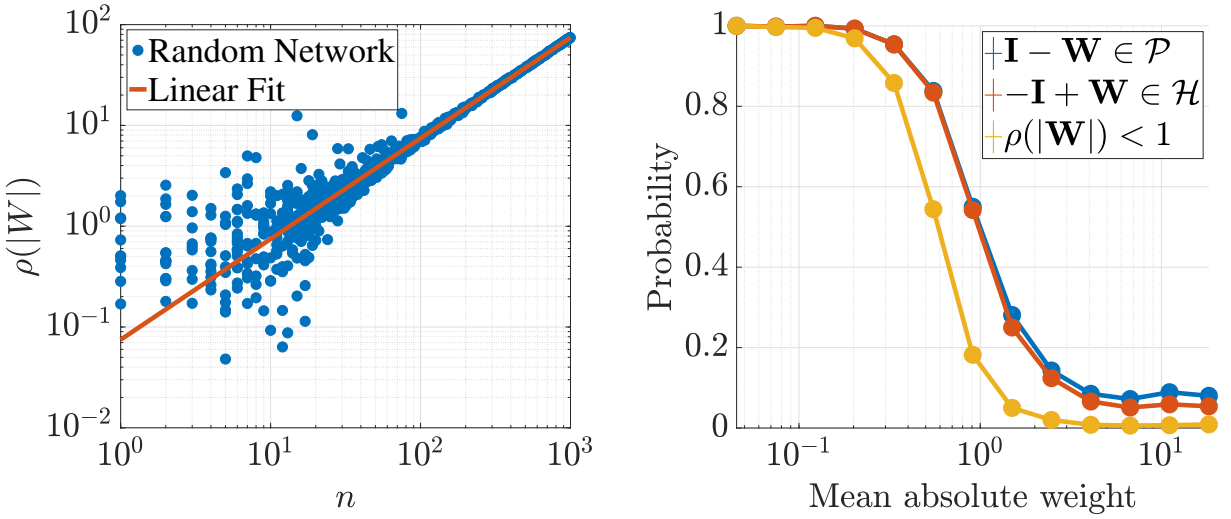
Theorems 8.4.2 and 8.4.3 use completely different mechanisms for inhibition of $\mathcal{N}_2^{\circ}$, yet they are strikingly similar in one conclusion, namely, that the dynamical properties achievable under selective inhibition are precisely those satisfied by the task-relevant part $\mathcal{N}_2^{1}$. This has important implications for the size and structure of the part $\mathcal{N}_2^{1}$ that can remain active at any instance of time

without resulting in instability. The next remark elaborates on this implication.

Remark 8.4.7. ***(Implications for the size of $\mathcal{N}_2^1$).*** Existing experimental evidence suggest that the synaptic weights $\mathbf{W}$ in cortical networks are sparse, approximately follow a log-normal distribution, and have a pairwise connection probability that is independent of physical distance between neurons within short distances [96]. Figure 8.3(a) shows the value of $\rho(|\mathbf{W}|)$ for random matrices with such statistics, which grows linearly with n. Notably, the network (representing $\mathcal{N}_2$ here) rapidly loses stability as its size grows. On the other hand, recent advances in machine learning suggest that the expressivity of a neuronal network (often loosely defined as its capacity to reproduce complex trajectories) is maximized when it operates at the boundary between stability and instability, commonly referred to as the *edge of chaos* [97–99]. While determining the optimal size of a network that leads to maximal expressivity is beyond the scope of this work, our results suggest a critical role for selective inhibition in keeping only a limited number of nodes in $\mathcal{N}_2$ active at any given time while inhibiting others. In other words, while the overall size of subnetworks in a brain network (corresponding to, e.g., the number of neuronal populations with distinct preferred stimuli in a brain region) is inevitably large, selective inhibition offers a plausible explanation for the mechanism by which the brain keeps the number of active populations at any given time bounded ($O(1)$), thus maintaining its local dynamics close to the "edge of chaos".

Similarly, Figure 8.3(b) shows the probability of the three stability related conditions $\mathbf{I} - \mathbf{W} \in \mathcal{P}$, $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$, and $\rho(|\mathbf{W}|) < 1$ (cf. Section 8.3) as a function of average absolute weight in the network, showing a rapid drop from 1 to less than 0.1 around unit average absolute weight. Interestingly, several works in the neuroscience literature have shown that neuronal networks maintain stability by re-scaling their synaptic weights that change during learning, a process

commonly referred to as *homeostatic synaptic plasticity* [100]. Our results therefore open the way to provide rigorous and quantifiable measures of the optimal size and weight distribution of neuronal subnetworks that may be active at any given time in order to maintain any desired level of network stability. □



(a) (b)

Figure 8.3: The effects of network size and weight distribution on its stability. (a) Linear growth of $\rho(|\mathbf{W}|)$ with network size n. Each circle represents a random matrix with 10% sparsity and synaptic weights log-normally distributed with parameters $\mu = -0.7$ and $\sigma = 0.9$ as given in [96]. As in cortical networks, 80% of nodes are excitatory and 20% inhibitory. The line illustrates a fit of the form $\log \rho(|\mathbf{W}|) = \alpha \log n + \beta$ with $\alpha = 1$ and $\beta = -1.2$, showing a linear growth of $\rho(|\mathbf{W}|)$ with n. (b) Probability of $\mathbf{I} - \mathbf{W} \in \mathcal{P}$, $-\mathbf{I} + \mathbf{W} \in \mathcal{H}$, and $\rho(|\mathbf{W}|) < 1$ as a function of the average absolute weight of the networks with the same statistics as in (a) but fixed size $n = 10$ and varying $\mu = -3.5, -3, \ldots, 2.5$. The probabilities are estimated empirically with 10^3 sample networks for each value of μ.

Complementary to selective inhibition is selective recruitment, as analyzed next.

8.5 Selective Recruitment in Bilayer Networks

In this section we tackle the analysis of simultaneous selective inhibition and top-down recruitment in a two-layer network. We consider the same dynamics as in (8.3) for the lower-

level subnetwork $\mathcal{N}_2$, but temporarily allow the dynamics of $\mathcal{N}_1$ to be arbitrary. This setup allows us to study the key ingredients of selective recruitment without the extra complications that arise from the multilayer interconnections of linear-threshold subnetworks and is the basis for our later developments. Further, by keeping the higher-level dynamics arbitrary, the results presented here are also of independent interest beyond HSR, as they allow for a broader range of external inputs $\tilde{\mathbf{p}}$ than those generated by linear-threshold dynamics. This can be of interest in, for example, brain-computer interface (BCI) applications, where $\tilde{\mathbf{p}}_2$ is generated and applied by a computer ($\mathcal{N}_1$, not necessarily possessing linear-threshold dynamics) in order to control the activity of certain areas of the brain ($\mathcal{N}_2$).

For our subsequent analysis we need the equilibrium map h (cf. equation (8.8)) to be Lipschitz, as stated next. The proof of this result is a special case of Lemma 8.6.2 and thus omitted.

Lemma 8.5.1. ***(Lipschitzness of** h**).*** *Let h be as in* (8.8) *and single-valued*[5] *on $\mathbb{R}^n$. Then, it is globally Lipschitz on $\mathbb{R}^n$.*

The main result of this section is as follows.

Theorem 8.5.2. ***(Selective recruitment in bilayer hierarchical networks).*** *Consider the multilayer dynamics* (8.3) *where $N = 2$, $n_1 = n_2 - r_2$, $\mathbf{W}_{2,1} = [\mathbf{0}_{n_1 \times r_2}\ \mathbf{I}_{n_1}]^T$, $\mathbf{c}_2 = \mathbf{0}$, but $\mathbf{x}_1(t)$ is generated by some arbitrary dynamics*

$$\tau_1 \dot{\mathbf{x}}_1(t) = \gamma(\mathbf{x}_1(t), \mathbf{x}_2(t), t). \tag{8.37}$$

Let $h_2^1 = h_{\mathbf{W}_{2,2}^{11}}$ as in (8.8)*. If*

[5] It is indeed possible to show, using the same proof technique, that h is Lipschitz in the Hausdorff metric even when it is multiple-valued.

(i) γ *is measurable in* t*, locally bounded, and locally Lipschitz in* $(\mathbf{x}_1, \mathbf{x}_2)$ *uniformly in* t*;*

(ii) (8.37) *has bounded solutions uniformly in* $\mathbf{x}_2(t)$*;*

(iii) $m_2 \geq r_2$*;*

(iv) $\mathbf{W}_{2,2}^{11}$ *is such that* $\tau\dot{\mathbf{x}}_2^1 = -\mathbf{x}_2^1 + [\mathbf{W}_{2,2}^{11}\mathbf{x}_2^1 + \mathbf{x}_1]^+$ *is GES towards a unique equilibrium for any constant* $\mathbf{x}_1$*;*

then there exists $\mathbf{K}_2 \in \mathbb{R}^{m_2 \times n_2}$ *such that by using the feedback control* $\mathbf{u}_2(t) = \mathbf{K}_2\mathbf{x}_2(t)$*, one has*

$$\lim_{\epsilon_1 \to 0} \sup_{t \in [\underline{t}, \bar{t}]} \left\| \mathbf{x}_2(t) - \left(\mathbf{0}_{r_2}, h_2^1(\mathbf{x}_1(t))\right) \right\| = 0, \tag{8.38}$$

for any $0 < \underline{t} < \bar{t} < \infty$*. Further, if the dynamics of* $\mathbf{x}_2$ *is monotonically bounded*[6]*, there also exists a feedforward control* $\mathbf{u}_2(t) \equiv \bar{\mathbf{u}}_2$ *such that* (8.38) *holds for any* $0 < \underline{t} < \bar{t} < \infty$*.*

Proof. First we prove the result for feedback control. By *(iii)*, there exists $\mathbf{K}_2 \in \mathbb{R}^{m_2 \times n_2}$ Lebesgue-almost always (i.e., for Lebesgue-almost all $(\mathbf{W}_{2,2}^{00}, \mathbf{W}_{2,2}^{01}, \mathbf{B}_2^0)$) such that

$$\mathbf{W}_{2,2} + \mathbf{B}_2\mathbf{K}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{W}_{2,2}^{10} & \mathbf{W}_{2,2}^{11} \end{bmatrix}. \tag{8.39}$$

Further, by [101, Thm IV.7(ii) & Thm V.3(ii)], all the principal submatrices of $-\mathbf{I} + (\mathbf{W}_{2,2} + \mathbf{B}_2\mathbf{K}_2)$ are Hurwitz. Therefore, by [101, Thm IV.3 & Assump 1], h_2^1 is singleton-valued Lebesgue-almost

[6] See [101, Def V.1]

always (i.e., for Lebesgue-almost all $\mathbf{W}_{2,2}$). Thus, the dynamics of $\mathbf{x}_2$ simplifies to

$$\tau_2 \dot{\mathbf{x}}_2^{\circ} = -\mathbf{x}_2^{\circ}, \tag{8.40a}$$

$$\tau_2 \dot{\mathbf{x}}_2^{\mathbb{1}} = -\mathbf{x}_2^{\mathbb{1}} + [\mathbf{W}_{2,2}^{\mathbb{1}\circ}\mathbf{x}_2^{\circ} + \mathbf{W}_{2,2}^{\mathbb{1}\mathbb{1}}\mathbf{x}_2^{\mathbb{1}} + \mathbf{x}_1]^+, \tag{8.40b}$$

and has a unique equilibrium for any *fixed* $\mathbf{x}_1$. Assumption *(iv)* and [101, Lemma A.2] then ensure that (8.40) is GES relative to $(\mathbf{0}_{r_2}, h_2^{\mathbb{1}}(\mathbf{x}_1))$ for any fixed $\mathbf{x}_1$.

Based on assumption *(ii)*, let $D \subset \mathbb{R}^n$ be a compact set that contains the trajectory of the reduced-order model $\tau_1 \dot{\mathbf{x}}_1 = \gamma(\mathbf{x}_1, (\mathbf{0}_{r_2}, h_2^{\mathbb{1}}(\mathbf{x}_1)), t)$. By assumption *(i)*, γ is Lipschitz in $(\mathbf{x}_1, \mathbf{x}_2)$ on compacts uniformly in t. Let L_γ be its associated Lipschitz constant on $D \times \{\mathbf{0}_{r_2}\} \times h_2^{\mathbb{1}}(D)$. Using (8.11) and Lemma 8.5.1,

$$\begin{aligned}
\forall \mathbf{x}_1, \hat{\mathbf{x}}_1 \in D \quad \|\gamma(\mathbf{x}_1, h_2^{\mathbb{1}}(\mathbf{x}_1), t) - \gamma(\hat{\mathbf{x}}_1, h_2^{\mathbb{1}}(\hat{\mathbf{x}}_1), t)\| &\leq L_\gamma \|(\mathbf{x}_1 - \hat{\mathbf{x}}_1, h_2^{\mathbb{1}}(\mathbf{x}_1) - h_2^{\mathbb{1}}(\hat{\mathbf{x}}_1))\| \\
&\leq L_\gamma (\|\mathbf{x}_1 - \hat{\mathbf{x}}_1\| + \|h_2^{\mathbb{1}}(\mathbf{x}_1) - h_2^{\mathbb{1}}(\hat{\mathbf{x}}_1)\|) \\
&\leq L_\gamma (1 + L_h) \|\mathbf{x}_1 - \hat{\mathbf{x}}_1\|,
\end{aligned}$$

so $\gamma(\cdot, h_2^{\mathbb{1}}(\cdot), t) : \mathbb{R}^{n_{\mathbb{1}}} \to \mathbb{R}^{n_{\mathbb{1}}}$ is $L_\gamma(1 + L_h)$-Lipschitz on D. Using this fact, Lemma 8.6.2 again, and the change of variables $t' \triangleq t/\tau_1$, the claim follows from [71, Prop 1].

Next, we prove the result for constant feedforward control $\mathbf{u}_2(t) \equiv \bar{\mathbf{u}}_2$. Based on assumption *(ii)*, let $\bar{\mathbf{x}}_1 \in \mathbb{R}_{>0}^{n_{\mathbb{1}}}$ be the bound on the trajectories of (8.37) and $\bar{\mathbf{u}}_2$ be a solution of

$$\mathbf{B}_2^{\circ} \bar{\mathbf{u}}_2 = -[[\mathbf{W}_{2,2}^{\circ\circ}\ \mathbf{W}_{2,2}^{\circ\mathbb{1}}]]^+ \boldsymbol{\nu}(\bar{\mathbf{x}}_1),$$

where $\boldsymbol{\nu}$ comes from the monotone boundedness of the dynamics of $\mathbf{x}_2$. This solution Lebesgue-almost always exists by assumption *(ii)*. Then, the dynamics of $\mathbf{x}_2$ simplifies to (8.40), and [101, Lemma A.2] guarantees that it is GES relative to $(\mathbf{0}_{r_2}, h_2^{\scriptscriptstyle 1}(\mathbf{x}_1))$ for any *fixed* $\mathbf{x}_1$. The claim then follows, similar to the feedback case, from [71, Prop 1]. □

Remark 8.5.3. ***(Validity of the assumptions of Theorem 8.5.2.).*** Assumption *(i)* is merely technical and is not a restriction in practice. In particular, this assumption is satisfied when using a linear-threshold model for (8.37). Likewise, assumption *(ii)* is always satisfied in reality, as the state of all biological neuronal networks are bounded by the inverse of the refractory period of their neurons. Even in theory, this assumption can be relaxed to only the boundedness of the reduced-order model in the case of feedback inhibition (cf. Theorem 8.6.3). Assumption *(iii)* requires that there exist sufficiently many inhibitory control channels to suppress the activity of the first r nodes of the lower-level subnetwork. The most critical requirement is assumption *(iv)*, which is not only sufficient but also necessary for inhibitory stabilization (cf. [101] for conditions on $\mathbf{W}_{2,2}^{11}$ that ensure this assumption as well as its necessity for inhibitory stabilization). □

The main conclusion of Theorem 8.5.2 is the Tikhonov-type singular perturbation statement in (8.38). According to this statement, for any $\theta > 0$,

$$|\mathbf{x}_2(t) - (\mathbf{0}_{r_2}, h_2^{\scriptscriptstyle 1}(\mathbf{x}_1(t)))| \leq \theta \mathbf{1}_{n_2}, \qquad \forall t \in [\underline{t}, \bar{t}], \tag{8.41}$$

provided that τ_2/τ_1 is sufficiently small, i.e., the higher-level dynamics is sufficiently slower than the lower-level one. As discussed in Section 8.1, this timescale separation is characteristic of biological neuronal networks.

An important observation regarding (8.41) is that the equilibrium map $h_2^{\scriptscriptstyle 1}$ does not have

a closed-form expression, so the reference trajectory $h_2^1(\mathbf{x}_1(t))$ of the lower-level network is only implicitly known for any given $\mathbf{x}_1(t)$. However, if a desired trajectory $\boldsymbol{\xi}_2^1(t) \in \mathbb{R}_{\geq 0}^{n_2 - r_2}$ for $\mathbf{x}_2^1$ is known a priori, one can specify the appropriate γ such that $h_2^1(\mathbf{x}_1(t)) = \boldsymbol{\xi}_2^1(t)$. To show this, let the dynamics of $\boldsymbol{\xi}_2^1(t)$ be given by

$$\tau_1 \dot{\boldsymbol{\xi}}_2^1(t) = \gamma_{\xi}(\boldsymbol{\xi}_2^1(t), t).$$

Then, choosing $\mathbf{x}_1(t) = (\mathbf{I} - \mathbf{W}_{2,2}^{11})\boldsymbol{\xi}_2^1(t)$, yields

$$[\mathbf{W}_{2,2}^{11}\boldsymbol{\xi}_2^1(t) + \mathbf{x}_1(t)]^+ = [\boldsymbol{\xi}_2^1(t)]^+ = \boldsymbol{\xi}_2^1(t),$$

which, according to (8.8), implies $\boldsymbol{\xi}_2^1(t) = h_2^1(\mathbf{x}_1(t))$.

Next, we use this result to illustrate the core concepts of the bilayer HSR in a synthetic but biologically-inspired example, where a subnetwork of inhibitory nodes generates oscillations which are then selectively induced on a lower-level excitatory subnetwork.

Example 8.5.4. ***(HSR of an excitatory subnetwork by inhibitory oscillations).*** Consider the dynamics (8.3) with $N = 2$, a 3-dimensional excitatory subnetwork at the lower level, and a 3-

dimensional inhibitory subnetwork at the higher level. Let

$$\mathbf{W}_{1,1} = \begin{bmatrix} 0 & -0.8 & -1.7 \\ -1 & 0 & -0.5 \\ -0.7 & -1.8 & 0 \end{bmatrix}, \quad \mathbf{c}_1 = \begin{bmatrix} 11 \\ 10 \\ 10 \end{bmatrix},$$

$$\mathbf{W}_{2,2} = \begin{bmatrix} 0 & 0.9 & 1.2 \\ 0.7 & 0 & 1 \\ 0.8 & 0.2 & 0 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{c}_2 = \begin{bmatrix} 2 \\ 3.5 \\ 2.5 \end{bmatrix},$$

$$\mathbf{W}_{1,2} = \mathbf{0}, \quad \mathbf{W}_{2,1} = -\mathbf{I}, \quad u_2 = -5. \tag{8.42}$$

It is straightforward to verify that this example satisfies all the assumptions of Theorem 8.5.2. Therefore, we expect the actual $\mathbf{x}_2$-trajectory to be close to the *desired* $\mathbf{x}_2$-trajectory $(0, h_2^1(\mathbf{x}_1(t))$ provided that $\epsilon_1 \ll 1$. Figure 8.4 shows the trajectories of this system for $\epsilon_1 = 0.5$ together with a schematic of the interconnections. We see that even with this mild separation of timescales, $\mathbf{x}_2(t)$ and $(0, h_2^1(\mathbf{x}_1(t))$ are remarkably close.

It is easy to see that the complete $\mathbf{x}_2$-subsystem is unstable by itself. However, when $x_{2,1}$ is inhibited, the remaining $x_{2,2}$-$x_{2,3}$ subnetwork becomes GES. Therefore, the higher-lever inhibitory network (which is oscillatory itself) has selectively inhibited $x_{2,1}$ while simultaneously recruiting (by inducing an oscillation in) the $x_{2,2}$-$x_{2,3}$ part.[7] Note that although $x_{2,1}$ is not effectively used here, it can be replaced by $x_{2,2}$ or $x_{2,3}$ at other times. In other words, while the full $\mathbf{x}_2$-dynamics is unstable, any two-node part of it is GES. Therefore, different "tasks" can be accomplished at different times through the selective inhibition of one of $\{x_{2,1}, x_{2,2}, x_{2,3}\}$ and top-down recruitment of the other

[7] Coherent oscillatory activity has been widely shown to be involved in transfer of information between cortical circuits, see, e.g., [102–104].

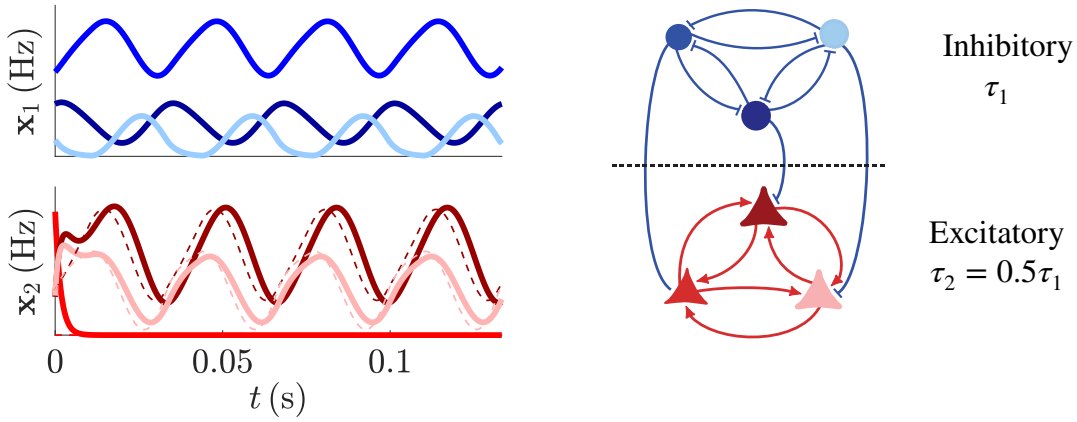


Figure 8.4: The network structure (right) and trajectories (left) of the two-timescale network in (8.42). The red pyramids and blue circles depict excitatory and inhibitory nodes, respectively, and the trajectory colors on the left correspond to node colors on the right. The dashed lines show the desired reference trajectories $\left(0, h_2^1(\mathbf{x}_1(t))\right)$.

two. Generalizing this to larger networks results in more flexible selective recruitment of different subsets of nodes at different times, as observed in nature and formulated in Theorem 8.5.2. □

Remark 8.5.5. ***(Biological relevance of Example 8.5.4).*** In addition to providing a simple illustration of the hierarchical selective recruitment framework developed here, the model (8.42) captures a number of well-known aspects of selective attention in brain dynamics. First, extensive human and animal studies have demonstrated a robust correlation between oscillatory activity, particularly in the gamma frequency band ($\sim 30 - 100^{\text{Hz}}$), and selective attention in a variety of contexts [105–108]. Furthermore, gamma oscillations in the cortex are shown to be primarily generated by networks of inhibitory neurons, which then recruit the excitatory populations (see [109] and the references therein), as captured by the network structure of Figure 8.4. Interestingly, the oscillations generated by the higher-level inhibitory subnetwork fall within the gamma band by setting $\tau_1 \sim 3^{\text{ms}}$ which lies within the decay time constant range of GABA_A inhibitory receptors.[8]

[8] See, e.g., the Neurotransmitter Time Constants database of the CNRGlab at the University of Waterloo, `http://compneuro.uwaterloo.ca/research/constants-constraints/`

Finally, the timescale of the dynamics of inhibitory subnetworks is in general slower than that of excitatory dynamics in the brain [110–112], supporting the about 2-fold separation of timescales in Figure 8.4. □

8.6 Selective Recruitment in Multilayer Networks

In this section, we tackle the problem stated in Section 8.2 in its general form and consider an N-layer hierarchical structure of subnetworks with linear-threshold dynamics. Given the model (8.3), let

$$h_i^{\scriptscriptstyle 1} : \mathbf{c}_i^{\scriptscriptstyle 1} \rightrightarrows \{\mathbf{x}_i^{\scriptscriptstyle 1} \mid \mathbf{x}_i^{\scriptscriptstyle 1} = [\mathbf{W}_{i,i+1}^{\scriptscriptstyle 11} h_{i+1}^{\scriptscriptstyle 1}(\mathbf{W}_{i+1,i}^{\scriptscriptstyle 11}\mathbf{x}_i^{\scriptscriptstyle 1} + \mathbf{c}_{i+1}^{\scriptscriptstyle 1}) + \mathbf{W}_{i,i}^{\scriptscriptstyle 11}\mathbf{x}_i^{\scriptscriptstyle 1} + \mathbf{c}_i^{\scriptscriptstyle 1}]^+\},\ i = 2, \dots, N-1,$$

$$h_N^{\scriptscriptstyle 1} = h_{\mathbf{W}_{N,N}^{\scriptscriptstyle 11}},$$

be the recursive definition of the (set-valued) equilibrium maps of the task-relevant part of the layers. The maps $\{h_i^{\scriptscriptstyle 1}\}_{i=2}^N$ play a central role in the multiple-timescale dynamics of (8.3). Therefore, our first step is to study their properties carefully. Our first result characterizes their piecewise affinity nature.

Lemma 8.6.1. ***(Piecewise affinity of equilibrium maps is preserved along the layers of a hierarchical linear-threshold network).*** *Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be a piecewise affine function of the form*

$$h(\mathbf{c}) = \mathbf{F}_\lambda \mathbf{c} + \mathbf{f}_\lambda, \qquad \forall \mathbf{c} \in \Psi_\lambda \triangleq \{\mathbf{c} \mid \mathbf{G}_\lambda \mathbf{c} + \mathbf{g}_\lambda \geq \mathbf{0}\},$$

$$\forall \lambda \in \Lambda,$$

neurotransmitter-time-constants-pscs.html.

where Λ is a finite index set and $\bigcup_{\lambda\in\Lambda}\Psi_\lambda = \mathbb{R}^n$. Given matrices $\mathbf{W}_\ell, \ell = 1,2,3$ and a vector $\bar{\mathbf{c}}$, assume

$$\mathbf{x} = [\mathbf{W}_1\mathbf{x} + \mathbf{W}_2 h(\mathbf{W}_3\mathbf{x} + \bar{\mathbf{c}}) + \mathbf{c}']^+, \tag{8.43}$$

is known to have a unique solution $\mathbf{x} \in \mathbb{R}^{n'}$ for all $\mathbf{c}' \in \mathbb{R}^{n'}$ and let $h'(\mathbf{c}')$ be this unique solution. Then, there exists a finite index set Λ' and $\{(\mathbf{F}'_{\lambda'}, \mathbf{f}'_{\lambda'}, \mathbf{G}'_{\lambda'}, \mathbf{g}'_{\lambda'})\}_{\lambda'\in\Lambda'}$ such that

$$h'(\mathbf{c}') = \mathbf{F}'_{\lambda'}\mathbf{c}' + \mathbf{f}'_{\lambda'}, \quad \forall \mathbf{c}' \in \Psi'_{\lambda'} \triangleq \{\mathbf{c}' \mid \mathbf{G}'_{\lambda'}\mathbf{c}' + \mathbf{g}'_{\lambda'} \geq \mathbf{0}\},$$
$$\forall \lambda' \in \Lambda',$$

and $\bigcup_{\lambda'\in\Lambda'}\Psi'_{\lambda'} = \mathbb{R}^{n'}$.

Proof. Pick any $\mathbf{c}' \in \mathbb{R}^{n'}$ and let $\mathbf{x}^*$ be the unique solution of (8.43). Since $\bigcup_{\lambda\in\Lambda}\Psi_\lambda = \mathbb{R}^n$, there exists $\lambda \in \Lambda$ such that

$$\mathbf{W}_3\mathbf{x}^* + \bar{\mathbf{c}} \in \Psi_\lambda. \tag{8.44}$$

If $\mathbf{W}_3\mathbf{x}^* + \bar{\mathbf{c}}$ lies on the boundary of more than one Ψ_λ, pick one arbitrarily. Therefore, $\mathbf{x}^*$ satisfies

$$\mathbf{x}^* = [(\mathbf{W}_1 + \mathbf{W}_2\mathbf{F}_\lambda\mathbf{W}_3)\mathbf{x}^* + \mathbf{W}_2(\mathbf{F}_\lambda\bar{\mathbf{c}} + \mathbf{f}_\lambda) + \mathbf{c}']^+.$$

Similar to (8.11), we have

$$\begin{aligned}\mathbf{x}^* &= \big(\mathbf{I} - \boldsymbol{\Sigma}(\mathbf{W}_1 + \mathbf{W}_2\mathbf{F}_\lambda\mathbf{W}_3)\big)^{-1}\boldsymbol{\Sigma}\big(\mathbf{c}' + \mathbf{W}_2(\mathbf{F}_\lambda\bar{\mathbf{c}} + \mathbf{f}_\lambda)\big)\\ &\triangleq \mathbf{F}'_{\lambda'}\mathbf{c}' + \mathbf{f}'_{\lambda'},\end{aligned} \tag{8.45}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma}), \boldsymbol{\sigma} \in \{0,1\}^{n'}$ is such that

$$\begin{aligned}&\mathbf{c}' \in \Delta_{\boldsymbol{\sigma}} = \{\mathbf{c}' \mid \mathbf{M}_{\boldsymbol{\sigma}}\big(\mathbf{W}_2(\mathbf{F}_\lambda\bar{\mathbf{c}} + \mathbf{f}_\lambda) + \mathbf{c}'\big) \geq \mathbf{0}\},\\ &\mathbf{M}_{\boldsymbol{\sigma}} \triangleq (2\boldsymbol{\Sigma} - \mathbf{I})(\mathbf{I} - (\mathbf{W}_1 + \mathbf{W}_2\mathbf{F}_\lambda\mathbf{W}_3)\boldsymbol{\Sigma})^{-1},\end{aligned}$$

and $\lambda' \triangleq (\lambda, \boldsymbol{\sigma})$. Using this new representation of $\mathbf{x}^*$ in (8.45), we see that (8.44) holds if and only if

$$\mathbf{W}_3(\mathbf{F}'_{\lambda'}\mathbf{c}' + \mathbf{f}'_{\lambda'}) + \bar{\mathbf{c}} \in \Psi_\lambda \Leftrightarrow \mathbf{G}_\lambda(\mathbf{W}_3\mathbf{F}'_{\lambda'}\mathbf{c}' + \mathbf{W}_3\mathbf{f}'_{\lambda'} + \bar{\mathbf{c}}) + \mathbf{g}_\lambda \geq \mathbf{0}.$$

Therefore, (8.45) holds if and only if $\mathbf{c}' \in \Psi'_{\lambda'}$ with

$$\mathbf{G}'_{\lambda'} \triangleq \begin{bmatrix}\mathbf{G}_\lambda\mathbf{W}_3\mathbf{F}'_{\lambda'}\\ \mathbf{M}_{\boldsymbol{\sigma}}\end{bmatrix}, \qquad \mathbf{g}'_{\lambda'} \triangleq \begin{bmatrix}\mathbf{G}_\lambda(\mathbf{W}_3\mathbf{f}'_{\lambda'} + \bar{\mathbf{c}}) + \mathbf{g}_\lambda\\ \mathbf{M}_{\boldsymbol{\sigma}}\mathbf{W}_2(\mathbf{F}_\lambda\bar{\mathbf{c}} + \mathbf{f}_\lambda)\end{bmatrix}.$$

The proof is therefore complete by letting $\Lambda' = \Lambda \times \{0,1\}^{n'}$ and noticing that $\bigcup_{\lambda' \in \Lambda'} \Psi'_{\lambda'} = \mathbb{R}^{n'}$ since any $\mathbf{c}' \in \mathbb{R}^{n'}$ must belong to at least one $\Psi'_{\lambda'}$ by construction. □

Note that a special case of Lemma 8.6.1 is when $\mathbf{W}_2 = \mathbf{0}$, in which case h' becomes, like $h^{\mathbb{1}}_N$, the standard equilibrium map (8.8) of linear-threshold dynamics. Our next result characterizes

the global Lipschitzness property of the equilibrium maps.

Lemma 8.6.2. ***(Piecewise affine equilibrium maps are globally Lipschitz).*** *Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be a piecewise affine function of the form*

$$h(\mathbf{c}) = \mathbf{F}_\lambda \mathbf{c} + \mathbf{f}_\lambda, \qquad \forall \mathbf{c} \in \Psi_\lambda \triangleq \{\mathbf{c} \mid \mathbf{G}_\lambda \mathbf{c} + \mathbf{g}_\lambda \geq \mathbf{0}\},$$
$$\forall \lambda \in \Lambda,$$

where Λ is a finite index set and $\bigcup_{\lambda\in\Lambda} \Psi_\lambda = \mathbb{R}^n$. Then, h is globally Lipschitz.

Proof. Pick any $\mathbf{c}, \hat{\mathbf{c}} \in \mathbb{R}^n$. Since all the sets Ψ_λ are convex, the line segment $\gamma \triangleq \left\{\left(\theta, (1-\theta)\mathbf{c} + \theta\hat{\mathbf{c}}\right) \mid \theta \in [0,1]\right\}$ joining $\mathbf{c}$ and $\hat{\mathbf{c}}$ can be broken into $k \leq |\Lambda| < \infty$ pieces such that $\gamma = \bigcup_{i=1}^k \gamma_i, \gamma_i \triangleq \left\{\left(\theta, (1-\theta)\mathbf{c} + \theta\hat{\mathbf{c}}\right) \mid \theta \in [\theta_{i-1}, \theta_i]\right\}, \theta_0 = 0, \theta_k = 1$ and each $\gamma_i \subset \Psi_{\lambda_i}$ for some $\lambda_i \in \Lambda$. Let $\mathbf{c}_i \triangleq (1-\theta_i)\mathbf{c} + \theta_i\hat{\mathbf{c}}$. Then,

$$\begin{aligned}
\|h(\mathbf{c}) - h(\hat{\mathbf{c}})\| &= \left\| \sum_{i=1}^k \left(h(\mathbf{c}_{i-1}) - h(\mathbf{c}_i)\right) \right\| \\
&\leq \sum_{i=1}^k \|h(\mathbf{c}_{i-1}) - h(\mathbf{c}_i)\| = \sum_{i=1}^k \|\mathbf{F}_{\lambda_i}(\mathbf{c}_{i-1} - \mathbf{c}_i)\| \\
&\leq \left[\max_{\lambda\in\Lambda} \|\mathbf{F}_\lambda\|\right] \sum_{i=1}^k \|\mathbf{c}_{i-1} - \mathbf{c}_i\| = \left[\max_{\lambda\in\Lambda} \|\mathbf{F}_\lambda\|\right] \|\mathbf{c} - \hat{\mathbf{c}}\|,
\end{aligned}$$

completing the proof. □

We are now ready to generalize Theorem 8.5.2 to an N-layer architecture while at the same time relaxing several of its simplifying assumptions in favor of generality.

Theorem 8.6.3. ***(Selective recruitment in multilayer hierarchical networks).*** *Consider the dynamics* (8.3)*. If*

(i) The reduced-order model (ROM)

$$\tau_1 \dot{\bar{\mathbf{x}}}_1^{\mathbb{1}} = -\bar{\mathbf{x}}_1^{\mathbb{1}} + [\mathbf{W}_{1,1}^{\mathbb{1}\mathbb{1}} \bar{\mathbf{x}}_1^{\mathbb{1}} + \mathbf{W}_{1,2}^{\mathbb{1}\mathbb{1}} h_2^{\mathbb{1}}(\mathbf{W}_{2,1}^{\mathbb{1}\mathbb{1}} \bar{\mathbf{x}}_1^{\mathbb{1}} + \mathbf{c}_2^{\mathbb{1}}) + \mathbf{c}_1^{\mathbb{1}}]^+,$$

of the first subnetwork has bounded solutions (recall $\mathbf{x}_1 \equiv \mathbf{x}_1^{\mathbb{1}}$ *since* $r_1 = 0$*);*

(ii) For all $i = 2, \ldots, N$*,*

$$\tau_i \dot{\mathbf{x}}_i^{\mathbb{1}}(t) = -\mathbf{x}_i^{\mathbb{1}}(t) + [\mathbf{W}_{i,i}^{\mathbb{1}\mathbb{1}} \mathbf{x}_i^{\mathbb{1}}(t) + \mathbf{W}_{i,i+1}^{\mathbb{1}\mathbb{1}} h_{i+1}^{\mathbb{1}}(\mathbf{W}_{i+1,i}^{\mathbb{1}\mathbb{1}} \mathbf{x}_i^{\mathbb{1}}(t) + \mathbf{c}_{i+1}^{\mathbb{1}}) + \mathbf{c}_i^{\mathbb{1}}]^+,$$

is GES towards a unique equilibrium for any $\mathbf{c}_{i+1}^{\mathbb{1}}$ *and any* $\mathbf{c}_i^{\mathbb{1}}$*;*

then there exists $\mathbf{K}_i \in \mathbb{R}^{m_i \times n_i}$ *and* $\bar{\mathbf{u}}_i : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}^{m_i}, i \in \{2, \ldots, N\}$ *such that using the feedback-feedforward control*

$$\mathbf{u}_i(t) = \mathbf{K}_i \mathbf{x}_i(t) + \bar{\mathbf{u}}_i(t), \qquad i \in \{2, \ldots, N\}, \tag{8.46}$$

we have, for any $0 < \underline{t} < \bar{t} < \infty$*,*

$$\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [\underline{t}, \bar{t}]} \|\mathbf{x}_i^{\circ}(t)\| = \mathbf{0}, \qquad \forall i \in \{2, \ldots, N\}, \tag{8.47a}$$

and

$$\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [0, \bar{t}]} \|\mathbf{x}_1^{\scriptscriptstyle 1}(t) - \bar{\mathbf{x}}_1^{\scriptscriptstyle 1}(t)\| = 0, \tag{8.47b}$$

$$\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [\underline{t}, \bar{t}]} \|\mathbf{x}_2^{\scriptscriptstyle 1}(t) - h_2^{\scriptscriptstyle 1}(\mathbf{W}_{2,1}^{\scriptscriptstyle 11} \mathbf{x}_1^{\scriptscriptstyle 1}(t) + \mathbf{c}_2^{\scriptscriptstyle 1})\| = 0, \tag{8.47c}$$

$$\vdots$$

$$\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [\underline{t}, \bar{t}]} \|\mathbf{x}_N^{\scriptscriptstyle 1}(t) - h_N^{\scriptscriptstyle 1}(\mathbf{W}_{N,N-1}^{\scriptscriptstyle 11} \mathbf{x}_{N-1}^{\scriptscriptstyle 1}(t) + \mathbf{c}_N^{\scriptscriptstyle 1})\| = 0. \tag{8.47d}$$

Proof. For any 2×2 block-partitioned matrix $\mathbf{W}$, we introduce the convenient notation $\mathbf{W}^{\ell,\mathrm{all}} \triangleq [\mathbf{W}^{\ell \scriptscriptstyle 0} \ \mathbf{W}^{\ell \scriptscriptstyle 1}]$ and $\mathbf{W}^{\mathrm{all},\ell} \triangleq [(\mathbf{W}^{\scriptscriptstyle 0 \ell})^T \ (\mathbf{W}^{\scriptscriptstyle 1 \ell})^T]^T$ for $\ell = \scriptscriptstyle 0, 1$. Further, for any $i \in \{2, \dots, N\}$, let $\mathbf{x}_{1:i} = [\mathbf{x}_1^T \ \dots \ \mathbf{x}_i^T]^T$. To begin with, let $\mathbf{K}_N$ and $\bar{\mathbf{u}}_N$ be such that

$$\mathbf{B}_N^{\scriptscriptstyle 0} \mathbf{K}_N \leq -\mathbf{W}_{N,N}^{\scriptscriptstyle 0,\mathrm{all}}, \tag{8.48a}$$

$$\bar{\mathbf{u}}_N(t) \leq -\mathbf{W}_{N,N-1}^{\scriptscriptstyle 0,\mathrm{all}} \mathbf{x}_{N-1}(t) - \mathbf{c}_N^{\scriptscriptstyle 0}, \qquad \forall t, \tag{8.48b}$$

Note that, if $m_N \geq r_N$, then (8.48a) can be satisfied with equality. Otherwise, (8.48a) can still be satisfied since all the rows of $\mathbf{B}_N^{\scriptscriptstyle 0}$ are nonzero, but may require excessive amounts of inhibition. Also, notice that $\bar{\mathbf{u}}_N$ is set by the subnetwork $N-1$, which has access to $\mathbf{x}_{N-1}(t)$ and can thus fulfill (8.48b). As a result, the nodes in $\mathbf{x}_N^{\scriptscriptstyle 0}$ are fully inhibited and evolve according to $\tau_N \dot{\mathbf{x}}_N^{\scriptscriptstyle 0} = -\mathbf{x}_N^{\scriptscriptstyle 0}$,

and the overall network dynamics become

$$
\begin{aligned}
\tau_1 \dot{\mathbf{x}}_1 &= -\mathbf{x}_1 + [\mathbf{W}_{1,1}\mathbf{x}_1 + \mathbf{W}_{1,2}\mathbf{x}_2 + \mathbf{c}_1]^+, \\
&\vdots \\
\tau_{N-1}\dot{\mathbf{x}}_{N-1} &= -\mathbf{x}_{N-1} + [\mathbf{W}_{N-1,N-1}\mathbf{x}_{N-1} + \mathbf{B}_{N-1}\mathbf{u}_{N-1} + \mathbf{W}_{N-1,N}\mathbf{x}_N \\
&\qquad + \mathbf{W}_{N-1,N-2}\mathbf{x}_{N-2} + \mathbf{c}_{N-1}]^+, \\
\epsilon_{N-1}\tau_{N-1}\dot{\mathbf{x}}_N^{\circ} &= -\mathbf{x}_N^{\circ}, \\
\epsilon_{N-1}\tau_{N-1}\dot{\mathbf{x}}_N^{\scriptscriptstyle 1} &= -\mathbf{x}_N^{\scriptscriptstyle 1} + [\mathbf{W}_{N,N}^{\scriptscriptstyle 1,\mathrm{all}}\mathbf{x}_N + \mathbf{W}_{N,N-1}^{\scriptscriptstyle 1,\mathrm{all}}\mathbf{x}_{N-1} + \mathbf{c}_N^{\scriptscriptstyle 1}]^+.
\end{aligned}
$$

Letting $\epsilon_{N-1} \to 0$, we get our first separation of timescales between $\mathbf{x}_N$ and $\mathbf{x}_{1:N-1}$, as follows. For any constant $\mathbf{x}_{N-1}$, the $\mathbf{x}_N$ dynamics is GES by assumption *(ii)* and [101, Lemma A.2]. Further, the equilibrium map $h_N = (\mathbf{0}_{r_N}, h_N^{\scriptscriptstyle 1})$ of the N'th subnetwork is globally Lipschitz by Lemmas 8.6.1 and 8.6.2, and the entire vector field of network dynamics is globally Lipschitz due to the Lipschitzness of $[\cdot]^+$. Therefore, it follows from [71, Prop 1] that for any $0 < \underline{t} < \bar{t} < \infty$,

$$
\begin{aligned}
&\lim_{\epsilon_{N-1}\to 0} \sup_{t\in[\underline{t},\bar{t}]} \|\mathbf{x}_N^{\circ}(t)\| = 0, \\
&\lim_{\epsilon_{N-1}\to 0} \sup_{t\in[\underline{t},\bar{t}]} \|\mathbf{x}_N^{\scriptscriptstyle 1}(t) - h_N^{\scriptscriptstyle 1}(W_{N,N-1}^{\scriptscriptstyle 1,\mathrm{all}}\mathbf{x}_{N-1}(t) + \mathbf{c}_N^{\scriptscriptstyle 1})\| = 0, \\
&\lim_{\epsilon_{N-1}\to 0} \sup_{t\in[0,\bar{t}]} \|\mathbf{x}_{1:N-1}(t) - \mathbf{x}_{1:N-1}^{(1)}(t)\| = 0.
\end{aligned}
$$

Here, $\mathbf{x}^{(1)}_{1:N-1}$ is the solution of the "first-step ROM"

$$\begin{aligned}
\tau_1 \dot{\mathbf{x}}^{(1)}_1 &= -\mathbf{x}^{(1)}_1 + [\mathbf{W}_{1,1}\mathbf{x}^{(1)}_1 + \mathbf{W}_{1,2}\mathbf{x}^{(1)}_2 + \mathbf{c}_1]^+, \\
&\vdots \\
\tau_{N-1} \dot{\mathbf{x}}^{(1)}_{N-1} &= -\mathbf{x}^{(1)}_{N-1} + [\mathbf{W}_{N-1,N-1}\mathbf{x}^{(1)}_{N-1} + \mathbf{W}^{\text{all},\mathfrak{1}}_{N-1,N} h^{\mathfrak{1}}_N(W^{\mathfrak{1},\text{all}}_{N,N-1}\mathbf{x}^{(1)}_{N-1}(t) + \mathbf{c}^{\mathfrak{1}}_N) \\
&\qquad + \mathbf{W}_{N-1,N-2}\mathbf{x}^{(1)}_{N-2} + \mathbf{B}_{N-1}\mathbf{u}_{N-1} + \mathbf{c}_{N-1}]^+,
\end{aligned}$$

which results from replacing $\mathbf{x}_N$ with its equilibrium value. Except for technical adjustments, the remainder of the proof essentially follows by repeating this process $N-2$ times. In particular, for $i = N-1, \ldots, 2$, let $\mathbf{K}_i$ and $\bar{\mathbf{u}}_i$ be such that

$$\begin{aligned}
\mathbf{B}^{\mathfrak{o}}_i \mathbf{K}_i &\leq -|\mathbf{W}^{\mathfrak{o},\text{all}}_{i,i}| - |\mathbf{W}^{\mathfrak{o}\mathfrak{1}}_{i,i+1}|\bar{\mathbf{F}}_{i+1}|\mathbf{W}^{\mathfrak{1},\text{all}}_{i+1,i}|, \\
\bar{\mathbf{u}}_i(t) &\leq -\mathbf{W}^{\mathfrak{o}:}_{i,i-1}\mathbf{x}_{i-1}(t) - \mathbf{c}^{\mathfrak{o}}_i, \qquad \forall t,
\end{aligned}$$

where $\bar{\mathbf{F}}_i \in \mathbb{R}^{(n_i-r_i)\times(n_i-r_i)}$ is the entry-wise maximal gain of the map $h^{\mathfrak{1}}_i$ over $\mathbb{R}^{n_i-r_i}$ (cf. Theo-

rem 8.6.4). This results in the "$(N-i)$'th-step ROM"

$$\begin{aligned}
\tau_1 \dot{\mathbf{x}}_1^{(N-i)} &= -\mathbf{x}_1^{(N-i)} + [\mathbf{W}_{1,1}\mathbf{x}_1^{(N-i)} + \mathbf{W}_{1,2}\mathbf{x}_2^{(N-i)} + \mathbf{c}_1]^+, \\
&\vdots \\
\tau_{i-1} \dot{\mathbf{x}}_{i-1}^{(N-i)} &= -\mathbf{x}_{i-1}^{(N-i)} + [\mathbf{W}_{i-1,i-1}\mathbf{x}_{i-1}^{(N-i)} + \mathbf{W}_{i-1,i}\mathbf{x}_i^{(N-i)} + \mathbf{W}_{i-1,i-2}\mathbf{x}_{i-2}^{(N-i)} + \mathbf{B}_{i-1}\mathbf{u}_{i-1} \\
&\qquad + \mathbf{c}_{i-1}]^+, \\
\epsilon_{i-1}\tau_{i-1} \dot{\mathbf{x}}_i^{(N-i)^{\mathfrak{o}}} &= -\mathbf{x}_i^{(N-i)^{\mathfrak{o}}}, \\
\epsilon_{i-1}\tau_{i-1} \dot{\mathbf{x}}_i^{(N-i)^{\mathfrak{1}}} &= -\mathbf{x}_i^{(N-i)^{\mathfrak{1}}} + [\mathbf{W}_{i,i}^{\mathfrak{1},\mathrm{all}}\mathbf{x}_i^{(N-i)^{\mathfrak{1}}} + \mathbf{W}_{i,i+1}^{\mathrm{all},\mathfrak{1}} h_{i+1}^{\mathfrak{1}}(\mathbf{W}_{i+1,i}^{\mathfrak{1},\mathrm{all}}\mathbf{x}_i^{(N-i)}(t) + \mathbf{c}_{i+1}^{\mathfrak{1}}) \\
&\qquad + \mathbf{W}_{i,i-1}^{\mathfrak{1},\mathrm{all}}\mathbf{x}_{i-1}^{(N-i)} + \mathbf{c}_i^{\mathfrak{1}}]^+.
\end{aligned}$$

Similarly to above, invoking [71, Prop 1] then ensures that

$$\begin{aligned}
&\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [\underline{t}, \bar{t}]} \|\mathbf{x}_i^{(N-i)^{\mathfrak{o}}}(t)\| = 0, \\
&\lim_{\epsilon \to \mathbf{0}} \sup_{t \in [\underline{t}, \bar{t}]} \|\mathbf{x}_i^{(N-i)^{\mathfrak{1}}}(t) - h_i^{\mathfrak{1}}(\mathbf{W}_{i,i-1}^{\mathfrak{1},\mathrm{all}}\mathbf{x}_{i-1}^{(N-i)}(t) + \mathbf{c}_i^{\mathfrak{1}})\| = 0, \\
&\lim_{\epsilon \to 0} \sup_{t \in [0, \bar{t}]} \|\mathbf{x}_{1:i-1}^{(N-i)}(t) - \mathbf{x}_{1:i-1}^{(N-i+1)}(t)\| = 0.
\end{aligned}$$

Notice that after every invocation of [71, Prop 1], the super-index inside the parenthesis increases by 1, showing one more replacement of a fast dynamics by its equilibrium state. In particular, after the last (i.e., $(N-1)$'th) invocation of [71, Prop 1], we reach $\mathbf{x}_1^{(N-1)^{\mathfrak{1}}}$, which is the same as $\bar{\mathbf{x}}_1^{\mathfrak{1}}$ in the statement. Together, these results (and sufficiently many applications of the triangle inequality and Lemma 8.6.2) ensure (8.47), completing the proof. □

Unlike the result in Theorem 8.5.2, (8.46) uses a combination of feedback and feedforward

inhibition. While using only feedforward or feedback inhibition has the advantage of a simpler implementation, their combination results in more flexibility and less conservativeness: in pure feedforward inhibition, countering local excitations requires monotone boundedness and a sufficiently large $\bar{\mathbf{u}}$ providing inhibition under the worst-case scenario, a goal that is achieved more efficiently using feedback. On the other hand, pure feedback inhibition needs to dynamically cancel local excitations at all times and is also unable to counter the effects of constant background excitation, limitations that are easily addressed when combined with feedforward inhibition.

Similar to Remark 8.5.3, assumption *(ii)* of Theorem 8.6.3 is its only critical requirement, which is both necessary and sufficient for selective inhibition. The next result relates this condition to the joint structure of the subnetworks, serving as a vital step in the practical utilization of Theorem 8.6.3.

Theorem 8.6.4. ***(Sufficient condition for existence and uniqueness of equilibria and GES in multilayer linear-threshold networks).*** *Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be a piecewise affine function of the form*

$$h(\mathbf{c}) = \mathbf{F}_\lambda \mathbf{c} + \mathbf{f}_\lambda, \qquad \forall \mathbf{c} \in \Psi_\lambda \triangleq \{\mathbf{c} \mid \mathbf{G}_\lambda \mathbf{c} + \mathbf{g}_\lambda \geq \mathbf{0}\}, \quad \forall \lambda \in \Lambda, \tag{8.49}$$

where Λ is a finite index set and $\bigcup_{\lambda\in\Lambda} \Psi_\lambda = \mathbb{R}^n$. Further, let $\bar{\mathbf{F}} \triangleq \max_{\lambda\in\Lambda} |\mathbf{F}_\lambda|$ be the matrix whose elements are the maximum of the corresponding elements from $\{|\mathbf{F}_\lambda|\}_{\lambda\in\Lambda}$. For arbitrary matrices $\mathbf{W}_\ell$, $\ell = 1, 2, 3$, if $\rho\big(|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|\big) < 1$, then the linear-threshold dynamics

$$\tau\dot{\mathbf{x}}(t) = -\mathbf{x}(t) + [\mathbf{W}_1\mathbf{x}(t) + \mathbf{W}_2 h(\mathbf{W}_3\mathbf{x}(t) + \bar{\mathbf{c}}) + \mathbf{c}]^+,$$

is GES towards a unique equilibrium for all $\bar{\mathbf{c}}$ *and* $\mathbf{c}$*.*

Proof. We use the same proof technique as in [22, Prop. 3]. For simplicity, assume that $|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|$ is irreducible (i.e., the network topology induced by it is strongly connected)[9]. Then, the left Perron-Frobenius eigenvector $\boldsymbol{\alpha}$ of $|\mathbf{W}_1|+|\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|$ has positive entries [76, Fact 4.11.4], making the map $\|\cdot\|_{\boldsymbol{\alpha}} : \mathbf{v} \to \|\mathbf{v}\|_{\boldsymbol{\alpha}} \triangleq \boldsymbol{\alpha}^T|\mathbf{v}|$ a norm on $\mathbb{R}^n$. Further, it can be shown, similar to the proof of Lemma 8.6.2, that for all $\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^n$,

$$|h(\mathbf{c}_1) - h(\mathbf{c}_2)| \leq \bar{\mathbf{F}}|\mathbf{c}_1 - \mathbf{c}_2|,$$

where the inequality is entrywise. Thus, for any $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^n$,

$$\begin{aligned}
&\Big\|[\mathbf{W}_1\mathbf{x}+\mathbf{W}_2 h(\mathbf{W}_3\mathbf{x}+\mathbf{w})+\mathbf{c}]^+ - [\mathbf{W}_1\hat{\mathbf{x}}+\mathbf{W}_2 h(\mathbf{W}_3\hat{\mathbf{x}}+\mathbf{w})+\mathbf{c}]^+\Big\|_{\boldsymbol{\alpha}} \\
&\qquad = \boldsymbol{\alpha}^T\Big|[\mathbf{W}_1\mathbf{x}+\mathbf{W}_2 h(\mathbf{W}_3\mathbf{x}+\mathbf{w})+\mathbf{c}]^+ - [\mathbf{W}_1\hat{\mathbf{x}}+\mathbf{W}_2 h(\mathbf{W}_3\hat{\mathbf{x}}+\mathbf{w})+\mathbf{c}]^+\Big| \\
&\qquad \leq \boldsymbol{\alpha}^T\Big|\mathbf{W}_1(\mathbf{x}-\hat{\mathbf{x}}) + \mathbf{W}_2(h(\mathbf{W}_3\mathbf{x}+\mathbf{w}) - h(\mathbf{W}_3\hat{\mathbf{x}}+\mathbf{w}))\Big| \\
&\qquad \leq \boldsymbol{\alpha}^T\big(|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|\big)|\mathbf{x}-\hat{\mathbf{x}}| \\
&\qquad = \rho\big(|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|\big)\boldsymbol{\alpha}^T|\mathbf{x}-\hat{\mathbf{x}}| \\
&\qquad = \rho\big(|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|\big)\|\mathbf{x}-\hat{\mathbf{x}}\|_{\boldsymbol{\alpha}}.
\end{aligned}$$

This proves that $\mathbf{x} \mapsto [\mathbf{W}_1\mathbf{x} + \mathbf{W}_2 h(\mathbf{W}_3\mathbf{x}+\mathbf{w}) + \mathbf{c}]^+$ is a contraction on $\mathbb{R}^n_{\geq 0}$ and has a unique fixed point, denoted $\mathbf{x}^*$, by the Banach Fixed-Point Theorem [113, Thm 9.23].

[9] If $|\mathbf{W}_1|+|\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|$ is not irreducible, it can be "upper-bounded" by the irreducible matrix $|\mathbf{W}_1|+|\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3| + \mu\mathbf{1}_n\mathbf{1}_n^T$, with $\mu > 0$ sufficiently small such that $\rho(|\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3| + \mu\mathbf{1}_n\mathbf{1}_n^T) < 1$. The same argument can then be employed for this upper bound.

To show GES, let $\boldsymbol{\xi}(t) \triangleq (\mathbf{x}(t) - \mathbf{x}^*)e^t$, satisfying

$$\tau\dot{\boldsymbol{\xi}}(t) = \mathbf{M}(t)\mathbf{W}\boldsymbol{\xi}(t), \tag{8.50}$$

where $\mathbf{M}(t)$ is a diagonal matrix with diagonal entries

$$m_{ii}(t) \triangleq \begin{cases} \frac{\left([\mathbf{W}_1\mathbf{x}(t)+\mathbf{W}_2 h(\mathbf{W}_3\mathbf{x}(t)+\mathbf{w})+\mathbf{c}]^+ - \mathbf{x}^*\right)_i}{\xi_i(t)} & \text{if } \xi_i(t) \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is straightforward to show that

$$|\mathbf{M}(t)| \leq |\mathbf{W}_1| + |\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|, \qquad \forall t \geq 0,$$

where the inequality is entry-wise. Then, by using [21, Lemma] (which is essentially a careful application of Gronwall-Bellman's Inequality [62, Lemma A.1] to (8.50)),

$$\|\boldsymbol{\xi}(t)\|_\alpha \leq \|\boldsymbol{\xi}(0)\|_\alpha e^{\rho(|\mathbf{W}_1|+|\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|)t} \Rightarrow \|\mathbf{x}(t)-\mathbf{x}^*\|_\alpha \leq \|\mathbf{x}(0)-\mathbf{x}^*\|_\alpha e^{-(1-\rho(|\mathbf{W}_1|+|\mathbf{W}_2|\bar{\mathbf{F}}|\mathbf{W}_3|))t},$$

establishing GES by the equivalence of norms on $\mathbb{R}^n$. □

Note that Theorem 8.6.4 applies to each layer of (8.3) separately. When put together, as-

sumption *(ii)* of Theorem 8.6.3 is satisfied if

$$\begin{aligned}
&\rho\big(|\mathbf{W}^{11}_{2,2}| + |\mathbf{W}^{11}_{2,3}|\bar{\mathbf{F}}^{1}_{3}|\mathbf{W}^{11}_{3,3}|\big) < 1,\\
&\qquad\vdots\\
&\rho\big(|\mathbf{W}^{11}_{N-1,N-1}| + |\mathbf{W}^{11}_{N-1,N}|\bar{\mathbf{F}}^{1}_{N}|\mathbf{W}^{11}_{N,N-1}|\big) < 1,\\
&\rho\big(|\mathbf{W}^{11}_{N,N}|\big) < 1,
\end{aligned} \tag{8.51}$$

where $\bar{\mathbf{F}}^{1}_{i}, i = 3, \dots, N$ is the matrix described in Theorem 8.6.4 corresponding to h^{1}_{i}, and the affine form (8.49) of h^{1}_{i} is computed recursively using Lemma 8.6.1.

8.7 Case Study: Selective Listening in Rodents

We present an application of our framework to a specific real-world example of goal-driven selective attention using measurements of single-neuron activity in the brain. Beyond the conceptual illustration of our results in Example 8.5.4 above, we argue that the cross-validation of theoretical results with real data performed here is a necessary step to make a credible case for neuroscience research and significantly enhances the relevance of the developed analysis. We have been fortunate to have access to data from a novel and carefully designed experimental paradigm [114, 115] that involves goal-driven selective listening in rodents and displays the key features of hierarchical selective recruitment noted here.

8.7.1 Description of Experiment and Data

A long standing question in neuroscience involves our capability to selectively listen to specific sounds in a crowded environment [2, 116–118]. Similar phenomena also exists in animals such as birds and rodents [116, 119]. To understand the neuronal basis of this phenomena, the work [114] has rats simultaneously presented with two sounds and trains them to selectively respond to one sound while actively suppressing the distraction from the other. In brief, the experimental procedure is as follows. In each trial, the animal simultaneously hears a white noise burst and a narrow-band warble. The noise burst may come from the left or the right while the warble may have low or high pitch, both chosen at random. Which of the two sounds (noise burst or warble) is relevant and which is a distraction depends on the "rule" of the trial: in "localization" (LC) and "pitch discrimination" (PD) trials, the animal has to make a motor choice based on the location of the noise burst (left/right) or the pitch of the warble (low/high), respectively, to receive a reward. Each rat performs several blocks of LC and PD trials during each session (with each block switching randomly between the 4 possible stimulus pairs), requiring it to quickly switch its response following the rule changes.

While the rats perform the task, spiking activity of single neurons is recorded in two brain areas: the primary auditory cortex (A1) and the medial prefrontal cortex (PFC). A1 is the first region in the cortex that receives auditory information (from subcortical areas and ears), thus forming a (relatively) low level of the hierarchy. PFC, on the other hand, is composed of multiple regions that form the top of the hierarchy, and serve functions such as imagination, planning, decision-making, and attention [120]. Overall, spike times of 211 well-isolated and reliable neurons are recorded in 5 rats, 112 in PFC and 99 in A1, see [115].

Using statistical analysis, it was shown in [114] that (i) the rule of the trial and the stimulus sounds are more strongly encoded by PFC and A1 neurons, respectively, (ii) electrical disruption of PFC significantly impairs task performance, and (iii) PFC activity temporally precedes A1 activity. These findings are all consistent with a model where PFC controls the activity of A1 based on the trial rule in order to achieve GDSA. We next build on these observations to define an appropriate network structure and rigorously analyze it using HSR.

8.7.2 Choice of Neuronal Populations

In order to form meaningful populations among the recorded neurons, we perform three classifications of them:

(i) first, we classify the neurons into excitatory and inhibitory. The standard procedure for this classification is based on the spike waveform of each neuron: excitatory neurons have slower and wider spikes while inhibitory neurons have faster and narrower ones [121]. Since the spike waveforms of neurons are high-dimensional (24 samples per waveform, recorded at 30^{kHz}), we first perform a t-SNE dimensionality reduction and then used k-means clustering to identify the 174 excitatory and 37 inhibitory neurons (Figure 8.5(a)). These results conform with spike width difference of excitatory and inhibitory neurons (Figure 8.5(b)) and the fact that about 80% of cortical neurons are excitatory.

(ii) Second, we classify the PFC neurons based on their rule-encoding (RE) property. This classification was also done in [114], so we briefly review the method for completeness. A neuron is said to have a RE property if its firing rate is significantly different during the LC and PD trials *before the stimulus onset*. In the absence of stimulus, any such difference is attributable to the

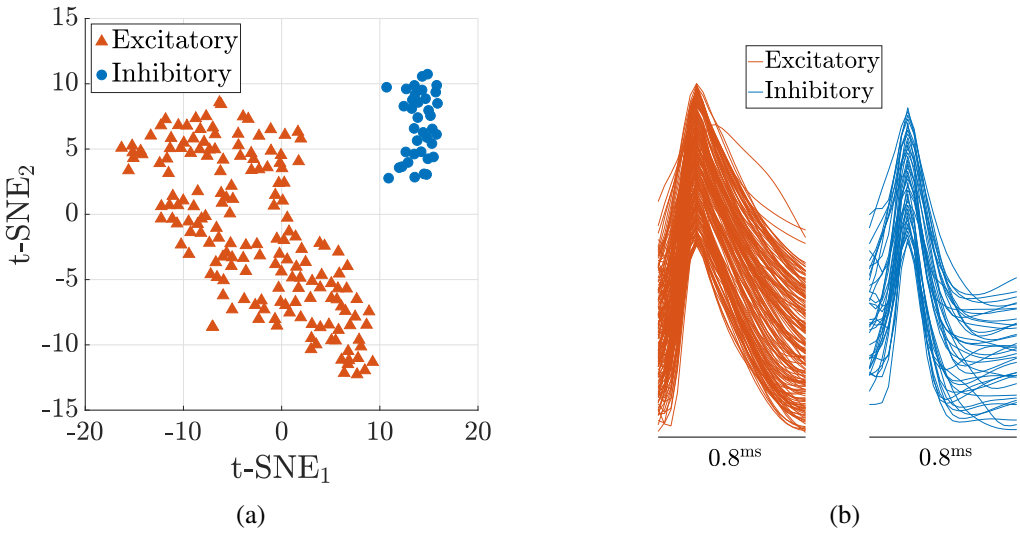


Figure 8.5: Excitatory/inhibitory classification of neurons. (a) Clustering of neuronal spike waveforms in the two-dimensional space arising from t-SNE dimensionality reduction. The excitatory and inhibitory neurons form clearly distinct clusters that are identified using the k-means clustering algorithm. (b) The spike waveforms of clustered neurons. As expected, the inhibitory neurons have faster and narrower spikes, verifying the outcome of the t-SNE + k-means clustering procedure.

animal's knowledge of the task rule (i.e., which upcoming stimulus it has to pay attention to in order to get the reward). Thus, it is standard to assess neurons' RE property during the *hold period*, namely, the time interval between the initiation of each trial and the stimulus onset of that trial. Therefore for each PFC neuron, we calculate its mean firing rate during the hold period of each trial and then statistically compare the results for LC and PD trials ($p < 0.05$, one-sided MWW rank-sum test). Among the 112 neurons in PFC, 40 encoded for LC while 44 encoded for PD (the remaining PFC neurons with no RE property are discarded from further analysis).

(iii) Finally, we classify the A1 neurons based on their evoked response (ER) property. In contrast to RE, a neuron has an ER property if its firing rate is significantly different in response to the white noise (LC stimulus) and warble (PD stimulus) *after the stimulus onset*. Since the white noise and warble are always presented simultaneously, it is not possible to make such a distinction based on normal trials. However, before each LC or PD block, the animal is only presented with the

respective stimulus for a few *cue trials* (which is how the animal realizes the rule change). Thus, for each A1 neuron, we compare its mean firing rate during the *listening period* of each cue trial (namely, the interval between the stimulus onset and the time that the animal commits to a decision) and statistically compare the distribution of the results for LC and PD cue trials ($p < 0.05$, one-sided MWW rank-sum test). Among the 99 A1 neurons, 21 had an ER for LC while another 21 had an ER for PD (the remaining A1 neurons with no ER property are discarded from further analysis).

Remark 8.7.1. ***(RE vs. ER detection).*** It is noteworthy that a smaller fraction of PFC and A1 neurons also have ER and RE properties, respectively. However, we know from systems neuroscience that these properties most likely arise from the PFC-A1 reciprocal connection, as auditory and attention/decision making information disseminate from A1 and PFC, respectively. This motivates our classification of A1 and PFC neurons based on ER and RE, respectively, and their reciprocal connection in the proposed network structure below. Further, we note that our ER detection has a difference with respect to [114]. In [114], the difference between the post-stimulus and pre-stimulus firing rates (the latter being RE) is used for ER detection, with the motivation of removing the portion of post-stimulus firing rate that is due to RE (and thus independent of stimulus). However, this relies on the strong assumption that the RE and ER responses superimpose linearly, which we found likely not to be true based on the statistical analysis of the present dataset, perhaps since RE drives many neurons close to their maximum firing rate, leaving little room for *additional* ER. We thus use the complete post-stimulus firing rate for ER detection, as above. □

As a result of the classifications described above, we group the neurons into 8 populations based on the PFC/A1, excitatory/inhibitory, and LC/PD classifications. The firing rate of each population (as a function of time) is then calculated as follows. For each neuron and each trial, the

interval $[-10, 10]$ (with time 0 corresponding to stimulus onset) is decomposed into 100^{ms}-wide bins and the firing rate of each bin (spike count divided by bin width) is assigned to the bin's center time. This time series is then averaged over all trials with the same stimulus pair and all the neurons within each population, and finally smoothed with a Gaussian kernel with 1^{s} standard deviation. This results in one firing rate time series for each neuron and each stimulus pair.

For the purpose of this work, we limit our choice of stimulus pairs as follows. Recall that each of LC and PD blocks contains 4 stimulus pairs (left-low, left-high, right-low, right-high). In each block, these 4 pairs are divided into two *go* and two *no-go* pairs. When the animal hears a go stimulus pair, his correct response is to go to a nearby food port to receive his reward. In no-go trials, on the other hand, the correct response is simply inaction (action is punished by a delay before the animal can do the next trial). Due to strong motor and reward-consumption artifacts in go trials (cf. [114, Fig. S4]), we limit our analysis here to no-go trials. Further, we also discard the no-go stimulus pair that is shared between LC and PD blocks, since the correct decision (no-go) is independent of the block and thus does not require selective attention. Therefore, our analysis hereafter only involves one firing rate time series for each neuronal population in each block.

8.7.3 Network Binary Structure

We next describe our proposed network binary structure[10]. In each of the two regions (PFC and A1), the 4 populations are connected to each other according to the following physiological properties (see [122–124] and [124–126] for evidence of these properties in PFC and A1, resp.):

(i) each excitatory population projects to (i.e., makes synapses on) the inhibitory population with

[10] We here make a distinction between the binary structure of the network, composed of only the connectivity pattern among nodes, and its full structure, that also includes the connection weights.

the same LC/PD preference (RE in PFC or ER in A1);

(ii) neurons in each excitatory population project to each other (captured by the excitatory self-loops in Figure 8.6).

(iii) each inhibitory population projects to the populations (both excitatory and inhibitory) with *opposite* LC/PD preference (the so-called *lateral inhibition* property);

While within-region connections are both excitatory and inhibitory, between-region connections in the cortex (including PFC and A1) are almost entirely excitatory, completing the binary structure shown in Figure 8.6.

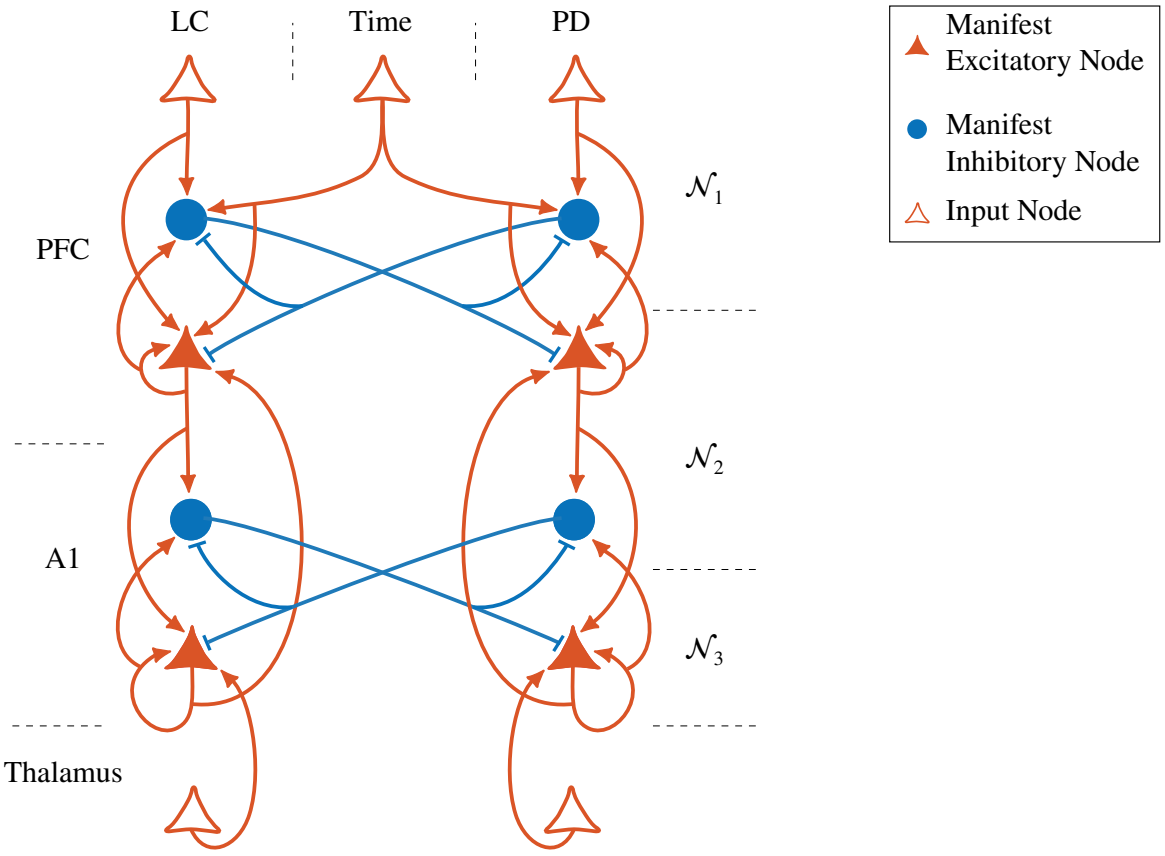


Figure 8.6: The proposed network binary structure. The physiological region, hierarchical layer, and encoding properties of nodes are indicated on the left, right, and above the figure, respectively.

Hierarchical Structure

To apply the HSR framework to the network of Figure 8.6, we still need to assign the nodes to hierarchical layers. This assignment is in general arbitrary except for two critical requirements, (i) the existence of timescale separation between layers and (ii) the existence of both excitatory and inhibitory projections from any layer to the layer below (to allow for simultaneous inhibition and recruitment). The trivial choice here is to consider each region as a layer, which also satisfies (i) (since PFC has slower dynamics than A1) but not (ii) (since there would be no inhibitory connection between regions). We thus propose an alternative 3-layer choice, as shown in Figure 8.6.[11] This choice clearly satisfies (ii), and we next show that it also satisfies (i).

Computation of Timescales

To assess the intrinsic timescales of each population, we employ the common method in neuroscience based on the decay rate of the correlation coefficient [49,50]. In brief, for each neuron ℓ, we partition the time window *before* the stimulus onset[12] into small bins (200^{ms}-wide here) and compute the smoothed mean firing rate of this neuron during each bin and each trial. This yields a set $\{r^{\ell}_{i,k}\}_{i,k,\ell}$, where $r^{\ell}_{i,k}$ denotes the mean firing rate of neuron ℓ in the k'th time bin of trial i. The Pearson correlation coefficient between two time bins k_1 and k_2 is estimated as

$$\rho^{\ell}_{k_1,k_2} = \frac{\sum_i (r^{\ell}_{i,k_1} - \bar{r}^{\ell}_{k_1})(r^{\ell}_{i,k_2} - \bar{r}^{\ell}_{k_2})}{\sqrt{\sum_i (r^{\ell}_{i,k_1} - \bar{r}^{\ell}_{k_1})^2 \sum_i (r^{\ell}_{i,k_2} - \bar{r}^{\ell}_{k_2})^2}} \in [-1, 1],$$

[11] The bottom-most layer $\mathcal{N}_4$ represents "external" inputs from sub-cortical areas. Since we have no recordings from these areas, we do not consider any dynamics for $\mathcal{N}_4$ and accordingly do not include it in HSR analysis.

[12] In general, the time interval used for timescale estimation should not include stimulus presentation in order to reduce the effects of external factors on the internal neuronal dynamics.

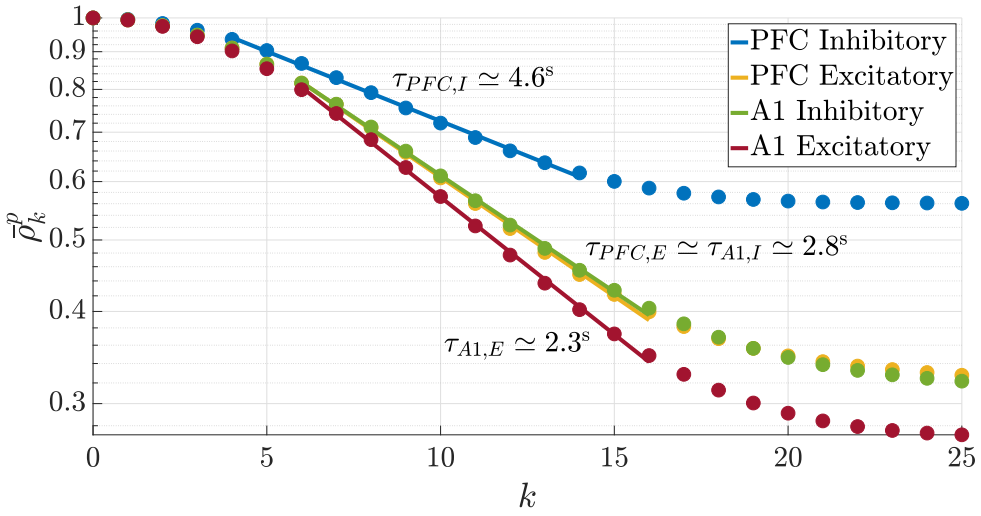


Figure 8.7: Timescale separation among the layers $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$ in Figure 8.6. The circles illustrate the values of the average auto-correlation coefficient $\bar{\rho}_k^p$ as a function of time lag k, whereas the lines represent the best exponential fit over the range of time lags where each $\bar{\rho}_k^p$ -ecays exponentially (note the logarithmic scale on the y-axis).

where $\bar{r}_k^\ell$ is the average of $r_{i,k}^\ell$ across all the trials for neuron ℓ. Let ρ_k^ℓ be the average of ρ_{k_1,k_2}^ℓ over all k_1, k_2 such that $|k_1 - k_2| = k$ and $\bar{\rho}_k^p$, for any population p, be the average of ρ_k^ℓ for all the neurons ℓ in the population p. Figure 8.7 shows this function for populations of excitatory and inhibitory neurons in PFC and A1 (we do not split the neurons based on their LC/PD preference because it is not relevant for timescale separation). Fitting $\bar{\rho}_k^p$ by an exponential function of the form $Ae^{-k/\tau}$ gives an estimate of the intrinsic timescale τ of this population, which becomes exact for spikes generated by a Poisson point process under certain regularity conditions [49]. Here, we use the range of k values for which the decay of $\bar{\rho}_k^p$ is approximately exponential for calculating the fit. As seen in Figure 8.7, there is a clear timescale separation between the layer of A1 excitatory neurons, the layer of A1 inhibitory and PFC excitatory neurons, and the layer of PFC inhibitory neurons, satisfying the requirement (i) above.[13]

[13] Also note that this method inherently underestimates the timescale separation between layers due to the mutual dynamical interactions between them.

Exogenous Inputs and Latent Nodes

The last step in specifying the binary structure of the network involves the exogenous inputs to the prescribed neuronal populations (nodes). Clearly, nodes at the bottom layer (layer 3) receive auditory inputs from subcortical areas which we represent as two input signals x_1^4 and x_2^4 coming from layer 4 and corresponding to the white noise and warble, respectively. Both these signals are constructed by smoothing a square pulse that equals 1 during stimulus presentation and 0 otherwise with the same Gaussian window used for smoothing the firing rate time-series.

The choice of the inputs to the PFC populations is more intricate. PFC is itself composed of a complex network of several regions, each involved in some aspects of high-level cognitive functions. The RE properties of the recorded PFC populations is only one outcome of such complex PFC dynamics that also host the animal's overall understanding of how the task works, his perception of time, etc. In order to capture the effects of such unrecorded PFC dynamics, we consider 3 additional excitatory PFC populations, as follows. Two input populations x_3^1 and x_4^1 simply encode the rule of each block[14]:

$$x_3^1 \equiv \begin{cases} 1, & \text{if in LC block,} \\ 0, & \text{if in PD block,} \end{cases} \qquad x_4^1 \equiv \begin{cases} 0, & \text{if in LC block,} \\ 1, & \text{if in PD block.} \end{cases}$$

Populations with such a sustained constant activity only as a function of task parameters are indeed observed during GDSA in PFC [127]. The third additional PFC population encodes the time relative to the stimulus onset, which is critical for the functioning of the recorded PFC populations. Among

[14]Note that this static response is different from, and much simpler than, the RE of the recorded PFC neurons, which is greatly dynamic.

the various forms of encoding time, we consider a population x_5^1 with firing rate

$$x_5^1(t) = \begin{cases} |t_0| - t & t \in [t_0, 0), \\ 0 & t \in (0, t_f], \end{cases}$$

where $[t_0, t_f] = [-7, 7]$ is the duration of each trial, since populations with such activity patterns have been observed in PFC [128].[15] Since these three populations have very slow dynamics but are excitatory, following the same logic as before, we position them in the layer 1 together with the recorded inhibitory PFC populations x_1^1, x_2^1.

Finally, to capture the effects of the large populations of neurons whose activity is not recorded, we consider one *latent* node for each of the 8 *manifest* nodes in the network[16] with the same in- and out-neighbors as their respective manifest node (the latent nodes are not displayed in Figure 8.6 to avoid cluttering the plot of the network structure). We let $\{x_{1,j}\}_{j=6,7}$, $\{x_{2,j}\}_{j=5}^{8}$, and $\{x_{3,j}\}_{j=3,4}$ denote these nodes in $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_3$, respectively.

8.7.4 Identification of Network Parameters

Having established the binary structure of the network, we next seek to determine its unknown parameters $\mathbf{W}^{i,j}$. While there are physiological methods for measuring the synaptic weight between a pair of neurons in vitro, they are not applicable in vivo and thus not available for our dataset. Also, our nodes consist of several neurons, making their aggregate synaptic weight an abstract quantity. Therefore, we resort to system identification/machine learning techniques to "learn"

[15] Even though both [127] and [128] involve primates, populations with similar activity patterns are expected to exist in rodents.

[16] A node is called *manifest* if its activity is recorded during the experiment and *latent* otherwise.

the structure of the network given its input-output signals. For this purpose, the choice of objective function is crucial, for which we propose

$$
\begin{aligned}
f(z) &= f_{\text{SSE}}(z) + \gamma_1 f_{\text{corr}}(z) + \gamma_2 f_{\text{var}}(z), \qquad (8.52)\\
f_{\text{SSE}}(z) &= \sum_{\ell=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{n_{m,i}}\sum_{k}(\hat{x}_{i,j}(kT;\ell) - x_{i,j}(kT;\ell))^2,\\
f_{\text{corr}}(z) &= 1 - \frac{1}{2n_m}\sum_{\ell=1}^{2}\frac{1}{n_m}\sum_{i=1}^{3}\sum_{j=1}^{n_{m,i}}\frac{1}{K-1}\sum_{k}\frac{(\hat{x}_{i,j}(kT;\ell) - \hat{\mu}_{i,j,\ell})(x_{i,j}(kT;\ell) - \mu_{i,j,\ell})}{\hat{\sigma}_{i,j,\ell}\sigma_{i,j,\ell}},\\
f_{\text{var}}(z) &= \Big(\sum_{\ell=1}^{2}\sum_{i=1}^{3}\sum_{j=1}^{n_{m,i}}(\hat{\sigma}_{i,j,\ell} - \sigma_{i,j,\ell})^4\Big)^{1/4},
\end{aligned}
$$

where,

– z is the vector of all unknown network parameters consisting of not only the synaptic weights but also the time constants τ_i, the background inputs $\mathbf{c}_i$, and the initial states $\mathbf{x}_i(0), i = 1, 2, 3$;

– $n_{m,i}$ is the number of manifest nodes in layer i (so $n_{m,1} = 2, n_{m,2} = 4, n_{m,3} = 2$) and $n_m = 8$ is the total number of manifest nodes;

– $x_{i,j}(t;\ell)$ is the measured state of j'th node in the i'th layer in response to the ℓ'th stimulus at time t (where $\ell = 1$ indicates the LC block and $\ell = 2$ the PD block) and $\hat{x}_{i,j}(t;\ell)$ is its model estimate;

– $T = 0.1$ is the sampling time; and

– $\mu_{i,j,\ell}, \sigma_{i,j,\ell}, \hat{\mu}_{i,j,\ell}, \hat{\sigma}_{i,j,\ell}$ are the means and standard deviations of $x_{i,j}(\cdot;\ell)$ and $\hat{x}_{i,j}(\cdot;\ell)$, respectively.

The rationale behind the objective function (8.52) is as follows. $f_{\text{SSE}}(z)$ is the standard sum of squared error (SSE) function. In HSR, an important property of nodal state trajectories is the sign of their derivatives, which *transiently* indicate recruitment (positive derivative) or inhibition (negative derivative). This is captured by the average correlation coefficient $f_{\text{corr}}(z)$, which is added to $f_{\text{SSE}}(z)$ to enforce similar recruitment and inhibition patterns between measured states and their estimates. Nevertheless, correlation coefficient between a pair of signals is invariant to the amount of variation in them, requiring us to add the third term $f_{\text{var}}(z)$. The use of 4-norm in $f_{\text{var}}(z)$ particularly weights the nodes with large standard deviation mismatches. We use $\gamma_1 = 250$ and $\gamma_2 = 150$ to approximately balance the size of the 3 terms in f.

The objective function f is highly nonconvex and we thus use the GlobalSearch algorithm from the MATLAB Optimization Toolbox to minimize it. Figure 8.8 shows the manifest nodal states as well as their best model estimates. In order to quantify the similarity between these states and their estimates, we use the standard R^2 measure given by

$$R^2 = 1 - \frac{\sum_{\ell,i,j,k}(x_{i,j}(kT;\ell) - \hat{x}_{i,j}(kT;\ell))^2}{\sum_{\ell,i,j,k}(x_{i,j}(kT;\ell) - \mu_{i,j,\ell})^2} \simeq 92.7\%.$$

This high value is indeed remarkable, especially given the very small size of the network and the limited availability of measurements in the experiment.

8.7.5 Concurrence of the Identified Network with Analysis

To conclude, we verify here whether the identified network structure satisfies the requirements of the HSR framework in terms of timescale separation and stability. Regarding the former,

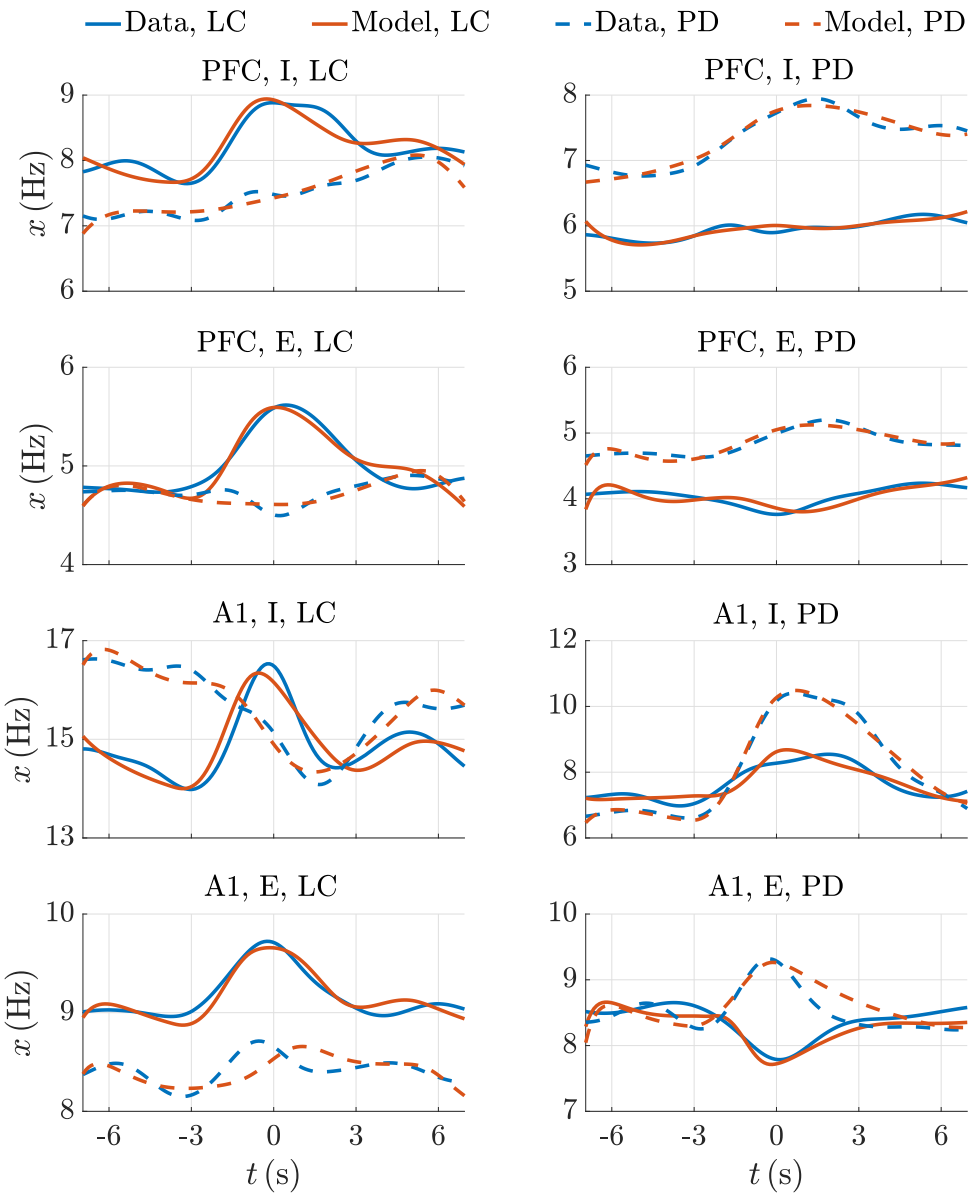


Figure 8.8: State trajectories of manifest nodes in the network of Figure 8.6 (blue: measured, red: model estimate). $t = 0$ indicates stimulus onset. Solid and dashed lines correspond to LC and PD blocks, respectively. The description of each node is indicated above its corresponding panel. The LC/PD in the legend refers to the trial rule, while the LC/PD above each panel refers to the preference of that particular node.

the identified time constants are given by

$$\tau_1 = 4.70, \qquad \tau_2 = 2.33, \qquad \tau_3 = 1.07,$$

yielding an almost twofold separation of timescales conforming to Figure 8.7. Regarding stability, we have to consider the LC and PD blocks separately (as the definition of task-relevant (^{1}) and task-irrelevant ($^{\circ}$) nodes changes according to the block).

In the LC block, the (manifest) LC nodes are task-relevant and the (manifest) PD nodes are task-irrelevant. Therefore, under this condition,

$$W_{3,3}^{11} = 0.17, \quad \mathbf{W}_{3,2}^{11} = \begin{bmatrix} 6.7 \times 10^{-3} & 0 \end{bmatrix}, \quad \mathbf{W}_{2,2}^{11} = \begin{bmatrix} 0.42 & 0 \\ 0.96 & 0 \end{bmatrix}, \quad \mathbf{W}_{2,3}^{11} = 10^{-2} \begin{bmatrix} 6 \\ 2.5 \end{bmatrix}.$$

It is then straightforward to see that

$$h_3^1(c_3^1) = \begin{cases} 0 & ; \quad c_3^1 \leq 0 \\ c_3^1/(1 - W_{3,3}^{11}) & ; \quad c_3^1 \geq 0 \end{cases} \Rightarrow \bar{F}_3^1 = \frac{1}{1 - W_{3,3}^{11}}.$$

Therefore,

$$\rho(|W_{3,3}^{11}|) = 0.17 < 1,$$

$$\rho\big(|\mathbf{W}_{2,2}^{11}| + |\mathbf{W}_{2,3}^{11}|\bar{F}_3^1|\mathbf{W}_{3,2}^{11}|\big) = \rho\Big(\begin{bmatrix} 0.42 & 0 \\ 0.96 & 0 \end{bmatrix}\Big) = 0.42 < 1,$$

satisfying the sufficient conditions for GES in (8.51). Similarly, in the PD block, we have

$$W_{3,3}^{11} = 0.14 < 1, \quad \mathbf{W}_{3,2}^{11} = \begin{bmatrix} 0.36 & 0 \end{bmatrix}, \quad \mathbf{W}_{2,2}^{11} = 10^{-2} \begin{bmatrix} 7.6 & 0 \\ 2.2 & 0 \end{bmatrix}, \quad \mathbf{W}_{2,3}^{11} = \begin{bmatrix} 0.13 \\ 0.95 \end{bmatrix},$$

$$\rho\big(|\mathbf{W}_{2,2}^{11}| + |\mathbf{W}_{2,3}^{11}|\bar{F}_3^1|\mathbf{W}_{3,2}^{11}|\big) = \rho\Bigg(\begin{bmatrix} 0.13 & 0 \\ 0.42 & 0 \end{bmatrix}\Bigg) = 0.13 < 1,$$

also satisfying the GES conditions of (8.51).

Given the concurrence between the identified network structure and the hypotheses of our results, Theorems 8.5.2 and 8.6.3 provide strong analytical support to explain the conclusions drawn in [114, 115] from experimental data and statistical analysis. We believe HSR constitutes a rigorous framework for the analysis of the multiple-timescale network interactions underlying GDSA, complementing the conventional statistical and computational analyses in neuroscience.

Appendix

8.A Auxiliary Results

Here we provide auxiliary results that are used in the proofs of main results of the chapter. The following result is used in the proof of Theorem 8.3.4 on the EUE for the dynamics (8.5).

Lemma 8.A.1. *Consider a matrix* $\bar{\mathbf{M}} \in \mathbb{R}^{n\times n}$ *with the block form*

$$\bar{\mathbf{M}} = \begin{bmatrix} \mathbf{I}_{n_1} & \star & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma} & \mathbf{0} \\ \mathbf{0} & \star & \mathbf{I}_{n_4} \end{bmatrix}, \tag{8.53}$$

where $\star$ *means an arbitrary block and* $\mathbf{\Gamma} \in \mathbb{R}^{n_{23}\times n_{23}}$. *Then, there exists nonzero* $\mathbf{y} \in \mathbb{R}^n$ *such that* $\mathbf{x} \triangleq \bar{\mathbf{M}}\mathbf{y}$ *and* $\mathbf{y}$ *belong to the same orthant(s) if and only if there exists nonzero* $\mathbf{y}_{23} \in \mathbb{R}^{n_{23}}$ *such that* $\mathbf{x}_{23} \triangleq \mathbf{\Gamma}\mathbf{y}_{23}$ *and* $\mathbf{y}_{23}$ *belong to the same orthant(s) or* $\mathbf{y}_{23} = \mathbf{0}$, *where* $\mathbf{x}_{23}$ *and* $\mathbf{y}_{23}$ *denote the middle* n_{23}*-dimensional sub-vectors of* $\mathbf{x}$ *and* $\mathbf{y}$, *respectively.*

Proof. It follows from (8.53) that

$$\bar{\mathbf{M}} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_{23} \\ \mathbf{y}_4 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 + \star \cdot \mathbf{y}_{23} \\ \mathbf{\Gamma}\mathbf{y}_{23} \\ \mathbf{y}_4 + \star \cdot \mathbf{y}_{23} \end{bmatrix}.$$

Therefore, the ($\Rightarrow$) implication is immediate. For the ($\Leftarrow$) implication, note that if $\mathbf{y}_{23}, \mathbf{\Gamma}\mathbf{y}_{23} \neq \mathbf{0}$ belong to the same orthant(s), then choosing $\mathbf{y}_1 \in \{\pm c\}^{n_1}, \mathbf{y}_4 \in \{\pm c\}^{n_4}$ with sufficiently large $c > 0$ puts $\bar{\mathbf{M}}\mathbf{y}$ in the same orthant(s) as $\mathbf{y}$. If $\mathbf{y}_{23} = \mathbf{0}$, then $\bar{\mathbf{M}}\mathbf{y} = \mathbf{y}$ and the result is trivial. □

Finally, the following result is used in the proof of Theorem 8.4.2.

Lemma 8.A.2. ***(GES of cascaded interconnections).*** *Consider the cascaded dynamics*

$$\begin{aligned} \tau\dot{\mathbf{x}}^{\circ} &= -\mathbf{x}^{\circ}, \\ \tau\dot{\mathbf{x}}^{1} &= -\mathbf{x}^{1} + [\mathbf{W}^{1\circ}\mathbf{x}^{\circ} + \mathbf{W}^{11}\mathbf{x}^{1} + \tilde{\mathbf{p}}^{1}]^{+}, \end{aligned} \tag{8.54}$$

where $\mathbf{x}^{\circ} \in \mathbb{R}^r$ *and* $\mathbf{x}^{1} \in \mathbb{R}^{n-r}$. *If* $\mathbf{W}^{11}$ *is such that*

$$\tau \dot{\mathbf{x}}^{1} = -\mathbf{x}^{1} + [\mathbf{W}^{11}\mathbf{x}^{1} + \tilde{\mathbf{p}}^{1}]^{+}, \tag{8.55}$$

is GES for any constant $\tilde{\mathbf{p}}^{1} \in \mathbb{R}^{n-r}$, *then the whole dynamics* (8.54) *is also GES for any constant* $\tilde{\mathbf{p}}^{1}$.

Proof. We only prove the result for $\tilde{\mathbf{p}}^{1} = \mathbf{0}$. This is without loss of generality, since for $\tilde{\mathbf{p}}^{1} \neq \mathbf{0}$, we can apply the change of variables $\boldsymbol{\xi} = \mathbf{x} - h([\mathbf{0}^T \ (\tilde{\mathbf{p}}^{1})^T]^T)$ and shift the equilibrium to the origin. Since (8.55) is GES, Theorem 8.B.1 guarantees that there exists $\mathbf{x}^{1} \mapsto V^{1}(\mathbf{x}^{1})$ such that

$$c_1\|\mathbf{x}^{1}\|^2 \leq V^{1}(\mathbf{x}^{1}) \leq c_2\|\mathbf{x}^{1}\|^2, \tag{8.56a}$$

$$\left\| \frac{\partial V^{1}(\mathbf{x}^{1})}{\partial \mathbf{x}^{1}} \right\| \leq c_3\|\mathbf{x}^{1}\|, \tag{8.56b}$$

for some $c_1, c_2, c_3 > 0$, and, if $\mathbf{x}^{1}(t)$ is the solution of (8.55),

$$\tau \frac{d}{dt} V^{1}(\mathbf{x}^{1}(t)) \leq -c_4\|\mathbf{x}^{1}\|^2, \tag{8.56c}$$

for some $c_4 > 0$. Since $[\cdot]^{+}$ is Lipschitz continuous, it follows from (8.56b) and (8.56c) that if $\mathbf{x}^{1}(t)$ is the solution of (8.54),

$$\begin{aligned} \tau \frac{d}{dt} V^{1}(\mathbf{x}^{1}(t)) &\leq -c_4\|\mathbf{x}^{1}\|^2 + c_3\|\mathbf{x}^{1}\|\|\mathbf{W}^{1\circ}\mathbf{x}^{\circ}\| \\ &\leq -\frac{c_4}{2}\|\mathbf{x}^{1}\|^2 + \frac{c_3^2\|\mathbf{W}^{1\circ}\|^2}{2c_4}\|\mathbf{x}^{\circ}\|^2, \end{aligned}$$

where the second inequality follows from Young's inequality [129]. Now, let

$$V(\mathbf{x}) = \frac{c_3^2 \|\mathbf{W}^{1o}\|^2}{2c_4} \|\mathbf{x}^o\|^2 + V^1(\mathbf{x}^1).$$

It is straightforward to verify that V satisfies all the assumptions of [62, Thm 4.10] with $a = 2$, completing the proof. □

8.B A Converse Lyapunov Theorem for GES Switched-Affine Systems

The existence of a converse Lyapunov function for asymptotically/exponentially stable switched linear systems has been extensively studied for the case of time-dependent (arbitrary) switching, see, e.g. [29, 130–133] and references therein. Similar results, however, are missing for state-dependent switching. In this appendix, we prove a converse Lyapunov theorem for continuous GES switched affine systems with state-dependent switching that is used in both Parts I and II of this work via [101, Lemma A.2]. The considered dynamics are general and subsume the linear-threshold dynamics of interest to us.

Theorem 8.B.1. ***(Converse Lyapunov theorem for GES switched affine systems).*** *Consider a*

state-dependent switched affine system of the form

$$\begin{aligned} \tau\dot{\mathbf{x}} &= f(\mathbf{x}), & \mathbf{x}(0) &= \mathbf{x}_0, \\ f(\mathbf{x}) &= \mathbf{A}_\lambda \mathbf{x} + \mathbf{b}_\lambda, & \forall \mathbf{x} &\in \Omega_\lambda = \{\mathbf{x} \in D \mid \mathbf{N}_\lambda \mathbf{x} + \mathbf{p}_\lambda \leq \mathbf{0}\}, \\ & & \forall \lambda &\in \Lambda, \end{aligned} \tag{8.57}$$

where Λ is a finite index set, $\mathbf{A}_\lambda$ is nonsingular for all $\lambda \in \Lambda$, $D = \bigcup_{\lambda\in\Lambda} \Omega_\lambda \subseteq \mathbb{R}^n$ is an (open) domain, and $\{\Omega_\lambda\}_{\lambda\in\Lambda}$ have mutually disjoint interiors. Assume that f is continuous. If (8.57) *is GES towards a unique equilibrium $\mathbf{x}^*$, then there exists a C^∞-function $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}$ and positive constants c_1, c_2, c_3, c_4 such that for all $\mathbf{x} \in D$,*

$$c_1\|\mathbf{x} - \mathbf{x}^*\|^2 \leq V(\mathbf{x}) \leq c_2\|\mathbf{x} - \mathbf{x}^*\|^2, \tag{8.58a}$$

$$\frac{\partial V}{\partial \mathbf{x}} f \leq -c_3\|\mathbf{x} - \mathbf{x}^*\|^2, \tag{8.58b}$$

$$\left\| \frac{\partial V}{\partial \mathbf{x}} \right\| \leq c_4\|\mathbf{x} - \mathbf{x}^*\|. \tag{8.58c}$$

Proof. We structure the proof in three steps: (i) showing that the solutions of (8.57) are continuously differentiable with respect to $\mathbf{x}_0$ *along its trajectories*, (ii) construction of a (not necessarily smooth) Lyapunov-like function that satisfies (8.58) along the trajectories of (8.57), and (iii) construction of V from this Lyapunov-like function (smoothening). We only prove the result for $\mathbf{x}^* = \mathbf{0}$ as the general case can be reduced to it with the change of variables $\mathbf{x} \leftarrow \mathbf{x} - \mathbf{x}^*$.

(i) Let $\psi(t; \mathbf{x}_0)$ denote the unique solution of (8.57) at time $t \in \mathbb{R}$ (note that we let $t < 0$). In this step, we prove that ψ is continuously differentiable with respect to $\mathbf{x}_0$ on D *if* $\mathbf{x}_0$ moves along

ψ. Precisely, that

$$\frac{\partial}{\partial \tau}\psi(t;\psi(\tau;\mathbf{x}_0)) \text{ exists and is continuous at } \tau = 0, \tag{8.59}$$

for all $\mathbf{x}_0 \in D$. First, assume that $\mathbf{x}_0 \notin H$, where $H \subset D$ is the union of all the switching hyperplanes.[17] Thus, $\mathbf{x}_0$ belongs to the interior of a switching region, say Ω_{λ_1}. Let $\{\lambda_j\}_{j=1}^{J}$, with $J = J(t) \geq 1$, be the indices of the regions visited by $\psi(\tau;\mathbf{x}_0)$ during $\tau \in [0,t]$. With a slight abuse of notation, let $\mathbf{A}_j \triangleq \mathbf{A}_{\lambda_j}$ and $\mathbf{b}_j \triangleq \mathbf{b}_{\lambda_j}$, for $j = 1,\ldots,J$. Then,

$$\psi(\tau;\mathbf{x}_0) = \begin{cases} e^{\mathbf{A}_1\tau}(\mathbf{x}_0 + \mathbf{A}_1^{-1}\mathbf{b}_1) - \mathbf{A}_1^{-1}\mathbf{b}_1; & \tau \in [0,t_1], \\ e^{\mathbf{A}_2(\tau-t_1)}(\psi(t_1;\mathbf{x}_0) + \mathbf{A}_2^{-1}\mathbf{b}_2) - \mathbf{A}_2^{-1}\mathbf{b}_2; & \tau \in [t_1,t_2], \\ \quad\vdots & \\ e^{\mathbf{A}_J(\tau-t_{J-1})}(\psi(t_{J-1};\mathbf{x}_0) + \mathbf{A}_J^{-1}\mathbf{b}_J) - \mathbf{A}_J^{-1}\mathbf{b}_J; & \tau \in [t_{J-1},t], \end{cases} \tag{8.60}$$

where $t_j = t_j(\mathbf{x}_0)$ is the time at which $\psi(\tau;\mathbf{x}_0)$ crosses the boundary between Ω_{λ_j} and $\Omega_{\lambda_{j+1}}$. This expression for ψ is valid for all $\mathbf{x}$ near $\mathbf{x}_0$ that undergo the same sequence of switches. To be precise, let $S \subset D$ be the set of points that lie at the intersection of two or more switching hyperplanes and

$$S_{(-\infty,0]} = \{\mathbf{x} \in D \mid \exists t \in [0,\infty) \quad \text{s.t.} \quad \psi(t;\mathbf{x}) \in S\}.$$

In words, $S_{(-\infty,0]}$ is the set of all points that, when evolving according to (8.57), will pass through S at some point in time. Since S is composed of a finite number of affine manifolds of dimensions

[17] Recall that for each λ, each row of $\mathbf{N}_\lambda\mathbf{x} + \mathbf{p}_\lambda = \mathbf{0}$ defines a switching hyperplane.

$n-2$ or smaller, $S_{(-\infty,0]}$ is in turn the union of a finite number of manifolds of dimensions $n-1$ or smaller, and thus has Lebesgue measure zero.

If $\mathbf{x}_0 \notin S_{(-\infty,0]}$, then it follows from the continuity of ψ with respect to $\mathbf{x}_0$ on D, see e.g., [62, Thm 3.5], that (8.60) is valid over a sufficiently small neighborhood of $\mathbf{x}_0$. Clearly, $\frac{\partial \psi}{\partial \mathbf{x}_0}$ then exists and is continuous if and only if t_j's are continuously differentiable with respect to $\mathbf{x}_0$. Consider t_1 and let $\mathbf{n}^T\mathbf{x} + p = 0$ be the corresponding switching surface, where $\mathbf{n}^T$ is equal to some row of $\mathbf{N}_{\lambda_1}$ and equal to minus some row of $\mathbf{N}_{\lambda_2}$. t_1 is the (smallest) solution to

$$\mathbf{n}^T\left(e^{\mathbf{A}_1\tau}(\mathbf{x}_0 + \mathbf{A}_1^{-1}\mathbf{b}_1) - \mathbf{A}_1^{-1}\mathbf{b}_1\right) + p = 0, \quad \tau \geq 0. \tag{8.61}$$

The derivative of the lefthand side of (8.61) with respect to τ equals $\mathbf{n}^T f(\psi(t_1; \mathbf{x}_0))$, which is nonzero if and only if the curve of ψ is not tangent to $\mathbf{n}^T\mathbf{x} + p = 0$. If so, then the continuous differentiability of t_1 with respect to $\mathbf{x}_0$ follows from the implicit function theorem [134]. Otherwise, it is not difficult to show that $\psi(t; \mathbf{x}_0)$ remains in Ω_{λ_1} after t_1[18], contradicting the fact that t_1 is a switching time. The same argument guarantees that $t_j, j = 2, \ldots, J$ are also continuously differentiable with respect to $\mathbf{x}_0$, and so is $\psi(t; \mathbf{x}_0)$.

Before moving on to the case when $\mathbf{x}_0 \in S_{(-\infty,0]}$, we analyze the case where still $\mathbf{x}_0 \notin S_{(-\infty,0]}$ but $\mathbf{x}_0 \in H$, i.e., $\mathbf{x}_0$ belongs to a switching hyperplane, say $\mathbf{n}^T\mathbf{x} + p = 0$ between Ω_{λ_1} from Ω_{λ_2}, as above. For simplicity, assume t is small enough such that $\psi(\tau; \mathbf{x}_0)$ remains within Ω_{λ_2} for all

[18] This is a general fact about the solutions of linear systems and can be shown using the series expansion of the matrix exponential.

$\tau \in [0, t]$.[19] Let $\mathbf{x}$ belong to a sufficiently small neighborhood of $\mathbf{x}_0$ such that for $\tau \in [0, t]$,

$$\psi(\tau; \mathbf{x}) = \begin{cases} e^{\mathbf{A}_2 \tau}(\mathbf{x} + \mathbf{A}_2^{-1}\mathbf{b}_2) - \mathbf{A}_2^{-1}\mathbf{b}_2; & \mathbf{x} \in \Omega_{\lambda_2}, \\ e^{\mathbf{A}_1 \tau}(\mathbf{x} + \mathbf{A}_1^{-1}\mathbf{b}_1) - \mathbf{A}_1^{-1}\mathbf{b}_1; & \mathbf{x} \in \Omega_{\lambda_1}, \tau \leq t_1, \\ e^{\mathbf{A}_2(\tau - t_1)}(\psi(t_1; \mathbf{x}) + \mathbf{A}_2^{-1}\mathbf{b}_2) - \mathbf{A}_2^{-1}\mathbf{b}_2; & \mathbf{x} \in \Omega_{\lambda_1}, \tau \geq t_1, \end{cases} \tag{8.62}$$

where $t_1 = t_1(\mathbf{x})$ is now the solution to $\mathbf{n}^T \psi(t_1; \mathbf{x}) + p = 0$. It is not difficult to show that for $\mathbf{x} \in \Omega_{\lambda_1}$,

$$\frac{\partial \psi(t; \mathbf{x})}{\partial x_i} = e^{\mathbf{A}_2 t}\Big[e^{-\mathbf{A}_2 t_1} e^{\mathbf{A}_1 t_1} e_i + \frac{\partial t_1}{\partial x_i}\big(- \mathbf{A}_2 e^{-\mathbf{A}_2 t_1} e^{\mathbf{A}_1 t_1}(\mathbf{x} + \mathbf{A}_1^{-1}\mathbf{b}_1) \\ + e^{-\mathbf{A}_2 t_1}\mathbf{A}_1 e^{\mathbf{A}_1 t_1}(\mathbf{x} + \mathbf{A}_1^{-1}\mathbf{b}_1) + \mathbf{A}_2 e^{-\mathbf{A}_2 t_1}(\mathbf{A}_2^{-1}\mathbf{b}_2 - \mathbf{A}_1^{-1}\mathbf{b}_1)\big)\Big],$$

where e_i is the i'th column of $\mathbf{I}_n$. Taking the limit of this expression as $\mathbf{x} \to \mathbf{x}_0$ and using the facts that $\lim_{\mathbf{x} \to \mathbf{x}_0} t_1 = 0$ and $\mathbf{A}_1 \mathbf{x}_0 + \mathbf{b}_1 = \mathbf{A}_2 \mathbf{x}_0 + \mathbf{b}_2$, we get

$$\lim_{\substack{\Omega_{\lambda_1} \\ \mathbf{x} \to \mathbf{x}_0}} \frac{\partial \psi(t; \mathbf{x})}{\partial x_i} = e^{\mathbf{A}_2 t} e_i, \qquad \forall i \in \{1, \ldots, n\},$$

$$\Rightarrow \lim_{\substack{\Omega_{\lambda_1} \\ \mathbf{x} \to \mathbf{x}_0}} \frac{\partial \psi(t; \mathbf{x})}{\partial \mathbf{x}} = e^{\mathbf{A}_2 t} = \lim_{\substack{\Omega_{\lambda_2} \\ \mathbf{x} \to \mathbf{x}_0}} \frac{\partial \psi(t; \mathbf{x})}{\partial \mathbf{x}},$$

where the second equality follows directly from (8.62). Therefore, $\psi(t; \mathbf{x}_0)$ is continuously differentiable with respect to $\mathbf{x}_0$ on the entire $D \setminus S_{(-\infty, 0]}$.

Finally, if $\mathbf{x}_0 \in S_{(-\infty,0]}$, the same expression as (8.60) or (8.62) (depending on whether $\mathbf{x}_0 \in H$ or not) holds for $\mathbf{x}_0$ and also for all $\mathbf{x}$ within a sufficiently small neighborhood of it *that lie*

[19] Note that if t is larger, then subsequent switches to Ω_{λ_j}, $j \geq 3$ are similar to the case above (where $\mathbf{x}_0$ was not on a switching hyperplane) and thus do not violate continuous differentiability of ψ with respect to $\mathbf{x}_0$.

on the same system trajectory as $\mathbf{x}_0$. This curve can be parameterized in many ways, one of which is given by $\psi(\tau; \mathbf{x}_0)$. Together with the analysis of the case $\mathbf{x}_0 \notin S_{(-\infty,0]}$ above, this proves that (8.59) exists and is continuous at τ_0, as claimed.[20]

(ii) In this step we introduce a function $\hat{V}$ that may not be smooth but satisfies properties similar to (8.58). Let

$$\hat{V}(\mathbf{x}) \triangleq \int_0^\delta \|\psi(t; \mathbf{x})\|^2 dt, \qquad \forall \mathbf{x} \in D,$$

where δ is a constant to be chosen. It is straightforward to show that f is globally Lipschitz. Using this and the GES of (8.57), the same argument as in [62, Thm 4.14] shows that

$$2c_1\|\mathbf{x}\|^2 \leq \hat{V}(\mathbf{x}) \leq \frac{2}{3}c_2\|\mathbf{x}\|^2, \tag{8.63}$$

for some $c_1, c_2 > 0$. Further, let

$$D_{\psi\circ\psi}(t; \tau; \mathbf{x}) \triangleq \frac{\partial}{\partial\tau}\psi(t; \psi(\tau; \mathbf{x})), \qquad t, \tau \in \mathbb{R}, \mathbf{x} \in D.$$

By the definition of ψ, we have the identity

$$\psi(t; \psi(s - t; \mathbf{x})) = \psi(s, \mathbf{x}), \qquad t, s \in \mathbb{R}, \mathbf{x} \in D.$$

[20] We have indeed proved a slightly stronger result than (8.59) for $\mathbf{x}_0 \notin S_{(-\infty,0]}$, which we use in step (ii) below.

Taking $\frac{d}{dt}$ of both sides, we get

$$\psi_t(t;\psi(s-t;\mathbf{x})) - D_{\psi\circ\psi}(t;s-t;\mathbf{x}) = 0,$$

where $\psi_t(t;\mathbf{x}) = \frac{\partial\psi(t;\mathbf{x})}{\partial t}$. Setting $s = t+\tau$, this yields

$$D_{\psi\circ\psi}(t;\tau;\mathbf{x}) = \psi_t(t;\psi(\tau;\mathbf{x})). \tag{8.64}$$

For the parallel of (8.58b), we then have

$$\begin{aligned}
\frac{d}{d\tau}\hat{V}(\psi(\tau;\mathbf{x})) &= \int_0^\delta 2\psi(t;\psi(\tau;\mathbf{x}))^T D_{\psi\circ\psi}(t;\tau;\mathbf{x})dt \\
&\overset{(8.64)}{=} \int_0^\delta 2\psi(t;\psi(\tau;\mathbf{x}))^T \psi_t(t;\psi(\tau;\mathbf{x}))dt \\
&= \int_0^\delta \frac{\partial}{\partial t}\|\psi(t;\psi(\tau;\mathbf{x}))\|^2 dt \\
&= \|\psi(\delta;\psi(\tau;\mathbf{x}))\|^2 - \|\psi(\tau;\mathbf{x})\|^2.
\end{aligned}$$

Thus

$$\frac{d}{d\tau}\hat{V}(\psi(\tau;\mathbf{x}))\bigg|_{\tau=0} = \|\psi(\delta;\mathbf{x})\|^2 - \|\mathbf{x}\|^2 \leq -2c_3\|\mathbf{x}\|^2, \tag{8.65}$$

where the last inequality holds, as shown in [62, Thm 4.14], for an appropriate choice of δ and $c_3 = \frac{1}{4}$. Finally, for the parallel of (8.58c), recall from step (i) that $\frac{\partial}{\partial\mathbf{x}}\psi(t;\mathbf{x})$ exists and is continuous

on $D \setminus S_{(-\infty,0]}$. Therefore, from (8.57), we have

$$\frac{\partial}{\partial t}\frac{\partial \psi(t;\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}}(\psi(t;\mathbf{x}))\frac{\partial \psi(t;\mathbf{x})}{\partial \mathbf{x}}, \qquad \left.\frac{\partial \psi(t;\mathbf{x})}{\partial \mathbf{x}}\right|_{t=0} = \mathbf{I}_n,$$

on $D \setminus (S_{(-\infty,0]} \cup H)$. Using the global Lipschitzness of f and the fact that $D \setminus S_{(-\infty,0]}$ is invariant under (8.57), we have

$$\left\| \frac{\partial \psi(t;\mathbf{x})}{\partial \mathbf{x}} \right\| \leq e^{Lt}, \qquad \forall x \in D \setminus S_{(-\infty,0]},$$

where L is the Lipschitz constant of f. The same argument as in [62, Thm 4.14] then yields

$$\left\| \frac{\partial \hat{V}}{\partial \mathbf{x}} \right\| \leq \frac{2}{3} c_4 \|\mathbf{x}\|, \qquad \forall x \in D \setminus S_{(-\infty,0]}, \tag{8.66}$$

for some $c_4 > 0$.

(iii) In this step, we follow [135, Thm 3 & 4] to construct V as an smooth approximation to $\hat{V}$ and show that it satisfies (8.58). Since f is globally Lipschitz, $\psi(t;\mathbf{x})$ is Lipschitz in $\mathbf{x}$ (see, e.g., [136, Ch 5]) and so is $\hat{V}$. This, together with (8.65), satisfies all the assumptions of [135, Thm 4], which in turn guarantees the existence of an infinitely smooth V such that

$$|V(\mathbf{x}) - \hat{V}(\mathbf{x})| < \frac{1}{2}\hat{V}(\mathbf{x}), \qquad \forall \mathbf{x} \in D, \tag{8.67a}$$

$$\frac{\partial V}{\partial \mathbf{x}} f(\mathbf{x}) < -c_3 \|\mathbf{x}\|^2, \tag{8.67b}$$

for all $\mathbf{x} \in D$. Equation (8.58a) follows immediately from (8.67b) and (8.63). To prove (8.58c), we

note that the same construction of V as in [135, Thm 3 & 4] satisfies

$$\left\| \frac{\partial V}{\partial \mathbf{x}} - \frac{\partial \hat{V}}{\partial \mathbf{x}} \right\| < \frac{1}{2} \left\| \frac{\partial \hat{V}}{\partial \mathbf{x}} \right\|, \qquad \forall \mathbf{x} \in D \setminus S_{(-\infty,0]},$$

if the constants $\xi_{i,k}$ and $\zeta_{i,k}, i, k = \dots, -2, 0, 2, \dots$ (and consequently the corresponding $\bar{r}_{i,k}, i, k = \dots, -2, 0, 2, \dots$) are chosen sufficiently small. This, together with (8.66), guarantees (8.58c), completing the proof.

□

Acknowledgements: This chapter is taken, in part, from the work which has been submitted for publication as "Hierarchical selective recruitment in linear-threshold brain networks. Part I: Intra-layer dynamics and selective inhibition" by E. Nozari and J. Cortés in *IEEE Transactions on Automatic Control*, as well as the work which has been submitted for publication as "Hierarchical selective recruitment in linear-threshold brain networks. Part II: Inter-layer dynamics and top-down recruitment" by E. Nozari and J. Cortés in *IEEE Transactions on Automatic Control*. The dissertation author was the primary investigator and author of these papers.

Chapter Bibliography

[1] J. Sully, "The psycho-physical process in attention," *Brain*, vol. 13, no. 2, pp. 145–164, 1890.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[4] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, p. 194, 2001.

[5] N. Lavie, A. Hirst, J. W. DeFockert, and E. Viding, "Load theory of selective attention and cognitive control." *Journal of Experimental Psychology: General*, vol. 133, no. 3, p. 339, 2004.

[6] A. Gazzaley and A. C. Nobre, "Top-down modulation: bridging selective attention and working memory," *Trends in Cognitive Sciences*, vol. 16, no. 2, pp. 129–135, 2012.

[7] D. E. Broadbent, Ed., *Perception and communication*. Pergamon, 1958.

[8] A. M. Treisman, "Strategies and models of selective attention." *Psychological review*, vol. 76, no. 3, p. 282, 1969.

[9] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, no. 4715, pp. 782–784, 1985.

[10] B. C. Motter, "Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli," *Journal of Neurophysiology*, vol. 70, no. 3, pp. 909–919, 1993.

[11] S. Kastner, P. DeWeerd, R. Desimone, and L. G. Ungerleider, "Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI," *Science*, vol. 282, no. 5386, pp. 108–111, 1998.

[12] M. A. Pinsk, G. M. Doniger, and S. Kastner, "Push-pull mechanism of selective attention in human extrastriate cortex," *Journal of Neurophysiology*, vol. 92, no. 1, pp. 622–629, 2004.

[13] N. Lavie, "Distracted and confused?: Selective attention under load," *Trends in Cognitive Sciences*, vol. 9, no. 2, pp. 75–82, 2005.

[14] J. J. Foxe and A. C. Snyder, "The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention," *Frontiers in Psychology*, vol. 2, p. 154, 2011.

[15] H. Pashler, *Attention*. Psychology Press, 2016.

[16] M. Gomez-Ramirez, K. Hysaj, and E. Niebur, "Neural mechanisms of selective attention in the somatosensory system," *Journal of neurophysiology*, vol. 116, no. 3, pp. 1218–1231, 2016.

[17] F. Ratliff and H. K. Hartline, *Studies on Excitation and Inhibition in the Retina*. Rockefeller University Press, 1974.

[18] K. P. Hadeler, "On the theory of lateral inhibition," *Kybernetik*, vol. 14, no. 3, pp. 161–165, 1973.

[19] R. H. R. Hahnloser, H. S. Seung, and J. J. Slotine, "Permitted and forbidden sets in symmetric threshold-linear networks," *Neural Computation*, vol. 15, no. 3, pp. 621–638, 2003.

[20] Z. Yi, L. Zhang, J. Yu, and K. K. Tan, "Permitted and forbidden sets in discrete-time linear threshold recurrent neural networks," *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 952–963, 2009.

[21] K. P. Hadeler and D. Kuhn, "Stationary states of the Hartline-Ratliff model," *Biological Cybernetics*, vol. 56, no. 5-6, pp. 411–417, 1987.

[22] J. Feng and K. P. Hadeler, "Qualitative behaviour of some simple networks," *Journal of Physics A: Mathematical and General*, vol. 29, no. 16, pp. 5019–5033, 1996.

[23] Z. Yi and K. K. Tan, "Multistability of discrete-time recurrent neural networks with unsaturating piecewise linear activation functions," *IEEE Transactions on Neural Networks*, vol. 15, no. 2, pp. 329–336, 2004.

[24] W. Zhou and J. M. Zurada, "A new stability condition for discrete time linear threshold recurrent neural networks," in *Fifth Int. Conf. on Intellig. Control and Inf. Proc.*, Aug 2014, pp. 96–99.

[25] T. Shen and I. R. Petersen, "Linear threshold discrete-time recurrent neural networks: Stability and globally attractive sets," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2650–2656, 2016.

[26] K. C. Tan, H. Tang, and W. Zhang, "Qualitative analysis for recurrent neural networks with linear threshold transfer functions," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 5, pp. 1003–1012, 2005.

[27] H. Wersing, W. Beyn, and H. Ritter, "Dynamical stability conditions for recurrent neural networks with unsaturating piecewise linear transfer functions," *Neural Computation*, vol. 13, no. 8, pp. 1811–1825, 2001.

[28] K. Morrison, A. Degeratu, V. Itskov, and C. Curto, "Diversity of emergent dynamics in competitive threshold-linear networks: a preliminary report," *arXiv preprint arXiv:1605.04463*, 2016.

[29] H. Lin and P. J. Antsaklis, "Stability and stabilizability of switched linear systems: A survey of recent results," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 308–322, 2009.

[30] D. Liberzon, *Switching in Systems and Control*, ser. Systems & Control: Foundations & Applications. Birkhäuser, 2003.

[31] M. K. J. Johansson, *Piecewise Linear Control Systems: A Computational Approach*, ser. Lecture Notes in Control and Information Sciences. Springer Berlin Heidelberg, 2003.

[32] K. N. Seidl, M. V. Peelen, and S. Kastner, "Neural evidence for distracter suppression during visual search in real-world scenes," *Journal of Neuroscience*, vol. 32, no. 34, pp. 11 812–11 819, 2012.

[33] S. Kastner, P. DeWeerd, M. A. Pinsk, M. I. Elizondo, R. Desimone, and L. G. Ungerleider, "Modulation of sensory suppression: implications for receptive field sizes in the human visual cortex," *Journal of Neurophysiology*, vol. 86, no. 3, pp. 1398–1411, 2001.

[34] G. Rees, C. D. Frith, and N. Lavie, "Modulating irrelevant motion perception by varying attentional load in an unrelated task," *Science*, vol. 278, no. 5343, pp. 1616–1619, 1997.

[35] D. H. O'Connor, M. M. Fukui, M. A. Pinsk, and S. Kastner, "Attention modulates responses in the human lateral geniculate nucleus," *Nature Neuroscience*, vol. 5, no. 11, p. 1203, 2002.

[36] N. Tinbergen, "The hierarchical organization of nervous mechanisms underlying instinctive behaviour," in *Symposium for the Society for Experimental Biology*, vol. 4, no. 305-312, 1950.

[37] A. R. Luria, "The functional organization of the brain," *Scientific American*, vol. 222, no. 3, pp. 66–79, 1970.

[38] D. J. Felleman and D. C. V. Essen, "Distributed hierarchical processing in the primate cerebral cortex," *Cerebral Cortex*, vol. 1, no. 1, pp. 1–47, 1991.

[39] A. Krumnack, A. T. Reid, E. Wanke, G. Bezgin, and R. Kötter, "Criteria for optimizing cortical hierarchies with continuous ranges," *Frontiers in Neuroinformatics*, vol. 4, p. 7, 2010.

[40] G. Zamora-López, C. Zhou, and J. Kurths, "Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks," *Frontiers in Neuroinformatics*, vol. 4, p. 1, 2010.

[41] N. T. Markov, J. Vezoli, P. Chameau, A. Falchier, R. Quilodran, C. Huissoud, C. Lamy, P. Misery, P. Giroud, S. Ullman, P. Barone, C. Dehay, K. Knoblauch, and H. Kennedy, "Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex," *Journal of Comparative Neurology*, vol. 522, no. 1, pp. 225–259, 2014.

[42] P. Lennie, "Single units and visual cortical organization," *Perception*, vol. 27, no. 8, pp. 889–935, 1998.

[43] D. Badre and M. D'Esposito, "Is the rostro-caudal axis of the frontal lobe hierarchical?" *Nature Reviews Neuroscience*, vol. 10, no. 9, pp. 659–669, 2009.

[44] U. Hasson, E. Yang, I. Vallines, D. J. Heeger, and N. Rubin, "A hierarchy of temporal receptive windows in human cortex," *Journal of Neuroscience*, vol. 28, no. 10, pp. 2539–2550, 2008.

[45] C. J. Honey, T. Thesen, T. H. Donner, L. J. Silbert, C. E. Carlson, O. Devinsky, W. K. Doyle, N. Rubin, D. J. Heeger, and U. Hasson, "Slow cortical dynamics and the accumulation of information over long timescales," *Neuron*, vol. 76, no. 2, pp. 423–434, 2012.

[46] B. Gauthier, E. Eger, G. Hesselmann, A. Giraud, and A. Kleinschmidt, "Temporal tuning properties along the human ventral visual stream," *Journal of Neuroscience*, vol. 32, no. 41, pp. 14 433–14 441, 2012.

[47] U. Hasson, J. Chen, and C. J. Honey, "Hierarchical process memory: memory as an integral component of information processing," *Trends in Cognitive Sciences*, vol. 19, no. 6, pp. 304–313, 2015.

[48] M. G. Mattar, D. A. Kahn, S. L. Thompson-Schill, and G. K. Aguirre, "Varying timescales of stimulus integration unite neural adaptation and prototype formation," *Current Biology*, vol. 26, no. 13, pp. 1669–1676, 2016.

[49] J. D. Murray, A. Bernacchia, D. J. Freedman, R. Romo, J. D. Wallis, X. Cai, C. Padoa-Schioppa, T. Pasternak, H. Seo, D. Lee, and X. Wang, "A hierarchy of intrinsic timescales across primate cortex," *Nature Neuroscience*, vol. 17, no. 12, p. 1661, 2014.

[50] R. Chaudhuri, K. Knoblauch, M. Gariel, H. Kennedy, and X. Wang, "A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex," *Neuron*, vol. 88, no. 2, pp. 419–431, 2015.

[51] J. P. Gilman, M. Medalla, and J. I. Luebke, "Area-specific features of pyramidal neurons?a comparative study in mouse and rhesus monkey," *Cerebral Cortex*, vol. 27, no. 3, pp. 2078–2094, 2016.

[52] C. Cioli, H. Abdi, D. Beaton, Y. Burnod, and S. Mesmoudi, "Differences in human cortical gene expression match the temporal properties of large-scale functional networks," *PLOS One*, vol. 9, no. 12, p. e115913, 2014.

[53] C. A. Runyan, E. Piasini, S. Panzeri, and C. D. Harvey, "Distinct timescales of population coding across cortex," *Nature*, vol. 548, no. 7665, p. 92, 2017.

[54] S. J. Kiebel, J. Daunizeau, and K. J. Friston, "A hierarchy of time-scales and the brain," *PLOS Computational Biology*, vol. 4, no. 11, p. e1000209, 2008.

[55] Y. Yamashita and J. Tani, "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment," *PLOS Computational Biology*, vol. 4, no. 11, p. e1000220, 2008.

[56] D. S. Bassett, E. T. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg, "Hierarchical organization of human cortical networks in health and schizophrenia," *Journal of Neuroscience*, vol. 28, no. 37, pp. 9239–9248, 2008.

[57] D. Meunier, R. Lambiotte, A. Fornito, K. Ersche, and E. T. Bullmore, "Hierarchical modularity in human brain functional networks," *Frontiers in Neuroinformatics*, vol. 3, p. 37, 2009.

[58] D. Meunier, R. Lambiotte, and E. T. Bullmore, "Modular and hierarchically modular organization of brain networks," *Frontiers in Neuroscience*, vol. 4, p. 200, 2010.

[59] Z. Zhen, H. Fang, and J. Liu, "The hierarchical brain network for face recognition," *PLOS One*, vol. 8, no. 3, p. e59886, 2013.

[60] P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, and C. E. Schroeder, "An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex," *Journal of Neurophysiology*, vol. 94, no. 3, pp. 1904–1911, 2005.

[61] A. N. Tikhonov, "Systems of differential equations containing small parameters in the derivatives," *Matematicheskii Sbornik*, vol. 73, no. 3, pp. 575–586, 1952.

[62] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.

[63] A. B. Vasilieva, "On the development of singular perturbation theory at moscow state university and elsewhere," *SIAM Review*, vol. 36, no. 3, pp. 440–452, 1994.

[64] D. Naidu, "Singular perturbations and time scales in control theory and applications: an overview," *Dynamics of Continuous Discrete and Impulsive Systems Series B*, vol. 9, pp. 233–278, 2002.

[65] J. K. Kevorkian and J. D. Cole, *Multiple scale and singular perturbation methods*. Springer Science & Business Media, 2012, vol. 114.

[66] R. E. O'Malley, *Singular perturbation methods for ordinary differential equations*. Springer Science & Business Media, 2012, vol. 89.

[67] A. L. Dontchev and V. M. Veliov, "Singular perturbation in mayer's problem for linear systems," *SIAM Journal on Control and Optimization*, vol. 21, no. 4, pp. 566–581, 1983.

[68] A. L. Dontchev and I. I. Slavov, "Upper semicontinuity of solutions of singularly perturbed differential inclusions," in *System Modelling and Optimization*, H. J. Sebastian and K. Tammer, Eds. Springer Berlin Heidelberg, 1990, pp. 273–280.

[69] M. Quincampoix, "Singular perturbations for systems of differential inclusions," *Banach Center Publications*, vol. 32, no. 1, pp. 341–348, 1995.

[70] A. Dontchev, T. Donchev, and I. I. Slavov, "A Tikhonov-type theorem for singularly perturbed differential inclusions," *Nonlinear Analysis, Theory, Methods & Applications*, vol. 26, no. 9, pp. 1547–1554, 1996.

[71] V. Veliov, "A generalization of the Tikhonov theorem for singularly perturbed differential inclusions," *Journal of Dynamical & Control Systems*, vol. 3, no. 3, pp. 291–319, 1997.

[72] F. Watbled, "On singular perturbations for differential inclusions on the infinite interval," *Journal of Mathematical Analysis and Applications*, vol. 310, no. 2, pp. 362–378, 2005.

[73] G. Grammel, "Exponential stability of nonlinear singularly perturbed differential equations," *SIAM Journal on Control and Optimization*, vol. 44, no. 5, pp. 1712–1724, 2005.

[74] D. Kuhn and R. Löwen, "Piecewise affine bijections of $\mathbb{R}^n$, and the equation $Sx^+ - Tx^- = y$," *Linear Algebra and its Applications*, vol. 96, pp. 109–129, 1987.

[75] C. Curto, J. Geneson, and K. Morrison, "Fixed points of competitive threshold-linear networks," *arXiv preprint arXiv:1804.00794*, 2018.

[76] D. S. Bernstein, *Matrix Mathematics*, 2nd ed. Princeton University Press, 2009.

[77] G. E. Coxson, "The p-matrix problem is co-np-complete," *Mathematical Programming*, vol. 64, no. 1, pp. 173–178, 1994.

[78] S. M. Rump, "On p-matrices," *Linear Algebra and its Applications*, vol. 363, pp. 237–250, 2003.

[79] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, vol. 12, no. 1, pp. 1–24, 1972.

[80] A. C. E. Onslow, M. W. Jones, and R. Bogacz, "A canonical circuit for generating phase-amplitude coupling," *PLOS One*, vol. 9, no. 8, p. e102591, 2014.

[81] M. P. Jadi and T. J. Sejnowski, "Regulating cortical oscillations in an inhibition-stabilized network," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 830–842, 2014.

[82] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[83] H. S. Seung, D. D. Lee, B. Y. Reis, and D. W. Tank, "Stability of the memory of eye position in a recurrent network of conductance-based model neurons," *Neuron*, vol. 26, no. 1, pp. 259–271, 2000.

[84] D. Durstewitz, J. K. Seamans, and T. J. Sejnowski, "Neurocomputational models of working memory," *Nature Neuroscience*, vol. 3, no. 11s, p. 1184, 2000.

[85] R. Cossart, D. Aronov, and R. Yuste, "Attractor dynamics of network up states in the neocortex," *Nature*, vol. 423, no. 6937, pp. 283–288, 2003.

[86] J. J. Knierim and K. Zhang, "Attractor dynamics of spatially correlated neural activity in the limbic system," *Annual Review of Neuroscience*, vol. 35, no. 1, pp. 267–285, 2012.

[87] A. Pavlov, N. van de Wouw, and H. Nijmeijer, "Convergent piecewise affine systems: analysis and design Part I: continuous case," in *IEEE Conf. on Decision and Control*, Dec 2005, pp. 5391–5396.

[88] P. Jiruska, J. Csicsvari, A. D. Powell, J. E. Fox, W. Chang, M. Vreugdenhil, X. Li, M. Palus, A. F. Bujan, R. W. Dearden, and J. G. R. Jefferys, "High-frequency network activity, global increase in neuronal activity, and synchrony expansion precede epileptic seizures in vitro," *Journal of Neuroscience*, vol. 30, no. 16, pp. 5690–5701, 2010.

[89] S. K. Y. Nikravesh, *Nonlinear Systems Stability Analysis: Lyapunov-Based Approach*. CRC Press, 2013.

[90] J. S. Isaacson and M. Scanziani, "How inhibition shapes cortical activity," *Neuron*, vol. 72, no. 2, pp. 231–243, 2011.

[91] E. Nozari and J. Cortés, "Hierarchical selective recruitment in linear-threshold brain networks. Part II: Inter-layer dynamics and top-down recruitment," *IEEE Transactions on Automatic Control*, 2018, submitted.

[92] E. D. Sontag, "On the input-to-state stability property," *European Journal of Control*, vol. 1, pp. 24–36, 1995.

[93] C. R. Johnson and H. A. Robinson, "Eigenvalue inequalities for principal submatrices," *Linear Algebra and its Applications*, vol. 37, pp. 11–22, 1981.

[94] H. L. Royden and P. Fitzpatrick, *Real Analysis*. Prentice Hall, 2010.

[95] C. T. Chen, *Linear System Theory and Design*, 3rd ed. New York, NY, USA: Oxford University Press, Inc., 1998.

[96] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, "Highly nonrandom features of synaptic connectivity in local cortical circuits," *PLOS Biology*, vol. 3, no. 3, p. e68, 2005.

[97] D. Sussillo and L. F. Abbott, "Generating coherent patterns of activity from chaotic neural networks," *Neuron*, vol. 63, no. 4, pp. 544–557, 2009.

[98] N. Bertschinger and T. Natschläger, "Real-time computation at the edge of chaos in recurrent neural networks," *Neural Computation*, vol. 16, no. 7, pp. 1413–1436, 2004.

[99] J. Aljadeff, M. Stern, and T. Sharpee, "Transition to chaos in random networks with cell-type-specific connectivity," *Physical Review Letters*, vol. 114, no. 8, p. 088101, 2015.

[100] G. Turrigiano, "Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function," *Cold Spring Harbor perspectives in biology*, p. a005736, 2012.

[101] E. Nozari and J. Cortés, "Hierarchical selective recruitment in linear-threshold brain networks. Part I: Intra-layer dynamics and selective inhibition," *IEEE Transactions on Automatic Control*, 2018, submitted.

[102] P. Fries, "A mechanism for cognitive dynamics: neuronal communication through neuronal coherence," *Trends in Cognitive Sciences*, vol. 9, no. 10, pp. 474–480, 2005.

[103] T. J. Buschman and E. K. Miller, "Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices," *Science*, vol. 315, no. 5820, pp. 1860–1862, 2007.

[104] D. Rubino, K. A. Robbins, and N. G. Hatsopoulos, "Propagating waves mediate information transfer in the motor cortex," *Nature Neuroscience*, vol. 9, no. 12, pp. 1549–1557, 2006.

[105] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone, "Modulation of oscillatory neuronal synchronization by selective visual attention," *Science*, vol. 291, no. 5508, pp. 1560–1563, 2001.

[106] A. Sokolov, M. Pavlova, W. Lutzenberger, and N. Birbaumer, "Reciprocal modulation of neuromagnetic induced gamma activity by attention in the human visual and auditory cortex," *NeuroImage*, vol. 22, no. 2, pp. 521–529, 2004.

[107] S. Ray, E. Niebur, S. S. Hsiao, A. Sinai, and N. E. Crone, "High-frequency gamma activity (80–150 hz) is increased in human cortex during selective attention," *Clinical Neurophysiology*, vol. 119, no. 1, pp. 116–133, 2008.

[108] N. Kahlbrock, M. Butz, E. S. May, and A. Schnitzler, "Sustained gamma band synchronization in early visual areas reflects the level of selective attention," *NeuroImage*, vol. 59, no. 1, pp. 673–681, 2012.

[109] J. A. Cardin, M. Carlén, K. Meletis, U. Knoblich, F. Zhang, K. Deisseroth, L. Tsai, and C. I. Moore, "Driving fast-spiking cells induces gamma rhythm and controls sensory responses," *Nature*, vol. 459, no. 7247, p. 663, 2009.

[110] R. Gao, E. J. Peterson, and B. Voytek, "Inferring synaptic excitation/inhibition balance from field potentials," *NeuroImage*, vol. 158, pp. 70–78, 2017.

[111] X. Chen and R. Dzakpasu, "Observed network dynamics from altering the balance between excitatory and inhibitory neurons in cultured networks," *Physical Review E*, vol. 82, no. 3, p. 031907, 2010.

[112] R. Srinivasan, S. Thorpe, and P. L. Nunez, "Top-down influences on local networks: basic theory with experimental implications," *Frontiers in Computational Neuroscience*, vol. 7, p. 29, 2013.

[113] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill, 1976.

[114] C. C. Rodgers and M. R. DeWeese, "Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents," *Neuron*, vol. 82, no. 5, pp. 1157–1170, 2014.

[115] ——, "Spiking responses of neurons in rodent prefrontal cortex and auditory cortex during a novel stimulus selection task," CRCNS.org, 2014. [Online]. Available: http://dx.doi.org/10.6080/K0W66HPJ

[116] M. A. Bee and C. Micheyl, "The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it?" *Journal of Comparative Psychology*, vol. 122, no. 3, p. 235, 2008.

[117] J. Ahveninen, M. Hämäläinen, I. P. Jääskeläinen, S. P. Ahlfors, S. Huang, F. Lin, T. Raij, M. Sams, C. E. Vasios, and J. W. Belliveau, "Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise," *Proceedings of the National Academy of Sciences*, vol. 108, no. 10, pp. 4182–4187, 2011.

[118] A. W. Bronkhorst, "The cocktail-party problem revisited: early processing and selection of multi-talker speech," *Attention, Perception, & Psychophysics*, vol. 77, no. 5, pp. 1465–1487, 2015.

[119] D. B. Geissler and G. Ehret, "Time-critical integration of formants for perception of communication calls in mice," *Proceedings of the National Academy of Sciences*, vol. 99, no. 13, pp. 9021–9025, 2002.

[120] J. Fuster, *The Prefrontal Cortex*. Elsevier Science, 2015.

[121] R. M. Bruno and D. J. Simons, "Feedforward mechanisms of excitatory and inhibitory cortical receptive fields," *Journal of Neuroscience*, vol. 22, no. 24, pp. 10 966–10 975, 2002.

[122] P. S. Goldman-Rakic, "Cellular basis of working memory," *Neuron*, vol. 14, no. 3, pp. 477–485, 1995.

[123] A. F. T. Arnsten, M. J. Wang, and C. D. Paspalas, "Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses," *Neuron*, vol. 76, no. 1, pp. 223–239, 2012.

[124] P. Somogyi, G. Tamasab, R. Lujan, and E. H. Buhl, "Salient features of synaptic organisation in the cerebral cortex," *Brain Research Reviews*, vol. 26, no. 2, pp. 113–135, 1998.

[125] G. K. Wu, R. Arbuckle, B. Liu, H. W. Tao, and L. I. Zhang, "Lateral sharpening of cortical frequency tuning by approximately balanced inhibition," *Neuron*, vol. 58, no. 1, pp. 132–143, 2008.

[126] H. K. Kato, S. K. Asinof, and J. S. Isaacson, "Network-level control of frequency tuning in auditory cortex," *Neuron*, vol. 95, no. 2, pp. 412–423, 2017.

[127] N. P. Bichot, M. T. Heard, E. M. DeGennaro, and R. Desimone, "A source for feature-based attention in the prefrontal cortex," *Neuron*, vol. 88, no. 4, pp. 832–844, 2015.

[128] A. Mita, H. Mushiake, K. Shima, Y. Matsuzaka, and J. Tanji, "Interval time coding by neurons in the presupplementary and supplementary motor areas," *Nature Neuroscience*, vol. 12, no. 4, p. 502, 2009.

[129] G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*. Cambridge, UK: Cambridge University Press, 1952.

[130] W. P. Dayawansa and C. F. Martin, "A converse Lyapunov theorem for a class of dynamical systems which undergo switching," *IEEE Transactions on Automatic Control*, vol. 44, no. 4, pp. 751–760, 1999.

[131] D. Liberzon and A. S. Morse, "Basic problems in stability and design of switched systems," *IEEE Control Systems*, vol. 19, no. 5, pp. 59–70, 1999.

[132] F. Wirth, "A converse Lyapunov theorem for linear parameter-varying and linear switching systems," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 210–239, 2005.

[133] F. M. Hante and M. Sigalotti, "Converse Lyapunov theorems for switched systems in Banach and Hilbert spaces," *SIAM Journal on Control and Optimization*, vol. 49, no. 2, pp. 752–770, 2011.

[134] S. G. Krantz and H. R. Parks, *The Implicit Function Theorem: History, Theory, and Applications*. Birkhäuser, 2002.

[135] J. Kurzweil, "On the inversion of Lyapunov's second theorem on stability of motion," *American Mathematical Society Translations*, vol. 24, no. 2, pp. 19–77, 1963.

[136] P. Hartman, *Ordinary Differential Equations*, 2nd ed., ser. Classics in Applied Mathematics. SIAM, 1982.

Chapter 9

Oscillations and Coupling in Brain Networks

Oscillations in the brain are one of the most ubiquitous and robust patterns of activity and correlate with various cognitive phenomena. Since Berger's groundbreaking discovery of oscillatory activity in the brain [1], oscillations have been found in a wide range of species and brain regions and multiple studies have shown the correlation between their properties (amplitude, phase, shape, coupling, etc.) and various neurocognitive processes. Despite their importance, our understanding of brain oscillations is far from complete. Here, we take an analytical approach and study the existence and properties of oscillations in simple mean-field models of brain activity with bounded linear-threshold rate dynamics. This reveals the relationship between network structure and oscillatory behavior, both within a single region and when coupled between multiple regions.

First, we obtain exact conditions for the existence of limit cycles in two-dimensional excitatory-inhibitory networks (E-I pairs). Building on this result, we study networks of multiple E-I pairs, provide exact conditions for the lack of stable equilibria, and numerically show that

this is a tight proxy for the existence of oscillatory behavior. Finally, we study cross-frequency coupling between pairs of oscillators each consisting of an E-I pair. We find that while both phase-phase coupling (synchronization) and phase-amplitude coupling (PAC) monotonically increase with inter-oscillator connection strength, there exists a tradeoff in increasing frequency mismatch between the oscillators as it de-synchronizes them while enhancing their PAC.

9.1 Prior Work

Oscillations have been the subject of extensive research in the neuroscience literature, see, e.g. [2, 3]. In addition to the vast number of experimental and computational works, several efforts have pursued analytical model-based approaches, particularly using mean-field models such as the Wilson-Cowan model [4]. However, the sigmoidal nonlinearity in the Wilson-Cowan model has not allowed more than partial characterizations [5–8] of structural conditions giving rise to oscillations. Motivated by this, [9] studies oscillations and synchronization in Wilson-Cowan models with bounded linear-threshold nonlinearities, but relies on unrealistic assumptions (excluding interaction terms in the nonlinearities, having mixed excitatory-inhibitory nodes (i.e., violating Dale's law), and a chain network topology) to obtain rigorous results. Linear-threshold networks are indeed capable of modeling a wide range of (nonlinear) phenomena such as mono-, bi-, and multi-stability, limit cycles, and chaos [10]. While the existence and uniqueness of equilibria and asymptotic stability are reasonably well understood, see [11] and references therein, our understanding of their oscillatory behavior has remained limited.

A growing body of research has also studied brain oscillations using models of phase oscillators such as the Kuramoto model, see [12–14] and references therein. This is motivated by the

fact [15] that the Kuramoto model is a local approximation to the Wilson-Cowan model (around zero interconnection strength) and has the advantage of having smaller state dimensions. Nevertheless, this also comes at the expense of different global behaviors (when coupling is large), cf. [16], and the exclusion of amplitude dynamics that are essential to neuronal phenomena such as PAC.

9.2 Problem Statement

Consider a neuronal network composed of a large number of neurons that communicate with each other via asynchronous sequences of spikes. Grouping together neurons with similar firing rates, under standard assumptions[1], the mean-field dynamics of the network can be described by the linear-threshold model

$$\tau\dot{\mathbf{x}}(t) = -\mathbf{x}(t) + [\mathbf{W}\mathbf{x}(t) + \mathbf{p}]_{\mathbf{0}}^{\mathbf{m}}, \qquad \mathbf{x}(0) \in [\mathbf{0}, \mathbf{m}], \tag{9.1}$$

where $\mathbf{x} \in \mathbb{R}^N_{\geq 0}$ is the state vector with components x_i denoting the average firing rate of the i'th neuronal population, $i \in \{1, \ldots, N\}$, $\mathbf{W} \in \mathbb{R}^{N\times N}$ is the matrix of average synaptic connectivities, $\mathbf{p} \in \mathbb{R}^N$ is the vector of average external (background) inputs to the populations, $\mathbf{m} \in \mathbb{R}^N_{>0}$ is the vector of average maximum firing rates, and $\tau > 0$ is the network time constant. Note that $[\mathbf{0}, \mathbf{m}]$ is invariant under (9.1), ensuring, in particular, that all solutions are bounded.

Our previous work [11] has characterized the existence and uniqueness of equilibria and asymptotic stability for a variant of (9.1) with an unbounded activation function ($\mathbf{m} = \infty \cdot \mathbf{1}_N$), and these results are readily extensible to arbitrary finite $\mathbf{m}$. However, the existence of oscillations in linear-threshold dynamics is not as well understood. Further, brain networks often contain in-

[1] See, e.g., [17, Ch 7] for a comprehensive exposition or [11] for a brief discussion.

terconnections of multiple coupled oscillators, but our understanding is even slimmer about the oscillatory behavior of interconnections of linear-threshold networks of the form (9.1).

Our goal is to characterize the relationship between network structure and the oscillatory behavior observed in linear-threshold dynamics modeling brain networks. We formalize the problem of interest as follows.

Problem 6. *For the bounded linear-threshold network dynamics* (9.1)*, characterize the relationship between network structure* ($\mathbf{W}, \mathbf{m}, \mathbf{p}$) *and*

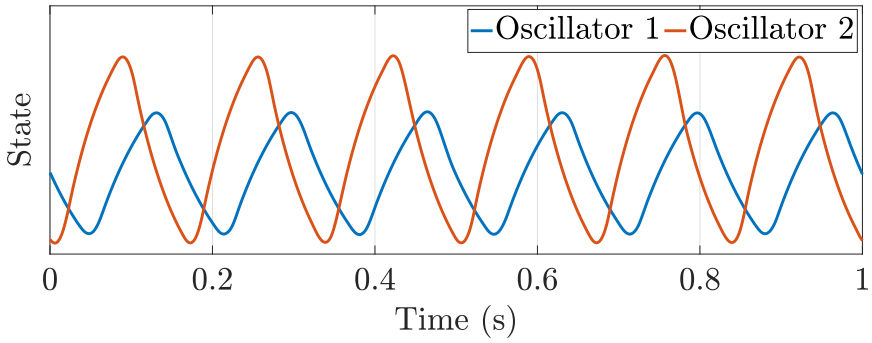
(i) existence of oscillations in a single network (9.1)*;*

(ii) existence/preservation of oscillations in a network of oscillatory networks, each modeled by (9.1)*;*

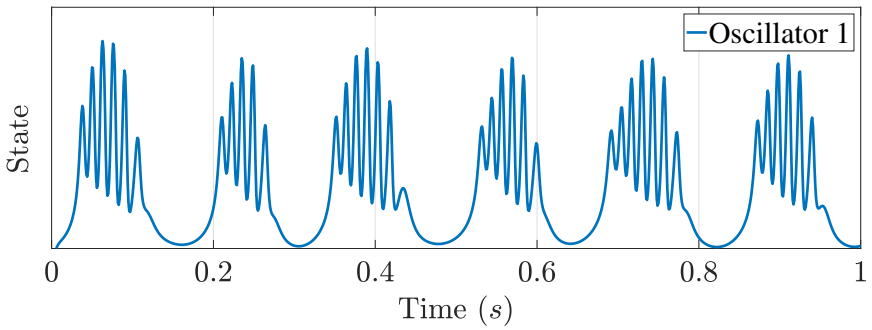
(iii) phase-phase coupling (synchronization) and phase-amplitude coupling (PAC) between pairs of oscillators. □

Questions (i) and (ii) arise naturally as the first steps towards understanding oscillatory behavior of (9.1). On the other hand, synchronization (i.e., the phase-locking of two oscillators with the same frequency) and PAC (i.e., the dependence of the amplitude of a high-frequency oscillator on the phase of a low-frequency one), are of specific interest as they are the most widely observed and studied oscillatory coupling phenomena in brain networks. Examples of these phenomena are shown in Figure 9.1. We address (i) and (ii) in Section 9.3 and (iii) in Section 9.4.

Following common practice in computational neuroscience, we here adopt a broad notion of oscillations that includes both periodic oscillations (limit cycles) and chaotic ones. In the latter case, a chaotic behavior is oscillatory if its state trajectories are near-periodic or, equivalently, have power spectra with distinct and pronounced resonance peaks.



(a)



(b)

Figure 9.1: Examples of (a) synchronization and (b) PAC in models of neuronal activity. Note that while both phenomena occur as a result of the interaction of two oscillators, synchronization is defined (and measured) between the trajectories of both oscillators but PAC is defined (and measured) between two frequency components of each trajectory.

9.3 Existence of Oscillations

In this section we analyze the dynamics (9.1) and derive conditions on the network structure $(\mathbf{W}, \mathbf{m}, \mathbf{p})$ giving rise to oscillatory behavior. The analytical tools in the study of oscillations are generally limited to 2-dimensional systems (cf. the Poincaré-Bendixson theory [18, Ch 3]) or higher-dimensional systems that are essentially confined to 2-dimensional manifolds (see, e.g., [19, 20]). Thus, we start our analysis by 2-dimensional networks and then extend the results to arbitrarily large interconnections of 2-dimensional oscillators.

9.3.1 Two-Dimensional Excitatory-Inhibitory Oscillators

An important property of biological neuronal networks, known as Dale's law [4, 17], is that each node has either an excitatory (E) or inhibitory (I) effect on other nodes, but not both. This means that each column of $\mathbf{W}$ is either nonnegative or nonpositive. Thus, a 2-dimensional network can be either E-E, I-I, or E-I. The latter, hereafter called an *E-I pair*, is also known as the Wilson-Cowan model and has been widely used in computational neuroscience for decades [4–8]. Unlike the standard Wilson-Cowan model that uses sigmoidal activation functions, we show in the following that a complete characterization of limit cycles can be obtained for E-I pairs with bounded linear-threshold nonlinearities.

According to the Poincaré-Bendixson theory [18, Ch 3], in a two-dimensional system ($N = 2$), the lack of stable equilibria is, under mild conditions, necessary and sufficient for the existence of almost globally (excluding trajectories starting at an unstable equilibrium) asymptotically stable limit cycles. To study the equilibria of (9.1), we use its representation as a switched affine system. It is straightforward to show [11] that $\mathbb{R}^N$ can be decomposed into 3^N switching regions $\{\Omega_{\boldsymbol{\sigma}}\}_{\boldsymbol{\sigma}\in\{0,\ell,\mathrm{s}\}^N}$ defined by

$$\begin{aligned}
\Omega_{\boldsymbol{\sigma}} = \{\mathbf{x} \mid (\mathbf{W}\mathbf{x}+\mathbf{p})_i \in (-\infty, 0], \qquad & \forall i \text{ s.t. } \sigma_i = 0, \text{ and} \\
(\mathbf{W}\mathbf{x}+\mathbf{p})_i \in [0, m_i], \qquad & \forall i \text{ s.t. } \sigma_i = \ell, \text{ and} \\
(\mathbf{W}\mathbf{x}+\mathbf{p})_i \in [m_i, \infty), \qquad & \forall i \text{ s.t. } \sigma_i = \mathrm{s}\},
\end{aligned}$$

where 0, ℓ, and s denote inactive, active (linear), and saturated nodes, respectively. Thus, (9.1) can

be rewritten in the switched affine form

$$\tau\dot{\mathbf{x}} = (-\mathbf{I} + \mathbf{\Sigma}^{\ell}\mathbf{W})\mathbf{x} + \mathbf{\Sigma}^{\ell}\mathbf{p} + \mathbf{\Sigma}^{\mathrm{s}}\mathbf{m}, \qquad \forall \mathbf{x} \in \Omega_{\boldsymbol{\sigma}}, \tag{9.2}$$

where for any $\boldsymbol{\sigma} \in \{0, \ell, \mathrm{s}\}^N$, $\mathbf{\Sigma}^{\ell} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with diagonal entries

$$\Sigma^{\ell}_{ii} = \begin{cases} 1 & \text{if} \quad \sigma_i = \ell, \\ 0 & \text{if} \quad \sigma_i = 0, \mathrm{s}, \end{cases}$$

and, likewise, $\mathbf{\Sigma}^{\mathrm{s}} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with diagonal entries

$$\Sigma^{\mathrm{s}}_{ii} = \begin{cases} 1 & \text{if} \quad \sigma_i = \mathrm{s}, \\ 0 & \text{if} \quad \sigma_i = 0, \ell. \end{cases}$$

Each $\Omega_{\boldsymbol{\sigma}}$ then has a corresponding *equilibrium candidate*

$$\mathbf{x}^*_{\boldsymbol{\sigma}} = (\mathbf{I} - \mathbf{\Sigma}^{\ell}\mathbf{W})^{-1}(\mathbf{\Sigma}^{\ell}\mathbf{p} + \mathbf{\Sigma}^{\mathrm{s}}\mathbf{m}),$$

and the equilibria of (9.1) consist of all $\mathbf{x}^*_{\boldsymbol{\sigma}}$ that belong to their respective switching regions. This allows us to derive an exact characterization of limit cycles for E-I pairs, as stated next.

Theorem 9.3.1. ***(Limit cycles in E-I pairs).*** *Consider the dynamics* (9.1) *with* $N = 2$ *and*

$$\mathbf{W} = \begin{bmatrix} a & -b \\ c & -d \end{bmatrix}, \qquad a, b, c, d \geq 0.$$

All network trajectories (except those starting at an unstable equilibrium, if any) converge to a limit cycle if and only if

$$d + 2 < a, \tag{9.3a}$$

$$(a-1)(d+1) < bc, \tag{9.3b}$$

$$(a-1)m_1 < bm_2, \tag{9.3c}$$

$$0 < p_1 < bm_2 - (a-1)m_1, \tag{9.3d}$$

$$0 < (d+1)p_1 - bp_2 < \big[bc - (a-1)(d+1)\big]m_1. \tag{9.3e}$$

Proof. By [21, Thm 4.1], all the trajectories (except those starting at unstable equilibria, if any) converge to a limit cycle if and only if the network does not have any stable equilibria. If $a < 1$, then all the regions $\Omega_{\boldsymbol{\sigma}}, \boldsymbol{\sigma} \in \{0, \ell, \mathrm{s}\}^2$ are stable, ensuring the existence of a stable equilibrium (since the existence of an equilibrium is always guaranteed by the Brouwer fixed point theorem [22]). Thus, assume $a \geq 1$. An exhaustive inspection of switching region dynamics shows that the trivially stable regions $(\sigma', j), \sigma' \in \{0, \mathrm{s}\}, j \in \{0, \ell, \mathrm{s}\}$ do not contain their equilibrium candidates if and only if $\mathbf{p} \in Y^c$ where

$$Y = \Big\{ (p_1, p_2) \mid p_1 \leq \max\Big\{0, \min\{bm_2, \frac{b}{d+1}p_2\}\Big\} \text{ or } \\ p_1 \geq -(a-1)m_1 + \min\Big\{bm_2, \max\{0, \frac{b(p_2 + cm_1)}{d+1}\}\Big\}\Big\}.$$

Therefore, $\mathbf{p} \in Y^c$ if and only if

$$p_1 > 0, \tag{9.4a}$$

$$p_1 < bm_2 - (a-1)m_1, \tag{9.4b}$$

$$p_1 > \min\{bm_2, \frac{b}{d+1}p_2\}, \tag{9.4c}$$

$$p_1 < -(a-1)m_1 + \max\{0, \frac{b(p_2 + cm_1)}{d+1}\}. \tag{9.4d}$$

For (9.4) to be feasible, it is necessary and sufficient that

$$\text{(9.4a) and (9.4b)} : bm_2 - (a-1)m_1 > 0, \tag{9.5a}$$

$$\text{(9.4a) and (9.4d)} : p_2 > -\frac{bc - (a-1)(d+1)}{b}m_1, \tag{9.5b}$$

$$\text{(9.4b) and (9.4c)} : p_2 < \frac{d+1}{b}(bm_2 - (a-1)m_1), \tag{9.5c}$$

$$\text{(9.4c) and (9.4d)} : bc > (a-1)(d+1). \tag{9.5d}$$

Conditions (9.5a) and (9.5d) are the same as (9.3c) and (9.3b), respectively. Furthermore, under (9.5), (9.4) simplifies to (9.3d) and (9.3e), which in turn ensure (9.5b) and (9.5c). In conclusion, $\mathbf{p} \in Y^c$ if and only if (9.3b)-(9.3e) hold.

What remains to study are the regions $(\ell, 0)$, (ℓ, s), and (ℓ, ℓ). The first two are not stable since $a \geq 1$. Also, though not needed, they do not include their equilibrium candidates due to (9.3d). On the other hand, for $\boldsymbol{\sigma} = (\ell, \ell)$, we have the equilibrium candidate

$$\mathbf{x}^*_{\boldsymbol{\sigma}} = \frac{1}{bc - (a-1)(d+1)} \begin{bmatrix} (d+1)p_1 - bp_2 \\ cp_1 - (a-1)p_2 \end{bmatrix} = \mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}} + \mathbf{p}.$$

The first component of $\mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}} + \mathbf{p}$ clearly belongs to $[0, m_1]$ by (9.3b) and (9.3e). For the second component of $\mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}} + \mathbf{p}$, we have[2]

$$(9.3b) \Rightarrow \frac{c}{a-1} > \frac{d+1}{b} \overset{(9.3e)}{\Rightarrow} cp_1 > (a-1)p_2,$$

and

$$\begin{aligned}(9.3d) \Rightarrow \frac{d+1}{b}p_1 - \frac{bc-(a-1)(d+1)}{b}m_1 \\ > \frac{c}{a-1}p_1 - \frac{bc-(a-1)(d+1)}{a-1}m_2 \\ \overset{(9.3e)}{\Rightarrow} p_2 > \frac{c}{a-1}p_1 - \frac{bc-(a-1)(d+1)}{a-1}m_2,\end{aligned}$$

ensuring that $\boldsymbol{\sigma} = (\ell, \ell)$ always contains its equilibrium candidate. Therefore, this region must be unstable which, under (9.3b), happens if and only if $a > d + 2$. This completes the proof. □

The conditions of Theorem 9.3.1 have simple biological intuitions. Equation (9.3a) requires the positive feedback among the neurons of the excitatory population[3] to be sufficiently stronger than the negative feedback among the inhibitory population. This, together with the strong mutual coupling (9.3b) between the two populations, ensures local instability of the equilibrium point $\mathbf{x}^*_{(\ell,\ell)}$ and prevents the oscillations from damping. On the other hand, condition (9.3c) ensures that the *upper bound* on the inhibitory input to the excitatory population (bm_2) is high enough to balance the strong self-excitation. This is consistent with thin spike widths and high firing rates of the inhibitory "fast-spiking interneurons" in the cortex and the theory of excitatory-inhibitory (E-I) balance [23].

[2]We assume $a \neq 1$ because $(\mathbf{W}\mathbf{x}^*_{\boldsymbol{\sigma}} + \mathbf{p})_2 \in [0, m_2]$ trivially if $a = 1$.

[3]Recall that each node of the network dynamics (9.1) represents one population of neurons with similar activity patterns.

Finally, the conditions (9.3d) and (9.3e) require that the external inputs to the two nodes are neither excessively low nor excessively high, as it would keep the respective nodes in negative or positive saturation, resp., which would reduce the effective dimensionality of the network to less than two and make oscillations impossible. We build on this result next to study the oscillatory behavior of a network of oscillators, each represented by an E-I pair.

9.3.2 Networks of Two-Dimensional Oscillators

Consider n oscillators, each modeled by an E-I pair, connected over a network with adjacency matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{n \times n}$ via their excitatory nodes [24]. Since $\mathbf{A}$ captures inter-oscillator connections, its diagonal entries are zero. Thus, the dynamics of the resulting network of networks is

$$\mathbf{T}\dot{\mathbf{x}} = -\mathbf{x} + [\mathbf{W}\mathbf{x} + \mathbf{p}]_{\mathbf{0}}^{\mathbf{m}}, \tag{9.6a}$$

where

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1^T & \dots & \mathbf{x}_n^T \end{bmatrix}^T, \quad \mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ \\ x_{i,2} \end{bmatrix}, \tag{9.6b}$$

$$\mathbf{T} = \text{diag}(\tau_1, \tau_1, \tau_2, \tau_2, \dots, \tau_n, \tau_n), \tag{9.6c}$$

$$\mathbf{W} = \text{diag}(\mathbf{W}_1, \dots, \mathbf{W}_n) + \mathbf{A} \otimes \mathbf{E}, \ \mathbf{E} = \begin{bmatrix} 1 & 0 \\ \\ 0 & 0 \end{bmatrix}, \tag{9.6d}$$

$$\mathbf{W}_i = \begin{bmatrix} a_i & -b_i \\ \\ c_i & -d_i \end{bmatrix}, \quad A_{ii} = 0, \quad i \in \{1, \dots, n\}, \tag{9.6e}$$

p and **m** have similar decompositions to **x**, and $\otimes$ denotes the Kronecker product.

We consider the case where each E-I pair oscillates on its own. The first question we address is whether the pairs maintain any oscillatory behavior after their interconnection. Since conditions for the existence of limit cycles in systems with higher than two dimensions are in general unknown, we use the lack of stable equilibria (which constitutes the main condition in the Poincaré-Bendixson theory for existence of limit cycles) as a proxy for oscillations. Later in Section 9.3.2, we show numerically that this proxy is indeed a tight characterization of oscillatory dynamics.

Theorem 9.3.2. ***(Lack of stable equilibria in networks of E-I pairs).*** *Consider the dynamics* (9.6) *and assume that each* $\mathbf{W}_i$ *satisfies the conditions of Theorem 9.3.1. Then, the overall network does*

not have any stable equilibria if and only if

$$\sum_{j=1}^{n} A_{ij} m_{j,1} < \bar{p}_{i,1} - p_{i,1}, \tag{9.7}$$

$$\bar{p}_{i,1} \triangleq b_i \min\left\{ m_{i,2}, \frac{p_{i,2} + c_i m_{i,1}}{d_i + 1} \right\} - (a_i - 1) m_{i,1},$$

holds for at least one $i \in \{1, \ldots, n\}$. *Moreover, the state of any E-I pair for which* (9.7) *holds may not converge to a fixed value (except for trivial solutions starting at unstable equilibria, if any) irrespective of the validity of* (9.7) *for other pairs.*

Proof. Consider an arbitrary $\boldsymbol{\sigma} \in \{0, \ell, \mathrm{s}\}^{2n}$ and let $L \subseteq \{1, \ldots, n\}, |L| = r$ be the set of pairs whose respective switching region from $\boldsymbol{\sigma}$ is unstable (i.e., $\boldsymbol{\sigma}_i = (\ell, j), j \in \{0, \ell, \mathrm{s}\}, i \in L$). Let $\boldsymbol{\Pi} = \bar{\boldsymbol{\Pi}} \otimes \mathbf{I}_2$ be the permutation matrix that permutes the pairs such that these r pairs are placed first. Then,

$$\boldsymbol{\Pi}(-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W})\boldsymbol{\Pi}^T = \begin{bmatrix} \mathbf{R} & \star \\ \mathbf{0} & \mathbf{N} \end{bmatrix},$$

$$\mathbf{R} = -\mathbf{I} + \boldsymbol{\Sigma}_L(\mathrm{diag}(\{\mathbf{W}_i\}_{i \in L}) + \mathbf{A}_L \otimes \mathbf{E}),$$

$$\mathbf{N} = -\mathbf{I} + \boldsymbol{\Sigma}_{L^c} \mathrm{diag}(\{\mathbf{W}_i\}_{i \in L^c}),$$

where L^c is the complement of L, and $\boldsymbol{\Sigma}_L$ is the $2r \times 2r$ principal submatrix of $\boldsymbol{\Sigma}$ consisting of rows and columns corresponding to the pairs in L. $\mathbf{A}_L$ and $\boldsymbol{\Sigma}_{L^c}$ are defined similarly. Therefore, the eigenvalues of $-\mathbf{I} + \boldsymbol{\Sigma}\mathbf{W}$ consist of those of $\mathbf{R}$ and $\mathbf{N}$.

$\mathbf{N}$ has $n-r$ eigenvalues equal to -1 and $n-r$ eigenvalues that equal $-1-d_i$ or -1, depending

on whether $\sigma_{i,2} = \ell$ or not for each $i \in L^c$. On the other hand, if $r > 0$, then

$$\begin{aligned}\operatorname{tr}(\mathbf{R}) &= \operatorname{tr}(-\mathbf{I} + \boldsymbol{\Sigma}_L \operatorname{diag}(\{\mathbf{W}_i\}_{i\in L})) \\ &\geq \operatorname{tr}(-\mathbf{I} + \operatorname{diag}(\{\mathbf{W}_i\}_{i\in L})) = \sum_{i=1}^{r} a_i - d_i - 2 > 0.\end{aligned}$$

Therefore, any switching region $\Omega_{\boldsymbol{\sigma}}$ is stable *if and only if* $\sigma_{i,1} \neq \ell$ for all $i \in \{1, \dots, n\}$. To prove the sufficiency of (9.7), consider any stable $\Omega_{\boldsymbol{\sigma}}$. Then, if (9.7) holds for even one $i \in \{1, \dots, n\}$,

$$p_{i,1} + \sum_{j=1}^{n} A_{ij}(\mathbf{x}^*_{\boldsymbol{\sigma}})_{j,1} \leq p_{i,1} + \sum_{j=1}^{n} A_{ij} m_{j,1} \overset{(9.7)}{<} \bar{p}_{i,1},$$

ensuring $\mathbf{x}^*_{\boldsymbol{\sigma}} \notin \Omega_{\boldsymbol{\sigma}}$ (by Theorem 9.3.1) and the sufficiency of (9.7). Regarding the last statement of the theorem, note that for $\mathbf{x}_i$ to converge to a fixed value, $\sum_j A_{ij}\mathbf{x}_{j,1}(t)$ must either also converge to a fixed value or be greater than or equal to $\bar{p}_{i,1} - p_{i,1}$ for sufficiently large t, both contradicting (9.7).

To prove the necessity of (9.7), assume that it does not hold for any i or, in other words, at least one of

$$p_{i,1} + \sum_{j=1}^{N} A_{ij} m_{j,1} > b_i m_{i,2} - (a_i - 1) m_{i,1}, \tag{9.8a}$$

or

$$p_{i,1} + \sum_{j=1}^{N} A_{ij} m_{j,1} > \frac{b_i(p_{i,2} + c_i m_{i,1})}{d_i + 1} - (a_i - 1) m_{i,1}, \tag{9.8b}$$

holds for all $i \in \{1, \dots, n\}$. Now, define $\boldsymbol{\sigma} \in \{0, \ell, \mathrm{s}\}^n$ by

$$\boldsymbol{\sigma}_i = \begin{cases} (\mathrm{s}, \mathrm{s}) & ; \quad \text{if } p_{i,2} \geq (d_i + 1)m_{i,2} - c_i m_{i,1}, \\ (\mathrm{s}, \ell) & ; \quad \text{if } p_{i,2} < (d_i + 1)m_{i,2} - c_i m_{i,1}. \end{cases}$$

Note that (9.8b) implies (9.8a) if $p_{i,2} \geq (d_i+1)m_{i,2}-c_i m_{i,1}$ and (9.8a) implies (9.8b) otherwise. Given that all the excitatory nodes are at saturation in $\boldsymbol{\sigma}$, it is not difficult to show that $\Omega_{\boldsymbol{\sigma}}$ (which is stable, by the reasoning above) contains its equilibrium, completing the proof of necessity of (9.7). □

Theorem 9.3.2 provides a precise characterization of the lack of stable equilibria for the network dynamics (9.6). Even though the lack of stable equilibria is in principle neither necessary nor sufficient for the existence of limit cycles, we show next that it is in fact almost necessary and sufficient for the existence of oscillatory behavior. Nevertheless, such oscillatory behavior is often chaotic, not a limit cycle, which may have more relevance for neuronal oscillations [25].

9.4 Oscillatory Properties and Coupling

In this section, we focus on the properties of oscillations generated by (9.6) under the conditions of Theorem 9.3.2. First, we show that the lack of stable equilibria (and thus (9.7)) is indeed a tight proxy for existence of oscillations. Then, motivated by the experimental and computational evidence in brain networks, we study the phenomena of synchronization and phase-amplitude coupling.

9.4.1 Regularity of Oscillations

To assess the oscillatory behavior of the networks that satisfy (9.7), we construct random networks according to

$$\begin{aligned}
&d_i \sim \mathcal{U}(0, d_{\max}), \quad a_i \sim \mathcal{U}(a_{\min}, a_{\max}),\ a_{\min} > d_{\max} + 2,\\
&b_i = c_i \sim \mathcal{U}(b_{\min}, b_{\max}),\ b_{\min} > \sqrt{(a_{\max} - 1)(d_{\max} + 1)},\\
&m_{j,i} \sim \mathcal{U}(m_{j,\min}, m_{j,\max}),\ m_{2,\min} > \frac{a_{\max} - 1}{b_{\min}} m_{1,\max},\\
&\tau_i \sim \mathcal{U}(\tau_{\min}, \tau_{\max}),\ \text{ i.i.d. } \forall j = 1, 2, i \in \{1, \dots, n\},
\end{aligned} \tag{9.9}$$

all satisfying (9.3a)-(9.3c). The values of $p_{i,1}$ and $p_{i,2}$ are always chosen at the center of their respective ranges in (9.3d)-(9.3e) in order for the E-I pairs to oscillate at their maximum amplitude before interconnection. For $\mathbf{A}$, we first generate a random $\mathbf{G} \in \mathbb{R}_{\geq 0}^{n \times n}$ with zero diagonal and i.i.d. $\mathcal{U}(0, 1)$-distributed off-diagonal entries and set

$$\mathbf{A} = \eta \bar{\mathbf{A}}, \quad \bar{\mathbf{A}} = \operatorname{diag}(\bar{\mathbf{p}}_1 - \mathbf{p}_1)\mathbf{G}[\operatorname{diag}(\mathbf{G}\mathbf{1}_n)\operatorname{diag}(\mathbf{m}_1)]^{-1}.$$

$\mathbf{A}$ then satisfies (9.7) for all $i \in \{1, \dots, n\}$ if and only if $\eta \in [0, 1)$.

To measure the existence of oscillations, we use the notion of regularity of oscillations. Given a *zero-mean* signal $x(t)$, we construct a *regularity index* as follows. Let $X(f)$ be the Fourier transform of $x(t)$, $f_{\max} = \arg\max_f |X(f)|$, and

$$\chi_{\text{reg}} = \frac{|X(f_{\max})|}{\max\{|X((1 - \epsilon) f_{\max})|, |X((1 + \epsilon) f_{\max})|\}} \in [1, \infty),$$

where $\epsilon \in (0, 1)$. $\chi_{\text{reg}} = 1$ indicates a flat power spectrum (lack of oscillations) whereas $\chi_{\text{reg}} \to \infty$ indicates a Dirac delta at $f_{\max}$ (perfectly regular oscillations). In practice, values of $\chi_{\text{reg}} \gtrsim 2$ for $\epsilon \lesssim 0.1$ capture oscillatory behavior, with more regularity (less chaotic behavior) as χ_{reg} grows.

Figure 9.2(a) shows the probability distribution of χ_{reg} for random networks of $n = 10$ oscillators ($N = 20$ nodes), $\epsilon = 0.1$, and varying interconnection strength η. For disconnected oscillators ($\eta = 0$), each oscillator has a perfectly regular oscillation (by Theorem 9.3.1) and thus very large χ_{reg} (though finite due to finite signal length and numerical error). These oscillations lose their regularity as we increase the connection strength η towards 1, but still persist up to $\eta = 0.99$, showing the almost sufficiency of (9.7). Further, moving beyond $\eta = 1$, about 10% of oscillations persist at $\eta = 1.01$ but all disappear at $\eta = 1.05$ due to convergence to the stable equilibria ensured by Theorem 9.3.2. This shows that (9.7) is also almost necessary for existence of oscillations in the network dynamics (9.6).

In addition to η, the regularity of oscillations also depend on the network size. Figure 9.2(b) shows the distribution of χ_{reg} for networks of varying size at $\eta = 0.9$. Interestingly, network oscillations lose regularity as we increase network size, which is in line with existing observations on the relation between chaos and network size [26].

Figure 9.2 suggests, indirectly via the regularity of oscillations, that the network dynamics (9.6) become increasingly chaotic as either n or η increases. To assess this more directly, we compute the maximal Lyapunov exponent (MLE) for random networks with the same statistics as (9.9), cf. Figure 9.3. MLE measures the exponential rate at which the norm of the solutions of the linearization of the dynamics around a certain trajectory (network attractor in this case) grow or decay. Therefore, a positive MLE is traditionally used as an indication of chaos [27]. As expected, Figure 9.3 shows a clear increase in MLE both as a function of $\eta < 1$ and n, while moving η beyond

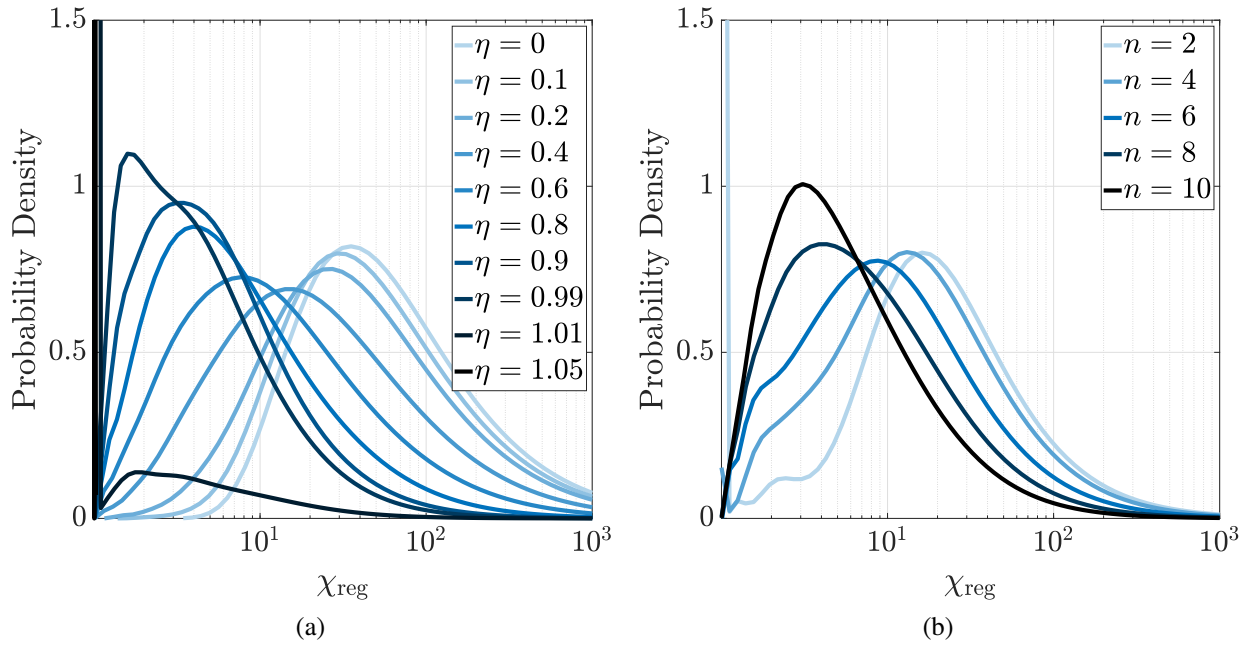


Figure 9.2: Regularity of oscillations as a function of network size (n) and inter-oscillator connection strength (η). The probability density function of $\log \chi_{\text{reg}}$ is plotted for **(a)** $n = 10$ and varying η and **(b)** $\eta = 0.9$ and varying n. Each distribution is based on 500 random networks (9.9) with $d_{\max} = 1$, $a_{\min} = 3.5$, $a_{\max} = 5$, $b_{\min} = \sqrt{8} + 0.5$, $b_{\max} = \sqrt{8} + 2$, $m_{1,\min} = 1$, $m_{1,\max} = 2$, $m_{2,\min} = 8/b_{\min} + 0.5$, $m_{2,\max} = 8/b_{\min} + 2$, $\tau_{\min} = 1$, $\tau_{\max} = 10$.

1 rapidly decreases MLE.

Somewhat surprisingly, even though at $\eta = 0$ each E-I pair has a perfectly regular oscillation (limit cycle) giving an individual MLE of 0 (see also [28, 29]), the network dynamics (9.6) is still slightly chaotic, potentially due to the mismatch between the periods of the individual oscillators. Interestingly, increasing η up to ~ 0.2 *enhances* order among the oscillators due to their effort to synchronize. This further motivates the analysis of synchronization within the network (cf. Problem 6(iii)), which we tackle next.

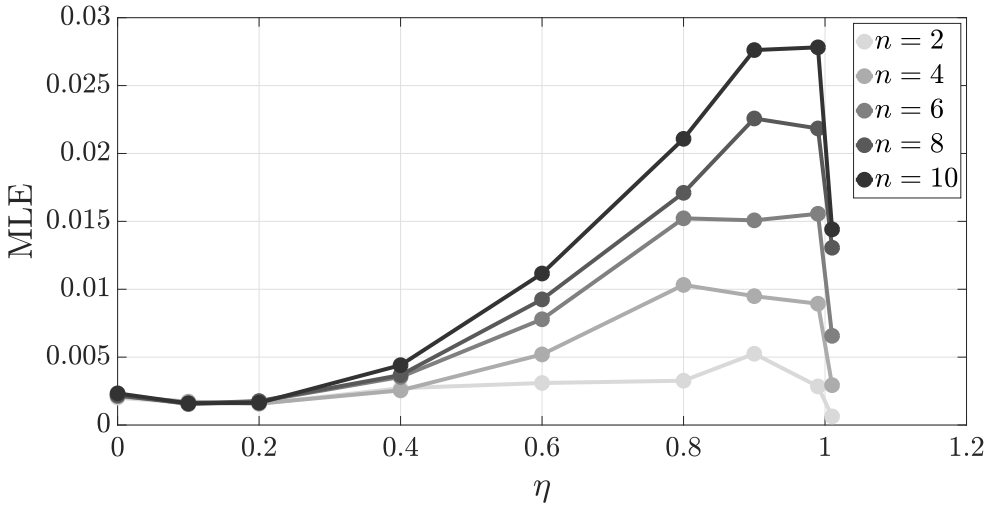


Figure 9.3: Maximal Lyapunov exponent for varying network size n and inter-oscillator connection strength η. Each point is the average MLE of 200 networks with the same statistics as in Figure 9.2.

9.4.2 Synchronization and Phase-Amplitude Coupling

The literature is rich in measures of synchronization, see e.g., [30] for a review and comparison of different methods. Given two discrete signals $z_1(k), z_2(k), k \in \{1, \dots, K\}$, we use the measure of phase synchronization

$$\chi_{\text{sync}} = \left| \frac{1}{K} \sum_{k=1}^{K} e^{j(\phi_1(k) - \phi_2(k))} \right| \in [0, 1], \tag{9.10}$$

where $\phi_i(k)$ is the instantaneous phase[4] of $\{z_i(k)\}_{k=1}^{K}$. This is simply a circular average of the phase difference between the two oscillators, giving a value of 1 if the two oscillators are phase-locked and about 0 if they oscillate independently.[5]

Figure 9.4(a) shows the average value of χ_{sync} as a function of interconnection strength η for pairs of oscillators ($n = 2, N = 4$) with the same statistic as in (9.9), except that the values of

[4]We here use Hilbert transform to obtain the instantaneous phase. For a review and comparison of different methods, see [31].

[5]To avoid edge effects and initial transients, we always compute (9.10) over a middle portion of $\{\phi_i(k)\}_{k=1}^{K}, i = 1, 2$, for K equal to 10^3 times the period of the slower oscillator.

the time constants τ_1 and τ_2 are chosen precisely to obtain a desired ratio ω_1/ω_2 of their natural frequencies. Similar to networks of phase oscillators such as the Kuramoto model [12], networks with $\omega_1 = \omega_2$ are always synchronized irrespective of η, while synchronization increases with η and decreases with frequency mismatch ω_1/ω_2. However, the important distinction with the Kuramoto model is that here it is not possible to fully synchronize arbitrary pairs of oscillators by increasing their connection strength since oscillations vanish for $\eta > 1$ (the so-called *oscillator death* due to saturation [16]). This results in a more realistic synchronization scheme and is consistent with the fact [15] that the Kuramoto model approximates E-I dynamics similar to (9.3) only locally around $\eta = 0$ (a.k.a. weakly coupled oscillators).

Next, we move to the analysis of PAC as given in Problem 6(iii). Here, we study the same random networks of $n = 2$ oscillators as above and see how strongly the phase of the slower oscillator affects the frequency of the faster one. To measure PAC in any signal[6] $\{z(k)\}_{k=1}^{K}$, we use the measure

$$\chi_{\text{PAC}} = \frac{D_{\text{KL}}(P_A \parallel \mathcal{U}(-\pi, \pi))}{\log(N_{\text{bin}})} \in [0, 1],$$

recommended in [32] following a comparison of several measures available in the literature. Here, we first bandpass-filter z around the two frequency ranges of interest to obtain a slow component z_{slow} and a fast one z_{fast}. Then, we bin the instantaneous phases of z_{slow} into N_{bin} bins and for each bin, compute the average instantaneous amplitude of z_{fast} over that bin.[7] This gives a phase distri-

[6] Note that unlike synchronization, PAC is defined and measured for a single signal, even though it arises as a result of the interaction between two oscillators. Throughout, we measure PAC using the state of the excitatory node of the faster oscillator.

[7] We compute both the instantaneous phases and amplitudes using the Hilbert transform and use $N_{\text{bin}} = 10$ throughout.

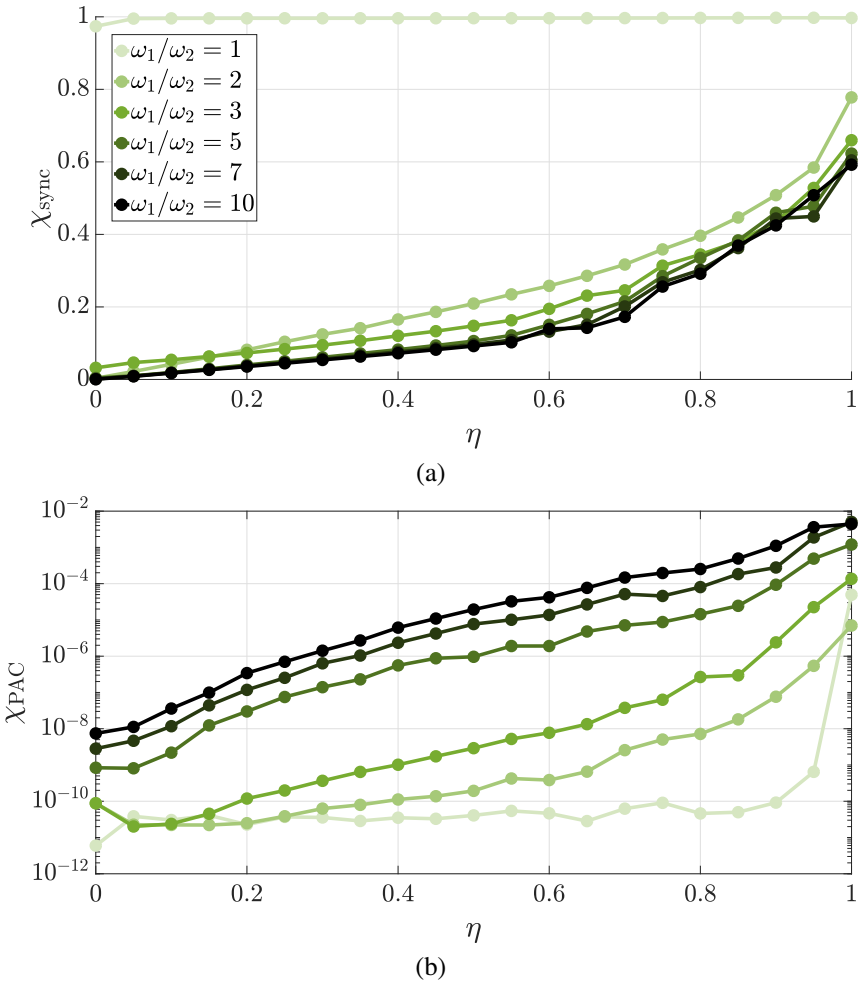


Figure 9.4: Cross-frequency coupling between pairs of oscillators. The average value of **(a)** χ_{sync} and **(b)** χ_{PAC} is plotted for pairs of oscillators with varying ratios of natural frequencies ω_1/ω_2 and connection strength η and the same statistics as in Figure 9.2.

bution P_A over $[-\pi, \pi]$ that is uniform in the absence of PAC but is centered around a "preferred phase" if the amplitude of z_{fast} is larger at a certain phase of z_{slow}. The measure χ_{PAC} then computes the KL divergence of P_A from the uniform distribution, normalized by its maximum possible value.[8]

Figure 9.4(b) shows the value of χ_{PAC} for the same networks as in Figure 9.4(a). Interestingly, χ_{PAC} also increases as a function of η, similarly to χ_{sync} but it *increases* as a function of

[8] As a reference, $\chi_{\text{PAC}} \sim 10^{-4}$ for θ-γ coupling in rodents hippocampus [32] (being a prominent example of PAC in neural data).

frequency mismatch between the oscillators. This shows, for the first time, a clear trade-off between synchronization and PAC, with χ_{PAC} reaching in vivo values of $\sim 10^{-4}$ only for large values of frequency mismatch $\omega_1/\omega_2 \gtrsim 5$ and strong coupling $\eta \gtrsim 0.7$. Note that this tradeoff (and PAC in general) cannot be observed or explained using models of phase oscillators that exclude amplitude dynamics, such as Kuramoto. These results also match observations in the brain, where the most prominent examples of PAC are between theta (4-8$^{\text{Hz}}$) and gamma (30-100$^{\text{Hz}}$) frequency ranges with $\omega_1/\omega_2 \gtrsim 5$ [32], providing an exciting and promising encouragement for further analysis and understanding of the structure of the underlying brain networks.

Appendix

9.A Auxiliary Result

Lemma 9.A.1. ***(Union of intersections with a partition).*** *Consider a collection of sets $\{A_j\}_{j=1}^J$ in a universal set U and a corresponding collection of sets $\{B_j\}_{j=1}^J$ that partition U. Then*

$$\left[\bigcup_{j=1}^{J}(A_j \cap B_j)\right]^c = \bigcup_{j=1}^{J}(A_j^c \cap B_j). \tag{9.11}$$

Proof. Let C denote the left hand side of (9.11). Then for any $j \in \{1, \dots, J\}$,

$$\begin{aligned} C \cap B_j &= \bigcap_{i=1}^{J}\left[(A_i^c \cap B_j) \cup (B_i^c \cap B_j)\right] \\ &= (A_j^c \cap B_j) \cap \bigcap_{i \neq j}\left[(A_j^c \cap B_j) \cup B_j\right] = A_j^c \cap B_j. \end{aligned}$$

The result follows by taking $\bigcup_{j=1}^{J}$ of both sides and using the fact that $\{\boldsymbol{B}_j\}_{j=1}^{J}$ is a partition of $\boldsymbol{U}$. □

Acknowledgements: This chapter is taken, in part, from the work which is to appear as "Oscillations and coupling in interconnections of two-dimensional brain networks" by E. Nozari and J. Cortés in *American Control Conference*, Philadelphia, PA, July 2019. The dissertation author was the primary investigator and author of this paper.

Chapter Bibliography

[1] H. Berger, "Über das elektrenkephalogramm des menschen," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 87, no. 1, pp. 527–570, Dec 1929.

[2] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[3] X. Wang, "Neurophysiological and computational principles of cortical rhythms in cognition," *Physiological reviews*, vol. 90, no. 3, pp. 1195–1268, 2010.

[4] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, vol. 12, no. 1, pp. 1–24, 1972.

[5] B. Baird, "Nonlinear dynamics of pattern formation and pattern recognition in the rabbit olfactory bulb," *Physica D: Nonlinear Phenomena*, vol. 22, no. 1-3, pp. 150–175, 1986.

[6] R. M. Borisyuk and A. B. Kirillov, "Bifurcation analysis of a neural network model," *Biological Cybernetics*, vol. 66, no. 4, pp. 319–325, 1992.

[7] L. H. A. Monteiro, M. A. Bussab, and J. G. C. Berlinck, "Analytical results on a Wilson-Cowan neuronal network modified model," *Journal of Theoretical Biology*, vol. 219, no. 1, pp. 83–91, 2002.

[8] A. C. E. Onslow, M. W. Jones, and R. Bogacz, "A canonical circuit for generating phase-amplitude coupling," *PLOS One*, vol. 9, no. 8, p. e102591, 2014.

[9] S. Campbell and D. Wang, "Synchronization and desynchronization in a network of locally coupled Wilson-Cowan oscillators," *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 541–554, 1996.

[10] K. Morrison, A. Degeratu, V. Itskov, and C. Curto, "Diversity of emergent dynamics in competitive threshold-linear networks: a preliminary report," *arXiv preprint arXiv:1605.04463*, 2016.

[11] E. Nozari and J. Cortés, "Hierarchical selective recruitment in linear-threshold brain networks. Part I: Intra-layer dynamics and selective inhibition," *IEEE Transactions on Automatic Control*, 2018, submitted.

[12] M. Breakspear, S. Heitmann, and A. Daffertshofer, "Generative models of cortical oscillations: neurobiological implications of the Kuramoto model," *Frontiers in human neuroscience*, vol. 4, p. 190, 2010.

[13] L. Tiberi, C. Favaretto, M. Innocenti, D. S. Bassett, and F. Pasqualetti, "Synchronization patterns in networks of Kuramoto oscillators: A geometric approach for analysis and control," in *IEEE Conf. on Decision and Control*, Melbourne, Australia, Dec. 2017, pp. 481–486.

[14] T. Menara, G. Baggio, D. S. Bassett, and F. Pasqualetti, "Stability conditions for cluster synchronization in networks of Kuramoto oscillators," *IEEE Transactions on Control of Network Systems*, 2018, submitted.

[15] H. G. Schuster and P. Wagner, "A model for neuronal oscillations in the visual cortex. 1. mean-field theory and derivation of the phase equations," *Biological Cybernetics*, vol. 64, no. 1, pp. 77–82, 1990.

[16] G. B. Ermentrout and N. Kopell, "Oscillator death in systems of coupled neural oscillators," *SIAM Journal on Applied Mathematics*, vol. 50, no. 1, pp. 125–146, 1990.

[17] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*, ser. Computational Neuroscience. Cambridge, MA: MIT Press, 2001.

[18] L. Perko, *Differential Equations and Dynamical Systems*, 3rd ed., ser. Texts in Applied Mathematics. New York: Springer, 2000, vol. 7.

[19] W. Grasman, "Periodic solutions of autonomous differential equations in higher-dimensional spaces," *The Rocky Mountain Journal of Mathematics*, vol. 7, no. 3, pp. 457–466, 1977.

[20] L. A. Sanchez, "Existence of periodic orbits for high-dimensional autonomous systems," *Journal of Mathematical Analysis and Applications*, vol. 363, no. 2, pp. 409–418, 2010.

[21] S. Simic, K. H. Johansson, J. Lygeros, and S. Sastry, "Hybrid limit cycles and hybrid Poincaré-Bendixson," in *IFAC World Congress*. IFAC, 2002.

[22] L. E. J. Brouwer, "Über abbildung von mannigfaltigkeiten," *Mathematische Annalen*, vol. 71, no. 1, pp. 97–115, 1911.

[23] S. Deneve and C. K. Machens, "Efficient codes and balanced networks," *Nature Neuroscience*, vol. 19, no. 3, p. 375, 2016.

[24] S. F. Muldoon, F. Pasqualetti, S. Gu, M. Cieslak, S. T. Grafton, J. M. Vettel, and D. S. Bassett, "Stimulation-based control of dynamic brain networks," *PLOS Computational Biology*, vol. 12, no. 9, p. e1005076, 2016.

[25] M. Mannino and S. L. Bressler, "Freeman's nonlinear brain dynamics and consciousness," *Journal of Consciousness Studies*, vol. 25, no. 1-2, pp. 64–88, 2018.

[26] I. Ispolatov, V. Madhok, S. Allende, and M. Doebeli, "Chaos in high-dimensional dissipative dynamical systems," *Scientific Reports*, vol. 5, p. 12506, 2015.

[27] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, "The analysis of observed chaotic data in physical systems," *Reviews of Modern Physics*, vol. 65, no. 4, p. 1331, 1993.

[28] H. Haken, "At least one Lyapunov exponent vanishes if the trajectory of an attractor does not contain a fixed point," *Physics Letters A*, vol. 94, no. 2, pp. 71–72, 1983.

[29] P. C. Müller, "Calculation of Lyapunov exponents for dynamic systems with discontinuities," *Chaos, Solitons & Fractals*, vol. 5, no. 9, pp. 1671–1681, 1995.

[30] T. Kreuz, F. Mormann, R. G. Andrzejak, A. Kraskov, K. Lehnertz, and P. Grassberger, "Measuring synchronization in coupled model systems: A comparison of different approaches," *Physica D: Nonlinear Phenomena*, vol. 225, no. 1, pp. 29–42, 2007.

[31] M. L. van Quyen, J. Foucher, J. P. Lachaux, E. Rodriguez, A. Lutz, J. Martinerie, and F. J. Varela, "Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony," *Journal of Neuroscience Methods*, vol. 111, no. 2, pp. 83–98, 2001.

[32] A. B. L. Tort, R. Komorowski, H. Eichenbaum, and N. Kopell, "Measuring phase-amplitude coupling between neuronal oscillations of different frequencies," *Journal of Neurophysiology*, vol. 104, no. 2, pp. 1195–1210, 2010.

Chapter 10

Conclusions

10.1 Summary

In this dissertation, we have studied several challenges that arise in the implementation of networked dynamical systems theory. Our results are motivated by and applicable to networked dynamical systems in a wide range of domains, from cyber-physical systems (Part I) to small-scale industrial and a variety of large-scale complex systems (Part II) and the brain (Part III). Despite the long-standing and well-studied nature of the problems studied, a common thread throughout the dissertation is the proposition of pioneering solutions to these problems rather than incremental improvements over the existing literature.

A further core is the interdisciplinary theme of the dissertation, bridging systems and control theory, computer science, network science, and cognitive and computational neuroscience. We have sought to simultaneously respect both the mathematical rigor of control theory and the practical considerations (and the fine boundary of “acceptable assumptions") of the respective application domains, as exemplified in our theory of goal-driven selective attention in Chapter 8. This cross-

disciplinary theme is not only a hallmark of contemporary scientific research, but is also a defining pillar of systems and control theory which is critical for filling in the yawning gap between rigorous mathematical and applied/experimental sciences.

In Part I, we have focused on the problem of privacy preservation in networked dynamical systems that perform distributed computations. While examples of distributed computations are numerous, we have studied average consensus and convex optimization as two of the most fundamental computations that have widespread applications *per se* and are building blocks of more complex ones. In both cases, we have employed the popular concept of differential privacy due to its mathematically elegant formulation, independence from side information, independence to the adversarial algorithms or capabilities, and immunity to post-processing. Also in both cases, we have followed a similar theme of first proving an impossibility result that bounds the space of possible differentially private algorithms and then designing a "best achievable" algorithm in this space. These two steps can also be regarded as a necessary and a sufficient condition, respectively, for each problem. We have further provided rigorous analysis of the outcome of our designed algorithms and characterized its relationship with the true (non-private) outcome of the respective computation.

In Part II, we have tackled three problems that all sought to solve an otherwise solved problem under resource constraints. These problems pertain the stabilization (Chapter 5), control (Chapter 6), and identification (Chapter 7) of networked systems with respective constraints on the bandwidth and latency of the network communications, the number of nodes that can be actuated as well as the control energy, and the number of nodes that can be sensed. In the first case, we have built upon the two elegant frameworks of event-triggered and predictor-feedback control to design the first event-triggered controller for nonlinear systems that can tolerate arbitrarily large delays. In

the second case, we have moved beyond the conventional time-invariant control schedules and provided a demonstration of the significant benefits of time-varying control scheduling as well as the connections between this benefit and network structure. In the latter case, we have re-adopted the well-established least-square autoregressive models but proposed their inherent time-lagged structure as a natural and asymptotically exact alternative to existing identification methods for networks with latent nodes.

Finally in Part III, we have focused on the applications of networked control theory to cognition. While the domain-unspecific results of Part II are applicable to brain networks in particular, the latter involves various specific challenges that call for a focused study. One such challenge is the existence of nonlinearities that are essential to the function of the nervous system and cannot be simplified via linearization. To incorporate this essential nonlinearity without losing analytical tractability, we have employed the well-established switched-affine class of linear-threshold models but used them to develop a first axiomatic approach to the study of goal-driven selective attention (Chapter 8) and neural oscillations (Chapter 9). In both cases, the main theme was to draw connections between stereotypical observations in experimental neuroscience and the structure of the underlying brain networks that are generating them. This led us to various necessary and/or sufficient conditions in terms of network structure that we have then proceeded to verify against real data in the case of selective attention. These studies pave the road for future research on various aspects of networked dynamical systems, as we briefly describe next.

10.2 Future Directions

Our results, although comprehensive within their scope, are only first steps towards complete solutions of their respective problems. Regarding differentially private distributed computations, a decently explored but still incomplete avenue of research involves the extension of our results to other distributed computations such as dynamic average consensus, non-convex optimization, filtering and estimation, and identification. A similar line of research can seek the design of algorithms for privacy preservation of the network structure and other parameters such as edge weights and vertex degrees. Moreover, the privacy-accuracy tradeoff in differential privacy is of paramount importance in its real-world applications and characterizing the optimal privacy-accuracy trade-off curve for any of these problems has remained elusive (while numerical estimates for specific algorithms can be obtained, as done here). Along the same lines, a fundamental and highly warranted direction is the analysis of potential alternative definitions of privacy for networked dynamical systems that impose less costs in terms of computational accuracy in exchange of potentially weaker privacy guarantees. This can be beneficial for applications where the significant strength of differential privacy is not necessarily required and can be traded off with higher computational accuracy.

Our design of event-triggered feedback stabilizers relied on the strong assumption of perfect knowledge of system dynamics (i.e., known structural parameters, lack of disturbances, known delay, and error-free numerical implementation) in order to compensate for arbitrarily large delays. To probe how critical these assumptions are, we presented various simulations that suggested rather high degrees of robustness (including input-to-state stability) for the closed-loop system. While encouraging, rigorous characterization of such robustness has remained elusive and remains open

for future research. In addition, the relaxation of our results to the use of output feedback will be a further warranted step towards wider practical applicability.

Our analysis of the benefits of time-varying control scheduling presents an even larger set of open questions. An important limitation of our work was the linearity of dynamics. As we saw in our analysis in Part III (and well-known from the theory of nonlinear systems), nonlinearity is essential for various fundamental aspects of complex systems. Characterizing the interplay between nonlinearity and time-varying actuation is, however, an unexplored topic but invaluable in practice. As explained before, we expect time-varying control scheduling to have an even larger benefit in the presence of nonlinearities such as state saturations due to the additional limitations and complexities that such nonlinearities impose on the spread of control energy from any single node in the network. Also open is the question of how increasing the number of control inputs will affect the benefits of time-varying actuation, as is the question of whether this benefit is dependent on the specific dynamics that evolve over complex networks beyond their static structural connectivity. Finally, a more fundamental quest seeks the definition of controllability metrics for large-scale networked dynamical systems that do not suffer from numerical instabilities and the intrinsic conservatism of the smallest eigenvalue, determinant, and trace of inverse of the Gramian, but do not restrict attention to the few most easy-to-control nodes (as done by the trace of the Gramian) either.

A similar line of research, i.e., the importance and effects of network structure, is warranted for network structure identification. While generic methods such as ours applies to any network structure, some networks may be easier to identify than others and different structural properties may even make networks suitable to be identified with one method better than others. Also our analysis relied on the fundamental assumption that a “manifest" node can be both sensed and actuated while a “latent" node is not only unavailable for sensing but is also subject to no external input

from the environment. Relaxing these conditions and allowing for the existence of nodes that can be sensed but not actuated or vice versa can indeed extend the applicability of the results presented here.

Our development of the hierarchical selective recruitment is also only a first step, which opens the door to numerous further investigations. A limitation of our analysis of feedback selective inhibition is the need for full state feedback, which is unlikely to exist in real brain networks (though is not impossible dependent on the definition and granularity of node definitions). While the natural extension of output feedback remains open for future exploration, a promising starting point is our parallel results on feedforward selective inhibition. The latter can provide significant levels of robustness that can help both state and output feedback mechanisms. Further, two intertwined questions pertain the transfer of (sensory) information along the hierarchy and the encoding of information in more complex (than equilibrium) attractors. Various experimental studies highlight the role of neural oscillations in selective attention which may be the link between these questions.

This was the motivation for our subsequent study of neural oscillations. With regard to this preliminary analysis, various extensions are in order, including generalizations to arbitrary network structures with higher than two dimensions, the analytical characterization of the effects of inter-oscillator connectivity strength and frequency mismatch on synchronization and phase-amplitude coupling, the verification of these results against real in vivo recordings, and generalizations to incorporate conduction delays and noise. Finally, an important but almost unexplored question pertains to the controllability and observability of linear-threshold networks as well as their optimal sensor and actuator placement.