Scalable Large-Scale Control of Network Aggregates

Denis Nikitin

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Automatique - productique**

Arrêté ministériel : 25 mai 2016

Présentée par
Denis NIKITIN

Thèse dirigée par **Carlos CANUDAS DE WIT** et
codirigée par **Paolo FRASCA**

préparée au sein du
Laboratoire Grenoble Images Parole Signal Automatique (GIPSA)
dans **l'École Doctorale Électronique, Électrotechnique, Automatique, Traitement du Signal (EEATS)**

Scalable large-scale control of network aggregates

Thèse soutenue publiquement le **02 septembre 2021**,
devant le jury composé de:

Emmanuel WITRANT
Professeur, Université Grenoble Alpes, Président du jury
Mario DI BERNARDO
Professeur, University of Naples Federico II, Rapporteur
Bassam BAMIEH
Professeur, University of California, Santa Barbara, Rapporteur
Miroslav KRSTIC
Professeur, University of California, San Diego, Examinateur
Brigitte D'ANDREA-NOVEL
Professeure, Institut de Recherche et de Coordination Acoustique/Musique, Examinatrice
Ursula EBELS
Professeure, CEA Grenoble, Invitée

UNIVERSITÉ DE GRENOBLE ALPES

ÉCOLE DOCTORALE EEATS

Électronique, Électrotechnique, Automatique, Traitement du Signal

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble
Mention : Automatique

Présentée et soutenue par

Denis NIKITIN

Scalable large-scale control of network aggregates

Thèse dirigée par Carlos CANUDAS-DE-WIT

et codirigée par Paolo FRASCA

préparée au laboratoire Grenoble Images Parole Signal Automatique (GIPSA-Lab)

soutenue le 02/09/2021

Jury :

Rapporteurs :	Mario DI BERNARDO	-	University of Naples Federico II
	Bassam BAMIEH	-	University of California, Santa Barbara
Président :	Emmanuel WITRANT	-	Université Grenoble Alpes
Examinateurs :	Miroslav KRSTIC	-	University of California, San Diego
	Brigitte D'ANDREA-NOVEL	-	IRCAM
Invitée :	Ursula EBELS	-	CEA Grenoble

Résumé — Cette recherche est réalisée dans le cadre du projet de subvention avancée Scale-FreeBack du Conseil européen de la recherche (ERC). L'objectif du projet Scale-FreeBack est de développer une approche holistique de contrôle sans échelle des systèmes complexes, et de poser de nouvelles bases pour une théorie traitant des réseaux physiques complexes avec une dimension arbitraire. Les contributions du présent travail de thèse sont principalement liées aux problèmes de modélisation et de conception de commandes pour les systèmes à grande échelle. Nous recherchons des représentations de modèles simplifiées à des fins de contrôle pour différentes classes de systèmes à grande échelle, des réseaux aux EDP. Dans cette thèse de doctorat, nous proposons des techniques de conception de commandes qui reposent entièrement sur des modèles agrégés de systèmes originaux à grande échelle. Tout d'abord, nous traitons de grands réseaux linéaires en contrôlant leur état moyen et l'écart de tous les états par rapport à la moyenne. Le problème du contrôle de l'état moyen avec contrôleur intégral est étudié, et une relation simple entre la positivité du système et sa passivité est établie. L'écart est ensuite minimisé via la méthode de recherche d'extremum contraint. Cette approche est généralisée pour contrôler une sortie linéaire multidimensionnelle générale et minimiser simultanément une sortie quadratique scalaire générale. Ensuite, nous tournons notre attention vers les systèmes EDP et une représentation simplifiée de leurs solutions. À savoir, nous développons une technique de réduction de modèle basée sur la forme applicable aux lois de conservation 1D, qui suppose une paramétrisation de forme particulière des solutions de la EDP, puis transforme la EDP en un système d'EDO décrivant l'évolution de ces paramètres de forme. Enfin, nous étudions le problème de la dérivation de représentations continues de systèmes spatialement distribués à grande échelle. À savoir, nous développons une méthode de continuation qui transforme tout système non linéaire général avec une structure spatiale en un modèle EDP. Nous proposons en outre une analyse de la précision et de la convergence d'une telle représentation dans le cas linéaire. La méthode est utile car elle ouvre de nouvelles possibilités pour l'analyse et la conception de contrôle dans le domaine continu pour les systèmes intrinsèquement discrets. Dans la thèse, nous élaborons diverses applications de la méthode de continuation. En particulier, nous appliquons la méthode à plusieurs problèmes de réseaux de transport et de systèmes multi-agents, fournissant des dérivations de modèles continus pour les systèmes de trafic, une solution originale au 6ème problème de Hilbert de la dérivation d'équations d'Euler à partir de systèmes newtoniens de particules, et un contrôle technique de conception d'une grande formation robotique au niveau de la densité. Enfin, nous appliquons la méthode aux réseaux d'oscillateurs à grande échelle (tels que les lasers ou les oscillateurs spin-couple). Les modèles EDP obtenus sont utilisés à des fins de contrôle (telles que la stabilisation des limites via un backstepping basé sur EDP) et pour l'analyse, en dérivant des conditions pour l'existence et la stabilité de solutions synchrones dans des systèmes avec des oscillateurs à la fois homogènes et inhomogènes.

Mots clés : Contrôle de grands systèmes, Réseaux à grande échelle, Équations aux dérivées partielles, Réduction de modèle, Systèmes multi-agents

Abstract — This research is done in the context of European Research Council's (ERC) Advanced Grant project Scale-FreeBack. The aim of Scale-FreeBack project is to develop a holistic scale-free control approach to complex systems, and to set new foundations for a theory dealing with complex physical networks with arbitrary dimension. The contributions of the present PhD work are mainly related to the problems of modeling and control design for large-scale systems. We seek simplified model representations for control purposes for different classes of large-scale systems, from networks to PDEs. Within this PhD thesis, we propose control design techniques that completely rely on aggregated models of original large-scale systems. First of all, we deal with large linear networks by controlling their average state and the deviation of all the states from the average. The problem of controlling the average state with integral controller is studied, and a simple relation between positivity of the system and its passivity is established. The deviation is then minimized via constrained extremum seeking method. This approach is generalized to control a general multidimensional linear output and simultaneously minimize a general scalar quadratic output. Then, we turn our attention to the PDE systems and a simplified representation of their solutions. In particular, we develop a shape-based model reduction technique applicable to 1D conservation laws, which assumes a particular shape parametrization of the PDE's solutions and then transforms the PDE into a system of ODEs describing the evolution of these shape parameters. Finally, we study the problem of deriving continuous representations of large-scale spatially-distributed systems. Namely, we develop a continuation method which transforms any general nonlinear system with spatial structure into a PDE model. We further provide an analysis of accuracy and convergence of such representation in the linear case. The method is useful since it opens new possibilities for analysis and control design in continuous domain for intrinsically discrete systems. In the thesis we elaborate various applications of the continuation method. In particular, we apply the method to several problems of transportation networks and multi-agent systems, providing derivations of continuous models for traffic systems, an original solution to the Hilbert's 6th problem of the derivation of Euler equations from Newtonian systems of particles, and a control design technique for a large robotic formation on a density level. Finally we apply the method to the large-scale networks of oscillators (such as lasers or spin-torque oscillators). The obtained PDE models are used for control purposes (such as boundary stabilization via PDE-based backstepping) and for the analysis, deriving conditions for the existence and stability of synchronous solutions in systems with both homogeneous and inhomogeneous oscillators.

Keywords: Control of Large Systems, Large-Scale Networks, Partial Differential Equations, Model Reduction, Multi-Agent Systems

Grenoble Images Parole Signal Automatique (GIPSA-Lab)
11 Rue des Mathématiques, 38400 Saint-Martin-d'Hères

Résumé

Nous vivons dans un monde complexe dans lequel tous les processus sont interconnectés. L'approche scientifique initiale en physique, ingénierie et théorie du contrôle consistait à isoler les sous-systèmes individuels, à simplifier les modèles autant que possible et à les explorer dans leur forme la plus pure, mais dans le monde d'aujourd'hui, nous devons faire face à des systèmes complexes qui ne peuvent pas être considérés comme la somme de leurs sous-systèmes indépendants constitutifs. Les structures créées par l'homme, comme le trafic urbain, les réseaux sociaux, les réseaux électriques ou les réseaux de lasers, peuvent avoir des milliers, voire des millions de degrés de liberté. Pour étudier des systèmes de cette taille et pour des applications pratiques, il est nécessaire de développer de bons modèles et des moyens de les analyser.

Il existe de nombreuses façons de décrire les systèmes à grande échelle. L'une d'elles est un réseau d'équations différentielles ordinaires (EDO), qui décrit l'évolution d'un grand système comme une évolution des états des nœuds dans le temps, en tenant compte des interactions par paires. Dans le même temps, le nombre d'états peut encore être énorme ; il devient nécessaire de développer des méthodes d'analyse évolutives des grands systèmes d'EDO. Une autre solution consiste à décrire l'état du grand système comme un continuum et à utiliser le langage des équations aux dérivées partielles (EDP) qui prédit l'évolution des champs continus dans le temps en fonction des dérivées partielles de ces quantités par rapport à la position. De nombreuses EDP ont été créées et constituent des modèles pour la description de différents processus physiques.

Le contrôle des systèmes dynamiques à grande échelle est un problème difficile pour la théorie moderne du contrôle. Sa difficulté provient de la grande dimensionnalité de ces systèmes du monde réel, où le nombre d'états peut atteindre des millions. Au lieu de développer des algorithmes de contrôle sophistiqués directement pour les grands systèmes, une approche fondamentalement différente du problème du contrôle de ces systèmes est la simplification du modèle. Dans ce paradigme, le modèle d'un système complexe est remplacé par un modèle de plus petite taille et/ou une structure d'interactions plus simple. Un tel processus peut entraîner une perte d'informations sur la dynamique du système original, mais la représentation du système sous une forme plus simple permet d'appliquer des algorithmes de contrôle standard.

Le principal intérêt de notre travail est le problème du contrôle évolutif des grands systèmes. En utilisant différents modèles et structures de grands systèmes comme points de départ, nous étudions différentes options sur la façon dont les systèmes peuvent être simplifiés et utilisés pour l'analyse et la conception du contrôle, en démontrant les résultats dans divers problèmes pratiques.

Les principaux chapitres de cette thèse sont résumés ci-dessous.

Contrôle de la moyenne et de la déviation dans les réseaux linéaires

Dans le Chapitre 2, nous nous concentrons sur le problème du contrôle de l'état moyen du réseau, ainsi que sur la minimisation simultanée de sa déviation. En utilisant un régulateur pour l'état moyen, il est naturel de souhaiter que les états du système soient proches de la moyenne : ce comportement peut être obtenu en minimisant leur déviation au carré. De plus, nous nous assurons que le modèle du système n'est pas utilisé dans le régulateur. Ainsi, le régulateur n'utilise directement que les sorties et le point de référence du système et l'équilibre des états internes n'est jamais calculé explicitement. Le fait de ne pas utiliser le modèle du système permet de contourner tous les problèmes de complexité de calcul et d'incertitudes qui affectent les grands réseaux.

Nous étudions d'abord un problème de contrôle de sortie linéaire et examinons les propriétés générales des fonctions de transfert du système et du régulateur. Nous étudions ensuite le régulateur intégral pour la régulation linéaire de la sortie et formulons une condition suffisante $CA^2 > 0$ pour la convergence de tout régulateur intégral positif, en montrant que la stabilisation de la sortie est atteinte lorsque la fonction de transfert du système est réelle strictement positive (Strictly Positive Real - SPR) et en donnant en plus un exemple montrant le conservatisme de cette condition. Si le système satisfait à cette condition, les paramètres du régulateur peuvent être choisis arbitrairement, et il n'est pas nécessaire de connaître le vecteur d'état ou les valeurs des éléments de la matrice A. Nous avons étendu notre analyse aux systèmes multi-sorties dans le but de contrôler les états moyens de plusieurs grappes et avons dérivé une condition suffisante sur les matrices du système pour que le système multi-sorties soit SPR.

Le contrôle de l'état moyen ne signifie pas que les états des systèmes individuels seront proches de l'état moyen. Par conséquent, en plus de contrôler la moyenne, il est utile de minimiser la déviation des états du système. Pour résoudre ce problème, nous utilisons l'algorithme de Extremum Seeking augmenté de la méthode primal-dual pour la minimisation sous contrainte. La stabilité de ce schéma est prouvée et sa performance, ainsi que celle de plusieurs versions modifiées, est testée dans les simulations numériques.

Réduction des modèles basés sur la forme pour les lois de conservation

Le Chapitre 3 est consacré à une méthode de description d'un système de loi de conservation 1D basée sur la notion de forme de la solution. La fonction de forme décrit la forme de la solution en fonction de plusieurs paramètres bien traitables. Nous réduisons l'état du système à un ensemble de ces paramètres de forme et dérivons leur dynamique, fournissant une solution à forme fermée. Nous analysons ensuite ses propriétés, montrant en particulier que cette solution minimise la distance de Wasserstein entre le système original et le système

réduit et que les points d'équilibre du système original sont préservés. L'idée de représenter la solution du système par des paramètres de forme spécifiques peut potentiellement conduire à de nouveaux types de conception de régulateurs basés sur les caractéristiques agrégées du système.

Méthode de continuation pour la modélisation et le contrôle des systèmes à grande échelle : des EDO aux EDP

Dans le Chapitre 4, nous nous concentrons sur le problème inverse rarement étudié de la transformation d'un système d'EDO en EDP, dans le but de combler cette lacune et de fournir une contrepartie à la procédure de discrétisation. Ceci peut être utile car les EDP fournissent une manière beaucoup plus compacte de décrire le système, qui dans de nombreux cas est plus facile à analyser analytiquement que le système d'EDO correspondant. Nous nous intéressons en particulier aux systèmes qui sont distribués dans l'espace et qui ont une interaction dépendant de la position, comme le trafic en ville, les réseaux électriques, les formations de robots, etc.

Notre idée est de remplacer le système original d'EDO spatialement distribué par une EDP continue dont les variables d'état et d'espace préservent les variables d'état et d'espace du système original. Nous développons une méthode pour les EDO linéaires spatialement invariantes qui les transforme en EDP à l'aide de différences finies. Nous appelons cette méthode une *continuation*, car elle est exactement opposée à la procédure de discrétisation. De plus, nous montrons que le spectre des EDP converge vers le spectre de l'EDO originale lorsque l'ordre de continuation augmente, et que cette convergence fournit une limite sur la déviation entre les solutions des systèmes. En utilisant le formalisme des graphes computationnels, nous étendons la méthode aux systèmes non linéaires, puis aux systèmes multidimensionnels, aux systèmes variant dans l'espace et dans le temps, aux systèmes multi-agents indexés et aux systèmes avec frontières. L'avantage de la méthode de continuation est qu'elle permet de récupérer une EDP qui décrit le même système physique que le réseau d'EDO original.

Applications de la méthode de continuation aux systèmes multi-agents

Une description basée sur les EDP du système physique décrit à l'origine par le réseau d'EDO peut être très utile non seulement pour l'analyse, mais aussi à des fins de contrôle. En effet, on peut utiliser une EDP obtenue pour concevoir une commande continue qui, discrétisée à nouveau, donne une loi de contrôle pour le système ODE original. Dans le Chapitre 5, nous montrons que sur la base de la méthode de continuation, de nouveaux modèles continus peuvent être dérivés et utilisés à des fins d'analyse et de contrôle. Une attention particulière est accordée aux systèmes multi-agents, pour lesquels la méthode de continuation peut être appliquée en utilisant une notion de densité définie comme l'inverse de la dérivée partielle

d'une position par rapport à la fonction d'indexation.

A titre d'exemple, nous utilisons la continuation pour montrer comment divers modèles d'EDP de trafic peuvent être récupérés à partir de représentations discrètes du trafic. Ensuite, nous nous concentrons sur la question de savoir comment les équations d'Euler pour un fluide compressible peuvent être dérivées des interactions entre particules newtoniennes, ce qui permet de mieux comprendre le sixième problème de Hilbert. Enfin, la même suite est ensuite utilisée pour décrire une formation de robots volant à travers une fenêtre. Nous développons un algorithme de contrôle pour stabiliser une trajectoire désirée basée sur une représentation continue de la formation. Cet algorithme est distribué car chaque robot n'a besoin d'information que sur les robots voisins. Une simulation numérique montre que le régulateur proposé est capable d'amener la formation de robots à effectuer les manœuvres souhaitées à la fois en 2D et en 3D.

Applications de la méthode de continuation aux systèmes oscillatoires

Le Chapitre 6 montre comment la méthode de continuation peut être utilisée pour transformer des réseaux oscillants en modèles d'EDP non linéaires, ce qui ouvre de nouvelles possibilités pour l'analyse et le contrôle des phénomènes de synchronisation.

Tout d'abord, un réseau laser est synchronisé en supprimant les oscillations indésirables grâce au fait que le modèle d'EDP du système laser est adapté à un backstepping basé sur les EDP. Nous démontrons par des simulations numériques que l'application d'un régulateur basé sur les EDP au système initialement discret assure la stabilité, tandis que la dérivation d'un tel contrôle continu est simple et explicite.

De plus, nous présentons un réseau d'oscillateurs non linéaires avec des interactions locales, couplés sur un anneau 1D. Nous introduisons une approximation d'EDP pour ce système en utilisant la méthode de continuation. Cette représentation EDP peut être plus appropriée pour l'analyse de la même manière que les systèmes dynamiques continus peuvent être plus faciles à traiter que les systèmes discrets. La question de la dérivation des conditions de synchronisation est ensuite abordée pour le cas particulier des oscillateurs de Kuramoto, puis pour un cas général d'oscillateurs non isochrones. Il apparaît que les EDP non linéaires apparaissant dans ce cas peuvent être analysées pour retrouver des solutions d'équilibre et vérifier leur stabilité. La validation par simulation numérique démontre que les solutions synchrones ainsi obtenues coïncident avec celles vers lesquelles converge le système réel.

Conclusion et perspectives

Bien que les méthodes décrites dans ce travail de recherche aient fourni de bons résultats initiaux, il reste des problèmes et des questions en suspens qui peuvent apporter des amélio-

rations significatives à notre compréhension des méthodes et à leur applicabilité pratique, ce qui peut constituer une base pour une recherche future.

En résolvant le problème du contrôle de l'état moyen et de la déviation des grands réseaux, notre objectif de contrôle était uniquement de conduire l'état moyen du réseau à une valeur fixe désirée en régime permanent, alors que dans le monde réel, la tâche de suivre la valeur au fur et à mesure qu'elle change dans le temps est beaucoup plus importante. En même temps, en supprimant l'hypothèse de l'état stationnaire, il est possible d'améliorer la minimisation de l'écart type dans les processus transitoires. Enfin, nous avons supposé que l'état moyen et l'écart-type pouvaient être mesurés directement, ce qui est une hypothèse relativement forte qui pourrait être relâchée.

La méthode de réduction de modèle basée sur la forme n'est actuellement limitée dans son application qu'à la classe des lois de conservation EDP 1D et seulement pour des périodes de temps limitées jusqu'à ce que la forme sélectionnée devienne dégénérée. La dégénérescence de la forme est une limitation très sérieuse qui pourrait être supprimée si une procédure de reparamétrisation était développée qui corrige automatiquement la forme chaque fois qu'elle devient dégénérée. Il serait également possible d'étudier l'extension de la méthode à d'autres classes de systèmes, y compris divers modèles d'EDP et des réseaux d'EDO avec une structure spatiale.

La suite la plus directe du travail qui a été décrit dans cette thèse de doctorat est une étude plus détaillée de la méthode de continuation, ainsi qu'un développement plus détaillé d'une théorie générale de son application à divers systèmes d'analyse et de contrôle. Premièrement, les résultats analytiques du Chapitre 4 garantissent la convergence des solutions des systèmes EDP vers les solutions ODE tant que l'ordre d'EDP tend vers l'infini. En réalité, en raison du manque de méthodes d'analyse des EDP pour les ordres élevés, ainsi que du risque d'instabilités artificielles, il est logique de limiter la dérivation des EDP aux premier et deuxième ordres. Ainsi, il serait hautement souhaitable de développer des critères pour l'applicabilité de la méthode aux approximations d'ordre inférieur. Deuxièmement, les résultats analytiques ont été dérivés pour des systèmes linéaires spatialement invariants. En réalité, cependant, la méthode est surtout appliquée à des systèmes non linéaires dépendant de l'espace, il est donc intéressant d'étudier les garanties de convergence pour de tels systèmes. Troisièmement, les Chapitres 5 et 6 ont montré le potentiel de l'application de la méthode de continuation à la conception de régulateurs utilisant une représentation continue du système. Une telle procédure nécessite non seulement l'application de la méthode de continuation pour dériver un modèle d'EDP du système, mais aussi la discrétisation de la loi de contrôle obtenue afin de pouvoir l'implémenter dans le système réel. Dans le futur, il serait souhaitable de trouver quelles conditions la continuation doit satisfaire pour que le régulateur soit capable d'accomplir cette tâche.

Acknowledgement

First of all, I would like to say that all the three years of my PhD were wonderful. I experienced many unforgettable moments and met many amazing people. The scene for all of this was the encircled by mountains beautiful city of Grenoble, where for three years I witnessed the most wonderful nature of my life.

This thesis would not be possible without the efforts of my supervisors, Carlos Canudas de Wit and Paolo Frasca. It was they who showed me the direction and taught me to follow it, and I thank them for all the help and lessons I learned from them.

I am very grateful to the jury members Emmanuel Witrant, Mario di Bernardo, Bassam Bamieh, Miroslav Krsic and Brigitte d'Andrea-Novel for their valuable comments and suggestions. It was their feedback and comments that gave this thesis its final form. I would also like to thank Ursula Ebels, who invited me to work on the topic of spin-torque oscillators, which resulted in Chapter 6 of this thesis.

The students from our team became my friends. Thank you Martin, Umar, Ujjwal, Stephane, Nicolas, Vadim and Leo for your friendship, your help and support! We have had many pleasant and even surprising moments together. I also want to say thanks to the other wonderful members of our team, Maria-Laura, Federica and Alain, who were always there to advise and to discuss any issue.

Finally, all this would have been impossible without my wife Liudmila, who was always supportive, helpful, creative, and shared both the work and the home with me.

Contents

List of acronyms

DTFT Discrete-Time Fourier Transform

LQR Linear-Quadratic Regulator

LWR Lighthill-Whitham-Richards traffic model

ODE Ordinary Differential Equation

PDE Partial Differential Equation

SPR Strictly Positive Real transfer funciton

STO Spin-Torque Oscillator

List of symbols

$\mathbb{N}$	Set of natural numbers.
$\mathbb{Z}$	Set of integer numbers.
$\mathbb{Z}^+$	Set of non-negative integer numbers.
$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^+$	Set of non-negative real numbers.
$\mathbb{C}$	Set of complex numbers.
$\mathbb{C}^+$	Set of complex numbers with non-negative real parts.
$\mathbb{S}$	Unit circle.
I	Identity matrix.
$\mathbf{1}$	Vector of ones.
$\|x\|$	Euclidean norm of a vector x.
$\|f(\cdot)\|$	L_2-norm of a function $f(x)$.
$\|a\|_{l_2}$	l_2-norm of a sequence $\{a_i\}_{i\in\mathbb{Z}}$.
$\lambda(A)$	Set of all eigenvalues of a matrix A.
$\lambda_{min}(A)$	Smallest eigenvalue of a matrix A.
$\lambda_{max}(A)$	Largest eigenvalue of a matrix A.
$\rho(A)$	Spectral radius of a matrix A.
$U(a,b)$	Uniform probability distribution on interval $[a,b] \in \mathbb{R}$.
$\mathcal{F}\{f\}(\omega)$	Fourier transform of a function $f(x)$.
$\mathcal{D}\{a\}(\omega)$	Discrete-time Fourier transform (DTFT) of a sequence $\{a_i\}_{i\in\mathbb{Z}}$.
∇	Row-vector of partial derivatives with respect to spatial coordinates.
e_j	Unit basis vector in $\mathbb{R}^n$ in the direction $j \in \{1,...,n\}$.

Chapter 1

Introduction

Contents

1.1 Large-scale systems modelling

We live in a complex world in which all processes are interconnected. The early scientific approach in physics, engineering and control theory was to isolate individual subsystems, simplify models as much as possible, and explore them in their purest form, but in today's world we have to deal with complex systems that cannot be viewed as the independent sum of their constituent subsystems. Man-made structures such as city traffic, social networks, power networks or laser arrays can have thousands or even millions of degrees of freedom. To study systems of this size and for practical applications it is necessary to develop good models and ways to analyze them.

1.1.1 Networks modelling

There are four fundamental interactions in physics, and they all work in such a way that objects interact with each other in pairs. In other words, the force acting on a particle can always be defined as a sum of independent forces acting from all other particles. It turns out, and perhaps this is not a coincidence, that in practical applications the same structure is often traced — subsystems interact with each other in pairs, and the total external influence

on any subsystem is the sum of all influences from other subsystems. Such systems, which are a collection of pairwise interacting subsystems, are described by the structure of networks. Throughout this work we will call the interacting subsystems as nodes or agents.

Using the language of ordinary differential equations (ODEs) it is possible to write down the evolution of any large network as an evolution of nodes' states in time, taking into account pairwise interactions. At the same time the number of states can still be enormous; it becomes necessary to develop methods for scalable analysis of large ODE systems. Like other ODEs, models describing networks can be linear or nonlinear.

Linear networks are essentially very large linear systems. Their distinctive feature is that the system matrices of linear networks have a clear structure, because an element of the matrix will have zero if the two corresponding nodes are not directly connected to each other. This property is investigated by the theory of structural controllability, see Lin 1974; Dion, Commault, and Woude 2003; Leitold, Vathy-Fogarassy, and Abonyi 2017. In addition, in many systems the influence of other nodes on a particular node is positive, and then such systems belong to the class of positive systems. Finally, linear networks describing real-world systems are often stable. One particular and important example is opinion dynamics and consensus networks, where the states of all nodes converge to a mean value, see Tanner 2004; Mirtabatabaei and Bullo 2012.

The analysis of nonlinear networks is much more difficult. This problem is now at the forefront of science in the theory of nonlinear dynamics. One of the most important issues in this problem is the investigation of various modes of possible network functioning. Besides stable and unstable equilibrium states, there are many other modes, such as synchronous oscillations, traveling waves, chaotic behavior and even chimera states. Therefore in the study of nonlinear networks the analysis of general patterns, rather than the specific trajectories of specific nodes, comes to the fore.

1.1.2 Partial differential equations

The problem of describing systems with a large number of interacting particles first appeared quite a long time ago. Probably the first famous equations to describe a large system were the Euler equations derived by Euler 1761 describing the behavior of fluids and gases. This model is written in the language of partial differential equations (PDEs) and predicts the evolution of continuous fields of density and velocity of a fluid over time as a function of the partial derivatives of these quantities, as well as pressure, with respect to the position. Since then, many PDEs have been created that are models for describing different physical processes, such as the Maxwell equations for electromagnetic field by Maxwell 1873. We can distinguish three prototypical linear partial differential equations that describe physical processes:

- **Transport equation:**

$$\frac{\partial \rho(t,x)}{\partial t} = -v\frac{\partial \rho(t,x)}{\partial x}.$$

This equation describes pure transportation of some quantity $\rho(t, x)$ through the space with velocity v. In particular, shape of the solution does not change with time and can be uniquely identified from initial conditions, $\rho(t, x) = \rho(0, x - vt)$.

- **Heat equation:**
$$\frac{\partial \rho(t,x)}{\partial t} = -\alpha \frac{\partial^2 \rho(t,x)}{\partial x^2}.$$
Here $\rho(x, t)$ is understood as the temperature at a particular point in space at a particular time, and α is the heat transfer coefficient. In the absence of external heat sources, the heat conduction equation tends to locally "average" the solution, so after some time the solution converges to thermal equilibrium.

- **Wave equation:**
$$\frac{\partial^2 \rho(t,x)}{\partial t^2} = \beta^2 \frac{\partial^2 \rho(t,x)}{\partial x^2}.$$
This model describes the vibrations of a stretched string. In this equation, $\rho(x, t)$ is the displacement of the string relative to its equilibrium position, and β is the rate of propagation of mechanical disturbances along the string. Since the wave equation contains a second derivative with respect to time, it gives the acceleration as a function of the displacement along the position, and hence the solution can be written in terms of harmonic functions in space and time.

Many other partial differential equations for real physical systems can be seen as more advanced nonlinear versions of these three equations. Interestingly, the first two of these equations belong also to the class of conservation laws models: quantity in any domain can be changed only by flows through boundaries of this domain.

It is important to note that fluids and gases, whose behavior is modeled by Euler's equations, still have an underlying structure defined in terms of particles (molecules). Nevertheless, trying to describe such a system using a set of ODEs is unimaginably difficult: for example, even a small room contains about 10^{27} of air molecules, each of which in turn has 6 positional degrees of freedom (3 for coordinate and 3 for velocity), not to mention rotational dynamics. That is why PDEs are the only reasonable way to describe such a system. The same logic is true for most other physical systems that are historically described by partial differential equations. In the modern world of large systems, such a representation is also beginning to make sense for describing large man-made structures. For example, in the 1950s Lighthill and Whitham 1955 and Richards 1956 developed the LWR model for highway traffic. Instead of independently writing equations for each of the thousands of cars, they assumed that the cars could be treated as small particles like a fluid, and that the density of the cars at each point could be determined. The LWR model is a nonlinear PDE conservation law describing the evolution of this density at each point depending on how many machines are around. With this formulation, the whole system has been reduced to just one equation, and thus traffic analysis has been greatly simplified.

Nevertheless, mathematical analysis of PDEs is very complicated, since the states of the system at each moment of time are functions, which requires sophisticated tools such as func-

tional analysis to study the systems. Still, in many cases this approach is more realistic than solving the original system of a huge number of equations describing the behavior of independent particles. A possible solution to the difficulty of working with PDEs is to discretize them, turning PDEs back into a set of ODEs for fixed points in space called cells. In this case, the designer has a possibility to adjust the choice of the number of cells in order to balance the accuracy of the model with the complexity of working with it.

1.2 Large-scale systems control

Control of large-scale dynamical systems is a challenging problem for modern control theory. Its difficulty originates from the large dimensionality of these real-world systems, where the number of states can reach millions. These large systems challenge the scalability of control methods from several points of view. First, the computation of traditional control algorithms becomes too expensive. Indeed, imagining a large-scale linear network with n nodes, even an eigenvalue stability check would require $O(n^3)$ operations, while algorithms like linear-quadratic regulator (LQR) are substantially more computationally demanding. Second, the structure and the detailed dynamics of a system may not be fully known. Third, the number of actuators and sensors is often much lower than the number of nodes, so that state feedback is not possible: see for instance biological neural networks, where only an average neuronal activity is measured by electrodes.

The paradigms for controlling large-scale systems can be divided into three groups:

- **Centralized control** Traditional control theory assumes the existence of a single decisive device, the controller, which measures the state of the system and produces control commands that are applied to the whole system. This situation can be imagined in large systems as well. For example, city traffic lights can be controlled from a single control center that analyzes traffic conditions in the city. However, due to the problems listed above, such a scheme is not always applicable.

  Another difficulty is that the amount of energy needed to control all elements of the system can grow unbounded as the state-space increases: in case of networks the growth is actually exponential for some network structures, see Yan et al. 2012; Liu, Slotine, and Barabási 2011; Cowan et al. 2012. A possible solution to this problem may be that the centralized controller does not have to worry about the state of each specific element, instead solving some more general control problem and thus minimizing the control energy. By limiting ourselves to controlling the mean state and standard deviation, we can control the system "on average", to which Chapter 2 of this work is devoted.

- **Boundary control** There is a particular type of centralized control that makes special sense in real physical systems, namely boundary control. In this case a physical system evolves in its domain, and control can be applied only from boundaries (Belishev 2007; Coron, d'Andrea-Novel, and Bastin 2007). This situation is very typical and realistic for physical systems, examples include controlling a road section where you can only

limit the flow of cars at the entrance or exit, but not inside the road segment itself, or controlling the temperature in a room which is done with a heater located by the wall, but which has no direct effect on the temperature at any point in the middle of the room (Krstic and Smyshlyaev 2008). In a sense, this type of control is "weak". This is easy to understand by imagining a system evolving on a continuous bounded manifold and noticing that the dimension of the manifold's boundary is less by one than the dimension of the interior of this manifold. Due to this problem even questions of controllability become non-trivial (Lasiecka and Triggiani 1989). Nevertheless, it is still often possible to achieve control goals by controlling systems through boundaries. For example, Tumash, Canudas-de-Wit, and Delle Monache 2021a showed that the state of traffic on a road can be driven to the desired time-dependent state by restricting the flow on the boundaries, and Prieur, Winkin, and Bastin 2008 developed a robust boundary control scheme for fluid networks.

- **Decentralized control** One approach to the problem of complexity of controlling large systems is to abandon the idea that the system is controlled by a single global controller and imagine instead that many smaller controllers are used to control only parts of the system. The advantage of such controllers is that they are assumed to have only local information about the system available to them. In the limit it is possible to imagine controllers independently controlling the state of each agent and measuring only the state of that agent and its neighbors. In decentralized control, each controller turns out to be simple, so this approach is well applicable to real-world problems where each agent can implement its own control but has limited computational resources and only local ability to measure the state of other agents. Examples of this approach include platooning control as in Jovanovic and Bamieh 2005; Barooah, Mehta, and Hespanha 2009; Bamieh et al. 2012. However, in most cases showing that decentralized controllers fulfill the global control goal requires analysis of the system as a whole.

1.3 Approaches to model simplification

Instead of developing sophisticated control algorithms directly for large systems, a fundamentally different approach to the problem of controlling such systems is model simplification. In this paradigm, the model of a complex system is replaced by a model of a smaller size and/or a simpler structure of interactions. Such a process possibly leads to a loss of information about the dynamics of the original system, but the representation of the system in a simpler form gives a possibility of applying standard control algorithms. Thus, the complexity shifts from the issue of control design to the process of model simplification itself. Several specific approaches to simplifying models of large systems can be distinguished:

- **Balanced realizations for linear systems.** In the middle of XX century Kalman 1965 analysed irreducible realizations (or minimal realizations), linear systems with very clear structure and minimal number of state variables preserving the original behaviour. Much later, Moore 1981 gave an algorithm of transforming every linear system into its

minimal realization, using the idea similar to principal component analysis (PCA) from statistics. This algorithm became known as balanced realization. Moore's idea was not only to transform the system into its minimal realization, but to do the similar thing as PCA does in statistics: to identify the most important directions in the state space and get rid of the others, thus obtaining some "reduced model". For this he diagonalized infinity-time Gramians. To extend this approach for unstable systems as well, in Zhou, Salomon, and Wu 1999 a method for creating balanced realization (and thus model reduction) was presented using Gramian defined in frequency space.

Gramian-based model reduction techniques are not very suitable for networks, because the reduced system loses network properties like sparsity and tractability. So, current research is more focused on other reduction methods which are trying to preserve some physical properties instead of copying the original system behaviour. But there are modern papers like Rossi and Frasca 2018 devoted to network reduction which preserves network structure using generalized Gramians, defined for semi-stable systems with zero eigenvalues, for example as for Laplacian dynamics systems.

- **Clustering.** Perhaps the most popular method of simplifying models for ODE networks is clustering, as in Cheng, Kawano, and Scherpen 2017; Martin, Frasca, and Canudas-de-Wit 2019; Niazi et al. 2019. The idea of this method is to combine several connected network nodes and represent them as one "large" node. In this case, the new node is connected to the same nodes with which its constituent nodes were connected, and is subject to the same dynamic equations. The logic behind this simplification is that possibly nearby nodes have similar dynamics, and due to the connection between them, they interact with the rest of the nodes in the same way. Different approaches and versions of this method use different criteria to determine the "similarity" of nodes to cluster them into a single node.

- **Aggregated parameters and moments.** Simplification of the model by clustering implies that the system retains the same physical meaning but reduces the number of considered nodes. An alternative to this approach may be to represent the behavior of the system through completely different parameters with a different physical meaning, which nevertheless allow to preserve the dynamics of the system and describe it in a more compact form. Some of the most popular aggregated parameters are moments: the average state of the system, the standard deviation, and so on. For most systems it is easy to define moments, but building a general principle of controlling a system exclusively through moments requires solving the problem of moments closure, since in general the dynamics of each moment depends on the next one and this series continues to infinity, see the review by Kuehn 2016 devoted to this problem. Moment closure can be performed explicitly only for some systems, e.g. for crowd control as in Yang, Dimarogonas, and Hu 2015.

- **Particular class of solutions.** Many models are difficult to study, as they describe all the solutions that can arise in the system. Sometimes in the process of analysis it is possible to restrict the class of considered solutions, and thus simplify the system. Perhaps the most striking example of such a simplification is the derivation of Newton's

classical mechanics from Einstein's General Theory of Relativity. Indeed, General Relativity describes any behavior of systems with gravitational interactions. However, if we assume that we are only interested in solutions in which all speeds are much less than the speed of light and gravity remains small enough (does not create black holes), the General Relativity can be simplified to Newton's second law, supplemented by Newton's law of universal gravity. The idea that we can use a priori knowledge of a solution's belonging to a certain class forms the basis of Chapter 3 of this work, where the system is described in terms of the parameters of the solution shape.

- **Continuous approximations.** Another way to simplify a large system of interacting particles is to represent their dynamics through a continuous model. This is how Euler's equations describe the behavior of molecules in fluids and gases, and this is how the LWR equation describes the motion of cars on the highway. In spite of the fact that in terms of system dimensionality this method does not look like a simplification (we replace a very large but finite-dimensional system with an infinite-dimensional system), the resulting equations look much more homogeneous, that is their form does not depend on a specific point in space, whereas in the original system the equation of each particle could depend on the number and position of other particles interacting with it. The method of approximating the dynamics of systems through continuous models is the subject of Chapters 4, 5, and 6 of this manuscript.

Any simplification of a model is essentially an approximation, and there is always the question of how much this approximation guarantees the accuracy of the solutions, especially if the goal is to control the original system. Often such methods guarantee only that when the size of the simplified model tends toward the original, the solutions will also tend toward the original, that is, the control will work "for a sufficiently complex system". The question of when control will work for truly small systems remains open.

1.4 Problems and contributions

The main interest of our work is the problem of scalable control of large systems. Using different models and structures of large systems as starting points, we investigate different options on how systems can be simplified and used for control analysis and design, demonstrating results in various practical problems.

The first problem we tackle in Chapter 2 is controlling a large linear network. The control goal is to bring the average state of the network to a certain desired value, while keeping the states of all nodes as close to the average as possible. It is assumed that the average state of the network and the standard deviation of all states from the average can be measured. First, we study the problem of controlling only the average state using an integral controller and show that control can be performed using arbitrarily large gains if the system is positive and satisfies simple conditions on the matrices. Thus, a simple relationship has been established between positivity and passivity of linear systems. We then minimize the standard deviation

using the constrained extremum seeking algorithm. This approach is generalized for cases of general multidimensional linear output control and simultaneous minimization of general scalar quadratic output.

In Chapter 3 we turn our attention to systems whose models are given through partial differential equations. We develop a shape-based model reduction technique applicable to 1D PDE conservation laws. The main idea is that we assume that the solution of the system can be approximated by a solution of some specific shape that can be parametrized. And thus the PDE dynamics of the original system turns into a simplified ODE dynamics for the parameters of the solution shape.

Noticing that there are many similarities between systems modeled by ODE networks and PDEs, for the third problem we focus on a way of combining these two worlds together, in particular deriving continuous PDE representations of large-scale spatially distributed ODE systems. In Chapter 4 we develop a continuation method which transforms any general nonlinear system with spatial structure into a PDE model. In addition, we show that in the linear case, taking a sufficiently large PDE order, one can approximate a solution of the original ODE system as closely as desired. This method can be used to derive PDE models of originally discrete systems, which can then be used for control and analysis.

The problem of the connection between the discrete and continuous world has a long history, and in this context the Hilbert's problem 6 is very famous for raising the question of strict derivation of the Euler equations from the equations of motion of individual particles. We derive an original solution to this problem for the case of long-range potentials in Chapter 5 using the continuation method. The same technique makes it possible to derive models of traffic motion and even to control large swarms of robots. In addition, in Chapter 6, turning ODE networks into continuous models allows large networks of oscillators to be treated as nonlinear PDEs, which in turn opens up many possibilities for synchronization analysis and for stabilization. Various systems such as Kuramoto oscillators or non-isochronous spin-torque oscillators (STO) are being studied, and the conditions for their synchronization are derived from a continuous model of the original discrete system.

Thesis organization. Chapters 2, 3, and 4 introduce various transformation, simplification and control methods for large systems. Chapters 5 and 6 focus on the particular continuation method developed in Chapter 4 and show how it can be applied to analyze and control various multi-agent and coupled oscillator systems, respectively. Each chapter is preceded by an introduction, a relevant review of the literature and the state of the art, and concludes with a description of the contributions, open issues and possible research areas. All contributions and perspectives are summarized and discussed in the conclusion, followed by Appendix A, containing technical proofs of some of the lemmas used in the main body of this work.

1.5 Publications

1.5.1 Journals

- Denis Nikitin, Carlos Canudas-de-Wit, Paolo Frasca and Ursula Ebels. "Synchronization of Spin-Torque Oscillators via Continuation Method". Submitted to *IEEE Transactions on Automatic Control*. Preprint: https://hal.archives-ouvertes.fr/hal-03315718. Corresponds to Sections 6.4 and 6.3 of this thesis.
- Denis Nikitin, Carlos Canudas-de-Wit and Paolo Frasca. "A Continuation Method for Large-Scale Modeling and Control: from ODEs to PDE, a Round Trip". Conditionally accepted to *IEEE Transactions on Automatic Control*. Preprint: https://hal.archives-ouvertes.fr/hal-03140368. Corresponds to Chapter 4, Sections 5.3 and 5.4 of this thesis.
- Denis Nikitin, Carlos Canudas-de-Wit and Paolo Frasca. "Control of Average and Deviation in Large-Scale Linear Networks". Accepted to *IEEE Transactions on Automatic Control*, April 2022. Accessible at: https://hal.archives-ouvertes.fr/hal-03170606. Corresponds to Chapter 2 of this thesis except for Section 2.3.

1.5.2 International conferences

- Denis Nikitin, Carlos Canudas-de-Wit and Paolo Frasca. "Boundary Control for Stabilization of Large-Scale Networks through the Continuation Method". Accepted to *60th IEEE Conference on Decision and Control (CDC)*, Austin, USA, December 2021. Preprint: https://hal.archives-ouvertes.fr/hal-03211021. Corresponds to Sections 6.2 and 4.5 of this thesis.
- Denis Nikitin, Carlos Canudas-de-Wit and Paolo Frasca. "Shape-Based Nonlinear Model Reduction for 1D Conservation Laws". *IFAC World Congress 2020*, Berlin, Germany, July 2020, pp. 5309-5314. Accessible at: https://hal.archives-ouvertes.fr/hal-02952161. Corresponds to Chapter 3 of this thesis.
- Denis Nikitin, Carlos Canudas-de-Wit and Paolo Frasca. "Boundary Control for Output Regulation in Scale-Free Positive Networks". *58th IEEE Conference on Decision and Control (CDC)*, Nice, France, December 2019, pp. 5050-5055. Accessible at: https://hal.archives-ouvertes.fr/hal-02335142. Corresponds to Sections 2.1 and 2.2 of this thesis.

1.5.3 Publications related to COVID-19

Due to the recent outbreak of COVID-19 pandemic I participated in a research related to the development of optimal testing policies helping to mitigate the infection spreading. This work resulted into several publications, however it is unrelated to my main direction of research and therefore is not included into this thesis.

- Muhammad Umar B. Niazi, Carlos Canudas-de-Wit, Alain Kibangou, Denis Nikitin, Liudmila Tumash and Pierre-Alexandre Bliman. "Testing Policies for Epidemic Control". Accepted to *60th IEEE Conference on Decision and Control (CDC)*, Austin, USA, December 2021. Preprint: https://hal.archives-ouvertes.fr/hal-03185142
- Muhammad Umar B. Niazi, Carlos Canudas-de-Wit, Alain Kibangou, Denis Nikitin, Liudmila Tumash and Pierre-Alexandre Bliman. "Modeling and Control of COVID-19 Epidemic through Testing Policies". Submitted to *Annual Reviews in Control*. Preprint: https://hal.archives-ouvertes.fr/hal-02986566

Chapter 2

Control of average and deviation in linear networks

Contents

2.1 Introduction

Control community has often approached the issue of network control by looking for distributed control algorithms, in which the control is applied locally at all nodes and uses only local information. Instead, in this chapter we choose to work in a centralized setting, where an external operator has limited information about the network and limited access to few nodes for sensing or actuation purposes. In view of these limitations, the operator shall aim at

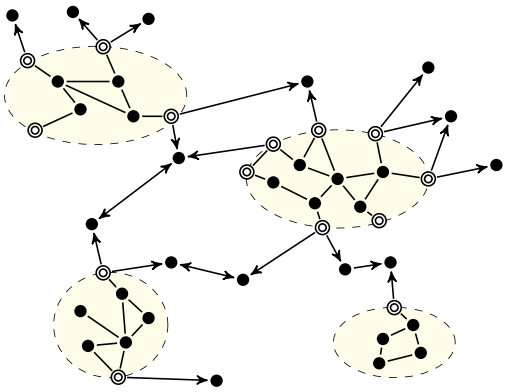


Figure 2.1: Scale-Free network with "hub" regions (shaded in yellow) controlled from the boundary nodes (double circles)

controlling some aggregate function of the network state, rather than controlling all of its individual nodes. A natural choice for such aggregate function is the average of the node states, which has indeed been defined as a control objective in some prior work that was motivated by opinion dynamics in social networks, see Vassio et al. 2014; Rossi and Frasca 2018. More broadly, the control of a generic output of a large-scale network was studied by Klickstein, Shirin, and Sorrentino 2017a; Wittmuess, Heidingsfeld, and Sawodny 2016; Klickstein, Shirin, and Sorrentino 2017b; Commault, Woude, and Frasca 2019; Casadei, Canudas-de-Wit, and Zampieri 2018. In particular it was shown that the energy required to control aggregated outputs instead of all states is much less.

In this chapter we focus on the problem of controlling the average state of the network, together with the concurrent minimization of its deviation. The average state of the network is defined as an average over all node states of the network, while the squared deviation is defined as an average over all squared differences between node states and the average state. While using a controller for the average state, it is natural to desire the system states to be close to the average: this behavior can be obtained by minimizing their squared deviation. In opposition to previous work, it is assumed that *the only values that are measured and regulated are the values of the system outputs, i.e. the average state and the squared deviation.* Moreover, we make sure that *the system model is not used in the controller.* Thus, the controller directly utilizes only system outputs and reference point and the equilibrium of internal states is never computed explicitly. Not using the system model circumvents all issues about computational complexity and uncertainties that affect large networks. Another relevant setup is a scale-free control approach to large-scale networks (see the Scale-FreeBack project of Canudas-de-Wit 2015) as in Fig. 2.1: in this approach, the goal is to control the average state and the deviation of the "hub" regions (such as large cities in a multi-city transportation network) and the control is applied to the boundaries of the hubs. Thus several linear outputs should be controlled independently which poses a problem of multi-output control.

Section 2.2 of this chapter is devoted to the problem of controlling the average state of the

linear network, where three stability results are presented. Theorem 2.1 provides conditions on the integral controller gains for the stability of the closed-loop system, and Theorem 2.2 simplifies the conditions under the assumption that the system is positive. Then, Theorem 2.3 is a main contribution of this section which gives a simple sufficient sign condition on the system matrices which guarantees stability of *any* positive integral controller for controlling the system output to a constant reference point without knowledge of the system matrices.

Most of our results regarding the output regulation problem of a large linear network system are presented under the assumption that the system is stable and positive (that is, the system matrix has positive elements outside the main diagonal). Network systems with stable dynamics and positive edge weights belong to this class. More generally, positive systems are an important class of systems for which the synthesis of large-scale control algorithms can be greatly simplified. Their impulse response is bounded by their static gain (Rantzer 2011), optimal (Rantzer 2015) and robust (Briat 2013) feedback control laws can be easily designed using linear programming, and the state feedback output regulation problem can be explicitly solved (Nogueira 2013). From the passivity analysis in the classical control theory it is known that the feedback interconnection between a linear operator with an integral controller is stable irrespective of the gain (has an infinite gain margin) if the linear operator is strictly positive real (SPR), see Sepulchre, Jankovic, and Kokotovic 2012; Kottenstette et al. 2014. From this point of view our analysis provides a new simple sufficient condition for the positive system to be SPR, which is summarized in Theorem 2.4.

In Section 2.3 this result is extended to include multi-output systems such that average states of different parts of a system could be regulated independently. Using the passivity formulation, we prove Theorem 2.5 which provides sufficient conditions on system matrices to assure that a MIMO system has a SPR transfer function.

Finally, Section 2.4 focuses on the deviation minimization problem, when the system should be driven to the particular average state while the control inputs should be balanced in such a way that the squared deviation of the states takes the smallest possible value. To solve this problem we use an extremum seeking scheme as in Ariyur and Krstić 2003; Tan et al. 2010 which is an adaptive model-free algorithm for the minimization/maximization of a nonlinear steady-state output characteristic. We augment this algorithm with an additional subsystem such that both tasks are accomplished simultaneously: the average is driven to the particular value while the squared deviation is minimized. Theorem 2.6 proves that the system approaches any small neighbourhood of the optimum state provided the gains of the controller are small enough.

2.1.1 Examples of physical systems

In our problem formulation we assume that the network operator has knowledge of the average and the squared deviation values. There are many physical examples of systems where the average and the squared deviation can be measured without measuring the states of the nodes. Here we briefly mention four examples:

Urban traffic networks. Consider a network of roads in a city, where the state of each node is the number of cars on the corresponding road section. The total number of cars in the city can be estimated either directly or indirectly. A direct estimation of the number of cars can be performed by vision-based methods, by processing images taken from satellites, as in Eslami and Faez 2010; Palubinskas, Kurz, and Reinartz 2010. Although every car is counted independently, the estimation error is defined as a discrepancy in the overall number of cars, therefore these methods effectively reconstruct the total number of cars. An indirect estimation can be based on the vehicle emissions: combustion engines produce CO_2, which then goes into the atmosphere. The polluted atmosphere changes its reflection properties based on the amount of CO_2, thus this amount can be measured using infrared sensors mounted on satellites, see Boynard et al. 2014. Therefore, the number of cars can be reconstructed from the satellite measurements. The total number of cars divided by the number of road sections equals to the average state of the network.

Biological neural networks. A widely known method of monitoring the brain activity is the *electroencephalography*, with electrodes placed usually along the sculp of the person being monitored. Each electrode measures voltage fluctuations of group of neurons under the surface, therefore it is directly related to the average of individual states of neurons, which obviously cannot be measured independently (Van Veen et al. 1997).

Dynamics of gas. Every gas consists of a huge number of particles colliding with each other, therefore it can be seen as a dynamical network with neighbouring particles whose interaction depends on their velocities. Thus, we can define the states being the velocities of each individual particle. The gas temperature can be easily measured, but at the same time it corresponds to the internal kinetic energy: $E_k = \frac{3}{2}k_BT = \frac{1}{2}mv_{rms}^2$. Here k_B is the Boltzmann constant, T is the temperature and m is the mass of one particle. The variable v_{rms}^2 represents the mean squared deviation of the velocities of particles with respect to the flow velocity. The flow velocity itself is a "wind speed", which represents the average state of the system and can be also directly measured.

Density of a fluid. Fluids also consist of a huge number of particles, and one way to write the dynamical model of a fluid is to consider a space partitioned into individual cells with states being defined as the densities of the fluid inside each cell. In this case, the average state would be the average density in the system: Hunt et al. 2021 showed that this density can indeed be measured for cryogenic fluids by measuring permittivity by a technique called electrical capacitance tomography.

2.1.2 Problem formulation

We start posing the problem assuming the system we need to control is the network given by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges. The number of vertices $|\mathcal{V}|$ is denoted by n.

On each node $v_i \in \mathcal{V}$ the state x_i is defined. Each edge $e \in \mathcal{E}$, where $e = \{v_i, v_j\}$, corresponds to the flow between nodes v_i and v_j. Matrix $A \in \mathbb{R}^{n\times n}$ represents flow ratio. The

set of nodes $\mathcal{V}$ is split into two parts $\mathcal{V}_1$ and $\mathcal{V}_2$ with state vectors x_1 and x_2 respectively (see Fig. 2.2). The set $\mathcal{V}_1$ consists of the nodes which are directly controlled by the control action u. We call these nodes "boundary". The set $\mathcal{V}_2$ is a set of uncontrolled nodes, which we call "inner nodes". The average state $y = \mathbf{1}^T x/n$ and the squared deviation $V = \|x\|^2/n - y^2$ are measured. Thus the network depicted on Fig. 2.2 can be viewed as one particular hub from the scale-free network on Fig. 2.1.

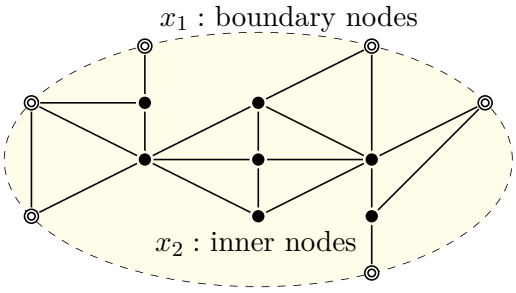


Figure 2.2: Network with boundary and inner nodes separation

The evolution of the states x, the average state y and the squared deviation V is given by the following linear time-invariant system

$$\begin{cases} \dot{x}_1 = A_{11}x_1 + A_{12}x_2 + u, \\ \dot{x}_2 = A_{21}x_1 + A_{22}x_2, \\ y = \dfrac{1}{n}\mathbf{1}^T x, \\ V = \|x\|^2/n - y^2. \end{cases} \tag{2.1}$$

Most of real-world networks are internally stable, so we further assume A being stable. Also in most of our analysis we will assume A is a Metzler matrix (defined as a matrix with its off-diagonal elements being non-negative), which means all edges have positive weights. Such choice of system matrix together with the fact that $B > 0$ and $C > 0$ means that the system (2.1) belongs to the class of positive systems.

It is useful to analyse more general case than (2.1), with general stable matrix $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times k}$, and symmetric positive semi-definite matrix $P \in \mathbb{R}^{n\times n}$ defining the quadratic output:

$$\begin{cases} \dot{x} = Ax + Bu, \\ y = Cx, \\ V = x^T P x. \end{cases} \tag{2.2}$$

System (2.1) can be written in form of (2.2) using

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} I \\ 0 \end{pmatrix},$$
$$C = \frac{1}{n}\mathbf{1}^T, \qquad P = \frac{1}{n}I - \frac{1}{n^2}\mathbf{1}\mathbf{1}^T$$

In general, the control goal is to stabilize average state y over the whole network to some desired constant state y_d without the explicit knowledge of system matrices. It is assumed that the number of states is too large that it is impossible to use full-state feedback or to use matrix A explicitly.

In the following two problems will be addressed:

1. **Control of average**: find the control law $u = u(y)$ for the system (2.2) such that
$$\lim_{t\to\infty} y(t) = y_d.$$
2. **Control of average and deviation**: Find the control law $u = u(y, V)$ for the system (2.2) such that
$$\lim_{t\to\infty} y(t) = y_d, \quad \text{and } V \text{ is minimized.}^1$$

For notational simplicity and expressiveness we will often use the word "average" when speaking about linear outputs and the word "deviation" for a general scalar quadratic output.

It appears that the problem of linear output control leads to different types of conditions for the case when the output is scalar with respect to the multiple output case. Therefore Section 2.2 is devoted to the first problem of average control, treating scalar output y. Section 2.3 presents an extension of the first problem for multidimensional output y. Finally, Section 2.4 shows the solution for the second problem of simultaneous control of linear and quadratic outputs.

Notation. Along this chapter several types of vector and matrix inequalities are used:

- $x \geqslant 0$ for $x \in \mathbb{R}^n$ means $x_i \geqslant 0\ \forall i \in \{1, ..., n\}$.
- $x > 0$ for $x \in \mathbb{R}^n$ means $x_i \geqslant 0\ \forall i \in \{1, ..., n\}$ and there exists $j \in \{1, ..., n\} : x_j > 0$.
- $x \gg 0$ for $x \in \mathbb{R}^n$ means $x_i > 0\ \forall i \in \{1, ..., n\}$.
- $P \succ 0$ for $P \in \mathbb{R}^{n\times n}$ means that $P = P^T$ and $x^T P x > 0\ \forall x \in \mathbb{R}^n : x \neq 0$.

2.2 Control of average

In this section, we solve Problem 1 of controlling the average as a scalar linear output of system (2.2). We first describe a general controller structure and then prove stability conditions for low and high gains. Further we discuss these conditions and present examples of the controller performance.

[1] Note that the problems are formulated in the steady-state, therefore we will not pursue any optimization of the transient process.

2.2.1 Controller structure

Define transfer function of the system (2.2)

$$W_s(s) = C(sI - A)^{-1}B, \tag{2.3}$$

thus $y(s) = W_s(s)u(s)$. Denote error between desired output and system output: $e = y_d - y$. Then we can define controller transfer function $W_c(s)$ such that $u(s) = W_c(s)e(s)$. System control loop is depicted on Fig. 2.3.

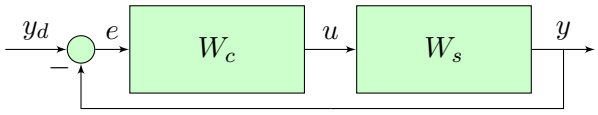


Figure 2.3: Control loop given by closed-loop transfer function (2.6)

Thus the input-output relation is

$$y(s) = W_s(s)W_c(s)e(s), \tag{2.4}$$

or, solving for y,

$$y(s) = \frac{W_s(s)W_c(s)}{1 + W_s(s)W_c(s)}\, y_d. \tag{2.5}$$

Define closed-loop transfer function

$$W(s) = \frac{W_s(s)W_c(s)}{1 + W_s(s)W_c(s)}. \tag{2.6}$$

In the following we investigate what properties W_s and W_c do have and what properties W should have in order to be stable. Values of W_s are row-vectors and values of W_c are column-vectors, because controller input e and system output y are scalars, while u which is controller output and system input is a vector, $u \in \mathbb{R}^k$. Let us look at the i-th component of W_s and W_c, where $i \in \{1, ..., k\}$, and define polynomials $\alpha(s), \beta_i(s), \delta(s), \gamma_i(s)$ such that

$$W_s(s)_i = \frac{\beta_i(s)}{\alpha(s)} \quad \text{and} \quad W_c(s)_i = \frac{\gamma_i(s)}{\delta(s)}. \tag{2.7}$$

It is obvious that $\alpha(s)$ is a polynomial of degree n. Moreover, our system is strictly stable, casual and have no direct influence of u on y, thus

$$\deg \beta_i(s) < \deg \alpha(s) = n \quad \text{and} \quad \alpha(s) \neq 0 \ \ \forall s \in \mathbb{C}^+. \tag{2.8}$$

We can choose $\alpha(s)$ and $\beta_i(s)$ such that $\alpha(s) \in \mathbb{R}$ for $s \in \mathbb{R}$ and $\alpha(s) > 0$ for $s \geqslant 0$.

Then, the controller W_c should also be stable and casual, which means

$$\deg \gamma_i(s) \leqslant \deg \delta(s) \quad \text{and} \quad \delta(s) \neq 0 \ \ \forall s \in \mathbb{C}^+. \tag{2.9}$$

Again, it is possible to choose $\delta(s)$ and $\gamma_i(s)$ such that $\delta(s) \in \mathbb{R}$ for $s \in \mathbb{R}$ and $\delta(s) > 0$ for $s > 0$. Now we can rewrite $W(s)$ in terms of polynomials:

$$W(s) = \frac{\sum\limits_i \beta_i(s)\gamma_i(s)}{\alpha(s)\delta(s) + \sum\limits_i \beta_i(s)\gamma_i(s)}. \tag{2.10}$$

The closed-loop transfer function $W(s)$ should have the following property: for a constant input y_d it should give the same output y, thus $W(0) = 1$. This means that $\alpha(0)\delta(0) = 0$, which is possible only if $\delta(0) = 0$, so $\delta(s)$ cannot contain free term. The simplest possible controller that satisfies this necessary condition is an integral controller given by

$$W_c(s)_i = \kappa\frac{\gamma_i}{s}, \tag{2.11}$$

where $\gamma \in \mathbb{R}^k$ is the vector of gains, defining relative control force applied to different actuated nodes (in other words in can be seen as a "control direction"), and κ is the overall gain. The following sections will be devoted to the integral controller and its properties.

2.2.2 Stability of integral controller

Assume we apply the integral controller (2.11) to the system (2.2). The closed-loop system may be unstable, and in general in order to prevent this one needs to carefully choose controller gains κ and γ in (2.11).

Theorem 2.1. *System (2.2) with applied integral controller (2.11) is asymptotically stable if* $-CA^{-1}B\gamma > 0$ *and* $\kappa \in (0, \kappa^*)$ *for some small* $\kappa^* \in \mathbb{R}$.

Proof of Theorem 2.1. Applying integral controller, a transfer function of the closed-loop system is given by

$$W(s) = \frac{\kappa\sum\limits_i \beta_i(s)\gamma_i}{\alpha(s)s + \kappa\sum\limits_i \beta_i(s)\gamma_i}. \tag{2.12}$$

For stability of the closed-loop system $W(s)$ should have no poles on the right-hand side of the complex plane $\mathbb{C}^+$. Decompose the denominator:

$$\alpha(s)s\left(1 + \kappa\frac{\sum\limits_i \beta_i(s)\gamma_i}{\alpha(s)s}\right) \neq 0.$$

Denote $Q(s) = \left(\sum\limits_i \beta_i(s)\gamma_i\right) / (\alpha(s)s)$. Any point such that $\alpha(s) = 0$ or $s = 0$ leads to $W(s) = 1$, thus poles of the transfer function can land only in positions of roots of $1 + \kappa Q(s)$. We will prove that there exists κ^* such that

$$\forall \kappa \in (0, \kappa^*) \,:\, \forall s \in \mathbb{C}^+ \setminus \{0\} \quad \mathrm{Re}\{1 + \kappa Q(s)\} > 0. \tag{2.13}$$

Choose $\varepsilon, R \in \mathbb{R}$ such that $\varepsilon < |\lambda_i(A)|$ and $R > |\lambda_i(A)|$ for all $i \in \{1..n\}$. Thus all the roots of $\alpha(s)$ lie in a ring between ε and R in the left half-plane. We split the complex right half-plane C^+ into three parts:

$$H^+_{0,\varepsilon} = \{s : \operatorname{Re} s \geqslant 0, |s| < \varepsilon\}$$
$$H^+_{\varepsilon,R} = \{s : \operatorname{Re} s \geqslant 0, |s| \geqslant \varepsilon, |s| \leqslant R\}$$
$$H^+_{R,\infty} = \{s : \operatorname{Re} s \geqslant 0, |s| > R\}$$

First we analyse $H^+_{0,\varepsilon}$. Function $Q(s)$ has a pole at zero, thus it can be written using Laurent series with coefficients Q_n:

$$Q(s) = \frac{Q_{-1}}{s} + \sum_{n=0}^{\infty} Q_n s^n = \frac{Q_{-1}}{s} + P(s),$$

where $P(s)$ is an analytic function. The residual $Q_{-1} = \left(\sum_i \beta_i(0)\gamma_i\right)/\alpha(0) = -CA^{-1}B\gamma > 0$, thus

$$\operatorname{Re} \frac{Q_{-1}}{s} \geqslant 0 \quad \forall s \in H^+_{0,\varepsilon} \setminus \{0\},$$

while $P(s)$ is analytic in $\mathbb{C}$ and thus has a minimum in $H^+_{0,\varepsilon}$.

Next we analyse $H^+_{R,\infty}$. This set is contained into a set $H_{R,\infty} = \{s : |s| > R\}$. If R is big enough, $Q(s)$ is analytic in $H_{R,\infty}$, but it vanishes at infinity, therefore by the maximum modulus principle $Q(s)$ is bounded from below in $H_{R,\infty}$ by values on its boundary, and consequently it is bounded in $H^+_{R,\infty}$.

Finally, set $H^+_{\varepsilon,R}$ is compact and does not contain zeros or roots of $\alpha(s)$. Therefore $Q(s)$ is analytic in it and thus bounded. We obtained that $\operatorname{Re} Q(s)$ is bounded from below in $\mathbb{C}^+ \setminus \{0\}$. Denoting this bound as Q_{inf}, we see that choosing $\kappa^* = -1/Q_{inf}$ in case $Q_{inf} < 0$ or $\kappa^* = +\infty$ in case $Q_{inf} \geqslant 0$ assures satisfaction of (2.13) and therefore proves the theorem. □

Theorem 2.2. *System (2.2) with applied integral controller (2.11) is asymptotically stable if the system (2.2) is positive, $\gamma > 0$, $-CA^{-1}B\gamma \neq 0$ and $\kappa \in (0, \kappa^*)$ for some small $\kappa^* \in \mathbb{R}$.*

Proof of Theorem 2.2. Positivity of the system (2.2) means that all elements of matrices B and C are greater or equal than zero, and matrix A is a Metzler matrix. Now we introduce a notion of M-matrix:

Definition 2.1 (M-matrix, Plemmons 1977)**.** An $n \times n$ matrix M that can be expressed in the form $M = \alpha I - L$, where $L = (l_{ij})$ with $l_{ij} \geqslant 0$, $1 \leqslant i, j \leqslant n$, and $\alpha \geqslant \rho(L)$ where $\rho(L)$ is the maximum of the moduli of the eigenvalues of L, is called an M-matrix.

From this definition it follows immediately that a negative of a Metzler stable matrix is an M-Matrix. The main property of any M-matrix M is that its inverse M^{-1} is a positive matrix, thus $(M^{-1})_{ij} \geqslant 0$ for all i, j (Fan 1958).

Matrix $-A$ is an M-Matrix which means that $-A^{-1}$ has its all elements nonnegative, therefore $-CA^{-1}B$ is a positive vector. By the theorem statement $\gamma > 0$, and having

$-CA^{-1}B\gamma \neq 0$ means $-CA^{-1}B\gamma > 0$ and, applying Theorem 2.1, this leads to an asymptotic stability of the closed-loop system for small enough κ. □

By Theorem 2.2, if the system is positive it is enough to choose $\gamma = \mathbf{1}$ (or any $\gamma \gg 0$, just to satisfy $-CA^{-1}B\gamma \neq 0$), and then pick up small enough overall gain κ.

2.2.3 Control with arbitrary large gains

It appears although that there exists simple criteria on the system matrices which says whether the closed-loop system will converge irrespectively of controller gain values, provided that they are positive. This result is one of the main contributions of this chapter, and it is formulated as follows:

Theorem 2.3. *System (2.2) with applied integral controller (2.11) is asymptotically stable for arbitrary large positive controller gains κ and γ if the system (2.2) is positive, $CA^2 > 0$ and $CA^2B\gamma > 0$.*

This result means that, irrespective of the gains, an integral controller will preserve stability for a very large class of systems. One of the important types of large-scale networks for which Theorem 2.3 is satisfied is a general consensus network (e.g. for social interactions), see the example below.

Example 2.1 (Damped consensus)**.** Assume system (2.2) is given by matrices $A = -L - \alpha I$, where L is a Laplacian matrix of some network with n nodes, $\alpha > 0$ means additional damping to the system to preserve stability, and $C = \mathbf{1}^T/n$ represents average state of the network. Then A is a Metzler stable matrix, and C is the eigenvector of A with corresponding eigenvalue $-\alpha$, thus $CA^2 = \alpha^2 C > 0$. Then any controller with positive gains κ and γ will lead to the convergence, provided $B\gamma > 0$.

One should notice that the condition $CA^2B\gamma > 0$ on the control matrix B is very non-restrictive, because by choosing appropriate vector gain γ it is always possible to make $B\gamma > 0$, and hence, provided $CA^2 > 0$ and $CA^2B \neq 0$, we will have $CA^2B\gamma > 0$. A reason for this is the fact that the regulation variable is a single scalar output.

Proposition 2.1. *Condition $CA^2B\gamma > 0$ is a sufficient condition for the output controllability of the system (2.2).*

Proof. Indeed, Kalman rank test for the output controllability of (2.2) can be written as

$$\operatorname{rank}\{C\begin{pmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{pmatrix}\} = 1,$$

and by $CA^2B\gamma > 0$ we have $CA^2B \neq 0$, which means that the rank test is satisfied. □

Note that the analogue of this Proposition can be proven for Theorems 2.1 and 2.2.

Corollary 2.1. *Positive system (2.2) with $CA^2 \gg 0$ is asymptotically stable for any integral controller (2.11) with positive gains applied to any single boundary node. Therefore, it is enough to control only one node.*

Before proving Theorem 2.3 we need to state three technical lemmas.

Lemma 2.1. *Suppose we have a matrix $\mathcal{M} = M + ibI$, which is a complex matrix with real part M and imaginary part bI, with $b \in \mathbb{R}$ and I an identity matrix. Assume M being invertible and having no eigenvalues on the imaginary axis. Denote $\mathcal{L} = \mathcal{M}^{-1} = L + i\bar{L}$. Then the real part of $\mathcal{L}$ is given by*

$$\operatorname{Re}\mathcal{L} = L = (M + b^2 M^{-1})^{-1}. \tag{2.14}$$

Proof. See Appendix A.1. □

Lemma 2.2. *Let M be an M-matrix. Let C be a row-vector such that $CM^2 > 0$. Then*

$$C(M + tM^{-1})^{-1} > 0 \tag{2.15}$$

for any $t \geqslant 0$.

Proof. See Appendix A.2. □

Lemma 2.3. *Let M be an M-matrix. Let C be a row-vector such that $CM^2 > 0$ and $CM^2B\gamma > 0$. Then*

$$C(M + tM^{-1})^{-1}B\gamma > 0 \tag{2.16}$$

for any $t \geqslant 0$.

Proof. See Appendix A.3. □

Proof of Theorem 2.3. Applying the integral controller and multiplying nominator and denominator by s, transfer function of the closed-loop system is given by

$$W(s) = \frac{\kappa C(sI - A)^{-1}B\gamma}{s + \kappa C(sI - A)^{-1}B\gamma}. \tag{2.17}$$

It is sufficient to show that real part of the denominator is strictly greater than zero in the right half-plane. Since $\operatorname{Re} s > 0$ in the right half-plane, it is enough to show that $\operatorname{Re}\{\kappa C(sI - A)^{-1}B\gamma\} > 0$.

Denote $\operatorname{Re} s = \alpha$ and $\operatorname{Im} s = \beta$, so matrix $(sI - A)^{-1} = ((\alpha I - A) + i\beta I)^{-1}$. Denote $M = \alpha I - A$. Matrix A is a Metzler stable matrix, thus $(-A)$ is an M-matrix and matrix M is an M-matrix too. Moreover, condition $CA^2 > 0$ implies $CM^2 > 0$ and $CA^2B\gamma > 0$ implies $CM^2B\gamma > 0$. Applying Lemma 2.1 we conclude that

$$\operatorname{Re}\kappa C(M + i\beta I)^{-1}B\gamma = \kappa C(M + \beta^2 M^{-1})^{-1}B\gamma. \tag{2.18}$$

By Lemma 2.3 $C(M + \beta^2 M^{-1})^{-1}B\gamma > 0$ for any $\beta \in \mathbb{R}$, and assuming $\kappa > 0$ we trivially obtain a sufficient condition on positivity of the real part of the denominator. □

2.2.4 Passivity formulation

Lemmas 2.1-2.3 allow us to formulate a more general theorem, applicable to a general SISO linear system:

Theorem 2.4. *Any positive stable SISO system with control matrix $B \in \mathbb{R}^{n\times 1}$ and observation matrix $C \in \mathbb{R}^{1\times n}$ such that $CA^2 > 0$ and $CA^2B > 0$ has a strictly positive real (SPR) transfer function, thus it is strictly-input passive.*

Proof. A *strictly positive real* (SPR) transfer function should by definition satisfy $\mathrm{Re}\{C(sI - A)^{-1}B\} > 0$ for $\mathrm{Re}\, S > 0$, see Sepulchre, Jankovic, and Kokotovic 2012; Kottenstette et al. 2014. Therefore the proof of this theorem follows the same steps as the second part of the proof of Theorem 2.3. □

From a point of view of linear systems theory, Theorem 2.4 is a main result of this chapter. Indeed, such a simple condition for passivity for positive SISO systems appears in literature for the first time.

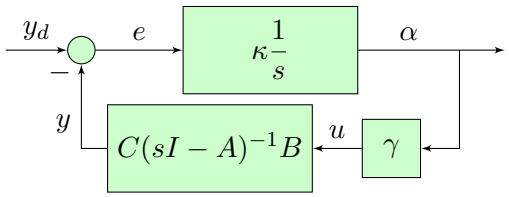


Figure 2.4: Feedback interconnection of passive systems

Moreover, it appears that Theorem 2.3 is a direct consequence of Theorem 2.4, as the following reasoning shows. Assume we fix an input gain vector γ and define a new controller output α such that $u = \gamma\alpha$. Then system (2.2) becomes SISO with respect to input variable α. Define $H_1(s) = \kappa/s$ and $H_2(s) = C(sI - A)^{-1}B\gamma$. It is possible to construct control loop with feedback interconnection as depicted on Fig. 2.4. The closed-loop system input is defined as y_d and the system output is α. It is known that for L_2 stability of a system with feedback interconnection it is sufficient that a transfer function of one of the blocks is *positive real* (PR, which is equivalent to passivity) and another is SPR, see Sepulchre, Jankovic, and Kokotovic 2012. Passivity of an integral controller $H_1(s)$ is obvious, and Theorem 2.4 is used to prove that $H_2(s)$ is SPR. Therefore the closed-loop system is L_2 stable. Now, it remains to prove that $y \to y_d$, which is obvious if one recalls that an output of a stable system with constant input converges to a constant value, thus for any constant y_d there exists α^* such that $\alpha \to \alpha^*$. But convergence of an output of an integral controller means that its input converges to zero, which reads as $e \to 0$, which is exactly $y \to y_d$.

2.2.5 Interpretation of conditions

Theorems 2.1-2.3 presented in the previous sections provide the same result, stability of the closed-loop system (2.2) with controller (2.11). However, they differ in their assumptions. During the derivation of the controller (2.11) we assumed that the system matrices are not known. However, usually one has a knowledge about some general properties of the system, such as positivity. These properties can sometimes be induced from the nature of the problem itself and do not rely on the particular topology. Therefore the results presented in our work can be used to analyse stability based on these properties.

Theorem 2.1 requires $-CA^{-1}B\gamma > 0$ for the integral controller to be stable for $\kappa \in (0, \kappa^*)$. This scalar condition essentially means that direction of adaptation of the integral controller forms an acute angle with a zero frequency gain of the system. In practice one usually knows direction of the zero frequency gain. At worst, it is enough to change a sign of γ once.

Theorem 2.2 exploits positivity of the system: zero frequency gain of the positive system is positive. Therefore it is enough to use positive gains for the integral controller, and the condition $-CA^{-1}B\gamma > 0$ can be loosened just to $-CA^{-1}B \neq 0$. However the gain κ still should satisfy $\kappa \in (0, \kappa^*)$.

In Theorem 2.3 a small-gain condition $\kappa \in (0, \kappa^*)$ is removed at the cost of adding a vector inequality $CA^2 > 0$. This inequality can be used to determine stability without knowing particular matrices for some classes of systems, such as damped consensus, see the example after Theorem 2.3. For other systems, this condition should be interpreted as a constraint on the system parameters.

In the remainder of this section we will analyse the condition $CA^2 > 0$ more closely. Namely, first of all we will prove that this condition cannot be relaxed, since weaker conditions would not assure the stability for all gains. Then, we will provide some graph-theoretical intuition and rewrite the condition in terms of quadratic constraint on node self-dampings.

If A is a Metzler stable matrix, all elements of A^{-1} are nonpositive. Multiplication of a positive vector by a matrix with nonpositive elements renders negative vector, therefore right multiplying the condition $CA^2 > 0$ by A^{-1} one obtains $CA < 0$, and the same argument provides $C > 0$. The condition $CA^2 > 0$ is new and it is used in Lemmas 2.2 and 2.3 (substituting $M = \alpha I - A$ as in the proof of the theorem). When one looks at the statement of Lemma 2.2, one might think that it would be enough to require a less restrictive condition $CA < 0$ (This condition can be obtained from the statement of Lemma 2.2 by letting $t \to +\infty$) and has been proposed for a full state static feedback output control of positive systems by Nogueira 2013).

However, let us show that condition $CA^2 > 0$ is significant and $CA < 0$ is not sufficient. An example of a positive system with $CA < 0$ but $CA^2 \not> 0$ would be

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad C = \mathbf{1}^T. \tag{2.19}$$

For this system $CA = (0, 0, -1)$, but $CA^2 = (0, -1, 1)$. We can then show that this system is not SPR. To check this, by definition we take a pole in the complex right half-plane $s = 0.01 + 2i$, which results in

$$\operatorname{Re} C(sI - A)^{-1}B = -0.0047. \tag{2.20}$$

Since the transfer function value is negative, the system is not positive real and thus it is not passive. Moreover, there exists an integral controller which makes this system unstable, for example one with a control vector $\gamma = 1$ (since only one node can be controlled) and a gain $\kappa = 3$ (although with $\kappa = 2$ the system is still stable). This confirms our understanding that the novel $CA^2 > 0$ condition is meant to ensure stability using any arbitrary boundary node and arbitrary positive gain κ.

Going deeper to understand topological properties of the condition $CA^2 > 0$, we first start with more intuitive one, $CA < 0$, which is implied by $CA^2 > 0$.

Define matrices D and E such that $A = E - D$, with D being diagonal and E having all diagonal elements zero. Thus both D and E have all their entries positive. Matrix E can be viewed as adjacency matrix of the network, with element E_{ij} meaning influence of node v_j on node v_i. Matrix D consists of self-damping powers on the diagonal. Therefore condition $CA < 0$ reads as $CD > CE$. This condition states some kind of diagonal dominance in the network.

Assume some $C_i = 0$. Then $(CD)_i = 0$ because D is diagonal. Thus $(CE)_i$ should be also zero, which means that for every index j either $C_j = 0$ or $E_{ji} = 0$.

Corollary 2.2. *If node v_i is not included in the aggregated output ($C_i = 0$), then its reachable set should not be included either.*

- *For a strongly connected graph this means that all nodes should be included in the aggregated output.*
- *If a network is divided into "boundary" nodes and "inner" nodes, and the goal is to control an average of the inner nodes, then at least one of the boundary nodes should also be included into the average.*

In the same manner it is possible to see this condition as a lower bound on the damping of each node: $D_{ii} \geqslant \sum_j C_j E_{ji}/C_i$. Thus the bigger is the influence of node's neighbours in the output, the bigger should be the node's damping.

We can use the same decomposition $A = E - D$ in order to understand the condition $CA^2 > 0$ and conclude that

$$CE^2 + CD^2 > C(ED + DE). \tag{2.21}$$

Being a quadratic inequality, this condition bounds damping of each node from above and below with respect to dampings of other nodes. We will see examples in the following section.

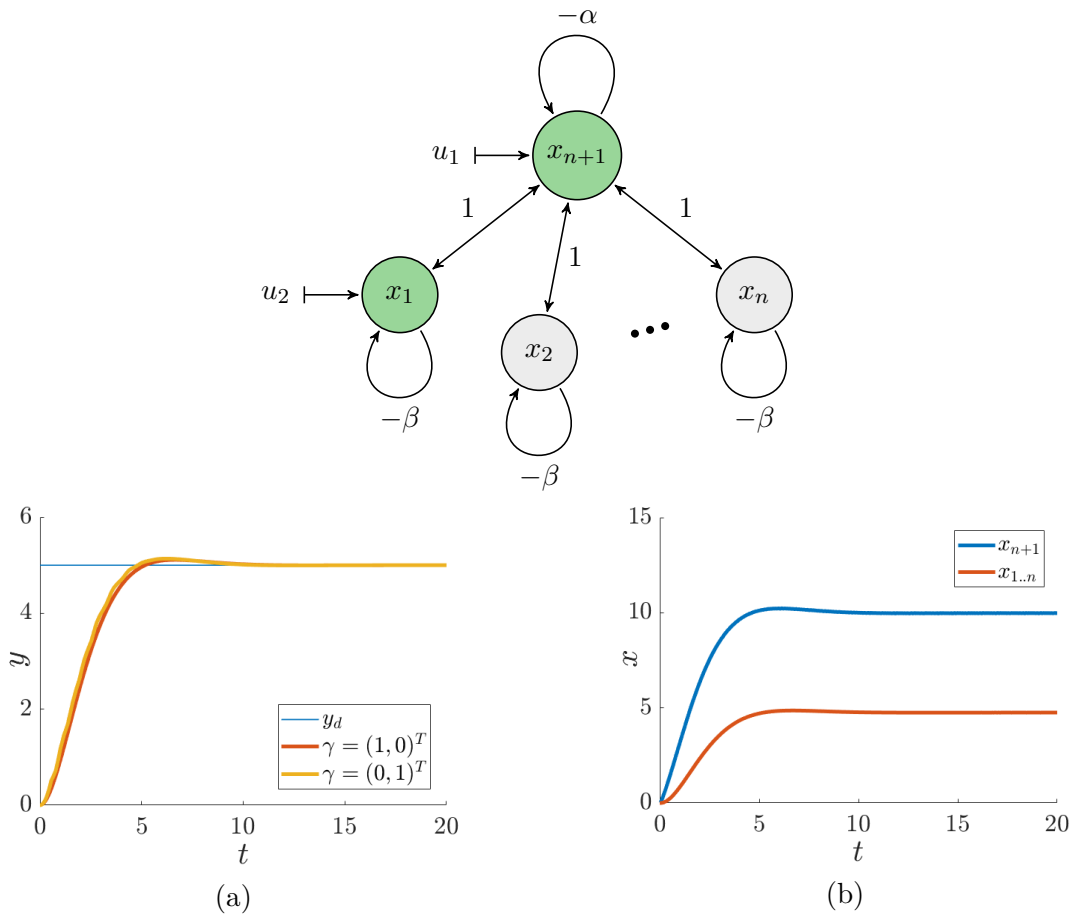


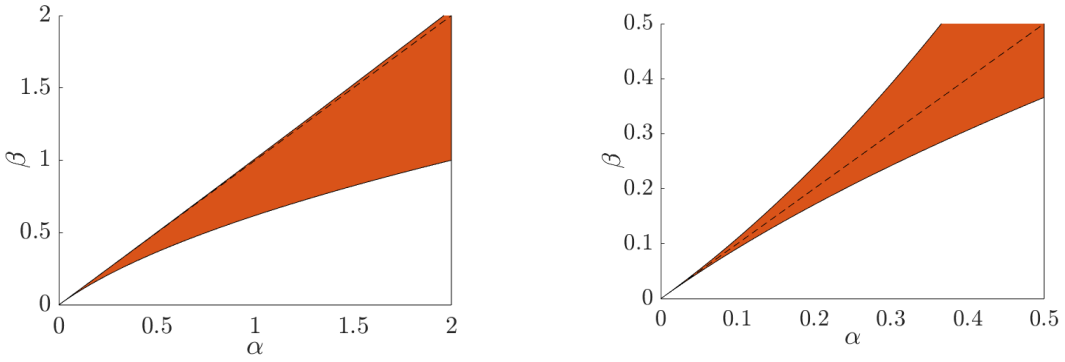
Figure 2.5: **Scheme:** Network with star topology with n leafs. Boundary nodes are in green. **Plots:** Output control of the star network with $n = 20$ leafs. $\alpha = 2$, $\beta = 1.1$, $\kappa = 12$, $y_d = 5$. **(a):** Output y for different γ vectors. **(b):** Spread of states x for $\gamma = (1, 0)^T$ corresponding to the control of the central node. All the leaf states $x_1 ... x_{20}$ have the same asymptotic value 4.751 (which is obvious from the symmetry), while the central state converges to 9.978.

2.2.6 Examples

Here we present three examples of networks, namely a star, a line and an Erdős-Rényi graph, and analyse the condition $CA^2 > 0$ for them.

2.2.6.1 Network with star topology

To begin with we choose network with star topology with one central node and n leafs, average state of which we want to control. Let nodes $1, ..., n$ be the leafs and node $n + 1$ be the center. Assume the center and the first leaf belong to the boundary node set and thus can be controlled (see Fig. 2.5).



(a) (b)

Figure 2.6: Regions in α-β space, **(a):** satisfying (2.24) as $n \to \infty$ for a star network, **(b):** satisfying (2.27) for a line network.

Dynamics of this network can be written as system (2.2) with matrices

$$A = \begin{pmatrix} -1-\beta & 0 & \cdots & 1 \\ 0 & -1-\beta & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & -n-\alpha \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}, \quad C = \frac{1}{n+1}\mathbf{1}^T. \tag{2.22}$$

Such choice of system matrices corresponds to the undirected network with star topology and damping $\alpha > 0$ for central node and $\beta > 0$ for all other nodes. The choice of B explores both cases of controlling leaf and center. It allows for maximum generality, moreover the controllability is guaranteed by Corollary 2.1. Integral controller (2.11) with $\gamma = (1,0)^T$ would correspond to a control applied only to the center, and controller with $\gamma = (0,1)^T$ would correspond to a control of the first leaf.

Calculating CA and CA^2 gives

$$\begin{aligned} CA &= \begin{pmatrix} -\beta & -\beta & \cdots & -\alpha \end{pmatrix} / (n+1) < 0, \\ (CA^2)_{1,\dots,n} &= \left(\beta^2 + (\beta - \alpha)\right) / (n+1), \\ (CA^2)_{n+1} &= \left(\alpha^2 + n(\alpha - \beta)\right) / (n+1). \end{aligned} \tag{2.23}$$

$CA^2 > 0$ means then $\alpha^2 + n(\alpha - \beta) \geqslant 0$ and $\beta^2 + (\beta - \alpha) \geqslant 0$ with at least one of these inequalities being strict. Solving this for damping of leaf nodes we obtain

$$\sqrt{\alpha + \frac{1}{4}} - \frac{1}{2} \leqslant \beta \leqslant \alpha + \frac{\alpha^2}{n}, \tag{2.24}$$

thus β is bounded from both sides with respect to α. Moreover, as $n \to \infty$, we obtain a limit inequality $\beta \leqslant \alpha$, which means that damping for leafs should be lower than damping for the center. The region satisfying (2.24) is depicted in Fig. 2.6a.

Simulation results for both cases, $\gamma = (1,0)^T$ and $\gamma = (0,1)^T$, and for $n = 20$ leafs are given in Fig. 2.5, with dampings $\alpha = 2$, $\beta = 1.1$, desired output value $y_d = 5$ and integral

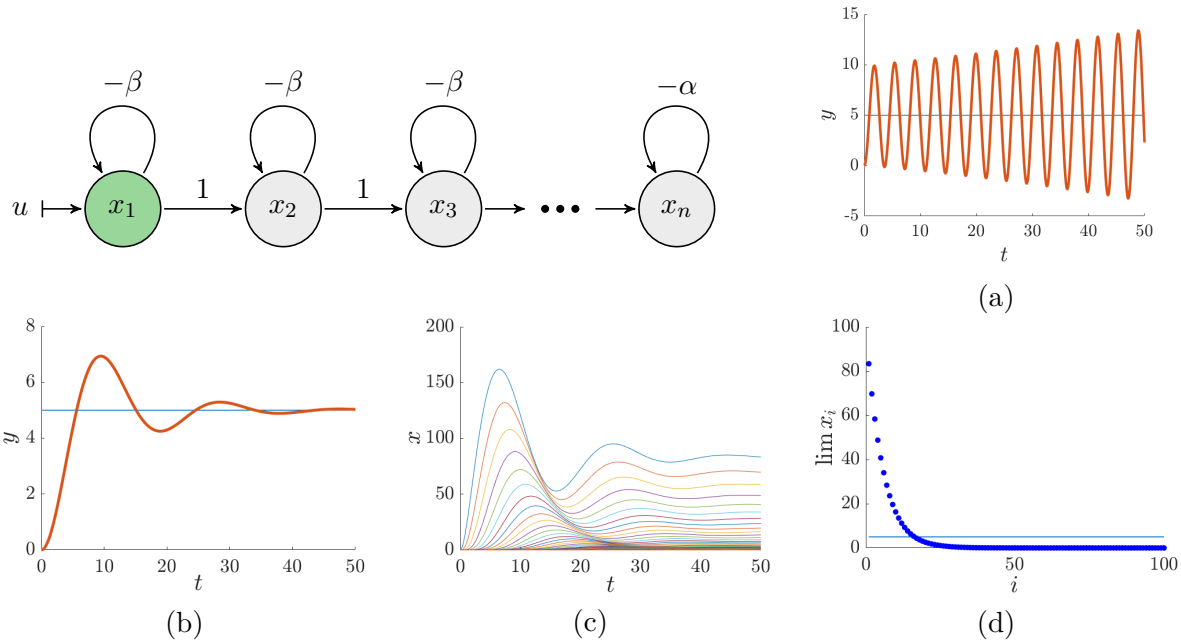


Figure 2.7: **Scheme:** Directed line network with n nodes. Boundary nodes are in green. **Plots:** Output control of the directed line network, $\alpha = 0.2$, $\kappa = 12$, $y_d = 5$. **(a):** $n = 4$, $\beta = 0.002$. Output y of the network is unstable, $CA^2 \not\succ 0$. **(b), (c), (d):** $n = 100$, $\beta = 0.2$. Network is stable. **(b):** Output y. **(c):** Spread of states x. **(d):** Values of $\lim_{t\to\infty} x_i$ depending on the number $i \in \{1, ..., 100\}$, which is the distance from the controlled node.

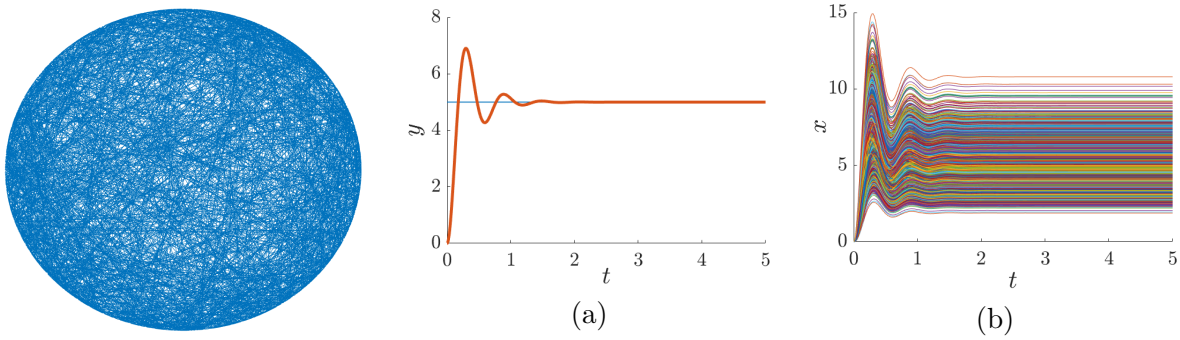
controller gain $\kappa = 12$. On Fig. 2.5a it is clearly seen that controlling the central node and controlling the leaf has almost the same effect on the output y.

2.2.6.2 Line network

Now we explore an example of a directed line network with n nodes. This network is depicted on Fig. 2.7. As usual, we are interested in controlling average state of the network, and it is assumed that we can control only the input node x_1 of the system. System matrices for n nodes are given as follows:

$$A = \begin{pmatrix} -1-\beta & 0 & \cdots & 0 \\ 1 & -1-\beta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\alpha \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad C = \frac{1}{n}\mathbf{1}^T, \quad \gamma = 1. \tag{2.25}$$

This choice of system matrices corresponds to the directed line network with damping $\alpha > 0$ for the last node and $\beta > 0$ for all other nodes.



(a) (b)

Figure 2.8: Output control of Erdős-Rényi graph for $n = 4000$ nodes, $\kappa = 250$. **(a):** Dynamics of output y. **(b):** Dynamics of states x.

Calculating CA and CA^2 gives

$$\begin{aligned} CA &= \begin{pmatrix} -\beta & -\beta & \cdots & -\alpha \end{pmatrix} /n < 0, \\ (CA^2)_{1,\ldots,n-2} &= \beta^2/n > 0, \\ (CA^2)_{n-1} &= (\beta^2 + \beta - \alpha)/n, \\ (CA^2)_n &= (\alpha^2 + \alpha - \beta)/n. \end{aligned} \tag{2.26}$$

$CA^2 > 0$ means then $\alpha^2 + (\alpha - \beta) \geqslant 0$ and $\beta^2 + (\beta - \alpha) \geqslant 0$. Solving this for damping of leaf nodes we obtain

$$\sqrt{\alpha + \frac{1}{4}} - \frac{1}{2} \leqslant \beta \leqslant \alpha + \alpha^2, \tag{2.27}$$

thus β is bounded from both sides with respect to α. The region satisfying (2.27) is depicted in Fig. 2.6b.

In order to validate our conclusions about this example, we take directed line networks with 4 and 100 nodes and check whether they are stable or unstable for different κ.

Fix $\alpha = 0.2$, therefore for condition $CA^2 > 0$ to hold one needs $\sqrt{0.45} - 0.5 \leqslant \beta \leqslant 0.24$. On Fig. 2.7 simulation results are shown for $\kappa = 12$, $y_d = 5$ and for two values of β, the first, $\beta = 0.2$, satisfies the condition, and the second $\beta = 0.002$ does not. In the case $\beta = 0.2$ and $n = 100$ it is very interesting to see what are limit values of the state variables x. It appears that they decrease exponentially starting from the controlled node x_1, while preserving their average equal to y_d. This is due to the fact that in the steady state all nodes' states except the first one and the last one should satisfy relation $x_{i-1} - (1 + \beta)x_i = 0$.

2.2.6.3 Random Erdős-Rényi graph

Here we present a simulation results for an integral controller for random Erdős-Rényi graph with $n = 4000$ nodes and probability of creating an edge $p = 0.01$. Vector $C = \mathbf{1}^T/n$ represents the average, and the system matrix A is a negative of the Laplacian of this ER

graph with an additional random damping on every node taken from the uniform distribution $U(6,7)$ such that $CA^2 > 0$. Matrix B is chosen to be a random vector of zeros and ones with equal probability. With such setup for any $\kappa > 0$ the system converges to the desired output reference $y_d = 5$, see Fig. 2.8.

2.3 Control of multiple linear aggregates

In this section we will enlarge our results of control of linear outputs to their reference values to include also a multidimensional output setup. Ideally we would like to obtain a condition similar to Theorem 2.3 which could describe classes of systems for which integral controller with arbitrary large gains would be stable.

2.3.1 Problem formulation

Assume again the system (2.2) has a special structure, corresponding to a network controlled from boundaries. Namely, let a state vector be divided into two parts, $x^T = (x_1^T, x_2^T)$. States $x_1 \in \mathbb{R}^k$ correspond to the boundary nodes, which can be directly controlled, and states $x_2 \in \mathbb{R}^{n-k}$ are inner nodes, thus no control is applied to them. Assume further that the subnetwork corresponding to the inner nodes is undirected, while an interconnection between inner and boundary nodes exists only in the direction from boundaries to inner nodes, thus there is no influence from inner nodes to boundaries. Schematically this structure is depicted in Fig. 2.9.

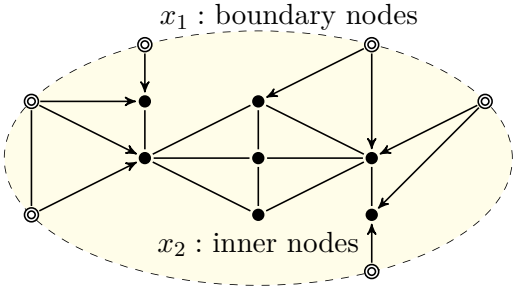


Figure 2.9: Network with boundary and inner nodes separation for multidimensional output control. Inner nodes have no influence on boundary nodes and their subnetwork is undirected.

In contrast to the previous section, the goal is to control multiple outputs to their desired values. Note that using direct control of boundary nodes, the number of outputs m should be the same as the number of inputs k in order to use the passivity formalism. The system model is then:

$$\begin{cases} \dot{x}_1 = A_{11}x_1 + u, \\ \dot{x}_2 = A_{21}x_1 + A_{22}x_2, \\ y = C_1x_1 + C_2x_2, \end{cases} \tag{2.28}$$

with an additional assumptions that $A_{22} = A_{22}^T \in \mathbb{R}^{(n-k)\times(n-k)}$ is a symmetric negative-definite matrix, corresponding to the undirected stable subnetwork, and that A_{21} is of full rank, meaning that the boundary nodes act independently.

Define the integral controller

$$\dot{u} = \Gamma(y_d - y), \tag{2.29}$$

where $\Gamma \in \mathbb{R}^{k\times k}$ is a symmetric positive-definite matrix. Then closed-loop system has a structure as in Fig. 2.10.

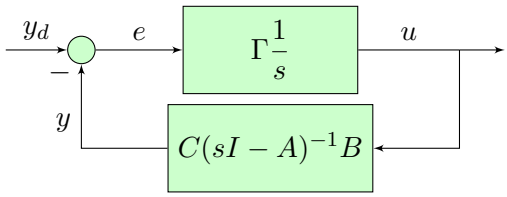


Figure 2.10: Feedback interconnection of passive systems in MIMO case

As before, we will use passivity decomposition of feedback interconnection similar to Section 2.2.4 and Theorem 2.4. Accomplishment of the control goal follows from $H_1(s) = \Gamma/s$ being PR and $H_2(s) = C(sI - A)^{-1}B$ being SPR. If Γ is positive-definite, $H_1(s)$ is PR. Thus we will focus on proving that under certain conditions the system (2.28) is SPR.

2.3.2 Conditions on passivity

Now we present a theorem which is the main result of this section:

Theorem 2.5. *If the matrix C_1 is symmetric positive-definite and a matrix inequality (2.30) is satisfied*

$$4H + \delta K - JK^{-1}J \succ 0, \tag{2.30}$$

with matrices H, J, K and a positive scalar δ defined as follows:

$$\begin{aligned} H &= C_2A_{22}C_2^T + C_2C_2^TA_{11} + A_{11}^TC_2C_2^T + A_{11}^TC_2A_{22}^{-1}C_2^TA_{11}, \\ J &= 2C_1A_{11} + 2A_{11}^TC_1 + C_2A_{21} + A_{21}^TC_2^T - A_{21}^TA_{22}^{-1}C_2^TA_{11} - A_{11}^TC_2A_{22}^{-1}A_{21}, \\ K &= A_{21}^TA_{22}^{-1}A_{21}, \end{aligned} \tag{2.31}$$

$$\delta = \begin{cases} \frac{1}{4}\left(\lambda_{max}(JK^{-1}) - \lambda_{min}(JK^{-1})\right)^2, \\ \qquad when \quad 4\lambda_{max}(C_2^TC_1^{-1}C_2) \leqslant \lambda_{min}(JK^{-1}) + \lambda_{max}(JK^{-1}) \\ 4\left(\lambda_{max}(C_2^TC_1^{-1}C_2) - \frac{1}{2}\lambda_{min}(JK^{-1})\right)^2 \\ \qquad when \quad 4\lambda_{max}(C_2^TC_1^{-1}C_2) > \lambda_{min}(JK^{-1}) + \lambda_{max}(JK^{-1}) \end{cases} \tag{2.32}$$

then the system (2.28) has a strictly positive real (SPR) transfer function.

Proof. A known result (Narendra 2014) from passivity theory says that a stable system is SPR iff there exists $P = P^T \succ 0$ such that

$$\begin{cases} A^T P + PA \prec 0, \\ PB = C^T. \end{cases} \tag{2.33}$$

Note that it is not possible to test this property directly for a given system, since it requires finding a feasible solution P, which can be done for example by optimization techniques such as LMIs. However it was also shown (Tao and Ioannou 1990, Theorem 3.4) that if such matrix P exists, then

$$\begin{cases} CB = (CB)^T \succ 0, \\ CAB + (CAB)^T \prec 0. \end{cases} \tag{2.34}$$

It is easy to show that the conditions (2.34) are also sufficient if matrices B and C are square and non-singular (all n nodes in the system are controlled). Indeed, take $P = C^T B^{-1}$. Then $B^T P B = (CB)^T = CB = B^T P^T B \succ 0$, from which it is clear that $P = P^T \succ 0$. Then, $CAB + (CAB)^T \prec 0$ implies $B^T(PA + A^T P)B \prec 0$, which is possible only if $A^T P + PA \prec 0$.

In our case only $k < n$ nodes are controlled, thus the control matrix is not square. But we can assume that there exist also $n - k$ "virtual" controls, acting on the inner nodes, such that the modified control matrix is square. Moreover, by the structure of the system real controls form the identity matrix of rank k, thus the reasonable choice for virtual controls is the identity matrix of rank $n - k$.

The observation matrix C should also be augmented, and from the condition $CB = (CB)^T \succ 0$ with identity controls it follows that $C = C^T \succ 0$. Therefore we can define all the matrices,

$$C = \begin{pmatrix} C_1 & C_2 \\ C_2^T & D \end{pmatrix}, \qquad A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}, \qquad B = I,$$

where $C_1 = C_1^T \succ 0$ and $D = D^T \in \mathbb{R}^{(n-k)\times(n-k)}$ is some positive-definite matrix which corresponds to the "virtual" observations.

The main reason to augment the system with virtual controls and observations is that once SPR-ness of the augmented system is proven, it immediately implies the that the transfer function of the original system is also SPR. If there exists P such that (2.33) holds for the augmented system, the same P can be used to prove (2.33) for the original system. Indeed, the first condition $A^T P + PA \prec 0$ is the same for both systems, and $PB = C^T$ can be decomposed into $PB_i = C_i^T$, where B_i is the i-th column of the control matrix and C_i is the i-th row of the observation matrix, thus for every subset of controls and observations the equality holds.

Now we define a matrix G representing the second condition (2.34) for the augmented system:

$$\begin{aligned} G =& CAB + (CAB)^T = CA + A^T C = \\ & \begin{pmatrix} C_1 A_{11} + A_{11}^T C_1 + C_2 A_{21} + A_{21}^T C_2^T & C_2 A_{22} + A_{11}^T C_2 + A_{21}^T D \\ A_{22} C_2^T + C_2^T A_{11} + D A_{21} & A_{22} D + D A_{22} \end{pmatrix} \end{aligned} \tag{2.35}$$

The main question is if there exists such matrix $D = D^T \succ 0$ that $C \succ 0$ and $G \prec 0$. In general this question is very hard to answer, but we can restrict our attention to a special class of matrices D such that the sufficient conditions on C and A can be obtained. Namely, let us choose $D = \alpha I$ for some positive scalar α. Then if we can find α such that $C \succ 0$ and $G \prec 0$, the system is SPR. With the new variable α the matrix G becomes

$$G = \begin{pmatrix} C_1A_{11} + A_{11}^TC_1 + C_2A_{21} + A_{21}^TC_2^T & C_2A_{22} + A_{11}^TC_2 + \alpha A_{21}^T \\ A_{22}C_2^T + C_2^TA_{11} + \alpha A_{21} & 2\alpha A_{22} \end{pmatrix}. \tag{2.36}$$

By Schur's Complement, $C \succ 0$ leads to the condition $\alpha I - C_2^TC_1^{-1}C_2 \succ 0$, thus α should satisfy

$$\alpha > \lambda_{max}(C_2^TC_1^{-1}C_2) \tag{2.37}$$

And applying the Schur's Complement to the matrix G (and recalling that A_{22} is negative-definite) we see that $G \prec 0$ is equivalent to

$$\begin{aligned} &2\alpha(C_1A_{11} + A_{11}^TC_1 + C_2A_{21} + A_{21}^TC_2^T) - \\ &(C_2A_{22} + A_{11}^TC_2 + \alpha A_{21}^T)A_{22}^{-1}(A_{22}C_2^T + C_2^TA_{11} + \alpha A_{21}) \prec 0 \end{aligned} \tag{2.38}$$

Using the definitions (2.31) and removing brackets we get

$$\alpha^2K - \alpha J + H \succ 0, \tag{2.39}$$

where $K \prec 0$ because $A_{22} \prec 0$ and A_{21} is of full rank, and J and H are in general sign-indefinite.

Define $L = (-K)^{-1/2}$, where square root is chosen such that L is positive definite. Multiplying (2.39) from both sides by L, we obtain

$$\alpha^2I + \alpha LJL - LHL \prec 0, \tag{2.40}$$

which can be rearranged as

$$\left(\alpha I + \frac{1}{2}LJL\right)^2 \prec L\left(H - \frac{1}{4}JK^{-1}J\right)L \tag{2.41}$$

If it had been a scalar quadratic equation, it would be possible to make the left-hand side zero by choosing appropriate α. In our case it is not possible, but we can find an upper bound on this term. Namely,

$$\left(\alpha I + \frac{1}{2}LJL\right)^2 \prec \lambda_{max}\left[\left(\alpha I + \frac{1}{2}LJL\right)^2\right] I, \tag{2.42}$$

where

$$\begin{aligned} &\lambda_{max}\left[\left(\alpha I + \frac{1}{2}LJL\right)^2\right] = \max\left\{\lambda_{max}\left(\alpha I + \frac{1}{2}LJL\right), -\lambda_{min}\left(\alpha I + \frac{1}{2}LJL\right)\right\}^2 = \\ &= \max\left\{\alpha + \frac{1}{2}\lambda_{max}(LJL), -\alpha - \frac{1}{2}\lambda_{min}(LJL)\right\}^2 = \\ &= \max\left\{\alpha - \frac{1}{2}\lambda_{min}(JK^{-1}), -\alpha + \frac{1}{2}\lambda_{max}(JK^{-1})\right\}^2. \end{aligned} \tag{2.43}$$

The minimal value is achieved when two arguments of maximum are equal. Define $\alpha_1^* = \frac{1}{4}\left(\lambda_{min}(JK^{-1}) + \lambda_{max}(JK^{-1})\right)$. Then the bound is

$$\delta_1 := \lambda_{max}\left[\left(\alpha_1^* I + \frac{1}{2}LJL\right)^2\right] = \frac{1}{16}\left(\lambda_{max}(JK^{-1}) - \lambda_{min}(JK^{-1})\right)^2. \tag{2.44}$$

This value is optimal, but it assumes that α_1^* can be used, which is possible only if (2.37) is satisfied. Denote $\alpha_2^* = \lambda_{max}(C_2^T C_1^{-1} C_2)$. If $\alpha_2^* > \alpha_1^*$, then the bound is

$$\begin{aligned}\delta_2 := \lambda_{max}\left[\left(\alpha_2^* I + \frac{1}{2}LJL\right)^2\right] &= \left(\alpha_2^* - \frac{1}{2}\lambda_{min}(JK^{-1})\right)^2 = \\ &= \left(\lambda_{max}(C_2^T C_1^{-1} C_2) - \frac{1}{2}\lambda_{min}(JK^{-1})\right)^2.\end{aligned} \tag{2.45}$$

Finally, defining $\delta := 4\delta_1$ if $\alpha_2^* \leqslant \alpha_1^*$ and $\delta := 4\delta_2$ otherwise, we can rewrite the quadratic equation (2.41) as

$$\left(\alpha I + \frac{1}{2}LJL\right)^2 \prec \frac{1}{4}\delta I \prec L\left(H - \frac{1}{4}JK^{-1}J\right)L, \tag{2.46}$$

and, multiplying both sides by L^{-1}, get a sufficient condition

$$-\delta K \prec 4H - JK^{-1}J, \tag{2.47}$$

which is exactly the condition (2.30) that we aimed to obtain. □

Remark 2.1. The result of the theorem is only a sufficient condition for the transfer function to be SPR. The augmentation of the system with virtual controls and observations is still an equivalence operation, as one can always choose such additional columns for B and C^T that $PB = C^T$ holds. The equivalence is lost when the matrix D is substituted with αI. For the future work it would be possible to add an additional degree of freedom to this procedure by considering an augmentation of matrix B in the form

$$B = \begin{pmatrix} I & 0 \\ 0 & \beta I \end{pmatrix},$$

thus obtaining a system of two quadratic matrix inequalities on α and β. Possibly tighter sufficient conditions could be recovered as a result.

2.3.3 Examples

Inequality (2.30) is a straightforward condition to check, given the system, but it is hard to interpret in general. However from the definitions of matrices H, J and K it is clear that the condition is easier to satisfy for bigger C_1, for C_2 closer to zero and for bigger K and $-K^{-1}$. The latter happens when A_{22} is strongly negative, for example when inner nodes have strong negative self-loops.

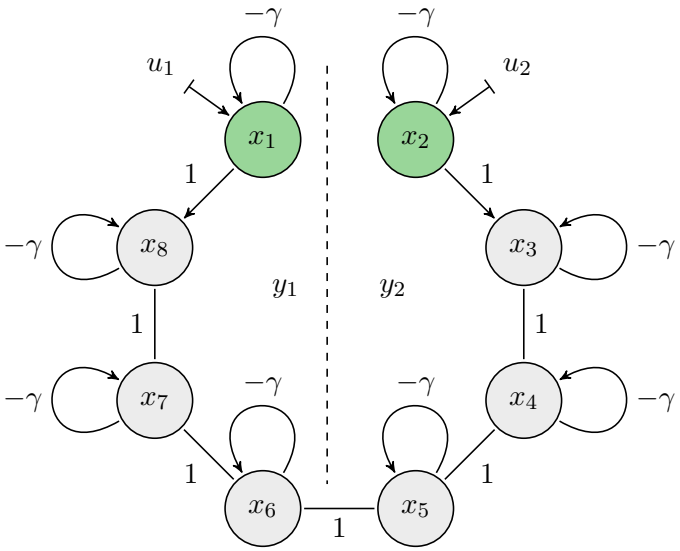


Figure 2.11: Undirected line network, controlled from two sides, with $n = 8$ nodes. Boundary nodes are in green. Network is splitted into two parts by dahsed line, denoting two separate outputs y_1 and y_2.

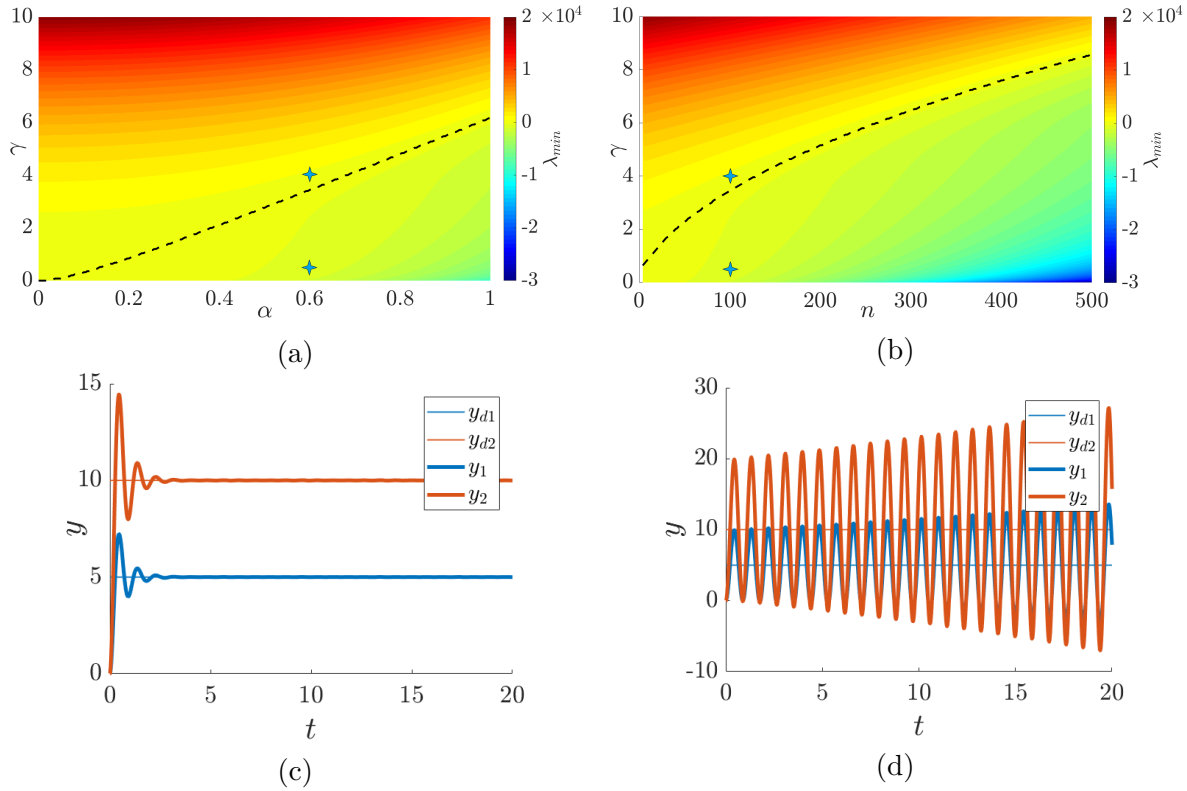


Figure 2.12: **Top:** Smallest eigenvalue $\lambda_{min}(4H + \delta K - JK^{-1}J)$, depending on **(a):** α and γ for $n = 100$, **(b):** n and γ for $\alpha = 0.6$. Dashed line denotes zero level. All points above dashed line satisfy (2.30). Blue stars denote points in which bottom images are obtained. **Bottom:** Multi-output control of the undirected line network, $n = 100$, $\kappa = 50$, $\alpha = 0.6$. **(c):** $\gamma = 4$, the closed-loop system is stable, **(d):** $\gamma = 0.5$, the closed-loop system is unstable.

It can be shown on the example of an undirected line network, controlled from two sides, with the aim to stabilize two halves of the network to different average values. The network for $n = 8$ nodes is depicted in Fig. 2.11.

Dynamics of this network are given by the matrices

$$A = \left(\begin{array}{cc|ccccc} -\gamma & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -\gamma & 0 & 0 & \cdots & 0 & 0 \\ \hline 0 & 1 & -2-\gamma & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -2-\gamma & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -2-\gamma & 1 \\ 1 & 0 & 0 & 0 & \cdots & 1 & -2-\gamma \end{array}\right), \tag{2.48}$$

$$C = \left(\begin{array}{cc|ccc|ccc} 1 & 0 & 0 & \cdots & 0 & \alpha & \cdots & \alpha \\ 0 & 1 & \alpha & \cdots & \alpha & 0 & \cdots & 0 \end{array}\right), \qquad B = \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{2.49}$$

This system can be either SPR or not for given n, depending on parameters α and γ. Intuitively, lower is α, smaller is the influence of C_2, and easier is to obtain passivity. In the same way if γ is large enough, A_{22} is strongly negative and thus system is passive. Indeed, in Fig. 2.12a the smallest eigenvalue λ_{min} of the matrix $4H + \delta K - JK^{-1}J$ is depicted for $n = 100$ depending on both α and γ. The condition (2.30) is satisfied when $\lambda_{min} > 0$.

Dependence of the condition (2.30) on the size of a network is presented in Fig. 2.12b, where the same $\lambda_{min}(4H + \delta K - JK^{-1}J)$ is depicted for $\alpha = 0.6$ for various γ and n. It is clear that for a longer line network to be passive its negative self-loops should be stronger.

Now we apply an integral controller to this system, with a goal to stabilize the outputs to the desired values $y_{d1} = 5$ and $y_{d2} = 10$. The controller has the form

$$\dot{u} = \begin{pmatrix} \kappa & 0 \\ 0 & \kappa \end{pmatrix} \begin{pmatrix} y_{d1} - y_1 \\ y_{d2} - y_2 \end{pmatrix}. \tag{2.50}$$

Simulation results for $n = 100$ and $\kappa = 50$ are shown in Fig. 2.12c,d. Indeed, for fixed value of $\alpha = 0.6$, small value of $\gamma = 0.5$ leads to unstable behaviour of the closed-loop system, while $\gamma = 4$ is stable.

2.4 Minimization of deviation

Previous sections showed that control of linear output can be performed rather easily by integral controller and without any knowledge of the system. But in some cases controlling an average with arbitrary control direction γ can lead to a poor performance: although the average state $y \to y_d$, states themselves can be very far from y_d. As an example one can look

at Fig. 2.7c-d, where a spread of steady states x is shown. Although the average value is 5, most of the states are almost zero, while some states are much larger, around 80.

This dispersion between states is captured by the squared deviation V. The smaller it is, the closer are the states to their average value. Therefore, it makes sense to find a control law $u = u(y, V)$ for the system (2.2) which solves simultaneously two problems: assures $\lim_{t\to\infty} y(t) = y_d$ and minimizes the squared deviation V.

Preliminary, let us make the following observation. Controlling a linear output $y \in \mathbb{R}^m$ to the desired value $y_d \in \mathbb{R}^m$ in a steady state means that the system should satisfy m-dimensional constraint $-CA^{-1}Bu^* = y_d$, thus if the dimension of the steady-state control vector u^* is $k > m$, there are still $k - m$ degrees of freedom left for optimizing the control direction in sense of minimization of the squared deviation.

2.4.1 Explicit solution

Let us assume that the desired steady state is reached and try to find it. Denote x^* and u^* as the state vector and the control vector respectively in the steady state. Also denote the steady-state squared deviation as V^*. Then the equations for the steady state, obtained from the system (2.2) in assumption that $y \to y_d$ are

$$\begin{cases} 0 = Ax^* + Bu^*, \\ y_d = Cx^*, \\ V^* = x^{*T} P x^*. \end{cases} \tag{2.51}$$

Our problem can be seen as a linear constrained quadratic minimization problem:

$$\begin{gathered} \text{minimize } V^* = x^{*T} P x^*, \\ \text{subject to } Ax^* + Bu^* = 0, \\ Cx^* = y_d. \end{gathered} \tag{2.52}$$

In comparison to the standard linear-quadratic regulator, note that problem (2.52) is formulated for the steady state, thus there is no more dynamics in it, as well as no optimization of the transient process.

Assume for a moment that all the system matrices are known. Using the fact that the matrix A is stable, we can take the inverse and thus obtain the steady state vector $x^* = -A^{-1}Bu^*$. Denoting $S = B^T A^{-T} P A^{-1} B$ and $\eta = -B^T A^{-T} C^T$, we can write the minimization problem (2.52) in terms of u^*:

$$\begin{gathered} \text{minimize } V^* = u^{*T} S u^*, \\ \text{subject to } \eta^T u^* = y_d. \end{gathered} \tag{2.53}$$

Solution for the constrained problem is found using the Lagrangian:

$$L(u) = u^T S u + \lambda^T (\eta^T u - y_d). \tag{2.54}$$

Minimizing it over the control variable and solving for the Lagrange multiplier λ, we find that the explicit solution to the minimization problem is given by

$$\begin{cases} \lambda^* = -2\left(\eta^T S^{-1}\eta\right)^{-1} y_d, \\ u^* = S^{-1}\eta\left(\eta^T S^{-1}\eta\right)^{-1} y_d, \\ x^* = -A^{-1}BS^{-1}\eta\left(\eta^T S^{-1}\eta\right)^{-1} y_d. \end{cases} \tag{2.55}$$

Without loss of generality we will assume that S is positive definite for the future analysis. This property corresponds to the fact that the minimizing control is unique.

The solution (2.55) cannot be used explicitly due to the fact that the system matrices are assumed to be unknown. But the next section introduces an algorithm which is able to stabilize the system in the arbitrary small neighbourhood of this solution.

2.4.2 Extremum seeking

Extremum seeking is *a form of adaptive control where the steady-state input-output characteristic is optimized, without requiring any explicit knowledge about this input-output characteristic other than that it exists and that it has an extremum* (Ariyur and Krstić 2003; Tan et al. 2010). This algorithm, developed in the first part of XX century, explores the control space with small oscillations and provides an approximation of the gradient, which then can be integrated in order to find the optimum.

In standard realisations of extremum seeking, one adds to a current control input an oscillating signal, which should be small and slow in comparison with the system dynamics. Further, multiplying the output by the same oscillating signal, it is possible to recover an estimate of the gradient of the output with respect to the input.

This standard algorithm is unfortunately not usable for us, since we want to perform a constrained optimization (2.53) with a constraint that the average steady state should be equal to the desired one. However, if we modify the algorithm so to minimize the Lagrangian (2.54) instead of the squared deviation itself, we will optimize the original squared deviation while preserving the average state constraint. This modification leads to an introduction of a vector of Lagrange multipliers λ, which can be reconstructed by an additional integrator.

Assume the control law for the system (2.2) is given by

$$\begin{cases} \dot{\bar{u}} = -\kappa\omega r(\omega t)(V + \lambda^T(y - y_d)), \\ \dot{\lambda} = \kappa a\omega\kappa_\lambda(y - y_d), \\ u = \bar{u} + ar(\omega t), \end{cases} \tag{2.56}$$

where a and κ are small gains, ω is a small frequency, κ_λ is a relative Lagrange multiplier adaptation gain, and the oscillating signal $r(\omega t)$ is defined as

$$r(\omega t) = \sqrt{2}\cdot\begin{pmatrix}\sin(2\pi\omega t) & \cos(2\pi\omega t) & \sin(4\pi\omega t) & \cdots\end{pmatrix}^T.$$

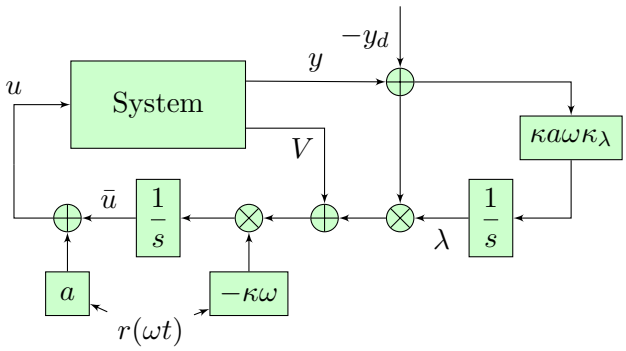


Figure 2.13: Extremum seeking scheme for constrained minimization

Inputs to this control scheme (Fig. 2.13) are the output average y and output squared deviation V. Therefore, the control law does not use any state feedback or the system matrices. This is a multi-variable (Moase et al. 2011) extremum seeking control scheme, augmented with an additional integrator for the adaptation of the Lagrange multiplier as in primal-dual method (Nedić and Ozdaglar 2009; Simpson-Porco et al. 2019) with κ_λ being the relative speed of adaptation.

To begin with the proof of the stability of the control scheme (2.56), we first present a preliminary analysis, treating ω, κ and a as "small" parameters. Assuming ω is small, we define a new time-scale $\tau := \omega t$, which should be slow enough such that the dynamics of the system (2.2) is much faster than the dynamics of the adaptation. Under this time-scale the closed-loop system equations are

$$\begin{cases} \omega \dfrac{dx}{d\tau} = Ax + B(\bar{u} + ar(\tau)), \\ \dfrac{d\bar{u}}{d\tau} = -\kappa r(\tau) \left(x^T P x + \lambda^T (Cx - y_d) \right), \\ \dfrac{d\lambda}{d\tau} = \kappa a \kappa_\lambda (Cx - y_d). \end{cases} \tag{2.57}$$

With small ω the singular perturbation analysis (Kokotović, Khalil, and O'reilly 1999) can be performed, thus system dynamics is substituted with its steady-state input-state mapping, i.e. $x^* = x^*(u) = -A^{-1}Bu$. The reduced dynamics is then approximated by

$$\begin{cases} \dfrac{d\bar{u}}{d\tau} = -\kappa r(\tau) \big[(\bar{u} + ar(\tau))^T S (\bar{u} + ar(\tau)) + \lambda^T \left(\eta^T (\bar{u} + ar(\tau)) - y_d \right) \big], \\ \dfrac{d\lambda}{d\tau} = \kappa a \kappa_\lambda (\eta^T (\bar{u} + ar(\tau)) - y_d). \end{cases} \tag{2.58}$$

As a next step we introduce an additional time-scale $\theta := \kappa\tau$, using which the system becomes

$$\begin{cases} \dfrac{d\bar{u}}{d\theta} = -r(\tau) \big[(\bar{u} + ar(\tau))^T S (\bar{u} + ar(\tau)) + \lambda^T \left(\eta^T (\bar{u} + ar(\tau)) - y_d \right) \big], \\ \dfrac{d\lambda}{d\theta} = a \kappa_\lambda (\eta^T (\bar{u} + ar(\tau)) - y_d). \end{cases} \tag{2.59}$$

This system is periodic in τ with unit period. When κ is small, the reduced dynamics can be approximated well by the dynamics averaged over the unit period:

$$\begin{cases} \dfrac{d\bar{u}_{av}}{d\theta} = -\displaystyle\int_0^1 \left\{ r(\sigma)\left[(\bar{u}_{av} + ar(\sigma))^T S(\bar{u}_{av} + ar(\sigma)) + \lambda_{av}^T \left(\eta^T(\bar{u}_{av} + ar(\sigma)) - y_d\right)\right]\right\} d\sigma, \\ \dfrac{d\lambda_{av}}{d\theta} = \displaystyle\int_0^1 \left\{ a\kappa_\lambda(\eta^T(\bar{u}_{av} + ar(\sigma)) - y_d)\right\} d\sigma. \end{cases}$$

Recall that by the definition of $r(\cdot)$ the oscillating signal has the following properties: $\int_0^1 r(\sigma)d\sigma = 0$ and $\int_0^1 r(\sigma)r(\sigma)^T d\sigma = I$. Then we can rewrite the system:

$$\begin{cases} \dfrac{d\bar{u}_{av}}{d\theta} = -2aS\bar{u}_{av} - a\eta\lambda_{av} - a^2 R, \\ \dfrac{d\lambda_{av}}{d\theta} = a\kappa_\lambda \eta^T \bar{u}_{av} - a\kappa_\lambda y_d, \end{cases} \tag{2.60}$$

where $R = \int_0^1 r(\sigma)r(\sigma)^T S r(\sigma)d\sigma$. It can be seen that $2Su+\eta\lambda$ is the gradient of the Lagrangian with respect to control, therefore

$$\begin{cases} \dfrac{d\bar{u}_{av}}{d\tau} = -a\nabla_{\bar{u}_{av}} L + O(a^2), \\ \dfrac{d\lambda_{av}}{d\tau} = a\kappa_\lambda \nabla_{\lambda_{av}} L, \end{cases} \tag{2.61}$$

which converges to $O(a)$ of the explicit solution (u^*, λ^*). Concretely, analysing steady state we obtain

$$\begin{aligned} y_{av}^* &= \eta^T \bar{u}_{av}^* \equiv y_d, \\ \lambda_{av}^* &= \lambda^* - a\left(\eta^T S^{-1}\eta\right)^{-1} \eta^T S^{-1} R, \\ \bar{u}_{av}^* &= u^* + \frac{a}{2}\left[S^{-1}\eta\left(\eta^T S^{-1}\eta\right)^{-1}\eta^T S^{-1} R - S^{-1}R\right]. \end{aligned} \tag{2.62}$$

The rigorous stability proof is based on the notion of semi-global practical asymptotic stability:

Definition 2.2 (SPA stability, Tan, Nešić, and Mareels 2006)**.** Consider the parametrized family of systems:

$$\dot{\mathbf{x}} = f(t, \mathbf{x}, \varepsilon_1, \varepsilon_2, ..., \varepsilon_l), \tag{2.63}$$

where $\mathbf{x} \in \mathbb{R}^n$ and parameters of the system $\varepsilon_i > 0 \quad \forall i = 1, 2, ..., l$. The system (2.63) is said to be semi-globally practically asymptotically (SPA) stable in $[\varepsilon_1, \varepsilon_2, ..., \varepsilon_l]$ at $\mathbf{x}^*$, if there exists $\beta \in \mathcal{KL}$ (Khalil and Grizzle 2002) such that the following holds: for each pair of strictly positive numbers (Δ, ν), there exists $\varepsilon_1^* > 0$ and for any $\varepsilon_1 \in (0, \varepsilon_1^*)$ there exists $\varepsilon_2^* = \varepsilon_2^*(\varepsilon_1) > 0$ and for any $\varepsilon_2 \in (0, \varepsilon_2^*)$ there exists $\varepsilon_3^* = \varepsilon_3^*(\varepsilon_1, \varepsilon_2) > 0$, ..., there exists $\varepsilon_l^* = \varepsilon_l^*(\varepsilon_1, \varepsilon_2, ..., \varepsilon_{l-1}) > 0$ such that for any $\varepsilon_l \in (0, \varepsilon_l^*)$ the solutions of (2.63) with the parameters $[\varepsilon_1, \varepsilon_2, ..., \varepsilon_l]$ satisfy:

$$|\mathbf{x} - \mathbf{x}^*| \leqslant \beta(|\mathbf{x}_0 - \mathbf{x}^*|, (\varepsilon_1 \cdot \varepsilon_2 \cdots \varepsilon_l)(t - t_0)) + \nu \tag{2.64}$$

for all $t \geqslant t_0 \geqslant 0$, $\mathbf{x}(t_0) = \mathbf{x}_0$ with $|\mathbf{x}_0 - \mathbf{x}^*| \leqslant \Delta$.

Remark 2.2. Note that the order of the parameters $[\varepsilon_1, \varepsilon_2, ..., \varepsilon_l]$ is very important, because the bound for every parameter depends on the choice of all previous parameters, i.e. ε_3^* depends on the chosen ε_1 and ε_2.

Theorem 2.6. *System* (2.2) *with applied control law* (2.56) *is SPA stable in* $[a, \kappa, \omega]$ *at* (x^*, u^*, λ^*).

Proof. First we see that the system (2.60) is a linear system with the system matrix $M = \begin{pmatrix} -2S & -\eta \\ \kappa_\lambda \eta^T & 0 \end{pmatrix}$, multiplied by a. The matrix M is stable, which can be shown by analysing its eigenvalues. Assuming μ with $\operatorname{Re}\mu \geqslant 0$ being eigenvalue of M, we show that the characteristic polynomial can have no roots. By the Schur complement:

$$\det(M - \mu I) = \det\left(-\kappa_\lambda \eta^T (2S + \mu I)^{-1} \eta - \mu I\right) \det(-2S - \mu I). \tag{2.65}$$

Matrix S is positive definite, thus $2S + \mu I$ has eigenvalues with strictly positive real parts, which means $\det(-2S - \mu I) \neq 0$. Further, defining $\bar{S} = 2S + \operatorname{Re}\mu I$, by Lemma 2.1 we rewrite $\operatorname{Re}(2S + \mu I)^{-1} = \left(\bar{S} + (\operatorname{Im}\mu)^2 \bar{S}^{-1}\right)^{-1}$ which is clearly positive definite, therefore $\operatorname{Re}\eta^T (2S + \mu I)^{-1} \eta \succ 0$. Finally, since $\operatorname{Re}\mu \geqslant 0$, the real part of the matrix inside of the determinant of the first multiplier in (2.65) has negative eigenvalues, thus the determinant cannot be zero, which means μ cannot be an eigenvalue of M.

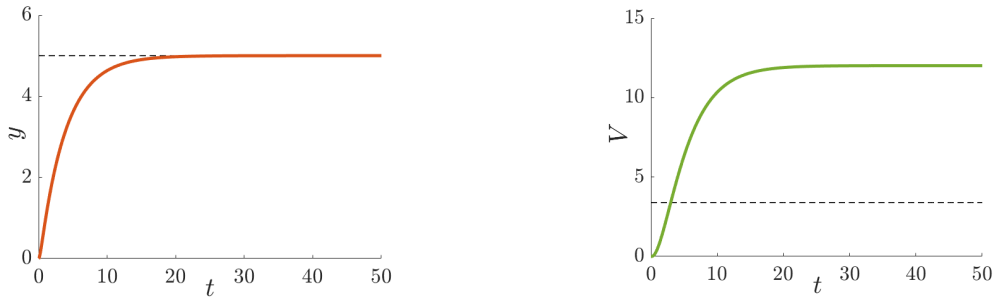
We see that the matrix M has no non-negative eigenvalues, which means that the system (2.60) converges to its steady state (2.62). In particular the system (2.60) is SPA stable in a. Then, using Lemma 1 from Tan, Nešić, and Mareels 2006, we see that the system (2.58) is SPA stable in $[a, \kappa]$, and finally using Lemma 2 from Tan, Nešić, and Mareels 2006, we conclude that the original closed-loop system is SPA stable in $[a, \kappa, \omega]$, which concludes the proof. □

Note that the stability of the closed-loop system (2.2)-(2.56) heavily depends on the chosen parameters. We proved SPA stability in $[a, \kappa, \omega]$, which by definition means that the bound for κ depends on the chosen α, and the bound for ω depends on the chosen α and κ. Therefore, it is difficult to find any rigorous bounds for how small these parameters should be. We can make only heuristic assumptions, such as requiring that all these parameters should be an order of magnitude smaller than the system impulse response.

Remark 2.3 (Additional integral controller). The extremum seeking scheme (2.56) can be enhanced by an array of possible modifications. For instance, an additional integral controller can be added:

$$\begin{cases} \dot{\bar{u}} = -\kappa a \omega \Gamma (y_d - y) - \kappa \omega r(\omega t)(V + \lambda^T (y - y_d)), \\ \dot{\lambda} = \kappa a \omega \kappa_\lambda (y - y_d), \\ u = \bar{u} + a r(\omega t), \end{cases} \tag{2.66}$$

where the matrix $\Gamma \in \mathbb{R}^{k \times m}$ such that $\eta^T \Gamma > 0$ plays the same role as in the first sections of this paper regarding average control. I.e. for positive system and in the case of scalar output it is enough to take $\Gamma > 0$, as in Theorem 2.2. Stability proof for the controller (2.66) follows exactly the same steps as in Theorem 2.6, replacing $S > 0$ by $S + \Gamma \eta^T / 2 > 0$.



(a) (b)

Figure 2.14: Integral controller (2.11) for average control, $T = 3.84$. **(a):** Average state y, black dashed line denotes y_d. **(b):** Squared deviation V, black dashed line denotes V^*.

This scheme can provide much faster convergence of the linear output (see the examples), which means that the bigger adaptation gains can be used without the possibility for λ to diverge. The equilibrium point for the averaged reduced model of the closed-loop system with this scheme is (2.62), exactly the same as in the previous case.

There are a lot of other possible modifications of the extremum seeking scheme that can be usefully included, for example the high- and low-pass filters that are added to pick up the adaptation signal, see Tan et al. 2010.

2.4.3 Examples

All the algorithms presented before were tested on a graph constructed as a random Erdős-Rényi graph with $n = 40$, probability of creating an edge (with weight 1) $p = 0.1$ and self-loops with weight -5. The dimension of the control vector was chosen $k = 3$, with matrix $B \in \mathbb{R}^{n\times k}$ being filled randomly: each element was set either to 0 or 1 with equal probability. Matrices C and P were chosen such that scalar linear output y corresponds to the average and V to the squared deviation of the states of the system. Desired value for the average was set $y_d = 5$.

In order to compare the speed of different algorithms we calculated characteristic times for the dynamics of average and squared deviation, defined as a negative inverse of the largest eigenvalue of the closed-loop system. Results of the simulations of different algorithms are presented in Figures 2.14-2.18.

To begin with, we apply the integral controller (2.11) to the system, and the dynamics of the average state y and the squared deviation V are shown on Fig. 2.14. The gain values are $\kappa = 1$ and $\gamma = [0, 0, 3]^T$. It is clearly seen that the squared deviation V does not reach its minimal value V^*, although this controller is the fastest one: its characteristic time is $T = 3.84$.

Now we aim to minimize the deviation of the system states together with controlling the average. If the extremum seeking scheme (2.56) is used, the goal is achieved, and the

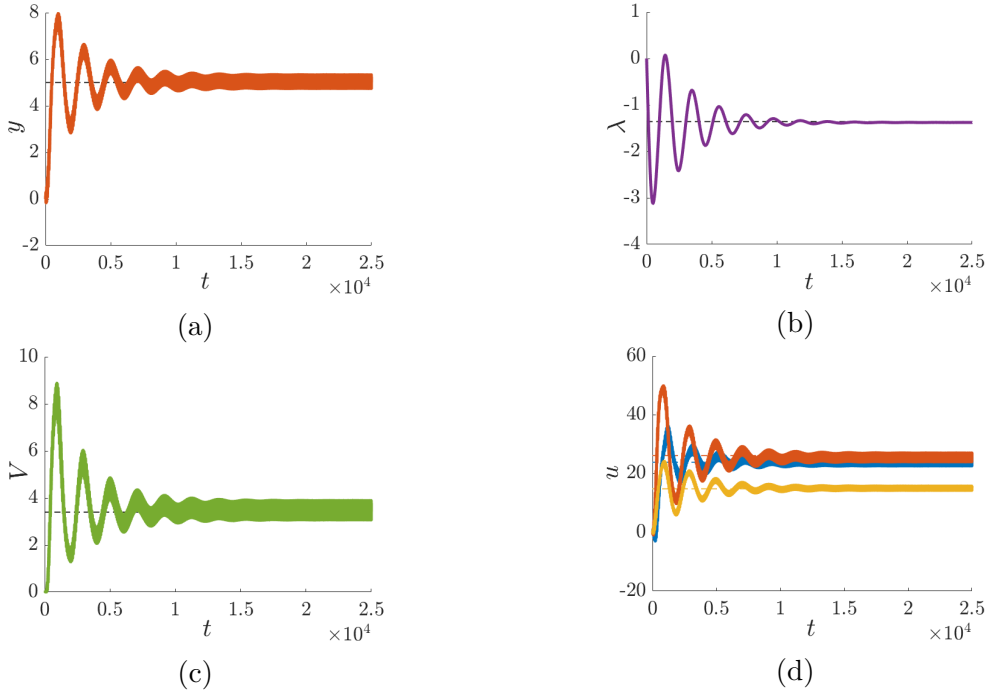


Figure 2.15: Extremum seeking control (2.56), $T = 2350$. **(a):** Average state y, black dashed line denotes y_d. **(b):** Lagrange multiplier λ, black dashed line denotes λ^*. **(c):** Squared deviation V, black dashed line denotes V^*. **(d):** Control vector u, dashed lines denote u^*.

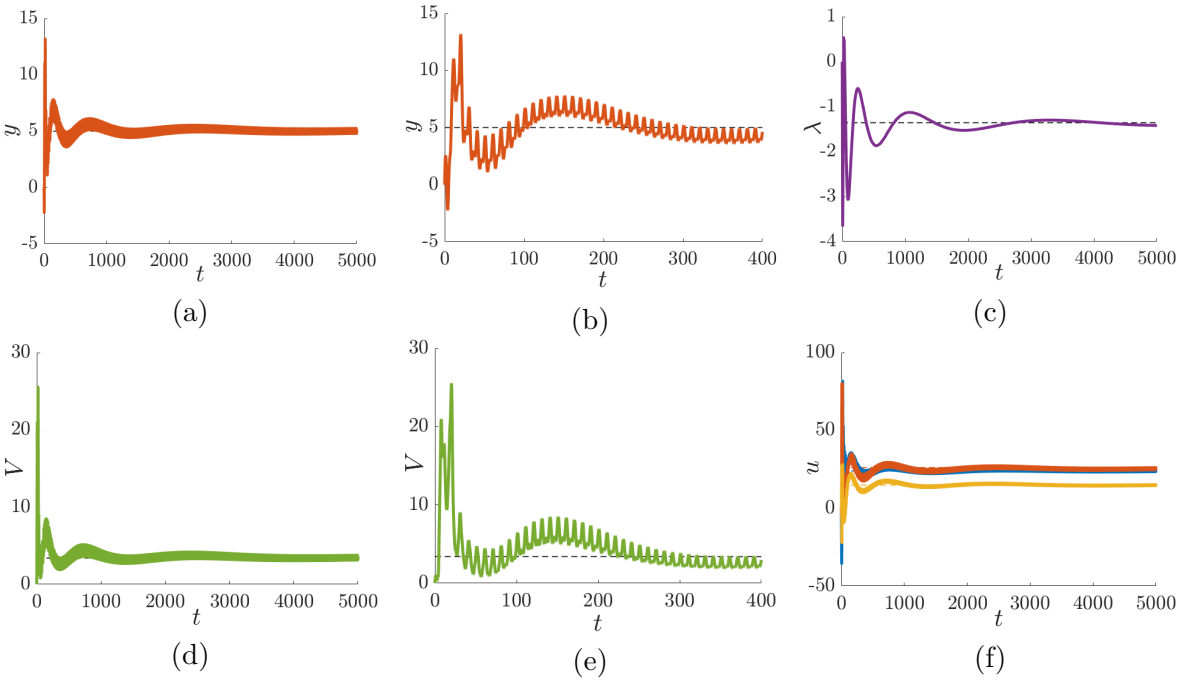


Figure 2.16: Extremum seeking control (2.56) with the gains decreasing over time, $T = 434$. **(a):** Average state y. **(d):** Squared deviation V. **(b)** and **(e):** Short-term plots for y and V. **(c):** Lagrange multiplier λ. **(f):** Control vector u.

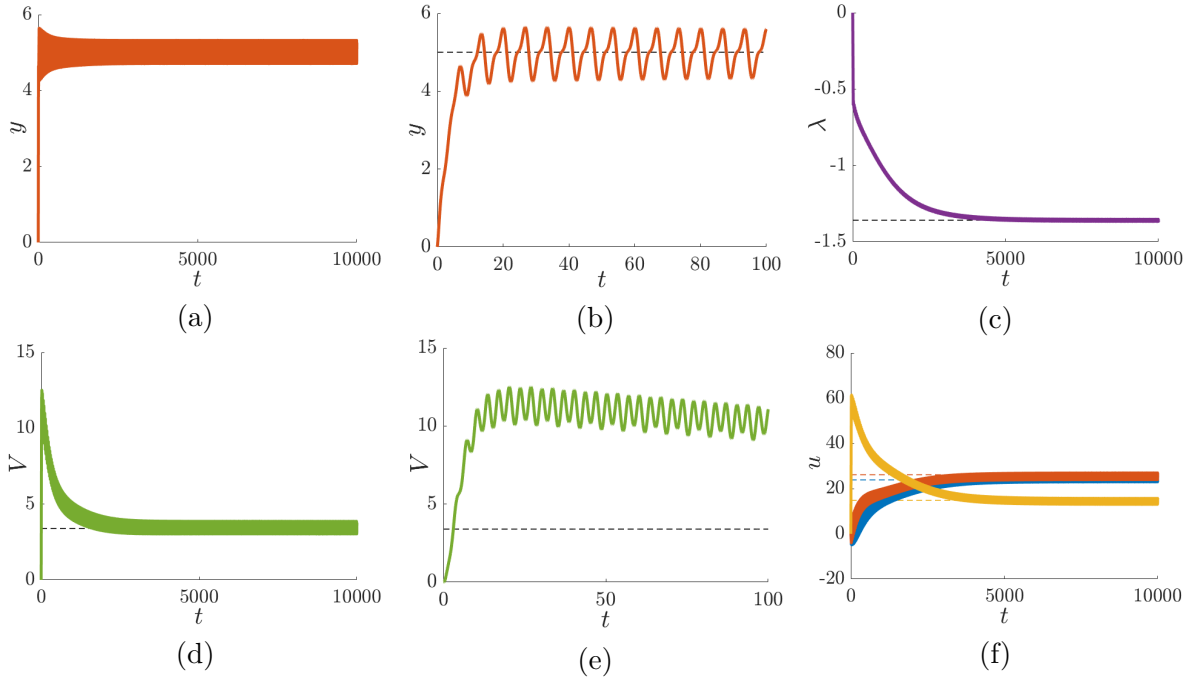


Figure 2.17: Extremum seeking control (2.66), $T_V = 732$ for squared deviation and $T_y = 3.9$ for average. **(a):** Average state y. **(d):** Squared deviation V. **(b)** and **(e):** Short-term plots for y and V. **(c):** Lagrange multiplier λ. **(f):** Control vector u.

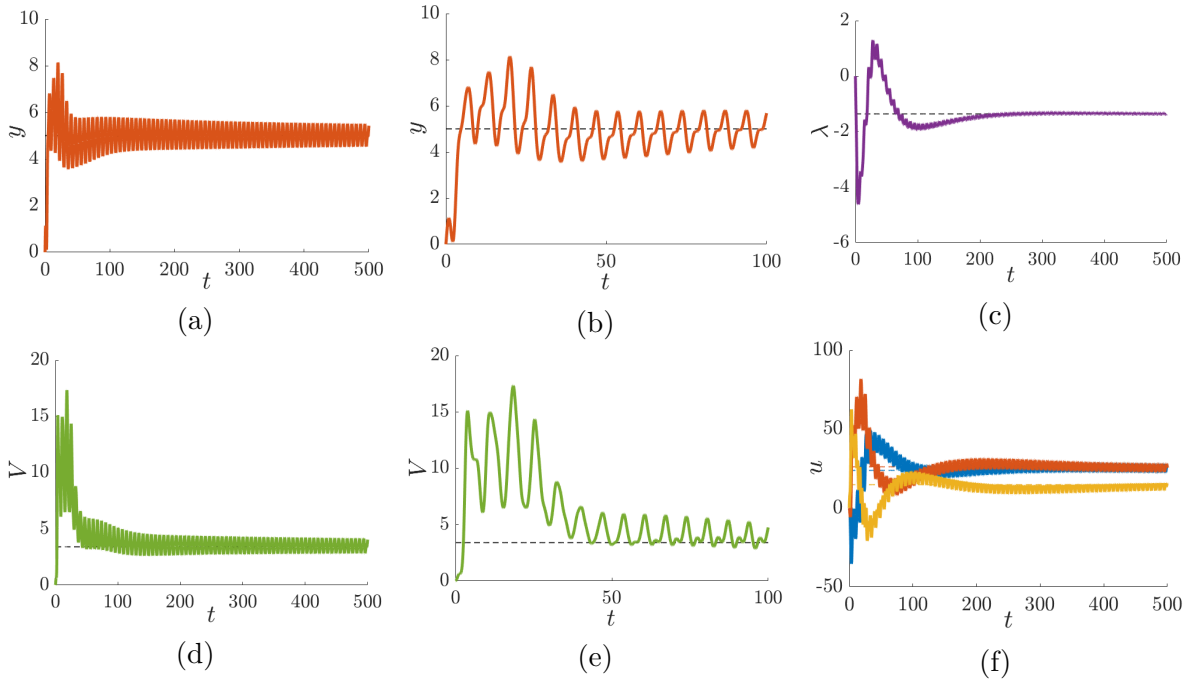


Figure 2.18: Extremum seeking control (2.66) with the gains decreasing over time, $T_V = 41$ and $T_y = 18$. **(a):** Average state y. **(d):** Squared deviation V. **(b)** and **(e):** Short-term plots for y and V. **(c):** Lagrange multiplier λ. **(f):** Control vector u.

performance is shown in Fig. 2.15. The gain values are $\omega = 0.1$, $\kappa = 2$, $a = 1$ and $\kappa_\lambda = 0.01$. This scheme is rather difficult to tune and also very slow, the characteristic time is $T = 2350$, three orders of magnitude higher than in Fig. 2.14. Also, significant oscillations in y, V and u can be seen even after the convergence of the system.

To minimize the oscillations in the extremum seeking, one needs to minimize the gains, but this would lead to an increased convergence time. Therefore we may try to improve the extremum seeking controller by adding the time-dependence to the gains, making them large at the beginning and decreasing them over time. Usually in adaptation algorithms the gains should decrease slower than $1/t$ (Borkar 2009), otherwise the algorithm does not converge. In the extremum seeking (2.56) the gains a and κ are multiplied together, therefore their product should decrease slower than $1/t$. In this example we set $a = \kappa = \frac{12}{(t/10+1)^{0.4}}$ together with $\omega = 0.1$ and $\kappa_\lambda = 0.01$. Performance of this scheme is shown in Fig. 2.16. It works much faster than the original one (the characteristic time is $T = 434$), and the oscillations are smaller, although the overshoot is larger.

Performance of the extremum seeking scheme (2.66) is shown on Fig. 2.17. The gain values are $\omega = 0.15$, $\kappa = 2$, $a = 1$ and $\kappa_\lambda = 0.1$. With respect to the scheme (2.56), the gain κ_λ can be chosen larger. This leads to the faster adaptation of the Lagrange multiplier, thus this scheme works faster than (2.56). The dynamics for average and deviation now behave differently due to the additional controller for the average, therefore it makes sense to find separate characteristic times for them. The characteristic time for average is just $T_y = 3.9$, while for squared deviation it is $T_V = 732$. The parameter of the integral controller is the same as in the case of the integral controller for average, $\kappa a \omega \gamma = [0, 0, 3]^T$.

Finally, the implementation of the scheme (2.66) with time-decreasing gains is presented in Fig. 2.18. The parameters are $\omega = 0.15$, $\kappa_\lambda = 0.1$, $\kappa a \omega \gamma = [0, 0, 3]^T$, and the dependent gains are $a = \frac{3}{(t/100+1)^{0.4}}$ and $\kappa = \frac{7}{(t/100+1)^{0.4}}$. It is clearly seen that this scheme is much faster than all previous ones, with the characteristic time for squared deviation being $T_V = 41$ and for average $T_y = 18$.

We see that the scheme (2.66) in general works faster that the scheme (2.56), but both of them are too slow to compare with the simple average controller (2.11). Their performance can be significantly increased using time-varying gains, although this leads to a large overshoot at the beginning.

2.5 Concluding remarks

In this chapter we considered a problem of control of aggregates of a large-scale network system (in particular its average and standard deviation). First we studied a linear output control problem and examined the general properties of the transfer functions of the system and the controller. We then studied the integral controller for the linear output regulation and formulated sufficient condition $CA^2 > 0$ for the convergence of any positive integral controller, showing that the output stabilization is achieved when the transfer function of the system is

SPR and giving in addition an example showing the conservatism of this condition. If the system satisfies this condition, the parameters of the controller can be chosen arbitrarily, and there is no need to have knowledge of the state vector or of the values of the elements of the A matrix. We extended our analysis to multi-output systems for the purpose of controlling average states of several clusters and derived a sufficient condition on the system matrices for the multi-output system to be SPR.

Control of the average state does not mean that the individual system states will be close to the average state. Therefore, in addition to controlling the average it is worth to minimize the deviation of the system states. To solve this problem we used the extremum seeking algorithm augmented with the primal-dual method for the constrained minimization. The stability of this scheme was proven and its performance, as well as that of several modified versions, was tested in the numerical simulations.

To conclude, we would like to further discuss the scope of application of this work. In the introduction, we argued that average and deviation can be directly measured in several practical examples. In an even broader range of cases, however, average and deviation can be estimated through sampling some nodes and constructing suitable observers, as recently illustrated by Niazi, Canudas-de-Wit, and Kibangou 2020a; Niazi, Canudas-de-Wit, and Kibangou 2020b: the inclusion of such observers in our control scheme should be a topic of future work.

Chapter 3

Shape-based model reduction for conservation laws

Contents

3.1 Introduction

Mathematical models of large physical systems can be described in various ways, for example, partial differential equations, conservation laws, or networks. Regardless of the way it is modelled, the problem of controlling state of the entire system is usually highly complex. In the previous chapter we presented one particular example of an aggregated characteristics control, where the average state of the system was stabilized to a desired value and the standard deviation of states was minimized. However, general methods for controlling higher moments of the system require solving the problem of moments closure (Kuehn 2016). In a particular case, this problem is solved if the system is homogeneous, that is, if the evolution

equation for each state equally depends on other states (Zhang et al. 2021), however this solution can be applied only to specific systems.

Other aggregated characteristics of the state of the system may be parameters that describe the spatial properties of the solution. For example, if there is a clear peak in the solution, it would be desirable to be able to describe the dynamics of the position and size of the peak. Often when describing the state of the system it is enough to know a simplified shape of the solution described by several parameters. Moreover, depending on various tasks, various basic shapes may be assumed. In this chapter we will present a model reduction method for conservation laws, where the model is reduced to the dynamics of user-defined aggregated characteristics that describe the simplified shape of the solution. Conservation laws are an important class of systems as they can describe various real processes. For example, road traffic is often modelled by LWR model (Lighthill and Whitham 1955; Richards 1956), which is a hyperbolic conservation law, and heat distribution is modelled by the parabolic heat equation.

For the model reduction of PDEs, the Galerkin approximation (see Li and Qi 2010) is often used. In this method, the solution is projected onto a set of basis functions, then a finite subset of these functions is selected, and then the final ODE system for projection gains is constructed. For a recent work on controlling PDEs using the Galerkin method and B-splines see Tol, Visser, and Kotsonis 2019. The Galerkin method is applicable also to nonlinear systems, however, the process of model reduction itself is linear. The state vector of the obtained ODE system in general does not have any clear physical meaning, and its dimension often turns out to be very large to describe the solution. Many methods have been proposed to refine the solution and find good basis functions, see for example Baker and Christofides 2000 and Barrault et al. 2004. Hyperbolic conservation laws can create shocks and discontinuities in finite time, so conventional projective methods do not work. For their approximation, discontinuous Galerkin methods (Cockburn, Karniadakis, and Shu 2012) were developed, which can be easily parallelized for the efficient computation. Nevertheless, the dimension of the state vector in this case is enormous.

In this chapter we propose a novel nonlinear model reduction method, in which just one function is used instead of a set of basis functions. This function describes the form of the solution depending on several parameters. The dynamics of the system turns into the dynamics of the shape parameters. The resulting system can be used for estimation and control tasks. We have also shown that the model reduction process minimizes the derivative of the Wasserstein distance (Villani 2009) between the original and reduced systems.

In Section 3.2 the general derivation of the reduced model is presented. It is based on an optimal projection of the system's flow on the desired shape via least squares. Further, the analysis of the Wasserstein distance between the original and reduced systems is given together with a behaviour of the deviation of the integral solutions. Section 3.3 discusses different shapes and presents one relevant choice of parametrization of the solution shape. Then, an example of applying our method to the reduction of the LWR traffic model is shown. Finally, Section 3.4 suggests a method for approximating boundary conditions and gives more examples based on the LWR model as well as on the parabolic heat equation.

3.2 Model reduction

3.2.1 Problem formulation

Let the original system be a one-dimensional conservation law PDE described by the following model:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \phi(t,x)}{\partial x} = 0, \tag{3.1}$$

where a state of the model is a density $\rho(t,x)$. The flow is described by

$$\phi(t,x) = \phi(x, \rho(t,x), \rho_x(t,x))$$

and can depend on the position, density or its derivative. If the flow depends only on the density, (3.1) is a first-order PDE, and if a dependence on the derivative of the density is also considered, it is a second-order PDE. System (3.1) is assumed to be defined on a domain $x \in [0, L]$. Boundary conditions for (3.1) are not specified in advance (one can think of a state continued infinitely in both directions), but we will introduce specific boundary conditions in Section 3.4.

We aim to create a reduced system, which is also a conservation law:

$$\frac{\partial \hat{\rho}(t,x)}{\partial t} + \frac{\partial \hat{\phi}(t,x)}{\partial x} = 0, \tag{3.2}$$

where $\hat{\rho}(t,x)$ is an approximated density and $\hat{\phi}(t,x)$ is an approximated flow. At each time we set the approximated density to have a form $\hat{\rho}(t,x) = g(x, \theta(t))$, where $g(x,\theta)$ is a function which describes the desired shape based on m parameters $\theta \in \mathbb{R}^m$. This function is assumed to be Lipschitz continuous in both x and θ. We will discuss shape functions in more details in Section 3.3, but for now as an example one can imagine $g(x,\theta)$ being a Gaussian kernel with $\theta = (\mu, \sigma)$, where μ is a position of the peak and σ is a standard deviation. The parameters θ will constitute the state of the reduced system. Our goal is to find an evolution of θ such that (3.2) approximates the original system (3.1) as close as possible.

3.2.2 Formal solution

Therefore, we assume that there exists some ODE system which drives the dynamics of the parameters θ:

$$\dot{\theta} = F(\theta). \tag{3.3}$$

From $\hat{\rho}(t,x) = g(x, \theta(t))$ it is possible to write a time evolution equation for the approximated density by the chain rule:

$$\frac{\partial \hat{\rho}(t,x)}{\partial t} = \frac{\partial g(x,\theta)}{\partial \theta} F(\theta). \tag{3.4}$$

Here we denote by $\partial g(x,\theta)/\partial\theta$ the *generalized derivative*, which is bounded due to the Lipschitz continuity of $g(x,\theta)$. We can imagine that change of the density (3.4) was caused by

some flow $\hat{\phi}(t,x)$ which we call an approximated flow. To satisfy (3.2), this flow should obey a conservation law:

$$\frac{\partial \hat{\phi}(t,x)}{\partial x} := -\frac{\partial \hat{\rho}(t,x)}{\partial t}, \tag{3.5}$$

which we will use as a definition for $\hat{\phi}(t,x)$. Taking an integral and substituting (3.4), we obtain

$$\hat{\phi}(t,x) := \hat{\phi}_0(t) - \left(\int_0^x \frac{\partial g(s,\theta)}{\partial \theta}\, ds \right) F(\theta), \tag{3.6}$$

where $\hat{\phi}_0(t)$ is an integration constant and does not affect the dynamics. Finally, assume we fix an initial time point t_0 and we set the density of the reduced system $\hat{\rho}(t_0,x)$ to be equal to the density $\rho(t_0,x)$ of the original system.

Now we are ready to define the model reduction procedure. If both conservation laws start from the same initial condition, the natural way to minimize the difference between them is to minimize the L_p-difference between the flows. In particular, in Section 3.2.3 it is shown that minimization of difference between flows in L_1 norm coincides with the Wassershein distance derivative minimization. However, for computational purposes we prefer to choose L_2 norm, which leads to the least squares minimization. This particular choice appears to be related to the minimization of L_2 norm of integral solutions' deviations as it is shown in Section 3.2.4. Performing least squares minimization, the dynamics $F(\theta)$ for the reduced system can be found as

$$F(\theta) = \operatorname{argmin}_{f\in\mathbb{R}^m} \left(\min_{\hat{\phi}_0\in\mathbb{R}} J(f,\hat{\phi}_0) \right), \tag{3.7}$$

where

$$J(f,\hat{\phi}_0) = \int_0^L \left| \hat{\phi}(t_0,x) - \phi(t_0,x) \right|^2 dx = \int_0^L \left| \hat{\phi}_0 - \left(\int_0^x \frac{\partial g(s,\theta(t_0))}{\partial \theta} ds \right) f - \phi(t_0,x) \right|^2 dx. \tag{3.8}$$

The minimization parameters are $f \in \mathbb{R}^m$ and $\hat{\phi}_0 \in \mathbb{R}$, where the first one is a value of $F(\theta)$ at the moment of optimization $f = F(\theta(t_0))$, and the second one is an additional parameter $\hat{\phi}_0 = \hat{\phi}_0(t_0)$ which is used to define approximated flow in (3.6) but is redundant for the dynamics of θ in (3.3).

The flow $\phi(t,x)$ in general depends on $\rho(t,x)$, but by our assumption at time point t_0 the density of the original system is the same as the approximated density, thus

$$\phi(t_0,x) = \phi(x, \hat{\rho}(t_0,x), \hat{\rho}_x(t_0,x)) = \phi(x, g(x,\theta), g_x(x,\theta)). \tag{3.9}$$

We further define a vector of decision variables $\xi = (f^T, \hat{\phi}_0)^T$, a function $h : [0,L] \times \mathbb{R}^m \to \mathbb{R}^{1\times(m+1)}$

$$h(x,\theta) = \left(-\int_0^x \frac{\partial g(s,\theta)}{\partial \theta} ds\ , 1 \right), \tag{3.10}$$

and then two new functions based on $h(x,\theta)$:

$$H(\theta) = \int_0^L h(x,\theta)^T h(x,\theta)\, dx, \tag{3.11}$$

$$\psi(\theta) = \int_0^L h(x,\theta)^T \phi(x, g(x,\theta), g_x(x,\theta))\, dx. \tag{3.12}$$

With this notation the cost functional (3.8) can be written as

$$J(\xi) = \xi^T H(\theta)\xi - 2\xi^T \psi(\theta) + const, \tag{3.13}$$

and its minimization is performed by setting $\frac{\partial J}{\partial \xi} = 0$. Minimization of the quadratic function is achieved by solving a linear equation $H(\theta)\xi = \psi(\theta)$, and its solution is just $\xi = H(\theta)^{-1}\psi(\theta)$.

Finally, we are interested in the first m components of the decision vector ξ, therefore the optimal dynamics for the reduced system is

$$\dot{\theta} = F(\theta) = \left[H(\theta)^{-1}\psi(\theta)\right]_{1,\dots,m}. \tag{3.14}$$

Note that knowing the flow $\phi(x, \rho, \rho_x)$ and class of functions $g(x, \theta)$ one can compute $H(\theta)$ and $\psi(\theta)$ symbolically, thus obtaining a closed-form solution to the problem. Moreover, the matrix $H(\theta)$ depends only on the parametrization $g(x, \theta)$ and not on the particular flow $\phi(x, \rho, \rho_x)$, therefore it is necessary to symbolically compute it (and its inverse) only once for each chosen parametrization.

3.2.3 Relation to Wasserstein distance minimization

It is possible to show that the minimization of flow discrepancy leads to the minimization of Wasserstein distance divergence between real and reduced solution. The L_p-Wasserstein distance between two nonnegative densities $\rho^0(x)$ and $\rho^1(x)$ of equal mass on the domain $x \in [0, L]$ is defined as

$$W_p(\rho^0, \rho^1) = \min_{T \in \mathcal{T}} \left(\int_0^L |T(x) - x|^p \, \rho^0(x)\, dx \right)^{1/p}, \tag{3.15}$$

where $\mathcal{T}$ is the set of all possible transformations over the domain $[0, L]$ that transfer the mass from one configuration to another. In other words, for any $x \in [0, L]$ the position defined by $T(x)$ means that the mass $\rho^1(x)$ consolidates the mass $\rho^0(T(x))$. More precisely,

$$\mathcal{T} := \left\{ T : [0,L] \to [0,L] \;\middle|\; \int_a^b \rho^1(x)\, dx = \int_{T(a)}^{T(b)} \rho^0(x)\, dx \quad \forall a, b \in [0,L] \right\}. \tag{3.16}$$

We will show that the L_1-minimization of flows is equivalent to the minimization of the time derivative of L_1-Wasserstein distance.

Assume at some time moment t_0 the state of the original system $\rho(t_0, x)$ and the reconstructed state of the reduced system $\hat{\rho}(t_0, x)$ are equal. Then the L_1-Wasserstein distance is

zero and the transformation T which achieves minimum is identity, $T(x) = x$. Equivalently we can define time-dependent transformation $T(t, x)$ between two densities, and since they coincide at t_0, we have $T(t_0, x) = x$. In the same way the time-dependent Wasserstein distance can be defined. Now take the time derivative of the L_1-Wasserstein distance (3.15) for this particular transformation:

$$\dot{W}_1(\rho, \hat{\rho}, t_0) = \int_0^L |\dot{T}(t_0, x)|\rho(t_0, x)\, dx = \int_0^L |\dot{T}(t_0, x)\rho(t_0, x)|\, dx, \tag{3.17}$$

where we used the fact that $\rho(t_0, x) \geqslant 0$.

Using the definition (3.16) of the transformation $T \in \mathcal{T}$ and taking its time derivative by the Leibniz rule, we get

$$\int_a^b \frac{\partial \hat{\rho}(t_0, x)}{\partial t} dx = \int_{T(t_0,a)}^{T(t_0,b)} \frac{\partial \rho(t_0, x)}{\partial t} dx + \dot{T}(t_0, a)\rho(t_0, a) - \dot{T}(t_0, b)\rho(t_0, b). \tag{3.18}$$

Both $\rho(t, x)$ and $\hat{\rho}(t, x)$ obey the conservation laws (3.1) and (3.2) with the flows $\phi(t, x)$ and $\hat{\phi}(t, x)$ respectively. Therefore (3.18) can be rewritten as

$$\begin{aligned} \hat{\phi}(t_0, a) - \hat{\phi}(t_0, b) = \phi(t_0, T(t_0, a)) - \phi(t_0, T(t_0, b)) + \\ + \dot{T}(t_0, a)\rho(t_0, a) - \dot{T}(t_0, b)\rho(t_0, b). \end{aligned} \tag{3.19}$$

This condition should be satisfied for all $a, b \in [0, L]$, therefore

$$\hat{\phi}(t_0, x) = \phi(t_0, x) + \dot{T}(t_0, x)\rho(t_0, x), \tag{3.20}$$

where we also used that $T(t_0, x) = x$ for all x. Finally, substituting this into (3.17) we obtain

$$\dot{W}_1(\rho, \hat{\rho}, t_0) = \int_0^L |\hat{\phi}(t_0, x) - \phi(t_0, x)|\, dx, \tag{3.21}$$

which is minimized exactly by the L_1-minimization of the flows discrepancy.

3.2.4 Relation to deviations between integral solutions

Let us introduce a special function $M(t, x)$ called *Moskowitz function*, which is an integral solution to the conservation law (3.1) and which has a definition:

$$\frac{\partial M(t, x)}{\partial t} = \phi(t, x), \qquad \frac{\partial M(t, x)}{\partial x} = -\rho(t, x). \tag{3.22}$$

It follows that the system (3.1) is just an equality of the second mutual derivatives of $M(x, t)$. Choosing $M(L, 0) = 0$, we can write the integral form as

$$M(t, x) = \int_0^L \rho(0, s)ds + \int_0^t \phi(\tau, 0)d\tau - \int_0^x \rho(t, s)ds. \tag{3.23}$$

If the system's state represents density (mass in a unit length), then $M(t,x)$ can be seen as an overall mass which was transfered through the point x up to the time t. In particular, in traffic modeling this function corresponds to the number of vehicles passed through a fixed point, see Newell 1993. Now define

$$J^*(\theta) = \min_{f \in \mathbb{R}^m} \left(\min_{\hat{\phi}_0 \in \mathbb{R}} J(f, \hat{\phi}_0) \right), \tag{3.24}$$

which is a minimal achievable value of the cost functional (3.8). Then, introducing a Moskowitz function $\hat{M}(t,x)$ for the reduced system (3.2) and using the minimization of (3.8) in Section 3.2.2, one can see from the definition (3.22) that

$$\left\| \dot{\hat{M}}(t,\cdot) - \dot{M}(t,\cdot) \right\|_2^2 = J^*(\theta(t)), \tag{3.25}$$

because the norm of difference of time derivatives of Moskowitz functions is exactly the norm of difference of flows. Therefore equation (3.25) provides an additional interpretation of the result in Section 3.2.2 as a minimization of integral solutions' deviations.

3.2.5 Equilibrium points

It appears that the model reduction procedure defined in Section 3.2.2 preserves equilibrium solutions while transforming the original system to the reduced one. In particular, it is possible to show that any equilibrium solution to the original system (3.1), which can be exactly reconstructed via chosen parametrization, is by itself an equilibrium solution to the reduced system (3.14):

Theorem 3.1. *Let $\rho^*(x)$ be an equilibrium solution to (3.1), and assume there exists θ^* such that $\rho^*(x) = g(x,\theta^*)$ for all $x \in [0,L]$. Then θ^* is an equilibrium point for the reduced system (3.14).*

Proof. First of all, $\rho(t,x) \equiv \rho^*(x)$ means that $\rho(t,x)$ does not depend on time, or $\partial\rho/\partial t = 0$, which by (3.1) leads to $\partial\phi/\partial x = 0$ in the equilibrium case. Therefore the flow should be constant in space: $\phi(t,x) \equiv \phi^*(t)$ for some $\phi^*(t)$. Now it becomes clear that the functional $J(f,\hat{\phi}_0)$ in (3.8) can be minimized to zero by setting $f = 0$ and $\hat{\phi}_0 = \phi^*(t)$ at each moment t, which by (3.7) means that the dynamics of the reduced system satisfy $F(\theta^*) = 0$. Therefore, θ^* is itself an equilibrium point. □

Note that the opposite property does not hold in general: if $F(\theta^*) = 0$ for some θ^*, the reduced system will be in equilibrium, however the original system can still evolve "orthogonally" if the flow $\phi(t,x)$ vanishes being projected on a subspace formed by $\partial g(x,\theta^*)/\partial\theta$.

3.3 Shape parametrization

The most important question which arises while designing the reduced system is the choice of the class of reduced solutions $g(x, \theta)$. One possible solution which is known as Galerkin projection is to take a countable set of basis functions $\psi_i(x)$ for $i \in \{1, ..., m\}$ and to define θ to be their multipliers:

$$g(x, \theta) = \sum_{i=1}^{m} \theta_i \psi_i(x). \tag{3.26}$$

Popular examples for the basis functions $\psi_i(x)$ are the set of all polynomials or the set of all harmonic functions. However this leads to a large number of parameters θ which need to be maintained, especially when the density profile cannot be easily described as a finite sum of basis functions.

Galerkin projection (3.26) is linear in θ, which explains its popularity. However the method in Section 3.2.2 is designed with nonlinear dependence of shapes $g(x, \theta)$ on θ in mind. Therefore what we suggest is to find a parametrization for each particular case. One of such cases when the traditional approach struggles is in describing densities with single peak or spike, since they require a large number of basis functions. Instead of Galerkin projection (3.26) one could use a shape specially designed for single peak functions. One obvious choice of shape in this case would be a Gaussian-type function:

$$g(x, \theta) = \gamma e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.27}$$

which has three parameters $\theta = (\gamma, \mu, \sigma)$, describing the height, the position and the spread of the peak respectively. It is possible to derive a reduced system (3.14) for the parametrization (3.27), however it appears that since $g(x, \theta)$ is required to be only Lipschitz continuous, we can use a more simple class of piecewise-linear functions, which at the same time has much more degrees of freedom.

3.3.1 A piecewise-linear approximation

Let $\theta = (\gamma, \mu, k_1, k_2, c_1, c_2)$, where the meaning of the parameters is as follows: γ is the height of the peak, μ is the position of the peak, k_1 is the slope to the left, k_2 is the slope to the right, c_1 is the constant level to the left, c_2 is the constant level to the right. We can define a piecewise-linear function $g(x, \theta)$ as follows:

$$g(x, \theta) = \begin{cases} c_1, & \text{if } x < p_1, \\ \gamma + k_1(x - \mu), & \text{if } p_1 \leqslant x < \mu, \\ \gamma + k_2(x - \mu), & \text{if } \mu \leqslant x < p_2, \\ c_2, & \text{if } p_2 \leqslant x, \end{cases} \tag{3.28}$$

where $p_1 = \mu - (\gamma - c_1)/k_1$ and $p_2 = \mu - (\gamma - c_2)/k_2$. This parametrization is depicted in Fig. 3.1.

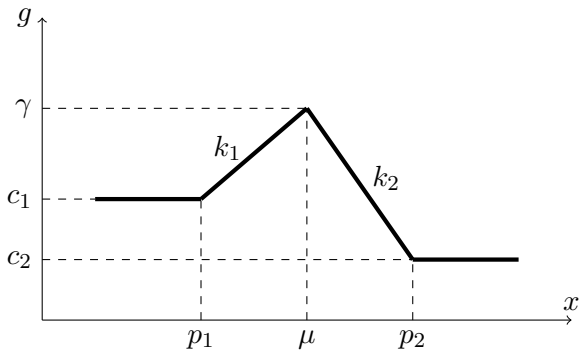


Figure 3.1: The parametrization of the piecewise-linear peak functions

Advantage of the parametrization (3.28) over the Gaussian-type function (3.27) is in the simplicity of computing $h(x, \theta)$ in (3.10) and further $H(\theta)$ in (3.11), which can be done analytically and renders polynomial functions. Moreover, piecewise-linear parametrization (3.28) can have many different degrees of freedom (in particular here we set its number to 6), while (3.27) has only three, and adding additional degrees of freedom would require using other types of complicated nonlinear functions. Finally, in theory it is possible to use higher-order polynomials and define spline-type approximations instead of (3.28), which would still be computationally feasible since all integrals in (3.10) and (3.11) would be still analytic.

Note that contrary to the Galerkin projections such parametrizations can lead to the situation when $\det H(\theta) = 0$, therefore the system (3.14) can no longer be solved. This happens for example in (3.28) when $c_1 = \gamma$. It is clear that in this case the shape becomes degenerate, the parameters become dependent and it is no longer possible to resolve which one of them should be varied to give the smallest flow discrepancy. Thus the system works as long as it preserves its shape, which is rather expected if one thinks that the particular class of functions was chosen based on the assumed shape of the real density.

3.3.2 Application to LWR system

We will show the capabilities of our method on the example of the LWR system:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \phi(\rho(t,x))}{\partial x} = 0, \tag{3.29}$$

where

$$\phi(\rho(t,x)) = v_{max}\left(1 - \frac{\rho(t,x)}{\rho_{max}}\right)\rho(t,x).$$

Such choice of the flow corresponds to the Greenshields fundamental diagram (Greenshields et al. 1935). This system models the flow of cars on a highway in assumption that the velocity of each car decreases linearly with the density of vehicles nearby. More comprehensive description of LWR system will be presented in Section 5.2. Here we set the length of the road to $L = 1000$m with $\rho_{max} = 0.181$ veh/m (one vehicle per 5.5m) and $v_{max} = 60$ km/h.

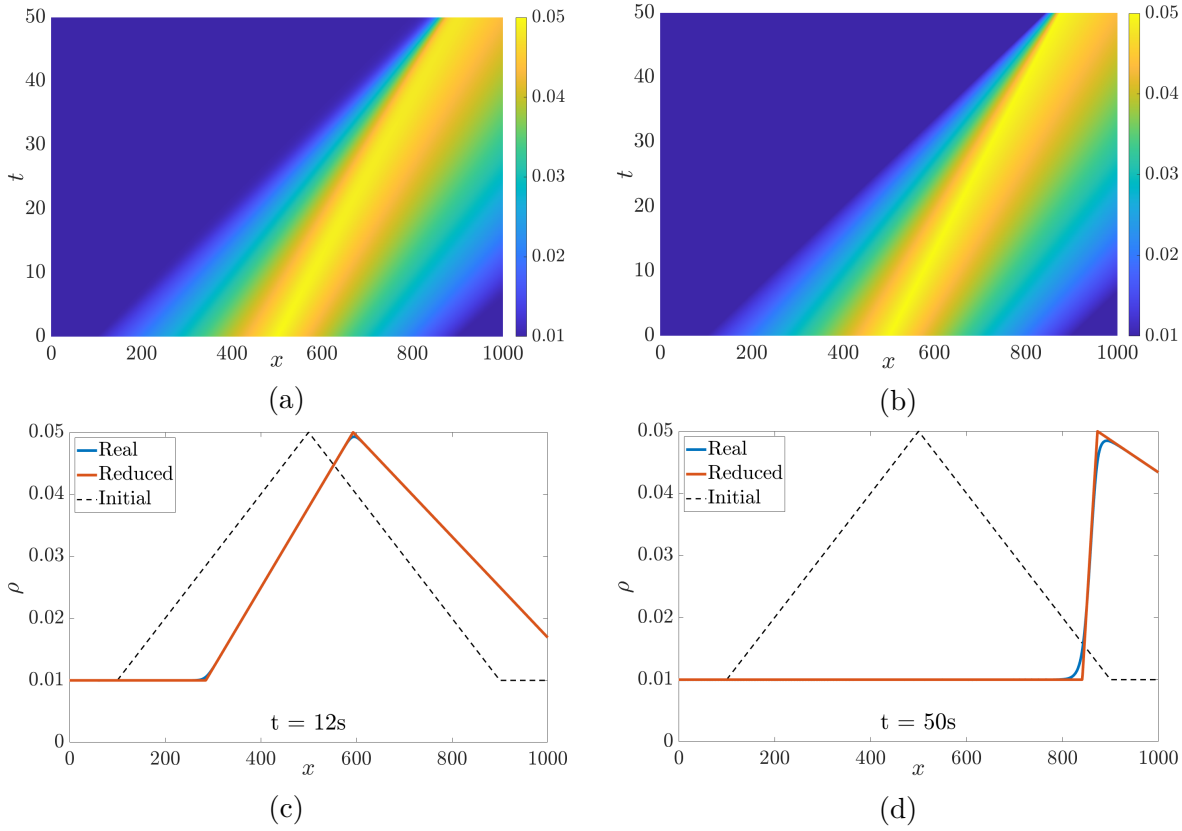


Figure 3.2: Comparison of real and reduced solutions to the LWR equation (3.29). **(a):** Evolution of real density $\rho(t, x)$. **(b):** Evolution of approximated density $\hat{\rho}(t, x)$. **(c):** Comparison of densities at $t = 12$s. Real density $\rho(t, x)$ is shown with blue line, approximated density $\hat{\rho}(t, x)$ with red line and initial density $\rho_0(x)$ with black dashed line. **(d):** Comparison of densities at $t = 50$s.

The results of the comparison are shown in Fig. 3.2. We simulated $T = 50$ seconds of both systems' behaviours. The original system (3.29) was numerically solved using Godunov method (Godunov 1959). Matrix $H(\theta)$ and vector $\psi(\theta)$ were symbolically computed using MATLAB Symbolic Toolbox, and the system (3.14) was numerically solved using Euler method. We used a space grid with 200 cells for the original system simulation, and the number of time steps was 500 for both systems. We calculated the time needed to simulate both systems: on average, simulation of the original system took 0.207241 seconds, and simulation of the reduced system took 0.027709 seconds, thus being almost 10 times faster. We believe that by designing a specialized software instead of using a general toolbox one can achieve much higher performance.

It is clear that the reduced solution perfectly tracks the position and the slopes of the peak, and the difference between the real density and the approximated density arises only because of the non-smoothness of the reduced solution. It is also interesting to note that if one continues simulation further, the reduced system will fail at time moment $t = 54$s,

because the shock arises on the left slope and k_1 becomes infinity. It is a known property of the LWR system which can produce shocks in a finite time. In our assumption $g(x, \theta)$ should be continuous both in x and θ for the correct definition of the artificial flow (3.6). Using a different kind of discontinuous parametrization it is possible to reduce the LWR system to a system similar to wave-front tracking algorithm, see Baiti and Jenssen 1998.

3.4 Boundary problems

3.4.1 Formal solution

Up to now not a word was said about boundary conditions which affect the solution to the original system and which should be taken into account properly in the reduced system. Essentially all the analysis performed in the previous sections was based on the assumption that the solutions evolve in $\mathbb{R}$, with only exception being the cost functional (3.8). Therefore boundary conditions were not taken into account either in original or reduced systems.

Now assume that the original system is given by the equation (3.1) defined on the domain $[0, L]$. Let one of the boundary flows $\phi(t, 0) = \phi_{in}(t)$ or $\phi(t, L) = \phi_{out}(t)$ (or possibly both of them) be given. The flows can be either given explicitly as functions of time or they can depend on the state of the system itself, as if one would like to control the system via feedback.

The given boundary flows work as constraints for the flow discrepancy minimization problem. Namely, if the inflow $\phi_{in}(t)$ is given, the solution ξ to the problem of minimization of the cost functional J should satisfy the constraint

$$h(0, \theta)\xi = \phi_{in}(t), \tag{3.30}$$

and similarly for outflow in case $\phi_{out}(t)$ is given.

The constraints can be written in a unified manner if one defines matrix $C(\theta)$ and column-vector d such that they have one row if only one condition is given and two rows if both boundary conditions are set. Consider both boundary conditions are set, then we define

$$C(\theta) := \begin{pmatrix} h(0, \theta) \\ h(L, \theta) \end{pmatrix}, \qquad d := \begin{pmatrix} \phi_{in}(t) \\ \phi_{out}(t) \end{pmatrix}.$$

Then the constrained minimization of J is equivalent to the minimization of the Lagrangian

$$L = \xi^T H(\theta)\xi - 2\xi^T \psi(\theta) + 2\lambda^T (C(\theta)\xi - d), \tag{3.31}$$

with $\lambda \in \mathbb{R}^2$ being a vector of Lagrange multipliers. Solution to this problem is given by

$$\begin{aligned} \lambda &= \left(CH^{-1}C^T\right)^{-1} \left(CH^{-1}\psi - d\right), \\ \xi &= H^{-1}\left(\psi - C^T\lambda\right), \end{aligned} \tag{3.32}$$

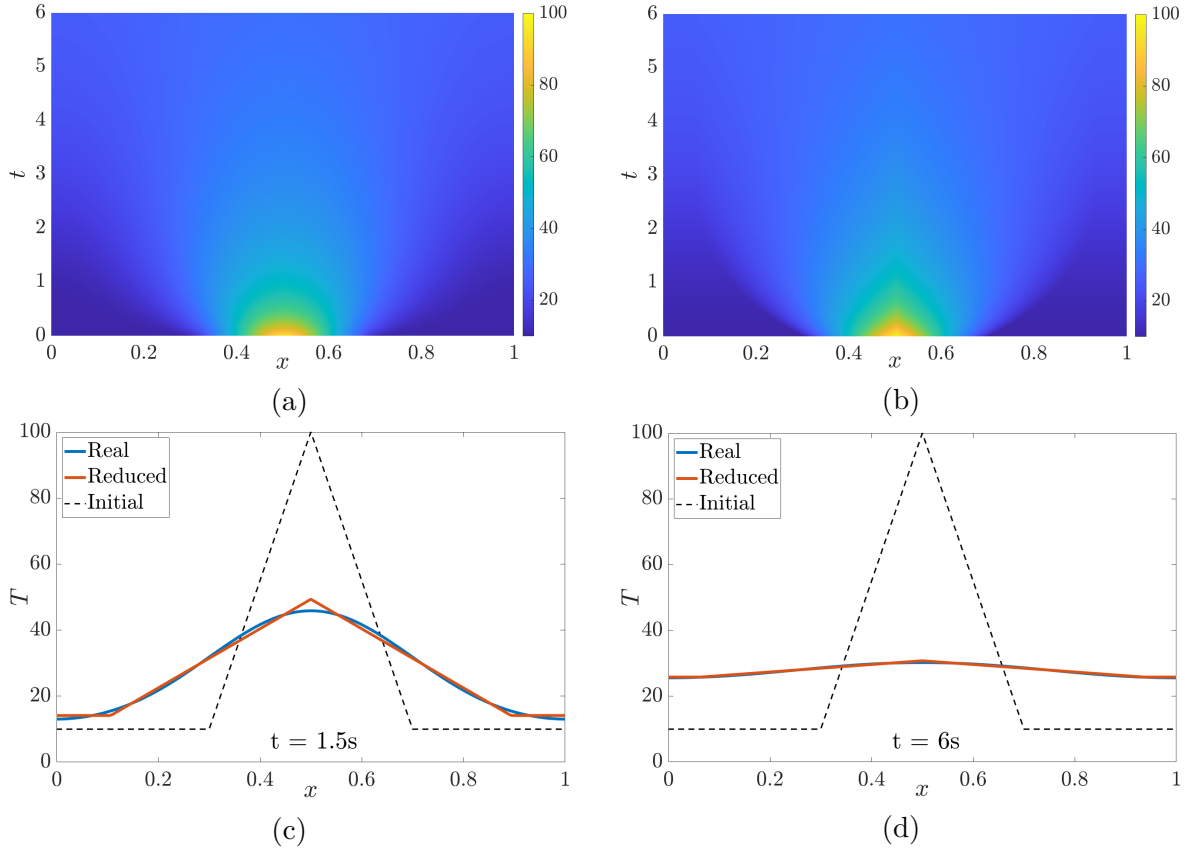


Figure 3.3: Comparison of real and reduced solutions to the heat equation (3.34) for the temperature of a copper rod in Celsius assuming the system is closed: $\phi_{in} = \phi_{out} = 0\,\mathrm{K{\cdot}m/s}$. **(a):** Evolution of real temperature $\rho(t, x)$. **(b):** Evolution of approximated temperature $\hat{\rho}(t, x)$. **(c):** Comparison of temperatures at $t = 1.5$s. Real temperature $\rho(t, x)$ is shown with blue line, approximated temperature $\hat{\rho}(t, x)$ with red line and initial temperature $\rho_0(x)$ with black dashed line. **(d):** Comparison of temperature at $t = 6$s.

where we omit dependencies on θ for simplicity of writing. From ξ the dynamics of θ can be easily recovered by discarding the last row:

$$\dot{\theta} = \left[H^{-1}\left(\psi - C^T\lambda\right)\right]_{1\ldots m}. \tag{3.33}$$

It is interesting to node that by putting the constraints on both boundaries one guaranties the conservation of mass, therefore the overall mass in the original and the reduced system are always equal.

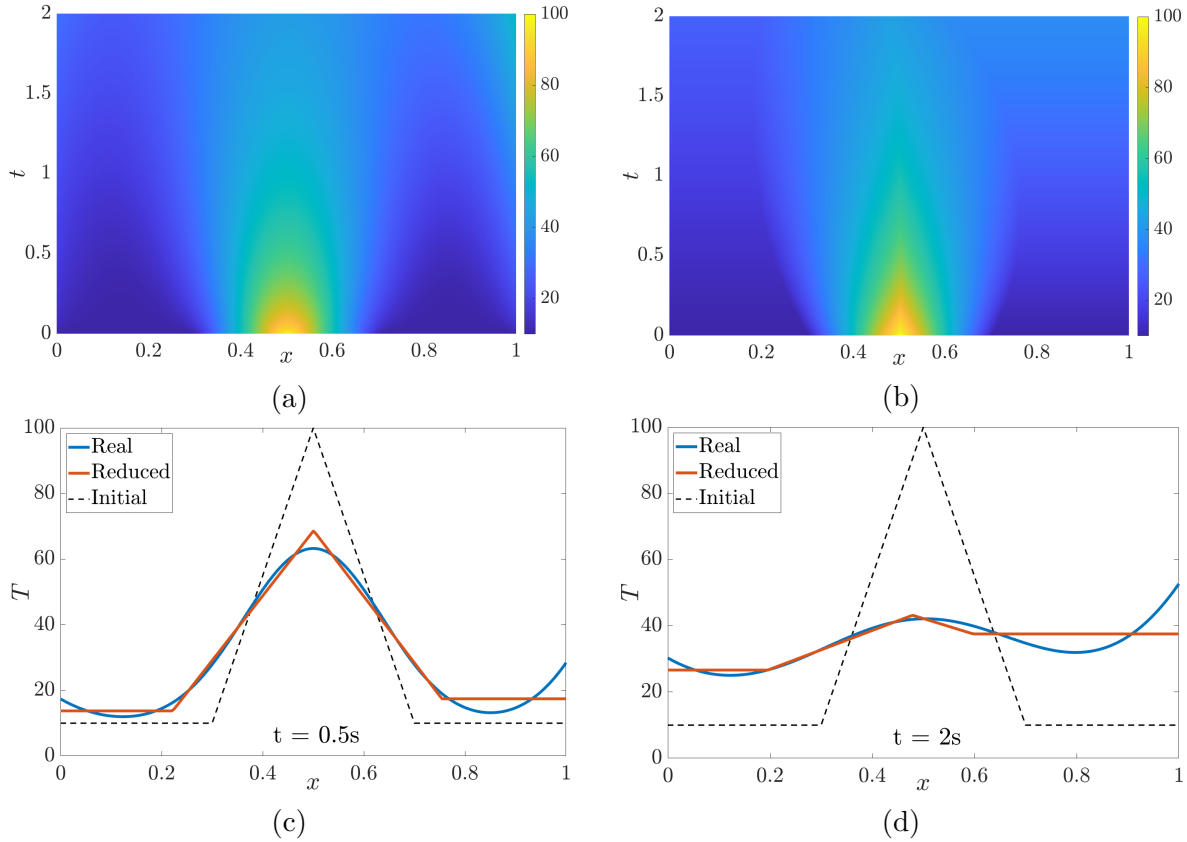


Figure 3.4: Comparison of real and reduced solutions to the heat equation (3.34) for the temperature of a copper rod in Celsius in an externally heated system: $\phi_{in} = 1$ K·m/s and $\phi_{out} = -2.5$ K·m/s. **(a):** Evolution of real temperature $\rho(t, x)$. **(b):** Evolution of approximated temperature $\hat{\rho}(t, x)$. **(c):** Comparison of temperature at $t = 0.5$s. Real temperature $\rho(t, x)$ is shown with blue line, approximated temperature $\hat{\rho}(t, x)$ with red line and initial temperature $\rho_0(x)$ with black dashed line. **(d):** Comparison of temperature at $t = 2$s.

3.4.2 Application to the heat equation

We can demonstrate how the method works for a bounded system by considering the heat equation:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \phi(\rho(t,x))}{\partial x} = 0, \tag{3.34}$$

where $\phi(\rho(t, x)) = -\alpha \rho_x(t, x)$, thus (3.34) is a linear second-order parabolic PDE. Let system (3.34) represent a heated rod of copper of a length $L = 1$m. The state $\rho(t, x)$ denotes temperature of the rod in Celsius at the point $x \in [0, L]$ at time t. Parameter $\alpha = 0.0111\text{m}^2/\text{s}$ is a thermal conductivity of copper. We consider approximation of the copper rod's temperature using parametrization (3.28). The simulations were performed with the same space and time discretization as in the previous section. Initial conditions were set in such way that the center of the rod was heated to 100° Celsius, while the ends were kept at 10° Celsius.

Assuming that there is no heat transfer between the domain and the environment, we set the boundary conditions $\phi_{in} = \phi_{out} = 0$ (measured as Kelvin·meter/second). The results are presented in Fig. 3.3. We can see that the system comes close to the thermal equilibrium in just 6 seconds, and the final temperature is around 28° Celsius. Further, it is clear that the reduced model approximates the original one almost perfectly.

Alternatively, we can set non-zero boundary conditions, heating the rod at the ends. Let us introduce heat flows $\phi_{in} = 1\,\mathrm{K{\cdot}m/s}$ and $\phi_{out} = -2.5\,\mathrm{K{\cdot}m/s}$. Note that the negative value of ϕ_{out} corresponds to the flow propagating "backwards", which for the boundary at $x = L$ means the flow is going into the system. The results are shown in Fig. 3.4. The rod is heated very fast. It is clear that although the reduced system cannot capture the shape of the real solution near the right boundary at time $t = 2$s, it still averages it in terms of the chosen shape. In this setup at $t = 2.15$s the shape becomes degenerate as c_2 reaches γ.

3.4.3 Application to LWR system

Time-dependent control over only one boundary can be demonstrated on the LWR example. Using the model (3.29) and the same parameters as in Section 3.3.2, assume that we set an additional inflow $\phi_{in}(t) = 0.2 + 0.15\sin(t/2)$ vehicles per second. Results of the numerical simulation are presented in Fig. 3.5. It appears that due to the shape limitations the reduced system averages the high frequency components, while precisely tracking the initial peak.

3.5 Concluding remarks

In this chapter we presented a method of describing 1D conservation law system based on notion of solution shape. We reduced system state to a set of well-tractable shape parameters and derived their dynamics, providing closed-form solution. We further analyzed its properties, showing in particular that this solution minimizes Wasserstein distance between the original and the reduced system and that the equilibrium points of the original system are preserved. The idea of representing system's solution by specific shape parameters can potentially lead to new types of control design based on the aggregated characteristics of the system. There are thought some open problems that should be considered:

- It was shown in Sections 3.3.2 and 3.4.2 that solutions to the reduced system can become degenerate once the chosen shape is violated. To fight with degeneracy of solutions it is interesting to develop a methods for reparametrizations and changes of shape, as well as to allow discontinuous shape functions.

- Since the reduced system is an approximation to the original one, it would be desirable to find bounds on a difference between systems' solutions to guarantee the performance and help in choosing the particular shape function.

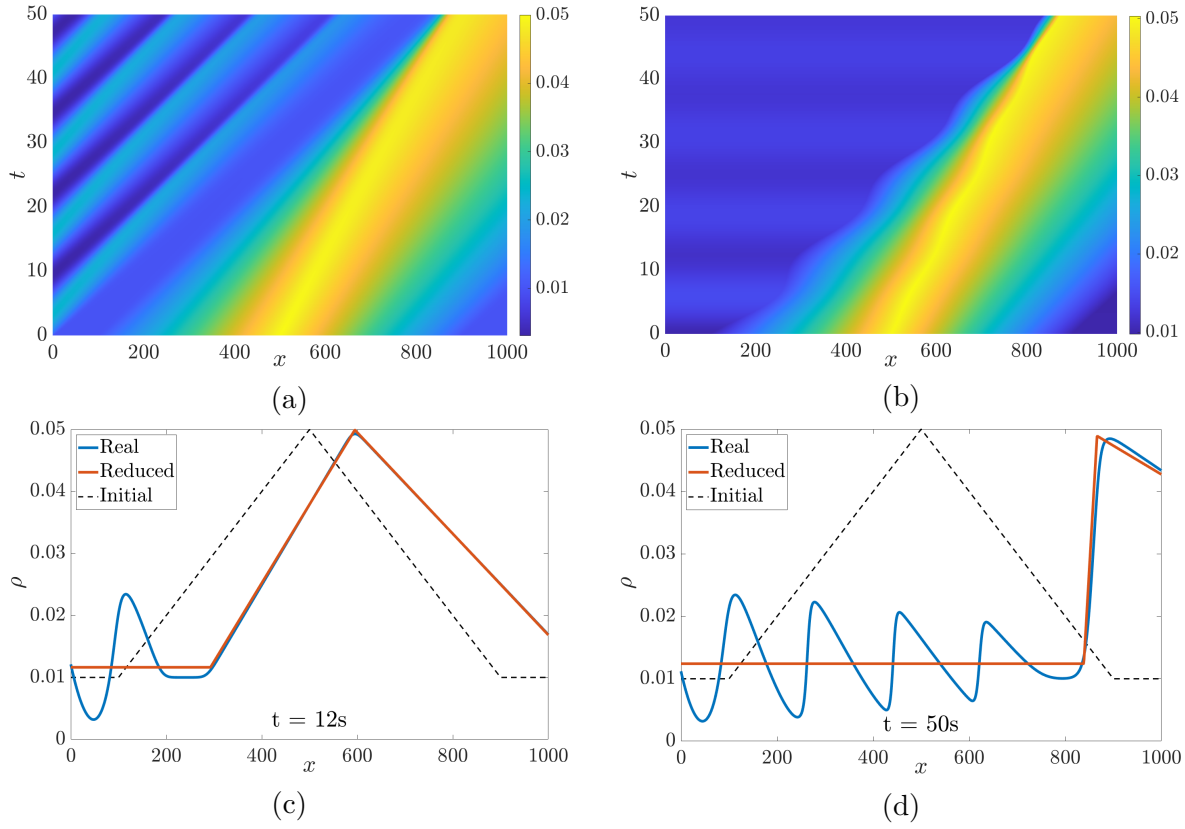


Figure 3.5: Comparison of real and reduced solutions to the LWR equation (3.29) with $\phi_{in}(t) = 0.2 + 0.15\sin(t/2)$ vehicles per second. **(a):** Evolution of real density $\rho(t,x)$. **(b):** Evolution of approximated density $\hat{\rho}(t,x)$. **(c):** Comparison of densities at $t = 12$s. Real density $\rho(t,x)$ is shown with blue line, approximated density $\hat{\rho}(t,x)$ with red line and initial density $\rho_0(x)$ with black dashed line. **(d):** Comparison of densities at $t = 50$s.

- Here we presented the method of the shape-based model reduction for 1D conservation laws. The method could be generalized to include more classes of systems, starting from multidimensional conservation laws and adding then general PDE models. Moreover, it is straightforward to generalize the idea of shape functions to systems with finite-dimensional state space, therefore including ODE networks in scope of the method.

- In Section 3.4.3 periodic inflow was applied to LWR system modeling traffic. However LWR model cannot be controlled directly, since the inflow which enters the system is equal to the minimum between demand on the boundary, which can be directly controlled, and supply in the system, which depends directly on the state of the system itself. Therefore in fact one can pose only inequality-type constraints on the in- and outflow instead of equality-type constraints discussed in Section 3.4. Generalization of the method to inequality constraints could be useful in application of the method to such systems with weak boundary conditions.

Chapter 4

Continuation method for large-scale systems modeling and control: from ODEs to PDE

Contents

4.1 Introduction

Most of the systems we encounter in real life consist of such a large number of particles that the direct analysis of their interaction is impossible. In such cases, simplified models are used that aggregate the behavior of a set of particles and replace them with a continuous representation. In general, discrete and continuous system descriptions often share a lot of common properties, which was noticed a long time ago. A common theory for discrete and continuous boundary problems was developed in Atkinson 1964, and properties of continuous wave-type oscillatory systems in a limiting case of discrete systems were derived in Gantmakher and Krein 1941. However even if the discretization procedure transforming PDEs to ODEs is a widely known and widely used method, the inverse problem of transforming an ODE system into PDE is more rarely studied.

In this chapter we focus on this particular problem, with the aim of filling this gap and providing a counterpart to the discretization procedure. This can be useful since PDEs provide a much more compact way of describing the system, which in many cases is easier to analyze analytically than the corresponding ODE system. In particular we are interested in systems which are spatially distributed and which have a position-dependent interaction, such as traffic in the city, power networks, robot formations, etc.

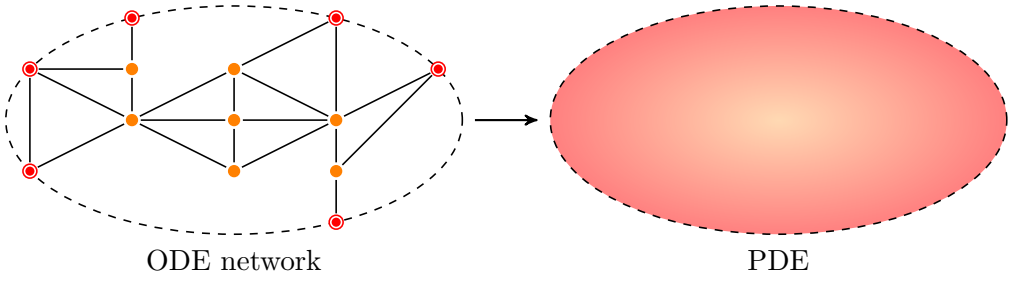


Figure 4.1: Transformation of ODE network into a single PDE model via continuation method.

Our idea is to replace the original spatially distributed ODE system by a continuous PDE whose state and space variables preserve the state and space variables of the original system, see Fig. 4.1. We develop a method for linear spatially invariant ODEs which transforms them into PDEs with the help of finite differences. We name this method as a *continuation*, since it is exactly opposite to the discretization procedure. Further we show how the continuation converges to the original system in sense of spectrum. Using computational graph formalism (see review Baydin et al. 2018) we extend the method to nonlinear systems and further to space-dependent systems and systems with boundaries. The advantage of the continuation method is that it allows to recover a PDE which describes the same physical system as the original ODE network. Such a description can be very helpful both for analysis and control purposes.

4.1.1 State of the art

The idea of replacing the system with its compact and simplified representation is widely used, especially for the ODE systems describing large-scale networks. Probably the most known approach of this type is a *clustering* technique which transforms a network into a smaller one while conserving the properties and the dynamics, see Aoki 1968; Niazi et al. 2019. Apart from various clustering techniques large-scale networks are studied by *mean field* methods in case of the *all-to-all* interaction topology. In this situation the effect of the network on each node is the same, therefore it is enough to use an equation for a single agent together with parameters of a state of the whole network, see Acebrón et al. 2005a for a review with application to Kuramoto networks. The idea of mean field can be further extended to track not only a single agent's state, but the whole probability distribution over all agents' states in the network. This method is called *population density* or *probability density* approach (Grabert 2006) and it can be used for example to model large biological neural networks (Nykamp and Tranchina 2000). Large-scale systems can be also simplified by studying the approximations to their probability densities, represented by *moments*. E.g., Yang, Dimarogonas, and Hu 2015; Zhang et al. 2021 took a moment-based approach to control crowds dynamics. A similar approach of network control via its first two moments is used as a starting point in Chapter 2 of this document. Different applications and issues of the method of moments are covered by Kuehn 2016. The idea of moment-based description of distributions is closely related to the shape-based model reduction for PDEs presented in Chapter 3.

Mean field and population density approaches are suitable in the case when the interaction topology between nodes is all-to-all. In other cases, the continuous representation of a system requires more sophisticated tools. A recently emerged theory of *graphons* studies graph limits, i.e. structural properties that the graph possesses if the number of nodes tends to infinity while preserving interaction topology. Using graphons it is possible to describe any dense graph as a linear operator in continuum space, see Lovász 2012. This method was further used to control large-scale linear networks (Gao and Caines 2019) and to study sensitivity of epidemic networks (Vizuete, Frasca, and Garin 2020). However, the resulting operator is non-local and requires the original network to have very dense connections. For example, Medvedev 2014 studied a dense network of Kuramoto oscillators using a continuous representation with integral coupling operator.

It worth noticing that by applying population density method or graphon theory to a system with position-dependent interactions, such as traffic in the city, power networks, robot formations, etc, we would end up with a continuous model which either looses spatial structure of the problem or describes it using non-local operators such as in partial-integral differential equations. Our idea is to assume predominance of local interactions and derive a single PDE describing the system. We start in Section 4.2 by defining a continuation for linear ODEs, discussing questions of accuracy, convergence and choice of the particular model. In particular it appears that the PDE approximation can capture all the effects of the original ODE provided the order of continuation (the highest spatial derivative) is high enough. Section 4.3 continues to nonlinear models, utilizing the computational graph formalism. In Section 4.4 the method

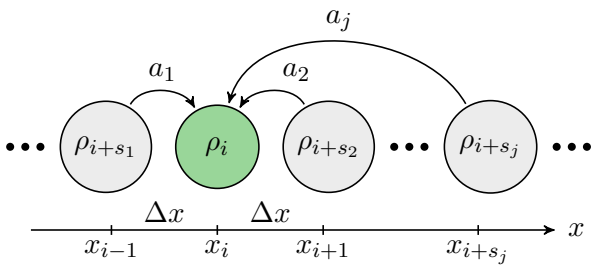


Figure 4.2: System of nodes aligned in 1D line with dynamics given by (4.1) with $s_1 = -1$ and $s_2 = 1$.

is extended to much broader class of systems, including multidimensional or space- and time-varying systems and also discussing boundary conditions. We also show that the method can be applied to multi-agent systems where positions of agents are themselves included into a state vector. This technique gives possibilities to derive density-based models for such systems, which is demonstrated further in Chapter 5. Finally, in Section 4.5 application of the method to general linear networks is covered and several particular structures are recovered. PDE representations of networks are useful for control and analysis purposes of oscillatory networks, among others, with several possible applications presented in Chapter 6.

4.2 Continuation for linear spatially invariant systems

The simplest class of systems for which the transformation of ODE into PDE can be performed is given by linear ODE systems corresponding to the dynamics of states of nodes, which are aligned on the 1D line in space and depend only on some fixed set of their neighbours. Let the node $i \in \mathbb{Z}$ have a state $\rho_i(t) \in \mathbb{R}$ and a fixed geographical position $x_i \in \mathbb{R}$ such that for every i the distance between two consecutive nodes in space is constant, $x_{i+1} - x_i = \Delta x$ (the assumption of Δx being constant will be relaxed later on). The number of nodes is assumed to be infinite. Then the systems of our interest take the form

$$\dot{\rho}_i(t) = \sum_{j=1}^{N} a_j \rho_{i+s_j}(t), \tag{4.1}$$

where $\dot{\rho}_i(t)$ denotes time derivative. That is $\dot{\rho}_i(t)$ linearly depends only on N neighbouring nodes shifted by $s_j \in \mathbb{Z}$ for $j \in \{1, ..., N\}$, and $a_j \in \mathbb{R}$ are the system gains, see Fig. 4.2. In other words, $\dot{\rho}_i(t)$ is defined as a convolution of $\rho_i(t)$ with a sequence of values a_j situated at indices $-s_j$. Thus the relation between $\dot{\rho}_i(t)$ and $\rho_i(t)$ is shift-invariant with respect to index shifts. This type of systems belongs to the class of linear *spatially invariant systems* (see e.g. Bamieh, Paganini, and Dahleh 2002), which is a natural class for distributed control. In the future we will omit writing the dependence on t whenever this is the only argument of ρ.

4.2.1 Motivating example

We start by considering the most simple ODE system of class (4.1) which has spatial dependence:

$$\dot{\rho}_i = \frac{1}{\Delta x}\left(\rho_{i+1} - \rho_i\right). \tag{4.2}$$

Comparing with (4.1), here $N = 2$, $a_1 = 1/\Delta x$, $a_2 = -1/\Delta x$, $s_1 = 1$ and $s_2 = 0$. This equation describes a transport of some quantity along the line, and is usually referred as a Transport ODE. Equation (4.2) often comes as a result of a discretization process applied to another equation,

$$\frac{\partial \rho}{\partial t}(t, x) = \frac{\partial \rho}{\partial x}(t, x). \tag{4.3}$$

This equation belongs to a class of PDEs, which is usually thought to be more difficult class of equations to study than ODEs. However, equation (4.3) describes a perfect transport of information with finite propagation speed along the line, which can be studied much more easily in PDE form than in ODE, as it perfectly conserves the form of a solution, performing only a shift along the line as time increases. We will refer to this equation as a Transport PDE.

Equation (4.2) can be obtained from (4.3) by the discretization process, which has been a well-established mathematical tool. Nevertheless, up to now there was no strict procedure describing a general process which could render equation (4.3) from (4.2). In the next subsections we explore more how the discretization procedure is defined for linear systems and how it should be inverted to obtain a continuation process.

4.2.2 Discretization

The discretization of PDEs is usually performed by a finite difference method, where the partial derivatives are approximated by finite differences. For example, in the case of Transport ODE,

$$\frac{\partial \rho}{\partial x} \approx \frac{1}{\Delta x}\left(\rho_{i+1} - \rho_i\right).$$

This approximation is valid in case when Δx is small. Indeed, assuming that the solution to PDE is given by a smooth function $\rho(x)$ and using Taylor series, we can write

$$\rho_{i+1} = \rho(x_{i+1}) = \rho(x_i) + \frac{\partial \rho}{\partial x}\Delta x + \frac{\partial^2 \rho}{\partial x^2}\frac{\Delta x^2}{2} + ..., \tag{4.4}$$

where all partial derivatives are calculated in x_i. Thus, subtracting ρ_i and dividing by Δx, we get

$$\frac{\partial \rho}{\partial x} = \left[\frac{1}{\Delta x}\left(\rho_{i+1} - \rho_i\right)\right] - \frac{\partial^2 \rho}{\partial x^2}\frac{\Delta x}{2} - ..., \tag{4.5}$$

which means that the residual belongs to the class $O(\Delta x)$, which is a class of all functions which go to zero at least as fast as Δx. Thus, taking Δx sufficiently small, one can ensure the arbitrary accuracy of the approximation, provided all the partial derivatives are bounded.

Accuracy can be further increased by taking different points where the function is sampled, called stencil points. For example, writing

$$\rho_{i-1} = \rho(x_{i-1}) = \rho(x_i) - \frac{\partial \rho}{\partial x}\Delta x + \frac{\partial^2 \rho}{\partial x^2}\frac{\Delta x^2}{2} - ..., \tag{4.6}$$

subtracting (4.6) from (4.4) and dividing by $2\Delta x$, we get

$$\frac{\partial \rho}{\partial x} = \left[\frac{1}{2\Delta x}\left(\rho_{i+1} - \rho_{i-1}\right)\right] - \frac{\partial^3 \rho}{\partial x^3}\frac{\Delta x^2}{6} + ... \tag{4.7}$$

Thus, using stencil points $\{i-1, i+1\}$ to approximate the first-order derivative in the point i the obtained residual belongs to the class $O(\Delta x^2)$, which means that this discretization of the Transport PDE has order of accuracy 2.

In general, if one wants to approximate the derivative of order m in point i using N stencil points $\{i+s_1, i+s_2, ..., i+s_N\}$ with $m < N$ in form

$$\frac{\partial^m \rho}{\partial x^m} \approx \sum_{j=1}^{N} a_j \rho_{i+s_j} \tag{4.8}$$

where coefficients a_j are unknown, one can define $S_{N,N} \in \mathbb{R}^{N\times N}$, $a \in \mathbb{R}^N$ and $c \in \mathbb{R}^N$ by

$$S_{N,N} = \begin{pmatrix} 1 & \cdots & 1 \\ s_1 & \cdots & s_N \\ \vdots & \ddots & \vdots \\ s_1^{N-1} & \cdots & s_N^{N-1} \end{pmatrix}, \; a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}, \; c = \frac{m!}{\Delta x^m}\begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix},$$

where c is nonzero on the position $m+1$, and solve a linear system

$$a = S_{N,N}^{-1}\, c. \tag{4.9}$$

The system (4.9) can be trivially obtained by writing Taylor series for all points $\rho_{i+s_1}...\rho_{i+s_N}$ and summing them in a linear combination as in (4.8). The obtained order of accuracy is at least $O(\Delta x^{(N-m)})$, and sometimes can be higher if some of the higher derivatives are also eliminated (as in case of (4.7)).

4.2.3 Continuation

Essentially the same process can be applied to the equation (4.1) to get the PDE version. For every term in a sum we can write

$$\rho_{i+s_j} = \rho(x_{i+s_j}) = \rho(x_i) + \frac{\partial \rho}{\partial x}\Delta x s_j + \frac{\partial^2 \rho}{\partial x^2}\frac{\Delta x^2 s_j^2}{2} + ... \tag{4.10}$$

Thus, assume we state the problem of finding the PDE approximation of (4.1) in form

$$\sum_{j=1}^{N} a_j \rho_{i+s_j} \approx \sum_{k=0}^{d} c_k \frac{\Delta x^k}{k!}\frac{\partial^k \rho}{\partial x^k}, \tag{4.11}$$

where d is the highest order of derivative (*order of continuation*) we want to use. Note that zero is also included in the right summation, since the function itself can be used in the resulting PDE. Then, introducing $S_{d+1,N} \in \mathbb{R}^{(d+1)\times N}$, $a \in \mathbb{R}^N$ and $c \in \mathbb{R}^{d+1}$ as

$$S_{d+1,N} = \begin{pmatrix} 1 & \cdots & 1 \\ s_1 & \cdots & s_N \\ \vdots & \ddots & \vdots \\ s_1^d & \cdots & s_N^d \end{pmatrix}, \quad a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix}, \quad c = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_d \end{pmatrix}, \tag{4.12}$$

substituting (4.10) in (4.11) we see that the vector of unknown coefficients c can be found by direct multiplication,

$$c = S_{d+1,N}\, a, \quad \text{or} \quad c_k = \sum_{j=1}^{N} a_j s_j^k \quad \forall k \in \{0, .., d\}. \tag{4.13}$$

Once (4.13) is solved, we write the PDE approximation to (4.1):

$$\frac{\partial \rho}{\partial t} = \sum_{k=0}^{d} c_k \frac{\Delta x^k}{k!} \frac{\partial^k \rho}{\partial x^k}. \tag{4.14}$$

As an example, applying (4.13) to the Transport ODE (4.2) with $d = 1$ we obtain the Transport PDE (4.3).

4.2.4 Analysis of reversibility

Procedures (4.9) and (4.13) look very similar from the algebraic point of view, however they are qualitatively different in the way how the problem is formulated and how we should interpret their results.

The discretization procedure tries to find the best approximation to a *continuous and smooth* function $\rho(t, x)$ and its derivatives. What is most important, the discretization step Δx is usually an adjustable parameter which can be set by a system engineer *arbitrarily small* to satisfy the desired performance. Thus the notion of accuracy of a discretization is used to describe how fast the solution of the discretized equation tends to the solution of the original equation when Δx tends to zero. In some sense this means quality of the discretization, since the higher order of accuracy means that the engineer can take larger Δx to achieve the same error and thus use the smaller number of states in the discretized system.

Contrary, when the original system is given by the ODE, the nodes have fixed locations, thus Δx is a *true constant* representing properties of an underlying physical system and it cannot be changed by an engineer. In turn this means that the accuracy defined as a class $O(\Delta x^{(N-m)})$ cannot measure quality of the approximation as Δx does not behave as an infinitely small value. Moreover, in the ODE case the system state ρ_i is known only on a given set of points i, thus in general the continuation $\rho(t, x)$ can be *non-smooth or discontinuous*.

Even if we assume the smoothness at the initial moment of time, the dynamics can render its derivatives unbounded. As a result the series in (4.10) can be non-convergent.

We know however that the systems (4.1) and (4.14) are connected by finite difference methods (4.9) and (4.13). We will use this fact as a definition of a reversible PDE approximation to an ODE.

Definition 4.1. Discretization of PDE to ODE is called *valid* if it is performed according to the finite difference method (4.9).

Definition 4.2. Continuation of ODE to PDE is called *reversible* if there exists a valid discretization of the obtained PDE to the original ODE.

These definitions basically mean that we assume a reversible PDE representation to be more natural, more intrinsic way to describe the system, and that the original ODE is just a particular discrete realization in physical world that we encountered. In particular definitions 4.1 and 4.2 say that if we use continuation on some ODE and then perform discretization at the same stencil points, we should arrive at the same ODE. This procedure sets a constraint on the minimum order of the reversible PDE:

Theorem 4.1. *The order of continuation d of the PDE which can be obtained from the ODE with N stencil points should satisfy the following constraint to be reversible:*

$$d + 1 \geqslant N. \tag{4.15}$$

Proof. Indeed, assume $d + 1 < N$. This means that the PDE approximation (4.14) has the highest derivative at most of the order d. Thus we can augment the vector of coefficients c by $N - d - 1$ zeros corresponding to higher-order derivatives obtaining a new vector $\bar{c} \in \mathbb{R}^N$. Augmenting c with $N - d - 1$ zeros to obtain $\bar{c}$ is equivalent to the augmentation of $S_{d+1,N}$ with $N - d - 1$ zero rows since c was defined by (4.13). Now, applying the discretization process (4.9) to $\bar{c}$ we should arrive at the same vector a of the parameters of the ODE system. Since this should be true for any a, we substitute $S_{N,N}^{-1}$ and augmented $S_{d+1,N}$ and obtain a condition

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_N \\ \vdots & \vdots & \ddots & \vdots \\ s_1^{N-1} & s_2^{N-1} & \cdots & s_N^{N-1} \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ s_1 & s_2 & \cdots & s_N \\ \vdots & \vdots & \ddots & \vdots \\ s_1^d & s_2^d & \cdots & s_N^d \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \end{pmatrix} = I,$$

which is impossible to satisfy since the second matrix is singular. Therefore there is no valid discretization process for the PDE obtained by continuation with order d such that $d+1 < N$, which by definition means that such continuation is not reversible.

Case $d + 1 = N$ is trivial, since the equations (4.9) and (4.13) are equivalent in this situation. This obviously provides a validity of the continuation procedure.

Now assume $d+1 > N$. Then the obtained vector of coefficients c is of higher dimension than a. The discretization (4.9) cannot be applied directly, since there is not enough stencil points to express the finite differences for the derivatives of order higher than $N-1$. However we can increase the set of stencil points. Let us choose additional $d+1-N$ stencil points $\bar{s}_{N+1}, \bar{s}_{N+2}, ..., \bar{s}_{d+1}$. Applying continuation (4.13) to the original ODE (4.1) and then (4.9) to the obtained PDE using the augmented set of stencil points we get a new ODE gains $\bar{a}$ which are expressed as

$$\bar{a} = \begin{pmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ s_1 & \cdots & s_N & \bar{s}_{N+1} & \cdots & \bar{s}_{d+1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ s_1^d & \cdots & s_N^d & \bar{s}_{N+1}^d & \cdots & \bar{s}_{d+1}^d \end{pmatrix}^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ s_1 & \cdots & s_N \\ \vdots & \ddots & \vdots \\ s_1^d & \cdots & s_N^d \end{pmatrix} a.$$

We can show that first N elements of $\bar{a}$ are exactly a and the rest is zero, irrespective of the chosen additional points $\bar{s}_j$. This means that the artificially introduced stencil points do not appear in the discretized PDE, rendering the same ODE as the original one.

Indeed, for any matrix $S \in \mathbb{R}^{d\times N}$ for $d > N$ and $\bar{S} \in \mathbb{R}^{d\times(d-N)}$ such that the matrix $\begin{pmatrix} S & \bar{S} \end{pmatrix}$ is invertible one can prove that

$$\begin{pmatrix} S & \bar{S} \end{pmatrix}^{-1} S = \begin{pmatrix} I \\ 0 \end{pmatrix}. \tag{4.16}$$

To prove (4.16) one can just multiply it by invertible matix $\begin{pmatrix} S & \bar{S} \end{pmatrix}$ from both sides and obtain the trivial equality $S = S$. □

The latter part of the proof of Theorem 4.1 means that the PDE obtained by the process (4.13) with $d+1 > N$ has more information that one with $d+1 = N$, since it provides exact Taylor approximations not only on the given set of points, but in the additional $d+1-N$ points which can be chosen arbitrary. This property can be used to define the excessive accuracy as $d+1-N$.

Definition 4.3. Excessive accuracy of a reversible continuation process of ODE to PDE is defined as the number of additional points in which the corresponding discretization process can be made exact simultaneously, i.e. $d+1-N$.

For example, a continuation

$$\rho_{i+1} - \rho_i \quad \to \quad \Delta x \frac{\partial \rho}{\partial x}$$

is of excessive accuracy 0, since trying to discretize the PDE on any larger set of stencil points except from $\{i, i+1\}$ will give different ODE. At the same time a continuation

$$\rho_{i+1} - \rho_{i-1} \quad \to \quad 2\Delta x \frac{\partial \rho}{\partial x} + \left(0 \cdot \frac{\partial^2 \rho}{\partial x^2}\right)$$

has excessive accuracy 1 because the second derivative vanishes (thus $d = 2$), and it is possible to discretize the PDE on a set of stencil points of size 3 (with one additional point), for example $\{i-1, i, i+1\}$.

4.2.5 Analysis of convergence

It is clear that the higher order of continuation is taken, the better the original ODE operator (4.1) is approximated by the PDE (4.14). It is possible to study the convergence properties by shifting the problem to the frequency domain using the Fourier transform. In this section we will perform a spectrum analysis and then derive a bound on solutions' deviation.

For simplicity of writing without loss of generality assume in this section $\Delta x = 1$. Let us define a function $a(x)$ as

$$a(x) = \sum_{j=1}^{N} a_j \delta\left(x + s_j\right), \tag{4.17}$$

where $\delta(x)$ is the Dirac delta function. Further, assume that the state $\rho_i(t)$ of (4.1) was sampled from some integrable function $\rho_i(t) := \rho(t, x_i)$. Then, equation (4.1) is equivalent to the following system with convolution

$$\frac{\partial \rho}{\partial t}(t, x) = (a \star \rho(t, \cdot))(x). \tag{4.18}$$

Use now the Fourier transform, defined as

$$\mathcal{F}\left\{f\right\}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-ix\omega} dx \tag{4.19}$$

for any integrable function $f(x)$ and for any frequency $\omega \in \mathbb{R}$. It is known that the Fourier image of a convolution is a multiplication. Therefore the system (4.18) is just

$$\frac{\partial \mathcal{F}\left\{\rho\right\}}{\partial t}(t, \omega) = \mathcal{F}\left\{a\right\}(\omega) \mathcal{F}\left\{\rho\right\}(t, \omega), \qquad \mathcal{F}\left\{a\right\}(\omega) = \sum_{j=1}^{N} a_j e^{is_j\omega}, \tag{4.20}$$

where $\mathcal{F}\left\{a\right\}(\omega)$ was found by direct calculation of Fourier transform. To interpret (4.20), let us introduce an operator T over integrable functions such that $T\rho$ is a right-hand side of the original ODE system (4.18). A spectrum of an operator T is defined as a closed set of points $\lambda \in \mathbb{C}$ for which $T - \lambda I$ is not invertible. Thus finding a spectrum of $(a \star \rho(t, \cdot))(x)$ is equivalent to finding a closure of a set of all λ such that for some $v(x)$ there is no solution to $(a \star \rho(t, \cdot))(x) - \lambda\rho(t, x) = v(x)$. Taking Fourier transform one arrives at $\left(\mathcal{F}\left\{a\right\}(\omega) - \lambda\right) \mathcal{F}\left\{\rho\right\}(t, \omega) = \mathcal{F}\left\{v\right\}(\omega)$, which clearly has no solution for $\mathcal{F}\left\{v\right\}(\omega) \neq 0$ if and only if $\lambda = \mathcal{F}\left\{a\right\}(\omega)$ for some ω. Therefore we have just shown that the spectrum of T is parametrized by the closure of the image of $\mathcal{F}\left\{a\right\}(\omega)$. In the case of (4.20) $\mathcal{F}\left\{a\right\}(\omega)$, being an image of the unit circle, coincides with its closure, therefore the spectrum is simply $\left\{\mathcal{F}\left\{a\right\}(\omega) | \omega \in \mathbb{R}\right\}$. In fact, this result is well-known, since the system (4.1) on an infinite line belongs to the class of Laurent systems, whose spectrum is known to be (4.20), see Frazho and Bhosri 2010.

Now let us calculate the spectrum of the right-hand side of the continualized system (4.14). Denote the state of the continualized system as $\rho^c(t, x)$. By another property of the Fourier transform, if the function $\rho^c(t, x)$ is sufficiently smooth and its derivatives are integrable, we

can recover their Fourier images by

$$\mathcal{F}\left\{\frac{\partial^k \rho^c}{\partial x^k}\right\}(t,\omega) = (i\omega)^k \mathcal{F}\left\{\rho^c\right\}(t,\omega).$$

Therefore (4.14) is read in frequency domain as

$$\frac{\partial \mathcal{F}\left\{\rho^c\right\}}{\partial t}(t,\omega) = \mathcal{F}\left\{c\right\}(\omega)\mathcal{F}\left\{\rho^c\right\}(t,\omega), \qquad \mathcal{F}\left\{c\right\}(\omega) = \sum_{k=0}^{d} c_k \frac{1}{k!}(i\omega)^k. \tag{4.21}$$

Substituting (4.13), we can rewrite $\mathcal{F}\left\{c\right\}(\omega)$ in (4.21) as

$$\mathcal{F}\left\{c\right\}(\omega) = \sum_{j=1}^{N} a_j \sum_{k=0}^{d} \frac{(is_j\omega)^k}{k!}. \tag{4.22}$$

Now, comparing (4.22) with (4.20), it is clear that (4.22) uses the first $d+1$ terms of the Taylor expansion of the exponential function in (4.20). In fact, since the exponential function is analytic on the whole complex plane, we have just proven the following result:

Theorem 4.2. *The spectrum of the PDE operator* (4.14) *converges to the spectrum of the original ODE operator* (4.1) *pointwise as* $d \to \infty$.

Define now a Discrete-Time Fourier Transform (or DTFT, although taken along the coordinate axis) for an infinite sequence f_n for $n \in \mathbb{Z}$ as

$$\mathcal{D}\left\{f\right\}(\omega) = \sum_{n=-\infty}^{+\infty} f_n e^{-in\omega}, \qquad \omega \in [-\pi, \pi]. \tag{4.23}$$

This transform is also known as the *z-transform*, evaluated on the unit circle. We can use Theorem 4.2 to prove that the sampled trajectory of the PDE converges to the solution of the ODE as d increases:

Theorem 4.3. *Let* $\tilde{\rho}_i(t) := \rho_i(t) - \rho^c(t, x_i)$ *be a deviation between the original and the continualized systems' solutions at the nodes' positions. Assume that at initial moment*

$$\begin{aligned} \mathcal{D}\left\{\rho\right\}(0,\omega) &= \mathcal{F}\left\{\rho^c\right\}(0,\omega) \quad \forall |\omega| \leqslant \pi, \\ 0 &= \mathcal{F}\left\{\rho^c\right\}(0,\omega) \quad \forall |\omega| > \pi, \end{aligned} \tag{4.24}$$

which defines the initial state of the PDE with respect to the original ODE. Then for $\forall t \geqslant 0$

$$\begin{gathered} \left\|\tilde{\rho}(t)\right\|_{l_2} \leqslant e^{\operatorname{Re}\lambda_{max} t}\left(e^{\gamma_d t} - 1\right)\left\|\rho(0)\right\|_{l_2}, \quad \textit{where} \\ \operatorname{Re}\lambda_{max} = \max_{|\omega|\leqslant\pi} \operatorname{Re}\mathcal{F}\left\{a\right\}(\omega), \qquad \gamma_d = \sum_{j=1}^{N} |a_j| \frac{|\pi s_j|^{d+1}}{(d+1)!} e^{|\pi s_j|}. \end{gathered} \tag{4.25}$$

In particular for any fixed $t \geqslant 0$ $\left\|\tilde{\rho}(t)\right\|_{l_2} \to 0$ *as* $d \to \infty$.

Proof. First of all, by definition (4.23) of DTFT it is clear that $\mathcal{D}\{a\}(\omega) \equiv \mathcal{F}\{a\}(\omega)$, where the former is taken for the sequence a_i in (4.1) and the latter is taken for the function $a(x)$ in (4.17). Since the right-hand side of (4.1) represents a convolution of $\rho_i(t)$ with the sequence a_i, we can use this equality and write the evolution of the DTFT image of ρ_i as

$$\frac{\partial \mathcal{D}\{\rho\}}{\partial t}(t,\omega) = \mathcal{F}\{a\}(\omega)\mathcal{D}\{\rho\}(t,\omega). \tag{4.26}$$

Further it is easy to show (e.g. by Fischer 2018) that the sampling of $\rho^c(x,t)$ induces *periodization* on its Fourier image:

$$\mathcal{D}\{\rho^c(t,x_i)\}(\omega) = \sum_{n=-\infty}^{+\infty} \mathcal{F}\{\rho^c(t,x)\}(\omega + 2\pi n).$$

By (4.24) and by (4.21) $\mathcal{F}\{\rho^c\}(t,\omega) = 0$ for $|\omega| > \pi, t \geqslant 0$, which means that $\mathcal{D}\{\rho^c(t,x_i)\}(\omega) \equiv \mathcal{F}\{\rho^c(t,x)\}(\omega)$ for $|\omega| \leqslant \pi$. Therefore, the DTFT of $\tilde{\rho}_i(t)$ is

$$\mathcal{D}\{\tilde{\rho}\}(t,\omega) := \mathcal{D}\{\rho\}(t,\omega) - \mathcal{F}\{\rho^c\}(t,\omega) \qquad \forall \omega \in [-\pi,\pi].$$

Let us now use Parseval's identity for DTFT, see Frazho and Bhosri 2010:

$$\|\tilde{\rho}(t)\|_{l_2}^2 = \sum_{i=-\infty}^{+\infty} |\tilde{\rho}_i|^2(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\mathcal{D}\{\tilde{\rho}\}(t,\omega)\right|^2 d\omega. \tag{4.27}$$

The integral is taken over the bounded interval of frequencies since the transformed sequence is discrete.

One can now notice that (4.21) and (4.26) are just scalar linear time-invariant ODEs for each ω, thus it is possible to write their explicit solutions as

$$\begin{aligned}\mathcal{D}\{\rho\}(t,\omega) &= e^{\mathcal{F}\{a\}(\omega)t}\mathcal{D}\{\rho\}(0,\omega),\\ \mathcal{F}\{\rho^c\}(t,\omega) &= e^{\mathcal{F}\{c\}(\omega)t}\mathcal{F}\{\rho^c\}(0,\omega).\end{aligned}$$

Using the condition (4.24) on initial conditions $\mathcal{D}\{\rho\}(0,\omega) = \mathcal{F}\{\rho^c\}(0,\omega)\ \forall|\omega| \leqslant \pi$ we write the Fourier image of $\tilde{\rho}_i(t)$:

$$\mathcal{D}\{\tilde{\rho}\}(t,\omega) = \left(e^{\mathcal{F}\{a\}(\omega)t} - e^{\mathcal{F}\{c\}(\omega)t}\right)\mathcal{D}\{\rho\}(0,\omega) = e^{\mathcal{F}\{a\}(\omega)t}\left(1 - e^{(\mathcal{F}\{c\}(\omega)-\mathcal{F}\{a\}(\omega))t}\right)\mathcal{D}\{\rho\}(0,\omega) \tag{4.28}$$

which holds $\forall|\omega| \leqslant \pi$. Inserting (4.28) in (4.27) and using Hölder's inequality we get

$$\begin{aligned}\|\tilde{\rho}(t)\|_{l_2}^2 \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\mathcal{D}\{\rho\}(0,\omega)\right|^2 d\omega \times\\ \times \max_{|\omega|\leqslant\pi}\left|e^{\mathcal{F}\{a\}(\omega)t}\right|^2 \times \max_{|\omega|\leqslant\pi}\left|1 - e^{(\mathcal{F}\{c\}(\omega)-\mathcal{F}\{a\}(\omega))t}\right|^2.\end{aligned} \tag{4.29}$$

The first multiplier is just $\|\rho(0)\|_{l_2}^2$. Further it is evident that $\max_{|\omega|\leqslant\pi}\left|e^{\mathcal{F}\{a\}(\omega)t}\right|^2 = e^{2\operatorname{Re}\lambda_{max}t}$. Thus we will concentrate on the third multiplier.

Let $z = u + iv$ be any complex number. Then by Mitrinovic and Vasic 1970-(3.8.23) we can write $|e^z - 1|^2 \leqslant \left(e^{|z|} - 1\right)^2$. This bound increases with respect to $|z|$, therefore

$$\max_{|\omega| \leqslant \pi} \left|1 - e^{(\mathcal{F}\{c\}(\omega) - \mathcal{F}\{a\}(\omega))t}\right|^2 \leqslant \left(e^{\gamma t} - 1\right)^2$$

for any $\gamma \geqslant \max_{|\omega|\leqslant\pi} |\mathcal{F}\{c\}(\omega) - \mathcal{F}\{a\}(\omega)|$. We can find the lowest bound on γ denoted as γ_d using the definitions of $\mathcal{F}\{a\}(\omega)$ and $\mathcal{F}\{c\}(\omega)$ in (4.20) and (4.22). Namely,

$$\begin{aligned} |\mathcal{F}\{c\}(\omega) - \mathcal{F}\{a\}(\omega)| &= \left|\sum_{j=1}^{N} a_j \sum_{k=d+1}^{+\infty} \frac{(is_j\omega)^k}{k!}\right| \leqslant \\ &\leqslant \sum_{j=1}^{N} |a_j| \sum_{k=d+1}^{+\infty} \frac{|s_j\omega|^k}{k!} \leqslant \sum_{j=1}^{N} |a_j| \frac{|s_j\omega|^{d+1}}{(d+1)!} \sum_{k=0}^{+\infty} \frac{|s_j\omega|^k}{k!}, \end{aligned} \tag{4.30}$$

and the last summation is just $e^{|s_j\omega|}$. Finally, since (4.30) increases with $|\omega|$, we can substitute the maximal value $|\omega| = \pi$ and thus obtain γ_d as in (4.25). Finally the bound (4.25) is recovered by taking square root of (4.29).

The final statement of the theorem can be proven if one notices that $\gamma_d \to 0$ as $d \to \infty$, which leads to $(e^{\gamma_d t} - 1) \to 0$ as $d \to \infty$ for any fixed $t \geqslant 0$. □

Remark 4.1. Condition (4.24) means that the continuous system should be initialized with the low-frequency continuation of the original ODE initial state. Note that this can always be done since (4.24) uniquely determines the Fourier image of $\rho^c(0, x)$. For example an initial state $\rho_0 = 1$ and $\rho_i = 0$ for $i \neq 0$ results in $\mathcal{D}\{\rho\}(\omega) \equiv 1$ which by (4.24) sets $\rho^c(0, x) = \operatorname{sinc}(\pi x)$. Moreover, bound (4.25) at $t = 0$ ensures that $\tilde{\rho}_i(0) \equiv 0$, therefore the continuation coincides with the ODE initial state.

4.2.6 Analysis of stability

We can now turn to the discussion of stability of the obtained PDE. Due to the simple nature of scalar equations (4.20) and (4.21) we can say that the system (4.20) is *stable* if and only if $\operatorname{Re}\mathcal{F}\{a\}(\omega) \leqslant 0\ \forall\omega \in \mathbb{R}$, otherwise it is *unstable*. A simple corollary of Theorem 4.3 can be derived:

Corollary 4.1. *If* $\operatorname{Re}\mathcal{F}\{a\}(\omega) \leqslant \operatorname{Re}\lambda_{max} < 0\ \forall\omega \in \mathbb{R}$, *then* $\|\tilde{\rho}(t)\|_{l_2} \to 0$ *as* $t \to \infty$ *for all high enough* d.

Proof. Indeed, for high enough d we have $\gamma_d < -\operatorname{Re}\lambda_{max}$, which means that (4.25) is bounded by an exponential $e^{(\operatorname{Re}\lambda_{max} + \gamma_d)t} \to 0$ as $t \to \infty$. □

Note that although Theorem 4.3 states the convergence of sampled trajectories, in Theorem 4.2 the convergence of spectrums is not uniform. Moreover, the spectrum (4.20) is an image of the unit circle and thus is a compact set, while the spectrum (4.22) for any d is

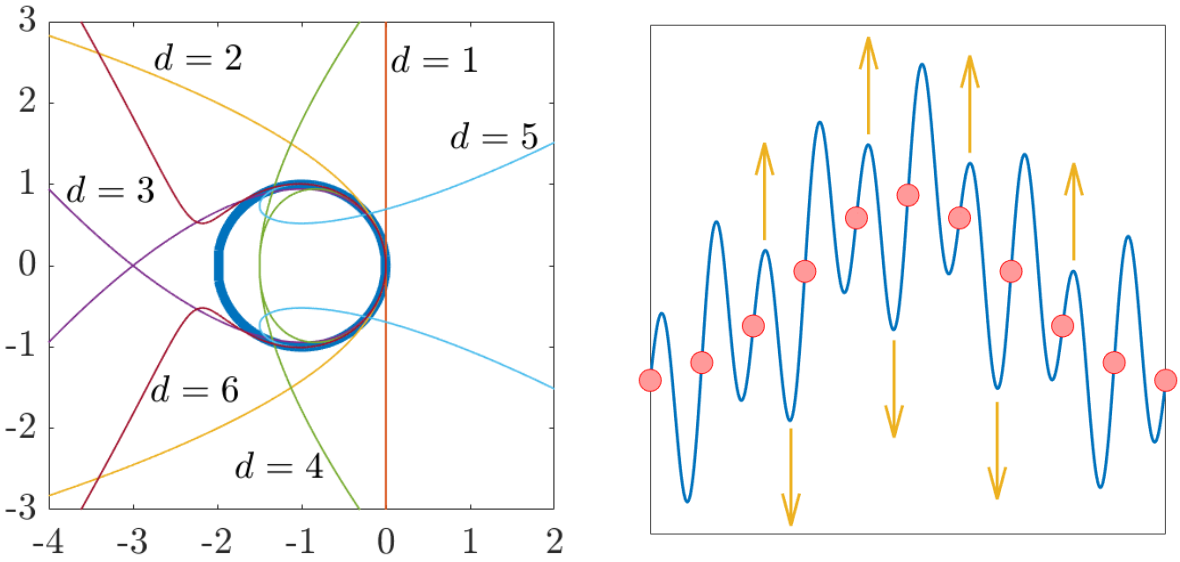


Figure 4.3: **Left:** spectrum for the Transport ODE (4.31) $e^{i\omega} - 1$ (blue circle) together with spectrums of the continuations up to the order 6, according to (4.32). As d increases, spectrums converge to the blue circle, however for some orders (such as 4 or 5) they can become unstable. **Right:** Schematic picture of an artificial instability for high order d. Although the continuation (blue) coincides with the original solution (red) at the nodes' positions, high-frequency components can be unstable.

a polynomial and thus unbounded. This can lead to an undesirable effect which we call an *artificial instability*, meaning that the tails of the image of the polynomial (4.22) happen to lie in the positive complex half-plane, as in the left panel of Fig. 4.3 for $d = 4$ or 5. Essentially this means that the PDE becomes unstable on high frequencies, see the right panel of Fig. 4.3. We can though induce several corollaries from Theorems 4.2 and 4.3 which can help in understanding stability properties of the obtained PDE.

Corollary 4.2. *If the original ODE (4.1) is unstable, there exists $D \geqslant 0$ such that for all $d \geqslant D$ the continualized system (4.14) will also be unstable.*

Proof. Since the original system is unstable, there exists ω_0 such that $\operatorname{Re}\mathcal{F}\{a\}(\omega_0) > 0$. Now, by Theorem 4.2 there exists $D \geqslant 0$ such that for all $d \geqslant D$ $\operatorname{Re}\mathcal{F}\{c\}(\omega_0) > 0$. □

Corollary 4.3. *PDE (4.14) with an odd order of continuation d has the same stability properties as a PDE with the order of continuation $d - 1$.*

Proof. All odd terms in the spectrum (4.22) are purely imaginary and thus have no impact on the stability. □

Corollary 4.4. *Artificial instability is introduced when the last even term in the PDE (4.14) has $c_k > 0$ if $k = 4m$ or $c_k < 0$ if $k = 4m + 2$ for some $m \in \mathbb{Z}^+$.*

Proof. Artificial instability comes if the term of the polynomial (4.22) with the highest even power is positive, which leads to a positive real part of the spectrum on high frequencies. Positivity of the highest even term is exactly equivalent to the statement of the corollary since $i^{4m} = 1$ and $i^{4m+2} = -1$ for any $m \in \mathbb{Z}^+$. □

We will demonstrate the convergence of spectrums on the Transport ODE

$$\dot{\rho}_i = \rho_{i+1} - \rho_i. \tag{4.31}$$

With $\Delta x = 1$, the continuation of (4.31) is:

$$\frac{\partial \rho}{\partial t}(t,x) = \sum_{k=1}^{d} \frac{1}{k!} \frac{\partial^k \rho}{\partial x^k}(t,x), \tag{4.32}$$

Spectrum of (4.31) equals $e^{i\omega} - 1$ by (4.20), which is depicted as a blue circle in the left panel of Fig. 4.3 together with the spectrums of the continuations up to the order $d = 6$. It is clear that as the order increases, the approximations become better.

The original Transport ODE is stable. Moreover, it has an intrinsic diffusion in it, which can be captured by the continuation of the second order. However, the continuations of orders 4 and 5 are unstable. It happens because of an artificial instability as described in Corollary 4.4, since $c_4 = 1 > 0$. In general all stable continuations of the Transport ODE are given by the orders $\{1, 2, 3, ..., 4m + 2, 4m + 3, ...\}$ for all $m \in \mathbb{Z}^+$.

Theorems 4.2 and 4.3 say that increasing order of continuation leads to the more correct capture of the behavior of the original ODE. Further, Theorem 4.1 shows that high enough order of continuation guarantees that the original model can be reconstructed back from the approximated PDE system. However, from the practical point of view, low-order PDEs capture low frequency effects very well, while high orders can cause artificial instability. Moreover, lack of tools for control and analysis of high-order PDEs makes impractical their derivation. Therefore it usually makes sense to stick to the orders $d = 1$ or $d = 2$, which will be used in examples throughout this thesis.

4.3 Continuation for nonlinear spatially invariant systems

Finite differences give us a complete tool for linear systems, but for nonlinear systems they should be applied in composition with nonlinearities. Using an additional concept of computational graph it is possible to elaborate the case of general nonlinear ODE systems.

As in the previous case we assume without loss of generality that the nodes are equally spaced along the 1D line, a node i having a state ρ_i and a position x_i. Then the general nonlinear ODE with space dependence takes form of

$$\dot{\rho}_i = F(\rho_{i+s_1}, \rho_{i+s_2}, ..., \rho_{i+s_N}). \tag{4.33}$$

We further assume that the function F is continuous.

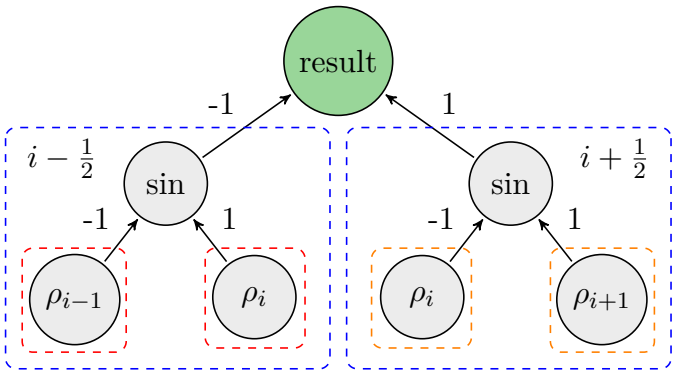


Figure 4.4: Computational graph for the system (4.34). Similar subgraphs are outlined by dashed rectangles of the same color. Possible choices of sinus subgraph' positions are written in the corners of blue rectangles.

4.3.1 Computational graph

Kolmogorov 1957 showed that every multidimensional continuous function can be written as a composition of functions of one variable and additions. This work laid the basis for the neural networks function approximation, which is now a major branch of modern machine learning.

Here we will use this idea and assume that the function F is given in the form of computational graph (see Baydin et al. 2018 for review). This is a directed acyclic graph, every node of which represents a one-dimensional function, applied to a weighted sum of inputs coming to this node. We assume that the leaves of this graph are the states of the system ρ_{i+s_j} and the root node computes the resulting value of F.

As an example of the computational graph we will consider a system

$$\dot{\rho}_i = \sin(\rho_{i+1} - \rho_i) - \sin(\rho_i - \rho_{i-1}) \tag{4.34}$$

which is a system of Kuramoto oscillators coupled on a ring. The computational graph for (4.34) is presented in Fig. 4.4.

4.3.2 Similar subgraphs and their positions

Now let us introduce an original notion of *similar subgraphs*. Subgraph is a computational graph which computes subexpression of the original computational graph. Every node in a computational graph serves as the root of a subgraph computing expression defined in this node. The leaf nodes are also the subgraphs "computing" themselves.

Definition 4.4. We call two subgraphs *similar* if

1. they serve as an input to the same node,

2. they differ only in the positions of the leaf nodes, and this difference can be represented by a single shift.

This is an equivalence relation, therefore we can speak about equivalence classes which we call sets of similar subgraphs.

For example, in Fig. 4.4 there are three sets of similar subgraphs:

1. ρ_{i-1} and ρ_i for the left sinus node,
2. ρ_i and ρ_{i+1} for the right sinus node,
3. $\sin(\rho_i - \rho_{i-1})$ and $\sin(\rho_{i+1} - \rho_i)$ for the root node, because they differ by a single shift which equals 1.

Finally we will define a *position* of a subgraph:

Definition 4.5. Position of a subgraph is defined as a coordinate in space where the expression of this subgraph is calculated.

The leaf nodes by definition are the states of the system, thus their positions are uniquely specified. For example for the leaf node ρ_{i+1} in Fig. 4.4 we say that its position is $i+1$. The root node by definition has a position i, since it is exactly the position of the left-hand side term in (4.33).

For other subgraphs defining the position is ambiguous. One choice of this *position function* could be an average of all positions of inputs, as in Fig. 4.4. Position function should have the only property to be well-defined: similar subgraphs, which differ by some shift s, should have their positions differ also by s. In general, this function could affect performed computations (and thus change the obtained PDE), but in a linear case it is possible to prove that the particular choice doesn't matter:

Theorem 4.4. *For any linear combination of states computed at point i*

$$E_i = \sum_{j=1}^{N} a_j \rho_{i+s_j}, \tag{4.35}$$

its continuous approximation $\hat{E}_i$ defined by (4.11) at point i coincides with an approximation $\left[\hat{E}_{i+k}\right]_i$ made at some position $i+k$ and then reapproximated again at i, if all approximations were done up to the order d.

Proof. By (4.11), we can write the approximation of E_i made directly at point i as

$$\hat{E}_i = \sum_{m=0}^{d} \left(\left(\sum_{j=1}^{N} a_j \frac{s_j^m}{m!} \right) \frac{\partial^m \rho_i}{\partial x^m} \Delta x^m \right), \tag{4.36}$$

where $\frac{\partial^m \rho_i}{\partial x^m}$ is a derivative of order m calculated at point i. Now, using the same approximation rule at $i+k$, we obtain that

$$\hat{E}_{i+k} = \sum_{m=0}^{d} \left(\left(\sum_{j=1}^{N} a_j \frac{(s_j - k)^m}{m!} \right) \frac{\partial^m \rho_{i+k}}{\partial x^m} \Delta x^m \right). \tag{4.37}$$

Finally, every partial derivative, calculated at point $i+k$, can be written again as a Taylor series at point i:

$$\frac{\partial^m \rho_{i+k}}{\partial x^m} = \sum_{q=0}^{d-m} \frac{k^q}{q!} \frac{\partial^{m+q} \rho_i}{\partial x^{m+q}} \Delta x^q, \tag{4.38}$$

where the summation is truncated such that the maximal order of derivative does not exceed d. The resulting approximation is

$$\begin{aligned} \left[\hat{E}_{i+k}\right]_i &= \sum_{m=0}^{d} \left(\left(\sum_{j=1}^{N} a_j \frac{(s_j - k)^m}{m!} \right) \sum_{q=0}^{d-m} \frac{k^q}{q!} \frac{\partial^{m+q} \rho_i}{\partial x^{m+q}} \Delta x^{m+q} \right) \\ &= \sum_{m=0}^{d} \left(\left(\sum_{j=1}^{N} a_j \left[\sum_{q=0}^{m} \frac{(s_j - k)^{m-q}}{(m-q)!} \frac{k^q}{q!} \right] \right) \frac{\partial^m \rho_i}{\partial x^m} \Delta x^m \right), \end{aligned} \tag{4.39}$$

and it is clear that the value inside the square brackets is exactly $\frac{s_j^m}{m!}$ by binomial expansion. Therefore, (4.39) coincides with (4.36), which concludes the proof. □

In the following we will stick to the choice of an averaging position function. Since the position of a subgraph represents a position on the line, it is natural to have non-integer position values, although the leaf nodes and the root have only integer positions. As an example, with the averaging position function, in Fig. 4.4 the node $\sin(\rho_{i+1} - \rho_i)$ has its position $i + 1/2$.

4.3.3 Continuation

When system (4.33) is expressed in a form of computational graph with similar subgraphs being found and their positions being defined, one can perform a continuation procedure described in section 4.2 to obtain a PDE.

Continuation should be performed recursively, starting from the leaves. Each set of similar subgraphs by definition is used in their common ancestor node as a linear combination of equivalent elements shifted by some distance. Continuation of this linear combination by (4.11) replaces a set of similar subgraphs by a weighted sum of partial derivatives of subexpressions, calculated at the position of the ancestor node.

Let $\Delta x = x_{i+1} - x_i$ be a distance between two neighbouring nodes. Elaborating example (4.34) and using $d = 1$ for each set of subgraphs, we perform the continuation in three steps:

1. $\sin(\rho_{i+1} - \rho_i) \quad \to \quad \sin\left(\Delta x \frac{\partial \rho}{\partial x}(x_{i+1/2})\right),$

2. $\sin(\rho_i - \rho_{i-1}) \quad \to \quad \sin\left(\Delta x \frac{\partial \rho}{\partial x}(x_{i-1/2})\right)$,

3. $\sin_{i+1/2} - \sin_{i-1/2} \quad \to \quad \Delta x \frac{\partial}{\partial x} \sin$.

which finally gives a nonlinear PDE representation of (4.34):

$$\frac{\partial \rho}{\partial t}(t,x) = \Delta x \frac{\partial}{\partial x} \sin\left(\Delta x \frac{\partial \rho}{\partial x}(t,x)\right). \tag{4.40}$$

To obtain higher-order PDE approximations it makes sense to specify the desired order of the equation d and then get rid of all the terms which consist of composition of derivatives of combined order higher than d.

4.4 Continuation for general ODE systems

Until now we discussed systems with nodes which were uniformly placed on the infinite 1D line and which had common space-independent dynamics. The method can be extended to include more classes of systems.

4.4.1 Trivial extensions

Spatially invariant systems (Bamieh, Paganini, and Dahleh 2002) such as periodic ones can be tackled by choosing different index spaces. In the periodic case we can assume that the positions $x \in \mathbb{S}$ are placed on the unit circle and indices $i \in \mathbb{Z} \setminus n\mathbb{Z}$ form a ring of integers modulo n, where n is the number of states of the original ODE. Since any function on $\mathbb{S}$ can be mapped to a periodic function on $\mathbb{R}$, the analysis in Sections 4.2 and 4.3 remain the same.

Time dependence can be introduced into system gains both in the ODE and in the PDE, where the continuation is performed independently for every fixed t. This allows to use the method for time-varying systems and switching networks. Also systems whose state is vector-valued can be continualized using the same finite differences based scheme, thus in the following we will assume that the state of a system is scalar.

In the following subsections we will explore how the method can be extended to include systems with several spatial dimensions, systems with space dependence or nonuniform placing and systems with boundaries. Further we introduce a concept of PDE with index derivatives which can be applied to systems whose states coincide with the positions in space, for example particle systems. Finally, all kinds of systems are covered by the general continuation algorithm presented in the end of this section.

4.4.2 Multidimensional systems

In Sections 4.2 and 4.3 we assumed that the nodes were placed on the 1D line, with integer indices and scalar positions. Now we describe how to extend the method for the space with n dimensions.

Let a position of a node ρ_i be described by $x_i \in \mathbb{R}^n$. Moreover, a node ρ_i is referenced by a multi-index $i = (i_1, ..., i_n) \in \mathbb{Z}^n$. We assume that the position difference between two neighbour nodes $i = (i_1, ..., i_k, ..., i_n)$ and $i' = (i_1, ..., i_k + 1, ..., i_n)$ is

$$x_{i'} - x_i = (0, ..., \Delta x_k, ..., 0) \qquad \forall k \in \{1..n\},$$

and that there exists a vector $\Delta x = (\Delta x_1, ..., \Delta x_k, ..., \Delta x_n)$.

Nonlinear multidimensional system can be treated by the same computational graph, so the only difference between 1D and multi-dimensional case is in a transformation from a linear weighted sum into PDE, as in (4.11). Therefore assume that the system F is linear:

$$\dot{\rho}_i = F(\rho_{i+s_1}, \rho_{i+s_2}, ..., \rho_{i+s_N}) = \sum_{j=1}^{N} a_j \rho_{i+s_j}. \tag{4.41}$$

For nonnegative multi-index h we define an absolute value $|h| = \sum_{k=1}^{n} h_k$. Further, we define multi-index power h of a vector x as $x^h = \prod_{k=1}^{n} x_k^{h_k}$, with an assumption $0^0 = 1$. By Taylor series

$$\rho_{i+s_j} = \rho_i + \sum_{|h|=1} s_j^h \Delta x^h \frac{\partial \rho_i}{\partial x^h} + \sum_{|h|=2} \frac{s_j^h}{2} \Delta x^h \frac{\partial^2 \rho_i}{\partial x^h} + ..., \tag{4.42}$$

and the dynamics thus is

$$\begin{aligned} \frac{\partial \rho}{\partial t} = &\left(\sum_{j=1}^{N} a_j \right) \rho + \sum_{|h|=1} \left(\sum_{j=1}^{N} a_j s_j^h \right) \Delta x^h \frac{\partial \rho}{\partial x^h} + \\ &+ \sum_{|h|=2} \left(\sum_{j=1}^{N} a_j \frac{s_j^h}{2} \right) \Delta x^h \frac{\partial^2 \rho}{\partial x^h} + ..., \end{aligned} \tag{4.43}$$

where the approximation is truncated up to the first $d+1$ terms representing the derivatives of order $|h| \leqslant d$.

4.4.3 Space-dependent and unequally spaced systems

Let us now look at the linear system (4.1) with one important difference: the system gains a_j, the shifts s_j and the number of neighbours N become space-dependent:

$$\dot{\rho}_i = \sum_{j=1}^{N_i} a_{ij} \rho_{i+s_{ij}}. \tag{4.44}$$

Notice that equation (4.44) describes in fact any linear system.

Now one can perform a continuation (4.13) at every point x_i up to the order d and obtain a PDE (4.14) with space dependent gains c_{ik}. This means that we know the gains c_{ik} at points with coordinates x_i, which can be seen as a sampling of some function $c_k(x)$ at x_i.

Non-uniform placing of nodes can be tackled in the same way. Indeed, assuming distance $x_{i+1} - x_i$ can be arbitrary, continuation can be performed by defining $c_{ik} = \sum_{j=1}^{N_i} a_{ij}(x_{i+s_{ij}} - x_i)^k$ instead of (4.13).

We can now perform either an interpolation or an approximation based on this sampling. In the first case we seek for $c_k(x)$ such that $c_k(x_i) = c_{ik}$, while in the second case it is enough to satisfy this relation approximately. In either case, the resulting continuation of (4.44) is given by

$$\frac{\partial \rho}{\partial t}(t,x) = \sum_{k=1}^{d} c_k(x) \frac{1}{k!} \frac{\partial^k \rho}{\partial x^k}(t,x). \tag{4.45}$$

For nonlinear systems the continuation can be performed if computational graphs for every node compute the same dynamics. We can formalize it with the following property:

Definition 4.6. We say that two computational graphs *have the same structure* if

1. their root nodes compute the same expression,
2. any child subgraph of the root node of the first graph *has the same structure* with some child subgraph of the root node of the second graph and vice versa.

This definition, formulated through recursion, essentially means that the order of nonlinearities which is hidden in two computational graphs should coincide, see Fig. 4.5.

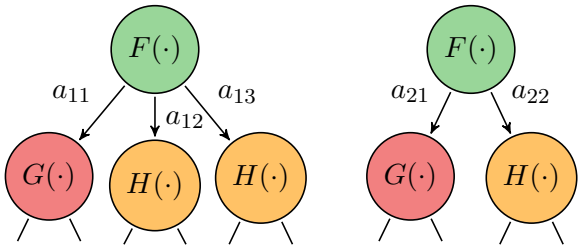


Figure 4.5: Illustration of two computational graphs having the same structure.

Finally, a continuation of a nonlinear ODE system can be performed if all the computational graphs computing the dynamics for all states ρ_i have the same structure. Indeed, in this case it is possible to perform a continuation for any set of similar subgraphs for each node as in the linear case of (4.44)-(4.45). Moreover, by Definition 6 these sets of similar subgraphs for different positions serve as inputs to the same nonlinearities, therefore a unique PDE with space-dependent coefficients can be obtained.

Remark 4.2. In theory, it is possible to satisfy Definition 6 for any nonlinear system formulated through computational graphs. Indeed, assume two computational graphs have two different root node expressions, denoted as $F(\cdot)$ and $G(\cdot)$ respectively. Then we can artificially create a new common root node which will compute $1 \cdot F(\cdot) + 0 \cdot G(\cdot)$ for the first graph and $0 \cdot F(\cdot) + 1 \cdot G(\cdot)$ for the second. Thus we can satisfy the first condition of Definition 6, and recursively applying this idea one can transform any pair of computational graphs into the pair which has the same structure. However, if the computational graphs of the system are too different in different points, it can make no sense to represent a system as a PDE, since it means that the dynamics of different parts of the system has nothing in common.

4.4.4 Systems with boundaries

Now let us look at the Heat PDE:

$$\frac{\partial \rho}{\partial t} = \frac{\partial^2 \rho}{\partial x^2}. \tag{4.46}$$

Imagine that this equation is defined on an interval $x \in [0, +\infty)$, that is there is a boundary in the point $x = 0$.

There are two types of boundary conditions (or BC) which can be supplied to provide a well-posed boundary value problem. For example for some $a \in \mathbb{R}$,

$$\begin{aligned} &1) \quad \textit{Dirichlet} \text{ BC: } \rho(0) = a, \\ &2) \quad \textit{Neumann} \text{ BC: } \partial\rho/\partial x\,(0) = a. \end{aligned} \tag{4.47}$$

There can also exist a linear combination of these boundary conditions, called *Robin* BC.

If the Heat Equation (4.46) is discretized in stencil points $\{i-1, i, i+1\}$, the result is

$$\dot{\rho}_i = \frac{1}{\Delta x^2}\left(\rho_{i-1} - 2\rho_i + \rho_{i+1}\right). \tag{4.48}$$

Assume now that there exists $i_0 = 1$ such that $x_{i_0-1} = 0$. Depending on the type of boundary conditions, the equation for the state ρ_1 can be obtained by the discretization of a boundary value problem (4.46)-(4.47) in two ways:

$$\begin{aligned} &1) \quad \text{Dirichlet BC: } \dot{\rho}_1 = \left(a - 2\rho_1 + \rho_2\right)/\Delta x^2, \\ &2) \quad \text{Neumann BC: } \dot{\rho}_1 = \left(\rho_2 - \rho_1\right)/\Delta x^2 - a/\Delta x. \end{aligned} \tag{4.49}$$

Now imagine the system (4.46) being obtained by the continuation process from the system (4.48). We can notice that every state of (4.48) is governed by the same dynamics except for the boundary state ρ_1. The question is how to recover the boundary conditions (4.47) for the PDE from the dynamics of ρ_1 in (4.49).

This indeed can be done if one assumes that there exists a "ghost cell" ρ_0 such that it has no dynamics, but is algebraically connected with adjacent states. With a proper definition of ρ_0 the equation for $\dot{\rho}_1$ can be represented in the same way as for other states (4.48) and

thus has the same continuation (4.46). For example, algebraic equations for ρ_0 representing (4.48)-(4.49) are

$$\begin{aligned} &1) \quad \text{Dirichlet BC: } \rho_0 = a, \\ &2) \quad \text{Neumann BC: } \rho_0 = \rho_1 - a\Delta x. \end{aligned} \tag{4.50}$$

The ghost cell $\rho_0 = a$ for the case of Dirichlet BC is depicted in Fig. 4.6. Notice that equations (4.50) can be directly continualized, obtaining (4.47).

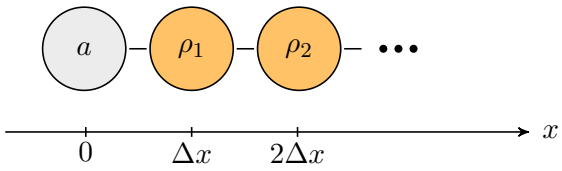


Figure 4.6: Boundary of the system (4.48) with Dirichlet boundary condition (4.49), represented by a ghost cell $\rho_0 = a$.

This procedure can be generalized to any ODE system: once the states near boundaries change their dynamics with respect to the general governing equation, this change can be represented by "ghost cells" with algebraic dependences on the "real" states. Continualizing these algebraic equations leads to the boundary conditions for the obtained PDE.

4.4.5 Systems with moving agents

Usually PDEs have derivatives written with respect to the time and space variables, thus their physical meaning is in the function continuously varying in time and space. However, in general no one prevents us from writing a PDE with derivatives with respect to some other variables.

Assume a physical system is given by a set of interacting agents, with agents being indexed by an integer index $i \in \mathbb{Z}$ (a general multiindex space $\mathbb{Z}^n$ can also be used). Let an agent i have a state ρ_i. The index variable i is by definition discrete. However we can make an assumption that in between of two agents with consecutive indexes i and $i+1$ there is a continuum of virtual agents having state varying from ρ_i to ρ_{i+1}. Denoting this continuously varying index by $M \in \mathbb{R}$ we can say that the state of the system ρ is a continuous and smooth function $\rho(t, M)$ with the property $\rho(t, i) = \rho_i(t)$. In fact it appears that this definition of M coincides with the definition of Moskowitz function which is used to describe the number of vehicles passed through a fixed point in traffic modeling (see Newell 1993) and which was already introduced in Section 3.2.4 to describe density evolution in conservation laws.

Once the index variable is continuous, we can think about it as a new space variable. Thus it is possible to use a continuation described in previous sections, where the distance between two consecutive agents is obviously $\Delta M = 1$. The derivatives of the state with respect to the index can be obtained by continuation, for example $\rho_{i+1} - \rho_i \to \partial\rho/\partial M$.

PDEs with index derivatives are very useful in multi-agent setups, when states of the

agents are represented by their positions. Examples include traffic systems as in Molnár et al. 2019 with agents being cars, or systems of interacting particles and robot formations which will be discussed in Chapter 5.

4.4.6 Algorithm for general continuation procedure

The general continuation procedure for different kinds of systems can be summarized in the Algorithm 4.1. It checks for boundedness and space-dependency of the system and uses nonlinear continuation based on computational graphs. In case of multi-agent systems indices are treated as space coordinates. Linear systems are also covered by the algorithm since their computational graph is trivial.

4.5 Continuation for large-scale linear networks

Large-scale networks are often used to describe physical systems such as urban traffic, brain activity or power networks. Entities in these systems have a predefined position in the real-world space representing individual nodes in the network, and interconnections between different nodes often have a property of locality meaning that nodes which are close in the real world are connected stronger than those which are far apart. Due to these properties various types of large-scale networks can be transformed to PDEs by continuation.

In this section we focus on the linear network analysis. It is widely known (Frihauf and Krstic 2010; Jafarizadeh 2020) that the Laplacian consensus networks are closely related to the diffusion PDEs. Performing continuation of a general linear spatially-distributed network we show how the network dynamics can be written as a linear second-order PDE with space-dependent coefficients. In particular we show that several additional properties such as absence of self-loops, regularity or undirected topology of the network can simplify the resulting PDE.

4.5.1 Model

Assume the system is given by a linear model with $\rho_i \in \mathbb{R}$ being the state of i-th agent and $x_i \in \mathbb{R}^n$ being its spatial position. We can assume that every agent i is influenced by its neighbourhood $\mathcal{N}_i$ and also has its own dynamics:

$$\dot{\rho}_i = a_{ii}\rho_i + \sum_{j \in \mathcal{N}_i} a_{ij}\rho_j. \tag{4.51}$$

Note that (4.51) is a space-dependent multidimensional generalization of (4.1). Also we assume that ghost cells are added to the system (4.51) to ensure boundary conditions as in Section 4.4.4. A particular choice would be to have a set of boundary nodes, placed on the boundaries of the domain, with either fixed or controlled states.

Algorithm 4.1: General continuation procedure

```
Input: System of ODEs, d
    if system consists of moving agents then
        treat indices as coordinate space;                          // Sec 4.4.5
    end
    if system has boundaries then                                   // Sec 4.4.4
        create ghost cells on boundaries such that equations
            for all nodes become homogeneous;
    end
    if system is space-dependent then                               // Sec 4.4.3
        for each node do
            build computational graph;
        end
        find the most general structure of the computational graph;
        for each node do
            adjust computational graph such that it has
                the same structure with others;
            continuation();
        end
        approximate PDE coefficients by space-dependent functions;
    else
        build computational graph;
        continuation();                          // same continuation for all nodes
    end
Procedure continuation()
    Input: Computational graph of ODE, d
    for each node in graph, starting from leaves, do
        for each group of children with similar subgraphs do        // Sec 4.3
            compute PDE coefficients by (4.13) using d;
            replace group by PDE;
        end
    end
```

Based on Theorems 4.2 and 4.3 and on the discussion in Section 4.2.6, we choose the order of continuation $d = 2$ to study transportation and diffusion properties of large-scale network.

4.5.2 Continuation

Using a second-order approximation, continuation of the state ρ_j at the point x_i can be performed in the following way:

$$\rho_j = \rho(x_j) \approx \rho(x_i) + (x_j - x_i)^T \cdot \nabla\rho + \frac{1}{2}(x_j - x_i)^T \frac{\partial^2 \rho}{\partial x^2}(x_j - x_i), \tag{4.52}$$

or using the property of trace:

$$\rho_j = \rho(x_j) \approx \rho(x_i) + (x_j - x_i)^T \cdot \nabla\rho + \frac{1}{2}\operatorname{Tr}\left((x_j - x_i)(x_j - x_i)^T \frac{\partial^2 \rho}{\partial x^2}\right), \tag{4.53}$$

which leads to the PDE transformation of (4.51), which can be written at agents' positions as

$$\begin{aligned} \frac{\partial \rho}{\partial t} = &\left[a_{ii} + \sum_{j \in \mathcal{N}_i} a_{ij}\right] \rho + \left[\sum_{j \in \mathcal{N}_i} a_{ij}(x_j - x_i)^T\right] \cdot \nabla\rho + \\ &+ \operatorname{Tr}\left(\left[\frac{1}{2}\sum_{j \in \mathcal{N}_i} a_{ij}(x_j - x_i)(x_j - x_i)^T\right] \frac{\partial^2 \rho}{\partial x^2}\right). \end{aligned} \tag{4.54}$$

Define $\lambda(x) \in \mathbb{R}$, $b(x) \in \mathbb{R}^n$ and $\varepsilon(x) \in \mathbb{R}^{n \times n}$ such that

$$\begin{aligned} \lambda(x_i) \approx \left[a_{ii} + \sum_{j \in \mathcal{N}_i} a_{ij}\right], \qquad b(x_i) \approx \left[\sum_{j \in \mathcal{N}_i} a_{ij}(x_j - x_i)\right], \\ \varepsilon(x_i) \approx \left[\frac{1}{2}\sum_{j \in \mathcal{N}_i} a_{ij}(x_j - x_i)(x_j - x_i)^T\right], \end{aligned} \tag{4.55}$$

thus these functions are found by a continuous approximation of coefficients of (4.54). With the help of these functions we finally formulate the main continuation result:

Theorem 4.5. *The continuation of a linear network (4.51) is given by*

$$\frac{\partial \rho}{\partial t} = \lambda(x)\rho + b(x)^T \cdot \nabla\rho + \operatorname{Tr}\left(\varepsilon(x)\frac{\partial^2 \rho}{\partial x^2}\right), \tag{4.56}$$

where $\lambda(x)$, $b(x)$ and $\varepsilon(x)$ are given by (4.55).

Remark 4.3. Note that if $a_{ij} > 0$ then the matrix inside of the trace is positive-semidefinite, which means that under suitable affine transformation of local coordinates the second-order term can be represented as a stable Laplacian diffusion. This corresponds to $c_2 > 0$ in (4.14), required for absence of artificial instability by Corollary 4.4.

4.5.3 Particular network structures

It is possible to derive several important corollaries for different classes of networks:

Corollary 4.5 (Laplacian network)**.** *If the original system* (4.51) *depends only on the differences of states*

$$\dot{\rho}_i = \sum_{j \in \mathcal{N}_i} a_{ij}(\rho_j - \rho_i),$$

then (4.56) *has* $\lambda(x) \equiv 0$.

Proof. This property corresponds to the fact that the network has no self-loops. For the Laplacian network $a_{ii} = -\sum_{j \in \mathcal{N}_i} a_{ij}$, thus by (4.55) $\lambda(x) \equiv 0$. □

Corollary 4.6 (Symmetric network)**.** *If the original system is symmetric, that is for every* $j \in \mathcal{N}_i$ *there exists such* $j' \in \mathcal{N}_i$ *that* $x_j - x_i = -(x_{j'} - x_i)$ *and* $a_{ij} = a_{ij'}$*, then* $b(x) \equiv 0$.

Proof. Straightforward by (4.55). □

Corollary 4.7 (Undirected regular network)**.** *If* $a_{ij} = a_{ji}$ *for all* i, j *and if for every* $j \in \mathcal{N}_i$ *there exists such* $j' \in \mathcal{N}_i$ *that* $x_j - x_i = -(x_{j'} - x_i)$*, then* (4.56) *can be represented in the form*

$$\frac{\partial \rho}{\partial t} = \lambda(x)\rho + \nabla \cdot \left(\varepsilon(x) \frac{\partial \rho}{\partial x} \right), \tag{4.57}$$

Proof. Indeed, by taking the derivative we see that

$$\nabla \cdot \left(\varepsilon(x) \frac{\partial \rho}{\partial x} \right) = (\nabla \cdot \varepsilon(x)) \cdot \nabla \rho + \operatorname{Tr} \left(\varepsilon(x) \frac{\partial^2 \rho}{\partial x^2} \right), \tag{4.58}$$

and it remains to prove that $b^T = \nabla \cdot \varepsilon$.

Since $a_{ij} = a_{ji}$, we can assume that there exists some continuous function $\alpha(x, \bar{n})$ dependent on the coordinate x and the direction $\bar{n}$ which is even with respect to the direction such that

$$a_{ij} = \alpha \left(\frac{x_i + x_j}{2}, \frac{x_j - x_i}{||x_j - x_i||} \right) = \alpha \left(\frac{x_i + x_j}{2}, \frac{x_i - x_j}{||x_i - x_j||} \right) = a_{ji}$$

Denote $\bar{n}_j = (x_i - x_j)/||x_i - x_j||$. Further, define $y = x_i$ to be the point where the function α is investigated, thus $\alpha\left((x_i + x_j)/2, \bar{n}_j\right) = \alpha\left(y + (x_j - x_i)/2, \bar{n}_j\right)$. We can now take the Taylor expansion of this function with respect to the coordinate:

$$\alpha \left(y + \frac{x_j - x_i}{2}, \bar{n}_j \right) \approx \alpha(y, \bar{n}_j) + \frac{1}{2} \nabla \alpha(y, \bar{n}_j) \cdot (x_j - x_i).$$

Inserting this expansion into the definition of $b(x)$ we obtain

$$\begin{aligned} b(y)^T &= \sum_{j \in \mathcal{N}_i} \alpha(y, \bar{n}_j)(x_j - x_i)^T + \frac{1}{2} \sum_{j \in \mathcal{N}_i} \nabla \alpha(y, \bar{n}_j) \cdot (x_j - x_i)(x_j - x_i)^T = \\ &= \frac{1}{2} \sum_{j \in \mathcal{N}_i} \nabla \alpha(y, \bar{n}_j) \cdot (x_j - x_i)(x_j - x_i)^T, \end{aligned} \tag{4.59}$$

since the first sum vanishes because for every j there exists j' such that $(x_j - x_i) = -(x_{j'} - x_i)$ and $\alpha(y, \bar{n}_j) = \alpha(y, \bar{n}_{j'})$. Now, analyzing $\varepsilon(x)$, we get

$$\begin{aligned}\varepsilon(y) =& \frac{1}{2} \sum_{j \in \mathcal{N}_i} \alpha(y, \bar{n}_j)(x_j - x_i)(x_j - x_i)^T + \\ &+ \frac{1}{4} \sum_{j \in \mathcal{N}_i} (x_j - x_i) \nabla \alpha(y, \bar{n}_j)(x_j - x_i)(x_j - x_i)^T = \\ =& \frac{1}{2} \sum_{j \in \mathcal{N}_i} \alpha(y, \bar{n}_j)(x_j - x_i)(x_j - x_i)^T,\end{aligned} \tag{4.60}$$

where the second sum vanishes by the same reasons. Now it is clear that taking the divergence of (4.60) with respect to y one ends up with (4.59), which finishes the proof. □

4.6 Concluding remarks

In this chapter we presented a general process of transformation of ODE systems into their PDE counterparts via the continuation method. Performing analysis of the continuation for linear systems, we found conditions for the continuation to be reversible, meaning that the original ODE system could be obtained from the PDE version by a correct discretization. We have further shown that the spectrum of PDE converges to the spectrum of the original ODE as the order of continuation grows, and that this convergence provides a bound on the deviation between systems' solutions. The continuation method was then elaborated for many classes of systems including nonlinear, multidimensional, space- and time-varying systems, indexed multi-agent systems and systems with boundaries. Based on this method, new continuous models can be derived and further utilized for analysis and control purposes.

In the next two chapters we will focus on applications of the continuation method to various systems. In Chapter 5 it will be shown that the continuation can be a helpful tool for the analysis and control design for multi-agent systems since it allows to recover their density-based continuous representations. In Chapter 6 the method will be applied to networks of oscillators, and resulting PDE models will be used to derive conditions which assure existence and stability of synchronous oscillatory behaviours in the networks.

As a future direction of research it would be desirable to investigate the continuation method in more details, as there are many problems that were not covered in the present manuscript:

- Convergence properties of the continuation method could be analysed for more classes of systems. In particular a generalization of Theorem 4.3, which was proven here only for linear spatially invariant systems, could provide guarantees on solutions for much broader class of nonlinear and space-dependent systems.
- Theorems 4.1, 4.2 and 4.3 show that the obtained system approximates the original ODE well enough as soon as the order of continuation is sufficiently high. However

due to possible artificial instabilities and difficulties in analysis of high-order PDEs it is often makes more sense to use only first- and second-order PDEs. Future research could provide deeper understanding of applicability of the method for these particular orders, stating explicitly which properties can be reconstructed with low-order PDEs and which systems are better suitable for this continuation.

- Finally, in Section 4.3 a computational graph was introduced to describe dynamics of a nonlinear system. In addition to the graph itself it was shown that a *position function* should be specified for every node in the graph to be able to perform continuation. It appears that a designer has some freedom in choosing this position function, which in turn leads to different PDE models. It would be desirable to study the dependence of the quality of the obtained model on the chosen position function and to provide some justified guidelines on how to make this choice.

Chapter 5

Applications of the continuation method to multi-agent systems

Contents

5.1 Where and why the method is useful

In the previous chapter we presented the continuation method for ODE-based systems which are spatially distributed and which have position-dependent interactions. The method replaces the original spatially distributed ODE system by a continuous PDE whose state and space variables preserve the state and space variables of the original system. Here we will focus on applications of this method to different classes of systems.

The idea of substituting finite differences with partial derivatives was already used in several particular applications for analysis purposes. For example, Barooah, Mehta, and

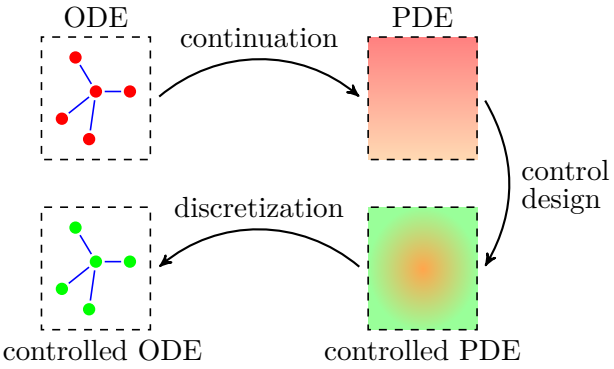


Figure 5.1: Proposed framework for control design based on the continuation method and a continuous representation of the system.

Hespanha 2009 derived a PDE model for the controlled platooning system, and consensus lattice networks were transformed into PDEs by Biccari, Ko, and Zuazua 2019.

The advantage of the continuation method is that it allows to recover a PDE which describes the same physical system as the original ODE network. Such a description can be very helpful not only for analysis, but also for control purposes. Indeed, one can use an obtained PDE to design a continuous control which, being discretized back, results in a control law for the original ODE system: this design framework is illustrated in Fig. 5.1.

Multi-agent systems are of particular interest here. If the system is given by a set of indexed agents with given interaction topology, than by Section 4.4.5 we can treat their interaction as a PDE with respect to indices. This can have long-lasting implications and allow for derivation of many important results. Indeed, one can imagine a system with agents whose state vector includes a one-dimensional position x. Introducing continuous index function M as in Section 4.4.5, we can write index derivative of a state as $\partial x/\partial M$. But its inverse is $\partial M/\partial x$, which in other words can be read as "a number of agents present in a unit of length". Therefore we can use it to define a *density*, $\rho := \partial M/\partial x$, and write a PDE describing density evolution for the system's dynamics.

In this chapter we will cover several examples of this transformation, using the continuation method to obtain density evolution PDEs from the original agent-defined systems. In Section 5.2, which is devoted to the urban traffic applications, we show with an example how a simple car-following law can be transformed to a density-based PDE conservation law describing traffic evolution, i.e. LWR model. We also show that LWR model can be recovered from a discrete Cell Transmission Model (CTM), and that a similar technique can lead to a derivation of a two-dimensional model for urban traffic. Next, in Section 5.3 we show that this transformation can be also useful for multidimensional systems. Here we present a transformation of a system of infinite number of interacting particles in n-dimensional space, recovering Euler equations for the fluid dynamics. Thus our method can be seen as an original solution to the Hilbert's 6th problem. Finally, Section 5.4 is devoted to the practical

application of the method: reusing the derivation of Euler equations, we obtain a PDE model for a large formation of flying drones. We use this PDE to design a control law on a density level and then discretize it back to be able to implement it on every drone in accordance with the scheme in Fig. 5.1. It is finally shown in a numerical simulation that it is possible to control the formation to perform desired maneuvers both in 2D and in 3D.

5.2 Applications for traffic systems

In 1950s Lighthill and Whitham 1955 and Richards 1956 presented a first model for traffic description, which is now called LWR model. It describes traffic as a fluid using a conservation law PDE. State of the system is a density of cars at every time and space point $\rho(t,x)$ for $x \in [0,L]$ where L is a length of the road, and $t \in \mathbb{R}^+$. LWR model predicts that evolution of the density is caused by a flow $\phi(t,x)$:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial \phi(t,x)}{\partial x} = 0. \tag{5.1}$$

The main assumption of this model is the existence of the *fundamental diagram* which couples density with flow:

$$\phi(t,x) = \Phi(\rho(t,x)). \tag{5.2}$$

Function $\Phi(\rho)$ is a concave function, which equals zero either at $\rho = 0$ (meaning there is no cars) or at $\rho = \rho_{max}$ (meaning cars are in complete traffic jam). It possesses a unique maximum at $\phi_{max} = \Phi(\rho_{crit})$, with ρ_{crit} being called a *critical density*. Since it produces a maximal flow, the critical density serves as an optimal operation point for traffic systems. The fundamental diagram relation is often established experimentally. Popular analytical choices include:

- Triangular diagram: $\Phi(\rho) = \min\{v_{max}\rho, \omega(\rho_{max} - \rho)\}$,
- Greenshields diagram: $\Phi(\rho) = v_{max}\rho\left(1 - \frac{\rho}{\rho_{max}}\right)$,
- Exponential diagram: $\Phi(\rho) = v_{max}\rho\left(1 - e^{\alpha\left(1-\frac{\rho_{max}}{\rho}\right)}\right)$.

In all these diagrams ρ_{max} denotes the maximal possible density and v_{max} denotes the free velocity of cars in absence of other cars. Also ω denotes the backward kinematic wave speed (which is a speed with which a traffic jam propagates), and α defines the skewness of the diagram.

5.2.1 From single car model to LWR model

Here we will derive LWR equation from the motion of individual vehicles. Assume each car has a number i, a position x_i (i.e. measured at a front bumper) and a length l. We will assume that all cars have the same length.

A car can control its velocity depending on a difference between its position and a position of the rear bumper of the next car, which is $x_{i+1} - l$. The car i uses very simple driving law: if the next car is far away, it drives with velocity v_{max}, but if the next car is closer that some distance s_{crit}, the car i starts decreasing its speed linearly with a gain γ, reaching zero velocity at the safety distance h. Denote $s_{stop} = l + h$, that is s_{stop} is a stopping distance between front bumpers of two successive cars. Therefore, the equation of motion can be written as

$$\dot{x}_i = \min\{v_{max}, \gamma(x_{i+1} - x_i - s_{stop})\}, \tag{5.3}$$

and s_{crit} satisfies the relation $v_{max} = \gamma(s_{crit} - s_{stop})$. We also need to assume that all cars are initially spaced with distances greater or equal than s_{stop}.

Now, according to Section 4.4.5, let us define a continuous function to numerate cars as $M(t, x)$, which defines an index of a car at the position x at time t, that is $M(t, x_i) = i$. Equivalently, $M(t, x)$ is a number of cars from the beginning of the road up to a position x. Then we can apply the continuation method for $x(t, M)$, i.e. with position as a state and with index and time as dependent variables. In particular, we substitute $x_{i+1} - x_i$ with $\partial x/\partial M$.

First-order PDE approximation for (5.3) is then just

$$\frac{\partial x}{\partial t} = \min\left\{v_{max}, \gamma\left(\frac{\partial x}{\partial M} - s_{stop}\right)\right\}. \tag{5.4}$$

Now by definition of functions $M(t, x)$ and $x(t, M)$ we have that

$$x(t, M(t, x)) \equiv x, \tag{5.5}$$

from which we immediately obtain

$$\frac{\partial x}{\partial M}\frac{\partial M}{\partial x} = 1, \qquad \frac{\partial x}{\partial t} + \frac{\partial x}{\partial M}\frac{\partial M}{\partial t} = 0. \tag{5.6}$$

Using these relations, we obtain a PDE for the index function:

$$\frac{\partial M}{\partial t} = -\min\left\{v_{max}, \gamma\left(\frac{1}{\frac{\partial M}{\partial x}} - s_{stop}\right)\right\}\frac{\partial M}{\partial x} = -\min\left\{v_{max}\frac{\partial M}{\partial x}, \gamma\left(1 - s_{stop}\frac{\partial M}{\partial x}\right)\right\}, \tag{5.7}$$

where the second equality comes from the fact that index can only increase with position. Finally, let us define the density and the flow:

$$\rho(t, x) := \frac{\partial M}{\partial x}, \quad \phi(t, x) := -\frac{\partial M}{\partial t}. \tag{5.8}$$

Then, for consistency, these new variables should satisfy

$$\frac{\partial^2 M}{\partial x \partial t} = \frac{\partial^2 M}{\partial t \partial x}, \tag{5.9}$$

which reads as

$$\frac{\partial \rho}{\partial t} + \frac{\partial \phi}{\partial x} = 0. \tag{5.10}$$

Moreover, by the definition of $\phi(x,t)$ and by (5.7) the flow should satisfy

$$\phi(t,x) = \min\left\{v_{max}\rho(t,x), \gamma\left(1 - s_{stop}\rho(t,x)\right)\right\}, \tag{5.11}$$

or

$$\phi(t,x) = \Phi(\rho(t,x)) = \min\{v_{max}\rho(t,x), \omega(\rho_{max} - \rho(t,x))\}, \tag{5.12}$$

where $\omega = \gamma s_{stop}$ and $\rho_{max} = 1/s_{stop}$. Therefore, we obtain equation (5.10), which is exactly the LWR equation, and the relation between flow and density (5.12), which corresponds to the Triangular fundamental diagram. Similar idea of substituting finite differences with partial derivatives was used in Molnár et al. 2019; Molnár et al. 2020 to derive LWR model with delayed interaction between cars.

An example of the transformation of a driver model to the fundamental diagram is shown in the first row of Fig. 5.2. Function $V(s)$ corresponds to the driver model in (5.3) with $s_{stop} =$ 5.5m, $s_{crit} =$ 20m and $v_{max} =$ 60km/h. After the transformation we obtain the Triangular fundamental diagram with $\phi_{max} =$ 3000veh/h, $\rho_{crit} =$ 50veh/km and $\rho_{max} =$ 182veh/km.

This method can be applied to any driver model, not only to (5.3), assuming the car's velocity can be directly controlled. In general, we can say that the driver control policy depends on the difference between the positions of two cars, thus given a function $V(s)$ for the desired velocity at each distance,

$$\dot{x}_i = V(x_{i+1} - x_i). \tag{5.13}$$

Using the first-order PDE approximation, one obtains

$$\frac{\partial x}{\partial t} = V\left(\frac{\partial x}{\partial M}\right). \tag{5.14}$$

Now, using (5.6) we rewrite this PDE with $M(t,x)$ as a state:

$$\frac{\partial M}{\partial t} = -V\left(\frac{1}{\frac{\partial M}{\partial x}}\right)\frac{\partial M}{\partial x}. \tag{5.15}$$

Finally, defining $\rho(t,x)$ and $\phi(t,x)$ as in (5.8), we obtain LWR equation with $\phi(t,x) = \Phi(\rho(t,x))$ and with a fundamental diagram

$$\Phi(\rho) = V\left(\frac{1}{\rho}\right)\rho. \tag{5.16}$$

Note that the derivation of (5.16) from (5.13) can be reversed, thus it is possible to obtain driver model from a fundamental diagram. To support this, let us derive the driver model for the Greenshields diagram,

$$\Phi(\rho) = v_{max}\rho\left(1 - \frac{\rho}{\rho_{max}}\right). \tag{5.17}$$

Assuming equality of (5.16) and (5.17),

$$V\left(\frac{1}{\rho}\right)\rho = v_{max}\rho\left(1 - \frac{\rho}{\rho_{max}}\right), \tag{5.18}$$

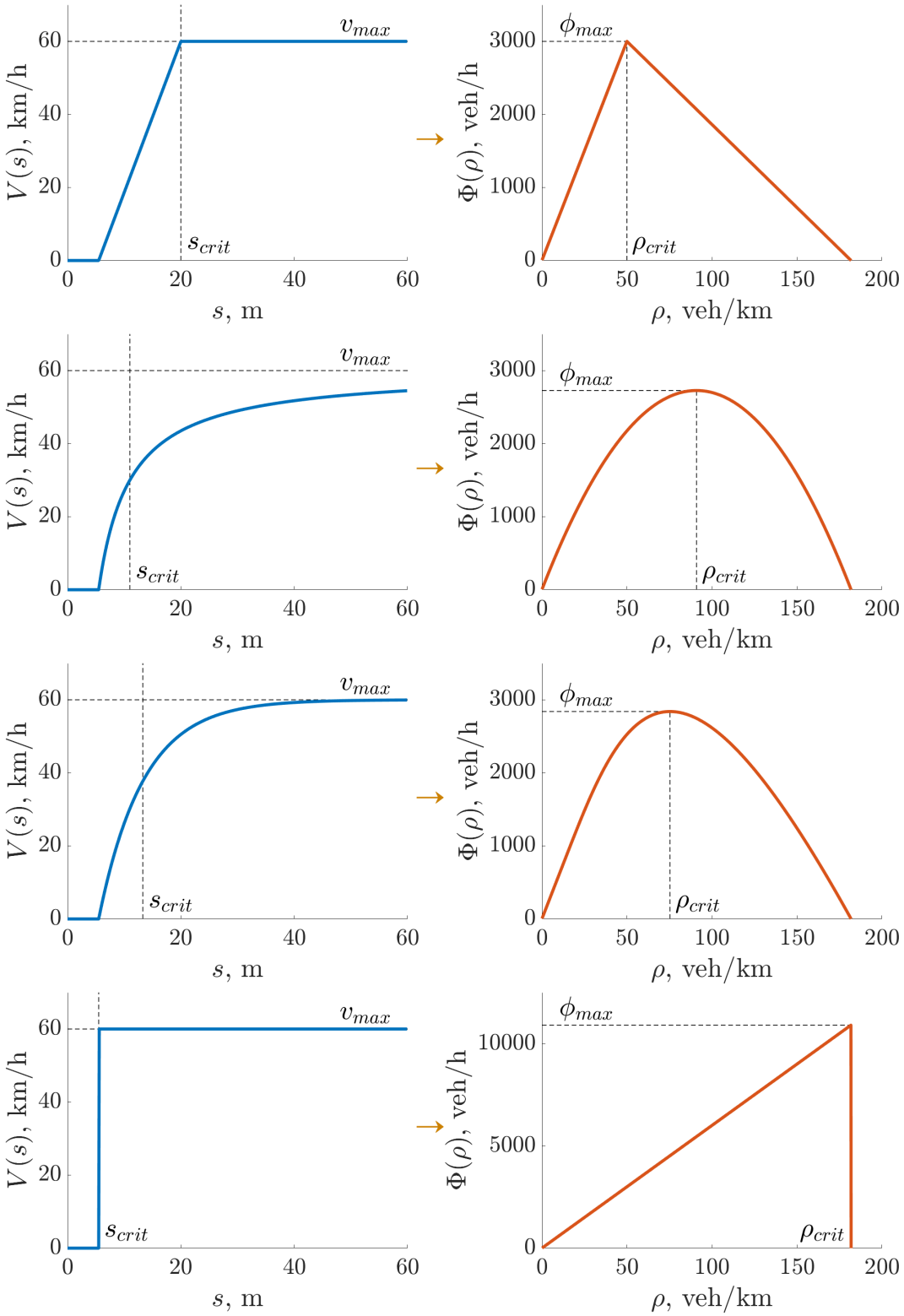


Figure 5.2: Driver models and their corresponding fundamental diagrams for LWR model. **First row:** Triangular fundamental diagram (5.3)-(5.12). **Second row:** Greenshields fundamental diagram (5.17)-(5.20). **Third row:** Exponential fundamental diagram (5.21)-(5.22). **Fourth row:** Triangular fundamental diagram producing maximal possible flow assuming autonomous cars can stop immediately.

we get (using the definition of $\rho_{max} = 1/s_{stop}$)

$$V(s) = v_{max}\left(1 - \frac{s_{stop}}{s}\right), \tag{5.19}$$

which than by (5.13) leads to a driver model

$$\dot{x}_i = v_{max}\left(1 - \frac{s_{stop}}{x_{i+1} - x_i}\right). \tag{5.20}$$

This is a continuous function which strictly increases with a distance between cars, which has an asymptotic value v_{max} and which reaches zero when the distance is $x_{i+1} - x_i = s_{stop}$.

For the Exponential fundamental diagram the similar result is obtained. Starting from the fundamental diagram itself

$$\Phi(\rho) = v_{max}\rho\left(1 - e^{\alpha\left(1 - \frac{\rho_{max}}{\rho}\right)}\right), \tag{5.21}$$

we end up with a driver model

$$\dot{x}_i = v_{max}\left(1 - e^{\alpha\left(1 - \frac{x_{i+1}-x_i}{s_{stop}}\right)}\right). \tag{5.22}$$

Transformations for the Greenshields and for the Exponential fundamental diagrams are shown in the second and third rows of Fig. 5.2. For the Exponential fundamental diagram skewness $\alpha = 0.7$ was chosen.

One can establish a distance $s_{crit} = 1/\rho_{crit}$ corresponding to the critical density ρ_{crit} which produces maximal flow. Thus s_{crit} is an optimal distance for cars to keep. Interesting to notice that for the Triangular diagram $V(s_{crit})$ always corresponds to the maximal velocity, while for the other diagrams it is significantly less.

Moreover, one can ask a question what should be a driver model such that the flow reaches maximum possible value. Keeping v_{max} and ρ_{max} fixed, the optimal fundamental diagram would be the one where cars can stop immediately at s_{stop}, driving with maximal velocity v_{max} otherwise (see the fourth row of Fig. 5.2). Their reaction distance s_{crit} should be as small as possible, which in turn translates as ρ_{crit} coinciding with ρ_{max}. If one takes $v_{max} = 60$km/h and $s_{stop} = 5.5$m, using this diagram the resulting flow reaches a value almost four times larger than the flows using other diagram types. Therefore it is always preferable to use as low reaction distance s_{crit} and as high velocity $V(s_{crit})$ as possible, provided it is still safe. This design paradigm can be implemented on large platoons of communicating autonomous vehicles, which can follow one another at very small distances and at very high velocities.

Continuation method gives a straightforward way of obtaining LWR model from single car dynamics, provided cars can directly control their velocities. In some scenarios this assumption is unrealistic, therefore in theory it would be possible to use the same method to derive second-order PDEs describing traffic flow of acceleration-controlled cars. We are not digging into this problem here, but this would constitute a promising future direction of research.

5.2.2 From Cell Transmission Model to LWR model

Another way of discrete representation of traffic on a road is a Cell Transmission Model (CTM) describing by the following equation:

$$\dot{\rho}_i = \frac{1}{\Delta x}\left(\min(D(\rho_{i-1}), S(\rho_i)) - \min(D(\rho_i), S(\rho_{i+1}))\right). \tag{5.23}$$

The road is assumed to be split into cells of length Δx, each cell i having a density of cars ρ_i inside. Further, function $D(\rho)$ is an increasing demand function, and $S(\rho)$ is a decreasing supply function. Term $\min(D(\rho_{i-1}), S(\rho_i))$ represents a flow which can physically pass from cell $i-1$ to cell i, since the flow of cars that want to enter the cell i is defined by $D(\rho_{i-1})$, and the flow of cars which can be accepted by the cell i is defined by $S(\rho_i)$.

This model is often used to describe propagation of traffic in a discrete way for optimization purposes and can be considered as a dicretization of the LWR model. We will show that this model can also be trivially transformed back into the LWR model by performing a continuation process up to the order $d = 1$. Using computation graph formalism we can define subexpression

$$g_{i-1/2} = g(\rho_{i-1}, \rho_i) = \min(D(\rho_{i-1}), S(\rho_i)), \tag{5.24}$$

computed at the average position of its leaf nodes. With the help of this subexpression we rewrite (5.23) as

$$\dot{\rho}_i = \frac{1}{\Delta x}\left(g_{i-1/2} - g_{i+1/2}\right) \tag{5.25}$$

which is approximated up to the first order by

$$\frac{\partial \rho_i}{\partial t} = -\frac{\partial g_i}{\partial x}. \tag{5.26}$$

Then,

$$g_i = \min(D(\rho_{i-1/2}), S(\rho_{i+1/2})) \approx \min\left(D\left(\rho_i - \frac{\Delta x}{2}\frac{\partial \rho}{\partial x}\right), S\left(\rho_i + \frac{\Delta x}{2}\frac{\partial \rho}{\partial x}\right)\right). \tag{5.27}$$

But the equation (5.26) already has a first-order derivative, thus we should keep only zeroth-order terms inside of the subexpression g_i. Therefore,

$$\frac{\partial \rho}{\partial t} = -\frac{\partial \min(D(\rho), S(\rho))}{\partial x} = -\frac{\partial \Phi(\rho)}{\partial x}, \tag{5.28}$$

where $\Phi(\rho) = \min(D(\rho), S(\rho))$. Finally we can notice that equation (5.28) is indeed the LWR model describing the continuous evolution of traffic, and $\Phi(\rho)$ defines a fundamental diagram which is concave since it is defined as a minimum between increasing demand and decreasing supply functions.

5.2.3 From urban traffic network to multidimensional PDE model

While the previous sections suggested a way of transformation of discrete models into continuous ones for a single road, it is possible to apply the continuation method to obtain a continuous model describing an evolution of traffic in the 2D plane.

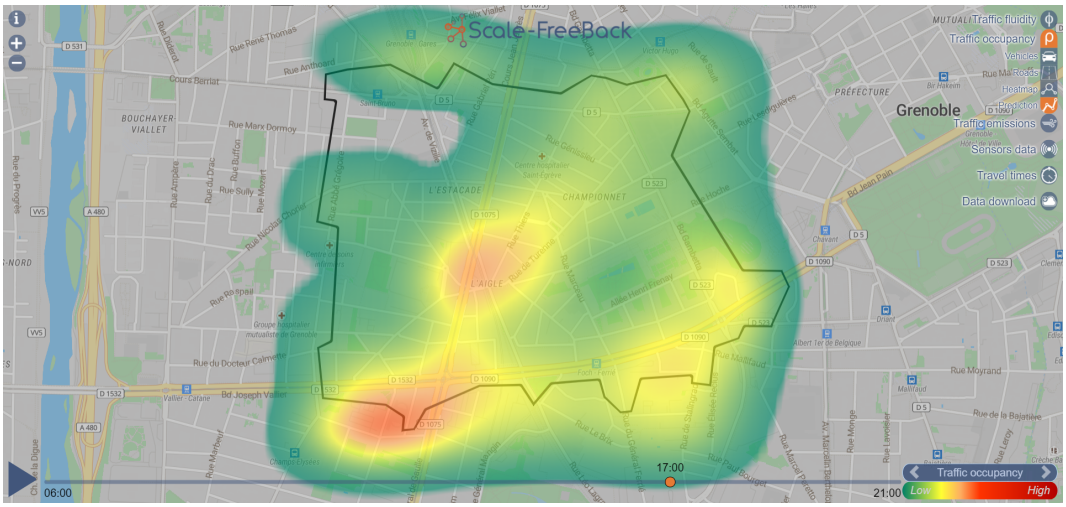


Figure 5.3: Prediction of traffic in Grenoble with NEWS PDE model (5.30). Screenshot is taken from GTL Ville web-application developed within Scale-FreeBack project, see gtlville.inrialpes.fr.

Given a network of roads, we can assume that the evolution of traffic is governed by a system of CTM equations for every road. Namely, for a road i evolution of its density ρ_i is

$$\begin{aligned} \dot{\rho}_i &= \frac{1}{L_i}\left(\phi_i^{in} - \phi_i^{out}\right), \\ \phi_i^{in} &= \min\left\{\sum_{j=1}^{N^{in}} \alpha_{ji} D(\rho_j), S(\rho_i)\right\}, \\ \phi_i^{out} &= \min\left\{D(\rho_i), \sum_{k=1}^{N^{out}} \beta_{ik} S(\rho_k)\right\}, \end{aligned} \tag{5.29}$$

where L_i is a length of the road i and flows ϕ_i^{in} and ϕ_i^{out} define the number of cars per second which enter and exit the road respectively. It is also assumed that cars can enter the road i from roads $j \in \{1, ..., N^{in}\}$, and that the proportion of cars leaving the road j to enter the road i is defined by a *turning ratio* α_{ji}. The same happens with cars leaving the road i to one of the roads $k \in \{1, ..., N^{out}\}$, whose splitting is defined by *supply ratios* β_{ik}.

It appears that under suitable conditions it is possible to approximate behaviour of (5.29) by a continuous model. First, a unique method of intersections representation is designed. This can be done by choosing four cardinal directions N (North), E (East), W (West) and S (South), and projecting every road's dynamics on these directions. Then equations for all roads become similar and a clear spatial structure arises. Note that for every coordinate direction two opposite cardinal directions are used (both N and S, and both E and W) in order to maintain independent descriptions of positive flows on opposite lanes. Using the continuation method this representation can be transformed into a space-dependent PDE

with four-dimensional state $\rho = (\rho_N, \rho_E, \rho_W, \rho_S)$ with evolution

$$\frac{\partial \rho}{\partial t} = \underbrace{\frac{1}{L}\left(\Phi^T - \Phi\right)\mathbf{1}}_{\text{mixing term}} - \underbrace{\frac{\partial(\overline{\cos\theta}\ \phi)}{\partial x} - \frac{\partial(\overline{\sin\theta}\ \phi)}{\partial y}}_{\text{transportation term}}, \tag{5.30}$$

where $L = L(x, y) \in \mathbb{R}$, $\overline{\cos\theta} = \overline{\cos\theta}(x, y) \in \mathbb{R}^{4\times4}$ and $\overline{\sin\theta} = \overline{\sin\theta}(x, y) \in \mathbb{R}^{4\times4}$ are space-dependent parameters describing average length of roads and average projection coefficients respectively in the neighbourhood of the point (x, y). Matrix $\Phi \in \mathbb{R}^{4\times4}$ consists of *partial flows*, e.g. $\Phi_{NE} = \min\{\alpha_{NE}D(\rho_N), \beta_{NE}S(\rho_E)\}$, and flows $\phi \in \mathbb{R}^4$ denote pure directional flows, e.g. $\phi_N = \min\{D(\rho_N), S(\rho_N)\}$. Thus the *mixing term* represents density exchange between different layers of different flow directions, and the *transportation term* represents density movement withing the layer.

For the detailed derivation of (5.30), its analytical properties and its performance validation see Tumash, Canudas-de-Wit, and Delle Monache 2021c; Tumash, Canudas-de-Wit, and Delle Monache 2021b. The model (5.30) was utilized to describe and predict the traffic in the city of Grenoble in the web-application GTL Ville within Scale-FreeBack project. Screenshot with an example of the prediction within GTL Ville is shown in Fig. 5.3.

5.3 Euler equations and Hilbert's 6th problem

5.3.1 Overview

The problem of description of systems of discrete interacting objects by continuous models has a long history. In the beginning of the XX century Hilbert posed his 6th problem (Hilbert 1902), where he suggested to develop a rigorous way leading from the atomistic view to the laws of motion of continua. In particular, the problem can be formulated as a derivation of Euler equations for compressible fluids from the Newton's dynamics of individual particles.

For the most famous case of particles interacting through collision the Boltzmann equation was developed, describing evolution of the joint position-velocity probability distribution of particles. The method of how to transform individual's dynamics into Boltzmann equation is based on the Boltzmann-Grad limit (Gallagher, Saint-Raymond, and Texier 2013), assuming velocities of colliding particles being independent. The following transformation from the Boltzmann equation to the Euler equations uses either Hilbert or Chapman-Erskog expansions with space contration limits (Saint-Raymond 2009; Chapman and Cowling 1990), Grad moments by Grad 1949 or the method of invariant manifolds by Gorban and Karlin 2014.

Another situation arises when the particles interact through long-range forces. In this case the Vlasov equation can be used instead of the Boltzmann equation to describe the joint position-velocity probability distribution. The derivation of the Euler equations from the Vlasov equation was performed by Caprino et al. 1993 using space-contracting limit. In particular it was shown that the resulting system has zero temperature, i.e. the velocities of

individual particles coincide with the velocity field. However, due to the space contraction the particular form of the potential function was lost and the obtained pressure was just a square of the density.

Here we present a derivation of Euler equations directly from the dynamics of individual particles interacting through long-range forces using the continuation method described in previous sections. Contrary to other works, we do not use any kind of limits and we use only one assumption on the isotropy of the space. The assumption requires that for any particle its nearest neighbours are distributed around uniformly in every direction, which can be seen as a counterpart to the molecular chaos hypothesis for the standard derivation of the Boltzmann equation.

5.3.2 System of particles

It is assumed that the fluid consists of small particles interacting with each other, with every particle following simple Newton laws. We will study the system with n space dimensions, and the particles are assumed to have unit mass.

We further assume that there is an interaction between each pair of particles which is given by a force

$$F(x_i - x_j) = \frac{x_i - x_j}{\|x_i - x_j\|} f(\|x_i - x_j\|) = (x_i - x_j)\phi(\|x_i - x_j\|), \tag{5.31}$$

thus the force acts along the line connecting two particles with the smooth magnitude f depending only on the distance between particles. For simplicity we also define a function $\phi(s) = f(s)/s$ representing the scaled magnitude. We will consider an infinite number of particles and an infinitely large space, therefore we should assume that the cumulative force on any particle is finite. In particular for an equally distributed grid this implies that the magnitude of the force should satisfy

$$\int_{\varepsilon}^{+\infty} s^{n-1} f(s) ds < \infty \quad \forall \varepsilon > 0, \tag{5.32}$$

thus the interaction should be fast-decaying.

We then need to enumerate all particles. For this we will use multiindex $i \in \mathbb{Z}^n$. Now let us write the dynamics of a particle with multiindex i using the second Newton's Law:

$$\begin{cases} \dot{x}_i = v_i, \\ \dot{v}_i = \sum\limits_{q \neq 0} F(x_i - x_{i+q}), \end{cases} \tag{5.33}$$

where the summation is performed among all multiindices q in $\mathbb{Z}^n \setminus \{0\}$, since all the particles interact with each other. Both the position x_i and the velocity v_i are vectors in $\mathbb{R}^n$.

5.3.3 Derivation in the Euclidean space

Treating the coordinate x_i as a state and using the idea written in section 4.4.5 we define a multiindex function $M(t,x)$ which is the inverse function of the coordinate: $M(t,x_i) := i$. Likewise, $x(t,i) = x_i(t)$ and thus $x(t,M(t,x)) \equiv x \quad \forall x \in \mathbb{R}^n$. Now let us write a property of inverse function of multiindex as $M(t,x(t,M)) \equiv M \quad \forall M \in \mathbb{R}^n$, where the space for multiindices is continuous by the assumption in section 4.4.5. Taking the time and the index derivatives, we obtain the following very useful relations on Jacobians:

$$\frac{\partial M}{\partial t} + \frac{\partial M}{\partial x}\frac{\partial x}{\partial t} = 0, \tag{5.34}$$

$$\frac{\partial M}{\partial x}\frac{\partial x}{\partial M} = I. \tag{5.35}$$

Equation (5.34) can be seen as a PDE where the function M depends both on t and x. Recalling that the multiindex is assumed to be continuous, we can further utilize the first equation of (5.33) written in a form $\partial x(t,M)/\partial t = v(t,M)$, substitute it in (5.34) and obtain the following equation on the multiindex evolution:

$$\frac{\partial M}{\partial t} = -\frac{\partial M}{\partial x}v(t,M(t,x)) = -\frac{\partial M}{\partial x}u(t,x), \tag{5.36}$$

where the velocity function $u(t,x) = v(t,M(t,x))$ is defined as a velocity of a particle at some given point in space. Finally, taking the derivative with respect to space, we obtain

$$\frac{\partial}{\partial t}\left(\frac{\partial M}{\partial x}\right) = -\frac{\partial}{\partial x}\left(\frac{\partial M}{\partial x}u\right). \tag{5.37}$$

The Jacobian matrix $\frac{\partial M}{\partial x}(t,x)$ represents a *compression tensor*, which measures how close are neighbour particles with respect to different directions in the euclidean space. Evolution of this Jacobian in the euclidean space is described by the matrix PDE (5.37), which is essentially a transport equation with flow velocity given by $u(t,x)$.

Now we approach the second equation in (5.33). It would be desirable to transform it in such a way that we could obtain an evolution equation for the flow velocity $u(t,x)$. First of all, let us rewrite the second equation of (5.33) in a way more suitable for continuation, namely

$$\dot{v}_i = -\sum_{q>0}\left(F(x_{i+q}-x_i) - F(x_i - x_{i-q})\right), \tag{5.38}$$

where the summation is performed among all multiinidices which are greater than zero in lexicographical order, i.e. the first nonzero element of q should be positive.

We can now use the continuation of order 1 on a multidimensional system such that

$$x_{i+q} - x_i \;\to\; \frac{\partial x}{\partial M}\left(t, x_{i+q/2}\right)q, \quad x_i - x_{i-q} \;\to\; \frac{\partial x}{\partial M}\left(t, x_{i-q/2}\right)q,$$

which means that (5.38) becomes

$$\dot{v}_i = -\sum_{q>0}\left(F\left(\frac{\partial x}{\partial M}q\right)_{i+q/2} - F\left(\frac{\partial x}{\partial M}q\right)_{i-q/2}\right).$$

Applying the continuation further to the forces, we obtain

$$F_{i+q/2} - F_{i-q/2} \quad \to \quad \frac{\partial F}{\partial M}(t, x_i)q.$$

Thus (5.38) transforms into

$$\frac{\partial v}{\partial t} = -\sum_{q>0} \frac{\partial}{\partial M} \left(\left[\frac{\partial x}{\partial M} q \right] \phi \left(\left\| \frac{\partial x}{\partial M} q \right\| \right) \right) q, \tag{5.39}$$

where we used a definition of the force (5.31).

Now, we state the following result:

Proposition 5.1. *For any $q \in \mathbb{Z}^n$ and for any smooth scalar field ϕ the following identity holds:*

$$\left[\frac{\partial}{\partial M} \left(\frac{\partial x}{\partial M} q\phi \right) q \right]^T = \nabla \cdot \left(\frac{\partial x}{\partial M} q q^T \frac{\partial x}{\partial M}^T \phi \right) - \left(\nabla \cdot \left(\frac{\partial x}{\partial M} \right) q q^T \frac{\partial x}{\partial M}^T \right) \phi, \tag{5.40}$$

where ∇ denotes a row vector of derivatives with respect to x.

Proof. First, for convenience denote the left-hand side as a vector Q:

$$Q := \frac{\partial}{\partial M} \left(\frac{\partial x}{\partial M} q\phi \right) q = \frac{\partial}{\partial x} \left(\frac{\partial x}{\partial M} q\phi \right) \frac{\partial x}{\partial M} q. \tag{5.41}$$

Also define $h = (\partial x/\partial M)q$. Expanding $\partial(h\phi)/\partial x$, we get

$$Q = h\frac{\partial \phi}{\partial x} h + \frac{\partial h}{\partial x} h\phi = hh^T \frac{\partial \phi}{\partial x}^T + \frac{\partial h}{\partial x} h\phi. \tag{5.42}$$

Now, for any $h \in \mathbb{R}^n$

$$\nabla \cdot (hh^T) = \left(\sum_i h_1 \tfrac{\partial h_i}{\partial x_i} + \sum_i h_i \tfrac{\partial h_1}{\partial x_i} \quad \cdots \quad \sum_i h_n \tfrac{\partial h_i}{\partial x_i} + \sum_i h_i \tfrac{\partial h_n}{\partial x_i} \right),$$

which means that

$$\left(\nabla \cdot (hh^T) \right)^T = \frac{\partial h}{\partial x} h + (\nabla \cdot h) h. \tag{5.43}$$

Therefore the transpose of (5.42) is

$$Q^T = \frac{\partial \phi}{\partial x} hh^T + \nabla \cdot (hh^T)\phi - (\nabla \cdot h) h^T \phi. \tag{5.44}$$

Since for any matrix J and for any scalar field α

$$\nabla \cdot (\alpha J) = \frac{\partial \alpha}{\partial x} J + (\nabla \cdot J)\, \alpha, \tag{5.45}$$

we can simplify (5.44) as $Q^T = \nabla \cdot (hh^T \phi) - (\nabla \cdot h) h^T \phi$. The result of the proposition follows by substituting h and noticing that $\nabla \cdot ((\partial x/\partial M)q) = (\nabla \cdot (\partial x/\partial M))q$. □

Proposition 5.1 allows us to rewrite (5.39) as being dependent only on the euclidean space divergences and the inverse of the compression tensor $\partial M/\partial x$. To finalize the derivation of a complete set of equations, recall the definition of the velocity field $u(t,x) = v(t, M(t,x))$. Taking the time derivative:

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial t} + \frac{\partial v}{\partial M}\frac{\partial M}{\partial t},$$

which by (5.36) is

$$\frac{\partial u}{\partial t} = -\frac{\partial v}{\partial M}\frac{\partial M}{\partial x}u + \frac{\partial v}{\partial t}.$$

This equation can be simplified by $\partial u/\partial x = \partial v/\partial M \cdot \partial M/\partial x$. Finally, substituting (5.39) and (5.40) and combining the result with (5.37) we obtain a system

$$\begin{cases} \dfrac{\partial}{\partial t}\left(\dfrac{\partial M}{\partial x}\right) = -\dfrac{\partial}{\partial x}\left(\dfrac{\partial M}{\partial x}u\right), \\ \dfrac{\partial u}{\partial t} = -\dfrac{\partial u}{\partial x}u - \displaystyle\sum_{q>0}\left[\nabla\cdot\left(\frac{\partial x}{\partial M}qq^T\frac{\partial x}{\partial M}^T\phi\right) - \left(\nabla\cdot\left(\frac{\partial x}{\partial M}\right)qq^T\frac{\partial x}{\partial M}^T\right)\phi\right]^T, \end{cases} \tag{5.46}$$

where $\phi = \phi(\|(\partial x/\partial M)q\|)$.

The system (5.46) has 12 states in 3-dimensional space, 9 for $\partial M/\partial x$ (t,x) and 3 for $u(t,x)$. It resembles the famous Grad 13-moment system by Grad 1949, which extends the Euler equations by considering directional-dependent pressure tensor. The last state of the Grad 13-moment system is the inner energy, which does not appear in (5.46). The reason for this is that we derive a continuous interaction term explicitly from the interaction forces, which is possible only if the forces are defined by long-range potentials. As it was shown by Caprino et al. 1993, expressing a system with long-range potentials by the Euler equations leads to the solution with zero temperature, therefore the inner energy becomes functionally dependent on the velocity field and its evolution equation can be omitted.

5.3.4 Dimensionality reduction

It appears that in some special cases it is possible to reduce the system (5.46) by considering only one scalar characteristic of a compression in any space point instead of the whole compression tensor. Indeed, we define a *density* as a determinant of the compression tensor:

$$\rho(t,x) := \det\left(\partial M/\partial x\right)(t,x).$$

Not only the compression tensor itself, but also its determinant satisfies (5.37). This nontrivial fact holds because the compression tensor is a Jacobian, which is shown by the following lemma:

Lemma 5.1. *Let $J(t,x) \in \mathbb{R}^{n\times n}$ be the Jacobian matrix of function $M(t,x)$. Let $J(t,x)$ satisfies the dynamic equation*

$$\frac{\partial J}{\partial t} = -\frac{\partial (Ju)}{\partial x}, \tag{5.47}$$

where $u = u(t,x)$ is some vector field. Then the determinant $\det J$ *satisfies the same equation:*

$$\frac{\partial \det J}{\partial t} = -\frac{\partial}{\partial x} \cdot (\det J \cdot u)\,. \tag{5.48}$$

Proof. See Appendix A.4. □

Therefore from (5.37)

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho u)\,. \tag{5.49}$$

This equation is the first of the complete set of Euler equations. Unfortunately, the second equation of (5.46) depends on the whole compression tensor and thus it cannot be described only by the means of density. This is reasonable since in general the system can have different forces in different directions in response to different compressions. Therefore in order to simplify the system we need to assume that the compression can be represented by a single number, i.e. that it is compressed equally in all directions.

Assumption 5.1 (*Isotropy*)**.** Compression tensor $\partial M/\partial x(t,x)$ is isotropic (equal in all directions), thus it can be represented as a rotation matrix multiplied by a scalar.

This assumption looks restricting at first glance, but for the infinitely large chaotic system with infinitely many particles the system indeed "looks the same" in all directions at every point, thus we can say it is isotropic.

Assumption 5.1 has long-lasting implications. Define $l(t,x) := \lambda(\partial x/\partial M(t,x))$, since all the eigenvalues are equal. This variable, called *specific distance*, represents an average distance between two neighbouring particles at point x. By definition of the density $\rho = l^{-n}$. Further, $\left\| \frac{\partial x}{\partial M} q \right\| = l \left\| q \right\|$. Since q is a multiindex vector, its squared length should be a natural number. Therefore we can define its length $r = \|q\|$ such that $r^2 \in \mathbb{N}$. Breaking the summation in (5.46) in a sum of all possible lengths r of multiindex vectors, we can rewrite the summation term as

$$\sum_{r^2 \in \mathbb{N}} \left[\nabla \cdot \left(\phi(rl) \frac{\partial x}{\partial M} \sum_{\substack{q>0 \\ \|q\|=r}} \left(qq^T \right) \frac{\partial x}{\partial M}^T \right) - \phi(rl) \left(\nabla \cdot \left(\frac{\partial x}{\partial M} \right) \sum_{\substack{q>0 \\ \|q\|=r}} \left(qq^T \right) \frac{\partial x}{\partial M}^T \right) \right]^T . \tag{5.50}$$

Proposition 5.2. *Given r such that $r^2 \in \mathbb{N}$, the summation over all outer products of multiindices of a length r is proportional to the identity matrix, i.e. there exists $\beta(r)$ such that*

$$\sum_{\substack{q>0 \\ \|q\|=r}} qq^T = \beta(r) I. \tag{5.51}$$

Proof. First of all, we will show that all nondiagonal elements in (5.51) are zero. Indeed, for any positive q its contribution to kj-th element of matrix (5.51) is given by $q_k q_j$. But for any

$k \neq j$ we can pick $\bar{q}$ such that it equals q except $\bar{q}_{\max(k,j)} = -q_{\max(k,j)}$. In this case $\bar{q}$ is also positive and thus is included into the summation, while the contribution to kj-th element of (5.51) has opposite sign. Therefore all nondiagonal elements of (5.51) are zero.

Further, all diagonal elements of (5.51) are equal. This can be proven by analogous argument. Indeed, we can take a positive q and look at the elements q_k^2 and q_j^2. Then $\bar{q}$ which is equal to q except for $\bar{q}_k = \operatorname{sgn}(q_k)|q_j|$ and $\bar{q}_j = \operatorname{sgn}(q_j)|q_k|$ is also positive, but swaps the contributions between k-th and j-th diagonal elements. Thus all the contributions to the diagonal elements are equal. Finally,

$$\operatorname{Tr} \sum_{\substack{q>0 \\ \|q\|=r}} qq^T = \sum_{\substack{q>0 \\ \|q\|=r}} q^T q = r^2 \cdot \#_r q = n\beta(r), \tag{5.52}$$

where $\#_r q$ denotes the number of positive multiindices q with length r and we define $\beta(r) = r^2/n \cdot \#_r q$. It is worth noticing that by Takloo-Bighash 2018 the average approximate behaviour of the number of positive multiindices q with length r is $\#_r q \propto r^{n-1}$ as $r \to +\infty$, thus $\beta(r) \propto r^{n+1}$. □

By Assumption 5.1

$$\frac{\partial x}{\partial M}\frac{\partial x}{\partial M}^T = l^2 I. \tag{5.53}$$

Using Proposition 5.2 and (5.53), (5.50) becomes

$$\sum_{r^2 \in \mathbb{N}} \beta(r) \left[\nabla \cdot \left(\phi(rl) l^2 I \right) - \phi(rl) \left(\nabla \cdot \left(\frac{\partial x}{\partial M} \right) \frac{\partial x}{\partial M}^T \right) \right]^T.$$

The value inside of the square brackets can be simplified further. Indeed, by (5.45) it is possible to inject density inside, which gives

$$\begin{aligned} &\frac{1}{\rho} \left[\nabla \cdot \left(\rho \phi(rl) l^2 I \right) - \frac{\partial \rho}{\partial x} \phi(rl) l^2 I - \phi(rl) \rho \left(\nabla \cdot \left(\frac{\partial x}{\partial M} \right) \frac{\partial x}{\partial M}^T \right) \right]^T \\ &\quad = \frac{1}{\rho} \left[\nabla \cdot \left(\rho \phi(rl) l^2 I \right) - \phi(rl) \left(\nabla \cdot \left(\rho \frac{\partial x}{\partial M} \right) \frac{\partial x}{\partial M}^T \right) \right]^T. \end{aligned}$$

Finally, the second term in the square brackets appears to be zero due to the following result:

Lemma 5.2. *Let $\partial x/\partial M$ be isotropic, i.e. represented by a scalar multiplied by a rotation matrix, and let $\rho = \det(\partial M/\partial x)$. Then*

$$\nabla \cdot \left(\rho \frac{\partial x}{\partial M} \right) \frac{\partial x}{\partial M}^T = 0. \tag{5.54}$$

Proof. See Appendix A.5. □

Using this Lemma and the fact that $\nabla \cdot (\rho\phi(rl)l^2 I) = \nabla(\rho\phi(rl)l^2)$, we can define the *pressure*:

$$P = \sum_{r^2 \in \mathbb{N}} \beta(r)\rho\phi(lr)l^2 = \sum_{r^2 \in \mathbb{N}} \frac{\beta(r)}{r} l^{1-n} f(lr). \tag{5.55}$$

Note that the pressure is well-defined since the sum is convergent by the property (5.32). With this definition, the system (5.46) together with (5.49) turns into the famous *Euler equations*:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = -\nabla \cdot (\rho u)\,, \\ \dfrac{\partial u}{\partial t} = -\dfrac{\partial u}{\partial x} u - \dfrac{\nabla P^T}{\rho}\,. \end{cases} \tag{5.56}$$

Therefore the following theorem was proven:

Theorem 5.1. *There exists a valid continuation process which leads from the Newtonian system (5.33) to the Euler equations (5.56) under the assumption that the system is locally isotropic in every point in space.*

Remark 5.1 (Non-complete interaction topologies). In the original ODE system (5.33) we assumed that an interaction exists between every pair of particles, i.e. that the topology of interactions is all-to-all. In general in order to obtain (5.33) it would be sufficient to use any topology for which the isotropy required in Assumption 5.1 is possible. The difference in topologies would modify the definitions of density $P(t, x)$ in (5.55). For example, for the grid topology with equations given by

$$\begin{cases} \dot{x}_i = v_i, \\ \dot{v}_i = \sum_{k=1}^{n} \left(F(x_i - x_{i-e_k}) - F(x_i - x_{i+e_k})\right), \end{cases} \tag{5.57}$$

where e_k denotes the k-th basis vector of $\mathbb{R}^n$, the continuation renders the same Euler equations (5.56) with the pressure given by $P = f(l)/l^{n-1}$.

5.4 Control of robotic formation

5.4.1 Overview

In this section we will demonstrate how the continuation method can help in the analysis and design of control laws for large-scale systems. We will do it by using an example of a robotic swarm, i.e. a formation of robots whose goal is to follow some desired trajectory while passing through obstacles and preserving relative agents' positions.

Control of robotic formations is an extensively studied topic, see recent reviews by Oh, Park, and Ahn 2015; Chung et al. 2018. However most of the methods rely on the graph-theoretic properties of interaction topology and on simple linear controllers to provide stability. A PDE approach was taken by Toner and Tu 1995 who used the Euler PDE with

diffusion terms to model the flocks of birds. The authors proposed a PDE to describe the behaviour of agents and analyzed it to study a symmetry breaking which leads to a coherent movement of birds. Similar PDE model was used to control 3D agent formation with 2D disc communication topology via backstepping, see Qi, Vazquez, and Krstic 2014. Lattice-based spatially-invariant models for platooning were considered by Jovanovic and Bamieh 2005; Bamieh et al. 2012, where stability properties of infinite systems were studied in various space dimensions.

Works mentioned above which use PDE representations of multi-agent systems just assume a PDE model, which can be justified by a limiting case of the infinite number of agents. Contrary, we will base our analysis on the continuation procedure, rigorously introducing a PDE to describe a finite formation of drones. We will study this PDE and recover a nonlinear local control law which, being applied to the agents, forces the whole formation to follow the desired density profile.

5.4.2 System continuation and PDE control

Let us start from a system of drones having double integrator dynamics:

$$\ddot{x}_i = \tau_i. \tag{5.58}$$

Here $x_i \in \mathbb{R}^n$ is a position of the i-th drone in n-dimensional space and $\tau_i \in \mathbb{R}^n$ is a control we want to design. The drones are enumerated with multiindices $i \in \mathbb{Z}^n$. Define $v_i = \dot{x}_i$. Similarly to the previous section we introduce multiindex function $M(t,x)$ such that $M(t,x_i) \equiv i$ and then perform a continuation. The resulting system is

$$\begin{cases} \dfrac{\partial \rho}{\partial t} = -\nabla \cdot (\rho u), \\ \dfrac{\partial u}{\partial t} = -\dfrac{\partial u}{\partial x} u + \tau(x,t), \end{cases} \tag{5.59}$$

where $\tau(t,x) = \tau(t, M(t,x))$ is a continuation of the control τ_i.

Now let us formulate a desired system which will be used as a reference which the real formation should converge to. Given a velocity profile $u_d(x)$, we define the desired density $\rho_d(t,x)$ to follow this velocity profile. Essentially this means "desired agents" have single-integrator dynamics. Note that in general u_d can be dependent on time but we don't consider it for simplicity of writing.

Thus we assume the desired system is governed by

$$\frac{\partial \rho_d}{\partial t} = -\nabla \cdot (\rho_d u_d). \tag{5.60}$$

Our goal is to derive $\tau(t,x)$ such that $\rho \to \rho_d$. First, direct calculations from (5.59) and (5.60) lead to the following systems in terms of flows (ρu) and $(\rho_d u_d)$:

$$\begin{aligned} \frac{\partial (\rho u)}{\partial t} &= -\nabla \cdot (\rho u) u - \rho \frac{\partial u}{\partial x} u + \rho \tau(x,t), \\ \frac{\partial (\rho_d u_d)}{\partial t} &= -\nabla \cdot (\rho_d u_d) u_d. \end{aligned} \tag{5.61}$$

Define the deviation from the desired density $\tilde{\rho} = \rho - \rho_d$. Then the second-order equation for the deviation is

$$\frac{\partial^2 \tilde{\rho}}{\partial t^2} = \nabla \cdot \left[\nabla \cdot (\rho u) u - \nabla \cdot (\rho_d u_d) u_d + \rho \frac{\partial u}{\partial x} u - \rho \tau(x,t) \right].$$

In order to cancel the nonlinear terms, define now the control τ as

$$\tau = \frac{\partial u}{\partial x} u + \frac{1}{\rho} \Big[\nabla \cdot (\rho u) u - \nabla \cdot (\rho_d u_d) u_d + \alpha(\rho_d u_d - \rho u) + \beta \nabla (\rho_d - \rho)^T \Big], \tag{5.62}$$

where α and β are some positive gains. Then the equation for the density deviation transforms into

$$\frac{\partial^2 \tilde{\rho}}{\partial t^2} = -\alpha \frac{\partial \tilde{\rho}}{\partial t} + \beta \nabla^2 \tilde{\rho}. \tag{5.63}$$

This equation is a wave equation with damping and thus it is asymptotically stable if $\tilde{\rho} = 0$ on the boundary of the domain (Folland 2020). Choosing a desired system such that $\rho_d = 0$ on the boundary and using a continuation of ρ such that $\rho = 0$ on the boundary ensures satisfaction of the boundary condition.

5.4.3 Control discretization

Formula (5.62) for PDE (5.59) is local by its nature, but it should be discretized to be implemented on every agent of the original ODE (5.58). One particular discretization is described next.

First of all, for the agent i define a matrix G_i as a discretization of the compression tensor:

$$[G_i]_j = (x_{i+e_j} - x_{i-e_j})/2 \approx \frac{\partial x}{\partial M_j}(t, x_i), \tag{5.64}$$

where e_j is the j-th unit basis vector and $[G_i]_j$ represent the j-th column of G_i. The matrix G_i depends on the positions of $2n$ neighbouring agents of the i-th agent, thus the interaction topology is a lattice. In the same way as G_i we define a matrix W_i representing a velocity Jacobian:

$$[W_i]_j = (v_{i+e_j} - v_{i-e_j})/2 \approx \frac{\partial u}{\partial M_j}(t, x_i). \tag{5.65}$$

Now we can write formulas for all terms inside of (5.62) depending on the real system:

$$\begin{aligned}
&1). \quad \frac{\partial u}{\partial x} u = \frac{\partial u}{\partial M} \frac{\partial M}{\partial x} u \approx W_i G_i^{-1} v_i, \\
&2). \quad \nabla \cdot u = \sum_{j=1}^{n} \frac{\partial u_j}{\partial M} \frac{\partial M}{\partial x_j} \approx \sum_{j=1}^{n} [W_i^T]_j \cdot [G_i^{-1}]_j, \\
&3). \quad \rho \approx 1/\det G_i, \\
&4). \quad \nabla \rho = -\rho^2 \nabla \left(\det \frac{\partial x}{\partial M} \right) = -\rho^2 \frac{\partial \left(\det \frac{\partial x}{\partial M} \right)}{\partial M} \frac{\partial M}{\partial x} \approx -\rho^2 \frac{\partial (\det G_i)}{\partial M} G_i^{-1},
\end{aligned} \tag{5.66}$$

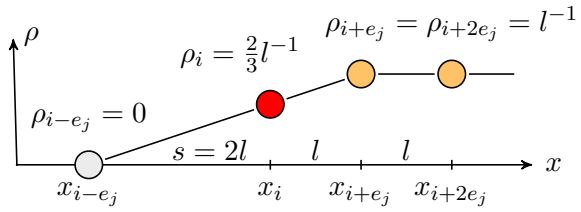


Figure 5.4: Left boundary of the system (5.58) with control (5.67). Agent i is on the boundary, the position of the "ghost agent" $i - e_j$ is chosen such that ρ linearly goes to zero at x_{i-e_j}.

where the gradient of the determinant $\det G_i$ in the last equation should be computed according to the determinant formula, using second derivatives of the positions discretized similarly to (5.64):

$$\frac{\partial^2 x}{\partial M_j \partial M_k}(x_i, t) \approx (x_{i+e_j+e_k} + x_{i-e_j-e_k} - x_{i+e_j-e_k} - x_{i-e_j+e_k})/4,$$
$$\frac{\partial^2 x}{\partial M_j^2}(x_i, t) \approx x_{i+e_j} - 2x_i + x_{i-e_j}.$$

Since the gradient of the determinant depends on the second derivatives, in total each agent requires information about the velocities of its $2n$ neighbouring agents and the positions of its $2n^2$ neighbouring agents, including diagonal ones.

Finally, substituting (5.66) into (5.62), the formula for the control action τ_i appears as

$$\tau_i = \left[W_i G_i^{-1} + \sum_{j=1}^{n} [W_i^T]_j \cdot [G_i^{-1}]_j - \alpha\right] v_i + \left[\beta I - v_i v_i^T\right] \frac{1}{\det G_i} G_i^{-T} \frac{\partial (\det G_i)}{\partial M}^T + \\ + \det G_i \Big[\alpha \rho_d u_d + (\beta I - u_d u_d^T) \nabla \rho_d^T - \rho_d (\nabla \cdot u_d) u_d\Big]. \tag{5.67}$$

5.4.4 Boundary conditions

For the system (5.63) to converge to zero proper boundary conditions should be used. Namely, the continuation should be chosen such that $\rho = 0$ outside of the formation. As it was shown in Section 4.4.4, boundary conditions for PDE correspond to "ghost agents" in the ODE case. In particular, information about neighbour agents is used in (5.64) and (5.65). Therefore, if an agent with index i is on the boundary with respect to the j-th axis direction, specifying boundary conditions means specifying position x_{i-e_j} and velocity v_{i-e_j} for the nonexisting agent $i - e_j$ (contrary, if the agent i is on the other side of the formation, nonexisting agent would have an index $i + e_j$ respectively).

Proposition 5.3. *Assume agent $i - e_j$ is a ghost agent. Then*

$$x_{i-e_j} = 3x_i - 2x_{i+e_j}, \quad v_{i-e_j} = 2v_i - v_{i+e_j} \tag{5.68}$$

ensures $\rho_{i-e_j} = 0$ and a correct computation of $[W_i]_j$ by (5.65).

Proof. Choice (5.68) for velocities is natural, since being substituted in (5.65) this leads to an approximation of the velocity gradient based solely on the existing i and $i + e_j$ agents.

For the position we want that the compression tensor (5.64) "feels" that the drone i is on the border. For this we can use such an approximation that the density near the border will linearly diminish to zero, see Fig. 5.4. Namely, let us look at 1D case and fix i-th agent to be on the left border, with $\rho_{i-e_j} = 0$ for the ghost agent. Assume further that the distance between each pair of existing agents is constant and equal to l. Then $\rho_{i+e_j} = l^{-1}$. Define an unknown distance $s := x_i - x_{i-e_j}$. Then asking for a linear dependency of a density on position, we have necessarily

$$\rho_i = \frac{l\rho_{i-e_j} + s\rho_{i+e_j}}{l+s} = \frac{s}{l(l+s)}.$$

But by (5.64) $\rho_i = 2/(l+s)$, which immediately gives the answer $s = 2l$, or $x_{i-e_j} = x_i + 2(x_i - x_{i+e_j})$, which is (5.68). □

Proposition 5.3 finalizes the formulation of the boundary conditions and thus the correct implementation of (5.67).

5.4.5 Numerical simulation

To demonstrate the control policy (5.67) we used a numerical simulation. The simulation was performed both in 2D and in 3D to show that the derived controller can handle arbitrary space dimensions. In 2D a formation of $7 \times 7 = 49$ drones was simulated, while in 3D space we used a cubic formation of $8 \times 8 \times 8 = 512$ drones. Starting from a random initial position, the drones' goal was to reach a regular formation, fly through a window and restore the regular formation after the maneuver.

Assume the center of the window is placed at the point $(x_0, 0, 0)$, and the formation should fly through it starting from the origin. The desired velocity field $u_d(x, y, z)$ able to fulfill the task was constructed as

$$u_{d_x} = 1, \quad u_{d_{y|z}} = 0.05 \operatorname{atan}(x - x_0) e^{-\frac{(x-x_0)^2}{100}} y|z,$$

where $y|z$ denotes y or z, see the left panel of Fig. 5.5 for the streamlines projected on the x-y plane.

For simplicity the desired system (5.60) was simulated by first-order integrators following the desired velocity profile, and the density $\rho_d(x, t)$ was interpolated between agents. Both the desired system (5.60) and the real system (5.58) were simulated using the Euler method for the regular formations of $7 \times 7 = 49$ drones in 2D and of $8 \times 8 \times 8 = 512$ drones in 3D. The initial positions for the real system were multiplied by 2 in comparison to the desired system and a uniform noise $U(-2, 2)$ was added. The control gains were chosen as $\alpha = 3$ and $\beta = 100$. The convergence of the real density to the desired one is shown on the right panel of Fig. 5.5 and snapshots of the simulation are presented in Fig. 5.6 for 2D system and in

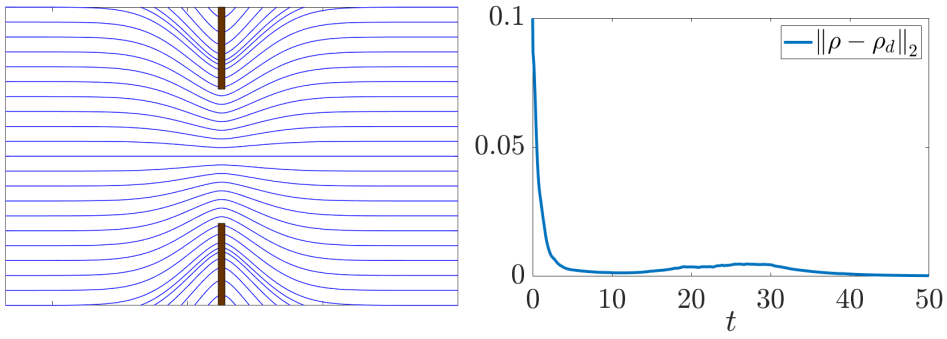


Figure 5.5: **Left:** streamlines of the desired velocity field $u_d(x, y)$. **Right:** convergence of the L_2 norm of the density deviation for 3D formation (similar picture can be obtained for 2D formation).

Fig. 5.7 for 3D system respectively. It is clear that the real formation, being heavily disturbed in the beginning, converges to the desired shape in less than 5 seconds and then follows the desired pattern, successfully passing through the window.

5.5 Concluding remarks

In this chapter we showed that based on the continuation method, new continuous models can be derived and further utilized for analysis and control purposes. A special attention was paid to multi-agent systems, which can be continualized using a notion of density defined as an inverse of a partial derivative of a position with respect to the indexing function.

As an example we used the continuation to show how various traffic PDE models can be recovered from discrete traffic representations and how the Euler equations for compressible fluid can be derived from the Newtonian particle interactions, providing more intuition into the Hilbert's 6th problem. The same continuation was then used to describe a robot formation flying through a window. We developed a control algorithm to stabilize a desired trajectory based on a continuous representation of the formation. This algorithm is distributed as every robot requires information only about neighbouring robots.

A general method for derivation of continuous models for multi-agent systems is a promising method that can lead to development of new boundary and distributed control approaches and algorithms in the future. To facilitate this process, it would be beneficial to explicitly derive conditions on the boundaries of applicability of the proposed framework and on the required order of continuation which should be used for the system description.

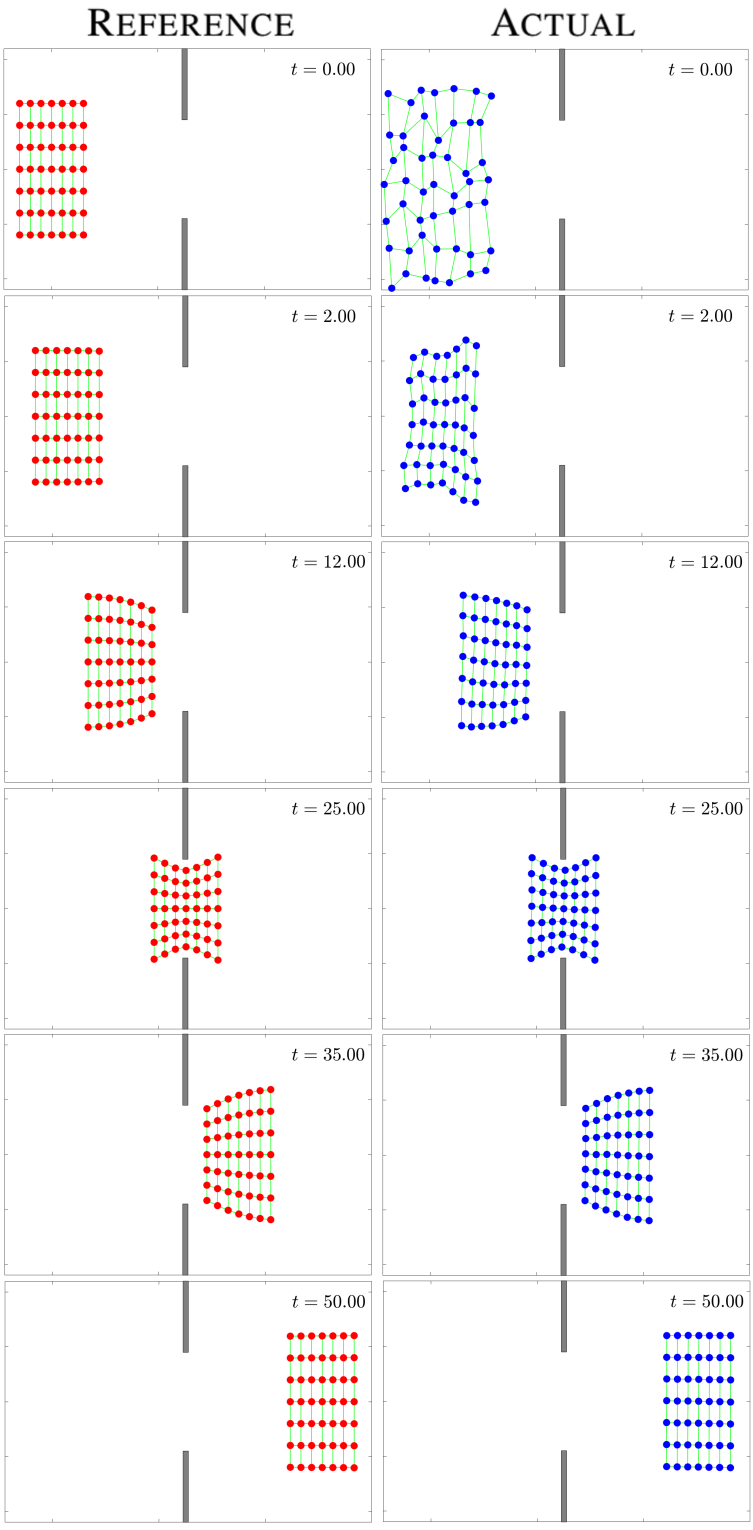


Figure 5.6: Simulation of 2D formation of $7 \times 7 = 49$ drones flying through window. Rows correspond to times $t = \{0s, 2s, 12s, 25s, 35s, 50s\}$. **Left column, reference:** desired system (5.60), governed by single integrators. **Right column, actual:** heavility perturbed real system (5.58) with control (5.67) which converges to the desired one.

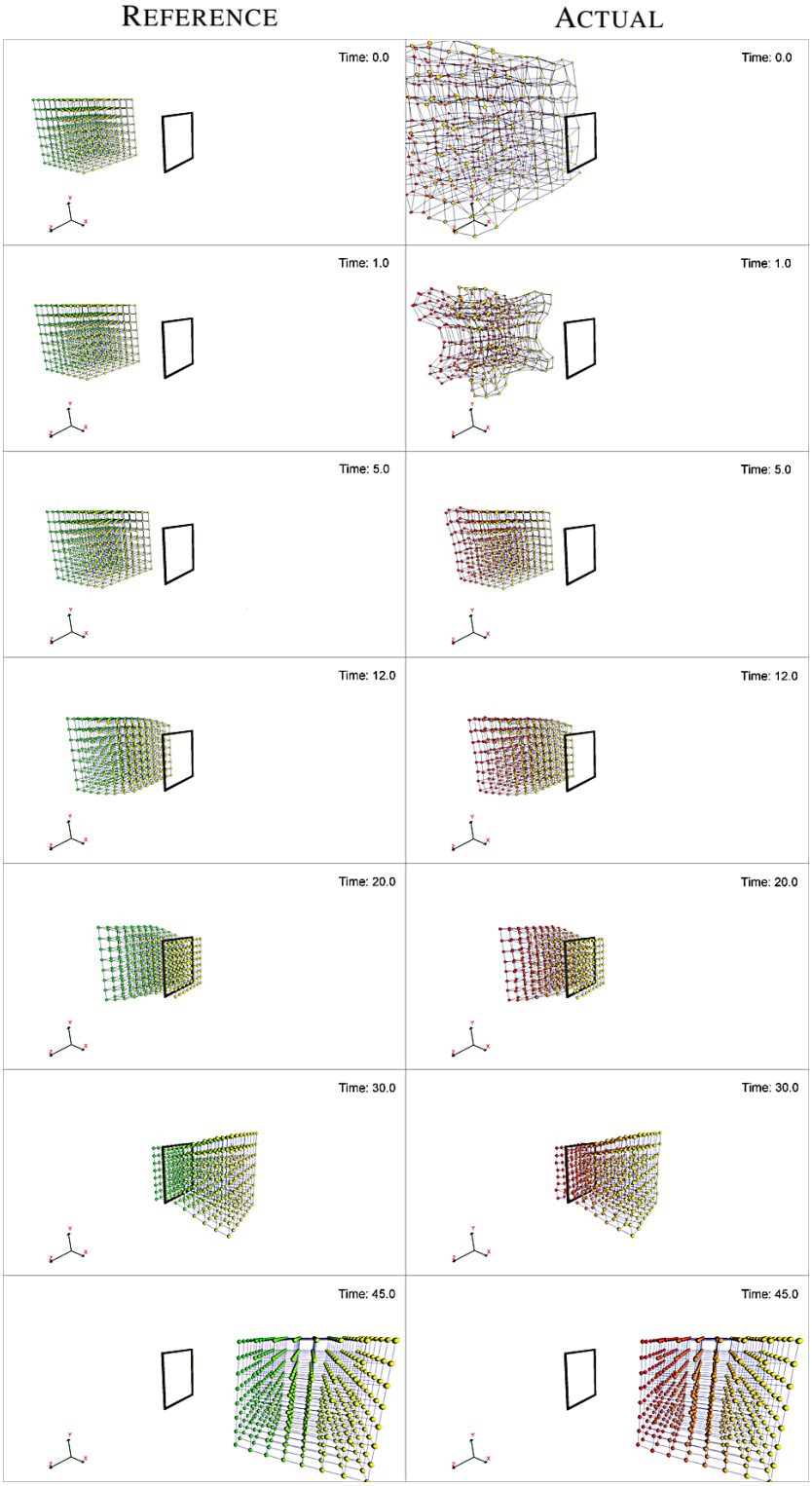


Figure 5.7: Simulation of 3D formation of $8 \times 8 \times 8 = 512$ drones flying through window. Rows correspond to times $t = \{0s, 1s, 5s, 12s, 20s, 30s, 45s\}$. **Left column, reference:** desired system (5.60), governed by single integrators. **Right column, actual:** heavility perturbed real system (5.58) with control (5.67) which converges to the desired one.

Chapter 6

Applications of the continuation method to oscillatory systems

Contents

6.1 Introduction to networks of oscillators

Networks are an astonishing subject of multidisciplinary research, since they are ubiquitous. Electric power grids and traffic networks are only few examples of systems that consist of

a large number of connected dynamical units. The large-scale collective behaviour of such systems is determined by the interplay of network spatial topology and individual dynamics. Oscillatory networks are a particular type of networks where every node exhibits oscillatory dynamics. In these networks major studies are devoted to synchronization phenomena, which is of special importance in systems such as laser arrays, biological neural networks, power grids, electrical and magnetic systems. For example, in laser arrays, in power or magnetic systems synchronization is a desirable mode of operation since it drastically increases the amount of energy produced by system. At the same time it was shown that in biological neural networks synchronization plays a negative role since it is related to the Parkinson disease.

In Chapter 5 we demonstrated how the continuation method derived in Chapter 4 can be helpful in the analysis and control design for large-scale nonlinear systems. In this chapter we will further apply the continuation method to solve various problems related to the synchronization oscillatory networks. One could argue: why do we want to use a PDE instead of ODEs, if PDEs are generally considered to be harder to analyze and to control? From the point of view of the control design the answer is that a suitable use of PDEs can lead to explicit and scalable algorithms. Indeed, the centralized computation of feedback control gains for a large-scale linear system with n agents requires at least $O(n)$ operations by methods such as ODE-based backstepping (Kanellakopoulos, Kokotovic, and Morse 1991) and at least $O(n^3)$ operations by methods like LQR, which require solving a Riccati matrix equation. On the contrary, in Section 6.2 we will give an example of such situation where the continuation helps to design a control with gains computed in $O(1)$ operations. In particular, in case of unstable 1-dimensional PDEs we can use the result by Smyshlyaev and Krstic 2005, where the general second-order linear space-dependent system is stabilized to zero state using backstepping control, based in turn on Smyshlyaev and Krstic 2004. From the point of view of system analysis it appears that synchronization of oscillations is a highly nonlinear effect and thus its analysis poses big challenges for large ODE networks. In Section 6.3 we show that the nonlinear continuation method allows to recover synchronization threshold for a ring of coupled Kuramoto oscillators which plays a role of a prototypical model for a power grid. Although this can be considered as a toy example (since the same result was recently discovered in the ODE setup), we show in Section 6.4 that the continuation method can be applied to a much more general class of oscillators which change their frequency depending on a magnitude of oscillations for which, up to the author's knowledge, no synchronization conditions exist for ODE-based coupled large networks.

6.2 Synchronization of a laser chain

6.2.1 Overview

The first problem in the domain of the networks of oscillators that we draw our attention to is a problem of stabilization of a chain of coupled semiconductor lasers. Coupled laser systems are important for high-precision power transmission applications such as welding,

laser surgery or fusion research as well as many others (Saxena, Prasad, and Ramaswamy 2012). Recently, Pietrzak et al. 2015 showed that large arrays of semiconductor laser diodes are more power- and cost-efficient compared to single crystal lasers due to lower electrical resistance and optical load. A typical array of coupled lasers is depicted in Fig. 6.1.

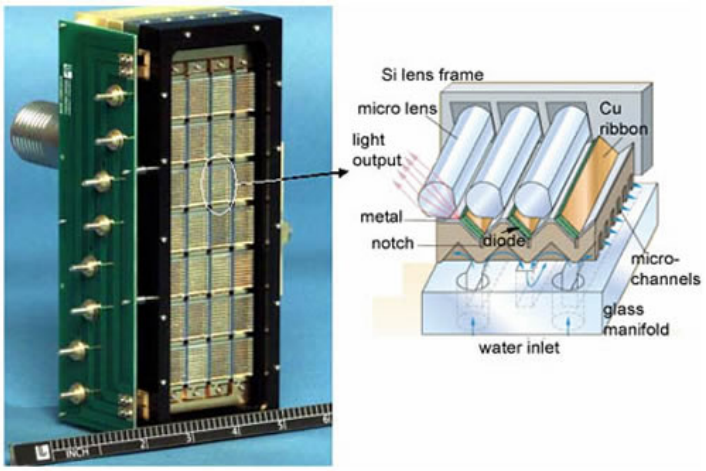


Figure 6.1: High-average-power laser-diode 41kW array, composed of 28 silicon monolithic microchannels (SiMMs) each consisting of thousands of diodes. Image from Lawrence Livermore National Laboratory, https://lasers.llnl.gov/science/photon-science/highpowered-lasers/hapl, licensed under CC BY-NC-SA 4.0.

It was shown by Carr, Taylor, and Schwartz 2006 that coupling of several Class-B lasers (Arecchi et al. 1984) can lead to a resonance effect, greatly increasing intensity comparing to the uncoupled laser system. However, such a system is prone to instabilities: electrical fields of lasers start to oscillate around the operating point, destroying resonance effect. Carr, Taylor, and Schwartz 2006 further showed that these oscillations, up to the first order, are described by coupled Stuart-Landau oscillators.

Stuart-Landau oscillators are prototypical models for Andronov-Hopf bifurcation, and apart from laser applications they are used to describe many oscillatory systems such as electronic oscillators (Bergner et al. 2012) or biological neural networks (Aoyagi 1995). Usually in laser analysis Stuart-Landau model describes electrical field of one laser, thus the oscillating behaviour is the desired one. However we base our analysis on the results of Carr, Taylor, and Schwartz 2006, where Stuart-Landau model is used to describe deviation from the synchronized steady state: thus, oscillations should be suppressed.

6.2.2 Model

Let the deviation of one laser be $c \in \mathbb{C}$, then one Stuart-Landau oscillator is described by the evolution equation

$$\dot{c} = \left(\Gamma + i\omega - \eta|c|^2\right) c, \tag{6.1}$$

where $\Gamma > 0$ is an excitation gain, ω is a natural frequency and $\eta \in \mathbb{C}$ is a nonlinear damping coefficient. For $\Gamma < 0$ the system has one stable equilibrium point $c = 0$, while for $\Gamma > 0$ zero equilibrium point is unstable, and system has a stable limit cycle with frequency ω and with amplitude $|c| = \sqrt{\Gamma/\eta}$.

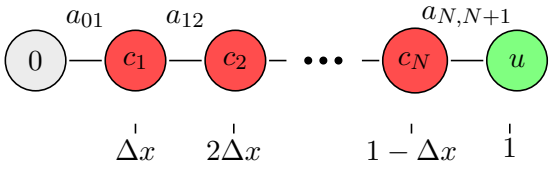


Figure 6.2: System of coupled semiconductor lasers, modeled by (6.2).

Saxena, Prasad, and Ramaswamy 2012 proposed to design laser hardware having in mind an effect called *amplitude death* to suppress laser electrical field's undesirable oscillations and thus prevent loss of efficiency. This effect appears when many inhomogeneous oscillators are strongly coupled, thus making their limit cycles unstable and the zero fixed point stable. Contrarily to this approach of hardware-designed amplitude death, we propose to use an active feedback stabilization from one boundary to suppress oscillations. We consider here a chain of $N + 2$ coupled Stuart-Landau oscillators. Let the position of i-th oscillator for $i \in \{0, .., N + 1\}$ be $x_i = i\Delta x$ with $\Delta x = 1/(N + 1)$ being distance between two neighbours, thus $x_0 = 0$ and $x_{N+1} = 1$. The state of i-th oscillator is $c_i \in \mathbb{C}$. We assume that the oscillators on the boundaries are directly controllable, namely the left boundary oscillator has fixed zero state $c_0 = 0$ and the state of the right boundary oscillator is a control variable $c_{N+1} := u$. We also assume that the coupling of lasers is realized by an overlapping of their evanescent fields (Winful and Rahman 1990), thus the evolution equation of i-th oscillator depends on the nearest neighbours' states with gains $a_{i,i-1}$ and $a_{i,i+1}$. Since it is a conservative force, $a_{i,i+1} = a_{i+1,i}$, thus the network is undirected. The system is given by

$$\begin{aligned} &\dot{c}_i = (\mu - \eta|c_i|^2)c_i + a_{i,i-1}(c_{i-1} - c_i) + a_{i,i+1}(c_{i+1} - c_i), \\ &c_0 = 0, \quad c_{N+1} = u, \end{aligned} \tag{6.2}$$

with $\mu = \Gamma + i\omega$. In general, coupling $a_{i,i+1}$ can be space-dependent. We assume that it is monotone, for example as in case of an increasing electrical permeability of the medium along the laser chain. Therefore, to approximate monotone dependencies we restrict ourselves to a class of coupling gains $a_{i,i+1} \approx \alpha(x_i - \beta)^2$ with $\alpha > 0$, $\beta \in \mathbb{R} \setminus [0, 1]$. Note that this class includes also homogeneous couplings in case $\beta \to \pm\infty$ and $\alpha \to 0$ such that $\alpha\beta^2 \equiv const$.

6.2.3 Continuation and boundary control using backstepping

Since $\Gamma > 0$, system (6.2) has unstable zero equilibrium. Our goal is to design a feedback control law $c_{N+1} = u(c)$ such that zero solution is stabilized, thus suppressing oscillations. Linearizing system (6.2) around zero and assuming $|c_i|$ is small, we get

$$\begin{aligned} &\dot{c}_i = \mu c_i + a_{i,i-1}(c_{i-1} - c_i) + a_{i,i+1}(c_{i+1} - c_i), \\ &c_0 = 0, \quad c_{N+1} = u. \end{aligned} \tag{6.3}$$

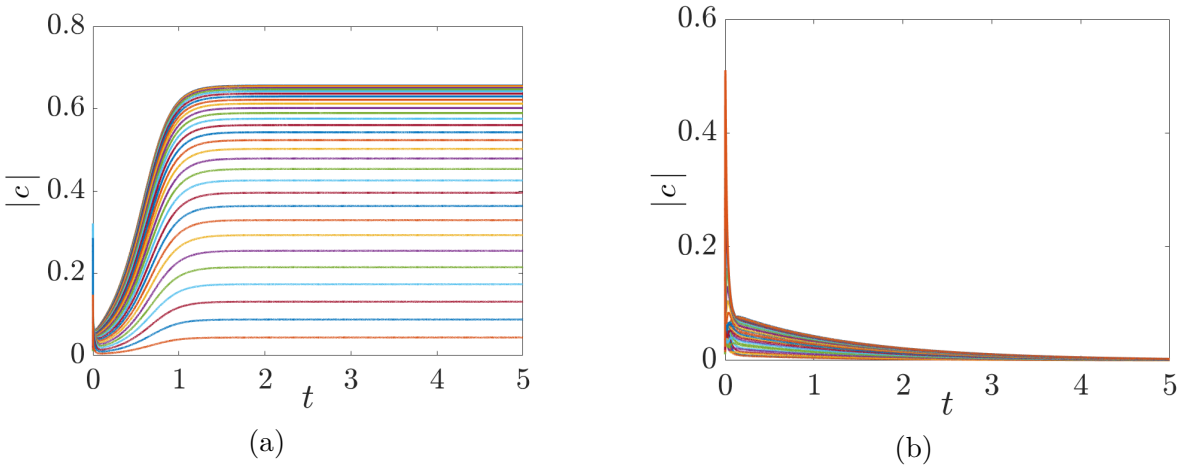


Figure 6.3: Numerical simulation of system (6.2) with $N = 30$ oscillators with parameters $\mu = 5 + 4i$, $\eta = 10$, $\alpha = 5$ and $\beta = -10$. The absolute values of all states $|c_i(t)|$ for $i \in \{1..N\}$ are depicted. **(a):** Uncontrolled system, $u = 0$. **(b):** Controlled system with controller (6.7) with $\gamma = 10$.

In case of thousands of coupled laser diodes, the implementation of traditional control algorithms for the system (6.3) would require a lot of computational power. Instead, we can use the continuation method as explained in Section 4.5, Corollary 4.7 for undirected linear networks and obtain a continuous model of (6.3) as a second-order PDE:

$$\frac{\partial c(x,t)}{\partial t} = \mu c(x) + \frac{\partial}{\partial x}\left(\alpha \Delta x^2 (x-\beta)^2 \frac{\partial c(x,t)}{\partial x}\right) \tag{6.4}$$

for $x \in (0,1)$ and with boundary conditions $c(0,t) = 0$ and $c(1,t) = u$.

Although system (6.4) is formulated in complex domain, one can use backstepping method by Smyshlyaev and Krstic 2005 to stabilize it. Indeed, the stabilizing controller is given by

$$u := \int_0^1 k(x) c(x,t) dx, \tag{6.5}$$

and the kernel is found by formula (44) from the work of Smyshlyaev and Krstic 2005:

$$k(x) = -\bar{x}\frac{(\mu+\gamma)}{\alpha|\beta|}\frac{(1-\beta)^{3/2}}{(x-\beta)^{5/2}} \times \frac{I_1\left(\sqrt{(\mu+\gamma)(\bar{y}^2-\bar{x}^2)/(\alpha\beta^2)}\right)}{\sqrt{(\mu+\gamma)(\bar{y}^2-\bar{x}^2)/(\alpha\beta^2)}}, \tag{6.6}$$

where $\gamma > 0$ is an adjustable gain, $I_1(s)$ is the modified Bessel function of order one, $\bar{x} = -\beta \log(1 - x/\beta)$ and $\bar{y} = -\beta \log(1 - 1/\beta)$.

6.2.4 Control discretization and numerical simulation

Finding a control law for the original ODE system (6.2) can be easily done by performing a numerical integration of (6.5) using the trapezoidal rule

$$u := \Delta x \sum_{i=1}^{N} k(x_i) c_i, \tag{6.7}$$

Each control gain $k(x_i)$ can be computed directly in $O(1)$ operations.

We validated the obtained control law by numerical simulation of system (6.2) with $N = 30$ coupled oscillators. We took $\Gamma = 5$ for excitation gain, $\omega = 4$ for natural frequency and $\eta = 10$ for damping, thus a steady state magnitude of an uncoupled oscillator would be $|c| = 1/\sqrt{2} \approx 0.7071$. Further, we took $\alpha = 5$ and $\beta = -10$ as parameters for the coupling coefficients $a_{i,i+1}$. Due to the coupling, steady state magnitudes of the network (6.2) diminish, which can be seen on the graph in Fig. 6.3(a) depicting simulation of the uncontrolled system (6.2) with $u = 0$. However, the system still oscillates. The oscillations can be suppressed by applying control (6.7) with kernel (6.6), where we took $\gamma = 10$. Successful suppression is depicted in Fig. 6.3(b).

6.3 Analysis of synchronization for a ring of Kuramoto oscillators

6.3.1 Overview

In the previous section we were focused on oscillators governed by Stuart-Landau equation, describing oscillations in the complex domain. But there is another special class of dynamical models, the Kuramoto phase oscillator model (Kuramoto 2003), which is a base model that serves to describe synchronization phenomena due to its simplicity. Compared to Stuart-Landau oscillator the Kuramoto oscillator tracks only phase dynamics of a node, assuming amplitude of oscillations being constant. This first-order model captures phenomena arising in coupled Josephson arrays (Wiesenfeld, Colet, and Strogatz 1996), quasi-optical oscillators (York and Compton 1991), etc. The original first-order Kuramoto model was extended by Tanaka, Lichtenberg, and Oishi 1997a; Tanaka, Lichtenberg, and Oishi 1997b who included an additional inertial term, making the whole model second-order. They were inspired by the work of Ermentrout 1991, who introduced a pulse coupled phase oscillator model with inertia to mimic synchronization mechanisms, observed among the fireflies *Pteroptix Malaccae*. Filatrella, Nielsen, and Pedersen 2008 showed that the second-order Kuramoto model can be used as well to describe the collective behaviour of power networks. Other uses include synchronization phenomena in crowd synchrony on London's Millennium Bridge (Strogatz et al. 2005), Huygens pendulum clocks (Bennett et al. 2002), and self-synchronization of smart grids (Salam, Marsden, and Varaiya 1984; Rohden et al. 2012; Dörfler, Chertkov, and Bullo 2013).

The collective behaviour of limit-cycle oscillators was first studied by Winfree 1967 who proposed a mean-field model of coupled phase oscillators with distributed natural frequencies. This work revealed that despite having different natural frequencies, oscillators spontaneously synchronize to some common frequency if the coupling strength exceeds a critical value. Conditions and effects of synchronization in complex networks were first analysed by Watts 2018 via the numerical study of the Kuramoto model in small-world networks and Barahona and Pecora 2002, who considered analytically the conditions for complete synchronization of identical chaotic systems on different kinds of graphs. Synchronization of Kuramoto oscillators in globally coupled networks was extensively analysed in the mean-field sense, see the detailed review by Acebrón et al. 2005b. At the same time, different effects in other topologies were treated mostly numerically, as by Tumash, Olmi, and Schöll 2019. General reviews on synchronization phenomena in complex networks include works of Arenas et al. 2008; Dörfler and Bullo 2014.

In this section we present a network of Kuramoto oscillators with local interactions, namely coupled on a 1D ring. We introduce a PDE approximation for this system using the continuation method. This PDE representation of Kuramoto system can be more appropriate for analysis (in the same way as continuous dynamical systems can be more tractable than the discrete ones). As a toy example, here we present one possible application of this representation, namely we analytically find a synchronization threshold for a 1D ring topology. Problem of computation of a general synchronization threshold for different topologies and frequency distributions was recently solved by Dörfler, Chertkov, and Bullo 2013 and Jafarpour and Bullo 2018 with the help of graph theory. However, the methods presented in these papers are sophisticated and not straightforward to extend to other types of oscillators. Contrary, the idea based on continuation which is presented in this section helps to find a synchronization condition in a very natural way. Moreover this method will be extended in the next Section 6.4 to a more general class of non-isochronous oscillators in the complex domain.

Apart from deriving synchronization threshold we demonstrate that the continuation method produces an accurate representation of the Kuramoto network by performing numerical simulations comparing the original ODE system with the obtained PDE. Finally, considering the Kuramoto system as a model for a power network, we give a design procedure for optimal power which should be produced by generators. We show that there exists a lower bound on the synchronization threshold depending on power loads in the system, and that this bound can be achieved by a careful design of generators. Such choice of generators provides the lowest possible required capacity of the transmission lines, which in theory can lower the construction costs and increase stability of the network.

6.3.2 Problem formulation and continuation

We start by analysing the Kuramoto oscillator system

$$\dot{\phi}_i = \omega_i + F\left(\sin(\phi_{i+1} - \phi_i) - \sin(\phi_i - \phi_{i-1})\right), \tag{6.8}$$

where ϕ_i is a phase angle of i-th oscillator, ω_i is its natural frequency and F is a coupling strength. Each oscillator is coupled with its two neighbours, thus a most natural topology for the system would be a ring (although intervals on a real line with boundary conditions could be also easily considered). We assume that there are n oscillators and that each oscillator has a position on a ring defined by $x_i \in [0, 2\pi)$, with $x_{i+1} - x_i = \Delta x$ and $x_1 - x_n + 2\pi = \Delta x$, meaning that the oscillators are spaced equally on the ring. Using these positions, we can further define a natural frequency function $\omega(x_i) = \omega_i$ and then a state function $\phi(x_i) = \phi_i$.

Continuation of system (6.8) was already performed as an example in Section 4.3 obtaining (4.40) with a change of notation more appropriate for oscillatory systems $\rho \to \phi$. Also a multiplication on the coupling strength F and an addition of the natural frequency function $\omega(x)$ should be performed comparing to (4.40), which leads to the following continuation of (6.8):

$$\frac{\partial \phi}{\partial t} = \omega(x) + F\Delta x \frac{\partial}{\partial x} \sin\left(\Delta x \frac{\partial \phi}{\partial x}\right) = \omega(x) + F \cos\left(\Delta x \frac{\partial \phi}{\partial x}\right) \Delta x^2 \frac{\partial^2 \phi}{\partial x^2}. \tag{6.9}$$

We validate this PDE approximation in the simulation with an ODE system with $n = 50$ oscillators, placed on a ring, such that the last oscillator is connected with the first one. For illustrative purposes the natural frequency function is set as $\omega(x) = 1 + x\sin(2x)$ for $x \in [0, 2\pi)$ (in general any function can be used, but for the future analysis we choose an integrable one) and the coupling strength $F = 4$. We numerically simulate the approximated PDE (6.9) on a grid with 500 points. The results of simulation are shown in Fig. 6.4.

From the figures 6.4(a),(b) we see that the ODE system converged to a partial synchronisation state, having 14 distinct clusters. At the same time PDE model continuously connects these clusters (Inset 6.4(c)), remaining rather accurate at the positions of the oscillators of the original ODE system (Inset 6.4(d)).

6.3.3 Synchronization threshold

The main advantage of describing the system in terms of partial derivatives is that now the space becomes a continuum, thus integrals can be taken (and in general integrals are much more tractable than series).

We will show how the obtained PDE (6.9) can be used to find a parameter F^* for which a phase transition from the complete synchronization to the emergence of clusters occurs. Namely, let us try to find an equilibrium solution ϕ^* of (6.9) in case of complete synchronization. It is clear that then there exists $\bar{\omega} = \frac{1}{2\pi}\int_0^{2\pi} \omega(x)dx$ such that all oscillators share the same frequency:

$$\frac{\partial \phi^*}{\partial t} = \bar{\omega}.$$

One can validate that the synchronization frequency is given by an average value of all oscillators' frequencies by integrating (6.9) over whole space domain. Therefore, the equilibrium solution should satisfy

$$F\Delta x \frac{\partial}{\partial x} \sin\left(\Delta x \frac{\partial \phi^*}{\partial x}\right) = \bar{\omega} - \omega(x). \tag{6.10}$$

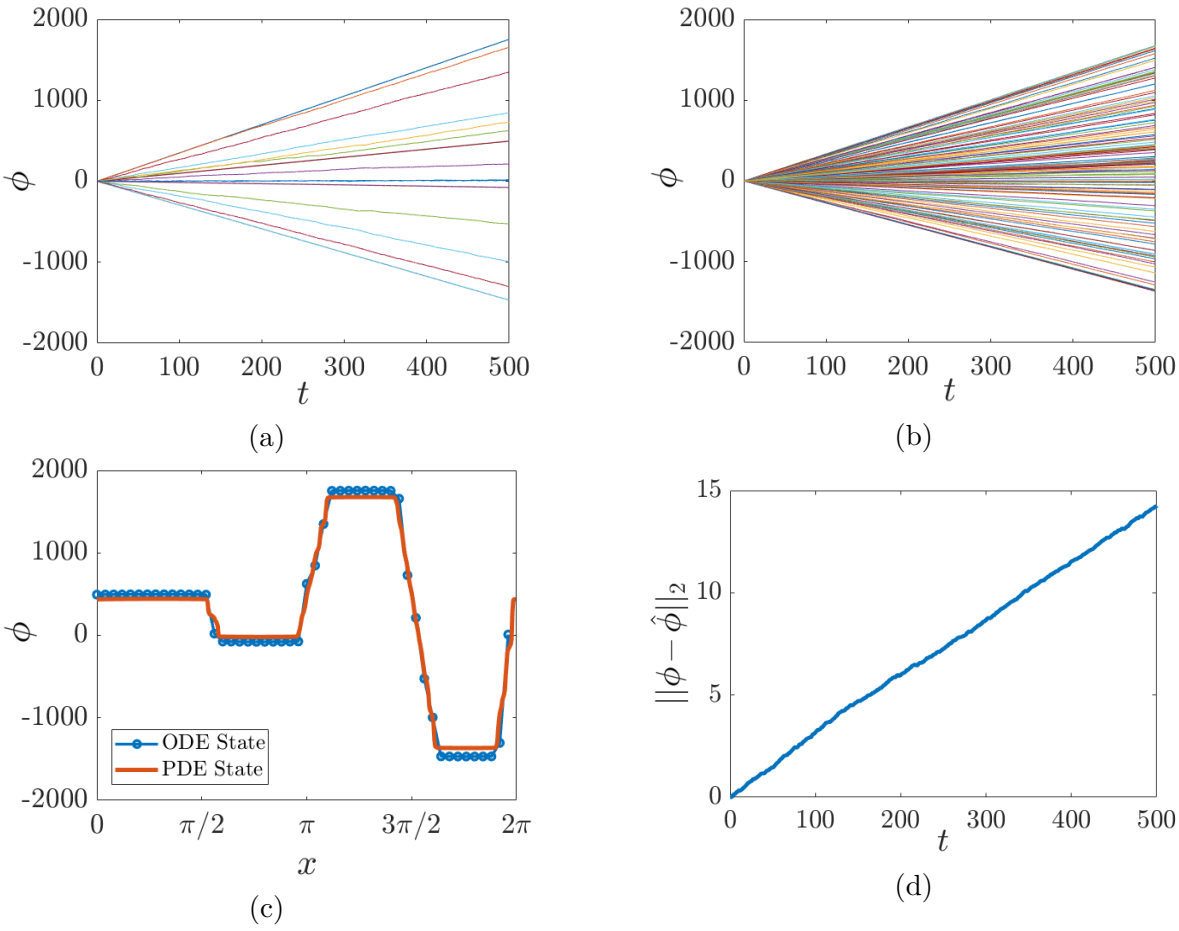


Figure 6.4: **(a):** Simulation of a Kuramoto ODE network (6.8) with $n = 50$ oscillators, different lines denote states of different nodes. **(b):** Simulation of a PDE approximation (6.9) discretized in 500 cells, different lines denote states of different cells. **(c):** Snapshot of profiles of both systems at time $T = 500$. **(d):** Evolution of a mean-square absolute divergence between solutions.

Let us integrate this equation from x_0 to x_1, where both are chosen arbitrary:

$$\sin\left(\Delta x \frac{\partial \phi^*}{\partial x}(x_1)\right) - \sin\left(\Delta x \frac{\partial \phi^*}{\partial x}(x_0)\right) = \frac{1}{F\Delta x}\left(H(x_1) - H(x_0)\right), \tag{6.11}$$

where $H(x)$ is some primitive function of $\bar{\omega} - \omega(x)$. Rearranging, we obtain

$$\sin\left(\Delta x \frac{\partial \phi^*}{\partial x}(x_1)\right) - \frac{1}{F\Delta x} H(x_1) = \sin\left(\Delta x \frac{\partial \phi^*}{\partial x}(x_0)\right) - \frac{1}{F\Delta x} H(x_0) =: C, \tag{6.12}$$

and since x_0 and x_1 were chosen arbitrary, C appears to be some constant independent of the choice of x_0 and x_1. We obtained that the existence of an equilibrium solution is equivalent to the existence of the primitive function $H(x)$ written in the form

$$\frac{1}{F\Delta x} H(x) = \sin\left(\Delta x \frac{\partial \phi^*}{\partial x}(x)\right) + C. \tag{6.13}$$

If such $H(x)$ exists, ϕ^* can be recovered by taking arcsine and then integrating.

Therefore, a complete synchronization for a given F is possible if and only if there exists $H(x)$ such that (6.13) is possible, in a sense that the sinus value lies in the interval $[-1, 1]$. Essentially this means that

$$H(x) \in [-F\Delta x + C, F\Delta x + C] \quad \forall x \in [0, 2\pi].$$

Recalling that $H(x)$ is a primitive function of $\bar{\omega} - \omega(x)$ and that in general it is defined up to a constant, this is equivalent to the condition

$$\max_{x\in[0,2\pi]} H(x) - \min_{x\in[0,2\pi]} H(x) \leqslant 2F\Delta x \tag{6.14}$$

for any $H(x)$. To recover synchronization threshold F^* it requires only to replace inequality with equality sign:

$$F^* = \frac{1}{2\Delta x}\left(\max_{x\in[0,2\pi]} H(x) - \min_{x\in[0,2\pi]} H(x)\right). \tag{6.15}$$

Synchronization threshold (6.15) provides a condition on the existence of equilibrium solutions. It appears that for all $F > F^*$ there will be a stable equilibrium solution:

Theorem 6.1. *For all $F > F^*$, there exists an equilibrium solution ϕ^*, satisfying (6.13), which is locally asymptotically stable.*

Proof. Without loss of generality we assume that $C = 0$ (because the primitive function $H(x)$ is defined up to a constant). Note that $H(x) \in [-F^*\Delta x, F^*\Delta x]$. Then the equilibrium solution can be recovered from (6.13) (up to a constant) as

$$\phi^*(x) = \frac{1}{\Delta x}\int_0^x \arcsin\left(\frac{1}{F\Delta x}H(x)\right)\,dx. \tag{6.16}$$

Now assume that the equilibrium solution is slightly perturbed: $\phi = \phi^* + \tilde{\phi}$. Then, by (6.9),

$$\frac{\partial\tilde{\phi}}{\partial t} = F\Delta x\frac{\partial}{\partial x}\sin\left(\Delta x\frac{\partial\phi^*}{\partial x} + \Delta x\frac{\partial\tilde{\phi}}{\partial x}\right) + \omega(x) - \bar{\omega}. \tag{6.17}$$

Rewriting sine, we get

$$\sin\left(\Delta x\frac{\partial\phi^*}{\partial x} + \Delta x\frac{\partial\tilde{\phi}}{\partial x}\right) = \sin\left(\Delta x\frac{\partial\phi^*}{\partial x}\right)\cos\left(\Delta x\frac{\partial\tilde{\phi}}{\partial x}\right) + \cos\left(\Delta x\frac{\partial\phi^*}{\partial x}\right)\sin\left(\Delta x\frac{\partial\tilde{\phi}}{\partial x}\right) \approx$$
$$\approx \sin\left(\Delta x\frac{\partial\phi^*}{\partial x}\right) + \cos\left(\Delta x\frac{\partial\phi^*}{\partial x}\right)\Delta x\frac{\partial\tilde{\phi}}{\partial x},$$

where the fact that $\tilde{\phi}$ is the small perturbation was used. Now (6.10) cancels the natural frequencies, therefore we arrive at

$$\frac{\partial\tilde{\phi}}{\partial t} = F\Delta x^2\frac{\partial}{\partial x}\left[\cos\left(\Delta x\frac{\partial\phi^*}{\partial x}\right)\frac{\partial\tilde{\phi}}{\partial x}\right]. \tag{6.18}$$

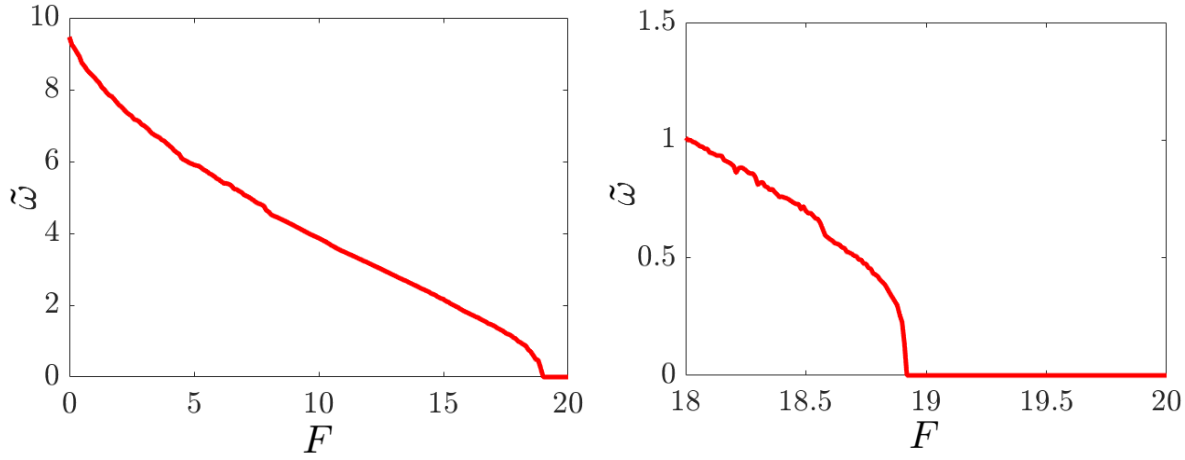


Figure 6.5: Desynchronization frequency $\tilde{\omega}$ depending on the coupling strength F. Other parameters as in Fig. 6.4. **Left:** $F \in [0, 20]$. **Right:** zoom in, $F \in [18, 20]$.

This is a standard linear diffusion equation with the diffusion coefficient $\cos\left(\Delta x \frac{\partial \phi^*}{\partial x}\right)$. For stability it remains to prove that this coefficient is always positive. Indeed,

$$\cos\left(\Delta x \frac{\partial \phi^*}{\partial x}\right) = \cos\left(\arcsin\left(\frac{1}{F\Delta x} H(x)\right)\right) = \sqrt{1 - \left(\frac{H(x)}{F\Delta x}\right)^2} > \sqrt{1 - \left(\frac{H(x)}{F^*\Delta x}\right)^2} \geqslant 0,$$

and thus the linearised system is locally asymptotically stable. □

To validate this analysis we use the parameters of the simulation made in Fig. 6.4: the length of the ring is $L = 2\pi$, the number of ODE nodes $n = 50$, and the natural frequency $\omega(x) = 1 + x\sin(2x)$ (which is an integrable function, thus can be analytically treated). By definition of the positions of nodes, $\Delta x = L/n = 2\pi/50$. Further,

$$\int_0^x \omega(s)ds = x - \frac{1}{2}x\cos(2x) + \frac{1}{4}\sin(2x),$$

thus the average frequency $\bar{\omega} = 1/2$. Primitive function $H(x)$ can be taken as

$$H(x) = \int_0^x (\bar{\omega} - \omega(s))ds = \frac{1}{2}x\cos(2x) - \frac{1}{4}\sin(2x) - \frac{1}{2}x, \tag{6.19}$$

with $\max H(x) = H(3.06) = 0.0203$ and $\min H(x) = H(4.765) = -4.726$. Substituting these values in (6.15) gives

$$F^* \approx 18.88. \tag{6.20}$$

The value (6.20) is the smallest F for which the equilibrium solution exists. To verify the result (6.20) for the original system (6.8), we simulated it for $F \in [0, 20]$ and calculated

$\tilde{\omega} = \max \dot{\phi}_i - \min \dot{\phi}_i$, which we call *desynchronization frequency*. In case of complete synchronization $\tilde{\omega}$ should be zero. Indeed, Fig. 6.5 shows that $\tilde{\omega}$ is zero for $F > 18.9$, and it increases when F becomes smaller.

Note that in general the formula analogous to (6.15) can be obtained also for the ODE case, with the primitive function $H(x)$ being replaced by the partial sum of $\omega(x)$. But for a very large network a computation complexity for the partial sum scales as $O(n)$, where n is the size of the network, while analytical computation for PDE has a complexity $O(1)$.

6.3.4 Kuramoto oscillators with inertia

The Kuramoto oscillator can be used to model the behaviour of power networks, see Filatrella, Nielsen, and Pedersen 2008; Rodrigues et al. 2016. In this case the coupling acts as an electromotive force, while the nodes itself have inertias, thus the second-order model should be considered:

$$\begin{cases} \dot{\phi}_i = \omega_i, \\ M\dot{\omega}_i = \Omega_i - \alpha\omega_i + F\left(\sin(\phi_{i+1} - \phi_i) - \sin(\phi_i - \phi_{i-1})\right), \end{cases} \tag{6.21}$$

where ω_i is a current frequency of the i-th oscillator, which is included in the state now. Physically (see Rodrigues et al. 2016), M represents an inertia coefficient, α is a damping (dissipation), Ω_i is a power supplied to (or taken from) the system, and F is the allowed maximum transferred power.

Performing the same approximation process, we arrive at the following PDE system

$$\begin{cases} \dfrac{\partial \phi}{\partial t} = \omega, \\ M\dfrac{\partial \omega}{\partial t} = \Omega - \alpha\omega + F\Delta x \dfrac{\partial}{\partial x} \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right), \end{cases} \tag{6.22}$$

or just

$$M\frac{\partial^2 \phi}{\partial t^2} + \alpha\frac{\partial \phi}{\partial t} = F\Delta x \frac{\partial}{\partial x} \sin\left(\Delta x \frac{\partial \theta}{\partial x}\right) + \Omega(x). \tag{6.23}$$

Note that the right-hand side of (6.23) coincides with the right-hand side of (6.9), thus the synchronization threshold F^* is exactly the same for (6.23) as for (6.9). One can simply use the formula (6.15) to obtain the synchronization threshold for the Kuramoto model with inertia substituting natural frequency $\omega(x)$ with supplied power $\Omega(x)$.

6.3.5 Design of generators in power networks

When the power network is modelled by (6.21), the value of $\Omega(x)$ represents the power which the node brings to the system (or consumes, if $\Omega(x)$ is negative). Without loss of generality, we can assume that $\bar{\Omega} = 0$ (which can be done by treating deviations from the normal behaviour).

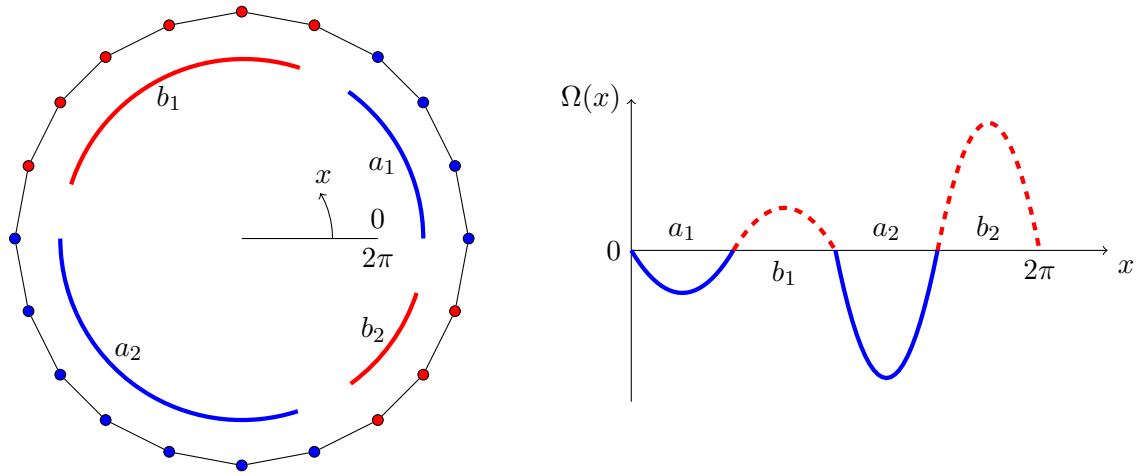


Figure 6.6: **Left:** Schematic representation of a power grid on a ring with $n = 20$ nodes. Red dots represent generators and blue dots represent loads. a_1 and a_2 define loads intervals, b_1 and b_2 define generators intervals. **Right:** Intervals of loads and generators. $\Omega(x)$ is negative on the intervals a_j of loads and positive on the intervals b_j of generators. Given loads, the goal is to design the generators.

Then it makes sense to split all the nodes into two classes: generators (with $\Omega(x) > 0$) and loads (with $\Omega(x) < 0$).

The desired behaviour of power networks is the synchronization, thus the system can be considered more stable if the synchronization threshold is smaller. Namely, Filatrella, Nielsen, and Pedersen 2008; Rodrigues et al. 2016 showed that the coupling strength F is the maximal power which the line connecting two oscillators should be able to transmit. And if this capacity is not enough, system desynchronizes, leading to a blackout.

Therefore, it is desirable to design such system that F^* required for synchronization is as small as possible. This leads to a cheaper construction of power system transmission lines and in the same time increases stability allowing for more fluctuations to happen before a blackout. As the large loads (such as factories) are usually known, one can ask a question what is the optimal design of generators such that the synchronization threshold is the smallest one.

Let the power network be the ring of n oscillators with coordinates $x_i \in [0, 2\pi)$, which can be represented by a PDE (6.9) (or (6.23) for the model with inertia). Then the ring can be split into $2K$ intervals a_1, b_1, a_2, ..., a_K, b_K (see Fig. 6.6), such that $\Omega(x)$ is negative on the intervals a_j and positive on the intervals b_j for $j \in \{1, ..., K\}$. Further let us define positions of intervals ends by x_{a_j} and x_{b_j}.

Denote

$$A_j = \int_{a_j} \Omega(x)dx \quad \text{and} \quad B_j = \int_{b_j} \Omega(x)dx, \quad j \in \{1, ..., K\}.$$

Then it is obvious that the primitive function $H(x) = \int \Omega(x)dx$ has its minimum points at

x_{a_j}, maximum points at x_{b_j}, and that

$$H(x_{a_j}) = \sum_{i=1}^{j} A_i + \sum_{i=1}^{j-1} B_i, \quad H(x_{b_j}) = \sum_{i=1}^{j} (A_i + B_i)\,. \tag{6.24}$$

Synchronization threshold F^* is defined via (6.15), and the way to minimize it is to minimize the difference between the maximal and the minimal values of $H(x)$.

Proposition 6.1. *Irrespectively of generator powers $\Omega(x)$ on intervals b_j,*

$$\max_{x\in[0,2\pi)} H(x) - \min_{x\in[0,2\pi)} H(x) \geqslant \max_{j\in\{1,\dots,K\}} |A_j|. \tag{6.25}$$

Proof. Indeed,

$$\max_{x\in[0,2\pi)} H(x) - \min_{x\in[0,2\pi)} H(x) \geqslant H(x_{b_{j-1}}) - H(x_{a_j}) = |A_j|,$$

and taking maximum for all $j \in \{1, ..., K\}$ results in (6.25). □

Proposition 6.1 shows that there is an intrinsic bound on the synchronization threshold which is defined by loads and cannot be lowered by any choice of generators. But this bound can be achieved in many different ways. One of them is to use generators to compensate for their neighbouring loads.

Proposition 6.2. *Let the integral power of generators be such that*

$$B_j = -\alpha A_j - (1-\alpha)A_{j+1}, \quad j \in \{1, ..., K\}, \tag{6.26}$$

where A_{K+1} is equivalent to A_1 and $\alpha \in [0, 1]$ is the splitting coefficient. Then the lower bound in inequality (6.25) is achieved.

Proof. By (6.24) and (6.26)

$$H(x_{a_j}) = (1-\alpha)A_1 + \alpha A_j, \qquad H(x_{b_j}) = (1-\alpha)A_1 - (1-\alpha)A_{j+1}.$$

One can see that such choice of generators assures

$$\bar{\Omega} =< \Omega >= 0.$$

Then, x_{a_j} are the minimum points, x_{b_j} are the maximum points, therefore

$$\begin{aligned}
&\max_{x\in[0,2\pi)} H(x) - \min_{x\in[0,2\pi)} H(x) = \max_{j\in\{1,\dots,K\}} H(x_{b_j}) - \min_{j\in\{1,\dots,K\}} H(x_{a_j}) = \\
&\max_{j\in\{1,\dots,K\}} ((1-\alpha)(A_1 - A_{j+1})) - \min_{j\in\{1,\dots,K\}} ((1-\alpha)A_1 + \alpha A_j) = \\
&= (1-\alpha) \max_{j\in\{1,\dots,K\}} |A_j| + \alpha \max_{j\in\{1,\dots,K\}} |A_j| = \max_{j\in\{1,\dots,K\}} |A_j|,
\end{aligned}$$

which gives the equality in (6.25). □

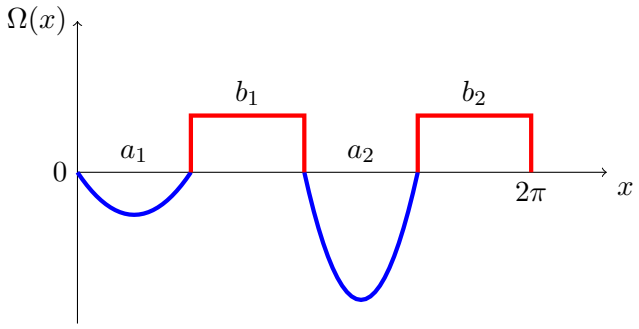


Figure 6.7: Intervals of loads and generators with generators designed according to the Proposition 6.2 with Remark 6.1. $\Omega(x)$ on intervals b_j is an average of the neighbouring loads.

Remark 6.1. By Proposition 6.1 one particular choice of the power of generators is

$$\Omega(x) = \frac{|A_j| + |A_{j+1}|}{2|b_j|}, \quad x \in b_j, \quad j \in \{1, ..., K\}, \tag{6.27}$$

where $|b_j|$ denotes the length of the interval b_j. Such generators equally compensate for the average of their neighbour loads, providing the optimal result (see Fig. 6.7).

Finally, synchronization threshold

$$F^* = \frac{\max_{j\in\{1,...,K\}} |A_j|}{2\Delta x} \tag{6.28}$$

should be large enough such that the coupling overcomes the strongest load.

6.4 Analysis of synchronization for a ring of non-isochronous oscillators

6.4.1 Overview

It was shown in the previous section that synchronization analysis can be easily performed for Kuramoto oscillators using PDE representation. The same synchronization conditions were already derived for ODE representation by Dörfler, Chertkov, and Bullo 2013 and Jafarpour and Bullo 2018, however these results were based on an extensive use of nontrivial graph theory and linear algebra. We can further show that PDE-based models allow for more natural analysis of systems by applying the continuation method for more complex class of oscillators, namely non-isochronous oscillators. Non-isochronous oscillators are characterized by a coupling of amplitude and phase. This class of models generalizes both Stuart-Landau oscillators in Section 6.2 and Kuramoto oscillators in Section 6.3 and includes many important physical systems, such as spin-torque oscillators (STO), Van der Pol oscillators, neuron models

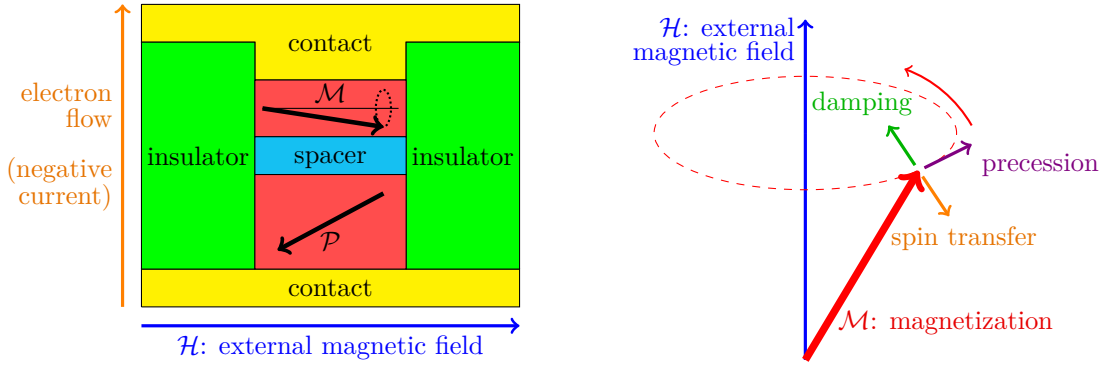


Figure 6.8: **Left:** Schematic representation of a possible geometry of spin-torque oscillator. Red blocks represent ferromagnetic layers with their magnetization directions denoted by black arrows. Electrons flow from bottom to top, first passing through the "fixed" magnetic layer which induces spin polarization coinciding with its magnetization direction $\mathcal{P}$. The magnetization $\mathcal{M}$ of the "free" magnetic layer then oscillates under the effect of polarized current and the external magnetic field $\mathcal{H}$. **Right:** Close view on the dynamics of the magnetization $\mathcal{M}$ of the "free" layer, governed by equation (6.29). Damping and current-induced spin-transfer torque compensate each other, stabilizing steady oscillations caused by precession around the magnetic field $\mathcal{H}$.

or many others. Here we will focus mostly on STO systems, however the analysis in this section can be applied to other types of non-isochronous oscillators.

Spin-torque oscillators are based on the spin-transfer torque effect discovered by Slonczewski 1996 and Berger 1996. It appears that an electric direct current which passes through a magnetized layer can become spin-polarized, and moreover this spin-polarized current can further transfer angular momentum to another magnetized layer. This transfer induces torque on the magnetization of the second layer, which can lead to switching of the magnetization direction. In presence of an external magnetic field a steady magnetization precession can be achieved instead of direction switching. Thus a typical spin-torque oscillator consists of two ferromagnetic layers, a thick one called "fixed" and a thin one called "free", see the left panel of Fig. 6.8. Magnetization direction of the "fixed" layer turns electrons' spin in the current in the same direction, which then induces torque on the magnetization of the "free" layer, thus creating precession, depicted in the right panel of Fig. 6.8. Denote the magnetization of the "free" magnetic layer by vector $\mathcal{M}$, the magnetization of the "fixed" magnetic layer by vector $\mathcal{P}$ and the external magnetic field by vector $\mathcal{H}$. Then the magnetization is governed by the Landau-Lifshitz-Gilbert equation:

$$\frac{\partial \mathcal{M}}{\partial t} = \underbrace{-\gamma\,(\mathcal{M}\times\mathcal{H})}_{\text{precession}} + \underbrace{\frac{\alpha}{|\mathcal{M}|}\left(\mathcal{M}\times\frac{\partial\mathcal{M}}{\partial t}\right)}_{\text{damping}} + \underbrace{\frac{\sigma I}{|\mathcal{M}|}\,(\mathcal{M}\times(\mathcal{M}\times\mathcal{P}))}_{\text{spin transfer}}, \tag{6.29}$$

where parameters γ, α and σ depend on system's geometry and materials, and I is a current which is applied to the system. For a review of the spin-transfer torque effect and STOs see Slavin and Tiberkevich 2009; Stiles and Miltat 2006; Dieudonné 2015.

Due to the fast precession of magnetization in a ferromagnetic layer, STOs produce microwaves. Thus large arrays of STOs can theoretically serve as very efficient microwave generators. This is why the question of synchronization of STOs is very important: synchronous oscillations of many oscillators amplify each other due to constructive interference, while asynchronous oscillations exhibit destructive interference and thus produce less power. In this section we will focus on the analysis of synchronization of a set of oscillators coupled in a ring topology, providing conditions which can guarantee synchronization depending on the parameters of independent oscillators.

Equation (6.29) can be simplified for analysis. Magnetization vector $\mathcal{M}$ oscillates around magnetic field vector $\mathcal{H}$. Let us project $\mathcal{M}$ on a plane orthogonal to $\mathcal{H}$ and denote the resulting projection via a complex variable c. Then, with some additional transformations and simplifications (see Slavin and Tiberkevich 2009 for details) it is possible to show that the magnetization dynamics (6.29) of an STO can be modelled through:

$$\dot{c} = i(\omega + Np)c - \Gamma_G(1 + Qp)c + \sigma I(1 - p)c, \tag{6.30}$$

where $p = |c|^2$ represents a squared amplitude of oscillations, ω is a base frequency, N is a frequency gain with respect to the amplitude, Γ_G is a base damping, Q is a damping amplitude gain, and I and σ are the same as in (6.29). Model (6.30) is nonlinear since the oscillations' frequency depends on the amplitude through the frequency gain N. In case of spin-torque oscillators this amplitude-related frequency shift happens to be very strong, thus these oscillations cannot be described by simpler linear models.

If $\sigma I \leqslant \Gamma_G$, the origin $c = 0$ is a stable equilibrium point. Oscillations will occur if $\sigma I > \Gamma_G$. Assuming it is true, define a linear part of sum of dissipative terms $\Gamma = \sigma I - \Gamma_G > 0$ and further a nonlinear gain of sum of dissipative terms $S = \Gamma_G Q + \sigma I$, thus the system (6.30) can be written as

$$\dot{c} = i(\omega + Np)c + (\Gamma - Sp)c. \tag{6.31}$$

System (6.31) will oscillate with amplitude $|c| = \sqrt{p} = \sqrt{\Gamma/S}$ and with frequency $\dot{\phi} = \omega + N\Gamma/S$, where ϕ is a phase of an oscillator. For the amplitude of oscillations to be well defined, we also require $S > 0$.

6.4.2 Logarithmic representation for a ring of non-isochronous oscillators

Model (6.31) is often studied in amplitude-phase representation $c = \sqrt{p}e^{i\phi}$, where $\sqrt{p}$ is an amplitude of oscillations and ϕ is a phase of an oscillator. Instead of writing two separate equations for them, we will write model (6.31) in logarithmic representation. Define $z = \ln c$. Then the real part of z will represent the amplitude, namely $\exp\{2\operatorname{Re} z\} = p$. Let us denote $r := \operatorname{Re} z = \frac{1}{2}\ln p$. The imaginary part of z is a phase of an oscillator, $\phi := \operatorname{Im} z$, thus such transformation allows to track phase information immediately. Since $dc = c \cdot dz$, the model (6.31) now becomes

$$\dot{z} = \Gamma + i\omega - (S - iN)e^{2\operatorname{Re} z}. \tag{6.32}$$

Now let us move to a system of coupled oscillators. We assume the oscillators are placed on a ring, and each oscillator is coupled with its two neighbours. As in the previous section, let n denote the number of oscillators and let $x_i \in [0, 2\pi)$ be a position on a ring of the i-th oscillator. The distance between oscillators is $x_{i+1} - x_i = \Delta x$ and $x_1 - x_n + 2\pi = \Delta x$, meaning that the oscillators are spaced equally on the ring. Coupling between oscillators means that each oscillator has its neighbors' states as an external force:

$$\dot{c}_i = i(\omega_i + N_i p_i)c_i + (\Gamma_i - S_i p_i)c_i + F_i(c_{i-1} + c_{i+1}). \tag{6.33}$$

Here F_i is a (possibly complex) coupling constant, with an amplitude representing coupling strength and a phase representing coupling phase.

Using logarithmic representation, the model (6.33) reads as

$$\dot{z}_i = \Gamma_i + i\omega_i - (S_i - iN_i)e^{2\operatorname{Re} z_i} + F_i\left(e^{z_{i-1}-z_i} + e^{z_{i+1}-z_i}\right). \tag{6.34}$$

6.4.3 Continuation and synchronization condition

It is now possible to perform continuation for the coupled system in the same way it was done for Kuramoto oscillators in the previous section. In theory one could use the continuation method for (6.33), but it appears that in this case an intrinsic difference in scales between the amplitude and the phase dynamics is lost. Indeed, in physical systems the phase of an oscillator usually changes much faster than its amplitude. Instead one should perform continuation in such a way that the amplitude and the phase dynamics are clearly separated. One way to do so is to consider separate equations for amplitude and phase. Another way, which we will choose, is to use logarithmic representation (6.34). Therefore, the continuation of (6.34) is performed in several steps:

1. $z_{i-1} - z_i \quad \to \quad -\Delta x\, \partial z/\partial x_{i-1/2}$
2. $z_{i+1} - z_i \quad \to \quad \Delta x\, \partial z/\partial x_{i+1/2}$
3. $e^{-\Delta x\, \partial z/\partial x_{i-1/2}} \quad \to \quad e^{-\Delta x\, \partial z/\partial x_i} - \Delta x \frac{1}{2}\frac{\partial}{\partial x} e^{-\Delta x\, \partial z/\partial x_i}$
4. $e^{\Delta x\, \partial z/\partial x_{i+1/2}} \quad \to \quad e^{\Delta x\, \partial z/\partial x_i} + \Delta x \frac{1}{2}\frac{\partial}{\partial x} e^{\Delta x\, \partial z/\partial x_i}$

Using these continuations, we finally get

$$e^{z_{i-1}-z_i} + e^{z_{i+1}-z_i} \;\to\; \left(e^{\Delta x \frac{\partial z}{\partial x}} + e^{-\Delta x \frac{\partial z}{\partial x}}\right) + \Delta x \frac{\partial}{\partial x}\left(\frac{e^{\Delta x \frac{\partial z}{\partial x}} - e^{-\Delta x \frac{\partial z}{\partial x}}}{2}\right),$$

or simply

$$e^{z_{i-1}-z_i} + e^{z_{i+1}-z_i} \;\to\; 2\cosh\left(\Delta x \frac{\partial z}{\partial x}\right) + \Delta x \frac{\partial}{\partial x}\sinh\left(\Delta x \frac{\partial z}{\partial x}\right).$$

Thus, system (6.34) can be written using PDE model as

$$\frac{\partial z}{\partial t} = \Gamma + i\omega - (S - iN)e^{2\operatorname{Re} z} + F\left[2\cosh\left(\Delta x \frac{\partial z}{\partial x}\right) + \Delta x \frac{\partial}{\partial x}\sinh\left(\Delta x \frac{\partial z}{\partial x}\right)\right], \tag{6.35}$$

where parameters Γ, ω, S, N and F are (possibly) varying functions of space, determined by approximating sampled values Γ_i, ω_i, S_i, N_i and F_i at points x_i.

Separating (6.35) into a system of two equations for $r = \operatorname{Re} z$ and $\phi = \operatorname{Im} z$, one gets

$$\begin{cases} \dfrac{\partial r}{\partial t} = \Gamma - Se^{2r} \\ \quad + \operatorname{Re} F \left[2\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right] \\ \quad - \operatorname{Im} F \left[2\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right], \\ \dfrac{\partial \phi}{\partial t} = \omega + Ne^{2r} \\ \quad + \operatorname{Re} F \left[2\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right] \\ \quad + \operatorname{Im} F \left[2\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right]. \end{cases} \tag{6.36}$$

In is interesting to note that (6.36) includes a standard Kuramoto PDE derived in Section 6.3.2 as a particular case. Indeed, assuming $r = r_0 = const$ both in space and time and assuming $F \in \mathbb{R}$, one gets an equation for ϕ as

$$\frac{\partial \phi}{\partial t} = \omega + Ne^{2r_0} + F\Delta x \frac{\partial}{\partial x} \sin\left(\Delta x \frac{\partial \phi}{\partial x}\right), \tag{6.37}$$

which exactly coincides with (6.9) changing ω to $\omega + Ne^{2r_0}$.

Similar to Section 6.3.3 we are interested in possible synchronized solutions of (6.35) and conditions for their existence and stability. A synchronized solution is such solution to (6.35) that $\partial z/\partial t = i\bar{\omega}$, where $\bar{\omega}$ is a synchronization frequency. Thus we are interested in a question when such a solution $z = z(x)$ exists for some $\bar{\omega}$. Then the condition for synchronization is

$$0 = \Gamma + i(\omega - \bar{\omega}) - (S - iN)e^{2\operatorname{Re} z} + F\left[2\cosh\left(\Delta x \frac{\partial z}{\partial x}\right) + \Delta x \frac{\partial}{\partial x}\sinh\left(\Delta x \frac{\partial z}{\partial x}\right)\right], \tag{6.38}$$

or in terms of $r(x)$ and $\phi(x)$

$$\begin{cases} 0 = \operatorname{Re}\left(F^{-1}\left[\Gamma + i(\omega - \bar{\omega}) - (S - iN)e^{2r}\right]\right) + \\ \quad + \left[2\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \cos\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right], \\ 0 = \operatorname{Im}\left(F^{-1}\left[\Gamma + i(\omega - \bar{\omega}) - (S - iN)e^{2r}\right]\right) + \\ \quad + \left[2\sinh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right) + \Delta x \dfrac{\partial}{\partial x}\left(\cosh\left(\Delta x \dfrac{\partial r}{\partial x}\right) \sin\left(\Delta x \dfrac{\partial \phi}{\partial x}\right)\right)\right]. \end{cases} \tag{6.39}$$

Note that we divided the equation by F before splitting real and imaginary parts such that the hyperbolic functions take the simplest form.

Exponential term e^{2r} in (6.39) can be removed by combining two equations together. Using amplitude-phase notation we can introduce $f = |F|$, $\beta = \arg F$, $G = |S + iN|$ and $\gamma = \arg(S + iN)$. With this notation

$$\mathrm{Re}\left(F^{-1}(S - iN)\right) = \frac{G}{f}\cos(\gamma + \beta), \qquad \mathrm{Im}\left(F^{-1}(S - iN)\right) = -\frac{G}{f}\sin(\gamma + \beta),$$

therefore defining $A = \tan(\gamma + \beta)$, multiplying the first equation in (6.39) by A and summing it with the second one we obtain

$$\begin{aligned} A\Delta x\frac{\partial}{\partial x}\left[\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right)\sinh\left(\Delta x\frac{\partial r}{\partial x}\right)\right] + \Delta x\frac{\partial}{\partial x}\left[\sin\left(\Delta x\frac{\partial\phi}{\partial x}\right)\cosh\left(\Delta x\frac{\partial r}{\partial x}\right)\right] + \\ + 2\left[A\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right)\cosh\left(\Delta x\frac{\partial r}{\partial x}\right) + \sin\left(\Delta x\frac{\partial\phi}{\partial x}\right)\sinh\left(\Delta x\frac{\partial r}{\partial x}\right)\right] + B = 0, \end{aligned} \tag{6.40}$$

where

$$B = \frac{1}{f\cos(\gamma + \beta)}\left[\cos\gamma\,(\omega - \bar{\omega}) + \sin\gamma\,\Gamma\right]. \tag{6.41}$$

Therefore, the synchronization condition is equivalent to (6.40) combined with one of the equations in (6.39) to determine connection between r and ϕ.

6.4.4 Identical oscillators case

In this section we will focus on the case when the ring consists of oscillators having identical parameters. Intuitively it is clear that in this case there exists a solution where all oscillators share the same amplitude r and the same phase ϕ. However it appears that depending on the number of oscillators and their parameters there can be more solutions, and that their stability properties are not trivial.

First let us assume that in the synchronized case the amplitudes of oscillators r are the same, namely $r(x) = r^* = const$. In Section 6.4.6 we will prove that this assumption is indeed valid, because it can be shown that $r(x)$ should be monotone with respect to x, which is possible on the ring only if $r(x)$ is constant. Since $\partial r/\partial x = 0$, we can use $\sinh(\Delta x\ \partial r/\partial x) = 0$ and $\cosh(\Delta x\ \partial r/\partial x) = 1$. With these simplifications the equation (6.40) depends only on $\phi(x)$ and thus can be solved independenty:

$$\Delta x\frac{\partial}{\partial x}\sin\left(\Delta x\frac{\partial\phi}{\partial x}\right) + 2A\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right) + B = 0. \tag{6.42}$$

If the parameters of oscillators would be non-identical, the equation (6.42) would be very difficult to solve analytically since A and B are varying functions of space (at most it can be converted to the Abel equation of the second kind). Therefore in this section we assume A and B being constant. A more general scenario of a piecewise constant functions A and B will be covered in the next section.

For constant A and B equation (6.42) is separable. We can notice that it depends only on the derivative of $\phi(x)$, not on the phase itself. Define $\theta = \Delta x\, \partial\phi/\partial x$. A physical meaning of θ is a difference in phases between two consecutive oscillators. With this definition, (6.42) becomes

$$\frac{\cos\theta}{-B - 2A\cos\theta} d\theta = \frac{1}{\Delta x} dx. \tag{6.43}$$

Integration of equation (6.43) is performed in Appendix A.6, admitting two solutions depending on $J := B/A$:

$$A\frac{x}{\Delta x} + C = \frac{J}{2\sqrt{4-J^2}} \ln \left| \frac{1 + \left(\frac{2-J}{\sqrt{4-J^2}} \tan\frac{\theta}{2}\right)}{1 - \left(\frac{2-J}{\sqrt{4-J^2}} \tan\frac{\theta}{2}\right)} \right| - \frac{1}{2}\theta \tag{6.44}$$

for $|J| < 2$ and

$$A\frac{x}{\Delta x} + C = \frac{J}{\sqrt{J^2-4}} \arctan\left(\frac{J-2}{\sqrt{J^2-4}} \tan\frac{\theta}{2}\right) - \frac{1}{2}\theta \tag{6.45}$$

for $|J| > 2$, with C being integration constant. Also in case $|J| < 2$ equation (6.42) has a constant solution $\cos\theta = -J/2$.

Apart from being a solution to (6.42), synchronization means that the solution $\phi(x)$ is a continuous angle, thus $\phi(x + 2\pi) - \phi(x) = 2\pi k$ for some $k \in \mathbb{Z}$. This implies two conditions which θ should satisfy:

$$\begin{aligned} &1). \quad \theta(x) = \theta(x + 2\pi) \quad \forall x \in \mathbb{R}, \\ &2). \quad \int_0^{2\pi} \theta(x)\, dx = 2\pi\Delta x k \quad \text{for some } k \in \mathbb{Z}. \end{aligned} \tag{6.46}$$

Finally, a nonconstant solution can't reach $\theta = \pm\pi/2$, since in this case left-hand side of (6.43) becomes zero and the solution becomes undefined. This corresponds to the fact that solutions (6.44) and (6.45) written in form $x = g(\theta)$ are strictly monotone functions such that the inverse function $\theta = g^{-1}(x)$ could exist.

In the case of identical oscillators with constant A and B it appears that the only possible solution to (6.42) is a constant one. Indeed, non-constant solutions (6.44) and (6.45) should be monotone with respect to coordinate, however the first condition in (6.46) requires θ to be periodic, which is not possible if θ is not constant. Thus all possible synchronized solutions for (6.43) are given by

$$\theta = \arccos\left(-\frac{B}{2A}\right). \tag{6.47}$$

6.4.4.1 Equilibrium points

Recall that $\theta = \Delta x \frac{\partial\phi}{\partial x}$. Since $\phi(x)$ is a phase, it is defined up to a constant. Assuming $\phi(x) = 0$ at $x = 0$, using (6.47) and the definitions of A and B, the solution for $\phi(x)$ is a linear function

$$\phi(x) = \frac{x}{\Delta x} \arccos\left(-\frac{\cos\gamma\,(\omega - \bar{\omega}) + \sin\gamma\,\Gamma}{2f\sin(\beta+\gamma)}\right). \tag{6.48}$$

Note that $\bar{\omega}$ is a synchronization frequency and is still unknown in this equation.

Position x is itself defined on a ring, thus $x \in [0, 2\pi)$. Moreover, since the equilibrium solution is a periodic function, $\phi(2\pi)$ should also be a multiple of 2π. We can define $k \in \mathbb{Z}_+$ such that $\phi(2\pi) = 2\pi k$. Therefore, the solution can exist for any $\bar{\omega}$ such that

$$k = \frac{1}{\Delta x} \arccos\left(-\frac{\cos\gamma\,(\omega - \bar{\omega}) + \sin\gamma\,\Gamma}{2f\sin(\beta+\gamma)}\right), \quad k \in \mathbb{Z}_+.$$

The case $k = 0$ corresponds to an *in-phase* synchronized system, meaning phases of all oscillators coincide, while the case $k = 1$ corresponds to the state where the phases of the oscillators do a round turn along the ring. It is clear that in general the phase difference between neighbours is

$$\theta^* = k\Delta x.$$

Note also that the system is symmetric for simultaneous substitution $k \to -k$ and $x \to -x$, thus phases can turn both clockwise and counter-clockwise along the ring.

The principal branch of arccos has a range of values $[0, \pi]$, therefore k should satisfy $k \leqslant \pi/\Delta x$ (other solutions will just copy the ones included in this range due to periodicity). Since Δx is defined as the distance between two oscillators and is assumed to be constant, $\Delta x = 2\pi/n$, where n is the number of oscillators in the system. Thus $k \leqslant n/2$, with $k = n/2$ corresponding to the case when two neighbor oscillators are in anti-phase.

The synchronization frequency is thus given by

$$\bar{\omega} = \omega + \tan\gamma\,\Gamma + 2f\frac{\sin(\beta+\gamma)}{\cos\gamma}\cos(k\Delta x), \quad k \in \left\{0, ..., \frac{n}{2}\right\}. \tag{6.49}$$

In particular, depending on the sign of $2f\frac{\sin(\beta+\gamma)}{\cos\gamma}$, the in-phase synchronized state is either the fastest or the slowest one.

6.4.4.2 Stability analysis

Assume the equilibrium solution is given by (6.48) with the frequency (6.49) for $k \in \{0, ..., n/2\}$. We want to study for which of these k the solution is stable.

Define $z^*(x,t) = r^* + i\phi^*(x) + i\bar{\omega}t$ to be an equilibrium solution for (6.35). Thus $\phi^*(x)$ is defined by (6.48) for a chosen k, $\bar{\omega}$ is a frequency of synchronized solution (6.49), and a constant r^* can be found from (6.38) by taking its real part:

$$e^{2r^*} = \frac{\Gamma + 2\,\mathrm{Re}\,F\cos(k\Delta x)}{S}. \tag{6.50}$$

Note that exponential should be positive to be well defined, therefore we require

$$\Gamma + 2f\cos\beta\cos(k\Delta x) > 0.$$

Now let us define a deviation from the equilibrium solution $\tilde{z}(x,t) = z(x,t) - z^*(x,t)$. It is governed by a difference of (6.35) for $z(x,t)$ and for $z^*(x,t)$, taking into account (6.38). Assuming $\tilde{z}(x,t)$ is small, linearization of (6.35) around $z^*(x,t)$ is given by

$$\begin{aligned}\frac{\partial \tilde{z}}{\partial t} = -2(S - iN)e^{2r^*}\operatorname{Re}\tilde{z} + 2F\sinh\left(\Delta x\frac{\partial z^*}{\partial x}\right)\Delta x\frac{\partial \tilde{z}}{\partial x} + \\ + F\Delta x\frac{\partial}{\partial x}\left[\cosh\left(\Delta x\frac{\partial z^*}{\partial x}\right)\Delta x\frac{\partial \tilde{z}}{\partial x}\right].\end{aligned} \tag{6.51}$$

Using

$$\Delta x\frac{\partial z^*}{\partial x} = i\Delta x\frac{\partial \phi^*}{\partial x} = i\theta^* = ik\Delta x,$$

we get

$$\cosh\left(\Delta x\frac{\partial z^*}{\partial x}\right) = \cos(k\Delta x), \quad \sinh\left(\Delta x\frac{\partial z^*}{\partial x}\right) = i\sin(k\Delta x),$$

which can be substituted in (6.51), resulting in

$$\frac{\partial \tilde{z}}{\partial t} = -2(S - iN)e^{2r^*}\operatorname{Re}\tilde{z} + 2iF\Delta x\sin(k\Delta x)\frac{\partial \tilde{z}}{\partial x} + F\Delta x^2\cos(k\Delta x)\frac{\partial^2 \tilde{z}}{\partial x^2}. \tag{6.52}$$

Separating (6.52) into real and imaginary parts $\tilde{z} = \tilde{r} + i\tilde{\phi}$ and using $F = fe^{i\beta}$:

$$\begin{cases}\dfrac{\partial \tilde{r}}{\partial t} = -2Se^{2r^*}\tilde{r} - 2f\Delta x\sin\beta\sin(k\Delta x)\dfrac{\partial \tilde{r}}{\partial x} - 2f\Delta x\cos\beta\sin(k\Delta x)\dfrac{\partial \tilde{\phi}}{\partial x} + \\ \qquad + f\Delta x^2\cos\beta\cos(k\Delta x)\dfrac{\partial^2 \tilde{r}}{\partial x^2} - f\Delta x^2\sin\beta\cos(k\Delta x)\dfrac{\partial^2 \tilde{\phi}}{\partial x^2}, \\ \dfrac{\partial \tilde{\phi}}{\partial t} = 2Ne^{2r^*}\tilde{r} + 2f\Delta x\cos\beta\sin(k\Delta x)\dfrac{\partial \tilde{r}}{\partial x} - 2f\Delta x\sin\beta\sin(k\Delta x)\dfrac{\partial \tilde{\phi}}{\partial x} + \\ \qquad + f\Delta x^2\sin\beta\cos(k\Delta x)\dfrac{\partial^2 \tilde{r}}{\partial x^2} + f\Delta x^2\cos\beta\cos(k\Delta x)\dfrac{\partial^2 \tilde{\phi}}{\partial x^2}.\end{cases} \tag{6.53}$$

System (6.53) is a system of linear equations, thus the method of separation of variables can be applied to solve it. Moreover, it is homogeneous, thus the basis functions should be exponential. Therefore stability of (6.53) can be checked by substituting exponential basis functions

$$\tilde{r} = r_0e^{\lambda t}e^{imx}, \quad \tilde{\phi} = \phi_0e^{\lambda t}e^{imx} \tag{6.54}$$

for some $\lambda \in \mathbb{C}$ and $m \in \mathbb{Z}$, since basis should be periodic in x along the ring. For asymptotic stability there should exist no solution of (6.53) with $\operatorname{Re}\lambda > 0$. Substituting (6.54) in (6.53) one gets

$$\begin{cases}\lambda r_0 = -2Se^{2r^*}r_0 - 2f\Delta x\sin\beta\sin(k\Delta x)imr_0 - 2f\Delta x\cos\beta\sin(k\Delta x)im\phi_0 - \\ \qquad - f\Delta x^2\cos\beta\cos(k\Delta x)m^2r_0 + f\Delta x^2\sin\beta\cos(k\Delta x)m^2\phi_0, \\ \lambda\phi_0 = 2Ne^{2r^*}r_0 + 2f\Delta x\cos\beta\sin(k\Delta x)imr_0 - 2f\Delta x\sin\beta\sin(k\Delta x)im\phi_0 - \\ \qquad - f\Delta x^2\sin\beta\cos(k\Delta x)m^2r_0 - f\Delta x^2\cos\beta\cos(k\Delta x)m^2\phi_0.\end{cases} \tag{6.55}$$

Define

$$\begin{aligned}P &= f\Delta x^2\cos\beta\cos(k\Delta x)m^2 + 2if\Delta x\sin\beta\sin(k\Delta x)m, \\ Q &= -f\Delta x^2\sin\beta\cos(k\Delta x)m^2 + 2if\Delta x\cos\beta\sin(k\Delta x)m.\end{aligned}$$

Note that as $m \to \infty$, first terms become dominating. With the help of these functions and with $\bar{S} = 2Se^{2r^*} > 0$ and $\bar{N} = 2Ne^{2r^*}$, (6.55) becomes

$$\lambda \begin{pmatrix} r_0 \\ \phi_0 \end{pmatrix} = \begin{pmatrix} -P - \bar{S} & -Q \\ \bar{N} + Q & -P \end{pmatrix} \begin{pmatrix} r_0 \\ \phi_0 \end{pmatrix}, \tag{6.56}$$

thus we are interested in the eigenvalues of the matrix in (6.56). It is trivial to show that they are given by

$$\lambda = \frac{1}{2}\left(-2P - \bar{S} \pm \sqrt{(2P+\bar{S})^2 - 4P(P+\bar{S}) - 4Q(Q+\bar{N})}\right). \tag{6.57}$$

Taking $m = 0$, one of the eigenvalues becomes zero, corresponding to the fact that the phase is defined up to a constant, and the other eigenvalue is $-\bar{S}$.

Further assume $m \neq 0$ and thus $P, Q \neq 0$. Condition for stability $\operatorname{Re}\lambda < 0$ translates as

$$\operatorname{Re}(2P+\bar{S}) > \operatorname{Re}\sqrt{(2P+\bar{S})^2 - 4P(P+\bar{S}) - 4Q(Q+\bar{N})}. \tag{6.58}$$

In particular, as $m \to \infty$, $\operatorname{Re} P$ should be positive. Now for simplicity define

$$H = (2P+\bar{S})^2, \quad D = 4P(P+\bar{S}) + 4Q(Q+\bar{N}).$$

Using complex relation for any $c \in \mathbb{C}$

$$2\,(\operatorname{Re} c)^2 = \operatorname{Re}\left(c^2\right) + |c|^2 \tag{6.59}$$

and taking square of inequality (6.58), it becomes

$$|H| + \operatorname{Re} D > |H - D|, \tag{6.60}$$

in particular $\operatorname{Re} D > -|H|$. Taking square once more we get

$$(\operatorname{Im} D)^2 - 2\operatorname{Im} H \operatorname{Im} D - 2|H|(\cos\upsilon + 1)\operatorname{Re} D < 0, \tag{6.61}$$

where $\upsilon = \arg H$. By (6.59) $|H|(\cos\upsilon + 1) = 2(\operatorname{Re}\sqrt{H})^2$, and thus (6.61) means that

$$(\operatorname{Im} D)^2 - 2\operatorname{Im} H \operatorname{Im} D < 4\operatorname{Re} D(\operatorname{Re}\sqrt{H})^2. \tag{6.62}$$

Defining sequences h_m, d_m for $m \in \mathbb{Z} \setminus \{0\}$ as

$$h_m = f\Delta x^2 \cos(k\Delta x)m^2 \quad \text{and} \quad d_m = 2f\Delta x \sin(k\Delta x)m, \tag{6.63}$$

we can express P, Q, D and H as

$$\begin{aligned} P =& h_m \cos\beta + i d_m \sin\beta, \quad Q = -h_m \sin\beta + i d_m \cos\beta, \\ D =& 4(h_m^2 - d_m^2 + \bar{S}h_m\cos\beta - \bar{N}h_m\sin\beta) + 4i(\bar{S}d_m\sin\beta + \bar{N}d_m\cos\beta), \\ H =& (4h_m^2\cos^2\beta - 4d_m^2\sin^2\beta + \bar{S}^2 + 4h_m\bar{S}\cos\beta) + 4i(\bar{S}d_m\sin\beta + 2h_m d_m\sin\beta\cos\beta), \end{aligned}$$

which then being inserted in (6.62) results in

$$\begin{aligned} (\bar{S}d_m\sin\beta + \bar{N}d_m\cos\beta)\cdot(\bar{N}d_m\cos\beta - \bar{S}d_m\sin\beta - 4h_m d_m\sin\beta\cos\beta) < \\ < (\bar{S} + 2h_m\cos\beta)^2(h_m^2 - d_m^2 + \bar{S}h_m\cos\beta - \bar{N}h_m\sin\beta). \end{aligned} \tag{6.64}$$

Finally, defining $\bar{G} = 2e^{2r^*}G$ and using $\bar{S} = \bar{G}\cos\gamma$, $\bar{N} = \bar{G}\sin\gamma$, we get

$$\begin{aligned} d_m^2\bar{G}\sin(\gamma+\beta)(\bar{G}\sin(\gamma-\beta) - 4h_m\sin\beta\cos\beta) < \\ < (\bar{G}\cos\gamma + 2h_m\cos\beta)^2(h_m^2 - d_m^2 + \bar{G}h_m\cos(\gamma+\beta)). \end{aligned} \tag{6.65}$$

Recall that by (6.58) $\operatorname{Re} P > 0$, and substituting P and h_m from (6.63) we imply also that $\cos(k\Delta x)\cos\beta > 0$. Thus we just proved theorem:

Theorem 6.2. *Necessary and sufficient condition for stability of a constant equilibrium state is given by the inequality* (6.65) *for all* $m \in \mathbb{Z} \setminus \{0\}$ *together with the requirement* $\cos(k\Delta x)\cos\beta > 0$.

Due to the dependence of (6.65) on m it is difficult to check this condition explicitly. Therefore we will state several corollaries for particular cases, providing explicit inequalities to check.

Corollary 6.1. *Necessary and sufficient conditions for in-phase synchronization are given by*

$$\cos\beta > 0, \quad \cos(\gamma+\beta) > -\frac{f\Delta x^2}{2e^{2r^*}G}.$$

Proof. Indeed, in-phase equilibrium solution satisfies $k = 0$, thus by (6.63) $d_m = 0$ and $h_m = f\Delta x^2 m^2 > 0$. From the second condition of Theorem 6.2 we recover $\cos\beta > 0$. Finally, (6.65) with $d_m = 0$ requires right-hand terms to be greater than zero, which is just $h_m(h_m + \bar{G}\cos(\gamma+\beta)) > 0$. Since this is always true as $h_m \to \infty$ with $m \to \pm\infty$, it is enough to satisfy this inequality for $m = \pm 1$, leading to $\cos(\gamma+\beta) > -f\Delta x^2/\bar{G}$. □

Notice that conditions required in Corollary 6.1 as well as in all other corollaries below immediately ensure the existence of exponential representation of the amplitude of oscillations defined in (6.50).

Corollary 6.2. *Necessary and sufficient conditions for anti-phase synchronization are*

$$\cos\beta < 0, \quad \cos(\gamma+\beta) < \frac{f\Delta x^2}{2e^{2r^*}G}.$$

Proof. The proof follows the same steps as the previous one, switching the sign of h_m. □

Corollary 6.3. *Sufficient conditions for synchronization with* $\sin(k\Delta x) \neq 0$ *are given by*

$$\cos(k\Delta x)\cos\beta > 0 \tag{6.66}$$

together with

$$\begin{aligned} &\Upsilon < \frac{\Delta x^2}{4}\cot(k\Delta x)^2 + \frac{Ge^{2r^*}\cos(\gamma+\beta)}{2f}\frac{\cos(k\Delta x)}{\sin(k\Delta x)^2} - 1, \\ &\text{where} \quad \Upsilon = \begin{cases} 0, & \cos^2\beta \leqslant \cos^2\gamma, \\ \dfrac{\cos^2\beta}{\cos^2\gamma} - 1, & \cos^2\beta > \cos^2\gamma. \end{cases} \end{aligned} \tag{6.67}$$

Condition (6.66) *is also a necessary condition for stability.*

Proof. First, condition (6.66) repeats the second condition of Theorem 6.2. Further, since $\sin(k\Delta x) \neq 0$, d_m is non-zero. Divide (6.65) by d_m^2 and by $(\bar{G}\cos\gamma + 2h_m\cos\beta)^2$, obtaining

$$\frac{\bar{G}\sin(\gamma+\beta)(\bar{G}\sin(\gamma-\beta) - 4h_m\sin\beta\cos\beta)}{(\bar{G}\cos\gamma + 2h_m\cos\beta)^2} < \frac{h_m^2 - d_m^2 + \bar{G}h_m\cos(\gamma+\beta)}{d_m^2}. \tag{6.68}$$

Inserting the definitions of h_m and d_m we see that right-hand side of (6.68) is strictly increasing with m^2, therefore it can be simplified by setting $m^2 = 1$ as in the worst-case, thus obtaining the right-hand side of (6.67).

Now, to find sufficient conditions for satisfaction of (6.68), let us bound the left-hand side from above. For this we will use the following Lemma:

Lemma 6.1. *Function $f(x)$, defined as*

$$f(x) = \frac{V + \mu x}{(U+x)^2} \tag{6.69}$$

with $U > 0$ and $x > 0$ is bounded from above by

$$f(x) \leqslant \begin{cases} 0, & V \leqslant 0 \text{ and } \mu \leqslant 0, \\ V/U^2, & V > 0 \text{ and } U\mu \leqslant 2V, \\ \dfrac{\mu^2}{4\mu U - 4V}, & \mu > 0 \text{ and } U\mu > 2V. \end{cases} \tag{6.70}$$

The proof of this lemma can be found in Appendix A.7. We apply this lemma to the left-hand side of (6.68), with $U = \bar{G}\cos\gamma > 0$, $x = 2h_m\cos\beta > 0$, $V = \bar{G}^2\sin(\gamma+\beta)\sin(\gamma-\beta)$ and $\mu = -2\bar{G}\sin(\gamma+\beta)\sin\beta$, obtaining (6.67). Note that due to the trigonometric properties the conditions $U > 0$, $\mu > 0$ and $U\mu > 2V$ are contradicting by definitions of variables, thus only the first two cases of (6.70) are present in (6.67). Further, $V \leqslant 0$ and $U > 0$ implies $\mu \leqslant 0$, while $V > 0$ and $U > 0$ implies $U\mu < 2V$, thus it is sufficient to check only V in (6.70). □

Assume $\cos\beta > 0$ such that the in-phase solution is stable. Then the second condition of Theorem 6.2 requires that $\cos(k\Delta x) > 0$. Thus for the stability $k\Delta x$ should be smaller than $\pi/2$. This means $k < n/4$, where n is the number of oscillators. In particular, the phase difference between two neighbouring oscillators should be smaller than $\pi/2$. Also this means that to observe a state with $k = 1$ one needs at least 5 coupled oscillators, and to observe higher-order states one needs at least 9 oscillators. As an example, all possible states in the system with 10 oscillators are shown in Fig. 6.9.

6.4.4.3 Numerical simulation

To compare predictions from the previous section we performed a numerical simulation of system of $n = 50$ coupled spin-torque oscillators. Simulation parameters were chosen according to Dieudonné 2015, namely, we set $\omega = 6.55 \cdot 2\pi$, $N = -3.82 \cdot 2\pi$, $\Gamma_G = 0.375 \cdot 2\pi$ (all

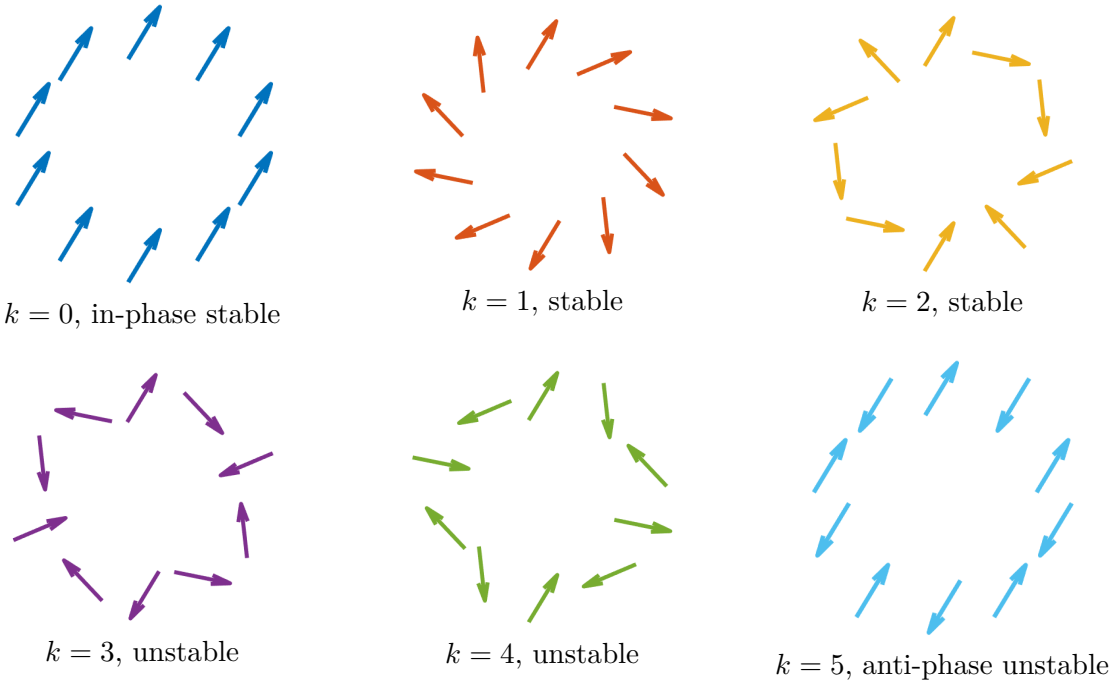


Figure 6.9: Six possible equilibrium solutions (6.48) for the ring of 10 spin-torque oscillators. Assuming $\cos\beta > 0$, the first three are stable and the second three are unstable.

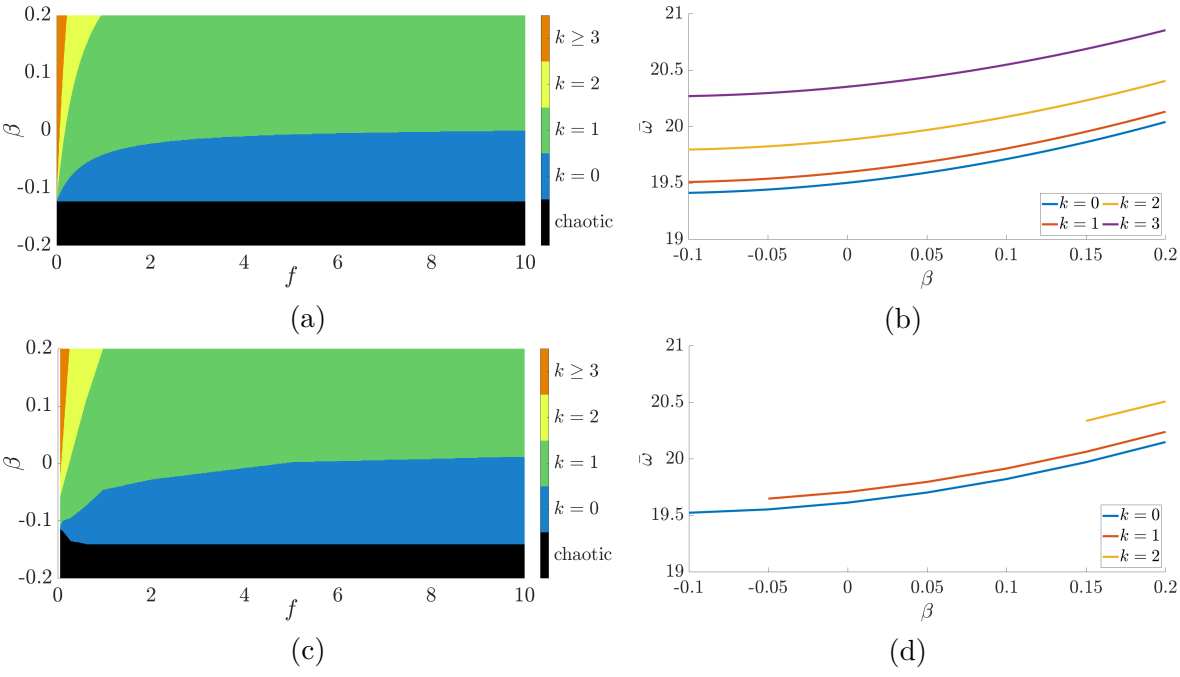


Figure 6.10: Synchronized solutions for system of $n = 50$ coupled identical spin-torque oscillators. **Top row**: analytic results for PDE (6.35). **(a):** diagram of possible regimes by Corollaries 6.1 and 6.3. Color code denotes the highest guaranteed existing regime, *chaotic* means that no stable solution exists. **(b):** Synchronization frequency $\bar{\omega}$ by (6.49) for different k depending on β for $f = 0.75$. **Bottom row:** numerical simulation of (6.33). **(c):** diagram of numerically estabilshed regimes. **(d):** experimentally measured synchronization frequency.

those measured in radians per nanosecond), $Q = -0.24$ and $\sigma = 5.48 \cdot 10^{-4} \cdot 2\pi$ for (6.30). In this case the critical current which is required to start oscillations is $I_c = \Gamma_G/\sigma = 684.3$. In our experiments we use a larger current $I = 1.5I_c$ to observe steady oscillations. With this setup the parameters of oscillator (6.31) are $\Gamma = 1.1781$ and $S = 2.9688$. Further, using definitions $G = |S + iN|$ and $\gamma = \arg(S + iN)$, we get $G = 24.1847$ and $\gamma = -82.95°$.

Due to large negative γ conditions in Corollaries 6.1 and 6.3 are not easy to satisfy. We can check which stable synchronized solutions are admitted by the coupled system depending on different coupling parameter F. Comparison between analytic predictions and numerical simulation results is shown in Fig. 6.10. We take different couplings $F = fe^{i\beta}$ with f changing from 0 to 10 and β changing from -0.2 to 0.2 radians. For each set of parameters we check the highest k for which conditions in Corollaries 6.1 and 6.3 are satisfied. These results are depicted in the diagram Fig. 6.10a. Further, we compare them with experimental results by simulating the original ODE system (6.33). We initialize all oscillators in this system using an amplitude $\sqrt{p_i} = \sqrt{\Gamma/S}$ and a phase $\phi_i = ik\Delta x$ for the i-th oscillator, such that the phase makes k turns along the ring. Finally a small Gaussian noise with a standard deviation of 0.05 is added to phases. The system is simulated for 5000 nanoseconds (corresponding roughly to 15000 periods of oscillation for $f = 0.75$). When simulation ends, we check if the system remained stable or it diverged from the corresponding equilibrium solution. The obtained highest possible stable regimes are depicted in the diagram Fig. 6.10c. Comparing it with the diagram Fig. 6.10a, we see that the analytic prediction almost perfectly reconstructs the experimental diagram, with deviations probably being attributed to the inaccuracies in the numerical stability check.

Finally we compare synchronization frequency $\bar{\omega}$ predicted by (6.49) with the one measured in simulation. To measure synchronization frequency in simulation we first notice that for every agent oscillating with constant amplitude its immediate frequency can be found as $\omega \approx \mathrm{Im}(\dot{c}/c)$. Then we average this frequency over all agents and over the last 1000 nanoseconds. The measured synchronization frequency for $f = 0.75$ and for $\beta \in [-0.1, 0.2]$ is depicted in Fig. 6.10d. It is clear that for the higher regimes for $k = 1$ and $k = 2$ stable solutions exist only for sufficiently high values of β. Comparing measured frequency with analytically predicted by (6.49) in Fig. 6.10b one can see that the trends and relative frequency differences between different regimes are reproduced correctly and that the measured frequency is about 0.1 rad/nanosec higher than the predicted one. This effect diminishes for higher values of f. This mismatch can have its origin in the fact that the analytic prediction was found for the PDE model (6.35), while the simulation was performed for the original ODE system (6.33).

6.4.5 Non-identical oscillators in small magnitude variation case

In the previous section we assumed that all oscillators are identical and that the solutions' magnitude is constant in space. In this section we relax a requirement on homogeneity but keep the assumption that $\partial r/\partial x \approx 0$. It was shown in Section 6.4.4 that under this assumption with piecewise constant parameters A and B the solution to the synchronization condition (6.42) is given by (6.44) and (6.45) with $\theta = \Delta x\ \partial\phi/\partial x$ and $J = B/A$, with a full solution

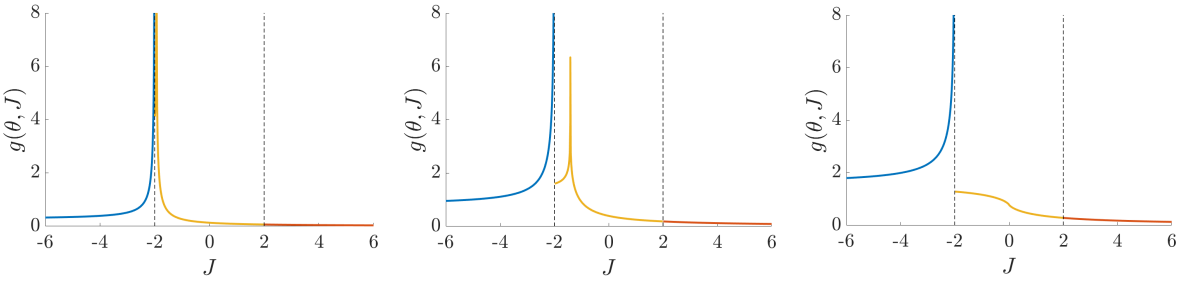


Figure 6.11: Function $g(\theta, J)$ defined in (6.71) with respect to J for different values of θ. Colors denote different branches in (6.71).

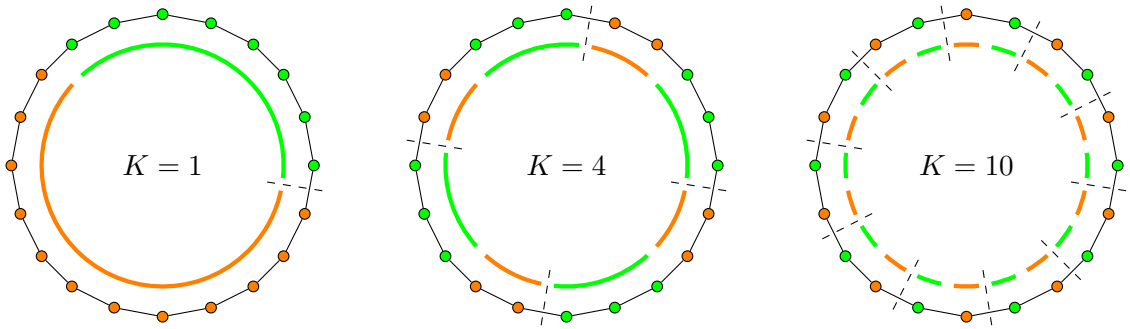


Figure 6.12: Examples of schematic representations of a ring with $n = 20$ oscillators with two different oscillator types placed periodically. K is a number of periods.

presented in Appendix A.6. In particular, let us define a function

$$g(\theta, J) = \begin{cases} \dfrac{J}{2\sqrt{4-J^2}} \ln \left| \dfrac{1 + \left(\frac{2-J}{\sqrt{4-J^2}} \tan \frac{\theta}{2} \right)}{1 - \left(\frac{2-J}{\sqrt{4-J^2}} \tan \frac{\theta}{2} \right)} \right| - \dfrac{1}{2}\theta, & |J| < 2, \\ \dfrac{J}{\sqrt{J^2-4}} \arctan \left(\dfrac{J-2}{\sqrt{J^2-4}} \tan \dfrac{\theta}{2} \right) - \dfrac{1}{2}\theta, & |J| > 2, \\ \dfrac{1}{2} \tan \dfrac{\theta}{2} - \dfrac{1}{2}\theta, & J = 2, \\ -\dfrac{1}{2} \cot \dfrac{\theta}{2} - \dfrac{1}{2}\theta, & J = -2. \end{cases} \tag{6.71}$$

It is interesting to note that there is a complex relation between arctangent and logarithm functions

$$\arctan s = -\frac{i}{2} \ln \left(\frac{1 + is}{1 - is} \right), \tag{6.72}$$

which means that the first two cases in (6.71) are essentially the same. In fact, the definition (6.71) defines a piecewise continuous function with at most two singularities with respect to J, see Fig. 6.11. Note further that $g(\theta, J)$ is an odd function with respect to θ.

With the help of this function we can define solutions to (6.42) as

$$A\frac{x}{\Delta x} + C = g(\theta, J), \tag{6.73}$$

where C is an integration constant. In particular (6.73) means that the constant solution (6.47) is captured by the singularity at $J = -2$.

Using the solution (6.73) it becomes possible to analyse systems with several different types of oscillators. Here for simplicity we will focus on the case of two types of oscillators. The first type of oscillators has a set of parameters ω_1, N_1, Γ_1, S_1 and F_1, and similarly the second type has a corresponding set of its own parameters. We further assume that oscillators' types are repeated K times along the ring, and that every continuous chunk of a particular oscillators' type consists of a fixed number of oscillators depending on its type (evidently this implies K is a divisor of the number of oscillators n). This means that the type of oscillators is a periodic function on the ring with period $2\pi/K$. For example if $K = 1$ this setup corresponds to one large set of oscillators of the first type followed by only one large set of oscillators of the second type, while if $K = n/2$ the types of oscillators alternate. We can define a set of switching points as y_j for $j \in \{0, ..., 2K-1\}$, with $y_0 = 0$ and $y_j = j/2 \cdot 2\pi/K$ for even j. Finally, for odd j we require $y_j - y_{j-1} = const$, thus the proportion of types is preserved. Oscillators placed in $[0, y_1) \cup [y_2, y_3) \cup ...$ are of the first type, and oscillators placed in $[y_1, y_2) \cup [y_3, y_4) \cup ...$ are of the second type. In particular this means that oscillators of the first type occupy proportion y_1/y_2 of the whole ring. Some possible examples of such distributions are schematically presented in Fig. 6.12.

Since oscillators are of different types, aggregated parameters A and B will have different values A_1, A_2, B_1 and B_2, leading to two different decision parameters J_1 and J_2. However an unknown synchronization frequency $\bar{\omega}$ should be common for both types, therefore by definition of B in (6.41) we can write $J_1 = \bar{J}_1 + \tau_1\bar{\omega}$ and $J_2 = \bar{J}_2 + \tau_2\bar{\omega}$, where

$$\bar{J}_1 = \frac{\cos\gamma_1\ \omega_1 + \sin\gamma_1\ \Gamma_1}{f_1 \sin(\gamma_1 + \beta_1)}, \qquad \tau_1 = -\frac{\cos\gamma_1}{f_1 \sin(\gamma_1 + \beta_1)}, \tag{6.74}$$

with $\bar{J}_2$ and τ_2 being defined in a similar way.

We are now interested in particular solutions $\theta(x)$ to (6.42). By (6.46) θ should be periodic. Since intervals of types of oscillators are equal, symmetry leads to the fact that θ should be periodic with period being equal to two intervals of different types of oscillators, namely $\theta(y_0) = \theta(y_2) = \theta(y_4) = ... = \theta(y_{2K-2})$. Further, one could expect to obtain continuous solutions, however performing numerical simulations of such systems we made an observation regarding possible synchronized solutions:

Observation 6.1. *Solution $\theta(x)$ behaves continuously and monotonically in the first type domain and is constant with discontinuity in the interior in the second type domain. Moreover, solution endpoints are symmetric about zero, namely $\theta(y_0) = -\theta(y_1)$.*

The set of all possible solutions is not covered only by those proposed by Observation 6.1, however each particular class of solutions heavily depends on properties of the function (6.71)

and thus requires special treatment. Further in this section we will stick to the class of solutions in agreement with Observation 6.1.

Defining $\theta^* = \theta(0)$ and assuming $\theta(y_1) = -\theta^*$ by Observation 6.1, we can compute (6.73) in points $x = y_0 = 0$ and $x = y_1$ for the first type of oscillators and subtract one from another, obtaining

$$A_1 \frac{y_1}{\Delta x} = 2g(\theta^*, J_1),$$

where we used the fact that the function $g(\theta, J)$ is odd with respect to θ. Substituting J_1 as in (6.74), we get a condition which should be satisfied for the first type of oscillators

$$2g(\theta^*, \bar{J}_1 + \tau_1\bar{\omega}) - A_1 \frac{y_1}{\Delta x} = 0, \tag{6.75}$$

which have two unknowns: θ^* and $\bar{\omega}$. The second condition comes from the assumption that for the second type domain the solution is constant and thus it is determined by (6.47). Using it for the second type domain we get

$$\theta^* = \arccos\left(-\frac{J_2}{2}\right). \tag{6.76}$$

Note that both θ^* and $-\theta^*$ are solutions to (6.47), which is consistent with Observation 6.1. Now, substituting J_2 by (6.74) in (6.76) and then substituting result in (6.75) we obtain an equation with a single unknown $\bar{\omega}$:

$$2g\left(\arccos\left(-\frac{\bar{J}_2 + \tau_2\bar{\omega}}{2}\right), \bar{J}_1 + \tau_1\bar{\omega}\right) - A_1 \frac{y_1}{\Delta x} = 0. \tag{6.77}$$

This equation can be solved for $\bar{\omega}$ using numerical methods such as Newton method for example. Once $\bar{\omega}$ is known, we can find J_1 and J_2 by (6.74) and then compute θ^* by (6.76). The full solution on the first domain can be then reconstructed by (6.73).

To determine the shape of solution $\theta(x)$ it remains only to find an exact position denoted by $y^* \in (y_1, y_2)$ where a discontinuous jump from θ^* to $-\theta^*$ happens in the second type domain. This position can be obtained if one recalls that $\theta = \Delta x\, \partial\phi/\partial x$ and thus integral of θ should have fixed value by (6.46) for some $k \in \mathbb{Z}$. In particular due to the periodic nature of the problem with K periods we have

$$\int_0^{y_2} \theta(x)dx = \frac{2\pi\Delta x}{K}k. \tag{6.78}$$

Since on the first type domain $\theta(x)$ is symmetric, its contribution to the integral is zero. Further, $\theta(x) = \theta^*$ on $x \in [y_1, y^*)$ and $\theta(x) = -\theta^*$ on $x \in (y^*, y_2]$, therefore (6.78) is just

$$(2y^* - y_1 - y_2)\theta^* = \frac{2\pi\Delta x}{K}k,$$

which leads to

$$y^* = \frac{\pi\Delta x}{K\theta^*}k + \frac{y_1 + y_2}{2}. \tag{6.79}$$

Thus the solution's shape $\theta(x)$ is fully reconstructed.

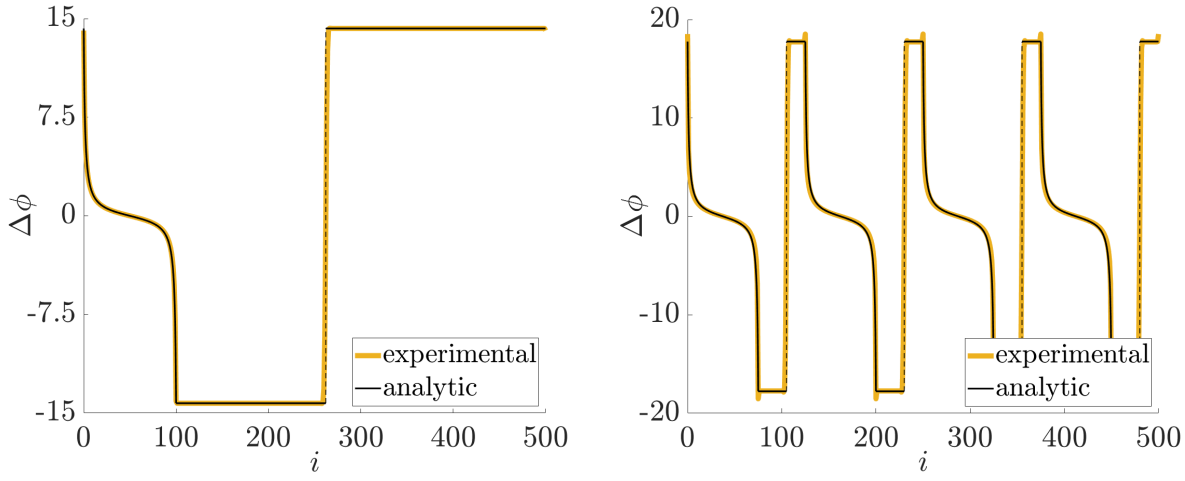


Figure 6.13: Comparison of numerical and analytical synchronized solutions of systems with $n = 500$ oscillators separated into two classes. Horizontal axis: index of oscillator. Vertical axis: phase difference between two consecutive oscillators in degrees. Yellow line denotes solution obtained by numerical simulation of (6.33), black line denotes analytic solution by (6.76)-(6.79). Parameters: **(a):** $K = 1$, $y_1/y_2 = 0.2$, $\Gamma_2 = 1.05 \cdot \Gamma_1$, $k = 3$. **(b):** $K = 4$, $y_1/y_2 = 0.6$, $N_2 = 1.03 \cdot N_1$, $k = -2$.

Remark 6.2. Observation 6.1 assumes the first part of the solution behaves continuously and the second part is piecewise constant. In real system these parts can be interchanged, which depends on the obtained values of J_1 and J_2: for the continuous part $|J| > 2$, while for the piecewise constant part $|J| < 2$ (while they are both usually negative and close to -2).

Remark 6.3. Other types of solutions except those presented in Observation 6.1 are also possible. In this case there is no piecewise constant domain and all solution's parts behave according to (6.73). It is then possible to formulate a system of nonlinear equations with several unknown variables which should be solved numerically. However we found that solutions to this system lie very close to singularities of $g(\theta, J)$, thus they cannot be found reliably by numerical methods without additional problem reformulation.

6.4.5.1 Numerical simulation

To demonstrate how solutions to the synchronization condition (6.42) found by (6.76)-(6.79) approximate synchronized solutions of the original system (6.33) we performed numerical simulations of (6.33) with $n = 500$ oscillators being split into two types as it was described earlier in this section. Parameters of the first type of oscillators were taken the same as in Section 6.4.4.3, and for the second type slight deviations in parameters were added. Oscillators were placed periodically on the ring with K periods, thus there were $2K$ groups of oscillators as it was shown in Fig. 6.12. Each group of oscillators of the first type occupies y_1/y_2 proportion of the period of the length y_2, and each group of oscillators of the second type

occupies $(y_2 - y_1)/y_2$ proportion. Numerical simulation was initialized in the same way as in Section 6.4.4.3 with k denoting initial shift in phases of consecutive oscillators such that the phase makes k turns along the ring.

We performed two simulations:

1. In the first simulation we altered damping parameter Γ for the second type of oscillators such that $\Gamma_2 = \Gamma_1 \cdot 1.05$. We used only two groups of oscillators, one of each type, thus $K = 1$. The first type occupies only 20% of the whole ring, thus $y_1/y_2 = 0.2$. Finally, oscillators were initialized such that the phase makes $k = 3$ turns along the ring.

2. In the second simulation we changed frequency gain parameter N for the second type of oscillators such that $N_2 = N_1 \cdot 1.03$. We used eight groups of oscillators, four of each type, thus $K = 4$. The first type occupies 60% of every period, thus $y_1/y_2 = 0.6$. In this simulation oscillators were initialized such that the phase makes $k = -2$ turns along the ring, rotating in opposite direction.

Results of the simulation are presented in Fig. 6.13. Simulation was performed for 2000 nanoseconds and then phase differences between consecutive oscillators were computed. The result was then compared with analytic predictions by (6.76)-(6.79). It is clear that the shape of solutions is reconstructed almost perfectly even though our analysis was based on the continualized PDE model of the network and a small magnitude variation assumption.

6.4.6 General large magnitude variation case

Analysis in Sections 6.4.4 and 6.4.5 was based on the assumption that the amplitude of oscillations is almost identical along the ring, i.e. $\partial r/\partial x \approx 0$. It appears that this assumption can be removed and that it is possible to equivalently transform an original synchronization condition (6.40) to the differential equation similar to (6.42) by using properties of trigonometric functions and moving the problem to the complex domain. To perform this transformation we pose another assumption that the parameter A is constant along the ring. With this assumption we can move A under the spatial derivative in (6.40) and obtain the following condition

$$\begin{aligned}\Delta x\frac{\partial}{\partial x}\left[A\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right)\sinh\left(\Delta x\frac{\partial r}{\partial x}\right)+\sin\left(\Delta x\frac{\partial\phi}{\partial x}\right)\cosh\left(\Delta x\frac{\partial r}{\partial x}\right)\right]+\\ +2\left[A\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right)\cosh\left(\Delta x\frac{\partial r}{\partial x}\right)+\sin\left(\Delta x\frac{\partial\phi}{\partial x}\right)\sinh\left(\Delta x\frac{\partial r}{\partial x}\right)\right]+B=0,\end{aligned} \tag{6.80}$$

Now let us define (possibly) complex variables ψ and L such that

$$A\cos\left(\Delta x\frac{\partial\phi}{\partial x}\right)=L\cos(i\psi),\quad i\sin\left(\Delta x\frac{\partial\phi}{\partial x}\right)=L\sin(i\psi). \tag{6.81}$$

In particular $L = \sqrt{A^2 - 1}$, which is purely real if $|A| \geqslant 1$ and purely imaginary if $|A| < 1$. Further, (6.81) results in ψ being defined in such way that

$$\tan(i\psi) = \frac{i}{A} \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right), \tag{6.82}$$

which in particular means that $\tan(i\psi)$ is purely imaginary. There can also be two cases:

1. If $|\tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)/A| < 1$, then (6.82) leads to

$$\psi = \frac{1}{i} \arctan\left[\frac{i}{A} \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)\right] = -\frac{1}{2} \ln\left(\frac{A - \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)}{A + \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)}\right) \in \mathrm{Re},$$

   since the argument of the logarithm is positive. Here a complex relation (6.72) between arctangent and logarithm was used.

2. If $|\tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)/A| > 1$, then

$$\psi = -\frac{1}{2} \ln\left(\frac{A - \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)}{A + \tan\left(\Delta x \frac{\partial \phi}{\partial x}\right)}\right) = -\frac{1}{2} \ln\left(\frac{\tan\left(\Delta x \frac{\partial \phi}{\partial x}\right) - A}{\tan\left(\Delta x \frac{\partial \phi}{\partial x}\right) + A}\right) \pm \frac{\pi}{2} i,$$

   since $\ln(-s) = \ln(s) \pm i\pi$.

Now let us simplify (6.80) using L and ψ in (6.81). First, with the help of identities $\cosh(x) = \cos(ix)$ and $\sinh(x) = -i \sin(ix)$ we can write the system (6.80) as

$$\Delta x \frac{\partial}{\partial x}\left[- iA \cos\left(\Delta x \frac{\partial \phi}{\partial x}\right) \sin\left(i\Delta x \frac{\partial r}{\partial x}\right) + \sin\left(\Delta x \frac{\partial \phi}{\partial x}\right) \cos\left(i\Delta x \frac{\partial r}{\partial x}\right)\right] +$$
$$+ 2\left[A \cos\left(\Delta x \frac{\partial \phi}{\partial x}\right) \cos\left(i\Delta x \frac{\partial r}{\partial x}\right) - i \sin\left(\Delta x \frac{\partial \phi}{\partial x}\right) \sin\left(i\Delta x \frac{\partial r}{\partial x}\right)\right] + B = 0,$$

which by substitution of (6.81) becomes

$$- iL\Delta x \frac{\partial}{\partial x} \sin\left(i\psi + i\Delta x \frac{\partial r}{\partial x}\right) + 2L \cos\left(i\psi + i\Delta x \frac{\partial r}{\partial x}\right) + B = 0. \tag{6.83}$$

Denoting $\theta := \psi + \Delta x \frac{\partial r}{\partial x}$, $\tilde{A} = i$ and $\tilde{B} = iB/L$, we arrive at the same type of equation as (6.42), although formulated now in the complex domain and with coefficients $\tilde{A}$ and $\tilde{B}$.

It is now possible to prove our original statement in Section 6.4.4 that $\partial r/\partial x = 0$ for any synchronized solution in the system with identical oscillators. Indeed, in case of identical oscillators the only solution to the equation (6.83) is given by $\cos(i\theta) = -B/2L$. It is clear that for the constant θ we should have $\frac{\partial r}{\partial x} = 0$, otherwise r could not be periodic along the ring. Thus, substituting $\theta = i\psi$, the constant solution is $\cos(i\psi) = -B/2L$ or by definitions of ψ and L in (6.81) it is just $\cos(\Delta x\, \partial\phi/\partial x) = -B/2A$. This solution exactly coincides with (6.47) which was obtained using small magnitude variation assumption in Section 6.4.4.

Moving to the case of non-identical oscillators, we can use the same methods as in Section 6.4.5 to find solutions to (6.83). In particular, for $J \neq \pm 2$ we can write the general solution (6.73) for (6.83) as

$$i\frac{x}{\Delta x} + C = \frac{J}{\sqrt{J^2-4}} \arctan\left(\frac{J-2}{\sqrt{J^2-4}} \tan\frac{i\theta}{2}\right) - \frac{i\theta}{2}, \tag{6.84}$$

where $J = B/L$. The first two cases in (6.71) become the same since the problem is now formulated in the complex plane and arctangent and logarithm functions are related by (6.72). Finally, using relations $\tan(i\theta) = i\tanh(\theta)$ and $\arctan(is) = i\,\mathrm{artanh}(s)$ formula (6.84) is simply

$$\frac{x}{\Delta x} + C = \frac{J}{\sqrt{J^2-4}} \mathrm{artanh}\left(\frac{J-2}{\sqrt{J^2-4}} \tanh\frac{\theta}{2}\right) - \frac{\theta}{2}. \tag{6.85}$$

Solutions to (6.85) can be checked numerically for several types of oscillators in a similar way it was done in Section 6.4.5. It is interesting to note that (6.85) depends only on the parameter J compared to the general solution (6.73) which depends both on J and A. This happens because unknown variable θ in (6.85) is scaled by (6.81) thus parameter A is already integrated inside. Finally, once (6.85) is solved and $\theta(x)$ is recovered, one can reconstruct $r(x)$ from the solution by using (6.39).

6.4.7 Open problems

Analysis of synchronization of spin-torque oscillators has a big practical importance since synchronous oscillations produce much more energy, therefore it is very important to realize when synchronized solutions do exist and what deviations in manufacturing (which result in deviations in parameters) they do tolerate. In previous sections we showed how the continuation method can help in the analysis of this problem and then we derived some results which could be useful in practical applications. Still, there are many questions that could be investigated in details regarding the system (6.33), its PDE approximation (6.35) and the synchronization condition (6.38).

- Case of identical oscillators was fully covered in Section 6.4.4 where synchronization condition was analysed to find equilibrium points and their stability conditions. Still Corollary 6.3 gives only sufficient conditions on stability and probably more rigorous statements could be made based on Theorem 6.2.
- Practically more important case of non-identical oscillators was discussed in Sections 6.4.5 and 6.4.6 but the results presented there cover only the question of a search for equilibrium solutions. Due to Observation 6.1 only specific class of equilibrium solutions was reconstructed, however it is not clear whether the obtained solution really persists in the system, e.g. whether it is stable. Also, numerical simulations have shown that synchronized solutions are very fragile in a sense that small deviations in parameters result in very large differences in phases between consecutive oscillators, although system still remains stable.

- In the original formulation of the system (6.33) we assumed that every oscillator is coupled only with its two neighbours, however in practical applications coupling between oscillators can occur mostly due to physical effects depending on distance, thus in general every oscillator is coupled to all others with distance-diminishing coupling coefficient. As it was shown in previous chapter in Section 5.3 systems with summation of distance-dependent forces are suitable for the continuation method, thus it would be possible to derive synchronization conditions for them as well.
- Instead of analysing inhomogeneous oscillators with known parameters one could assume stochastic parameter deviations and thus construct a Fokker-Plank-type PDE which generalizes (6.35), which could be then analysed for a search of synchronization conditions in probabilistic sense.
- In (6.33) we assumed a ring topology of oscillators. Analysis for more general topologies, especially 2-dimensional, would be of a great importance for practical applications.

6.5 Concluding remarks

In this chapter we demonstrated how the continuation method can be utilized to transform oscillatory networks into nonlinear PDE models which open new possibilities for analysis and control of synchronization phenomena.

First, a laser network was synchronized by suppressing undesirable oscillations due to the fact that the PDE model of the laser system was suitable for a PDE-based backstepping. We demonstrated by numerical simulations that application of a PDE-based control to the initially discrete system indeed provides stability, while derivation of such continuous control is simple and explicit. Question of derivation of synchronization conditions was then covered for the particular case of Kuramoto oscillators and then for a general case of non-isochronous oscillators. It appears that nonlinear PDEs appearing in this case can be analysed to recover equilibrium solutions and to check their stability. Validation by numerical simulation demonstrated that synchronized solutions obtained in this way coincide with the ones to which the real system converges.

It is interesting to note that while it is possible to duplicate the derivation of synchronization threshold for Kuramoto oscillators in the ODE-based setup (although PDE approach can still sometimes be more scalable), the true power of the continuation method becomes visible in the general problem of reconstructing synchronization conditions for non-isochronous oscillators. We have shown that in the case of two different types of oscillators synchronized solutions are given by (6.73), where function $g(\theta, J)$ in (6.71) is a solution to the differential equation (6.42). It is clear that it would not be possible to obtain an analytic formula for an analogue of $g(\theta, J)$ in the ODE-based setup since (6.42) would transform in a large system of nontrivial trigonometric difference equations. Therefore, application of the continuation method results in continuous conditions which are simpler to tackle analytically compared to discrete ones in the same way as ODEs are easier to solve than difference equations.

Conclusion and perspectives

Control and analysis of large-scale systems is a complex problem and there are many different ways to approach it. In this thesis, we examined in detail the various possibilities of analyzing systems through an aggregated simplified representation and the application of this approach to systems control. Below we will first summarize the main contributions of this work, and then propose some possible future directions of research.

Contributions

Large-scale network control

Control of large ODE networks differs from traditional control tasks in that in networks it is often not necessary to independently control the state of each node. In some real-world problems, for normal functioning of the network it is sufficient that the states of the nodes would be close enough to the average state, and this average state, in turn, would be maintained at a given desired value. This is exactly the kind of setup we considered in Chapter 2. Assuming that we can only measure the average state of the network and that the control goal is to stabilize the average state to a given value, we showed that it is possible to do this with an integral controller. We then showed that any integral controller with positive coefficients will work for a positive system if the system matrices satisfy the condition $CA^2 > 0$. This result was proved in Theorem 2.3. Due to the special structure, it can be shown that this result trivially holds for all systems with Laplacian dynamics. In terms of passivity theory, Theorem 2.4 showed that this condition is equivalent to the system transfer function being strictly positive real (SPR). The result was generalized to the more general case of controlling multiple outputs such as the average states of different clusters of the network.

To ensure that the network states are indeed close to the desired mean values we additionally proposed an algorithm that can minimize the standard deviation of the system states. This algorithm is based on the extremum seeking method, which is used to minimize steady-state input-output response. In order to perform standard deviation minimization in our case we developed our own modified version of the algorithm, namely a constrained extremum seeking, which minimizes a given function subject to a constraint that the average network state is held at the desired value.

Shape-based model reduction for PDEs

Usually model reduction methods are based on simplifying the original system in order to obtain a system that is in some sense ”close” to the original system but with a smaller

dimensionality of the state space. In Chapter 3 of this thesis we additionally assumed that the solution profile of the system has a certain shape, which can be approximately described by a small number of parameters. In this way we were able to perform a model reduction that transforms the dynamics of the original system into an evolution of the parameters of the shape of the solution. Using numerical simulations, we have shown that the reduced model is able to describe the behavior of the system very well. Even when the solution profile of the original system violates the chosen shape, the reduced system "averages" effects that the shape cannot describe, as long as solution of the reduced system does not become degenerate. The method in Chapter 3 was developed for 1D conservation laws, but the idea of reducing a system to the dynamics of parameters of some shape can be generalized to other classes of models of large systems.

Continuation method and its applications

Many large ODE systems have an underlying physical and spatial structure, that is, the nodes of the network have coordinates in space, and interactions in the network depend on the relative position of the nodes. For such systems we have developed a continuation method that allows us to turn ODE models of large spatial systems into PDE models. The method was described in Chapter 4 of this thesis. The key idea behind the method is to replace finite differences with corresponding partial derivatives through the Taylor series. At the same time Theorem 4.2 showed that for linear spatially invariant systems taking a sufficiently large PDE order one can approximate the spectrum of the original ODE system arbitrarily closely, and Theorem 4.3 gave an estimate for the deviation of solutions between ODE and PDE representations depending on the PDE order. Then the method was generalized to nonlinear systems using a computational graph formalization. Many additional extensions have been described such as multidimensional spaces, space-dependent systems or systems with boundaries. Thus essentially any ODE system that has a spatial structure can be transformed into a PDE using the continuation method.

A special type of ODE systems which can be turned into PDEs are systems in which the underlying space is index space, and the state of each node includes a spatial position. In other words, these are systems in which some indexed moving agents interact with their neighbors with adjacent indices. By performing a continuation in index space we can write such systems as PDEs, which determine the motion of virtual agents with non-integer indices. The reciprocal of the position derivative with respect to the index can be used to define the concept of density, which in turn transforms the resulting PDE into an equation for the evolution of the density of agents in space. We have thus derived a procedure for transforming the ODE dynamics of individual agents into a PDE for the density of agents. Chapter 5 of this manuscript is devoted to a discussion of this procedure and its application to various specific problems. More specifically, in Section 5.2 we showed how PDE models for car traffic density (e.g., the LWR model for car density on highways) can be derived from individual driver models. Section 5.3 generalized the method to probably the most famous problem of connecting particle and continuous worlds, the Hilbert's 6th problem, which is devoted to the derivation of the Euler equations from the dynamics of individual particles with dynamics

based on Newton's laws.

In addition to deriving new models the continuation method can be used for control design. To do this one must perform a continuation of the original discrete system, find a control law for the system in its continuous representation and then discretize the resulting control back to apply it to the original system. The application of such a technique was demonstrated in Section 5.4 for a robot formation transformed into a continuous representation similarly to the Euler equations and in Section 6.2, where a chain of lasers was synchronized using PDE-based backstepping.

Finally, we applied the continuation method to large networks of nonlinear oscillators in Chapter 6. One of the most important problems for such systems is the question of synchronization, the conditions for its existence and stability. In Chapter 6, we concentrated on deriving synchronization conditions for Kuramoto oscillator networks and for non-isochronous oscillator networks such as spin-torque oscillators. Working with PDE models of these networks we were able to take advantage of the analysis of continuous systems and therefore obtained conditions that could not be derived from the original ODE network. We have shown with the help of numerical simulations that the obtained conditions can indeed capture the behavior of the original systems and thus they can be used in real practical applications.

Perspective and extensions

Although the methods described in this research work have provided good initial results, there are still open problems and questions that can provide significant improvements to our understanding of the methods and to their practical applicability, which can be a base for a future research.

One clear path is an in-depth development of a method for controlling the average state of a network and the standard deviation of states in that network. In Chapter 2, our control goal was solely to drive the average network state to a fixed desired steady-state value, whereas in the real world the task of tracking the value as it changes over time is much more important. At the same time, removing the steady-state assumption it is possible to improve the minimization of the standard deviation in transient processes. Finally, we assumed that both average state and standard deviation could be measured directly. The method of controlling the average would be much more widely applicable if, for example, we could locally measure only some specific states of the network (for example, boundary states).

The shape-based model reduction method from Chapter 3 is currently limited in application only to the class of 1D PDE conservation laws and only for limited periods of time until the selected shape becomes degenerate. The degeneracy of the shape is a very serious limitation that could be removed if a reparameterization procedure were developed that automatically corrects the shape each time it becomes degenerate. It would be convenient to enclose such a system in a specially created software that is applicable to any system and automatically controls the parameterizations of the reduced system. In the future it would also

be possible to investigate the extension of the method to other classes of systems, including various PDE models and ODE networks with a spatial structure.

The most straightforward continuation of the work which was described in this PhD thesis is a more detailed study of the continuation method, as well as a more detailed development of a general theory of its application to various systems for analysis and control. First, the analytic results in Chapter 4 guarantee the convergence of PDE system solutions to ODE solutions as long as the PDE order tends to infinity. In reality, due to the lack of PDE analysis methods for high orders, as well as the risk of artificial instabilities, it makes sense to limit the PDE derivation to no more than the first and second orders. Thus, it would be highly desirable to develop criteria for the applicability of the method to low-order approximations. Second, the analytic results were derived for linear spatially invariant systems. In reality, however, the method is mostly applied to nonlinear space-dependent systems, so it is worth investigating convergence guarantees for such systems. Third, Chapters 5 and 6 showed the potential of applying the continuation method to control design using a continuous representation of the system. Such a procedure requires not only the application of the continuation method to derive a PDE model of the system, but also the discretization of the obtained control law in order to be able to implement it in the actual system. That is, the efficiency of the control law depends on the accuracy of the continuation-discretization two-way process. In the future, it would be desirable to find what conditions the continuation must satisfy for the control to be able to perform the task, and how these conditions might be related to Theorem 4.1 about reversibility of the continuation procedure as well as to the other theorems proved in Chapter 4.

Appendix A

Technical proofs

Contents

A.1 Proof of the Lemma 2.1

Lemma 2.1. *Suppose we have a matrix $\mathcal{M} = M + ibI$, which is a complex matrix with real part M and imaginary part bI, with $b \in \mathbb{R}$ and I an identity matrix. Assume M being invertible and having no eigenvalues on the imaginary axis. Denote $\mathcal{L} = \mathcal{M}^{-1} = L + i\bar{L}$. Then the real part of $\mathcal{L}$ is given by*

$$\operatorname{Re}\mathcal{L} = L = (M + b^2 M^{-1})^{-1}. \tag{A.1}$$

Proof. By the definition of inverse $\mathcal{M}\mathcal{L} = (M + ibI)(L + i\bar{L}) = I$, which decomposes into real and imaginary parts:

$$\begin{aligned} ML - b\bar{L} &= I, \\ M\bar{L} + bL &= 0. \end{aligned} \tag{A.2}$$

From the second equation $\bar{L} = -bM^{-1}L$, and substitution of $\bar{L}$ into the first equation gives

$$ML + b^2 M^{-1} L = I, \tag{A.3}$$

which means $L = \left(M + b^2 M^{-1}\right)^{-1}$. □

A.2 Proof of the Lemma 2.2

Lemma 2.2. *Let M be an M-matrix. Let C be a row-vector such that $CM^2 > 0$. Then*

$$C(M + tM^{-1})^{-1} > 0 \tag{A.4}$$

for any $t \geqslant 0$.

Proof. Denote $L(t) = (M + tM^{-1})^{-1}$. We need to prove that $CL(t) > 0$ for all $t \geqslant 0$. The idea of the proof is to provide series expansion for $CL(t)$ such that each term in the expansion is positive. First the coefficients of Taylor series will be computed and then summation by parts will be used twice to obtain series with positive terms. The proof of the Lemma is separated into subsections A.2.1-A.2.6.

A.2.1 Series expansion

Matrix M is an M-matrix, which by definition means that there exists some matrix P with $P_{i,j} \geqslant 0$, $\rho(P) < 1$ and scalar $s > 0$ such that $M = s(I - P)$. Now make the following transformations:

$$\begin{aligned} L(t) = (M + tM^{-1})^{-1} = M(M^2 + tI)^{-1} = s(I - P)\left(s^2(I - P)^2 + tI\right)^{-1} \\ = s(I - P)\left((s^2 + t)I - 2s^2P + s^2P^2\right)^{-1} = \frac{s}{s^2 + t}(I - P)\left(I - \frac{2s^2}{s^2 + t}P + \frac{s^2}{s^2 + t}P^2\right)^{-1}. \end{aligned} \tag{A.5}$$

Multiplier $\frac{s}{s^2+t}$ is always positive, thus it doesn't affect the sign of the result, so in future we will omit it. Now denote $\alpha = \frac{s^2}{s^2+t}$. By definition of t and s this variable satisfies $0 < \alpha \leqslant 1$. Case $\alpha = 1$ is trivial (it corresponds to the case $t = 0$), thus often in the following we will use $0 < \alpha < 1$. Then

$$L = (I - P)\left(I - 2\alpha P + \alpha P^2\right)^{-1}. \tag{A.6}$$

We aim to find a coefficients in formal series expansion of L in the powers of P:

$$L = \sum_{k=0}^{+\infty} L_k P^k. \tag{A.7}$$

A.2.2 Coefficients of the series expansion

We can introduce a scalar function $F(x)$ which has the same expansion as (A.7) and for which a recursive computation of series coefficients is possible. Concretely, define

$$f(x) = \frac{1 - x}{1 - 2\alpha x + \alpha x^2}, \tag{A.8}$$

where $x \in [0, 1)$. Writing the same expansion as L:

$$f(x) = (1 - x) \sum_{k=0}^{+\infty} (2\alpha x - \alpha x^2)^k. \tag{A.9}$$

At the same time, $f(x)$ can be expanded as Taylor series centered at 0:

$$f(x) = \sum_{k=0}^{+\infty} \frac{f^{(k)}(0)}{k!} x^k. \tag{A.10}$$

Power series expansion is unique, thus coefficients $L_k = \frac{f^{(k)}(0)}{k!}$.

The next step is to determine derivatives of $f(x)$ evaluated at $x = 0$. Let us introduce function

$$g(x) = 1 - 2\alpha x + \alpha x^2. \tag{A.11}$$

It is obvious that $f(x)g(x) = 1 - x$. Now take n-th derivative of this multiplication:

$$\frac{d^n}{dx^n}(1 - x) = \frac{d^n}{dx^n}(f(x)g(x)) = \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)}(x) g^{(k)}(x). \tag{A.12}$$

Function $g(x)$ is a polynomial of the degree 2, thus its derivatives can be explicitly written:

$$g^{(0)}(0) = 1, \quad g^{(1)}(0) = -2\alpha, \quad g^{(2)}(0) = 2\alpha, \tag{A.13}$$

and all higher derivatives are zero. Moreover, $(1 - x)^{(0)}(0) = 1$ and $(1 - x)^{(1)}(0) = -1$ with all higher derivatives also zero. Recall that $L_n = \frac{f^{(n)}(0)}{n!}$. Using (A.12) we have the following recurrent relation for L_n:

$$L_n - 2\alpha L_{n-1} + \alpha L_{n-2} = 0, \quad \forall n \geqslant 2, \tag{A.14}$$

with initial conditions $L_0 = 1$ and $L_1 = 2\alpha - 1$.

A.2.3 Solving the linear recurrent equation

Equation (A.14) is a linear recurrent equation, which solution is found by solving the characteristic polynomial

$$\lambda^2 - 2\alpha\lambda + \alpha = 0. \tag{A.15}$$

For $0 < \alpha < 1$ roots are complex conjugate pair (λ, λ^*) with

$$\lambda = \alpha + i\sqrt{\alpha(1 - \alpha)}, \quad |\lambda| = \sqrt{\alpha}. \tag{A.16}$$

The general solution to the equation (A.14) is given by $L_n = \operatorname{Re}\left[z\lambda^n\right]$, where z is a complex value that should be determined from the initial conditions. From $L_0 = 1$ we simply recover $\operatorname{Re} z = 1$, and from $L_1 = 2\alpha - 1$ it is found that $\operatorname{Im} z = \sqrt{\frac{1-\alpha}{\alpha}}$. Thus the solution to the equation (A.14) is given by

$$L_n = \operatorname{Re}\left[\left(1 + i\sqrt{\frac{1 - \alpha}{\alpha}}\right)\left(\alpha + i\sqrt{\alpha(1 - \alpha)}\right)^n\right]. \tag{A.17}$$

A.2.4 Back to the matrix equation

It is established that matrix L can be expressed by the series

$$L = \sum_{k=0}^{+\infty} L_k P^k, \tag{A.18}$$

where L_k are given by (A.17). Now it is evident that (A.18) is a convergent series due to $\rho(P) < 1$ and $|\lambda| = \sqrt{\alpha} < 1$.

Coefficients L_k can be both positive and negative, thus in general matrix L should not be positive. But we want to prove positivity of the vector CL:

$$CL = \sum_{k=0}^{+\infty} L_k C P^k > 0. \tag{A.19}$$

A.2.5 Properties of $\{CP^k\}$ sequence

Now it is time to use the condition $CM^2 > 0$. First of all, M is an M-matrix, thus for any vector x inequality $xM > 0$ implies $x > 0$. Therefore $CM > 0$ (and actually $C > 0$ automatically).

From $CM > 0$ we obtain $C(I - P) > 0$, which means $C > CP$. Moreover, matrix P is positive, thus multiplying both sides of this inequality on P preserves it. Thus the order relation holds:

$$C > CP > CP^2 > CP^3 > ... > 0. \tag{A.20}$$

Therefore sequence $\{CP^k\}$ is monotonically decreasing with a limit zero (because $\rho(P) < 1$).

Now let us use the next condition, $CM^2 > 0$. Essentially it means $C(I - P)(I - P) > 0$, or $(C - CP) > (CP - CP^2)$. Again, multiplication by P preserves order, so we have

$$C - CP > CP - CP^2 > CP^2 - CP^3 > ... > 0, \tag{A.21}$$

or

$$CP^k - 2CP^{k+1} + CP^{k+2} > 0. \tag{A.22}$$

This implies that sequence $\{CP^k - CP^{k+1}\}$ is also monotonically decreasing to zero. In some sense this is equivalent to the "convexity" of $\{CP^k\}$ sequence.

A.2.6 Summations by parts

For any series

$$\sum_{k=0}^{N} x_k y_k = x_N Y_N - \sum_{k=0}^{N-1} (x_{k+1} - x_k) Y_k, \tag{A.23}$$

where $Y_n = \sum\limits_{k=0}^{n} y_k$. This transformation is called Abel transformation or summation by parts. We will apply this procedure twice to obtain series with each term positive.

Denote $H_n = \sum\limits_{k=0}^{n} L_k$. Then H_n is bounded because L_k consists of powers of λ with $|\lambda| = \sqrt{\alpha} < 1$. By $\rho(P) < 1$ follows $\lim_{k\to+\infty} CP^k = 0$. Thus $\lim_{k\to+\infty} H_k CP^k = 0$ and we can write

$$CL = \sum_{k=0}^{+\infty} L_k CP^k = -\sum_{k=0}^{+\infty} H_k (CP^{k+1} - CP^k) = \sum_{k=0}^{+\infty} H_k (CP^k - CP^{k+1}). \tag{A.24}$$

Applying Abel transformation for the second time with $G_n = \sum\limits_{k=0}^{n} H_k$, we get

$$CL = \sum_{k=0}^{+\infty} G_k (CP^k - 2CP^{k+1} + CP^{k+2}). \tag{A.25}$$

Let us calculate H_n:

$$H_n = \sum_{k=0}^{n} L_k = \operatorname{Re}\left[z \sum_{k=0}^{n} \lambda^k\right] = \operatorname{Re}\left[z \frac{1-\lambda^{n+1}}{1-\lambda}\right], \tag{A.26}$$

where $\lambda = \alpha + i\sqrt{\alpha(1-\alpha)}$ and $z = 1 + i\sqrt{\frac{1-\alpha}{\alpha}}$. Multiply nominator and denominator by $(1-\lambda^*)$:

$$H_n = \frac{1}{1-\alpha} \operatorname{Re}\left[z(1-\lambda^*)(1-\lambda^{n+1})\right]. \tag{A.27}$$

Product $z(1-\lambda^*) = i\sqrt{\frac{1-\alpha}{\alpha}}$, which is purely imaginary, so

$$H_n = -\operatorname{Re}\left[i \frac{\lambda}{\sqrt{\alpha(1-\alpha)}} \lambda^n\right]. \tag{A.28}$$

Denote $w = -i\frac{\lambda}{\sqrt{\alpha(1-\alpha)}}$ and calculate G_n:

$$G_n = \sum_{k=0}^{n} H_k = \operatorname{Re}\left[w \sum_{k=0}^{n} \lambda^k\right] = \operatorname{Re}\left[w \frac{1-\lambda^{n+1}}{1-\lambda}\right]. \tag{A.29}$$

Multiply nominator and denominator by $(1-\lambda^*)$:

$$G_n = \frac{1}{1-\alpha} \operatorname{Re}\left[w(1-\lambda^*)(1-\lambda^{n+1})\right]. \tag{A.30}$$

Product $w(1-\lambda^*) = 1$, thus this function reads as

$$G_n = \frac{1}{1-\alpha} \left(1 - \operatorname{Re}\left[\lambda^{n+1}\right]\right). \tag{A.31}$$

By definition $|\lambda| = \sqrt{\alpha} < 1$, thus $\operatorname{Re}[\lambda^n] < 1$ for any $n > 0$. This means that $G_n > 0$ for any $n \geqslant 0$. Furthermore, by convexity of the sequence $\{CP^k\}$ for any $k \geqslant 0$: $CP^k - 2CP^{k+1} + CP^{k+2} > 0$.

Thus every term in (A.25) is greater than zero, which concludes the proof. □

A.3 Proof of the Lemma 2.3

Lemma 2.3. *Let M be an M-matrix. Let C be a row-vector such that $CM^2 > 0$ and $CM^2B\gamma > 0$. Then*

$$C(M + tM^{-1})^{-1}B\gamma > 0 \tag{A.32}$$

for any $t \geqslant 0$.

Proof. As in the previous proof, define $L = (M + tM^{-1})^{-1}$ and $M = s(I - P)$. By Lemma 2.2 $CL > 0$. Using the series expansion (A.25), we can write

$$CL = \sum_{k=0}^{+\infty} G_k(C - 2CP + CP^2)P^k > 0, \tag{A.33}$$

where all $G_k > 0$. Condition $CM^2B\gamma > 0$ reads as

$$(C - 2CP + CP^2)B\gamma > 0. \tag{A.34}$$

Then

$$\begin{aligned} CLB\gamma &= \sum_{k=0}^{+\infty} G_k(C - 2CP + CP^2)P^kB\gamma = \\ &= G_0(C - 2CP + CP^2)B\gamma + \sum_{k=1}^{+\infty} G_k(C - 2CP + CP^2)P^kB\gamma, \end{aligned} \tag{A.35}$$

where the first term is strictly greater than zero and all others a greater or equal than zero. Thus $CLB\gamma > 0$, which concludes the proof. □

A.4 Proof of the Lemma 5.1

Lemma 5.1. *Let $J(t,x) \in \mathbb{R}^{n\times n}$ be the Jacobian matrix of function $M(t,x)$. Let $J(t,x)$ satisfies the dynamic equation*

$$\frac{\partial J}{\partial t} = -\frac{\partial (Ju)}{\partial x}, \tag{A.36}$$

where $u = u(t,x)$ is some vector field. Then the determinant $\det J$ satisfies the same equation:

$$\frac{\partial \det J}{\partial t} = -\frac{\partial}{\partial x} \cdot (\det J \cdot u)\,. \tag{A.37}$$

Proof. First of all let us rewrite (A.36) for one element J_{ik} of the matrix J:

$$\frac{\partial J_{ik}}{\partial t} = -\frac{\partial (J_i\, u)}{\partial x_k} = -\sum_{j=1}^{n} \frac{\partial^2 M_i}{\partial x_k \partial x_j} u_j - \sum_{j=1}^{n} J_{ij}\frac{\partial u_j}{\partial x_k} = -\sum_{j=1}^{n} \frac{\partial J_{ik}}{\partial x_j} u_j - \sum_{j=1}^{n} J_{ij}\frac{\partial u_j}{\partial x_k}, \tag{A.38}$$

where we used the fact that $J = \partial M/\partial x$.

Now let us recall the definition of the determinant: $\det J = \sum_\sigma \operatorname{sgn}(\sigma) \prod_{i=1}^n J_{\sigma_i,i}$, where σ is a permutation of the set $\{1, 2, ..., n\}$ and $\sum_\sigma$ is taken over all possible permutations, with $\operatorname{sgn}(\sigma)$ being the sign of the permutation. Let us take the time derivative and then substitute (A.38):

$$\begin{aligned} \frac{\partial \det J}{\partial t} &= \sum_\sigma \operatorname{sgn}(\sigma) \sum_{k=1}^n \frac{\partial J_{\sigma_k,k}}{\partial t} \prod_{i=1, i\neq k}^n J_{\sigma_i,i} = \\ &= -\sum_{j=1}^n \sum_\sigma \operatorname{sgn}(\sigma) \sum_{k=1}^n \left[\frac{\partial J_{\sigma_k,k}}{\partial x_j} u_j + J_{\sigma_k,j} \frac{\partial u_j}{\partial x_k} \right] \prod_{i=1, i\neq k}^n J_{\sigma_i,i} \end{aligned} \tag{A.39}$$

We will investigate two parts of (A.39), corresponding to the two terms inside the square brackets. For the first term we have

$$-\sum_{j=1}^n \sum_\sigma \operatorname{sgn}(\sigma) \sum_{k=1}^n \frac{\partial J_{\sigma_k,k}}{\partial x_j} u_j \prod_{i=1, i\neq k}^n J_{\sigma_i,i} = -\sum_{j=1}^n \frac{\partial \det J}{\partial x_j} u_j = -\frac{\partial \det J}{\partial x} u. \tag{A.40}$$

The second term is a little more tricky:

$$\begin{aligned} &-\sum_{j=1}^n \sum_\sigma \operatorname{sgn}(\sigma) \sum_{k=1}^n J_{\sigma_k,j} \frac{\partial u_j}{\partial x_k} \prod_{i=1, i\neq k}^n J_{\sigma_i,i} = -\det J \sum_{j=1}^n \frac{\partial u_j}{\partial x_j} - \\ &-\sum_{j=1}^n \sum_\sigma \operatorname{sgn}(\sigma) \sum_{k=1, k\neq j}^n J_{\sigma_k,j} \frac{\partial u_j}{\partial x_k} \prod_{i=1, i\neq k}^n J_{\sigma_i,i}. \end{aligned}$$

Here we split the summation over k into the term with $k = j$ and all other terms. The former immediately gives the determinant multiplied by the divergence of the vector field, where the latter sum over all other terms is zero. Indeed, imagine a permutation $\bar\sigma$ such that it is equal to σ except σ_j and σ_k are swapped. Then the sign of $\bar\sigma$ is opposite to the sign of σ. Further, since the product $J_{\sigma_k,j} J_{\sigma_j,j}$ is the only way in which σ_k and σ_j enter the formula, the absolute value does not change with the change of permutation. Therefore for each j, k and for each permutation there exists a permutation which cancels them out.

Finally, substitution of the nonzero term of the last equation and (A.40) into (A.39) leads to (A.37). □

A.5 Proof of the Lemma 5.2

Lemma 5.2. *Let $\partial x/\partial M$ be isotropic, i.e. represented by a scalar multiplied by a rotation matrix, and let $\rho = \det(\partial M/\partial x)$. Then*

$$\nabla \cdot \left(\rho \frac{\partial x}{\partial M} \right) \frac{\partial x}{\partial M}^T = 0. \tag{A.41}$$

Proof. Define $\lambda = \lambda(\partial M/\partial x)$, thus $\rho = \lambda^n$. By isotropy,

$$\frac{\partial x}{\partial M} = \lambda^{-2} \frac{\partial M}{\partial x}^T$$

and therefore, by using (5.45), the left-hand side of (A.41) is

$$\nabla \cdot \left(\lambda^{n-2} \frac{\partial M}{\partial x}^T \right) \frac{\partial M}{\partial x} \lambda^{-2} = \lambda^{n-4} \nabla \cdot \left(\frac{\partial M}{\partial x}^T \right) \frac{\partial M}{\partial x} + (n-2)\lambda^{n-3} \frac{\partial \lambda}{\partial x}. \tag{A.42}$$

Now let us investigate the first term more closely. Taking the divergence and looking at j-th element, we see that

$$\left[\nabla \cdot \left(\frac{\partial M}{\partial x}^T \right) \frac{\partial M}{\partial x} \right]_j = \sum_{k=1}^{n} \frac{\partial^2 M}{\partial x_k^2}^T \frac{\partial M}{\partial x_j}. \tag{A.43}$$

Now, by isotropy

$$\frac{\partial M}{\partial x_j}^T \frac{\partial M}{\partial x_k} = 0 \ \ \forall j \neq k, \quad \frac{\partial M}{\partial x_k}^T \frac{\partial M}{\partial x_k} = \lambda^2. \tag{A.44}$$

Taking the derivative of the multiplication of basis vectors:

$$\frac{\partial}{\partial x_j} \left(\frac{\partial M}{\partial x_k}^T \frac{\partial M}{\partial x_k} \right) = 2 \frac{\partial^2 M}{\partial x_j \partial x_k}^T \frac{\partial M}{\partial x_k}, \tag{A.45}$$

but at the same time the value under the derivative is λ^2 by (A.44), therefore

$$\frac{\partial}{\partial x_j} \left(\frac{\partial M}{\partial x_k}^T \frac{\partial M}{\partial x_k} \right) = \frac{\partial \lambda^2}{\partial x_j} = 2\lambda \frac{\partial \lambda}{\partial x_j}. \tag{A.46}$$

Then, taking the derivative of multiplication of different basis vectors with $j \neq k$, by (A.44) we obtain zero:

$$\frac{\partial}{\partial x_k} \left(\frac{\partial M}{\partial x_j}^T \frac{\partial M}{\partial x_k} \right) = \frac{\partial^2 M}{\partial x_j \partial x_k}^T \frac{\partial M}{\partial x_k} + \frac{\partial M}{\partial x_j}^T \frac{\partial^2 M}{\partial x_k^2} = 0,$$

which by equality of (A.45) and (A.46) means that for $j \neq k$

$$\frac{\partial M}{\partial x_j}^T \frac{\partial^2 M}{\partial x_k^2} = -\frac{\partial^2 M}{\partial x_j \partial x_k}^T \frac{\partial M}{\partial x_k} = -\lambda \frac{\partial \lambda}{\partial x_j}. \tag{A.47}$$

In the case of $j = k$ by equality of (A.45) and (A.46) we have

$$\frac{\partial^2 M}{\partial x_j^2}^T \frac{\partial M}{\partial x_j} = \lambda \frac{\partial \lambda}{\partial x_j}. \tag{A.48}$$

Combination of (A.47) and (A.48) means that (A.43) is

$$\left[\nabla \cdot \left(\frac{\partial M}{\partial x}^T \right) \frac{\partial M}{\partial x} \right]_j = (2-n)\lambda \frac{\partial \lambda}{\partial x_j}. \tag{A.49}$$

Finally, substituting (A.49) in (A.42) gives zero. □

A.6 Solution to the equation (6.43)

Here we will solve equation (6.43), which is:

$$\frac{\cos\theta}{-B-2A\cos\theta}d\theta=\frac{1}{\Delta x}dx.$$

For simplicity let us define $J:=B/A$, thus the ODE becomes

$$\frac{\cos\theta}{-J-2\cos\theta}d\theta=\frac{A}{\Delta x}dx.$$

Define $t=\tan\frac{\theta}{2}$. Then

$$\cos\theta=\frac{1-t^2}{1+t^2},\qquad d\theta=\frac{2}{1+t^2}dt,$$

and thus

$$\int\frac{\cos\theta}{-J-2\cos\theta}d\theta=\int\frac{t^2-1}{J(1+t^2)+2(1-t^2)}\frac{2}{1+t^2}dt.$$

Define $p=J+2$ and $q=J-2$. Then

$$\frac{t^2-1}{J(1+t^2)+2(1-t^2)}=\frac{t^2-1}{p+qt^2}=\frac{1}{q}\left(1-\frac{2J}{p+qt^2}\right).$$

Further,

$$\frac{1}{1+t^2}\frac{1}{p+qt^2}=\frac{1}{4}\left[\frac{1}{1+t^2}-\frac{q}{p+qt^2}\right],$$

which leads to

$$\frac{2}{1+t^2}\frac{1}{q}\left(1-\frac{2J}{p+qt^2}\right)=J\frac{1}{p+qt^2}-\frac{1}{1+t^2}.$$

A.6.1 Case 1.

Let $pq<0$. This is equivalent to the statement $|J|<2$. Then

$$\int\frac{1}{p+qt^2}dt=\frac{1}{\sqrt{-pq}}\int\frac{1}{1-\left(\frac{q}{\sqrt{-pq}}t\right)^2}\frac{-q}{\sqrt{-pq}}dt=$$

$$=\frac{1}{2\sqrt{-pq}}\int\left[\frac{1}{1+\left(\frac{q}{\sqrt{-pq}}t\right)}+\frac{1}{1-\left(\frac{q}{\sqrt{-pq}}t\right)}\right]\frac{-q}{\sqrt{-pq}}dt=\frac{1}{2\sqrt{-pq}}\ln\left|\frac{1-\left(\frac{q}{\sqrt{-pq}}t\right)}{1+\left(\frac{q}{\sqrt{-pq}}t\right)}\right|,$$

and the solution with C being integration constant is

$$\frac{Ax}{\Delta x}+C=\frac{J}{2\sqrt{-pq}}\ln\left|\frac{1-\left(\frac{q}{\sqrt{-pq}}\tan\frac{\theta}{2}\right)}{1+\left(\frac{q}{\sqrt{-pq}}\tan\frac{\theta}{2}\right)}\right|-\arctan\left(\tan\frac{\theta}{2}\right),$$

which can be simplified to

$$A\frac{x}{\Delta x}+C=\frac{J}{2\sqrt{4-J^2}}\ln\left|\frac{1+\left(\frac{2-J}{\sqrt{4-J^2}}\tan\frac{\theta}{2}\right)}{1-\left(\frac{2-J}{\sqrt{4-J^2}}\tan\frac{\theta}{2}\right)}\right|-\frac{1}{2}\theta. \tag{A.50}$$

A.6.2 Case 2.

Let $pq > 0$. This is equivalent to the statement $|J| > 2$. Then

$$\int \frac{1}{p+qt^2}dt = \frac{1}{\sqrt{qp}} \int \frac{1}{1+\left(\frac{q}{\sqrt{qp}}t\right)^2} \frac{q}{\sqrt{qp}}dt = \frac{1}{\sqrt{qp}} \arctan\left(\frac{q}{\sqrt{qp}}t\right),$$

and the full solution is therefore

$$A\frac{x}{\Delta x} + C = \frac{J}{\sqrt{qp}} \arctan\left(\frac{q}{\sqrt{qp}} \tan\frac{\theta}{2}\right) - \arctan\left(\tan\frac{\theta}{2}\right),$$

or simply

$$A\frac{x}{\Delta x} + C = \frac{J}{\sqrt{J^2-4}} \arctan\left(\frac{J-2}{\sqrt{J^2-4}} \tan\frac{\theta}{2}\right) - \frac{1}{2}\theta. \tag{A.51}$$

Note that solutions (A.50) and (A.51) are essentially the same functions if one uses complex relation between arctangent and logarithm:

$$\arctan s = -\frac{i}{2} \ln\left(\frac{1+is}{1-is}\right). \tag{A.52}$$

A.6.3 Case 3.

Finally, if $pq = 0$, it means that $J = \pm 2$.

1. If $J = 2$ we get $p = 4$ and $q = 0$. Then the solution is given simply by
$$\frac{Ax}{\Delta x} + C = \frac{1}{2}\tan\frac{\theta}{2} - \frac{1}{2}\theta.$$
2. If $J = -2$ we get $p = 0$ and $q = -4$. Then the solution is
$$\frac{Ax}{\Delta x} + C = -\frac{1}{2}\cot\frac{\theta}{2} - \frac{1}{2}\theta.$$

A.7 Proof of the Lemma 6.1

Lemma 6.1. *Function $f(x)$, defined as*

$$f(x) = \frac{V + \mu x}{(U+x)^2} \tag{A.53}$$

with $U > 0$ and $x > 0$ is bounded from above by

$$f(x) \leqslant \begin{cases} 0, & V \leqslant 0 \text{ and } \mu \leqslant 0, \\ V/U^2, & V > 0 \text{ and } U\mu \leqslant 2V, \\ \dfrac{\mu^2}{4\mu U - 4V}, & \mu > 0 \text{ and } U\mu > 2V. \end{cases} \tag{A.54}$$

Proof. The function $f(x)$ is defined for $x \in [0, +\infty)$, thus its supremum is achieved either at $x = 0$, $x = +\infty$ or at $f'(x) = 0$. If $V \leqslant 0$ and $\mu \leqslant 0$, then the function is nonpositive with asymptotic value $f(+\infty) = 0$, thus we use 0 as a bound in this case. Let us now find its extremum:

$$f'(x) = \left(\frac{V + \mu x}{(U + x)^2}\right)' = \frac{\mu(U - x) - 2V}{(U + x)^3} = 0, \tag{A.55}$$

thus it is achieved at $x_{extr} = U - 2V/\mu$. Substituting it back in (A.53) we obtain

$$f(x_{extr}) = \frac{\mu^2}{4\mu U - 4V}. \tag{A.56}$$

Finally we notice that the extremum (A.55) is indeed maximum only if $\mu > 0$ and if $x_{extr} > 0$, otherwise the maximum is achieved at zero, $f(0) = V/U^2$. Therefore, combining the bounds together we get

$$f(x) \leqslant \begin{cases} 0, & V \leqslant 0 \text{ and } \mu \leqslant 0, \\ V/U^2, & V > 0 \text{ and } U\mu \leqslant 2V, \\ \dfrac{\mu^2}{4\mu U - 4V}, & \mu > 0 \text{ and } U\mu > 2V. \end{cases} \tag{A.57}$$

□

Bibliography

Acebrón, Juan A et al. (2005a). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Reviews of modern physics* 77.1, p. 137 (cit. on p. 69).

Acebrón, Juan A et al. (2005b). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Reviews of modern physics* 77.1, p. 137 (cit. on p. 127).

Aoki, Masanao (1968). "Control of large-scale dynamic systems by aggregation". In: *IEEE Transactions on Automatic Control* 13.3, pp. 246–253 (cit. on p. 69).

Aoyagi, Toshio (1995). "Network of neural oscillators for retrieving phase information". In: *Physical review letters* 74.20, p. 4075 (cit. on p. 123).

Arecchi, FT et al. (1984). "Deterministic chaos in laser with injected signal". In: *Optics communications* 51.5, pp. 308–314 (cit. on p. 123).

Arenas, Alex et al. (2008). "Synchronization in complex networks". In: *Physics reports* 469.3, pp. 93–153 (cit. on p. 127).

Ariyur, Kartik B and Miroslav Krstić (2003). *Real time optimization by extremum seeking control*. Wiley Online Library (cit. on pp. 15, 40).

Atkinson, Frederick Valentine (1964). *Discrete and continuous boundary problems*. Academic Press (cit. on p. 68).

Baiti, Paolo and Helge Kristian Jenssen (1998). "On the front-tracking algorithm". In: *Journal of mathematical analysis and applications* 217.2, pp. 395–404 (cit. on p. 61).

Baker, James and Panagiotis D Christofides (2000). "Finite-dimensional approximation and control of non-linear parabolic PDE systems". In: *International Journal of Control* 73.5, pp. 439–456 (cit. on p. 52).

Bamieh, Bassam, Fernando Paganini, and Munther A Dahleh (2002). "Distributed control of spatially invariant systems". In: *IEEE Transactions on automatic control* 47.7, pp. 1091–1107 (cit. on pp. 70, 85).

Bamieh, Bassam et al. (2012). "Coherence in large-scale networks: Dimension-dependent limitations of local feedback". In: *IEEE Transactions on Automatic Control* 57.9, pp. 2235–2249 (cit. on pp. 6, 114).

Barahona, Mauricio and Louis M Pecora (2002). "Synchronization in small-world systems". In: *Physical review letters* 89.5, p. 054101 (cit. on p. 127).

Barooah, Prabir, Prashant G Mehta, and Joao P Hespanha (2009). "Mistuning-based control design to improve closed-loop stability margin of vehicular platoons". In: *IEEE Transactions on Automatic Control* 54.9, pp. 2100–2113 (cit. on pp. 6, 97).

Barrault, Maxime et al. (2004). "An 'empirical interpolation'method: application to efficient reduced-basis discretization of partial differential equations". In: *Comptes Rendus Mathematique* 339.9, pp. 667–672 (cit. on p. 52).

Baydin, Atilim Gunes et al. (2018). "Automatic differentiation in machine learning: a survey". In: *Journal of machine learning research* 18 (cit. on pp. 68, 82).

Belishev, Michael I (2007). "Recent progress in the boundary control method". In: *Inverse problems* 23.5, R1 (cit. on p. 5).

Bennett, Matthew et al. (2002). "Huygens's clocks". In: *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 458.2019, pp. 563–579 (cit. on p. 126).

Berger, Luc (1996). "Emission of spin waves by a magnetic multilayer traversed by a current". In: *Physical Review B* 54.13, p. 9353 (cit. on p. 136).

Bergner, Andre et al. (2012). "Remote synchronization in star networks". In: *Physical Review E* 85.2, p. 026208 (cit. on p. 123).

Biccari, Umberto, Dongnam Ko, and Enrique Zuazua (2019). "Dynamics and control for multi-agent networked systems: A finite-difference approach". In: *Mathematical Models and Methods in Applied Sciences* 29.04, pp. 755–790 (cit. on p. 98).

Borkar, Vivek S (2009). *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer (cit. on p. 45).

Boynard, Anne et al. (2014). "First simultaneous space measurements of atmospheric pollutants in the boundary layer from IASI: A case study in the North China Plain". In: *Geophysical Research Letters* 41.2, pp. 645–651 (cit. on p. 16).

Briat, Corentin (2013). "Robust stability and stabilization of uncertain linear positive systems via integral linear constraints: L1-gain and L∞-gain characterization". In: *International Journal of Robust and Nonlinear Control* 23.17, pp. 1932–1954 (cit. on p. 15).

Canudas-de-Wit, Carlos (2015). *ERC: Scale-Free Control for Complex Physical Network Systems*. http://scale-freeback.eu (cit. on p. 14).

Caprino, S et al. (1993). "Hydrodynamic limits of the Vlasov equation". In: *Communications in partial differential equations* 18.5-6, pp. 805–820 (cit. on pp. 106, 110).

Carr, Thomas W, Michael L Taylor, and Ira B Schwartz (2006). "Negative-coupling resonances in pump-coupled lasers". In: *Physica D: Nonlinear Phenomena* 213.2, pp. 152–163 (cit. on p. 123).

Casadei, Giacomo, Carlos Canudas-de-Wit, and Sandro Zampieri (2018). "Controllability of large-scale networks: An output controllability approach". In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 5886–5891 (cit. on p. 14).

Chapman, Sydney and Thomas George Cowling (1990). *The mathematical theory of non-uniform gases: an account of the kinetic theory of viscosity, thermal conduction and diffusion in gases*. Cambridge university press (cit. on p. 106).

Cheng, Xiaodong, Yu Kawano, and Jacquelien MA Scherpen (2017). "Reduction of second-order network systems with structure preservation". In: *IEEE Transactions on Automatic Control* 62.10, pp. 5026–5038 (cit. on p. 6).

Chung, Soon-Jo et al. (2018). "A survey on aerial swarm robotics". In: *IEEE Transactions on Robotics* 34.4, pp. 837–855 (cit. on p. 113).

Cockburn, Bernardo, George E Karniadakis, and Chi-Wang Shu (2012). *Discontinuous Galerkin methods: theory, computation and applications*. Vol. 11. Springer Science & Business Media (cit. on p. 52).

Commault, Christian, Jacob van der Woude, and Paolo Frasca (2019). "Functional target controllability of networks: structural properties and efficient algorithms". In: *IEEE Transactions on Network Science and Engineering* 7.3, pp. 1521–1530 (cit. on p. 14).

Coron, Jean-Michel, Brigitte d'Andrea-Novel, and Georges Bastin (2007). "A strict Lyapunov function for boundary control of hyperbolic systems of conservation laws". In: *IEEE Transactions on Automatic control* 52.1, pp. 2–11 (cit. on p. 5).

Cowan, Noah J et al. (2012). "Nodal dynamics, not degree distributions, determine the structural controllability of complex networks". In: *PloS one* 7.6, e38398 (cit. on p. 5).

Dieudonné, Christophe (2015). "Synchronization of a Spin Transfer oscillator to a RF current: mechanisms and room-temperature characterization." PhD thesis. Université Grenoble Alpes (cit. on pp. 136, 146).

Dion, Jean-Michel, Christian Commault, and Jacob Van der Woude (2003). "Generic properties and control of linear structured systems: a survey". In: *Automatica* 39.7, pp. 1125–1144 (cit. on p. 2).

Dörfler, Florian and Francesco Bullo (2014). "Synchronization in complex networks of phase oscillators: A survey". In: *Automatica* 50.6, pp. 1539–1564 (cit. on p. 127).

Dörfler, Florian, Michael Chertkov, and Francesco Bullo (2013). "Synchronization in complex oscillator networks and smart grids". In: *Proceedings of the National Academy of Sciences* 110.6, pp. 2005–2010 (cit. on pp. 126, 127, 135).

Ermentrout, Bard (1991). "An adaptive model for synchrony in the firefly Pteroptyx malaccae". In: *Journal of Mathematical Biology* 29.6, pp. 571–585 (cit. on p. 126).

Eslami, Mehrad and Karim Faez (2010). "Automatic Traffic Monitoring from Satellite Images Using Artificial Immune System". In: *Structural, Syntactic, and Statistical Pattern Recognition*. Ed. by Edwin R. Hancock et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 170–179 (cit. on p. 16).

Euler, Leonhard (1761). "Principia motus fluidorum". In: *Novi commentarii academiae scientiarum Petropolitanae*, pp. 271–311 (cit. on p. 2).

Fan, Ky (1958). "Topological proofs for certain theorems on matrices with non-negative elements". In: *Monatshefte für Mathematik* 62.3, pp. 219–237 (cit. on p. 22).

Filatrella, Giovanni, Arne Hejde Nielsen, and Niels Falsig Pedersen (2008). "Analysis of a power grid using a Kuramoto-like model". In: *The European Physical Journal B* 61.4, pp. 485–491 (cit. on pp. 126, 132, 133).

Fischer, Jens V (2018). "Four particular cases of the Fourier transform". In: *Mathematics* 6.12, p. 335 (cit. on p. 78).

Folland, Gerald B (2020). *Introduction to partial differential equations*. Princeton university press (cit. on p. 115).

Frazho, Arthur E and Wisuwat Bhosri (2010). "Toeplitz and Laurent Operators". In: *An Operator Perspective on Signals and Systems*. Springer (cit. on pp. 76, 78).

Frihauf, Paul and Miroslav Krstic (2010). "Leader-enabled deployment onto planar curves: A PDE-based approach". In: *IEEE Transactions on Automatic Control* 56.8, pp. 1791–1806 (cit. on p. 90).

Gallagher, Isabelle, Laure Saint-Raymond, and Benjamin Texier (2013). *From Newton to Boltzmann: hard spheres and short-range potentials*. European Mathematical Society (cit. on p. 106).

Gantmakher, Feliks Ruvimovich and Mark Grigorevich Krein (1941). *Oscillation matrices and kernels and small vibrations of mechanical systems*. 345. American Mathematical Soc. (cit. on p. 68).

Gao, Shuang and Peter E Caines (2019). "Graphon control of large-scale networks of linear systems". In: *IEEE Transactions on Automatic Control* 65.10, pp. 4090–4105 (cit. on p. 69).

Godunov, Sergei Konstantinovich (1959). "A difference scheme for numerical solution of discontinuous solution of hydrodynamic equations". In: *Math. Sbornik* 47, pp. 271–306 (cit. on p. 60).

Gorban, Alexander and Ilya Karlin (2014). "Hilbert's 6th problem: exact and approximate hydrodynamic manifolds for kinetic equations". In: *Bulletin of the American Mathematical Society* 51.2, pp. 187–246 (cit. on p. 106).

Grabert, Hermann (2006). *Projection operator techniques in nonequilibrium statistical mechanics*. Vol. 95. Springer (cit. on p. 69).

Grad, Harold (1949). "On the kinetic theory of rarefied gases". In: *Communications on pure and applied mathematics* 2.4, pp. 331–407 (cit. on pp. 106, 110).

Greenshields, BD et al. (1935). "A study of traffic capacity". In: *Highway research board proceedings*. Vol. 1935. National Research Council (USA), Highway Research Board (cit. on p. 59).

Hilbert, David (1902). "Mathematical problems". In: *Bulletin of the American Mathematical Society* 8.10, pp. 437–479 (cit. on p. 106).

Hunt, Andrew et al. (2021). "High-speed density measurement for LNG and other cryogenic fluids using electrical capacitance tomography". In: *Cryogenics* 113, p. 103207 (cit. on p. 16).

Jafarizadeh, Saber (2020). "Weighted average consensus-based optimization of Advection-Diffusion Systems". In: *IEEE Transactions on Signal and Information Processing over Networks* (cit. on p. 90).

Jafarpour, Saber and Francesco Bullo (2018). "Synchronization of Kuramoto oscillators via cutset projections". In: *IEEE Transactions on Automatic Control* 64.7, pp. 2830–2844 (cit. on pp. 127, 135).

Jovanovic, Mihailo R and Bassam Bamieh (2005). "On the ill-posedness of certain vehicular platoon control problems". In: *IEEE Transactions on Automatic Control* 50.9, pp. 1307–1321 (cit. on pp. 6, 114).

Kalman, RE (1965). "Irreducible realizations and the degree of a rational matrix". In: *Journal of the Society for Industrial and Applied Mathematics* 13.2, pp. 520–544 (cit. on p. 6).

Kanellakopoulos, Ioannis, Petar V Kokotovic, and A Stephen Morse (1991). "Systematic design of adaptive controllers for feedback linearizable systems". In: *1991 American control conference*. IEEE, pp. 649–654 (cit. on p. 122).

Khalil, Hassan K and Jessy W Grizzle (2002). *Nonlinear systems*. Vol. 3. Prentice hall Upper Saddle River, NJ (cit. on p. 43).

Klickstein, Isaac, Afroza Shirin, and Francesco Sorrentino (2017a). "Energy scaling of targeted optimal control of complex networks". In: *Nature communications* 8.1, pp. 1–10 (cit. on p. 14).

Klickstein, Isaac, Afroza Shirin, and Francesco Sorrentino (2017b). "Locally optimal control of complex networks". In: *Physical review letters* 119.26, p. 268301 (cit. on p. 14).

Kokotović, Petar, Hassan K Khalil, and John O'reilly (1999). *Singular perturbation methods in control: analysis and design*. SIAM (cit. on p. 42).

Kolmogorov, Andrei Nikolaevich (1957). "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition". In: *Doklady Akademii Nauk*. Vol. 114. 5. Russian Academy of Sciences, pp. 953–956 (cit. on p. 82).

Kottenstette, Nicholas et al. (2014). "On relationships among passivity, positive realness, and dissipativity in linear systems". In: *Automatica* 50.4, pp. 1003–1016 (cit. on pp. 15, 25).

Krstic, Miroslav and Andrey Smyshlyaev (2008). *Boundary control of PDEs: A course on backstepping designs*. SIAM (cit. on p. 5).

Kuehn, Christian (2016). "Moment closure—a brief review". In: *Control of self-organizing nonlinear systems*, pp. 253–271 (cit. on pp. 7, 51, 69).

Kuramoto, Yoshiki (2003). *Chemical oscillations, waves, and turbulence*. Courier Corporation (cit. on p. 126).

Lasiecka, Irene and Roberto Triggiani (1989). "Exact controllability of the wave equation with Neumann boundary control". In: *Applied Mathematics and Optimization* 19.1, pp. 243–290 (cit. on p. 5).

Leitold, Dániel, Ágnes Vathy-Fogarassy, and János Abonyi (2017). "Controllability and observability in complex networks–the effect of connection types". In: *Scientific reports* 7.1, pp. 1–9 (cit. on p. 2).

Li, Han-Xiong and Chenkun Qi (2010). "Modeling of distributed parameter systems for applications—A synthesized review from time–space separation". In: *Journal of Process Control* 20.8, pp. 891–901 (cit. on p. 52).

Lighthill, Michael James and Gerald Beresford Whitham (1955). "On kinematic waves II. A theory of traffic flow on long crowded roads". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178, pp. 317–345 (cit. on pp. 4, 52, 99).

Lin, Ching-Tai (1974). "Structural controllability". In: *IEEE Transactions on Automatic Control* 19.3, pp. 201–208 (cit. on p. 2).

Liu, Yang-Yu, Jean-Jacques Slotine, and Albert-László Barabási (2011). "Controllability of complex networks". In: *nature* 473.7346, pp. 167–173 (cit. on p. 5).

Lovász, László (2012). *Large networks and graph limits*. Vol. 60. American Mathematical Soc. (cit. on p. 69).

Martin, Nicolas, Paolo Frasca, and Carlos Canudas-de-Wit (2019). "Large-scale network reduction towards scale-free structure". In: *IEEE Transactions on Network Science and Engineering* 6.4, pp. 711–723 (cit. on p. 6).

Maxwell, James Clerk (1873). *A treatise on electricity and magnetism*. Vol. 1. Oxford: Clarendon Press (cit. on p. 2).

Medvedev, Georgi S (2014). "The nonlinear heat equation on dense graphs and graph limits". In: *SIAM Journal on Mathematical Analysis* 46.4, pp. 2743–2766 (cit. on p. 69).

Mirtabatabaei, Anahita and Francesco Bullo (2012). "Opinion dynamics in heterogeneous networks: Convergence conjectures and theorems". In: *SIAM Journal on Control and Optimization* 50.5, pp. 2763–2785 (cit. on p. 2).

Mitrinovic, Dragoslav S and Petar M Vasic (1970). *Analytic inequalities*. Springer (cit. on p. 79).

Moase, William H et al. (2011). "Non-local stability of a multi-variable extremum-seeking scheme". In: *2011 Australian Control Conference*. IEEE, pp. 38–43 (cit. on p. 41).

Molnár, Tamás G et al. (2019). "Lagrangian models for controlling large-scale heterogeneous traffic". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, pp. 3152–3157 (cit. on pp. 90, 101).

Molnár, Tamás G et al. (2020). "Open and closed loop traffic control by connected automated vehicles". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 239–244 (cit. on p. 101).

Moore, Bruce (1981). "Principal component analysis in linear systems: Controllability, observability, and model reduction". In: *IEEE transactions on automatic control* 26.1, pp. 17–32 (cit. on p. 6).

Narendra, Kumpati S (2014). *Frequency domain criteria for absolute stability*. Elsevier (cit. on p. 34).

Nedić, Angelia and Asuman Ozdaglar (2009). "Subgradient methods for saddle-point problems". In: *Journal of optimization theory and applications* 142.1, pp. 205–228 (cit. on p. 41).

Newell, Gordon F (1993). "A simplified theory of kinematic waves in highway traffic, part I: General theory". In: *Transportation Research Part B: Methodological* 27.4, pp. 281–287 (cit. on pp. 57, 89).

Niazi, Muhammad Umar B, Carlos Canudas-de-Wit, and Alain Y Kibangou (2020a). "Average state estimation in large-scale clustered network systems". In: *IEEE Transactions on Control of Network Systems* 7.4, pp. 1736–1745 (cit. on p. 48).

Niazi, Muhammad Umar B, Carlos Canudas-de-Wit, and Alain Y Kibangou (2020b). "State Variance Estimation in Large-Scale Network Systems". In: *2020 59th IEEE Conference on Decision and Control (CDC)*, pp. 6052–6057 (cit. on p. 48).

Niazi, Muhammad Umar B et al. (2019). "Structure-based clustering algorithm for model reduction of large-scale network systems". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, pp. 5038–5043 (cit. on pp. 6, 69).

Nogueira, Filipa Nunes (2013). "Contributions to positive systems control: mathematical methods and applications". PhD thesis. Universidade do Porto (Portugal) (cit. on pp. 15, 26).

Nykamp, Duane Q and Daniel Tranchina (2000). "A population density approach that facilitates large-scale modeling of neural networks: Analysis and an application to orientation tuning". In: *Journal of computational neuroscience* 8.1, pp. 19–50 (cit. on p. 69).

Oh, Kwang-Kyo, Myoung-Chul Park, and Hyo-Sung Ahn (2015). "A survey of multi-agent formation control". In: *Automatica* 53, pp. 424–440 (cit. on p. 113).

Palubinskas, Gintautas, Franz Kurz, and Peter Reinartz (2010). "Model based traffic congestion detection in optical remote sensing imagery". In: *European Transport Research Review* 2.2, pp. 85–92 (cit. on p. 16).

Pietrzak, A et al. (2015). "Heading to 1 kW levels with laser bars of high-efficiency and emission wavelength around 880 nm and 940 nm". In: *High-Power Diode Laser Technology and Applications XIII*. Vol. 9348. International Society for Optics and Photonics, 93480E (cit. on p. 123).

Plemmons, Robert J (1977). "M-matrix characterizations. I—nonsingular M-matrices". In: *Linear Algebra and its Applications* 18.2, pp. 175–188 (cit. on p. 22).

Prieur, Christophe, Joseph Winkin, and Georges Bastin (2008). "Robust boundary control of systems of conservation laws". In: *Mathematics of Control, Signals, and Systems* 20.2, pp. 173–197 (cit. on p. 5).

Qi, Jie, Rafael Vazquez, and Miroslav Krstic (2014). "Multi-agent deployment in 3-D via PDE control". In: *IEEE Transactions on Automatic Control* 60.4, pp. 891–906 (cit. on p. 114).

Rantzer, Anders (2011). "Distributed control of positive systems". In: *2011 50th IEEE Conference on Decision and Control and European Control Conference*. IEEE, pp. 6608–6611 (cit. on p. 15).

Rantzer, Anders (2015). "Scalable control of positive systems". In: *European Journal of Control* 24, pp. 72–80 (cit. on p. 15).

Richards, Paul I (1956). "Shock waves on the highway". In: *Operations research* 4.1, pp. 42–51 (cit. on pp. 4, 52, 99).

Rodrigues, Francisco A et al. (2016). "The Kuramoto model in complex networks". In: *Physics Reports* 610, pp. 1–98 (cit. on pp. 132, 133).

Rohden, Martin et al. (2012). "Self-organized synchronization in decentralized power grids". In: *Physical review letters* 109.6, p. 064101 (cit. on p. 126).

Rossi, Wilbert Samuel and Paolo Frasca (2018). "On the convergence of message passing computation of harmonic influence in social networks". In: *IEEE Transactions on Network Science and Engineering* 6.2, pp. 116–129 (cit. on pp. 6, 14).

Saint-Raymond, Laure (2009). *Hydrodynamic limits of the Boltzmann equation*. 1971. Springer Science & Business Media (cit. on p. 106).

Salam, Fathi, J Marsden, and P Varaiya (1984). "Arnold diffusion in the swing equations of a power system". In: *IEEE Transactions on Circuits and Systems* 31.8, pp. 673–688 (cit. on p. 126).

Saxena, Garima, Awadhesh Prasad, and Ram Ramaswamy (2012). "Amplitude death: The emergence of stationarity in coupled nonlinear systems". In: *Physics Reports* 521.5, pp. 205–228 (cit. on pp. 123, 124).

Sepulchre, Rodolphe, Mrdjan Jankovic, and Petar V Kokotovic (2012). *Constructive nonlinear control*. Springer Science & Business Media (cit. on pp. 15, 25).

Simpson-Porco, John W et al. (2019). "Input–Output Performance of Linear–Quadratic Saddle-Point Algorithms With Application to Distributed Resource Allocation Problems". In: *IEEE Transactions on Automatic Control* 65.5, pp. 2032–2045 (cit. on p. 41).

Slavin, Andrei and Vasil Tiberkevich (2009). "Nonlinear auto-oscillator theory of microwave generation by spin-polarized current". In: *IEEE Transactions on Magnetics* 45.4, pp. 1875–1918 (cit. on pp. 136, 137).

Slonczewski, John C (1996). "Current-driven excitation of magnetic multilayers". In: *Journal of Magnetism and Magnetic Materials* 159.1-2, pp. L1–L7 (cit. on p. 136).

Smyshlyaev, Andrey and Miroslav Krstic (2004). "Closed-form boundary state feedbacks for a class of 1-D partial integro-differential equations". In: *IEEE Transactions on Automatic Control* 49.12, pp. 2185–2202 (cit. on p. 122).

Smyshlyaev, Andrey and Miroslav Krstic (2005). "On control design for PDEs with space-dependent diffusivity or time-dependent reactivity". In: *Automatica* 41.9, pp. 1601–1608 (cit. on pp. 122, 125).

Stiles, Mark D and Jacques Miltat (2006). "Spin-transfer torque and dynamics". In: *Spin dynamics in confined magnetic structures III*, pp. 225–308 (cit. on p. 136).

Strogatz, Steven H et al. (2005). "Crowd synchrony on the Millennium Bridge". In: *Nature* 438.7064, pp. 43–44 (cit. on p. 126).

Takloo-Bighash, Ramin (2018). "How many lattice points are there on a circle or a sphere?" In: *A Pythagorean Introduction to Number Theory*. Springer, pp. 151–164 (cit. on p. 112).

Tan, Ying, Dragan Nešić, and Iven Mareels (2006). "On non-local stability properties of extremum seeking control". In: *Automatica* 42.6, pp. 889–903 (cit. on pp. 43, 44).

Tan, Ying et al. (2010). "Extremum seeking from 1922 to 2010". In: *Proceedings of the 29th Chinese control conference*. IEEE, pp. 14–26 (cit. on pp. 15, 40, 44).

Tanaka, Hisa-Aki, Allan J Lichtenberg, and Shin'ichi Oishi (1997a). "First order phase transition resulting from finite inertia in coupled oscillator systems". In: *Physical review letters* 78.11, p. 2104 (cit. on p. 126).

Tanaka, Hisa-Aki, Allan J Lichtenberg, and Shin'ichi Oishi (1997b). "Self-synchronization of coupled oscillators with hysteretic responses". In: *Physica D: Nonlinear Phenomena* 100.3-4, pp. 279–300 (cit. on p. 126).

Tanner, Herbert G (2004). "On the controllability of nearest neighbor interconnections". In: *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*. Vol. 3. IEEE, pp. 2467–2472 (cit. on p. 2).

Tao, G and Petros A Ioannou (1990). "Necessary and sufficient conditions for strictly positive real matrices". In: *IEE Proceedings G (Circuits, Devices and Systems)* 137.5, pp. 360–366 (cit. on p. 34).

Tol, Henry J, Cornelis C de Visser, and Marios Kotsonis (2019). "Model reduction of parabolic PDEs using multivariate splines". In: *International Journal of Control* 92.1, pp. 175–190 (cit. on p. 52).

Toner, John and Yuhai Tu (1995). "Long-range order in a two-dimensional dynamical XY model: how birds fly together". In: *Physical review letters* 75.23, p. 4326 (cit. on p. 113).

Tumash, Liudmila, Carlos Canudas-de-Wit, and Maria Laura Delle Monache (2021a). "Boundary Control Design for Traffic with Nonlinear Dynamics". In: *IEEE Transactions on Automatic Control* (cit. on p. 5).

Tumash, Liudmila, Carlos Canudas-de-Wit, and Maria Laura Delle Monache (2021b). "Boundary Control for Multi-Directional Traffic on Urban Networks". In: *Submitted to CDC 2021-60th IEEE Conference on Decision and Control* (cit. on p. 106).

Tumash, Liudmila, Carlos Canudas-de-Wit, and Maria Laura Delle Monache (2021c). "Multi-Directional Continuous Traffic Model For Large-Scale Urban Networks". In: *Submitted to Transportation Research Part B: Methodological* (cit. on p. 106).

Tumash, Liudmila, Simona Olmi, and Eckehard Schöll (2019). "Stability and control of power grids with diluted network topology". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29.12, p. 123105 (cit. on p. 127).

Van Veen, Barry D et al. (1997). "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering". In: *IEEE Transactions on biomedical engineering* 44.9, pp. 867–880 (cit. on p. 16).

Vassio, Luca et al. (2014). "Message passing optimization of harmonic influence centrality". In: *IEEE Transactions on Control of Network Systems* 1.1, pp. 109–120 (cit. on p. 14).

Villani, Cédric (2009). *Optimal transport: old and new*. Vol. 338. Springer (cit. on p. 52).

Vizuete, Renato, Paolo Frasca, and Federica Garin (2020). "Graphon-based sensitivity analysis of SIS epidemics". In: *IEEE Control Systems Letters* 4.3, pp. 542–547 (cit. on p. 69).

Watts, Duncan J (2018). *Small worlds*. Princeton university press (cit. on p. 127).

Wiesenfeld, Kurt, Pere Colet, and Steven H Strogatz (1996). "Synchronization transitions in a disordered Josephson series array". In: *Physical review letters* 76.3, p. 404 (cit. on p. 126).

Winfree, Arthur T (1967). "Biological rhythms and the behavior of populations of coupled oscillators". In: *Journal of theoretical biology* 16.1, pp. 15–42 (cit. on p. 127).

Winful, Herbert G and Lutfur Rahman (1990). "Synchronized chaos and spatiotemporal chaos in arrays of coupled lasers". In: *Physical Review Letters* 65.13, p. 1575 (cit. on p. 124).

Wittmuess, Philipp, Michael Heidingsfeld, and Oliver Sawodny (2016). "Optimal Actuator Design for Optimal Output Controllability". In: *IFAC-PapersOnLine* 49.21, pp. 234–239 (cit. on p. 14).

Yan, Gang et al. (2012). "Controlling complex networks: How much energy is needed?" In: *Physical review letters* 108.21, p. 218703 (cit. on p. 5).

Yang, Yuecheng, Dimos V Dimarogonas, and Xiaoming Hu (2015). "Shaping up crowd of agents through controlling their statistical moments". In: *2015 European Control Conference (ECC)*. IEEE, pp. 1017–1022 (cit. on pp. 7, 69).

York, Robert A and Richard C Compton (1991). "Quasi-optical power combining using mutually synchronized oscillator arrays". In: *IEEE Transactions on Microwave Theory and Techniques* 39.6, pp. 1000–1009 (cit. on p. 126).

Zhang, Silun et al. (2021). "Modeling collective behaviors: A moment-based approach". In: *IEEE Transactions on Automatic Control* 66.1, pp. 33–48 (cit. on pp. 52, 69).

Zhou, Kemin, Gregory Salomon, and Eva Wu (1999). "Balanced realization and model reduction for unstable systems". In: *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal* 9.3, pp. 183–198 (cit. on p. 6).