Thèse n° 10 550

Stability: a search for structure

Présentée le 20 juin 2024

Collège du management de la technologie
Chaire analyse de risque et optimisation
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Wouter JONGENEEL

Acceptée sur proposition du jury

Prof. A. Billard, présidente du jury
Prof. D. Kuhn, directeur de thèse
Prof. F. Dörfler, rapporteur
Prof. M. Johansson, rapporteur
Prof. G. Ferrari Trecate, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

Preface

> "*And we hope that he will finish the book with the ability to apply to his system problems whatever technique is most appropriate, irrespective of tradition or habit.*"
>
> — Kalman, Falb and Arbib [KFA69, p. 12].

The mathematical notion of stability is practically important and usually intuitive, yet, an elusive concept to work with.

The Cambridge dictionary defines[1] *stability* as: "*a situation in which something is not likely to move or change*". This definition captures the above, we all have a feeling for what we mean by stability, yet to make this precise we must, mathematically, specify the "*situation*", the "*something*", what we mean by "*likely*" and what it means for this something to "*move*" or "*change*".

In this thesis we study the interplay between stability and structural properties of the problem at hand. Not only does this lead to a better understanding of stability and the problem, but knowledge of merely structural properties is commonly the most we can ask for. Hence, in an age dominated by the demand for guarantees, it is beneficial to be able to provide some.

A search for structure has always been there in dynamical systems and control theory, notably, due to influential works like the 1967 edition of the pivotal book by Abraham and Marsden on the "*Foundation of Mechanics*" [AM08] and the monograph by Wonham towards a geometric control theory [Won79]. Still, a lot of structure needs to be understood, especially regarding feedback [Lew14].

The same search for structure is heavily present in optimization. Nesterov writes: "*The main lesson of the development of our field in the last few decades is that efficient optimization methods can be developed only by intelligently employing the structure of particular instances of problems*" [Nes18, p. viii-ix].

In fact, he continues by writing that: "*In order to do this, it is always useful to look at*

[1] https://dictionary.cambridge.org/dictionary/english/stability

successful examples" [Nes18, p. ix]. This is precisely the aim of this thesis, where judging successfulness is left to the reader.

Indeed, exploiting structure is common in all of engineering and applied mathematics—science perhaps, yet, the aim of this thesis is to show that structure should be *derived*, not imposed. Linearity and convexity are convenient structural assumptions, but both are rarely found in nature. The *Control for Societal-scale Challenges: Road Map 2030* by the IEEE Control Systems Society also contains several pointers in this direction. To uncover the structure of our problems we must work with the fields we hope to apply our theory to: "*Just because you can formulate and solve an optimization problem does not mean that you have the correct or best cost function.*" [AJP23, p. 32]. Optimality is only as interesting as the objective is relevant.

In this thesis we study the interplay between *structure* and *desiderata*. We study ramifications of imposing structural assumptions:

$$\text{structure} \implies \text{desiderata}. \tag{1}$$

Indeed, (1) is a common direction of study, *e.g.*, given a controllable linear system $\dot{x} = Ax + Bu$ (structure), there is ($\implies$) a continuous feedback exponentially stabilizing the origin (desiderata). We think of (1) as a sufficient structural condition, or as the constructive direction.

The other direction is less frequently studied, but perhaps of more fundamental importance, that is, we study structural ramifications of certain desiderata:

$$\text{structure} \impliedby \text{desiderata}. \tag{2}$$

Condition (2) should be understood as necessary (structure). In that sense, (2) is frequently obstructive, *e.g.*, the desire of some set being a global attractor of some space puts ($\impliedby$) hard topological constraints on the underlying space, possibly contradicting and thereby obstructing the initial desire (desiderata).

Closing the gap between these sufficient and necessary (deliberately put in that order) conditions is of great interest towards a principled control theory. We hope to contribute towards this goal, may it be so slightly.

Acknowledgements

This thesis is the accumulation of being around the right people at the right time.

Largely due to Prof. Robert Babuška his passionate delivery of our undergraduate course on control theory I fell in love with the subject. It must be said that the same is probably true for hundreds of students in Delft. My Master's studies at the Delft Center for Systems and Control (DCSC) only propelled this interest, again, thanks to outstanding teachers. The continuous support of both Prof. Peyman Mohajerin Esfahani and Prof. Jan H. van Schuppen is kindly acknowledged. Continuing my studies in Lausanne, I thank the entire Risk Analytics and Optimization (RAO) team for the warm welcome. The mentorship of Dr. Tobias Sutter (now, Prof.) was outstanding and instrumental to this thesis. The PhD program is not only about writing (emails, papers, ...), but much more about thinking. What are fundamental open problems and why? What to aspire for? What is good science? And so forth. It is the difference between *doing* research

and *being* a researcher. Besides what it means to do rigorous research, to communicate research and how to teach, I finally learned how to be a researcher from my advisor Prof. Daniel Kuhn. I will be forever grateful for this invaluable experience. Besides the RAO team I was part of a second academic family: the National Centre of Competence in Research (NCCR) Automation. The established community is remarkable and I thank the entire team for their fantastic work and in particular the Laboratoire d'Automatique (LA) and the Institut für Automatik (IfA) for their open doors. Outside of Switzerland, I was fortunate enough to be able to work with Dr. Emmanuel Moulay and Dr. Man-Chung Yue. Their trust is greatly acknowledged. I am also grateful to the Marianne Bernadotte Centrum, Karolinska Institutet, for their exceptional hospitality. At last, I want to thank the committee members: Prof. Aude Billard, Prof. Florian Dörfler, Prof. Giancarlo Ferrari Trecate and Prof. Mikael Johansson, it was an honour.

Throughout all of this, the support of friends and family always meant the world to me, although I could have mentioned this more frequently. In particular, I could not have done this without you Esra. Senin aşkın benim enerjimdir.

A special thanks to my father, Theo, for finding a few mistakes in a preprint of this thesis. De appel valt niet ver van de boom.

I will refrain now from making this a "look who I know list" and rather propose to thank everyone in person, sooner or later, with a beer, or two.

The future is exciting and I consider myself lucky to be part of this scientific community.

Lausanne, Switzerland, *Wouter Jongeneel*
April 6, 2024

The thesis is written using LaTeX, with all figures being made using Inkscape with TexText[2].

Bibliography

[AJP23] A M Annaswamy, K H Johansson, and G J (eds.) Pappas. Control for societal-scale challenges: Road map 2030. *IEEE Control Systems Society*, 2023.

[AM08] R Abraham and J E Marsden. *Foundations of Mechanics.* American Mathematical Society, Providence, 2008.

[KFA69] R E Kalman, P L Falb, and M A Arbib. *Topics in Mathematical System Theory.* McGraw-Hill, New York, 1969.

[Lew14] A D Lewis. *Tautological Control Systems.* Springer, Cham, 2014.

[Nes18] Y Nesterov. *Lectures on Convex Optimization.* Springer, Cham, 2018.

[Won79] W M Wonham. *Linear Multivariable Control.* Springer, New York, 1979.

[2] See https://inkscape.org/ and https://textext.github.io/textext/.

Prologue

> “*The tantalizing possibility suggested by entropy ... is that there may be other “little somethings” around us the mathematical beauty of which we still fail to recognize because we see them in a curved mirror of our preconceptions*”
>
> — Gromov [Gro12, p. 11].

What have bees, starlings and white-spotted pufferfish in common? They all display a remarkable sense of *structure*. Bees construct honeycomb tiling that are isoperimetrically optimal in 2D [Hal01] (not in general [Tót64]), a flock of starlings create unprecedented collision-free dynamic sculptures in the air by just taking 6 to 7 neighbours into account [YSC+13] and the male white-spotted pufferfish creates a beautiful sand-sculpture to attract the attention of a female [Mat15].

The word “*structure*” is deliberately chosen over words like “*geometry*” and “*order*” as we believe that geometry is rather a manifestation of order, and order itself is some particular arrangement, whereas structure aims to capture the *raison d'être* of this arrangement within a system. The word “*principles*” could have been suitable, indeed, Newton was looking for the structure of nature.

The belief that *all* of nature is well-structured, and not just Newton's mechanics, was propelled through, for instance, the 1917 book by Thompson towards a mathematical biology: “*the flower for the bee, the berry for the bird*” [Tho17, p. 3]. This belief has never ceased to amaze.

Unsurprisingly, mathematics plays a key role in discovering and describing the structure of nature. Now, if you believe that mathematics itself is part of nature, in that it is discovered and not created, then we can argue that it must display inherent structure, beauty [Har92] or elegance if you like.

With this in mind, there is nothing else we could do, but looking for this structure.

Bibliography

[Gro12] M Gromov. In a search for a structure, part 1: On entropy. In *European Congress of Mathematics*, pages 51–78, 2012.

[Hal01] T C Hales. The honeycomb conjecture. *Discrete Comput. Geom.*, 25:1–22, 2001.

[Har92] G H Hardy. *A Mathematician's Apology.* Cambridge University Press, Cambridge, 1992.

[Mat15] K Matsuura. A new pufferfish of the genus Torquigener that builds "mystery circles" on sandy bottoms in the Ryukyu Islands, Japan (Actinopterygii: Tetraodontiformes: Tetraodontidae). *Ichthyol. Res.*, 62:207–212, 2015.

[Tho17] D'A W Thompson. *On Growth and Form.* Cambridge University Press, Cambridge, 1917.

[Tót64] L F Tóth. What the bees know and what they do not know. *Bull. Amer. Math. Soc.*, 70(6):468–481, 1964.

[YSC+13] G F Young, L Scardovi, A Cavagna, I Giardina, and N E Leonard. Starling flock networks manage uncertainty in consensus at low cost. *PLoS Comput. Biol.*, 9(1):e1002894, 2013.

Abstract

In this thesis we study *stability* from several viewpoints. After covering the practical importance, the rich history and the ever-growing list of manifestations of stability, we study the following.

(i) (Statistical identification of *stable* dynamical systems): Understanding stability of identified systems is of great practical- and theoretical importance. Even the simplest case, that of characterizing spectral properties of the least-squares estimator of a linear dynamical system has been largely open. To that end, we propose a principled method for projecting a system matrix to the nonconvex set of Schur stable matrices. Leveraging large deviations theory, we show that this projection is optimal in an information-theoretic sense and that the projection can be approximated, up to arbitrary precision, by simply adding a feedback term corresponding to the optimal gain matrix of a linear quadratic regulator problem. The estimator resulting through this projection is constructed from a single trajectory of state measurements, is guaranteed to be stable and offers non-asymptotic statistical bounds on the estimation error.

Going one step beyond stability, we further exploit large deviations theory to identify the topological class of an unknown stable system, again from a single trajectory of data. We prove that the probability of misclassification decays exponentially with the number of samples at a rate that is proportional to the square of the smallest singular value of the unknown matrix.

(ii) (Spaces of *stable* dynamical systems): Various technical domains are aided by better understanding spaces of stable dynamical systems. For instance, in many identification- and optimal control problems we effectively try to optimize over such a space and thus understanding its properties is important.

First, we study the linear quadratic regulator problem and finally provide an operational meaning for cross terms in the stage cost. In particular, we show that the topological class of the closed-loop system is invariant under a change of the stage cost, as long as no cross term is introduced, that is, when restricting system matrices to the general linear group, closed-loop matrices can jump to the opposite path-connected component of this group if and only if a particular cross term is introduced. Hence, formally speaking, one

can only “tune” the closed-loop behaviour by introducing a cross term.

Secondly, motivated by the learning community often employing convex Lyapunov functions to obtain stability certificates, we study the ramifications of the convexity assumption. We show that continuous dynamical systems, on Euclidean space, equipped with convex Lyapunov functions, asserting that the origin is globally asymptotically stable, can always be homotoped to each other such that along this homotopy stability is preserved. This means that the space of those dynamical systems is path-connected, which in its turn leads to obstructions and the necessity of rethinking convexity assumptions.

(iii) (Numerical *stability*): Besides stability of attractors and structural stability under perturbations, which are all classical topics of study, we also look at stability of the implementation of certain dynamical systems. Specifically, we look at zeroth-order optimization algorithms, a widely applicable class of algorithms understood as discrete-time dynamical systems on Euclidean space.

Most zeroth-order optimization algorithms mimic a first-order algorithm, that is, a discretized gradient flow, but replace the gradient of the objective with some estimator that can be computed from a number of function evaluations. This estimator is typically constructed randomly, and its expectation matches the gradient of a smooth approximation of the objective whose quality improves as some underlying smoothing parameter, usually a finite-difference parameter, is reduced. As such, most zeroth-order algorithms require this smoothing parameter to decay to zero as the algorithm proceeds. While estimators based on just a single function evaluation can be obtained via Stokes’ theorem, their variance is unbounded. Then, estimators based on multiple function evaluations, overcome the exploding variance, yet, they suffer from numerical cancellation once the smoothing parameter is sufficiently small. To combat both effects simultaneously, we extend the objective function to the complex domain and leverage the complex-step derivative to construct a new randomized estimator. This new estimator is immune to cancellation as it requires only one function evaluation, in addition, its variance remains bounded. We prove that zeroth-order algorithms that use our estimator offer the same theoretical convergence guarantees as the state-of-the-art methods. At the cost of complex lifting, our results remain true after implementation, the algorithm is numerically stable.

(iv) (Topological obstructions to stability and stabilization): Letting go of stylized structural assumptions, we study necessary conditions for stability and stabilizability of dynamical control system defined on topological spaces. We provide new insights in multistability and odd-dimensional dynamical systems, generalize the obstruction for continuous feedback to globally asymptotically stabilize a point on a fiber bundle, to the stabilization of embedded submanifolds and we fully characterize when a compact attractor is a strong deformation retract of its domain of attraction. Results of this nature display the synergy between topology and dynamical systems.

All of these studies focus on understanding the interplay between underlying structure and desiderata. Naturally emerging future directions close the thesis.

Résumé

Dans cette thèse, nous étudions la *stabilité* sous plusieurs angles. Après avoir couvert l'importance pratique, l'histoire riche et la liste toujours croissante des manifestations de la stabilité, nous étudions ce qui suit.

(i) (Identification statistique des systèmes dynamiques *stables*) : Comprendre la stabilité des systèmes identifiés revêt une grande importance pratique et théorique. Même dans le cas le plus simple, celui de la caractérisation des propriétés spectrales de l'estimateur des moindres carrés d'un système dynamique linéaire, la question reste largement ouverte. À cette fin, nous proposons une méthode fondée pour projeter une matrice système sur l'ensemble non convexe des matrices Schur. En utilisant la théorie des Large Deviations, nous montrons que cette projection est optimale d'un point de vue informationnel et que la projection peut être approximée, jusqu'à une précision arbitraire, en ajoutant simplement un terme de rétroaction correspondant à la matrice de gain optimale d'un problème de LQR. L'estimateur résultant de cette projection est construit à partir d'une seule trajectoire de mesures d'état, est garanti d'être stable et offre des bornes statistiques non asymptotiques sur l'erreur d'estimation. Allant au-delà de la stabilité, nous exploitons davantage la théorie des Large Deviations pour identifier la classe topologique d'un système stable inconnu, encore une fois à partir d'une seule trajectoire de données. Nous prouvons que la probabilité de mauvaise classification décroît exponentiellement avec le nombre d'échantillons à un taux proportionnel au carré de la plus petite valeur singulière de la matrice.

(ii) (Espaces de systèmes dynamiques *stables*) : Divers domaines techniques bénéficient d'une meilleure compréhension des espaces de systèmes dynamiques stables. Tout d'abord, nous étudions le problème du LQR et fournissons enfin une signification opérationnelle pour les termes croisés dans le coût de l'étape. En particulier, nous montrons que la classe topologique du système en boucle fermée est invariante en cas de changement du coût de l'étape, tant qu'aucun terme croisé n'est introduit. Ainsi, formellement parlant, on ne peut "ajuster" le comportement en boucle fermée qu'en introduisant un terme croisé. Deuxièmement, motivés par la communauté de Learning Theory utilisant souvent des fonctions de Lyapunov convexes pour obtenir des certificats de stabilité, nous étudions les

ramifications de l'hypothèse de convexité. Nous montrons que les systèmes dynamiques continus, sur l'espace Euclidien, équipés de fonctions de Lyapunov convexes, affirmant que l'origine est globalement asymptotiquement stable, peuvent toujours être homotopiques les uns aux autres de telle sorte que la stabilité est préservée le long de cette homotopie. Cela signifie que l'espace de ces systèmes dynamiques est path-connexe, ce qui entraîne des obstacles et la nécessité de repenser les hypothèses de convexité.

(iii) (Stabilité numérique) : Nous examinons également la stabilité de la mise en oeuvre de certains systèmes dynamiques. Nous examinons spécifiquement les algorithmes d'optimisation d'ordre zéro (Z-O), une classe largement applicable d'algorithmes compris comme des systèmes dynamiques en temps discret sur l'espace Euclidien. La plupart des algorithmes d'optimisation Z-O imitent un algorithme du premier ordre, mais remplacent le gradient de l'objectif par un estimateur qui peut être calculé à partir d'un certain nombre d'évaluations de la fonction. Cet estimateur est généralement construit de manière random, et son expectation correspond au gradient d'une approximation lisse de l'objectif dont la qualité s'améliore à mesure qu'un paramètre de lissage sous-jacent, généralement un paramètre de différence finie, est réduit. En tant que tels, la plupart des algorithmes Z-O nécessitent que ce paramètre de lissage décroisse à zéro à mesure que l'algorithme progresse. Alors que les estimateurs basés sur une seule évaluation de fonction peuvent être obtenus via le théorème de Stokes, leur variance est illimitée. Ensuite, les estimateurs basés sur plusieurs évaluations de fonction surmontent la variance explosive, mais ils souffrent d'une annulation numérique une fois que le paramètre de lissage est suffisamment petit. Pour lutter contre les deux effets simultanément, nous étendons la fonction objective au domaine complexe et exploitons la dérivée de complex-step pour construire un nouvel estimateur aléatoire. Cet nouvel estimateur est immunisé contre l'annulation car il ne nécessite qu'une seule évaluation de fonction, en outre, sa variance reste bornée. Nous prouvons que les algorithmes Z-O qui utilisent notre estimateur offrent les mêmes garanties de convergence théorique que les méthodes de pointe.

(iv) (Obstructions topologiques à la stabilité et à la stabilisation) : En abandonnant les hypothèses stylisées de structure, nous étudions les conditions nécessaires à la stabilité et à la stabilisabilité des systèmes de contrôle dynamique définis sur des espaces topologiques. Nous apportons de nouvelles perspectives sur la multistabilité et les systèmes dynamiques de dimension impaire, généralisons l'obstruction à la rétroaction continue pour stabiliser globalement asymptotiquement un point sur un fibré, à la stabilisation de submanifolds et caractérisons entièrement quand un attractor compact est une strong deformation retract de son domaine d'attraction. Des résultats de cette nature mettent en évidence la synergie entre la topologie et les systèmes dynamiques.

Toutes ces études se concentrent sur la compréhension de l'interaction entre la structure sous-jacente et les desiderata. Les orientations futures qui émergent naturellement concluent la thèse.

Contents

1

Introduction

> “*The procedure ordinarily used consists in neglecting, in the differential equations … all the terms of higher than first order … The only attempt, as far as I know, at a rigorous solution belongs to Poincaré, who, in the remarkable memoir … ‘Sur les courbes definies par les equations differentielles’ … considered questions of stability for the case of second order systems … the methods he used allow much more general applications and could still lead to many new results. This will be seen in what follows, for, in a large part of my researches, I was guided by the ideas developed in the above-mentioned memoir.*”
>
> — Lyapunov [Lia92] [Lya92, p. 531–532].

Early in the 1969 book by Kalman, Falb and Arbib one finds the following: “*The notion of a dynamical system as just defined is far too general. Such a definition is needed to set up terminology, to analyze and refine concepts, and to perceive unity in a diversity of applications, but it is not sufficiently coherent to bear a large array of deep mathematical theorems or useful practical deductions. To get good theorems and interesting applications, we must particularize and impose additional structure.*” [KFA69, p. 6-7]. We agree with them that the right level of generality is important to derive anything interesting, but we also believe that *imposing* structure should be done with utmost care and ideally, structure is not imposed but *derived*.

1.1 Impetus

The pivotal structural breakthrough in control theory during the last century was the work of Kalman "*On the General Theory of Control Systems*" [Kal60]. Inspired by Shannon and challenged by technological advances, Kalman set out to study when a plant is "*unalterable*", that is, to study *controllability*. Interestingly, as he writes it, the Kalman filter is somewhat of a by-product, due to duality. It is well-known that Kalman succeeds in providing a rather general theory of control systems, in the *linear* case. Both reviewers of [Kal60] comment on reality being nonlinear, however. Kalman responds by stating that the paper does not pretend being that general, but nevertheless, he writes: "*The remarks of ..., that the paper is of importance only for linear systems, are, in the final analysis correct.*" [Kal60, p. 492]. As we know now, controllability has far weaker ramifications in the nonlinear case.

Example 1.1.1 (Controllability and continuous feedback)**.** It is well-known that for a linear control system of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = Ax(t) + Bu \tag{1.1.1}$$

we can select a *continuous* feedback $x \mapsto \mu(x)$ for the input u such that the origin $0 \in \mathbb{R}^n$ is *globally asymptotically stable* when the pair (A, B) is *controllable*. Indeed, μ can be chosen to be simply linear, which follows from a standard canonical transformation and pole-placement argument [Won79, TSH01]. Going against populair belief, in the late 1970s, Jurdjevic and Quinn constructed an example of a controllable system that cannot be stabilized using differentiable feedback [JQ78]. A year later, Sussmann provided an example of a controllable system that cannot even be stabilized using continuous feedback [Sus79]. After further investigations by Sontag and Sussmann [SS80], it was Brockett who provided a general topological condition that must hold for continuous feedback to exist [Bro83]. It can be argued that his condition follows from earlier work by Krasnosel'skiĭ and Zabreiko [KZ84], see also [Zab89] and [JM23]. Simply put, for a continuous control system of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t), u) \tag{1.1.2}$$

with $f(0,0) = 0$ and $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, one must have that $f(0,0) \in \operatorname{int} f(U)$ for any open neighbourhood U of $(0,0)$, for continuous feedback to exist that *locally*, asymptotically, stabilizes the origin. Indeed, various controllable systems fail to satisfy this condition and hence, we see that controllability is *not* a sufficient structural condition for the existence of continuous feedback.

Next, we illustrate *necessary* topological structure for global, continuous, asymptotic stabilization.

Example 1.1.2 (Stabilization on the n-torus)**.** Suppose we have a continuous control system (the precise details are irrelevant at this point) defined over an n-torus $\mathbb{T}^n \simeq_t \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ and we are tasked with finding a continuous control law that globally,

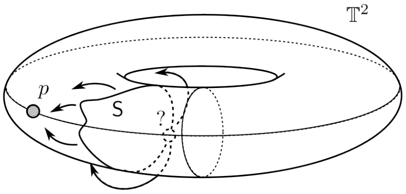


Figure 1.1: The 2-torus and the set S from Example 1.1.2.

asymptotically, stabilizes a point p on $\mathbb{T}^n$. Regardless of the control system, this means there must be a continuous (closed-loop) dynamical system, say, a vector field, with these qualitative properties. It is somewhat evident that this fails for $n = 1$, since we need to "rip apart" the circle $\mathbb{S}^1 \simeq_t \mathbb{T}^1$, see [JM23, Fig. 1.1]. However, perhaps for $n > 1$ we have sufficient degrees of freedom? To that end, consider a circle S, embedded in $\mathbb{T}^2$, that is not homotopic to a point, which always exists since $\pi_1(\mathbb{T}^2) \simeq \mathbb{Z}^2$ ($\pi_1(\cdot)$ being the fundamental group). Since we demand the closed-loop system to be continuous, we need to be able to continuously deform S into a point, which is obstructed precisely by S not being homotopic to a point, see Figure 1.1. As $\pi_1(\mathbb{T}^n) \simeq \mathbb{Z}^n$ this intuition extends for any $n \in \mathbb{N}_{>0}$. This obstruction is purely topological as it can be shown that the domain of attraction of a point that is globally asymptotically stable (under some continuous dynamical system), is contractible [Son98, Thm. 21]. Indeed, for a contractible set X (with X being a topological space), we have that $\pi_1(\mathsf{X}) \simeq 0$.

Recall from the preface that we are after understanding the interplay between structure and desiderata. Now see that Example 1.1.1 and Example 1.1.2 allude to a rather large gap between sufficient and necessary conditions for continuous feedback to exist (linear control system vs. a contractible state space). It is also evident that solving this problem (closing the gap) in full generality is somewhat futile, in line with the comment from Kalman, Falb and Arbib. To that end, we are inspired by the examples above to better understand structural properties of stability and stabilizability problems.

We are not only looking at questions from systems- and control theory, but also at optimization. There, an interesting question is to understand "practically relevant[1]" necessary conditions on the objective for gradient descent to converge globally. Although we will only study sufficient conditions for some optimization algorithm to convergence, see Chapter 5, we highlight one of the most powerful and relevant flavours of invexity.

Example 1.1.3 (Global optimization under a Polyak-Łojasiewicz condition)**.** A function $f \in C^1(\mathbb{R}^n;\mathbb{R})$ is said to satisfy a μ-Polyak-Łojasiewicz (PL) condition when there is a $\mu > 0$ such that

$$\tfrac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f(x^\star)) \quad \forall x \in \mathbb{R}^n, \tag{1.1.3}$$

where $x^\star$ denotes any critical point of f, that is, $\nabla f(x^\star) = 0$. Now suppose we want to solve

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x). \tag{1.1.4}$$

[1] The notion of "*invexity*" is appropriate to mention here, this property is necessary but also quickly meaningless as little structure is provided.

Since f only satisfies a μ-PL condition, we cannot exploit convexity. However, consider the dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -\nabla f(x(t)) \tag{1.1.5}$$

and the Lyapunov function (candidate) $V(x,t) = e^{2\mu t}(f(x) - f(x^\star))$. It is evident that V is positive and only zero at a critical point. Now see that by the μ-PL condition (1.1.3) we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}V(x(t),t) = 2\mu e^{2\mu t}(f(x(t)) - f(x^\star)) - e^{2\mu t}\|\nabla f(x(t))\|_2^2 \leq 0. \tag{1.1.6}$$

Hence, since $V(x(t),t) = V(x(0),0) + \int_0^t \dot{V}(x(s),s)\mathrm{d}s \leq V(x(0),0)$ we have that $f(x(t)) - f(x^\star) \leq e^{-2\mu t}f(x(0)) - f(x^\star)$. A more direct argument employs the Lyapunov function $V'(x) = f(x) - f(x^\star)$, leading to $\dot{V}' \leq -2\mu V'$, and by *Grönwall's inequality* to $V(x(t)) \leq e^{-2\mu t}V(x(0))$. Regardless, this construction serves as a motivation for the development of gradient descent algorithms to solve (1.1.4), without appealing to convexity arguments. Indeed, this PL-*structure* is frequently key in convergence of gradient descent applied to nonconvex problems [FGKM18]. See [Wil18] for more on Lyapunov techniques in the context of optimization.

The Examples from above have in common that interesting remarks can be made after looking at the *right* structure, *e.g.*, linearizing a nonlinear problem is usually not revealing (or preserving) the right structure. Focusing on precisely the right structure is something the main figure of the next section was widely praised for.

1.2 Historical comments

Space does not permit to cover the entire history of stability theory and to some extent, this has been done before, *e.g.*, see [Kel15, LP17] or see [JM23] and especially the references therein. We aim to highlight in this section how several notions of stability emerged in the context of control and we try to look slightly beyond the usual suspects. Here, we focus on how the 1960s could have been so fruitful, whereas other and more detailed accounts are presented in the appropriate chapters.

If we reconsider the definition of stability from the dictionary, then one might say that stability (in the context of differential equations) started with understanding local flows, that is, solutions depending continuously on initial conditions. However, we want to guarantee a notion of stability for *all* time. Inspired by Poincaré's work on the qualitative theory of differential equations, Aleksandr M. ***Lyapunov*** (1857-1918) pioneered this study of the *stability of motion*, although he was not the first. Motivated by questions from celestial mechanics, Lagrange, Dirichlet, Poisson, Poincaré and many of their contemporaries all studied stability (of the solar system). We point out that some stability notions with their name attached to it are in fact different from what they proposed themselves, *e.g.*, Lagrange studied a particular notion of convergence and not necessarily bounded solutions [LP17].

Going back to Lyapunov, regarding his notion of stability, he finally introduced the $\varepsilon - \delta$ notion of stability—nowadays known as *Lyapunov stability*, in the right order, that is, he requested that for each ε-neighbourhood of an equilibrium point there is a δ-neighbourhood (of initial conditions) such that the solutions starting from the δ-neighbourhood are contained in the ε-neighbourhood. Previous stability conditions required one to find an ε given a δ. Of course, then he went on to define his *first* (indirect, via linearization) and *second method* (direct, via Lyapunov functions) to study stability. In fact, he set up the *convention* of Lyapunov functions V being positive and introduced the notion of *asymptotic* stability. See [LP17] for further comments on how we eventually moved from asymptotic stability *in the large* (non-trivial region of attraction different from the entire space) and asymptotic stability *in the whole* (region of attraction is the entire space) to *global* asymptotic stability.

However, it took several generations of researchers to revitalize and then formalize Lyapunov's work. Regarding the revitalization, we largely credit Nikolay G. ***Chetaev*** (1902–1959), the doctoral advisor of Nikolai N. ***Krasovskii*** (1924–2012). Krasovskii himself was responsible for the formalization of a significant amount of Lyapunov stability theory. In particular, working with Chetaev, Barbashin and Pontryagin, he developed converse theory (different from Massera), invariance principles and stochastic notions of stability. A telling anecdote is the following, at the first IFAC world congress in Moscow (1960), foreign participants asked Barbashin, who gave a talk on behalf of Krasovskii, if "N. N. Krasovskii" was a group (we suppose similar to N. Bourbaki at the time) [KS15].

Besides aforementioned mentors, Krasovskii was also influenced by the work of José L. ***Massera*** (1915–2002), from Uruguay, whom, as early as 1949, pioneered Lyapunov (converse) theory [Mas49], with further elaborations in 1956 [Mas56]. It is due to him that we understand the relation between asymptotic- and equiasymptotic stability, have comparison functions (also due to Hahn [Hah67]) and frequently appeal to integral representation of candidate Lyapunov functions. Despite highly-influential work early on, Massera was faced with long political imprisonment and we can only wonder what would have been otherwise. We do mention that Massera, in his turn, credits Khalikoff, Malkin, Marachkoff and Persidskiĭ. With Persidskiĭ providing the first converse result and especially Malkin being important and influential regarding the focus on *uniformity.*

Despite the pivotal work in the East and contributions by Massera, it took a while before *their* work crossed the ocean and found its way in the West.

Solomon ***Lefschetz*** (1884–1972), an engineer turned mathematician, in part due to tragically losing both of his hands, was a pioneer in both topology and dynamical systems and it is largely due to him that we now have a mature stability theory (in the West).

Although retiring in 1953 from Princeton, his most influential work—that is, for this thesis, had yet to come. Lefschetz, as being born in Moscow, was aware of the Russian work on differential equations and with the advent of Sputnik (1957) he was convinced the rest of the world was lagging behind and something had to be done [Lef70]. He soon established a center for study and research on differential equations at the Research Institute for Advanced Studies (RIAS) in Baltimore, housed at the Martin Company (now Lockheed Martin). This center flourished and eventually transformed—since the Martin company wanted to focus on applications—in 1964 into the (Lefschetz) center for dynamical systems at Brown [CMP10], initially directed by LaSalle.

Joseph P. ***LaSalle*** (1916-1983) was asked by Lefschetz to join the center at the RIAS in 1958. Ten years prior to this request, LaSalle met Lefschetz (and Bellman) at Princeton and developed an interest in differential equations. Soon after joining the RIAS, LaSalle wrote one of the most influential texts on Lyapunov stability theory together with Lefschetz [LSL61]. Not only that, he established his[2] invariance principle [LaS60], he initiated the SIAM Journal on Control and he eventually found the Journal on Differential Equations [BHJ83] with Hale (see below).

Besides LaSalle, Lefschetz also found support in Lamberto ***Cesari*** (1910–1990), an expert on the calculus of variations and the author of several influential books, most notably his text on dynamical systems [Ces71], with the first edition appearing in 1959, praised by Lefschetz for a good overview of literature in the East.

Not only Cesari, but in particular one of his students: Jack K. ***Hale*** (1928–2009) was attracted to the center. Thanks to a recommendation from Cesari, Hale joined the center in 1958, after 4 years working in industry, he could finally work on his true passion again [Cha00]. Together with Lefschetz and LaSalle, Hale was instrumental in the development of dynamical systems theory in the West. Initially influenced by Krasovskii, he became an expert on the qualitative theory of infinite-dimensional dynamical systems (*e.g.*, due to time-delays and PDEs) and eventually replaced LaSalle as the director of the Lefschetz center for dynamical systems [CMP10].

The final member of the center at the RIAS was Rudolf E. ***Kalman*** (1930–2016), who joined after completing his doctoral studies in 1957. In the early 60s, while being at this center, Kalman developed all his groundbreaking work (the Kalman filter, the popularization of state space methods, minimal realization theory and an accessible exposition of Lyapunov stability theory). For more on Kalman developing his filter, the impact of the first IFAC world congress in Moscow and Kalman his relation to East, see [Ste11].

With all these names at the RIAS, it is hardly a surprise the center flourished, as we wrote before. This continued long after moving to Brown, attracting the likes of Fleming, Kushner, Wonham, and many others. It can also be argued that frequency domain methods from the 50s were pushed to the background in part due to the work at the RIAS [Bro14].

Lefschetz only established this center after retiring. Before doing so, he worked on algebraic geometry and later algebraic topology, both very successfully, he was praised for the right level of generality, see [Hod73] for an overview. Detailing his work here is too much of a detour, but we like to point out that in his topological work there is a strong link to dynamical systems. Namely, his fixed-point theorem is key in understanding the interplay between topology and dynamical systems, *e.g.*, see [GP10].

Before closing this section, we highlight one additional person. Lefschetz and LaSalle wrote in the preface of their 1961 book on Lyapunov stability theory the following: "*We are especially pleased that our book appears in a series inspired and edited by an old and cherished friend, Richard Bellman.*" [LSL61, p. vi]. Richard E. ***Bellman*** (1920–1984) will not be unfamiliar to the typical reader, but what is less well-known, is that before his seminal work on dynamic programming, he worked with no other than Lefschetz on stability theory. In fact, he wrote one of the first English, pedagogical, texts on stability theory, published in 1953 [Bel53]. Throughout the text, the influence of Poincaré and Lyapunov is acknowledged and Bellman emphasizes the importance of better understanding periodic systems, 70 years later still very much an open problem. It needs to be said,

[2] Although the principle was also discovered by Barbashin and Krasovskii.

however, that his motivation for writing this text is less romantic, he wanted to quickly obtain the PhD degree and finally work on analytic number theory [Bel84, p. 110-111].

We believe that the crux of this section is twofold, first, good research is a team effort and a vibrant research center greatly contributes, secondly, international contacts are crucial and isolation is to be avoided, which is unfortunately (again) far from obvious these days.

Control and optimization

In this thesis we discuss stability problems in the areas of *both* control and optimization. A somewhat naïve point of view would be that if you control a system, you better do it optimally, hence the control theorist must be a part-time optimizer. We believe there is more to this. Especially Lyapunov theory appears to provide for a fruitful bridge between the two fields.

To provide a miniscule bit of historical context. The "*Journal of the Society for Industrial and Applied Mathematics Series A Control*" was established in 1962 and the opening text by McMillan emphasizes the importance of bringing engineers and mathematicians together. Control theory is there to bring order to the chaos, to aid in design, "*Theory creates understanding.*" [McM62]. In 1966 the journal became the "*SIAM Journal on Control*" and in 1976 the name was again changed, this time to the "*SIAM Journal on Control and Optimization*". Mixing control and optimization was—and still is—classical, mainly due to the prominent role of mechanics and the calculus of variations at that time [Bry96, SW97]. Going far beyond optimal control, we owe some of this synergy to Brockett [BC98, HM12], who also featured in the first edition of the SIAM Journal on Control and Optimization [BF76].

At the time of writing this thesis, the ETH Zürich just opened a faculty position on "*algorithmic system theory*", the synergy continues.

1.3 Outline of the thesis

In this thesis we focus on addressing the following problems, as illustrated in Figure 1.2. Before the beginning (Ch. 2), we review mathematical stability theory. Then, first (Ch. 3), given an unknown linear dynamical system that is known to be *stable*, what should one do if an estimator, constructed from data, is *unstable*? This is non-trivial since collecting more data will only help asymptotically and the space of stable matrices is non-convex (so, a projection is not obvious). Secondly (Ch. 4), what are the ramifications of imposing a convex structure on Lyapunov functions? This assumption is populair in the learning community, but so far it is unclear how the space of dynamical systems is constrained by doing so. Then, third (Ch. 5), the vast majority of zeroth-order optimization algorithms need a "difference parameter" δ to vanish for these algorithms to provably work, but can this ever be implemented in a numerically stable way? At last (Ch. 6), what kind of structure is imposed on the domain of attraction of a compact attractor? This is important for stabilization tasks, where we usually start from a space and ask if it admits a certain global attractor. After the ending (Ch. 7), we provide commentary on future directions of research.

More specifically, after having set the stage in the preface and introduction, we provide an overview of stability theory in Chapter 2, with the emphasis on the topological and qualitative viewpoint.

Towards a principled data-driven control, in Chapter 3 we study the identification of Schur stable linear dynamical systems with a stability certificate and finite-sample guarantees (data-driven stability problem). We show how the large deviations rate function (derived structure) of the least squares estimator naturally leads to a tractable, statistically consistent, identification scheme that *guarantees* stability (desiderata). Chapter 3 is based (almost verbatim) on:

(3-i) Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. "*Efficient Learning of a Linear Dynamical System With Stability Guarantees*". IEEE Transactions on Automatic Control 68.5 (2023), pp. 2790–2804. doi: 10.1109/TAC.2022.3213770.

(3-ii) Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. "*Topological Linear System Identification via Moderate Deviations Theory*". IEEE Control Systems Letters 6 (2022), pp. 307–312. doi: 10.1109/LCSYS.2021.3072814.

Most control problems do not come with a natural optimization objective. In fact, optimization is often merely a principled means to arrive at something "*reasonable*" [SL12]. In Chapter 4 we study ramifications of common selections of the objective in the linear quadratic regulator (LQR) problem (imposed structure). We show that the topological class of the closed-loop system can only be altered by introducing cross-terms (between state and input) in the objective. Secondly, we discuss the ramifications of assuming (control) Lyapunov functions to be convex (imposed structure). Chapter 4 is based (almost verbatim) on:

(4-i) Wouter Jongeneel and Daniel Kuhn. "*On Topological Equivalence in Linear Quadratic Optimal Control*". European Control Conference (ECC). 2021, pp. 2002–2007. doi: 10.23919/ECC54610.2021.9654863.

(4-ii) Wouter Jongeneel and Roland Schwan. "*On Continuation and Convex Lyapunov Functions*". IEEE Transactions on Automatic Control (2024), pp. 1–12. doi: 10.1109/TAC.2024.3381913.

Regarding zeroth-order optimization, we show in Chapter 5 how the real analytic structure, present in most smooth objective functions (natural structure), can be exploited to derive a numerically stable randomized gradient estimator (desiderata). Chapter 5 is based (almost verbatim) on:

(5-i) Wouter Jongeneel, Man-Chung Yue, and Daniel Kuhn. "*Small errors in random zeroth-order optimization are imaginary*". SIAM Journal on Optimization (2024), pp. 1-32, in press. arXiv: 2103.05478.

(5-ii) Wouter Jongeneel. "*Imaginary Zeroth-Order Optimization*". (2021). arXiv: 2112.07488.

In Chapter 6 we contribute necessary conditions for stabilizing continuous feedback to exist (from desiderata to structure). We discuss multistability, obstructions to submanifold stabilization and we charachterize the homotopy type of the domain of attraction for compact attractors. Chapter 6 is based on:

(6-i) Wouter Jongeneel and Emmanuel Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, 2023. doi: 10.1007/978-3-031-30133-9.

(6-ii) Wouter Jongeneel. "*On topological properties of compact attractors on Hausdorff spaces*". European Control Conference (ECC). 2024, pp. 1–6, in press. arXiv: 2301.05932.

We close the thesis in Chapter 7, with commentary on future research.

We remark that parts of both Chapter 2 and Chapter 7 are also based on the aforementioned prints.

Comment on simulations Despite some work on (numerical) algorithms, in this thesis we try to stay away from arguments that extrapolate inherently discrete simulation results, insights should be obvious or proven. We do agree that simulations can be greatly helpful, *e.g.*, to visualize and gain intuition, but especially with stability in mind, we cannot rely on finite realizations to conclude. In fact, some of our work is the result of phenomena being attributed to simulation errors while in fact they were singularities embedded in the problem.

Comment on notation We introduce notation throughout the thesis and repeat this whenever convenient. We strive for consistency and standard notation, but already admit that ":=" and "=:" are used to emphasize and clarify, not according to a book of rules.

Bibliography

[BC98] A M Bloch and P E Crouch. Optimal control, optimization, and analytical mechanics. In *Mathematical control theory*, pages 268–321. Springer, New York, 1998.

[Bel53] R Bellman. *The stability theory of differential equations.* McGraw-Hill, New York, 1953.

[Bel84] R Bellman. *Eye of the Hurricane.* World Scientific, Singapore, 1984.

[BF76] R W Brockett and P A Fuhrmann. Normal symmetric dynamical systems. *SIAM J. Contr. Optim.*, 14(1):107–119, 1976.

[BHJ83] H T Banks, H G Hermes, and M Q Jacobs. In memoriam. *SIAM J. Contr. Optim.*, 21(6):vii–ix, 1983.

[Bro83] R W Brockett. Asymptotic stability and feedback stabilization. In *Differential Geometric Control Theory*, pages 181–191, Boston, 1983. Birkhäuser.

[Bro14] R W Brockett. The early days of geometric nonlinear control. *Automatica*, 50(9):2203–2224, 2014.

[Bry96] A E Bryson. Optimal control-1950 to 1985. *IEEE Contr. Syst. Mag.*, 16(3):26–33, 1996.

[Ces71] L Cesari. *Asymptotic behavior and stability problems in ordinary differential equations.* Springer-Verlag, New York, 1971.

[Cha00] N Chafee. Jack K. Hale: A brief biography. *J. Differ. Equ.*, 168(1):2–9, 2000.

[CMP10] S.-N Chow and J Mallet-Paret. Obituary of Jack K. Hale. *J. Dyn. Differ. Equ.*, 22:73–78, 2010.

[FGKM18] M Fazel, R Ge, S Kakade, and M Mesbahi. Global convergence of policy gradient methods for the Linear Quadratic Regulator. In *Proc. International Conference on Machine Learning*, pages 1467–1476, 2018.

[GP10] V Guillemin and A Pollack. *Differential topology.* American Mathematical Society, Providence, 2010.

[Hah67] W Hahn. *Stability of motion.* Springer, Berlin, 1967.

[HM12] U Helmke and J B Moore. *Optimization and dynamical systems.* Springer, London, 2012.

[Hod73] W Hodge. Solomon Lefschetz. 1884-1972. *Biographical Memoirs of Fellows of the Royal Society*, 19:433–453, 1973.

[JM23] W Jongeneel and E Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, Cham, 2023.

[JQ78] V Jurdjevic and J P Quinn. Controllability and stability. *J. Differ. Equ.*, 28(3):381–389, 1978.

[Kal60] R E Kalman. On the general theory of control systems. In *Proc. First International Conference on Automatic Control*, pages 481–492, 1960.

[Kel15] C M Kellett. Classical converse theorems in Lyapunov's second method. *Discrete Cont. Dyn.-B*, 20(8):2333–2360, 2015.

[KFA69] R E Kalman, P L Falb, and M A Arbib. *Topics in mathematical system theory.* McGraw-Hill, New York, 1969.

[KS15] M I Kuzin and O S Shkrob. Nikolai Nikolaevich Krasovskii (on the 90th anniversary of his birth). *Proc. Steklov Institute of Mathematics*, 291:S1–S21, 2015.

[KZ84] A Krasnosel'skiĭ and P P Zabreiko. *Geometrical methods of nonlinear analysis.* Springer, Berlin, 1984.

[LaS60] J LaSalle. Some extensions of Liapunov's second method. *IRE T. Circuit Theory*, 7(4):520–527, 1960.

[Lef70] S Lefschetz. Reminiscences of a mathematical immigrant in the united states. *Am. Math. Mon.*, 77(4):344–350, 1970.

[Lia92] A M Liapunov. *A general task about the stability of motion.* dissertation, University of Kharkov, 1892.

[LP17] A Loria and E Panteley. Stability, as told by its developers. In *Prof. IFAC World Congress*, pages 5219–5230, 2017.

[LSL61] J La Salle and S Lefschetz. *Stability by Liapunov's Direct Method with Applications.* Academic Press, New York, 1961.

[Lya92] A M Lyapunov. The general problem of the stability of motion. *Int. J. Control*, 55(3):531–773, 1992.

[Mas49] J L Massera. On Liapounoff's conditions of stability. *Ann. Math.*, 50(3):705–721, 1949.

[Mas56] J L Massera. Contributions to stability theory. *Ann. Math.*, 64(1):182–206, 1956.

[McM62] B McMillan. Introduction. *Journal of the Society for Industrial and Applied Mathematics Series A Control*, 1(1):1–2, 1962.

[SL12] H Schättler and U Ledzewicz. *Geometric optimal control: theory, methods and examples.* Springer, New York, 2012.

[Son98] E D Sontag. *Mathematical control theory: deterministic finite dimensional systems.* Springer, New York, 1998.

[SS80] E Sontag and H Sussmann. Remarks on continuous feedback. In *Proc. IEEE Conference on Decision and Control*, pages 916–921, 1980.

[Ste11] O A Stepanov. Kalman filtering: Past and present. an outlook from russia.(on the occasion of the 80th birthday of Rudolf Emil Kalman). *Gyroscopy Navig.*, 2(2):99–110, 2011.

[Sus79] H J Sussmann. Subanalytic sets and feedback control. *J. Differ. Equ.*, 31(1):31–52, 1979.

[SW97] H J Sussmann and J C Willems. 300 years of optimal control: from the brachystochrone to the maximum principle. *IEEE Contr. Syst. Mag.*, 17(3):32–44, 1997.

[TSH01] H L Trentelman, A A Stoorvogel, and M Hautus. *Control Theory for Linear Systems.* Springer-Verlag, London, 2001.

[Wil18] A Wilson. *Lyapunov arguments in optimization.* PhD thesis, University of California, Berkeley, 2018.

[Won79] W M Wonham. *Linear Multivariable Control.* Springer, New York, 1979.

[Zab89] J Zabczyk. Some comments on stabilizability. *Appl. Math. Opt.*, 19(1):1–9, 1989.

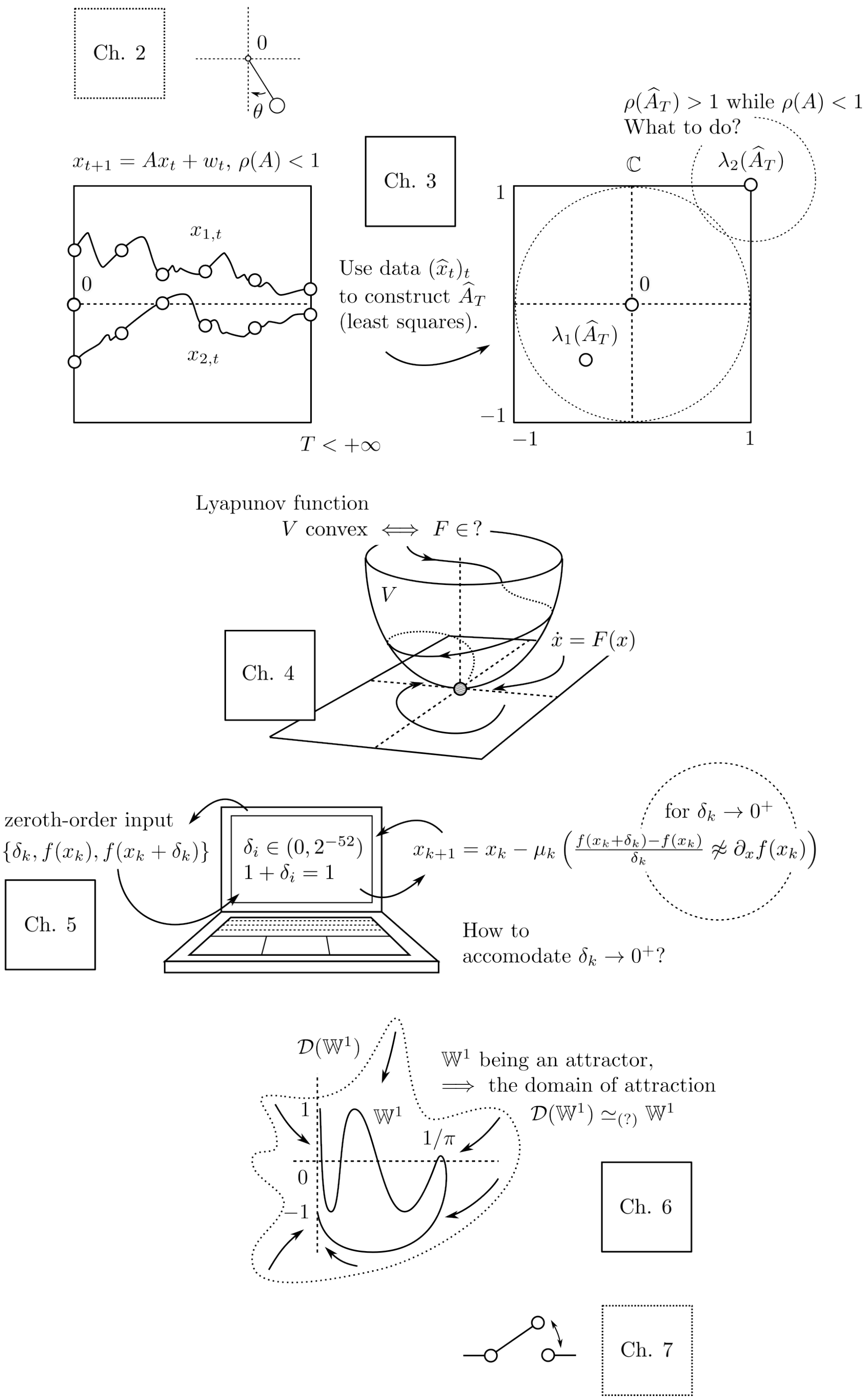

Ch. 2
0
θ
$x_{t+1} = Ax_t + w_t,\ \rho(A) < 1$
$x_{1,t}$
0
$x_{2,t}$
$T < +\infty$
Ch. 3
Use data $(\widehat{x}_t)_t$
to construct $\widehat{A}_T$
(least squares).
$\rho(\widehat{A}_T) > 1$ while $\rho(A) < 1$
What to do?
$\mathbb{C}$
$\lambda_2(\widehat{A}_T)$
1
0
$\lambda_1(\widehat{A}_T)$
-1
-1
1
Lyapunov function
V convex $\iff$ $F \in ?$
V
$\dot{x} = F(x)$
Ch. 4
zeroth-order input
$\{\delta_k, f(x_k), f(x_k + \delta_k)\}$
$\delta_i \in (0, 2^{-52})$
$1 + \delta_i = 1$
$x_{k+1} = x_k - \mu_k \left(\frac{f(x_k+\delta_k)-f(x_k)}{\delta_k} \not\approx \partial_x f(x_k)\right)$
for $\delta_k \to 0^+$
Ch. 5
How to
accomodate $\delta_k \to 0^+$?
$\mathcal{D}(\mathbb{W}^1)$
$\mathbb{W}^1$ being an attractor,
$\implies$ the domain of attraction
$\mathcal{D}(\mathbb{W}^1) \simeq_{(?)} \mathbb{W}^1$
1
$\mathbb{W}^1$
$1/\pi$
0
-1
Ch. 6
Ch. 7

Figure 1.2: Overview of the thesis contents.

2
Introduction to stability theory

> “*Stability theory is the study of systems under various perturbing influences. Since there are many systems, many types of influences, and many equations describing systems, this is an open-ended problem. A system is designed so that it will be stable under external influences. However, one cannot predict all external influences, nor predict the magnitude of those that occur. Consequently, we need control theory. If one is interested in stability theory, a natural result is a theory of control.*”
>
> — Bellman [Bel84, p. 110].

Despite the dictionary definition, *stability* does admit a precise mathematical definition, although “*a*” should be immediately replaced with “*many*”, not because stability is ill-defined, but because many different notions of stability exist. With that in mind we provide a brief *overview* of mathematical stability, with the emphasis on *stability of motion*. We assume some familiarity with dynamical systems, for further introductionary comments and far more details we point to the references, *e.g.*, see [Son98, Sas99, vS21].

Additionally, we comment on stabilizability and stabilization throughout, that is, when and how to enforce certain notions of stability.

Notation will be standard and always accompanied by explanatory text.

2.1 Stability through regularity

The word “*regular*” has a remarkable number of meanings in mathematics. In this section, regularity relates to smoothness properties of a map, that is, we follow the PDE meaning

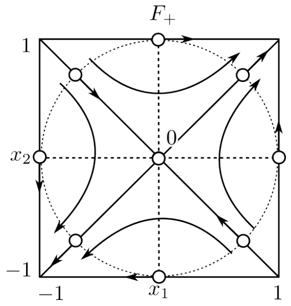


Figure 2.1: Example 2.1.1: integral curves of (2.1.2). The dashed unit circle aids in the computation of $\mathrm{ind}_0(F_+) = -1$, as used in Section 2.4.1 below.

of the word.

2.1.1 Continuity

Most mathematical notions of stability are intimately related to *continuity*, *e.g.*, in learning theory [BE02] (when considering what happens under a change of the training data) or in optimization [Ber63, SLG95] (to understand how the optimal value changes under a change in parameters), see also [HRS16].

We do point out that in those settings, stability frequently goes by different names, *e.g.*, *sensitivity*- or *perturbation analysis*. We will now elaborate on continuity in the context of the stability of motion ("objects changing over time"). What could even be continuous here and what does this imply?

Continuity and stability of motion

When speaking of stability with respect to some dynamical system, say, corresponding to a differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) \tag{2.1.1}$$

one might aim to characterize when solutions to (2.1.1) "are close". As continuity of solutions to (2.1.1) is well-understood (*e.g.*, when local flows exist), one might believe at first that for f being sufficiently regular, we are done. The important point is of course that claims of that nature are only somewhat sensible for *finite* time, while we demand stability for *all time*. An example that comes to mind is $\dot{x} = x^2$. Evidently, the right-hand side is continuous (even analytic), yet the integral curves are of the form

$$t \mapsto \xi(t) = \frac{x_0}{1 - t \cdot x_0}, \quad \xi(0) = x_0.$$

Hence, despite continuity we experience a *finite escape time* [Kha02, p. 93]. Perhaps then, the vector field should be *complete* [Lee12, pp. 215–217]?

Example 2.1.1 (Continuity with respect to initial conditions and stability)**.** Consider the linear dynamical system

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = F_{+}(x) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x(t). \tag{2.1.2}$$

It follows that the flow corresponding to (2.1.2) is defined globally and given by $(t, x_0) \mapsto \varphi^t(x_0) = e^{At}x_0$. Hence, the vector field is complete. It also readily follows from (2.1.2) that the origin is unstable (a saddle) under (2.1.2) with the stable mode being $v_s = (1, -1)$ and the unstable mode being $v_u = (1, 1)$. Hence, it follows that $\lim_{t\to+\infty} \varphi^t(v_s) = 0$, yet, for any (arbitrarily small) $\varepsilon > 0$ we get that

$$\lim_{t\to+\infty} \|\varphi^t(v_s + \varepsilon v_u)\|_2^2 = +\infty,$$

despite continuity of φ in x_0, see Figure 2.1. Note, this is not just a linear phenomenon, the same happens when starting on and around an unstable periodic orbit and so forth.

Example 2.1.1 shows that to capture the stability of motion, we need to go beyond the standard notion of continuity for difference- and differential equations, that is, we need to go beyond continuity with respect to initial conditions.

On the other hand, continuity of the right-hand side of a difference or differential equation is not necessary for stability either.

Example 2.1.2 (Switched linear systems)**.** Consider the switched linear system given by

$$x_{k+1} = A_{\sigma(k)}x_k, \quad A_1 = \begin{pmatrix} -1/2 & 1/2 \\ 0 & -1/2 \end{pmatrix}, A_2 = \begin{pmatrix} 1/2 & 1/10 \\ 1/5 & -4/5 \end{pmatrix}, \tag{2.1.3}$$

where $\sigma : \mathbb{Z} \to \{1, 2\}$ is the switching signal. Now, the origin is globally asymptotically stable under (2.1.3), for any particular realization of σ since $x \mapsto V(x) = \langle x, x\rangle$ is a common Lyapunov function for $\{A_1, A_2\}$, that is, for any $i \in \{1, 2\}$ we have $V(A_i x) < V(x)$ for all $x \neq 0$, which suffices to conclude on stability [LA09, Sec. II] (see below).

The examples above motivate the *asymptotic* notions of stability, as central to this thesis. Below, we elaborate on these notions, on Lyapunov functions to capture stability, as used in Example 2.1.2, but first, we discuss further manifestations of regularity and their ramifications for stability.

2.1.2 Lipschitz inequalities

In order to be able to prove convergence of zeroth-order optimization algorithms in Chapter 5, we need to exploit regularity of the objective function f. Here, it suffices for f to display certain *Lipschitz* continuity properties. Following [Nes03], for any integers $p, k \geq 0$ with $p \leq k$, we use $C_L^{k,p}(\mathcal{D})$ to denote the family of all k times continuously differentiable functions on $\mathcal{D}$ whose p^{th} derivative is Lipschitz continuous with Lipschitz constant $L \geq 0$. We sometimes write $L(f)$ to indicate which function we are discussing. Similarly, we use $C_L^{\omega,p}(\mathcal{D})$ to denote the family of all analytic functions in $C_L^{p,p}(\mathcal{D})$.

For example, if $f \in C^{1,1}_{L_1}(\mathcal{D})$, then f has a Lipschitz continuous gradient, that is,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_1\|x-y\|_2 \quad \forall x, y \in \mathcal{D}. \tag{2.1.4}$$

By [NS17, Eq. (6)], this condition is equivalent to the inequality

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle| \leq \tfrac{1}{2}L_1\|x-y\|_2^2 \quad \forall x, y \in \mathcal{D}. \tag{2.1.5}$$

To see how inequalities of the form (2.1.5) aid in stability, suppose we aim to find a minimizer of $f \in C^{1,1}_{L_1}(\mathcal{D})$, denoted by $x^\star$. Since it is unlikely we will ever find $x^\star$ exactly (by hand) we need to resort to numerical means and thus it is very important to get a grip on the *suboptimality gap* under perturbations of $x^\star$ (since the computer will also never find $x^\star$ exactly), that is, on $f(x^\star+\varepsilon) - f(x^\star)$ for some perturbation ε. Under the assumption that $x^\star$ is a critical point of f, it readily follows from (2.1.5) that $f(x^\star+\varepsilon) - f(x^\star) \leq \frac{1}{2}L_1\|\varepsilon\|_2^2$. Note that getting a grip on suboptimality is also an integral part of studying convergence of optimization algorithms, which is precisely how we will use these and other inequalities in Chapter 5.

To continue, if $f \in C^{1,1}_{L_1}(\mathcal{D})$ is also convex then, the Lipschitz condition (2.1.4) is also equivalent to

$$f(y) \geq f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2L_1}\|\nabla f(y) - \nabla f(x)\|_2^2 \quad \forall x, y \in \mathcal{D}, \tag{2.1.6}$$

see, *e.g.*, [Nes03]. In particular, if x is a critical point, *e.g.*, a local minimizer of f with $\nabla f(x) = 0$, then the estimate (2.1.6) simplifies to $2L_1\left(f(y) - f(x)\right) \geq \|\nabla f(y)\|_2^2$ for all $y \in \mathcal{D}$.

If $f \in C^{2,2}_{L_2}(\mathcal{D})$, then f has a Lipschitz continuous Hessian, *i.e.*,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L_2\|x-y\|_2 \quad \forall x, y \in \mathcal{D}. \tag{2.1.7}$$

By [Nes03, Lem. 1.2.4], this condition is equivalent to the inequality

$$|f(y) - f(x) - \langle \nabla f(x), y-x\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle| \leq \tfrac{1}{6}L_2\|x-y\|_2^3 \quad \forall x, y \in \mathcal{D}. \tag{2.1.8}$$

More generally, any $f \in C^{p,p}_{L_p}(\mathcal{D})$ has a Lipschitz continuous p^{th} derivative. Recalling the definitions of higher-order partial derivatives and multi-indices, this requirement can be expressed as

$$|\textstyle\sum_{|\alpha|=p} \partial_x^\alpha f(x)\cdot u^\alpha - \sum_{|\alpha|=p} \partial_x^\alpha f(y)\cdot u^\alpha| \leq L_p\|x-y\|_2 \quad \forall x, y \in \mathcal{D},\ u \in \mathbb{S}^{n-1}.$$

It is often referred to as a $(p+1)^{\text{th}}$-order smoothness condition [BP16, Sec. 1.1] as it implies that any $f \in C^{p+1,p}_{L_p}(\mathcal{D}) \subseteq C^{p,p}_{L_p}(\mathcal{D})$ has a bounded $(p+1)^{\text{th}}$ derivative, that is,

$$|\textstyle\sum_{|\alpha|=p+1} \partial_x^\alpha f(x)\cdot u^\alpha| = |\partial_t^{p+1} f(x+tu)|_{t=0}| \leq L_p \quad \forall x \in \mathcal{D},\ u \in \mathbb{S}^{n-1}. \tag{2.1.9}$$

Stability under maps

A more "*dynamical*" application of a map being Lipschitz is the following. Let $F : X \to X$ be a map on a metric space (X, d), then F is said to be a ***contraction*** when there is a $\gamma \in [0, 1)$ such that

$$d(F(x_1), F(x_2)) \leq \gamma d(x_1, x_2) \quad \forall x_1, x_2 \in X. \tag{2.1.10}$$

See that if $X = \mathbb{R}^n$ and d is the usual metric on $\mathbb{R}^n$, then we get that F is Lipschitz continuous with $L_0(F) = \gamma$. Now suppose that F relates to some algorithm, say an optimization algorithm. Then we usually think of F as in the update rule $x_k \mapsto F(x_k) = x_{k+1}$ and ideally $F(x^\star) = x^\star$ for $x^\star$ an optimizer of our problem at hand. In that case, F being a contraction readily implies that $d(x_{k+1}, x^\star) \leq \gamma^k d(x_1, x^\star)$, which under mild topological conditions on (X, d) implies that $x_k \to x^\star$. This observation appears frequently in optimization, however, we would like to stress, already, that results of this form say more than mere convergence. Consider a map $V : X \to \mathbb{R}_{\geq 0}$ defined through $V(x) = d(x, x^\star)$, then, since F is a contraction, it follows that $V(F(x)) < V(x)$ and $V(x) = 0 \iff x = x^\star$. Hence, through Lyapunov arguments (see below), we observe *stability*, which is very important since an algorithm only runs for finite time. Indeed, since the time of Malkin and Massera we know that asymptotic stability and uniform convergence are intimately related. More on that below.

2.2 Dynamical control systems

In this section we start by discussing (time-invariant) ***dynamical systems*** over $\mathbb{R}^n$ of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = F(x(t)) : \begin{cases} F : \mathbb{R}^n \to T\mathbb{R}^n \\ \pi \circ F = \mathrm{id}_{\mathbb{R}^n}, \end{cases} \tag{2.2.1}$$

where F is C^r-smooth with $r \geq 0$, $\pi : T\mathbb{R}^n \to \mathbb{R}^n$ defined by $(x, v) \mapsto \pi(x, v) = x$ is the canonical projection and for any $x \in \mathbb{R}^n$ we have with some abuse of notation $F(x) \in T_x\mathbb{R}^n$. Evidently, $T\mathbb{R}^n \simeq \mathbb{R}^n \times \mathbb{R}^n$, but (2.2.1) is useful to keep in mind when comparing objects to assess if generalizations beyond $\mathbb{R}^n$ are possible. *Integral curves* of (2.2.1) are differentiable curves $t \mapsto \xi(t) \in \mathbb{R}^n$ such that $\dot{\xi}(t) = F(\xi(t))$ for all $t \in \mathrm{dom}(\xi)$, which is non-empty by, for instance, assuming that $r \geq 1$. However, in general, such an assumption is too strong. We will not go into further regularity conditions and always assume, for simplicity and unless written otherwise, that $r = 0$ and that the vector field is *complete*, *i.e.*, a global flow (see below) is induced, such that we are allowed to make global statements[1], for further information we point the reader to [Son98, Hal09] and Example 2.2.3 below.

Going beyond *descriptions*, when aiming to *prescribe* the dynamics of a system we consider (time-invariant) ***dynamical control systems*** over $\mathbb{R}^n \times \mathbb{R}^m$ of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t), u) \tag{2.2.2}$$

such that $f(x, u) \in T_x\mathbb{R}^n \simeq \mathbb{R}^n$ for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$, where x and u denote the state and input, respectively. Again, with some abuse of notation, we will assume that $f \in C^0(\mathbb{R}^n \times \mathbb{R}^m; \mathbb{R}^n)$, but again omit integrability discussions. Input functions are of the form $t \mapsto \mu(t) \in \mathbb{R}^m$, *e.g.*, a state feedback is of the form $t \mapsto \mu(x(t))$. Note, we use μ

[1] We remark that *completeness* is the important property here as we will appeal to a global flow, *smoothness* of F (going beyond C^0), on the other hand, is rarely exploited. The only reason to potentially keep smoothness is that one can naturally relax completeness and make some local statements.

instead of u to differentiate between the function and the point. A subclass of (2.2.2) of interest are the so-called *control affine* systems of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) + \textstyle\sum_{i=1}^{m} g_i(x(t))u_i, \tag{2.2.3}$$

where u_i is the i^{th} element of $u \in \mathbb{R}^m$ and again $f, g_i \in C^0(\mathbb{R}^n;\mathbb{R}^n)$ for $i = 1,\dots,m$ [NvdS90]. Based on the control system at hand, one might say more about the space of allowable inputs $t \mapsto \mu(t)$, *e.g.*, one might consider *absolutely integrable* (L^1_{loc}) or *essentially bounded* (L^∞_{loc}) function spaces [Son98, App. C].

2.2.1 Stability

Now, we will define the central asymptotic notion of stability.

Let F parametrize a dynamical system of the form (2.2.1). By our standing completeness and smoothness assumptions, F will give rise to a continuous *flow*[2] $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, with its evaluation denoted by $\varphi^t(x_0) := \varphi(t, x_0)$, which is understood to describe a solution to (2.2.1) at time t, starting at time 0 from x_0. A point $x^\star \in \mathbb{R}^n$ is an *equilibrium point* of F when $F(x^\star) = 0$, *w.l.o.g.* we set $x^\star = 0$. Then, 0 is said to be ***globally asymptotically stable*** (GAS) (with respect to F) if

(s-i) 0 is *Lyapunov stable*, that is, for any open neighbourhood $U_\varepsilon \ni 0$ there is an open set $U_\delta \subseteq U_\varepsilon$ such that a solution (with respect to F) starting in U_δ stays in U_ε;

(s-ii) 0 is *globally attractive*, that is, $\lim_{t\to+\infty} \varphi^t(x_0) = 0$ for all $x_0 \in \mathbb{R}^n$.

Simply put, we speak of asymptotic stability of some equilibrium point, when solutions starting sufficiently close to this equilibrium converge to this point and while doing so do not first diverge. We do point out that Lyapunov stability and attractivity are independent, for instance, the origin is Lyapunov stable under a center[3], but fails to be attractive, on the other hand, the origin is attractive under Artstein's circles[4], yet, it fails to be Lyapunov stable.

We also point out that the focus on *global* stability should be understood as working with the domain of attraction and not "some" neighbourhood. Indeed, the focus on $(0, \mathbb{R}^n)$ is somewhat arbitrary for the moment.

As we focus on time-invariant dynamical systems, we will not further digress much into solutions and stability and refer to [Hah67, Son98]. In particular, we will not discuss *comparison functions*, see [Hah67, Sec. 24].

In general it is not straightforward to capture if 0 is GAS or not. A fruitful tool that *does* allow for conclusions of this form has been devised by Lyapunov in the late 1800s [Lia92]. A function $V \in C^\infty(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ is said to be a (smooth, strict and proper) ***Lyapunov function*** (with respect to F and 0) when

(V-i) $V(x) > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$ and $V(0) = 0$;

[2] Flows satisfy: (1) the *identity* $\varphi^0 = \mathrm{id}_{\mathbb{R}^n}$; and (2) *group* property $\varphi^{s+t} = \varphi^s \circ \varphi^t \ \forall s,t \in \mathbb{R}$.

[3] A system of the form $\dot{x}_1 = x_2$, $\dot{x}_2 = -x_1$.

[4] A system of the form $\dot{x}_1 = x_1^2 - x_2^2$, $\dot{x}_2 = 2x_1x_2$, mapped onto the sphere $\mathbb{S}^2$, see also [JM23, Rem. 6.1].

(V-ii) $\langle \nabla V(x), F(x)\rangle < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$;

(V-iii) and V is *radially unbounded*, that is, $V(x) \to +\infty$ for $\|x\| \to +\infty$.

Property (V-iii) implies sub-level set compactness, which is want we need, coercivity is just a convenient way to capture this. Now, based on work by Massera, Kurzweil and others [Kur63, FP19], we will exploit the celebrated theorem stating that 0 is GAS if and *only if* there is a (corresponding) smooth Lyapunov function [BR05, Thm. 2.4]. Note, we dropped the adjective "*strict and proper*" as we almost exclusively look at Lyapunov functions of that form. See also that, given that V satisfies Property (V-i), then $\langle \nabla V(x), F(x)\rangle \leq -V(x)$ implies Property (V-ii).

For further references on Lyapunov stability theory we point the reader to [BS70, Son98, BR05].

Remark 2.2.1 (On asymptotic- and uniform asymptotic stability)**.** Due to Massera we know that for time-invariant dynamical systems on finite-dimensional spaces (*e.g.*, like (2.2.1)) asymptotic- and uniform asymptotic stability are equivalent [Mas49, Thm. 7], [Mas56, p. 188]. In the time-invariant case, this simply means that the convergence from (s-ii) is *uniform*, that is, for any ϵ there is a $T(\epsilon)$ such that $\|\varphi^t(x_0)\|_2 \leq \epsilon$ for all $t \geq T(\epsilon)$ and any $x_0 \in \mathbb{R}^n$. This observation was very important regarding converse theory [Hah67, Sec. 47], looking at Lyapunov functions, the sub-level set compactness naturally connects to uniformity.

Remark 2.2.2 (On time-invariant converse theorems)**.** The vast majority of converse theorems provide a Lyapunov function of the form $(x,t) \mapsto V(x,t)$ (for possibly time-varying and non-smooth dynamical systems). For the full proof, that includes the time-invariant case, we point to [BR05, p. 146] and the theory that precedes that remark.

Example 2.2.3 (On completeness)**.** We assume completeness of our vector fields to avoid technical discussions on existence and uniqueness of solutions. However, since the time of Zubov an easy transformation is known that justifies our assumption in the qualitative study of dynamical systems, see [Hah67, Sec. 34]. Given some dynamical system (2.2.1), then define the differential (one-form) $\mathrm{d}s = (1+\|F(x)\|_2^2)^{1/2}\mathrm{d}t$. Now consider

$$\frac{\mathrm{d}}{\mathrm{d}s}x(s) = \varphi(x)F(x) := \frac{1}{(1+\|F(x(s))\|_2^2)^{1/2}}F(x(s)). \tag{2.2.4}$$

Since the right-hand side of (2.2.4) is globally bounded, the transformed vector field is complete under mild assumptions on F. As such a transformation is just a positive scaling of the original vector field F, stability properties are preserved (*e.g.*, consider $\langle \nabla V(x), F(x)\rangle$ and $\langle \nabla V(x), \varphi(x)F(x)\rangle$ for appropriate functions $x \mapsto \varphi(x)$).

Now, given a control system (2.2.2), when it comes to the task of *globally asymptotically stabilizing* 0 (we will exclusively focus on stabilization by means of time-invariant state feedback[5]), the Lyapunov function paradigm can be adjusted.

[5] Considering more general input functions, *e.g.*, of the form $t \mapsto \mu(t, x(t))$, integral curves of the corresponding closed-loop system are generally understood to be absolutely continuous curves $\xi : I \to \mathbb{R}^n$ such that the differential relation $\dot{\xi}(t) = F(\xi(t), \mu(t, \xi(t))) =: F'(t, \xi(t))$ holds for almost all $t \in I$, in the sense of Lebesgue. This requires rethinking some concepts, *e.g.*, global asymptotic stability and what a closed-loop *vector field* really is.

Given our stabilization goal, we seek a function $t \mapsto \mu(x(t))$ such that under $f(x,\mu(x)) =: F(x)$ the origin is GAS. Then, analogously to the definition of a Lyapunov function, one can define ***control Lyapunov functions*** (CLFs), yet, Property (V-ii) is now replaced by asking that for any $x \in \mathbb{R}^n \setminus \{0\}$ the following holds

$$\inf_{u\in\mathbb{R}^m} \langle \nabla V(x), f(x,u)\rangle < 0. \tag{2.2.5}$$

It is not evident that a choice of input function based on (2.2.5) can result in a continuous—let alone smooth—feedback. The next section elaborates on this problem.

Before doing so, we do remark that rendering a system stable is by no means always the goal, rather understanding how stability properties of our system relate to our desiderata. We must not forget to *respect the unstable*[6] [Ste03].

Chetaev functions and instability

When studying *in*stability it is imperative to understand that unstable systems are not simply generated by stable systems under time reversal, the class of unstable systems is larger. Hence, searching for Lyapunov functions to conclude on instability is frequently futile (consider a saddle). Instead, one can do with so-called ***Chetaev functions***. Intuitively, you just need "some direction to escape", like an eigenvector corresponding to an unstable eigenvalue.

Formally, let $x^\star = 0$ be an equilibrium point of some vector field on $\mathbb{R}^n$. Suppose there is a $V \in C^1(D;\mathbb{R})$ with $0 \in D$, $V(0) = 0$ and such that any neighbourhood U of 0 satisfies $V^{-1}((0,+\infty)) \cap U \neq \emptyset$. Then, $x^\star = 0$ is unstable if $\dot{V} > 0$ on $V^{-1}((0,+\infty)) \cap K$, for K some compact neighbourhood of 0, being a subset of D [Kha02, Ch. 4].

2.2.2 Control Lyapunov functions

Consider a dynamical control affine system with scalar input of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) + g(x(t))u, \tag{2.2.6}$$

Then, for V to be a smooth CLF for (2.2.6), we must have that for any $x \in \mathbb{R}^n \setminus \{0\}$ there exists a $u \in \mathbb{R}$ such that $L_f V(x) + u L_g V(x) < 0$. However, the existence of a *smooth control*-Lyapunov function is topologically strong in the sense that it generally implies (see below) that an asymptotically stabilizing *continuous* feedback exists [Son98, Ch. 5]. Indeed, the controller attributed to Sontag is

$$x \mapsto \mu_s(x) := \begin{cases} -\dfrac{\alpha(x) + \sqrt{\alpha(x)^2 + \beta(x)^4}}{\beta(x)} & \text{if } \beta(x) \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.2.7}$$

[6] The very first Bode lecture, by Stein, contains various examples of how delicate unstable plants are, but also that some engineering problems desire instability. For instance, for airplanes, instability in certain regimes is necessary to be efficient in other regimes. Similarly, neuroscientists actively study in which stability regime the brain operates.

for $\alpha(x) := L_f V(x)$ and $\beta(x) := L_g V(x)$, *e.g.*, see [Son98, p. 249]. Although (2.2.7) appears singular, $\mu_s(x)$ can be shown to be continuous under the following condition; we speak of the *small control property* when for all $\varepsilon > 0$ there is a $\delta > 0$ such that if $x \in \mathbb{R}^n \setminus \{0\}$ satisfies $\|x\| < \delta$, then, there is a u such that $\|u\| < \varepsilon$ and $L_f V(x) + L_g V(x) u < 0$ [Son89, p. 247]. As such, the existence of a *smooth* CLF is strictly stronger than being (globally) asymptotically controllable[7], *e.g.*, continuous feedback can be easily obstructed for globally controllable systems that even admit smooth CLFs[8]. To add, the small control property does not always hold and it is well-known that CLF-based-controllers can be singular, and ever since their inception so-called "*desingularization techniques*" emerged [Cor07, Sec. 12.5.1]. For instance, under structural assumptions a backstepping approach to handle CLF singularities is studied in [LK97] and a PDE reformulation to avoid singularities is presented in [YI00].

Nevertheless, in case the dynamical control system is affine in the input u, and u is constrained to a compact convex set, then, the *existence* of a C^∞ CLF is equivalent to the existence of a C^0 (on $\mathbb{R}^n \setminus \{0\}$) stabilizing feedback [Art83]. Indeed, the work by Sontag aimed at making the *construction* of such a feedback transparent. Further relaxing regularity of a CLF, it can be shown that the existence of a so-called "*proximal CLF*" is equivalent to asymptotic controllability. These proximal CLFs are C^r-smooth with $r \in [0,1)$, *e.g.*, see [Cla10] for more on non-smooth CLFs. Better yet, it can be shown that global asymptotic controllability implies the existence of a—possibly discontinuous—feedback [CLSS97]. Even more, Rifford showed that when the control system is globally asymptotically controllable, a—possibly nonsmooth—semiconcave[9] CLF always exists. Exploiting this structure, for control affine systems, Rifford could extend Sontag's formula (2.2.7) to this setting [Rif02, Thm. 2.7] and *get* again an explicit feedback.

The existence of a smooth CLF is not only topologically strong, it implies there exists a *robustly* stabilizing feedback [LS99]. We also point out that the CLF framework largely extends to switched systems [MBP07].

On learning-based stabilization

Neural networks are becoming increasingly popular in the context of controller synthesis [JWYM20, GZY21, MDT$^+$22, ZZL22]. A principled approach, however, that guarantees some form of stability is largely lacking. Progress has been made when it comes to handling side-information [AEK20], obtaining statistical stability guarantees [BTM$^+$21], in the context of input-state stability [YXRR22], in the context of input-output stability by exploiting the Hamilton-Jacobi inequality [OK22], by exploiting contraction theory [RKKM22] and by exploiting Koopman operator theory [ZB22], to name a few. As these methods are data-driven, errors inevitably slip in and great care must be taken when one aims to mimic CLF-based controllers, *i.e.*, if $L_g V(x) = 0 \implies L_f V(x) < 0$ holds for the estimated system, does it hold for the real system and what happens if it does not? In particular, recall (2.2.7). Moreover, in this setting the underlying dynamical

[7] See for example [Rif02, Sec. 2] and references therein for more on this notion.

[8] A well-known example attributed to Ledyaev and Sontag is of the form $\dot{x}_1 = u_2 u_3$, $\dot{x}_2 = u_1 u_3$, $\dot{x}_3 = u_1 u_2$ *cf.* [LS99].

[9] A continuous function f is said to be *semiconcave* when there is a $C > 0$ such that $x \mapsto f(x) - C\|x\|_2^2$ is concave.

control system is frequently unknown and a function class for V needs to be chosen *a priori*, what does this choice imply? These questions inspired Section 4.2. We also point out that these methods continue a long history of research on computational methods for Lyapunov functions, *e.g.*, see [GH15] for a review.

2.2.3 Topological perspective on level sets and singularities

We proceed by detailing (recalling) how level sets of smooth Lyapunov functions, with respect to points, look like topologically. This result has some ramifications and provides for motivation in the Section 4.2. For simplicity, we momentarily focus on (2.2.6).

In Section 2.2.2 we discussed why one might be interested in studying terms of the form $L_g V(x)^{-1} = \langle \nabla V(x), g(x)\rangle^{-1}$ *cf.* (2.2.7). In this section we show that for practical purposes, the properties of V frequently obstruct this term to be well-behaved (perhaps, to no surprise). Indeed, singularities are studied and shown to be unavoidable when $g(x) := g$ for some $g \in \mathbb{R}^n$.

To start, consider a C^0 dynamical system of the form (2.2.1) on $\mathbb{R}^n$, with $n \geq 2$, and assume that $0 \in \mathbb{R}^n$ is globally asymptotically stable (and hence isolated). This implies that there is a (strict) C^∞ Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$. In particular, this implies that V is also a Lyapunov function for the C^∞ auxiliary system

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = -\nabla V(z(t)). \tag{2.2.8}$$

Hence, $0 \in \mathbb{R}^n$ is also GAS under (2.2.8). By a classical topological result largely[10] due to Krasnosel'skiĭ & Zabreĭko [KZ84, Sec. 52] this directly implies that the corresponding *vector field* ***index*** (with respect to 0) satisfies

$$\mathrm{ind}_0(-\nabla V) = (-1)^n \neq 0.$$

As the vector field index is the (oriented) degree of the map $v : \partial U \to \mathbb{S}^{n-1}$ for any open neighbourhood U of 0 containing no other equilibrium points in its closure [Mil65, Sec. 6], [GP10, Ch. 3], this can only be true if

$$v : \partial U \ni z \mapsto \frac{-\nabla V(z)}{\|\nabla V(z)\|_2}$$

is surjective. As U is arbitrary, it follows that the (normalized) gradient of V along any non-trivial level set hits any vector in $\mathbb{S}^{n-1}$. Differently put, fix any $g \in \mathbb{S}^{n-1}$ then, for any $c > 0$ there is always a $z \in V^{-1}(c) =: V_c \subset \mathbb{R}^n$ such that $\langle \nabla V(z), g\rangle = 0$. Indeed, this is why we assumed $n \geq 2$, otherwise the claim is not true *cf.* [Son89, p. 121]. Summarizing, we have shown the following—which we attribute to Wilson [WJ67] and Byrnes [Byr08, Thm. 4.1].

Proposition 2.2.1 (Level sets of smooth Lyapunov functions (Wilson, Byrnes))**.** *Let $n \geq 2$ and fix some $g \in \mathbb{R}^n \setminus \{0\}$. Then, for any level set V_c, with $c > 0$, of any C^∞-smooth Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, asserting $0 \in \mathbb{R}^n$ to be GAS under some dynamical system* (2.2.1)*, there is an $x \in V_c$ such that $\langle \nabla V(x), g\rangle = 0$.*

[10] Earlier comments can be found in [BK74], see also [JM23].

Note, Proposition 2.2.1 implicitly assumes that $0 \in \mathbb{R}^n$ is the only equilibrium point as we assume the origin is *globally* asymptotically stable. If desired, one can adapt the statement and work with the domain of attraction. Also note that the discussion above detailed that the normalized vector $\nabla V(x)$ will hit *any* vector in $\mathbb{S}^{n-1}$, our focus on $\langle \nabla V(x), g\rangle$ being equal to 0 at some point is purely application-driven. Moreover, we see that the set of points that render the inner product zero is of codimension 1.

Indeed, Proposition 2.2.1 is itself classical as this result can also be understood more intuitively by directly appealing to work by Wilson. Namely, due to the work by Wilson, and later Perelman, we know that the level sets of (strict and proper) C^∞ Lyapunov functions $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ are homeomorphic to $\mathbb{S}^{n-1}$ [WJ67, Sti12]. Although we might assume that these level sets V_c and $\mathbb{S}^{n-1}$ come equipped with a smooth structure, this does not immediately imply the manifolds are diffeomorphic, *e.g.*, consider Milnor's *exotic spheres* [Mil56]. Nevertheless, one expects that the gradient of V along V_c hits any direction when seen as a vector in $\mathbb{S}^{n-1}$, as indeed succinctly shown above. Visualizations can be found in [Son99] and further comments of this nature are collected by Byrnes in [Byr08], in particular, the diffeomorphism question is addressed.

The ramifications for smooth CLFs are immediate as one observes that the argument with respect to the auxiliary system (2.2.8) extends *mutatis mutandis*.

These comments are motivated by renewed interest in CLFs from the neural network community. The following example highlights some work that arguably would benefit from Proposition 2.2.1.

Example 2.2.4 ((Almost) Singular CLF-based controllers)**.** In [KYK22, Sec. IV] the authors consider a dynamical control system of the form $\dot{x} = f(x)+gu$ with $f \in C^\infty(\mathbb{R}^2;\mathbb{R}^2)$, $g \in \mathbb{R}^2$ and $u \in \mathbb{R}$. Their to-be-learned CLF is of the form $V(x) = \sigma_k(\gamma(x)-\gamma(0))+\varepsilon\|x\|^2$ with $\gamma : \mathbb{R}^n \to \mathbb{R}$ being an input-convex neural network and $\sigma_i : \mathbb{R} \to \mathbb{R}_{\geq 0}$ C^1-smooth locally quadratic activation functions [KM19, Eq. (13)] for $i = 0,\dots,k$. Hence, $V \in C^1(\mathbb{R}^n;\mathbb{R}_{\geq 0})$. Indeed, the authors report that the learned CLF leads to large control values (under a Sontag-type controller (2.2.7)), they do not detail why. The above discussion provides a topological viewpoint.

One can also interpret Proposition 2.2.1 through the lens of feedback linearization. Consider some input-output system Σ of the form

$$\Sigma : \begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) + gu \\ \quad\; y(t) = h(x(t)) \end{cases} \tag{2.2.9}$$

for $h = V$, that is, h is given by the CLF V (with respect to f and g). Let the desired output be $y_d \equiv 0$ such that $e(t) = y(t) - y_d(t) = y(t)$. Hence, $\dot{e} = \dot{V}$. Now the standard (relative degree 1) feedback linearizing controller for (2.2.9) is of the form $u = (L_g V)^{-1}(v - L_f V)$ with v denoting the new auxiliary input [Isi85, NvdS90]. Indeed, under the choice

$$v = -\sqrt{(L_f V)^2 + (L_g V)^4}$$

one recovers Sontag's controller (2.2.7). Now Proposition 2.2.1 tells us that the *decoupling* term $(L_g V)^{-1}$ must be singular in any sufficiently small neighbourhood of 0, *i.e.*, the relative degree assumption fails to hold.

Remark 2.2.5 (Generalizations)**.** To go beyond input vector fields of the form $g(x) \equiv g \in \mathbb{R}^n$ we look at two scenarios.
(g-i) (*Dependency on* x): Introduce the function class

$$\mathscr{G}_n := \{g \in C^0(\mathbb{R}^n;\mathbb{R}^n) : g(x) = g_1 + g_2(x),\, g_1 \in \mathbb{R}^n \setminus \{0\},\, \lim_{x\to 0} g_2(x) = 0\}.$$

Indeed, for any $g \in \mathscr{G}_n$, with $n > 1$, it follows that for sufficiently small $c > 0$ there is an $x \in V_c$ such that $\langle \nabla V(x), g(x)\rangle = 0$. The reason being that since $g \in \mathscr{G}_n$ there are always $x_1, x_2 \in V_c$ such that $\langle \nabla V(x_1), g(x_1)\rangle < 0$ while $\langle \nabla V(x_2), g(x_2)\rangle > 0$. Then the claim follows from standing regularity assumptions[11] and the intermediate value theorem.
(g-ii) (*Multidimensional input*): Assume that $u \in \mathbb{R}^m$ with $1 < m < n$ and let the dynamical control system be of the form $\dot{x} = f(x) + \sum_{i=1}^m g_i u_i$ (dependence on x can be generalized as in (g-i)). Then, as $\operatorname{span}\{g_1,\dots,g_m\} \neq \mathbb{R}^n$ there is a nonzero $v \in \operatorname{span}\{g_1,\dots,g_m\}^\perp$.

Exploiting the remark from above, we recover a slightly weaker version of a well-known result *cf.* [Blo15, Prop. 6.1.4], better yet, one recovers (locally) a weaker version of the highly influential obstruction to continuous asymptotic stabilization of Brockett's nonholonomic integrator *e.g.*, see [Son98, Ex. 5.9.16].

Corollary 2.2.2 (Obstruction for nonholonomic systems)**.** *Assume that $u \in \mathbb{R}^m$ with $1 < m < n$ and let the dynamical control system be of the form $\dot{x} = \sum_{i=1}^m g_i(x)u_i$ with $g_i \in \mathscr{G}_n$ for $i = 1,\dots,m$, then, there is no smooth CLF with respect to $0 \in \mathbb{R}^n$.*

Proof. Indeed, this result follows from, for example, Brockett's condition [Bro83]. However, from Proposition 2.2.1 and Remark 2.2.5 we know there is a point $x' \in \mathbb{R}^n \setminus \{0\}$ such that $\langle \nabla V(x'), \sum_{i=1}^m g_i(x')\rangle = 0$. This implies that $L_f V(x') < 0$ must hold for V to be a CLF. As $f \equiv 0$, this is impossible and no smooth CLF can exist. □

2.2.4 Comments on discrete-time dynamical systems

So far, we looked at *continuous-time* dynamical systems, but similar results hold true for *discrete-time* dynamical systems. The discrete version of (2.2.1) would simply be a continuous map $F : \mathbb{R}^n \to \mathbb{R}^n$, usually interpreted as $x_k \mapsto F(x_k) = x_{k+1}$ for $k \in \mathbb{Z}$. If F is derived from a flow, that is, if F corresponds to the *time-one map* $x \mapsto \varphi^1(x)$, then F is a *homeomorphism* (since φ^{-1} exists and is continuous). Indeed, in that case we can still work with maps of the form $(t, x) \mapsto \varphi^t(x)$, yet, now t is taking values in $\mathbb{Z}$.

Again, we can capture global asymptotic stability, in this case of a *fixed point* of F, that is, a point $x^\star$ such that $F(x^\star) = x^\star$. For simplicity of exposition, suppose that $x^\star$ is again 0. In that case, the stability notions (s-i–s-ii) and Lyapunov characterization (V-i–V-iii) carry over almost immediately, with the only difference being that (V–ii) is replaced with $V(F(x)) < V(x)$ for all $x \neq 0$. We also remark that in this case we gain little by demanding that V is smooth. Continuity, however, is important regarding robustness.

It is interesting to point out that converse theory for the discrete case was not immediate an usually relies on *exponential stability* [Hah58, KB60]. It took until 2002, when

[11] It readily follows that $V_{c>0}$ is a codimension 1, C^∞ manifold.

Jiang and Wang provided a converse theory for discrete-time systems, similar to what was known for continuous-time systems [JW02, Thm. 1]. See also [GGLW14].

Discrete-time systems are frequently presented as a simple modification of continuous-time systems, however, some phenomena are purely discrete or purely continuous. For instance, in the discrete case, a finite escape time does not exist since $x \mapsto F^n(x)$ remains a continuous map for any finite $n \in \mathbb{Z}_{\geq 0}$, mapping compact sets to compact sets. Indeed, discretizing $\dot{x} = x^2$ will not result in a (global) continuous map $F : \mathbb{R} \to \mathbb{R}$.

Also see that a global flow for a vector field immediately translates to a time-one map $x \mapsto \varphi^1(x)$ (discrete-time system), yet, the other way around is less obvious. For instance, if $F \in C^0(\mathbb{R}^n; \mathbb{R}^n)$ should correspond to a time-one map of a global flow, then F must be a homeomorphism. Now consider $x \mapsto F(x) = Bx$ for some low-rank $B \in \mathbb{R}^{n\times n}$. Such a map does not correspond to a global flow and indeed see that $\dot{x} = Ax$ translates to $x_{k+1} = e^A x_k$ where $\det(e^A) = e^{\mathrm{Tr}(A)} > 0$ for any $A \in \mathbb{R}^{n\times n}$. We refer to work on *suspensions* for more on relations between maps and flows [KH95].

2.2.5 Further comments on linear dynamical systems

Largely due to outstanding texts on linear algebra, we have a beautifully developed theory of linear (dynamical) (control) systems, *e.g.*, see [Won79]. In this section we highlight a few aspects that are relevant for this thesis. In particular, we comment on dynamical systems of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = Ax(t) \tag{2.2.10a}$$

and

$$x_{k+1} = Bx_k \tag{2.2.10b}$$

for some matrices $A, B \in \mathbb{R}^{n\times n}$.

First of all, by looking at the flow corresponding to (2.2.10a), that is, $\varphi^t(x_0) = e^{At}x_0$, it follows somewhat directly (up to messy computations due to non-trivial Jordan blocks) that for linear dynamical systems there is no difference between *local* and *global* convergence, also, there is no difference between *attractivity* and *asymptotic stability*. In fact, there is no difference between *asymptotic stability* and *exponential stability*. A slightly different viewpoint is that isolated equilibria of linear systems of the form (2.2.10a)—which must be 0—always have their vector field index satisfying $\mathrm{ind}_0(Ax) \in \{-1, 1\}$ (this follows from a directly computation, recall Section 2.2.3).

Despite not parametrizing a wealth of behaviour, linear systems are populair as they ought to approximate nonlinear systems locally. The Hartman-Grobman theorem [Rob95, Ch. 5] formalizes this and states that if the linearization of a nonlinear system around an equilibrium point is locally hyperbolic, then, the linearized system is qualitatively equivalent to the original nonlinear system, locally. This is not completely satisfactory as the neighbourhood where this result holds true is frequently unknown and nothing can be said for systems as simple as $\dot{x} = -x^3$. We return to *hyperbolicity* in Section 2.3.

We also remark that more recent work looks at linear time-varying (LTV) and switched linear systems (SLSs), to leverage our understanding of linear systems, applied to nonlinear problems, *e.g.*, see [Lib03, VTHK21].

So far, we have discussed the Euclidean finite-dimensional case, which is largely intuitive and indeed well-understood. In the *infinite*-dimensional case, intuition rapidly break down.

Example 2.2.6 (Stability of infinite-dimensional linear systems)**.** Here, we elaborate on [VDKS93, Thm. 4.11], which we point to as a source of various other examples. Throughout, let $n \geq 1$. Now, define the matrix $A_n \in \mathbb{R}^{n \times n}$ via

$$A_n := \begin{pmatrix} -1 & -2 & \cdots & -2 \\ 0 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -2 \\ 0 & \cdots & 0 & -1 \end{pmatrix}.$$

By the triangular structure of A_n it follows immediately that $\lambda_i(A_n) = -1$ for all $i \in [n]$ and thus $\dot{x} = A_n x$ corresponds to a dynamical system with the origin $0 \in \mathbb{R}^n$ being globally asymptotically stable. In particular, by the definition of Lyapunov stability this implies that $\|\exp(tA_n)\|_p < +\infty$ for any $t > 0$ since $\varphi^t(x_0) = \exp(tA_n)x_0$. Such a bound is of interest since it directly relates to relative peaking phenomena through

$$\sup_{t \geq 0} \sup_{x_0 \in \mathbb{S}^{n-1}} \frac{\|\varphi^t(x_0)\|_p}{\|x_0\|_p} = \sup_{t \geq 0} \|\exp(A_n t)\|_p.$$

Interestingly, despite the structure of A_n, $\lim_{n \to +\infty} \|\exp(A_n t)\|_\infty = +\infty$ for any $t > 0$. The intuition is that the operator A_∞ does not map ℓ_∞ to ℓ_∞. To make this precise, define a *shift* matrix $S_n \in \mathbb{R}^{n \times n}$ via $S_{n,ij} := \delta_{i+1,j}$, that is, S_n is of the form

$$S_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ & 0 & 1 & \ddots & \\ \vdots & & 0 & \ddots & 0 \\ & & & \ddots & 1 \\ 0 & & \cdots & & 0 \end{pmatrix}.$$

The matrix S_n is *nilpotent*, meaning that $S_n^k = 0$ for sufficiently large $k \in \mathbb{N}_{\geq 0}$. In this case $S_n^k = 0$ for any $k \geq n$. Also, fix some $t > 0$ and define the map $\mathbb{C} \ni z \mapsto f(z) := \exp(-t(1+z)(1-z)^{-1}) \in \mathbb{C}$. Indeed, f has merely a pole at $e^{i0} \in \mathbb{C}$ such that $f \in H(\mathbb{C} \setminus \{e^{i0}\}; \mathbb{C})$ (holomorphic away from e^{i0}). This means that f admits a series representation around, for instance, $0 \in \mathbb{C}$ on the open complex unit disk, that is, $f(z) = \sum_{j=0}^{\infty} a_j(t) z^j$ for some t-dependent sequence $(a_j(t))_{j \in \mathbb{N}_{\geq 0}}$. Now see that $f(e^{i\theta}) = \exp(-it/\tan(\theta/2))$ such that $\lim_{\theta \to 0} f(e^{i\theta})$ fails to exist indeed. All of this tells us that $\sum_{j=0}^{\infty} |a_j(t)| = +\infty$. The claim is concluded by overloading f and observing that $-(I_n + S_n) = A_n(I_n - S_n)$ such that $\exp(tA_n) = f(S_n) = \sum_{j=0}^{n-1} a_j(t) S_n^j$ and $\|\exp(tA_n)\|_\infty = \sum_{j=0}^{n-1} |a_j(t)|$.

Next, we provide more commentary on discrete-time systems and stability, central to Chapter 3.

Schur stable matrices

A matrix $B \in \mathbb{R}^{n\times n}$ is said to be ***Schur*** stable when $\rho(B) < 1$. Consider such a matrix B in the context of a linear dynamical system, *e.g.*, (2.2.10b). Now if B is diagonalizable through $B = T\Lambda T^{-1}$, it immediately follows that $\|x_{k+1}\|_2 \leq \kappa(T)\rho(B)^k\|x_1\|_2$ and thus asymptotic stability of the origin readily follows. As usual, we remark that if B is not diagonalizable, the computation is more involved. Indeed, as in the continuous case, *exponential* stability also readily follows as we simply translate terms of the form α^k, with $\alpha \in (0,1)$ to $e^{-\beta k}$ with $\beta = -\log(\alpha) > 0$. For a more *quantitative* notion of stability, that is, (τ,ρ)-stability, we refer to Definition 3.1.11.

Regarding Lyapunov functions, we point out that since Lyapunov functions can always we chosen to be quadratic, we end up with the so-called "*discrete Lyapunov equation*":

$$B^{\mathsf{T}}PB - P + Q = 0. \tag{2.2.11}$$

for some $P, Q \in \mathcal{S}^n_{\succeq 0}$. As (2.2.11) is linear (affine) in P, we can solve for P when B is Schur and Q is fixed. To see this, it follows that

$$B^{\mathsf{T}}PB - P = -Q \iff (B^{\mathsf{T}} \otimes B^{\mathsf{T}} - I_{n^2})\mathrm{vec}(P) = -\mathrm{vec}(Q),$$

for "$\otimes$" denoting the Kronecker product and $P \mapsto \mathrm{vec}(P) \in \mathbb{R}^{n^2}$ being standard "column stacking", with $(B^{\mathsf{T}}\otimes B^{\mathsf{T}} - I_{n^2})$ being invertible when B is Schur. For a detailed exposition of (2.2.11) and related equations, consider [LR95] or [vS21, Ch. 22].

We end this subsection with a comment on mere stability. The set of Schur (asymptotically stable) matrices is open (the spectral radius is a continuous map), unbounded (rescale an off-diagonal term) and nonconvex (combinations of upper- and lower-triangular matrices are common counterexamples). The set of matrices that are merely stable (*i.e.*, such that 0 is Lyapunov stable under (2.2.10b)) is more deceptive. For instance, consider the matrices B_1 and B_2 given by

$$B_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{pmatrix}, \tag{2.2.12}$$

where $\varepsilon \in (0,1)$ is arbitrarily small. Both matrices have their spectral radius equal to 1, so one might jump to the conclusion that they correspond to stable matrices. However, now consider the limits $B_i^\infty := \lim_{k\to\infty} B_i^k$:

$$B_1^\infty = \begin{pmatrix} 1 & \infty \\ 0 & 1 \end{pmatrix}, \quad B_2^\infty = \begin{pmatrix} 1 & \varepsilon^{-1} \\ 0 & 0 \end{pmatrix}. \tag{2.2.13}$$

This means that the spectral properties are not enough and one needs to discuss *marginal stability*. In fact, the space of stable matrices is *neither* open nor closed [GKS19].

2.2.6 Dynamical control systems on topological spaces

The state space framework as illustrated above can be generalized from $\mathbb{R}^n$ to smooth manifolds and when working with flows directly, to topological Hausdorff spaces. The benefit of such a generalization is largely the contribution to our understanding. These studies usually reveal what we can do and most importantly, *why*. The price to pay for this generality is that these studies are largely qualitative, that is, explicit models are rarely available and since we are not imposing metrics, we have to do with asymptotic- instead of exponential notions of stability. The key difference, between working over $\mathbb{R}^n$ and some space that is not contractible, *e.g.* the circle $\mathbb{S}^1$, is that the topology of the space restricts dynamical behaviour, we saw this in Example 1.1.2. It is interesting to note that, intuitively, the stability notions and even the Lyapunov stability theory largely extends.

In this section we only scratch the surface and introduce some notation. We elaborate in Chapter 6.

First, there are several generalizations of the standard "Euclidean dynamical control system": $\dot{x} = f(x,u)$. Generalization of the following form are still actively studied, and doing so, Kvalheim recently generalized [Kva23] the seminal necessary conditions by Brockett [Bro83], Coron [Cor90] and Mansouri [Man07, Man10]. For more context, see also [JM23, Ch. 6]. For simplicity, we assume our manifold to be C^∞-smooth (simply, "smooth" from now on).

A ***continuous control system*** [WVdS82, Def. 6], [NvdS90, TP05, KK22] is the triple $\Sigma = (\mathsf{M}, F, \mathcal{U})$, consisting of a smooth manifold M, a topological space $\mathcal{U}$, a continuous surjective map $\pi_u : \mathcal{U} \to \mathsf{M}$, the canonical projection $\pi_x : T\mathsf{M} \to \mathsf{M}$ and a continuous fiber-preserving map $F : \mathcal{U} \to T\mathsf{M}$ such that the following diagram (solid lines) commutes:

$$\begin{tikzcd}
\mathcal{U} \arrow[rr, "F"] \arrow[dr, "\pi_u"] & & T\mathsf{M} \arrow[dl, "\pi_x"] \\
& \mathsf{M} \arrow[ul, dotted, bend left, "\mu"] &
\end{tikzcd}$$

Available inputs at $p \in \mathsf{M}$ are characterized by the sections (dashed line) $\Gamma(\mathcal{U})$, *i.e.*, the continuous maps $\mu : \mathsf{M} \to \mathcal{U}$ such that $\pi_u \circ \mu = \mathrm{id}_{\mathsf{M}}$. Hence, the closed-loop vector field is of the form $F \circ \mu : \mathsf{M} \to T\mathsf{M}$. For example, the populair input-affine framework corresponds to $\pi_u : \mathcal{U} \to \mathsf{M}$ being a *vector* bundle. Effectively, what is achieved, is that necessary conditions from global stabilization via feedback, can be studied through the existence of *sections*.

For instance, in aerospace, constraints in the engine thrust can result in input constraint sets that look like homotopy spheres [MRS$^+$22, p. 47] (*e.g.*, constraints are of the form $\underline{\rho} \leq \|u\|_2 \leq \overline{\rho}$). Hence, in that case it is far from obvious that a global section exists, *e.g.*, consider the *Hopf fibration*.

We continue along these lines in Chapter 6.

2.2.7 Further examples

An unwritten rule is that every work on dynamical systems must contain something on the pendulum. For us, this example helps in illuminating that our idealized notion of stability (asymptotic stability) appears where one would expect it.

Example 2.2.7 (The (mathematical) pendulum)**.** Normalizing mass, gravity and the length of the rod, the potential energy of the pendulum is given by $U(\theta) = 1 - \cos(\theta)$ whereas the kinetic energy becomes $T(\theta, \dot{\theta}) = \dot{\theta}^2$, see Figure 2.2. Hence, the Lagrangian

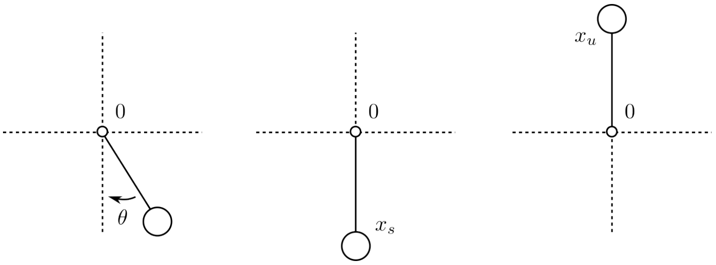


Figure 2.2: Example 2.2.7.

is simply $L(\theta,\dot{\theta}) = \dot{\theta}^2 - (1-\cos(\theta))$ such that the Euler-Lagrange equations yield $\ddot{\theta} = -\frac{1}{2}\sin(\theta)$ for the equations of motion. Now, we suppose that there is friction (again, normalizing any constants) such that the final equation becomes $\ddot{\theta} = -\frac{1}{2}\sin(\theta) - \dot{\theta}$, or as in first-order form with $(x_1, x_2) = (\theta, \dot{\theta})$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ -\frac{1}{2}\sin(x_1(t)) - x_2(t) \end{pmatrix}. \tag{2.2.14}$$

Although the notation for (2.2.14) might reminisce of $\mathbb{R}^2$, the dynamical system should be understood as living on $\mathbb{S}^1 \times \mathbb{R}$ (see [JM23, Fig. 1.2]). As the cylinder is not contractible, we cannot have *global* asymptotic stability and indeed we have two equilibrium points being $x_s = (0,0)$ and $x_u = (\pi, 0)$, see Figure 2.2. Linearizing (2.2.14) around these points reveals they are both hyperbolic such that we can infer *local* stability properties. Indeed, x_s is locally asymptotically stable and x_u is unstable (a saddle). Exactly as intuition would tell us. To make *almost* global statements, one can use the energy $E = T + U$ as a Lyapunov function and use the *Krasovskii-LaSalle invariance principle* [Kha02, Sec. 4.2] to conclude on *almost* global asymptotic stability of x_s.

Example 2.2.8 (Solving a linear system)**.** Suppose we want to solve the linear system $Ax = b$ for x, assuming that $A \in \mathsf{GL}(n, \mathbb{R})$. We can construct the negative gradient flow (with respect to the objective $f(x) = \frac{1}{2}\|Ax - b\|_2^2$):

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = -A^{\mathsf{T}}(Ax(t) - b).$$

Now consider the Lyapunov function $V = f$ ($V(x) = 0 \iff x = A^{-1}b$, $V(x) \geq 0$) and see that since $-\|\cdot\|_{AA^{\mathsf{T}}}^2 \leq -\|\cdot\|_2^2 \sigma_{\min}(A)^2$ we have that $\dot{V} \leq -V\sigma_{\min}(A)^2$ such that (by *Grönwall's inequality*)

$$\tfrac{1}{2}\|Ax(t) - b\|_2^2 = V(x(t)) \leq e^{-\sigma_{\min}(A)^2 t} V(x(0)) = \tfrac{1}{2} e^{-\sigma_{\min}(A)^2 t}\|Ax(0) - b\|_2^2.$$

Let us apply this to the discrete Lyapunov equation (2.2.11). In vectorized form, we aim to solve $(B^{\mathsf{T}} \otimes B^{\mathsf{T}} - I_{n^2})\mathrm{vec}(P) = -\mathrm{vec}(Q)$ for P. Unrolling vectorizations and exploiting properties of the Kronecker product (*e.g.*, $(B^{\mathsf{T}} \otimes B^{\mathsf{T}} - I_{n^2})^{\mathsf{T}} = (B \otimes B - I_{n^2})$ and $\mathrm{vec}(BXA^{\mathsf{T}}) = (A \otimes B)\mathrm{vec}(X)$) we readily find that the following (affine) differential

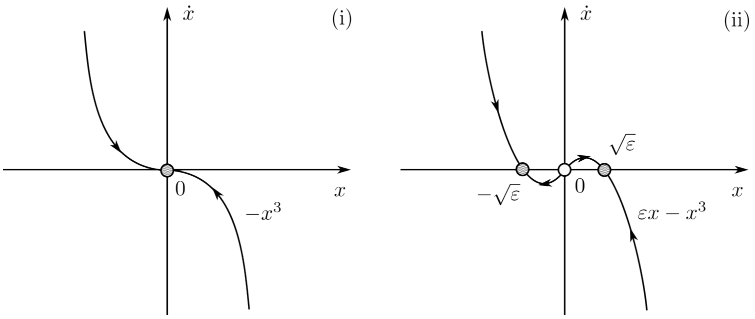


Figure 2.3: Phase portrait of (i) $\dot{x} = -x^3$ and (ii) $\dot{x} = \varepsilon x - x^3$.

equation

$$\frac{\mathrm{d}}{\mathrm{d}t}P(t) = -B(B^\mathsf{T}P(t)B - P(t) + Q)B^\mathsf{T} + B^\mathsf{T}P(t)B - P + Q$$

solves (2.2.11), in an exponentially *stable*[12] manner.

2.3 Structural stability

Stability is sometimes, and incorrectly, said to be important since it allows for modelling errors (one should discuss feedback instead [Kha16]). Consider the non-hyperbolic system $\dot{x} = -x^3$, with $0 \in \mathbb{R}$ being globally asymptotically stable. Indeed adding an arbitrarily small term of the form εx for some $\varepsilon > 0$ destroys stability, *e.g.*, consider Figure 2.3. The exact opposite can be done for $\dot{x} = x^3$, from 0 being globally unstable, to 0 being locally stable under $\dot{x} = x^3 - \varepsilon x$. The crux here is the lack of *structural stability*, which is as shown evidently independent of equilibria being stable or not.

Stability in the sense of Lyapunov is essentially about perturbations in the state (initial condition), away from an attractor. Structural stability is about perturbations in the system description itself. If somehow "most systems" would be structurally stable, then, sufficiently small modelling errors are forgiven. This motivated the likes of Poincaré, Birkhoff, Andronov, Pontryagin, Lefschetz, Smale, Thom, Peixoto(s) and many others to study this topic.

To define structural stability, we need to define *topological equivalence*.

2.3.1 Topological equivalence

Regarding topological equivalence in the context of linear control systems, [Wil80] Willems stated in 1980 that "*Because of the obvious ... practical importance of these concepts,*

[12] It is tempting to say "exponentially fast", but towards such a claim, one must talk about how exactly the differential equation is integrated. It is interesting to note the typical iterative scheme $P_{k+1} = B^\mathsf{T}P_kB + Q$ can be analyzed through contractions as highlighted in Section 2.1.2. To that end one does need to find an appropriate norm $\|\cdot\|_P$ (or change of coordinates), to link norms to spectral radii, this is also used in the proof of Proposition 3.1.6.

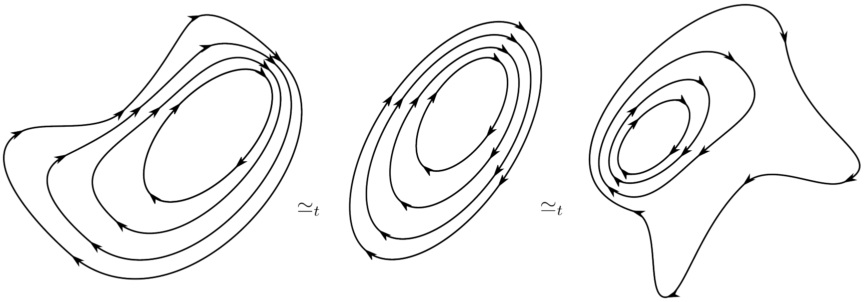


Figure 2.4: We speak of *topological equivalence*, denoted $\simeq_t$, when two phase portraits are homeomorphic, while preserving the direction of time.

... there is no doubt that they will become standard vocabulary among practitioners." Although Polderman [Pol87] provided many additional insights a few years later, there has been little recent follow-up work on topological properties of control systems.

In this thesis we will frequently study the qualitative behaviour of a dynamical system. We capture this via the notion of *topological equivalence*[13].

The advantage is that we can greatly simplify the study and pass from a continuum of systems to just a set of classes which is usually finite. The topological classification of linear flows and maps was pioneered by Kuiper and co-workers [KR73]. A range of these ideas were later extended into the system theoretic direction [Wil80] and some attention has been given to topological feedback linearization, *e.g.,* see [Če95], plus there is recent interest in structural stability in the context of systems biology [BCCG20]. We believe this lack of interest is in part due to the fact that topological classification seems to be rather coarse, and frankly, rather difficult. In this thesis we hope to provide some concrete motivation and perhaps contribute to re-instigating this beautiful field.

Topological equivalence in linear dynamical systems

This part mainly highlights the work of Kuiper and Robbin [Rob72, KR73], looking at dynamical systems of the form

$$\mathcal{V} \ni x \mapsto f(x) \in \mathcal{V} \tag{2.3.1}$$

where the *time-one map* $f : \mathcal{V} \to \mathcal{V}$ is a linear endomorphism[14] over a finite-dimensional topological vector space $\mathcal{V}$. Then, we say that two dynamical systems are ***topologically equivalent*** when their phase-portraits are homeomorphic[15], preserving the direction of time [Rob95], [Kuz04, Ch. 2] (see Figure 2.4). The purpose of this tool is to characterize classes of dynamical systems giving rise to *qualitatively* similar behaviour. This notion appears in the celebrated Hartman-Grobman theorem [Rob95, Thm. 5.3] and is the key

[13] To avoid confusion, we would like to stress that the vast majority of work on *topology* in the context of control, relates to *network* topology, which is *not* what this, or any other section, is about.

[14] For example, for $\mathcal{V} = (\mathbb{R}^n, +)$ linear maps are endomorphisms as they preserve the group-structure of $\mathbb{R}^n$. If these maps are invertible they are called *automorphisms*.

[15] To remind the reader, two topological spaces $\mathcal{X}$ and $\mathcal{Y}$ are *homeomorphic* when there exists a continuous bijective map $\varphi : \mathcal{X} \to \mathcal{Y}$ with a continuous inverse. Such a map φ is called a *homeomorphism*.

concept in *bifurcation theory* [Kuz04], which studies precisely this qualitative change. In fact, we speak of a *bifurcation* when a system, after some parameter change, is not (locally) topologically equivalent anymore to its initial configuration. The notion of topological equivalence has an explicit characterization in the discrete-time setting, there it coincides with the two time-one maps being *conjugates*.

Definition 2.3.1 (Topological equivalence)**.** *Two endomorphisms* $f:\mathcal{V}\to\mathcal{V}$ *and* $g:\mathcal{W}\to\mathcal{W}$ *over topological vector spaces* $\mathcal{V}$ *and* $\mathcal{W}$ *are topologically equivalent, denoted* $f\simeq_t g$*, if and only if there exists a homeomorphism* $\varphi:\mathcal{V}\to\mathcal{W}$ *such that* $g\circ\varphi=\varphi\circ f$*, that is, the diagram*

$$\begin{tikzcd}
\mathcal{V} \arrow[r, "f"] \arrow[d, "\varphi"'] & \mathcal{V} \arrow[d, "\varphi"] \\
\mathcal{W} \arrow[r, "g"] & \mathcal{W}
\end{tikzcd} \tag{2.3.2}$$

commutes.

Instead of Definition 2.3.1 one encounters the stronger notion of *linear equivalence* more often. Indeed, for any $T\in\mathsf{GL}(n,\mathbb{R})$ and $A\in\mathbb{R}^{n\times n}$ the diagram (2.3.2) commutes for $f(x)=Ax$, $g(y)=TAT^{-1}y$, *i.e.*, $\varphi(x)=Tx$. However, the quotient space under linear equivalence is still a continuum, whereas from a topological point of view, there are for example just 7 scalar linear systems (maps) [KR73, Prop. 1.5]. Hence, one can think of Definition 2.3.1 as a weaker change of coordinates. However, one should merely assume that the map φ is a homeomorphism, assuming φ to be a diffeomorphism implies that φ is linear [Rob95, Prop. 6.1].

To clarify Definition 2.3.1, examine the example from [KR73] given by $f(x)=2x$ and $g(y)=8y$. Although their eigenvalues are clearly different, *qualitatively*, f and g are the same. Indeed, $f\simeq_t g$ since $\varphi(x)=x^3$ is the corresponding homeomorphism. Observe that although $\varphi\in C^{\omega}(\mathbb{R})$, the inverse $\varphi^{-1}(x)=\sqrt[3]{x}\in C^0(\mathbb{R})$.

Then, Kuiper proposes several conditions on the (generalized) eigenspaces of the linear endomorphims f and g to show topological equivalence. We will mainly focus on two of them; stability, but most and for all: *orientation*.

Definition 2.3.2 (Orientation of linear maps)**.** *We call a linear automorphism* f *orientation preserving when the sign of the signed volume of the unit cube is invariant under the map* f*. This preservation (of orientation) is denoted by* $\mathrm{or}(f)=1$*, otherwise* $\mathrm{or}(f)=-1$*.*

For example, given $x\mapsto f(x)=Fx$ and $y\mapsto g(y)=Gy$ with $F\in\mathsf{GL}^+(n,\mathbb{R})$ and $G\in\mathsf{GL}^-(n,\mathbb{R})$, then, $\mathrm{or}(f)=1$ while $\mathrm{or}(g)=-1$. The intuitive reason why stability and orientation show up is as follows. Given time-one maps f and g, the trajectories they induce are homeomorphic when there is a homeomorphism φ such that $f=\varphi\circ g\circ\varphi^{-1}$. The link with stability follows from the observation that this definition implies that[16] $f^n=\varphi\circ g^n\circ\varphi^{-1}$ must hold for any n, that is, the direction of time is enforced. As φ will be either orientation preserving or reversing, Definition 2.3.1 implies that $\mathrm{or}(f)=\mathrm{or}(g)$. In fact, orientation is a *topological invariant* [Lee11, Ch. 6, 10], such that for two automorphisms f and g, $f\simeq_t g$, only if $\mathrm{or}(f)=\mathrm{or}(g)$. When f is not an automorphism,

[16] Where f^n is understood as $f\circ f\circ\cdots\circ f$, that is, applying the operator n times.

then, the orientation of f is only defined over its (invariant) automorphic domain. In the scalar case, the orientation can be interpreted as spring vs. damper-like behaviour, in higher dimensions one can think of the map relating to a *flow* or not, see Remark 2.3.1 below.

At last we state the main tool of this section, which supplies us with an easy sufficient condition to assess if $f \simeq_t g$ (in the case of linear maps).

Theorem 2.3.3 (Topological equivalence of asymptotically stable systems [Rob95, Thm 9.2, p. 117]). *Let $x \mapsto f(x) = Fx$ and $y \mapsto g(y) = Gy$ be asymptotically stable linear automorphisms on $\mathbb{R}^n$. Moreover, let $X(t)$ parametrize a path in $\mathsf{GL}(n,\mathbb{R})$, continuously depending on $t \in [0,1]$, such that $X(0) = F$ and $X(1) = G$, then, $f \simeq_t g$.*

The key in Theorem 2.3.3 is to demand that F and G are members of the same path-connected component[17] of $\mathsf{GL}(n,\mathbb{R})$, hence the maps f and g have the same orientation.

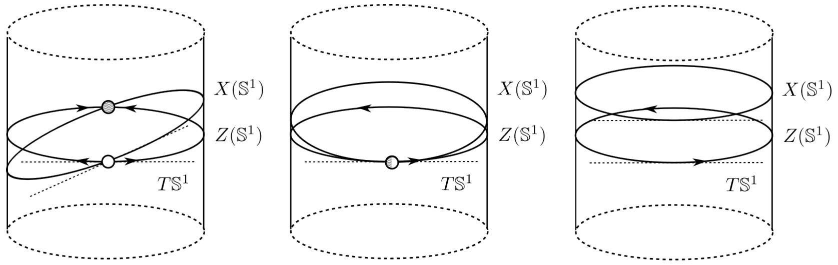
Remark 2.3.1 (Orientation in the wild). Orientation might seem like an esoteric property, especially for simple linear dynamical systems, but it makes its appearance especially often when one discretizes a continuous-time problem. For instance, sampling any solution to $\dot{x} = Ax$ yields the time-one map $x \mapsto \exp(sA)x$ for some sampling step $s > 0$. It can be seen from $\det(\exp(A)) = \exp(\mathrm{Tr}(A))$ that this map is always orientation-preserving. It is known that this observation extends to non-linear systems [Arn88], *e.g.,* the same holds (locally) for any Poincaré return map, which follows from the Liouville formula (*cf.* [Kuz04, Ch. 1]). This means that if one imposes a control law on these discretized maps which flips the orientation, then, this resulting map could never relate to some continuous flow.

Remark 2.3.2 (Continuous-time topological equivalence). Definition 2.3.1 allows to work with topological equivalence in the case of discrete-time systems (*cf.* Section 4.1). Regarding continuous-time systems, suppose we have two vector fields $\dot{x} = F(x)$, $\dot{x} = G(x)$, for simplicity on $\mathbb{R}^n$, and we want to understand their qualitative equivalence. Suppose we attach flows to those vector fields, say φ to F and ψ to G. Then, if there is a homeomorphism h such that $\varphi^t = h^{-1} \circ \varphi^t \circ h$, then, we speak of a *topological conjugacy cf* Definition 2.3.1. In the continuous-time case, this is stronger than topological equivalence, as a continuous, monotone, reparametrization of time does not break topological equivalence. That is, we speak of *topological equivalence*, we there is a map $\alpha \in C^0(\mathbb{R} \times \mathbb{R}^n;\mathbb{R})$, $(t,x) \mapsto \alpha(t,x)$, monotonically increasing in t for any fixed x, such that $\varphi^{\alpha(t,\cdot)} = h^{-1} \circ \varphi^t \circ h$ [Rob95, p. 115]. Evidently, as we appeal to flows, this is hard to practically work with. In fact, the previous remarks can be also be seen as further motivation for Lyapunov theory, that is, to understand qualitative dynamical behaviour without resorting to solutions (flows).

Comments on structural stability

Having briefly discussed topological equivalence, we can comment on structural stability. Simply put, a dynamical system is *structurally stable* when a *perturbation* to the system

[17] The general linear group $\mathsf{GL}(n,\mathbb{R})$ has two path-connected components, which follows from $\det : \mathbb{R}^{n\times n} \to \mathbb{R}$ being continuous and $\mathbb{R}^{n\times n} \setminus \mathsf{GL}(n,\mathbb{R}) = \det^{-1}(0)$.



(i) (ii) (iii)

Figure 2.5: Three vector fields on $\mathbb{S}^1$: (i) X having two hyperbolic equilibrium points and clearly $X \pitchfork Z(\mathbb{S}^1)$; (ii) X having a single index-0 equilibrium point and indeed X fails to be transversal to $Z(\mathbb{S}^1)$ at this point; (iii) X having no equilibrium points and trivially $X \pitchfork Z(\mathbb{S}^1)$. The visualization is possible by $T\mathbb{S}^1 \simeq_t \mathbb{S}^1 \times \mathbb{R}$.

preserves its *topological class*. Evidently, the type of perturbation must be made precise. As we will look several times at linear dynamical systems in this thesis, frequently motivated as being local representations of nonlinear dynamical systems, we discuss the structural stability of these local linear representations. We do this from a rather general viewpoint, that is, to see how the linearization behaves under perturbation to the nonlinear system directly.

For simplicity, we only consider *smooth* (*i.e.*, C^∞-smooth) spaces. Let M be a compact, smooth manifold. Then, a vector field $X \in \Gamma^r(T\mathsf{M})$ is ***C^r-structurally stable*** (typically, C^1) when $\Gamma^r(T\mathsf{M})$ contains an open neighbourhood U of X, with respect to the C^r-topology (Whitney topology), such that all vector fields in U are topologically equivalent. We recall that the C^r-topology is a vast generalization of, for instance, the topology generated by the C^k-norm on $C^k([0,1])$, *i.e.*,

$$\|f\|_{C^k} = \textstyle\sum_{i=0}^{k} \sup_{x\in[0,1]} |f^{(i)}(x)|.$$

For the details we point to [Hir76, PDM82]. We need another technical notion. Let $F : \mathsf{M} \to \mathsf{N}$ be a smooth map between smooth manifolds and let $\mathsf{S} \subseteq \mathsf{N}$ be a smooth submanifold. We say that F is *transversal* to S, denoted $F \pitchfork \mathsf{S}$, when $\mathrm{Im}(DF_p) + T_{F(p)}\mathsf{S} = T_{F(p)}\mathsf{N}$ for all p such that $F(p) \in \mathsf{S}$. This notion of transversality allows for generalizations of critical point theory, *i.e.*, F is transversal to its regular values. Thom provided several powerful results, for instance, suppose in addition to the above that S is closed and let $k \geq 1$, then, the subset $\{F \in C^k(\mathsf{M};\mathsf{N}) : F \pitchfork \mathsf{S}\}$ is open and dense, *e.g.*, see [PDM82, p. 25]. This is intuitive, when drawing two lines on a piece of paper, they will be "almost surely" transversal (note, if $F^{-1}(\mathsf{S}) = \emptyset$, then F is trivially transversal to S). Now, we say that an equilibrium point $p^\star$ of $X \in \Gamma^r(T\mathsf{M})$ is *simple* when $DX_{p^\star}$ is an isomorphism. It is not too surprising that $p^\star$ is simple if and only if the vector field $X : \mathsf{M} \to T\mathsf{M}$ is transversal to the zero section at $p^\star$, that is, the map $p \mapsto (p, X(p)) \in T\mathsf{M}$ should be transversal to $Z(\mathsf{M}) = \{(p,0) \in T_p\mathsf{M} : p \in \mathsf{M}\}$ at $p^\star$, see also Figure 2.5. Then, by exploiting Thom's transversality theorem, one can now show that $\{X \in \Gamma^r(T\mathsf{M}) : X \pitchfork Z(\mathsf{M})\}$ is open and dense in $\Gamma^r(T\mathsf{M})$ [PDM82, p. 56]. In fact, one can show that vector fields with their equilibrium points being *hyperbolic*, are open and dense in that particular set as well [PDM82, p. 58]. Now one can continue, along the lines of Theorem 2.3.3, and show that hyperbolic equilibria of $X \in \Gamma^r(T\mathsf{M})$ are locally structurally stable [PDM82, p. 67]. This means that in a fairly general sense, hyperbolic equilibria are structurally stable. Hence, the study of linear dynamical systems as local representations can be rigorously justified, however, *how local* is up for debate.

2.4 Other notions of stability

There are many interesting notions of stability we did not cover here.

First of all, we did not cover *input-to-state stability* (ISS) and "*practical stability*", see [Son01] for an introduction to ISS by Sontag himself. We also barely touched upon *time-varying* systems and hence did not cover any of the relevant stability notions in that context, most of the early texts already include non-autonomous systems, *e.g.*, see [Hah67], see [JL14] for further technical discussions on time-varying vector fields. Another stability notion we did not cover is *finite time* stability. To put our further work on asymptotic stability in the right perspective, we provide an explicit example.

Example 2.4.1 (Finite time stability)**.** Consider some Hurwitz matrix $A \in \mathbb{R}^{n\times n}$ and the linear ODE $\dot{x} = F_A(x) := Ax$. We know that the origin must be exponentially stable under this dynamical system. If instead, we consider $\dot{x} = F_1(x) := Ax/\|Ax\|_2$ with $F_1(0) := 0$, what would happen? We point in the exact same direction, yet, the magnitude of the tangent vectors is constant on $\mathbb{R}^n \setminus \{0\}$. First, as dynamical systems of this form fail to be continuous, we resort to analysis *in the sense of Filippov.* For the details we point to his book [Fil88] and in particular to Chapter 2 for the intuition behind his convexification method. In our case, we can study F_1 through the differential inclusion

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) \in F(x(t)) = \begin{cases} F_1(x) & \text{if } x \neq 0 \\ \cap_{\delta>0} \operatorname{cl}\operatorname{conv} F_1(\mathbb{B}^n_\delta(0)\setminus\{0\}) & \text{otherwise.} \end{cases} \tag{2.4.1}$$

Since $\|F(x)\|_2 \leq 1$, solutions to (2.4.1) (in the integral sense) can be guaranteed to exist [Kun00, Thm. 2.2.1] (this is why we consider $Ax/\|Ax\|_2$ and not $Ax/\|Ax\|_2^2$). Now, to study finite time stability of the origin under F_1, we aim for finding a Lyapunov function V, together with constants $k > 0$ and $\alpha \in [0, 1)$ such that under (2.4.1) we have that $\dot{V} \leq -kV^{\alpha}$ on $\mathbb{R}^n \setminus \{0\}$. The intuition follows from studying the scalar ODE $\dot{x} = -kx^{\alpha}$, for references we point to [MP05].

Now, let $x \mapsto V(x) = \langle Px, x\rangle$ be the Lyapunov function corresponding to F_A. In particular, we know there is a $Q \succ 0$ such that $PA + A^{\mathsf{T}}P \preceq -Q$. Now see that the

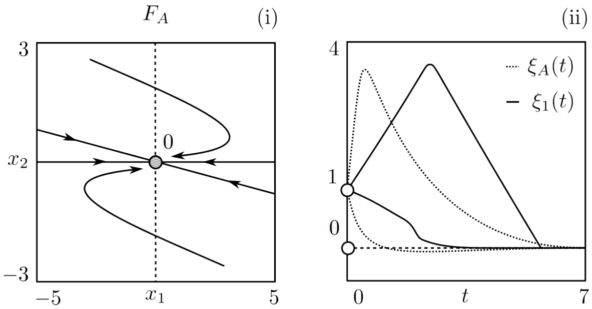


Figure 2.6: Example 2.4.1: (i) the vector field F_A for (2.4.2); and (ii) integral curves $t \mapsto \xi(t)$, emanating from $(1,1) \in \mathbb{R}$ at $t=0$, for both F_A and F_1.

following holds for any $x \in \mathbb{R}^n \setminus \{0\}$:

$$\begin{aligned}\langle \nabla V(x), F(x)\rangle = \frac{1}{\|Ax\|_2}\langle (PA + A^\mathsf{T} P)x, x\rangle &\leq -\frac{1}{\|Ax\|_2}\langle Qx, x\rangle \\ &\leq -\frac{\|x\|_2^2}{\|x\|_2\|A\|_2\|Q^{-1}\|_2} \\ &= -\frac{1}{\|A\|_2\|Q^{-1}\|_2}\frac{(\|x\|_2^2\|P\|_2)^{1/2}}{\|P\|_2^{1/2}} \\ &\leq -kV(x)^\alpha\end{aligned}$$

for $k = 1/(\|A\|_2\|Q^{-1}\|_2\|P\|_2^{1/2})$ and $\alpha = \frac{1}{2}$. Them, to compare F_A against F_1 we consider

$$A = \begin{pmatrix} -1 & 10 \\ 0 & -2 \end{pmatrix} \tag{2.4.2}$$

and show F_A, plus the convergence over time, in Figure 2.6. We see that, although finite time stability is very appealing, it is, for all practical purposes, not uniformly “better” than exponential- or even asymptotic stability. Of course, it does allow for the computation of settling times, which is of great interest.

We also did not touch on stability through optimal- or model predictive control (MPC) (although we will mention LQR several times in the thesis, see Chapters 3-4.). Especially in optimal control, care needs to be taken with discount factors [PBND16], which is the *de facto* paradigm in reinforcement learning. Stability in the context of MPC is also still an very active research area. Recently, the stability analysis of MPC—towards *necessary* and sufficient conditions, is generalized through the lens of dissipativity and the maximum principle [Fau21].

Another important notion is *incremental stability*. Throughout, we always assume some extra knowledge regarding our system in that we assume to know something about our attractor[18]. Suppose you do not have this knowledge, then, incremental stability

[18] Although, we like to emphasize that we are frequently after necessary conditions in that we start from a desirable attractor and then ask if a corresponding dynamical system exists.

allows you to still conclude on some notion of stability, that is, when trajectories are asymptotically (usually, exponentially) attracted to each other (hence the name). Or as Lohmiller and Slotine put it, “the initial conditions are exponentially forgotten” [LS98, p. 4]. Incremental stability is elegantly captured via *contraction analysis*. Very briefly, given a C^1-smooth dynamical system of the form $\dot{x} = F(x)$, construct the variational equation $\delta\dot{x} = \partial_x F(x)\delta x$ and see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \delta x(t), \delta x(t)\rangle = \left\langle \left(\partial_x F(x(t)) + \partial_x F(x(t))^{\mathsf{T}}\right)\delta x(t), \delta x(t)\right\rangle .$$

Hence, bounds on the symmetric part of the Jacobian $\partial_x F(x)$ allow for an exponentially fast contraction. For more details, see [LS98].

We also barely touched upon *hybrid-* or *switched systems* [Lib03]. The stability notions in that context are similar, yet, care needs to be taken in handling a lack of regularity [GST12], as we briefly saw in Example 2.4.1. Common approaches are to work with piecewise “regular”, multiple- or path-complete Lyapunov functions, instead of a common smooth Lyapunov function directly (we were lucky in Example 2.1.3), *e.g.*, see [Bra98, Joh03, AJPR14].

At last, we also did not touch upon stability of *systems*, which we cover in some more detail below.

2.4.1 Comments on stability of systems

Although we are mostly concerned with *dynamical systems*, that is, systems of the form $\dot{x} = F(x)$ or $x_{k+1} = G(x_k)$, we do briefly mention several viewpoints when working with non-trivial outputs. The key difference with other notions of stability is that now the dynamical (differential/difference) perspective is *internal*, that is, we effectively consider “simply” maps from inputs to outputs.

There are many different notions of input-output stability, we follow [Kha02, Ch. 5-6], but only scratch the surface.

L_p-stability

Suppose that the input $t \mapsto u(t) \in \mathbb{R}^m$ and output $t \mapsto y(t) \in \mathbb{R}^p$ are related through $y = H(u)$, that is, a map $H : L_p^e(\mathbb{R}^m) \to L_p^e(\mathbb{R}^p)$ (where the superscript e denotes that we deal with “extended” L_p-spaces, that is, we cannot constrain $y \in L_p(\mathbb{R}^p)$, *a priori*). Then we say, for instance, that H is ***L_p-stable*** when there is a class-$\mathcal{K}$ function[19] γ and a constant $\beta \geq 0$ such that

$$\|y_\tau\|_{L_p} \leq \gamma(\|u_\tau\|_{L_p}) + \beta \quad \forall \tau \in [0, \infty). \tag{2.4.3}$$

Indeed, when H is L_∞ stable, the input-output system is *bounded-input bounded-output* (BIBO) stable.

Now, given a linear time-invariant (LTI) (input-output) system of the form

$$\Sigma^{\mathrm{LTI}}_{\mathrm{i/o}} : \begin{cases} \dot{x} = & Ax + Bu \\ y = & Cx + Du, \end{cases} \tag{2.4.4}$$

[19] The function γ is said to be of class$-\mathcal{K}$ when $\gamma(0) = 0$ and $\gamma(s) > \gamma(t)$ for all $s > t$.

then clearly, input-output stability is not immediately related to stability of the state (internal stability), *e.g.*, uncontrollable unstable modes might live in the kernel of the observability matrix. Moreover, interpret (2.4.4) as a controlled system with u a disturbance signal and suppose that A is Hurwitz. It can be shown that for the transfer matrix $G(s)=C(sI-A)^{-1}B+D$ we have that the L_2-gain of (2.4.4) equals $\sup_{\omega\in\mathbb{R}}\|G(j\omega)\|_2=\|G(j\omega)\|_{H_\infty}$, which one might try to optimize towards desired performance (H_∞ control).

It should be evident that in general, it is hard to say anything about L_p stability of an input-output system. A fruitful approach relates again to Lyapunov stability theory and is called ***passivity***. A nonlinear input-output system

$$\Sigma_{\mathrm{i/o}}:\begin{cases}\dot{x}=&f(x,u)\\ y=&h(x,u),\end{cases}\tag{2.4.5}$$

with $f(0,0)=0$, $h(0,0)=0$, f and h both sufficiently regular, $x\in\mathbb{R}^n$ and $u,y\in\mathbb{R}^m$ is then said to be *passive* when there is a *storage function* $S\in C^1(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ such that $\langle u,y\rangle\geq\dot{S}(x)$. Now suppose that (2.4.5) is not just passive, but "*strictly output passive*", meaning in this case that there is a $\delta>0$ such that $\langle u,y\rangle\geq\dot{S}(x)+\delta\langle y,y\rangle$. One readily computes (see [Kha02, Lem. 6.5]) that

$$\|y_\tau\|_{L_2}\leq\tfrac{1}{\delta}\|u_\tau\|_{L_2}+\sqrt{\tfrac{2}{\delta}S(x(0))},$$

and thus, (2.4.5) is L_2-stable with a L_2-gain bound of $1/\delta$. This line of reasoning can be continued and one can even use passivity to reason about internal stability, which is not so surprising, since given a passive system with storage function S, we know that regarding stability of the origin under $\dot{x}=f(x,0)$ we can simply use $V=S$, since $\dot{S}(x)\leq 0$ $(u=0)$.

Stability through output feedback

Stability of systems is rather different from stability of dynamical systems, the same is true for stabilizability.

Consider (2.4.4). Typically, the internal states cannot be directly measured and as such a standard control problem is to regulate the system through output feedback. When (2.4.4) is *minimal*, this can always be done, but it is important to note that controllability just guarantees there is a stabilizing feedback of the form $u=Kx$, not $u=K'y$. Indeed, a stabilizing static output feedback is not guaranteed to exist when (2.4.4) is minimal.

Consider the single-input single-output (SISO) LTI system

$$\Sigma^{\mathrm{LTI}}_{\mathrm{i/o}}:\begin{cases}\dot{x}_1=x_2\\ \dot{x}_2=u\\ \;\;y=x_1.\end{cases}\tag{2.4.6}$$

Note, (2.4.6) is controllable and observable (simply compute, for the appropriate triple (A,b,c), the matrices $(b\,Ab)$ and $(c^\mathsf{T}\,(Ac)^\mathsf{T})$. One might wonder if there is a continuous output feedback $u=k(y)=k(x_1)$ that asymptotically stabilizes the origin of the internal system. Sontag provides an elegant Lyapunov-based counterargument to the existence of such a feedback (2.4.6) [Son98, Ex. 7.2.1], we elaborate.

First, reconsider (2.1.2). Then, since k is continuous and by assumption $k(x_1) = 0 \iff x_1 = 0$, we know that either (i) the sign of k agrees with x_1 or (ii) the sign of k is the opposite.

Regarding case (i), we use a topological argument to argue that the closed-loop behaviour cannot be stable. We know that since 0 is a saddle under F_+, the vector field index evaluates to $\mathrm{ind}_0(F_+) = -1$, see Figure 2.1, whereas for an asymptotically stable vector field this would be 1 instead. Now the index of the closed-loop system under k is not obvious, however, vector fields that "point in the same direction" have the same index (formalized in [JM23, Ex. 3.2]). Taking the inner product between $(x_2, k(x_1)$ and (x_2, x_1) yields $x_2^2 + x_1 k(x_1) > 0$ for all $(x_1, x_2) \neq 0$. Hence, the index of 0 under the closed-loop vector field will be -1 as well.

Then, regarding case (ii), we take a Hamiltonian[20] point of view and recover the example by Sontag. Recall that if we have a (normalized) harmonic oscillator, then the Hamiltonian is $H(p,q) = \frac{1}{2}p^2 + \frac{1}{2}q^2$, resulting in the center

$$\dot{q} = \frac{\partial}{\partial p} H(q,p) = p, \quad \dot{p} = -\frac{\partial}{\partial q} H(q,p) = -q,$$

that is, $\ddot{q} = -q$. Now, given that k must have the opposite sign of x_1 it readily follows that the closed-loop system must be of the form $\ddot{q} = -K(q)q$, for $K \in C^0(\mathbb{R}; \mathbb{R}_{\geq 0})$ with $K(0) = 0$. Differently put, K is a state-dependent spring constant. With this in mind, the corresponding Hamiltonian readily follows as $H(q,p) = \frac{1}{2}p^2 + \int_0^q K(s)s\mathrm{d}s$ *cf.* [Son98, Ex. 7.2.1]. By construction, the Hamiltonian is conserved along trajectories of the system and hence the origin cannot be asymptotically stable (as there are points (q,p) such that $H(q,p) \neq 0$).

Evidently, since (2.4.6) is minimal, we can stabilize the system, however, we must use *dynamic* feedback, that is, construct an observer $\dot{\widehat{x}} = (A+BK)\widehat{x} + L(C\widehat{x} - y)$ such that we recover the current state through the output and effectively apply standard static state feedback $u = K\widehat{x}$, see [Son98, Thm. 32]. For instance, we can pick $L = K^\mathsf{T} = (-1,-1) \in \mathbb{R}^2$.

2.4.2 Numerical stability

Very much related to stability through regularity, we briefly comment on *numerical stability*. Rigorously defining numerical stability in great generality is somewhat futile as it is algorithm (problem) dependent. Technically speaking, it is even system dependent. Yet, we do highlight what we will discuss in Chapter 5: *numerical cancellation*.

In general, we can think of numerical stability as statements of the following form [Hig02, p. 7], suppose we only work with scalars, let $u \mapsto h(u) = y$ be an algorithm with the numerical output evaluating to $\widehat{y}$. Now we say that the algorithm is ("*mixed*") stable

[20] We do not cover Lagrangian/Hamiltonian/Routhian mechnical systems and their stability. In general, those systems do exhibit significant structure. We point the reader to [AM08, Blo15] for the mechanical point of view and to [GFXFT23] for a modern exploitation of this structure in the context of neural networks.

when the following holds

$$h(u+\Delta u)=\widehat{y}+\Delta y,\quad |\Delta u|\leq \alpha|u|,\, |\Delta y|\leq \beta|y|$$

for sufficiently small $\alpha,\beta>0$. This is very similar to Lyapunov stability.

However, for this thesis, when it comes to numerical stability, we will be mostly interested in how *rounding errors* complicate or even obstruct the numerical implementation of algorithms. As a typical example, consider the standard rules to compute the roots of a quadratic equation $\alpha x^2+\beta x+\gamma=0$. For instance, a triple (α,β,γ) with $\beta^2\approx 4\alpha\gamma$, yet, $\beta^2\neq 4\alpha\gamma$, can lead to *numerical cancellation*, that is, $\pm\sqrt{\beta^2-4\alpha\gamma}$ might be rounded to 0, incorrectly giving the impression one has two equal roots.

To elaborate on the problem, consider the subtraction of two scalars $a-b$, for $\widehat{a}=a(1+\Delta a)$ and $\widehat{b}=b(1+\Delta b)$ the numerical evaluations of a and b, respectively. Suppose that $a\neq b$, then we have

$$\widehat{a}-\widehat{b}=(a-b)\left(1+\frac{a\Delta a-b\Delta b}{a-b}\right),$$

such that for non zero $(\Delta a,\Delta b)$ and $a\approx b$, the evaluation $\widehat{a}-\widehat{b}$ can be far away from $a-b$. This should be understood in the rounding context, the exact value of a might require precision not available on the system at hand, hence, the numerical representation becomes $\widehat{a}=a(1+\Delta a)$ where Δa is such that $\widehat{a}$ is one of the nearest floating points to a.

The notion of cancellation becomes more intricate when functions are involved, *e.g.*, $f(a)-f(b)$. In Chapter 5 we consider precisely this phenomenon, that is, typical zeroth-order optimization algorithms work with expressions of the form

$$\frac{f(x+\delta)-f(x)}{\delta},\tag{2.4.7}$$

for f differentiable and δ asymptotically vanishing ($\delta\to 0^+$). On paper, we can let δ vanish and recover the derivative of f, however, numerically, for δ sufficiently small, $f(x+\delta)\approx f(x)$ and the numerical evaluation of (2.4.7) becomes useless[21]. We point out that in such a case we do not simply have $0/0$, due to the representation of floating point numbers[22].

To see how properties of f can make the situation better or worse, suppose that f is $L_0(f)$-globally Lipschitz. Then we have that $|f(x+\delta)-f(x)|\leq L_0(f)\delta$. Therefore, if $\delta\leq \mu_{\mathsf{M}}/L_0(f)$, for μ_{M} denoting the machine precision, we are in numerical trouble. This means that flat functions are numerically challenging, which is precisely how the most important areas of an optimization landscape look like.

Bibliography

[AEK20] A A Ahmadi and B El Khadir. Learning dynamical systems with side information. In *Proc. Learning for Dynamics and Control*, pages 718–727, 2020.

[21] Although, the output is not random [Hig02, Sec. 1.17].

[22] On a standard 64-bit system we have a precision $\approx 10^{-16}$ and a smallest number $\approx 10^{-308}$.

[AJPR14] A A Ahmadi, R M Jungers, P A Parrilo, and M Roozbehani. Joint spectral radius and path-complete graph Lyapunov functions. *SIAM J. Contr. Optim.*, 52(1):687–717, 2014.

[AM08] R Abraham and J E Marsden. *Foundations of mechanics.* American Mathematical Society, Providence, 2008.

[Arn88] V I Arnold. *Geometrical methods in the theory of ordinary differential equations.* Springer, New York, 1988.

[Art83] Z Artstein. Stabilization with relaxed controls. *Nonlinear Anal.-Theor.*, 7(11):1163–1173, 1983.

[BCCG20] F Blanchini, G Chesi, P Colaneri, and G Giordano. Checking structural stability of bdc-decomposable systems via convex optimisation. *IEEE Control Syst. Lett.*, 4(1):205–210, 2020.

[BE02] O Bousquet and A Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

[Bel84] R Bellman. *Eye of the Hurricane.* World Scientific, Singapore, 1984.

[Ber63] C Berge. *Topological spaces.* Oliver & Boyd, Edinburgh, 1963.

[BK74] N A Bobylev and M A Krasnosel'skiĭ. Deformation of a system into an asymptotically stable system. *Automat. Remote Control*, 35(7):1041–1044, 1974.

[Blo15] A Bloch. *Nonholonomic Mechanics and Control.* Springer, New York, 2015.

[BP16] F Bach and V Perchet. Highly-smooth zero-th order online optimization. In *Proc. Conference on Learning Theory*, pages 1–27, 2016.

[BR05] A Bacciotti and L Rosier. *Liapunov functions and stability in control theory.* Springer Science & Business Media, Berlin, 2005.

[Bra98] M S Branicky. Multiple Lyapunov functions and other analysis tools for switched and hybrid systems. *IEEE T. Automat. Contr.*, 43(4):475–482, 1998.

[Bro83] R W Brockett. Asymptotic stability and feedback stabilization. In *Differential Geometric Control Theory*, pages 181–191, Boston, 1983. Birkhäuser.

[BS70] N P Bhatia and G P Szegö. *Stability theory of dynamical systems.* Springer, Berlin, 1970.

[BTM$^+$21] N Boffi, S Tu, N Matni, J.-J Slotine, and V Sindhwani. Learning stability certificates from data. In *Proc. Conference on Robot Learning*, pages 1341–1350, 2021.

[Byr08] C I Byrnes. On Brockett's necessary condition for stabilizability and the topology of Liapunov functions on $\mathbb{R}^n$. *Commun. Inf. Syst.*, 8(4):333–352, 2008.

[Cla10] F Clarke. Discontinuous feedback and nonlinear systems. *IFAC Proc. Vol.*, 43(14):1–29, 2010.

[CLSS97] F H Clarke, Y S Ledyaev, E D Sontag, and A I Subbotin. Asymptotic controllability implies feedback stabilization. *IEEE T. Automat. Contr.*, 42(10):1394–1407, 1997.

[Cor90] J.-M Coron. A necessary condition for feedback stabilization. *Syst. Control Lett.*, 14(3):227–232, 1990.

[Cor07] J.-M Coron. *Control and Nonlinearity.* American Mathematical Society, Providence, 2007.

[Fau21] T Faulwasser. Towards necessary and sufficient stability conditions for NMPC. *IFAC Conf. on Nonlinear Model Predictive Control*, 54(6):139–146, 2021.

[Fil88] A F Filippov. *Differential Equations with Discontinuous Righthand Sides.* Springer Science & Business Media, Dordrecht, 1988.

[FP19] A Fathi and P Pageault. Smoothing Lyapunov functions. *T. Am. Math. Soc.*, 371(3):1677–1700, 2019.

[GFXFT23] C L Galimberti, L Furieri, L Xu, and G Ferrari-Trecate. Hamiltonian deep neural networks guaranteeing nonvanishing gradients by design. *IEEE T. Automat. Control*, 68(5):3155–3162, 2023.

[GGLW14] R Geiselhart, R H Gielen, M Lazar, and F R Wirth. An alternative converse lyapunov theorem for discrete-time systems. *Syst. Control Lett.*, 70:49–59, 2014.

[GH15] P Giesl and S Hafstein. Review on computational methods for Lyapunov functions. *Discrete Contin. Dyn. Syst. Ser. B*, 20(8):2291, 2015.

[GKS19] N Gillis, M Karow, and P Sharma. Approximating the nearest stable discrete-time system. *Linear Algebra Appl.*, 573:37–53, 2019.

[GP10] V Guillemin and A Pollack. *Differential topology.* American Mathematical Society, Providence, 2010.

[GST12] R Goebel, R G Sanfelice, and A R Teel. Hybrid dynamical systems. In *Hybrid dynamical systems.* Princeton University Press, Princeton, 2012.

[GZY21] N Gaby, F Zhang, and X Ye. Lyapunov-net: A deep neural network architecture for Lyapunov function approximation. *arXiv e-print:2109.13359*, 2021.

[Hah58] W Hahn. Über die anwendung der methode von ljapunov auf differenzengleichungen. *Math. Ann.*, 136(5):430–441, 1958.

[Hah67] W Hahn. *Stability of motion.* Springer, Berlin, 1967.

[Hal09] J K Hale. *Ordinary Differential Equations.* Dover Publications, New York, 2009.

[Hig02] N J Higham. *Accuracy and stability of numerical algorithms.* SIAM, Philadelphia, 2002.

[Hir76] M W Hirsch. *Differential topology.* Springer, New York, 1976.

[HRS16] M Hardt, B Recht, and Y Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proc. International Conference on Machine Learning*, pages 1225–1234, 2016.

[Isi85] A Isidori. *Nonlinear control systems: an introduction.* Springer, Berlin, 1985.

[JL14] S Jafarpour and A D Lewis. *Time-varying vector fields and their flows.* Springer, Cham, 2014.

[JM23] W Jongeneel and E Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, Cham, 2023.

[Joh03] M Johansson. *Piecewise linear control systems: a computational approach.* Springer-Verlag, Berlin, 2003.

[JW02] Z.-P Jiang and Y Wang. A converse Lyapunov theorem for discrete-time systems with disturbances. *Syst. Control Lett.*, 45(1):49–58, 2002.

[JWYM20] W Jin, Z Wang, Z Yang, and S Mou. Neural certificates for safe control policies. *arXiv e-print:2006.08465*, 2020.

[KB60] R E Kalman and J E Bertram. Control System Analysis and Design Via the “Second Method” of Lyapunov: II—Discrete-Time Systems. *J. Basic Eng.-T. ASME*, 82(2):394–400, 1960.

[KH95] A Katok and B Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems.* Cambridge University Press, New York, 1995.

[Kha02] H K Khalil. *Nonlinear systems.* Prentice Hall, Upper Saddle River, 2002.

[Kha16] M Khammash. An engineering viewpoint on biological robustness. *BMC biology*, 14(1):22, 2016.

[KK22] M D Kvalheim and D E Koditschek. Necessary conditions for feedback stabilization and safety. *J. Geom. Mech.*, 14(4):659–693, 2022.

[KM19] J Z Kolter and G Manek. Learning stable deep dynamics models. In *Proc. Neural Inf. Process. Syst.*, pages 11126–11134, 2019.

[KR73] N H Kuiper and J W Robbin. Topological classification of linear endomorphisms. *Inventiones math. Springer Verlag*, 19:83–106, 1973.

[Kun00] M Kunze. *Non-Smooth Dynamical Systems.* Springer-Verlag, Berlin Heidelberg, 2000.

[Kur63] J Kurzweil. On the inversion of Ljapunov's second theorem on stability of motion. *AMS Transl. Ser. 2*, 24:19–77, 1963.

[Kuz04] Y A Kuznetsov. *Elements of Applied Bifurcation Theory.* Springer, New York, 2004.

[Kva23] M D Kvalheim. Obstructions to asymptotic stabilization. *SIAM J. Contr. Optim.*, 61(2):536–542, 2023.

[KYK22] K Kashima, R Yoshiuchi, and Y Kawano. Learning stabilizable deep dynamics models. *arXiv e-prints*, 2022.

[KZ84] A Krasnosel'skiĭ and P P Zabreiko. *Geometrical methods of nonlinear analysis.* Springer, Berlin, 1984.

[LA09] H Lin and P J Antsaklis. Stability and stabilizability of switched linear systems: a survey of recent results. *IEEE T. Automat. Contr.*, 54(2):308–322, 2009.

[Lee11] J M Lee. *Introduction to Topological Manifolds.* Springer, New York, 2011.

[Lee12] J M Lee. *Introduction to Smooth Manifolds.* Springer, New York, 2012.

[Lia92] A M Liapunov. *A general task about the stability of motion.* dissertation, University of Kharkov, 1892.

[Lib03] D Liberzon. *Switching in systems and control.* Birkhäuser, Boston, 2003.

[LK97] Z.-H Li and M Krstić. Maximizing regions of attraction via backstepping and CLFs with singularities. *Syst. Control Lett.*, 30(4):195–207, 1997.

[LR95] P Lancaster and L Rodman. *Algebraic Riccati Equations.* Oxford University Press, Oxford, 1995.

[LS98] W Lohmiller and J.-J E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[LS99] Y S Ledyaev and E D Sontag. A Lyapunov characterization of robust stabilization. *Nonlinear Anal. Theory Methods Appl.*, 37(7):813–840, 1999.

[Man07] A.-R Mansouri. Local asymptotic feedback stabilization to a submanifold: Topological conditions. *Syst. Control Lett.*, 56(7-8):525–528, 2007.

[Man10] A.-R Mansouri. Topological obstructions to submanifold stabilization. *IEEE T. Automat. Contr.*, 55(7):1701–1703, 2010.

[Mas49] J L Massera. On Liapounoff's conditions of stability. *Ann. Math.*, 50(3):705–721, 1949.

[Mas56] J L Massera. Contributions to stability theory. *Ann. Math.*, 64(1):182–206, 1956.

[MBP07] E Moulay, R Bourdais, and W Perruquetti. Stabilization of nonlinear switched systems using control Lyapunov functions. *Nonlinear Anal.: Hybrid Syst.*, 1(4):482–490, 2007.

[MDT$^+$22] S Mukherjee, J Drgoňa, A Tuor, M Halappanavar, and D Vrabie. Neural Lyapunov differentiable predictive control. *arXiv e-print:2205.10728*, 2022.

[Mil56] J Milnor. On manifolds homeomorphic to the 7-sphere. *Ann. Math.*, 64(2):399–405, 1956.

[Mil65] J Milnor. *Topology from the differentiable viewpoint.* Princeton University Press, Princeton, 1965.

[MP05] E Moulay and W Perruquetti. Finite time stability of differential inclusions. *IMA J. Math. Control. Inf.*, 22(4):465–475, 2005.

[MRS$^+$22] D Malyuta, T P Reynolds, M Szmuk, T Lew, R Bonalli, M Pavone, and B Açıkmeşe. Convex optimization for trajectory generation: A tutorial on generating dynamically feasible trajectories reliably and efficiently. *IEEE Contr. Syst. Mag.*, 42(5):40–113, 2022.

[Nes03] Y Nesterov. *Introductory lectures on convex optimization: A basic course.* Springer Science & Business Media, New York, 2003.

[NS17] Y Nesterov and V Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, 2017.

[NvdS90] H Nijmeijer and A J van der Schaft. *Nonlinear Dynamical Control Systems.* Springer, New York, 1990.

[OK22] Y Okamoto and R Kojima. Learning deep input-output stable dynamics. *arXiv e-print:2206.13093*, 2022.

[PBND16] R Postoyan, L Buşoniu, D Nešić, and J Daafouz. Stability analysis of discrete-time infinite-horizon optimal control with discounted cost. *IEEE T. Automat. Contr.*, 62(6):2736–2749, 2016.

[PDM82] J Jr Palis and W De Melo. *Geometric theory of dynamical systems: an introduction.* Springer, New York, 1982.

[Pol87] J W Polderman. *Adaptive Control & Identification: Conflict or Conflux.* dissertation, University of Groningen, 1987.

[Rif02] L Rifford. Semiconcave control-Lyapunov functions and stabilizing feedbacks. *SIAM J. Contr. Optim.*, 41(3):659–681, 2002.

[RKKM22] N Rezazadeh, M Kolarich, S S Kia, and N Mehr. Learning contraction policies from offline data. *IEEE Robot. Autom. Lett.*, 7(2):2905–2912, 2022.

[Rob72] J W Robbin. Topological conjugacy and structural stability for discrete dynamical systems. *Bull. Amer. Math. Soc.*, 78(6):923–952, 11 1972.

[Rob95] C Robinson. *Dynamical systems: stability, symbolic dynamics, and chaos.* CRC Press, Boca Raton, 1995.

[Sas99] S Sastry. *Nonlinear Systems.* Springer, New York, 1999.

[SLG95] Y N Sotskov, V K Leontev, and E N Gordeev. Some concepts of stability analysis in combinatorial optimization. *Discrete Applied Mathematics*, 58(2):169–190, 1995.

[Son89] E D Sontag. A 'universal'construction of Artstein's theorem on nonlinear stabilization. *Syst. Control Lett.*, 13(2):117–123, 1989.

[Son98] E D Sontag. *Mathematical control theory: deterministic finite dimensional systems.* Springer, New York, 1998.

[Son99] E D Sontag. Stability and stabilization: discontinuities and the effect of disturbances. In *Nonlinear Analysis, Differential Equations and Control*, pages 551–598. Springer, Dordrecht, 1999.

[Son01] E D Sontag. The ISS philosophy as a unifying framework for stability-like behavior. In *Nonlinear control in the year 2000 volume 2*, pages 443–467, London, 2001. Springer London.

[Ste03] G Stein. Respect the unstable. *IEEE Contr. Syst. Mag.*, 23(4):12–25, 2003.

[Sti12] J Stillwell. Poincaré and the early history of 3-manifolds. *B. Am. Math. Soc.*, 49(4):555–576, 2012.

[TP05] P Tabuada and G J Pappas. Quotients of fully nonlinear control systems. *SIAM J. Contr. Optim.*, 43(5):1844–1866, 2005.

[VDKS93] J L M Van Dorsselaer, J F B M Kraaijevanger, and M N Spijker. Linear stability analysis in the numerical solution of initial value problems. *Acta Numer.*, 2:199–237, 1993.

[vS21] J H van Schuppen. *Control and system theory of discrete-time stochastic systems.* Springer, Cham, 2021.

[VTHK21] C Verhoek, R Tóth, S Haesaert, and A Koch. Fundamental lemma for data-driven analysis of linear parameter-varying systems. In *Proc. IEEE Conference on Decision and Control*, pages 5040–5046, 2021.

[Wil80] J C Willems. Topological classification and structural stability of linear systems. *J. Differ. Equ.*, 35(3):306 – 318, 1980.

[WJ67] F W Wilson Jr. The structure of the level surfaces of a Lyapunov function. *J. Differ. Equ.*, 3(3):323–329, 1967.

[Won79] W M Wonham. *Linear Multivariable Control.* Springer, New York, 1979.

[WVdS82] J C Willems and A J Van der Schaft. Modelling of dynamical systems using external and internal variables with applications to Hamilton systems'. In *Dynamical Systems and Microphysics*, pages 233–264. 1982.

[YI00] Y Yamashita and A Isidori. Global output regulation through singularities. In *Proc. IEEE Conference on Decision and Control*, volume 2, pages 1295–1300, 2000.

[YXRR22] A Yang, J Xiong, M Raginsky, and E Rosenbaum. Input-to-state stable neural ordinary differential equations with applications to transient modeling of circuits. In *Proc. Learning for Dynamics and Control Conference*, pages 663–675, 2022.

[ZB22] V Zinage and E Bakolas. Neural koopman Lyapunov control. *arXiv e-print:2201.05098*, 2022.

[ZZL22] J Zhang, Q Zhu, and W Lin. Neural stochastic control. *arXiv e-print:2209.07240*, 2022.

[Če95] S Čelikovský. Topological equivalence and topological linearization of controlled dynamical systems. *Kybernetika*, 31(2):141–150, 1995.

3
Identifying stability

> “*The concept of a least-squares estimator was first discovered in 1795 by the eighteen-year-old Gauss. He needed a practical tool for analyzing survey data … Gauss thought it such a simple idea that he did not publish it … On January 1, 1801 Giuseppe Piazzi discovered the asteroid Ceres. Ceres was only visible for forty days before it was lost to view behind the sun … Gauss, using three observations, extensive analysis, and the method of least squares, was able to determine the orbit with such accuracy that Ceres was easily found when it reappeared in late 1801.*”
>
> —Kahaner, Moler and Nash [KMN89, Sec. 6.6].

Understanding stability of identified systems is of great practical- and theoretical importance. Even the simplest case, that of characterizing spectral properties of the least-squares estimator of a linear dynamical system has been largely open. To that end, we propose a principled method for projecting a system matrix to the nonconvex set of Schur stable matrices. Leveraging large deviations theory, we show that this projection is optimal in an information-theoretic sense and that the projection can be approximated, up to arbitrary precision, by simply adding a feedback term corresponding to the optimal gain matrix of a linear quadratic regulator problem. The estimator resulting through this projection is constructed from a single trajectory of state measurements, is guaranteed to be stable and offers non-asymptotic statistical bounds on the estimation error.

Going one step beyond stability, we further exploit large deviations theory to identify the topological class of an unknown stable system, again from a single trajectory of data. We prove that the probability of misclassification decays exponentially with the number of samples at a rate that is proportional to the square of the smallest singular value of the unknown matrix.

3.1 Identifying a linear dynamical system with stability guarantees

In this section we study how to enforce stability while preserving statistical guarantees. To our best knowledge, we present the first method for the identification of a linear dynamical system with stability guarantees that is both computationally efficient and offers asymptotic consistency and tight statistical error bounds.

3.1.1 Introduction

We study the problem of identifying an asymptotically stable linear dynamical system (*i.e.*, the system matrix is Schur stable) from a single trajectory of correlated state observations. This problem is of fundamental importance in various disciplines such as adaptive control [ÅW73], system identification [KV86, VV07], reinforcement learning [SB18, Ber19, MPRT19, Rec19] and approximate dynamic programming [BT96, Pow07]. Specifically, we consider a discrete-time linear time-invariant system of the form

$$x_{t+1} = \theta x_t + w_t, \quad x_0 \sim \nu, \tag{3.1.1}$$

where $x_t \in \mathbb{R}^n$ and $w_t \in \mathbb{R}^n$ denote the state and the exogenous noise at time $t \in \mathbb{Z}_{\geq 0}$, respectively, while θ represents a fixed system matrix[1], and ν stands for the marginal distribution of the initial state x_0. We assume that θ is *asymptotically stable*, that is, it belongs to $\Theta := \{\theta \in \mathbb{R}^{n\times n} : \rho(\theta) < 1\}$, where $\rho(\theta)$ denotes the spectral radius of θ, see Section 2.2.5. For ease of terminology, we will usually refer in this section to Θ as the set of *stable* matrices and to its complement in $\mathbb{R}^{n\times n}$ as the set of *unstable* matrices. We assume that nothing is known about θ except for its membership in Θ, and we aim to identify θ from a single-trajectory of data $\{\widehat{x}_t\}_{t=0}^T$ generated by (3.1.1). To this end, one can use the *least squares estimator*

$$\widehat{\theta}_T = \left(\textstyle\sum_{t=1}^T \widehat{x}_t \widehat{x}_{t-1}^\intercal\right) \left(\textstyle\sum_{t=1}^T \widehat{x}_{t-1} \widehat{x}_{t-1}^\intercal\right)^{-1}, \tag{3.1.2}$$

which may take any value in $\Theta' := \mathbb{R}^{n\times n}$ under standard assumptions on the noise distribution. It is therefore possible that $\widehat{\theta}_T \notin \Theta$ even though $\theta \in \Theta$. This is troubling because stability is important in many applications, for example, when the estimated model is used for prediction, filtering or control, *e.g.*, see the discussions in [VODM96, pp. 53–60, 125–129].

Given the prior structural information that θ is stable, we thus seek an estimator that is guaranteed to preserve stability. A natural approach to achieve this goal would be to "*project*" the least squares estimator $\widehat{\theta}_T$ to the nearest stable matrix with respect to

[1] Although the standard notation would be $x_{t+1} = Ax_t + w_t$, we use θ since this is the common symbol to denote a parametrization in several statistical communities.

some "*discrepancy function*" on $\mathbb{R}^{n\times n}$. This seems challenging, however, because Θ is open, unbounded and non-convex; see [JSK23, Fig.1 (a)]. To circumvent this difficulty, we introduce a new discrepancy function that adapts to the geometry of Θ and is thus ideally suited for projecting unstable matrices onto Θ. We will characterize the statistical properties of this projection when applied to the least squares estimator, and we will show that it can be computed efficiently even for systems with $O(10^3)$ states.

The following example shows that naïve heuristics to project θ' into the interior of Θ could spectacularly fail.

Example 3.1.1 (Projection by eigenvalue scaling)**.** A naïve method to stabilize a matrix $\theta' \notin \Theta$ would be to scale its unstable eigenvalues into the complex unit circle. To see that the output of this transformation may not retain much similarity with the input θ', consider the matrices

$$\theta' = \begin{pmatrix} 1.01 & 10 \\ .01 & 1 \end{pmatrix}, \; \theta'_a = \begin{pmatrix} .84 & 4.77 \\ .005 & .84 \end{pmatrix}, \; \theta'_b = \begin{pmatrix} .99 & 10 \\ 0 & .99 \end{pmatrix}.$$

Clipping off the unstable eigenvalues of θ' at $|\lambda| = .99$ yields θ'_a with $\rho(\theta'_a) = .99$ and $\|\theta' - \theta'_a\|_2 \gtrsim 5$. However, the matrix θ'_b also has spectral radius $\rho(\theta'_b) = .99$ but is much closer to θ'. Indeed, we have $\|\theta' - \theta'_b\|_2 \approx 0.02$. Even more so, if this clipping procedure would work, what is a desirable clipping-point?

Learning stable systems

The problem of learning a stable dynamical system is widely studied in system identification, while the problem of projecting an unstable matrix onto Θ with respect to some norm has attracted considerable interest in matrix analysis.

In the context of identification, Maciejowski proposed one of the first methods to project a possibly unstable estimator onto Θ by using subspace methods [Mac95]. This pioneering approach has significant practical merits [VODM96] but may also significantly distort the original estimator. To overcome this deficiency, Lacy and Bernstein approximate Θ by the set of *contractive* matrices whose operator norm is at most 1 [LB02]. While this set is convex, it offers but a conservative approximation of Θ. Several related methods have since been proposed to enforce stability [LB03, BGS08, TBQ$^+$13], which are all either conservative or computationally expensive. Moreover, these methods do not provide any statistical guarantees. Van Gestel *et al.* regularize the least squares objective and show that the spectral radius of the resulting estimator is bounded by a function of the regularization weights [vSvd00, vSvd01]. As Θ is an open set, however, the tuning of these weights remains a matter of taste. More recently, Umenberger *et al.* propose a maximum likelihood approach that is attractive from a statistical point of view but can be computationally challenging in certain applications [UWMS18]. On the other hand, several authors use Lyapunov theory to provide stability guarantees for deterministic vector fields; see, *e.g.*, [MB14, BTSK17, KM19, UH20, BTM$^+$21]. There is also a substantial body of literature on (sub-)optimal finite-sample concentration bounds for linear systems identified via least squares estimation [SMT$^+$18, JP19, SR19, JP20, SRD21]. These approaches offer learning rates but cannot guarantee stability of the identified systems for finite sample sizes, the point being that bounds of the form $\mathbb{P}_\theta(\|\widehat{\theta}_T - \theta\|_2 \leq \varepsilon) \geq 1 - \beta_T$ tell us very little about *stability* (spectral properties) of $\widehat{\theta}_T$.

Much like in dynamical systems theory, in matrix analysis one seeks algorithms for projecting an unstable matrix θ' onto Θ, which is equivalent to finding the smallest additive perturbation that stabilizes θ'. More specifically, matrix analysis studies the *nearest stable matrix problem*

$$\Pi_\Theta(\theta') \in \arg\min_{\theta \in \operatorname{cl}\Theta} \|\theta' - \theta\|^2, \tag{3.1.3}$$

where $\|\cdot\|$ represents a prescribed norm on $\mathbb{R}^{n\times n}$. Note that optimizing over the closure of Θ is necessary for (3.1.3) to be well-defined because, for $\theta' \notin \Theta$, any minimizer lies on the boundary of the open set Θ. Unfortunately, solving (3.1.3) is challenging because Θ is non-convex. Existing numerical solution procedures rely on successive convex approximations [ONv13], on local optimization schemes based on the solution of low-rank matrix differential equations [GL17] or on an elegant reparametrization of the set of stable matrices, which simplifies the numerics of the projection operation [GKS19, CGS20]. The latter approach was recently used for learning stable systems [MXM20, MAM23]. Nesterov and Protasov solve (3.1.3) for certain polyhedral norms and non-negative matrices θ' [NP20], which allows them to find exact solutions. See [Hig89] for a general discussion on matrix nearness problems.

Optimal control offers a promising alternative perspective on problem (3.1.3), which is closely related to the approach advocated in this chapter: one could try to design a linear quadratic regulator (LQR) problem, see also Section 4.1.2, whose optimal feedback gain $K^\star \in \mathbb{R}^{n\times n}$ renders $\theta' + K^\star$ stable. By proposing an LQR objective that is inversely proportional to the sample covariance matrix of the measurement noise, Tanaka and Katayama show that this idea is indeed valid, but they provide no error analysis or statistical guarantees [TK05]. As we show in Section 4.1 below, these optimal control techniques also naturally preserve structure of the underlying system matrix. Such a structure-preserving approach seems preferable over the plain nearest stable matrix problem (3.1.3), which merely seeks stability at minimal cost. Appealing to the theory of *large deviations*, we will give such approaches a statistical underpinning.

We already bring some notation forward to make the upcoming sections easier to read.

Notation: For a matrix $A \in \mathbb{C}^{n\times n}$, we denote by $\rho(A)$ the largest absolute eigenvalue and by $\kappa(A)$ the (ℓ_2) condition number of A. For a set $\mathcal{D} \subset \mathbb{R}^n$, we denote by $\mathcal{D}^c$ the complement, by $\operatorname{cl}\mathcal{D}$ the closure and by $\operatorname{int}\mathcal{D}$ the interior of $\mathcal{D}$. For a real sequence $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ we use $1 \ll a_T \ll T$ to express that $a_T/T \to 0$ and $a_T \to \infty$ as $T \to \infty$. We also use the soft-O notation $\widetilde{O}(f(T))$ as a shorthand for $O(f(T)\log(T)^c)$ for some $c \in \mathbb{Z}_{\geq 0}$, that is, $\widetilde{O}(\cdot)$ ignores polylogarithmic factors.

Contributions

Throughout this section we assume that all random objects are defined on a measurable space $(\Omega, \mathcal{F})$ equipped with a probability measure $\mathbb{P}_\theta$ that depends parametrically on the *fixed* yet *unknown* system matrix θ, and the system equations (3.1.1) are assumed to hold $\mathbb{P}_\theta$-almost surely; see also the discussion below Assumption 3.1.1. The expectation operator with respect to $\mathbb{P}_\theta$ is denoted by $\mathbb{E}_\theta[\cdot]$. Even though the least squares estimator $\widehat{\theta}_T$ is strongly consistent and thus converges $\mathbb{P}_\theta$-almost surely to θ [CK98], it differs $\mathbb{P}_\theta$-almost surely from θ for any finite T. To quantify estimation errors, we introduce a discrepancy function $I : \Theta' \times \Theta \to [0, \infty]$ defined through

$$I(\theta', \theta) = \tfrac{1}{2} \mathrm{Tr} \left(S_w^{-1} (\theta' - \theta) S_\theta (\theta' - \theta)^\top \right). \tag{3.1.4}$$

Here, $S_w \succ 0$ stands for the time-independent noise covariance matrix, and S_θ denotes the covariance matrix of x_t under the stationary state distribution, which exists for $\theta \in \Theta$ but diverges as θ approaches the boundary of Θ; see [JSK23, Fig.1 (b)]. Note that since $S_w \succ 0$ and hence $S_\theta \succ 0$, $I(\theta', \theta)$ vanishes if and only if $\theta' = \theta$. In this sense I behaves like a distance. Note, however, that $I(\theta', \theta)$ is not symmetric in θ and θ'.

Now, we propose to use the discrepancy function (3.1.4) for projecting an unstable matrix θ' onto Θ. Specifically, we define the *reverse I-projection* of any $\theta' \in \mathbb{R}^{n \times n}$ as

$$\mathcal{P}(\theta') \in \arg \inf_{\theta \in \Theta} I(\theta', \theta). \tag{3.1.5}$$

We will see that the discrepancy function (3.1.4) has a natural statistical interpretation, which enables us to derive strong statistical guarantees for the reverse I-projection of the least squares estimator. We will actually show that the discrepancy function (3.1.4) determines the speed at which the probability of the least squares estimator $\widehat{\theta}_T$ being sufficiently different from the true system matrix θ decays with the sample size T.

Specifically, we will prove that the *transformed* estimator $\widehat{\vartheta}_T = \sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta$ satisfies a moderate deviations principle with rate function (3.1.4). By exploiting the relation $I(\widehat{\theta}_T, \theta) = (a_T/T) I(\widehat{\vartheta}_T, \theta)$, one can then show that the probability density function $\varrho_{\theta,T}$ of the original least squares estimator $\widehat{\theta}_T$ with respect to the probability measure $\mathbb{P}_\theta$ decays exponentially with T, that is,

$$\varrho_{\theta,T}(\widehat{\theta}_T) \approx \exp(-I(\widehat{\theta}_T, \theta) \cdot T). \tag{3.1.6}$$

Thus, the reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$ maximizes the right-hand-side of (3.1.6) across all $\theta \in \Theta$. Therefore, one can interpret $\mathcal{P}(\widehat{\theta}_T)$ as a *maximum likelihood estimator*, that is, the most likely *asymptotically stable model* in view of the data. In addition, by using ideas due to Jedra and Proutiere [JP20], one can readily show that if the exogenous noise is Gaussian, then the discrepancy function $I(\theta', \theta)$ defined in (3.1.4) can be interpreted as the long-run average expected log-likelihood ratio between observations generated under $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$.

Our main contributions can be summarized as follows.

(i) We prove that the discrepancy function (3.1.4) has a natural statistical interpretation as the rate function of a moderate deviation principle for the transformed least squares estimators $\sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta$, $T \in \mathbb{Z}_{\geq 0}$.

(ii) We derive finite-sample and asymptotic statistical error bounds on the operator norm distance between the reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$ of the least squares estimator $\widehat{\theta}_T$ and the unknown true system matrix θ.

(iii) We show that the reverse I-projection $\mathcal{P}(\theta')$ can be computed to within any desired accuracy by solving a standard LQR problem, *e.g.*, via readily available numerical routines.

We also note that the derivation of the explicit rate function (3.1.4) is of independent interest in the context of statistical learning of linear dynamical systems.

3.1.2 Efficient identification with stability guarantees

From now on we impose the following assumption.

Assumption 3.1.1 (Linear system)**.** *The following hold.*

(i) The linear system (3.1.1) *is stable, i.e.,* $\theta \in \Theta$.

(ii) For each $\theta \in \Theta$ *the disturbances* $\{w_t\}_{t\in\mathbb{Z}_{\geq 0}}$ *are independent and identically distributed (i.i.d.) and independent of* x_0 *under* $\mathbb{P}_\theta$*. The marginal noise distributions are unbiased (*$\mathbb{E}_\theta[w_t] = 0$*), non-degenerate (*$S_w = \mathbb{E}_\theta[w_t w_t^\intercal] \succ 0$ *is finite) and have an everywhere positive probability density function.*

Assumption 3.1.1 ensures that the linear system (3.1.1) admits an invariant distribution ν_θ [MT09, Sec. 10.5.4]. This means that $x_t \sim \nu_\theta$ implies $x_{t+1} \sim \nu_\theta$ for any $t \in \mathbb{Z}_{\geq 0}$. Moreover, as the probability density function of w_t is everywhere positive, $\{x_t\}_{t\in\mathbb{Z}_{\geq 0}}$ represents a uniformly ergodic Markov process, which implies that the marginal distribution of x_t under $\mathbb{P}_\theta$ converges weakly to ν_θ as t tends to infinity [MT09, Thm. 16.2.1, 16.5.1]. Assumption 3.1.1 then implies that the mean vector of ν_θ vanishes and that the covariance matrix S_θ of ν_θ coincides with the unique solution of the discrete Lyapunov equation

$$S_\theta = \theta S_\theta \theta^\intercal + S_w, \tag{3.1.7}$$

which provides for a convenient way to compute S_θ; see, *e.g.*, [AM06, Sec. 6.10 E]. Recall that S_θ critically enters the discrepancy function $I(\theta', \theta)$ defined in (3.1.4) and thus also the reverse I-projection defined in (3.1.5). Given the existence of an invariant distribution, we impose another standard regularity condition.

Assumption 3.1.2 (Light-tailed noise and stationarity)**.** *The following hold for every* $\theta \in \Theta$.

(i) The disturbances $\{w_t\}_{t\in\mathbb{Z}_{\geq 0}}$ *are light-tailed, i.e., there exists* $\alpha > 0$ *with* $\mathbb{E}_\theta[e^{\alpha\|w_t\|^2}] < \infty$ *for all* $t \in \mathbb{Z}_{\geq 0}$.

(ii) The initial distribution ν *coincides with the invariant distribution* ν_θ *of the linear system* (3.1.1).

Assumption 3.1.2 (i) essentially requires the noise to have no heavier tails than a normal distribution and is equivalent to the requirement for the noise to be sub-Gaussian [Ver18, Prop. 2.5.2 (iv)]. Assumption 3.1.2 (ii) stipulates that the linear system is in the stationary regime already at time $t = 0$.

To motivate the remainder of this chapter, we bring the key result already forward at this point. That is, the following theorem summarizes the key statistical and computational properties of the reverse I-projection that will be proved in the remainder. This theorem involves the function $\mathsf{dlqr}(A, B, Q, R)$, which outputs the optimal feedback gain matrix of an infinite-horizon deterministic LQR problem in discrete time, *e.g.*, see [Ber05, Sec. 4].

Theorem 3.1.3 (Efficient identification with stability guarantees)**.** *Suppose that Assumptions 3.1.1 and 3.1.2 hold, and that $\widehat{\theta}_T$ is the least squares estimator* (3.1.2)*. Then, for any $\theta \in \Theta$ the reverse I-projection defined in* (3.1.5) *has the following properties.*

(i) ***Asymptotic consistency.***

$$\lim_{T\to\infty} \mathcal{P}(\widehat{\theta}_T) = \theta \quad \mathbb{P}_\theta\text{-}a.s.$$

(ii) ***Finite sample guarantee.*** *There are constants $\tau \geq 0$ and $\rho \in (0, 1)$ that depend only on θ such that*

$$\mathbb{P}_\theta\left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \kappa(S_w) \frac{2\varepsilon n^{1/2}\tau}{\sqrt{1-\rho^2}} \right) \geq 1 - \beta$$

for all $\beta, \varepsilon \in (0, 1)$ and $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$.

(iii) ***Efficient computation.*** *For any $\theta' \in \Theta' \setminus \partial\Theta$ and $S_w, Q \succ 0$ there is a $p \geq 1$, such that for all $\delta > 0$ we have that*

$$\theta^\star_\delta = \theta' + \mathsf{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1})$$

is stable and satisfies $\|\mathcal{P}(\theta') - \theta^\star_\delta\|_2 \leq O(\delta^p)$.

The asymptotic consistency (i) formalizes the intuitive requirement that more data is preferable to less data. We emphasize that the reverse I-projection does not introduce unnecessary bias because $\mathcal{P}(\theta') = \theta'$ if θ' is already stable. The finite sample guarantee (ii) stipulates that the projected least squares estimator $\mathcal{P}(\widehat{\theta}_T)$ is guaranteed to be close to the (unknown) true stable matrix θ with high probability $1 - \beta$. Note that if the observed state trajectory $\{\widehat{x}_t\}_{t=0}^T$ is generated under $\mathbb{P}_\theta$, then the inverse matrix appearing in (3.1.2) exists $\mathbb{P}_\theta$-almost surely for any sample size $T \geq n$ thanks to Assumption 3.1.1 (ii). The efficient computability property (iii), finally, shows that computing the reverse I-projection to within high accuracy is no harder than solving a standard LQR problem.

We also emphasize that setting $Q = I_n$ works well in practice, that is, no tuning is required to compute $\mathcal{P}(\theta')$. However, tuning Q can nevertheless improve the conditioning of the optimization problem and speed up the computation of $\mathcal{P}(\theta')$. Guidelines on choosing Q and the results of extensive numerical experiments are reported in [Jon22].

Recall that the reverse I-projection exhibits optimism in the face of uncertainty, a decision-making paradigm that is used with great success in various reinforcement learning applications [LS20]. In general, however, optimism in the face of uncertainty leads to computational intractability [CK98]. Thus, the tractability result of Theorem 3.1.3 (iii) is a perhaps unexpected exception to this rule; see Proposition 3.1.16 below for further details.

3.1.3 The reverse I-projection

We now demonstrate that the discrepancy function (3.1.4) underlying the reverse I-projection has a natural statistical interpretation, which is crucial for the proof of Theorem 3.1.3.

Moderate Deviations Theory

We leverage recent results from moderate deviations theory to show that the discrepancy function (3.1.4) is intimately related to the least squares estimator (3.1.2). To this end, we first introduce the basic notions[2] of a *rate function* and a *moderate deviation principle*. For a comprehensive introduction to moderate- and large deviations theory we refer to [dH08, DZ09].

Definition 3.1.4 (Rate function)**.** *An extended real-valued function $I : \Theta' \times \Theta \to [0, \infty]$ is called a rate function if it is lower semi-continuous in its first argument.*

Definition 3.1.5 (Moderate deviation principle)**.** *A sequence of estimators $\{\widehat{\vartheta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ is said to satisfy a moderate deviation principle with rate function I if for every sequence $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ of real numbers with $1 \ll a_T \ll T$, for every Borel set $\mathcal{D} \subset \Theta'$ and for every $\theta \in \Theta$ all of the following inequalities hold.*

$$
\begin{aligned}
-\inf_{\theta'\in\operatorname{int}\mathcal{D}} I(\theta',\theta) &\leq \liminf_{T\to\infty} \frac{1}{a_T}\log \mathbb{P}_\theta\left(\widehat{\vartheta}_T \in \mathcal{D}\right) && \text{(3.1.8a)}\\
&\leq \limsup_{T\to\infty} \frac{1}{a_T}\log \mathbb{P}_\theta\left(\widehat{\vartheta}_T \in \mathcal{D}\right) && \text{(3.1.8b)}\\
&\leq -\inf_{\theta'\in\operatorname{cl}\mathcal{D}} I(\theta',\theta) && \text{(3.1.8c)}
\end{aligned}
$$

If the rate function $I(\theta', \theta)$ is continuous in θ' and the interior of $\mathcal{D}$ is dense in $\mathcal{D}$, then the infima in (3.1.8a) and (3.1.8c) coincide, which implies that all inequalities in (3.1.8) collapse to equalities. In this case, (3.1.8) can be paraphrased as $\mathbb{P}_\theta(\widehat{\vartheta}_T \in \mathcal{D}) = e^{-ra_T+o(a_T)}$, where $r = \inf_{\theta'\in\mathcal{D}} I(\theta', \theta)$ represents the I-distance between the system matrix θ and the set $\mathcal{D}$ of estimator realizations. Thus, r represents the decay rate of the probability $\mathbb{P}_\theta(\widehat{\vartheta}_T \in \mathcal{D})$, while $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ can be viewed as the speed of convergence. The condition $1 \ll a_T \ll T$ is satisfied, for example, if $a_T = \sqrt{T}$, $T \in \mathbb{Z}_{\geq 0}$. However, many other choices are possible. It is perhaps surprising that if a sequence of estimators satisfies a moderate deviations principle, then the choice of the speed $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ has no impact on the decay rate r but may only influence the coefficients of the higher-order terms hidden in $o(a_T)$. We also remark that if the inequalities in (3.1.8) hold for $a_T = T$, $T \in \mathbb{Z}_{\geq 0}$ (in which case the speed of convergence violates the condition $1 \ll a_T \ll T$), then $\{\widehat{\vartheta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ is said to satisfy a *large* deviation principle [DZ09]. It is also customary to talk about a moderate deviation principle as being a large deviation principle with reduced speed $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ such that $1 \ll a_T \ll T$.

[2] The theory is applicable to far more general settings, we define the objects only in our setting: that of estimators on subsets of Euclidean space.

We now show that the transformed least squares estimators

$$\widehat{\vartheta}_T = \sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta \tag{3.1.9}$$

satisfy a moderate deviation principle, where the discrepancy function (3.1.4) plays the role of the rate function.

Proposition 3.1.6 (Moderate deviation principle)**.** *If Assumptions 3.1.1 and 3.1.2 hold, $\{\widehat{\theta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ denote the least squares estimators defined in* (3.1.2) *and $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ is a real sequence with $1 \ll a_T \ll T$, then the transformed least squares estimators $\{\widehat{\vartheta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ defined in* (3.1.9) *satisfy a moderate deviation principle with rate function* (3.1.4)*.*

Proof. Fix any $\theta \in \mathbb{R}^{n\times n}$ and assume that $\|\theta\|_2 < 1$. This condition is stronger than Assumption 3.1.1 (i) because the spectral radius $\rho(\theta)$ is bounded above by the spectral norm $\|\theta\|_2$. Together with Assumptions 3.1.1 (ii) and 3.1.2, this condition implies via [YS09, Prop. 2.2], that the transformed least squares estimators $\{\sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta\}_{T\in\mathbb{Z}_{\geq 0}}$ satisfy a moderate deviation principle with rate function

$$\sup_{L\in\mathbb{R}^{n\times n}} \left\{\langle L, \theta' - \theta\rangle - \tfrac{1}{2}\mathbb{E}_\theta\left[\langle L, w_1 x_0^\top S_\theta^{-1}\rangle^2\right]\right\}, \tag{3.1.10}$$

where we recall that the inner product of two matrices $A, B \in \mathbb{R}^{n\times n}$ is defined as $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$.

As an intermediate step, we derive the analytical solution of the following unconstrained convex quadratic maximization problem over the matrix space $\mathbb{R}^{n\times n}$,

$$\max_{X\in\mathbb{R}^{n\times n}} \left\{\langle C, X\rangle - \tfrac{1}{2}\mathrm{Tr}\left(XB_1X^\top B_2\right)\right\}, \tag{3.1.11}$$

which is parameterized by $B_1, B_2 \in \mathcal{S}^n_{\succ 0}$ and $C \in \mathbb{R}^{n\times n}$. As the trace term $\mathrm{Tr}(XB_1X^\top B_2)$ is convex in X, we can solve (3.1.11) by setting the gradient of the objective function to zero. Hence, we find

$$\mathrm{grad}_X\left(\langle C, X\rangle - \tfrac{1}{2}\mathrm{Tr}(XB_1X^\top B_2)\right) = C^\top - B_1X^\top B_2.$$

As $B_1, B_2 \succ 0$, ones verifies that this gradient vanishes at $X^\star = B_2^{-1}CB_1^{-1}$, which implies that the optimal value of problem (3.1.11) is $\frac{1}{2}\mathrm{Tr}(B_2^{-1}CB_1^{-1}C^\top)$. Next, we rewrite the expectation in (3.1.10) as

$$\begin{aligned}\mathbb{E}_\theta\left[\langle L, w_1x_0^\top S_\theta^{-1}\rangle^2\right] &= \mathbb{E}_\theta\left[(w_1^\top LS_\theta^{-1}x_0)^2\right]\\ &= \mathbb{E}_\theta\left[w_1^\top LS_\theta^{-1}S_\theta S_\theta^{-1}L^\top w_1\right]\\ &= \mathbb{E}_\theta\mathrm{Tr}(LS_\theta^{-1}L^\top w_1w_1^\top)\\ &= \mathrm{Tr}(LS_\theta^{-1}L^\top S_w),\end{aligned}$$

where the second equality follows from Assumption 3.1.1 (ii), which implies that x_0 and w_1 are independent, and from Assumption 3.1.2 (ii), which implies that x_0 is governed by the invariant state distribution ν_θ and thus has zero mean and covariance matrix S_θ. Substituting the resulting trace term into (3.1.10) yields

$$\begin{aligned}\max_{L\in\mathbb{R}^{d\times d}} \left\{\langle\theta'-\theta, L\rangle - \tfrac{1}{2}\mathrm{Tr}\left(LS_\theta^{-1}L^\mathsf{T}S_w\right)\right\} &= \tfrac{1}{2}\mathrm{Tr}\left(S_w^{-1}(\theta'-\theta)S_\theta(\theta'-\theta)^\mathsf{T}\right)\\ &= I(\theta',\theta),\end{aligned}$$

where the first equality follows from our analytical solution of problem (3.1.11) in the special case where $B_1 = S_\theta^{-1}$, $B_2 = S_2$ and $C = \theta' - \theta$.

At last, we show that the moderate deviations principle established for $\|\theta\|_2 < 1$ remains valid for all asymptotically stable system matrices. To this end, fix any θ with $\rho(\theta) < 1$. By standard Lyapunov stability theory, there exists $P \succ 0$ with $P - \theta^\mathsf{T}P\theta \succ 0$; see, *e.g.*, [LR95, Thm. 5.3.5]. Using P, we can apply the change of variables $\bar{x}_t = P^{1/2}x_t$ and $\bar{w}_t = P^{1/2}w_t$ to obtain the auxiliary linear dynamical system

$$\bar{x}_{t+1} = \bar{\theta}\,\bar{x}_t + \bar{w}_t, \quad \bar{x}_0 \sim \bar{\nu},$$

with system matrix $\bar{\theta} = P^{1/2}\theta P^{-1/2}$, where the noise $\bar{w}_t$ has zero mean and covariance matrix $S_{\bar{w}} = P^{1/2}S_wP^{1/2}$ for all $t\in\mathbb{Z}_{\geq 0}$, and $\bar{\nu} = \nu\circ P^{-1/2}$ is the pushforward distribution of ν under the coordinate transformation $P^{1/2}$. Note also that the invariante state covariance matrix is given by $S_{\bar{\theta}} = P^{1/2}S_\theta P^{1/2}$. By construction, the auxiliary linear system is equivalent to (3.1.1) and satisfies Assumptions 3.1.1 (ii) and 3.1.2. Moreover, multiplying $P - \theta^\mathsf{T}P\theta \succ 0$ from both sides with $P^{-1/2}$ yields $I_n - \bar{\theta}^\mathsf{T}\bar{\theta} \succ 0$, which means that the largest eigenvalue of $\bar{\theta}^\mathsf{T}\bar{\theta}$ is strictly smaller than 1 or, equivalently, that $\|\bar{\theta}\|_2 < 1$. If we denote by $\widehat{\bar{\theta}}_T$ the least squares estimator for $\bar{\theta}$ based on T state observations of the auxiliary linear system, we may then conclude from the first part of the proof that the estimators $\{\sqrt{T/a_T}(\widehat{\bar{\theta}}_T - \theta) + \theta\}_{T\in\mathbb{Z}_{\geq 0}}$ satisfy a moderate deviations principle with rate function

$$\bar{I}(\bar{\theta}',\bar{\theta}) = \tfrac{1}{2}\mathrm{Tr}\left(S_{\bar{w}}^{-1}(\bar{\theta}'-\bar{\theta})S_{\bar{\theta}}(\bar{\theta}'-\bar{\theta})^\mathsf{T}\right).$$

One also readily verifies from (3.1.2) that the least squares estimators pertaining to the original and the auxiliary linear systems are related through the continuous transformation $\widehat{\bar{\theta}}_T = P^{1/2}\widehat{\theta}_TP^{-1/2}$. The corresponding *transformed* estimators evidently obey the same relation. By the contraction principle [DZ09, Thm. 4.2.1], the estimators $\{\sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta\}_{T\in\mathbb{Z}_{\geq 0}}$ thus satisfy a moderate deviations principle with rate function $\bar{I}(P^{-1/2}\theta'P^{1/2}, P^{-1/2}\theta P^{1/2}) = I(\theta',\theta)$. This observation completes the proof. □

Unlike the standard least squares estimators (3.1.2), the transformed estimators (3.1.9) depend on the unknown parameter θ. However, as we will explain below, they are useful for theoretical considerations. Proposition 3.1.6 can be viewed as a corollary of [YS09, Thm. 2.1], which uses ideas from [Wor99] to show that the transformed least squares estimators satisfy a moderate deviation principle with a rate function that is defined implicitly in variational form. Proposition 3.1.6 shows that this rate function admits the explicit representation (3.1.4) and allows for showing (3.1.6). It also relaxes the restrictive condition $\|\theta\|_2 < 1$ from [YS09, Prop. 2.2] to $\rho(\theta) < 1$.

By identifying the discrepancy function (3.1.4) with the rate function of a moderate deviation principle, Proposition 3.1.6 justifies our terminology, whereby $\mathcal{P}(\theta')$ is called the reverse I-projection of θ'. Indeed, Csiszar and Matus use this term to denote any projection with respect to an information divergence $I(\theta', \theta)$ [CM03]. Note that swapping the arguments θ' and θ of the (asymmetric) function $I(\theta', \theta)$ would give rise to an ordinary *I-projection* [Csi84]. Proposition 3.1.6 also suggests that the reverse I-projection is intimately related to maximum likelihood estimation, as already alluded to in the introduction. Indeed, for i.i.d. data it is well-known that every maximum likelihood estimator can be regarded as a reverse I-projection with respect to the rate function of some large deviation principle [CS04, Lem. 3.1].

The power of Proposition 3.1.6 lies in its generality. Indeed, a moderate deviation principle provides tight bounds on the probability of *any* Borel set of estimator realizations. A simple direct application of the moderate deviation principle established in Proposition 3.1.6 is described below.

Example 3.1.2 (System identification)**.** Consider a scalar system with $S_w = 1$ that satisfies Assumptions 3.1.1 and 3.1.2. In this case $\Theta = (-1, 1)$ with the rate function (3.1.4) reducing to $I(\theta', \theta) = \frac{1}{2}(\theta' - \theta)^2/(1 - \theta^2)$. Using the least squares estimators (3.1.2) to identify θ, Proposition 3.1.6 reveals that

$$\begin{aligned}
&\mathbb{P}_\theta(|\widehat{\theta}_T - \theta| > \varepsilon\sqrt{a_T/T}) \\
&\qquad = \mathbb{P}_\theta(\theta + \sqrt{T/a_T}(\widehat{\theta}_T - \theta) \in \mathcal{D}) \\
&\qquad = \exp\left(-\inf_{\theta' \in \mathcal{D}} I(\theta', \theta) \cdot a_T + o(a_T)\right) \\
&\qquad = \exp\left(-\tfrac{1}{2}\varepsilon^2\, a_T/(1 - \theta^2) + o(a_T)\right)
\end{aligned}$$

for any $\varepsilon > 0$ and $T \in \mathbb{Z}_{\geq 0}$, where $\mathcal{D} = \{\theta' \in \mathbb{R} : |\theta' - \theta| > \varepsilon\}$. This result confirms the insight that stable systems with $|\theta| \approx 1$ are easier to identify than systems with $|\theta| \approx 0$, *e.g.*, see [SMT$^+$18].

Next, we establish[3] several structural properties of the rate function (3.1.4).

Proposition 3.1.7 (Properties of $I(\theta', \theta)$)**.** *The rate function $I(\theta', \theta)$ defined in* (3.1.4) *has the following properties.*

(i) $I(\theta', \theta)$ is real analytic in $(\theta', \theta) \in \Theta' \times \Theta$.

(ii) If $\theta' \in \Theta' \backslash \partial\Theta$, then the sublevel set $\{\theta \in \Theta : I(\theta', \theta) \leq r\}$ is compact for every $r \geq 0$.

(iii) If $\theta' \in \Theta' \setminus \partial\Theta$, then $I(\theta', \theta)$ tends to infinity as θ approaches the boundary of Θ.

Proof. The proof of the first item follows directly from [Pol86, Lem. 3.2].

The proof of assertion (ii) consists of two steps. We first prove that if a sequence $\{\theta_k\}_{k\in\mathbb{N}}$ in Θ has an unstable limit θ (*i.e.*, $\rho(\theta) = 1$), then there exists a subsequence $\{\theta_{k_l}\}_{l\in\mathbb{N}}$ with $\lim_{l\to\infty} I(\theta', \theta_{k_l}) = \infty$ for all $\theta' \in \Theta' \setminus \partial\Theta$ (Step 1). We then use this result to show that the set $\{\theta \in \Theta : I(\theta', \theta) \leq r\}$ is compact for all $r \geq 0$ (Step 2).

Step 1: We first derive an easily computable lower bound on the rate function $I(\theta', \theta)$ for any asymptotically stable matrix $\theta \in \Theta$ and $\theta' \in \Theta' \setminus \partial\Theta$. To this end, we denote by

[3] This result is slightly more general than the original result from [JSK23], that is, with essentially no modification of the proof, we can work with $\theta' \in \Theta' \setminus \partial\Theta$ instead of $\theta' \in \Theta$.

$\lambda \in \mathbb{C}$ an eigenvalue of θ whose modulus $|\lambda|$ matches the spectral radius $\rho(\theta) < 1$. We further denote by $v \in \mathbb{C}^n$ a normalized eigenvector corresponding to the eigenvalue λ, that is, $\|v\| = 1$ and $\theta v = \lambda v$. We also use $\beta = \lambda_{\min}(S_w)/\lambda_{\max}(S_w) > 0$ as a shorthand for the inverse condition number of the noise covariance matrix $S_w \succ 0$. Recalling that for any $A, B, C \in \mathcal{S}^n_{\succeq 0}$ the semidefinite inequality $A \succeq B$ implies $\operatorname{Tr}(AC) \geq \operatorname{Tr}(BC)$, we find the following estimate.

$$\begin{aligned}
2I(\theta', \theta) &= \operatorname{Tr}\left(S_w^{-1}(\theta' - \theta)S_\theta(\theta' - \theta)^\mathsf{T}\right)\\
&\geq \lambda_{\max}^{-1}(S_w)\operatorname{Tr}\left((\theta' - \theta)S_\theta(\theta' - \theta)^\mathsf{T}\right)\\
&\geq \beta \textstyle\sum_{k=0}^\infty \operatorname{Tr}\left((\theta' - \theta)\theta^k(\theta^k)^\mathsf{T}(\theta' - \theta)^\mathsf{T}\right)\\
&= \beta \textstyle\sum_{k=0}^\infty \operatorname{Tr}\left((\theta^k)^\mathsf{T}(\theta' - \theta)^\mathsf{T}(\theta' - \theta)\theta^k\right)\\
&\geq \beta \textstyle\sum_{k=0}^\infty v^\mathsf{H}(\theta^k)^\mathsf{T}(\theta' - \theta)^\mathsf{T}(\theta' - \theta)\theta^k v\\
&= \beta \textstyle\sum_{k=0}^\infty |\lambda|^{2k} v^\mathsf{H}(\theta' - \theta)^\mathsf{T}(\theta' - \theta)v\\
&= \beta\|(\theta' - \theta)v\|_2^2 \frac{1}{1 - |\lambda|^2}
\end{aligned}$$

Here, the first equality follows from the definition of the rate function in (3.1.4), and the first inequality exploits the bound $\lambda_{\max}(S_w)I_n \succeq S_w$. The second inequality holds due to the series representation $S_\theta = \sum_{t=0}^\infty \theta^t S_w(\theta^t)^\mathsf{T}$, the bound $S_w \succeq \lambda_{\min}(S_w)I_n$ and the definition of β. The second equality exploits the cyclicity property of the trace, and the third inequality holds because any (real) matrix $C \in \mathcal{S}^n_{\succeq 0}$ satisfies

$$\operatorname{Tr}(C) \geq w^\mathsf{H} C w \quad \forall w \in \mathbb{C}^n : \|w\| = 1.$$

The third equality then uses the eigenvalue equation $\theta v = \lambda v$, and the last equality holds because $|\lambda| = \rho(\theta) < 1$. We thus conclude that the rate function admits the lower bound

$$I(\theta', \theta) \geq \frac{\beta}{2}\|(\theta' - \theta)v\|_2^2 \frac{1}{1 - |\lambda|^2}. \tag{3.1.12}$$

Consider now a converging sequence $\{\theta_k\}_{k\in\mathbb{N}}$ in Θ whose limit θ satisfies $\rho(\theta) = 1$. Define $\lambda_k \in \mathbb{C}$ as an eigenvalue of θ_k with $|\lambda_k| = \rho(\theta_k) < 1$ and let $v_k \in \mathbb{C}^n$ be a normalized eigenvector corresponding to λ_k, that is, $\|v_k\| = 1$ and $\theta_k v_k = \lambda_k v_k$. As the spectral radius is a continuous function, we then have

$$\lim_{k\to\infty} |\lambda_k| = \lim_{k\to\infty} \rho(\theta_k) = \rho(\lim_{k\to\infty} \theta_k) = \rho(\theta) = 1.$$

In addition, as the unit spheres in $\mathbb{C}$ and in $\mathbb{C}^n$ are both compact, there exists a subsequence $\{(\lambda_{k_l}, v_{k_l})\}_{l\in\mathbb{N}}$ converging to a point $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n$ with $|\lambda| = 1$ and $\|v\| = 1$. This limit satisfies the eigenvalue equation

$$\theta v = \lim_{l\to\infty} \theta_{k_l} v_{k_l} = \lim_{l\to\infty} \lambda_{k_l} v_{k_l} = \lambda v, \tag{3.1.13}$$

which implies that v is an eigenvector of θ corresponding to the eigenvalue λ with $|\lambda| = 1 = \rho(\theta)$.

The above reasoning allows us to conclude that

$$\begin{aligned}\lim_{l\to\infty} I(\theta',\theta_{k_l}) &\geq \lim_{l\to\infty} \frac{\beta}{2}\|(\theta_{k_l}-\theta')v_{k_l}\|_2^2 \frac{1}{1-|\lambda_{k_l}|^2} \\ &= \lim_{l\to\infty} \frac{\beta}{2}\|\lambda_{k_l}v_{k_l}-\theta' v_{k_l}\|_2^2 \frac{1}{1-|\lambda_{k_l}|^2} \\ &= \frac{\beta}{2}\|\lambda v-\theta' v\|_2^2 \lim_{l\to\infty} \frac{1}{1-|\lambda_{k_l}|^2} = \infty,\end{aligned}$$

where the inequality follows from (3.1.12), the first equality holds because $\theta_k v_k = \lambda_k v_k$, and the second equality exploits (3.1.13). Finally, the last equality holds because $\lim_{k\to\infty}|\lambda_k| = 1$ and because the term $\frac{\beta}{2}\|\lambda v - \theta' v\|^2$ is strictly positive. Indeed, this non-negative term can only vanish if $\theta' v = \lambda v$, which would imply that θ' has unimodular eigenvalues ($|\lambda| = 1$), which contradicts our standing assumption $\theta' \in \Theta' \setminus \partial\Theta$. This observation completes Step 1.

Step 2: Select now any $\theta' \in \Theta' \setminus \partial\Theta$ and $r \geq 0$, and define $\mathcal{A} = \{\theta \in \Theta : I(\theta',\theta) \leq r\}$. As we work with subsets of finite-dimensional Euclidean spaces, in order to prove that $\mathcal{A}$ is compact, we need to show that it is bounded and closed. This is potentially difficult because Θ itself is unbounded and open. In order to prove boundedness of $\mathcal{A}$, note that every $\theta \in \mathcal{A}$ satisfies

$$\begin{aligned} r \geq I(\theta',\theta) &= \tfrac{1}{2}\mathrm{Tr}\left(S_w^{-1}(\theta'-\theta)S_\theta(\theta'-\theta)^\intercal\right) \\ &\geq \tfrac{1}{2}\mathrm{Tr}\left(S_w^{-1}(\theta'-\theta)S_w(\theta'-\theta)^\intercal\right),\end{aligned}$$

where the second inequality follows from the trivial bound $S_\theta \succeq S_w$, which is implied by the Lyapunov equation (3.1.7). Thus, the sublevel set $\mathcal{A}$ is contained in a bounded ellipsoid,

$$\mathcal{A} \subset \left\{\theta \in \mathbb{R}^{n\times n} : \tfrac{1}{2}\mathrm{Tr}\left(S_w^{-1}(\theta'-\theta)S_w(\theta'-\theta)^\intercal\right) \leq r\right\},$$

and thus $\mathcal{A}$ is bounded. To show that $\mathcal{A}$ is closed, consider a converging sequence $\{\theta_k\}_{k\in\mathbb{N}}$ in $\mathcal{A}$ with limit θ. We first prove that $\theta \in \Theta$. Suppose for the sake of argument that $\theta \notin \Theta$. As θ is the limit of a sequence in $\mathcal{A} \subset \Theta$, this implies that θ must reside on the boundary of Θ (*i.e.*, $\rho(\theta) = 1$). By the results of Step 1, we may thus conclude that there exists a subsequence $\{\theta_{k_l}\}_{l\in\mathbb{N}}$ with $\lim_{l\to\infty} I(\theta',\theta_{k_l}) = \infty$. Clearly, we then have $I(\theta',\theta_{k_l}) > r$ for all sufficiently large l, which contradicts the assumption that $\theta_{k_l} \in \mathcal{A}$ for all $l \in \mathbb{N}$. Thus, our initial hypothesis was wrong, and we may conclude that $\theta \in \Theta$. In addition, we have

$$r \geq \lim_{k\to\infty} I(\theta',\theta_k) = I(\theta', \lim_{k\to\infty}\theta_k) = I(\theta',\theta),$$

where the inequality holds because $\theta_k \in \mathcal{A}$ for all $k \in \mathbb{N}$. Here, the first equality follows from assertion (i), which ensures that the rate function is analytic and thus continuous. Hence, we find that $\theta \in \mathcal{A}$. As the sequence $\{\theta_k\}_{k\in\mathbb{N}}$ was chosen arbitrarily, we conclude that $\mathcal{A}$ is closed. In summary, we have shown that $\mathcal{A}$ is bounded and closed and thus compact. This observation completes Step 2. Hence, assertion (ii) follows.

As for assertion (iii), fix $\theta' \in \Theta' \setminus \partial\Theta$ and consider a sequence $\{\theta_k\}_{k\in\mathbb{N}}$ in Θ whose limit θ resides on the boundary of the open set Θ. This implies that $\theta \notin \Theta$. Next, choose any $r \geq 0$. We know from assertion *(ii)* that $\mathcal{A} = \{\theta \in \Theta : I(\theta', \theta) \leq r\}$ is a compact subset of Θ, and thus $\theta \notin \mathcal{A}$. Hence, the complement of $\mathcal{A}$ represents an open neighborhood of θ, and thus there exists $k(r) \in \mathbb{N}$ such that $\theta_k \notin \mathcal{A}$ and $I(\theta', \theta_k) \geq r$ for all $k \geq k(r)$. As r was chosen freely, this means that $\lim_{k\to\infty} I(\theta', \theta_k) = \infty$.

□

Lemma 3.1.8 (Pinsker-type inequality)**.** *For any $\theta' \in \Theta'$ and $\theta \in \Theta$ we have $\|\theta' - \theta\|_2^2 \leq 2\kappa(S_w) \cdot I(\theta', \theta)$.*

Proof. By the definition of the rate function we have

$$\begin{aligned} 2I(\theta', \theta) &= \mathrm{Tr}\left(S_w^{-1}(\theta' - \theta)S_\theta(\theta' - \theta)^\mathsf{T}\right) \\ &\geq \sigma_{\min}(S_w^{-1})\sigma_{\min}(S_\theta)\|\theta' - \theta\|_F^2 \geq \frac{1}{\kappa(S_w)}\|\theta' - \theta\|_2^2, \end{aligned}$$

where the third inequality holds because $S_\theta \succeq S_w$ and $\sigma_{\min}(S_w^{-1}) = 1/\sigma_{\max}(S_w)$. □

Lemma 3.1.8 provides a direct link between the nearest stable matrix problem (3.1.3) and the reverse I-projection (3.1.5) as

$$\inf_{\theta\in\Theta} \|\theta' - \theta\|_2^2 \leq 2\kappa(S_w) \cdot I(\theta', \mathcal{P}(\theta')).$$

Statistics of the reverse I-projection

In the following we apply the reverse I-projection to the least squares estimator $\widehat{\theta}_T$. An elementary calculation shows that $I(\widehat{\theta}_T, \theta) = (a_T/T)I(\widehat{\vartheta}_T, \theta)$, and thus $I(\widehat{\theta}_T, \theta)$ inherits any statistical interpretations from $I(\widehat{\vartheta}_T, \theta)$.

We first show that $\mathcal{P}(\widehat{\theta}_T)$ is asymptotically consistent.

Proposition 3.1.9 (Asymptotic consistency)**.** *Suppose that Assumption 3.1.1 holds and that $\widehat{\theta}_T$ is the least squares estimator. Then, for any $\theta \in \Theta$ the reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$ of $\widehat{\theta}_T$ satisfies $\lim_{T\to\infty} \mathcal{P}(\widehat{\theta}_T) = \theta$ $\mathbb{P}_\theta$-almost surely.*

Proof. Recall that $\lim_{T\to\infty} \widehat{\theta}_T = \theta$ $\mathbb{P}_\theta$-almost surely [CK98]. Therefore, we have $\mathbb{P}_\theta$-almost surely that

$$\begin{aligned} \lim_{T\to\infty} \mathcal{P}(\widehat{\theta}_T) &= \lim_{T\to\infty} \arg\min_{\bar{\theta}\in\Theta} I(\widehat{\theta}_T, \bar{\theta}) \\ &= \arg\min_{\bar{\theta}\in\Theta} \lim_{T\to\infty} I(\widehat{\theta}_T, \bar{\theta}) \\ &= \arg\min_{\bar{\theta}\in\Theta} I\left(\lim_{T\to\infty} \widehat{\theta}_T, \bar{\theta}\right) \\ &= \arg\min_{\bar{\theta}\in\Theta} I(\theta, \bar{\theta}) = \theta, \end{aligned}$$

where the first equality exploits the definition of $\mathcal{P}(\widehat{\theta}_T)$ in (3.1.5). The second equality follows from the strict convexity of the rate function in its first argument and [Sun96, Thm. 9.17], which imply that the reverse I-projection is continuous. The third equality follows from the continuity of the rate function established in Proposition 3.1.7 (i), and the last equality holds because the rate function vanishes if and only if its arguments coincide. This proves the proposition. □

Next, we can use the results of Section 3.1.3 to establish probabilistic bounds. Specifically, the following lemma provides two *implicit* finite-sample bounds involving random error estimates. These bounds are both structurally identical to existing finite-sample bounds for $\widehat{\theta}_T$; see, *e.g.*, [SR19, Sec. 6]. In Proposition 3.1.12 below, these implicit bounds will be used to establish *explicit* finite sample bounds involving deterministic error estimates.

Lemma 3.1.10 (Implicit finite sample bounds)**.** *Suppose that Assumptions 3.1.1 and 3.1.2 hold and that $\widehat{\theta}_T$ and $\mathcal{P}(\widehat{\theta}_T)$ represent the least squares estimator and its reverse I-projection, respectively. Setting $\widehat{\varepsilon}_T = (2\kappa(S_w) I(\widehat{\theta}_T, \mathcal{P}(\widehat{\theta}_T)))^{1/2}$, we then have $\|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \widehat{\varepsilon}_T$ $\mathbb{P}_\theta$-almost surely. In addition, the following finite sample bounds hold for all $\beta, \varepsilon \in (0,1)$.*

(i) We have

$$\mathbb{P}_\theta \left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \varepsilon + \widehat{\varepsilon}_T \right) \geq 1 - \beta \tag{3.1.14a}$$

for all $T \in \mathbb{Z}_{\geq 0}$ with $T \geq \kappa(S_w) \widetilde{O}(n) \log(1/\beta)/\varepsilon^2$.

(ii) If $\{a_T\}_{T \in \mathbb{Z}_{\geq 0}}$ is a real sequence satisfying $1 \ll a_T \ll T$, then we have

$$\mathbb{P}_\theta \left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \varepsilon \sqrt{a_T/T} + \widehat{\varepsilon}_T \right) \geq 1 - \beta \tag{3.1.14b}$$

for all $T \in \mathbb{Z}_{\geq 0}$ with $a_T \geq 2\kappa(S_w)(\log(1/\beta) + o(a_T))/\varepsilon^2$.

Proof. Lemma 3.1.8 and the monotonicity of the square root function imply that $\|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \widehat{\varepsilon}_T$ is a $\mathbb{P}_\theta$-almost sure event. As for assertion (i), we thus have

$$\begin{aligned}
\mathbb{P}_\theta \left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \varepsilon + \widehat{\varepsilon}_T \right) &\geq \mathbb{P}_\theta \left(\|\theta - \widehat{\theta}_T\|_2 + \|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \varepsilon + \widehat{\varepsilon}_T \right) \\
&\geq \mathbb{P}_\theta \left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon, \ \|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \widehat{\varepsilon}_T \right) \\
&= \mathbb{P}_\theta \left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon \right).
\end{aligned}$$

Hence, to estimate the probability, we can leverage tools developed in [SR19, Sec. 6]. To this end, assume first that the noise is isotropic, *i.e.*, assume that $S_w = \alpha I_n$ for some $\alpha > 0$. In this case, [SR19, Thm. 1] implies that $\mathbb{P}_\theta(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon) \geq 1 - \beta$ for all $\beta, \varepsilon \in (0,1)$ and sample sizes $T \geq \widetilde{O}(n) \log(1/\beta)/\varepsilon^2$. As $\kappa(S_w) = \kappa(\alpha I_n) = 1$, this settles assertion (i) when the noise is isotropic.

Assume now that the noise is anisotropic with an arbitrary convariance matrix $S_w \succ 0$. The change of coordinates $\bar{x}_t = S_w^{-1/2} x_t$ and $\bar{w}_t = S_w^{-1/2} w_t$ then yields the auxiliary system

$$\bar{x}_{t+1} = \bar{\theta}\, \bar{x}_t + \bar{w}_t, \quad \bar{x}_0 \sim \nu \circ S_w^{1/2},$$

with $\bar{\theta} = S_w^{-1/2}\theta S_w^{1/2}$ and isotropic noise $\bar{w}_t$ having zero mean and unit covariance matrix for all $t \in \mathbb{Z}_{\geq 0}$. Denoting by $\widehat{\bar{\theta}}_T$ the least squares estimator for the auxiliary system, we find

$$\begin{aligned}\mathbb{P}_\theta\left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon\right) &= \mathbb{P}_\theta\left(\|S_w^{1/2}(\bar{\theta} - \widehat{\bar{\theta}}_T)S_w^{-1/2}\|_2 \leq \varepsilon\right)\\ &\geq \mathbb{P}_\theta\left(\|S_w^{1/2}\|_2\|\bar{\theta} - \widehat{\bar{\theta}}_T\|_2\|S_w^{-1/2}\|_2 \leq \varepsilon\right)\\ &= \mathbb{P}_\theta\left(\|\bar{\theta} - \widehat{\bar{\theta}}_T\|_2 \leq \varepsilon\kappa(S_w)^{-1/2}\right),\end{aligned}$$

where the last equality holds because $\|S_w^{1/2}\|_2\|S_w^{-1/2}\|_2 = \kappa(S_w^{1/2}) = \kappa(S_w)^{1/2}$. From the first part of the proof for linear systems driven by isotropic noise we know that the resulting probability is no less than $1-\beta$ whenever $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$. This observation completes the proof of assertion (i).

The proof of assertion (ii) first parallels that of assertion (i). In particular, multiplying ε with $\sqrt{a_T/T}$ yields

$$\mathbb{P}_\theta\left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \varepsilon\sqrt{a_T/T} + \widehat{\varepsilon}_T\right) \geq \mathbb{P}_\theta\left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon\sqrt{a_T/T}\right).$$

However, now we use the moderate deviations principle from Section 3.1.3 to bound the resulting probability. To this end, define $\mathcal{D} = \{\theta' \in \mathbb{R}^{n\times n} : \|\theta' - \theta\|_2 > \varepsilon\}$. By Lemma 3.1.8, we have $I(\theta', \theta) > \varepsilon^2/(2\kappa(S_w))$ for any estimator realization $\theta' \in \mathcal{D}$, and thus $\inf_{\theta' \in \mathrm{cl}\,\mathcal{D}} I(\theta', \theta) \geq \varepsilon^2/(2\kappa(S_w))$. Recall now from Proposition 3.1.6 that the transformed least squares estimators $\widehat{\vartheta}_T = \sqrt{T/a_T}(\widehat{\theta}_T - \theta) + \theta$ obey a moderate deviations principle with rate function I. Hence, we have

$$\begin{aligned}\limsup_{T\to\infty} \frac{1}{a_T}\log\mathbb{P}_\theta\left(\|\widehat{\theta}_T - \theta\|_2 > \varepsilon\sqrt{a_T/T}\right) &= \limsup_{T\to\infty}\frac{1}{a_T}\log\mathbb{P}_\theta(\widehat{\vartheta}_T \in \mathcal{D})\\ &\leq -\inf_{\theta' \in \mathrm{cl}\,\mathcal{D}} I(\theta', \theta) \leq -\varepsilon^2/(2\kappa(S_w)),\end{aligned}$$

where the equality exploits the definitions of $\mathcal{D}$ and $\widehat{\vartheta}_T$, and the first inequality follows from Proposition 3.1.6. By passing over to complementary events, we therefore obtain

$$\mathbb{P}_\theta\left(\|\widehat{\theta}_T - \theta\|_2 \leq \varepsilon\sqrt{a_T/T}\right) \geq 1 - e^{-\varepsilon^2 a_T/(2\kappa(S_w)) + o(a_T)}.$$

For all sufficiently large sample sizes T satisfying the inequality $a_T \geq 2\kappa(S_w)(\log(1/\beta) + o(a_T))/\varepsilon^2$ this implies that

$$\mathbb{P}_\theta\left(\|\widehat{\theta}_T - \theta\|_2 \leq \varepsilon\sqrt{a_T/T}\right) \geq 1 - \beta.$$

This observation completes the proof of assertion (ii). □

Note that the finite sample bound (3.1.14a), which leverages sophisticated results from [SR19, Sec. 6], and the bound (3.1.14b), which follows almost immediately from the moderate deviations principle of Section 3.1.3, are qualitatively similar. They both hold for all T that exceed a critical sample size depending on an unknown deterministic function of the order $\widetilde{O}(n)$ or $o(a_T)$, respectively. Both bounds also involve a random error estimate $\widehat{\varepsilon}_T$. As $\widehat{\theta}_T$ as well as $\mathcal{P}(\widehat{\theta}_T)$ converge $\mathbb{P}_\theta$-almost surely to $\theta \in \Theta$, and as I is continuous in both of its arguments, it is easy to show that the random variable $\widehat{\varepsilon}_T$ as defined in Proposition 3.1.10 converges $\mathbb{P}_\theta$-almost surely to 0 as T grows. Therefore, the bounds (3.1.14a) and (3.1.14b) improve with T. As the inequalities in (3.1.8) are asymptotically tight, we conjecture that the bound (3.1.14b) is statistically optimal.

In the following we will show that the implicit finite sample bounds of Lemma 3.1.10 can be used to derive explicit finite sample bounds involving deterministic error estimates. To this end, we recall a more nuanced *quantitative* notion of stability.

Definition 3.1.11 ((τ,ρ)-stability [KTR19, Def. 1])**.** *We say that the system matrix $\theta \in \Theta$ is (τ,ρ)-stable for some $\tau \geq 1$ and $\rho \in (0,1)$ if $\|\theta^k\|_2 \leq \tau\rho^k$ for all $k \in \mathbb{Z}_{\geq 0}$.*

We emphasize that any stable matrix $\theta \in \Theta$ is in fact (τ,ρ)-stable for some $\tau \geq 1$ and $\rho \in (0,1)$. If θ is diagonalizable with spectral decomposition $\theta = T\Lambda T^{-1}$, for example, then $\|\theta^k\| = \|T\Lambda^k T^{-1}\| \leq \kappa(T)\rho(\theta)^k$, which implies that θ is (τ,ρ) stable for $\tau = \kappa(T)$ and $\rho = \rho(\theta)$. If θ is not diagonalizable, a similar but more involved argument is used, akin to showing Schur stability.

Proposition 3.1.12 (Explicit finite sample bounds)**.** *Suppose that Assumptions 3.1.1 and 3.1.2 hold and that $\widehat{\theta}_T$ and $\mathcal{P}(\widehat{\theta}_T)$ are the least squares estimator and its reverse I-projection, respectively. The following finite sample bounds hold for all $\beta, \varepsilon \in (0,1)$ and for all parameters $\tau \geq 1$ and $\rho \in (0,1)$ such that θ is (τ,ρ)-stable, which are guaranteed to exist.*

(i) We have

$$\mathbb{P}_\theta\left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \kappa(S_w)\frac{2\varepsilon n^{1/2}\tau}{\sqrt{1-\rho^2}}\right) \geq 1-\beta$$

for all $T \in \mathbb{Z}_{\geq 0}$ with $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$.

(ii) If $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ is a real sequence satisfying $1 \ll a_T \ll T$ and $T \in \mathbb{Z}_{\geq 0}$, then we have

$$\mathbb{P}_\theta\left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \kappa(S_w)\frac{2\varepsilon n^{1/2}\tau}{\sqrt{1-\rho^2}}\sqrt{\frac{a_T}{T}}\right) \geq 1-\beta$$

for all $T \in \mathbb{Z}_{\geq 0}$ with $a_T \geq 2\kappa(S_w)(\log(1/\beta) + o(a_T))/\varepsilon^2$.

Proof. As θ is (τ,ρ)-stable, the defining properties of the reverse I-projection imply that

$$\begin{aligned} I(\widehat{\theta}_T, \mathcal{P}(\widehat{\theta}_T)) \leq I(\widehat{\theta}_T, \theta) =& \tfrac{1}{2}\mathrm{Tr}\left(S_w^{-1}(\widehat{\theta}_T - \theta)S_\theta(\widehat{\theta}_T - \theta)^\top\right) \\ \leq& \tfrac{1}{2}\mathrm{Tr}(S_w^{-1})\|\widehat{\theta}_T - \theta\|_2^2\|S_\theta\|_2 \\ \leq& \tfrac{1}{2}n\kappa(S_w)\|\widehat{\theta}_T - \theta\|_2^2\frac{\tau^2}{1-\rho^2}, \end{aligned}$$

where the second inequality holds because $\mathrm{Tr}(AB) \leq \mathrm{Tr}(A)\|B\|_2$ for any two symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, while the third inequality follows from [KTR19, Prop. E.5]. Hence, up to problem-dependent constants, $I(\widehat{\theta}_T, \mathcal{P}(\widehat{\theta}_T))$ decays as least as fast as $\|\widehat{\theta}_T - \theta\|_2^2$. Combining the above estimate with Lemma 3.1.8 and taking square roots then yields

$$\|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \|\widehat{\theta}_T - \theta\|_2 \kappa(S_w) \frac{n^{1/2}\tau}{\sqrt{1-\rho^2}}. \tag{3.1.16}$$

Setting $\eta = \kappa(S_w)n^{1/2}\tau/\sqrt{1-\rho^2} \geq 1$, we may use a similar reasoning as in the proof Lemma 3.1.10 to obtain

$$\begin{aligned}\mathbb{P}_\theta\left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq 2\eta\varepsilon\right) &\geq \mathbb{P}_\theta\left(\|\theta - \widehat{\theta}_T\|_2 \leq \eta\varepsilon,\ \|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \eta\varepsilon\right)\\ &\geq \mathbb{P}_\theta\left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon,\ \|\widehat{\theta}_T - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \eta\|\theta - \widehat{\theta}_T\|_2\right)\\ &= \mathbb{P}_\theta\left(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon\right),\end{aligned}$$

where the second inequality holds because $\eta \geq 1$, the equality follows from (3.1.16), which holds with certainty. However, from the proof of Lemma 3.1.10 (i) we already know that $\mathbb{P}_\theta(\|\theta - \widehat{\theta}_T\|_2 \leq \varepsilon) \geq 1-\beta$ whenever $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$. This observation completes the proof of assertion (i).

The proof of assertion (ii) widely parallels that of assertion (i) and is thus omitted for brevity. □

The explicit finite-sample bounds of Proposition 3.1.12 refine the implicit bounds of Lemma 3.1.10 and notably expose the dependence of the approximation error on the stability parameters τ and ρ. Of course, these parameters are unknown under our standing assumption that θ is unknown, as such we cannot adapt the projection (3.1.5) to incorporate (τ, ρ)-stability. In contrast, the implicit finite-sample bounds of Lemma 3.1.10 involve approximation errors that are random but known.

As the finite-sample bounds established in Lemma 3.1.10 (i) and Proposition 3.1.12 (i) critically rely on [SR19], they depend on the sub-Gaussianity parameter α from Assumption 3.1.2 (i). In particular, the sample complexity deteriorates for small α. The exact dependency can be inferred from [SR19, Sec. 8–10].

Computation of the reverse I-projection

We now address the numerical computation of $\mathcal{P}(\theta')$ as defined in (3.1.5) for any given estimator realization $\theta' \in \Theta' \setminus \partial\Theta$. To this end, we fix $Q \succ 0$ and show that solving (3.1.5) is equivalent to finding a minimizer of the optimization problem

$$\min_{\theta\in\mathbb{R}^{n\times n}} \{\mathrm{Tr}(QS_\theta) : I(\theta', \theta) \leq r\} \tag{3.1.17}$$

for the smallest radius $r = \underline{r}$ that renders (3.1.17) feasible. Note that $\underline{r}$ exists because the optimal value of (3.1.17) is lower semi-continuous in r (*e.g.*, by Berge's theorem). In addition, problem (3.1.17) admits a minimizer for any $r \geq \underline{r}$ due to Proposition 3.1.7. Moreover, the proposed procedure is computationally attractive because we will prove below that (3.1.17) is equivalent to a standard LQR problem. We emphasize that the exact choice of Q has no effect on the validity and hardly any effect on the numerical performance of this procedure. Summarizing the discussion, we have the following.

Proposition 3.1.13 (Reformulation of (3.1.5))**.** *If $\theta' \in \Theta' \setminus \partial\Theta$, $Q \succ 0$ and $\underline{r}$ is the smallest $r \geq 0$ for which (3.1.17) is feasible, then any minimizer of (3.1.17) at $r = \underline{r}$ is a reverse I-projection.*

Note that if we consider $r \geq \overline{r} = I(\theta', 0)$, then problem (3.1.17) has simply the trivial solution $\theta = 0$, and its optimal value reduces to $\mathrm{Tr}(QS_w)$.[4] In this case, the rate constraint is not binding at optimality. If $r < \overline{r}$, then problem (3.1.17) is infeasible if additionally $r < \underline{r}$, the problem admits a quasi-closed form solution for $r > \underline{r}$ as explained in the following proposition.

Proposition 3.1.14 (Optimal solution of (3.1.17))**.** *Suppose that Assumption 3.1.1 holds. Then, for every $\theta' \in \Theta' \setminus \partial\Theta$ there exists an analytic function $\varphi : (\underline{r}, \overline{r}) \to (0, \infty)$ that is increasing and bijective such that the following hold for all $r \in (\underline{r}, \overline{r})$.*

(i) For any $\delta \in (0, \infty)$ the matrix $P_\delta \in \mathcal{S}^n$ is the unique positive definite solution of the Riccati equation

$$P_\delta = Q + {\theta'}^\top P_\delta \left(I_n + 2\delta S_w P_\delta\right)^{-1} \theta'. \tag{3.1.18}$$

(ii) The matrix $\theta^\star_\delta = (I_n + 2\delta S_w P_\delta)^{-1}\theta'$ is the unique solution of problem (3.1.17) at $r = \varphi^{-1}(\delta)$, and the rate constraint is binding at optimality, i.e., $I(\theta', \theta^\star_\delta) = r$.

The approximate computation of the reverse I-projection exploits standard results on infinite-horizon dynamic programming (see, *e.g.*, [BB95, Ch. 3] or [Ber07]) as well as the following exact constraint relaxation (Lagrangian) result borrowed from [Jon19, Lem. A-0.1]; see also [JSM19]. We repeat this result here, again, to keep the thesis somewhat self-contained.

Lemma 3.1.15 (Exact constraint relaxation)**.** *Let f and g be two arbitrary functions from Θ to $(-\infty, \infty]$, and consider the two minimization problems*

$$\begin{aligned} \mathcal{P}_1(r) \ &: \ \inf_{\theta \in \Theta} \left\{ f(\theta) : g(\theta) \leq r \right\} \\ \mathcal{P}_2(\delta) \ &: \ \inf_{\theta \in \Theta} f(\theta) + \frac{g(\theta)}{\delta} \end{aligned}$$

parametrized by $r \in \mathbb{R}$ and $\delta \in (0, \infty)$, respectively. If the penalty-based minimization problem $\mathcal{P}_2(\delta)$ admits an optimal solution $\theta^\star_2(\delta)$ for the parameter values δ within some set $\Delta \subset (0, \infty)$, then the following hold.

(i) The function $h(\delta) = g(\theta^\star_2(\delta))$ is non-decreasing in the parameter $\delta \in \Delta$.

(ii) If there exits $\delta \in \Delta$ with $h(\delta) = r$, then the constrained minimization problem $\mathcal{P}_1(r)$ is solved by $\theta^\star_2(\delta)$.

[4] One readily verifies that $\overline{r} = \frac{1}{2}\|{S_w}^{-1/2}\theta' S_w^{1/2}\|_F^2$, where $\|\cdot\|_F$ stands for the Frobenius norm.

Proof. In order to prove assertion (i), choose any parameters $\delta_1, \delta_2 \in \Delta$ with $\delta_1 > \delta_2$. As $\theta_2^\star(\delta_1)$ is optimal in $\mathcal{P}_2(\delta_1)$ and $\theta_2^\star(\delta_2)$ is optimal in $\mathcal{P}_2(\delta_2)$, one can readily verify that

$$f\big(\theta_2^\star(\delta_1)\big) + \frac{g(\theta_2^\star(\delta_1))}{\delta_1} \leq f(\theta_2^\star(\delta_2)) + \frac{g(\theta_2^\star(\delta_2))}{\delta_1}$$

and

$$f(\theta_2^\star(\delta_2)) + \frac{g(\theta_2^\star(\delta_2))}{\delta_2} \leq f\big(\theta_2^\star(\delta_1)\big) + \frac{g(\theta_2^\star(\delta_1))}{\delta_2}.$$

Summing up these two inequalities yields

$$\begin{aligned} &\left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right) g\big(\theta_2^\star(\delta_2)\big) \leq \left(\frac{1}{\delta_2} - \frac{1}{\delta_1}\right) g\big(\theta_2^\star(\delta_1)\big) \\ \iff & h(\delta_2) = g\big(\theta_2^\star(\delta_2)\big) \leq g\big(\theta_2^\star(\delta_1)\big) = h(\delta_1), \end{aligned}$$

where the equivalence holds because $\delta_1 > \delta_2$. This completes the proof of assertion (i).

As for assertion (ii), fix any $r \in \mathbb{R}$ and assume that there exists $\delta \in \Delta$ with $r = h(\delta)$. We need to show that the optimizer $\theta_2^\star(\delta)$ of $\mathcal{P}_2(\delta)$ is also optimal in $\mathcal{P}_1(r)$. To this end, observe that $\theta_2^\star(\delta)$ is feasible in $\mathcal{P}_1(r)$ because $r = h(\delta) = g(\theta_2^\star(\delta))$. It then suffices to prove optimality. Assume for the sake of contradiction that there exists $\theta_1' \in \Theta$ with $f(\theta_1') < f(\theta_2^\star(\delta))$ and $g(\theta_1') \leq g(\theta_2^\star(\delta)) = r$. In this case, we have

$$f(\theta_1') + \frac{g(\theta_1')}{\delta} < f\big(\theta_2^\star(\delta)\big) + \frac{g(\theta_2^\star(\delta))}{\delta},$$

which contradicts the optimality of $\theta_2^\star(\delta)$ in $\mathcal{P}_2(\delta)$. We thus conclude that $\theta_2^\star(\delta)$ must indeed solve $\mathcal{P}_1(r)$. □

Proof of Proposition 3.1.14. Fix any $\theta' \in \Theta' \setminus \partial\Theta$, and identify the reverse I-projection problem (3.1.17) with problem $\mathcal{P}_1(r)$ from Lemma 3.1.15, that is, set $f(\theta) = \mathrm{Tr}(QS_\theta)$ and $g(\theta) = I(\theta', \theta)$. By the definition of the rate function I in (3.1.4), the corresponding unconstrained problem $\mathcal{P}_2(\delta)$ is equivalent to

$$\begin{aligned} &\min_{\theta\in\Theta} \mathrm{Tr}(QS_\theta) + \frac{1}{\delta} I(\theta', \theta) \\ &= \min_{\theta\in\Theta} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\theta\Big[\textstyle\sum_{k=0}^{T-1} x_k^\mathsf{T} Q x_k + \frac{1}{2\delta} x_k^\mathsf{T} (\theta' - \theta)^\mathsf{T} S_w^{-1} (\theta' - \theta) x_k\Big] \\ &= \min_{L\in\mathbb{R}^{n\times n}} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\theta'+L}\Big[\textstyle\sum_{k=0}^{T-1} x_k^\mathsf{T} \left(Q + \frac{1}{2\delta} L^\mathsf{T} S_w^{-1} L\right) x_k\Big], \end{aligned}$$

where the first equality exploits the Markov law of large numbers. The second equality follows from the variable substitution $L \leftarrow \theta' - \theta$. Note that the constraint $\theta' + L \in \Theta$ can be relaxed because $\mathbb{E}_{\theta'+L}[x_k^\mathsf{T} Q x_k]$ diverges with k whenever $\theta' + L$ is unstable. Indeed, in this case the trace of the covariance matrix of x_k explodes. Also, we remark again that $\mathbb{E}_{\theta'+L}[\cdot]$ merely indicates that the distribution is parametric in $\theta' + L$, the variables θ' and L are not random in the above.

As any infinite horizon LQR problem with average cost criterion is solved by a linear control policy of the form $u_k = Lx_k$ for some $L \in \mathbb{R}^{n\times n}$, problem $\mathcal{P}_2(\delta)$ is equivalent to

$$\begin{array}{ll} \min\limits_{\varphi_k(\cdot)} & \lim\limits_{T\to\infty} \dfrac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T-1} x_k^\intercal Q x_k + u_k^\intercal R u_k\right] \\ \text{s.t.} & x_{k+1} = \theta' x_k + u_k + w_k, \quad x_0 \sim \nu, \\ & u_k = \varphi_k(x_k), \end{array} \tag{3.1.19}$$

where $R = (2\delta S_w)^{-1}$. As standard stabilizability and detectability assumptions are trivially satisfied [Ber05, Ch. 4], the LQR problem (3.1.19) is solvable for every $\delta > 0$. Its optimal solution is a stationary linear control policy with state feedback gain $L_\delta = -(P_\delta + R)^{-1}P_\delta\theta'$, where P_δ is the unique positive definite solution of the Riccati equation

$$P_\delta = Q + \theta'^\intercal P_\delta \theta' - \theta'^\intercal P_\delta (P_\delta + R)^{-1} P_\delta \theta'.$$

Note that this equation is equivalent to (3.1.18) by standard matrix inversion results [VV07, p. 19] and the definition of R. Hence, problem $\mathcal{P}_2(\delta)$ is solved by

$$\theta_2^\star(\delta) = \theta' + L_\delta = \theta' - (P_\delta + R)^{-1}P_\delta\theta' = (I_n + R^{-1}P_\delta)^{-1}\theta'$$

for any $\delta > 0$. Note that the last expression is equivalent to $\theta_\delta^\star$ from the proposition statement. By using [Pol86, Lem. 3.2], one can show that P_δ and consequently also $\theta_\delta^\star$ are real-analytic in $\delta > 0$. As the rate function I is analytic thanks to Proposition 3.1.7 (i), the function $\varphi^{-1}(\delta) = I(\theta', \theta_\delta^\star)$ is thus analytic as a composition of two analytic functions. In addition, $\varphi^{-1}(\delta)$ is non-decreasing thanks to Lemma 3.1.15 (i). As any non-decreasing analytic function that is not constant must be strictly monotonically increasing, we may conclude that $\varphi^{-1} : (0,\infty) \to (\underline{r}, \overline{r})$ is bijective, where

$$\underline{r} = \lim_{\delta\downarrow 0} \varphi^{-1}(\delta) \quad \text{and} \quad \overline{r} = \lim_{\delta\uparrow\infty} \varphi^{-1}(\delta).$$

Note that if δ tends to 0, then problem $\mathcal{P}_2(\delta)$ just minimizes $g(\theta) = I(\theta', \theta)$ over Θ, in which case the reverse I-projection $\mathcal{P}(\theta')$ is optimal. Recall that $\mathcal{P}(\theta')$ is also optimal in problem $\mathcal{P}_1(r)$ for $r = I(\theta', \mathcal{P}(\theta')) = \underline{r}$. Note also that if δ tends to ∞, then problem $\mathcal{P}_2(\delta)$ just minimizes $f(\theta) = \mathrm{Tr}(QS_\theta)$ over Θ, in which case the trivial solution $\theta = 0$ is optimal. Clearly, 0 is also optimal in problem $\mathcal{P}_1(r)$ for $r = I(\theta', 0) = \overline{r}$. Hence, we may define $\varphi : (\underline{r}, \overline{r}) \to (0,\infty)$ as the inverse of φ^{-1}. By construction, φ is analytic, strictly increasing and bijective.

In summary, Lemma 3.1.15 implies that for each $r \in (\underline{r}, \overline{r})$ we may set $\delta = \varphi(r)$ such that $\theta_\delta^\star$ is the unique optimal solution of problem $\mathcal{P}_1(r)$, and this solution satisfies $I(\theta', \theta_\delta^\star) = r$. □

We have seen that evaluating $\mathcal{P}(\theta')$ is equivalent to solving (3.1.17) at $r = \underline{r}$. Unfortunately, $\underline{r}$ is unknown, and Proposition 3.1.14 only characterizes solutions of (3.1.17) for $r > \underline{r}$. However, by the properties of φ established in Proposition 3.1.14, we also have $\lim_{r\downarrow\underline{r}} \varphi(r) = 0$, which is equivalent to $\lim_{\delta\downarrow 0} \varphi^{-1}(\delta) = \underline{r}$. A standard continuity argument therefore implies that $\lim_{\delta\downarrow 0} \theta_\delta^\star$ solves (3.1.17) at $r = \underline{r}$. In practice, we may simply set δ to a small positive number and compute $\theta_\delta^\star$ by solving (3.1.18) to find a high-accuracy approximation for the reverse I-projection $\mathcal{P}(\theta')$.

Proposition 3.1.16 (Computing the reverse I-projection)**.** *If Assumption 3.1.1 holds, $\theta' \in \Theta' \setminus \partial\Theta$ and $Q \succ 0$, then there exists a $p \geq 1$ such that for all $\delta > 0$ the matrix $\theta^\star_\delta$ from Proposition 3.1.14 (ii) is stable and satisfies*

$$\|\mathcal{P}(\theta') - \theta^\star_\delta\|_2 = O(\delta^p). \tag{3.1.20}$$

In addition, $\theta^\star_\delta$ can be computed as

$$\theta^\star_\delta = \theta' + \mathsf{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1}), \tag{3.1.21}$$

where the standard LQR routine $\mathsf{dlqr}(\cdot)$ has time and memory complexity of the order $O(n^3)$ and $O(n^2)$, respectively.

Proof. Fix any $\theta' \in \Theta' \setminus \partial\Theta$. Proposition 3.1.14 implies that $\lim_{\delta\downarrow 0} \theta^\star_\delta = \mathcal{P}(\theta')$. However, one cannot simply evaluate $\theta^\star_\delta$ at $\delta = 0$ (in general). Nevertheless, the error bound (3.1.20) follows directly from the Pinsker-type inequality established in Lemma 3.1.8 and from the analyticity of $I(\theta', \theta^\star_\delta)$ in $\delta \in (0, \infty)$. For $\delta > 0$, the proof of Proposition 3.1.14 reveals that $\theta^\star_\delta$ can be computed by solving problem (3.1.19), which can be addressed with standard LQR routines. Hence, the computational bottleneck is the solution of the Riccati equation (3.1.18). The state-of-the-art methods to solve (3.1.18) utilize a QZ algorithm that has time and memory complexity of the order $O(n^3)$ and $O(n^2)$, respectively; see, *e.g.*, [PLS80] and [GvL13, Alg. 7.7.3]. However, large problem instances should be addressed with alternative schemes such as the ones proposed in [GL91, BF11]. □

As mentioned before, we refer to [Jon22] for more on the computation. In particular, we propose to use a QZ algorithm to compute $\theta^\star_\delta$.

Corollary 3.1.17 ($\mathcal{P}(\theta')$ and $\theta^\star_\delta$ preserve the structure of θ')**.** *For any $\theta' \in \Theta' \setminus \partial\Theta$ there exist invertible matrices $\Lambda, \Lambda_\delta \in \mathbb{R}^{n\times n}$ such that $\mathcal{P}(\theta') = \Lambda^{-1}\theta'$ and $\theta^\star_\delta = \Lambda_\delta^{-1}\theta'$.*

Corollary 3.1.17 follows directly from the above in combination with Section 4.1 below and alludes to the preservation of structure, which is precisely what we exploit in Section 3.2.

Final remarks

First, to conclude, we have all the tools in place to prove Theorem 3.1.3.

Proof of Theorem 3.1.3. The three assertions follow directly from Propositions 3.1.9, 3.1.12 and 3.1.16, respectively. □

Secondly, we were pedantic in pointing out that $\theta' \in \Theta' \setminus \partial\Theta$. From a practical and statistical point of view this is of course irrelevant. In fact, as our results hold either asymptotically or for a sufficiently large T, this constraint can be omitted since there is a $\bar{T}$ such that $\mathbb{P}_\theta(\widehat{\theta}_T \in \partial\Theta) = 0$ for all $T \geq \bar{T}$ (due to $\mathbb{P}_\theta$-almost surely convergence of $\widehat{\theta}_T$ and Θ being open). We also point out that during the time of writing, a new approach (SIMBa) to optimize over (not just project) stable matrices has been proposed, which appears promising [DNZH+23].

3.2 Topological linear system identification via moderate deviations theory

Elaborating on Section 3.1 we show in this section how the flexibility of large (moderate) deviations theory helps us in further refining qualitatively correct system identification.

3.2.1 Introduction

We consider the same setting as in Section 3.1, that is, we work with the least squares estimators $\{\widehat{\theta}_T\}_T$ (3.1.2) for the system (3.1.1). Previously we exclusively focused on (asymptotic) stability. However, stability is not the only property of θ that impacts the qualitative behaviour of a linear system; see Section 2.3.1. In fact—as we elaborate on in Section 4.1, the closed-loop system corresponding to LQ optimal regulation (with a block-diagonal cost) preserves the structure of the system matrix. Hence, if the least squares estimator $\widehat{\theta}_T$ is structurally different from θ itself, even despite stability, then implementing optimal linear feedback designed for $\widehat{\theta}_T$ results in a *closed-loop* system that is structurally different from the predicted closed-loop system, *e.g.*, you predict a damper but get a spring.

As we discussed before, linear system identification—especially by means of least squares techniques—has a rich history [VODM96, VV07]. In this section we are, however, not only interested in finding estimators that fall into the vicinity of the unknown true model θ. In addition, the estimators should give rise to the same qualitative behaviour as θ. This requirement relates to some extent to the work on *qualitative identification* pioneered by Kuipers [Kui94].

More recently, the focus in linear system identification shifted towards ensuring the efficient use of data. General *informativity* of data is discussed in [VWETC20], which justifies the identification pipeline for a class of control problems. Moreover, sharp statistical characterizations of the effectiveness of the least squares estimator (3.1.2) are presented in [SMT$^+$18, SR19]. These statistical results usually quantify the likelihood that θ lies in some *ball* around $\widehat{\theta}_T$. However, the models residing within this ball may be *qualitatively* different. Again, leveraging recent results from the theory of large and moderate deviations [DE97, dH08, DZ09], we will be able to characterize the likelihood that the estimated system is qualitatively equivalent to the unknown true system.

Contribution

A high-level goal is to showcase how topological insights can benefit the control community. More specifically, we establish topological properties of the *reverse I-projection* $\mathcal{P}(\cdot)$ introduced above. We will characterize here the probability that the reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$ of the least-squares estimator $\widehat{\theta}_T$ is topologically different from θ. Formally, we show that

$$\mathbb{P}_\theta(\mathcal{P}(\widehat{\theta}_T) \not\simeq_t \theta) \lesssim e^{-\mathcal{O}(\sigma_{\min}(\theta)^2\sqrt{T})},$$

where '$\simeq_t$' denotes topological equivalence (see Section 2.3.1). Thus, the probability that $\mathcal{P}(\widehat{\theta}_T)$ misrepresents the topological properties of θ decays exponentially with T at a rate $\propto \sigma_{\min}(\theta)^2$.

Notation: We remind the reader that we denote the real n-dimensional general linear group by $\mathsf{GL}(n,\mathbb{R}) = \{A \in \mathbb{R}^{n\times n} : \det(A) \neq 0\}$. The sets $\mathsf{GL}^{+}(n,\mathbb{R})$ and $\mathsf{GL}^{-}(n,\mathbb{R})$ contain all matrices in $\mathsf{GL}(n,\mathbb{R})$ with a strictly positive or negative determinant.

3.2.2 Topological linear system identification

Given two linear maps $f(x) = Fx$ and $g(y) = Gy$ we recall from Theorem 2.3.3 that f and g are topologically equivalent only if the signs of the determinants of F and G match. By slight abuse of notation, we henceforth define the orientation $\mathrm{or}(F)$ of an invertible matrix F as the sign of $\det(F)$.

Now we also recall that the MDP framework giving rise to the reverse I-projection is rather flexible, that is, Definition 3.1.5 holds for *any* Borel set. In addition, the reverse I-projection preserves orientation, *i.e.*, $\mathrm{or}(\mathcal{P}(\theta')) = \mathrm{or}(\theta')$ for any $\theta' \in \mathsf{GL}(n,\mathbb{R})$, see, Corollary 3.1.17. In fact, the same is true for the numerical approximation, for any $\delta > 0$. Due to its desirable statistical and computational properties, our proposed approach to estimate the topological class of $\theta \in \Theta$ will critically rely on the reverse I-projection $\widehat{\theta}_T \mapsto \mathcal{P}(\widehat{\theta}_T)$.

To aid the presentation we assume from now on, without much loss of generality, that θ is invertible. Then, we are equipped to demonstrate that the MDP of Proposition 3.1.6 allows us via the reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$ to derive sharp bounds on the decay rate of the probability of the event $\mathcal{P}(\widehat{\theta}_T) \not\simeq_t \theta$. The approach is as follows. Recall again that the two stable and ($\mathbb{P}_\theta$-almost surely) invertible matrices $\mathcal{P}(\widehat{\theta}_T)$ and θ are topologically equivalent if and only if they have the same orientation. Recall also that $\mathrm{or}(\mathcal{P}(\widehat{\theta}_T)) = \mathrm{or}(\widehat{\theta}_T)$ because the reverse I-projection preserves orientation. Checking whether $\mathcal{P}(\widehat{\theta}_T)$ is topologically equivalent to θ is thus tantamount to checking whether the determinants of $\widehat{\theta}_T$ and θ have the same signs.

Theorem 3.2.1 (Probability of misclassification). *Assume that $\theta \in \Theta \cap \mathsf{GL}(n,\mathbb{R})$, $\{\widehat{\theta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ are the least squares estimators (3.1.2) and $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ is a sequence with $1 \ll a_T \ll T$. If $S_{\theta^\circ} = S_w^{-1/2} S_\theta S_w^{-1/2}$ and $r = \frac{1}{2}\lambda_{\min}(S_{\theta^\circ} - I_n)$, then*

$$\limsup_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\big(\mathrm{or}(\widehat{\theta}_T) \neq \mathrm{or}(\theta)\big) \leq -r, \tag{3.2.1a}$$

$$\limsup_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\big(\mathcal{P}(\widehat{\theta}_T) \not\simeq_t \theta\big) \leq -r. \tag{3.2.1b}$$

Proof. As for (3.2.1a), note that $\widehat{\theta}_T \in \mathsf{GL}(n,\mathbb{R})$ $\mathbb{P}_\theta$-almost surely for all sufficiently large T, and therefore we have

$$\begin{aligned}\mathbb{P}_\theta\big(\mathrm{or}(\widehat{\theta}_T) \neq \mathrm{or}(\theta)\big) &= \mathbb{P}_\theta\big(\mathrm{or}(\widehat{\theta}_T) = -\mathrm{or}(\theta)\big)\\ &= \mathbb{P}_\theta\big(\exists G \in \mathsf{GL}^{-}(n,\mathbb{R}) : \widehat{\theta}_T = G\theta\big)\\ &= \mathbb{P}_\theta\big(\widehat{\vartheta}_T \in \mathcal{D}_T(\theta)\big)\end{aligned}$$

where the second equality holds because any invertible matrices $\widehat{\theta}_T$ and θ whose determinants have opposite signs satisfy $\widehat{\theta}_T = G\theta$ for some $G \in \mathsf{GL}^{-}(n,\mathbb{R})$. The third equality follows from the definition of the transformed least squares estimators $\{\widehat{\vartheta}_T\}_{T\in\mathbb{Z}_{\geq 0}}$ in (3.1.9) and the construction of the set

$$\mathcal{D}_T(\theta) := \{\sqrt{(T/a_T)}(G-I_n)\theta + \theta : G \in \mathsf{GL}^-(n,\mathbb{R})\}.$$

As this set is time-*dependent*, we cannot directly use it. That is, it is not admissible in the sense of Definition 3.1.5. To sidestep this complication, we consider instead the larger set

$$\mathcal{D}(\theta) = \bigcup_{T\in\mathbb{Z}_{\geq 0}} \mathcal{D}_T(\theta).$$

The above reasoning then implies that

$$\begin{aligned}\limsup_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\big(\mathrm{or}(\widehat{\theta}_T) \neq \mathrm{or}(\theta)\big) &\leq \limsup_{T\to\infty} \frac{1}{a_T} \log \mathbb{P}_\theta\big(\widehat{\vartheta}_T \in \mathcal{D}(\theta)\big)\\ &\leq - \inf_{\theta'\in \mathrm{cl}\,\mathcal{D}(\theta)} I(\theta',\theta),\end{aligned}$$

where the first inquality holds because $\mathcal{D}_T(\theta) \subset \mathcal{D}(\theta)$ for all $T \in \mathbb{Z}_{\geq 0}$, while the second inequality follows from the MDP established in Proposition 3.1.6. In the remainder of the proof we derive an analytical lower bound on the minimization problem on the right hand side of the above expression. To this end, assume first that $S_w = I_n$. Evaluating the rate function (3.1.4) at an arbitrary $\theta' \in \mathcal{D}(\theta)$ then yields

$$I(\theta',\theta) = \frac{T}{2a_T}\mathrm{Tr}\left((G-I_n)\theta S_\theta \theta^\mathsf{T}(G-I_n)^\mathsf{T}\right) \tag{3.2.2}$$

for some $G \in \mathsf{GL}^-(n,\mathbb{R})$ and some $T \in \mathbb{Z}_{\geq 0}$, and the Lyapunov equation (3.1.7) implies that $\theta S_\theta \theta^\mathsf{T} = S_\theta - I_n$. In addition, our assumptions about the sequence $\{a_T\}_{T\in\mathbb{Z}_{\geq 0}}$ imply that $T/a_T \geq 1$. We may thus conclude that

$$\begin{aligned}&\min_{\theta'\in\mathrm{cl}\,\mathcal{D}(\theta)} I(\theta',\theta)\\ &\geq \inf_{\det(G)\leq 0} \tfrac{1}{2}\mathrm{Tr}\left((G-I_n)(S_\theta - I_n)(G-I_n)^\mathsf{T}\right)\\ &\geq \tfrac{1}{2}(\lambda_{\min}(S_\theta)-1) \inf_{\det(G)\leq 0} \|G-I_n\|_F^2 = \tfrac{1}{2}(\lambda_{\min}(S_\theta)-1),\end{aligned}$$

where the first inequality holds because $\det(G) \leq 0$ for every $G \in \mathrm{cl}\,\mathsf{GL}^-(n,\mathbb{R})$, the second inequality uses the bound $\mathrm{Tr}(AB) \geq \sigma_{\min}(A)\mathrm{Tr}(B)$ for any $A, B \succeq 0$, and the third inequality follows from the Eckart-Young Theorem [GvL13, Thm. 2.4.8].

This establishes (3.2.1a) for $S_w = I_n$. If $S_w \succ 0$ is arbitrary, one may first apply a change of coordinates $x^\circ = S_w^{-1/2}x$, under which the noise covariance matrix simplifies to I_n, while the system matrix and the invariant state covariance matrix become $\theta^\circ = S_w^{-1/2}\theta S_w^{1/2}$ and $S_{\theta^\circ} = S_w^{-1/2} S_\theta S_w^{-1/2}$, respectively. As $S_w \succ 0$, we further have $\theta^\circ \simeq_t \theta$. Applying the results of the first part of the proof to the transformed system finally yields (3.2.1a). As for (3.2.1b), recall from Sections 2.3.1 and 3.1.3 that $\mathcal{P}(\theta')$ is asymptotically stable for $\theta' \in \Theta' \setminus \partial\Theta$ and that $\mathrm{or}(\theta') = \mathrm{or}(\mathcal{P}(\theta'))$ for any $\theta' \in \mathsf{GL}(n,\mathbb{R})$. Therefore, we can focus on orientation, *i.e.,* if $\mathrm{or}(\theta') = \mathrm{or}(\theta)$ and $\theta' \in \Theta' \setminus \partial\Theta$, then $\mathcal{P}(\theta') \simeq_t \theta$. Then again, as $\mathbb{P}_\theta\big(\widehat{\theta}_T \notin \mathsf{GL}(n,\mathbb{R}) \cap \Theta\big) = 0$ for sufficiently large T, the claim follows from (3.2.1a). □

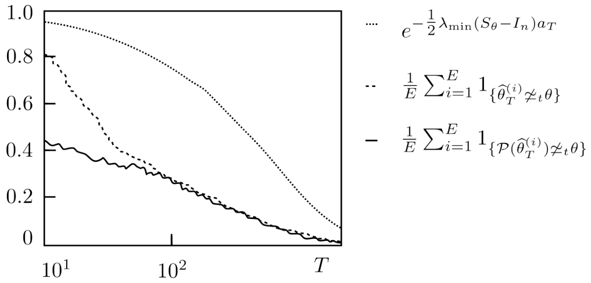


Figure 3.1: Example 3.2.1: empirical versus theoretical convergence rates. Adapted from the original Matlab figure [JSK22, Fig. 4].

We point out that the rate $r = \frac{1}{2}\lambda_{\min}(S_{\theta^\circ} - I_n)$ established in Theorem 3.2.1 is non-trivial (strictly positive) for any $\theta \in \Theta \cap \mathsf{GL}(n, \mathbb{R})$ as

$$S_{\theta^\circ} = S_w^{-1/2} S_\theta S_w^{-1/2} \overset{(3.1.7)}{=} S_w^{-1/2} \sum_{k=1}^{\infty} \theta^k S_w (\theta^k)^\mathsf{T} S_w^{-1/2} + I_n.$$

One can continue and further study this term, as is done in [JSK22, Sec. III A]. We only point out that since $S_{\theta^\circ} - I_n \succeq \theta^\circ(\theta^\circ)^\mathsf{T}$ it follows that $\frac{1}{2}\lambda_{\min}(S_{\theta^\circ} - I_n) \geq \frac{1}{2}\sigma_{\min}(\theta^\circ)^2$. Hence, an increase in $\sigma_{\min}(\theta^\circ)$ improves the rate r from Theorem 3.2.1. So as in Example 3.1.2, very fast dynamics are perhaps desired for the application, not for the identification. We end with a short numerical example.

Example 3.2.1 (Numerical topological identification)**.** We now compare the theoretical decay rate of topological misclassification derived in Theorem 3.2.1 against the empirical decay rate for the nominal least squares estimator $\widehat{\theta}_T$ and its reverse I-projection $\mathcal{P}(\widehat{\theta}_T)$. To this end, we set

$$\theta = \begin{pmatrix} Y & I_2 \\ 0 & Y \end{pmatrix}, \quad \text{for} \quad Y = \begin{pmatrix} -0.1 & 1 \\ 0.1 & 0.05 \end{pmatrix},$$

and simulate (3.1.1) under this θ, starting from $E = 10^3$ initial conditions $x_0 \overset{i.i.d.}{\sim} \mathcal{N}(0, I_4)$ with $S_w = I_4$. Each initial condition leads to a trajectory under θ from which we construct the corresponding least squares estimators $\widehat{\theta}_T^{(i)}$, $i = 1, \ldots, E$, $T = 0, \ldots, 10^3$ (of course, only for a sufficiently large T we have that $\widehat{\theta}_T$ is well-defined). Averaging over the E simulation runs yields the empirical probability that $\widehat{\theta}_T$ or its reverse I-projection are topologically equivalent to the true system matrix θ. Figure 3.1 compares the bounds on the misclassification probability derived in Theorem 3.2.1 for $a_T = T^{\frac{1}{1+\epsilon}}$ with $\epsilon = 10^{-9}$ against the empirical probabilities. Here, $\mathcal{P}(\widehat{\theta}_T^{(i)})$ is computed via (3.1.21) for $\delta = 10^{-9}$. As expected, the projection accelerates topological identification.

Closing this chapter, we saw the least squares estimators give rise to a "*natural*" discrepancy function that can be derived (and not just simply imposed), which can be exploited to construct probabilistic bounds with respect to qualitative system properties.

Bibliography

[AM06] P J Antsaklis and A N Michel. *Linear Systems.* Birkhäuser, Basel, 2006.

[ÅW73] K J Åström and B Wittenmark. On self tuning regulators. *Automatica*, 9(2):185–199, 1973.

[BB95] T Başar and P Bernhard. *H_∞-Optimal Control and Related Minimax Design Problems A Dynamic Game Approach.* Birkhäuser, Basel, 1995.

[Ber05] D P Bertsekas. *Dynamic Programming and Optimal Control (I).* Athena Scientific, Nashua, 2005.

[Ber07] D P Bertsekas. *Dynamic Programming and Optimal Control (II).* Athena Scientific, Nashua, 2007.

[Ber19] D P Bertsekas. *Reinforcement Learning and Optimal Control.* Athena Scientific, Nashua, 2019.

[BF11] P Benner and H Fassbender. On the numerical solution of large-scale sparse discrete-time Riccati equations. *Adv. Comput. Math.*, 35:119–147, 11 2011.

[BGS08] B Boots, G J Gordon, and S M Siddiqi. A constraint generation approach to learning stable linear dynamical systems. In *Proc. Neural Information Processing Systems*, pages 1329–1336. 2008.

[BT96] D P Bertsekas and J Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Nashua, 1996.

[BTM$^+$21] N Boffi, S Tu, N Matni, J.-J Slotine, and V Sindhwani. Learning stability certificates from data. In *Prof. Conference on Robot Learning*, volume 155, pages 1341–1350, 2021.

[BTSK17] F Berkenkamp, M Turchetta, A P Schoellig, and A Krause. Safe model-based reinforcement learning with stability guarantees. In *Proc. Neural Information Processing Systems*, pages 908–919, 2017.

[CGS20] N Choudhary, N Gillis, and P Sharma. On approximating the nearest Ω-stable matrix. *Numer. Linear Algebra Appl.*, 27(3):e2282, 2020.

[CK98] M C Campi and P R Kumar. Adaptive linear quadratic Gaussian control: The cost-biased approach revisited. *SIAM J. Control Optim.*, 36(6):1890–1907, 1998.

[CM03] I Csiszar and F Matus. Information projections revisited. *IEEE T. Inf. Theory*, 49(6):1474–1490, 2003.

[CS04] I Csiszar and P C Shields. Information theory and statistics: A tutorial. *Found. Trends Commun. Inf. Theory*, (4):417–528, 2004.

[Csi84] I Csiszar. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.

[DE97] P Dupuis and R S Ellis. *A weak convergence approach to the theory of large deviations.* John Wiley & Sons, Inc., New York, 1997.

[dH08] F den Hollander. *Large Deviations.* American Mathematical Society, Providence, 2008.

[DNZH$^+$23] L Di Natale, M Zakwan, P Heer, G Ferrari Trecate, and C N Jones. SIMBa: System identification methods leveraging backpropagation. *arXiv e-print:2311.13889*, 2023.

[DZ09] A Dembo and O Zeitouni. *Large Deviations Techniques and Applications.* Springer-Verlag, Berlin, 2009.

[GKS19] N Gillis, M Karow, and P Sharma. Approximating the nearest stable discrete-time system. *Linear Algebra Appl.*, 573:37–53, 2019.

[GL91] J D Gardiner and A J Laub. Parallel algorithms for algebraic Riccati equations. *Int. J. Control*, 54(6):1317–1333, 1991.

[GL17] N Guglielmi and C Lubich. Matrix stabilization using differential equations. *SIAM J. Numer. Anal.*, 55(6):3097–3119, 2017.

[GvL13] G H Golub and C F van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 2013.

[Hig89] N J Higham. Matrix nearness problems and applications. In *Applications of Matrix Theory*, pages 1–27, Oxford, 1989. Oxford University Press.

[Jon19] W Jongeneel. Controlling the unknown: A game theoretic perspective. Master's thesis, Systems & Control, Delft University of Technology, 2019.

[Jon22] W Jongeneel. Stability via reverse I-projections, a symplectic perspective on the computation. 2022.

[JP19] Y Jedra and A Proutiere. Sample complexity lower bounds for linear system identification. In *Proc. IEEE Conference on Decision and Control*, pages 2676–2681, 2019.

[JP20] Y Jedra and A Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *Proc. IEEE Conference on Decision and Control*, pages 996–1001, 2020.

[JSK22] W Jongeneel, T Sutter, and D Kuhn. Topological linear system identification via moderate deviations theory. *IEEE Control Syst. Lett.*, 6:307–312, 2022.

[JSK23] W Jongeneel, T Sutter, and D Kuhn. Efficient learning of a linear dynamical system with stability guarantees. *IEEE T. Automat. Contr.*, 68(5):2790–2804, 2023.

[JSM19] W Jongeneel, T Summers, and P Mohajerin Esfahani. Robust linear quadratic regulator: Exact tractable reformulation. In *Proc. IEEE Conference on Decision and Control*, pages 6742–6747, 2019.

[KM19] J Z Kolter and G Manek. Learning stable deep dynamics models. In *Proc. Neural Information Processing Systems*, pages 11128–11136. 2019.

[KMN89] D Kahaner, C Moler, and S Nash. *Numerical methods and software.* Prentice-Hall, Inc., Hoboken, 1989.

[KTR19] K Krauth, S Tu, and B Recht. Finite-time analysis of approximate policy iteration for the Linear Quadratic Regulator. In *Proc. Neural Information Processing Systems*, pages 8514–8524. 2019.

[Kui94] B Kuipers. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge.* MIT Press, Cambridge, 1994.

[KV86] P R Kumar and P Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control.* Prentice-Hall, Hoboken, 1986.

[LB02] S L Lacy and D S Bernstein. Subspace identification with guaranteed stability using constrained optimization. In *Proc. American Control Conference*, pages 3307–3312, 2002.

[LB03] S L Lacy and D S Bernstein. Subspace identification with guaranteed stability using constrained optimization. *IEEE T. Automat. Contr.*, 48(7):1259–1263, 2003.

[LR95] P Lancaster and L Rodman. *Algebraic Riccati Equations.* Oxford University Press, Oxford, 1995.

[LS20] T Lattimore and C Szepesvári. *Bandit Algorithms.* Cambridge University Press, Cambridge, 2020.

[Mac95] J M Maciejowski. Guaranteed stability with subspace methods. *Syst. Control Lett.*, 26(2):153 – 156, 1995.

[MAM23] Giorgos Mamakoukas, Ian Abraham, and Todd D Murphey. Learning stable models for prediction and control. *IEEE Trans. Robot.*, 2023.

[MB14] S Mohammad Khansari-Zadeh and A Billard. Learning control Lyapunov function to ensure stability of dynamical system-based robot reaching motions. *Rob. Auton. Syst.*, 62(6):752 – 765, 2014.

[MPRT19] N Matni, A Proutiere, A Rantzer, and S Tu. From self-tuning regulators to reinforcement learning and back again. In *Proc. IEEE Conference on Decision and Control*, pages 3724–3740, 2019.

[MT09] S Meyn and R L Tweedie. *Markov Chains and Stochastic Stability.* Cambridge University Press, Cambridge, 2009.

[MXM20] G Mamakoukas, O Xherija, and T Murphey. Memory-efficient learning of stable linear dynamical systems for prediction and control. In *Proc. Neural Information Processing Systems*, pages 13527–13538. 2020.

[NP20] Y Nesterov and V Y Protasov. Computing closest stable nonnegative matrix. *SIAM J. Matrix Anal. Appl.*, 41(1):1–28, 2020.

[ONv13] F.-X Orbandexivry, Y Nesterov, and P van Dooren. Nearest stable system using successive convex approximations. *Automatica*, 49(5):1195 – 1203, 2013.

[PLS80] T Pappas, A Laub, and N Sandell. On the numerical solution of the discrete-time algebraic Riccati equation. *IEEE T. Automat. Contr.*, 25(4):631–641, 1980.

[Pol86] J W Polderman. A note on the structure of two subsets of the parameter space in adaptive control problems. *Syst. Control Lett.*, 7(1):25–34, 1986.

[Pow07] W B Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* John Wiley & Sons, Hoboken, 2007.

[Rec19] B Recht. A tour of reinforcement learning: The view from continuous control. *Annu. Rev. Control Robot. Auton. Syst.*, 2:253–279, 2019.

[SB18] R S Sutton and A G Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, 2018.

[SMT$^+$18] M Simchowitz, H Mania, S Tu, M I Jordan, and B Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proc. Conference On Learning Theory*, pages 439–473, 2018.

[SR19] T Sarkar and A Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *Proc. International Conference on Machine Learning*, pages 5610–5618, 2019.

[SRD21] T Sarkar, A Rakhlin, and M A Dahleh. Finite time LTI system identification. *J. Mach. Learn. Res.*, 22(1):1186–1246, 2021.

[Sun96] R K Sundaram. *A First Course in Optimization Theory.* Cambridge University Press, Cambridge, 1996.

[TBQ$^+$13] K Turksoy, E S Bayrak, L Quinn, E Littlejohn, and A Cinar. Guaranteed stability of recursive multi-input-single-output time series models. In *Proc. American Control Conference*, pages 77–82, 2013.

[TK05] H Tanaka and T Katayama. Stochastic subspace identification guaranteeing stability and minimum phase. *Proc. IFAC World Congress*, 38(1):910 – 915, 2005.

[UH20] J Umlauft and S Hirche. Learning stochastically stable Gaussian process state–space models. *IFAC J. Syst. Control*, 12:100079, 2020.

[UWMS18] J Umenberger, J Wågberg, I R Manchester, and T B Schön. Maximum likelihood identification of stable linear dynamical systems. *Automatica*, 96:280 – 292, 2018.

[Ver18] R Vershynin. *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press, Cambridge, 2018.

[VODM96] P Van Overschee and B De Moor. *Subspace Identification for Linear Systems: Theory-Implementation-Applications.* Kluwer Academic Publishers, Dordrecht, 1996.

[vSvd00] T van Gestel, J A K Suykens, P van Dooren, and B de Moor. Imposing stability in subspace identification by regularization. In *Proc. IEEE Conference on Decision and Control*, volume 2, pages 1555–1560, 2000.

[vSvd01] T van Gestel, J A K Suykens, P van Dooren, and B. de Moor. Identification of stable models in subspace identification by using regularization. *IEEE T. Automat. Contr.*, 46(9):1416–1420, 2001.

[VV07] M Verhaegen and V Verdult. *Filtering and System Identification.* Cambridge University Press, Cambridge, 2007.

[VWETC20] H J Van Waarde, J Eising, H L Trentelman, and M K Camlibel. Data informativity: a new perspective on data-driven analysis and control. *IEEE T. Automat. Contr.*, 65(11):4753–4768, 2020.

[Wor99] J Worms. Moderate deviations for stable Markov chains and regression models. *Electron. J. Probab.*, 4:28 pp., 1999.

[YS09] M Yu and S Si. Moderate deviation principle for autoregressive processes. *J. Multivar. Anal.*, 100(9):1952–1961, 2009.

4
Space of stable systems and structural stability

> “*A well-known early example of the occurance of convexity, though originally in a more implicite role, is Euler’s theorem on polyhedra. … the theorem … aims at a classification of the polyhedra. … There appeared many investigations about this topic, in particular in the second half of the last century. In most of them only convex polyhedra are considered, presumably often because the problems otherwise seemed insuperable.*”
>
> —Fenchel [Fen83, p. 121].

Various domains are aided by better understanding spaces of stable dynamical systems. For instance, in many identification- and optimal control problems we effectively try to optimize over such a space and thus understanding its properties is important.

First, we study the linear quadratic regulator problem and finally provide an operational meaning for cross terms in the stage cost. In particular, we show that the topological class of the closed-loop system is invariant under a change of the stage cost, as long as no cross term is introduced, that is, when restricting system matrices to the general linear group, closed-loop matrices can jump to the opposite path-connected component of this group if and only if a cross term is introduced. Hence, formally speaking, one can only “tune” the closed-loop behaviour by introducing a cross term.

Secondly, motivated by the learning community often employing convex Lyapunov functions to obtain stability certificates, we study the ramifications of the convexity assumption. We show that continuous dynamical systems, on Euclidean space, equipped with convex Lyapunov functions, asserting that the origin is globally asymptotically stable, can always be homotoped to each other such that along this homotopy stability is preserved. This means that the space of those dynamical systems is path-connected, which in its turn leads to obstructions and the necessity of rethinking convexity assumptions.

4.1 On topological equivalence in linear quadratic optimal control

In this section we will show that in general, closed-loop systems resulting from discrete-time linear quadratic optimal control problems are all topologically equivalent. As such, we provide new insights in structural "*tuning*" of controlled behaviour.

4.1.1 Introduction

Ever since its inception and celebration, the theory of optimal control also received critique. The cost is commonly scalar-valued, making the optimal control selection solely dependent on a single performance criteria, which limits practicality [Zad63]. To improve our understanding of optimality in the context of control, Kalman set out to understand the *inverse* problem [Kal64], that is, given a control policy, does there exist an optimal control problem giving rise to this policy? To quote his motivation "*...discover general properties shared by all optimal control laws. We might be able to separate control laws which are optimal in some sense from those which are not optimal in any sense.*". In this section we add to this investigation, with an emphasis on the *controlled* behaviour.

In particular, we will consider the discrete-time Linear Quadratic Regulation (LQR) problem. This is a classical setting which made its appearance in many real-world systems. There, one usually encounters the notion of "*tuning*" the cost function such that the system is "*sufficiently*" stable. Success-stories of this tuning can be found throughout, with the catch being that there, linear feedback is designed for a *locally*-linear system, where tuning might be needed indeed. This section shows that, for the better or worse, if one *does* have a linear system how to change the closed-loop system behaviour *structurally.*

To classify the behaviour of a controlled dynamical system we take again a topological approach, see Section 2.3.1. As in the original work by Kuiper and Robbin, we focus on linear dynamical systems, but this time, in line with Kalman, driven by some optimal Linear Quadratic (LQ) regulator, or any other policy originating from the family of LQ optimal control problems. LQ theory is well-understood, especially in the context of classical engineering [AM89] and currently in the context of statistical reinforcement learning [DMM$^+$20] and optimization [FGKM18]. In the context of adaptive control, several interesting topological results, with respect to the underlying model, are made in [Pol87, vS94, CL19]. Topological insights in the *resulting* closed-loop systems are less known, or at least, not described as much in the modern literature. We try to fill in this gap and provide a new interpretation of how one can change the dynamical behaviour, *structurally*, via selecting appropriate cost-matrices.

Contributions

The main result of this section is to show that a well-known class of optimization problems have structurally equivalent minimizers (closed-loop systems), *i.e.*, in optimization parlance, without defining the class $\mathcal{F}$ and *equivalence relation* $\sim$, we have

$$\arg\min_{x\in\mathcal{X}} f_1(x) \sim \arg\min_{x\in\mathcal{X}} f_2(x) \quad \forall f_1, f_2 \in \mathcal{F}.$$

Specifically, we highlight that the most common class of LQ Optimal Control (OC) problems result in topologically equivalent closed-loop behaviour[1]. In particular, we show that by means of tuning the cost matrices, a bifurcation in the controlled system can only be induced by the introduction of cross-terms. Concurrently, building on [Pol87], we see that a lot of structure of the underlying system is LQ feedback invariant. These observations have some implications, for example to reduce the dimension of the optimal control problem, to give a new interpretation of cross-terms in the cost or to preserve structure from a corresponding continuous-time problem. Although the arguments are simple, to the best of our knowledge, this was not observed before.

Notation: To remind the reader, we denote the real n-dimensional *General Linear group* by $\mathsf{GL}(n,\mathbb{R}) := \{A \in \mathbb{R}^{n\times n} : \det(A) \neq 0\}$. The group $\mathsf{GL}(n,\mathbb{R})$ can be written as $\mathsf{GL}^-(n,\mathbb{R}) \cup \mathsf{GL}^+(n,\mathbb{R})$, which is the disjoint union of two path-connected sets. Here, the superscript denotes the sign of the determinant, *e.g.,* $T \in \mathsf{GL}^+(n,\mathbb{R}) \iff \det(T) > 0$. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be *asymptotically stable* (Schur) when $\rho(A) := \max_i |\lambda_i(A)| < 1$.

4.1.2 Linear quadratic optimal control

In this subsection we introduce the control problem at hand. Consider the deterministic linear discrete-time system

$$x \mapsto Ax + Bu =: \sigma(x,u), \quad x \in \mathbb{R}^n \tag{4.1.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ comprise a stabilizable pair, that is, there exists a $K \in \mathbb{R}^{m\times n}$ such that $\rho(A+BK) < 1$. We will write this as $\sigma \in \Sigma$, for Σ parametrized by the set of stabilizable pairs (A,B). Then, for some triple $(Q,R,S) \in \mathcal{S}^n_{\succeq 0} \times \mathcal{S}^m_{\succ 0} \times \mathbb{R}^{n\times m}$ define the corresponding stage-cost $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by

$$(x,u) \mapsto c(x,u) := \begin{pmatrix} x \\ u \end{pmatrix}^\mathsf{T} \begin{pmatrix} Q & S \\ S^\mathsf{T} & R \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix}. \tag{4.1.2}$$

When $S = 0$, we refer to the (stage-)cost as being *block-diagonal.* The block-diagonal form is well-understood and dominates the practical- and theoretical literature. We try to understand why. Now, following [LR95, Ch. 16], we can easily bring (4.1.2) to such a block-diagonal form. Specifically, by defining $v := R^{-1}S^\mathsf{T}x + u$, $Q' := Q - SR^{-1}S^\mathsf{T}$ and $A' := A - BR^{-1}S^\mathsf{T}$, we can, equivalently to (4.1.2) under (4.1.1), consider the stage-cost

[1] Preliminary arguments appeared in [Jon19].

$c'(x,v) := x^\mathsf{T} Q' x + v^\mathsf{T} R v$, under the time-one map $x \mapsto A'x + Bv$. This transformation allows for applying all celebrated block-diagonal tools.

In what follows we will assume (without loss of generality), that the input u is linear state-feedback, that is, for some $K \in \mathbb{R}^{m\times n}$, $u := Kx$. Then, fixing some $\sigma \in \Sigma$, we define the cost function $J : \mathbb{R}^n \times \mathbb{R}^{m\times n} \to \mathbb{R} \cup \{\pm\infty\}$ by

$$\begin{aligned} J(x', K) :=\ & \textstyle\sum_{k=0}^{\infty} c(x_k, Kx_k), \\ \text{subject to}\ & x_{k+1} = \sigma(x_k, Kx_k)\ \forall k \geq 0,\ x_0 = x'. \end{aligned} \tag{4.1.3}$$

To optimize this cost over K and have a meaningful solution, assume the stage-cost is non-negative and that (A, C) is a detectable pair for $C \in \mathbb{R}^{p\times n}$ defined by $C^\mathsf{T} C := Q$ with $\operatorname{rank}(Q) = \operatorname{rank}(Q')$[2]. It is known that under the aforementioned conditions (*cf.* [LR95, Ch. 13-16]), $\arg\min_K J(x', K)$ is given by $K^\star = -(R + B^\mathsf{T} P B)^{-1} B^\mathsf{T} P A' - R^{-1} S^\mathsf{T}$, where $P \in \mathcal{S}^n_{\succeq 0}$ is the unique solution to the *Algebraic Riccati Equation* (ARE)

$$P = Q' + A'^\mathsf{T} P A' - A'^\mathsf{T} P B (R + B^\mathsf{T} P B)^{-1} B^\mathsf{T} P A', \tag{4.1.4}$$

such that the *optimal closed-loop* time-one map $x \mapsto \sigma(x, K^\star x) =: \sigma^\star(x)$ is asymptotically stable. With respect to (4.1.3), the aforementioned domain of (Q, R, S) and the definition of Q', we define the set of cost-matrices $\mathcal{C}(\sigma)$ by

$$\mathcal{C}(\sigma) := \left\{ (Q, R, S) : \begin{array}{l} Q - SR^{-1}S^\mathsf{T} \succeq 0, \\ \operatorname{rank}(Q) = \operatorname{rank}(Q'), \\ \exists\, C \in \mathbb{R}^{p\times n} : C^\mathsf{T} C = Q, \\ (A, C) \text{ detectable} \end{array} \right\}. \tag{4.1.5}$$

Linear quadratic dynamic games

One of the insights of this section is that the qualitative behaviour of the whole *family* of block-diagonal Linear Quadratic Optimal Control problems is the same. To exemplify what we mean by this "*family*", we introduce a different, but analogous, block-diagonal cost function to (4.1.3). We introduce what is called a *two-player zero-sum dynamic game*, *e.g.,* see [BB95]. There, given some $\delta \in \mathbb{R}_{\geq 0}$, the stage-cost is defined by the function $g : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$

$$g(x, u, w) := x^\mathsf{T} Q x + u^\mathsf{T} R u - \delta^{-1} w^\mathsf{T} w. \tag{4.1.6}$$

The variable w will act as an *adversary*. Given some $D \in \mathbb{R}^{n\times d}$, let $\sigma_w(x, u) := \sigma(x, u) + Dw$ and again, without loss of generality, assume w to be linear in x, that is $w := Lx$ for some $L \in \mathbb{R}^{d\times n}$. such that we can define the cost function $J : \mathbb{R}^n \times \mathbb{R}^{m\times n} \times \mathbb{R}^{d\times n} \to \mathbb{R} \cup \{\pm\infty\}$ by

$$\begin{aligned} J(x', K, L) :=\ & \textstyle\sum_{k=0}^{\infty} g(x_k, Kx_k, Lx_k), \\ \text{subject to}\ & x_{k+1} = \sigma_{Lx_k}(x_k, Kx_k)\ \forall k \geq 0,\ x_0 = x'. \end{aligned} \tag{4.1.7}$$

[2] See [LR95, Sec. 16.2] for the intuition regarding this condition. It is essentially there to assure we can carry out standard LQR practices while avoiding degenerate examples. Under this condition we can be concerned with detectability, simply, with respect to Q and not Q'.

Under conditions analogous to the ones from before (see [BB95, Ch. 3]), a solution to $\min_K \max_L J(x', K, L)$ exists and is given by the static gains $K^\star(\delta) := -R^{-1}B^\mathsf{T}P(\delta)\Lambda(\delta)^{-1}A$ and $L^\star(\delta) := \delta D^\mathsf{T}P(\delta)\Lambda(\delta)^{-1}A$. Here, the pair $(P(\delta), \Lambda(\delta))$ compromises a solution to the *Generalized Algebraic Riccati Equation* (GARE):

$$\begin{aligned} P &= Q + A^\mathsf{T}P\Lambda^{-1}A, \\ \Lambda &= \left(I_n + \left(BR^{-1}B^\mathsf{T} - \delta DD^\mathsf{T}\right)P\right). \end{aligned} \tag{4.1.8}$$

The parameter δ relates to how much adversarial action we allow for. Crossing what is called the "*breakdown-point*" $\bar{\delta}$ means it is "*affordable*" for the adversary to destabilize the system and hence this scenario is avoided by selecting $\delta \in [0, \bar{\delta})$ (see [BB95, Whi90]). Moreover, it can be shown that the closed-loop system matrix is asymptotically stable and can be written as $\Lambda(\delta)^{-1}A$. This observation is the key in showing topological equivalence in the next section.

Summarizing, we parametrize a Linear Quadratic Optimal Control problem by the pair (σ, J), where one seeks a sequence of inputs to the linear dynamical system σ such that the quadratic cost function J, subject to σ, is minimized. Then, we will be interested in understanding to which topological class the corresponding optimal closed-loop systems $x \mapsto \sigma^\star(x)$ belong.

4.1.3 Topological equivalence in linear quadratic optimal control

The main result of this section can be stated as follows: given any two closed-loop maps f, g resulting from LQ optimal control problems with block-diagonal cost, then f and g are topologically equivalent. This means that the *qualitative* behaviour of these controlled systems is invariant under "*tuning*" of the cost. To proceed, we introduce a *orientation-dependent* version of the set in (4.1.5). Given a $\sigma \in \Sigma$ such that $A \in \mathsf{GL}(n, \mathbb{R})$ then define $\mathcal{C}^{(i)}(\sigma)$, $(i) \in \{+, -\}$ by

$$\mathcal{C}^{(i)}(\sigma) := \left\{ (Q, R, S) \in \mathcal{C}(\sigma) : \begin{matrix} A \in \mathsf{GL}^{(i)}(n, \mathbb{R}), \\ A' \in \mathsf{GL}^{(i)}(n, \mathbb{R}) \end{matrix} \right\}. \tag{4.1.9}$$

Here, $\mathsf{GL}^{(i)}(n, \mathbb{R})$ relates to either $\mathsf{GL}^{+}(n, \mathbb{R})$ or $\mathsf{GL}^{-}(n, \mathbb{R})$. Now we can state the main result.

Theorem 4.1.1 (Topological equivalence in LQ regulation)**.** *Fix some $\sigma \in \Sigma$ with $A \in \mathsf{GL}^{(i)}(n, \mathbb{R})$, $(i) \in \{+, -\}$. Let $x \mapsto \sigma_1^\star(x)$ be the optimal LQ regulated closed-loop time-one map corresponding to an arbitrary triple $(Q_1, R_1, S_1) \in \mathcal{C}^{(i)}(\sigma)$, that is $x \mapsto (A+BK_1^\star)x = \sigma_1^\star(x)$ with $K_1^\star$ the minimizing argument in (4.1.3). Analogously, define $\sigma_2^\star$ for some arbitrary triple $(Q_2, R_2, S_2) \in \mathcal{C}^{(i)}(\sigma)$. Then, $\sigma_1^\star \simeq_t \sigma_2^\star$.*

Proof. Since the cost-matrices are elements of $\mathcal{C}^{(i)}(\sigma)$ we use the transformations as set forth in Section 4.1.2. Then, it is known (see [LR95, Ch. 12]) that the closed-loop system matrices can be written as $(I_n + BR_j^{-1}B^\mathsf{T}P_j)^{-1}A'_j =: \Lambda_j^{-1}A'_j$ for $P_j \in \mathcal{S}^n_{\succeq 0}$ the solution to the algebraic Riccati equation (4.1.4) under (Q_j, R_j, S_j) $j \in \{1, 2\}$. Now we claim that $\Lambda_j \in \mathsf{GL}^+(n,\mathbb{R})$. Let $N := BR^{-1}B^\mathsf{T}$ and $M := P_j$, which are both symmetric positive semidefinite. Then observe that given some eigenpair (λ, v) such that $NMv = \lambda v$[3], multiplying from the left with $v^\mathsf{T}M^\mathsf{T}$ implies that $v^\mathsf{T}Mv\lambda = v^\mathsf{T}M^\mathsf{T}NMv \geq 0$ and hence, by construction of M and N, that $\lambda \geq 0$. Therefore, the eigenvalues of Λ_j are all strictly positive such that $\det(\Lambda_j) > 0$. Since an application of such a Λ_j does not alter the membership of A' to $\mathsf{GL}^{(i)}(n,\mathbb{R})$, *i.e.*, $\mathsf{GL}^+\mathsf{GL}^{(i)} = \mathsf{GL}^{(i)}$, we can directly appeal to Theorem 2.3.3 and conclude the proof. □

We see from Theorem 4.1.1 that if $(Q,R,0) \in \mathcal{C}(\sigma)$, then, since $A = A'$ we have $(Q,R,0) \in \mathcal{C}^{(i)}(\sigma)$ and indeed we see that *all* block-diagonal problems result in closed-loop maps being topologically equivalent. In particular, if $\rho(A) < 1$, then $Ax \simeq_t (A + BK^\star)x$. Moreover, when $A \in \mathsf{GL}^+(n,\mathbb{R})$, then, block-diagonal LQ feedback leaves the group-structure intact, *i.e.,* $(A+BK^\star) \in \mathsf{GL}^+(n,\mathbb{R})$. Moreover, if A is singular, the gist of Theorem 4.1.1 remains true, yet, we need to restrict our discussion to the automorphic part of $\sigma^\star$, which is remarkably simple since $\ker(A)$ is preserved under block-diagonal LQ feedback (since the closed-loop matrix is of the form $\Lambda^{-1}A$ or see [Pol86a, Lem. 3.4]). When we introduce a non-zero S, however, the kernel of A and the optimally LQ controlled closed-loop system matrix $\Lambda^{-1}A'$ do not necessarily match anymore since $\ker(A)$ and $\ker(A - BR^{-1}S^\mathsf{T})$ can be different.

The form of Theorem 4.1.1 is chosen since it captures the central message: without constructing explicit LQR solutions one can easily assess *a priori* if some closed-loop maps will be topologically equivalent. Next, we construct another indirect characterization of these distinct topological classes.

When $A' \in \mathsf{GL}(n,\mathbb{R})$, then, a minimizing solution to the standard LQR cost (4.1.3) can be characterized via a Symplectic matrix. In particular, define $\Omega \in \mathbb{R}^{2n\times 2n}$ by

$$\Omega := \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}.$$

Then, define the real *Symplectic group* by $\mathsf{Sp}(2n,\mathbb{R}) := \{M \in \mathbb{R}^{2n\times 2n} : M^\mathsf{T}\Omega M = \Omega\}$. Moreover, we speak of a subspace $\mathcal{Y}$ being M-invariant, when $M\mathcal{Y} \subseteq \mathcal{Y}$. Next, define $M \in \mathsf{Sp}(2n,\mathbb{R})$ by

$$M := \begin{pmatrix} A' + BR^{-1}B^\mathsf{T}A'^{-\mathsf{T}}Q' & -BR^{-1}B^\mathsf{T}A'^{-\mathsf{T}} \\ -A'^{-\mathsf{T}}Q' & A'^{-\mathsf{T}} \end{pmatrix}. \tag{4.1.10}$$

A celebrated result—as for example communicated for an even more general setting in [PLS80]—is that eigenspaces of M in (4.1.10) directly map to solutions of (4.1.4) and in fact, the spectrum of M relates directly to the spectrum of the optimal LQ regulated time-one map. Better yet, the relation between M and $\Lambda^{-1}A'$ is well-understood. Now, assume that the triple (Q,R,S) is parametrized by some scalar $\gamma \in [0,1]$, that is, let $A'(\gamma) := A - BR(\gamma)^{-1}S(\gamma)^\mathsf{T}$, $Q'(\gamma) := Q(\gamma) - S(\gamma)R(\gamma)^{-1}S(\gamma)^\mathsf{T}$ and define $M(\gamma) \in \mathsf{Sp}(2n,\mathbb{R})$ accordingly. Then, loosely speaking, it turns out that when $M(\gamma)$

[3] We point out that, in general, one cannot assume that NM would have real eigenvalues. However, since both matrices are symmetric positive definite, this does immediately follow. The intuition being that, under the assumption that $M \in \mathsf{GL}(n,\mathbb{R})$ we have that NM is similar to $M^{1/2}NM^{1/2}$. Then the claim follows from $\mathsf{GL}(n,\mathbb{R})$ being dense in $\mathbb{R}^{n\times n}$.

is a continuous curve in $\mathsf{Sp}(2n,\mathbb{R})$, then, all closed-loop maps it parametrizes (see Section 4.1.2) are topologically equivalent. We formalize this in a Corollary to Theorem 4.1.1:

Corollary 4.1.2 (Topological equivalence via the Symplectic Group)**.** *Fix some $\sigma \in \Sigma$ and let $\gamma \in [0,1]$ parametrize a curve $(Q,R,S)(\gamma) \subset \mathcal{C}(\sigma)$ such that both $(Q,R,S)(0)$ and $(Q,R,S)(1)$ correspond to feasible LQR problems with optimal closed-loop maps $\sigma^\star(x)(0)$ and $\sigma^\star(x)(1)$. Then, $\sigma^\star(0) \simeq_t \sigma^\star(1)$ if there exists a continuous path $[0,1] \mapsto M[0,1] \subset \mathsf{Sp}(2n,\mathbb{R})$ from $M(0)$ to $M(1)$.*

Proof. Since $\mathsf{Sp}(2n,\mathbb{R}) \subset \mathsf{SL}(2n,\mathbb{R})$ and we can continuously deform $M(0)$ into $M(1)$ this must mean we do not drop rank along the path $M(\gamma)$, $\gamma \in [0,1]$. Moreover, we know that for any $M \in \mathsf{Sp}(2n,\mathbb{R})$ $\mu \in \lambda(M) \Rightarrow 1/\mu \in \lambda(M)$. Also, for any feasible LQR problem leading to (4.1.10) it is known that $1 \not\in |\lambda(M)|$, hence, when one constructs the Jordan normal form related to such a M, there are n-dimensional M-invariant *stable* (s) and *unstable* (u) subspaces, that is $M = XJX^{-1}$ with $X = [X^s \ X^u]$, $J = \mathrm{diag}(J^s, J^u)$. In fact, it can be shown that $\lambda(J^s) = \lambda(\Lambda^{-1}A')$, where $\Lambda^{-1}A'$ is the optimal LQ regulated closed-loop system matrix from Section 4.1.2. See [PLS80] and references therein for more on the aforementioned results. The prior discussion implies that if $0 \not\in \lambda(M(\gamma))\ \forall \gamma \in [0,1]$, then, $0 \not\in \lambda(J^s(\gamma))$ and as such $0 \not\in \lambda(\Lambda(\gamma)^{-1}A'(\gamma))$. This however means that $A'(\gamma) \in \mathsf{GL}^{(i)}(n,\mathbb{R})\ \forall \gamma \in [0,1]$, thereby, the result follows after an application of Theorem 2.3.3. □

So, although $\mathsf{Sp}(2n,\mathbb{R})$ is connected, by varying the triple $(Q,R,S) \in \mathcal{C}(\sigma)$ we effectively generate disjoint connected sets of matrices $M \in \mathsf{Sp}(2n,\mathbb{R})$ through (4.1.10) corresponding to distinct topological classes of closed-loop maps they generate. Since this result provides the link with the far more general *Maximum Principle*, the hope is that similar constructions are possible for different Hamiltonians.

Now, we are not just interested in the OC problem related to (4.1.3), but into the whole "*family of LQ problems*". To that end, we give one example and use Section 4.1.2 to extend Theorem 4.1.1.

Corollary 4.1.3 (Topological equivalence in dynamic games)**.** *Fix $\sigma \in \Sigma$, let $A \in \mathsf{GL}^{(i)}(n,\mathbb{R})$ and set $D := I_n$ in (4.1.7)-(4.1.8). Moreover, consider J as in (4.1.7) for some pair (Q,R) being such that for any $\delta \in [0,\bar{\delta})$ the extremizers in $\min_{K\in\mathbb{R}^{m\times n}} \max_{L\in\mathbb{R}^{d\times n}} J(x',K,L)$ denoted by $K^\star(\delta)$ and $L^\star(\delta)$, exist. Then, the "nominal"-, "robust"- and "worst-case robust" optimal closed-loop maps given by $f(x) := \big(A + BK^\star(\delta)\big)x|_{\delta=0}$, $g(x) := \big(A + BK^\star(\delta)\big)x|_{\delta\in(0,\bar{\delta})}$, $h(x) := \big(A + BK^\star(\delta) + L^\star(\delta)\big)x|_{\delta\in(0,\bar{\delta})}$, respectively, are topologically equivalent.*

Proof. Let the pair $(P(\delta),\Lambda(\delta))$ correspond to a solution to (4.1.8). Recall, for example, from [BB95, Ch. 3] that $f(x) = \Lambda^{-1}(\delta)Ax|_{\delta=0}$, $g(x) = \big(I_n - \delta P(\delta)\big)\Lambda^{-1}(\delta)Ax|_{\delta\in(0,\bar{\delta})}$ with $(\delta^{-1}I_n - P(\delta)) \succ 0$ and $h(x) = \Lambda^{-1}(\delta)Ax|_{\delta\in(0,\bar{\delta})}$. Then, all that we need to show is that $\Lambda(\delta)|_{(0,\bar{\delta})} \in \mathsf{GL}^+(n,\mathbb{R})$. It follows from Theorem 4.1.1 that $\lim_{\delta\downarrow 0}\Lambda(\delta) = (I_n + BR^{-1}B^\mathsf{T}P) \in \mathsf{GL}^+(n,\mathbb{R})$. Since $\mathsf{GL}(n,\mathbb{R})$ has two connected components, $\Lambda(\delta)$ is continuous in δ (this can be shown as in [Pol86b]) and starts in $\mathsf{GL}^+(n,\mathbb{R})$, it must remain in that group. This concludes the proof. □

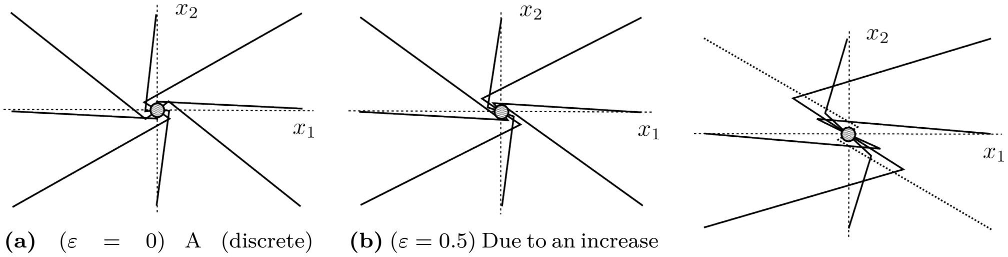
Corollary 4.1.3 indicates that the adversaries these dynamic games hedge against are somewhat *natural*.

Remark 4.1.1 (Beyond standard LQR)**.** Corollary 4.1.3 shows that Theorem 4.1.1 is not limited to the "*standard*" LQR problem. Hence, we would like to point to related problems, displaying similar, if not equivalent, structure. When considering an exponential utility function in (4.1.3) subject to a linear Gaussian system, which is called the LEQR problem, then, its optimal policy coincides with that of a dynamic game [Jac73, Whi90, BB95]. Similar algebraic structures are also seen in distributionally robust control and estimation [Yan20]. Note that in the stochastic case, the topological equivalence is with respect to the closed-loop mean state processes. Also, to be able to use Theorem 2.3.3 in a discounted setting, stability must be explicitly verified.

We can conclude, however, without a formal proof, that given any two optimal closed-loop time-one maps resulting from any two block-diagonal LQ OC problems, they are topologically equivalent. The crux is that all these optimal closed-loop maps are of the form $x \mapsto \Lambda^{-1}Ax$ for some $\Lambda \in \mathsf{GL}^+(n,\mathbb{R})$. Now, if the cost is not block-diagonal, the cross-terms will determine, for the better or worse, to which topological class the closed-loop map belongs.

This section also showed again the importance of correctly identifying $\ker(A)$ and $\operatorname{or}(A)$. If some estimate of A, say $\widehat{A}$, satisfies $\operatorname{or}(A) \neq \operatorname{or}(\widehat{A})$, then, for standard (block-diagonal) LQR, no matter the tuning, the simulated and real behaviour will always be structurally different, which was our partial motivation for the previous chapter.

4.1.4 Concluding comment on "tuning"



(a) ($\varepsilon = 0$) A (discrete) stable counter-clockwise spiral with $\operatorname{or}(\sigma^\star) = 1$.

(b) ($\varepsilon = 0.5$) Due to an increase in ε the spiral starts to shear, yet $\operatorname{or}(\sigma^\star) = 1$.

(c) ($\varepsilon = 2$) For $\varepsilon > 1$ $\operatorname{or}(\sigma^\star) = -1$ and the spiral deteriorates.

Figure 4.1: Given the parameters from Example 4.1.2, starting from 8 initial conditions on $\partial[-1,1]^2$, we show a few closed-loop trajectories as a function of $S(\varepsilon)$. In Figure 4.1c, we see that emergence of the dashed trajectory, breaking the spiral.

In the vast majority of work on LQ optimal control the stage-cost (4.1.2) is *diagonal* (*cf.* [Kal64, PCC$^+$15]). However, all of the above emphasizes that one should not underestimate the use of $S \neq 0$. One successful example is presented in [TD12] ("the balancing cube"), where the authors exploit S with the purpose of penalizing subsequent input deviations.

Looking at this from the tuning point of view, by excluding S, you are restricting the behaviour of the closed-loop system to maps with at least the same orientation as the automorphic part of $x \mapsto Ax$, *e.g.*, in the scalar case, by changing the pair (Q, R), one cannot go from *spring*- to *damper*-like behaviour. Therefore, we propose that if one wants to tune, if a change in behaviour is desired, change S. At last, we briefly visualize the effect of S.

Example 4.1.2 (Structural tuning)**.** Consider the general LQR problem from Section 4.1.2 parametrized by $B = I_2$, $R = I_2$, $Q = 10 \cdot I_2$ and

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad S(\varepsilon) = \begin{pmatrix} \varepsilon & 0 \\ 0 & -\varepsilon \end{pmatrix},$$

for some $\varepsilon \in \mathbb{R}_{\geq 0}$. We see that for all $\varepsilon < 1$, $(Q, R, S(\varepsilon)) \in \mathcal{C}^+(\sigma)$, *i.e.*, $\det(A' = A - S(\varepsilon)) = -\varepsilon^2 + 1$. To illustrate the structural change, we vary ε from 0 and 2, construct $K^\star$ accordingly and show a few closed-loop trajectories in Figure 4.1. Indeed, once S (formally, (Q, R, S)) leaves $\mathcal{C}^+(\sigma)$, the behaviour changes structurally (*cf.* Figure 4.1c).

4.2 On continuation and convex Lyapunov functions

In this section we will discuss how the structure of Lyapunov functions can be exploited towards a study of the space of stable dynamical systems.

4.2.1 Introduction

Ever since the time of Descartes, *convexity* has been recognized as an important notion to study structural properties of various mathematical objects [Fen83]. In this section, we aim to improve our understanding of the topology of spaces of stable systems and show that again convexity plays an important role. In particular, we study if vector fields on $\mathbb{R}^n$ with a common global attractor can be *continuously transformed* (formally, homotoped, see Section 4.2.2) into each other while preserving the attractor along the transformation. Before making this precise, we briefly elaborate on the relevance of this question.

One can argue that, originally, this question emerged in the dynamical systems community. That is, well over 40 years ago, Conley asked if dynamical systems with qualitatively similar properties can be continuously transformed into each other while preserving those properties along the transformation [Con78, p. 83]. In this work we address—in arguably the most simple setting—those *continuation* (see Section 4.2.3) questions as posed by Conley [Con78] and later Kvalheim [Kva23]. Our setting is simple in the sense that we largely focus on stability of equilibrium points instead of general attractors and spaces.

Then, in the context of linear optimal control, policy gradient methods have been recently shown to be a powerful controller synthesis paradigm as data and constraints can be naturally incorporated [HZL$^+$23]. Omitting details, these algorithms are frequently studied as being discretizations of a continuous-time gradient flow [BMFM19]. Here, the common assumption is that the optimal control cost is only finite under a stabilizing controller. Now, the hope is that, if initialized properly, gradient flow gives rise to a curve of *stabilized* closed-loop systems, moving from some initial closed-loop system to an optimal closed-loop system. As such, it is of great importance to understand *a priori* when such a curve exists, especially when moving beyond linear systems. For instance, when such a curve does not exist, a step size cannot be made arbitrarily small *cf.* [HZL+23, Thm. 1] as one might need to "*jump*", similarly, initialization becomes of critical importance. Indeed, the importance of understanding the topology of the space of stable systems has been recognized early on, *e.g.*, see [Bro76, Obe87] and motivated by these gradient-based methods this question received renewed interest, *e.g.*, see [FL19, BMFM19, BMM20].

A more surprising motivating example can be found in the context of switched systems. It turns out that if we have two vector fields such that the origin is globally asymptotically stable (GAS) (see Section 2.2.1), then, the origin remains GAS under arbitrary switching between those two vector fields only if those two vector fields can be continuously transformed into each other such that along the transformation the origin remains GAS (see Proposition 4.2.6 below).

Motivated by the above, this work aims to illustrate how the path-connectedness of spaces of dynamical systems can be studied via structural properties of Lyapunov functions. In particular, motivated by recent advances in learning [AXK17, KM19], we focus on the ramifications of assuming (control) Lyapunov functions—as pioneered by Artstein [Art83] and Sontag [Son89]—to be convex. Overall, this work is also in the spirit of the work by Arnold [Arn73, Sec. 22], Zabczyk [Zab89], Reineck [Rei91], Sepulchre and Aeyels [SA96], Grüne, Sontag and Wirth [GSW99], Coron [Cor07, Ch. 11], Byrnes [Byr08] and Cieliebak and Eliashberg [CE12, Ch. 9].

For this section we recall Section 2.2 on Lyapunov functions for dynamical control systems on $\mathbb{R}^n$. In particular, we recall Section 2.2.3, where we highlight topological properties of level sets of Lyapunov functions. These observations are the motivation for Sections 4.2.2-4.2.3 where we infer continuation results by considering several notions of convexity.

Notation is standard, but for simplicity we again highlight the main objects below. *Notation:* Let $r \in \mathbb{N} \cup \{\infty\}$, then, $C^r(U;V)$ denotes the set of C^r-smooth functions from U to V. The inner product on $\mathbb{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$ and $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. The Lie derivative of a smooth function h over some open set $U \subseteq \mathbb{R}^n$ with respect to a smooth vector field X over U is denoted by $L_X h$ and is defined pointwise by $L_X h(p) := \langle \nabla h(p), X(p) \rangle$ for any $p \in U$ [Lee12, Prop. 12.32]. By $\mathrm{cl}(W)$ we denote the closure of W and by $\mathrm{int}(W)$ we denote its interior. The map $x \mapsto x$ on $\mathbb{R}^n$ is denoted by $\mathrm{id}_{\mathbb{R}^n}$ and tangent spaces of appropriate sets M are denoted by T_pM, for $p \in M$, with TM denoting the corresponding tangent bundle [Lee12, p. 65].

4.2.2 On convexity

Section 2.2.3 illustrated why level sets of Lyapunov functions are topological spheres. As such, this motivates the hope that all those Lyapunov functions can be transformed—in some sense—to the canonical Lyapunov function $V(x) = \frac{1}{2}\langle x, x \rangle$. Indeed, Grüne, Sontag and Wirth [GSW99] showed that when V is a C^∞ Lyapunov function corresponding to 0 being GAS, then, there is a C^1 homeomorphism T such that $\widetilde{V}(T(y)) = V(y)$ for $\widetilde{V}(y) = \frac{1}{2}\langle y, y \rangle$. However, it is not clear if their arguments can be extended to construct a homotopy from $-\nabla V(y)$ to $-y$ along vector fields such that 0 remains GAS throughout.

The complication here is the topology of the homeomorphism- and diffeomorphism groups used in their line of arguments. Those spaces are not necessarily path-connected, similar to $\{X \in \mathbb{R}^{n\times n} : \det(X) \neq 0\}$ not being path-connected, see [Kup19, Ch. 9].

To continue, we start this study of transformations by looking at *convex* Lyapunov functions, as this class is particularly simple to handle. Better yet, by exploiting this structure, it follows that any convex Lyapunov function also asserts stability of the "*canonical*" inward pointing vector field on $\mathbb{R}^n$ indeed, which we will denote with some abuse of notation by the map $-\mathrm{id}_{\mathbb{R}^n}$, *i.e.*, giving rise to $\dot{x} = -x$. Exactly this observation will be formalized and further studied below.

Convex Lyapunov functions

Convexity in the context of Lyapunov stability theory has been an active research area. For example, convexity in linear optimal control [LR95], convexity in the dual density formulation due to Rantzer [PPR04], convexity of the set of Lyapunov functions due to Moulay [Mou10] and recently, component-wise convexity of vector fields to construct Chetaev functions due to Sassano and Astolfi [SA23]. We are, however, interested in understanding convexity of Lyapunov functions themselves. It is known that simple asymptotically stable dynamical systems do not always admit polynomial Lyapunov functions. For instance

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} -x_1(t) + x_1(t)x_2(t) \\ -x_2(t) \end{pmatrix} \tag{4.2.1}$$

does not admit a (global) polynomial Lyapunov function [AKP11], but one can show that $V(x) = \log(1 + x_1^2) + x_2^2$ is a Lyapunov function asserting $0 \in \mathbb{R}^2$ is GAS. Indeed, V is smooth, yet *not* convex. We will come back to this several times below. Similar obstructions can be found for analytic or rational Lyapunov functions [BR05, AEK18].

The (computational) assumption to look for *convex* Lyapunov functions is a popular one in the learning community, *e.g.*, propelled by [AXK17, KM19]. However, this assumption evidently restricts the problem class that can be handled. The ramifications of assuming Lyapunov functions to be convex are understood in the context of linear systems, even for linear differential inclusions [GTHL06] and linear switched systems [MCS23], but not completely in the C^0 nonlinear setting. An exception is [AJ18], where the authors consider nonlinear difference inclusions of the form $x_{k+1} \in \mathrm{conv}\{f_1(x_k), \dots, f_n(x_k)\}$ with $k \in \mathbb{N}$, $f_i \in C^0(\mathbb{R}^n; \mathbb{R}^n)$ $f_i(0) = 0$ for $i = 1, \dots, n$ and $\mathrm{conv}(\cdot)$ denoting the convex hull. Then, assuming that the maps $f_1, \dots, f_n$ share a common *convex* Lyapunov function allows for concluding on 0 being GAS[4]. Concurrently, they show that relaxing convexity is not possible in general, that is, counterexamples exist [AJ18, Ex. 1].

Similarly, in our setting, for $n > 1$, one can construct vector field examples $\dot{x} = F(x)$ over $\mathbb{R}^n$ such that 0 is globally asymptotically stable, F is smooth, yet no smooth convex Lyapunov function exists. To see why, for the sake of contradiction, one can exploit that by convexity[5] we must have $\langle \nabla V(x), x\rangle \geq 0$ $\forall x \in \mathbb{R}^n$ and due to the stability assumption we have $\langle \nabla V(x), F(x)\rangle < 0$ $\forall x \in \mathbb{R}^n \setminus \{0\}$ such that the function V must satisfy $\langle \nabla V(x), F(x) - x\rangle < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. Hence, if there is a non-zero fixed

[4] This should be understood in the discrete-time sense.

[5] Combine the inequalities $V(x) \geq V(x^\star) + \langle \nabla V(x^\star), x - x^\star\rangle$ and $V(x^\star) \geq V(x) + \langle \nabla V(x), x^\star - x\rangle$.

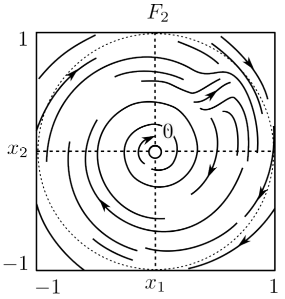


Figure 4.2: Example 4.2.1: integral curves of a smooth dynamical system that obstructs the existence of a smooth *convex* Lyapunov function, yet, 0 is GAS. Adapted from the original figure made with Python.

point[6] of F, we contradict the existence of such a V. See Figure 4.2 for a phase portrait illustrating a dynamical system with a fixed point obstructing the existence of a *smooth, convex* Lyapunov function. As one will be able to infer from the results below, F cannot point (radially) outward. Indeed, it is known that for *homogeneous* Lyapunov functions this can also not be true [SA96, Prop. 1]. We also remark that for convex Lyapunov functions Property (V-iii) is implied by Property (V-i) [AJ18, Lem. 4.1].

We will now formalize this observation. To do so, we recall the notion of a homotopy. The functions $f, g \in C^0(X;Y)$ are said to be *homotopic* when there is a continuous map $H : [0,1] \times X \to Y$ such that for any $x \in X$ we have that $x \mapsto H(0,x) = f(x)$ while $x \mapsto H(1,x) = g(x)$. The homotopy is said to be a *straight-line homotopy* when H is simply of the form $H(s,x) = (1-s)f(x) + sg(x)$. Note that homotopies only become interesting beyond $X = Y = \mathbb{R}^n$, *e.g.*, on manifolds *or* when requiring more structure to be preserved along the homotopy, as is done in this section. See [JM23] for more on homotopies in the context of control theory. See also Section 4.2.4 for more on how homotopies allow us to discuss path-connectedness.

Theorem 4.2.1 (Convex Lyapunov functions)**.** *Let $F \in C^0(\mathbb{R}^n;\mathbb{R}^n)$ give rise to $\dot{x} = F(x)$ with $0 \in \mathbb{R}^n$ globally asymptotically stable (GAS) under F. Then, if there is a convex C^∞ Lyapunov function asserting 0 is GAS, the vector field F is straight-line homotopic to $-\mathrm{id}_{\mathbb{R}^n}$ such that 0 is GAS throughout the homotopy.*

Proof. By assumption there is a C^∞ Lyapunov function V such that $\langle \nabla V(x), F(x)\rangle < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. By the convexity of V we also know that

$$V(y) \geq V(x) + \langle \nabla V(x), y - x\rangle, \quad \forall y, x \in \mathbb{R}^n. \tag{4.2.2}$$

In particular, (4.2.2) must hold for $y = 0$, which yields $\langle \nabla V(x), -x\rangle \leq -V(x)$, that is, V is also a Lyapunov function for $\dot{x} = -x$. Hence, we find that 0 is also GAS under $sF(x) - (1-s)x$ (or $(1-s)F(x) - sx$ for that matter) for all $s \in [0,1]$ since for any such s

$$\langle \nabla V(x), sF(x) - (1-s)x\rangle < 0 \quad \forall x \in \mathbb{R}^n \setminus \{0\}.$$

[6] Note, here we heavily exploit the underlying vector space structure to be able to compare x and $F(x)$.

Hence, $H(s,x)=sF(x)-(1-s)x$ is the homotopy. □

To illustrate the homotopy resulting from Theorem 4.2.1, two vector fields F_1 and F_2 on $\mathbb{R}^n$ such that 0 is GAS—asserted via possibly different smooth convex Lyapunov functions—are homotopic through a continuous map $H:[0,1]\times\mathbb{R}^n\to\mathbb{R}^n$ of the form

$$H(s,x)=\begin{cases}-2sx+(1-2s)F_1(x) & s\in[0,\frac{1}{2}]\\ -(2-2s)x+(2s-1)F_2(x) & s\in(\frac{1}{2},1].\end{cases}$$

A variety of known topological conditions capture the existence of a (local) homotopy (in far more general settings), but not that *along* the homotopy stability is preserved *cf.* [Kva23].

Similar statements can be made about *control* Lyapunov functions.

Corollary 4.2.2 (Convex control Lyapunov functions)**.** *Let $f\in C^0(\mathbb{R}^n\times\mathbb{R}^m;\mathbb{R}^n)$ give rise to the control system $\dot{x}=f(x,u)$. If there is a convex control Lyapunov function (CLF) $V\in C^\infty(\mathbb{R}^n;\mathbb{R}_{\geq 0})$ for this control system with respect to 0, then, V is a CLF for any control system on the straight-line homotopy between f and the map $(x,u)\mapsto -x$.*

Proof. The proof is identical to that of Theorem 4.2.1, yet, now we start from V satisfying

$$\forall x\in\mathbb{R}^n\setminus\{0\}\,\exists u\in\mathbb{R}^m\ :\ \langle\nabla V(x),f(x,u)\rangle<0.$$

and again exploit convexity of V to conclude. □

As remarked above, we see from Theorem 4.2.1 that a necessary condition for $\dot{x}=F(x)$ to admit a smooth, convex Lyapunov function, asserting 0 is GAS, is that

$$F(x)\neq\lambda x\quad\forall\lambda\in\mathbb{R}_{\geq 0},\,\forall x\in\mathbb{R}^n\setminus\{0\}.\tag{4.2.3}$$

Differently put, if V is a Lyapunov function for $\dot{x}=F(x)$, it is also a Lyapunov function for $\dot{x}=F(x)-\lambda x$, with $\lambda\geq 0$. The next example shows we can find families of dynamical systems that do not obey Condition (4.2.3).

Example 4.2.1 (Necessarily nonconvex)**.** The system as depicted in Figure 4.2 can be made explicit. Consider a C^∞ dynamical system of the form (2.2.1) on $\mathbb{R}^2$ as given by

$$\frac{\mathrm{d}}{\mathrm{dt}}\begin{pmatrix}x_1(t)\\x_2(t)\end{pmatrix}=\begin{pmatrix}\alpha & 1\\-1 & \alpha\end{pmatrix}\begin{pmatrix}x_1(t)\\x_2(t)\end{pmatrix}+\gamma\left(\exp(-\beta\|x(t)-p\|_2^2)-\exp(-\beta\|p\|_2^2)\right)\begin{pmatrix}1\\1\end{pmatrix}\tag{4.2.4}$$

where $x=(x_1,x_2)\in\mathbb{R}^2$, and $\alpha=-0.1$, $\beta=100$, $\gamma=10$ and $p=(0.5,0.5)\in\mathbb{R}^2$ correspond to Figure 4.2. Indeed, one can show that only $0\in\mathbb{R}^2$ is an equilibrium point of this dynamical system. Note, (4.2.4) can also be understood as a stabilizable linear system under a (bounded) nonlinear perturbation. Regarding our necessary condition for convexity, we find, for instance, that $x=(0.51,0.45)\in\mathbb{R}^2$ and $\lambda=1/0.0629$ provide for a (numerical)[7] invalidation of (4.2.3). Regarding 0 being GAS, we first note that 0 is

[7] We write "*numerical*" since the obstruction is not provided in closed-form. However, our argument is as follows. Due to the symmetry in the nonlinear part, a necessary condition for (4.2.4) to satisfy $F_2(x)=\lambda x$ is that $(\lambda-\alpha-1)x_1=(\lambda-\alpha+1)x_2$ holds for some appropriate tuple (λ,α,x_1,x_2). For instance, *fix* (α,λ,x_2) such that $(\lambda-\alpha-1)\neq 0$ and solve for x_1. Let $\zeta(x)=\exp(-\beta\|x-p\|_2^2)-\exp(-\beta\|p\|_2^2)$, then, we can simply set $\gamma=(\lambda x_1-\alpha x_1-x_2)/\zeta(x)$ (to avoid dividing by 0 we can simply adjust λ) such that $F_2(x)=\lambda x$ holds. The provided numerical values are obtained accordingly and hence provide for a valid obstruction.

hyperbolic and LAS. Then, it is not too hard to show that $\mathbb{D}^2 \subset \mathbb{R}^2$ is forward invariant under F_2, *e.g.*, construct a quadratic Lyapunov function for the linear part of (4.2.4), see Figure 4.2. Since we have a single equilibrium point at 0, the only obstruction to global asymptotic stability is the existence of a periodic orbit , which must live, if it exists, within $\mathbb{D}^2$ (the closed unit ball). Now we can also find a smaller ball $r\mathbb{D}^2$ of radius $r = 1/4$ that is again forward invariant. Hence, if a periodic orbit exists, it is either a semi-stable orbit, comprised of stable/unstable pairs, or a combination. It turns out that γ parametrizes such a bifurcation, with 0 being GAS for our choice of parameters. Details are provided elsewhere.

Remark 4.2.2 (A nonconvex conic structure)**.** Let $0 \in \mathbb{R}^n$ be GAS under $\dot{x} = F(x)$, then 0 is also GAS under $\dot{x} = \theta F(x)$ for any $\theta > 0$, *e.g.*, consider $\langle \nabla V(x), F(x)\rangle$ and $\langle \nabla V(x), \theta F(x)\rangle$ for some Lyapunov function V with respect to F. Hence, if F is convex, then by Theorem 4.2.1, all θF are straight-line homotopic to $-\mathrm{id}_{\mathbb{R}^n}$. Despite the conic structure, convexity of the set of these vector fields (vector fields such that 0 is GAS, verified via a convex Lyapunov functions) breaks down as already the set of Hurwitz stable matrices is nonconvex, *e.g.*,

$$s\begin{pmatrix} -1 & 10 \\ 0 & -1 \end{pmatrix} + (1-s)\begin{pmatrix} -1 & 0 \\ 10 & -1 \end{pmatrix}$$

becomes unstable (not all eigenvalues lie in $\mathbb{C}_{\Re<0}$) for $s = 1/2$.

Similarly, from Corollary 4.2.2 we see that the control system $\dot{x} = f(x,u)$ admits a smooth, convex CLF only when

$$\forall x \in \mathbb{R}^n \setminus \{0\}\, \exists u \in \mathbb{R}^m \;:\; f(x,u) \neq \lambda x \quad \forall \lambda \in \mathbb{R}_{\geq 0}. \tag{4.2.5}$$

Indeed, one can replace $\lambda \in \mathbb{R}_{\geq 0}$ in (4.2.5) by $\lambda(x) \in \mathbb{R}_{\geq 0}$, for example, $\lambda \in C^0(\mathbb{R}^n; \mathbb{R}_{\geq 0})$ *cf.* [SA96].

Example 4.2.3 (Linear dynamical systems)**.** Consider the linear dynamical system $\dot{x} = Ax$ for some matrix $A \in \mathbb{R}^{n\times n}$. Theorem 4.2.1 implies that for a convex Lyapunov function to exist (asserting 0 is GAS) the expression $sAx-(1-s)x$ cannot vanish for some $s \in [0,1]$ and $x \in \mathbb{R}^n \setminus 0$. Reformulating, we get (4.2.3), *i.e.*, $Ax = \lambda x$ cannot have a solution for some $\lambda \geq 0$ and $x \in \mathbb{R}^n \setminus \{0\}$. However, this is precisely stating that A cannot have an unstable eigenvalue of the form $\lambda \in \mathbb{R}_{\geq 0}$. Indeed, for globally asymptotically stable linear systems a convex (quadratic) Lyapunov function of the form $V(x) = \langle Px, x\rangle$ always exists [Son98, Thm. 18].

Remark 4.2.4 (On sufficiency)**.** Example 4.2.1 showed that for dynamical systems of the form (4.2.4) (for certain parameters) no convex Lyapunov function can exist. Going back to (4.2.1), the provided Lyapunov function is nonconvex. Concurrently, one can check that (4.2.3) holds, so a convex Lyapunov function is not ruled out. We come back to this below.

To elaborate on Example 4.2.3, for controllable linear systems, *e.g.*, of the form $\dot{x} = Ax+Bu$, one can always parameterize a quadratic Lyapunov function for the LQ optimally controlled closed-loop system by the positive definite solution to the corresponding Riccati equation (for any appropriate cost) [Son98, Thm. 42, Ex. 8.5.4].

Example 4.2.5 (Linear dynamical control systems and Hautus' test)**.** A celebrated condition largely attributed to Hautus (plus Belevitch and Popov) states that a linear dynamical control system of the form $\dot{x} = Ax + Bu$, is stabilizable when

$$\operatorname{rank}\left(\begin{pmatrix} A - \lambda I_n & B \end{pmatrix}\right) = n \quad \forall \lambda \in \sigma(A) \cap \mathbb{C}_{\Re \geq 0}, \tag{4.2.6}$$

where $\sigma(A)$ denotes the spectrum of A. See for instance [TSH12, Ch. 3]. Now, elementary algebraic arguments show that Hautus' condition (4.2.6) implies that (4.2.5) holds, as it should for linear control systems.

Using the above, one can readily verify that, for example

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} x_1(t)u \\ x_1(t)x_2(t)u \end{pmatrix} \tag{4.2.7}$$

does not admit a smooth, convex CLF. Indeed, for (4.2.7), controllability is lost at $(0, x_2) \in \mathbb{R}^2$.

Example 4.2.3 and Example 4.2.5 show that conditions (4.2.3) and (4.2.5) are to some extent generalizations of known conditions for linear systems, yet, lifted to nonlinear systems under convexity assumptions. These conditions are, however, weak.

A stronger set of conditions one can derive from Theorem 4.2.1 is of the form: $\dot{x} = f(x, u)$ admits a smooth, convex CLF only if $\dot{x} = f(x, u) - \lambda(x)x$ does, for any $\lambda \in C^0(\mathbb{R}^n; \mathbb{R}_{\geq 0})$. We are not the first to observe something of this form, *e.g.*, Sepulchre and Aeyels [SA96, Sec. 4.1] look at homogeneous CLFs and recover a similar condition.

We close this subsection with a comment on mere *Lyapunov stability*, that is, Property (V-ii) is replaced with the weaker notion $\langle \nabla V(x), F(x) \rangle \leq 0$. This notion of stability is understood as local, yet, for practical purposes such a Lyapunov function allows for concluding trajectories to remain bounded.

Remark 4.2.6 (Lyapunov stability)**.** At the time of writing, several examples of Lyapunov stable dynamical systems surfaced that provably fail to admit a smooth, convex Lyapunov function [APH23]. We show that our line of arguments offers an arguably simpler means of reaching such a conclusion. Following the same reasoning as for Theorem 4.2.1, when 0 is Lyapunov stable under some vector field F and comes equipped with a C^∞ convex Lyapunov function V, then, we must have that $\langle \nabla V(x), F(x) - \lambda x \rangle \leq 0$ $\forall x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}_{\geq 0}$. We claim that the existence of a point $x' \in \mathbb{R}^n \setminus \{0\}$ such that $F(x')_i = \lambda' x'_i \neq 0$ for $i = 1, \dots, n$ for some $\lambda' \in \mathbb{R}_{>0}$ contradicts the existence of such a function V. To see this, suppose that such a pair (x', λ') exists, then we can find $\lambda_1, \lambda_2 \in \mathbb{R}_{>0}$ such that $2F(x') = (\lambda_1 + \lambda_2)x'$. In particular, we have that $F(x') - \lambda_1 x' = (-1)(F(x') - \lambda_2 x')$. We can select $\lambda_1 \neq \lambda_2$ such that by construction no element of the equation above equals 0. However, that means that when we move λ from λ_1 to λ_2, the sign of $\langle \nabla V(x'), F(x') - \lambda x' \rangle$ flips, which is a contradiction. One can employ precisely this argument to show that for $k > 1$ the origin of the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} r(t) \\ \theta(t) \end{pmatrix} = \begin{pmatrix} r^{2k+1}\sin(1/r) \\ 1 \end{pmatrix}$$

is Lyapunov stable, yet, no smooth, convex Lyapunov function exists to assert this *cf.* [APH23, Thm. 1].

On compact convex sets

We briefly show that without too much effort the results extend from $0 \in \mathbb{R}^n$ being GAS under some dynamical system parametrized by $F \in C^0(\mathbb{R}^n; \mathbb{R}^n)$ to a compact convex set $A \subseteq \mathbb{R}^n$ being GAS[8] under F. As A is homotopy equivalent to a point, this is perhaps not surprising. Define the projection operator by

$$\Pi_A(x) := \arg\min_{y \in A} \|x - y\|_2.$$

We have the following.

Corollary 4.2.3 (Convex Lyapunov functions for convex compact sets)**.** *Let $F \in C^0(\mathbb{R}^n; \mathbb{R}^n)$ give rise to $\dot{x} = F(x)$ with a compact convex set $A \subseteq \mathbb{R}^n$ being globally asymptotically stable (GAS) under F. Then, if there is a convex C^∞ Lyapunov function asserting A is GAS, the vector field F is straight-line homotopic to $\Pi_A - \mathrm{id}_{\mathbb{R}^n}$ on $\mathbb{R}^n \setminus A$ such that A is GAS throughout the homotopy.*

Proof. The Lyapunov function is such that $V(x) = 0 \iff x \in A$, hence for the convexity condition $V(y) \geq V(x) + \langle \nabla V(x), y - x \rangle$ we pick $y = \Pi_A(x)$ such that for all $x \in \mathbb{R}^n \setminus A$ we have $\langle \nabla V(x), \Pi_A(x) - x \rangle < 0$. We can conclude. □

Some comments are in place, we do not need $F(A) = 0$, A merely needs to be invariant[9]. This is why we cannot say anything about the homotopy on A itself. Moreover, settings like these easily obstruct $V \in C^\omega(\mathbb{R}^n; \mathbb{R}_{\geq 0})$ (real-analyticity), not to contradict real-analytic function theory (bump functions cannot be C^ω). Also, when A is not convex, Π_A is potentially set-valued, obstructing our vector field construction and perhaps $\Pi_A - \mathrm{id}_{\mathbb{R}^n}$ is not the expected "*canonical*" inward vector field (due to the non-scaled offset $-x$). At last, we point out that although V is smooth, this does not imply that ∂A must be a smooth manifold. For instance, consider $V(x_1, x_2) = (x_1 - x_2)^2(x_1 + x_2)^2$ (although here, $V^{-1}(0)$ is clearly not convex), or the ℓ_2 distance from $[-1, 1]^n$. We direct the reader to [FP19] and references therein for more on Lyapunov theory with respect to sets.

Geodesic convexity

To go beyond vanilla convexity, we follow [Udr13, Ch. 3], [Bou23, Ch. 11] and show how the situation is hardly different in the context of *geodesic* convexity. We will be brief, for the details on geodesic convexity we point the reader to the references above and for background information on Riemannian geometry we suggest [Lee97].

Let $(\mathbb{R}^n, g)$ be a C^∞ *Riemannian manifold* for some Riemannian metric g. One can think of g as inducing a change of coordinates via the inner product $\langle \cdot, \cdot \rangle_g$, in particular, this metric has an effect on gradients, that is, the (Riemanian) gradient of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, with respect to g, satisfies $Df(x)[v] = \langle \operatorname{grad} f(x), v \rangle_g$ for any $(x, v) \in T\mathbb{R}^n$, with $Df(x)[v]$ being the directional derivative in the direction $v \in T_x\mathbb{R}^n$. For example, let g be parametrized by a symmetric positive definite matrix P, that is,

[8] For more on the generalization of stability notions from points to sets we point the reader to [Hur82] for a topological treatment.

[9] Let φ be the flow corresponding to F, then A is said to be invariant (under φ) when $\varphi(\mathbb{R}, A) = A$.

$\langle v,w\rangle_g := \langle Pv,w\rangle$ for any $v,w\in T_x\mathbb{R}^n$ and $x\in\mathbb{R}^n$, then, $\operatorname{grad} f(x) = P^{-1}\nabla f(x)$. Indeed, for a practical application of this in $\mathbb{R}^n$, we point the reader to a discussion of Newton's method as used in second-order optimization [BV04, Sec. 9.5]. The metric g also has ramifications for "*straight lines*", a C^1 curve $[0,1]\ni s\mapsto\gamma(s)$ is a *geodesic*, with respect to g, when it is an *extremal* of the energy functional $\gamma\mapsto E(\gamma):=\frac{1}{2}\int_{[0,1]}\langle\dot\gamma(\tau),\dot\gamma(\tau)\rangle_g\mathrm{d}\tau$. This implies geodesics are *locally* minimizing length and in that sense they generalize straight lines. As this statement is local, geodesics are by no means always unique. Then, a subset $U\subseteq\mathbb{R}^n$ is called *geodesically convex* (g-convex) when for all points $x,y\in U$ there is a *unique*[10] geodesic $\gamma:[0,1]\to\mathbb{R}^n$ (with respect to g) connecting x to y such that $\gamma([0,1])\subseteq U$. A function $f: U\subseteq\mathbb{R}^n\to\mathbb{R}$, over some g-convex domain U, is said to be ***geodesically convex*** (***g-convex***) when

$$(1-t)f(x)+tf(y)\geq f(\gamma(t))\quad\forall t\in[0,1] \tag{4.2.8}$$

for $\gamma:[0,1]\to\mathbb{R}^n$ a geodesic, with $\gamma([0,1])\subseteq U$ connecting the point x to y. Indeed, (4.2.8) generalizes the standard C^0 definition of convexity. A C^1 condition is now given by

$$f(\operatorname{Exp}_x(tv))\geq f(x)+t\langle\operatorname{grad} f(x),v\rangle_g\quad\forall t\in[0,1],$$

where $v\in T_x\mathbb{R}^n$, Exp_x is the (Riemannian) *exponential map* at $x\in U\subseteq\mathbb{R}^n$ and $\operatorname{grad} f(x)$ is the *Riemannian gradient* of f. Here, the exponential map is defined, locally, by $\operatorname{Exp}_x(v)=\gamma(1)$ for γ the unique geodesic with $\gamma(0)=x$ and $\dot\gamma(0)=v$.

Similarly, for a C^2 condition, a function f is g-convex when the *Riemannian Hessian* satisfies $\operatorname{Hess} f(x)\succeq 0$ for all $x\in U\subseteq\mathbb{R}^n$ (see [AMS09, ch. 5] for a correct interpretation). The interest in g-convex functions stems from the fact that local minima are again global minima, as with standard convex functions.

We are now equipped to generalize Theorem 4.2.1.

Theorem 4.2.4 (Geodesically convex Lyapunov functions)**.** *Let $(\mathbb{R}^n,g)$ be a Riemannian manifold and let $U\subseteq\mathbb{R}^n$ be open and g-convex. Let $F\in C^0(U;\mathbb{R}^n)$ give rise to $\dot x=F(x)$ with $0\in U$ globally asymptotically stable (GAS) (on U) under F. Then, if there is a g-convex C^∞ Lyapunov function asserting 0 is GAS, the vector field F is straight-line homotopic to $\operatorname{Exp}^{-1}(0)$ such that 0 is GAS throughout the homotopy.*

In Theorem 4.2.4, $\operatorname{Exp}^{-1}(0)$ should be understood as the map being defined by $x\mapsto\operatorname{Exp}_x^{-1}(0)\in T_x\mathbb{R}^n$.

Proof. By assumption, there is a C^∞ Lyapunov function V such that $\langle\nabla V(x),F(x)\rangle<0$ for all $x\in U\setminus\{0\}$. By the g-convexity of V we also know that for all $t\in[0,1]$ and $(x,v)\in TU$ we have

$$V(\operatorname{Exp}_x(tv))\geq V(x)+t\langle\operatorname{grad} V(x),v\rangle_g=V(x)+t\langle\nabla V(x),v\rangle,$$

where we removed the dependency on the metric g by identifying both inner products with the directional derivative $DV(x)[v]$. We consider $t=1$ and pick $v:=\operatorname{Exp}_x^{-1}(0)$. This map is always well-defined since our geodesics are unique. Now we proceed exactly as in the proof of Theorem 4.2.1 and conclude. □

[10] See the discussion in [Bou23, Sec. 11.3] on various slightly different definitions of *geodesic convexity* and their implications.

Indeed, we recover Theorem 4.2.1 for the identity metric on $\mathbb{R}^n$ and $U = \mathbb{R}^n$. In particular, in that case we can define our Riemannian exponential map as $\mathrm{Exp}_x(v) = x+v$ for $v \in T_xU$. Hence, the tangent vector v such that $\mathrm{Exp}_x(v) = 0$ is simply $-x$ (now seen as a tangent vector), *i.e.*, $\mathrm{Exp}_x^{-1}(0) = -x$. Recall, formally speaking, $-\mathrm{id}_{\mathbb{R}^n}$ should be understood as $x \mapsto (x, -x) \in TU$ while ignoring the first component of the image. With this in mind we can again understand $\mathrm{Exp}^{-1}(0)$ as the canonical "*inward*" vector field, yet now on a subset of $(\mathbb{R}^n, g)$.

Generalizing to compact manifolds and so forth (beyond contractible sets) is somewhat nonsensical as no smooth function with a single critical point exists on those spaces. This restriction comes from the demand that our geodesics are unique, obstructing nontrivial topologies. See [Udr13, Ch. 4] for more pointers.

A similar generalization can be achieved through the lens of *contraction analysis* [LS98]. See in particular [WS20] for a relation between g-convexity and contraction metrics.

We end this section by returning to Remark 4.2.4, the Lyapunov function with respect to (4.2.1) is nonconvex, yet the dynamical system satisfies the necessary condition (4.2.3). Indeed, fixating on each quadrant separately, the function *is* g-convex[11] (under quadrant-wise exponential geodesics, which suffices thanks to the invariance properties of the vector field), and the necessary condition effectively extends (as inferred from Theorem 4.2.4).

4.2.3 On continuation

The existence of a mere homotopy is not immediately informative. Often, only when the homotopy itself satisfies certain properties, one can draw nontrivial conclusions.

In our case the homotopies as detailed in Theorem 4.2.1, Corollary 4.2.2, Corollary 4.2.3 and Theorem 4.2.4 all preserve qualitative properties of the underlying dynamical system. More formally, this construction provides a *continuation* in the sense of Conley, albeit from a different perspective. Again, we are decidedly brief, but we point the reader to [Con78, MM02] for more details on Conley index theory and suggest [Hat02] as a reference on algebraic topology.

Recall that, under our assumptions, a dynamical system of the form (2.2.1) gives rise to a global flow $\varphi : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Let $S \subset \mathbb{R}^n$ be an *isolated invariant set* (with respect to φ), that is,

$$S = \mathrm{Inv}(M, \varphi) := \{x \in M : \varphi(\mathbb{R}, x) \subseteq M\} \subseteq \mathrm{int}(M)$$

for some compact set $M \subset \mathbb{R}^n$. Note that not every invariant set is isolated, *e.g.* consider an equilibrium point of the *center*-type. Then, a pair of compact sets $(N, L) \subset \mathbb{R}^n \times \mathbb{R}^n$ is an *index pair* for S when

(I-i) $S = \mathrm{Inv}(\mathrm{cl}(N \setminus L), \varphi)$ and $N \setminus L$ is a neighbourhood of S;

(I-ii) L is positively invariant in N;

(I-iii) L is an exit set for N (a trajectory that leaves N, must leave through L).

Now, the (homotopy) *Conley index* of S is the homotopy type of the pointed (quotient) space $(N/L, [L])$, *e.g.*, for $N = \mathbb{B}^n$, $L = \partial\mathbb{B}^n = \mathbb{S}^{n-1}$, we have that $N/L \simeq_t \mathbb{S}^n$ such

[11] In fact, the function $x \mapsto \log(1 + x^2)$ is also semiconcave.

that $(N/L,[L])$ is the pointed n-sphere. As this object is hard to computationally work with, let $H^k(A,B;\mathbb{Z})$ denote the k^{th} singular cohomology group of A relative to $B\subseteq A$, then, the *homological* ***Conley index*** defined as $\mathrm{CH}^k(S,\varphi):=H^k(N/L,[L];\mathbb{Z})$ is of larger interest, *e.g.*, as computational tools *are* available [KMM04]. Going back to our setting, assume for that moment that $0\in\mathbb{R}^n$ is a GAS hyperbolic fixed point of the flow φ. By hyperbolicity (local linearity), we can pick $N=\varepsilon\mathbb{B}^n$ (a sufficiently small closed ball in $\mathbb{R}^n$) and $L=\emptyset$. Now see that

$$\begin{aligned}\mathrm{CH}^k(0,\varphi)&=H^k(\varepsilon\mathbb{B}^n/\emptyset,[\emptyset];\mathbb{Z})\\&\simeq H^k(\varepsilon\mathbb{B}^n;\mathbb{Z})\\&\simeq\begin{cases}\mathbb{Z} & \text{if } k=0\\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

since $\varepsilon\mathbb{B}^n$ is homotopic (homotopy equivalent) to a point. If 0 is not hyperbolic, pick N to be a sub-level set of a smooth Lyapunov function that asserts 0 is GAS, this set is compact by Property (V-iii). Indeed, constructions like these provide for topological obstructions [MH11].

Now, if some N can be chosen to be an isolating neighbourhood *throughout* a homotopy, then the Conley index is preserved along that homotopy [Mro94, Thm. 1.10]. Simply put, we speak in this case of a ***continuation*** between the dynamical system at the beginning and the end of the homotopy. A question asked by Conley concerns the opposite [Con78, p. 83], to what extent do equivalent Conley indices relate to the existence of such a continuation. See also the discussion in [MRS00, Kva23]. Indeed, we see that if there is a homotopy through flows $[0,1]\ni\lambda\mapsto\varphi_\lambda$ such that 0 is GAS along the homotopy, then $\mathrm{CH}^k(0,\varphi_0)\simeq\mathrm{CH}^k(0,\varphi_1)$.

For the other direction, based on the above we have the following. One can extend the statement to compact convex sets or g-convexity if desired.

Corollary 4.2.5 (On continuation and convex Lyapunov functions)**.** *Let $0\in\mathbb{R}^n$ be GAS under two dynamical systems of the form* (2.2.1) *parametrized by F_0 and F_1, giving rise to the flows φ_0 and φ_1. Assume that 0 being GAS is asserted by—possibly different—smooth, convex Lyapunov functions V_0 and V_1. Then 0 (with respect to φ_0) and 0 (with respect to φ_1) are related by continuation where N can be chosen to be of the form $N=N_0\cap N_1$ (based on sub-level sets of V_0 and V_1).*

A further study of this observation is the topic of future work.

To return to similarities pointed out in the introduction, the work by Reineck [Rei91] and the proof of [Cor07, Thm. 11.4] provide the homotopy (preserving the Conley index) between F and the (a) negative gradient flow $-\nabla V$. However, how to link—if at all—multiple dynamical systems is unclear. The book by Cieliebak and Eliashberg does contain results in this direction, yet under C^k-nearness assumptions [CE12, Ch. 9], not in general.

Then, this work alludes to convexity being a simple structural ingredient to actually link several dynamical systems together via some canonical dynamical system.

4.2.4 Appendix

In this appendix we present auxiliary results on topology and switched systems.

Homotopies and path-connectedness

We frequently refer to spaces of stable dynamical systems and ask if such a space is path-connected or not, however, without making precise how to think of continuous curves in such a space. In this appendix, we briefly highlight how to go about this. We recall that we identify continuous vector fields on $\mathbb{R}^n$ with elements of $C^0(\mathbb{R}^n;\mathbb{R}^n)$. We will address in which sense the homotopy $H:[0,1]\times\mathbb{R}^n\to\mathbb{R}^n$ provides a *continuous path* in the space $C^0(\mathbb{R}^n;\mathbb{R}^n)$. Due to space constraints, the discussion is brief, but for more details, we refer the reader to [Hir76, Mun14].

First, a topological space X is said to be *path-connected* if for any two points $x_0,x_1\in X$ there exists a continuous map $\gamma:[0,1]\to X$, a path, such that $\gamma(0)=x_0$ and $\gamma(1)=x_1$. Then, to reason about continuous curves in $C^0(\mathbb{R}^n;\mathbb{R}^n)$ we need to endow it with a topology, that is, we need to decide when two continuous maps are "*close*". Leaving $\mathbb{R}^n$ for the moment, given two topological spaces X and Y, then, for K a compact subset of X and U an open subset of Y, sets of the form $\mathcal{O}(K,U):=\{f:f\in C^0(X;Y),\,f(K)\subset U\}$ comprise a *subbasis* for the *compact-open topology* on $C^0(X;Y)$. It turns out that this is the appropriate topology, as one can show the following. Let X,Y and Z be topological spaces with X locally compact Hausdorff and endow $C^0(X;Y)$ with the compact-open topology, then, the map $H:Z\times X\to Y$ is continuous if and only if the map $h:Z\to C^0(X;Y)$ is continuous, where h is defined by $(h(z))(x)=H(z,x)$ [Mun14, Thm. 46.11]. In particular, pick $Z=[0,1]$ and $X=Y=\mathbb{R}^n$, then, the existence of a homotopy $H:[0,1]\times\mathbb{R}^n\to\mathbb{R}^n$ is equivalent to a continuous path in $C^0(\mathbb{R}^n;\mathbb{R}^n)$.

Switched systems

Suppose we have a finite set of locally Lipschitz vector fields $\mathcal{F}=\{F_1,\dots,F_n\}$ on $\mathbb{R}^n$ such that the origin $0\in\mathbb{R}^n$ is GAS under any $F_i\in\mathcal{F}$. Now one might be interested in understanding if 0 is still GAS under arbitrary switching between elements of $\mathcal{F}$, that is, to understand if 0 is GAS under the *switched system*

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t)=F_{\sigma(t)}(x(t)),\tag{4.2.9}$$

where $t\mapsto\sigma(t)$ is a piecewise constant function taking values in $\{1,\dots,n\}$. It is known that for this to be true a common C^∞ Lyapunov function V must exist [MAG00]. However, by the proceeding arguments we know that this implies that any $F_i\in\mathcal{F}$ can be homotoped to $-\nabla V$ such that 0 remains GAS along the homotopy. Then, by the transitive properties of homotopies, this implies that for any two elements of $\mathcal{F}$ there must be a homotopy between them such that along the homotopy 0 remains GAS. Hence, a somewhat counterintuitive statement is the following.

Proposition 4.2.6 (Necessary condition for switched stability)**.** *The origin $0\in\mathbb{R}^n$ is GAS under (4.2.9) only if all elements of $\mathcal{F}$ belong to the same path-connected component of the space of vector fields on $\mathbb{R}^n$ for which 0 is GAS.*

Hence, Proposition 4.2.6 further motivates studying the topology of the space of vector fields with a common attractor, *e.g.*, vector fields on $\mathbb{R}^n$ such that 0 is GAS. It is interesting to note that under the aforementioned conditions, the switched system (4.2.9) can be continuously transformed to a negative gradient flow, without sacrificing stability. To that end, simply construct the maps $H_\sigma:[0,1]\times\mathbb{R}^n\to\mathbb{R}^n$ defined by $H_\sigma(s,x)=(1-s)F_\sigma(x)-s\nabla V(x)$ and observe that for any $s\in[0,1]$ the origin is GAS under

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t)=H_{\sigma(t)}(s,x(t)).$$

For more on switched systems we refer the reader to [Lib03]. In particular, we point the reader to [Lib03, Rem. 2.1] for subtleties with respect to Lyapunov functions for switched systems. In fact, from the same point of view, one observes that a necessary condition for 0 to be GAS under (4.2.9) is that 0 is GAS under any element of the convex hull of $\mathcal{F}$, *e.g.*, consider $\langle\theta F_i+(1-\theta)F_j,\nabla V\rangle$ for $\theta\in[0,1]$ [Lib03, Cor. 2.3].

Bibliography

[AEK18] A A Ahmadi and B El Khadir. A globally asymptotically stable polynomial vector field with rational coefficients and no local polynomial Lyapunov function. *Syst. Control Lett.*, 121:50–53, 2018.

[AJ18] A A Ahmadi and R M Jungers. SOS-convex Lyapunov functions and stability of difference inclusions. *arXiv e-print:1803.02070*, 2018.

[AKP11] A A Ahmadi, M Krstic, and P A Parrilo. A globally asymptotically stable polynomial vector field with no polynomial Lyapunov function. In *Proc. IEEE Conference on Decision and Control and European Control Conference*, pages 7579–7580, 2011.

[AM89] B D O Anderson and J B Moore. *Optimal Control: Linear Quadratic Methods.* Prentice-Hall, Hoboken, 1989.

[AMS09] P.-A Absil, R Mahony, and R Sepulchre. *Optimization algorithms on matrix manifolds.* Princeton University Press, Princeton, 2009.

[APH23] M Akbarian, N Pariz, and A Heydari. Introducing some classes of stable systems without any smooth Lyapunov functions. *Commun. Nonlinear Sci. Numer. Simul.*, page 107485, 2023.

[Arn73] V I Arnold. *Ordinary differential equations.* MIT Press, Cambridge, 1973.

[Art83] Z Artstein. Stabilization with relaxed controls. *Nonlinear Anal.-Theor.*, 7(11):1163–1173, 1983.

[AXK17] B Amos, L Xu, and J Z Kolter. Input convex neural networks. In *Proc. International Conference on Machine Learning*, pages 146–155, 2017.

[BB95] T Basar and P Bernhard. *H_∞-Optimal Control and Related Minimax Design Problems A Dynamic Game Approach.* Birkhauser, 1995.

[BMFM19] J Bu, A Mesbahi, M Fazel, and M Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv e-print:1907.08921*, 2019.

[BMM20] J Bu, A Mesbahi, and M Mesbahi. On topological properties of the set of stabilizing feedback gains. *IEEE T. Automat. Contr.*, 66(2):730–744, 2020.

[Bou23] N Boumal. An introduction to optimization on smooth manifolds. Cambridge University Press, Cambridge, 2023.

[BR05] A Bacciotti and L Rosier. *Liapunov functions and stability in control theory.* Springer Science & Business Media, Berlin, 2005.

[Bro76] R W Brockett. Some geometric questions in the theory of linear systems. *IEEE T. Automat. Contr.*, 21(4):449–455, 1976.

[BV04] S Boyd and L Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, 2004.

[Byr08] C I Byrnes. On Brockett's necessary condition for stabilizability and the topology of Liapunov functions on $\mathbb{R}^n$. *Commun. Inf. Syst.*, 8(4):333–352, 2008.

[CE12] K Cieliebak and Y Eliashberg. *From Stein to Weinstein and back: symplectic geometry of affine complex manifolds.* American Mathematical Society, Providence, 2012.

[CL19] P E Caines and D Levanony. Stochastic ε-optimal Linear Quadratic adaptation: An alternating controls policy. *SIAM J. Control Optim.*, 57(2):1094–1126, 2019.

[Con78] C C Conley. *Isolated invariant sets and the Morse index.* American Mathematical Society, Providence, 1978.

[Cor07] J.-M Coron. *Control and Nonlinearity.* American Mathematical Society, Providence, 2007.

[DMM$^+$20] S Dean, H Mania, N Matni, B Recht, and S Tu. On the sample complexity of the Linear Quadratic Regulator. *Found. Comput. Math.*, 20(4):633–679, 2020.

[Fen83] W Fenchel. Convexity through the ages. In *Convexity and its Applications*, pages 120–130. Birkhäuser, Basel, 1983.

[FGKM18] M Fazel, R Ge, S Kakade, and M Mesbahi. Global convergence of policy gradient methods for the Linear Quadratic Regulator. In *Proc. International Conference on Machine Learning*, pages 1467–1476, 2018.

[FL19] H Feng and J Lavaei. On the exponential number of connected components for the feasible set of optimal decentralized control problems. In *Proc. IEEE American Control Conference*, pages 1430–1437, 2019.

[FP19] A Fathi and P Pageault. Smoothing Lyapunov functions. *T. Am. Math. Soc.*, 371(3):1677–1700, 2019.

[GSW99] L Grüne, E D Sontag, and F R Wirth. Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes. *Syst. Control Lett.*, 38(2):127–134, 1999.

[GTHL06] R Goebel, A R Teel, T Hu, and Z Lin. Conjugate convex Lyapunov functions for dual linear differential inclusions. *IEEE T. Automat. Contr.*, 51(4):661–666, 2006.

[Hat02] A Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, 2002.

[Hir76] M W Hirsch. *Differential topology.* Springer, New York, 1976.

[Hur82] M Hurley. Attractors: persistence, and density of their basins. *T. Am. Math. Soc.*, 269(1):247–271, 1982.

[HZL$^+$23] B Hu, K Zhang, N Li, M Mesbahi, M Fazel, and T Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annu. Rev. Control Robot. Auton.*, 6:123–158, 2023.

[Jac73] D Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE T. Automat. Contr.*, 18(2):124–131, 1973.

[JM23] W Jongeneel and E Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, Cham, 2023.

[Jon19] W Jongeneel. Controlling the unknown: A game theoretic perspective. Master's thesis, Systems & Control, Delft University of Technology, 2019.

[Kal64] R E Kalman. When Is a Linear Control System Optimal? *J. Fluids Eng.*, 86(1):51–60, 1964.

[KM19] J Z Kolter and G Manek. Learning stable deep dynamics models. In *Proc. Neural Information Processing Systems*, pages 11126–11134, 2019.

[KMM04] T Kaczynski, K M Mischaikow, and M Mrozek. *Computational homology.* Springer, New York, 2004.

[Kup19] A Kupers. Lectures on diffeomorphism groups of manifolds, 2019. https://people.math.harvard.edu/~kupers/teaching/272x/book.pdf.

[Kva23] M D Kvalheim. Obstructions to asymptotic stabilization. *SIAM J. Control Optim.*, 61(2):536–542, 2023.

[Lee97] J M Lee. *Riemannian Manifolds.* Springer, New York, 1997.

[Lee12] J M Lee. *Introduction to Smooth Manifolds.* Springer, New York, 2012.

[Lib03] D Liberzon. *Switching in systems and control.* Birkhäuser, Boston, 2003.

[LR95] P Lancaster and L Rodman. *Algebraic Riccati Equations.* Oxford University Press, Oxford, 1995.

[LS98] W Lohmiller and J.-J E Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

[MAG00] J L Mancilla-Aguilar and R A Garcıa. A converse Lyapunov theorem for nonlinear switched systems. *Syst. Control Lett.*, 41(1):67–71, 2000.

[MCS23] P Mason, Y Chitour, and M Sigalotti. On universal classes of Lyapunov functions for linear switched systems. *Automatica*, 155:111155, 2023.

[MH11] E Moulay and Q Hui. Conley index condition for asymptotic stability. *Nonlinear Anal.-Theor.*, 74(13):4503–4510, 2011.

[MM02] K Mischaikow and M Mrozek. Conley index. In *Handbook of dynamical systems*, volume 2, pages 393–460. Elsevier, Amsterdam, 2002.

[Mou10] E Moulay. Some properties of Lyapunov function sets. *Proc. Am. Math. Soc.*, 138(11):4067–4073, 2010.

[Mro94] M Mrozek. Shape index and other indices of Conley type for local maps on locally compact Hausdorff spaces. *Fundam. Math.*, 145(1):15–37, 1994.

[MRS00] M Mrozek, J Reineck, and R Srzednicki. The Conley index over a base. *Trans. Am. Math. Soc.*, 352(9):4171–4194, 2000.

[Mun14] J Munkres. *Topology.* Pearson Education Limited, Essex, 2014.

[Obe87] R J Ober. Topology of the set of asymptotically stable minimal systems. *Int. J. Control*, 46(1):263–280, 1987.

[PCC$^+$15] M C Priess, R Conway, J Choi, J M Popovich, and C Radcliffe. Solutions to the inverse LQR problem with application to biological systems analysis. *IEEE Trans. Control Syst. Technol.*, 23(2):770–777, 2015.

[PLS80] T Pappas, A Laub, and N Sandell. On the numerical solution of the discrete-time algebraic Riccati equation. *IEEE T. Automat. Contr.*, 25(4):631–641, 1980.

[Pol86a] J W Polderman. A note on the structure of two subsets of the parameter space in adaptive control problems. *Syst. Control Lett.*, (7):25–34, 1986.

[Pol86b] J W Polderman. On the necessity of identifying the true parameter in adaptive LQ control. *Syst. Control Lett..*, (8):87–91, 1986.

[Pol87] J W Polderman. *Adaptive Control & Identification: Conflict or Conflux.* PhD thesis, University of Groningen, 1987.

[PPR04] S Prajna, P A Parrilo, and A Rantzer. Nonlinear control synthesis by convex optimization. *IEEE T. Automat. Contr.*, 49(2):310–314, 2004.

[Rei91] J F Reineck. Continuation to gradient flows. *Duke Math. J.*, 64(2):261–269, 1991.

[SA96] R Sepulchre and D Aeyels. Homogeneous Lyapunov functions and necessary conditions for stabilization. *Math. Control. Signals, Syst.*, 9(1):34–58, 1996.

[SA23] M Sassano and A Astolfi. On the role of convexity/concavity in vector fields, flows and stability/stabilizability. *IEEE T. Automat. Contr.*, 2023.

[Son89] E D Sontag. A ‘universal’construction of Artstein’s theorem on nonlinear stabilization. *Syst. Control Lett.*, 13(2):117–123, 1989.

[Son98] E D Sontag. *Mathematical control theory: deterministic finite dimensional systems.* Springer, New York, 1998.

[TD12] S Trimpe and R D’Andrea. The balancing cube: A dynamic sculpture as test bed for distributed estimation and control. *IEEE Contr. Syst. Mag.*, 32(6):48–75, 2012.

[TSH12] H L Trentelman, A A Stoorvogel, and M Hautus. *Control Theory for Linear Systems.* Springer Science & Business Media, London, 2012.

[Udr13] C Udriste. *Convex functions and optimization methods on Riemannian manifolds*, volume 297. Springer Science & Business Media, Dordrecht, 2013.

[vS94] J H van Schuppen. Tuning of Gaussian stochastic control systems. *IEEE T. Automat. Contr.*, 39(11):2178–2190, 1994.

[Whi90] P Whittle. *Risk-sensitive Optimal Control.* Wiley, Hoboken, 1990.

[WS20] P M Wensing and J.-J Slotine. Beyond convexity—contraction and global convergence of gradient descent. *Plos one*, 15(8):e0236661, 2020.

[Yan20] I Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE T. Automat. Contr.*, 66(8):3863–3870, 2020.

[Zab89] J Zabczyk. Some comments on stabilizability. *Appl. Math. Opt.*, 19(1):1–9, 1989.

[Zad63] L Zadeh. Optimality and non-scalar-valued performance criteria. *IEEE T. Automat. Contr.*, 8(1):59–60, 1963.

5
Numerical stability in optimization

> “*It has been written that the shortest and best way between two truths of the real domain often passes through the imaginary one.*”
>
> — Hadamard [Had45, p. 123].

Besides classical dynamical systems topics, like stability of attractors and structural stability of a system itself, as discussed elsewhere in the thesis, we look in this chapter at stability of the implementation of certain dynamical systems. Specifically, we look at zeroth-order optimization algorithms, a widely applicable class of algorithms understood as discrete-time dynamical systems on Euclidean space.

Zeroth-order optimization methods are developed to overcome the practical hurdle of having knowledge of explicit derivatives. Instead, these schemes work with merely access to noisy functions evaluations. One of the predominant approaches is to mimic first-order methods by means of some gradient estimator. The theoretical limitations are well-understood, yet, as most of these methods rely on finite-differencing for shrinking differences, numerical cancellation can be catastrophic. The numerical community developed an efficient method to overcome this by passing to the complex domain. In this chapter we show how this “*imaginary*” method can be adopted for optimization.

5.1 Introduction to imaginary zeroth-order optimization

In this fist section we introduce our framework, in the next section we highlight how to cope with noise and further details.

5.1.1 Introduction

We study optimization problems of the form

$$\underset{x\in\mathcal{X}}{\text{minimize}} \quad f(x), \tag{5.1.1}$$

where $f : \mathcal{D} \to \mathbb{R}$ is a real analytic objective function defined on an open set $\mathcal{D} \subseteq \mathbb{R}^n$, and $\mathcal{X} \subseteq \mathcal{D}$ is a non-empty closed feasible set. Throughout, we assume that problem (5.1.1) admits a global minimizer $x^\star$ (if not, restrict to an appropriate neighbourhood) and that the objective function f can only be accessed through a deterministic (Section 5.2 considers noise) zeroth-order oracle, which outputs function evaluations at prescribed test points. Under this premise, we aim to develop optimization algorithms that generate a (potentially randomized) sequence of iterates $x_1, x_2, \ldots, x_K \in \mathcal{X}$ approximating $x^\star$. As they only have access to a zeroth-order oracle, these algorithms fall under the umbrella of *zeroth-order optimization*, *derivative-free optimization* or, more broadly, *black-box optimization*, see, *e.g.,* [AH17]. As we will explain below and in contrast to (almost) all prior work on zeroth-order optimization, we will assume that our zeroth-order oracle also accepts *complex* inputs beyond $\mathcal{D}$.

Zeroth-order optimization algorithms are needed when problem (5.1.1) cannot be addressed with first- or higher-order methods. This is the case when there is no simple closed-form expression for f and its partial derivatives or when evaluating the gradient of f is expensive. In simulation-based optimization, for example, the function f can be evaluated via offline- or online simulation methods, but its gradient is commonly inaccessible. Zeroth-order optimization algorithms can also be used for addressing minimax, bandit or reinforcement learning problems, and they lend themselves for hyperparameter tuning in supervised learning [Spa05, CSV09, NS17]. As they can only access function values, zeroth-order optimization methods are inevitably somewhat crude. This simplicity is both a curse and a blessing. On the one hand, it has a detrimental impact on the algorithms' ability to converge to local minima, on the other hand—and this requires further formalization [Sch22], it may enable zeroth-order methods to escape from saddle points and thus makes them attractive for non-convex optimization.

Zeroth-order optimization algorithms can be categorized into direct search methods, model-based methods and random search methods [LMW19]. Direct search methods evaluate the objective function at a set of trial points without the goal of approximating the gradient. A representative example of a direct search method is the popular Nelder–Mead algorithm [NM65]. Model-based methods use zeroth-order information acquired in previous iterations to calibrate a C^r-smooth model for some $r \in \mathbb{Z}_{\geq 0}$ that approximates the black-box function f locally around the current iterate and then construct the next iterate via r^{th}-order optimization methods. These approaches typically attain a higher accuracy than the direct and random search methods, and they have the additional advantage that function evaluations can be re-used. In general, however, they require at least $O(n)$ function evaluations in each iteration to construct a well-defined local model [BCCS21]. Examples of commonly used models include polynomial models, interpolation models and regression models [LMW19]. In contrast to model-based methods, random search methods estimate the gradients of f at the iterates directly from finitely many function evaluations and use the resulting estimators as surrogates for the actual gradients in a first-order algorithm. More precisely, random search methods typically approximate f by a smooth function f_δ that is close to f for small δ and construct an unbiased estimator $g_\delta(x)$ for $\nabla f_\delta(x)$ by sampling f at test points in the vicinity of x [FKM04, NS17]. For many popular approximations f_δ there exists $p \geq 1$ such that $\|\nabla f_\delta(x) - \nabla f(x)\| \leq O(\delta^p)$. In analogy to the model-based methods, $g_\delta(x)$ can thus be used as a surrogate for the actual gradient in a first-order algorithm. A striking advantage of these random search methods over model-based methods is that the computation of $g_\delta(x)$ requires only $O(1)$ function evaluations, yet at the expense of weaker approximation guarantees [LCK$^+$20, BCCS21, Sch22]. In principle, the approximation quality of the surrogate gradients (and therefore also the convergence rate of the first-order method at hand) can be improved by reducing the smoothing parameter δ. As $g_\delta(x)$ is often reminiscent of a difference quotient with increment δ, however, its evaluation is plagued by numerical cancellation. This means that if δ drops below a certain threshold, innocent round-off errors in the evaluations of f have a dramatic impact on the evaluations of g_δ. Hence, the actual numerical performance of a random search zeroth-order algorithm may fall significantly short of its theoretical performance [SXON22], however, the awareness for this phenomenon seems to be somewhat lacking.

Inspired by techniques for numerically differentiating analytic functions, we propose here a new smoothed approximation f_δ as well as a corresponding stochastic gradient estimator g_δ that can be evaluated rapidly and faithfully for arbitrarily small values of δ without suffering from cancellation effects. Integrating the new estimator into the gradient-descent-type algorithm

$$x_{k+1} \leftarrow x_k - \mu_k \cdot g_{\delta_k}(x_k) \tag{5.1.2}$$

with adaptive stepsize μ_k and smoothing parameter δ_k gives rise to new randomized zeroth-order algorithms. The performance of such algorithms is measured by the decay rate of the regret $R_K = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[f(x_k) - f(x^\star)\right]$ as K grows. Here, $x^\star$ is a global minimizer, and the expectation $\mathbb{E}[\cdot]$ is taken with respect to the randomness introduced by the algorithm. Note that if f is convex, then, Jensen's inequality ensures that the expected suboptimality gap (or expected optimization error) of the *averaged* iterate $\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$ satisfies $\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq R_K$. The main goal of this section is to understand how R_K scales with the total number K of iterations and with critical problem parameters such as the dimension of x or Lipschitz moduli of f. Whenever possible (*e.g.,* when f is strongly convex), we also analyze the expected suboptimality gap $\mathbb{E}[f(x_K) - f(x^\star)]$ of the last iterate x_K. The scaling behavior of R_K with respect to K reflects the algorithm's *convergence rate.* We will show that algorithms of the form (5.1.2) equipped with the new gradient estimator offer provable convergence rates, are numerically stable, and empirically outperform algorithms that exploit existing smoothed approximations both in terms of accuracy and runtime.

Before continuing, let us briefly comment on the emphasis on *gradient descent* in optimization.

Remark 5.1.1 (On gradient descent)**.** Besides being simple to implement, there is a strong link with control and stability. Suppose that $f \in C^1(\mathbb{R}^n)$ is strongly convex (or at least such that critical points are global minimizers, *i.e.*, $\nabla f(x) = 0 \iff x = x^\star$). In this case, we can construct a function $V(x) = f(x) - f(x^\star) \geq 0$ and study convergence of the negative gradient flow $\dot{x} = -\nabla f(x)$. Indeed, V acts as a Lyapunov function since f is strongly convex and $\langle \nabla V(x), -\nabla f(x) \rangle = -\|\nabla f(x)\|_2^2$, which is strictly negative for all $x \neq x^\star$. Hence, the gradient flow converges globally, in a stable manner, to $x^\star$. This is solid motivation to look for an efficient discretization of such a flow and to see how far this continuous viewpoint extends. That the gradient points in the direction of steepest descent is just static motivation.

We use standard notation, but we emphasize the use of complex numbers throughout this chapter.
Notation: We reserve the symbol $i = \sqrt{-1}$ for the imaginary unit. The real and imaginary parts of a complex number $z = a + ib$ for $a, b \in \mathbb{R}$ are denoted by $\Re(z) = a$ and $\Im(z) = b$. In addition, V_n stands for the volume of the closed unit ball $\mathbb{B}^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$, and S_{n-1} stands for the surface area of the unit sphere $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. The family of all r times continuously differentiable real-valued functions on an open set $\mathcal{D} \subseteq \mathbb{R}^n$ is simply denoted by $C^r(\mathcal{D})$ instead of $C^r(\mathcal{D}; \mathbb{R})$, and similarly, the family of all real analytic functions on $\mathcal{D}$ is denoted by $C^\omega(\mathcal{D})$.

More on zeroth-order optimization

Given a deterministic zeroth-order oracle, one could address problem (5.1.1) with a gradient-descent algorithm that approximates the gradient of f with a vector of coordinate-wise finite differences [KW52, KY03, Spa05, BCCS21]. The corresponding finite-difference methods for zeroth-order optimization are reminiscent of inexact gradient methods [d'A08, DGN14]. Maybe surprisingly, there is merit in using *stochastic* gradient estimates even if a *deterministic* zeroth-order oracle is available [NS17]. The randomness not only helps to penetrate previously unexplored parts of the feasible set but also simplifies the convergence analysis. Specifically, if f is convex, then it is often easy to show that $f(x_k)$ converges *in expectation* to the global minimum $f(x^\star)$ [NS17].

Zeroth-order optimization algorithms that mimic gradient descent algorithms can be categorized by the number of oracle calls needed for a single evaluation of the gradient estimator. The most efficient algorithms of this kind make do with one single oracle call. Arguably the first treatise on zeroth-order optimization with a random single-point gradient estimator appeared in [NY83, Sec. 9.3], where the objective function $f(x)$ is approximated by the smoothed function $f_\delta(x) = V_n^{-1} \int_{\mathbb{B}^n} f(x + \delta y)\, \mathrm{d}y$, and the degree of smoothing is controlled by the parameter $\delta > 0$. By leveraging the dominated convergence theorem and the classical divergence theorem, one can then derive the following integral representation for the gradient of $f_\delta(x)$,

$$\nabla f_\delta(x) = \frac{n}{\delta} \int_{\mathbb{S}^{n-1}} f(x + \delta y) y \, \sigma(\mathrm{d}y),$$

where σ represents the uniform distribution on the unit sphere $\mathbb{S}^{n-1}$ (see also the proof of Proposition 5.1.8). Hence, the gradient of the smoothed function f_δ admits the unbiased stochastic estimator

$$g_\delta(x) = \frac{n}{\delta} f(x + \delta y) y \quad \text{with} \quad y \sim \sigma, \tag{5.1.3}$$

which can be accessed with merely a single function evaluation. Stochastic gradient estimators of this kind have been used as surrogate gradients in gradient descent algorithms, for example, in the context of bandit problems [FKM04]. However, as already pointed out in [NY83], the variance of the gradient estimator (5.1.3) is of the order $O(n^2/\delta^2)$ for small δ even if the function f is constant. This is inconvenient because a smaller δ reduces the bias of f_δ *vis-à-vis* f. To improve this bias-variance trade-off, it has been proposed to subtract from $g_\delta(x)$ the control variate $n\delta^{-1}f(x)y$, which has a vanishing mean but is strongly correlated with $g_\delta(x)$ and therefore leads to a variance reduction [ADX10, NS17]. The resulting unbiased stochastic gradient is representable as

$$g'_\delta(x) = \frac{n}{\delta}\left(f(x+\delta y) - f(x)\right) y \quad \text{with} \quad y \sim \sigma, \tag{5.1.4}$$

which is reminiscent of a directional derivative and can be accessed via two function evaluations. Now, under mild conditions on f, the variance of $g'_\delta(x)$ remains bounded as δ tends to 0. If we aim to solve problem (5.1.1) to an arbitrary precision, however, the smoothing parameter δ needs to be made arbitrarily small, in which case $f(x+\delta y)$ and $f(x)$ become numerically indistinguishable. Subtractive cancellation therefore makes it impossible to evaluate estimators of the form (5.1.4) to an arbitrarily high precision. This phenomenon is exacerbated when the function evaluations are noisy, which commonly happens in simulation-based optimization [LZH+16]. Generalized stochastic gradient estimators requiring multiple function evaluations are discussed in [HL14], and in [DJWW15, LLZ21] various optimality properties of zeroth-order schemes with multi-point gradient estimators are discussed.

Stochastic gradient estimators akin to (5.1.4) with u following a Gaussian instead of a uniform distribution are studied in [NS17]. The corresponding stochastic gradient descent algorithms may converge as fast as $O(n/K)$ if f is convex and has a Lipschitz continuous gradient, but they are typically $O(n)$ times slower than their deterministic counterparts. Convergence can be accelerated by leveraging central finite-difference schemes or by adding random perturbations to the gradient estimators [DJWW15, GKL+17, Sha17]. Local convergence results for nonconvex optimization problems are investigated in [GL13], and second-order algorithms similar to (5.1.2), which use a Stein identity to estimate the Hessian matrix, are envisioned in [BG22]. Lower bounds on the convergence rate of algorithm (5.1.2) are established in [AWBR09, JNR12, Sha13].

Another stream of related research investigates zeroth-order optimization methods that have only access to a *stochastic* zeroth-order oracle, which returns function evaluations contaminated by noise. The performance of these methods critically depends on the smoothness properties of f. Indeed, the higher its degree of smoothness, the more terms in the Taylor series of f can be effectively averaged out [PT90]. Improved convergence results for zeroth-order optimization methods under convexity assumptions are derived in [BP16, APT20, NG22]. When function evaluations are noisy, the smoothing parameter δ controls a bias-variance tradeoff. Indeed, reducing δ reduces the bias introduced by smoothing f, while increasing δ reduces the variance of the gradient estimator induced by the noisy oracle, which scales as $1/\delta$ for small δ. The variance can be further reduced by mini-batching [JWZL19]. The impact of exact line search methods and adaptive stepsize selection schemes is discussed in [SMG13, BCS21]. Better stepsize rules are available if f displays a latent low-dimensional structure [GKK+20].

Generalized zeroth-order methods for optimizing functions defined on Riemannian manifolds are proposed in [LBM22], and algorithms that have only access to a *comparison oracle*, which is less informative than a zeroth-order oracle, are investigated in [CMYZ22].

For comprehensive surveys of zeroth-order optimization and derivative-free optimization we refer to [LMW19, LCK$^+$20]. Abstract zeroth-order methods for convex optimization are studied in [HPGS16]. The minimax regret bounds derived in that work reveal the importance of having control over the randomness of the zeroth-order oracle. Accordingly, most existing methods rely on the assumption that the noise distribution is light-tailed. In contrast, if the zeroth-order oracle is affected by adversarial noise, then optimization is easily obstructed [SV15, Thm 3.1].

Contributions

Most existing zeroth-order schemes approximate the gradient of f in a way that makes them susceptible to numerical instability. For example, if $f \in C^1(\mathbb{R})$ is Lipschitz continuous with Lipschitz constant L, then, in theory, the finite-difference approximation $(f(x+\delta) - f(x))/\delta$ converges to $\partial_x f(x)$ as $\delta > 0$ tends to zero. In practice, however, f can only be evaluated to within machine precision, which means that $f(x+\delta)$ and $f(x)$ become indistinguishable for sufficiently small δ. More precisely, as f is Lipschitz continuous, we have $|f(x+\delta) - f(x)| \leq L \cdot |\delta|$, and thus cancellation errors[1] are prone to occur when $L \cdot |\delta|$ approaches machine precision [Ove01, Sec. 11]. Other gradient estimators that are based on multiple function evaluations or that involve interpolation schemes suffer from similar cancellation errors. Nevertheless, the convergence guarantees of the corresponding zeroth-order methods require that the smoothing parameter δ must be driven to zero. For example, [APT20, Thm. 3.1] establishes regret bounds under the assumption that the smoothing parameter of a multi-point estimator scales as $\delta_k = O(1/\sqrt{k})$.

The randomized gradient estimator (5.1.3) avoids cancellation errors because it requires only one single function evaluation—an attractive feature that has, to the best of our knowledge, gone largely unnoticed to date. However, as pointed out earlier, the variance of this estimator diverges as δ decays, which leads to suboptimal convergence rates. In this section we propose a numerically stable gradient estimator that enables competitive convergence rates and is immune to cancellation errors. More precisely, we will use complex arithmetic to construct a one-point estimator akin to (5.1.3) that offers similar approximation and convergence guarantees as state-of-the-art two-point estimators. Maybe surprisingly, we will see that computing this new estimator is not significantly more expensive than evaluating (5.1.3). Our results critically rely on the assumption that the objective function f is real analytic on its domain $\mathcal{D}$. Recall that f is real analytic if it locally coincides with its multivariate Taylor series. We emphasize that real analyticity does not imply β^{th}-order smoothness for some $\beta \in \mathbb{Z}_{>0}$ in the sense of [BP16, Sec. 1.1], which means that f is almost surely $\beta - 1$ times differentiable and that the $(\beta-1)^{\text{th}}$-order term of its Taylor series is globally Lipschitz continuous. We will recall that f can be extended to a complex analytic function $f : \Omega \to \mathbb{C}$ defined on some open set $\Omega \subseteq \mathbb{C}^n$ that covers $\mathcal{D} \subseteq \mathbb{R}^n$. By slight abuse of notation, this extension is also denoted by f. Given an

[1] It is true that cancellation errors are not *always* catastrophic, the typical example being $x + y - z$ with $y \approx z$ and $|x| \gg y, z$ [Hig02]. However, since we just subtract two numbers, the cancellation will be catastrophic in that signs flip and so forth.

oracle that evaluates f at any query point in Ω, we will devise new zeroth-order methods that combine the superior convergence rates and low variances of multi-point schemes reported in [DJWW15, LLZ21] with the numerical robustness of single-point approaches.

We now use $R = \|x_1 - x^\star\|_2$ to denote the distance from the initial iterate x_1 to a[2] minimizer $x^\star$ and $F = f(x_1) - f(x^\star)$ to denote the suboptimality of x_1. Assuming that the objective function f is real analytic and has an L-Lipschitz continuous gradient, we will devise zeroth-order methods that offer the following convergence guarantees. If (5.1.1) represents a convex optimization problem with $x^\star \in \operatorname{int}(\mathcal{X})$, then our algorithm's regret decays as $O(nLR^2/K)$ with the iteration counter K. If, in addition, f is τ-strongly convex for some $\tau > 0$, then the expected suboptimality decays at the linear rate $O((1 - \tau/(4nL))^K LR^2)$. If (5.1.1) represents a non-convex optimization problem, finally, we establish local convergence to a stationary point and prove that $\min_{k\in[K]} \mathbb{E}[\|\nabla f(x_k)\|_2^2] \leq O(nLF/K)$. All of these convergence rates are qualitatively equivalent to the respective rates reported in [NS17, Thm. 8], and they are sharper than the rates provided in [APT20, Sec. 3] in the noise-free limit. The latter rely on higher-order smoothness properties of f but do not require f to be analytic. The key difference to all existing methods is that we can drive the smoothing parameter to 0, *e.g.*, as $\delta_k = \delta/k$, without risking numerical instability.

As highlighted in the recent survey article [LMW19], an important open question in zeroth-order optimization is whether single-point estimators enable equally fast convergence rates as multi-point estimators. The desire to reap the benefits of multi-point estimators at the computational cost of using single-point estimators has inspired multi-point estimators with memory, which only require a single new function evaluation per call [ZZJZ22]. However, this endeavor has not yet led to algorithms that improve upon the theoretical and empirical performance of the state-of-the-art methods in [NS17]. Filtering techniques inspired by ideas from extremum seeking control can be leveraged to improve the convergence rates obtained in [ZZJZ22] to $O(n/K^{2/3})$ [CTL22]. However, this rate is still inferior to the ones reported in [NS17]. To our best knowledge, we propose here the first single-point zeroth-order algorithm that enjoys the same convergence rates as the multi-point methods in [NS17] but often outperforms these methods in experiments. The price we pay for these benefits is the assumption that there exists a zeroth-order oracle accepting complex queries. This assumption is restrictive as it rules out oracles that depend on performing a physical experiment or timing a computational run etc. Nevertheless, the method can be shown to excel in the context of simulation-based optimization [JYK21, Sec. 7].

5.1.2 Elements from complex analysis

Before detailing our algorithms, we highlight a few elements from complex analysis.

Multivariate complex analysis

For any multi-index $\alpha \in \mathbb{Z}^n_{\geq 0}$ and vector $x \in \mathbb{R}^n$, we use x^α as a shorthand for the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, and we denote the degree of x^α by $|\alpha| = \sum_{i=1}^n \alpha_i$. The factorial of α is

[2] As becomes clear later on, we assume that $x^\star$ is unique for simplicity, this is not necessary.

defined as $\alpha! = \prod_{i=1}^n \alpha_i!$, and ∂_x^α stands for the higher-order partial derivative $\partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$. Multi-index notation facilitates a formal definition of real analytic functions.

Definition 5.1.1 (Real analytic function)**.** *The function $f : \mathcal{D} \to \mathbb{R}$ is real analytic on $\mathcal{D} \subseteq \mathbb{R}^n$, denoted $f \in C^\omega(\mathcal{D})$, if for every $x' \in \mathcal{D}$ there exist $f_\alpha \in \mathbb{R}$, $\alpha \in \mathbb{Z}_{\geq 0}^n$, and an open set $U \subseteq \mathcal{D}$ containing x' such that*

$$f(x) = \textstyle\sum_{\alpha \in \mathbb{Z}_{\geq 0}^n} f_\alpha \cdot (x - x')^\alpha \quad \forall x \in U. \tag{5.1.5}$$

Whenever we write that a series has a finite value, we mean that it converges absolutely, that is, it converges when the summands of the series are replaced by their absolute values. In this case any ordering of the summands results in the same value.

One can show that any real analytic function is infinitely differentiable and that the coefficients of its power series are given by $f_\alpha = \frac{1}{\alpha!} \partial_x^\alpha f(x')$ for every $\alpha \in \mathbb{Z}_{\geq 0}^n$. This implies that the power series is unique and coincides with the multivariate Taylor series of f around x' [KP02, Sec. 2.2]. We will now recall that every real analytic function admits a complex analytic extension.

Definition 5.1.2 (Complex analytic function)**.** *The function $f : \Omega \to \mathbb{C}$ is complex analytic on $\Omega \subseteq \mathbb{C}^n$, denoted $f \in H(\Omega)$, if for every $z' \in \Omega$ there exist $f_\alpha \in \mathbb{C}$, $\alpha \in \mathbb{Z}_{\geq 0}^n$, and an open set $U \subseteq \Omega$ containing z' such that*

$$f(z) = \textstyle\sum_{\alpha \in \mathbb{Z}_{\geq 0}^n} f_\alpha \cdot (z - z')^\alpha \quad \forall z \in U. \tag{5.1.6}$$

Complex analytic functions are intimately related to holomorphic functions.

Definition 5.1.3 (Holomorphic function)**.** *The function $f : \Omega \to \mathbb{C}$ is holomorphic on an open set $\Omega \subseteq \mathbb{C}^n$ if the complex partial derivatives $\partial_{z_j} f$, $j = 1, \ldots, n$, exist and are finite at every $z \in \Omega$.*

The requirement that Ω is open is essential, f may fail to be holomorphic on a neighborhood of a point z even if it is complex differentiable at z. For example, the Cauchy-Riemann equations reviewed below imply that $f(z) = |z|^3$ is complex differentiable at $z = 0$ but fails to be complex differentiable on any neighborhood of 0. Holomorphic functions are in fact infinitely often differentiable [Leb20, Prop. 1.1.3]. Moreover, a function is holomorphic if and only if it is complex analytic [Leb20, Thm. 1.2.1].

It is common to identify any complex vector $z \in \mathbb{C}^n$ with two real vectors $x, y \in \mathbb{R}^n$ through $z = x + iy$ ($\mathbb{C}^n \simeq \mathbb{R}^{2n}$). Similarly, we may identify any complex function $f : \mathbb{C}^n \to \mathbb{C}$ with two real functions $u : \mathbb{R}^n \to \mathbb{R}$ and $v : \mathbb{R}^n \to \mathbb{R}$ through the relation $f(x + iy) = u(x, y) + iv(x, y)$. Clearly, u and v inherit the differentiability properties of f and vice versa. In particular, one can show that if f is holomorphic, then the partial derivatives of u and v exist and satisfy the multivariate *Cauchy-Riemann equations.*

Theorem 5.1.4 (Multivariate Cauchy-Riemann equations)**.** *If $f(x + iy) = u(x, y) + iv(x, y)$ is a holomorphic function on an open set $\Omega \subseteq \mathbb{C}^n$, then the multivariate Cauchy-Riemann equations*

$$\partial_{x_j} u(x, y) = \partial_{y_j} v(x, y) \quad \textit{and} \quad -\partial_{x_j} v(x, y) = \partial_{y_j} u(x, y) \quad \forall j = 1, \ldots, n \tag{5.1.7}$$

hold for all $x, y \in \mathbb{R}^n$ with $x + iy \in \Omega$.

Theorem 5.1.4 is a standard result in complex analysis; see, *e.g.*, [Rud87, Thm. 11.2] or [Kra00]. Nevertheless, we provide here a short proof to keep this chapter self-contained.

Proof of Theorem 5.1.4. We use e_j to denote the j^{th} standard basis vector in $\mathbb{R}^n$. By the definition of the complex partial derivative, for any $z \in \Omega$ we have

$$\partial_{z_j} f(z) = \lim_{\delta \in \mathbb{C},\, \delta \to 0} \frac{1}{\delta}(f(z+\delta e_j) - f(z)),$$

where the limit exists and is independent of how $\delta \in \mathbb{C}$ converges to 0 because f is holomorphic on Ω. In particular, δ may converge to 0 along the real or the imaginary axis without affecting the result. Using our conventions that $z = x + iy \in \Omega$ and $f(x+iy) = u(x,y) + iv(x,y)$, we thus have

$$\begin{aligned}\partial_{x_j}(u(x,y) + iv(x,y)) &= \lim_{\delta \in \mathbb{R},\, \delta \to 0} \frac{f((x+\delta e_j) + iy) - f(x+iy)}{\delta} \\ &= \lim_{\delta \in \mathbb{R},\, \delta \to 0} \frac{f(x + i(y+\delta e_j)) - f(x+iy)}{i\delta} \\ &= \frac{1}{i}\partial_{y_j}(u(x,y) + iv(x,y))\end{aligned}$$

for all $x, y \in \mathbb{R}^n$, where the second equality holds because both limits are equal to $\partial_{z_j} f(z)$. Matching the real and imaginary parts of the above equations yields (5.1.7). □

Under additional assumptions one can further show that the Cauchy-Riemann equations imply that f is holomorphic [GM78]. However, this reverse implication will not be needed in this thesis. The following lemma based on [Kra00, Sec. 2.3] establishes that any real analytic function defined on an open set $\mathcal{D} \subseteq \mathbb{R}^n$ admits a complex analytic extension defined on an open set $\Omega \subseteq \mathbb{C}^n$ that covers $\mathcal{D}$.

Lemma 5.1.5 (Complex analytic extensions)**.** *If $f \in C^\omega(\mathcal{D})$, then there exists an open set $\Omega \subseteq \mathbb{C}^n$ and a complex analytic function $g \in H(\Omega)$ such that $\mathcal{D} \subseteq \Omega$ and $f(x) = g(x)$ for every $x \in \mathcal{D}$, with $\mathcal{D}$ understood as embedded in $\mathbb{C}^n$.*

Proof. Select any $x' \in \mathcal{D}$. As $f \in C^\omega(\mathcal{D})$, there exists a neighborhood $U \subseteq \mathcal{D}$ of x' such that f admits a power series representation of the form (5.1.6) on U. Also, as U is open, there exists $x \in U$ with $r_j = |x_j - x'_j| > 0$ for every $j = 1, \ldots, n$. By Abel's lemma [Kra00, Prop. 2.3.4], the power series (5.1.6) extended to $\mathbb{C}^n$ is thus guaranteed to converge on the open polydisc $\Delta(x') = \{z \in \mathbb{C}^n : |z_j - x'_j| < r_j \; \forall j = 1, \ldots, n\}$. This reasoning implies that f extends locally around x' to a complex analytic function, which we henceforth denote as $g_{x'}$. It remains to be shown that the local extensions corresponding to different reference points $x' \in \mathcal{D}$ are consistent. To this end, select any $x', x'' \in \mathcal{D}$ such that the polydiscs $\Delta(x')$ and $\Delta(x'')$ overlap. We need to prove that $g_{x'}$ and $g_{x''}$ coincide on the open convex set $\Delta = \Delta(x') \cap \Delta(x'')$, which has a non-empty intersection with $\mathbb{R}^n$. For ease of exposition, we will equivalently prove that the holomorphic function $h = g_{x'} - g_{x''}$ vanishes on Δ. We first notice that h vanishes on $\Delta \cap \mathbb{R}^n$ because $g_{x'}$ and $g_{x''}$ are constructed to coincide with f on $\Delta \cap \mathbb{R}^n$. This implies that $\partial_{z_j} h = \partial_{x_j} h = 0$ on $\Delta \cap \mathbb{R}^n$, where the first equality follows from standard arguments familiar from the proof of Theorem 5.1.4. As any partial derivative of a holomorphic function is also holomorphic, one can use induction to show that all higher-order partial derivatives of h must vanish on $\Delta \cap \mathbb{R}^n$. Hence, the Taylor series of h around any reference point in $\Delta \cap \mathbb{R}^n$ vanishes, too. We have thus shown that $h = 0$ on an open subset of Δ. By standard results in complex analysis, this implies that h vanishes throughout Δ; see, *e.g.*, [Leb20, Thm. 1.2.2]. In summary, this reasoning confirms that all local complex analytic extensions $g_{x'}$, $x' \in \mathcal{D}$, of f are consistent and thus coincide with a complex analytic function g defined on the open set $\Omega = \cup_{x' \in \mathcal{D}} \Delta(x')$. This observation completes the proof. □

Lemma 5.1.5 implies that the complex extension of a real analytic function is unique. For example, $f(x) = \log(x)$ is real analytic on the positive real line. Representing $z = re^{i\theta} \in \mathbb{C}$ in polar form with $r \geq 0$ and $\theta \in (-\pi, \pi]$, the complex logarithm has countably many branches, that is, $\log(z)$ can be defined as $g_k(z) = \log(r) + i(\theta + 2\pi k)$ for any $k \in \mathbb{Z}$. However, only the branch g_0 corresponding to $k = 0$ matches f on the positive reals. We will henceforth use the same symbol f to denote both the given real analytic function as well as its unique complex analytic extension g. We now explicitly derive the complex analytic extensions of a few simple univariate functions.

Example 5.1.2 (Complex analytic extensions)**.** The unique complex analytic extension of $f(x) = e^x$ is the entire function $g(z) = g(x + iy) = e^x(\cos(y) + i \sin(y))$. Similarly, the unique complex analytic extension of the even polynomial $f(x) = x^{2p}$ with $p \in \mathbb{Z}_{\geq 0}$ is the entire function

$$g(z) = g(x + iy) = \textstyle\sum_{k=0}^{p} (-1)^k \binom{2p}{2k} y^{2k} x^{2(p-k)} + i \sum_{k=0}^{p-1} (-1)^k \binom{2p}{2k+1} y^{2k+1} x^{2(p-k)-1}.$$

Finally, the unique solution $f(x)$ to the Lyapunov equation $f(x) = x^2 f(x) + 1$ parametrized by $x \in \mathbb{R}$ is real analytic on $\mathbb{R} \setminus \{1\}$. It admits the extension

$$g(z) = g(x + iy) = \frac{1 - x^2 + y^2 - 2ixy}{(1 - x^2 + y^2)^2 + 4x^2y^2},$$

which is analytic throughout $\mathbb{C} \setminus \{(1, 0)\}$.

The next example shows that the domain Ω of the complex analytic extension is not always representable as $\mathbb{R}^n + i \cdot (-\bar{\delta}, \bar{\delta})^n$ for some $\bar{\delta} > 0$ even if $\mathcal{D} = \mathbb{R}^n$.

Example 5.1.3 (Non-trivial extension)**.** Consider the function $f(x) = \sum_{k=1}^{\infty} 2^{-k}(1 + k^2(x - k)^2)^{-1} \in C^\omega(\mathbb{R})$, which admits a unique complex analytic extension with domain $\Omega = \mathbb{C} \setminus \{k + ik^{-1} : k \in \mathbb{Z}_{>0}\}$. In addition, f can be extended to a meromorphic function on $\mathbb{C}$ with countably many poles $k + ik^{-1}$, $k \in \mathbb{Z}_{>0}$. As these poles approach $\mathbb{R}$ arbitrarily closely, however, Ω cannot contain any strip of the form $\mathbb{R} \times i \cdot (-\bar{\delta}, \bar{\delta})$.

To avoid technical discussions of limited practical impact, we will from now on restrict attention to functions $f \in C^\omega(\mathcal{D})$ that admit a complex analytic extension to $\Omega = \mathcal{D} \times i \cdot (-\bar{\delta}, \bar{\delta})^n$ for some $\bar{\delta} > 0$. One can show that such an extension always exists if $f \in C^\omega(\mathbb{R}^n)$ and $\mathcal{D}$ is bounded or if f is *entire*, that is, if f has a globally convergent power series representation. The latter condition is restrictive, however, because it rules out simple functions such as $f(x) = 1/(1 + x^2)$. Provided there is no risk of confusion, we will sometimes call a real analytic function $f \in C^\omega$ and its complex analytic extension simply an *analytic* function.

Complex-step approximation

The *finite-difference* method [VVvGV23, Ch. 3] is arguably the most straightforward approach to numerical differentiation. It simply approximates the derivative of any sufficiently smooth function $f \in C^2(\mathbb{R})$ by a difference quotient. For example, the forward-difference method uses the approximation

$$\partial_x f(x) = \frac{1}{\delta}(f(x+\delta) - f(x)) + O(\delta). \tag{5.1.8}$$

The continuity of the second derivative of f allows for a precise formula for the $O(\delta)$ remainder term[3]. However, as explained earlier, typical finite difference methods suffer from cancellation errors when δ becomes small. The *complex-step* approximation proposed in [LM67, ST98] and further refined in [MSA03, ASM15, ASKG18] leverages ideas from complex analysis to approximate the derivative of any real analytic function $f \in C^\omega(\mathbb{R})$ on the basis of one single function evaluation only, thereby offering an elegant remedy for numerical cancellation. Denoting by u and v as usual the real and imaginary parts of the unique complex analytic extension of f, which exists thanks to Lemma 5.1.5, we observe that $\partial_x f(x)$ equals

$$\begin{aligned}\partial_x u(x,0) = \partial_y v(x,0) &= \lim_{\delta\to 0^+} \frac{v(x,\delta) - v(x,0)}{\delta} \\ &= \lim_{\delta\to 0^+} \frac{1}{\delta} v(x,\delta) = \lim_{\delta\to 0^+} \frac{1}{\delta}\Im\big(f(x+i\delta)\big),\end{aligned}$$

where the first and the fourth equalities hold because $f(x)$ must be a real number, which implies that $v(x,0) = 0$, while the second equality follows from the Cauchy-Riemann equations. The derivative $\partial_x f(x)$ can thus be approximated by the fraction $\Im(f(x+i\delta))/\delta$, which requires merely a single function evaluation. To estimate the approximation error, we consider the Taylor expansion

$$f(x+i\delta) = f(x) + \partial_x f(x) i\delta - \tfrac{1}{2}\partial_x^2 f(x)\delta^2 - \tfrac{1}{6}\partial_x^3 f(x) i\delta^3 + O(\delta^4) \tag{5.1.9}$$

of the unique complex analytic extension of f, which exists thanks to Lemma 5.1.5. Separating the real and imaginary parts of (5.1.9) then yields

$$f(x) = \Re(f(x+i\delta)) + O(\delta^2) \quad \text{and} \quad \partial_x f(x) = \delta^{-1}\Im\big(f(x+i\delta)\big) + O(\delta^2).$$

This reasoning shows that a single *complex* function evaluation $f(x+i\delta)$ is sufficient to approximate both $f(x)$ as well as $\partial_x f(x)$ without the risk of running into numerical instability caused by cancellation effects. In addition, the respective approximation errors scale quadratically with δ and are thus one order of magnitude smaller than the error incurred by (5.1.8). Note also that the complex-step approximation recovers the derivatives of quadratic functions *exactly* irrespective of the choice of δ. For example, if $f(x) = x^2$, then $\Im(f(x+i\delta))/\delta = 2x = \partial_x f(x)$. This insight suggests that the approximation is numerically robust for locally quadratic functions.

[3] Of course, f need not be C^2-smooth, this is merely an easy sufficient condition.

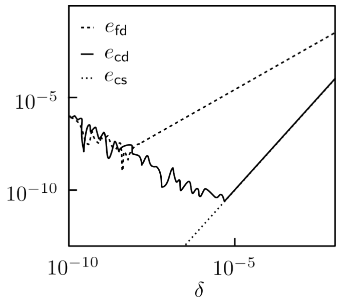


Figure 5.1: Comparison of the gradient estimators of Example 5.1.4 at $x = -1$, with $e_{\mathsf{i}} = |\nabla f(x) - f_{\mathsf{i}}(x, \delta)|$, figure adapted from the original Matlab figure [JYK21, Fig. 2.1].

The error of the complex-step approximation can be further reduced to $O(\delta^4)$ by enriching it with a finite-difference method [ASM15, HS23]. However, the resulting scheme requires multiple function evaluations and is thus again prone to cancellation errors. Unless time is expensive, the standard complex-step approximation therefore remains preferable. The complex-step approximation can also be generalized to handle matrix functions [AMH10] or to approximate higher-order derivatives [LRD12]. For a discussion concerning automatic differentiation (AD) and the ramifications of its underlying *ring-structure*, we refer to [JYK21, Sec. 2.3], the crux being that although AD is powerful, the representation of the objective is critical.

Being immune to cancellation effects, the complex-step approach offers approximations of almost arbitrary precision. For example, software by the UK's National Physical Laboratory is reported to use smoothing parameters as small as $\delta = 10^{-100}$ [CH04, p. 44]. The complex-step approach also emerges in various other domains. For example, it is successfully used in airfoil design [GWX17]. However, its potential for applications in optimization has not yet been fully exploited. Coordinate-wise complex-step approximations with noisy function evaluations show promising performance in line search experiments [NS18] but come without a rigorous convergence analysis. In addition, the complex-step approach is used to approximate the gradients and Hessians in deterministic Newton algorithms for blackbox optimization models [HS23]. The potential of leveraging complex arithmetic in mathematical optimization is also mentioned in [SW18, BBN19]. In this work we use the complex-step method to construct an estimator akin to (5.1.3) and provide a full regret analysis. Our approach is most closely related to the recent works [WS21, WZS21], which integrate the complex-step and *simultaneous perturbation stochastic approximation* (SPSA) [Spa92] into a gradient-descent algorithm and offer a rigorous *asymptotic* convergence theory. In contrast, we will derive convergence *rates* for a variety of zeroth-order optimization problems.

In optimization, the ability to certify that the gradient of an objective function is sufficiently small (*i.e.*, smaller than a prescribed tolerance) is crucial to detect local optima. The following example shows that, with the exception of the complex-step approach, standard numerical schemes to approximate gradients fail to offer such certificates—at least when a high precision is required.

Example 5.1.4 (Numerical stability of gradient estimators)**.** To showcase the power of the complex-step method and to expose the numerical difficulties encountered by finite-difference methods, we approximate the derivative of $f(x) = x^3$ at $x = -1$ via a forward-difference (fd), central-difference (cd) and complex-step (cs) method, that is, for small values of δ we compare

$$\begin{aligned} f_{\mathsf{fd}}(x,\delta) &= \delta^{-1}(f(x+\delta) - f(x)) \\ f_{\mathsf{cd}}(x,\delta) &= (2\delta)^{-1}(f(x+\delta) - f(x-\delta)) \\ f_{\mathsf{cs}}(x,\delta) &= \delta^{-1}\Im(f(x+i\delta)). \end{aligned}$$

Figure 5.1 visualizes the absolute approximation errors as a function of δ. We observe that f_{cd} and f_{cs} offer the same approximation quality and incur an error of $O(\delta^2)$ for all sufficiently large values of δ. However, only the complex-step approximation reaches machine precision ($\approx 10^{-16}$), whereas both finite-difference methods deteriorate below $\delta \approx 10^{-6}$ due to subtractive cancellation errors and suboptimal difference parameters[4]. As most existing zeroth-order optimization methods use finite-difference-based gradient estimators, we conclude that there is room for numerical improvements by leveraging complex arithmetic.

5.1.3 A smoothed complex-step approximation

We now use ideas from [NY83, NS17] to construct a new gradient estimator, which can be viewed as a complex-step generalization of the estimators proposed in [NY83, FKM04]. Our construction is based on the following assumption, which we assume to hold throughout the rest of this section.

Assumption 5.1.6 (Analytic extension). *The function $f : \mathcal{D} \to \mathbb{R}$ of problem (5.1.1) admits an analytic extension to the strip $\mathcal{D} \times i \cdot (-\bar{\delta}, \bar{\delta})^n$ for some $\bar{\delta} \in (0,1)$.*

Recall from Lemma 5.1.5 that f admits an analytic extension to some open set $\Omega \subseteq \mathbb{C}^n$ covering $\mathcal{D}$ whenever $f \in C^\omega(\mathcal{D})$. However, unless f is entire or $\mathcal{D}$ is bounded, Ω may not contain a strip of the form envisaged in Assumption 5.1.6. Hence, this assumption is *not* automatically satisfied for any real analytic function $f \in C^\omega(\mathcal{D})$. The requirement $\bar{\delta} \in (0,1)$ is unrestrictive and has the convenient consequence that $\delta^p \leq \delta^{p-1}$ for any $\delta \in (0,\bar{\delta})$ and $p \in \mathbb{Z}_{\geq 0}$. All subsequent results are based on a *smoothed complex-step approximation* f_δ of f, which is defined through

$$x \mapsto f_\delta(x) := V_n^{-1} \int_{\mathbb{B}^n} \Re\big(f(x+i\delta y)\big)\mathrm{d}y. \tag{5.1.10}$$

Here, the radius $\delta \in (0,\bar{\delta})$, of the ball used for averaging, represents a tuneable smoothing parameter. Given prior structural knowledge about f, one could replace $\mathbb{B}^n$ with a

[4] Regarding (sub)optimal difference parameters, let μ_{M} denote machine precision, then, we recall that for the forward-difference method one commonly selects $\delta = O(\sqrt{\mu_{\mathsf{M}}}) \approx 10^{-8}$ and for the central-difference method $\delta = O(\sqrt[3]{\mu_{\mathsf{M}}}) \approx 10^{-5}$, which follows from optimizing the numerical approximation error over δ. To be slightly more concrete, we provide an example in case of the forward-difference method. Consider the Taylor series of f around x, but suppose that the numerical evaluation of $f(x+\delta)$ comes with an error $O(\mu_{\mathsf{M}})$. Then, we have $f(x+\delta) = f(x) + \partial_x f(x)\delta + O(\mu_{\mathsf{M}} + \delta^2)$. Now, if we optimize the approximation error $|(f(x+\delta) - f(x))/\delta - \partial_x f(x)|$ over δ we recover the aforementioned.

different compact set [HL14, Jon21], see also Section 5.2. We emphasize that the integral in (5.1.10) is well-defined whenever $\delta \in (0, \bar{\delta})$, which ensures that f has no singularities in the integration domain. Next, we address the approximation quality of f_δ.

Proposition 5.1.7 (Approximation quality of f_δ)**.** *If $f \in C^{\omega,1}_{L_1}(\mathcal{D})$ satisfies Assumption 5.1.6, then for f_δ defined as in (5.1.10) and for any fixed $x \in \mathcal{D}$ and $\kappa \in (0,1)$ there exists $C_\kappa \geq 0$ with*

$$|f_\delta(x) - f(x)| \leq \tfrac{1}{2}L_1\delta^2 + C_\kappa\delta^4 \quad \forall \delta \in (0, \kappa\bar{\delta}].$$

Proof. By the definition of f_δ in (5.1.10), we have $|f_\delta(x) - f(x)| \leq V_n^{-1} \int_{\mathbb{B}^n} |\Re\big(f(x + i\delta y)\big) - f(x)| \mathrm{d}y$. The Taylor series of $f(x + i\delta y)$ around x then yields

$$\begin{aligned}\Re\left(f(x + i\delta y)\right) - f(x) &= \sum_{k=0}^{\infty} \frac{(-1)^k \delta^{2k}}{(2k)!} \sum_{|\alpha|=2k} \partial_x^\alpha f(x) y^\alpha - f(x) \\ &= -\tfrac{1}{2}\delta^2 \langle \nabla^2 f(x) y, y\rangle + \delta^4 R(y, \delta),\end{aligned}$$

where the real-valued remainder term $R(y, \delta)$ is continuous in $y \in \mathbb{B}^n$ and $\delta \in [0, \bar{\delta})$. Substituting the last expression into the above estimate and using (2.1.9), we obtain

$$|f_\delta(x) - f(x)| \leq V_n^{-1} \int_{\mathbb{B}^n} \tfrac{1}{2}\delta^2 L_1 + \delta^4 |R(y, \delta)| \mathrm{d}y \leq \tfrac{1}{2}\delta^2 L_1 + C_\kappa \delta^4 \quad \forall \delta \in (0, \kappa\bar{\delta}],$$

where the non-negative constant $C_\kappa = \max_{y \in \mathbb{B}^n} \max_{\delta \in [0, \kappa\bar{\delta}]} |R(y, \delta)|$ is finite due to continuity of $R(y, \delta)$ and compactness of $\mathbb{B}^n$ and $[0, \kappa\bar{\delta}]$ (the role of κ is to enforce compactness). Hence, the claim follows. □

Note that if f is affine, then $f_\delta = f$. Note also that $[0, \kappa\bar{\delta}]$ is a compact subset of the set $[0, \bar{\delta})$ on which $R(y, \delta)$ is continuous in δ and that $R(y, \delta)$ may be unbounded on $[0, \bar{\delta})$. The following proposition provides an integral representation for the gradient of f_δ. It extends [NY83, Sec. 9.3] and [FKM04, Lem. 1] to the realm of complex arithmetic. This is the main result of this chapter.

Proposition 5.1.8 (Gradient of the smoothed complex-step function)**.** *If $f \in C^\omega(\mathcal{D})$ satisfies Assumption 5.1.6, then f_δ defined as in (5.1.10) is differentiable, and we have*

$$\nabla f_\delta(x) = \frac{n}{\delta} \mathbb{E}_{y\sim\sigma} \left[\Im\left(f(x + i\delta y)\right) y\right] \quad \forall x \in \mathcal{D},\ \delta \in (0, \bar{\delta}), \tag{5.1.11}$$

where σ denotes the uniform distribution on $\mathbb{S}^{n-1}$.

Proof. Any function $g \in C^1(\mathbb{R}^n)$ and vector $w \in \mathbb{R}^n$ define a vector field $v(y) = g(y) \cdot w$. The divergence theorem [Lee13, Thm. 16.32] then implies that

$$\begin{aligned}\int_{\mathbb{B}^n} \langle w, \nabla g(y)\rangle \, \mathrm{d}y = \int_{\mathbb{B}^n} \mathrm{div}(v(y)) \, \mathrm{d}y &= S_{n-1} \int_{\mathbb{S}^{n-1}} \langle v(y), y\rangle \, \sigma(\mathrm{d}y) \\ &= S_{n-1} \int_{\mathbb{S}^{n-1}} g(y)\langle w, y\rangle \, \sigma(\mathrm{d}y),\end{aligned}$$

where the scaling factor S_{n-1} accounts for the fact that the uniform distribution σ is normalized on $\mathbb{S}^{n-1}$. Note also that the outward-pointing unit normal vector of $\mathbb{S}^{n-1}$ at any point $y \in \mathbb{S}^{n-1}$ is exactly y itself. As the above equation holds for all vectors $w \in \mathbb{R}^n$ and as both the leftmost and rightmost expressions are linear in w, their gradients (in w) must coincide. This reasoning implies that

$$\int_{\mathbb{B}^n} \nabla g(y)\, \mathrm{d}y = S_{n-1} \int_{\mathbb{S}^{n-1}} g(y) y\, \sigma(\mathrm{d}y). \tag{5.1.12}$$

We are now ready to prove (5.1.11) by generalizing tools developed in [NY83, FKM04] to the complex domain. Specifically, by the definition of f_δ in (5.1.10) we have

$$\begin{aligned}\nabla f_\delta(x) &= V_n^{-1} \int_{\mathbb{B}^n} \nabla_x \Re\left(f(x+i\delta y)\right) \mathrm{d}y \\ &= (V_n\delta)^{-1} \int_{\mathbb{B}^n} \nabla_y \Im\left(f(x+i\delta y)\right) \mathrm{d}y \\ &= S_{n-1}(V_n\delta)^{-1} \int_{\mathbb{S}^{n-1}} \Im\left(f(x+i\delta y)\right) y\, \sigma(\mathrm{d}y),\end{aligned}$$

where the interchange of the gradient and the integral in the first equality is permitted by the dominated convergence theorem, which applies because $\mathbb{B}^n$ is compact and because any continuously differentiable function on a compact set is Lipschitz continuous. The second equality is a direct consequence of the Cauchy-Riemann equations, and the third equality, finally, holds thanks to (5.1.12) with $g(y) = \Im\left(f(x+i\delta y)\right)$. At last, we observe that the volume of the unit ball and the surface of the unit sphere satisfy $V_n = \int_{\mathbb{B}^n} \mathrm{d}y = S_{n-1}\int_0^r r^{n-1}\mathrm{d}r = S_{n-1}/n \implies S_{n-1}/V_n = n$. Thus, the claim follows. □

Proposition 5.1.8 reveals that ∇f_δ admits the unbiased single-point estimator

$$g_\delta(x) = \frac{n}{\delta}\Im\left(f(x+i\delta y)\right) y \quad \text{with} \quad y \sim \sigma. \tag{5.1.13}$$

Now we show that $\nabla f_\delta(x)$ approximates $\nabla f(x)$ arbitrarily well as δ drops to 0.

Proposition 5.1.9 (Approximation quality of ∇f_δ)**.** *If $f \in C^{\omega,2}_{L_2}(\mathcal{D})$ satisfies Assumption 5.1.6, then for f_δ defined as in (5.1.10) and for any fixed $x \in \mathcal{D}$ and $\kappa \in (0,1)$ there exists $C_\kappa \geq 0$ with*

$$\|\nabla f_\delta(x) - \nabla f(x)\|_2 \leq \tfrac{1}{6} n L_2 \delta^2 + n C_\kappa \delta^4 \quad \forall \delta \in (0, \kappa\bar{\delta}]. \tag{5.1.14}$$

Proof. If we denote as usual by I_n the identity matrix in $\mathbb{R}^n$, then the covariance matrix of the uniform distribution σ on the unit sphere $\mathbb{S}^{n-1}$ can be expressed as

$$\int_{\mathbb{S}^{n-1}} yy^\mathsf{T} \sigma(\mathrm{d}y) = \int_{\mathbb{S}^{n-1}} \|y\|_2^2\, \sigma(\mathrm{d}y) \cdot \tfrac{1}{n} I_n = \tfrac{1}{n} I_n, \tag{5.1.15}$$

where the two equalities hold because the sought covariance matrix must be isotropic and because $\|y\|_2 = 1$ for all $y \in \mathbb{S}^{n-1}$, respectively, see also Lemma 5.3.3. Thus, the gradient of f can be represented as

$$\nabla f(x) = n \int_{\mathbb{S}^{n-1}} \langle \nabla f(x), y\rangle y\, \sigma(\mathrm{d}y)$$

Together with Proposition 5.1.8, this yields the estimate

$$\begin{aligned}\|\nabla f_\delta(x) - \nabla f(x)\|_2 &= \frac{n}{\delta}\left\|\int_{\mathbb{S}^{n-1}} \Im\left(f(x+i\delta y)\right) y - \delta\langle \nabla f(x), y\rangle y \,\sigma(\mathrm{d}y)\right\|_2 \\ &\le \frac{n}{\delta}\int_{\mathbb{S}^{n-1}} \left|\Im\left(f(x+i\delta y)\right) - \delta\langle \nabla f(x), y\rangle\right| \|y\|_2 \,\sigma(\mathrm{d}y).\end{aligned}$$

By using the Taylor series representation of $f(x+i\delta y)$ around x, we find

$$\begin{aligned}\Im\left(f(x+i\delta y)\right) - \delta\langle \nabla f(x), y\rangle &= \sum_{k=0}^{\infty} \frac{(-1)^k \delta^{2k+1}}{(2k+1)!} \sum_{|\alpha|=2k+1} \partial_x^\alpha f(x) y^\alpha - \delta\langle \nabla f(x), y\rangle \\ &= \sum_{k=1}^{\infty} \frac{(-1)^k \delta^{2k+1}}{(2k+1)!} \sum_{|\alpha|=2k+1} \partial_x^\alpha f(x) y^\alpha \\ &= -\tfrac{1}{6}\delta^3 \sum_{|\alpha|=3} \partial_x^\alpha f(x) y^\alpha + \delta^5 R(y,\delta),\end{aligned}$$

where the real-valued remainder term $R(y,\delta)$ is continuous in $y \in \mathbb{B}^n$ and $\delta \in [0,\bar{\delta})$. Substituting the last expression into the above and using (2.1.9), we obtain

$$\begin{aligned}\|\nabla f_\delta(x) - \nabla f(x)\|_2 \le &\frac{n}{\delta}\int_{\mathbb{S}^{n-1}} \left(\tfrac{1}{6}\delta^3 L_2 + \delta^5 \left|R(y,\delta)\right|\right) \|y\|_2 \,\sigma(\mathrm{d}y) \\ \le &\tfrac{1}{6}\delta^2 n L_2 + n C_\kappa \delta^4 \quad \forall \delta \in (0, \kappa\delta],\end{aligned}$$

where the non-negative constant $C_\kappa = \max_{y\in\mathbb{S}^{n-1}} \max_{\delta\in[0,\kappa\bar{\delta}]} |R(y,\delta)|$ is again finite due to the continuity of $R(y,\delta)$ and the compactness of $\mathbb{S}^{n-1}$ and $[0,\kappa\bar{\delta}]$. □

Proposition 5.1.9 implies that the *single-point* estimator (5.1.13) incurs only errors of the order $O(\delta^2)$ on average. Equally small errors were attained in [NS17] for $f \in C^{2,2}_{L_2}$ by using Gaussian smoothing and a *multi-point* estimator. Unfortunately, the latter is susceptible to cancellation effects. Proposition 5.1.9 also implies that $\lim_{\delta\to 0^+} \nabla f_\delta(x) = \nabla f(x)$. In addition, one readily verifies that if f is quadratic (that is, if $L_2 = 0$), then $\nabla f_\delta(x) = \nabla f(x)$ for all $x \in \mathcal{D}$ and $\delta \in (0,\bar{\delta})$. The single-point estimator $g_\delta(x)$ introduced in (5.1.13) is unbiased by construction. In addition, as for the multi-point estimator proposed in [NS17], the second moment of $g_\delta(x)$ admits a convenient bound.

Corollary 5.1.10 (Second moment of $g_\delta(x)$)**.** *If $f \in C^{\omega,2}_{L_2}(\mathcal{D})$ satisfies Assumption 5.1.6, then for g_δ as in (5.1.13) and for any fixed $x \in \mathcal{D}$ and $\kappa \in (0,1)$ we have*

$$\begin{aligned}\mathbb{E}_{y\sim\sigma}\left[\|g_\delta(x)\|_2^2\right] \le n^2(\tfrac{1}{6}L_2\delta^2 + C_\kappa\delta^4)^2 + n\|\nabla f(x)\|_2^2 \\ + 2n^2\left(\tfrac{1}{6}L_2\delta^2 + C_\kappa\delta^4\right)\|\nabla f(x)\|_2,\end{aligned} \tag{5.1.16}$$

where $C_\kappa \ge 0$ is the same constant as in Proposition 5.1.9.

Proof. Using the definition of g_δ and the fact that $\|y\|_2 = 1 \; \forall y \in \mathbb{S}^{n-1}$, we find

$$\mathbb{E}_{y\sim\sigma}\left[\|g_\delta(x)\|_2^2\right] = \frac{n^2}{\delta^2}\mathbb{E}_{y\sim\sigma}[(\Im\left(f(x+i\delta y)\right))^2]. \tag{5.1.17}$$

By essentially the same arguments as in the proof of Proposition 5.1.9, we further have

$$\begin{aligned}|\Im\left(f(x+i\delta y)\right)| &= |\Im\left(f(x+i\delta y)\right) - \langle\nabla f(x),\delta y\rangle + \langle\nabla f(x),\delta y\rangle| \\ &\leq \left|\tfrac{1}{6}\delta^3 L_2 + \delta^5 C_\kappa\right| + |\langle\nabla f(x),\delta y\rangle|\,.\end{aligned}$$

Squaring the above and applying the Cauchy-Schwarz inequality yields

$$\begin{aligned}|\Im\left(f(x+i\delta y)\right)|^2 \leq \left(\tfrac{1}{6}\delta^3 L_2 + \delta^5 C_\kappa\right)^2 &+ \langle\nabla f(x),\delta y\rangle^2 \\ &+ 2\delta\left(\tfrac{1}{6}\delta^3 L_2 + \delta^5 C_\kappa\right)\|\nabla f(x)\|_2\|y\|_2.\end{aligned}$$

The claim then follows from substituting the above into (5.1.17) and using (5.1.15). □

In analogy to Proposition 5.1.9, one readily verifies that if f is quadratic (*i.e.*, if $L_2 = 0$), then the right hand side of (5.1.16) vanishes. Under a third-order smoothness condition, there exist multi-point estimators that satisfy a bound akin to (5.1.16) [NS17, Thm. 4.3].

Unlike the smooth approximations proposed in [NS17], the smoothed complex-step approximation f_δ does frequently *not* belong to the same function class as f. For example, even though the Lorentzian function $f(x) = 1/(1+x^2)$ has a Lipschitz continuous gradient with $L_1 = 2$, the Lipschitz modulus of its approximation f_δ strictly exceeds 2 for some values of δ close to 1 because f has two poles at i and $-i$. Similarly, f_δ does not necessarily inherit convexity from f.

Example 5.1.5 (Loss of convexity)**.** If $f \in C^\omega(\mathbb{R})$ is entire, then it has a globally convergent power series representation with real coefficients. Consequently, f satisfies

$$\Re(f(x+i\delta y)) = \sum_{k=0}^{\infty}(-1)^k \frac{f^{(2k)}(x)}{(2k)!}(\delta y)^{2k}.$$

In the special case when $f(x) = x^2$, the complex-step approximation $\Re(f(x+i\delta y)) = x^2 - (\delta y)^2$ inherits convexity from f regardless of the choice of $\delta > 0$ and $y \in \mathbb{R}$. Thus, f_δ is also convex because convexity is preserved by integration. However, if $f(x) = x^4$, then we find $\Re(f(x+i\delta y)) = x^4 - 6x^2(\delta y)^2 + (\delta y)^4$, which fails to be convex in x for any $\delta > 0$ and $y \neq 0$. In this case, f_δ remains non-convex despite the smoothing. Finally, if f is strongly convex (*e.g.*, if $f(x) = x^2+x^4$), then one readily verifies that $\Re(f(x+i\delta y))$ is convex in x provided that δ is sufficiently small (*e.g.*, exploit second-order convexity conditions).

If f_δ inherited convexity from f, one could simply incorporate the estimator (5.1.13) into the algorithms studied in [NS17, § 5], and the corresponding convergence analysis would carry over with minor modifications. As the smoothed complex-step approximation may destroy convexity, however, a different machinery is needed here.

5.1.4 Optimization: algorithms and analysis

In what follows, we study zeroth-order optimization under the randomized (gradient) estimator (5.1.13).

Algorithm 1 Imaginary zeroth-order optimization

```
Input: initial iterate $x_1 \in \mathcal{X}$, stepsizes $\{\mu_k\}_{k\in\mathbb{Z}_{>0}}$, smoothing parameters $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$
for $k = 1, 2, \dots, K-1$ do
    sample $y_k \sim \sigma$
    set $g_{\delta_k}(x_k) = \frac{n}{\delta_k}\Im\left(f(x_k + i\delta_k y_k)\right) y_k$
    set $x_{k+1} = \Pi_{\mathcal{X}}\left(x_k - \mu_k\, g_{\delta_k}(x_k)\right)$
end for
Output: last iterate $x_K$ and averaged iterate $\bar{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$
```

Convex optimization

Now, we study the convergence properties of zeroth-order algorithms for solving problem (5.1.1) under the assumption that f is a convex function on $\mathcal{D}$ and $\mathcal{X}$ is a non-empty closed convex subset of $\mathcal{D}$. Our methods mimic existing algorithms developed in [NS17] but use the single-point estimator g_δ defined in (5.1.13) instead of a multi-point estimator that may suffer from cancellation effects. Our method is described in Algorithm 1, where $\Pi_{\mathcal{X}} : \mathcal{D} \to \mathcal{X}$ denotes the Euclidean projection onto $\mathcal{X}$. Note that $\Pi_{\mathcal{X}}$ reduces to the identity operator if $\mathcal{X} = \mathcal{D}$.

In the remainder we will assume that the iterates $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ generated by Algorithm 1 as well as all samples $\{y_k\}_{k\in\mathbb{Z}_{>0}}$ and the corresponding gradient estimators $\{g_{\delta_k}(x_k)\}_{k\in\mathbb{Z}_{>0}}$ represent random objects on an abstract filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\in\mathbb{Z}_{>0}}, \mathbb{P})$, where $\mathcal{F}_k$ denotes the σ-algebra generated by the independent and identically distributed samples $y_1, \dots, y_{k-1}$. Therefore, x_k is $\mathcal{F}_k$-measurable. In the following, we use $\mathbb{E}[\cdot]$ to denote the expectation operator with respect to $\mathbb{P}$.

The proofs of our convergence results rely on the following lemma borrowed from [SRB11].

Lemma 5.1.11 ([SRB11, Lem. 1])**.** *If $\{t_k\}_{k\in\mathbb{Z}_{>0}}$ and $\{\nu_k\}_{k\in\mathbb{Z}_{>0}}$ are two sequence of non-negative real numbers, while $\{T_K\}_{K\in\mathbb{Z}_{>0}}$ is a non-decreasing sequence of real numbers with $T_1 \geq t_1^2$ such that $t_K^2 \leq T_K + \sum_{k=1}^K \nu_k t_k$ $\forall K \in \mathbb{Z}_{>0}$, then we have*

$$t_K \leq \tfrac{1}{2}\textstyle\sum_{k=1}^K \nu_k + \Big(T_K + (\tfrac{1}{2}\sum_{k=1}^K \nu_k)^2\Big)^{\frac{1}{2}} \quad \forall k \in \mathbb{Z}_{>0}.$$

In addition, several proofs in the main text make use of the inequalities

$$\textstyle\sum_{j=1}^J j^{-2} \leq \zeta(2) = \frac{1}{6}\pi^2 \quad \text{and} \quad \sum_{j=1}^J j^{-4} \leq \zeta(4) = \frac{1}{90}\pi^4 \quad \forall J \in \mathbb{Z}_{>0}, \tag{5.1.18}$$

which are obtained by truncating the series that defines the Riemann zeta function.

Theorem 5.1.12 (Convergence rate of Algorithm 1 for convex optimization)**.** *Suppose that f is a convex, real analytic function satisfying Assumption 5.1.6 as well as the Lipschitz conditions (2.1.4) and (2.1.7) with $L_1 > 0$ and $L_2 \geq 0$. Also assume that $\mathcal{X}$ is non-empty, closed and convex and that there exists $x^\star \in \mathcal{X}$ with $\nabla f(x^\star) = 0$. Denote by $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ the iterates generated by Algorithm 1 with constant stepsize $\mu_k = \mu = 1/(2nL_1)$ and adaptive smoothing parameter $\delta_k \in (0, \kappa\bar{\delta}]$ for all $k \in \mathbb{Z}_{>0}$, where $\kappa \in (0,1)$, and define $R = \|x_1 - x^\star\|_2$. Then, the following hold for all $K \in \mathbb{Z}_{>0}$.*

(i) There is a constant $C_1 \geq 0$ such that

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq & \tfrac{1}{\mu K} R^2 + \tfrac{1}{K} C_1 n R \textstyle\sum_{k=1}^K \delta_k^2 + \tfrac{1}{K}\mu C_1^2 n^2 (\textstyle\sum_{k=1}^K \delta_k^2)^2 \\ & + \tfrac{1}{K}\mu C_1 C_2 n^2 (\textstyle\sum_{k=1}^K \delta_k^2)(\textstyle\sum_{k=1}^K \delta_k^4)^{\frac{1}{2}} + \tfrac{1}{K}\mu C_2^2 n^2 \textstyle\sum_{k=1}^K \delta_k^4.\end{aligned}$$

(ii) If $\delta_k = \delta$ for all $k \in \mathbb{Z}_{>0}$, then we have

$$\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq \tfrac{1}{K} 2 n L_1 R^2 + C_1 n R \delta^2 + \tfrac{1}{L_1}(1 + \sqrt{K})^2 C_1^2 n \delta^4.$$

(iii) If $\delta_k = \delta/k$ for all $k \in \mathbb{Z}_{>0}$, then there is a constant $C_2 \geq 0$ such that

$$\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq \tfrac{n}{K}\left(\sqrt{2L_1} R + C_2 \delta^2\right)^2.$$

Under the assumptions of Theorem 5.1.12, problem (5.1.1) is convex and $x^\star$ represents a global minimizer. Note, however, that $\mathcal{X}$ may not contain any $x^\star$ with $\nabla f(x^\star) = 0$ even if $\mathcal{X}$ is compact. This is usually the case if the global minimum of (5.1.1) is attained at the boundary of $\mathcal{X}$. If $x^\star$ is not unique, one should set $R = \|x_1 - P^\star(x_1)\|_2$ for the bounds not to be trivial, with $P^\star(x_1) = \operatorname{argmin}_{x^\star} \|x^\star - x_1\|_2^2$, which is well-defined since f is convex, real analytic. Explicit formulas for C_1 and C_2 in terms of κ, L_2 etc. are derived in the proof of Theorem 5.1.12.

Proof of Theorem 5.1.12. For ease of notation, we define $r_k = \|x_k - x^\star\|_2$ for all $k \in \mathbb{Z}_{>0}$. We prove the theorem first under the simplifying assumption that $\mathcal{X} = \mathcal{D}$, which implies the projection onto $\mathcal{X}$ becomes obsolete, that is, $x_{k+1} = x_k - \mu_k \cdot g_{\delta_k}(x_k)$. Thus, we have

$$\begin{aligned}\mathbb{E}\left[r_{k+1}^2 \,\middle|\, \mathcal{F}_k\right] &= \mathbb{E}\left[r_k^2 - 2\mu_k \langle g_{\delta_k}(x_k), x_k - x^\star\rangle + \mu_k^2 \|g_{\delta_k}(x_k)\|_2^2 \,\middle|\, \mathcal{F}_k\right] \\ &= r_k^2 - 2\mu_k \langle \nabla f_{\delta_k}(x_k), x_k - x^\star\rangle + \mu_k^2\, \mathbb{E}\left[\|g_{\delta_k}(x_k)\|_2^2 \,\middle|\, \mathcal{F}_k\right],\end{aligned}$$

where the second equality follows from (5.1.11), the definition of $g_{\delta_k}(x_k)$ and the $\mathcal{F}_k$-measurability of x_k and r_k. The Cauchy-Schwartz inequality then implies that

$$\begin{aligned}&\mathbb{E}\left[r_{k+1}^2 \,\middle|\, \mathcal{F}_k\right] \\ &\leq r_k^2 - 2\mu_k \langle \nabla f(x_k), x_k - x^\star\rangle + 2\mu_k \|\nabla f_{\delta_k}(x_k) - \nabla f(x_k)\|_2\, r_k + \mu_k^2\, \mathbb{E}\left[\|g_{\delta_k}(x_k)\|_2^2 \,\middle|\, \mathcal{F}_k\right] \\ &\leq r_k^2 - 2\mu_k\left(f(x_k) - f(x^\star)\right) + 2\mu_k\left(\tfrac{1}{6} n L_2 \delta_k^2 + n C_\kappa \delta_k^4\right) r_k \\ &\qquad + \mu_k^2 n^2 \left(\left(\tfrac{1}{6} L_2 \delta_k^2 + C_\kappa \delta_k^4\right)^2 + \tfrac{1}{n}\|\nabla f(x_k)\|_2^2 + 2\left(\tfrac{1}{6} L_2 \delta_k^2 + C_\kappa \delta_k^4\right)\|\nabla f(x_k)\|_2\right) \\ &\leq r_k^2 - 2\mu_k\left(f(x_k) - f(x^\star)\right) + 2n\mu_k \delta_k^2\left(\tfrac{1}{6} L_2 + C_\kappa \delta_k^2 + n L_1 \mu_k\left(\tfrac{1}{6} L_2 + C_\kappa \delta_k^2\right)\right) r_k \\ &\qquad + \mu_k^2 n^2 \left(\delta_k^4\left(\tfrac{1}{6} L_2 + C_\kappa \delta_k^2\right)^2 + \tfrac{1}{n} 2 L_1 (f(x_k) - f(x^\star))\right),\end{aligned}$$

where the second inequality exploits the convexity of f as well as Proposition 5.1.9 and Corollary 5.1.10, while the third inequality follows from the estimates (2.1.4) and (2.1.6), which imply that $\|\nabla f(x_k)\|_2 \leq L_1 \|x_k - x^\star\|_2$ and $2L_1(f(x_k) - f(x^\star)) \geq \|\nabla f(x_k)\|_2^2$, respectively. To simplify notation, we now introduce the constant $C_1 = \frac{1}{2} L_2 + 3C_\kappa$, which upper bounds $\frac{1}{2} L_2 + 3C_\kappa \delta_k^2$ and $\frac{1}{6} L_2 + C_\kappa \delta_k^2$ for any $k \in \mathbb{Z}_{>0}$ because all smoothing parameters belong to the interval $[-1,1]$. Recalling that the stepsize is constant and equal to $\mu = 1/(2nL_1)$, the above display equation thus simplifies to

$$\mathbb{E}\left[r_{k+1}^2 \,\middle|\, \mathcal{F}_k\right] \leq r_k^2 - \mu\left(f(x_k) - f(x^\star)\right) + n\mu\delta_k^2 C_1 r_k + \mu^2 n^2 C_1^2 \delta_k^4. \tag{5.1.19}$$

Taking unconditional expectations and rearranging terms then yields

$$\begin{aligned}\mathbb{E}\left[f(x_k) - f(x^\star)\right] &\leq \tfrac{1}{\mu}\left(\mathbb{E}\left[r_k^2\right] - \mathbb{E}\left[r_{k+1}^2\right]\right) + nC_1\delta_k^2\mathbb{E}\left[r_k\right] + \mu n^2 C_1^2 \delta_k^4 \\ &\leq \tfrac{1}{\mu}\left(\mathbb{E}\left[r_k^2\right] - \mathbb{E}\left[r_{k+1}^2\right]\right) + nC_1\delta_k^2\sqrt{\mathbb{E}\left[r_k^2\right]} + \mu n^2 C_1^2 \delta_k^4.\end{aligned}$$

Next, choose any $k' \in \mathbb{Z}_{>0}$ and sum the above inequalities over all $k \leq k'-1$ to obtain

$$\begin{aligned}\textstyle\sum_{k=1}^{k'-1} \mathbb{E}\left[f(x_k) - f(x^\star)\right] \leq &\tfrac{1}{\mu}\left(r_1^2 - \mathbb{E}\left[r_{k'}^2\right]\right) + C_1 n \textstyle\sum_{k=1}^{k'-1} \delta_k^2 \sqrt{\mathbb{E}\left[r_k^2\right]} \\ &+ \mu C_1^2 n^2 \textstyle\sum_{k=1}^{k'-1} \delta_k^4.\end{aligned} \tag{5.1.20}$$

Clearly, the inequality (5.1.20) remains valid if we lower bound its left hand side by 0 and upper bound its right hand side by increasing the upper limits of the two sums to k'. We then obtain $\mathbb{E}[r_{k'}^2] \leq r_1^2 + \mu C_1 n \sum_{k=1}^{k'} \delta_k^2 \sqrt{\mathbb{E}\left[r_k^2\right]} + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4$. Setting $t_k = \sqrt{\mathbb{E}\left[r_k^2\right]}$ and $\nu_k = \mu C_1 n \delta_k^2$ for all $k \in \mathbb{Z}_{>0}$ and defining $T_{k'} = r_1^2 + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4$ for all $k' \in \mathbb{Z}_{>0}$, we may use Lemma 5.1.11 to conclude that

$$\begin{aligned}\sqrt{\mathbb{E}\left[r_{k'}^2\right]} &\leq \tfrac{1}{2}\mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 + \left(r_1^2 + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4 + (\tfrac{1}{2}\mu C_1 n \sum_{k=1}^{k'} \delta_k^2)^2\right)^{\frac{1}{2}} \\ &\leq \mu C_1 n \textstyle\sum_{k=1}^{K} \delta_k^2 + r_1 + (\mu^2 C_1^2 n^2 \sum_{k=1}^{K} \delta_k^4)^{\frac{1}{2}} \quad \forall k \leq K,\end{aligned}$$

where the second inequality holds because $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$ for all $a, b, c \geq 0$ and because the sums increase when we increase their upper limits from k' to K. Next, consider the estimate (5.1.20) for $k' = K+1$, replace $\mathbb{E}[r_{K+1}^2]$ with its trivial lower bound 0 and replace $\sqrt{\mathbb{E}[r_k^2]}$ with the above upper bound for every $k \leq K$. Noting that $r_1 = R$ and dividing by K then yields

$$\begin{aligned}&\tfrac{1}{K}\textstyle\sum_{k=1}^{K} \mathbb{E}\left[f(x_k) - f(x^\star)\right] \\ &\leq \tfrac{1}{\mu K} R^2 + \tfrac{1}{K}\mu C_1^2 n^2 \textstyle\sum_{k=1}^{K} \delta_k^4 \\ &\quad + \tfrac{1}{K} C_1 n \textstyle\sum_{k=1}^{K} \delta_k^2 \left(\mu C_1 n \sum_{k=1}^{K} \delta_k^2 + R + (\mu^2 C_1^2 n^2 \sum_{k=1}^{K} \delta_k^4)^{\frac{1}{2}}\right) \\ &= \tfrac{1}{\mu K} R^2 + \tfrac{1}{K} C_1 n R \textstyle\sum_{k=1}^{K} \delta_k^2 + \tfrac{1}{K}\mu C_1^2 n^2 (\sum_{k=1}^{K} \delta_k^2)^2 \\ &\quad + \tfrac{1}{K}\mu C_1^2 n^2 (\textstyle\sum_{k=1}^{K} \delta_k^2)(\sum_{k=1}^{K} \delta_k^4)^{\frac{1}{2}} + \tfrac{1}{K}\mu C_1^2 n^2 \sum_{k=1}^{K} \delta_k^4.\end{aligned}$$

As $\mathbb{E}[f(\bar{x}_K) - f(x^\star)] \leq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[f(x_k) - f(x^\star)]$ by Jensen's inequality, assertion *(i)* thus follows. If $\delta_k = \delta \in (0, \kappa\bar{\delta}]$ for all $k \in \mathbb{Z}_{>0}$, then assertion *(i)* implies that

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] &\leq \tfrac{1}{\mu K} R^2 + C_1 n R \delta^2 + C_1^2 K \mu n^2 \delta^4 + C_1^2 \sqrt{K} \mu n^2 \delta^4 + C_1^2 \mu n^2 \delta^4 \\ &\leq \tfrac{1}{\mu K} R^2 + C_1 n R \delta^2 + (C_1\sqrt{K} + C_1)^2 \mu n^2 \delta^4 \\ &\leq \tfrac{1}{K} 2nL_1 R^2 + C_1 n R \delta^2 + (C_1\sqrt{K} + C_1)^2 \tfrac{1}{L_1} n \delta^4 \\ &\leq \tfrac{1}{K} 2nL_1 R^2 + C_1 n R \delta^2 + C_1^2 (1 + \sqrt{K})^2 \tfrac{1}{L_1} n \delta^4,\end{aligned}$$

where the last two inequalities exploit the assumption $\mu = 1/(2nL_1)$. Thus, assertion *(ii)* follows. Next, assume that $\delta_k = \delta/k$ for all $k \in \mathbb{Z}_{>0}$. In analogy to the proof of assertion *(ii)*, we combine assertion *(i)* with the standard zeta function inequalities (5.1.18) to conclude that

$$\begin{aligned}\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq& \tfrac{1}{\mu K}R^2 + \tfrac{1}{6}\pi^2 C_1 \tfrac{1}{K} nR\delta^2 + \tfrac{1}{90}\pi^4 C_1^2 \tfrac{1}{K}\mu n^2\delta^4 \\ &+ \tfrac{1}{36}\pi^4 C_1^2 \tfrac{1}{K}\mu n^2\delta^4 + \tfrac{1}{6\sqrt{90}}\pi^4 C_1^2 \tfrac{1}{K}\mu n^2\delta^4 \\ \leq& \tfrac{n}{K}2L_1R^2 + \tfrac{n}{K}R\tfrac{1}{6}\pi^2 C_1\delta^2 + \tfrac{n}{K}R\pi^4(\tfrac{1}{6}C_1 + \tfrac{1}{\sqrt{90}}C_1)^2\tfrac{1}{2L_1}\delta^4 \\ \leq& \tfrac{n}{K}(\sqrt{2L_1}R + C_2\delta^2)^2,\end{aligned}$$

where $C_2 = \pi^2(C_1/3 + C_1/\sqrt{90})/\sqrt{2L_1}$. The third inequality holds because $\mu = 1/(2nL_1)$. Thus, assertion *(iii)* follows. This completes the proof for $\mathcal{X} = \mathcal{D}$.

In the last part of the proof we show that the three assertions remain valid when $\mathcal{X}$ is a non-empty closed convex subset of $\mathcal{D}$. Indeed, as the projection $\Pi_{\mathcal{X}}$ onto $\mathcal{X}$ is contractive, we have

$$\begin{aligned}r_{k+1}^2 = \|x_{k+1} - x^\star\|_2^2 &= \|\Pi_{\mathcal{X}}(x_k - \mu_k g_{\delta_k}(x_k)) - \Pi_{\mathcal{X}}(x^\star)\|_2^2 \\ &\leq \|x_k - \mu_k g_{\delta_k}(x_k) - x^\star\|_2^2.\end{aligned}$$

Thus, all arguments used above carry over trivially to situations where $\mathcal{X} \neq \mathcal{D}$. □

Theorem 5.1.12 *(iii)* shows that if δ_k decays as $O(1/k)$, then one needs $O\left(nL_1R^2/\epsilon\right)$ iterations to guarantee that $\mathbb{E}\left[f(\bar{x}_K) - f(x^\star)\right] \leq \epsilon$. This is the first-order complexity scaled by n [Nes03, Sec. 2.1.5]. Theorem 5.1.12 can be extended to a larger class of convex optimization problems by relaxing the assumption of constant stepsizes [Jon21]. In particular, it can be extended to constrained optimization problems whose constraints are binding at optimality, in which case $\nabla f(x^\star) \neq 0$; see also Section 5.2 below.

Strongly convex optimization

We now extend the results from Section 5.1.4 to analytic objective functions f that are τ-strongly convex over their domain $\mathcal{D}$ for some $\tau > 0$, *i.e.*, we assume that $f(y) \geq f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2}\tau\|y - x\|_2^2 \ \forall x, y \in \mathcal{D}$. If y is a stationary point with $\nabla f(y) = 0$, then τ-strong convexity ensures that

$$f(y) - f(x) \geq \tfrac{1}{2}\tau\|y - x\|_2^2 \quad \forall x \in \mathcal{D}, \tag{5.1.21}$$

which in turn implies via the Polyak-Łojasiewicz inequality $\|\nabla f(x)\|_2^2 \geq 2\tau(f(x) - f(y))$ for τ-strongly convex functions [Nes03, Eq. 2.1.19] that

$$\|\nabla f(x)\|_2 \geq \tau\|y - x\|_2. \tag{5.1.22}$$

Theorem 5.1.13 (Convergence rate of Algorithm 1 for strongly convex optimization)**.** *Suppose that all assumptions of Theorem 5.1.12 (iii) are satisfied and that f is τ-strongly convex for some $\tau > 0$. Then, there is a constant $C \geq 0$ such that the following inequality holds for all $K \in \mathbb{Z}_{>0}$.*

$$\mathbb{E}[f(x_K) - f(x^\star)] \leq \tfrac{1}{2}L_1\left(\delta^2 C + (1 - \tfrac{\tau}{4nL_1})^{K-1}\left(R^2 - \delta^2 C\right)\right) \tag{5.1.23}$$

An explicit formula for C in terms of n, L_1, L_2 and τ is derived in the proof.

Proof of Theorem 5.1.13. As in the proof of Theorem 5.1.12, we set $C_1 = 3(\frac{1}{6}L_2 + C_\kappa)$ and $r_k = \|x_k - x^\star\|_2$ for all $k \in \mathbb{Z}_{>0}$, and we initially assume that $\mathcal{X} = \mathcal{D}$. Combining the estimate (5.1.19) from the proof of Theorem 5.1.12 with the strong convexity condition (5.1.21) yields $\mathbb{E}\left[r_{k+1}^2 | \mathcal{F}_k\right] \leq \left(1 - \frac{\mu\tau}{2}\right) r_k^2 + \mu C_1 n \delta_k^2 r_k + \mu^2 C_1^2 n^2 \delta_k^4$. By taking unconditional expectations and applying Jensen's inequality, we then find

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &\leq \left(1 - \tfrac{\mu\tau}{2}\right) \mathbb{E}[r_k^2] + \mu C_1 n \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \delta_k^4 && (5.1.24a) \\ &\leq \mathbb{E}[r_k^2] + \mu C_1 n \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \delta_k^4. && (5.1.24b) \end{aligned}$$

Next, choose any $k' \in \mathbb{Z}_{>0}$ and sum the above inequalities over all $k \leq k' - 1$ to obtain

$$\begin{aligned} \mathbb{E}[r_{k'}^2] &\leq r_1^2 + \mu C_1 n \textstyle\sum_{k=1}^{k'-1} \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'-1} \delta_k^4 \\ &\leq r_1^2 + \mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 \sqrt{\mathbb{E}[r_k^2]} + \mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4. \end{aligned}$$

By using the same reasoning as in the proof of Theorem 5.1.12, the last bound implies

$$\sqrt{\mathbb{E}\left[r_{k'}^2\right]} \leq \mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 + r_1 + (\mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4)^{\frac{1}{2}}.$$

Substituting this inequality into (5.1.24a) for $k = k'$ and noting that $r_1 = R$ yields

$$\begin{aligned} \mathbb{E}[r_{k'+1}^2] \leq \left(1 - \tfrac{\mu\tau}{2}\right) \mathbb{E}[r_{k'}^2] &+ \mu^2 C_1^2 n^2 \delta_{k'}^4 \\ &+ \mu C_1 n \delta_{k'}^2 \left(\mu C_1 n \textstyle\sum_{k=1}^{k'} \delta_k^2 + R + (\mu^2 C_1^2 n^2 \sum_{k=1}^{k'} \delta_k^4)^{\frac{1}{2}}\right). \end{aligned}$$

As $\delta_k = \delta/k$ for all $k \in \mathbb{Z}_{>0}$ and as the constant stepsize satisfies $\mu = 1/(2nL_1)$, we may then use the standard zeta function inequalities (5.1.18) to obtain

$$\begin{aligned} \mathbb{E}[r_{k'+1}^2] &\leq (1 - \tfrac{\tau}{4nL_1})\mathbb{E}[r_{k'}^2] + C_1^2 \tfrac{\delta^4}{4L_1^2(k')^4} + C_1^2 \tfrac{\pi^2\delta^4}{24L_1^2(k')^2} + C_1 R \tfrac{\delta^2}{2L_1(k')^2} + C_1^2 \tfrac{\pi^2\delta^4}{4\sqrt{90}L_1^2(k')^2} \\ &\leq (1 - \tfrac{\tau}{4nL_1})\mathbb{E}[r_{k'}^2] + C_1 R \tfrac{\delta^2}{L_1} + 3C_1^2 \tfrac{\delta^4}{L_1^2}, \end{aligned}$$

where the last inequality follows from the elementary bounds $\frac{1}{2(k')^2} < 1$, $\frac{1}{4(k')^4} < 1$, $\frac{\pi^2}{24(k')^2} < 1$ and $\pi^2/(4\sqrt{90}(k')^2) < 1$. As $|\delta| < 1$, we may set $C = \frac{4n}{\tau}(C_1 R + 3C_1^2/L_1)$ to obtain

$$\mathbb{E}[r_{k'+1}^2] \leq (1 - \tfrac{\tau}{4nL_1})\mathbb{E}[r_{k'}^2] + \tfrac{\tau}{4nL_1}\delta^2 C.$$

Taken together, the Lipschitz inequality (2.1.4) and the strong convexity inequality (5.1.22) imply that $\tau \leq L_1$, which in turn ensures that $\tau/(4nL_1) < 1$. Hence, the above inequality implies $\mathbb{E}[r_{k'+1}^2] - \delta^2 C \leq (1 - \frac{\tau}{4nL_1})\left(\mathbb{E}([r_{k'}^2] - \delta^2 C\right)$. As this estimate holds for all $k' < K$, we may finally conclude that

$$\mathbb{E}[r_K^2] - \delta^2 C \leq (1 - \tfrac{\tau}{4nL_1})\left(\mathbb{E}[r_{K-1}^2] - \delta^2 C\right) \leq \cdots \leq (1 - \tfrac{\tau}{4nL_1})^{K-1}(R^2 - \delta^2 C).$$

The claim then follows by combining this inequality with the estimate $\mathbb{E}[f(x_K) - f(x^\star)] \leq \frac{1}{2}L_1\mathbb{E}[r_K^2]$, which follows from the Lipschitz condition (2.1.5). This completes the proof for $\mathcal{X} = \mathcal{D}$. To show that the claim remains valid when $\mathcal{X}$ is a non-empty closed convex subset of $\mathcal{D}$, we may proceed as in the proof of Theorem 5.1.12. Details are omitted for brevity. □

By Theorem 5.1.13 and the construction of C, we can enforce $\mathbb{E}[f(x_K) - f(x^\star)] \leq \epsilon$ for a given tolerance $\epsilon > 0$ by selecting a sufficiently small smoothing parameter $\delta \leq O(\sqrt{\epsilon\tau/(nL_1^2)})$ and by running Algorithm 1 over
$O(nL_1/\tau \log(L_1 R^2/\epsilon))$ iterations.

Remark 5.1.6 (Stochastic stability)**.** The commentary of Section 2.1.2 applies to the convergence result from Theorem 5.1.13. In particular, the proof of that theorem reveals we have

$$\mathbb{E}\left[\|x_{k+1} - x^\star\|_2^2 - \delta^2 C\right] \leq \gamma \mathbb{E}\left[\|x_k - x^\star\|_2^2 - \delta^2 C\right] \leq \gamma^k (R^2 - \delta^2 C), \tag{5.1.25}$$

for C as given in the proof and $\gamma = (1 - \tau/(4nL_1)) \in (0,1)$. Now define the function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ by $V(x) = 0$ for all $x \in \mathbb{D}^n_{\delta\sqrt{C}}(x^\star)$ and $V(x) = \|x - x^\star\|_2^2 - C\delta^2$ otherwise. Let F_k be the usual (stochastic) update rule of the algorithm (Algorithm 1), that is, $x_{k+1} = F_k(x_k)$, then it follows that for any $k \geq 1$ we have $\mathbb{E}[V(F_k(x))] < V(x)$ for all $x \notin \mathbb{D}^n_{\delta\sqrt{C}}(x^\star)$ and $V(x) \geq 0$ with $V(x) = 0 \iff x \in \mathbb{D}^n_{\delta\sqrt{C}}(x^\star)$. See [Mor68] for more on precisely defining stochastic stability of discrete-time systems.

Non-convex optimization

We now extend the convergence guarantees for Algorithm 1 to unconstrained non-convex optimization problems. Our proof strategy differs from the one in [NS17] as the smoothed objective function f_δ does not necessarily admit a Lipschitz continuous gradient. In this setting, convergence can still be guaranteed if the initial iterate x_1 is sufficiently close to some global minimizer $x^\star$.

Theorem 5.1.14 (Convergence rate of Algorithm 1 for nonconvex optimization)**.** *Suppose that all assumptions of Theorem 5.1.12 (iii) hold, but assume that f may be non-convex, $\mathcal{X} = \mathcal{D}$ and $\mu_k = \mu = 1/(nL_1)$ for all $k \in \mathbb{Z}_{>0}$. Define $F = f(x_1) - f(x^\star)$, where $x^\star$ is a global minimizer of problem (5.1.1). If $\|\nabla f(x_1)\|_2^2 \leq 2nL_1 F$, then there is a constant $C \geq 0$ such that for all $K \in \mathbb{Z}_{>0}$ we have*

$$\tfrac{1}{K}\textstyle\sum_{k=1}^K \mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \leq \tfrac{n}{K}\left(2L_1 F + \delta^2 C\right).$$

The dependence of C on n, L_1, L_2 and F can be derived from the proof of Theorem 5.1.14.

Proof of Theorem 5.1.14. As $\mathcal{X} = \mathcal{D}$, the iterates of Algorithm 1 satisfy $x_{k+1} = x_k - \mu_k \cdot g_{\delta_k}(x_k)$. In addition, as f has a Lipschitz continuous gradient, the Lipschitz inequality (2.1.5) implies that

$$\begin{aligned} f(x_{k+1}) \leq& f(x_k) - \mu_k \langle \nabla f(x_k), g_{\delta_k}(x_k)\rangle + \tfrac{1}{2}\mu_k^2 L_1 \|g_{\delta_k}(x_k)\|_2^2 \\ =& f(x_k) - \mu_k \|\nabla f(x_k)\|_2^2 - \mu_k \langle \nabla f(x_k), g_{\delta_k}(x_k) - \nabla f(x_k)\rangle + \tfrac{1}{2}\mu_k^2 L_1 \|g_{\delta_k}(x_k)\|_2^2. \end{aligned}$$

Taking conditional expectations on both sides of this expression, recalling that g_{δ_k} is an unbiased estimator for ∇f_{δ_k} conditional on $\mathcal{F}_k$ and applying the Cauchy-Schwarz inequality then yields

$$\begin{aligned} \mathbb{E}\left[f(x_{k+1})|\mathcal{F}_k\right] \leq& f(x_k) - \mu_k \|\nabla f(x_k)\|_2^2 \\ &+ \mu_k \|\nabla f(x_k)\|_2 \|\nabla f_{\delta_k}(x_k) - \nabla f(x_k)\|_2 + \tfrac{1}{2}\mu_k^2 L_1 \mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2 | \mathcal{F}_k]. \end{aligned}$$

Defining $C_0 = \frac{1}{6}L_2 + C_\kappa$, we may use the estimates (5.1.14) and (5.1.16) to obtain

$$\begin{aligned}\mathbb{E}\left[f(x_{k+1})|\mathcal{F}_k\right] \leq & f(x_k) - \mu_k\|\nabla f(x_k)\|_2^2 + \mu_k C_0 n\delta_k^2\|\nabla f(x_k)\|_2 \\ & + \tfrac{1}{2}\mu_k^2 L_1\left(n\|\nabla f(x_k)\|_2^2 + C_0^2 n^2\delta_k^4 + 2C_0 n^2\delta_k^2\|\nabla f(x)\|_2\right) \\ = & f(x_k) - \tfrac{1}{2nL_1}\|\nabla f(x_k)\|_2^2 + \tfrac{1}{2L_1}C_0^2\delta_k^4 + \tfrac{2}{L_1}C_0\delta_k^2\|\nabla f(x_k)\|_2,\end{aligned}$$

where the equality holds because the stepsize is constant and equal to $\mu_k = 1/(nL_1)$. By taking unconditional expectations, applying Jensen's inequality and rearranging terms, we then find

$$\begin{aligned}\mathbb{E}[\|\nabla f(x_k)\|_2]^2 &\leq \mathbb{E}[\|\nabla f(x_k)\|_2^2] \\ &\leq 2nL_1\mathbb{E}[f(x_k) - f(x_{k+1})] + 4nC_0\delta_k^2\mathbb{E}\left[\|\nabla f(x_k)\|_2\right] + nC_0^2\delta_k^4.\end{aligned} \tag{5.1.26}$$

Next, choose any $k' \in \mathbb{Z}_{>0}$ and sum the left- and rightmost terms in (5.1.26) over all $k \leq k'$ to obtain

$$\begin{aligned}&\mathbb{E}[\|\nabla f(x_{k'})\|_2]^2 \\ &\leq \textstyle\sum_{k=1}^{k'}\mathbb{E}[\|\nabla f(x_k)\|_2]^2 \\ &\leq 2nL_1\mathbb{E}[f(x_1) - f(x_{k'+1})] + 4nC_0\textstyle\sum_{k=1}^{k'}\delta_k^2\mathbb{E}[\|\nabla f(x_k)\|_2] + nC_0^2\textstyle\sum_{k=1}^{k'}\delta_k^4 \\ &\leq 2nL_1F + 4nC_0\textstyle\sum_{k=1}^{k'}\delta_k^2\mathbb{E}[\|\nabla f(x_k)\|_2] + nC_0^2\textstyle\sum_{k=1}^{k'}\delta_k^4,\end{aligned}$$

where the third inequality holds because $x^\star$ is a global minimizer of problem (5.1.1), which implies $\mathbb{E}[f(x_1) - f(x_{k'+1})] = \mathbb{E}[f(x_1) - f(x^\star)] + \mathbb{E}[f(x^\star) - f(x_{k'+1})] \leq F$. Setting $t_k = \mathbb{E}[\|\nabla f(x_k)\|_2]$ and $\nu_k = 4nC_0\delta_k^2$ for all $k \in \mathbb{Z}_{>0}$, and defining $T_{k'} = 2nL_1F + nC_0^2\sum_{k=1}^{k'}\delta_k^4$ for all $k' \in \mathbb{Z}_{>0}$, we may then use Lemma 5.1.11, which applies because $\|\nabla f(x_1)\|_2^2 \leq 2nL_1F$, to find

$$\mathbb{E}[\|\nabla f(x_{k'})\|_2] \leq 2nC_0\textstyle\sum_{k=1}^{k'}\delta_k^2 + \left(2nL_1F + nC_0^2\textstyle\sum_{k=1}^{k'}\delta_k^4 + (2nC_0\textstyle\sum_{k=1}^{k'}\delta_k^2)^2\right)^{\frac{1}{2}}.$$

As $\delta_k = \delta/k$ for all $k \in \mathbb{Z}_{>0}$, the standard zeta function inequalities (5.1.18) imply that

$$\begin{aligned}\mathbb{E}[\|\nabla f(x_{k'})\|_2] &\leq nC_0\delta^2\tfrac{\pi^2}{3} + \left(2nL_1F + nC_0^2\delta^4\tfrac{\pi^4}{90} + n^2C_0^2\delta^4\tfrac{\pi^4}{9}\right)^{\frac{1}{2}} \\ &\leq \sqrt{2nL_1F} + nC_0\delta^2(\tfrac{2\pi^2}{3} + \tfrac{\pi^2}{\sqrt{90}}),\end{aligned}$$

where the second inequality holds because $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$ for all $a, b, c \geq 0$ and because $\sqrt{n} \leq n$ for all $n \in \mathbb{Z}_{\geq 0}$. Averaging the second inequality in (5.1.26) across all $k \leq K$ and using the above upper bound on $\mathbb{E}[\|\nabla f(x_k)\|_2]$ for each $k \leq K$ finally yields

$$\begin{aligned}\tfrac{1}{K}\textstyle\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \leq & \tfrac{n}{K}\Big[2L_1F + 4C_0\textstyle\sum_{k=1}^{K}\delta_k^2\left(\sqrt{2nL_1F} + nC_0\delta^2(\tfrac{2\pi^2}{3} + \tfrac{\pi^2}{\sqrt{90}})\right) \\ & + C_0^2\textstyle\sum_{k=1}^{K}\delta_k^4\Big].\end{aligned}$$

Applying the zeta function inequalities (5.1.18) once again and recalling that $\delta_k^2 \leq 1$ for all $k \in \mathbb{Z}_{>0}$, it is then easy to construct a constant $C \geq 0$ such that
$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(x_k)\|_2^2] \leq \frac{n}{K}(2L_1F + \delta^2C)$. □

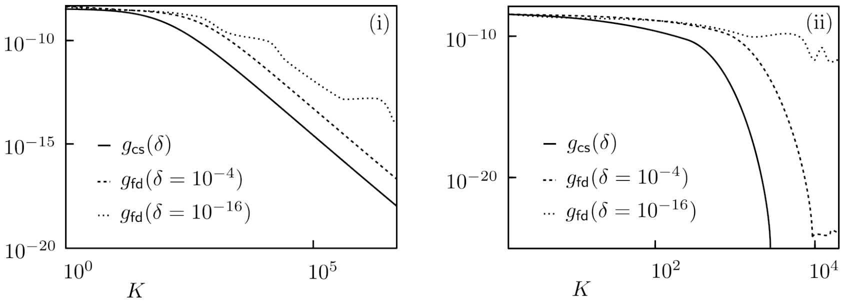


Figure 5.2: Example 5.1.7, (i) suboptimality gap $f(\bar{x}_K) - f^\star$ for (5.1.27) and (ii) suboptimality gap $f(x_K) - f^\star$ for (5.1.27). Figure adapted from the original Matlab figure [JYK21, Fig. 7.1].

By Theorem 5.1.14, we can enforce $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(x_k)\|_2^2] \leq \epsilon$ for a given $\epsilon > 0$ by selecting a smoothing parameter $\delta \leq O(\sqrt{K\epsilon/n})$ and by running Algorithm 1 over $O(nL_1F/\epsilon)$ iterations.

Numerical example

For an extensive numerical section, we point to [JYK21, Sec. 7], here, we showcase just a simple example.

Specifically, we will compare the proposed complex-step estimator g_{cs} defined in (5.1.13) against the forward-difference estimator

$$g_{\mathsf{fd}}(x, \delta) = \frac{1}{\delta}(f(x + \delta y) - f(x))y \quad \text{with} \quad y \sim \mathcal{N}(0, I_n)$$

which relies on Gaussian smoothing [NS17, Eq. (30)]. When using g_{fd}, we set the stepsize of Algorithm 1 to $\mu_k = 1/(4(n+4)L_1)$ as recommended in [NS17, Eq. (55)]. When using g_{cs}, on the other hand, we select the stepsize in view of the structural properties of the given objective function f in accordance with Theorem 5.1.12. The experiments are performed in MATLAB on a x86_64 machine with a 4 GHz CPU and 16 GB RAM, using double precision, that is, machine precision is $2^{-52} \approx 2.2204 \cdot 10^{-16}$.

Example 5.1.7 (Quadratic test function)**.** Assume that $\mathcal{X} = \mathbb{R}^n$ and f is an ill-conditioned version of what Nesterov calls the '*worst function in the world*' [Nes03, Sec. 2.1.2], that is, assume that

$$f(x) = L\left(\tfrac{1}{2}\left[(x^{(1)})^2 + \textstyle\sum_{j=1}^{n-1}(x^{(j+1)} - x^{(j)})^2 + (x^{(n)})^2\right] - x^{(1)}\right), \tag{5.1.27}$$

where $n = 5$, $L = 10^{-8}$, and $x^{(j)}$ denotes the j^{th} component of x for any $j \leq n$. One can show that ∇f has Lipschitz modulus $L_1 = 4L$ and that the unique global minimizer $x^\star$ of f has coordinates $(x^\star)^{(j)} = 1 - j/(n+1)$. In this case, the theoretical convergence guarantees of Algorithm 1 are independent of whether g_{cs} or g_{fd} is used. However, starting from $x_1 = 0$ and gradually reducing the smoothing parameter δ towards machine precision exposes the advantages of the one-point estimator g_{cs} over the multi-point estimator g_{fd}.

Figure 5.2 visualizes the suboptimality gap of $\bar{x}_K$ and x_K as a function of K along a single sample trajectory, respectively. For g_{cs} the performance is independent of selecting $\delta \in [10^{-16}, 10^{-4}]$. Note that especially the performance of x_K is significantly better when g_{cs} is used.

5.2 Further topics in imaginary zeroth-order optimization

The previous section introduced our zeroth-order optimization framework. In this section we comment on noise and several other generalizations. For the full details and further references, we point to [Jon21].

In contrast to the section above, we now allow for the presence of (computational) noise.

Assumption 5.2.1 (Stochastic complex oracle)**.** *Consider some unknown function $f \in C^{\omega}(\mathcal{D})$ which admits a holomorphic extension to $\Omega \subseteq \mathbb{C}^n$. We assume to have access to an oracle which outputs $\Re(f(z)) + \xi$ and $\Im(f(z)) + \xi$ for any $z \in \Omega$ with ξ a zero-mean random variable supported on $\Xi \subseteq \mathbb{R}$ with $\mathbb{E}[\xi^2] \leq \sigma_\xi$ for some $\sigma_\xi > 0$.*

Assumption 5.2.1 is particularly important in the simulation-based context. As there the evaluation of $f(z)$ might pertain to millions of floating-point operations, chopping and round-off errors are easily introduced. Again, the set Ω will be specified later on. We will make no further assumptions regarding the distribution of ξ.

In what follows we show that catastrophic numerical cancellation errors are also inevitable in the widely used noisy multi-point case. We will show that this non-deterministic setting also benefits from the imaginary gradient estimator as proposed in Section 5.1. Using this *single*-point estimator and building upon [HRB08, APT20], we provide the non-asymptotic analysis for a variety of algorithms. Specifically, we consider for strongly convex functions the *unconstrained*, *constrained*, *online* and *quadratic* cases. In the last setting we can show that the algorithm is *rate-optimal*. To comply with zeroth-order knowledge we also propose an estimation scheme for the strong-convexity parameter. As an outlook we provide a local result in the nonconvex case. Besides, we generalize some results from Section 5.1.

Notation will be largely the same as in Section 5.1, however, as we not only consider $\mathbb{B}^n$ and its boundary to smooth over, we need more general equipment. Let $\mathsf{Y} \subset \mathbb{R}^n$ be a Borel measurable set such that $\partial\mathsf{Y}$ is an orientable compact differentiable manifold. We write $y \sim \mathsf{Y}$ to declare that y is a random vector following the uniform distribution on Y, and for any Borel measurable function $g : \mathsf{Y} \subset \mathbb{R}^n \to \mathbb{R}$ we denote by

$$\mathbb{E}_{y\sim\mathsf{Y}}[g(y)] = 1/\mathrm{vol}(\mathsf{Y}) \int_{\mathsf{Y}} g(y)\mathrm{d}V(y)$$

the expected value of $g(y)$, where $\mathrm{d}V$ represents the Borel measure induced by the ***volume form*** on Y, and $\mathrm{vol}(\mathsf{Y})$ represents the volume of Y.

5.2.1 On the gradient estimator

Under Assumption 5.2.1, Proposition 5.1.8 provides us immediately with a (noisy) single-point estimator of $\nabla f_\delta(x)$, namely

$$g_\delta(x) = \frac{n}{\delta}\Im\left(f(x+i\delta u)\right)u + \frac{n}{\delta}\xi u, \quad u \sim \mathbb{S}^{n-1} \tag{5.2.1}$$

for some noise term $\xi \in \Xi$. In contrast to the noise-free setting in Section 5.1, equation (5.2.1) immediately reveals the delicacy in selecting $\delta \in \mathbb{R}_{>0}$. Note, the term n/δ follows from our choice to average over $\mathbb{B}^n$, *i.e.*, by (5.1.10). Below we will clarify that this term, and thereby the offset due to the noise, cannot be decreased by any other choice of solid. In that sense, $\mathbb{B}^n$ is *geometrically optimal.* Again, we will use (5.2.1) in gradient descent algorithms of the form $x_{k+1} = x_k - \mu_k g_{\delta_k}(x_k)$, as detailed in Algorithm 1, for $\mu_k \in \mathbb{R}$ a stepsize and $\delta_k \in \mathbb{R}_{>0}$ the smoothing parameter.

The next assumption on the (computational) noise will be assumed throughout.

Assumption 5.2.2 (Independence)**.** *The random variable ξ is drawn independently of $u \sim \mathbb{S}^{n-1}$.*

In general $\mathbb{E}[g_\delta(x)] \neq \nabla f(x)$, thus, there will be a bias, controlled in part by selecting the sequence $\{\delta_k\}_{k\in\mathbb{Z}_{\geq 0}}$ and unfortunately, a fixed bias prohibits (local) convergence in general [AS21]. However, by looking at (5.2.1), it can be shown that to overcome this, a selection of $\{\mu_k\}_{k\in\mathbb{Z}_{>0}}$ and $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ should satisfy the following;

(i) As $\mu_k = \Theta(k^{-1})$ [RSS12], for fixed $\delta_k = \delta > 0$ a bias term prevails of the form $\sum_{k=1}^{K}\mu_k\delta = O(\log(K)+1)$. This can be avoided by selecting δ_k to be asymptotically vanishing.

(ii) However, as the data is noisy, a term of the form μ_k/δ_k also accumulates. As such, by (i) $\delta_k \to 0$, but slower than $\mu_k \to 0$.

With this in mind we see that when $\mathbb{E}[g_\delta] \neq \nabla f$ zeroth-order optimization algorithms resort to selecting the smoothing-parameter sequence $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ such that δ_1 converges to 0, but sufficiently slow, *cf.* [NG22, Thm. 1], [BG22, Thm. 3]. See also [Fab71], [Spa05, Ch. 6], [WZS21, Assump. 1] for similar assumptions from the stochastic approximation viewpoint. Motivated by the observation that $\delta_k \to 0$ is necessary for an abundance of algorithms, this section provides a framework that can handle this requirement numerically. That means, a framework where δ_k can be made arbitrarily small[5].

At last, to characterize the effectiveness of our algorithms, we need to bound the second moment of the estimator (5.2.1) again.

Lemma 5.2.3 (Estimator second moment)**.** *Let $f \in C^{\omega,2}_{L_2(f)}(\mathcal{D})$ satisfy Assumption 5.1.6 for some $\bar{\delta} \in (0,1)$ and $L_2(f) \geq 0$. Then, for any fixed $x \in \mathcal{D}$, $\kappa \in (0,1)$ and $g_\delta(x)$ as in (5.2.1) there are constants $C_a, C_b \geq 0$, vanishing with $L_2(f)$, such that for any $\delta \in (0,\kappa\bar{\delta}]$ one has*

$$\mathbb{E}\left[\|g_\delta(x)\|_2^2\right] \leq C_a n^2\delta^4 + C_b n^2\delta^2\|\nabla f(x)\|_2 + n\|\nabla f(x)\|_2^2 + \frac{n^2}{\delta^2}\sigma_\xi. \tag{5.2.2}$$

[5] Again, this means up to what the machine at hand can produce, usually $2^{-1023} \approx 10^{-308}$.

Proof. First, observe from Algorithm 1, Assumption 5.2.1 and Assumption 5.2.2 that

$$\mathbb{E}\left[\|g_\delta(x)\|_2^2\right] = \frac{n^2}{\delta^2}\mathbb{E}_{u\sim\mathbb{S}^{n-1}}\left[\left(\Im\left(f(x+i\delta u)\right)\right)^2\right] + \frac{n^2}{\delta^2}\mathbb{E}_\xi[\xi^2].$$

Then, the claim follows directly by the same reasoning as for Corollary 5.1.10. □

As with standard gradient-descent, the more isotropic the level sets of the objective are, the better. The common way to enforce this is by means of changing the underlying metric via the Hessian, *i.e.*, Newton's method. With this in mind, averaging over some solid ellipsoid might appear more beneficial than averaging over the ball. In the spirit of [HL14] and [HPGS16, Prop. 3, Lem. 4] we generalize Proposition 5.1.8 to more generic solids and show—perhaps unsurprisingly—that spherical smoothing is optimal in the sense that it minimizes the offset due to noise in (5.2.2).

To be in line with Assumption 5.1.6 we assume that this generic solid M is a subset of $[-1,1]^n$.

Lemma 5.2.4 (The gradient of the complex-step function for smooth solids)**.** *Let* $\mathsf{M} \subset [-1,1]^n \subset \mathbb{R}^n$ *be diffeomorphic to* $\mathbb{B}^n$*. Let* $f \in C^\omega(\mathcal{D})$ *satisfy Assumption 5.1.6 for some* $\bar{\delta} \in (0,1)$*, then,* $f_{\delta,\mathsf{M}}$ *as in*

$$f_{\delta,\mathsf{M}}(x) = \mathbb{E}_{v\sim\mathsf{M}}\left[\Re(f(x+i\delta v)\right] \tag{5.2.3a}$$

is differentiable and for any $x \in \mathcal{D}$ *we have for any* $\delta \in (0,\bar{\delta})$

$$\nabla f_{\delta,\mathsf{M}}(x) = \frac{\mathrm{vol}(\delta\partial\mathsf{M})}{\mathrm{vol}(\delta\mathsf{M})}\cdot\mathbb{E}_{u\sim\partial\mathsf{M}}\left[\Im\left(f(x+i\delta u)\right)N(u)\right]. \tag{5.2.3b}$$

for $N(u)$ *a unit normal in* $T_u^\perp\partial\mathsf{M}$*.*

Proof. As $\mathsf{M} \subset \mathbb{R}^n$ is a compact oriented manifold with boundary, we can appeal to the *Divergence theorem* [Lee13, Thm. 16.32] (under the Euclidean metric), which states that for any smooth vector field X on M one has

$$\int_{\mathsf{M}} \mathrm{div}(X(v))dV(v) = \int_{\partial\mathsf{M}} \langle X(u), N(u)\rangle dV(u), \tag{5.2.4}$$

for N denoting the unit normal vector (field) along $\partial\mathsf{M}$. That is, $\mathbb{R}^n = T_p\partial\mathsf{M} \oplus T_p^\perp\partial\mathsf{M}$ for all $p \in \partial\mathsf{M}$ and $N(p) \in T_p^\perp\partial\mathsf{M}$.

Using the same reasoning as for example in [JYK21], since one can select $X = h\cdot C$ for h some smooth function and C some constant vector field on M, then, as $\mathrm{div}(C) = 0$ and we can select C to be aligned with any coordinate axis, (5.2.4) implies that

$$\int_{\mathsf{M}} \nabla h(v)dV(v) = \int_{\partial\mathsf{M}} h(u)N(u)dV(u). \tag{5.2.5}$$

Now we obtain the generalization of the result in [JYK21], that is, by compactness, the Dominated Convergence theorem [Fol99, Sec. 2.3], the Divergence theorem (5.2.4) and the Cauchy-Riemann equations [Kra00] we get

$$\nabla_x \int_{\delta\mathsf{M}} \Re\left(f(x+iv)\right)\mathrm{d}V(v) = \int_{\delta\partial\mathsf{M}} \Im\left(f(x+iu)\right)N(u)\mathrm{d}V(u),$$

e.g., see [JYK21] for more on this line of reasoning. Then, due to the distributional assumption (uniformity), we write

$$f_{\delta,\mathsf{M}}(x) = \mathbb{E}_{v\sim\mathsf{M}}\left[\Re\left(f(x+i\delta v)\right)\right] = 1/\mathrm{vol}(\delta\mathsf{M})\int_{\delta\mathsf{M}}\Re\left(f(x+iv)\right)\mathrm{d}V(v),$$

and similarly,

$$\mathbb{E}_{u\sim\partial\mathsf{M}}\left[\Im\left(f(x+i\delta u)\right)N(u)\right] = 1/\mathrm{vol}(\delta\partial\mathsf{M})\int_{\delta\partial\mathsf{M}}\Im\left(f(x+iu)\right)N(u)\mathrm{d}V(u).$$

Combining it all yields (5.2.3b). □

As $N(u)\in T_u^{\perp}\partial\mathsf{M}$ is a unit vector, the offset term in the variance (5.2.2) is minimized when we select M as

$$\arg\min_{\mathsf{M}\in\mathscr{M}}\frac{\mathrm{vol}(\delta\partial\mathsf{M})}{\mathrm{vol}(\delta\mathsf{M})}, \tag{5.2.6}$$

where $\mathscr{M}$ is the set of manifolds in $[-1,1]^n$ diffeomorphic to $\mathbb{B}^n$ and $\delta\in\mathbb{R}_{>0}$. To retrieve the optimizer, consider the isoperimetric inequality in $\mathbb{R}^n$ [Oss78] which implies that $\mathsf{M}^\star=\mathbb{B}^n$ is optimal in the sense of (5.2.6).

To get (the complex-step version of) [HL14, Cor. 6] from Lemma 5.2.4, let $\mathcal{E}^n_Q=\{x\in\mathbb{R}^n:\langle Q^{-1}x,x\rangle\leq 1\}$ for some $Q\in\mathcal{S}^n_{\succ 0}$. Now, $T_p\partial\mathcal{E}^n_Q=\{v\in\mathbb{R}^n:\langle Q^{-1}p,v\rangle=0\}$. As $\mathcal{E}^n_Q=Q^{1/2}\mathbb{B}^n$ one can write

$$f_{\delta,\mathcal{E}^n_Q}(x)=\mathbb{E}_{v\sim\mathcal{E}^n_Q}\left[f(x+i\delta v)\right]=\mathbb{E}_{v\sim\mathbb{B}^n}\left[f(x+i\delta Q^{1/2}v)\right]. \tag{5.2.7a}$$

Via the rightmost term in (5.2.7a) and the proof of Lemma 5.2.4 it follows immediately that

$$\nabla f_{\delta,\mathcal{E}^n_Q}(x)=\mathbb{E}_{u\sim\mathbb{S}^{n-1}}\frac{n}{\delta}\left[f(x+i\delta Q^{1/2}u)Q^{-1/2}u\right]. \tag{5.2.7b}$$

Equivalently, one can directly appeal to (5.2.3b). However, here one needs to appeal to the isoperimetric ratio for ellipsoids [Riv07].

5.2.2 On the addition of computational noise

In this section we will utilize the imaginary gradient estimator g_δ as given by (5.2.1) in the context of zeroth-order optimization algorithms. We will not focus (again) on fully generic convex optimization problems as the flat parts of non-trivial real-analytic convex functions must have measure zero [Kra00]. Hence, without too much loss of generality we omit convex functions which are not strongly convex[6]. See also [KSST09] for more on strong-convexity in the context of generalization.

In this section we relax some of the assumptions in Section 5.1, not only can we handle computational noise, the algorithms demand less knowledge of the problem compared to other work. This is possible by introducing a time-varying stepsize and a construction

[6] Future work will highlight the intimate relation between convex and strongly convex functions under the assumption that both are real analytic. Initial work was done together with Helia Atarod.

very much in line with [APT20]. In fact, recall from [RSS12] that $\mu_k = \Theta(k^{-1})$ to allow for optimal rates. The edge our results have, however, over these existing works is that our sequence of smoothing parameters $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ is *never*[7] catastrophic.

In contrast to Section 5.1, our algorithms "*only*" demand knowledge of the strong-convexity parameter. We comment in [Jon21, Sec. A.2] on the estimation of $\tau(f)$. With some abuse of notation we will again use $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_{k\in\mathbb{Z}_{>0}}, \mathbb{P})$ to denote our probability space at hand.

Generic convergence rates

As in [APT20], we start with the constrained case.

Theorem 5.2.5 (Convergence rate of Algorithm 1 (constrained) with noise)**.** *Let $f \in C^{\omega}(\mathcal{D})$ be a $\tau(f)$-strongly convex function satisfying Assumption 5.1.6, let $\mathcal{K} \subset \mathcal{D}$ be a compact convex set and suppose that f has a Lipschitz Hessian over $\mathcal{K}$. Let $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ be the sequence of iterates generated by Algorithm 1 with stepsize $\mu_k = 2/(\tau(f)k)$ and smoothing parameters $\delta_k = \delta k^{-1/6}$ with $\delta \in (0, \kappa\bar{\delta}]$ for $\kappa \in (0,1)$. Then, if the oracle satisfies Assumption 5.2.1 and $K \geq 1$, $\bar{x}_K = K^{-1}\sum_{k=1}^{K} x_k$ satisfies*

$$\mathbb{E}[f(\bar{x}_K) - f(x^\star)] \leq \widetilde{O}\left(\frac{n^2}{\tau(f)}\delta^{-1/3}\sigma_\xi K^{-2/3}\right).$$

Proof. We mainly follow [APT20]. To that end, let $\sup_{x\in\mathcal{K}}\|\nabla f(x)\|_2 \leq G$. As $\mathcal{K}$ is convex and compact we have by the contractive property of $\Pi_{\mathcal{K}}$ that $\|x_{k+1} - x^\star\|_2^2 \leq \|x_k - \mu_k g_{\delta_k}(x_k) - x^\star\|_2^2$. This can be written conveniently as

$$\langle g_{\delta_k}(x_k), x_k - x^\star\rangle \leq \tfrac{1}{2\mu_k}\left(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2\right) + \tfrac{\mu_k}{2}\|g_{\delta_k}(x_k)\|_2^2. \tag{5.2.8}$$

After reordering the standard strong $\tau(f)$-convexity expression, one obtains

$$f(x_k) - f(x^\star) \leq \langle \nabla f(x_k), x_k - x^\star\rangle - \tfrac{\tau(f)}{2}\|x_k - x^\star\|_2^2. \tag{5.2.9}$$

Set $a_k = \|x_k - x^\star\|_2^2$, then, an application of the Cauchy-Schwarz inequality after combining (5.2.8) with (5.2.9) and taking the expectation over u_k and ξ_k conditioned on x_k yields

$$\begin{aligned}
\mathbb{E}[f(x_k) - f(x^\star)|\mathcal{F}_k] &\leq \|\mathbb{E}[g_{\delta_k}(x_k)|\mathcal{F}_k] - \nabla f(x_k)\|_2\,\|x_k - x^\star\|_2 \\
&\quad -\tfrac{\tau(f)}{2}\mathbb{E}[a_k|\mathcal{F}_k] + \tfrac{1}{2\mu_k}\mathbb{E}[a_k - a_{k+1}|\mathcal{F}_k] \\
&\quad +\tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2|\mathcal{F}_k] \\
&\overset{(5.1.14)}{\leq} C_1 n\delta_k^2\|x_k - x^\star\|_2 + \tfrac{1}{2\mu_k}\mathbb{E}[a_k - a_{k+1}|\mathcal{F}_k] \\
&\quad +\tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2|\mathcal{F}_k] - \tfrac{\tau(f)}{2}a_k,
\end{aligned}$$

[7] Up to the limits of the machine.

for some $C_1 > 0$. Now, use $ab \leq \frac{1}{2}(a^2+b^2)$, in particular $ab \leq \frac{1}{2}(\gamma a^2 + \gamma^{-1}b^2)$ for $\gamma \neq 0$, to construct

$$n\delta_k^2\|x_k - x^\star\|_2 \leq \tfrac{1}{2}\left(\tfrac{2C_1}{\tau(f)}n^2\delta_k^4 + \tfrac{\tau(f)}{2C_1}\|x_k - x^\star\|_2^2\right).$$

Next, take unconditional expectation and let $r_k = \mathbb{E}[a_k]$ such that we can write

$$\mathbb{E}[f(x_k) - f(x^\star)] \leq \tfrac{1}{2\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{4}r_k + \tfrac{1}{\tau(f)}C_1^2n^2\delta_k^4 + \tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2]. \quad (5.2.10)$$

Summing (5.2.10) over k yields

$$\begin{aligned}\textstyle\sum_{k=1}^K \mathbb{E}[f(x_k) - f(x^\star)] \leq{}& \tfrac{1}{2}\textstyle\sum_{k=1}^K \left(\tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{2}r_k\right)\\ &+ \textstyle\sum_{k=1}^K\left(\tfrac{1}{\tau(f)}C_1^2n^2\delta_k^4 + \tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2]\right).\end{aligned}$$

As we selected $\mu_k = 2/(\tau(f)k)$ we can simplify the above by using the same reasoning as in [APT20], that is

$$\begin{aligned}\textstyle\sum_{k=1}^K\left(\tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{2}r_k\right) \leq{}& r_1\left(\tfrac{1}{\mu_1} - \tfrac{\tau(f)}{2}\right)\\ &+ \textstyle\sum_{k=2}^K r_k\left(\tfrac{1}{\mu_k} - \tfrac{1}{\mu_{k-1}} - \tfrac{\tau(f)}{2}\right) = 0.\end{aligned}$$

Note that we rely on the $\tau(f)$-strong convexity. Using the observation from above and plugging in the stepsize μ_k elsewhere yields by (5.2.2)

$$\begin{aligned}\textstyle\sum_{k=1}^K \mathbb{E}[f(x_k) - f(x^\star)] \leq{}& \tfrac{1}{\tau(f)}\textstyle\sum_{k=1}^K\left(C_1^2n^2\delta_k^4 + \tfrac{1}{k}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2]\right)\\ \leq{}& \tfrac{n^2}{\tau(f)}\textstyle\sum_{k=1}^K\Big(C_1^2\delta_k^4 + \tfrac{1}{k}\Big[C_2\delta_k^4 + C_3\delta_k^2\|\nabla f(x_k)\|_2\\ &+ \tfrac{1}{n}\|\nabla f(x_k)\|_2^2 + \tfrac{1}{\delta_k^2}\sigma_\xi\Big]\Big),\end{aligned}$$

for some $C_2, C_3 > 0$. Now, minimizing over $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ is possible but yields smoothing parameters as a function of unknown constants. Instead, we retain the "*optimal*" root[8] and propose

$$\widetilde{\delta}_k = \left(\frac{\alpha\sigma_\xi}{k}\right)^{1/6},$$

for some $\alpha \in (0,1)$ to be specified. Using this smoothing parameter sequence, that is, $\delta_k = \widetilde{\delta}_k$, together with $\sum_{k=1}^K k^{-1} \leq 1 + \log(K)$ (Lemma 5.3.1) yields

$$\begin{aligned}\textstyle\sum_{k=1}^K \mathbb{E}[f(x_k) - f(x^\star)] \quad\leq\quad& \tfrac{n^2}{\tau(f)}\textstyle\sum_{k=1}^K\left(C_1^2\left(\tfrac{\alpha\sigma_\xi}{k}\right)^{2/3} + \tfrac{1}{k}\left[C_2\left(\tfrac{\alpha\sigma_\xi}{k}\right)^{2/3} + \left(\tfrac{\alpha\sigma_\xi}{k}\right)^{-1/3}\sigma_\xi\right]\right)\\ &+\tfrac{n}{\tau(f)}G^2(1+\log(K)) + \tfrac{n^2}{\tau(f)}GC_3\textstyle\sum_{k=1}^K\tfrac{1}{k}\left(\tfrac{\alpha\sigma_\xi}{k}\right)^{1/3}\\ =\quad& \tfrac{n^2}{\tau(f)}\textstyle\sum_{k=1}^K\left(C_1^2\left(\tfrac{\alpha\sigma_\xi}{k}\right)^{2/3} + \tfrac{1}{k}\left[C_2\left(\tfrac{\alpha\sigma_\xi}{k}\right)^{2/3} + k^{1/3}\sigma_\xi^{2/3}\alpha^{-1/3}\right]\right)\\ &+\tfrac{n}{\tau(f)}G^2(1+\log(K)) + \tfrac{n^2}{\tau(f)}GC_3\textstyle\sum_{k=1}^K k^{-2/3}(\alpha\sigma_\xi)^{1/3}\\ \leq\quad& \tfrac{n^2}{\tau(f)}\textstyle\sum_{k=1}^K C_4k^{-2/3}\sigma_\xi^{2/3}\alpha^{-1/3}\\ &+\tfrac{n}{\tau(f)}G^2(1+\log(K)) + \tfrac{n^2}{\tau(f)}GC_3\textstyle\sum_{k=1}^K k^{-2/3}(\alpha\sigma_\xi)^{1/3}.\end{aligned}$$

[8] Let $a, b \in \mathbb{R}_{>0}$, then, see that $(b/(2a))^{1/6} = \arg\min_{\delta\in\mathbb{R}_{\geq 0}}\{a\delta^4 + b\frac{1}{\delta^2}\}$.

Now, as $\sum_{k=1}^{K} k^{-2/3} \leq 3K^{1/3}$ (Lemma 5.3.2) we can continue and write

$$\begin{aligned}\textstyle\sum_{k=1}^{K} \mathbb{E}[f(x_k) - f(x^\star)] &\leq \tfrac{n^2}{\tau(f)} C_5 K^{1/3} \sigma_\xi^{2/3} \alpha^{-1/3} + \tfrac{n}{\tau(f)} G^2 (1+\log(K)) \\ &\quad + \tfrac{n^2}{\tau(f)} G C_6 K^{1/3} (\alpha\sigma_\xi)^{1/3}.\end{aligned}$$

and as such we obtain the optimization error

$$\begin{aligned}\mathbb{E}[f(\bar{x}_K) - f(x^\star)] &\leq \tfrac{n^2}{\tau(f)} C_5 K^{-2/3} \sigma_\xi^{2/3} \alpha^{-1/3} + \tfrac{n}{\tau(f)} G^2 K^{-1} (1+\log(K)) \\ &\quad + \tfrac{n^2}{\tau(f)} G C_6 K^{-2/3} (\alpha\sigma_\xi)^{1/3}.\end{aligned}$$

As $\alpha \in (0,1)$ was arbitrary, we can set $\delta = (\alpha\sigma_\xi)^{1/6}$ such that $\delta_k = \delta k^{-1/6}$ for some $\delta \in (0, \bar{\delta})$. □

The edge Theorem 5.2.5 has over existing work is that the requested sequence $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ *can* always be safely implemented. With respect to optimality, we highlight *α-suffix averaging*, as proposed in [RSS12], as a general method to pass from $\widetilde{O}(\cdot)$ to $O(\cdot)$ complexities.

Next we consider the unconstrained case. Here, we cannot appeal to a uniform bound on $\nabla f(x)$. Instead, we use the idea from [APT20, Thm. 3.2] and bound a subset of iterates before strong-convexity kicks in. The intuition being, when $\tau(f)$ is small, the first few stepsizes will be relatively large and can lead to overflow. This is to be avoided. In some sense one could interpret the multiphase algorithm as some restarting mechanism.

Theorem 5.2.6 (Convergence rate of Algorithm 1 (unconstrained) with noise)**.** *Let $f \in C^\omega(\mathcal{D})$ be a $\tau(f)$-strongly convex function satisfying Assumption 5.1.6. Suppose that f has a Lipschitz gradient and Hessian, that is,* (2.1.4) *and* (2.1.7) *hold, for non-zero constants $L_1(f)$ and $L_2(f)$, respectively. Let $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ be the sequence of iterates generated by Algorithm 1 (a) for*

$$\begin{aligned}\mu_k &= \tfrac{1}{\tau(f)K}, \quad \delta_k = \delta K^{-1/6}, \quad k = 1,\dots,K_0, \\ \mu_k &= \tfrac{2}{\tau(f)k}, \quad \delta_k = \delta k^{-1/6}, \quad k = K_0+1,\dots,K,\end{aligned}$$

with $K_0 = \left\lfloor \frac{8n^2 L_1(f)^2}{\tau(f)^2} \right\rfloor$ and $\delta \in (0, \kappa\bar{\delta}]$ for some $\kappa \in (0,1)$. Then, if the oracle satisfies Assumption 5.2.1 *and $K \geq 2K_0$, we have for $\bar{x}_{K_0,K} = \frac{1}{K-K_0}\sum_{k=K_0+1}^{K} x_k$ that*

$$\mathbb{E}[f(\bar{x}_{K_0,K}) - f(x^\star)] \leq O\left(\frac{n^2 L_1(f)^2}{\tau(f)} \|x_1 - x^\star\|_2^2 K^{-1}\right) + O\left(\frac{n^2 \sigma_\xi}{\tau(f)\delta^2} K^{-2/3}\right). \tag{5.2.11}$$

Proof. The proof will be similar to that of [APT20, Thm. 3.2]. Again, set $a_k = \|x_k - x^\star\|_2^2$, then, similar to the proof of Theorem 5.2.5, use $ab \leq \frac{1}{2}(a^2+b^2)$ together with $\tau(f)$-strong convexity, *i.e.*, (5.1.21), to construct

$$n\delta_k^2 \|x_k - x^\star\|_2 \leq \tfrac{1}{2}\left(\tfrac{2C_1}{\tau(f)} n^2 \delta_k^4 + \tfrac{\tau(f)}{2C_1}\|x_k - x^\star\|_2^2\right) \leq \tfrac{C_1}{\tau(f)} n^2 \delta_k^4 + \tfrac{1}{2C_1}(f(x_k) - f(x^\star)).$$

Next, let $r_k = \mathbb{E}[a_k]$ such that by $\|\nabla f(x_k)\|_2^2 \leq L_1(f)^2\|x_k - x^\star\|_2^2$ we can write

$$\begin{aligned}\mathbb{E}[f(x_k) - f(x^\star)] &\leq \tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tau(f)r_k + \tfrac{2}{\tau(f)}C_1^2 n^2\delta_k^4 + \mu_k\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2] \\ &\overset{(5.2.2)}{\leq} \tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tau(f)r_k + \tfrac{2}{\tau(f)}C_1^2 n^2\delta_k^4 \\ &\quad +\mu_k\left(C_2 n^2\delta_k^4 + C_3 n^2\delta_k^4 + 2n^2 L_1(f)^2 r_k + \tfrac{n^2}{\delta_k^2}\sigma_\xi\right)\end{aligned} \tag{5.2.12}$$

Where in the last step we used

$$\delta_k^2\|x_k - x^\star\|_2 \leq \tfrac{1}{2}\left(\tfrac{C_3}{2L_1(f)}\delta_k^4 + \tfrac{2L_1(f)}{C_3}\|x_k - x^\star\|_2^2\right)$$

to rewrite (5.2.2). Now we use the step- and smoothing parameters for $k = 1, \dots, K_0$, that is, $\mu_k = 1/(\tau(f)K)$ and $\delta_k = \delta K^{-1/6}$ to observe from (5.2.12) that

$$\begin{aligned}r_{k+1} &\leq r_k - \tau(f)\mu_k r_k + \tfrac{2\mu_k}{\tau(f)}C_1^2 n^2\delta_k^4 + \mu_k^2\left(C_2 n^2\delta_k^4 + C_3 n^2\delta_k^4 + 2n^2 L_1(f)^2 r_k + \tfrac{n^2}{\delta_k^2}\sigma_\xi\right) \\ &= \left(1 - \tfrac{1}{K} + \tfrac{2L_1(f)^2}{(\tau(f)K)^2}n^2\right) r_k + \nu_K = A_K r_k + \nu_K\end{aligned}$$

for A_k as between brackets and ν_K defined as

$$\begin{aligned}\nu_K =& \tfrac{2}{\tau(f)^2 K}C_1^2 n^2\delta_k^4 + \tfrac{1}{(\tau(f)K)^2}\left(C_2 n^2\delta_k^4 + C_3 n^2\delta_k^4 + \tfrac{n^2}{\delta_k^2}\sigma_\xi\right) \\ \leq& \tfrac{1}{\tau(f)^2 K}\left(n^2\delta^4 C_4 + \tfrac{n^2}{\delta^2}\sigma_\xi\right) K^{-2/3}.\end{aligned}$$

We now proceed with bounding r_{K_0+1}. As in [APT20], set

$$q_K = 1 + \tfrac{2L_1(f)^2}{(\tau(f)K)^2}n^2 \geq A_K,$$

then, by iterating over r_k it follows from a geometric series argument that

$$r_{K_0+1} \leq A_K^{K_0} r_1 + \textstyle\sum_{i=0}^{K_0-1} A_K^i \nu_K \leq \left(r_1 + \tfrac{(\tau(f)K)^2}{2L_1(f)^2 n^2}\nu_K\right) q_K^{K_0}.$$

Now for $\lfloor \cdot \rfloor$ being the floor function, let K_0 be as in the theorem. Then, as $\log(1+x) \leq x$, on $\mathbb{R}_{\geq 0}$

$$\begin{aligned}q_K^{K_0} = \exp\left(K_0 \log\left(1 + \tfrac{2L_1(f)^2}{(\tau(f)K)^2}n^2\right)\right) &\leq \exp\left(\tfrac{8n^2 L_1(f)^2}{\tau(f)^2}\log\left(1 + \tfrac{2L_1(f)^2}{(\tau(f)K)^2}n^2\right)\right) \\ &\leq \exp\left(\tfrac{16n^4 L_1(f)^4}{\tau(f)^4 K^2}\right).\end{aligned}$$

Fix any $C_e \in (0, \frac{1}{32})$, when

$$K = \sqrt{\tfrac{8n^4 L_1(f)^4}{\tau(f)^4 C_e}}$$

then $K \geq 2K_0$ and $q_K^{K_0} \leq e^{2C_e} = C_5$. As such,

$$\begin{aligned}r_{K_0+1} &\leq \left(r_1 + \tfrac{(\tau(f)K)^2}{2L_1(f)^2 n^2}\nu_K\right) C_5 \\ &\leq \left(r_1 + \tfrac{(\tau(f)K)^2}{L_1(f)^2 n^2}\tfrac{1}{\tau(f)^2 K}\left(n^2\delta^4 C_4 + \tfrac{n^2}{\delta^2}\sigma_\xi\right) K^{-2/3}\right) C_5 \\ &= \left(r_1 + \tfrac{1}{L_1(f)^2 n^2}\left(n^2\delta^4 C_4 + \tfrac{n^2}{\delta^2}\sigma_\xi\right) K^{1/3}\right) C_5.\end{aligned}$$

Now we return to our normal step- and smoothing parameters, that is $\mu_k = 2/(\tau(f)k)$, $\delta_k = \delta k^{-1/6}$, for $k \geq K_0 + 1$. By plugging this into (5.2.12) we get

$$\begin{aligned}(K-K_0)\mathbb{E}[f(\bar{x}_{K_0,K}) - f(x^\star)] \leq & \textstyle\sum_{k=K_0+1}^{K} \frac{\tau(f)k}{2}(r_k - r_{k+1}) - \tau(f)r_k + \frac{4}{\tau(f)k}n^2 L_1(f)^2 r_k \\ & + \textstyle\sum_{k=K_0+1}^{K} \frac{2}{\tau(f)} C_6 n^2 \delta^4 k^{-2/3} \\ & + \textstyle\sum_{k=K_0+1}^{K} \frac{2}{\tau(f)k}\frac{n^2}{\delta^2} k^{1/3}\sigma_\xi\end{aligned}$$

By construction of K_0 we have that for $k \geq K_0+1$, $\tau(f)/2 \geq (4n^2 L_1(f)^2)/(\tau(f)k)$. Hence

$$(K-K_0)\mathbb{E}[f(\bar{x}_{K_0,K}) - f(x^\star)] \leq \tfrac{\tau(f)}{2}\left(\textstyle\sum_{k=K_0+1}^{K} k(r_k - r_{k+1}) - r_k\right) + U_{K_0,K}.$$

where by Lemma 5.3.2

$$\begin{aligned}U_{K_0,K} =& \textstyle\sum_{k=K_0+1}^{K} \frac{2}{\tau(f)} C_6 n^2 \delta^4 k^{-2/3} + \frac{2}{\tau(f)k}\frac{n^2}{\delta^2} k^{1/3}\sigma_\xi \\ =& \tfrac{2n^2}{\tau(f)}\left(C_6\delta^4 + \tfrac{1}{\delta^2}\sigma_\xi\right)\textstyle\sum_{k=K_0+1}^{K} k^{-2/3} \\ \leq& \tfrac{2n^2}{\tau(f)}\left(C_6\delta^4 + \tfrac{1}{\delta^2}\sigma_\xi\right)3K^{1/3}.\end{aligned}$$

As demonstrated in [APT20], one can now construct the bound $\sum_{k=K_0+1}^{K} k(r_k - r_{k+1}) - r_k \leq K_0 r_{K_0+1}$ where the last term is exactly the term we could bound above. In combination with the bound on K_0 itself, we find that

$$\begin{aligned}(K-K_0)\mathbb{E}[f(\bar{x}_{K_0,K}) - f(x^\star)] \leq & \tfrac{\tau(f)}{2}\tfrac{8n^2 L_1(f)^2}{\tau(f)^2}\left(r_1 + \tfrac{1}{L_1(f)^2 n^2}\left(n^2\delta^4 C_4 + \tfrac{n^2}{\delta^2}\sigma_\xi\right)K^{1/3}\right)C_5 \\ & + \tfrac{2n^2}{\tau(f)}\left(C_6\delta^4 + \tfrac{1}{\delta^2}\sigma_\xi\right)3K^{1/3}.\end{aligned}$$

By our selection of C_e we have that $K \geq 2K_0$ and as such

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{K_0,K}) - f(x^\star)] \leq & \tfrac{8n^2 L_1(f)^2}{\tau(f)K}\left(r_1 + \tfrac{1}{L_1(f)^2}\left(\delta^4 C_3 + \tfrac{1}{\delta^2}\sigma_\xi\right)K^{1/3}\right)C_5 \\ & + \tfrac{4n^2}{\tau(f)K}\left(C_6\delta^4 + \tfrac{1}{\delta^2}\sigma_\xi\right)3K^{1/3}.\end{aligned}$$

Now, reordering terms yields (5.2.11). □

Optimal convergence rates

Now we consider the special case of f being quadratic. Here we improve upon the previous section due to exploitation of the quadratic nature of f, that is, by using $\Im(f(x+i\delta u)) = \delta\langle\nabla f(x), u\rangle$ for any $\delta > 0$.

Better yet, we see that for quadratic functions we incur optimal regret. Optimality can be shown along the lines of [Sha13], or along the lines of [AWBR09] after observing that in the quadratic case the gradient estimator $g_\delta(x)$ becomes an unbiased estimator for $\nabla f(x)$. The test function used in [APT20] is smooth but unfortunately not analytic[9]. We start by providing the bound from below.

[9] Section 5.2.3 highlights that this might not be an obstruction.

Theorem 5.2.7 (Bound from below)**.** *Any possibly randomized zeroth-order algorithm of fixed length $K \geq 1$, applying the estimator* (5.2.1) *under Assumption 5.2.1, cannot achieve a rate faster than*

$$\Omega\left(\frac{n^2}{\tau(f)K}\right),$$

uniformly over all $\tau(f)$-strongly convex quadratic (real-analytic) functions.

Proof. We largely follow [Sha13, Thm. 3], but for the sake of completeness we highlight the main arguments.

Recall that based on $x_1, x_2, \ldots, x_K$, in particular the function evaluations at those points, we compute some point x'_K (this could be a non-uniform average estimator). In our case the function queries correspond to $v_k = \Im(f(x_k + i\delta u))$ for some choice of $\delta > 0$, $u \in \mathbb{S}^{n-1}$ and with the possibility of being corrupted by additive noise ξ.

Now, consider the following C^ω function over $\mathbb{R}^n$

$$x \mapsto f_z(x) = \tfrac{\tau}{2}\|x\|_2^2 - \langle z, x\rangle. \tag{5.2.13}$$

The unique minimizer of $f_z(x)$ is given by $x^\star = \frac{1}{\tau}z$. Moreover, assume z is drawn uniformly from $\{-\nu, \nu\}^n$ for some ν that will be specified later. It follows from the strong τ-convexity of (5.2.13) that $f_z(x) - f_z(x^\star) \geq \frac{\tau}{2}\|x - \frac{1}{\tau}z\|_2^2$. As such, let $x'_{i,K}$ denote the i^{th} element of the vector x'_K, then, for any randomized strategy

$$\begin{aligned}\mathbb{E}_z[f_z(x'_K) - f_z(x^\star)] \geq \tfrac{\tau}{2}\mathbb{E}_z[\|x'_K - \tfrac{1}{\tau}z\|_2^2] &= \tfrac{\tau}{2}\mathbb{E}_z\left[\textstyle\sum_{i=1}^n (x'_{i,K} - \tfrac{1}{\tau}z_i)^2\right] \\ &\geq \tfrac{\nu^2}{2\tau}\mathbb{E}_z\left[\mathbb{1}_{x'_{i,K}z_i<0}\right],\end{aligned}$$

where the expectation is taken over the quadratic functions of the form (5.2.13), that is, over z. This means that we can construct a bound from below if we can get a grip on the signs of each z_i. To that end, we follow the proof of [Sha13, Thm. 3]. The idea is to consider deterministic strategies that have only access to a sequence of function evaluations. The KL-divergence will allow for relating these function evaluations and the sign of z_i.

The key difference with respect to [Sha13], however, is the estimator. Given some point x_k, our function evaluation v_k is of the form $v_k = \Im(f(x_k + i\delta u)) + \xi$ for some $\delta > 0$, $u \in \mathbb{S}^{n-1}$ and noise realization ξ. Now observe that $\Im(f_z(x + i\delta u)) = \delta(\tau\langle x, u\rangle - \langle z, u\rangle)$. Hence, conditioning on $z_i > 0$ we get $v_k = \delta(\tau\langle x_k, u\rangle - \sum_{j\neq i} z_j u_j) - \nu u_i + \xi$ whereas conditioning on $z_i < 0$ yields $v_k = \delta(\tau\langle x_k, u\rangle - \sum_{j\neq i} z_j u_j) + \nu u_i + \xi$. Under the assumption that ξ is Gaussian, which complies with Assumption 5.2.1, one can now bound the KL-divergence (between these two functions evaluations) by $(2\nu u_i)^2/(2\sigma_\xi)$, *e.g.*, see [Sha13, Lem. 5]. Using the fact that $u \in \mathbb{S}^{n-1}$ one can then exploit [Sha13, Lem. 4] and show that

$$\tfrac{\nu^2}{2\tau}\mathbb{E}_z[\mathbb{1}_{x'_{i,K}z_i<0}] \geq \tfrac{n\nu^2}{4\tau}\left(1 - \sqrt{\tfrac{2\nu^2 K}{n\sigma_\xi}}\right).$$

As such, selecting $\nu = \sqrt{(n\sigma_\xi)/(4K)}$ yields the desired result. □

In the light of Theorem 5.2.7 and [RSS12] (α-suffix averaging), the following algorithms are *rate* optimal. More specifically, one can show that the dependence on σ_ξ is also optimal.

Note that for quadratic functions we should not simply appeal to Theorem 5.2.5, as that result is true for a larger class of objective functions.

Theorem 5.2.8 (Convergence rate of Algorithm 1 (constrained) with noise, f being quadratic)**.** *Let $f \in C^{\omega}(\mathcal{D})$ be a $\tau(f)$-strongly convex function satisfying Assumption 5.1.6, let $\mathcal{K} \subset \mathcal{D}$ be a compact convex set and suppose that f has a constant Hessian over $\mathcal{K}$. Let $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ be the sequence of iterates generated by Algorithm 1 (b) with stepsize $\mu_k = 2/(\tau(f)k)$ and constant smoothing parameter $\delta_k = \delta$ with $\delta \in (0, \kappa\bar{\delta}]$ for some $\kappa \in (0,1)$. Then, if the oracle satisfies Assumption 5.2.1 and $K \geq 1$, $\bar{x}_K = K^{-1}\sum_{k=1}^{K} x_k$ satisfies*

$$\mathbb{E}[f(\bar{x}_K) - f(x^{\star})] \leq O\left(\frac{n}{\tau(f)}K^{-1}\right) + \widetilde{O}\left(\frac{n^2\sigma_{\xi}}{\tau(f)\delta^2}K^{-1}\right).$$

Proof. We can mainly follow the proof of Theorem 5.2.5, which relies itself largely on [APT20]. To that end, let again $\sup_{x\in\mathcal{K}}\|\nabla f(x)\|_2 \leq G$ and set $a_k = \|x_k - x^{\star}\|_2^2$ such that

$$\begin{aligned}
\mathbb{E}[f(x_k) - f(x^{\star})|\mathcal{F}_k] \quad &\leq \quad \|\mathbb{E}[g_{\delta_k}(x_k)|\mathcal{F}_k] - \nabla f(x_k)\|_2 \, \|x_k - x^{\star}\|_2 \\
&\quad\quad -\tfrac{\tau(f)}{2}\mathbb{E}[a_k|\mathcal{F}_k] + \tfrac{1}{2\mu_k}\mathbb{E}[a_k - a_{k+1}|x_k] \\
&\quad\quad +\tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2|\mathcal{F}_k] \\
&\overset{(5.1.14)}{\leq} \quad \tfrac{1}{2\mu_k}\mathbb{E}[a_k - a_{k+1}|\mathcal{F}_k] + \tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2|\mathcal{F}_k] - \tfrac{\tau(f)}{2}a_k.
\end{aligned}$$

Note, here we already use that f is quadratic *cf.* (5.1.14). Next, let $r_k = \mathbb{E}[a_k]$ such that we can write

$$\mathbb{E}[f(x_k) - f(x^{\star})] \leq \tfrac{1}{2\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{2}r_k + \tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2]. \tag{5.2.14}$$

To allow for an identical stepsize as before, we replace $-\tau(f)/2$ with $-\tau(f)/4$. Summing (5.2.14) over k yields

$$\textstyle\sum_{k=1}^{K}\mathbb{E}[f(x_k) - f(x^{\star})] \leq \tfrac{1}{2}\sum_{k=1}^{K}\left(\tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{2}r_k\right) + \sum_{k=1}^{K}\tfrac{\mu_k}{2}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2].$$

As we selected $\mu_k = 2/(\tau(f)k)$ we can simplify the above by using (again) the same reasoning as in [APT20]:

$$\begin{aligned}
\textstyle\sum_{k=1}^{K}\left(\tfrac{1}{\mu_k}(r_k - r_{k+1}) - \tfrac{\tau(f)}{2}r_k\right) \leq & r_1\left(\tfrac{1}{\mu_1} - \tfrac{\tau(f)}{2}\right) \\
& + \textstyle\sum_{k=2}^{K} r_k\left(\tfrac{1}{\mu_k} - \tfrac{1}{\mu_{k-1}} - \tfrac{\tau(f)}{2}\right) = 0.
\end{aligned}$$

Indeed, without the scaling of $\tau(f)$ our stepsize would have been $\mu_k = 1/(\tau(f)k)$. Note that we rely on the $\tau(f)$-strong convexity. Using the observation from above and plugging in the stepsize μ_k elsewhere yields

$$\begin{aligned}
\textstyle\sum_{k=1}^{K}\mathbb{E}[f(x_k) - f(x^{\star})] \quad &\leq \quad \tfrac{1}{\tau(f)}\textstyle\sum_{k=1}^{K}\tfrac{1}{k}\mathbb{E}[\|g_{\delta_k}(x_k)\|_2^2] \\
&\overset{(5.2.2)}{\leq} \quad \tfrac{n^2}{\tau(f)}\textstyle\sum_{k=1}^{K}\tfrac{1}{k}\left[\tfrac{1}{n}\|\nabla f(x_k)\|_2^2 + \tfrac{1}{\delta_k^2}\sigma_{\xi}\right].
\end{aligned}$$

Now, minimizing over $\{\delta_k\}_{k\in\mathbb{Z}_{>0}}$ clearly yields a desire to pick a larger and constant δ_k *cf.* Theorem 5.2.5. Combining this with the bound on $\nabla f(x_k)$ yields by (5.3.1)

$$K\mathbb{E}[f(\bar{x}_k)-f(x^\star)] \leq \textstyle\sum_{k=1}^{K}\mathbb{E}[f(x_k)-f(x^\star)] \leq \frac{n^2}{\tau(f)}\left[\frac{1}{n}G^2+\frac{1}{\delta^2}\sigma_\xi\right](\log(K)+1).$$

□

See [Jon21, Ex. 4.5] for a comparison between Theorem 5.2.8 (CS algorithm) against a state-of-the-art multi-point method [APT20, Thm. 5.1]. It turns out that for sufficiently small δ_k, the multi-point method diverges. Similar to Theorem 5.2.6, we can analyze unconstrained zeroth-order optimization when f is quadratic and obtain a rate-optimal algorithm [Jon21, Thm. 4.6].

Further remarks

Another somewhat straight-forward extension is to consider *online* optimization, see [BP16].

Less straight-forward is the extension to nonconvex objective functions. We provide the first steps in [Jon21, Thm. 5.1], with [Jon21, Ex. 5.2] showing the importance of taking the noise into account.

5.2.3 Discussion

On the necessity of leaving the real numbers

Given the results from the previous section, one might wonder if this "*complex-lifting*" is needed. Real single-point gradient estimators evidently exist, *cf.* [FKM04], but with problematic variance bounds for $\delta\to 0^+$. The common solution is to bring back some relation with the (directional) derivative [ADX10, NS17]. Hence, one might wonder if there is a purely real analogue to the complex-step derivative. The next proposition strongly hints at a negative answer.

Proposition 5.2.9 (On the necessity of leaving the real numbers)**.** *Consider some non-empty, open, set $\mathcal{D}\subseteq\mathbb{R}^n$. Then, there does not exist a continuous map $G:\mathbb{R}\to\mathbb{R}$ such that for all real-analytic functions $f:\mathcal{D}\to\mathbb{R}$*

$$G\left(f(x+\delta y)\right)=\langle\nabla f(x),\delta y\rangle+o(\delta)\quad\forall x\in\mathcal{D},\,y\in\mathbb{S}^{n-1}.$$

Proof. As $f\in C^\omega(\mathcal{D})$ we can construct, for sufficiently small δ and any $y\in\mathbb{S}^{n-1}$, the convergent Taylor series of $f(x+\delta y)$ around x and as such G must satisfy

$$G\left(f(x)+\langle\nabla f(x),\delta y\rangle+o(\delta)\right)=\langle\nabla f(x),\delta y\rangle+o(\delta)\quad\forall x\in\mathcal{D},\,y\in\mathbb{S}^{n-1}.$$

Evidently, there is no map G that can satisfy the above for all $f\in C^\omega(\mathcal{D})$. □

We remark that there is recent work studying zeroth-order optimization without assuming bounded variance [KSL$^+$23].

Comment on C^∞-smooth imaginary zeroth-order optimization

Consider the smooth function $\psi : \mathbb{R} \to \mathbb{R}$ defined by $\psi(x) = |x|^2$. When evaluating ψ at some complex point $z = x + iy \in \mathbb{C}$ one finds that $\psi(z) = x^2 + y^2$, as such, ψ does not satisfy the Cauchy-Riemann equations and is *nowhere* (complex) analytic. This, however, means that one cannot appeal to the complex-step framework from above. On the other hand, the function $\varphi : \mathbb{R} \to \mathbb{R}$ defined by

$$\varphi(x) = \begin{cases} \exp\left(-x^{-4}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}.$$

is widely known not to be analytic, yet, φ does satisfy the Cauchy-Riemann equations, see [GM78]. Hence, although $\varphi \in C^\infty \setminus C^\omega$, the complex-step framework is not obstructed.

It turns out that from a topological point of view, the function φ is somewhat of a special case. Let X be a topological space. Then the set $M \subset \mathsf{X}$ is of the ***first category***, in the sense of Baire, when M is a countable union of nowhere dense sets in X. A set $A \subseteq \mathsf{X}$ is said to be ***nowhere dense*** when $\mathrm{cl}(A)^c$ is dense in X, or equivalently, when $\mathrm{int}(\mathrm{cl}(A)) = \emptyset$. Now one can show that under a sup-metric, the complement to the space of nowhere differentiable functions in $C^0([0,1])$ is of the first category [Fol99, Ch. 5]. Differently put, almost every continuous function on $[0,1]$ is nowhere differentiable. A similar topological statement can be made about nowhere analytic functions in the space of smooth functions $C^\infty([0,1])$ under a sup-metric, *e.g.*, see[10] [Dar73, Cat84]. Again, bluntly put, almost every smooth function is nowhere analytic. An important question that comes with such an observation is where in the space of smooth functions optimization takes place? What is the right topology?

5.3 Appendix

This appendix contains auxiliary results related to the work above.

The following two results are well-known and follow from a Darboux argument, that is, $\sum_{j=1}^{J}\frac{1}{j} \leq \int_1^J (1/x)\mathrm{d}x + 1$ and $\sum_{j=1}^{J} j^{-1+1/\beta} \leq \int_1^J x^{-1+1/\beta}\mathrm{d}x + 1$.

Lemma 5.3.1 (Logarithm bound)**.** *For any $J \in \mathbb{Z}_{>0}$ one has*

$$\textstyle\sum_{j=1}^{J}\frac{1}{j} \leq \log(J) + 1. \tag{5.3.1}$$

Lemma 5.3.2 (Fractional bound)**.** *For any $\beta \geq 1$ one has*

$$\textstyle\sum_{j=1}^{J} j^{-1+1/\beta} \leq \beta J^{1/\beta}. \tag{5.3.2}$$

The following result allows for showing consistency, *i.e.*, $\lim_{\delta\to 0^+}\nabla f_\delta(x) = \nabla f(x)$.

Lemma 5.3.3 (Integration over the $(n-1)$-sphere)**.** *Given any $x \in \mathbb{R}^n$, then*

$$\frac{n}{\mathrm{vol}(\mathbb{S}^{n-1})} \cdot \int_{\mathbb{S}^{n-1}} \langle x, u\rangle u \mathrm{d}V(u) = x. \tag{5.3.3}$$

Although this result is well-known, for completeness we also provide the proof.

[10] See in particular this post `https://web.archive.org/web/20161009194815/mathforum.org/kb/message.jspa?messageID=387148` by Dave L. Renfro for more context.

Proof. First, rewrite part of (5.3.3) as $n \cdot \int_{\mathbb{S}^{n-1}} uu^{\mathsf{T}} dV(u)\, x$ by recalling that $uu^{\mathsf{T}}x = \langle x, u\rangle u$. Now we would like to show that $n \cdot \int_{\mathbb{S}^{n-1}} uu^{\mathsf{T}} dV(u) = \mathrm{vol}(\mathbb{S}^{n-1}) \cdot I_n$. To that end, use the "*geometric tracing identity*"[11] $n \cdot \int_{\mathbb{S}^{n-1}} \langle Xu, u\rangle dV(u) = \mathrm{Tr}(X) \cdot \mathrm{vol}(\mathbb{S}^{n-1})$ [GHL04, Lem. 3.100], differentiating both sides with respect to $X \in \mathcal{S}^n$ yields $n \cdot \int_{\mathbb{S}^{n-1}} uu^{\mathsf{T}} dV(u) = \mathrm{vol}(\mathbb{S}^{n-1}) \cdot I_n$ indeed. □

Bibliography

[ADX10] A Agarwal, O Dekel, and L Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proc. Conference on Learning Theory*, pages 28–40, 2010.

[AH17] C Audet and W Hare. *Derivative-Free and Blackbox Optimization.* Springer, Cham, 2017.

[AMH10] A Al-Mohy and N Higham. The complex step approximation to the Fréchet derivative of a matrix function. *Numer. Algorithms*, 53:133–148, 2010.

[APT20] A Akhavan, M Pontil, and A Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Proc. Neural Information Processing Systems*, pages 9017–9027, 2020.

[AS21] A Ajalloeian and S U Stich. On the convergence of SGD with biased gradients. *arXiv e-print:2008.00051*, 2021.

[ASKG18] R Abreu, Z Su, J Kamm, and J Gao. On the accuracy of the complex-step-finite-difference method. *J. Comput. Appl. Math.*, 340:390–403, 2018.

[ASM15] R Abreu, D Stich, and J Morales. The complex-step-finite-difference method. *Geophys. J. Int.*, 202(1):72–93, 2015.

[AWBR09] A Agarwal, M J Wainwright, P Bartlett, and P Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Proc. Neural Information Processing Systems*, pages 1–9, 2009.

[BBN19] A S Berahas, R H Byrd, and J Nocedal. Derivative-free optimization of noisy functions via quasi-Newton methods. *SIAM J. Optim.*, 29(2):965–993, 2019.

[BCCS21] A S Berahas, L Cao, K Choromanski, and K Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Found. Comput. Math.*, pages 1–54, 2021.

[BCS21] A S Berahas, L Cao, and K Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM J. Optim.*, 31(2):1489–1518, 2021.

[BG22] K Balasubramanian and S Ghadimi. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. *Found. Comput. Math.*, 22(1):35–76, 2022.

[BP16] F Bach and V Perchet. Highly-smooth zero-th order online optimization. In *Proc. Conference on Learning Theory*, pages 257–283, 2016.

[Cat84] F S Cater. Differentiable, nowhere analytic functions. *Am. Math. Mon.*, 91(10):618–624, 1984.

[CH04] M G Cox and P M Harris. *Numerical Analysis for Algorithm Design in Metrology.* Software Support for Metrology Best Practice Guide No. 11. National Physical Laboratory, Teddington, 2004.

[CMYZ22] H Cai, D Mckenzie, W Yin, and Z Zhang. A one-bit, comparison-based gradient estimator. *Appl. Comput. Harmon. Anal.*, 60:242–266, 2022.

[11] Let ϕ be a symmetric bilinear form on $\mathbb{R}^n$, then $\int_{\mathbb{S}^{n-1}} \phi(u,u) dV(u) = \int_{\mathbb{S}^{n-1}} u^{\mathsf{T}} Q \Lambda Q^{\mathsf{T}} u dV(u)$ for some orthogonal decomposition of ϕ. However, now see that the inner part simplifies to $\langle \Lambda u', u'\rangle$ for $u' = Q^{\mathsf{T}} u$. Also see that we can write $\Lambda = \sum_{i=1}^{n} \lambda_i E_{ii}$ for E_{ii} the usual basis matrix element in $\mathbb{R}^{n\times n}$. Hence, all we need to do is to compute $\int_{\mathbb{S}^{n-1}} u^{\mathsf{T}} E_{ii} u dV(u)$ for any $i \in [n]$ (overloading the meaning of u). By symmetry, we can consider $I_n(1/n)$ instead of E_{ii} and obtain the identity.

[CSV09] A R Conn, K Scheinberg, and L N Vicente. *Introduction to Derivative-Free Optimization.* SIAM, Philadelphia, 2009.

[CTL22] X Chen, Y Tang, and N Li. Improved single-point zeroth-order optimization using high-pass and low-pass filters. In *Proc. International Conference on Machine Learning*, pages 3603–3620, 2022.

[d'A08] A d'Aspremont. Smooth optimization with approximate gradient. *SIAM J. Optim.*, 19(3):1171–1183, 2008.

[Dar73] R B Darst. Most infinitely differentiable functions are nowhere analytic. *Can. Math. Bull.*, 16(4):597–598, 1973.

[DGN14] O Devolder, F Glineur, and Y Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Math. Program.*, 146(1):37–75, 2014.

[DJWW15] J C Duchi, M I Jordan, M J Wainwright, and A Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE T. Inf. Theory*, 61(5):2788–2806, 2015.

[Fab71] V Fabian. Stochastic approximation. In *Symposium on Optimizing Methods in Statistics*, pages 439–470, Cambridge, 1971. Academic Press.

[FKM04] A Flaxman, A T Kalai, and H B McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. *arXiv e-print:0408007*, 2004.

[Fol99] G B Folland. *Real Analysis.* Wiley-Interscience, Hoboken, 1999.

[GHL04] S Gallot, D Hulin, and J Lafontaine. *Riemannian Geometry.* Springer, Berlin, 2004.

[GKK$^+$20] D Golovin, J Karro, G Kochanski, C Lee, X Song, and Q Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In *Proc. International Conference on Learning Representations*, 2020.

[GKL$^+$17] A V Gasnikov, E A Krymova, A A Lagunovskaya, I N Usmanova, and F A Fedorenko. Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. *Autom. Remote. Control*, 78(2):224–234, 2017.

[GL13] S Ghadimi and G Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

[GM78] J D Gray and S A Morris. When is a function that satisfies the Cauchy-Riemann equations analytic? *Am. Math. Mon.*, 85(4):246–256, 1978.

[GWX17] Y Gao, Y Wu, and J Xia. Automatic differentiation based discrete adjoint method for aerodynamic design optimization on unstructured meshes. *Chinese J. Aeronaut.*, 30(2):611–627, 2017.

[Had45] J Hadamard. *The psychology of invention in the mathematical field.* Princeton University Press, Princeton, 1945.

[Hig02] N J Higham. *Accuracy and stability of numerical algorithms.* SIAM, Philadelphia, 2002.

[HL14] E Hazan and K Levy. Bandit convex optimization: Towards tight bounds. In *Proc. Neural Information Processing Systems*, pages 784–792, 2014.

[HPGS16] X Hu, L A Prashanth, A György, and C Szepesvari. (Bandit) convex optimization with biased noisy gradient oracles. In *Proc. Artificial Intelligence and Statistics*, pages 819–828, 2016.

[HRB08] E Hazan, A Rakhlin, and P Bartlett. Adaptive online gradient descent. In *Proc. Neural Information Processing Systems*, pages 65–72, 2008.

[HS23] W Hare and K Srivastava. A numerical study of applying complex-step gradient and Hessian approximations in blackbox optimization. *Pac. J. Optim.*, 19(3):391–410, 2023.

[JNR12] K G Jamieson, R Nowak, and B Recht. Query complexity of derivative-free optimization. In *Proc. Neural Information Processing Systems*, pages 2681–2689, 2012.

[Jon21] W Jongeneel. Imaginary zeroth-order optimization. *arXiv e-print:2112.07488*, 2021.

[JWZL19] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109, 2019.

[JYK21] W Jongeneel, M.-C Yue, and D Kuhn. Small errors in random zeroth-order optimization are imaginary. *arXiv e-print:2103.05478*, 2021.

[KP02] S G Krantz and H R Parks. *A Primer of Real Analytic Functions*. Birkhäuser, Boston, 2002.

[Kra00] S G Krantz. *Function Theory of Several Complex Variables*. AMS Chelsea Publishing, Providence, 2000.

[KSL$^+$23] N Kornilov, O Shamir, A Lobanov, D Dvinskikh, A Gasnikov, I A Shibaev, E Gorbunov, and S Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. In *Proc. Neural Information Processing Systems*, 2023.

[KSST09] S M Kakade, S Shalev-Shwartz, and A Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. 2009. https://home.ttic.edu/~shai/papers/KakadeShalevTewari09.pdf.

[KW52] J Kiefer and J Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.*, 23(3):462–466, 1952.

[KY03] H Kushner and G G Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2003.

[LBM22] J Li, K Balasubramanian, and S Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. *Math. Oper. Res.*, 48(2):1183–1211, 2022.

[LCK$^+$20] S Liu, P.-Y Chen, B Kailkhura, G Zhang, A O Hero III, and P K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Process. Mag.*, 37(5):43–54, 2020.

[Leb20] J Lebl. Tasty Bits of Several Complex Variables. https://www.jirka.org/scv/scv.pdf, 2020.

[Lee13] J M Lee. *Introduction to Smooth Manifolds*. Springer, New York, 2013.

[LLZ21] H Lam, H Li, and X Zhang. Minimax efficient finite-difference stochastic gradient estimators using black-box function evaluations. *Oper. Res. Lett.*, 49(1):40–47, 2021.

[LM67] J N Lyness and C B Moler. Numerical differentiation of analytic functions. *SIAM J. Numer. Anal.*, 4(2):202–210, 1967.

[LMW19] J Larson, M Menickelly, and S M Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[LRD12] G Lantoine, R P Russell, and T Dargent. Using multicomplex variables for automatic computation of high-order derivatives. *ACM Trans. Math. Softw.*, 38(3):1–21, 2012.

[LZH$^+$16] X Lian, H Zhang, C.-J Hsieh, Y Huang, and J Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *Proc. Neural Information Processing Systems*, pages 3062–3070, 2016.

[Mor68] T Morozan. Stability of stochastic discrete systems. *J. Math. Anal. Appl.*, 23(1):1–9, 1968.

[MSA03] J R R A Martins, P Sturdza, and J J Alonso. The complex-step derivative approximation. *ACM Trans. Math. Softw.*, 29(3):245–262, 2003.

[Nes03] Y Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, New York, 2003.

[NG22] V Novitskii and A Gasnikov. Improved exploitation of higher order smoothness in derivative-free optimization. *Optim. Lett.*, 16(7):2059–2071, 2022.

[NM65] J A Nelder and R Mead. A simplex method for function minimization. *Comput. J.*, 7(4):308–313, 1965.

[NS17] Y Nesterov and V Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, 2017.

[NS18] F Nikolovski and I Stojkovska. Complex-step derivative approximation in noisy environment. *J. Comput. Appl. Math.*, 327:64–78, 2018.

[NY83] A S Nemirovsky and D B Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, New York, 1983.

[Oss78] R Osserman. The isoperimetric inequality. *Bull. Am. Math. Soc.*, 84(6):1182–1238, 1978.

[Ove01] M L Overton. *Numerical Computing with IEEE Floating Point Arithmetic.* SIAM, Philadelphia, 2001.

[PT90] B T Polyak and A B Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.

[Riv07] I Rivin. Surface area and other measures of ellipsoids. *Adv. Appl. Math.*, 39(4):409–427, 2007.

[RSS12] A Rakhlin, O Shamir, and K Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proc. International Conference on Machine Learning*, pages 1571–1578, 2012.

[Rud87] W Rudin. *Real and Complex Analysis.* McGraw-Hill Education, New York, 1987.

[Sch22] K Scheinberg. Finite difference gradient approximation: To randomize or not? *INFORMS J. Comput.*, 34(5):2384–2388, 2022.

[Sha13] O Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proc. Conference on Learning Theory*, pages 3–24, 2013.

[Sha17] O Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *J. Mach. Learn. Res.*, 18(1):1703–1713, 2017.

[SMG13] S U Stich, C L Müller, and B Gärtner. Optimization of convex functions with random pursuit. *SIAM J. Optim.*, 23(2):1284–1309, 2013.

[Spa92] J C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE T. Autom. Control*, 37(3):332–341, 1992.

[Spa05] J C Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* John Wiley & Sons, Hoboken, 2005.

[SRB11] M Schmidt, N Roux, and F Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Proc. Neural Information Processing Systems*, pages 1458–1466, 2011.

[ST98] W Squire and G Trapp. Using complex variables to estimate derivatives of real functions. *SIAM Rev.*, 40(1):110–112, 1998.

[SV15] Y Singer and J Vondrák. Information-theoretic lower bounds for convex optimization with erroneous oracles. In *Proc. Neural Information Processing Systems*, pages 3204–3212, 2015.

[SW18] J A Snyman and D N Wilke. *Practical Mathematical Optimization: Basic Optimization Theory and Gradient-Based Algorithms.* Springer, Cham, 2018.

[SXON22] H.-J M Shi, M Q Xuan, F Oztoprak, and J Nocedal. On the numerical performance of derivative-free optimization methods based on finite-difference approximations. *Optim. Methods Softw.*, 38(2):289–311, 2022.

[VVvGV23] C Vuik, F J Vermolen, M B van Gijzen, and M J Vuik. *Numerical Methods for Ordinary Differential Equations.* TU Delft Open, Delft, 2023.

[WS21] L Wang and J C Spall. Improved SPSA using complex variables with applications in optimal control problems. In *Proc. American Control Conference*, pages 3519–3524, 2021.

[WZS21] L Wang, J Zhu, and J C Spall. Model-free optimal control using SPSA with complex variables. In *Proc. Conference on Information Sciences and Systems*, pages 1–5, 2021.

[ZZJZ22] Y Zhang, Y Zhou, K Ji, and M M Zavlanos. A new one-point residual-feedback oracle for black-box learning and control. *Automatica*, 136, 2022.

6

Topological obstructions to stability

"*To appreciate what we are hinting at it might be good to reflect that if electronics was purely linear, ..., we would have no fun with transistors, ..., in fact you would not be reading these lines.*"

— Fliess [FH86, p. xi].

The previous century saw a surge in nonlinear analysis, as illustrated by the edited book containing the quotation from above. In contrast to the linear case, a substantial amount of results are not constructive and instead try to improve our understanding by looking at *necessary* conditions, *e.g.*, for stability. For these conditions to be useful, they cannot rely on very precise knowledge of the problem, hence we put emphasis on non-trivial conditions only relying on *coarse* information, *e.g.*, one needs only knowledge of attractors, state spaces, vector fields and so forth, *up to homotopy equivalences.*

The purpose of one of our works was to show that this point of view leads to a principled, yet indeed coarse, nonlinear analysis [JM23]. In fact, the crux was that most results can either can be shown via Borsuk's *retraction theory* [Bor67] or via a generalization of the *index theoretic* work from Krasnosel'skiĭ and Zabreiko [KZ84].

In what follows we provide new retraction-based results in Section 6.1 and—given the length of [JM23]—we highlight only very briefly the index theoretic way of thinking in Section 6.2, providing a few new examples. The vector field index already appeared in Section 2.2.3 and Section 2.4.1, however, in Section 6.2 we elaborate on further theory and ramifications.

6.1 Topological properties of compact attractors on Hausdorff spaces

In this section we characterize when a compact, invariant, asymptotically stable attractor on a locally compact Hausdorff space is a strong deformation retract of its domain of attraction.

6.1.1 Introduction

The purpose of this section is to improve our understanding of topological properties of compact asymptotically stable attractors and their respective domain of attraction. Here, we will almost exclusively appeal to topological tools pioneered by Borsuk [Dyd12]. In particular, we will elaborate on the *retraction theoretic* work by Moulay and Bhat [MB10], which itself is a generalization of the seminal works [Son98, Thm. 21], [BB00] and [Bha67, Thm. 4.1].

After Poincaré and Lyapunov, the modern qualitative study of attractors was largely propelled through the monographs by Birkhoff [Bir27] and Nemytskii and Stepanov [NS60], with influential follow-up works by Auslander, Bhatia and Siebert [ABS64], Wilson [WJ67], Hahn [Hah67], Bhatia and Szegö [BS70], Conley [Con78], Milnor [Mil85] and many others, *e.g.*, see [JM23, Ch. 1].

Lately, attractors have been extensively studied through the lens of *shape theory*, *e.g.*, see [Gar91, GMDPS01, GS09] and [KK22, Prop. 1], with the seminal work of Günther and Segal showing that a finite-dimensional compact subset of a manifold can be an attractor if and only if it has the shape of a finite polyhedron [GS93].

The interest in understanding topological properties of attractors and their respective domain of attraction stems from the simple observation that if a certain dynamical system does not exist, then certainly there is no feedback law resulting in a closed-loop dynamical system with precisely those dynamics.

Indeed, this type of study often provides *necessary* conditions of the form that an attractor must be *equivalent* in some sense to its domain of attraction. With that in mind, one seeks a notion of equivalence that is weak enough to cover many dynamical systems, yet also strong enough to obtain insights, *e.g.*, obstructions. Hence, although shape equivalence is more widely applicable [KR00] and in that sense more fundamental, we focus on *homotopy equivalence* with the aim of recovering stronger necessary conditions.

In the same spirit, by further restricting the problem class, one could even look for stronger notions of equivalence as recently done in [YLAC23]. There, in the context of a vector-field guided path-following problem, homotopy equivalences have been strengthened to topological equivalences (homeomorphisms).

Although we focus on continuous dynamical systems, one can link work of this form to families of differential inclusions [MT11]. Indeed, further partial generalizations of [MB10] to nonsmooth dynamical system are presented in [LQ12].

Notation and technical preliminaries: The *identity map* on a space X is denoted by id_X, that is, $\mathrm{id}_X : x \mapsto x \ \forall x \in X$. The (embedded) n-sphere is the set $\mathbb{S}^n := \{x \in \mathbb{R}^{n+1} : \langle x, x\rangle = 1\}$, with the (closed) n-disk being $\mathbb{D}^n := \{x \in \mathbb{R}^n : \langle x, x\rangle \leq 1\}$. The topological boundary of a space X is denoted by ∂X, *e.g.*, $\partial\mathbb{D}^{n+1} = \mathbb{S}^n$. We use $\simeq_h$ to denote homotopy (equivalence), see Section. 6.1.2.

A topological space X is said to be a ***locally compact Hausdorff*** space when: (i) for any $x \in X$ there is a compact set $K \subseteq X$ containing an open neighbourhood U of x and; (ii) for any $x_1, x_2 \in X$ there are open neighbourhoods $U_1 \ni x_1$ and $U_2 \ni x_2$ such that $U_1 \cap U_2 = \emptyset$, *e.g.*, see [Lee11, p. 31, 104]. Examples of locally compact Hausdorff spaces are: $\mathbb{R}^n$, topological manifolds, the Hilbert cube, any discrete space and so forth. In particular, any compact Hausdorff space is locally compact. Regarding counterexamples, a space X with the trivial topology $\tau = \{X, \emptyset\}$ is not Hausdorff and any infinite-dimensional Hilbert space is a Hausdorff topological vector space, yet, it fails to be locally compact, see also [Mun14, Thm. 29.1].

6.1.2 Continuous dynamical systems

In this section we study continuous (global) ***semi-dynamical systems*** compromised of the triple $\Sigma := (\mathsf{M}, \varphi, \mathbb{R}_{\geq 0})$. Here, M is a locally compact Hausdorff space and $\varphi : \mathbb{R}_{\geq 0} \times \mathsf{M} \to \mathsf{M}$ is a (global) semi-flow, that is, a continuous map that satisfies for any $x \in \mathsf{M}$:

(i) $\varphi(0, x) = x$ (*identity* axiom); and

(ii) $(\varphi(s, \varphi(t, x)) = \varphi(t + s, x)$ $\forall s, t \in \mathbb{R}_{\geq 0} := \{t \in \mathbb{R} : t \geq 0\}$ (*semi-group* axiom).

We will usually write φ^t instead of $\varphi(t, \cdot)$.

We say that a point $x \in \mathsf{M}$ is a ***start point*** (under Σ) if $\forall (t, y) \in \mathbb{R}_{>0} \times \mathsf{M}$ we have that $\varphi^t(y) \neq x$. Differently put, $x \in \mathsf{M}$ is a starting point when a flow starting from x cannot be extended backwards, see [BH06, Ex. 5.14] for an example. To avoid confusion, the evaluation of an integral curve at 0 is sometimes called a "*starting point*" [Lee12, p. 206], which is not what we are talking about here. Then, to eliminate the existence of start points we appeal to [BH06, Prop. 1.7], for instance, we can consider semi-flows generated by a smooth vector field. Concretely, let $F \in \Gamma^\infty(T\mathsf{M})$ be a smooth vector field on a smooth manifold M. It is well-known that under these conditions, for each $p \in \mathsf{M}$ there is a $\varepsilon > 0$ such that $\gamma : (-\varepsilon, \varepsilon) \to \mathsf{M}$ is an integral curve of F with $\gamma(0) = p$ [Lee12, Prop. 9.2], that is, in terms of the (local) flow $\varphi : (-\varepsilon, \varepsilon) \times \mathsf{M} \to \mathsf{M}$ we have

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\varphi^t(p)\right|_{t=s} = F\left(\varphi^s(p)\right), \quad s \in (-\varepsilon, \varepsilon).$$

Hence, with the previous observation in mind we will assume the following throughout the remainder of this section.

Assumption 6.1.1 (Start points)**.** *The set of start points under the semi-dynamical system Σ is empty.*

Stability

We will exclusively focus on a subclass of semi-dynamical systems with practically relevant stability properties, as introduced in Chapter 2.

Definition 6.1.2 (Attractor)**.** *Given some global semi-dynamical system $\Sigma = (\mathsf{M}, \varphi, \mathbb{R}_{\geq 0})$, then, a compact set $A \subseteq \mathsf{M}$ is said to be an invariant, local asymptotically stable, attractor when*

(i) *$\varphi(\mathbb{R}, A) = A$ (invariance); and*

(ii) *for any open neighbourhood $U_\varepsilon \subseteq \mathsf{M}$ of A there is an open neighbourhood $U_\delta \subseteq U_\varepsilon \subseteq \mathsf{M}$ of A such that $\varphi(\mathbb{R}_{\geq 0}, U_\delta) \subseteq U_\varepsilon$ (Lyapunov stability), plus, there is an open neighbourhood $W \subseteq \mathsf{M}$ of A such that all semi-flows initialized in W converge to A (local attractivity), that is, for any $p \in W$ and any open neighbourhood $V \subseteq \mathsf{M}$ of A there is a $T \geq 0$ such that $\varphi^t(p) \in V$ $\forall t \geq T$.*

The combination of Lyapunov stability and local attractivity is referred to as *local asymptotic stability*. When the neighbourhood W in Item (ii) can be chosen to be all of M we speak of *global* asymptotic stability. Local asymptotic stability is also captured by the existence of an open neighbourhood $U \subseteq \mathsf{M}$ of A such that $\cap_{t\geq 0}\varphi^t(U) = A$ [Hur82, Lem. 1.6]

On can find several definitions of "*attractors*" in the literature, see for instance [BS70, Def. V.1.5], [Mil85] and [KR00, Sec. 2.2].

Definition 6.1.3 (Domain of attraction)**.** *Let the compact set $A \subseteq \mathsf{M}$ be an invariant, local asymptotically stable attractor under the semi-dynamical system $\Sigma = (\mathsf{M}, \varphi, \mathbb{R}_{\geq 0})$, then, its domain of attraction is*

$$\mathcal{D}_\Sigma(A) = \{p \in \mathsf{M} : \textit{for any open neighbourhood } U \subseteq \mathsf{M} \textit{ of } A \\ \textit{there is a } T \geq 0 \textit{ such that } \varphi^t(p) \in U\, \forall t \geq T\}.$$

Definition 6.1.3 can be equivalently written in terms of convergent subsequences. Topological properties of attractors $A \subseteq \mathsf{M}$ and their respective domain of attraction $\mathcal{D}_\Sigma(A) \subseteq \mathsf{M}$ are an active topic of study since the 1960s [BS70, JM23].

To elaborate on the introduction, the interest stems from the observation that the (numerical) analysis or synthesis, *e.g.*, via feedback control, of dynamical systems Σ can be involved, while topological properties of the pair $(\mathcal{D}_\Sigma(A), A)$ might be readily available. Here, topological knowledge of $\mathcal{D}_\Sigma(A)$ is frequently used to study if some "*desirable*" domain of attraction is admissible. For instance, one can show that no point $p \in \mathbb{S}^1$ can be a global asymptotically stable attractor under any $\Sigma = (\mathbb{S}^1, \varphi, \mathbb{R}_{\geq 0})$, *e.g.*, see Theorem 6.1.5 below. The intuition being that for this to be true the circle $\mathbb{S}^1$ needs to be torn apart, which is obstructed by demanding φ to be continuous, see also [JM23, Fig. 1.1]. Again, we emphasize that conclusions of this form emerge without involved analysis of any particular system Σ.

Retraction theory

The previous example can be understood through $\mathbb{S}^1$ not being *contractible*, that is, $\mathbb{S}^1$ is not *homotopy equivalent* to a point p. Formally, two topological spaces X and Y are said to have the same *homotopy type* when they are homotopy equivalent[1], that is, there are

[1] More abstractly, homotopies are isomorphisms in the homotopy category of topological spaces.

continuous maps $f : X \to Y$ and $g : Y \to X$ such that $f \circ g \simeq_h \mathrm{id}_Y$ and $g \circ f \simeq_h \mathrm{id}_X$. In some sense, this notion is a more general version of a deformation retract—which we recall below, and is most naturally understood through invariance in differential- [GP10] and algebraic topology [Hat02]. As alluded to, now, we recall that $A \subseteq X$ is a ***retract*** of X when there is a map $r : X \to A$ such that $r \circ \iota_A = \mathrm{id}_A$, for $\iota_A : A \hookrightarrow X$ the inclusion map. The set A is said be a ***deformation retract*** of X when A is a retract and additionally $\iota_A \circ r \simeq_h \mathrm{id}_X$, implying that X is homotopy equivalent to A. When, additionally, the homotopy is *stationary relative to* A, we speak of a ***strong deformation retract***.

Remark 6.1.1 (Deformation retracts)**.** The literature does not agree on what a "*deformation retract*" is. For instance, Hatcher calls strong deformation retracts simply deformation retracts and speaks of "deformation retracts in the weak sense" where we would speak of simply a deformation retract [Hat02, Ch. 0]. This should be contrasted with for instance the 1965 text of Hu [Hu65, Sec. 1.11].

Next, a set $A \subseteq X$ is said to be a ***weak deformation retract*** of X when every open neighbourhood $U \supseteq A$ contains a strong deformation retract $V \supseteq A$ of X.

In this section we will elaborate on the following result due to Moulay and Bhat. In particular, we aim to understand when $\mathcal{D}_\Sigma(A)$ strongly deformation retracts onto A and not just to a subset of a neighbourhood around A.

Theorem 6.1.4 ([MB10, Thm. 5])**.** *Suppose that the compact set $A \subseteq \mathsf{M}$ is an invariant, local asymptotically stable attractor under the semi-dynamical system $\Sigma = (\mathsf{M}, \varphi, \mathbb{R}_{\geq 0})$, then, A is a weak deformation retract of $\mathcal{D}_\Sigma(A)$.*

It is well-known that when M is a smooth manifold and A is an embedded submanifold of M, then, A is a strong neighbourhood deformation retract of M and thus A is homotopic to $\mathcal{D}_\Sigma(A)$ [MB10, Prop. 10], see also [LYC22]. Our aim is to provide further, especially weaker, conditions for this to be true.

Indeed, Theorem 6.1.4 can be seen as a generalization of the following well-known result due to Sontag.

Theorem 6.1.5 ([Son98, Thm. 21])**.** *Suppose that the point $A := \{p\} \subseteq \mathsf{M}$ is an invariant, global asymptotically stable attractor under the semi-dynamical system $\Sigma = (\mathsf{M}, \varphi, \mathbb{R}_{\geq 0})$, then, $\mathcal{D}_\Sigma(A)$ is contractible.*

We aim to strengthen this generalization. We do this by appealing to neighbourhood retracts. A set $A \subseteq X$ is a ***neighbourhood retract*** of X when there is an open neighbourhood $U \subseteq X$ of A such that A is retract of U. This definition extends naturally to (strong) deformation retracts. Now indeed, if for instance M weakly deformation retracts onto $A \subseteq \mathsf{M}$ while A is a neighbourhood deformation retract of M, then M deformation retracts onto A and thus $A \simeq_h \mathsf{M}$ [MB10, Thm. 4].

In this section we focus on homotopy equivalence, a similar but weaker notion is that of *shape equivalence*, understood as being for Čech (co)homology what homotopy theory is for singular (co)homology. Indeed, only for sufficiently "*nice*" spaces, singular cohomology and Čech cohomology agree. For a concise introduction to shape theory, in the sense of Fox [Fox72], we refer the reader to [KR00, Sec. 3]. The crux is to work with open neighbourhoods of a set and not solely with the set itself. For instance, the Warsaw circle $\mathbb{W}^1$, as studied below in Example 6.1.3 is not homotopy equivalent to the circle $\mathbb{S}^1$ but the two spaces are shape equivalent.

6.1.3 Cofibrations

It follows from Theorem 6.1.4 that for A to be homotopy equivalent to $\mathcal{D}_{\Sigma}(A)$, it suffices for A to be a neighbourhood deformation retract of $\mathcal{D}_{\Sigma}(A)$. We will appeal to *cofibrations* to capture this property.

The theory of cofibrations emerges from the so-called "*extension problem*", that is, understanding when a continuous map $f : A \subseteq X \to Y$ can be extended from A to all of X. A typical counterexample is any map $f : \partial\mathbb{D}^n \to Y$ of nonzero (topological) degree, that is, when $\deg(f) \neq 0$, f cannot be extended from the n-sphere $\mathbb{S}^n \simeq \partial\mathbb{D}^{n+1}$ to the $(n+1)$-disk $\mathbb{D}^{n+1}$ [GP10, Sec. 2.4].

Now, cofibrations tell us, loosely speaking, if maps that can be extended, lead to homotopies that can be extended. To define this, we need the following. Let X be a topological space and $A \subseteq X$, then, a pair (X, A) has the *homotopy extension property* (HEP) when, for any Y, the diagram

$$\begin{tikzcd}
(A \times [0,1]) \cup (X \times \{0\}) \arrow[r] \arrow[d, hook] & Y \\
X \times [0,1] \arrow[ur, dotted]
\end{tikzcd} \tag{6.1.1}$$

can always be completed (the *dotted arrow* $\dashrightarrow$ can be found) to be commutative. Differently put, given a a homotopy $H : A \times [0,1] \to Y$ and some map $g : X \to Y$ such that $H(\cdot, 0) = g|_A$, one needs to be able to extend the homotopy from A to X. Pick $Y = (A \times [0,1]) \cup (X \times \{0\})$, then we see that (X, A) having the HEP implies that $(A \times [0,1]) \cup (X \times \{0\})$ is a retract of $X \times [0,1]$. On the other hand, one can show that the existence of such a retract implies that (X, A) has the HEP, that is, these two notions are equivalent, see Theorem 6.1.7 below. See also [Str66, p. 13] for a stronger result.

Then, a continuous map $i : A \to X$ is said be a *cofibration*[2] if the following commutative diagram

$$\begin{tikzcd}
A \times \{0\} \arrow[r, hook] \arrow[d] & A \times [0,1] \arrow[d] \arrow[ddr, bend left] & \\
X \times \{0\} \arrow[r, hook] \arrow[drr, bend right] & X \times [0,1] \arrow[dr, dotted] & \\
 & & Y
\end{tikzcd} \tag{6.1.2}$$

can be completed for any Y. Loosely speaking, the map i is a cofibration when it has the HEP[3].

[2] Our notion of cofibration is aligned with the so-called *Hurewicz cofibration*. We will not discuss *Serre cofibrations*, which are most easily understood through *fibrations*, a notion dual to that of cofibrations, *e.g.*, see [May99, Ch. 7].

[3] We focus on pairs (X, A) such that $A \subseteq X$, this inclusion is, however, not required for a cofibration to be well-defined. Nevertheless, to make sense of (6.1.1) one should work with $(A \times [0,1]) \cup_i X$ instead of $(A \times [0,1]) \cup (X \times \{0\})$, that is, with the so-called "*mapping cylinder*" as further discussed below.

Next, we need a slight variation of the aforementioned notions of retraction, that of a *neighbourhood deformation retract pair* (NDR pair).

Definition 6.1.6 (NDR pair)**.** *A pair (X, A) is said to be an NDR pair if:*

(i) there is a continuous map $u : X \to [0, 1]$ such that $A = u^{-1}(0)$; and

(ii) there is a homotopy $H : X \times [0, 1] \to X$ such that $H(\cdot, 0) = \mathrm{id}_X$, $H(x, s) = x\ \forall x \in A$, $\forall s \in [0, 1]$ and $H(x, 1) \in A$ if $u(x) < 1$.

See that if $u(x) < 1\ \forall x \in X$, then, A is a strong deformation retract of X. In general, however, we cannot assume u to be of this form. See that for (X, A) to be an NDR pair, A must be closed. Now, a useful result is the following.

Theorem 6.1.7 ([May99, Ch. 6])**.** *Let A be closed in X, then, the following are equivalent:*

(i) the inclusion $\iota_A : A \hookrightarrow X$ is a cofibration;

(ii) $(A \times [0, 1]) \cup (X \times \{0\})$ is a retract of $X \times [0, 1]$;

(iii) (X, A) is an NDR pair.

To be somewhat self-contained, we provide intuition regarding Item (ii) and Item (iii). In particular, we highlight continuity.

Proof (sketch). Suppose we have the retract $r : X \times [0, 1] \to (A \times [0, 1]) \cup (X \times \{0\})$. Define the projections $\pi_1 : X \times [0, 1] \to X$ and $\pi_2 : X \times [0, 1] \to [0, 1]$. Next, define the map $u : X \to [0, 1]$ via

$$u(x) = \sup_{\tau \in [0,1]} \{\tau - \pi_2(r(x, \tau))\}, \tag{6.1.3}$$

where the supremum is attained since $[0, 1]$ is compact and π_2 and r are continuous. Now define the homotopy $H : X \times [0, 1] \to X$ by $H(x, s) = \pi_1(r(x, s))$. Indeed, one can readily check that the pair (u, H) satisfies the properties required for (X, A) to be an NDR pair. Note that since the retract r is continuous, we have that u is not identically 0 when $X \setminus A \neq \emptyset$. The map u constructed through (6.1.3) is in fact continuous since $[0, 1]$ is compact and both π_2 and r are continuous, *e.g.*, one can appeal to the simplest setting of Berge's maximum theorem [Ber63]. □

Note that in general, $(A \times [0, 1]) \cup (X \times \{0\})$ will be equivalent to the *mapping cylinder* under the inclusion map $\iota_A : A \hookrightarrow X$, that is, $M\iota_A = ((A \times [0, 1]) \cup X)/\sim$ with $(a, 0) \sim \iota_A(a)$ for all $a \in A$, also denoted by $(A \times [0, 1]) \cup_{\iota_A} X$. Equivalence can possibly fail when the product and quotient topologies under consideration do not match.

For illustrative purposes, we end this section with the collection of a powerful result. Omitting the details, we recall that X is a *CW complex* when X can be constructed via iteratively "*glueing*" n-cells, being (closed) topological disks $\mathbb{D}^n$, along their boundary to a $(n-1)$-dimensional CW complex, with a 0-dimensional CW complex being simply a set of discrete points. For instance, the circle $\mathbb{S}^1$ can be constructed from a single point and the interval. Then, a set $A \subseteq X$ is a *subcomplex* of the CW complex X when it is closed and a union of (open) cells of X. For more on CW complices, we refer to [Hat02, Ch. 0] and [Lee11, Ch. 5].

Proposition 6.1.8 (CW complices [Hat02, Prop. 0.16])**.** *Let X be a CW complex and $A \subseteq X$ a subcomplex, then, the inclusion map $\iota_A : A \hookrightarrow X$ is a cofibration.*

Proposition 6.1.8 hinges on $\iota_{\mathbb{S}^{n-1}} : \mathbb{S}^{n-1} \hookrightarrow \mathbb{D}^n$ being a cofibration, which can be shown via showing that $(\mathbb{D}^n, \mathbb{S}^{n-1})$ is an NDR pair, however, using a strategy of more general use, one can show there is a (strong deformation) retract from $\mathbb{D}^n \times [0,1]$ onto $(\partial\mathbb{D}^n \times [0,1]) \cup (\mathbb{D}^n \times \{0\})$ [Hat02, p. 15], *e.g.*, consider some point $(0,c) \in \mathbb{D}^n \times \mathbb{R}_{\geq 2}$ and project $(0,c)$ onto $(\partial\mathbb{D}^n \times [0,1])$, then, the line between $(0,c)$ and this projection provides for the homotopy.

CW complices are fairly general, yet, properties that obstruct X admitting a CW decomposition are for instance: (i) X failing to be locally contractible [Hat02, Prop. A4]; and (2) X failing to adhere to Whitehead's theorem *cf.* Example 6.1.3.

Main result

Now, we have collected all ingredients to prove the following.

Lemma 6.1.9 ($\Leftarrow$ Cofibration)**.** *Let $A \subseteq \mathsf{M}$ be a compact, invariant, asymptotically stable, attractor with domain of attraction $\mathcal{D}_\Sigma(A)$. If $\iota_A : A \hookrightarrow \mathcal{D}_\Sigma(A)$ is a cofibration, then, A is a strong deformation retract of $\mathcal{D}_\Sigma(A)$.*

Proof. We know from Theorem 6.1.4 that A is a weak deformation retract of $\mathcal{D}_\Sigma(A)$. Since $\iota_A : A \hookrightarrow \mathcal{D}_\Sigma(A)$ is a cofibration, we also know from Theorem 6.1.7 that $(\mathcal{D}_\Sigma(A), A)$ is an NDR pair. Then, recall Definition 6.1.6 and recall the proof of Theorem 6.1.7. Now, let $W := u^{-1}([0,1)) \supset A$, which is open since $u : \mathcal{D}_\Sigma(A) \to [0,1]$ is continuous, and consider the map $H|_{W\times[0,1]}$. It is imperative to remark that this map does *not* provide a strong deformation retract from W onto A in general. The reason why we cannot conclude on the existence of such a map is that we cannot guarantee that throughout the homotopy we have $H(x,s) \in W$ for any $(x,s) \in W \times [0,1]$. Indeed, we have a map $H|_{W\times[0,1]} : W \times [0,1] \to \mathcal{D}_\Sigma(A) \supseteq W$, the codomain cannot be assumed to be W. Precisely this detail was already known to Strøm *cf.* [Str66, Thm. 2], see also [Bre93, p. 432]. Nevertheless, since A is a weak deformation retract of $\mathcal{D}_\Sigma(A)$ we know that W contains a set $V \supseteq A$ such that $\mathcal{D}_\Sigma(A)$ strongly deformation retracts onto V, that is, there is map $\bar{H} : \mathcal{D}_\Sigma(A) \times [0,1] \to \mathcal{D}_\Sigma(A)$ such that $\bar{H}(x,0) = x \ \forall x \in \mathcal{D}_\Sigma(A)$, $\bar{H}(x,1) \in V \ \forall x \in \mathcal{D}_\Sigma(A)$ and $\bar{H}(x,s) = x \ \forall (x,s) \in V \times [0,1]$. Hence, the continuous map $\widetilde{H} : \mathcal{D}_\Sigma(A) \times [0,1] \to \mathcal{D}_\Sigma(A)$ defined by

$$(x,s) \mapsto \widetilde{H}(x,s) := \begin{cases} \bar{H}(x,2s) & s \in [0,\frac{1}{2}] \\ H\left(\bar{H}(x,1), 2s-1\right) & s \in (\frac{1}{2},1] \end{cases}$$

is a homotopy and provides for the strong deformation retract of $\mathcal{D}_\Sigma(A)$ onto A. □

To continue, we need a converse result.

Lemma 6.1.10 ($\Rightarrow$ Cofibration)**.** *Suppose that M is a locally compact Hausdorff space, that Σ satisfies Assumption 6.1.1 and let $A \subseteq \mathsf{M}$ be a compact, invariant, asymptotically stable, attractor with domain of attraction $\mathcal{D}_\Sigma(A)$. If A is a strong deformation retract of $\mathcal{D}_\Sigma(A)$, then, $\iota_A : A \hookrightarrow \mathcal{D}_\Sigma(A)$ is a cofibration.*

Proof. We will appeal to the characterization of a cofibration as given by Theorem 6.1.7. As A is a strong deformation retract of $\mathcal{D}_\Sigma(A)$ by assumption, then, to conclude on $(\mathcal{D}_\Sigma(A), A)$ being an NDR pair, we need to construct the map $u : \mathcal{D}_\Sigma(A) \to [0,1]$. As A is a compact, invariant, asymptotically stable attractor, M is a locally compact Hausdorff space and Σ satisfies Assumption 6.1.1, there is a Lyapunov function of precisely this form [BH06, Thm. 10.6]. □

Theorem 6.1.11 (Cofibrations)**.** *Suppose that* M *is a locally compact Hausdorff space, that* Σ *satisfies Assumption 6.1.1 and let* $A \subseteq \mathsf{M}$ *be a compact, invariant, asymptotically stable, attractor with domain of attraction* $\mathcal{D}_\Sigma(A)$*. Then,* A *is a strong deformation retract of* $\mathcal{D}_\Sigma(A)$ *if and only if the inclusion* $\iota_A : A \hookrightarrow \mathcal{D}_\Sigma(A)$ *is a cofibration,*

Proof. The results follow directly by combining Lemma 6.1.9 and Lemma 6.1.10. □

Examples

Regarding ramifications of Theorem 6.1.11, we start with a sanity check. We know that for a a linear ODE $\dot{x} = Fx$ with $F \in \mathbb{R}^{n\times n}$ a Hurwitz matrix, $A = \{0\}$ and $\mathcal{D}_\Sigma(A) = \mathbb{R}^n$. Hence, we remark that: (i) $\iota_{\{0\}} : \{0\} \hookrightarrow \mathbb{R}^n$ is a cofibration, *e.g.*, since $(\mathbb{R}^n, \{0\})$ is an NDR pair under the map $x \mapsto u(x) := 1 - e^{-\langle x,x\rangle}$; and (ii) $\mathbb{R}^n$ strongly deformation retracts onto $0 \in \mathbb{R}^n$ via the map $\mathbb{R}^n \times [0,1] \ni (x,s) \mapsto H(x,s) := (1-s)\cdot x$.

Cofibrations that are not strong deformation retracts are abundant. We start with a well-known example.

Example 6.1.2 (Spheres and disks)**.** It can be shown that the inclusion $\iota_{\mathbb{S}^n} : \mathbb{S}^n \hookrightarrow \mathbb{D}^{n+1}$ is a cofibration, *e.g.*, see the remark on CW complices above. However, $\mathbb{D}^{n+1}$ cannot strongly deformation retract onto $\mathbb{S}^n$ since $\mathbb{S}^n$ and $\mathbb{D}^{n+1}$ are not even homotopy equivalent, *e.g.*, $\chi(\mathbb{S}^n) \neq \chi(\mathbb{D}^{n+1})$. Hence, $\mathbb{S}^n$ cannot be a global, asymptotically stable, attractor under any semi-dynamical system $\Sigma = (\mathbb{D}^{n+1}, \varphi, \mathbb{R}_{\geq 0})$.

We proceed with an example where we obtain topological insights through dynamical systems knowledge.

Example 6.1.3 (The Warsaw circle)**.** Let $\mathbb{W}^1 := \{(0,x_2) \in \mathbb{R}^2 : x_2 \in [-1,1]\} \cup \{(x_1, \sin(x_1^{-1})) \in \mathbb{R}^2 : x_1 \in (0,\pi^{-1})\} \cup \{\text{arc from } (0,-1) \text{ to } (\pi^{-1},0)\}$ denote the so-called "*Warsaw circle*" (*e.g.*, see Figure 1.2). The set $\mathbb{W}^1$ is compact, but not a manifold since $\mathbb{W}^1$ is not locally connected. Hastings showed[4] that $\mathbb{W}^1$ can be rendered a compact, invariant, locally asymptotically stable attractor with an annular neighbourhood $A \subset \mathbb{R}^2$ as $\mathcal{D}_\Sigma(\mathbb{W}^1)$ [Has79]. Although the circle $\mathbb{S}^1 \subset \mathbb{R}^2$ and $\mathbb{W}^1 \subset \mathbb{R}^2$ are shape equivalent, they are not homotopy equivalent since $\pi_1(\mathbb{W}^1) \simeq 0$ while $\pi_1(\mathbb{S}^1) \simeq \mathbb{Z}$ and the fundamental group $\pi_1(\cdot)$ is homotopy invariant [Lee11, Thm. 7.40]. As such, $\mathbb{W}^1$ cannot be a strong deformation retract of any annulus $A \subset \mathbb{R}^2$ it embeds in. Then, according to Theorem 6.1.11, $\iota_{\mathbb{W}^1} : \mathbb{W}^1 \hookrightarrow A$ cannot be a cofibration.

We recall that for inclusion maps $\iota_A : A \hookrightarrow X$ to not be a cofibration, the pair (X,A) cannot be too regular, *e.g.*, by Proposition 6.1.8 (X,A) cannot be a CW pair.

[4] Although a substantial part of the proof is left to the reader.

Indeed, by the Whitehead theorem [Hat02, Thm. 4.5], The Warsaw circle $\mathbb{W}^1$ is not homotopy equivalent to a CW complex. This could also be concluded by observing that CW complices must be locally path-connected.

Next, we provide an example inspired by an example from [tDKP70, p. 78–79]. Here we gain dynamical insights via topological knowledge. Before doing so, we recall the difference between the *box* and *product* topology. Let X_α be a topological space indexed by $a \in A$, then, if we endow a topological space of the form $X = \prod_{\alpha \in A} X_\alpha$ with the ***product topology***, open sets are of the form $\prod_{\alpha \in A} U_\alpha$ with U_α open in X_α and all but finitely many $U_\alpha = X_\alpha$. The box topology, on the other hand, does not require the last constraint to hold and open sets are simply of the form $\prod_{\alpha \in A} U_\alpha$ with U_α open in X_α. When A is finite, these topologies are equivalent, however, the box topology is finer than the product topology in general.

Example 6.1.4 (The Tychonoff cube)**.** Let $\Omega > \aleph_0$, then, define the Tychonoff cube as $[0,1]^\Omega$, that is, as a uncountably infinite product of the unit interval. Here we endow $[0,1]$ with the standard topology and $[0,1]^\Omega$ with the product topology. As such, $[0,1]^\Omega$ is a compact Hausdorff space by Tychonoff's theorem [Mun14, Thm. 37.3] and the fact that any product of Hausdorff spaces is Hausdorff [Mun14, Thm. 19.4]. Exploiting compactness, $\{0\}^\Omega \in [0,1]^\Omega$ can be shown to be a strong deformation retract of $[0,1]^\Omega$. Indeed, one can simply use the map $[0,1]^\Omega \times [0,1] \ni (x,s) \mapsto H(x,s) := (1-s)\cdot x$, which would be continuous in the product topology, but not in the box topology. Despite the strong deformation retraction, $\iota_0 : \{0\}^\Omega \hookrightarrow [0,1]^\Omega$ is not a cofibration since otherwise, by Definition 6.1.6 and Theorem 6.1.7, there must be a continuous map $u : [0,1]^\Omega \to [0,1]$ such that $\{0\}^\Omega = u^{-1}(0)$. However, it can be shown that such a map fails to exist due to Ω being uncountable [tDKP70, p. 78–79]. This is precisely where the product topology enters[5]. Hence, Theorem 6.1.11 implies that $\{0\}^\Omega \in [0,1]^\Omega$ cannot be an asymptotically stable attractor, for any continuous—with respect to the product topology on $[0,1]^\Omega$ —semi-dynamical system $\Sigma = ([0,1]^\Omega, \varphi, \mathbb{R}_{\geq 0})$. Note that if Ω would be finite, then, the map u does exist and can be chosen to be $u : (x_1, \dots, x_\Omega) \mapsto \max_{i=1,\dots,\Omega}\{x_i\}$.

Note that Example 6.1.4 is essentially saying that despite seemingly convenient properties of $\Sigma = ([0,1]^\Omega, \varphi, \mathbb{R}_{\geq 0})$, a Lyapunov function fails to exist for $\{0\}^\Omega \in [0,1]^\Omega$. Concurrently, this example shows that even a strong notion of homotopy equivalence can be insufficient to conclude on the existence of an asymptotically stable attractor.

Although, in general, a metrizable space must be merely countably locally finite (σ-locally finite) [Mun14, Thm. 40.3], compact metric spaces must be second countable. Hence, $[0,1]^{\Omega > \aleph_0}$ is not metrizable since $\Omega > \aleph_0$ obstructs second countability. Similarly, one can consider the topology of pointwise convergence. Regardless, Example 6.1.4 illustrates where to look for counterexamples. Indeed, as $[0,1]^{\Omega > \aleph_0}$ is not a normed space and

[5] As the reference is in German, we provide a sketch of the proof. Suppose that u does exist. Let $\{0\} = \cap_{n=1}^{\infty}[0,1/n) \in [0,1]$ and note that all $[0,1/n)$ are open in $[0,1]$. It follows that $u^{-1}(0) = \cap_{n=1}^{\infty} u^{-1}([0,1/n))$, with $u^{-1}([0,1/n))$ an open neighbourhood of $\{0\}^\Omega$. Then, consider the identification $[0,1]^\Omega \simeq [0,1]^{\Omega_n} \times [0,1]^{\Omega - \Omega_n}$. Now we know from the product topology, that there must be a finite Ω_n such that $u^{-1}([0,1/n)) \supseteq \{0\}^{\Omega_n} \times [0,1]^{\Omega-\Omega_n}$. However, if we do this for any n we find that $\cap_{n=1}^{\infty} u^{-1}([0,1/n)) \supseteq \{0\}^{\Omega'} \times [0,1]^{\Omega'-\Omega}$, for $\Omega' = \cup_{n=1}^{\infty}\Omega_n$, which is countable. This leads to a contradiction since for $u^{-1}(0) = \{0\}^\Omega$ we must have $\Omega = \Omega'$, where Ω is uncoutable.

in particular not a Hilbert space, it does not fit into common analysis frameworks, *e.g.*, [CZ12].

It turns out that Theorem 6.1.11 covers known results in case A is an embedded submanifold of M. We will assume all our manifolds under consideration to be C^∞-smooth and second countable. In that case, let $A \subseteq \mathsf{M}$ be a closed embedded submanifold, then, one appeals to the existence of a *tubular neighbourhood* [Lee12, Thm. 6.24] to show that $(\mathcal{D}_\Sigma(A), A)$ comprises an NDR pair. Hence, using the following proposition, [MB10, Prop. 10] follows as a corollary to Theorem 6.1.11, see also [LYC22].

Proposition 6.1.12 (Submanifolds)**.** *Let A be a compact embedded submanifold of M, then, $\iota_A : A \hookrightarrow \mathsf{M}$ is a cofibration.*

Our last example pertains to compositions, indicating that Theorem 6.1.11 can be applied to subsystems.

Example 6.1.5 (Compositions)**.** Cofibrations are closed under composition. Let $i_1 : A \to B$ and $i_2 : B \to C$ be cofibrations, then, $i := i_2 \circ i_1 : A \to C$ is a cofibration. To see this, consider the diagram

$$\begin{tikzcd}
A \times \{0\} \arrow[r, hook] \arrow[d] & A \times [0,1] \arrow[d] \arrow[dddr, bend left, "\alpha_1"] & \\
B \times \{0\} \arrow[r, hook] \arrow[d] \arrow[ddrr, bend right, "\alpha_2"'] & B \times [0,1] \arrow[d] \arrow[ddr, bend left, "\beta_1"] & \\
C \times \{0\} \arrow[r, hook] \arrow[drr, bend right, "\beta_2"'] & C \times [0,1] \arrow[dr, dotted] & \\
 & & Y
\end{tikzcd} . \tag{6.1.4}$$

For any triple (Y, α_1, α_2) there is some appropriate map β_1 completing (6.1.2) for the cofibration $i_1 : A \to B$, but for any triple (Y, β_1, β_2) the diagram (6.1.4) can be completed since $i_2 : B \to C$ is a cofibration.

6.2 Some applications of intersection theory

Now, we focus again on topological obstructions, yet based on results from differential topology and in particular intersection theory [GP10]. As mentioned before, we are decidedly brief.

6.2.1 Multistability

Early work concerned with *global* stabilization focused on topological obstructions with respect to *global*, continuous, asymptotic stabilization of a single point, *e.g.*, see [Son98, BB00], with the typical obstruction being that the state space is not contractible (*cf.* Example 1.1.2). Continuing this line of research, instead of stabilizing a single point, one could consider stabilizing a *connected* submanifold, *e.g.*, see [Man07, Man10] or look at general compact attractors, *e.g.*, see [MB10]. Another question—which is the one we focus on—is when simultaneous stabilization of *multiple* equilibrium points is possible.

Understanding this scenario is closely related to switched systems, do you need several local controllers and switch between them, or is there a single *continuous* controller with the same qualitative behaviour?

Loosely speaking, when dealing with multiple attractors we speak of *multistability*. What is more, having multiple attractors means that disturbances can qualitatively change the nominal behaviour, moving from one attractor to another.

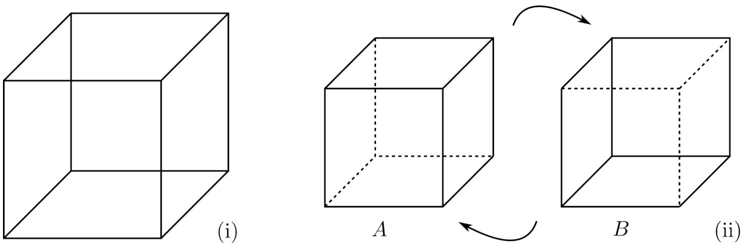


Figure 6.1: Cover (ii) and focus solely on (i) for a few seconds. Most likely, one experiences the uncontrolled switching between A and B.

As the literature does not provide a consistent definition, we follow Angeli, Forni and coworkers to further clarify the concept.

Multistability: "*While most contributions focus on Lyapunov stability of a single connected attractor, e.g. an equilibrium point, several applications in system biology, mechanics, and electronics, have called for a global analysis of the so-called "multistable" systems. The term encompasses a variety of non-trivial dynamical behaviours - almost global stability, multiple equilibria, periodicity, almost periodicity, chaos - and commonly refers to the existence of a compact invariant set which is simultaneously globally attractive and decomposable in a finite number of smaller compact invariant sets.*" [FA17].

Multistability appears for example in the study of laser dynamics [AMPT82] and neural networks [CLS06], with the importance of multistability being especially acknowledged in biology, *e.g.*, see [May77, LK99, AFS04, VNS07, CEA08] and also [PF14, FPS18] for an overview.

Example 6.2.1 (The Necker cube)**.** Multistability appears in a variety of contexts. In this example, originally due to Necker, the reader is invited to experience this phenomenon firsthand. Please find Figure 6.1. While looking at (i) one observes that configuration A and B are attractive, yet unstable, states. This phenomenon is commonly referred to as *bistable perception* and is a form of multistability. The switching dynamics can be modelled as a two-state Markov chain, but in a similar vein one could consider a model on the circle $\mathbb{S}^1$ (say, with some external disturbances), see Figure 6.2. It will be shown below that if one accepts the model on $\mathbb{S}^1$, then, by the topology of $\mathbb{S}^1$, both configurations A and B cannot be simultaneously locally stable indeed.

Akin to Example 6.2.1, a more profound illustration would be Parkinson's disease, were multiple (undesirable) stationary states coexist. In fact, multistability is commonly observed experimentally, *e.g.*, see [PF14] for an overview.

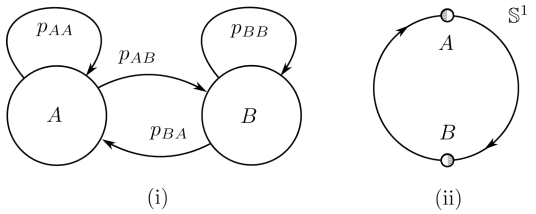


Figure 6.2: Two methods of modelling the behaviour from Figure 6.1.

Theoretically handling multistability is usually done via a variation of classical Lyapunov theory [Ang04, Efi12], *e.g.*, by passing to the dual density formulation as proposed in [Ran01]. There, global requirements are relaxed to *almost* global requirements. By doing so, topological obstructions are surmounted at the cost of introducing singularities.

This section briefly elaborates on the interplay between topology and the manifestation of multistability.

A global obstruction to multistability

Suppose we construct a vector field over some manifold and verify that some desirable equilibrium points are all asymptotically stable. In general, we cannot be certain about the behaviour of the dynamical system outside of the regions of attraction of these points. However, if we suppose that our manifold is *compact*, then, all local behaviour, must add up to certain *global invariants*. Think of squeezing a balloon on one side, pressure builds up elsewhere. On the contrary, throwing a stone into an infinitely large pool causes ripples, yet ripples that disappear into the infinite.

Poincaré was largely responsible for the initiation of this type of work, with important formalizations by the likes of Brouwer and Hopf. In particular, starting from the triangulization of spaces, a deep connection between the *Euler characteristic*, denoted $\chi(\cdot)$, of a compact set (global invariant) and the local behaviour around all equilibrium points of a vector field on this set, was discovered. This is the *Poincaré-Hopf theorem*. The vector field index was introduced before (see Section 2.2.3), but to comment on the Euler characteristic, for polyhedra $\mathsf{P}^2 \subset \mathbb{R}^3$ we have that $\chi(\mathsf{P}^2) = \#(V - E + F)$, that is, the number of vertices, plus the number of faces, minus the number of edges. For convex polyhedra this number is always equal to 2 (like $\mathbb{S}^2$). Now, this alternating formula can be generalized in several ways. For instance, let X^n be a finite *CW complex* (*e.g.*, see [Lee11, Ch. 5]), then, the ***Euler characteristic*** of X^n is defined as

$$\chi(\mathsf{X}^n) = \textstyle\sum_{k=0}^{n}(-1)^k n_k, \tag{6.2.1}$$

for n_k the number of (open) k-cells ($\simeq_t \mathbb{B}^k$) of X^n. Crucial to the wide applicability of this number is its homotopy invariance [Lee11, Thm. 13.36]. Note, for the homotopy invariance it is crucial that X^n is a *finite* CW complex, that is, such that X^n is compact. For otherwise, consider $(0,1)$, which is not a finite CW complex (*e.g.*, construct $[\frac{1}{2},1)$ via intervals $[1-\frac{1}{n-1}, 1-\frac{1}{n}]$ for $n = 3,4,\dots$), and see that $\chi([0,1]) = 1$ while $\chi((0,1)) = -1$, despite $[0,1] \simeq_h (0,1)$. Similar to (6.2.1), one can use the definition from homology [Hat02, Thm.

2.44]. Or, when X^n is an orientable closed manifold (compact and without boundary), the definition form intersection theory [GP10, p. 116].

Now, this global invariant $\chi(\mathsf{M})$ dictates how local behaviour of any vector field on M must be complementary in the following sense.

Theorem 6.2.1 (Poincaré-Hopf theorem [Mil65, p. 35])**.** *Let M be a smooth, compact, oriented, boundaryless manifold. Then, for any smooth vector field $X \in \Gamma^\infty(T\mathsf{M})$ with only isolated equilibrium points $\{p_i^\star\}_{i\in\mathcal{I}} \subset \mathsf{M}$ one has*

$$\chi(\mathsf{M}) = \textstyle\sum_{i\in\mathcal{I}} \mathrm{ind}_{p_i^\star}(X). \tag{6.2.2}$$

For simplicity we focus on the smooth setting, but note that under completeness, most results extend via standard homotopy arguments to the continuous setting. The theorem can also be adapted to work with non-trivial boundaries and to compensate for a lack of orientability (using a covering argument), see the commentary in [JM23].

The relation to stability is provided via the following result: *the Krasnosel'skiĭ-Zabreĭko (Bobylev) theorem.*

Theorem 6.2.2 (Index of local asymptotic stability [Kra68, Ch. II], [KZ84, Thm. 52.1].)**.** *Let $p^\star \in \mathsf{M}^n$ be an isolated asymptotically stable equilibrium point of $X \in \Gamma^\infty(T\mathsf{M}^n)$, then,* $\mathrm{ind}_{p^\star}(X) = (-1)^n$.

It is interesting to note that such a clear topological characterization is not yet known for other notions of stability, *e.g.*, for Lyapunov stability. Moreover, it was shown by Zabczyk [Zab89], that through Theorem 6.2.2 one recovers Brockett's seminal necessary condition for a local, *continuous*, asymptotically stabilizing controller to exist [Bro83].

Now, the following result is somewhat immediate.

Theorem 6.2.3 (A global necessary condition for local stability)**.** *Let M^n be a closed, smooth manifold. Then, M^n admits a smooth vector field X with $z \in \mathbb{N}$ zeroes, all of which are isolated and locally asymptotically stable, if and only if $\chi(\mathsf{M}^n) = z$.*

Proof. The special case of $z = 0$ is well-known, so let $z > 0$. For the "*only if*" direction, as the index of these locally asymptotically stable equilibrium points is $(-1)^n$, $\chi(\mathsf{M}^n)$ must equal $z(-1)^n$. This cannot be true for odd-dimensional manifolds (*i.e.*, $\chi(\mathsf{M}^{2n+1}) = 0$ for all $n \in \mathbb{N}_{>0}$ [Hat02, Cor. 3.37]). Therefore $\chi(\mathsf{M}^n) = z$.

For the "*if*" direction one can follow the same line of arguments as used to show existence of non-vanishing vector fields on M^n with $\chi(\mathsf{M}^n) = 0$, *e.g.*, see [GP10, p. 141–148]. □

An immediate ramification is that since $\chi(\mathbb{S}^1) = 0$, we cannot just have two locally asymptotically stable equilibrium points *cf.* Example 6.2.1. As for manifolds like any odd-dimensional sphere $\mathbb{S}^{2n+1}$, or a product thereof like the n-torus, and any Lie group, $\chi(\mathsf{M}) = 0$. Hence, for many spaces, local asymptotic multistability is simply impossible. This is of importance in many dynamical systems grounded in mechanics as they can be frequently identified with Lie groups [Arn88, Sas99, MLS94, BL04].

Note, results of this form go beyond simply looking at a gradient flow on a compact set and stating that since we have a maximum and minimum, we cannot have a globally asymptotically stable point. For instance, as $\chi(\mathbb{S}^2)=2$, you can have vector fields with just two equilibrium points which are both locally asymptotically stable. This is resolved by having an unstable periodic orbit in between.

Next, we consider a setting that is becoming increasingly relevant due to applications in system identification [UM14] and data-driven control, in particular, when taking the *behavioural* point of view [MD21, PCVW+22].

Let $\mathsf{Gr}(k,n)$ denote the (real) Grassmann manifold, that is, the set of k-dimensional subspaces of $\mathbb{R}^n$. Using *Schubert cells* [MS74, Sec. 5–6], one can show that

$$\chi(\mathsf{Gr}(k,n)) = \begin{cases} 0 & n \text{ even, } k \text{ odd} \\ \begin{pmatrix} \lfloor \frac{n}{2} \rfloor \\ \lfloor \frac{k}{2} \rfloor \end{pmatrix} & \text{otherwise.} \end{cases} \tag{6.2.3}$$

As $\mathsf{Gr}(k,n)$ is compact, global stabilization of any point is obstructed, yet, instead one could consider stabilizing some compact set, *e.g.*, think of optimizing over behaviours. Now, in this context of behavioural systems theory, we do usually have some freedom in selecting the precise Grassmannian. Then equation 6.2.3 shows how these degrees of freedom immediately impact the underlying topology, *e.g.*, when we change n. It is interesting future work to see how to align the precise choice of Grassmannian with system- and control objectives.

We discuss one particular Grassmannian setting in more detail, elaborating on [JM23].

Example 6.2.2 (The Grassmann manifold $\mathsf{Gr}(2,3)$)**.** By identifying points in $\mathsf{Gr}(2,3)$ with their *normals*, we see that $\mathsf{Gr}(2,3)\simeq_t \mathbb{RP}^2 \simeq_t \mathbb{S}^2/\sim$ for $p\sim q$ when p and q are *antipodal* ($p=-q$). This identification leads to a better understanding of admissible behaviour on $\mathsf{Gr}(2,3)$. Either from this identification ($\chi(\mathbb{S}^2)=2$ and $\mathbb{S}^2$ is a double cover of $\mathbb{RP}^2$) or from (6.2.3) we observe that $\chi(\mathsf{Gr}(2,3))=1$. Hence, by Theorem 6.2.3, there must be some vector field on $\mathsf{Gr}(2,3)$ such that we have a single equilibrium point being locally asymptotically stable. What happens elsewhere? To make this precise, we need to discuss limit sets. The ω-limit set of a point p, under a flow φ is defined as

$$\omega(\varphi,p)=\cap_{T\geq 0}\mathrm{cl}\cup_{t\geq T}\{\varphi^t(p)\},$$

by reversing time we get the α-limit set (instead of ω and α one might see ω_+ and ω_- instead). Now, let $p^\star$ be some locally asymptotically stable equilibrium point of $X\in\Gamma^\infty(\mathbb{RP}^2)$, that is, we consider a dynamical system $\Sigma=(\mathbb{RP}^2,\varphi_X,\mathbb{R})$. Although $\mathbb{RP}^2$ cannot be embedded in $\mathbb{R}^3$, we illustrate our envisioned dynamical system in Figure 6.3 via its flat representation. This gives the impression that besides the equilibrium point, we must have an unstable periodic orbit. To show this, recall that the domain of attraction $\mathcal{D}_\Sigma(p^\star)$ is open, so its complement is closed and even compact (since $\mathbb{RP}^2$ is compact). However, then we know that both the ω- and α-limit sets of $\mathcal{D}_\Sigma(p^\star)^c$ are non-empty, compact and connected [Tes12, Ch. 9]. This means that indeed, we will have an unstable limit set as alluded to by Figure 6.3.

An odd-dimensional obstruction

Although topological information is a mild request, it might happen that one has no access to $\chi(\mathsf{M})$. Then, the following condition is useful in situations where $\chi(\mathsf{M})$ is unavailable

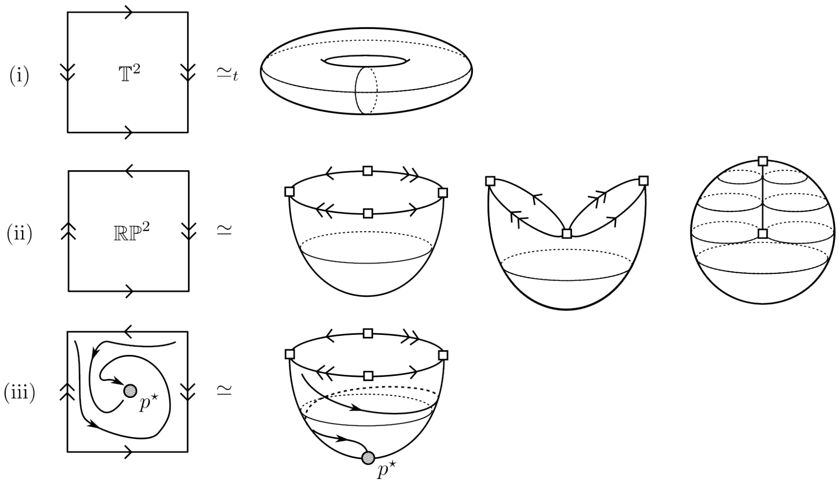


Figure 6.3: Example 6.2.2: (i) flat representation of the torus; (ii) flat representation of $\mathbb{RP}^2$ (see the we cannot simply embed $\mathbb{RP}^2$ into $\mathbb{R}^3$; and (iii) a sketch of $p^\star$ being locally asymptotically stable under some vector field on $\mathbb{RP}^2$, we observe the emergence of some unstable limit set.

and/or one has no knowledge of the number of equilibrium points of $X \in \Gamma^r(T\mathsf{M})$, but one does have *some* local information.

Motivated by the discussion on structural stability in Section 2.3, we solely consider the case of hyperbolic equilibrium points. Here, we use $W^u(\varphi_X, p^\star)$ to denote the unstable "manifold" corresponding to an equilibrium point $p^\star$ of some vector field X (*e.g.*, consider the eigenspaces after linearization).

Theorem 6.2.4 (Odd-number obstruction to local asymptotic multistabilization)**.** *Let* M *be a compact, smooth manifold for which all equilibrium points of some vector field* $X \in \Gamma^r(T\mathsf{M})$ *correspond to the non-empty set* $\{p_i^\star\}_{i\in\mathcal{I}}$ *of isolated hyperbolic equilibria. Given a control system* $\Sigma = (\mathsf{M}, F, \mathcal{U})$ *in the sense of Section 2.2.6, then, the set* $\{p_i^\star\}_{i\in\mathcal{I}}$ *can be locally asymptotically stabilized by continuous feedback* $\mu \in \Gamma^0(\mathcal{U})$*, without introducing new equilibria, only if* $\dim(W^u(\varphi_X, p_i^\star))$ *is even for all* $i \in \mathcal{I}$.

Proof. Without loss of generality, we will consider an even-dimensional manifold as the odd case cannot be handled regardless, *i.e.*, $\chi(\mathsf{M}^{2k+1}) = 0$ for all $k \in \mathbb{N}_{\geq 0}$. If $\dim(W^u(\varphi_X, p_i^\star))$ is odd, then $\mathrm{ind}_{p_i^\star}(X) = -1$ and as such, by the Poincaré-Hopf theorem and the hyperbolic index results from [KZ84, Sec. 6], $|\mathcal{I}| \neq \chi(\mathsf{M})$. Hence, by Theorem 6.2.3, this set cannot be locally asymptotically stabilized by means of continuous state-feedback. □

Remark 6.2.3 (Coordinate invariance)**.** As hinted at, the condition of Theorem 6.2.4 is equivalent to constraining the *orientation* (see Section 2.3.1) of the differential $DX_{p_i^\star} : T_{p_i^\star}\mathsf{M}^n \to T_{p_i^\star}\mathsf{M}^n$ for all $p_i^\star$. That is, instead of demanding that $\dim(W^u(\varphi_X, p_i^\star))$ is not odd, one could equivalently demand that, in coordinates, the differential of the uncontrolled vector field X satisfies $DX_{p_i^\star} \in \mathsf{GL}^+(n, \mathbb{R})$ for all $i \in \mathcal{I}$, that is, $\mathrm{ind}_{p_i^\star}(X) = 1$ for all $i \in \mathcal{I}$. As the orientation of a map is a *topological invariant* [Lee11, Ch. 6], this implies that the statement of Theorem 6.2.4 is topologically invariant, that is, it does not rely on a selection of coordinates for the computation of $DX_{p_i^\star}$.

Theorem 6.2.4 is somewhat counter-intuitive as it implies that there are situations where to be able to continuously render a set of equilibrium points $\{p_i^\star\}_{i\in\mathcal{I}} \subset \mathsf{M}$ locally asymptotically stable, one might need to enlarge the unstable manifold of some $p_j^\star$. As stated before, Theorem 6.2.4 is of use when $\chi(\mathsf{M})$ is unknown or one has partial knowledge about the points that need to be stabilized. For instance, let us be given the task of stabilizing all equilibrium points of a vector field $X \in \Gamma^r(\mathbb{S}^2)$ and suppose we do not know the Euler characteristic of $\mathbb{S}^2$, then, if just one if those points is a saddle, we know we cannot perform the task.

For more on similar phenomenona, see [HA12, AH13, dWS21].

See also [TWLC07, Cor. 5] for an odd-number obstruction in the context of network control and [Chr17, Thm. 1] for odd-number results in the context of optimization.

6.2.2 Submanifold stabilization

In the previous section we were concerned with points, now we provide commentary on the stabilization of manifolds.

The seminal necessary condition by Bhat and Bernstein states that there is no continuous dynamical system on a vector bundle, with a closed (*i.e.*, compact and without boundary) base manifold, such that some point is globally asymptotically stable [BB00]. The intuition being that a vector bundle deformation retracts to its base (formally, a set homeomorphic to the base), which itself is not contractible by assumption, hence, the overall space fails to be contractible, obstructing global asymptotic stability.

One might hope that the situation changes when we aim to stabilize more involved submanifolds, not just points. We will briefly argue why this is not the case.

First, recall the notions of retractions as defined in Section 6.1.2. In particular, recall the zero section $Z(\mathsf{M})$ of the tangent bundle $T\mathsf{M}$ from Section 2.3. One can show that the zero section of a general vector bundle $\pi : \mathsf{E} \to \mathsf{B}$—so, not just $T\mathsf{M}$—is a deformation retract of the total space E, see [JM23, Ex. 3.1]. The intuition being that you retract along the fibers. Next, using (mod 2) degree theory (intersection theory), one can show that a closed manifold M cannot deformatoin retract on any of its proper subsets [JM23, Lem. 3.2]. For instance, $\mathbb{S}^2$ cannot deformation retract onto $\mathbb{S}^1 \hookrightarrow \mathbb{S}^2$.

Now, let $\pi : \mathsf{M} \to \mathsf{B}$ be a vector bundle over a smooth, closed and connected base manifold B. For simplicity, suppose now that $\mathsf{A} \subseteq \mathsf{M}$ is a compact, embedded submanifold. We know from above (*e.g.*, see Proposition 6.1.12) that for A to be a global attractor under some dynamical system Σ, it must be a deformation retract of M. Suppose now that A is a subset of the zero section $Z(\mathsf{B})$. By the discussion above and the transitive properties of homotopies [JM23, Lem 2.1], all of this implies that $Z(\mathsf{B})$ deformation retracts onto A, but due to the properties of B, this can only be true if $\mathsf{A} = Z(\mathsf{B})$.

Summarizing the above, we have the following. We note that results of this form are particularly relevant as mechanical systems can be often identified with dynamical systems on vector bundles over compact spaces.

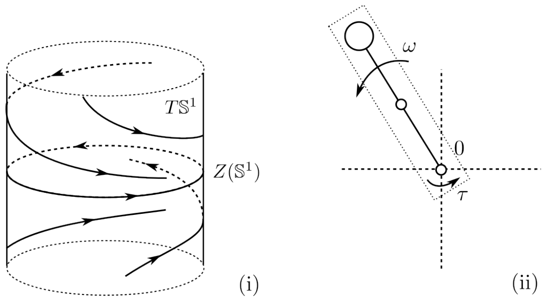


Figure 6.4: Example 6.2.4: (i) global asymptotic stabilization of a periodic orbit for the the single-link pendulum; and (ii) a multi-link pendulum, only actuated at the bottom link, moving with a constant angular velocity when seen as a whole.

Theorem 6.2.5 (Obstruction to submanifold stabilizaiton)**.** *Let $\pi : \mathsf{M} \to \mathsf{B}$ be a vector bundle over a smooth, closed and connected base manifold B. Then, the embedded submanifold $\mathsf{A} \subseteq Z(\mathsf{B})$ is a global attractor of some continuous dynamical system on M only if $\mathsf{A} = Z(\mathsf{B})$.*

Example 6.2.4 (Multi-link pendulum)**.** As an example, consider again the (mathematical) pendulum, that is, Example 2.2.7, yet, suppose that the pendulum is actuated. However, also attach an additional unactuated link on the top of the first link, that is, we create the so-called "*pendubot.* Clearly, for the single-link pendulum we can globally asymptotically stabilize a periodic orbit with constant angular velocity, *e.g.*, see Figure 6.4 (i). This is in line with Theorem 6.2.5 (*i.e.*, such a periodic orbit is to be understood as a zero section of the tangent bundle $T\mathbb{S}^1$, yet after a change of coordinates to accomodate a non-zero angular velocity), since if we consider the controlled pendulum

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ -\frac{1}{2}\sin(x_1(t)) - x_2(t) + \mu(x), \end{pmatrix}$$

for $\mu(x) = \frac{1}{2}\sin(x_1) + \omega$, we stabilize a periodic orbit with angular velocity ω. Going back to our pendubot, if we want to generalize the above, we might want to synthesize a continuous, global and asymptotically stabilizing controller such that the pendubot rotates with a constant angular velocity, while the multi-link system behaves as a solid rod (see Figure 6.4 (ii)). Such an attractor, as understood on $T\mathbb{T}^2$, is of the form $(\mathbb{S}^1 \times \{0\}) \times (\{0\} \times \{0\})$, and is clearly not even homotopy equivalent to the zero section of $T\mathbb{T}^2$. Hence, the stabilization task is obstructed via Theorem 6.2.5.

Bibliography

[ABS64] J Auslander, N Bhatia, and P Siebert. Attractors in dynamical systems. *Bol. Soc. Mat. Mex.*, 9:55–66, 1964.

[AFS04] D Angeli, J E Ferrell, and E D Sontag. Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. *Proc. Natl. Acad. Sci. U.S.A.*, 101(7):1822–1827, 2004.

[AH13] Andreas Amann and Edward W Hooton. An odd-number limitation of extended time-delayed feedback control in autonomous systems. *Philos. T. R. Soc. A*, 371(1999), 2013.

[AMPT82] F T Arecchi, R Meucci, G Puccioni, and J Tredicce. Experimental evidence of subharmonic bifurcations, multistability, and turbulence in a q-switched gas laser. *Phys. Rev. Lett.*, 49(17):1217, 1982.

[Ang04] D Angeli. An almost global notion of input-to-state stability. *IEEE T. Automat. Contr.*, 49(6):866–874, 2004.

[Arn88] V I Arnold. *Mathematical Methods of Classical Mechanics.* Springer, New York, 1988.

[BB00] S P Bhat and D S Bernstein. A topological obstruction to continuous global stabilization of rotational motion and the unwinding phenomenon. *Syst. Control Lett.*, 39:63–70, 2000.

[Ber63] C Berge. *Topological spaces.* Oliver & Boyd, Edinburgh, 1963.

[BH06] N P Bhatia and O Hájek. *Local semi-dynamical systems.* Springer, Berlin, 2006.

[Bha67] N P Bhatia. On asymptotic stability in dynamical systems. *Math. Syst. Theory*, 1(2):113–127, 1967.

[Bir27] George David Birkhoff. *Dynamical systems.* American Mathematical Society, Providence, 1927.

[BL04] F Bullo and A D Lewis. *Geometric Control of Mechanical Systems.* Springer, New York, 2004.

[Bor67] K Borsuk. *Theory of retracts.* Państwowe Wydawn. Naukowe, Warszawa, 1967.

[Bre93] G E Bredon. *Topology and geometry.* Springer, New York, 1993.

[Bro83] R W Brockett. Asymptotic stability and feedback stabilization. In *Differential Geometric Control Theory*, pages 181–191, Boston, 1983. Birkhäuser.

[BS70] N P Bhatia and G P Szegö. *Stability theory of dynamical systems.* Springer, Berlin, 1970.

[CEA08] M Chaves, T Eissing, and F Allgower. Bistable biological systems: A characterization through local compact input-to-state stability. *IEEE T. Automat. Contr.*, 53:87–100, 2008.

[Chr17] F Christensen. A necessary and sufficient condition for a unique maximum with an application to potential games. *Econ. Lett.*, 161:120–123, 2017.

[CLS06] C.-Y Cheng, K.-H Lin, and C.-W Shih. Multistability in recurrent neural networks. *SIAM J. Appl. Math.*, 66(4):1301–1320, 2006.

[Con78] C C Conley. *Isolated invariant sets and the Morse index.* American Mathematical Society, Providence, 1978.

[CZ12] R F Curtain and H Zwart. *An introduction to infinite-dimensional linear systems theory.* Springer, New York, 2012.

[dWS21] B de Wolff and I Schneider. Geometric invariance of determining and resonating centers: Odd-and any-number limitations of Pyragas control. *Chaos*, 31(6):063125, 2021.

[Dyd12] J Dydak. Ideas and influence of Karol Borsuk. *Wiadom. Mat.*, 48(2):81–95, 2012.

[Efi12] D Efimov. Global Lyapunov analysis of multistable nonlinear systems. *SIAM J. Contr. Optim.*, 50(5):3132–3154, 2012.

[FA17] P Forni and D Angeli. Smooth Lyapunov functions for multistable differential inclusions. *Proc. IFAC World Congress*, 50(1):1661–1666, 2017.

[FH86] M Fliess and M Hazewinkel. *Algebraic and geometric methods in nonlinear control theory.* Kluwer, Dordrecht, 1986.

[Fox72] R H Fox. On shape. *Fund. Math*, 74:47–71, 1972.

[FPS18] U Feudel, A N Pisarchik, and K Showalter. Multistability and tipping: From mathematics and physics to climate and brain—minireview and preface to the focus issue. *Chaos*, 28(3):033501, 2018.

[Gar91] B Garay. Strong cellularity and global asymptotic stability. *Fund. Math.*, 2(138):147–154, 1991.

[GMDPS01] A Giraldo, M A Morón, F R R Del Portal, and J M R Sanjurjo. Some duality properties of non-saddle sets. *Topol. Appl.*, 113(1-3):51–59, 2001.

[GP10] V Guillemin and A Pollack. *Differential topology.* American Mathematical Society, Providence, 2010.

[GS93] B Günther and J Segal. Every attractor of a flow on a manifold has the shape of a finite polyhedron. *Proc. Am. Math. Soc.*, 119(1):321–329, 1993.

[GS09] A Giraldo and J M R Sanjurjo. Singular continuations of attractors. *SIAM J. Appl. Dyn.*, 8(2):554–575, 2009.

[HA12] E W Hooton and A Amann. Analytical limitation for time-delayed feedback control in autonomous systems. *Phys. Rev. Lett.*, 109(15):154101, 2012.

[Hah67] W Hahn. *Stability of motion.* Springer, Berlin, 1967.

[Has79] H M Hastings. A higher-dimensional Poincaré-Bendixson theorem. *Glas. Mat.*, 14(34):263–268, 1979.

[Hat02] A Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, 2002.

[Hu65] S T Hu. *Theory of retracts.* Wayne State University Press, Detroit, 1965.

[Hur82] M Hurley. Attractors: persistence, and density of their basins. *T. Am. Math. Soc.*, 269(1):247–271, 1982.

[JM23] W Jongeneel and E Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, Cham, 2023.

[KK22] M D Kvalheim and D E Koditschek. Necessary conditions for feedback stabilization and safety. *J. Geom. Mech.*, 14(4):659–693, 2022.

[KR00] L Kapitanski and I Rodnianski. Shape and Morse theory of attractors. *Commun. Pure Appl. Math.*, 53(2):218–242, 2000.

[Kra68] M A Krasnosel'skiĭ. *The operator of translation along the trajectories of differential equations.* American Mathematical Society, Providence, 1968.

[KZ84] A Krasnosel'skiĭ and P P Zabreiko. *Geometrical methods of nonlinear analysis.* Springer, Berlin, 1984.

[Lee11] J M Lee. *Introduction to Topological Manifolds.* Springer, New York, 2011.

[Lee12] J M Lee. *Introduction to Smooth Manifolds.* Springer, New York, 2012.

[LK99] M Laurent and N Kellershohn. Multistability: a major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.*, 24(11):418–422, 1999.

[LQ12] D Li and A Qi. Morse equation of attractors for nonsmooth dynamical systems. *J. Differ. Equ.*, 253(11):3081–3100, 2012.

[LYC22] B Lin, W Yao, and M Cao. On Wilson's theorem about domains of attraction and tubular neighborhoods. *Syst. Control Lett.*, 167:105322, 2022.

[Man07] A.-R Mansouri. Local asymptotic feedback stabilization to a submanifold: Topological conditions. *Syst. Control Lett.*, 56(7-8):525–528, 2007.

[Man10] A.-R Mansouri. Topological obstructions to submanifold stabilization. *IEEE T. Automat. Contr.*, 55(7):1701–1703, 2010.

[May77] R M May. Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature*, 269(5628):471–477, 1977.

[May99] J P May. *A concise course in algebraic topology.* University of Chicago Press, Chicago, 1999.

[MB10] E Moulay and S P Bhat. Topological properties of asymptotically stable sets. *Nonlinear Anal.-Theor.*, 73(4):1093–1097, 2010.

[MD21] I Markovsky and F Dörfler. Behavioral systems theory in data-driven analysis, signal processing, and control. *Annu. Rev. Control*, 52:42–64, 2021.

[Mil65] J Milnor. *Topology from the differentiable viewpoint.* Princeton University Press, Princeton, 1965.

[Mil85] J Milnor. On the concept of attractor. *Comm. Math. Phys.*, 99:177–195, 1985.

[MLS94] R M Murray, Z Li, and S S Sastry. *A mathematical introduction to robotic manipulation.* CRC Press, Boca Raton, 1994.

[MS74] J W Milnor and J D Stasheff. *Characteristic classes.* Princeton University Press, Princeton, 1974.

[MT11] C G Mayhew and A R Teel. On the topological structure of attraction basins for differential inclusions. *Syst. Control Lett.*, 60(12):1045–1050, 2011.

[Mun14] J Munkres. *Topology.* Pearson Education Limited, Essex, 2014.

[NS60] V V Nemytskii and V V Stepanov. *Qualitative theory of differential equations.* Princeton University Press, Princeton, 1960.

[PCVW$^+$22] A Padoan, J Coulson, H J Van Waarde, J Lygeros, and F Dörfler. Behavioral uncertainty quantification for data-driven control. In *Proc. IEEE Conference on Decision and Control*, pages 4726–4731, 2022.

[PF14] A N Pisarchik and U Feudel. Control of multistability. *Phys. Rep.*, 540(4):167–218, 2014.

[Ran01] A Rantzer. A dual to Lyapunov's stability theorem. *Syst. Control Lett.*, 42(3):161–168, 2001.

[Sas99] S Sastry. *Nonlinear Systems.* Springer, New York, 1999.

[Son98] E D Sontag. *Mathematical control theory: deterministic finite dimensional systems.* Springer, New York, 1998.

[Str66] A Strøm. Note on cofibrations. *Math. Scand.*, 19(1):11–14, 1966.

[tDKP70] T tom Dieck, K H Kamps, and D Puppe. *Homotopietheorie.* Springer-Verlag, Berlin, 1970.

[Tes12] G Teschl. *Ordinary differential equations and dynamical systems.* American Mathematical Society, Providence, 2012.

[TWLC07] A Tang, J Wang, S H Low, and M Chiang. Equilibrium of heterogeneous congestion control: Existence and uniqueness. *IEEE/ACM T. Netw.*, 15(4):824–837, 2007.

[UM14] K Usevich and I Markovsky. Optimization on a grassmann manifold with application to system identification. *Automatica*, 50(6):1656–1662, 2014.

[VNS07] E H Van Nes and M Scheffer. Slow recovery from perturbations as a generic indicator of a nearby catastrophic shift. *Am. Nat.*, 169(6):738–747, 2007.

[WJ67] F W Wilson Jr. The structure of the level surfaces of a Lyapunov function. *J. Differ. Equ.*, 3(3):323–329, 1967.

[YLAC23] W Yao, B Lin, B D O Anderson, and M Cao. The domain of attraction of the desired path in vector-field guided path following. *IEEE Trans. Automat. Contr.*, 68(11):6812–6819, 2023.

[Zab89] J Zabczyk. Some comments on stabilizability. *Appl. Math. Opt.*, 19(1):1–9, 1989.

7
On the future

> “*... if such rough equations are to be of use it is necessary to study them in rough terms ...*”
>
> — Conley [Con78, p. 1].

Most of Kalman his pioneering work was done in the 1960s and around that time computers were not widely available. In fact, the first commercial microprocessor, the 4004 by Intel, only appeared in the early 70s and clocked well below 1MHz. In contrast, this thesis is written on a laptop with a 4 GHz CPU and 16 GB of RAM. As control theory cannot be understood without its applications one must see the popularity of linear state space tools also in the light of the available computational power at the time.

Or as put by Vidyasagar, almost 30 years ago, “*Up to now, the design of control laws has proceeded on the assumption that a linear time-invariant control law is easier to implement than any other. What happens if this fundamental assumption is challenged?*” [BGL95].

Willems’ behavioural systems theory is appealing in that respect. This author is particularly curious about applications to nonlinear problems. The first steps beyond a purely linear behaviour are taken in [PDL23], yet, there are no results beyond conic (contractible) behaviours.

However, this brings us to the main roadblock. Any form of stability analysis beyond physics (mechanics) is largely hampered by a lack of (sufficiently expressive) models and data. For instance, state-of-the-art models in ecology are practically linear, concurrently, measurements are scarce [CWL23]. Then, so-called “model-free” approaches are ought to be promising, yet, they are never model-free; they might not identify a particular model, but they work within a model class, *e.g.*, LTI models. We believe it is crucial to identify more relevant model classes, or better yet, strive for the same understanding of the problem at hand as is common in physics. We believe that a certain level of domain knowledge is crucial.

This also means that our tools must evolve to meet those applications. Several populair assumptions should be challenged. We need to focus more on *transients*, *constraints* (especially inputs), *digital implementations* (controller synthesis that matches), *optimal control objectives* (from performance criteria to objective, not the other way around), *modelling errors* (challenging linearity and time-invariance) and most and for all *model classes* (behaviour).

"*We must place renewed emphasis on stating and teaching the principles of our subject clearly and well. The applications out there are simply too serious for us to hide from responsibility under a cloak of mathematics.*"—Stein [Ste03, p. 25].

7.1 Future work

Now we highlight, for all the chapters, a few key directions of future work. It turns out that for most chapters the central question is what the right topology is.

7.1.1 Chapter 3, large deviations theory

One of the key benefits of the large deviations principle (LDP) as used in Chapter 3 was the exact and flexible characterization of "convergence" of the stochastic process at hand *cf.* Definition 3.1.5. However, large deviations theory also comes somewhat naturally with a mechanism to handle a rather weak notion of "a change of coordinates" (the contraction principle), as we used in the proof of Proposition 3.1.6. In principle, this allows to go beyond Θ, that is, consider further structural knowledge.

Recently, we developed a large deviation (upper bound), for the iterates of a policy gradient algorithm in the context of reinforcement learning (RL) [JKL23]. We recover known rates when applied to tabular, regularized RL with a softmax policy parametrization, however, the benefit of the large deviations viewpoint is that the ambiguity of the policy parametrization can somewhat be overcome through the contraction principle, that is, we recover large deviation upper bounds for all families of policy parametrizations that can be "continuously related" to the softmax parametrization. Of course, there is no free lunch, one needs to check that such a bound is non-trivial and relates to meaningful statistics.

In [JKL23]—by building heavily upon [BJK23]—we took the first steps, effectively providing a large deviations upper bound for stochastic gradient descent under a PL condition. Future work aims at getting a better grip on when a continuous transformation leads to a non-trivial LDP, studying different *classes* of algorithms (momentum, multiphase, adaptive and so forth) and different notions of uncertainty, *e.g.*, bounded noise instead of sub-Gaussian noise.

7.1.2 Chapter 4, spaces of stable systems

Given the results from Section 4.1, we believe we should strive for improving our understanding when it comes to selecting a cost. Mere optimality is not enough, but so is stability without understanding. Design-specifics and other performance metrics should enter the picture again. We write "again" since the early frequency-domain days of control saw a significantly more direct focus on performance (applications). Evidently, MPC is very promising here.

Regarding Section 4.2, we showed that the space of dynamical systems over $\mathbb{R}^n$ with 0 being GAS subject to the existence of a (generalized) convex Lyapunov function is path-connected, see Section 4.2.4. As a byproduct we derived necessary conditions for smooth, convex (control) Lyapunov functions to exist.

There is recent work regarding paths in the space of stable dynamical systems in the context of linear optimal control [BMFM19, JK21], but further extensions are largely lacking. We hope the chapter inspires more work.

We focused on the complete C^0 setting with the emphasis on $\mathbb{R}^n$, future work aims at studying dynamical control systems under weaker regularity assumptions in more general spaces with the focus on more general attractors.

Also, this work focused on the exploitation of a g-convex structure, however, more general structures have been proposed and studied, *e.g.*, a compositional structure [Grü21]. It seems worthwhile to study more structural assumptions along the lines of this article and previous work by Aeyels and Sepulchre [SA96]. For instance, one could consider *weak convexity*, *e.g.*, see [DD19], and similarly, one might consider other stability notions that provide for more structure, like *exponential* stability *cf.* [Vid22].

Another direction of future work is to elaborate on the work by Grüne, Sontag and Wirth [GSW99]. In the remaining subsections we identify more concrete directions of future work.

Invexity

Let $W \subseteq \mathbb{R}^n$ be open. A function $f \in C^1(W;\mathbb{R})$ is said to be *invex* when there is a map $\eta : W \times W \to \mathbb{R}^n$ such that $f(y) \geq f(x) + \langle \nabla f(x), \eta(x,y)\rangle$ for any $x, y \in W$. Differently put, invex functions are such that critical points, *i.e.*, $x^\circ \in W$ such that $\nabla f(x^\circ) = 0$, are *global* minima. The map η is sometimes referred to as the *kernel* and f is said to be invex with respect to this kernel η. Let x° be a critical point of some invex function $f \in C^1(W;\mathbb{R})$, then $f(x) \geq f(x^\circ)$ for all $x \in X$. Conversely, let $f \in C^1(W;\mathbb{R})$ be such that every critical point is a global minimizer, then we can define $\eta : W \times W \to \mathbb{R}^n$ by

$$\eta(x,y) = \begin{cases} 0 & \text{if } \nabla f(x) = 0 \\ \dfrac{(f(y) - f(x))\nabla f(x)}{\langle \nabla f(x), \nabla f(x)\rangle} & \text{otherwise.} \end{cases}$$

This construction shows that indeed $f \in C^1(W;\mathbb{R})$ is invex if and only if every critical point is a global minimizer. So far, we have not said anything about the space of kernels, and in that sense, η is unconstrained and not equipped with any structure. Indeed, invexity has been the subject of controversy [Zăl14, Bor17], mainly due to vacuous generalizations. Nevertheless, continuity of η has been studied [Sma96] and a further study might allow for generalizing several arguments from above. A similar viewpoint can be found in [BSH23], where the authors identify mild assumptions on η such that first-order invex optimization algorithms provably converge.

Convex envelopes

Suppose 0 is GAS under a vector field F on $\mathbb{R}^n$, hence, there is a C^∞ Lyapunov function V and we know that there is a homotopy between F and $-\nabla V(x)$ such that along the homotopy 0 remains GAS. Now, construct the *convex envelope* of V as $\operatorname{conv}(V)(x) := \sup\{g(x) : g \text{ is convex and } g \leq V \text{ on } \mathbb{R}^n\}$. It can be shown that $\operatorname{conv}(V)$ is C^1 by our assumptions on V [KK01]. It readily follows that, with respect to $\dot{z} = -\nabla\operatorname{conv}(V)(z)$, $\operatorname{conv}(V)$ satisfies Properties (V-i)-(V-iii), as such, 0 is GAS under $\dot{z} = -\nabla\operatorname{conv}(V)(z)$. Therefore, a homotopy between V and $\operatorname{conv}(V)$ that preserves regularity and invexity would allow for solving the main research question of Section 4.2 for equilibrium points of vector fields on $\mathbb{R}^n$. A similar question has been studied in the context of Hamilton-Jacobi equations. Omitting details, Vese showed in 1999 that the PDE

$$\frac{\partial u}{\partial t} = \sqrt{1 + \|\nabla_x u\|_2^2}\,\min\{0, \lambda_{\min}(\nabla_x^2 u)\} \tag{7.1.1}$$

converges to precisely the convex envelope of $u(0,x)$ [Ves99]. This observation has been used in the context of homotopy methods for nonconvex optimization [MF15], see also [STP21, HWFO23]. It is, however, not clear if regularity and invexity, perhaps after adapting (7.1.1), are indeed preserved along a solution $u(t,x)$. We believe this is interesting future work.

7.1.3 Chapter 5, zeroth-order optimization

Chapter 5 exploits smoothness to be able to appeal to the Cauchy-Riemann equations. Other work, like [PT90, BP16, APT20, NG22] exploit the knowledge of smoothness and construct kernels to (optimally) filter out (all) low-order errors. For increasing smoothness, however, we observe numerical instability in this approach, that is, the kernels become ill-defined. It would be worthwhile to further study how to exploit smoothness while taking the implementation into consideration.

Also, our work is mostly positioned within the scope of randomized methods via Proposition 5.1.8. Recent work indicated that in fact *non*-randomized methods can outperform their randomized/smoothed counterparts [BCCS21, Sch22]. This provides for interesting future work, especially when smoothness parameters are not (precisely) available and in the presence of noise. Estimating the noise statistics itself also provides for relevant future work as it allows for a more appropriately scaled sequence of smoothing parameters.

Then, as was recently (Nov. 2023) pointed out by Russ Tedrake[1], RL might be able to handle discontinuous objectives since the noise acts like smoothing through randomization. Formalizations are certainly exciting here.

Another exciting area is ODE/PDE-based optimization, as we briefly touched upon in [JYK21], by considering the Lorenz system (attractor).

At last, we point out that the algorithm deserves a *full* numerical study, since *no cancellation* does not imply the full pipeline is numerically stable [Hig02].

[1] See `https://www.youtube.com/watch?v=whpKOHDtOJ0&ab_channel=IntelligentRobotMotionLab`

7.1.4 Chapter 6, topological obstructions

Several directions of future work are: (i) discrete systems of the form $\Sigma = (\mathsf{M}, \varphi, \mathbb{Z}_{\geq 0})$; (ii) exploiting duality (fibrations); (iii) extensions to other notions of stability; (iv) developing computational tools (homology); (v) relaxing the invariance assumption; (vi) addressing stochastic systems in a meaningful way; and most importantly (vii) leveraging the obstructive insights to develop a constructive and practical nonlinear control theory.

We highlight two areas in particular.

Closed attractors

Several results hold when the compactness assumption on $A \subseteq \mathsf{M}$ is relaxed to A being merely *closed*, *e.g.*, see [LSW96, BH06], this generalization is future work. We do emphasize that the generalization is not trivial. Consider for instance [LYC22, Ex. 22] with $\mathsf{M} = \mathbb{R}^2 \setminus \{(1,0)\}$ and $A = \mathbb{S}^1 \setminus \{(1,0)\} \subset \mathsf{M}$. There, the authors construct a vector field on M such that A is an asymptotically stable attractor, with $\mathcal{D}_\Sigma(A) = \mathsf{M} \setminus \{(0,0)\}$. So, although A is an attractor and $\iota_A : A \hookrightarrow \mathcal{D}_\Sigma(A)$ is a cofibration due to Proposition 6.1.12, A cannot be a strong deformation retract of $\mathcal{D}_\Sigma(A)$ since those sets are not homotopy equivalent. Indeed, A is *not* compact and one cannot simply appeal to Theorem 6.1.4. The intuition is that for attractors of this form, limits need not be attained and as such stability does not provide for a homotopy between A and $\mathcal{D}_\Sigma(A)$. Formally speaking, the proof of Theorem 6.1.4 exploits compactness of the sublevel sets of the corresponding Lyapunov function and implicitly *Cantor's intersection theorem*, which fails to be generally true for closed sets. See also [YLAC23, Counterex. 1].

Hybrid systems

As we wrote in the introduction, in times where guanrantees are increasingly important we should develop tools that can provide some. To that end, a global analysis becomes important, or at least, an analysis not concerned with some unknown arbitarily small neighbourhood. We saw that to work within the realm of continuous feedback is restrictive. Solutions can be found by means of time-varying feedback or within the framework of hybrid systems. Typically, hybrid systems are motivated by practical applications that are naturally hybrid (*e.g.* a walking robot), however, topological obstructions to stability can also be naturally linked to a hybrid system; a hybrid system that overcomes these obstructions. With this in mind, to have a solid basis, recent work aims at bringing hybrid systems theory on equal footing with ($C^{r\geq 0}$) dynamical systems theory [KGK21]. Then, towards a principled controller synthesis, we must understand how to bring in these hybrid elements, *e.g.*, "how many switches do we need?" [Bar21] and how to do this robustly [May10]? In general we lack computational tools, a fruiful approach appears to be a combinatorial abstraction of the nonlinear system [VGS$^+$22].

Most and for all, and this is true with respect to all chapters and our field in general, the future is exciting, but we might want to reconsider *the art of the state* [Sep22], pause, and accept some *feedback* [SBB$^+$20].

Bibliography

[APT20] A Akhavan, M Pontil, and A Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Proc. Neural Information Processing Systems*, pages 9017–9027, 2020.

[Bar21] Y Baryshnikov. Topological perplexity in feedback stabilization, 2021. `http://publish.illinois.edu/ymb/files/2021/08/tp.pdf`.

[BCCS21] A S Berahas, L Cao, K Choromanski, and K Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Found. Comput. Math.*, pages 1–54, 2021.

[BGL95] V Blondel, M Gevers, and A Lindquist. Survey on the state of systems and control. *Eur. J. Control*, 1(1):5–23, 1995.

[BH06] N P Bhatia and O Hájek. *Local semi-dynamical systems.* Springer, Berlin, 2006.

[BJK23] D Bajovic, D Jakovetic, and S Kar. Large deviations rates for stochastic gradient descent with strongly convex functions. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 10095–10111, 2023.

[BMFM19] J Bu, A Mesbahi, M Fazel, and M Mesbahi. LQR through the lens of first order methods: Discrete-time case. *arXiv e-print:1907.08921*, 2019.

[Bor17] J M Borwein. Generalisations, examples, and counter-examples in analysis and optimisation: In honour of Michel Théra at 70. *Set-Valued Var. Anal.*, 25(3):467–479, 2017.

[BP16] F Bach and V Perchet. Highly-smooth zero-th order online optimization. In *Proc. Conference on Learning Theory*, pages 1–27, 2016.

[BSH23] A Barik, S Sra, and J Honorio. Invex programs: First order algorithms and their convergence. *arXiv e-print:2307.04456*, 2023.

[Con78] C C Conley. *Isolated invariant sets and the Morse index.* American Mathematical Society, Providence, 1978.

[CWL23] C Chen, X.-W Wang, and Y.-Y Liu. Stability of ecological systems: A theoretical review. *arXiv e-print:2312.07737*, 2023.

[DD19] D Davis and D Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

[Grü21] L Grüne. Computing Lyapunov functions using deep neural networks. *J. Comput. Dyn.*, 8(2):131–152, 2021.

[GSW99] L Grüne, E D Sontag, and F R Wirth. Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes. *Syst. Control Lett.*, 38(2):127–134, 1999.

[Hig02] N J Higham. *Accuracy and stability of numerical algorithms.* SIAM, Philadelphia, 2002.

[HWFO23] H Heaton, S Wu Fung, and S Osher. Global solutions to nonconvex problems by evolution of Hamilton-Jacobi PDEs. *Commun. Appl. Math. Comput.*, pages 1–21, 2023.

[JK21] W Jongeneel and D Kuhn. On topological equivalence in Linear Quadratic optimal control. In *Proc. European Control Conference*, pages 2002–2007, 2021.

[JKL23] W Jongeneel, D Kuhn, and M Li. A large deviations perspective on policy gradient algorithms. *arXiv e-print:2311.07411*, 2023.

[JYK21] W Jongeneel, M.-C Yue, and D Kuhn. Small errors in random zeroth-order optimization are imaginary. *arXiv e-print:2103.05478*, 2021.

[KGK21] M D Kvalheim, P Gustafson, and D E Koditschek. Conley's fundamental theorem for a class of hybrid systems. *SIAM J. Appl. Dyn. Syst.*, 20(2):784–825, 2021.

[KK01] B Kirchheim and J Kristensen. Differentiability of convex envelopes. *C. R. Acad. sci. Ser. I Math.*, 333(8):725–728, 2001.

[LSW96] Y Lin, E D Sontag, and Y Wang. A smooth converse Lyapunov theorem for robust stability. *SIAM J. Control Optim.*, 34(1):124–160, 1996.

[LYC22] B Lin, W Yao, and M Cao. On Wilson's theorem about domains of attraction and tubular neighborhoods. *Syst. Control Lett.*, 167:105322, 2022.

[May10] C G Mayhew. *Hybrid control for topologically constrained systems*. PhD thesis, University of California, Santa Barbara, 2010.

[MF15] H Mobahi and J W Fisher. On the link between Gaussian homotopy continuation and convex envelopes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 43–56. Springer, 2015.

[NG22] V Novitskii and A Gasnikov. Improved exploitation of higher order smoothness in derivative-free optimization. *Optim. Lett.*, 16(7):2059–2071, 2022.

[PDL23] A Padoan, F Dörfler, and J Lygeros. Data-driven representations of conical, convex, and affine behaviors. In *Proc. IEEE Conference on Decision and Control*, pages 596–601, 2023.

[PT90] B T Polyak and A B Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.

[SA96] R Sepulchre and D Aeyels. Homogeneous Lyapunov functions and necessary conditions for stabilization. *Math. Control. Signals, Syst.*, 9(1):34–58, 1996.

[SBB$^+$20] T Samad, M Bauer, S Bortoff, S Di Cairano, L Fagiano, P F Odgaard, R R Rhinehart, R Sánchez-Peña, A Serbezov, F Ankersen, P Goupil, B Grosman, M Heertjes, I Mareels, and R Sosseh. Industry engagement with control research: Perspective and messages. *Annu. Rev. Control*, 49:1–14, 2020.

[Sch22] K Scheinberg. Finite difference gradient approximation: To randomize or not? *INFORMS J. Comput.*, 34(5):2384–2388, 2022.

[Sep22] R Sepulchre. The art of the state [from the editor]. *IEEE Contr. Syst. Mag.*, 42(6):4–4, 2022.

[Sma96] I Smart. On the continuity of the kernel of invex functions. *J. Math. Anal. Appl.*, 197(2):548–557, 1996.

[Ste03] G Stein. Respect the unstable. *IEEE Control Syst. Mag.*, 23(4):12–25, 2003.

[STP21] M Simões, A Themelis, and P Patrinos. Lasry-Lions envelopes and nonconvex optimization: A homotopy approach. In *Proc. European Signal Processing Conference*, pages 2089–2093. IEEE, 2021.

[Ves99] L Vese. A method to convexify functions via curve evolution. *Commun. Partial Differ.*, 24(9-10):1573–1591, 1999.

[VGS$^+$22] E R Vieira, E Granados, A Sivaramakrishnan, M Gameiro, K Mischaikow, and K E Bekris. Morse graphs: topological tools for analyzing the global dynamics of robot controllers. In *Proc. International Workshop on the Algorithmic Foundations of Robotics*, pages 436–453. Springer, 2022.

[Vid22] M Vidyasagar. A new converse Lyapunov theorem for global exponential stability and applications to stochastic approximation. In *Proc. IEEE Conference on Decision and Control*, pages 2319–2321, 2022.

[YLAC23] W Yao, B Lin, B D O Anderson, and M Cao. The domain of attraction of the desired path in vector-field guided path following. *IEEE Trans. Automat. Contr.*, 68(11):6812–6819, 2023.

[Zăl14] C Zălinescu. A critical view on invexity. *J. Optim. Theory Appl.*, 162:695–704, 2014.

Academic CV

Wouter Jongeneel [Aug. 1994, Voorburg (NL) - Today, Lausanne (CH))
Contact: wouter.jongeneel@epfl.ch, http://wjongeneel.nl/

Education
Doctoral program Electrical Engineering—EPFL [Jan. 2020 - Today)
Member of the Risk Analytics & Optimization (RAO) chair, supervised by Prof. D. Kuhn. Member of the NCCR *Automation*.
Doctoral Courses: *Electrical Engineering*: Optimal Control, Learning to Control, Theory and Methods for Reinforcement Learning. *Mathematics*: Some Aspects of Calculus of Variations. *Physics*: Introduction to Topological Phases. *Other*: Convex Optimization, Constrained discrete optimal control on Lie groups (EECI), Optimization on Manifolds.

MSc Systems & Control—TU Delft [Sept. 2017 - Nov. 2019] (*cum laude*)
Master Thesis project: *Controlling the Unknown: A Game Theoretic Perspective*, supervisor: Dr. P. Mohajerin Esfahani, chair: Prof. M. Verhaegen.
Courses: *Applied Sciences*: Geometry of Physics, *Applied Mathematics*: Scientific Computing, Control of Discrete-time Stochastic Systems, *Systems and Control*: Control Theory, Optimization, Filtering and Identification, Nonlinear Systems Theory, Philosophy, Robust and Multivariable Control, Stochastic Control and Dynamic Programming, Nonlinear Control (TU/e), *Practical:* Introduction project, Integration project: *LQR-funnel design.*

Selected publications
Largely interested in dynamical problems, *e.g.*, designing (optimization) algorithms and control systems. In particular, the goal is to understand the interplay between the underlying problem structure and achievable qualitative behaviour.

[P1] Wouter Jongeneel, Man-Chung Yue, and Daniel Kuhn. "*Small errors in random zeroth-order optimization are imaginary*". SIAM Journal on Optimization (2024), pp. 1-32, in press. arXiv: 2103.05478. *Winner of the the INFORMS Optimization Society Student Paper Prize 2023. Presented at SIOPT21.

[P2] Wouter Jongeneel. "*On topological properties of compact attractors on Hausdorff spaces*". European Control Conference (ECC). 2024, pp. 1–6, in press. arXiv: 2301.05932.

[P3] Wouter Jongeneel and Roland Schwan. "*On continuation and convex Lyapunov functions*". IEEE Transactions on Automatic Control (2024), pp. 1–12. doi: 10.1109/TAC.2024.3381913. *Presented at the 17th International Young Researchers Workshop on Geometry, Mechanics, and Control.

[P4] Wouter Jongeneel and Emmanuel Moulay. *Topological Obstructions to Stability and Stabilization: History, Recent Advances and Open Problems.* Springer Nature, 2023. doi: 10.1007/978-3-031-30133-9. *Received the Swiss National Science Foundation open access grant.

[P5] Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. "*Topological Linear System Identification via Moderate Deviations Theory*". IEEE Control Systems Letters 6 (2022), pp. 307–312. doi: 10.1109/LCSYS.2021.3072814. *Presented at the 2021 CDC (invited session).

[P6] Wouter Jongeneel, Tobias Sutter, and Daniel Kuhn. "*Efficient Learning of a Linear Dynamical System With Stability Guarantees*". IEEE Transactions on Automatic Control 68.5 (2023), pp. 2790–2804. doi: 10.1109/TAC.2022.3213770. *Runner up *IEEE CSS Swiss Chapter Young Author Best Journal Paper Award 2023.* Presented at SIOPT23.

[P7] Wouter Jongeneel and Daniel Kuhn. "*On Topological Equivalence in Linear Quadratic Optimal Control*". European Control Conference (ECC). 2021, pp. 2002–2007. doi: 10.23919/ECC54610.2021.9654863.

Selected preprints (under review)

[p1] Wouter Jongeneel. "*Imaginary Zeroth-Order Optimization*". (2021). arXiv: 2112.07488. *Presented at ICCOPT 2022.

Selected talks

[T1] *On continuation and convex Lyapunov functions*, Mar. 2023, KU Leuven, 17th International Young Researchers Workshop on Geometry, Mechanics, and Control.

[T2] *Control theory with Heinz Hopf*, Feb. 2023, ETH Zürich, Institut für Automatik (IfA) Coffee talk.

[T3] *Topological obstructions to stability and stabilization*, Mar. 2022, University of Warsaw, Department of Mathematical Methods in Physics, Theory of Duality seminar.

[T4] *On topological equivalence in Linear Quadratic Optimal Control*, Nov. 2020, Universiteit Utrecht, 15th International Young Researchers Workshop on Geometry, Mechanics, and Control.

[T5] *From correlated data to guarantees in stable identification*, Oct. 2020, ETH Zürich, Institut für Automatik (IfA) Coffee talk.

Reviewer [2019-today) *Journals*: IEEE Transactions on Automatic Control, IEEE Control Systems Letters, Mathematical Programming, Operations Research Letters, Mathematics of Operations Research, Automatica. *Conferences*: ECC, CDC, L4DC, ICML.

Teaching
Master: *Optimal Decision Making* (EPFL, Prof. D. Kuhn) (2019-2021): (graded) assignments, tutorial sessions, proctoring, examination, grading, office hours; *Convex Optimization* (EPFL, Prof. D. Kuhn) (2021-2023): (graded) assignments, forum, tutorial sessions, lecturing, proctoring, examination, grading, office hours; *System Identification* (EPFL, Prof. A. Karimi) (2021-2022): tutorial sessions, project grading.
Bachelor: *Analysis I* (EPFL, Prof. T. Mountford) (2022-2023): tutorial sessions, forum, proctoring, grading; *Analysis II* (EPFL, Dr. X. Fernandez-Real) (2022-2023): tutorial sessions, forum, proctoring, grading; *Linear Algebra* (EPFL, Dr. A. Iseli) (2023-2024): tutorial sessions, forum, proctoring, grading; *Dynamical Systems* (EPFL, Prof. S. Sakar) (2020-2021): forum, proctoring; *Mechatronics* (TU Delft, Dr. Z. Doubrovski) (2016-2017): tutorial sessions; *Statics* (TU Delft, Dr. A. Jansen) (2016-2017): (graded) assignments, tutorial sessions.

Organization
Organizer (with M. Gargiani, T. Sutter, C. Galimberti and B. Guo) of the 2021 EPFL-ETHZ summer school on *the Foundations and Mathematical Guarantees of Data-driven Control*, received funding from the ETH board and sponsorships from the NCCR Automation and the IEEE CSS.