Diss. ETH No. 24789

# ON THE STABILIZATION OF UNSTABLE SYSTEMS - SOME COMMENTS AND APPLICATIONS

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH Zurich

(Dr. sc. ETH Zurich)

presented by

MICHAEL MUEHLEBACH
M.Sc. ETH in Robotics, Systems, and Control

born on 27 March 1989
citizen of Tegerfelden, Aargau

accepted on the recommendation of

Prof. Dr. Raffaello D'Andrea, examiner
Prof. Dr. Melanie Zeilinger, co-examiner

2018

# On the Stabilization of Unstable Systems - Some Comments and Applications

Michael Muehlebach

Institute for Dynamic Systems and Control
ETH Zurich
2018

Institute for Dynamic Systems and Control
ETH Zurich
Switzerland

# Abstract

This thesis deals with the stabilization of open-loop unstable systems. The thesis is divided into four parts. The first two parts discuss the stabilization of two particular unstable systems: a three-dimensional inverted pendulum and a flying vehicle. The third part introduces a general approach for controlling systems that have input or state constraints. The last part presents an approach for the state estimation of distributed networked systems, which is particularly suited for feedback loops that stabilize unstable systems.

The inverted-pendulum system that is discussed in the first part has two main features: the ability to balance on its edge or corner and to jump from lying flat to its corner by suddenly braking its reaction wheels. It is an ideal testbed for nonlinear control algorithms. Applications include space exploration (locomotion in low-gravity environment), self-assembly, balance assistance, and inertially stabilized platforms that are used, for example, for sensor calibration and image stabilization. For realizing the jump-up we relied on a computationally efficient gradient-based learning algorithm that is shown to perform well in practice. Although the approach is discussed and illustrated on the jump-up example, the methods and intuition used generalize and can be translated to other learning problems. The first part concludes with the discussion of a nonlinear algorithm for determining the state (mainly tilt) of the inverted pendulum system based on accelerometer measurements. The approach applies to arbitrary rigid bodies that have a non-accelerated pivot point.

The flying vehicle presented in the second part is actuated by three electric ducted fans. Thrust vectoring is essential for stabilizing the vehicle. Controlling the vehicle is challenging due to the fact that it is open-loop unstable, non-minimum phase, and has limited control authority. The flaps used for thrust vectoring have a limited radius of movement leading to input constraints. The thesis discusses the design and the control of the flying vehicle. The control authority is optimized by a systematic trade-off between the lever arm of the actuation and the total inertia of the system. The low-complexity model that is derived from first principles is refined with a system identification. It is shown that the ducted fan actuation leads to aerodynamic effects that are not captured by the low-complexity model. The flying vehicle is used as a testbed for evaluating control schemes that take input and state constraints into account.

The third part discusses approximations of the constrained linear quadratic regulator problem that are obtained by representing input and state trajectories by a linear combination of basis functions. The constrained linear quadratic regulator problem represents the basis for model predictive control, which is, due to its ability of taking constraints explicitly into account, one of the most successful and widely used control techniques. We

show that with our parametrized approach an infinite prediction horizon can be retained, leading to inherent closed-loop stability and recursive feasibility. Compared to the standard model predictive control approach proposed in the literature, the presented approach provides a different trade-off between approximation quality (performance) and computation, leading to computational advantages in certain applications. We conjecture that the computational benefits are particularly apparent for unstable or marginally stable systems, since these system often require fast sampling and a relatively large prediction horizon when applying the standard approach.

The thesis concludes with presenting a state estimation algorithm tailored to distributed networked systems. Each agent reconstructs the entire state of the system by sporadically exchanging data via a common bus network, and as such, the method is particularly suitable for the stabilization of open-loop unstable systems. Compared to earlier work, we further reduce the network communication load by taking the distributed nature of the system into account. Furthermore, performance and stability guarantees are provided.

# Kurzfassung

Die vorliegende Arbeit befasst sich mit der Stabilisierung von instabilen Systemen. Die Arbeit ist in vier Teile gegliedert. Die ersten zwei befassen sich mit der Stabilisierung eines dreidimensionalen, invertiertem Pendel-Systems und einer Drohne. Der dritte Teil befasst sich mit einer allgemeinen Regelstrategie für Systeme, die Zustands- oder Stellgrössenbeschränkungen aufweisen. Der letzte Teil stellt einen Algorithmus zur effizienten Zustandsschätzung bei verteilten Systemen vor, welcher sich besonders zur Regelung instabiler Systeme eignet.

Das invertierte Pendel-System, welches im ersten Teil betrachtet wird, hat die folgenden zwei Eigenschaften: Es kann auf einer Kante oder Ecke balancieren und von flachliegender Position auf die Ecke aufspringen durch gezieltes und schnelles Bremsen seiner Schwungräder. Das System bildet eine ideale Testumgebung für nichtlineare Regelalgorithmen, die Anwendungen in der Raumfahrtstechnik, in der Medizin, und in der Technik finden könnten. Das Aufspringen wird durch einen recheneffizienten gradientbasierten Lernalgorithmus realisiert, der in der Praxis gut funktioniert. Auch wenn der Lernalgorithmus an einem spezifischen Beispiel vorgestellt wird, generalisiert der Ansatz und die Intuition darüber hinaus. Der erste Teil schliesst mit der Diskussion eines nichtlinearen Zustandsschätzungsalgorithmus ab, der auf Beschleunigungsmessungen basiert. Der Ansatz lässt sich zur Zustandsschätzung beliebiger Starrkörpersysteme die einen unbeschleunigten Fixpunkt aufweisen, erweitern.

Der zweite Teil befasst sich mit einer neu entwickelten Drohne, welche durch drei elektrische Impeller angetrieben wird. Die Drohne wird durch Schubvektorsteuerung stabilisiert. Das System ist instabil, nicht-minimalphasig, und hat eine geringe Steuerbarkeit. Die Stabilisierung eines solchen Systems stellt somit ein anspruchvolles Regelproblem dar. Die Steuerklappen weisen einen relativ geringen Aktuationsradius auf und führen damit zu einem beschränkten Systemeingang. Die Arbeit diskutiert die Konstruktion und die Regelung der Drohne. Es wird aufgezeigt, dass die Steuerbarkeit durch ein geschicktes Abwägen von Hebelarm der Aktuation und Gesamtträgheit optimiert werden kann. Ein physikalisches Modell wird hergeleitet, dessen Parameter durch eine Systemidentifikation bestimmt werden. Die Messergebnisse werden genutzt um das Modell um einen zusätzlichen Dämpfungsanteil zu erweitern, der eine wesentliche Eigenschaft der Impelleraktuation charakterisiert. Die Drohne eignet sich als Teststrecke für Regelalgorithmen, die Eingangs- und Zustandsschranken explizit berücksichtigen.

Im dritten Teil werden Approximationen zum beschränkten linear-quadratischen Regler vorgestellt, welche auf einer Parametrisierung von Eingangs- und Zustandstrajektorien mittels Basisfunktionen basieren. Der beschränkte linear-quadratische Regler wird oft als

Ausgangslage für die modellprädiktive Regelung verwendet und ist somit von zentraler Bedeutung. Aufgrund der expliziten Berücksichtigung der Eingangs- und Zustandsbeschränkungen mittels modellprädiktiver Regelung, geniesst der Ansatz eine grosse Beliebtheit und wurde erfolgreich für diverse Anwendungen eingesetzt. Es wird aufgezeigt, dass man dank der vorgeschlagenen Parametrisierung einen unendlichen Voraussagehorizont beibehalten kann, was zu inherenten Stabilitäts- und Lösbarkeitsgarantien führt. Im Gegensatz zu den geläufigen modellprädiktiven Regelansätzen, führt der vorgeschlagene Algorithmus zu anderen Kompromissen zwischen Approximationsqualität und Rechenaufwand, was zu kürzeren Ausführungszeiten in gewissen Anwendungen führt. Es wird vermutet, dass sich der vorgeschlagene Ansatz besonders zur Regelung instabiler oder grenzstabiler Systeme eignet, da diese im Allgemeinen hohe Abtastraten und einen relativ langen Voraussagehorizon (im standard Ansatz) benötigen.

Der letzte Teil der Arbeit stellt einen Algorithmus zur Zustandsschätzung bei verteilten Systemen vor. Jedes Teilsystem rekonstruiert den Zustand des Gesamtsystems durch einen geschickten Austausch von Daten über ein gemeinsames Bus-Netzwerk. Als solches ist der Algorithmus besonders zur Regelung von instabilen Systemen geeignet. Verglichen mit vorangegangenen Ansätzen wird durch die Berücksichtigung der (verteilten) Struktur des Systems die Kommunikation über das Bus-Netzwerk reduziert. Zusätzlich wird Stabilität garantiert und die Zustandschätzgüte quantifiziert.

[This page is intentionally left blank.]

[Diese Seite wurde absichtlich leer gelassen.]

# Acknowledgement

During the last four years I was supported by many people to which I am very grateful.

I would like to thank Raff for pushing me to work with hardware, which raised my awareness of real-world problems in control, for your critical assessment of new ideas, the seemingly uncountable number of constructive feedback rounds regarding articles and lecture notes, for your support, and above all, for your enthusiasm. Although I studied engineering, I believe that it is only due to your intensive training during my PhD that I truly acquired (some) engineering skills.

I would like to thank Melanie for carefully reading my thesis, providing me with feedback, and asking stimulating and critical questions. I thank you for your willingness to act as a co-examiner.

I would like to thank all the people at the Institute of Dynamic Systems and Control for contributing to a fruitful and productive working atmosphere. I was extremely privileged for being able to rely on good advice from all the intelligent people that surrounded me. In particular, I would like to thank Gajan for supervising my semester project, supporting my PhD application, and letting me write my first conference publication. I still remember the huge effort you put in revising our conference manuscript, culminating in a meeting at unispital hours before AT was born. Being able to attend and present at the International Conference on Decision and Control was a big step for me and exposed me for the first time to state-of-the-art research in control theory. I would like to thank Sebastian for the efficient and fruitful collaboration on the work on distributed estimation. I am grateful for your detailed feedback, for the advice, and the time we spent together during conferences. I would like to thank Philipp for your encouragement, mental support, the nice early morning runs, and contributing to a balanced and positive working atmosphere in the early days of my PhD. I would like to thank Kai and Rajan for being the best office mates that one could imagine. I am grateful to Mike for your grammar advice, to Tony for the nerve-racking chess duels, to Max for your help on embedded systems, to Mark for sharing your original and unique opinions, to Weixuan for diversifying our group, to Dario for introducing me to Superkondi, and to Robin for the fun times in Singapore. I would like to thank Matthias and Carlo for sharing your enthusiasm about the Flying Platform and putting so much effort into the project. I would like to thank Lukas for helping me out with teaching Recursive Estimation and the fun times at the Gotthard bar, Marcel for the nice trip through southern California, Andrea for teaching me the meaning of fishing in Italian, Rie for slowing down the eating pace during lunch, Andreas for the beautiful spring park skiing in Laax, Stijn for introducing me to the Irchel bouldering, and to Michael and Adrian for the numerous interesting discussions

during lunch. I am grateful to Katherina for taking care of the administrative work and for introducing me to the exhausting Tuesday Cardio sessions.

I would like to thank Daniel Vey and Prof. Jan Lunze for inviting me to the University of Bochum, where I had the opportunity to present my research. I am also grateful to Prof. Colin Jones who provided me the opportunity to visit the Automatic Control Laboratory at EPFL.

I would like to thank all the people that contributed to the flying machine arena. Your work saved me an incredible amount of time, and I am thankful that I could use and rely on parts of your software and hardware infrastructure. I would like to thank Michi and Mac for the effort you made in designing the Cubli and the Flying Platform. The same applies to numerous students contributing to hardware and software. In particular, I am grateful to Tobias who wrote a lot of the low-level functionality of the Flying Platform, characterized parts of the actuation, and let me use his results and diagrams for the article on the Flying Platform.

Finally, the encouragement, support, love, and friendship of Barbara, Anja, my family, and my friends were invaluable.

# Contents

# Foreword

This thesis documents the research carried out by the author during his doctoral studies under the supervision of Professor Raffaello D'Andrea at the Institute for Dynamic Systems and Control at ETH Zurich between January 2013 and December 2017.

The work is presented in the form of a cumulative thesis: its main content consists of five self-contained research journal articles that have been published or submitted for publication during the doctoral studies.

The work is divided into four parts: the control of a nonlinear inverted-pendulum system is presented in Part A, followed by the design, implementation, and control of a novel flying vehicle in Part B. Part C deals with approximations of the constrained linear quadratic regulator problem and Part D presents a state estimation algorithm for distributed networked systems.

The articles are put into context by three introductory chapters, which are structured as follows: Chapter 1 introduces and motivates this work, including the problems considered, related work, and the approaches used. Chapter 2 describes the key contributions of the research papers included in this thesis and how the individual papers relate to each other. Chapter 3 then provides a discussion of potential extensions and new directions of this research.

# 1

# Introduction

Feedback mechanisms are omnipresent in nature, technology and our everyday life. A prominent example is the earth's climate, where an increased amount of water vapor in the atmosphere leads to further warming, representing a positive feedback loop. Likewise, the higher radiation heat losses as the earth's temperature increases exemplifies a negative feedback loop. Other well-known examples are the insulin regulation of blood sugar, heat regulation (buildings, warm-blooded animals), voltage and frequency control in power grids, self-driving cars, and autopilots to name a few. Feedback mechanisms are particularly apparent for open-loop unstable systems, which would fail in the absence of regulation. Unstable modes impose constraints on the bandwidth of the feedback loop, which renders the control of these systems particularly challenging. As a result, unstable systems are often used as testbeds for control algorithms, see [1]–[5].

The research in this thesis is divided into four parts and discusses certain aspects related to the stabilization of unstable systems. The first part presents a nonlinear inverted pendulum system. Compared to the classical control benchmarks, [1]–[3], the system presented in the following evolves on the three-dimensional rotation group, which makes the system inherently nonlinear. In addition, the reaction-wheel based actuation leads to the conservation of the total angular momentum in yaw direction. Another unique feature is that by braking the reaction wheels, the inverted pendulum can jump-up to its upright equilibrium. Control algorithms that deal with equilibrium and non-equilibrium motion will be presented. The second part discusses the design, implementation, and control of a novel flying machine. Unlike other flying vehicles that are frequently used as testbeds, it is actuated by three electric ducted fans. Due to the fact that the ducted fans are all rotating in the same direction, thrust vectoring is required to stabilize yaw. The actuation limits resulting from the thrust vectoring render the control of the vehicle particularly challenging. This motivates the third part of the thesis, which deals with approximations of the constrained linear quadratic regular problem. The constrained linear quadratic regulator problem represents the foundation for model predictive control, a control strategy that takes input and state constraints explicitly into account. The last part discusses an estimation algorithm for distributed networked systems, which is particularly suitable for the control of unstable systems. In fact, the method extends previous work that was motivated by the inverted pendulum system presented in [5]. The context for each part is presented below, while the contributions made in the thesis (and specifically the contributions of the papers in this thesis) are discussed in Chapter 2.

## 1.1 A three-dimensional reaction-wheel based inverted pendulum system

Inverted pendulum systems have a long history as control benchmarks. A particularly well-known example is the inverted-pendulum-on-a-cart system, [6]. In addition to its conceptual simplicity, the inverted-pendulum-on-a-cart system often serves as an abstraction of problems encountered in the flight of rockets and missiles at low speeds, [7, p. 100]. Several variants have been proposed, and include multi-link pendulum systems, [8], parallel-type pendulum systems, [9], the Furuta pendulum, [2], the flying inverted pendulum, [10], and the reaction-wheel pendulum, [3].

The inverted pendulum system considered here is unique: It is able to balance in upright position, but can also jump up from lying flat to its corner by braking its reaction wheels with a mechanical braking system. Its configuration includes the three-dimensional rotation group and the total angular momentum about yaw is conserved while balancing. This must be taken into account by the control design, since, depending on the initial condition, it may be impossible to bring the system (reaction wheels and housing) to rest. For example, a yaw motion in the upright position can be slowed down by increasing the velocity of the reaction wheels. However, the yaw motion and the reaction wheel velocity cannot be driven to zero at the same time.

Practical applications that share certain aspects include space exploration, [11], self-assembly, [12], balance assistance, [13], and inertially stabilized platforms, [14].

The work on the inverted pendulum system motivated the design of an attitude estimation algorithm that exploits the fact that the system has a pivot point at rest. The algorithm uses only accelerometer measurements and is able to estimate tilt (pitch and roll), angular velocities, and angular accelerations.

The problem of determining the attitude of a rigid body relative to an inertial frame occurs in many engineering disciplines, and has applications in robotics, aeronautics, and space engineering. Traditional approaches include extended or unscented Kalman filtering and complementary filtering. In complementary filtering, a gyroscope and an accelerometer-based tilt estimate are combined, exploiting the fact that the gyroscope-based estimate is corrupted mainly by low frequency noise (drift of the gyroscope), whereas the accelerometer-based estimate is mainly accurate at low frequencies, see [15, p.290]. However, the accelerometer is typically assumed to be at rest in order to extract the attitude information form the accelerometer measurement, see for example [16]. Kalman filter approaches (as presented in [17]) exploit a dynamic model of the system that captures the temporal correlation of the sensor data. These models include a process noise model, and might require knowledge of physical parameters, such as the inertia, the mass, and the center of mass, which might not be available or only approximately known. Compared to these approaches, the proposed estimation algorithm maximizes the measurement likelihood without taking the temporal correlation of the sensor data into account, thereby not relying on a dynamic model of the system. The underlying

assumption that is exploited is that the rigid body has a non-accelerated pivot point. Compared to complementary filtering approaches, the accelerometers are therefore not assumed to be at rest, and the resulting angular and centripetal acceleration terms are explicitly taken into account.

## 1.2 A flying vehicle actuated by ducted fans

Advances in microelectromechanical systems (MEMS) technology contributed to an increasing interest in autonomous aerial vehicles in the last decades, not least because of the numerous applications ranging from surveillance, data acquisition, aerial photography, construction, and transportation to entertainment. The operation of these systems introduces a large diversity of engineering problems, for example related to the state estimation, the control of equilibrium and non-equilibrium motion, failsafe mechanisms, and the software architecture. This renders autonomous flying vehicles attractive research platforms. The flying vehicle introduced in this thesis, was built for two reasons: 1) as a testbed for control algorithms that take input and state constraints into account; 2) for investigating ducted fan actuation and thrust vectoring.

Ducted fans are an appealing propulsion system for flying machines, where size is limited, but high static thrusts are required. This includes flying vehicles combining efficient forward flight, high maneuverability with vertical take-off and landing capabilities, such as tailsitters and hovercrafts. The research findings might be also useful for actuated wingsuit flight, [18]. The high exit velocities can be exploited for thrust vectoring.

Previous work mainly focused on the design and control of a flying vehicle with a single duct, [19], [20]. The vehicle presented here comprises three ducted fans with relatively small diameters, each of which can vector the thrust. We also investigate the aerodynamic effects resulting from the ducted fan actuation. In particular, a system identification about hover reveals the presence of so-called momentum drag: During horizontal movements, the incoming air is redirected downwards and leaves the exit nozzle with a translational velocity component, explaining the presence of a drag force that is roughly linear in the translational velocity of the vehicle. As highlighted in [21], this force introduces a pitching moment on the center of gravity rendering the system open-loop unstable. The thrust vectoring mechanism is implemented with two orthogonally mounted flaps mounted at the exit of the duct. The fact that the flaps are constrained to an actuation radius of $\pm 18°$ limits the thrust vectoring capabilities. The resulting input constraints combined with the open-loop unstable dynamics render the control of the vehicle particularly challenging. In addition, the sheer size, weight, and power of the vehicle (weight 8kg, requires 6.6kW during hover) require a careful implementation and modular testing of the mechanical and electrical components, as well as the software running the control algorithms.

## 1.3 Approximations of the constrained linear quadratic regulator problem

The constrained linear quadratic regulator problem is central to control theory, as it represents the basis for model predictive control. Model predictive control is one of the most successful control strategies due to its ability of treating input and state constraints in a systematic way. It is widely used in various applications ranging from the process industry to the control of autonomous vehicles, see for example [22] and [23]. The underlying constrained linear quadratic regulator problem is often approximated by discretizing the dynamics and truncating the prediction horizon. The resulting quadratic program is then either solved online, [24] or offline using parametric programming techniques, [25]. Both, online and offline (parametric) solutions are challenging for high-dimensional systems involving long prediction horizons. In case of the parametric programming approach the difficulty stems from an exponential growth in complexity of the resulting feedback policy with the prediction horizon, [26]. In case of an online solution the challenge results from the limited execution time available for the numerical solution, which is often on the order of milliseconds. In addition, the truncation of the prediction horizon leads to issues regarding recursive feasibility and closed-loop stability. Several approaches addressing these issues are proposed in the literature. These include terminal equality constraints, a combination of terminal cost and terminal state constraints, and establishing contraction properties of the running cost, see for example [27] and [28].

In contrast, we propose to represent input and state trajectories using a linear combination of basis functions as an alternative to the standard approximation. By choosing exponentially decaying basis functions, an infinite prediction horizon can be retained, leading to inherent recursive feasibility and closed-loop stability. The basis functions can be used to encode a priori knowledge of the system's dynamics and therefore the approach typically results in a relatively small finite-dimensional optimization problem that is solved at every time step. Changing the number of basis functions leads to a trade-off between approximation quality and computational effort, and duality is exploited for quantifying the approximation quality. We further show that for well-chosen basis functions the approximate solutions converge to the solutions of the underlying constrained linear quadratic regulator problem.

We believe that the proposed approach is particularly suitable for the control of unstable systems, since these systems often require a fast sampling time combined with a relatively large prediction horizon in the classical model predictive control setting. Thus, the computational advantages obtained with a parametrization of input and state trajectories might be particularly apparent in these cases.

## 1.4 A state estimation algorithm for distributed networked systems

The advances in computation and the emerge of low-cost sensing capabilities has greatly improved the capabilities of embedded systems in the last decades. More and more, these system are connected with each other, leading, for example, to large sensor networks, whose sensing and monitoring capabilities exceed those of a single sensing device, [29]–[31]. However, the analysis of such systems is very challenging, due to the fact that the classical control and analysis tools mainly apply to centralized or hierarchical control architectures and do not scale well in the system's dimension.

Here, a distributed control system is considered, where multiple sensor and actuator agents observe and control a dynamic system. The agents are linked by a common bus network, over which they can communicate and exchange data. An event-based protocol is proposed for reducing, respectively averaging communication. Thereby, the agents transmit information only when necessary, instead of communicating periodically at fixed rates. We focused on the state estimation problem and applied Lyapunov-based techniques to provide stability and performance guarantees. Compared to previous work, [32] we came up with performance and stability guarantees that explicitly take the distributed structure into account. These guarantees are formulated in terms of linear matrix inequalities and can be used for the synthesis of stabilizing estimators.

# 2

# Contributions

This chapter describes the scientific contributions for each of the papers that are contained in this thesis. In total, five journal publications are discussed. A list of other contributions such as conference publications, results from unpublished student projects, and outreach activities are provided in this chapter.

## 2.1 A three-dimensional reaction-wheel based inverted pendulum system

**[1] M. Muehlebach and R. D'Andrea, "Nonlinear analysis and control of a recation-wheel-based 3-D inverted pendulum", *IEEE Transactions on Control Systems Technology*, vol. 25, no. 1, pp. 235–246, 2017**

The article presents a nonlinear analysis and several control strategies for a three-dimensional reaction-wheel based inverted pendulum. The equations of motion are conveniently expressed using generalized momenta and it is shown that the total angular momentum about yaw is conserved. In order to deal with the fundamental limits imposed by this conservation law, a reduced attitude description based on the gravity vector (expressed in the body-fixed frame) is introduced and the angular momentum is divided into a controllable part (orthogonal to gravity) and an uncontrollable part (in direction of gravity). The fact that the dynamics have strict-feedback form is exploited for designing a nonlinear controller based on backstepping. This leads to a smooth control law that stabilizes the upright equilibrium (in the almost-everywhere sense and in the absence of input constraints). The control law is parametrized by four tuning parameters, which are related to the closed-loop behavior. In addition, the system is shown to be feedback linearizable, which is exploited for tracking non-equilibrium motions. A low-complexity model is used to derive a gradient-based learning strategy for determining the reaction wheel velocities (before braking) that enable a successful jump up. To enhance robustness, a predefined jump-up trajectory is tracked.

**[2] M. Muehlebach and R. D'Andrea, "Accelerometer-based tilt determination for rigid bodies with a nonaccelerated pivot point", *IEEE Transactions on***

***Control Systems Technology*, 2017, accepted, to appear**

The article describes a tilt estimation algorithm that is based on accelerometer measurements. The estimate is obtained by maximizing the likelihood of the sensor measurements, taking the structure of the angular and centripetal acceleration terms into account. As a byproduct, an angular velocity estimate and an estimate of the rate of change of the angular velocities are obtained. The resulting constrained least-squares problem is solved with a dedicated optimization algorithm that takes advantage of the fact that projections on the feasible sets can be evaluated in closed-form. Moreover, the Fisher information matrix is derived and is used to characterize the information content in the accelerometer measurements and to deduce optimal sensor placements.

## 2.2 A flying vehicle actuated by ducted fans

**[3] M. Muehlebach and R. D'Andrea, "The Flying Platform - a testbed for ducted fan actuation and control design", *Mechatronics*, vol. 42, no. 1, pp. 52–68, 2017**

The article presents the design of a flying vehicle that is actuated by three electric ducted fans. Experimental results are presented, characterizing a single fan unit, comprising of an electric ducted fan, an exit nozzle, and two control flaps. Both static and dynamic measurement results are provided. A low-complexity model is used to investigate the controllability of the vehicle. The mechnical design is shown to maximize the determinant of the controllability Gramian that results from a systematic trade-off between the lever arm of the actuation and the total inertia. Moreover, the low-complexity model motivates a cascaded control structure that is shown to work reliably in flight experiments. A non-parametric system identification about hover reveals the limitations of the low-complexity model. The gyroscopic effects of the fans, as well as the so-called momentum drag are found to be two dominant unmodeled effects and are included in an augmented model. The augmented model is shown to roughly match the measured frequency response function of the system.

## 2.3 Approximations of the constrained linear quadratic regulator problem

**[4] M. Muehlebach and R. D'Andrea, "On the approximation of constrained linear quadratic regulator problems and their application to model predictive**

**control", *Automatica*, 2017, submitted, in review**

We discuss the approximation of the constrained linear quadratic regulator problem based on a parametrization of input and state trajectories with basis functions. A sequence of upper and lower bounds on the cost of the underlying problem is derived, providing a means to quantify the suboptimality. We provide conditions guaranteeing the convergence of these upper and lower bounds to the cost of the underlying problem. The approximations are applied in the context of model predictive control, where it is shown that an infinite prediction horizon can be retained if the basis functions are chosen to be decaying. As a result, closed-loop stability and recursive feasibility are shown to be inherent to the resulting model predictive control algorithm. Although, the optimization problem that is solved at every time step is finite dimensional, has a quadratic cost, and linear equality constraints, it includes linear semi-infinite inequality constraints. These originate from the continuous-time formulation, where it is necessary to impose input and state constraints over a compact time interval, rather than at a finite number of sampling instances. We propose a dedicated active-set-based optimization algorithm for dealing with these semi-infinite constraints and highlight its numerical effectiveness on the example of a quadruple integrator system. The approach is compared to the standard model predictive control solvers FORCES, [33] and qpOASES, [34].

## 2.4 A state estimation algorithm for distributed networked systems

**[5] M. Muehlebach and R. D'Andrea, "Distributed event-based state estimation for networked systems: An LMI-approach", *IEEE Transactions on Automatic Control*, 2017, accepted, to appear**

This article is concerned with the state estimation of a dynamic system that is controlled by multiple sensor-actuator agents. The agents exchange sporadically measurements over a common bus network. Each agent triggers a communication whenever the local measurement's prediction deviates too much from the actual local measurement. The closed-loop dynamics are brought in strict feedforward form by expressing them in terms of the agent errors (deviation of the agents' estimates from the real state), the inter-agent errors (the difference in the agents' state estimates), and the system's state. This enables a Lyapunov-based stability analysis that can also be used for the synthesis of stabilizing observer gains. A flexible performance objective is derived, such that the estimator design is formulated as an optimization problem. Compared to earlier work, [32], both the triggering threshold and the observer gains are obtained by solving convex optimization problems, whereby the distributed nature of the system is taken into account. A numerical example based on a vehicle platoon demonstrates the scalability of the proposed approach.

## 2.5 List of publications

**Publications in this Thesis**

[1] M. Muehlebach and R. D'Andrea, "Nonlinear analysis and control of a recation-wheel-based 3-D inverted pendulum", *IEEE Transactions on Control Systems Technology*, vol. 25, no. 1, pp. 235–246, 2017.

[2] M. Muehlebach and R. D'Andrea, "Accelerometer-based tilt determination for rigid bodies with a nonaccelerated pivot point", *IEEE Transactions on Control Systems Technology*, 2017, accepted, to appear.

[3] M. Muehlebach and R. D'Andrea, "The Flying Platform - a testbed for ducted fan actuation and control design", *Mechatronics*, vol. 42, no. 1, pp. 52–68, 2017.

[4] M. Muehlebach and R. D'Andrea, "On the approximation of constrained linear quadratic regulator problems and their application to model predictive control", *Automatica*, 2017, submitted, in review.

[5] M. Muehlebach and R. D'Andrea, "Distributed event-based state estimation for networked systems: An LMI-approach", *IEEE Transactions on Automatic Control*, 2017, accepted, to appear.

**Related publications**

[6] M. Muehlebach and R. D'Andrea, "Basis functions design for the approximation of constrained linear quadratic regulator problems encountered in model predictive control", *Proceedings of the International Conference on Decision and Control*, 2017, accepted.

[7] C. Sferrazza, M. Muehlebach, and R. D'Andrea, "Trajectory tracking of an unmanned aerial vehicle with a parametrized model predictive control approach", *Proceedings of the International Conference on Decision and Control*, 2017, accepted.

[8] M. Muehlebach, C. Sferrazza, and R. D'Andrea, "Implementation of a parametrized infinite-horizon model predictive control scheme with stability guarantees", *Proceedings of the International Conference on Robotics and Automation*, pp. 2723–2730, 2017.

[9] M. Muehlebach and R. D'Andrea, "Approximation of continuous-time infinite-horizon optimal control problems arising in model predictive control", *Proceedings of the International Conference on Decision and Control*, pp. 1464–1470, 2016.

[10] M. Hofer, M. Muehlebach, and R. D'Andrea, "Application of an approximate model predictive control scheme on an unmanned aerial vehicle", *Proceedings of the International Conference on Robotics and Automation*, pp. 2952–2957, 2016.

[11] M. Muehlebach and R. D'Andrea, "Parametrized infinite-horizon model predictive control for linear time-invariant systems with input and state constraints", *Proceedings of the American Control Conference*, pp. 2669–2674, 2016.

[12] M. Muehlebach and S. Trimpe, "Guaranteed $\mathcal{H}_2$ performance in distributed event-based state estimation", *Proceedings of the International Conference on Event-based Control, Communication, and Signal Processing*, 2015.

[13] M. Muehlebach and S. Trimpe, "LMI-based synthesis for distributed event-based state estimation", *Proceedings of the American Control Conference*, pp. 4060–4067, 2015.

[14] M. Muehlebach, G. Mohanarajah, and R. D'Andrea, "Nonlinear analysis and control of a reaction wheel-based 3D inverted pendulum", *Proceedings of the International Conference on Decision and Control*, pp. 1283–1288, 2013.

[15] M. Gajamohan, M. Muehlebach, and R. D'Andrea, "The Cubli: A reaction wheel based 3D inverted pendulum", *Proceedings of the European Control Conference*, pp. 268–274, 2013.

## 2.6 Student supervision

### Masters thesis

The masters thesis is a six-month, full-time project.

[1] S. Nacht, "Nonlinear MPC applied to the pendulum swing-up", Masters thesis, ETH Zurich, 2017.

[2] J. Kohler, "Aggressive quadrocopter maneuvers", Masters thesis, ETH Zurich, 2017.

[3] C. Sferrazza, "Parametrized model predictive control on the Flying Platform: Trajectory tracking and full constraint satisfaction", Masters thesis, ETH Zurich, 2016.

[4] M. Hofer, "Parametric model predictive control of the Flying Platform", Masters thesis, ETH Zurich, 2015.

### Semester project

The semester project is a semester-long, part-time project.

[1] Z. Zhejun, "Improving the trajectory tracking of a parametrized model predictive control approach", Semester project, ETH Zurich, 2017.

[2] E. Kaufmann, "Nonlinear infinite-horizon model predictive control using multi-interval polynomial trajectories", Semester project, ETH Zurich, 2016.

[3] L. Fröhlich, "Improvement of the parametric model predictive control on the Flying Platform", Semester project, ETH Zurich, 2016.

[4] J. Carius, "Nonlinear infinite-horizon model predictive control with parametric trajectories", Semester project, ETH Zurich, 2015.

[5] D. Dugas, "Cubli choreographer", Semester project, ETH Zurich, 2015.

[6] A. Widmer, "Feedback linearization of the Cubli", Semester project, ETH Zurich, 2014.

[7] T. Meier, "Implementation of the Flying Platform", Semester project, ETH Zurich, 2014.

**Internship**

[1] A. Ali, "Design of the One-Wheel Cubli (Octahedronli)", Internship, ETH Zurich, 2015.

[2] Y. Yih Tang, "Developing the wingsuit flying platform", Internship, ETH Zurich, 2014.

## 2.7 Outreach

**Talks**

Note that the talks at scientific conferences corresponding to the publications [8], [9], [11]–[14] are not listed.

| | |
|---|---|
| Jun. 2017 | *Seminar, Automatic Control Laboratory* (EPFL). |
| Nov. 2016 | *Seminar, Lehrstuhl Automatisierungstechnik und Prozessinformatik* (University of Bochum). |
| Mar. 2016 | *Coffee Talk, Automatic Control Laboratory* (ETH Zurich). |
| Oct. 2015 | *Lecture, IFM Institute for Facility Management* (ZHAW). |

**Demonstrations**

During the period of this thesis, the reaction-wheel based inverted pendulum was demonstrated at various events.

| | | |
|---|---|---|
| Jan. 2017 | Davos | World Economic Forum |
| Oct. 2016 | Zurich | National Council Switzerland |
| Oct. 2015 | Brussels | Soirée Suisse |
| Apr. 2014 | Zurich | Haus Konstruktiv |

In addition to the above, smaller demonstrations were also conducted for visitors (ranging from primary school students to distinguished professors) at the Institute for Dynamic Systems and Control.

### Youtube videos

The following videos were created as an addition to research articles and for consumption by the general public, demonstrating some of the research results.

[1] M. Muehlebach, C. Sferrazza, and R. D'Andrea, *Online model predictive control of the flying platform*, Sep. 2016. [Online]. Available: `https://youtu.be/GgIwrnoNvTY`.

[2] M. Hofer, M. Muehlebach, and R. D'Andrea, *Approximate model predictive control on the flying platform*, Mar. 2016. [Online]. Available: `https://youtu.be/_hE_bN1ylB4`.

[3] M. Muehlebach and R. D'Andrea, *Flying Platform*, Dec. 2015. [Online]. Available: `https://youtu.be/NYY9q-vs4Nw`.

### Media coverage

The reaction-wheel based inverted pendulum was featured in Galileo, a TV show of the German TV station Prosieben (on January 2016).

# 3

# Future work

This chapter provides an overview of potential future work based on the research presented in this thesis.

**A three-dimensional reaction-wheel based inverted pendulum system**

A controllability analysis revealed that a single reaction wheel is enough for stabilizing the upright equilibrium provided that the principle components of the pendulum's inertia (roll and pitch axis) are not equal and the reaction wheel is placed such that the motor's reaction torque affects both tilt directions. This is due to a separation of the time constants associated to the unstable poles. A rigid-body model is used to quantify and optimize the controllability in terms of the determinant of the controllability Gramian. It is found that for maximum controllability (in a natural set of coordinates) the principle inertia components $I_1$ and $I_2$ need to satisfy the relationship

$$\frac{I_1}{I_2} = (\sqrt{2} - 1)^2. \tag{3.1}$$

We were in the process of manufacturing a carefully designed prototype at the time of writing the thesis. Future work thus includes the experimental realization of a single-wheel three-dimensional inverted pendulum system.

**A flying vehicle actuated by ducted fans**

As highlighted in the introduction, the ducted fans have the property of redirecting crosswinds resulting in drag terms that are linear in the forward velocity. This drag term is commonly referred to as momentum drag. Compared to other flying vehicles (for example quadrotors) the duct renders the momentum-drag particularly pronounced. We conjecture that the momentum drag can be exploited for estimating the vehicle's translational velocities by means of accelerometer measurements. Thus, the addition of ducts for flying vehicles might not only improve the aerodynamic characteristics, but might also facilitate the on-board state estimation.

Potential future work could therefore aim at studying the influence of ducts and shrouded propellers on the on-board estimation capabilities of flying vehicles.

**Approximations of the constrained linear quadratic regulator problem**

The article discuss a numerical optimization routine that deals with semi-infinite constraints. As an alternative, a discrete-time point of view could be adopted, which provides a means to reduce the semi-infinite constraint (in continuous-time) to a finite number of inequality constraints, ultimately resulting in a standard quadratic program. Still, the number of inequality constraints that must be imposed for guaranteeing constraint satisfaction might be very large. Future work could aim at finding optimization algorithms that are particularly efficient for dealing with a large number of constraints and applying these in the proposed model predictive control framework.

In addition, the framework could be extended to time-varying systems and/or general nonlinear systems. The dynamics could likewise be encoded via a Galerkin approach. However, due to the fact that the dynamics are only approximated, the stability and recursive feasibility guarantees would most probably cease to hold in that case (without further assumptions). The student projects [1, masters thesis], [4, semester project] present first results of such an approach applied to an inverted-pendulum-on-a-cart system.

**A state estimation algorithm for distributed networked systems**

Each agent tries to estimate the whole state of the system with the proposed approach. Arguably, for large scale systems this might not be a very sensible approach, since certain states might be only weakly coupled. Potential future work might therefore aim at changing the architecture in such a way that each agent estimates only parts of the system's state. However, this renders the analysis of the closed-loop system much more complicated, and one must probably rely on robust control arguments (for example the small-gain theorem) to bound the effect of the states that are neglected.

# References for Chapters 1-3


[1] K. Åström and K. Furuta, "Swinging up a pendulum by energy control", *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

[2] K. Furuta, M. Yamakita, and S. Kobayashi, "Swing-up control of inverted pendulum using pseudo-state feedback", *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 206, no. 4, pp. 263–269, 1992.

[3] M. W. Spong, P. Corke, and R. Lozano, "Nonlinear control of the reaction wheel pendulum", *Automatica*, vol. 37, no. 11, pp. 1845–1851, 2001.

[4] M. Hehn and R. D'Andrea, "Real-time trajectory generation for quadrocopters", *IEEE Transactions on Robotics*, vol. 31, no. 4, pp. 877–892, 2015.

[5] S. Trimpe and R. D'Andrea, "The balancing cube: A dynamic sculpture as test bed for distributed estimation and control", *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 48–75, 2012.

[6] K. H. Lundberg and T. W. Barton, "History of inverted-pendulum systems", *Proceedings of the IFAC World Congress*, vol. 42, no. 24, pp. 131–135, 2010.

[7] K. J. Åström and R. M. Murray, *Feedback Systems*. Princeton University Press, 2009.

[8] T. Glück, A. Eder, and A. Kugi, "Swing-up control of a triple pendulum on a cart with experimental validation", *Automatica*, vol. 49, no. 3, pp. 801–808, 2013.

[9] J. Yi, N. Yubazaki, and K. Hirota, "A new fuzzy controller for stabilization of parallel-type double inverted pendulum system", *Fuzzy Sets and Systems*, vol. 126, no. 1, pp. 105–119, 2002.

[10] M. Hehn and R. D'Andrea, "A flying inverted pendulum", *Proceedings of the International Conference on Robotics and Automation*, pp. 763–770, 2011.

[11] B. Hockman, A. Frick, R. Reid, I. Nesnas, and M. Pavone, "Design, control, and experimentation of internally-actuated rovers for the exploration of low-gravity planetary bodies", *Journal of Field Robotics*, vol. 34, no. 1, pp. 5–24, 2017.

[12] J. Romanishin, K. Gilpin, and D. Rus, "M-blocks: Momentum-driven, magnetic modular robots", *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, pp. 4288–4295, 2013.

[13] A. Berry, D. Lemus, R. Babuska, and H. Vallery, "Directional singularity-robust torque control for gyroscopic actuators", *IEEE/ASM Transactions on Mechatronics*, vol. 21, no. 6, pp. 2755–2763, 2016.

[14] J. Hilkert, "Inertially stabilized platform technology", *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 26–46, 2008.

[15] R. G. Brown and P. Y. Hwang, *Introduction to Random Signal Analysis and Kalman Filtering*, Fourth. John Wiley & Sons, 2012.

[16] R. Mahony, T. Hamel, and J.-M. Pflimlin, "Nonlinear complementary filters on the special orthogonal group", *IEEE Transactions on Automatic Control*, vol. 53, no. 5, pp. 1203–1218, 2008.

[17] E. Lefferts, F. Markley, and M. Shuster, "Kalman filtering for spacecraft attitude estimation", *Journal of Guidance, Control, and Dynamics*, vol. 5, no. 5, pp. 417–429, 1982.

[18] G. Robson and R. D'Andrea, "Longitudinal stability analysis of a jet-powered wingsuit", *Proceedings of the AIAA Atmospheric Flight Mechanics Conference*, 2010.

[19] "Modeling, control, and flight testing of a small ducted-fan aircraft", *Journal of Guidance, Control, and Dynamics*, vol. 29, no. 4, pp. 769–779, 2006.

[20] "Control of aerial robots: Hybrid force and position feedback for a ducted fan", *IEEE Control Systems Magazine*, vol. 32, no. 4, pp. 43–65, 2012.

[21] J.-M. Pflimlin, P. Binetti, P. Souères, T. Hamel, and D. Trouchet, "Modeling and atttitude control analysis of a ducted-fan micro aerial vehicle", *Control Engineering Practice*, vol. 18, no. 3, pp. 209–218, 2010.

[22] J. Richalet, "Industrial applications of model based predictive control", *Automatica*, vol. 29, no. 5, pp. 1251–1274, 1993.

[23] P. Falcone, F. Borrelli, J. Asgari, H. E. Tseng, and D. Hrovat, "Predictive active steering control for autonomous vehicle systems", *IEEE Transactions on Control Systems Technology*, vol. 15, no. 3, pp. 566–580, 2007.

[24] M. Morari and J. H. Lee, "Model predictive control: Past, present and future", *Computers and Chemical Engineering*, vol. 23, no. 4, pp. 667–682, 1999.

[25] A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, "The explicit linear quadratic regulator for constrained systems", *Automatica*, vol. 38, no. 1, pp. 3–20, 2002.

[26] A. Alessio and A. Bemporad, "A survey on explicit model predictive control", in, L. Magni, D. M. Raimondo, and F. Allgöwer, Eds. Springer, 2009, pp. 345–369.

[27] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality", *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[28] L. Grüne and J. Pannek, *Nonlinear Model Predictive Control*, Second. Springer, 2017.

[29] S. Trimpe, PhD thesis, ETH Zurich, 2013.

[30] M. Hilbert and P. López, "The world's technological capacity to store, communicate, and compute information", *Science*, vol. 332, no. 6025, pp. 60–65, 2011.

[31] O. Kanoun and H.-R. Trankler, "Sensor technology advances and future trends", *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1497–1501, 2004.

[32] S. Trimpe, "Event-based state estimation: An emulation-based approach", *IET Control Theory & Applications*, vol. 11, no. 11, pp. 1684–1693, 2017.

[33] A. Domahidi and J. Jerez, *FORCES Professional*, embotech GmbH (`http://embotech.com/FORCES-Pro`), Jul. 2014.

[34] H. Ferreau, C. Kirches, A. Potschka, H. Bock, and M. Diehl, "qpOASES: A parametric active-set algorithm for quadratic programming", *Mathematical Programming Computation*, vol. 6, no. 4, pp. 327–363, 2014.

# Part A

## A three-dimensional reaction-wheel based inverted pendulum system

# Paper P1

# Nonlinear Analysis and Control of a Reaction-Wheel-Based 3D Inverted Pendulum

Michael Muehlebach and Raffaello D'Andrea

**Abstract**

This article presents control and learning algorithms for a reaction wheel-based 3D inverted pendulum. The inverted pendulum system has two main features: the ability to balance on its edge or corner and to jump from lying flat to its corner by suddenly braking its reaction wheels. Algorithms which address both features are presented. For balancing, a backstepping based controller providing global stability (almost everywhere) is derived, together with a simple tuning method based on the analysis of the resulting closed-loop system. For jump-up, a computationally efficient, gradient-based learning algorithm is provided, which is shown experimentally to converge to the correct angular velocities enabling a successful jump-up. Moreover, a controller based on feedback linearization is derived and used to track an ideal trajectory during jump-up, increasing robustness and reliability.

Published in the *IEEE Transactions on Control Systems Technology*.

## 1. Introduction

This article presents control and learning algorithms for a reaction wheel-based 3D inverted pendulum. The inverted pendulum system consists of three perpendicular reaction wheels embedded in a cubic housing. Due to its relatively small footprint, i.e. a side length of 150 mm, it is called Cubli, which is derived from the Swiss German diminutive for cube. Figure 1.1 shows the Cubli balancing on a corner. Unlike other inverted pendulum test beds, [1]–[7], and references therein, it has the ability to jump-up from a resting position without any external support by suddenly braking its reaction wheels rotating at high angular velocities. While the mechatronic design is covered in [8], and a linear controller is discussed in [9], this paper presents nonlinear control strategies and a learning algorithm enabling a successful jump-up.[1]

In [10] several design variants of a reaction wheel-based 3D inverted pendulum are compared. Moreover, a swing-up control strategy is presented based on feedforward and linear state feedback, for which local stability is shown. However, no braking system is used, which has the drawback that the design is not capable of swinging up from arbitrary positions, as the electric motors provide only limited torques.

Based on a reduced system description two nonlinear controllers are proposed herein. The first control design is based on backstepping and provides a smooth, globally (almost everywhere) stabilizing control law characterized by four tuning parameters. In contrast to earlier work, e.g. [10]–[12] the full 3D case is treated and global stability is proved (almost everywhere). The work presented in [13] is extended by relating these parameters to the closed-loop behavior, leading to a simple tuning strategy suitable for implementation.

The second control design is based on feedback linearization; an appropriate state transformation is introduced allowing for feedback linearization in the 3D case. This extends the result of [14], where the 1D (planar) case is discussed.

Both controllers are implemented on the Cubli: The controller based on backstepping is used for balancing. The controller based on feedback linearization is used for tracking predefined non-equilibrium motions; compared to other methods, such as time-varying LQR control, feedback linearization has the advantage of providing a time-invariant feedback law.

Additionally, a low-complexity model describing the jump-up is derived. The model is used to apply a gradient-based learning algorithm, similar to [15], to the Cubli and is shown experimentally to converge. To enhance the reliability of the jump-up, a predefined jump-up trajectory is tracked using the controller based on feedback linearization.

The remainder of this article is structured as follows: The dynamics are introduced in Section 2, followed by the control design in Section 3. Aspects related to the jump-up are covered in Section 4. Finally, experimental results are presented in Section 5, and the conclusions are summarized in Section 6.

[1] A video showing the Cubli can be found under `https://www.youtube.com/watch?v=n_6p-1J551Y`.

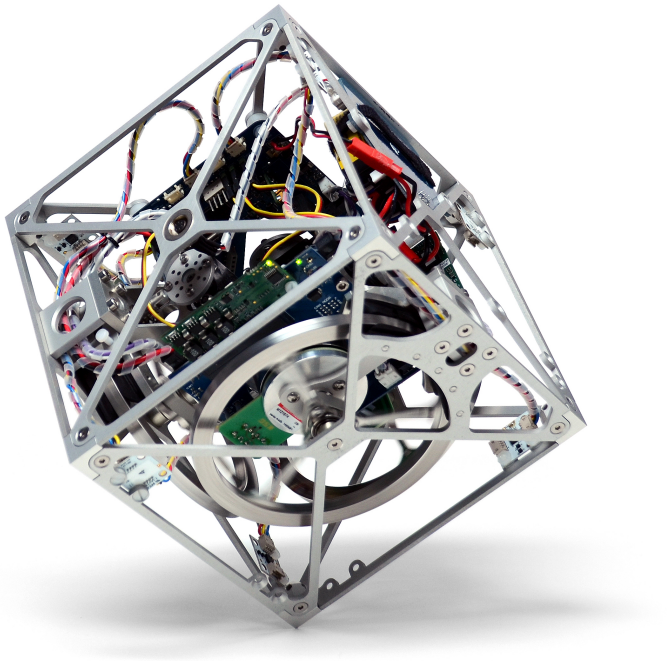

**Figure 1.1.** The Cubli balancing on a corner.

## 2. Dynamics of the Reaction Wheel-based 3D Inverted Pendulum

In this section the reaction wheel-based 3D inverted pendulum dynamics are briefly outlined. After introducing the notation, the equations of motion are presented and are used to demonstrate the conservation of angular momentum. As will be pointed out, this has important consequences for control design. Additionally, in the absence of motor torques energy is conserved. This will become important in Section 4, where an ideal jump-up trajectory is determined via the conservation of energy.

### 2.1 Notation

Let $\Theta_{wi},\ i = 1, 2, 3$ denote the moment of inertia of each reaction wheel (in the direction of the corresponding rotation axis, referred to the corresponding suspension point), and define $\Theta_w := \mathrm{diag}(\Theta_{w1}, \Theta_{w2}, \Theta_{w3})$. Let $\Theta_0 + \Theta_w$ denote the total moment of inertia of the Cubli around the pivot point $O$ (see Figure 1.2). Next, let $\vec{m}$ denote the position vector from the pivot point to the center of gravity multiplied by the total mass and $\vec{g}$ denote the gravity vector. The projection of a tensor onto a particular coordinate frame is denoted by a preceding superscript, i.e. ${}^{\mathrm{K}}\Theta_0 \in \mathbb{R}^{3\times3},\ {}^{\mathrm{K}}m \in \mathbb{R}^3$. The arrow notation is used to emphasize that a vector (and tensor) should be a priori thought of as a linear object in a normed vector space detached from its coordinate representation in a particular coordinate frame. The transformation matrix $R_{\mathrm{IK}} \in SO(3)$ relates vectors from the body-fixed frame to their representation in the inertial frame, that is ${}^{\mathrm{I}}v = R_{\mathrm{IK}}{}^{\mathrm{K}}v$,

for all vectors ${}^{\mathrm{K}}v \in \mathbb{R}^3$. Moreover, the skew symmetric matrix corresponding to a vector $a \in \mathbb{R}^3$, denoted by $\widetilde{a}$, is defined as $a \times b = \widetilde{a}b$, for all $b \in \mathbb{R}^3$, where $a \times b$ refers to the cross product of the two vectors $a$ and $b$. The Euclidean norm is referred as $|\cdot|$, i.e. $|a|^2 = a^{\mathsf{T}}a$, and $a \parallel b$ is used to indicate that the two vectors $a \in \mathbb{R}^3$ and $b \in \mathbb{R}^3$ are parallel (that is $a \times b = 0$). Additionally, the sphere of radius $|g|$ is denoted by $S^2$.

Since the body-fixed coordinate frame $\{\mathrm{K}\}$ is the most commonly projected coordinate frame, its preceding superscript is usually removed for ease of notation. That is, ${}^{\mathrm{K}}m = m$, ${}^{\mathrm{K}}\Theta_0 = \Theta_0$, etc.

Moreover, vectors are expressed as n-tuples $(x_1, x_2, \ldots, x_n)$ with dimension and stacking clear from context.

### 2.2 Equations of Motion

It was derived in [9] and [13] that the equations of motion are given by

$$\begin{aligned} \dot{p}_{\omega_h} &= -\widetilde{\omega}_h p_{\omega_h} + \widetilde{m}g, \qquad \dot{p}_{\omega_w} = T, \qquad \dot{R}_{\mathrm{IK}} = R_{\mathrm{IK}}\widetilde{\omega}_h, \\ p_{\omega_h} &:= \Theta_0\omega_h + \Theta_w(\omega_h + \omega_w), \qquad p_{\omega_w} := \Theta_w(\omega_h + \omega_w), \end{aligned} \tag{1.1}$$

where $\omega_h \in \mathbb{R}^3$ denotes the angular velocity of the Cubli housing, $\omega_w \in \mathbb{R}^3$ the angular velocity of the reaction wheels, and $T \in \mathbb{R}^3$ the motor torque applied to the reaction wheels. The fixed-body coordinate frame is aligned with the Cubli housing and therefore the first component of $\omega_w$ denotes the angular velocity of the reaction wheel pointing in ${}_K\vec{e}_1$ direction, the second component the reaction wheel pointing in ${}_K\vec{e}_2$, etc. The components of the motor torque $T$ have a similar interpretation.

The following observations are worth pointing out: The dynamics are invariant to the initial reaction wheel positions, leading to the conservation of the angular momentum $p_{\omega_w}$ in the absence of motor torques. Moreover, the evolution of all possible initial conditions over time[2] is symmetric around the gravity vector leading to the conservation of angular momentum $p_{\omega_h}^{\mathsf{T}}g$. This can be easily checked by explicit calculation:

$$\frac{d}{dt}\left(p_{\omega_h}^{\mathsf{T}}g\right) = \dot{p}_{\omega_h}^{\mathsf{T}}g + p_{\omega_h}^{\mathsf{T}}\dot{g} = p_{\omega_h}^{\mathsf{T}}\widetilde{\omega}_h g - p_{\omega_h}^{\mathsf{T}}\widetilde{\omega}_h g = 0, \tag{1.2}$$

where $\dot{g}$ is expressed by $\dot{g} = \dot{R}_{\mathrm{IK}}^{\mathsf{T}}{}^{\mathrm{I}}g = -\widetilde{\omega}_h g$, or by noting that gravity exerts no torque in direction ${}_{\mathrm{I}}\vec{e}_3$. The conservation of the angular momentum $g^{\mathsf{T}}p_{\omega_h}$ has an important consequence for control design: Independent of the control input applied, the momentum in direction $\vec{g}$ is conserved and, depending on the initial condition, it may be impossible to bring the system to rest. For example, a yaw motion in the upright position can be slowed down by increasing the velocity of the reaction wheels. However, the yaw motion and the reaction wheel velocity cannot be driven to zero at the same time. Note that the conservation of angular momentum in direction $\vec{g}$ is independent of the mass distribution

[2] Commonly referred to as the flow of the system.

or inertia of the Cubli and independent of the motor torque $T$.

In the presence of friction between the pivot point and the ground, exerting a friction torque about ${}_{\mathrm{I}}\vec{e}_3$, the angular momentum $p_{\omega_h}^{\mathsf{T}} g$ is no longer conserved, and as a result, a yaw motion in the upright position will slowly decay.

In addition, in the absence of motor torques, the total energy given by

$$\mathcal{H} = \frac{1}{2}{\omega_h}^{\mathsf{T}}\Theta_0\omega_h + \frac{1}{2}\,(\omega_h+\omega_w)^{\mathsf{T}}\Theta_w(\omega_h+\omega_w) - m^{\mathsf{T}}g - |m|\;|g|, \tag{1.3}$$

is conserved. Due to the fact that $p_{\omega_w} = \Theta_w(\omega_h + \omega_w)$ is constant for $T = 0$, the energy related to the Cubli housing, given by

$$\mathcal{H}_h = \frac{1}{2}{\omega_h}^{\mathsf{T}}\Theta_0\omega_h - m^{\mathsf{T}}g - |m|\;|g|, \tag{1.4}$$

is conserved as well. Note that the energy is normalized such that it attains zero for the upright equilibrium. The conservation of energy will become important in Section 4, where it will be used to derive an ideal jump-up trajectory.

Using the gravity vector expressed in the Cubli's body-fixed coordinate frame, i.e. $g = {R_{\mathrm{IK}}^{\mathsf{T}}}^{\mathrm{I}} g$, to represent the attitude, the dynamics given by (1.1) can be reduced to

$$\begin{aligned} \dot{p}_{\omega_h} &= -\widetilde{\omega}_h p_{\omega_h} + \widetilde{m} g, \qquad \dot{p}_{\omega_w} = T, \qquad \dot{g} = -\widetilde{\omega}_h g, \\ p_{\omega_h} &= \Theta_0\omega_h + \Theta_w(\omega_h+\omega_w), \qquad p_{\omega_w} = \Theta_w(\omega_h+\omega_w). \end{aligned} \tag{1.5}$$

This comes however at the cost of losing the yaw information. A formal treatment of this reduction step can, for example, be found in [16].

### 2.3 Equilibria

In this section the equilibria of the Cubli are briefly discussed. The reduced equations of motion (1.5) give rise to equilibria corresponding to limit cycles in the full configuration, so called relative equilibria, [17].

The relative equilibria are obtained by setting the right-hand side of (1.5) to zero, leading to

$$-\bar{\omega}_h \times \bar{p}_{\omega_h} + m \times \bar{g} = 0, \qquad \bar{T} = 0, \qquad -\bar{\omega}_h \times \bar{g} = 0, \tag{1.6}$$

where $\bar{g}$, $\bar{p}_{\omega_h}$, and $\bar{\omega}_h$ denote the equilibrium configurations. The last equation implies that $\bar{\omega}_h \parallel \bar{g}$ or likewise $\bar{\omega}_h = \lambda_1\bar{g}$, with $\lambda_1 \in \mathbb{R}$. Thus, the relative equilibria are characterized by

$$\bar{\omega}_h = \lambda_1\bar{g}, \quad \lambda_1\bar{p}_{\omega_h} + m = \lambda_2\bar{g}, \quad \bar{T} = 0, \tag{1.7}$$

with $\lambda_1, \lambda_2 \in \mathbb{R}$, $\bar{g} \in S^2$, and $\bar{\omega}_h, \bar{p}_{\omega_h}, \bar{T} \in \mathbb{R}^3$. The hanging and upright equilibria, which

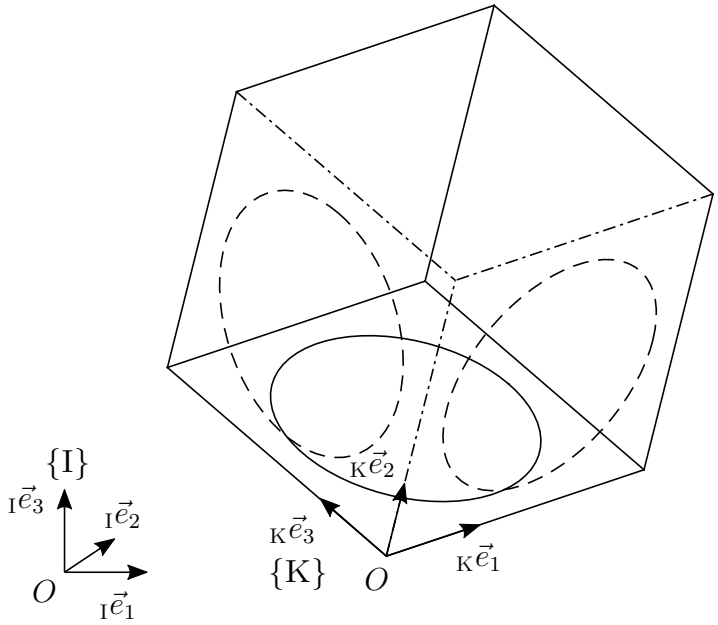


**Figure 1.2.** The Cubli balancing on its corner. The vectors $_{\mathrm{K}}\vec{e}_i$ and $_{\mathrm{I}}\vec{e}_i$, $i = 1, 2, 3$, denote the principle axes of the body fixed frame {K} and inertial frame {I}. The pivot point $O$ is the common origin of coordinate frames {I} and {K}. For illustration purposes the coordinate system {I} is shifted to the left.

are of interest for the remainder of this article, are obtained by setting $\lambda_1 = 0$ implying $\bar{g} \parallel m$. As expected, a linear analysis reveals that the upright equilibrium is unstable, while the hanging equilibrium is marginally stable.

## 3. Nonlinear Control

In the next section two different control strategies are presented, which asymptotically stabilize the upright equilibrium. The first approach is based on backstepping and provides a smooth control law characterized by four tuning parameters. In a subsequent step the tuning parameters are related to the closed-loop behavior, extending the result presented in [13]. The second approach is based on feedback linearization and extends the result in [14] to the 3D case.

For the control design and subsequent analysis the reduced dynamics (1.5) are used. The state space is chosen to be $(g, p_{\omega_h}, p_{\omega_w}) \in \mathcal{X} := S^2 \times \mathbb{R}^3 \times \mathbb{R}^3$. By using the reduced attitude representation, the feedback control laws derived next will naturally be invariant to the orientation around the gravity vector and to the reaction wheel positions.

Since the component of the angular momentum $p_{\omega_h}$ in the direction of gravity is a conserved quantity, only the component of $p_{\omega_h}$ that is orthogonal to $g$ can be affected by feedback control. Hence, it is convenient to split the angular momentum $p_{\omega_h}$ into two

parts: one in the direction of gravity, and one orthogonal to it, i.e.

$$p_{\omega_h} =: p^{\perp}_{\omega_h} + p^{g}_{\omega_h}\frac{g}{|g|}, \qquad p^{g}_{\omega_h} := p^{\mathsf{T}}_{\omega_h}\frac{g}{|g|}. \tag{1.8}$$

The control objective consists of balancing the Cubli in the upright position, and at the same time requiring $\omega_h \to 0$ together with $p^{\perp}_{\omega_h} \to 0$ as time goes to infinity. Thus, the control objective for balancing can be formulated as driving the system to the closed invariant set

$$\mathcal{T} = \{(g, p_{\omega_h}, p_{\omega_w}) \in \mathcal{X} \mid g^{\mathsf{T}} m = -|g|\ |m|,\ p^{\perp}_{\omega_h} = 0,\ p_{\omega_h} = p_{\omega_w}\}. \tag{1.9}$$

Note that $\omega_h$ is given by $\Theta_0^{-1}(p_{\omega_h} - p_{\omega_w})$ and therefore $p_{\omega_h} = p_{\omega_w}$ implies zero angular velocity of the Cubli housing.

### 3.1 Backstepping Approach

In the following section a nonlinear controller is presented, which stabilizes the set $\mathcal{T}$ asymptotically. In a subsequent step its closed-loop behavior is analyzed leading to a geometric interpretation of closed-loop trajectories and a simple tuning strategy.
For ease of notation, the hanging relative equilibria with $\omega_h = 0$ are denoted by $x^-$, i.e.

$$x^- = \{(g, p_{\omega_h}, p_{\omega_w}) \in \mathcal{X} \mid g = \frac{|g|}{|m|}m,\ p^{\perp}_{\omega_h} = 0,\ p_{\omega_h} = p_{\omega_w}\}.$$

Next, the control law

$$T = K_1 \widetilde{m} g + K_2 \omega_h + K_3 p_{\omega_h} - K_4 p_{\omega_w}, \tag{1.10}$$

with

$$\begin{aligned} K_1 =& I + (\alpha + \beta\gamma + \delta)\Theta_0, \\ K_2 =& \Theta_0\left(\alpha\widetilde{p}^{\perp}_{\omega_h} + \beta\widetilde{m}\widetilde{g}\right) + \widetilde{p}_{\omega_h}, \\ K_3 =& \gamma(I + \alpha\Theta_0(I - \frac{gg^{\mathsf{T}}}{|g|^2})), \\ K_4 =& \gamma I, \quad \alpha, \beta, \gamma, \delta > 0, \end{aligned}$$

and $I \in \mathbb{R}^{3\times 3}$ the identity matrix, is shown to asymptotically stabilize the upright equilibrium. More precisely:

**Theorem 1.** *The controller* (1.10) *renders the closed invariant set* $\mathcal{T}$ *of the system* (1.5) *stable and asymptotically stable on* $x \in \mathcal{X} \setminus x^-$.

*Proof.* Consider the following Lyapunov candidate function $V : \mathcal{X} \to \mathbb{R}$,

$$V(x) = \frac{1}{2}\alpha p_{\omega_h}^{\perp\mathsf{T}} p_{\omega_h}^{\perp} + m^{\mathsf{T}} g + |m|\ |g| + \frac{1}{2\delta} z^{\mathsf{T}} \Theta_0^{-2} z, \tag{1.11}$$

with $z := \Theta_0 \left(\alpha p_{\omega_h}^{\perp} + \beta \widetilde{m} g\right) + p_{\omega_h} - p_{\omega_w}$.

Clearly, there exists a $\mathcal{K}_\infty$ function[3] $a : [0, \infty) \to [0, \infty)$ such that $V(x) \geq a(|x - x_0|)$ for all $x \in \mathcal{X}$ and all $x_0 \in \mathcal{T}$. Furthermore $V(x = x_0) = 0$ implies $x = x_0$, where $x_0 \in \mathcal{T}$. Therefore $V$ is a positive definite function and a valid Lyapunov candidate.

Next, $\dot{V}$ is evaluated along trajectories of the closed-loop system:

$$\begin{aligned}\dot{V}(x) &= \alpha p_{\omega_h}^{\perp\mathsf{T}} \dot{p}_{\omega_h}{}^{\perp} + m^{\mathsf{T}} \dot{g} + \frac{1}{\delta} z^{\mathsf{T}} \Theta_0^{-2} \dot{z} \\ &= m^{\mathsf{T}} \widetilde{g}(\alpha p_{\omega_h}^{\perp} + \omega_h) + \frac{1}{\delta} z^{\mathsf{T}} \Theta_0^{-2} \dot{z}.\end{aligned}$$

From the identity $\Theta_0^{-1} z = \alpha p_{\omega_h}^{\perp} + \beta \widetilde{m} g + \omega_h$ it follows that

$$\begin{aligned}\dot{V}(x) &= m^{\mathsf{T}} \widetilde{g}(\beta \widetilde{g} m + \Theta_0^{-1} z) + \frac{1}{\delta} z^{\mathsf{T}} \Theta_0^{-2} \dot{z} \\ &= -\beta (\widetilde{g} m)^{\mathsf{T}} (\widetilde{g} m) + z^{\mathsf{T}} \Theta_0^{-1} \widetilde{m} g + \frac{1}{\delta} z^{\mathsf{T}} \Theta_0^{-2} \dot{z}.\end{aligned}$$

Moreover, the control input $T$ can be rewritten as

$$T = \frac{d}{dt}(z + p_{\omega_w}) + \gamma z + \delta \Theta_0 \widetilde{m} g. \tag{1.12}$$

Using the fact that $\dot{p}_{\omega_w} = T$, the closed loop evolution of the auxiliary variable $z$ is given by

$$\dot{z} = -\gamma z - \delta \Theta_0 \widetilde{m} g, \tag{1.13}$$

which can be used to simplify $\dot{V}$ to

$$\dot{V}(x) = -\beta (\widetilde{g} m)^{\mathsf{T}} (\widetilde{g} m) - \frac{\gamma}{\delta} z^{\mathsf{T}} \Theta_0^{-2} z \leq 0, \qquad \forall x \in \mathcal{X}.$$

Since $\dot{V}(x) \leq 0$, for all $x \in \mathcal{X}$, we conclude from Lyapunov's stability theorem, [18, Theorem 4.8] that the equilibria $x_0 \in \mathcal{T}$ are stable.

To prove asymptotic stability of the set $\mathcal{T}$ for $x \in \mathcal{X} \setminus x^-$, the set

$$\mathcal{R} := \{x \in \mathcal{X} \setminus x^- \mid \dot{V}(x) = 0\} \tag{1.14}$$

[3] A continuous function belongs to class $\mathcal{K}_\infty$ if it is strictly increasing and radially unbounded, see e.g. [18, Definition 4.2, p. 144].

is considered in more detail. From $\dot{V}(x) < 0$ for all $x \in \mathcal{X} \setminus (\mathcal{R} \bigcup x^-)$ it can be inferred that any trajectory in $\mathcal{X} \setminus x^-$ is converging to an invariant set contained in $\mathcal{R}$. The condition $\dot{V}(x) = 0$ leads to $z = 0$, $m$ parallel $g$, such that $\mathcal{R}$ can be rewritten as $\mathcal{R} = \{x \in \mathcal{X} \setminus x^- \mid m \parallel g,\ p_{\omega_w} = \alpha\Theta_0 p^{\perp}_{\omega_h} + p_{\omega_h}\}$. The dynamics on $\mathcal{R}$ can be simplified to:

$$\begin{aligned} g \parallel m \Rightarrow g = -\frac{m}{|m|}|g| \Rightarrow \dot{g} = 0 \\ \Rightarrow \omega_h \parallel g \quad \text{because} \quad \dot{g} = -\widetilde{\omega}_h g \end{aligned} \tag{1.15}$$

$$\begin{aligned} g \parallel m,\ z = 0 \Rightarrow \omega_h = \alpha p^{\perp}_{\omega_h} \\ \Rightarrow \omega_h \parallel p^{\perp}_{\omega_h} \end{aligned} \tag{1.16}$$

However, since $p^{\perp}_{\omega_h}$ is orthogonal to $g$ by definition, equations (1.15) and (1.16) imply $\omega_h = 0$ and $p^{\perp}_{\omega_h} = 0$. Therefore $\mathcal{T}$ is the largest invariant set contained in $\mathcal{R}$. This implies by the Krasovskii–LaSalle principle [18, Theorem 4.4], that for any trajectory $x(t)$,

$$\lim_{t\to\infty} x(t) = x_f, \quad x(0) \in \mathcal{X} \setminus x^-, \quad x_f \in \mathcal{T}.$$

□

*1) Remarks*

*a) Interpretation of the Lyapunov Function:* The Lyapunov function given by (1.11) can be found via a backstepping approach, see for example [18] or [19] for an introduction to backstepping. The reduced Lyapunov function

$$V_R(x) = \frac{1}{2}\alpha p^{\perp\mathsf{T}}_{\omega_h} p^{\perp}_{\omega_h} + m^{\mathsf{T}} g + |m|\ |g|, \tag{1.17}$$

which is independent of the momentum $p_{\omega_w}$ can be used to demonstrate stability given that $p_{\omega_w} = \alpha\Theta_0 p^{\perp}_{\omega_h} + p_{\omega_h} + \beta\widetilde{m}g$ (corresponding to $z = 0$). Therefore, $z$ accounts for the momentum $p_{\omega_w}$ and penalizes indirectly non-zero wheel velocities.

*b) Extension of the Controller:* In practice, modeling errors can cause steady-state deviations, e.g. an erroneous estimate of the center of gravity leads to non-vanishing steady-state reaction wheel velocities when balancing. Integral control can be used to prevent these steady-state deviations. Therefore the controller is extended with the state $z_{\text{int}}$, i.e. $\hat{u} = u + \nu z_{\text{int}}$, where

$$z_{\text{int}}(t) = z_0 + \int_0^t z(\tau)d\tau$$

and $\nu > 0$. In that case, closed-loop stability can be proved by augmenting the Lyapunov function given by (1.11):

$$V_I(x) = V(x) + \frac{\nu}{2\delta} z^{\mathsf{T}}_{\text{int}} \Theta_0^{-2} z_{\text{int}}.$$

In [13] an alternative approach to account for non-zero steady-state wheel velocities is presented, which has the advantage of directly providing an estimate of the center of gravity.

*c) Interpretation of the Control Law:* Rewriting (1.10) yields

$$u = \dot{p}_{\omega_h} + \gamma p_{\omega_h} + \alpha\Theta_0(\dot{p}^{\perp}_{\omega_h} + \gamma p^{\perp}_{\omega_h}) + \Theta_0\widetilde{m}(\beta\dot{g} + (\delta + \gamma\beta)g) - \gamma p_{\omega_w}, \tag{1.18}$$

where

$$p_{\omega_w} = u_0 + \int_0^t u(\tau)d\tau. \tag{1.19}$$

Therefore the controller given by (1.10) is a linear PID controller in the variables $p_{\omega_w}, p^{\perp}_{\omega_h}$ and $g$. The only nonlinearity of the controller lies in the projection of $p_{\omega_h}$ into $p^{\perp}_{\omega_h}$ and $p^{g}_{\omega_h}$. Nevertheless, the control law guarantees global asymptotic stability (almost everywhere) as has been shown previously.

*2) Closed-loop behavior* Due to its smoothness and its dependence on only four tuning parameters, the controller is well-suited for practical implementation. A simple tuning strategy based on the closed-loop behavior is outlined next. We will analyze the closed-loop response subject to two different initial conditions, providing an interpretation of the tuning parameters. In the first case, the Cubli will be released at rest, but with a non-zero inclination angle. For this specific initial condition the closed-loop dynamics of the inclination angle are given by a third-order differential equation, which allows for pole placement. It will be shown that there is a set of tuning parameters matching every desired pole location (provided that the desired poles have negative real parts). This determines three of the four tuning parameters ($\alpha, \beta$ and $\delta$). In the second case, a pure yaw motion will be analyzed and related to the remaining tuning parameter $\gamma$.

**Proposition 1.** *Consider the controller* (1.10) *applied to the system governed by* (1.5) *with initial conditions at* $t = 0$ *such that* $p_{\omega_h}(0)$ *and* $\omega_h(0)$ *are parallel to* $m \times g(0) \neq 0$. *Then it holds for all* $t > 0$ *that* $\omega_h(t)$, $p_{\omega_h}(t)$, *and* $m \times g(t)$ *remain parallel.*

*Proof.* Since $p_{\omega_h}(0) \parallel m \times g(0)$ it implies that $p^{\perp}_{\omega_h}(t) = p_{\omega_h}(t)$ for all $t > 0$. Moreover, by combining the control law given by (1.10) with the system dynamics it follows that

$$\begin{aligned}\dot{\omega}_h &= \Theta_0^{-1}(\dot{p}_{\omega_h} - T) \\ &= \alpha\omega_h \times p_{\omega_h} - (\alpha + \beta\gamma + \delta)m \times g + \beta m \times (\omega_h \times g) - \gamma(\alpha p_{\omega_h} + \omega_h),\end{aligned} \tag{1.20}$$

together with

$$\dot{p}_{\omega_h} = p_{\omega_h} \times \omega_h + m \times g \quad \text{and} \quad \frac{d}{dt}(m \times g) = m \times (g \times \omega_h).$$

Note also that from the Lagrange identity, [20],

$$m \times (g \times (m \times g)) = -m^{\mathsf{T}} g \; m \times g \tag{1.21}$$

follows. Assume that $p_{\omega_h}(t^*)$, $\omega_h(t^*)$, and $m \times g(t^*)$ are parallel at time $t = t^*$. Together with equations (1.20)-(1.21) these assumptions imply that

$$\frac{d}{dt}(m \times g(t^*)) \parallel m \times g(t^*), \tag{1.22}$$

$$\dot{\omega}_h(t^*) \parallel m \times g(t^*), \quad \text{and} \tag{1.23}$$

$$\dot{p}_{\omega_h}(t^*) \parallel m \times g(t^*). \tag{1.24}$$

Hence, $p_{\omega_h}(t)$, $\omega_h(t)$, and $m \times g(t)$ will remain parallel for an infinitesimal time increment $\mathrm{d}t$, that is at time $t = t^* + \mathrm{d}t$. By induction, the vectors $p_{\omega_h}(t)$, $\omega_h(t)$, and $m \times g(t)$ will therefore remain parallel for all times $t > t^*$. Note that the right-hand side of the closed-loop dynamics is locally Lipschitz, which implies the local existence and uniqueness of closed-loop trajectories, [21]. Since the initial conditions at $t = 0$ are such that $p_{\omega_h}(0)$, $\omega_h(0)$, and $m \times g(0)$ are parallel, the result follows. □

Note that the previous proposition applies especially in the case where the Cubli is initialized with zero body angular velocity and zero wheel velocity ($\omega_h(0) = \omega_w(0) = 0$), and states that the Cubli's center of mass will never leave the plane normal to $m \times g(0)$ for all times $t > 0$. This sets the stage for deriving a differential equation describing the inclination angle in closed-loop provided that $p_{\omega_h}$, $\omega_h$, and $m \times g$ are parallel at $t = 0$.

It is convenient to introduce the unit vector

$$e_\varphi := \frac{m \times g(0)}{|m \times g|}, \qquad \text{where} \quad m \times g(0) \neq 0, \tag{1.25}$$

and define the inclination angle by

$$\varphi := \arccos\left(-\frac{m^{\mathsf{T}} g}{|m| \; |g|}\right), \tag{1.26}$$

with $\varphi \in [0, \pi]$ for $g \in S^2$. Note that

$$\sin\varphi = \frac{|-g \times m|}{|m| \; |g|} = \frac{|m \times g|}{|m| \; |g|} \tag{1.27}$$

holds. By Proposition 1 it follows that $\omega_h$ is parallel to $m \times g$ and $e_\varphi$ for all times $t > 0$. Furthermore, from (1.26) and the system dynamics (1.5) it can be confirmed that

$\omega_h = \dot{\varphi} e_\varphi$. Rewriting (1.20) yields

$$\begin{aligned}\dot{\omega}_h &= \alpha\omega_h \times p_{\omega_h} - (\alpha + \beta\gamma + \delta)m \times g + \beta m \times (\omega_h \times g) - \gamma(\alpha p_{\omega_h} + \omega_h) \\ &= -e_\varphi(\alpha + \beta\gamma + \delta)|m|\ |g| \sin\varphi - e_\varphi \beta |m|\ |g| \dot{\varphi} \cos\varphi - \gamma(\alpha p_{\omega_h} + e_\varphi \dot{\varphi}).\end{aligned} \tag{1.28}$$

Taking the time derivative of the previous equation and using the fact that $\dot{e}_\varphi = 0$ and $\dot{p}_{\omega_h} = e_\varphi |m|\ |g| \sin\varphi$ results in

$$\begin{aligned}\dddot{\varphi} + (\beta|m|\ |g| \cos\varphi + \gamma)\ddot{\varphi} + (\alpha + \beta\gamma + \delta)|m|\ |g|\dot{\varphi}\cos\varphi \\ - \beta|m|\ |g|\dot{\varphi}^2 \sin\varphi + \gamma\alpha|m|\ |g| \sin\varphi = 0.\end{aligned} \tag{1.29}$$

Linearizing (1.29) around the upright equilibrium, i.e. $\varphi = 0$ yields

$$\dddot{\varphi} + (\beta|m|\ |g| + \gamma)\ddot{\varphi} + (\alpha + \beta\gamma + \delta)|m|\ |g|\dot{\varphi} + \gamma\alpha|m|\ |g|\varphi = 0 \tag{1.30}$$

and provides a method to relate the closed-loop poles to the parameters $\{\alpha, \beta, \gamma, \delta\}$. To simplify notation the following scaling is introduced, $\hat{\alpha} := \alpha|m|\ |g|$, $\hat{\beta} := \beta|m|\ |g|$, $\hat{\gamma} := \gamma$, and $\hat{\delta} := \delta|m|\ |g|$, such that (1.30) reads as

$$\dddot{\varphi} + (\hat{\beta} + \hat{\gamma})\ddot{\varphi} + (\hat{\alpha} + \hat{\beta}\hat{\gamma} + \hat{\delta})\dot{\varphi} + \hat{\gamma}\hat{\alpha}\varphi = 0. \tag{1.31}$$

Moreover, the parameter $\hat{\gamma}$ is related to the closed-loop yaw motion, by considering the case where the Cubli is initialized in an upright relative equilibrium, with non-zero angular velocity, $\omega_h(0) \neq 0$. Hence, it follows that $\omega_h(0) \parallel p_{\omega_h}(0) \parallel g(0) \parallel m$ and that the closed-loop dynamics read as

$$\dot{g} = 0, \quad \dot{p}_{\omega_h} = 0, \quad \text{and} \quad \dot{p}_{\omega_w} = -\hat{\gamma}(p_{\omega_w} - p_{\omega_h}). \tag{1.32}$$

This leads to the interpretation of $\hat{\gamma}$ as a time constant prescribing how fast the yaw rotation is slowed down.

Ideally, the parameters $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}$ are chosen such that the desired closed-loop poles of the inclination angle are matched and that a prescribed time constant of the closed-loop yaw motion is met. However, it turns out that depending on $\hat{\gamma}$, this might be impossible, i.e. for a fixed $\hat{\gamma} > 0$ it might be impossible to obtain $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}$ such that the pole configuration is met, while guaranteeing nonlinear closed-loop stability with the proposed controller. This fact is illustrated in the following.

For given closed-loop pole locations of the inclination angle, let the third order characteristic polynomial corresponding to (1.31) be denoted by

$$s^3 + As^2 + Bs + C = 0, \tag{1.33}$$

where the coefficients $\{A, B, C\}$ are related to the pole locations by a homeomorphism. Therefore it is sufficient to analyze the function[4] $f : \mathbb{R}_+^4 \to \mathbb{R}_+^3$ mapping the tuning parameters $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}\}$ to the constants $\{A, B, C\}$. According to the Routh-Hurwitz criterion the poles have strictly negative parts if and only if the conditions $A > 0, B > 0, AB > C > 0$ are fulfilled, see e.g. [22]. Clearly, if $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta} > 0$, it follows that $A > 0, B > 0, AB > C > 0$, which corresponds to a stable pole configuration (as expected, nonlinear closed-loop stability implies linear closed-loop stability). The converse is not true; for a fixed $\hat{\gamma} > 0$ there might be no $\hat{\alpha}, \hat{\beta}, \hat{\delta} > 0$, such that the desired pole location is matched. This fact is illustrated by expressing the level set of $f(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}) = (A, B, C)$ as

$$\{(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}) \in \mathbb{R}^4 \mid \hat{\alpha}\hat{\gamma} = C,\ \hat{\beta} = A - \hat{\gamma},\ \hat{\gamma}\hat{\delta} = \hat{\gamma}^3 - A\hat{\gamma}^2 + B\hat{\gamma} - C\}. \tag{1.34}$$

Hence, given that $A, B, C > 0$ the condition $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta} > 0$ reduces to

$$A > \hat{\gamma} > 0, \quad h(\hat{\gamma}) := \hat{\gamma}^3 - A\hat{\gamma}^2 + B\hat{\gamma} - C > 0. \tag{1.35}$$

Note that $h(-s) = -(s^3 + As^2 + Bs + C)$ holds, which implies that the zeros of $h(\hat{\gamma})$ are just the negative values of the desired poles. Thus, if the desired pole locations, $s_0$, $s_1$, and $s_2$, with $s_2 < s_1 < s_0 < 0$, are all real and distinct, then there are two different $\hat{\gamma}$-regions, e.g. $\hat{\gamma} \in (-s_0, -s_1)$, $\hat{\gamma} \in (-s_2, A)$, where $h(\hat{\gamma}) > 0$, see Fig. 1.3. If there are two complex conjugated poles or non-distinct poles then there might be only one $\hat{\gamma}$-region, where $h(\hat{\gamma}) > 0$, see Fig. 1.4.

Note that in all cases $\hat{\gamma}$ needs to be greater than $\min_i\{-\text{real}(s_i)\}$, where $s_0, s_1, s_2$ are the desired pole locations. Hence, the closed-loop yaw motion needs to have a time constant at least as fast as the smallest pole of the (closed-loop) inclination angle dynamics, in order to guarantee global closed-loop stability with the proposed controller.

Summarizing, the following tuning recipe is proposed:

1) Choose the desired pole locations of the closed-loop inclination angle dynamics, which determines possible intervals for $\hat{\gamma}$.

2) Choose $\hat{\gamma}$ within those intervals such that the time constant of the yaw-dynamics matches the desired one as close as possible. Solving (1.34) yields the parameters $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}\}$.

### 3.2 Feedback Linearization

Next an explicit input-to-state feedback linearization is found extending the result presented in [14]. The generalized momentum $p_{\omega_h}$ is chosen to be the virtual output. However, to remove the conserved component (in direction $\vec{g}$), it is convenient to project $p_{\omega_h}$ in the

[4]The positive real numbers are denoted by $\mathbb{R}_+ := \{x \in \mathbb{R} | x > 0\}$.

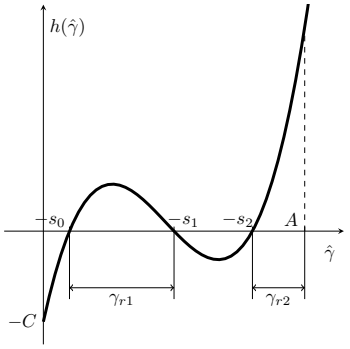


**Figure 1.3.** Example for a desired pole configuration with three real poles. The admissible regions for $\hat{\gamma}$ are denoted by $\gamma_{r1}$ and $\gamma_{r2}$.

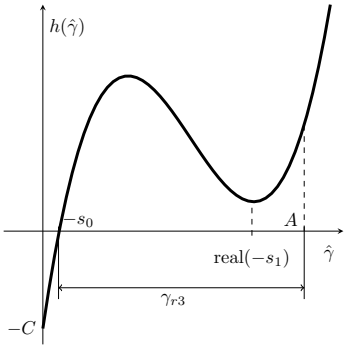


**Figure 1.4.** Example for a desired pole configuration with one complex pole pair. The admissible region for $\hat{\gamma}$ is denoted by $\gamma_{r3}$.

inertial frame, where the dynamics of the Cubli are given by

$$\begin{aligned} {}^{\mathrm{I}}\dot{m} &= {}^{\mathrm{I}}\omega_h \times {}^{\mathrm{I}}m, \\ {}^{\mathrm{I}}\dot{p}_{\omega_h} &= {}^{\mathrm{I}}m \times {}^{\mathrm{I}}g, \\ {}^{\mathrm{I}}\dot{p}_{\omega_w} &= {}^{\mathrm{I}}T + {}^{\mathrm{I}}\omega_h \times {}^{\mathrm{I}}p_{\omega_w}. \end{aligned} \tag{1.36}$$

The virtual output $y$ is formed by the first two elements of ${}^{\mathrm{I}}p_{\omega_h}$, i.e.

$$y := \left({}^{\mathrm{I}}p_{\omega_h 1}, {}^{\mathrm{I}}p_{\omega_h 2}\right), \tag{1.37}$$

since the third component of ${}^{\mathrm{I}}p_{\omega_h}$ is conserved. This choice can be motivated by the feedback linearization of the 1D reaction wheel-based inverted pendulum presented in [14]. Using the matrices

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad \text{and} \qquad P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{1.38}$$

the first two components of the cross product $a \times b$ with $a \in \mathbb{R}^3$, $b \in \mathbb{R}^3$ can be expressed by

$$P(a \times b) = -a_3 JPb + b_3 JPa. \tag{1.39}$$

Thus, $P({}^{\mathrm{I}}m \times {}^{\mathrm{I}}g)$ simplifies to $P({}^{\mathrm{I}}m \times {}^{\mathrm{I}}g) = -|g|JP\ {}^{\mathrm{I}}m$.

Taking the time derivative of $y$, $\dot{y}$, and $\ddot{y}$ leads to

$$\dot{y} = -|g|JP\ {}^{\mathrm{I}}m, \tag{1.40}$$

$$\ddot{y} = -|g|JP({}^{\mathrm{I}}\omega_h \times {}^{\mathrm{I}}m), \tag{1.41}$$

$$\dddot{y} = -|g|JP({}^{\mathrm{I}}\dot{\omega}_h \times {}^{\mathrm{I}}m + {}^{\mathrm{I}}\omega_h \times {}^{\mathrm{I}}\dot{m}). \tag{1.42}$$

Additionally, ${}^{\mathrm{I}}\dot{\omega}_h$ is given by

$$ {}^{\mathrm{I}}\dot{\omega}_h = R_{\mathrm{IK}}\Theta_0^{-1}(\dot{p}_{\omega_h} - \dot{p}_{\omega_w}) = R_{\mathrm{IK}}\Theta_0^{-1}(m \times g - \omega_h \times p_{\omega_h} - T). \tag{1.43} $$

Solving for the input torque $T$, i.e. using the change of variable $T \to {}^{\mathrm{I}}v$ with

$$ T = -\omega_h \times p_{\omega_h} + m \times g - \Theta_0 R_{\mathrm{IK}}^{\mathsf{T}}\, {}^{\mathrm{I}}v \tag{1.44} $$

leads to ${}^{\mathrm{I}}\dot{\omega}_h = {}^{\mathrm{I}}v$. Using the identity given by (1.39) allows us to rewrite (1.42) as

$$ \ddot{y} = |g| \left( {}^{\mathrm{I}}m_3 P\, {}^{\mathrm{I}}v - {}^{\mathrm{I}}v_3 P\, {}^{\mathrm{I}}m - JP\, {}^{\mathrm{I}}\widetilde{\omega}_h\, {}^{\mathrm{I}}\widetilde{\omega}_h {}^{\mathrm{I}}m \right). \tag{1.45} $$

Choosing the first two components of ${}^{\mathrm{I}}v$ to be

$$ P^{\mathrm{I}}v = \frac{1}{{}^{\mathrm{I}}m_3}({}^{\mathrm{I}}v_3 P\, {}^{\mathrm{I}}m + JP\, {}^{\mathrm{I}}\widetilde{\omega}_h\, {}^{\mathrm{I}}\widetilde{\omega}_h {}^{\mathrm{I}}m + \frac{1}{|g|}w) \tag{1.46} $$

with $w \in \mathbb{R}^2$ leads to $\ddot{y} = w$. Note that the transformation is not defined for ${}^{\mathrm{I}}m_3 = 0$. This parallels the 1D case, where it was shown that a feedback linearization exists only for an inclination angle $\varphi$ such that $\varphi \neq \pm\frac{\pi}{2}$, see [14].

Hence, by choosing the state transformation

$$ x = (y, \dot{y}, \ddot{y}, {}^{\mathrm{I}}\omega_{h3}), \tag{1.47} $$

together with the input transformation given by (1.44) and (1.46) the following linear system dynamics are obtained for the case ${}^{\mathrm{I}}m_3 \neq 0$:

$$ \dot{x} = \begin{pmatrix} 0_{2\times2} & I_{2\times2} & 0_{2\times2} & 0_{2\times1} \\ 0_{2\times2} & 0_{2\times2} & I_{2\times2} & 0_{2\times1} \\ 0_{2\times2} & 0_{2\times2} & 0_{2\times2} & 0_{2\times1} \\ 0_{1\times2} & 0_{1\times2} & 0_{1\times2} & 0 \end{pmatrix} x + \begin{pmatrix} 0_{4\times3} \\ I_{3\times3} \end{pmatrix} \begin{pmatrix} w \\ {}^{\mathrm{I}}v_3 \end{pmatrix}. \tag{1.48} $$

## 4. Jump Up

By suddenly braking its reaction wheels spinning at high angular velocities, the Cubli is able to "jump up" from lying flat to its upright equilibrium as shown in Figure 1.5.

The jump-up is divided into two parts: the *braking phase*, where the reaction wheels are almost instantaneously slowed down and the *guiding phase*, where additional control action is used to guide the Cubli to its upright equilibrium. Identifying and modeling the braking phase exactly is difficult due to large process uncertainties such as the friction

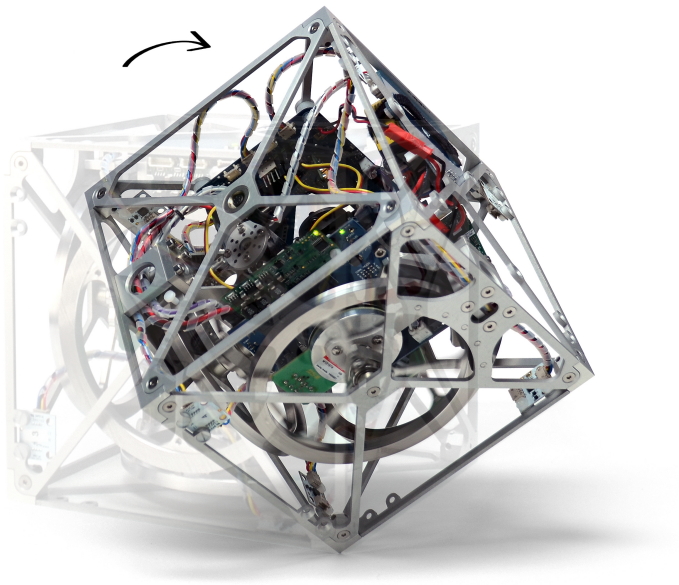

**Figure 1.5.** The Cubli jumping from lying flat to its upright equilibrium.

between the brake and the wheel, the timing of the different brakes and the inaccuracies in the state estimation due to high accelerations. However, these uncertainties are mostly time invariant and can therefore be circumvented by using a low-order model in combination with a learning algorithm. The learning algorithm accounts therefore for the repeatable modeling errors, and is used to adapt the initial wheel velocities of the reaction wheels.

To further improve the reliability of the jump-up, an ideal trajectory is tracked during the guiding phase using feedback linearization. Compared to a linear reference tracking approach, this has the advantage of providing a time-invariant control law.

The next section is divided into the following parts. First, a low-complexity model for the jump-up is outlined for both the braking and the guiding phase. Then, the learning framework is introduced and discussed in general, before being applied to the Cubli jump-up.

### 4.1 Impact-based Braking Model

The jump-up is modeled by assuming that the reaction wheels are stopped instantaneously. To simplify the analysis further, it is assumed that after braking the angular momentum associated with the reaction wheels is zero, that is $p_{\omega_w}(0)^+ = \Theta_w(\omega_h(0)^+ + \omega_w(0)^+) = 0$. This assumption is used to determine an ideal jump-up trajectory; it guarantees the conservation of angular momentum around the figure axis in the absence of control inputs, reducing the Cubli model to a symmetric spherical pendulum (see Section 4.2). Note that this assumption is not entirely fulfilled since in reality the wheel speed $\omega_w(0)^+$ is actually zero after braking. Compared to the reaction wheel momentum before braking, $p_{\omega_w}(0)^+$ is however negligible. The braking is assumed to happen at the

time instant 0; $\omega_w(0)^-$ and $\omega_w(0)^+$ denote the left and right limits of the reaction wheel angular velocity $\omega_w$. Note that the left and right limits of a discontinuous function $f$ (of locally bounded variation) are defined by

$$f(0)^- := \lim_{t \uparrow 0} f(t) \qquad \text{and} \qquad f(0)^+ := \lim_{t \downarrow 0} f(t). \tag{1.49}$$

The impact is modeled by using conservation of angular momentum. More formally, an impact torque density $\mathrm{d}\Lambda$ ($[\mathrm{d}\Lambda] = \mathrm{Nms}$) is introduced and the equations of motion given by (1.5) are integrated over the impact time singleton $\{0\}$. This yields

$$\int_{\{0\}} \mathrm{d}p_{\omega_h} = p_{\omega_h}(0)^+ - p_{\omega_h}(0)^- = \int_{\{0\}} (-\widetilde{\omega}_h p_{\omega_h} + \widetilde{m} g)\, \mathrm{d}t = 0, \tag{1.50}$$

$$\int_{\{0\}} \mathrm{d}p_{\omega_w} = p_{\omega_w}(0)^+ - p_{\omega_w}(0)^- = \int_{\{0\}} (T \mathrm{d}t + \mathrm{d}\Lambda) = \Lambda(0)^+ - \Lambda(0)^-, \tag{1.51}$$

where $\mathrm{d}p_{\omega_h}$ and $\mathrm{d}p_{\omega_w}$ are the differential measures of $p_{\omega_h}$ and $p_{\omega_w}$, containing a density with respect to the Lebesgue measure $\mathrm{d}t$ and the atomic measure $\mathrm{d}\eta$, i.e.

$$\begin{aligned} \mathrm{d}p_{\omega_h} &= \dot{p}_{\omega_h} \mathrm{d}t + (p^+_{\omega_h} - p^-_{\omega_h}) \mathrm{d}\eta, \\ \mathrm{d}p_{\omega_w} &= \dot{p}_{\omega_w} \mathrm{d}t + (p^+_{\omega_w} - p^-_{\omega_w}) \mathrm{d}\eta. \end{aligned} \tag{1.52}$$

The time singleton $\{0\}$ has zero Lebesgue measure. By assumption, it holds that $\Theta_w(\omega_h(0)^+ + \omega_w(0)^+) = p_{\omega_w}(0)^+ = 0$. Since the Cubli is at rest when activating the brakes, $\omega_h(0)^- = 0$, and therefore (1.50) yields

$$p_{\omega_h}(0)^+ = \Theta_0 \omega_h(0)^+ = p_{\omega_h}(0)^- = \Theta_w \omega_w(0)^-, \tag{1.53}$$

which relates the body angular momentum after braking to the initial wheel velocity.

### 4.2 Guiding Phase

During the guiding phase, the Cubli is guided along an "ideal" trajectory to the upright equilibrium. The trajectory is tracked using feedback linearization, resulting in a time-invariant control law. Next, this predefined trajectory is derived by using first integrals of the equations of motion.

To simplify the analysis, the following assumption is made:

*Assumption* 4.1. (Symmetric housing inertia) The inertia tensor $\Theta_0$ has an eigenvector in direction $m$. The associated eigenvalue is denoted by $I_3$. The remaining two eigenvalues are equal, i.e. $I_1 = I_2$.

In case $T = 0$ and $p_{\omega_w} = 0$, this assumption leads to an additional conserved quantity,

which is nothing but the angular momentum around the figure axis, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(m^{\mathsf{T}} p_{\omega_h}\right) = m^{\mathsf{T}}(\omega_h \times p_{\omega_h}) = m^{\mathsf{T}}\widetilde{\omega}_h \Theta_0 \omega_h + m^{\mathsf{T}}\widetilde{\omega}_h p_{\omega_w} = 0, \tag{1.54}$$

where the first term of the previous expression vanishes due to Assumption 4.1 and the second due to the fact that $p_{\omega_w} = 0$.

The "ideal" trajectory is defined as the trajectory leading from the state just after braking, i.e. the right limit at time $t = 0$, to the upright equilibrium without using any motor torque. By assumption, the right limit of $p_{\omega_w}$ vanishes at time $t = 0$, which implies that $p_{\omega_w}(t)$ remains zero for all $t > 0$, see Section 2. In the absence of motor torque, energy, the angular momentum in direction $\vec{g}$, and the angular momentum in direction $\vec{m}$ are conserved (see (1.2), (1.4), and (1.54)), that is

$$\begin{aligned} &\mathcal{H}_h = \frac{1}{2}\omega_h^{\mathsf{T}}\Theta_0\omega_h - m^{\mathsf{T}} g - |m|\ |g| = \text{const}, \\ &p^g_{\omega_h} = p^{\mathsf{T}}_{\omega_h}\frac{g}{|g|} = \text{const}, \qquad p^m_{\omega_h} = p^{\mathsf{T}}_{\omega_h}\frac{m}{|m|} = \text{const}. \end{aligned} \tag{1.55}$$

In other words, the Cubli is modeled as a symmetric spherical pendulum during the guiding phase. It has as many first integrals as degrees of freedom. This suggests to parametrize the attitude of the Cubli by the inclination angle

$$\varphi := \arccos\left(\frac{-m^{\mathsf{T}} g}{|m|\ |g|}\right) \in [0, \pi]. \tag{1.56}$$

Since the ideal trajectory is supposed to lead to the upright equilibrium, with $g_0 = -\frac{m}{|m|}|g|$, $p_{\omega_{h_0}} = 0$ and $p_{\omega_{w_0}} = 0$ it follows that $p^g_{\omega_h} = 0$, $p^m_{\omega_h} = 0$, and $\mathcal{H} = 0$ along the motion. Thus, the angular momentum can only have a component orthogonal to $g$ and $m$, and is therefore simplified to

$$p_{\omega_h} = p^{\varphi}_{\omega_h}\ e_\varphi,$$

where the unit vector $e_\varphi$ is given by

$$e_\varphi = \frac{m \times g}{|m \times g|}, \qquad \text{for} \quad m \times g \neq 0. \tag{1.57}$$

From the condition that the ideal trajectory lies on the zero energy surface it can be inferred that

$$(p^{\varphi}_{\omega_h})^2 = \frac{2(m^{\mathsf{T}} g - |m|\ |g|)}{e_\varphi^{\mathsf{T}}\Theta_0^{-1} e_\varphi} = 2I_1|m|\ |g|(1 - \cos\varphi), \tag{1.58}$$

with $I_1 = e_\varphi^{\mathsf{T}}\Theta_0 e_\varphi$, which is constant. Due to a vanishing wheel momentum $p_{\omega_w} = 0$, it

follows from $\omega_h = \Theta_0^{-1} p_{\omega_h}$ and the system dynamics that

$$\omega_h = \frac{1}{I_1} p^{\varphi}_{\omega_h} e_\varphi = \dot{\varphi}\ e_\varphi. \tag{1.59}$$

Hence, along the ideal trajectory the Cubli follows the great circle of $S^2$ passing through the upright equilibrium represented by the north pole. The trajectory is implicitly parametrized by (1.58), by prescribing the angular momentum as a function of the inclination angle $\varphi$.

This "ideal" trajectory is tracked using the controller presented in Section 3.2. To that extent the error $y - y_{\text{des}}$ is introduced, with $y$ defined according to (1.37). From Section 3.2 it can be inferred that

$$\begin{aligned} \ddot{e} = \ddot{y} - \ddot{y}_{\text{des}} &= {}^{I}w - \ddot{y}_{\text{des}} := u_1 \\ {}^{\text{I}}\dot{\omega}_{h3} - {}^{\text{I}}\dot{\omega}_{h3_{\text{des}}} &= {}^{\text{I}}v_3 - {}^{\text{I}}\dot{\omega}_{h3_{\text{des}}} := u_2. \end{aligned} \tag{1.60}$$

Using $x = (e, \dot{e}, \ddot{e}, {}^{\text{I}}\omega_{h3} - {}^{\text{I}}\omega_{h3_{\text{des}}})$, the error dynamics are rewritten as

$$\dot{x} = \begin{pmatrix} 0_{2\times2} & I_{2\times2} & 0_{2\times2} & 0_{2\times1} \\ 0_{2\times2} & 0_{2\times2} & I_{2\times2} & 0_{2\times1} \\ 0_{2\times2} & 0_{2\times2} & 0_{2\times2} & 0_{2\times1} \\ 0_{1\times2} & 0_{1\times2} & 0_{1\times2} & 0 \end{pmatrix} x + \begin{pmatrix} 0_{4\times3} \\ I_{3\times3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{1.61}$$

Thus, a time-invariant state feedback controller, e.g. $u = (u_1, u_2) = Kx$ can be used to stabilize the error dynamics. The controller gain $K \in \mathbb{R}^{3\times7}$ can be found by linear control strategies such as a linear quadratic regulator approach or pole placement. Once the virtual control inputs $u_1$ and $u_2$ are determined, the resulting input torque is calculated by solving ${}^{\text{I}}w$ and ${}^{\text{I}}v_3$ for $T$. This transformation, given by (1.44) and (1.46) is bijective, except when the Cubli is inclined by 90 degrees.[5]

For tracking the ideal jump-up trajectory we impose that ${}^{\text{I}}\omega_{h3_{\text{des}}} = 0$ and ${}^{\text{I}}\dot{\omega}_{h3_{\text{des}}} = 0$ together with

$$\begin{aligned} y_{\text{des}} &= I_1 \dot{\varphi}_{\text{des}}\ PR_{IK} e_\varphi, \\ \dot{y}_{\text{des}} &= |m|\ |g| \sin\varphi\ PR_{IK} e_\varphi, \\ \ddot{y}_{\text{des}} &= |m|\ |g| \cos\varphi\ \dot{\varphi}_{\text{des}}\ PR_{IK} e_\varphi, \\ \dddot{y}_{\text{des}} &= \frac{|m|^2 |g|^2}{I_1} \sin\varphi\ (3\cos\varphi - 2)\ PR_{IK} e_\varphi, \\ \dot{\varphi}_{\text{des}} &:= \sqrt{\frac{2|m|\ |g|}{I_1}(1 - \cos\varphi)}. \end{aligned} \tag{1.62}$$

[5] In practice, an inclination of 90 degrees can never occur.

The formulas are obtained by mere differentiation and using (1.58), which prescribes the desired angular momentum as a function of the inclination angle.

### 4.3 Learning Algorithm

For adapting the initial wheel velocities $\omega_w(0)^-$ a learning algorithm is used. The Cubli therefore makes multiple jump trials and evaluates the quality of each jump according to predefined criteria. The initial wheel velocities are adjusted using a model-based gradient descent method. In the next section the learning framework is elaborated in more detail.

*1) Gradient-based Learning* The learning strategy used can be seen as a variation of the Newton procedure for finding the roots of a differentiable function. It has recently been presented and successfully implemented in [15].

The underlying process, e.g. the Cubli jump-up, is assumed to be dependent on the parameter vector $\theta \in \mathbb{R}^p$, which can be adjusted, as well as the unknown parameters $s \in \mathbb{R}^q$.[6] The goal is to adjust the parameters $\theta$ such that a certain error $e \in \mathbb{R}^m$ vanishes. In the case of the Cubli jump-up, we would like to adapt the initial wheel velocities $\omega_w(0)^-$ such that the upright equilibrium is reached without using additional control. The dependence of the error on the parameters $(\theta, s)$ is described by the mapping $E : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^m$. The error dimension $m$ is assumed to be smaller or equal than the number of parameters $p$ that can be adjusted ($m \leq p$).

A model based on nominal parameters $s_0$ is assumed to be known, which predicts the error $E(\theta, s_0)$. Based on this model, the parameters $\theta^0$ leading to a vanishing error $E(\theta^0, s_0) = 0$ can be inferred, together with the gradient of $E$ with respect to $\theta$, evaluated at $\theta^0$ and $s_0$. Still, the parameters of the real system, $s^*$, are unknown. By performing experiments, e.g. jump-up attempts, we can access noisy measurements of the error, $E^i = E(\theta, s^*) + N^i$, where $N^i$ are bounded disturbances, $|N^i| < D$, $i = 0, 1, 2, \ldots$. The goal is therefore to iteratively find the zero of the function $E(\cdot, s^*)$ for unknown parameters $s^*$. A natural solution is to use Newton's method. However, since the gradient of $E$ with respect to $\theta$ is unknown for $s = s^*$, the model-based approximation is used instead.

This leads to the following, simple and computationally efficient update rule for the parameters $\theta$

$$\theta^{i+1} = \theta^i - \lambda^i \left.\frac{\partial E}{\partial \theta}\right|^{\dagger}_{\theta^0, s_0} E^i, \qquad i = 0, 1, 2, \ldots, \tag{1.63}$$

with $\lambda^i \in (0, 2)$ a predefined sequence of step sizes, $i = 0, 1, 2, \ldots$, and where $^\dagger$ denotes the pseudoinverse.

*2) Application to the Cubli* Next, the learning algorithm is applied to the Cubli jump-up. By suddenly braking its reaction wheels rotating at high speeds the Cubli is able to jump up from lying flat to the edge-balancing position, from the edge-balancing position to the corner balancing position, and from lying flat to the corner balancing position.

[6] As pointed out in [15] the vector of unknown parameters can be infinite dimensional.

The analysis is restricted to the face to the corner jump-up (initially lying flat, jump-up to the corner), as the other cases can be treated in a similar manner.

From the modeling in Section 4.1 and 4.2 it can be concluded that the Cubli has essentially three degrees of freedom. The analysis suggests further to split them into a rotation around its center of mass $\vec{m}$, a rotation around the gravity vector $\vec{g}$ and a rotation around the direction perpendicular to $\vec{m}$ and $\vec{g}$. For a successful jump-up, where the upright equilibrium is reached with zero angular velocity, each degree of freedom must be controlled. Therefore, the error is chosen to be composed of the angular momentum in direction $\vec{m}$, the angular momentum in direction $\vec{g}$ and the energy $\mathcal{H}_h$, each of them evaluated at the top point

$$E(\omega_w(0)^-, s) = \begin{pmatrix} p^m_{\omega_h}(t_t) \\ p^g_{\omega_h}(t_t) \\ \mathcal{H}_h(t_t) \end{pmatrix}.$$

The top point is defined as the time instant $t = t_t$ at which the Cubli has either reached the upright position

$$g(t_t) = -\frac{m}{|m|}|g|$$

or has no angular momentum in direction $\vec{m} \times \vec{g}$, i.e. $p_{\omega_h}(t_t)^\mathsf{T}(m \times g(t_t)) = 0$. The parameters to be adjusted are the initial wheel velocities $\omega_w(0)^- \in \mathbb{R}^3$, whereas the vector $s$ contains unknown system parameters, e.g. the inertia, the center of mass, the parameters related to the brake properties, etc. Clearly, the error vanishes only if the Cubli reaches the upright equilibrium.

According to the model derived in Section 4.1 and 4.2 the error components are all conserved quantities in the absence of the input torque $T$. Hence

$$\begin{aligned} p^m_{\omega_h}(t^t) &= p^m_{\omega_h}(0)^+ = p^m_{\omega_h}(0)^- = m^\mathsf{T}\Theta_w\omega_w(0)^- \\ p^g_{\omega_g}(t^t) &= p^g_{\omega_h}(0)^+ = p^g_{\omega_h}(0)^- = g(0)^\mathsf{T}\Theta_w\omega_w(0)^- \end{aligned} \tag{1.64}$$

and

$$\begin{aligned} \mathcal{H}_h(t^t) = \mathcal{H}_h(0)^+ &= \frac{1}{2}(\omega_h(0)^+)^\mathsf{T}\Theta_0\omega_h(0)^+ - m^\mathsf{T}g(0) - |m|\ |g| \\ &= \frac{1}{2}(\omega_w(0)^-)^\mathsf{T}\Theta_w\Theta_0^{-1}\Theta_w\omega_w(0)^- - m^\mathsf{T}g(0) - |m|\ |g|. \end{aligned} \tag{1.65}$$

This implies that the gradient with respect to $\omega_w(0)^-$ evaluated for the model parameters

$s_0$ yields

$$\left.\frac{\partial E}{\partial \omega_w(0)^-}\right|_{s_0} = \begin{pmatrix} m^\mathsf{T}\Theta_w \\ (g(0)^-)^\mathsf{T}\Theta_w \\ (\omega_w(0)^-)^\mathsf{T}\Theta_w\Theta_0^{-1}\Theta_w \end{pmatrix}. \tag{1.66}$$

The initial guess $\theta^0 = (\omega_w(0)^-)^0$ is calculated by requiring the model-based error to vanish. This yields according to Section 4.2

$$\begin{aligned} (\omega_w(0)^-)^0 &= \sqrt{2I_1|m|\ |g|(1-\cos\varphi_0)}\ e_\varphi(0), \\ e_\varphi(0) &= \frac{m\times g(0)}{|m\times g(0)|}, \end{aligned} \tag{1.67}$$

with $\varphi_0$ the inclination angle when the Cubli is lying on its face.

*3) Compensation for the Guiding Control Action* In the previous section the error function evaluating the quality of a jump-up trial has been introduced and its gradient based on the jump-up model has been derived. Therefore the update rule given by (1.63) can be applied to learn the initial wheel velocities, which lead the Cubli to its upright equilibrium without any control action.

In practice however, not every jump-up succeeds as the process noise, e.g. the randomness in the braking mechanism is too high. To increase the chances of a successful jump-up the guiding controller introduced in Section 4.2 is used. The controller tries to maintain the Cubli on a successful jump-up trajectory and is activated after releasing the brakes. Naturally, the control effort of the guiding controller must be considered when evaluating the error criterion $E(\theta^i, s^*)$. In other words, given the value $E(\theta^i, s^*)$, the jump-up performance $E'(\theta^i, s^*)$ which would have been obtained if no additional control action would have been applied needs to be determined. Since the error $E$ is composed of conserved quantities (in the absence of motor torque) it suffices to estimate their values shortly after braking, which yields

$$E'(\omega_w(0)^-, s) = \begin{pmatrix} p^m_{\omega_h}(t_t) - \int_0^{t_t} \dot{p}^m_{\omega_h}\mathrm{d}t \\ p^g_{\omega_h}(t_t) \\ \mathcal{H}_h(t_t) - \int_0^{t_t} \dot{\mathcal{H}}_h\mathrm{d}t \end{pmatrix}. \tag{1.68}$$

Note, that the momentum around the $g$ axis is constant, regardless of the motor torque. The time derivative of the momentum around $m$ is obtained from the reduced system dynamics, (1.5) and is given by

$$\dot{p}^m_{\omega_h} = \frac{m}{|m|}^\mathsf{T}(-\omega_h \times p_{\omega_h}).$$

Moreover the rate of change of the energy related to the Cubli housing, $\mathcal{H}_h$, can be calculated to be $\dot{\mathcal{H}}_h = -\omega_h^\mathsf{T} T$.

Clearly, if the jump-up is ideal (in the sense of Section 4.2), no correction is applied and therefore $E$ and $E'$ agree. Moreover, the error $E'$ can be simplified to

$$E'(\omega_w(0)^-, s) = \left(p^m_{\omega_h}(0)^+, p^g_{\omega_h}(0)^+, \mathcal{H}_h(0)^+\right), \tag{1.69}$$

leading to the conclusion that the gradient of $E'$ with respect to $\theta$ is likewise given by the right hand side of (1.66) for $s = s_0$.

The jump-up procedure is summarized by Algorithm 1.

**Procedure 1** Cubli Jump Up

**procedure** JumpUp(Initial guess $(\omega_w(0)^-)^0$, Step sizes $\lambda^i$)
  $\theta^0 \leftarrow (\omega_w(0)^-)^0$
  $i = 0$
  **while** Not converged **do**
    Set $\theta^i$ to be the initial wheel velocities
    Speed up wheels, brake and apply guiding controller
    **while** Top point is not reached **do**
      Approximate $\int_0^{t_t} \dot{p}^m_{\omega_h} dt$ and $\int_0^{t_t} \dot{\mathcal{H}}_h dt$ by trapezoidal integration
    **end while**
    Calculate $E'(\theta^i, s^*)$ according to (1.68)
    $\theta^{i+1} \leftarrow \theta^i - \lambda^i \left.\frac{\partial E'}{\partial \theta}\right|^\dagger_{\theta^0, s_0} E'(\theta^i, s^*)$ according to (1.63)
    $i \leftarrow i + 1$
  **end while**
**end procedure**

# 5. Experimental Results

In the following section the experimental results are discussed. The control algorithms are implemented on a Cortex M4 processor with a sampling time of 20 ms, except for the guiding controller, which runs at 10 ms. The algorithm presented in [2] is used for state estimation. The state estimation exploits the fact that there is a single pivot point being always at rest to derive a computationally light-weight, nonlinear attitude estimator. It is therefore "model free", in the sense that the estimation is solely based on a kinematic model and does not require knowledge of the center of gravity nor the inertia.

## 5.1 Balancing Performance

For balancing, an additional offset-correction filter is implemented, which accounts for modeling errors in the parameter $m$. Details of the implementation can be found in [13].

The controller parameters are tuned using the strategy presented in Section 3 and are chosen to be $\alpha = 15$, $\beta = 18$, $\gamma = 12$ and $\delta = 10^{-5}$. This yields closed-loop poles of the inclination angle located at $-32.7$ rad/s, $-12.0$ rad/s, and $-0.86$ rad/s and a time constant for the yaw motion of 0.083 s. With those parameters a root mean squared inclination angle error (at steady state) below $0.025\,^\circ$ can be observed.

Disturbance rejection measurements are depicted in Figure 1.6 and 1.7. The disturbance was chosen to be 0.17 Nm and was applied to a single wheel for 60 ms. After less than 1.8 s the inclination angle reaches steady state. Note that the reaction wheels are barely turning in steady state (the jitter visible in Figure 1.7 is due to measurement noise).

### 5.2 Tracking Performance

Next, the tracking performance is evaluated. Simple state feedback in the transformed error variable $e$ is used, that is $u = Kx$ with $x$ and $u$ defined according to (1.61). The feedback gain $K$ is chosen such that the linearization of the controller around the upright equilibrium agrees with the linearization of the balancing controller.

Figure 1.8 shows the evolution of $\dot{y}$. Note that according to (1.40), $\dot{y}_2$ is proportional to ${}^{\mathrm{I}}m_1$ and $-\dot{y}_1$ to ${}^{\mathrm{I}}m_2$. Therefore the graph can be interpreted as the time evolution of the center of mass in the inertial frame. Although the center of mass is initially away from the ideal trajectory, the tracking controller manages to guide the Cubli back to the desired path. As soon as the center of mass is close enough to the upright equilibrium, i.e. reaches the region indicated by the dotted arc in Figure 1.8, the balancing controller takes over. Figure 1.9 shows the time evolution of the controller states $y$, which is associated to the momentum ${}^{\mathrm{I}}p_{\omega_h}$ and $\ddot{y}$, which is proportional to ${}^{\mathrm{I}}\omega_h \times {}^{\mathrm{I}}m$. The reference trajectory is again depicted by the dashed curves. It follows from Figure 1.9 that the generalized momentum ${}^{\mathrm{I}}p_{\omega_h}$ is accurately tracked. The error in the second derivative $\ddot{y}$ is initially larger, but is decreased by the controller as time evolves. However, a slight overshoot can be observed.

### 5.3 Learning Performance

The learning algorithm proposed in the previous section is implemented for the face to corner jump. A constant step size of $\lambda^i = 0.8$ for all iterations $i = 0, 1, 2\ldots$ is used. Figure 1.10 shows the evolution of the initial wheel speeds $\omega_{w_1}$ and $\omega_{w_2}$. Due to the geometry of the Cubli, the third reaction wheel is only slightly used to correct for a non-zero momentum $p^g_{\omega_h}$ and is therefore not depicted. The initial wheel speeds were chosen to be around $100\,{}^{\mathrm{rad}}/_{\mathrm{s}}$ away from the angular velocities leading to a successful jump-up. Hence, for the initial wheel speeds the Cubli barely moves or falls on the opposite side. After around 5 trials, the error of the angular momentum $p_{\omega_h}(t_0)^+$ is small enough such that the guiding controller can lead the Cubli to its upright equilibrium. At this point the learning algorithm is only compensating for the control action of the guiding controller leading to small correction steps.

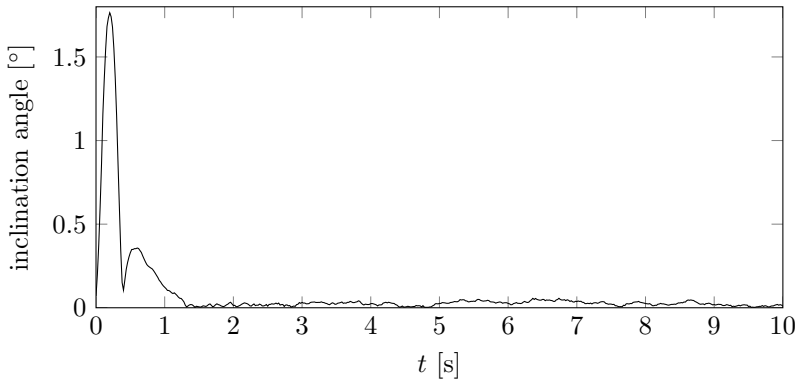


**Figure 1.6.** Disturbance rejection measurements. Depicted is the inclination angle over time. Note the inclination angle is not measured directly but estimated using the algorithm presented in [2].

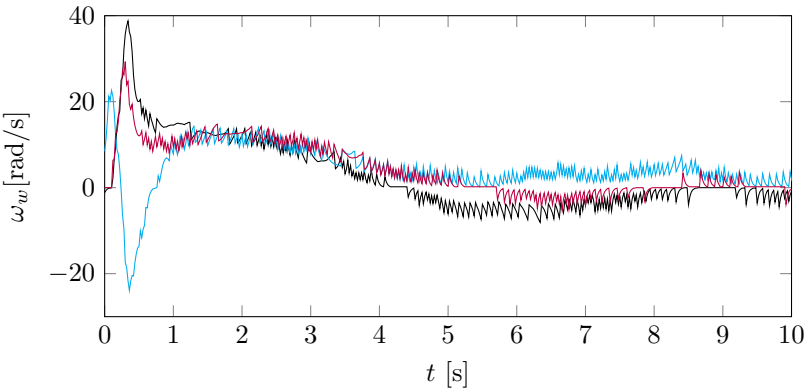


**Figure 1.7.** Disturbance rejection measurements. Depicted are the reaction wheel velocities over time, which are directly measured via a hall sensor. The different colors correspond to the different elements of the vector $\omega_w$.

## 6. Conclusion

This article presents aspects related to the dynamics and control of a reaction wheel-based 3D inverted pendulum. The analysis of the equations of motion revealed the existence of conserved quantities and relative equilibria, and allowed to find a reduced description of the dynamics. In particular, the reduced description was used for the control design. Two different nonlinear control approaches were presented and subsequently discussed. Finally, aspects related to the jump-up were presented, where the effect of repeatable disturbances was decreased by an iterative learning algorithm. To enhance robustness, feedback linearization was used to guide the inverted pendulum system to its upright equilibrium on a predefined trajectory. All control and learning algorithms were evaluated in experiments, which confirmed their effectiveness.

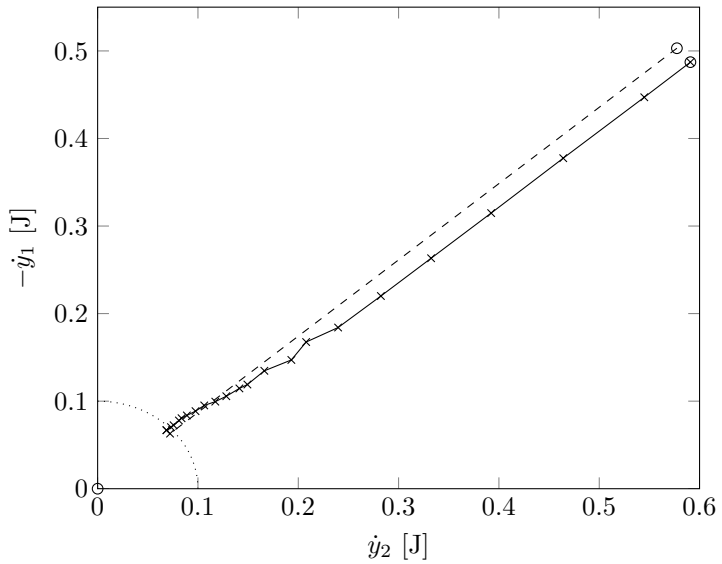


**Figure 1.8.** Trajectory tracking: Depicted is the evolution of $\dot{y}_1$ and $\dot{y}_2$ together with the ideal trajectory (dashed) for a successful jump-up. The black crosses indicate the sampling instants. The starting points (right after braking) of the ideal and actual trajectory are marked by black circles. The point $(0, 0)$ denotes the upright equilibrium. The area around the upright equilibrium separated by the dotted circle arc represents the balancing region, i.e. the region where the tracking controller is turned off and the balancing controller takes over.

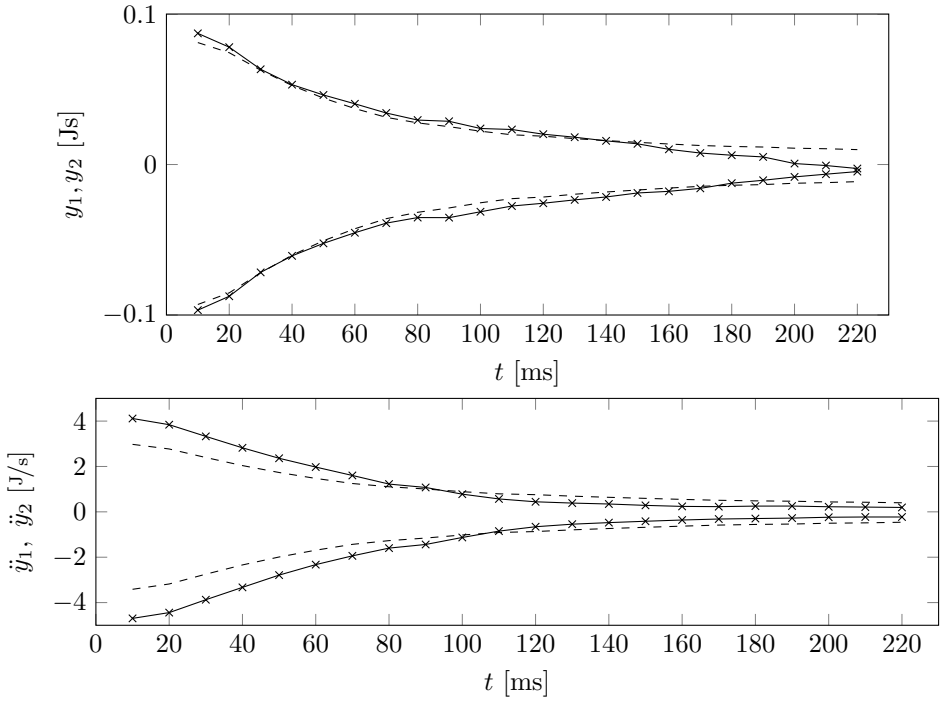


**Figure 1.9.** Trajectory tracking: Depicted is the evolution of $y$ and $\ddot{y}$ (solid), where the crosses indicate the sampling instants. The ideal trajectories, $y_{\text{des}}$ and $\ddot{y}_{\text{des}}$ are shown by the dashed curves.

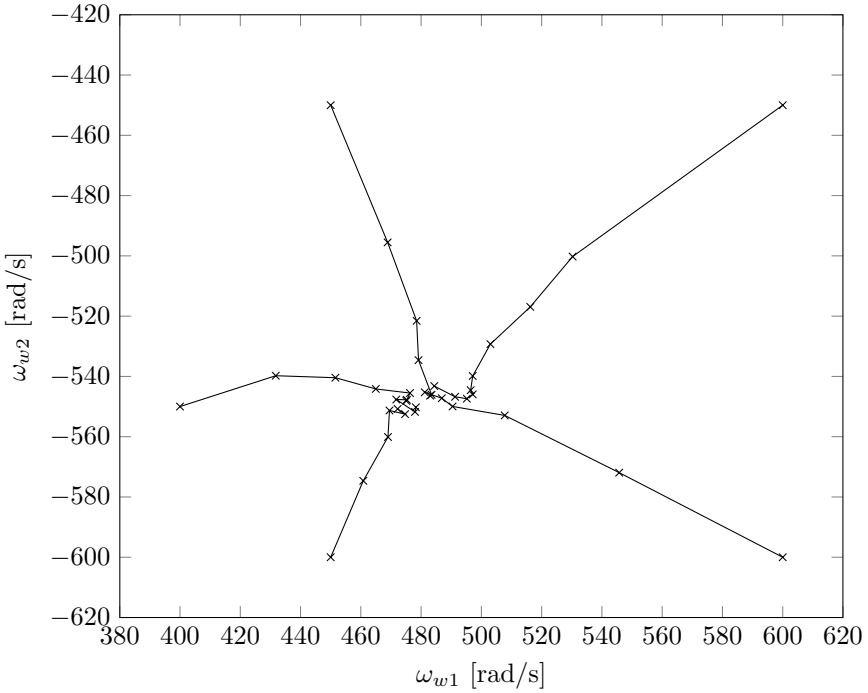


**Figure 1.10.** Depicted are the initial wheel speeds of the reaction wheels starting from five different initial conditions. The learning algorithm converges after few iterations to feasible wheel speeds resulting in a successful jump-up.

## References

[1] D. S. Bernstein, N. H. McClamroch, and A. Bloch, "Development of air spindle and triaxial air bearing testbeds for spacecraft dynamics and control experiments", *American Control Conference*, pp. 3967–3972, 2001.

[2] S. Trimpe and R. D'Andrea, "Accelerometer-based tilt estimation of a rigid body with only rotational degrees of freedom", *International Conference on Robotics and Automation*, pp. 2630–2636, 2010.

[3] M. W. Spong and D. J. Block, "The Pendubot: A mechatronic system for control research and education", *International Conference on Decision and Control*, pp. 555–556, 1995.

[4] D. J. Block, K. J. Åström, and M. W. Spong, *The Reaction Wheel Pendulum*. Morgan & Claypool Publishers, 2007.

[5] K. Åström and K. Furuta, "Swinging up a pendulum by energy control", *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

[6] J. Shen, A. K. Sanyal, N. A. Chaturvedi, D. Bernstein, and H. McClamroch, "Dynamics and control of a 3D pendulum", *International Conference on Decision and Control*, pp. 323–328, 2004.

[7] W. Zhong and H. Röck, "Energy and passivity based control of the double inverted pendulum on a cart", *International Conference on Control Applications*, pp. 896–901, 2001.

[8] M. Gajamohan, M. Merz, I. Thommen, and R. D'Andrea, "The Cubli: A cube that can jump up and balance", *International Conference on Intelligent Robots and Systems*, pp. 3722–3727, 2012.

[9] M. Gajamohan, M. Muehlebach, T. Widmer, and R. D'Andrea, "The Cubli: A reaction wheel based 3D inverted pendulum", *European Control Conference*, 2013.

[10] J. Mayr, F. Spanlang, and H. Gattringer, "Mechatronic design of a self-balancing three-dimensional inertia wheel pendulum", *Mechatronics*, vol. 30, pp. 1–10, 2015.

[11] R. Olfati-Saber, "Global stabilization of a flat underactuated system: The inertia wheel pendulum", *International Conference on Decision and Control*, pp. 3764–3765, 2001.

[12] B. Bapiraju, K. Srinivas, P. Prem. Kumar, and L. Behera, "On balancing control strategies for a reaction wheel pendulum", *Annual IEEE India Conference*, pp. 199–204, 2004.

[13] M. Muehlebach, M. Gajamohan, and R. D'Andrea, "Nonlinear analysis and control of a reaction wheel-based 3D inverted pendulum", *International Conference on Decision and Control*, pp. 1283–1288, 2013.

[14] M. W. Spong, P. Corke, and R. Lozano, "Nonlinear control of the reaction wheel pendulum", *Automatica*, vol. 37, no. 11, pp. 1845–1851, 2001.

[15] S. Lupashin and R. D'Andrea, "Adaptive fast open-loop maneuvers for quadrocopters", *Autonomous Robots*, vol. 33, no. 1, pp. 89–102, 2012.

[16] N. A. Chaturvedi, T. Lee, M. Leok, and N. H. McClamroch, "Nonlinear dynamics of the 3D pendulum", *Journal of Nonlinear Science*, vol. 21, no. 1, pp. 3–32, 2011.

[17] V. Arnold, *Mathematical methods of classical mechanics*. Springer, 1989.

[18] H. K. Khalil, *Nonlinear Systems*. Prentice Hall, 1996.

[19] M. Krstic, I. Kanellakopoulos, and P. Kokotovic, *Nonlinear and Adaptive Control Design*. John Wiley & Sons, 1995.

[20] K. Itô, *Encyclopedic dictionary of mathematics*. MIT press, 1993.

[21] F. Scheck, *Mechanics: From Newton's laws to deterministic chaos*. Springer, 2010.

[22] L. Meirovitch, *Methods of analytical dynamics*. Courier Dover Publications, 2010.

# Paper P2

# Accelerometer-Based Tilt Determination for Rigid Bodies with a Non-Accelerated Pivot Point

Michael Muehlebach and Raffaello D'Andrea

**Abstract**

An estimation algorithm is proposed for determining pitch and roll angles (tilt), angular velocities, and angular accelerations of a rigid body with a non-accelerated pivot point. The estimation uses only accelerometer measurements. It is based on a kinematic model of the rigid body and is therefore independent of its dynamics; only the mounting positions of the sensors need to be known. Simulation results indicate a significant performance increase compared to an existing method, a claim which is supported by experimental results.

Accepted for publication in *IEEE Transactions on Control Systems Technology*.

## 1. Introduction

The problem of determining the attitude of a rigid body relative to an inertial frame occurs in many engineering disciplines ranging from robotics to aeronautics and space engineering. We propose an algorithm that estimates the tilt (pitch and roll angles) of a rigid body based on accelerometer measurements only, by exploiting the assumption of a non-accelerated pivot point. The tilt estimate is obtained by maximizing the likelihood of the sensor measurements. Hence, the approach does not require temporal correlation in the accelerometer data (e.g. as given by a dynamical model) and only relies on a kinematic rigid-body model, where only the mounting positions of the accelerometers relative to the pivot need to be known. As such, the method is independent of physical parameters such as the inertia, the mass, and the center of mass, and less susceptible to modeling errors, as, for example, a process noise model is not required.

The method has been successfully applied to estimate the tilt of balancing robots, see for example [1], [2]. Other potential applications are inertially stabilized platforms, [3], including various gimbal mountings that are used, for example, for sensor calibration and image stabilization, [4], or to orient and stabilize optical elements such as mirrors, wedges, prisms, and lenses, [5].

A tri-axis accelerometer[7] is sufficient to determine the pitch and roll of a non-moving (or uniformly translating) rigid body directly from a single (tri-axis) accelerometer measurement. In case the rigid body is rotating or is accelerated, however, the body-fixed accelerometer also measures angular and centripetal acceleration terms.

One method to compensate for these effects is complementary filtering, where a gyroscope and an accelerometer-based tilt estimate are combined, see e.g. [6, p. 165]. Thereby the fact that the gyroscope-based estimate is corrupted mainly by low frequency noise (drift of the gyroscope) and the accelerometer-based estimate is mainly accurate at low frequencies is exploited. Various applications and extensions of complementary filtering can be found in the literature, for instance [7], [8], [9], [10]. Typically, the implicit assumption of the accelerometer being at rest is made in order to extract roll and pitch estimates from the accelerometer measurements, see for example [7]. In contrast to the non-accelerated pivot assumption made here, this assumption is even more stringent and might often be violated in practice.

Alternative approaches include for example extended or unscented Kalman filtering, see for example [11]–[14], [15]. These approaches are typically based on local approximations of the attitude dynamics. Moreover, the estimation relies on a dynamic model of the system in order to capture the temporal correlation of sensor data. The model includes a process noise model, and as such, the filter might be susceptible to modeling errors and/or might require careful tuning. In case a model based on the kinetics of the rigid body is used, as for example done in [13], physical parameters, such as the inertia, the mass, and the center of mass are needed. In case of a kinematics-based model,

[7] Throughout this article, we will refer to a tri-axis accelerometer simply as an accelerometer.

again (stringent) assumptions on the acceleration of the acceleremoters (e.g. zero mean, or at rest) are required. Most approaches therefore rely on additional sensors, such as magnetometers in [14] and vector sun sensors and star trackers in [12].

In contrast to these Kalman filter-based approaches, we obtain a tilt estimate by maximizing the measurement likelihood. Thereby, we do not rely on a dynamic model that describes the temporal correlation of the sensor data (although the method might be extended to account for temporal correlation by including prior distributions). As a result, the attitude estimation is formulated as an optimization problem that fully respects the nonlinearities of the attitude dynamics. We exploit the fact that measurements from multiple accelerometers are available to compensate for the angular and centripetal acceleration terms. However, if an accurate dynamic model of the system is available, our approach might be outperformed by strategies that indeed consider the temporal correlation of the accelerometer measurements. Moreover, the approach leads to an iterative optimization algorithm, where the number of iterations needed for convergence depends on the initial guess. Therefore, unless early termination is used, the proposed estimation algorithm might have a variable execution time, which might be undesirable in some applications.

In [16], it is shown that with the fixed pivot assumption one can determine the attitude by solving a linear least-squares problem. The algorithm is successfully implemented on various balancing robots, see for example [1], [2], [17]. However, the particular mathematical structure of the angular and centripetal acceleration terms is not taken into account, thereby sacrificing estimation performance.

In this work, we extend the approach from [16] by formulating the attitude estimation as a maximum likelihood estimate and taking the structure of the angular and centripetal acceleration terms explicitly into account. This leads to a constrained least-squares problem for which a dedicated solution algorithm is proposed. A criterion ensuring local convergence of the algorithm is presented. In addition, an estimate of the angular velocity and its rate of change is obtained. Simulation examples and real-world experiments indicate that a significantly higher estimation performance is achieved compared to [16].

*Outline*

In Section 2, the kinematic model of the rigid body is presented and used to formulate the attitude estimation in a maximum likelihood framework. Section 3 discusses the projection of an arbitrary matrix onto a particular non-convex set given by the structure of the angular and centripetal accelerations. An analytic solution to the projection is derived and is used in the later sections. An optimization algorithm for the resulting constrained least-squares problem is presented in Section 4 and is based on the augmented Lagrangian approach. A criterion for local convergence, and the generation of initial conditions for the solution algorithm is subsequently elaborated. The Fisher information matrix quantifying the information content of the accelerometer measurements is discussed in Section 5 and is used to optimize the accelerometer placements. Simulation results are presented in Section 6 and experimental results are provided in Section 7. The article concludes with

final remarks in Section 8.

## 2. Problem Formulation

In this section we formulate the problem of tilt estimation as a constrained least-squares problem. The derivation is similar to [16], except that the angular and centripetal acceleration terms are explicitly taken into account. In Section 2.1 the notation is briefly introduced, before deriving the kinematic model in Section 2.2 and discussing the assumption of a non-accelerated pivot in Section 2.3. The formulation of the tilt estimation in the maximum likelihood framework is presented in Section 2.4.

### 2.1 Notation

The representation of a tensor and vector in a particular coordinate frame is denoted by a preceding superscript, i.e. ${}^{\mathrm{B}}(A) = {}^{\mathrm{B}}A \in \mathbb{R}^{3\times 3}$, ${}^{\mathrm{B}}(v) = {}^{\mathrm{B}}v \in \mathbb{R}^3$. The body-fixed coordinate frame is denoted by $\{\mathrm{B}\}$. The rotation matrix $R_{\mathrm{IB}} \in SO(3)$ relates vectors from the body-fixed frame to their representation in the inertial frame $\{\mathrm{I}\}$, that is ${}^{\mathrm{I}}v = R_{\mathrm{IB}}{}^{\mathrm{B}}v$, for all vectors ${}^{\mathrm{B}}v \in \mathbb{R}^3$. The set $SO(3)$ denotes the special orthogonal group of rigid-body rotations. Moreover, the skew symmetric matrix corresponding to a vector $a \in \mathbb{R}^3$, denoted by $\tilde{a}$, is defined as $a \times b = \tilde{a}b$, for all $b \in \mathbb{R}^3$, where $a \times b$ refers to the cross product of the two vectors $a$ and $b$. The sphere of radius $g_0 := 9.81\,\mathrm{m/s^2}$ is denoted by $S^2$. The Frobenius scalar product of two matrices $A \in \mathbb{R}^{3\times 3}$ and $B \in \mathbb{R}^{3\times 3}$ is defined as

$$\langle A, B\rangle_F := \mathrm{tr}(B^{\mathsf{T}}A), \tag{2.1}$$

where $\mathrm{tr}(A)$ denotes the trace of the matrix $A$, that is, the sum of its diagonal elements. The induced norm (Frobenius norm) is then given by

$$||A||_F^2 := \langle A, A\rangle_F, \tag{2.2}$$

whereas the (induced) two norm is denoted by $||\cdot||_2$ and the (induced) maximum norm is denoted by $||\cdot||_\infty$.

Vectors are expressed as n-tuples $(x_1, x_2, \ldots, x_n)$ with dimension and stacking clear from context.

### 2.2 Kinematic Model

In this section we use a kinematic model to derive the maximum likelihood estimate of the tilt of the rigid body. Let ${}^{\mathrm{B}}p_i$ denote the position of the $i$'th accelerometer with respect to the pivot point represented in the body-fixed frame. We assume that there are $L$ sensors on the body, $i = 1, \ldots, L$. The variables are illustrated with the sketch shown in Fig. 2.1.

From the fact that the pivot point is not accelerated, it follows that (see [16])

$$^{\mathrm{I}}\ddot{p}_i = \ddot{R}_{\mathrm{IB}}\,{}^{\mathrm{B}}p_i \quad \text{and} \quad {}^{\mathrm{B}}\ddot{p}_i = R_{\mathrm{IB}}^{\mathsf{T}}\ddot{R}_{\mathrm{IB}}\,{}^{\mathrm{B}}p_i.$$

The kinematics yield additionally

$$\dot{R}_{\mathrm{IB}} = R_{\mathrm{IB}}\,{}^{\mathrm{B}}\widetilde{\omega}, \tag{2.3}$$

where ${}^{\mathrm{B}}\omega$ denotes the angular velocity of the body-fixed frame relative to the inertial frame, represented in the body-fixed frame. Taking the time derivative of (2.3) results in

$$\ddot{R}_{\mathrm{IB}} = \dot{R}_{\mathrm{IB}}\,{}^{\mathrm{B}}\widetilde{\omega} + R_{\mathrm{IB}}\,{}^{\mathrm{B}}\dot{\widetilde{\omega}} = R_{\mathrm{IB}}\left({}^{\mathrm{B}}\widetilde{\omega}^2 + {}^{\mathrm{B}}\dot{\widetilde{\omega}}\right). \tag{2.4}$$

Note that the matrix ${}^{\mathrm{B}}\dot{\widetilde{\omega}}$ is skew-symmetric, whereas ${}^{\mathrm{B}}\widetilde{\omega}^2$ is symmetric. Defining ${}^{\mathrm{B}}\Omega$ to be

$${}^{\mathrm{B}}\Omega := {}^{\mathrm{B}}\widetilde{\omega}^2 + {}^{\mathrm{B}}\dot{\widetilde{\omega}}, \tag{2.5}$$

leads to the following expression for the acceleration ${}^{\mathrm{B}}\ddot{p}_i$:

$${}^{\mathrm{B}}\ddot{p}_i = {}^{\mathrm{B}}\Omega\ {}^{\mathrm{B}}p_i. \tag{2.6}$$

The accelerometer measures the acceleration with respect to an observer in free fall, and therefore the acceleration measurement of the $i$'th sensor is given by

$$\begin{aligned} {}^{\mathrm{B}}m_i &= {}^{\mathrm{B}}\ddot{p}_i - {}^{\mathrm{B}}g + {}^{\mathrm{B}}n_i && (2.7)\\ &= {}^{\mathrm{B}}\Omega\ {}^{\mathrm{B}}p_i - {}^{\mathrm{B}}g + {}^{\mathrm{B}}n_i, && (2.8) \end{aligned}$$

where ${}^{\mathrm{B}}g \in S^2$ denotes the gravity vector in the body-fixed frame and ${}^{\mathrm{B}}n_i$ is the measurement noise. The measurement noise is assumed to be independent (for the different accelerometers), Gaussian, with zero mean and variance $\sigma_n^2 I$, where $I \in \mathbb{R}^{3\times 3}$ is the identity. We assume that the sensors are well-calibrated, such that the bias is negligible.

## 2.3 The Assumption of a Non-Accelerated Pivot

If the pivot point is not at rest, the acceleration of the $i$'th accelerometer is given by

$$^{\mathrm{I}}\ddot{p}_i = {}^{\mathrm{I}}\ddot{p}_0 + \ddot{R}_{\mathrm{IB}}\,{}^{\mathrm{B}}p_i, \tag{2.9}$$

where $\ddot{p}_0$ denotes the acceleration of the pivot point. Expressing the acceleration of the

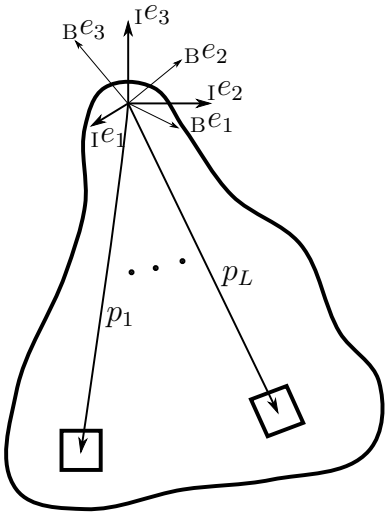


**Figure 2.1.** The sketch shows the inertial coordinate frame {I}, consisting of the vectors ${}_{\mathrm{I}}e_1$, ${}_{\mathrm{I}}e_2$, and ${}_{\mathrm{I}}e_3$, the body-fixed coordinate frame {B}, consisting of the vectors ${}_{\mathrm{B}}e_1$, ${}_{\mathrm{B}}e_2$, and ${}_{\mathrm{B}}e_3$, and the vectors $p_1, \ldots, p_L$. The pivot point is assumed to be at the origin of the coordinate frames.

$i$'th accelerometer in the body frame and inserting (2.6) yields

$$ {}^{\mathrm{B}}\ddot{p}_i = {}^{\mathrm{B}}\ddot{p}_0 + {}^{\mathrm{B}}\Omega\ {}^{\mathrm{B}}p_i, \tag{2.10} $$

and therefore

$$ {}^{\mathrm{B}}m_i = {}^{\mathrm{B}}\ddot{p}_0 + {}^{\mathrm{B}}\Omega\ {}^{\mathrm{B}}p_i - {}^{\mathrm{B}}g + {}^{\mathrm{B}}n_i \tag{2.11} $$

captures the measurement of the $i$'th accelerometer. As a result, given the accelerometer measurement ${}^{\mathrm{B}}m_i$, there is no possibility to distinguish between the measurement noise, the acceleration of the pivot, and the ${}^{\mathrm{B}}g$ vector capturing the tilt without further assumptions on the acceleration ${}^{\mathrm{B}}\ddot{p}_0$. Hence, if an extended (or unscented) Kalman filter based on rigid-body kinematics or a complementary filter is used, a potential, often implicit assumption is that ${}^{\mathrm{B}}\ddot{p}_i$ is white noise, e.g. [7]. As such, it may be combined with the measurement noise, and is nothing but an implicit non-accelerated pivot assumption. An alternative approach is to use rigid body kinetics, which provide a model for ${}^{\mathrm{B}}\ddot{p}_0$. However, as discussed in the introduction, such a model depends on the properties of the rigid body, e.g. the center of mass and the inertia, and is therefore susceptible to parameter errors. Moreover, such a model might not be available, as the forces and torques applied to the rigid body might not be known.

## 2.4 Maximum Likelihood Estimation

In the following, the superscript referring to the body-fixed coordinate frame {B} is omitted to simplify notation, e.g. ${}^{\mathrm{B}}g = g$, ${}^{\mathrm{B}}p_i = p_i$, etc.

The noise assumption yields the following combined likelihood for the $L$ sensors:

$$\begin{aligned} f(m_1, \ldots, m_L | g, \Omega) &= \prod_{i=1}^{L} \frac{1}{(2\pi\sigma_n^2)^{\frac{3}{2}}} \exp\left(-\frac{1}{2\sigma_n^2} ||m_i - \Omega\ p_i + g||_2^2\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{\frac{3L}{2}}} \exp\left(-\frac{1}{2\sigma_n^2} \sum_{i=1}^{L} ||m_i - \Omega\ p_i + g||_2^2\right). \end{aligned} \tag{2.12}$$

Additional gyroscope measurements, if available, could be included by extending the likelihood accordingly, provided that these measurements are assumed to be corrupted by uncorrelated Gaussian noise.

Note that $\Omega$ is given by (2.5) and contains the angular and centripetal acceleration. In particular, the symmetric part has the form $\widetilde{\omega}^2$ and therefore $\Omega \in \mathcal{M}$, where

$$\mathcal{M} := \{A \in \mathbb{R}^{3\times 3} | A + A^{\mathsf{T}} = \widetilde{a}^2, a \in \mathbb{R}^3\}. \tag{2.13}$$

Likewise, the gravity vector has length $g_0$ and is therefore an element of $S^2$.

The maximum likelihood estimate is thus obtained by maximizing $f(m_1, \ldots, m_L | g, \Omega)$ with respect to $g \in S^2$ and $\Omega \in \mathcal{M}$, which is equivalent to the following constrained least-squares problem

$$\min_{\Omega\in\mathcal{M}, g\in S^2} \sum_{i=1}^{L} ||m_i - \Omega\ p_i + g||_2^2. \tag{2.14}$$

In [16], the optimization was simplified to the following linear least-squares problem

$$\min_{\Omega\in\mathbb{R}^{3\times 3}, g\in\mathbb{R}^3} \sum_{i=1}^{L} ||m_i - \Omega\ p_i + g||_2^2. \tag{2.15}$$

Note that the linear least-squares problem has 12 unknowns, whereas the solutions of (2.14) are constrained to the manifolds $S^2$ and $\mathcal{M}$, which are 2 and 6-dimensional, respectively. Numerical examples, as presented in Sections 6 and 7, indicate that solving (2.14) instead of (2.15) improves the estimation performance significantly.

The optimization problem (2.14) is non-convex, since both $\mathcal{M}$ and $S^2$ are non-convex sets. Compared to (2.15), (2.14) is therefore computationally more demanding and there will be no guarantee that a global minimum is found. However, the projection (with respect to the Frobenius norm) of a matrix $A \in \mathbb{R}^{3\times 3}$ to $\mathcal{M}$ can be determined analytically. We will exploit this fact to derive a computationally tractable solution algorithm for (2.14) using the augmented Lagrangian approach. Subsequently, a criterion ensuring local convergence will be presented.

## 3. Projection to $\mathcal{M}$

In the following the projection

$$\text{prox}_{\mathcal{M}}(A) := \underset{A^* \in \mathcal{M}}{\text{argmin}} \, ||A - A^*||_F^2, \tag{2.16}$$

where $A$ is an arbitrary matrix in $\mathbb{R}^{3\times 3}$ is discussed. This projection will be used in the later sections to derive a computational efficient solution algorithm to the constrained least-squares problem (2.14).

The decomposition of $A - A^*$ into its symmetric (denoted $\text{symm}(\cdot)$) and skew-symmetric (denoted $\text{skew}(\cdot)$) parts, where

$$\text{symm}(A) := \frac{1}{2}(A + A^\mathsf{T}), \quad \text{skew}(A) := \frac{1}{2}(A - A^\mathsf{T}), \tag{2.17}$$

yields

$$\begin{aligned} ||A - A^*||_F^2 &= ||\text{symm}(A - A^*)||_F^2 + ||\text{skew}(A - A^*)||_F^2 \\ &\quad + 2\langle \text{symm}(A - A^*), \text{skew}(A - A^*) \rangle_F \\ &= ||\text{symm}(A - A^*)||_F^2 + ||\text{skew}(A - A^*)||_F^2, \end{aligned}$$

since symmetric and skew-symmetric matrices are orthogonal with respect to the Frobenius inner product. This decomposition implies that the minimizer of (2.16) reduces to

$$\underset{A^* \in \mathcal{M}}{\text{argmin}} \, ||A - A^*||_F^2 = \text{skew}(A) + \underset{A^* \in \mathcal{M}^s}{\text{argmin}} \, ||\text{symm}(A) - A^*||_F^2,$$

where the set $\mathcal{M}^s$ contains the symmetric elements of $\mathcal{M}$, i.e.

$$\mathcal{M}^s := \{A \in \mathcal{M} | A = A^\mathsf{T}\}. \tag{2.18}$$

Furthermore, the set $\mathcal{M}$ and the set $\mathcal{M}^s$ are invariant under rotations, that is

$$A^* \in \mathcal{M} \;\Leftrightarrow\; TA^*T^\mathsf{T} \in \mathcal{M}, \tag{2.19}$$

$$A^* \in \mathcal{M}^s \Leftrightarrow\; TA^*T^\mathsf{T} \in \mathcal{M}^s, \tag{2.20}$$

for all $T \in SO(3)$, which is due to the invariance of the cross product under rotations, or more specifically,

$$\widetilde{Ta} = T\tilde{a}T^\mathsf{T}, \tag{2.21}$$

for any vector $a \in \mathbb{R}^3$ and for all $T \in SO(3)$. The same is true for the Frobenius norm,

and therefore, the projection $\operatorname{argmin}_{A^* \in \mathcal{M}^s} ||\operatorname{symm}(A) - A^*||_F^2$ can be simplified to

$$\operatorname*{argmin}_{A^* \in \mathcal{M}^s} ||T^\mathsf{T} \operatorname{symm}(A) T - T^\mathsf{T} A^* T||_F^2 = \operatorname*{argmin}_{B^* \in \mathcal{M}^s} ||\Lambda - B^*||_F^2, \tag{2.22}$$

where $\Lambda$ is a diagonal matrix containing the eigenvalues of $\operatorname{symm}(A)$ and $T$ is the matrix containing the eigenvectors. An element $B^* \in \mathcal{M}^s$ can be parametrized by $\widetilde{a}^2$, for some $a = (a_1, a_2, a_3) \in \mathbb{R}^3$. Thus, writing the term $||\Lambda - B^*||_F^2$ out yields

$$\begin{aligned} ||\Lambda - B^*||_F^2 = \big[&(\lambda_1 + a_2^2 + a_3^2)^2 + a_1^2 a_2^2 + a_1^2 a_3^3 \\ &+ (\lambda_2 + a_1^2 + a_3^2)^2 + a_1^2 a_2^2 + a_2^2 a_3^2 \\ &+ (\lambda_3 + a_1^2 + a_2^2)^2 + a_1^2 a_3^2 + a_2^2 a_3^2 \big], \end{aligned}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$. This can be simplified further to a quadratic objective function by the change of variables $x = (a_1^2, a_2^2, a_3^2)$,

$$||\Lambda - B^*||_F^2 = 2x^\mathsf{T} \mathbf{1}\mathbf{1}^\mathsf{T} x + 2b^\mathsf{T} x + \lambda_1^2 + \lambda_2^2 + \lambda_3^2,$$

with

$$b = (\lambda_2 + \lambda_3,\ \lambda_1 + \lambda_3,\ \lambda_1 + \lambda_2),$$

and where $\mathbf{1} := (1, 1, 1)$. Evaluation of the KKT-conditions of $\operatorname{argmin}_{B^* \in \mathcal{M}^s} ||\Lambda - B^*||_F^2$ results in the following linear complementarity problem, [18, p. 4]:

$$\begin{aligned} 4x^\mathsf{T} \mathbf{1}\mathbf{1}^\mathsf{T} + 2b^\mathsf{T} &= \mu^\mathsf{T}, \\ \mu \geq 0, x \geq 0, \mu^\mathsf{T} x &= 0. \end{aligned}$$

These conditions can be interpreted in a geometric way, c.f. [18, p. 20]: The unit vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$, and the vector $-\mathbf{1} = (-1, -1, -1)$ partition $\mathbb{R}^3$ into four convex cones. A solution $x$ is found by decomposing $1/2\ b$ using the unit vectors spanning the cone that contains $1/2\ b$. The components in direction $e_1$, $e_2$, and $e_3$ yield the optimal multiplier $\mu$, whereas the component in direction $-\mathbf{1}$ represents the solution vector $x$, see Fig. 2.2.[8]

This implies that at most one component of $x$ is nonzero. Let $j$ denote the minimum

[8] We consider the non-degenerate case only. If $1/2\ b$ is colinear with $-\mathbf{1}$, then there are multiple solutions. Note however, that the set of vectors $1/2\ b$, which are colinear with $-\mathbf{1}$ is of (Lebesgue) measure zero.

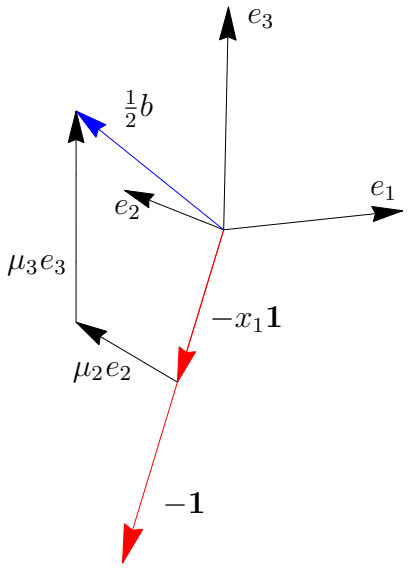


**Figure 2.2.** Decomposition of the vector $1/2\ b$ (blue) using the vectors $-\mathbf{1}$ (red), $e_2$ (black), and $e_3$ (black). In this case the solution is given by $x = (x_1, 0, 0)$, where $x_1$ is the component along $-\mathbf{1}$, with $\mu = (0, \mu_2, \mu_3)$, where $\mu_2$, $\mu_3$ are the components along $e_2$ and $e_3$, respectively.

element of $b$, i.e. $j = \text{argmin}_{i\in\{1,2,3\}} b_i$. If $b_j < 0$ it follows that

$$x = \begin{cases} x_1 = -\frac{1}{2}(\lambda_2 + \lambda_3), x_2 = x_3 = 0, & \text{for } j = 1 \\ x_2 = -\frac{1}{2}(\lambda_1 + \lambda_3), x_1 = x_3 = 0, & \text{for } j = 2 \\ x_3 = -\frac{1}{2}(\lambda_1 + \lambda_2), x_1 = x_2 = 0, & \text{for } j = 3. \end{cases}$$

In case $b_j \geq 0$, the vector $\frac{1}{2}b$ lies in the cone spanned by $e_1, e_2, e_3$ (the positive orthant) and therefore $x_1 = x_2 = x_3 = 0$.

Let $\lambda_1, \lambda_2, \lambda_3$ be the eigenvalues of $\text{symm}(A)$ with corresponding (normalized) eigenvectors $u_1, u_2, u_3$, such that $\lambda_1 \leq \lambda_2 \leq \lambda_3$. The solution of problem (2.16) is therefore given by

$$\begin{aligned} \underset{A^*\in\mathcal{M}}{\text{argmin}} ||A - A^*||_F^2 &= \text{skew}(A) + T^\mathsf{T} \begin{pmatrix} \frac{1}{2}(\lambda_1 + \lambda_2) & 0 & 0 \\ 0 & \frac{1}{2}(\lambda_1 + \lambda_2) & 0 \\ 0 & 0 & 0 \end{pmatrix} T \\ &= \text{skew}(A) - \frac{1}{2}(\lambda_1 + \lambda_2) T^\mathsf{T} \tilde{e}_3 \tilde{e}_3 T \\ &= \text{skew}(A) - \frac{1}{2}(\lambda_1 + \lambda_2) \tilde{u}_3 \tilde{u}_3, \end{aligned} \tag{2.23}$$

in case $\frac{1}{2}(\lambda_1 + \lambda_2) < 0$ and

$$\underset{A^* \in \mathcal{M}}{\operatorname{argmin}} \, ||A - A^*||_F^2 = \operatorname{skew}(A) \tag{2.24}$$

otherwise. Hence to compute the minimizer of (2.16) we only need to perform an eigendecomposition of $\operatorname{symm}(A)$, which can be performed very efficiently (for a symmetric $3 \times 3$ matrix there is an analytic solution, see [19]).

# 4. Proposed Solution Method

In the following, the augmented Lagrangian approach is used to derive a computationally efficient solution method to the constrained least-squares problem (2.14). Criteria ensuring local convergence are derived in Section 4.2 and the computation of initial conditions for the solution method is addressed in Section 4.3.

## 4.1 Augmented Lagrangian Approach

We use the so-called augmented Lagrangian approach, see e.g. [20, p. 515] to divide the optimization problem

$$\underset{\Omega \in \mathcal{M}, g \in S^2}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{L} ||m_i - \Omega \; p_i + g||_2^2 \tag{2.25}$$

into two subproblems. This provides a means to exploit the fact that one can easily project on $\mathcal{M}$ and on $S^2$ (see previous section). Compared to other solution methods (e.g. second-order methods), this approach leads to a simple solution algorithm which is straightforward to implement, even on embedded hardware. Numerical experiments indicate rapid convergence.

The optimization over $\mathcal{M}$ and $S^2$ is separated from the remaining optimization by adding additional artificial equality constraints. Thus, using the Lagrangian multipliers $\Lambda \in \mathbb{R}^{3\times3}$ and $\lambda \in \mathbb{R}^3$, the problem (2.25) is reformulated as

$$\min_{\substack{\Omega \in \mathbb{R}^{3\times3}, g \in \mathbb{R}^3, \\ A \in \mathcal{M}, a \in S^2}} \; \sup_{\substack{\Lambda \in \mathbb{R}^{3\times3}, \\ \lambda \in \mathbb{R}^3}} \mathcal{L}(\Omega, g, A, a, \Lambda, \lambda) \tag{2.26}$$

where the Lagrangian $\mathcal{L}$ is defined as

$$\begin{aligned} \mathcal{L}(\Omega, g, A, a, \Lambda, \lambda) := \frac{1}{2} \sum_{i=1}^{L} ||m_i - \Omega \; p_i + g||_2^2 + \lambda^{\mathsf{T}}(a - g) \\ + \langle \Lambda, A - \Omega \rangle_F + \frac{1}{2r}||A - \Omega||_F^2 + \frac{1}{2q}||a - g||_2^2, \end{aligned} \tag{2.27}$$

with $r$ and $q$ arbitrary positive scalars (they are used as tuning parameters in a later stage). Completing the squares yields

$$\begin{aligned}\mathcal{L}(\Omega, g, A, a, \Lambda, \lambda) = \frac{1}{2}\sum_{i=1}^{L} ||m_i - \Omega\ p_i + g||_2^2 - \frac{r}{2}||\Lambda||_F^2 \\ + \frac{1}{2r}||A - (\Omega - r\Lambda)||_F^2 + \frac{1}{2q}||a - (g - q\lambda)||_2^2 - \frac{q}{2}||\lambda||_2^2.\end{aligned} \tag{2.28}$$

Note that the term $\frac{1}{2r}||A-(\Omega-r\Lambda)||_F^2$ leads to a projection of $\Omega - r\Lambda$ on $\mathcal{M}$ and likewise the term $\frac{1}{2q}||a-(g-q\lambda)||_2^2$ to a projection of $g - q\lambda$ on $S^2$. Thus, equation (2.26) is stationary with respect to $A$ and $a$ if

$$A = \text{prox}_{\mathcal{M}}(\Omega - r\Lambda), \tag{2.29}$$

$$a = \text{prox}_{S^2}(g - q\lambda), \tag{2.30}$$

where

$$\text{prox}_{S^2}(g - q\lambda) := \underset{a^* \in S^2}{\text{argmin}}\, ||a^* - (g - q\lambda)||_2^2.$$

The stationary points of (2.26) with respect to $\Omega$, $g$, $\lambda$, and $\Lambda$ are given by

$$-\frac{1}{r}(A - \Omega + r\Lambda) - \sum_{i=1}^{L}(m_i - \Omega\ p_i + g)p_i^{\mathsf{T}} = 0, \tag{2.31}$$

$$-\frac{1}{q}(a - g + q\lambda) + \sum_{i=1}^{L}[m_i - \Omega\ p_i + g] = 0, \tag{2.32}$$

$$A - \Omega = 0, \tag{2.33}$$

$$a - g = 0. \tag{2.34}$$

Moreover, from (2.31) and (2.32) we obtain

$$\Omega - r\Lambda = A + r\sum_{i=1}^{L}(m_i - \Omega\ p_i + g)p_i^{\mathsf{T}}, \tag{2.35}$$

$$g - q\lambda = a - q\sum_{i=1}^{L}(m_i - \Omega\ p_i + g), \tag{2.36}$$

which can be combined with (2.29), (2.30), (2.33), and (2.34) to yield

$$\Omega = \text{prox}_{\mathcal{M}} \left( \Omega + r \sum_{i=1}^{L} (m_i - \Omega\ p_i + g) p_i^{\mathsf{T}} \right), \tag{2.37}$$

$$g = \text{prox}_{S^2} \left( g - q \sum_{i=1}^{L} (m_i - \Omega\ p_i + g) \right). \tag{2.38}$$

Equations (2.37) and (2.38) are necessary conditions for a minimum of (2.25). We propose to use fixed-point iteration to solve (2.37) and (2.38), that is

$$\Omega^{k+1} = \text{prox}_{\mathcal{M}} \left( \Omega^k + r \sum_{i=1}^{L} (m_i - \Omega^k\ p_i + g^k) p_i^{\mathsf{T}} \right), \tag{2.39}$$

$$g^{k+1} = \text{prox}_{S^2} \left( g^k - q \sum_{i=1}^{L} (m_i - \Omega^{k+1}\ p_i + g^k) \right). \tag{2.40}$$

### 4.2 Local Convergence

In the following section the convergence of the fixed-point iteration given by (2.39) and (2.40) will be elaborated. We will apply the Banach fixed-point theorem to provide conditions ensuring local convergence. The main result is summarized by Theorem 2. In addition, special cases leading to a simplification of the conditions for local convergence are discussed at the end of this section.

**Theorem 2.** *Assume that the matrix*

$$A_s := \begin{pmatrix} ||I - rP||_F & r||p||_2 \\ q||p||_2||I - rP||_F & |1 - qL| + qr||p||_2^2 \end{pmatrix},$$

*with*

$$p := \sum_{i=1}^{L} p_i, \quad P := \sum_{i=1}^{L} p_i p_i^{\mathsf{T}}, \quad 0 < q, r < \infty,$$

*has eigenvalues strictly within the unit circle, and let $g^* \in S^2$ and $\Omega^* \in \mathcal{M}$ be the fixed points of* (2.39)*, respectively* (2.40)*. Let $\Omega^* = (\widetilde{\omega}^*)^2 + \dot{\widetilde{\omega}}^*$ and assume that $\omega^* \neq 0$. Then, provided that $||g^0 - g^*||_2$, $||\Omega^0 - \Omega^*||_F$,*

$$q \Big|\Big| \sum_{i=1}^{L} m_i - \Omega^* p_i + g^* \Big|\Big|_2, \quad \textit{and} \quad r \Big|\Big| \sum_{i=1}^{L} (m_i - \Omega^* p_i + g^*) p_i^{\mathsf{T}} \Big|\Big|_F \tag{2.41}$$

*are small enough,* $||g^k - g^*||_2$ *and* $||\Omega^k - \Omega^*||_F$ *remain bounded for all* $k = 1, 2, \ldots,$ *and*

$$\lim_{k\to\infty} ||g^k - g^*||_2 \to 0, \quad \lim_{k\to\infty} ||\Omega^k - \Omega^*||_F \to 0.$$

*Proof.* We use the fact that $\text{prox}_{\mathcal{M}}$ and $\text{prox}_{S^2}$ are locally Lipschitz and apply the Banach fixed-point theorem. It is shown in the appendix (Prop. 4) that for every $\varepsilon > 0$ there exists a constant $\delta_{\mathcal{M}} > 0$, such that for all $\Omega_1, \Omega_2 \in \mathbb{R}^{3\times 3}$ with $||\Omega^* - \Omega_1||_F < \delta_{\mathcal{M}}, ||\Omega_1 - \Omega_2||_F < \delta_{\mathcal{M}}$ implies

$$||\text{prox}_{\mathcal{M}}(\Omega_1) - \text{prox}_{\mathcal{M}}(\Omega_2)||_F \leq (1+\varepsilon)||\Omega_1 - \Omega_2||_F. \tag{2.42}$$

In order for Prop. 4 to hold we must have $\omega^* \neq 0$, which is true by assumption. Likewise it is shown by Prop. 5 that for all $\varepsilon > 0$ there exists a $\delta_{S^2} > 0$ such that for all $g_1, g_2 \in \mathbb{R}^3$ with $||g^* - g_1|| < \delta_{S^2}$, $||g_1 - g_2||_2 < \delta_{S^2}$ implies

$$||\text{prox}_{S^2}(g_1) - \text{prox}_{S^2}(g_2)||_2 \leq (1+\varepsilon)||g_1 - g_2||_2. \tag{2.43}$$

By assumption $\Omega^*$ and $g^*$ fulfill (2.37), respectively (2.38). Combined with (2.39) and (2.40), it follows that

$$\begin{aligned}
\Omega^{k+1} - \Omega^* &= \text{prox}_{\mathcal{M}}\left(\Omega^k + r(\bar{m} - \Omega^k P + g^k p^\mathsf{T})\right) - \text{prox}_{\mathcal{M}}\left(\Omega^* + r(\bar{m} - \Omega^* P + g^* p^\mathsf{T})\right), \\
g^{k+1} - g^* &= \text{prox}_{S^2}\left(g^k - q(m - \Omega^{k+1} p + L g^k)\right) - \text{prox}_{S^2}\left(g^* - q(m - \Omega^* p + L g^*)\right),
\end{aligned}$$

where $m := \sum_{i=1}^L m_i$ and $\bar{m} := \sum_{i=1}^L m_i p_i^\mathsf{T}$. It holds therefore that $||\Omega^{k+1} - \Omega^*||_F$ can be bounded by

$$\begin{aligned}
(1+\varepsilon)||(\Omega^k - \Omega^*)(I - rP) + r(g^k - g^*)p^\mathsf{T}||_F& \\
&\leq (1+\varepsilon)\left(||I - rP||_F||\Omega^k - \Omega^*||_F + r||p||_2||g^k - g^*||_2\right),
\end{aligned}$$

provided that

$$(1+\varepsilon)\left(||I - rP||_F||\Omega^k - \Omega^*||_F + r||p||_2||g^k - g^*||_2\right) < \delta_{\mathcal{M}} \tag{2.44}$$

is fulfilled. Similarly, $||g^{k+1} - g^*||_2$ can be bounded by

$$\begin{aligned}
(1+\varepsilon)||(1-qL)(g^k - g^*) + q(\Omega^{k+1} - \Omega^*)p||_2& \\
&\leq (1+\varepsilon)\left(|1 - qL|||g^k - g^*||_2 + q||p||_2||\Omega^{k+1} - \Omega^*||_F\right) \\
&\leq (1+\varepsilon)\left(q||p||_2||I - rP||_F||\Omega^k - \Omega^*||_F + \left(|1 - qL| + rq||p||_2^2\right)||g^k - g^*||_2\right),
\end{aligned}$$

provided that (2.44) and

$$(1+\varepsilon)\left(q||p||_2||I-rP||_F||\Omega^k-\Omega^*||_F+\left(|1-qL|+rq||p||_2^2\right)||g^k-g^*||_2\right)<\delta_{S^2} \quad (2.45)$$

are fulfilled. The conditions can be simplified and written in compact form with $v^k := (||\Omega^k-\Omega^*||_F, ||g^k-g^*||_2)$, $\delta' := \min\{\delta_{\mathcal{M}}, \delta_{S^2}\}$, and $\hat{A}_s(\varepsilon)$ defined as

$$(1+\varepsilon)\begin{pmatrix} ||I-rP||_F & r||p||_2 \\ q||p||_2||I-rP||_F & |1-qL|+qr||p||_2^2 \end{pmatrix},$$

as

$$||\hat{A}_s(\varepsilon)v^k||_\infty < \delta' \quad \Rightarrow \quad v^{k+1} \leq \hat{A}_s(\varepsilon)v^k,$$

where $||\cdot||_\infty$ denotes the infinity-norm.

Note that $\hat{A}_s(0) = A_s$, $v^k \geq 0$, and by assumption, $A_s$ has eigenvalues strictly within the unit circle. The eigenvalues are continuous functions with respect to the matrix elements, [21, p. 26], and therefore, there exists an $\varepsilon' > 0$ such that the eigenvalues of $\hat{A}_s(\varepsilon)$ are strictly within the unit circle for all $\varepsilon < \varepsilon'$. Fixing $\varepsilon < \varepsilon'$ and using a spectral decomposition of $\hat{A}_s(\varepsilon)$ implies that the sequence $v^k$ can be bounded by

$$||v^k||_\infty \leq ||\hat{A}_s(\varepsilon)^k v^0||_\infty \leq ||T||_\infty||T^{-1}||_\infty||v^0||_\infty, \quad (2.46)$$

provided that $||T||_\infty||T^{-1}||_\infty||v^0||_\infty < \delta'$ holds, where $T$ is the matrix containing the eigenvectors of $\hat{A}_s(\varepsilon)$.[9] Therefore, given an $\varepsilon'' > 0$, we choose

$$||v^0||_\infty < \delta := \min\left\{\frac{\varepsilon''}{||T||_\infty||T^{-1}||_\infty}, \frac{\delta'}{||T||_\infty||T^{-1}||_\infty}\right\},$$

implying $||v^k||_\infty < \varepsilon''$ for all $k = 1, 2, \ldots$.

Moreover, the eigenvalues $\hat{A}_s(\varepsilon)$ are strictly within the unit circle ($\varepsilon$ was picked such that $\varepsilon < \varepsilon'$) and hence,

$$\lim_{k\to\infty} ||v^k||_\infty \leq \lim_{k\to\infty} ||\hat{A}_s(\varepsilon)^k v^0||_\infty = 0.$$

□

Given the problem data - the matrix $P$, the vector $p$, and the number of accelerometers $L$ - the parameters $q$ and $r$ can be chosen to minimize the absolute values of the eigenvalues of the matrix $A_s$. If eigenvalues with magnitudes less than 1 are obtained, local stability

[9] The Perron-Frobenius Theorem asserts that the matrix $\hat{A}_s(\varepsilon)$ is always diagonalizable, provided that all entries are strictly positive. In case some entries are zero, simple arguments ensure that the matrix $\hat{A}_s(\varepsilon)$ is still diagonalizable.

of the fixed-point iteration given by (2.39) and (2.40) is guaranteed, provided that the assumption $\omega^* \neq 0$ is fulfilled and that

$$q||\sum_{i=1}^{L} m_i - \Omega^* p_i + g^*||_2 \quad \text{and} \quad r||\sum_{i=1}^{L}(m_i - \Omega^* p_i + g^*)p_i^{\mathsf{T}}||_F$$

are small. Note that $\omega^* = 0$ is a set of Lebesgue measure zero and due to measurement noise it will occur with zero probability in practice. The other two requirements, (2.41), can be interpreted as bounds on the measurement noise. Simulation results indicate rapid convergence for realistic measurement noise, even when $\omega^* = 0$, see Section 6. The experimental results presented in Section 7 suggest that the fixed point iteration might even converge in case the matrix $A_s$ has eigenvalues with magnitude greater than one.

*Choosing the tuning variables q and r:* In case the tuning variable $q$ is chosen to be $1/L$, simple expressions for the eigenvalues of the matrix $A_s$ can be obtained. Note that the choice $q = 1/L$ might be suboptimal in the sense that the absolute values of the eigenvalues of $A_s$ might be further reduced if $q$ is chosen differently.

For $q = 1/L$ the eigenvalues of $A_s$ are given by

$$\lambda_1(A_s) = 0, \quad \lambda_2(A_s) = ||I - rP||_F + \frac{r}{L}||p||_2^2. \tag{2.47}$$

Minimizing $\lambda_2(A_s)$ with respect to $r$ yields the optimizer $r^*$,[10]

$$r^* := \begin{cases} \frac{\mathrm{tr}(P) - \sqrt{\mathrm{tr}(P)^2 - c||P||_F^2}}{||P||_F^2} & c \geq 0 \\ \frac{\mathrm{tr}(P) + \sqrt{\mathrm{tr}(P)^2 - c||P||_F^2}}{||P||_F^2} & c < 0, \end{cases} \qquad c := \frac{\mathrm{tr}(P)^2 - \frac{3||p||_2^4}{L^2}}{||P||_F^2 - \frac{||p||_2^4}{L^2}}. \tag{2.48}$$

Numerical experiments indicate that the choices $r = r^*$ and $q = 1/L$ provide reasonable initial guesses for the tuning parameters $r$ and $q$.

*Special Case p = 0:* In addition, further simplifications are obtained if the accelerometers are placed such that $p = 0$, i.e. the mean of the accelerometer positions relative to the fixed pivot is zero. In that case the matrix $A_s$ is diagonal. Furthermore its eigenvalues are minimal with $q = 1/L$ and $r = \mathrm{tr}(P)/||P||_F^2$, and are given by

$$\lambda_1(A_s) = 0, \quad \lambda_2(A_s) = \sqrt{3 - \frac{\mathrm{tr}(P)^2}{||P||_F^2}}. \tag{2.49}$$

According to Theorem 2, we must have $|\lambda_2(A_s)| < 1$ for local convergence, which is

[10]The additional condition $||P||_F > ||p||_2^2/L$ guaranteeing that $r^*$ is actually a minimizer is always fulfilled and can be verified using Jensen's inequality. Jensen's inequality implies further that $r^*$ is always real-valued.

equivalent to

$$\sqrt{2}||P||_F < \mathrm{tr}(P). \tag{2.50}$$

In case the accelerometers are placed such that $P$ has eigenvalues which are all equal, i.e. $P = \lambda(P)I$, it follows (directly from the definition of $A_s$) that choosing $r = 1/\lambda(P)$ results in $\lambda_2(A_s) = 0$. Hence, in that case, $\lambda_1(A_s) = \lambda_2(A_s) = 0$, and it can be concluded that the optimization routine given by (2.39) and (2.40) converges in one step, regardless of the initial condition (in the almost everywhere sense).

### 4.3 Generation of Initial Conditions

The algorithm presented in [16] is used to generate the initial conditions, i.e. $g^0$ and $\Omega^0$, for the fixed-point iteration given by (2.39) and (2.40). Let the solutions to the linear least-squares problem given by (2.15) be denoted by $(\hat{\Omega}, \hat{g})$. Then the initial conditions are obtained via the projections

$$\Omega^0 = \mathrm{prox}_{\mathcal{M}}(\hat{\Omega}), \quad g^0 = \mathrm{prox}_{S^2}(\hat{g}). \tag{2.51}$$

## 5. Information Content of the Accelerometer Data

In the following, the information content of the accelerometer measurements is quantified and analyzed using the Fisher information, [22, p. 196]. The analysis is motivated by the fact that in the high signal-to-noise ratio limit, the maximum likelihood estimate $(\Omega, g)$ reaches the Cramér-Rao bound, which is given by the inverse of the Fisher information matrix, [23]. We will determine the minimum number of accelerometers needed for the Fisher information matrix to have full rank. In addition, we will find optimal sensor placements by maximizing the determinant of the Fisher information matrix.

### 5.1 Derivation of the Fisher Information Matrix

We recall that the log-likelihood function for the $L$ sensors is proportional to

$$l(m_1, \ldots, m_L | g, \Omega) = -\sum_{i=1}^{L} ||m_i - \Omega p_i + g||_2^2. \tag{2.52}$$

We will express the Fisher information using a parametrization of $\mathcal{M}$ and $S^2$. More precisely, we choose the parameters $\omega \in \mathbb{R}^3, \dot{\omega} \in \mathbb{R}^3$, and $n \in \mathbb{R}^2$, with $\Omega = \widetilde{\omega}^2 + \dot{\widetilde{\omega}} \in \mathcal{M}$ and

$$g = \exp(\widetilde{\Pi n})\bar{g} \in S^2 \tag{2.53}$$

for a fixed $\bar{g} \in S^2$, where exp denotes the matrix exponential and $\Pi \in \mathbb{R}^{3\times 2}$ is a constant matrix containing two linearly independent vectors as columns, which are both orthogonal

to $\bar{g}$.[11] The derivative of the log-likelihood function with respect to the parameters $x := (\omega, \dot{\omega}, n)$ evaluated at $(\omega, \dot{\omega}, 0)$ is then found to be proportional to

$$\frac{\partial l}{\partial x} = -\sum_{i=1}^{L} (m_i - \Omega p_i + \bar{g})^{\mathsf{T}} \left(2\widetilde{\omega}\widetilde{p}_i, \widetilde{p}_i, -\widetilde{\bar{g}}\Pi\right), \tag{2.54}$$

which by virtue of (2.8) (with $g = \bar{g}$) reduces to

$$\frac{\partial l}{\partial x} = -\sum_{i=1}^{L} n_i^{\mathsf{T}} \underbrace{\left(2\widetilde{\omega}\widetilde{p}_i, \widetilde{p}_i, -\widetilde{\bar{g}}\Pi\right)}_{=:U_i}. \tag{2.55}$$

The evaluation at $(\omega, \dot{\omega}, 0)$ is without loss of generality, as the vector $\bar{g}$ can be chosen arbitrarily. As a result, the Fisher information evaluated at $(\omega, \dot{\omega}, 0)$ is proportional to

$$I(x) := \mathrm{E}\left[\frac{\partial l}{\partial x}^{\mathsf{T}} \frac{\partial l}{\partial x}\Big| g, \Omega\right] = \sum_{i=1}^{L} U_i^{\mathsf{T}} E[n_i n_i^{\mathsf{T}}] U_i = \sigma_n^2 \sum_{i=1}^{L} U_i^{\mathsf{T}} U_i. \tag{2.56}$$

### 5.2 Minimum Number of Accelerometers

In case of the (linear) least-squares solution given by (2.15), at least 4 accelerometers, which are not aligned on a plane are needed to obtain unique estimates. However, we will show that the Fisher information matrix is regular even in the case when 3 accelerometers are used. Due to the fact that the maximum likelihood estimator reaches the Cramér-Rao bound in the high signal-to-noise ratio limit, [23], this indicates that the proposed approach might be effective even for configurations with 3 accelerometers. This can be confirmed by the simulation shown in Section 6.

**Proposition 2.** *Provided that $\omega \neq 0$, the Fisher information matrix has full rank if and only if span$\{p_1, \ldots, p_L\} = \mathbb{R}^3$.*

*Proof.* The Fisher information was shown to be proportional to $I(x)$. We first note that $I(x)$ is symmetric and at least positive semi-definite. Moreover, $I(x)$ is singular if and only if a nonzero set of parameters $(v, h, n) \in \mathbb{R}^8$ exists such that $U_i(v, h, n) = 0$ for all $i = 1, 2, \ldots, L$, which is equivalent to

$$2\widetilde{\omega}\widetilde{p}_i v + \widetilde{p}_i h - \widetilde{\bar{g}}\Pi n = 0 \tag{2.57}$$

for all $i = 1, 2, \ldots, L$.

We first argue that in case the vectors $p_1, \ldots, p_L$ do not span $\mathbb{R}^3$, the matrix $I(x)$ is necessarily rank deficient. Indeed, if $p_1, \ldots, p_L$ do not span $\mathbb{R}^3$ there exists a nonzero

[11] In case the coordinate system is chosen such that $\bar{g} = (0, 0, -9.81\,\mathrm{m/s^2})$, the columns of $\Pi$ could be chosen as $(1, 0, 0)$ and $(0, 1, 0)$.

vector $v_0 \in \mathbb{R}^3$ that is orthogonal to $p_1, \ldots, p_L$, and hence, the set of equations (2.57) vanishes for $v = v_0$, $h = 0$, and $n = 0$, for example.

We now show that $I(x)$ has full rank provided that $p_1, \ldots, p_L$ spans $\mathbb{R}^3$. The set of equations (2.57) can be rewritten as

$$-(2\widetilde{\omega}\widetilde{v} + \widetilde{h})p_i - \widetilde{g}\Pi n = 0, \tag{2.58}$$

which can be used to conclude that

$$2\widetilde{\omega}\widetilde{v} + \widetilde{h} = 0 \tag{2.59}$$

must hold (by adding and subtracting (2.58) for different $i$ and using the fact that $p_1, \ldots, p_L$ span $\mathbb{R}^3$). Taking the trace of (2.59) reveals that $v$ must be orthogonal to $\omega$ in order for (2.59) to be fulfilled. Hence, the symmetric part of (2.59) reduces to

$$-4v^{\mathsf{T}}\omega I + 2(v\omega^{\mathsf{T}} + \omega v^{\mathsf{T}}) = 2(v\omega^{\mathsf{T}} + \omega v^{\mathsf{T}}) = 0,$$

which implies $v = 0$ (by assumption $\omega \neq 0$). From (2.59) and (2.58) it follows therefore that $h = 0$ and $n = 0$ must hold, which concludes the proof. □

### 5.3 Accelerometer Placement

The Fisher information can be used to optimize the accelerometer placement. The information content in the accelerometer data can be quantified, for example, using the trace (T-optimal), the determinant (D-optimal), or the trace of the inverse of the Fisher information (A-optimal), see e.g. [24]. We will focus here on the D-optimal design and treat all estimation variables, that is, $\dot{\omega}$, $\omega$, and $g$, on equal footing.

An alternative approach would be to focus on the variance (in the high signal-to-noise ratio limit) corresponding to the angular rate and the $g$ estimate only, or simply on the variance of the $g$ estimate only. This would naturally lead to an optimization of the determinant of the inverse Fisher information matrix, where only the parts corresponding to the angular rates and orientation, or the orientation alone, are considered. It turns out, however, that the results are similar; in case the optimization contains only the variance of the $g$ estimate, the optimum accelerometer configuration is one where the "center of mass" of the sensors (assuming all sensors have equal weight) lies at the pivot; in case the optimization is based on the combined angular rate and $g$ estimate, essentially the same result as in the derivation below is obtained.[12]

In order to proceed, we assume that the tilt $g$ is fixed (the choice of $g$ will be immaterial) and that the components of $\omega$ are mutually independent, zero mean, and have variance $\sigma_\omega^2$. The expected value (with respect to $\omega$) of the Fisher information matrix is

[12] In that case, (2.64) (see below) reduces to $\det(Lg_0^2\Theta - \widetilde{p}\widetilde{g}\Pi\,\Pi^{\mathsf{T}}\widetilde{g}\widetilde{p})$.

then proportional to

$$\sum_{i=1}^{L} \begin{pmatrix} 4\widetilde{p}_i \mathrm{E}[\widetilde{\omega}\widetilde{\omega}]\widetilde{p}_i & 2\widetilde{p}_i \mathrm{E}[\widetilde{\omega}]\widetilde{p}_i & -2\widetilde{p}_i \mathrm{E}[\widetilde{\omega}]\widetilde{g}\Pi \\ -2\widetilde{p}_i \mathrm{E}[\widetilde{\omega}]\widetilde{p}_i & -\widetilde{p}_i\widetilde{p}_i & \widetilde{p}_i\widetilde{g}\Pi \\ 2\Pi^{\mathsf{T}}\widetilde{g}\mathrm{E}[\widetilde{\omega}]\widetilde{p}_i & \Pi^{\mathsf{T}}\widetilde{g}\widetilde{p}_i & -\Pi^{\mathsf{T}}\widetilde{g}\widetilde{g}\Pi \end{pmatrix}. \tag{2.60}$$

Without loss of generality, we choose the coordinate system such that $g = (0, 0, -9.81\,\mathrm{m/s^2})$, and choose $(1, 0, 0)$ and $(0, 1, 0)$ as the column vectors of the matrix $\Pi$. The expression $\Pi^{\mathsf{T}}\widetilde{g}\widetilde{g}\Pi$ simplifies then to $-g_0^2 I_2$, where $I_2 \in \mathbb{R}^{2\times 2}$ is the identity matrix. From the fact that $E[\widetilde{\omega}] = 0$ and $E[\widetilde{\omega}\widetilde{\omega}] = -2\sigma_w^2 I$, we obtain

$$\mathrm{E}[I(x)] = \sum_{i=1}^{L} \begin{pmatrix} -8\widetilde{p}_i\widetilde{p}_i\sigma_w^2 & 0 & 0 \\ 0 & -\widetilde{p}_i\widetilde{p}_i & \widetilde{p}_i\widetilde{g}\Pi \\ 0 & \Pi^{\mathsf{T}}\widetilde{g}\widetilde{p}_i & g_0^2 I_2 \end{pmatrix}. \tag{2.61}$$

In order to simplify notation we introduce

$$\Theta := \sum_{i=1}^{L} -\widetilde{p}_i\widetilde{p}_i, \quad p := \sum_{i=1}^{L} p_i, \tag{2.62}$$

where $\Theta$ can be regarded as the inertia of the sensors (each sensor has unit mass) with respect to the pivot point. As a result, we obtain

$$\mathrm{E}[I(x)] = \begin{pmatrix} 8\sigma_w^2\Theta & 0 & 0 \\ 0 & \Theta & \widetilde{p}\widetilde{g}\Pi \\ 0 & \Pi^{\mathsf{T}}\widetilde{g}\widetilde{p} & Lg_0^2 I_2 \end{pmatrix}, \tag{2.63}$$

and by applying the Schur determinant identity, the determinant of (2.63) is therefore found to be proportional to

$$\det(\Theta)\det(Lg_0^2\Theta - \widetilde{p}\widetilde{g}\Pi\,\Pi^{\mathsf{T}}\widetilde{g}\widetilde{p}). \tag{2.64}$$

From the fact that the expression $\widetilde{g}\Pi\,\Pi^{\mathsf{T}}\widetilde{g}$ reduces to $-g_0^2\mathrm{diag}(1, 1, 0)$ it can be inferred that (2.64) is only dependent on the accelerometer placement, but not on the orientation $g$. In most applications there are specific requirements on the location of the sensors that must be taken into account. Given these requirements, one could then try to maximize (2.64) in order to determine suitable accelerometer placements. We emphasize that such an optimization optimizes a measure of the information content in the accelerometer measurements, which is independent of the estimation technique used for inferring the tilt. However, such an optimization tends to be application specific, due to the requirements on the sensor locations. We therefore conclude the section by discussing the simple case,

where the accelerometer placements are restricted to a unit ball around the pivot point. The main motivation lies not necessarily in the practical relevance of the problem, but rather in the fact that symmetric and geometrically appealing solutions are obtained.

**Proposition 3.** *Provided that accelerometer placements fulfilling*

$$\Theta = \frac{2L}{3} I, \quad p = 0, \tag{2.65}$$

*exist, then these conditions are necessary and sufficient for maximizing*

$$det(\Theta)\, det(L g_0^2 \Theta - \widetilde{p g} \Pi\, \Pi^{\mathsf{T}} \widetilde{g p}) \tag{2.66}$$

*subject to* $||p_i||_2^2 \leq 1$ *for all* $i = 1, 2, \ldots, L$.

*Proof.* From the fact that $\widetilde{p g} \Pi\, \Pi^{\mathsf{T}} \widetilde{g p}$ is positive semi-definite it follows that

$$\det(L g_0^2 \Theta - \widetilde{p g} \Pi\, \Pi^{\mathsf{T}} \widetilde{g p}) \leq \det(L g_0^2 \Theta), \tag{2.67}$$

where equality holds if and only if $p = 0$, see [25, p. 274]. Moreover, $\det(\Theta)$ is upper bounded by

$$\det(\Theta) \leq (1/3 \mathrm{tr}(\Theta))^3 = \left( 2/3 \sum_{i=1}^{L} ||p_i||_2^2 \right)^3 \leq (2L/3)^3,$$

which follows from the inequality of arithmetic and geometric means, and the constraint $||p_i||_2 \leq 1$. Moreover, the first inequality reduces to an equality if and only if $\Theta$ is proportional to the identity matrix, whereas the second inequality reduces to an equality if and only if $||p_i||_2 = 1$ for all $i = 1, 2, \ldots, L$. As a result, we therefore obtain that (2.66) is upper bounded by

$$(L g_0^2)^3 \; (2L/3)^6, \tag{2.68}$$

where the upper bound is attained if and only if the conditions (2.65) are met. Note that the condition $\Theta = 2L/3 \; I$ implies implicitly that $\mathrm{tr}(\Theta) = 2L = 2 \sum_{i=1}^{L} ||p_i||_2^2$, which is only fulfilled if $||p_i||_2 = 1$ for all $i = 1, 2, \ldots, L$. □

For most applications the optimality conditions given by Prop. 3 are too stringent, as particularly the condition $p = 0$ might be impossible to satisfy (see e.g. [2]). However, in applications where the rigid body is suspended in a gimbal (e.g. for inertially stabilized platforms), the "center of mass" of the sensors might be chosen to collide with the pivot resulting in $p = 0$. In that case, according to [26], $\Theta$ is diagonal if the placement of the accelerometers is such that at least two $n$-fold symmetry axes exist with $n \geq 3$ (or at least one $n$-fold symmetry axis ($n \geq 3$) and a non-orthogonal 2-fold symmetry axis). This

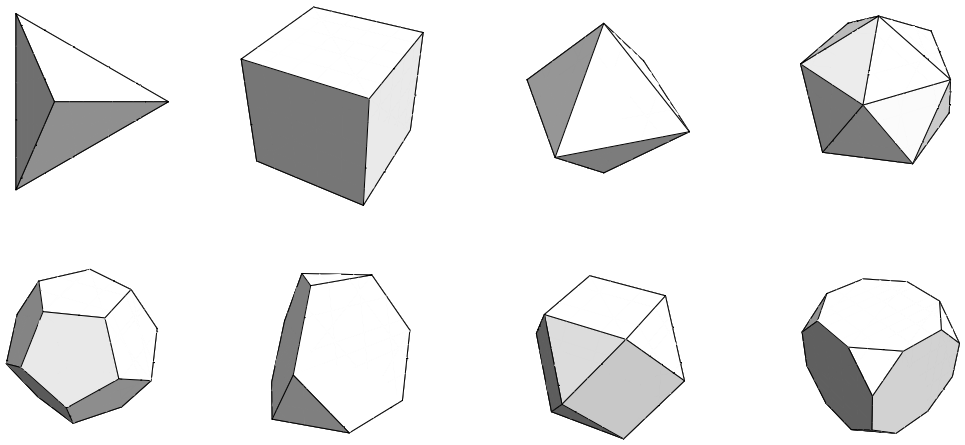

**Figure 2.3.** Depicted are the five Platonic solids and the first three Archimedean solids (from left to right: Tetrahedron, Cube, Octahedron, Icosahedron, Dodecahedron, truncated Tetrahedron, Cuboctahedron, truncated Cube). All these solids possess at least one of the point group symmetries $T$, $T_n$, $T_h$, $O$, $O_h$, $I$, $I_h$. Placing accelerometers at the vertices of any of these solids would therefore yield an optimal accelerometer configuration according to Prop. 3, provided that the pivot lies at the center.

implies that an optimal placement is achieved if the configuration has one of the following point group symmetries: $T$; $T_n$; $T_h$; $O$; $O_h$; $I$; $I_h$ (see e.g. [27] for an introduction to point group symmetries). Potential configurations therefore include accelerometer placements on the vertices of Platonic solids, the Archimedean solids or the Catalan solids, some of which are shown in Fig. 2.3.

## 6. Simulations

We consider a rigid body with four accelerometers, placed at

$$r_1 = (1,0,0)^\mathsf{T}, \quad r_2 = (0,1,0)^\mathsf{T},$$
$$r_3 = (0,0,1)^\mathsf{T}, \quad r_4 = \frac{1}{\sqrt{3}}(0.2,0.2,0.2)^\mathsf{T}.$$

The motion of the rigid body is generated by

$$\dot{\omega}(t) = q_k, \quad t \in [kT_s, (k+1)T_s),$$
$$\dot{R}_{\mathrm{IB}} = R_{\mathrm{IB}}\widetilde{\omega}, \qquad R_{\mathrm{IB}}(0) = I, \quad \omega(0) = 0,$$

with $q_k \sim \mathcal{N}(0, (10\ \mathrm{rad/s^2})^2 \cdot I)$, where $\mathcal{N}(0,\Sigma)$ denotes a multivariate Gaussian random

variable with zero mean and variance $\Sigma$. The differential equation $\dot{R}_{\mathrm{IB}} = R_{\mathrm{IB}}\widetilde{\omega}$ is integrated numerically with MATLAB's ode45 routine using a relative tolerance of $10^{-5}$. The noise on the accelerometer is assumed to be additive:

$$m_i(k) = \left(\widetilde{\omega}(kT_s)^2 + \dot{\widetilde{\omega}}(kT_s)\right)\ p_i - g(kT_s) + n_i(k), \tag{2.69}$$

$k = 0, 1, \ldots$, $i = 1, 2, 3, 4$, where $T_s = 20\mathrm{ms}$ denotes the sampling time, and $n_i(k)$ is independent Gaussian noise (independent across time and across the different sensors), with $n_i(k) \sim \mathcal{N}(0, (7.70 \cdot 10^{-3}\ \mathrm{m/s^2})^2 \cdot I)$.[13] A typical trajectory realization is depicted in Fig. 2.4.

The fixed-point iteration given by (2.39) and (2.40) is initialized with (2.51). We chose $q = 1/4$ and $r = 0.975$ according to (2.48) in Section 4.2. The resulting matrix $A_s$ as defined in Theorem 2 has eigenvalues located at $\lambda_1(A_s) = 0$, $\lambda_2(A_s) = 0.9479$, which according to Theorem 2 guarantees local convergence provided that the noise is small enough. As stopping criterion a relative tolerance of $10^{-8}$ was used, i.e. if

$$\frac{||g^k - g^{k-1}||_2}{||g^k||_2} \leq 10^{-8}, \quad \text{and} \quad \frac{||\Omega^k - \Omega^{k-1}||_F}{||\Omega^k||_F} \leq 10^{-8}$$

was met, the fixed-point iteration was stopped. For the trajectory shown in Fig. 2.4, an average of 28 iterations was needed for convergence (with a standard deviation of approximately 9 iterations), see Fig. 2.5.

Fig. 2.6 shows the two-norm of the estimation errors related to the gravity vector and the angular velocity. Note that the error of the angular velocity estimate decreases over time. This is due to the small angular velocities occurring initially. Since the angular velocity enters the estimation as $\widetilde{\omega}^2$, the sensitivity to the measurement noise is higher for small angular velocities. The error in the gravity vector is compared to the method proposed in [16]. On average, the estimation error of the gravity vector (two-norm) can be reduced by around 35%, for the trajectory depicted in Fig. 2.4. The estimation performance and the number of iterations needed is evaluated over a set of 50 randomized trajectories, see Tab. 1 and Tab. 2. On average, the tilt estimation error (two-norm) is reduced by approximately 45% compared to the approach presented in [16]. An average of 26 iterations with a standard-deviation of approximately 9 iterations is needed for convergence.

The influence of accelerometer-bias on the estimation performance (quantified by $\mathrm{E}[||\hat{g} - g||_2]$) is shown in Tab. 3. The estimation performance is again evaluated over a set of 50 randomized trajectories, where a different accelerometer-bias is introduced for each trajectory (uniformly sampled, for each sensor and each direction). The results indicate a linear correlation between the bias and the estimation performance. The re-

[13] The noise variance is taken from the specifications of the InvSense MPU6000 accelerometer, see `http://www.invensense.com`.

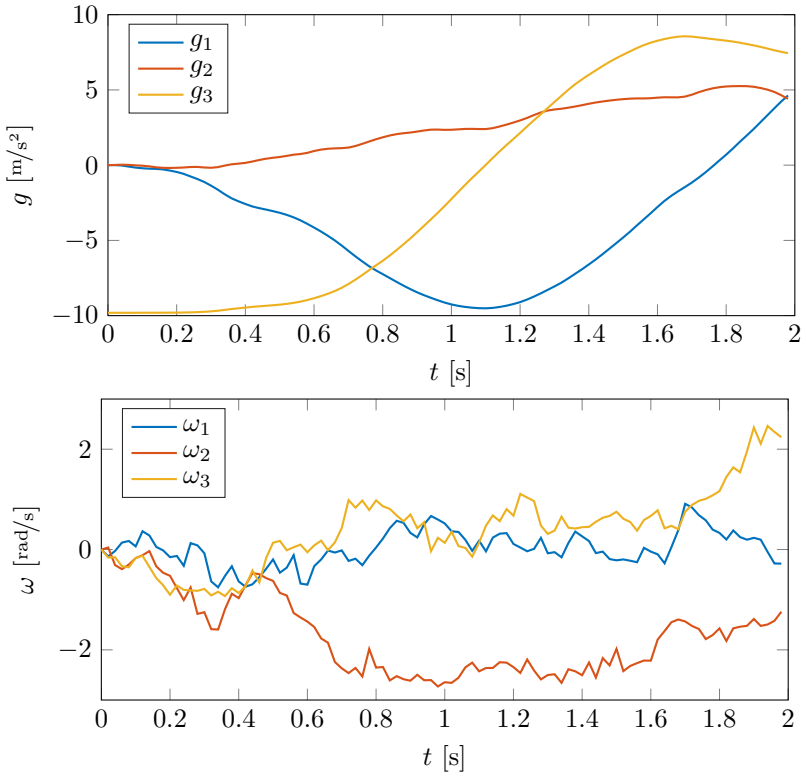


**Figure 2.4.** Trajectory of the gravity vector $g$ (top) and the angular velocity $\omega$ (bottom).

lative performance increase compared to the approach presented in [16] remains roughly constant.

We also study the influence of the magnitude of $\omega$ on the estimation performance. According to Thm. 2, local convergence of the algorithm can be guaranteed as long as $\omega^* \neq 0$. Thus, one might expect a deterioration of the estimation performance when $\omega$ approaches zero. Indeed, this can be confirmed numerically, as shown in Fig. 2.8, where we chose $g = (0, 0, -9.81\,\mathrm{m/s^2})$, $\dot{\omega} = 0$, $\omega = (\omega_x, 0, 0)$, and successively reduced $\omega_x$ from $1\,\mathrm{rad/s}$ to 0. In contrast, the performance of the tilt estimate and the estimate of the rate of change of the angular velocities are not affected by the magnitude of $\omega$.

In addition, the estimation is evaluated for a configuration with three accelerometers located at $(0.05, 0, 0)$, $(0, 0.05, 0)$, $(0, 0, 0.05)$, and the trajectory given by Fig. 2.4. The optimization is solved with $q = 1/3$ and $r = 400$, which is obtained from (2.48) in Section 4.2. The estimation performance decreases significantly when using only three accelerometers, as can be seen in Fig. 2.7. In particular, if one component of the angular velocity is close to zero, the estimation error increases by up to two magnitudes. Still, the algorithm given by (2.39) and (2.40) converges, yielding a unique gravity vector estimate and a unique angular velocity estimate (possibly a local minimum). This is in contrast to the method presented in [16] where a configuration with three accelerometers would not provide unique estimates.

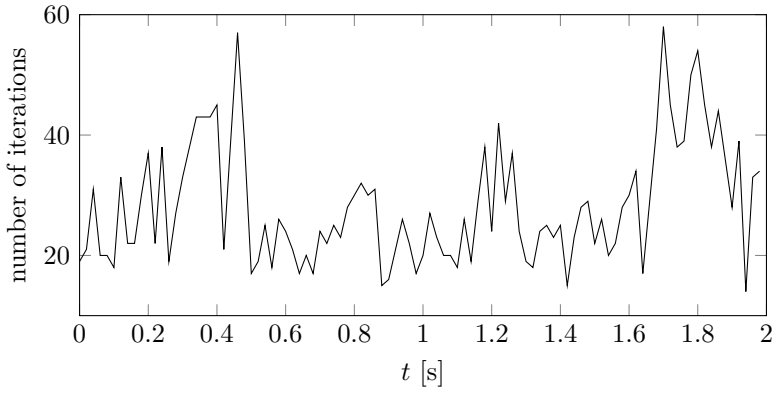


**Figure 2.5.** Number of fixed-point iterations needed to reach convergence for the trajectory depicted in Fig. 2.4.

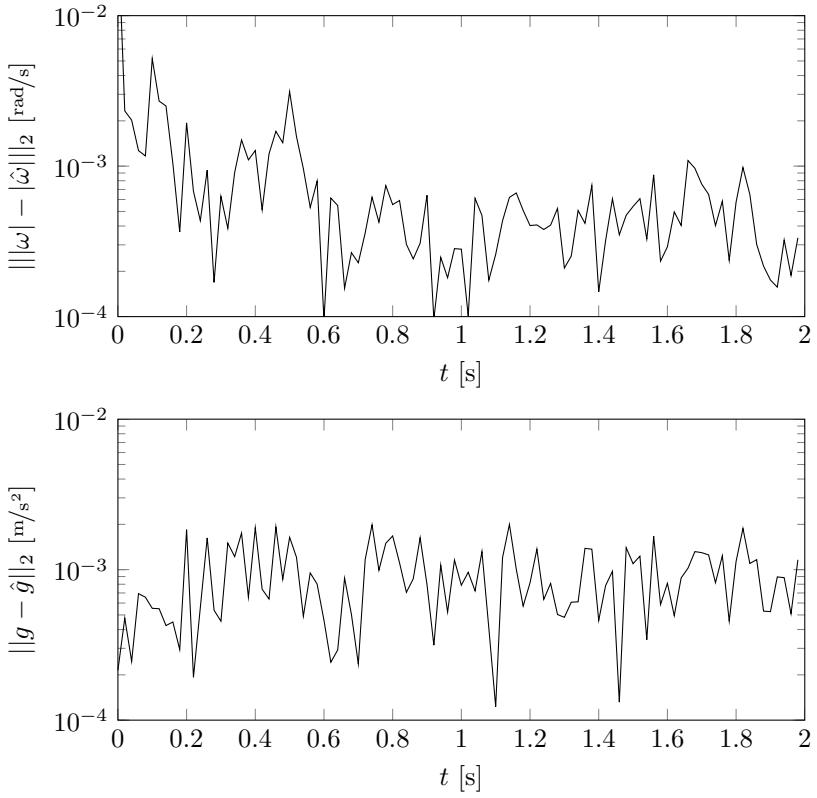


**Figure 2.6.** Estimation performance for the trajectory depicted in Fig. 2.4. The two-norm of the angular velocity error (absolute value, since we can only infer $\widetilde{\omega}^2$) is depicted on top. The two-norm of the error related to the gravity estimate is depicted on the bottom.

# 7. Experimental Results

We conducted experiments on a real-world testbed (we used a balancing robot, see e.g. [2]), where four inertial measurement units (of the type InvSense MPU6000), each con-

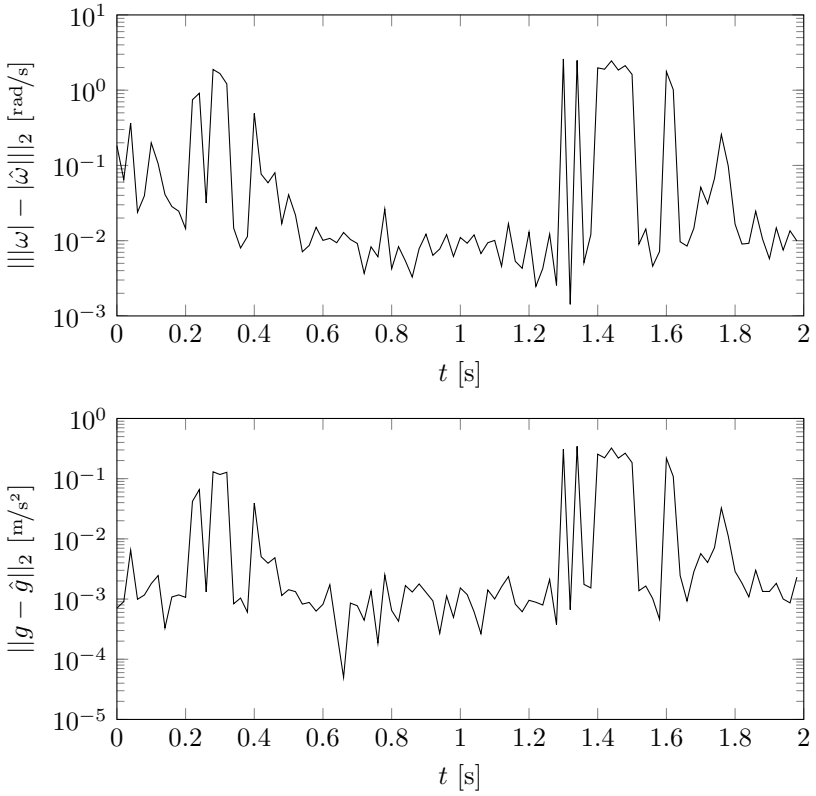


**Figure 2.7.** Estimation performance for the trajectory depicted in Fig. 2.4 for the configuration with three accelerometers. The two-norm of the angular velocity error (absolute value, since we can only infer $\widetilde{\omega}^2$) is depicted on top. The two-norm of the error related to the gravity estimate is depicted on the bottom.

**Table 1.** Evaluation of the tilt estimation for 50 randomized trajectories and comparison to the approach presented in [16] (denoted previous approach below).

| | $\mathrm{E}[\|\|\hat{g} - g\|\|_2]$ | $\mathrm{Var}[\|\|\hat{g} - g\|\|_2]$ |
|---|---|---|
| Approach herein | $0.8 \cdot 10^{-3}$ | $2.1 \cdot 10^{-7}$ |
| Previous approach | $1.5 \cdot 10^{-3}$ | $6.1 \cdot 10^{-7}$ |
| Reduction | 45% | 65% |

sisting of an accelerometer, gyroscope, and magnetometer, were placed at (in m)

$$r_1 = (0.122, 0.013, 0.145)^\mathsf{T}, \quad r_2 = (0.145, 0.122, 0.013)^\mathsf{T},$$
$$r_3 = (0.028, 0.013, 0.005)^\mathsf{T}, \quad r_4 = (0.005, 0.122, 0.013)^\mathsf{T}.$$

The ground-truth data for the tilt was obtained using a motion capture system, whereas the ground-truth data for the angular velocity was collected using the gyroscopes, which were part of the inertial measurement units. The ground-truth rate of change of the angular velocity was obtained by first-order finite differences with a step size of 20 ms. The

**Table 2.** Evaluation of the estimation errors and the number of iterations required for convergence for 50 randomized trajectories.

| | $\mathrm{E}[\|\|\hat{\omega}\| - \|\omega\|\|_2]$ | $\mathrm{E}[\|\|\dot{\hat{\omega}} - \dot{\omega}\|\|_2]$ | av. iter. |
|---|---|---|---|
| App. herein | $0.9 \cdot 10^{-3}$ | $1.1 \cdot 10^{-3}$ | 26 |

**Table 3.** Evaluation of the tilt estimation for 50 randomized trajectories and comparison to the approach presented in [16] (denoted previous approach below) with increasing accelerometer-bias. For each trajectory the accelerometer-bias is uniformly sampled and the estimation performance is quantified by $\mathrm{E}[\|\hat{g} - g\|_2]$.

| | $\mathrm{Unif}([-0.1, 0.1])$ | $\mathrm{Unif}([-0.5, 0.5])$ | $\mathrm{Unif}([-1, 1])$ |
|---|---|---|---|
| App. herein | 0.065 | 0.323 | 0.643 |
| Prev. app. | 0.116 | 0.582 | 1.165 |

accelerometers were calibrated prior to the experiments in order to account for the bias and the tuning parameters $q$ and $r$ were chosen to approximately minimize the eigenvalues of the matrix $A_s$, as defined in Theorem 2. This resulted in $q = 0.1$ and $r = 11.5$. The same stopping criterion as in the previous section was used. The resulting trajectories are depicted in Fig. 2.9. For space reasons only the first component of each vector is plotted. A good fit of the gravity vector and the rate of change of the angular velocity can be observed. The angular velocity enters the optimization as $\tilde{\omega}^2$, which inherently decreases its estimation accuracy. In addition, the quality of the estimate is further decreased due to uncertainties in the accelerometer placements. This was confirmed in experiments, where the angular velocity estimate was found to be relatively noisy.

Fig. 2.10 compares the error in the gravity estimate resulting from the algorithm presented here to the method proposed in [16], and indicates a significant performance increase. The corresponding numerical values are shown in Tab. 4. The angular velocity error, the error of the rate of change of the angular velocity, and the number of iterations

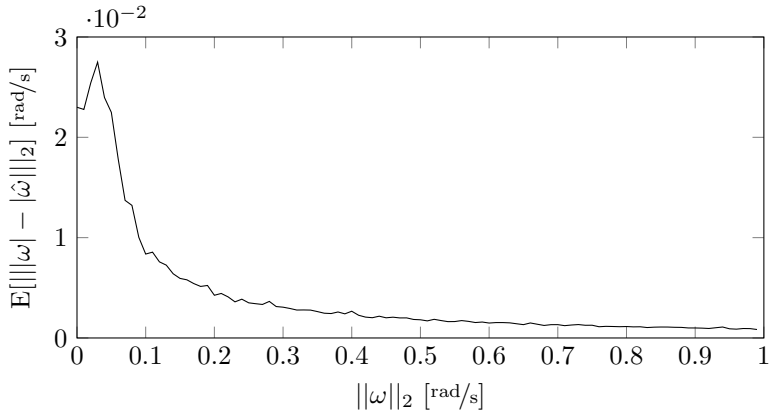


**Figure 2.8.** Influence of the magnitude of $\omega$ on the performance of the angular rates estimate.

**Table 4.** Evaluation of the tilt estimation on real-world measurements and comparison to the approach presented in [16] (denoted previous approach below).

| | $\mathrm{E}[\|\|\hat{g} - g\|\|_2]$ | $\mathrm{Var}[\|\|\hat{g} - g\|\|_2]$ |
|---|---|---|
| Approach herein | $1.1 \cdot 10^{-1}$ | $1.6 \cdot 10^{-3}$ |
| Previous approach | $1.7 \cdot 10^{-1}$ | $4.1 \cdot 10^{-3}$ |
| Reduction | 37% | 60% |

**Table 5.** Evaluation of the estimation errors on real-world measurements and the number of iterations required for convergence.

| | $\mathrm{E}[\|\|\hat{\omega} - \omega\|\|_2]$ | $\mathrm{E}[\|\|\dot{\hat{\omega}} - \dot{\omega}\|\|_2]$ | av. iter. |
|---|---|---|---|
| App. herein | $7.8 \cdot 10^{-1}$ | 1.1 | 209 |

needed for convergence are presented in Tab. 5. The time history of the error (two-norm) of the rate of change of the angular velocity is shown in Fig. 2.11. Compared to the simulation example, more iterations are needed in the real-world experiment. This can be explained by the accelerometer placement resulting in a matrix $A_s$ with eigenvalues of larger magnitude. In fact, the matrix $A_s$ has eigenvalues of magnitude greater than one, and therefore the requirements for guaranteeing local convergence according to Theorem 2 are not fulfilled. Nevertheless, the estimates were found to converge in practice, which highlights the robustness of the algorithm. Note that the accelerometer configuration used in the experiments was fixed and not specifically adjusted to our needs.

Compared to the simulations in Section 6, the overall performance of the algorithm is inferior in the real-world experiment although the noise characteristics in the simulation were chosen to match the experiment. We conjecture that this is partly related to the accelerometer placement, but also due to uncertainties in the accelerometer positions and calibration biases.

## 8. Conclusion

A method for accelerometer-based state determination of a rigid body with a single pivot is presented. The pivot is assumed to be at rest or to move uniformly. The proposed approach extends the method from [16] by accounting for the angular and centripetal acceleration, which allows estimation of the gravity vector and the angular velocity, as well as the rate of change of the angular velocity. Simulation results indicate that the method works reliably (convergence after few iterations) and outperforms [16] in terms of precision. This was confirmed in real-world experiments.

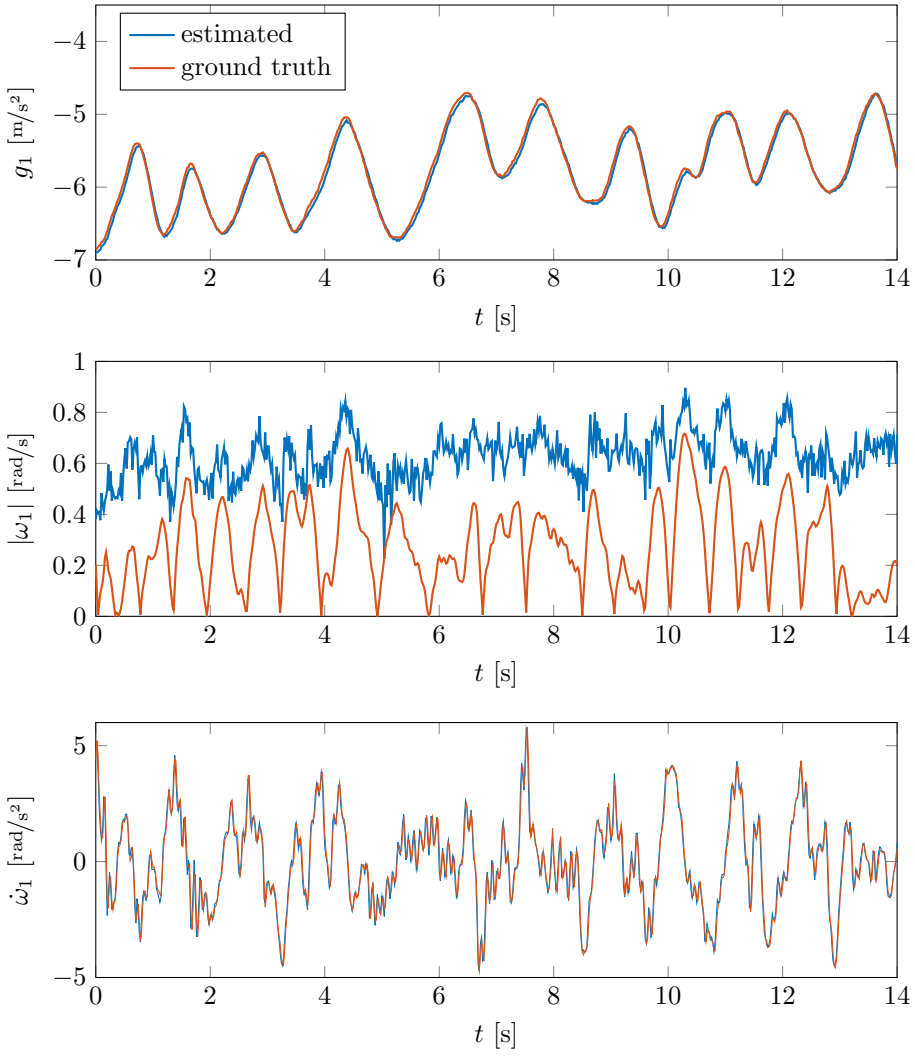


**Figure 2.9.** Depicted is the first component of the gravity vector $g$ (top), the absolute value of the first component of the angular velocity $\omega$ (middle), and the first component of rate of change of the angular velocity $\dot{\omega}$ (bottom). The estimated quantities are shown in blue, the ground-truth data is shown in red.

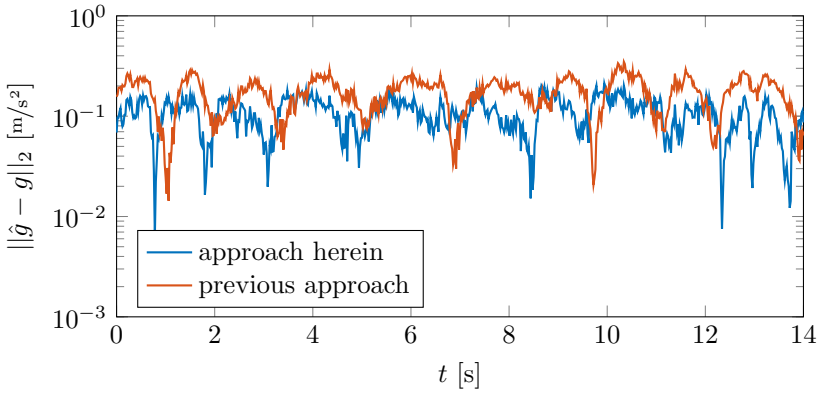


**Figure 2.10.** Estimation performance on real measurements. Depicted is the two-norm of the estimation error of the gravity vector.

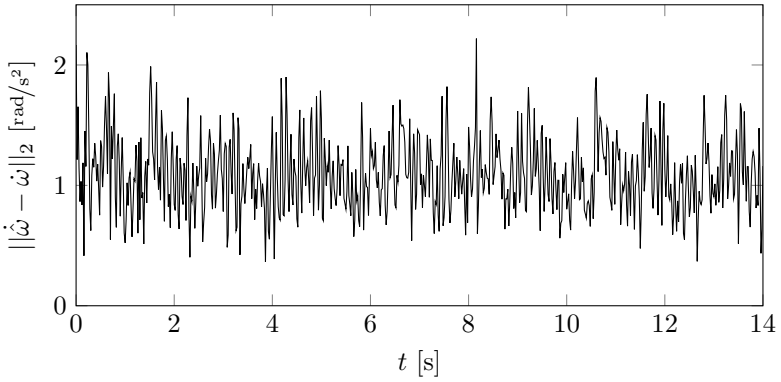


**Figure 2.11.** Estimation performance on real measurements. Depicted is the two-norm of the estimation error of the rate of change of the angular velocity.

## A. Bound on the Lipschitz constant of $\mathrm{prox}_{\mathcal{M}}$

A bound on the Lipschitz constant of $\mathrm{prox}_{\mathcal{M}}$, denoted by $L_{\mathcal{M}}$, is computed in four steps. The Einstein summation convention and index notation will be used in the following.

We first note that for a diagonal matrix $A_0 := \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^{3\times 3}$ with $\lambda_1 \leq \lambda_2 < \lambda_3$, $\lambda_1 + \lambda_2 < 0$, the partial derivatives of the projection on $\mathcal{M}$ can be calculated analytically and are given by

$$
\begin{aligned}
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{11}}\bigg|_{A_0} &= \frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{22}}\bigg|_{A_0} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{12}}\bigg|_{A_0} &= \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{21}}\bigg|_{A_0} &= \frac{1}{2}\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{13}}\bigg|_{A_0} &= \frac{1}{4}\frac{\lambda_1+\lambda_2}{\lambda_3-\lambda_1}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{23}}\bigg|_{A_0} &= \frac{1}{4}\frac{\lambda_1+\lambda_2}{\lambda_3-\lambda_2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{31}}\bigg|_{A_0} &= \frac{1}{4}\frac{\lambda_1+\lambda_2}{\lambda_3-\lambda_1}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{32}}\bigg|_{A_0} &= \frac{1}{4}\frac{\lambda_1+\lambda_2}{\lambda_3-\lambda_2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \\
\frac{\partial \mathrm{prox}_{\mathcal{M}}}{\partial A_{33}}\bigg|_{A_0} &= 0.
\end{aligned}
\tag{2.70}
$$

We show that

*Lemma* 3. Let $A \in \mathbb{R}^{3\times3}$, $B \in \mathbb{R}^{3\times3}$, $T \in SO(3)$, and $A_{ij}$ denote the $ij$'th entry of the matrix $A$. Then it holds that

$$\left\| \left.\frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\right|_A B_{ij} \right\|_F = \left\| \left.\frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\right|_{TAT^\mathsf{T}} (TBT^\mathsf{T})_{ij} \right\|_F,$$

provided that the partial derivative of $\text{prox}_{\mathcal{M}}$ evaluated for $A$ is well defined.

*Proof.* The projection onto the set $\mathcal{M}$ is rotationally invariant. Therefore it holds that for any matrix $A \in \mathbb{R}^{3\times3}$,

$$\text{prox}_{\mathcal{M}}(A) = T^\mathsf{T}\text{prox}_{\mathcal{M}}(TAT^\mathsf{T})T, \tag{2.71}$$

or equivalently (using index notation and the Einstein summation convention),

$$\text{prox}_{\mathcal{M}jl}(A) = T_{ij}\text{prox}_{\mathcal{M}ik}(TAT^\mathsf{T})T_{kl}. \tag{2.72}$$

Taking the derivative with respect to the element $A_{mn}$ yields

$$\left.\frac{\partial \text{prox}_{\mathcal{M}jl}}{\partial A_{mn}}\right|_A = T_{ij} \left.\frac{\partial \text{prox}_{\mathcal{M}ik}}{\partial A_{op}}\right|_{TAT^\mathsf{T}} \frac{\partial (TAT^\mathsf{T})_{op}}{\partial A_{mn}} T_{kl}. \tag{2.73}$$

The squared Frobenius norm is the sum of all squared entries of a matrix, and therefore

$$\left\| \left.\frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{mn}}\right|_A B_{mn} \right\|_F^2 = \left.\frac{\partial \text{prox}_{\mathcal{M}jl}}{\partial A_{mn}}\right|_A B_{mn} \left.\frac{\partial \text{prox}_{\mathcal{M}jl}}{\partial A_{qr}}\right|_A B_{qr}. \tag{2.74}$$

From (2.73) it follows that

$$\begin{aligned}
\left\| \left.\frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{mn}}\right|_A B_{mn} \right\|_F^2 &= \left.\frac{\partial \text{prox}_{\mathcal{M}ik}}{\partial A_{op}}\right|_{TAT^\mathsf{T}} \left.\frac{\partial \text{prox}_{\mathcal{M}st}}{\partial A_{uv}}\right|_{TAT^\mathsf{T}} T_{ij}T_{kl}T_{om}T_{pn}B_{mn}T_{sj}T_{tl}T_{uq}T_{vr}B_{qr} \\
&= \left.\frac{\partial \text{prox}_{\mathcal{M}ik}}{\partial A_{op}}\right|_{TAT^\mathsf{T}} \left.\frac{\partial \text{prox}_{\mathcal{M}ik}}{\partial A_{uv}}\right|_{TAT^\mathsf{T}} T_{om}B_{mn}T_{pn}T_{uq}B_{qr}T_{vr} \\
&= \left\| \left.\frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{uv}}\right|_{TAT^\mathsf{T}} (TBT^\mathsf{T})_{uv} \right\|_F^2,
\end{aligned}$$

where $T^\mathsf{T}T = I$ has been used. □

Next, an upper bound to the difference $\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)$ is derived. More precisely,

*Lemma* 4. Let $A_1, A_2 \in \mathbb{R}^{3\times3}$ be symmetric matrices such that $A_1t + A_2(1-t)$ has eigenvalues $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$ with $\lambda_1(t) + \lambda_2(t) < 0$ for all $t \in [0,1]$. Then it holds

that $\|\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)\|_F$ can be bounded by

$$\int_0^1 \left\| \frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\bigg|_{B_1(t)+(1-t)B_2(t)} (B_1(t) - B_2(t))_{ij} \right\|_F \mathrm{d}t,$$

where $B_1(t)$, $B_2(t)$ are such that

$$B_1(t) := T(t)A_1T(t)^\mathsf{T}, \quad B_2(t) := T(t)A_2T(t)^\mathsf{T},$$

for some $T(t) \in SO(3)$, $t \in [0, 1]$.

*Proof.* The matrix $A_1t + A_2(1-t)$ is symmetric and has eigenvalues $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$ with $\lambda_1(t)+\lambda_2(t) < 0$ for all $t \in [0, 1]$ and therefore the derivative of $\text{prox}_{\mathcal{M}}(A_1t+(1-t)A_2)$ with respect to $t$ is well defined, see also Lemma 3. Therefore

$$\begin{aligned} ||\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)||_F &= \Big|\Big| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\text{prox}_{\mathcal{M}}(A_1t + (1-t)A_2)\mathrm{d}t \Big|\Big|_F \\ &\leq \int_0^1 \Big|\Big| \frac{\mathrm{d}}{\mathrm{d}t}\text{prox}_{\mathcal{M}}(A_1t + (1-t)A_2) \Big|\Big|_F \mathrm{d}t \\ &= \int_0^1 \Big|\Big| \frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\bigg|_{A_1t+(1-t)A_2} (A_1 - A_2)_{ij} \Big|\Big|_F \mathrm{d}t. \end{aligned}$$

The result follows by invoking Lemma 3. □

*Lemma* 5. For symmetric matrices $A_1$, $A_2$, such that $A_1t + (1-t)A_2$ has eigenvalues $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$ with $\lambda_1(t) + \lambda_2(t) < 0$ for all $t \in [0, 1]$, it holds that

$$\|\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)\|_F \leq L_1||A_1 - A_2||_F, \tag{2.75}$$

with

$$L_1 := \max_{t\in[0,1]} \left(1, \frac{\sqrt{2}}{4}\frac{|\lambda_1(t) + \lambda_2(t)|}{\lambda_3(t) - \lambda_2(t)}\right). \tag{2.76}$$

*Proof.* From Lemma 4 it follows that $||\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)||_F$ can be bounded by

$$\int_0^1 \left\| \frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\bigg|_{B_1(t)+(1-t)B_2(t)} (B_1(t) - B_2(t))_{ij} \right\|_F \mathrm{d}t, \tag{2.77}$$

where $B_1(t)$, $B_2(t)$ are such that

$$B_1(t) := T(t)A_1T(t)^\mathsf{T}, \quad B_2(t) := T(t)A_2T(t)^\mathsf{T},$$

for $T(t) \in SO(3)$. The matrices $A_1$ and $A_2$ are symmetric and therefore $T(t)$ can be chosen such that for all $t \in [0,1]$, $B_1(t) + (1-t)B_2(t)$ is diagonal with real eigenvalues $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$, $\lambda_1(t) + \lambda_2(t) < 0$. By inserting (2.70) and accounting for the fact that $B_1(t)$ and $B_2(t)$ are symmetric, it follows that

$$\begin{aligned}\Big|\Big| \frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\Big|_{B_1(t)+(1-t)B_2(t)} (B_1(t) - B_2(t))_{ij}\Big|\Big|_F^2 &= \frac{1}{2}(\bar{B}_{11}(t) + \bar{B}_{22})^2 \\ + \left(\frac{1}{4}\frac{\lambda_1(t)+\lambda_2(t)}{\lambda_3(t)-\lambda_1(t)}(\bar{B}_{13}(t) + \bar{B}_{31}(t))\right)^2 &+ \left(\frac{1}{4}\frac{\lambda_1(t)+\lambda_2(t)}{\lambda_3(t)-\lambda_2(t)}(\bar{B}_{23}(t) + \bar{B}_{32}(t))\right)^2,\end{aligned} \tag{2.78}$$

where $\bar{B}(t) := B_1(t) - B_2(t)$. Using Jensen's inequality results in

$$\begin{aligned}&\Big|\Big| \frac{\partial \text{prox}_{\mathcal{M}}}{\partial A_{ij}}\Big|_{B_1(t)+(1-t)B_2(t)} (B_1(t) - B_2(t))_{ij}\Big|\Big|_F^2 \\ &\quad\leq \max\left\{1, 2\left(\frac{1}{4}\frac{\lambda_1(t)+\lambda_2(t)}{\lambda_3(t)-\lambda_1(t)}\right)^2, 2\left(\frac{1}{4}\frac{\lambda_1(t)+\lambda_2(t)}{\lambda_3(t)-\lambda_2(t)}\right)^2\right\}\left\|\bar{B}(t)\right\|_F^2, \\ &\quad= \max\left\{1, 2\left(\frac{1}{4}\frac{\lambda_1(t)+\lambda_2(t)}{\lambda_3(t)-\lambda_2(t)}\right)^2\right\}\left\|\bar{B}(t)\right\|_F^2.\end{aligned} \tag{2.79}$$

The Frobenius norm is invariant under rotations and therefore $||\bar{B}(t)||_F^2 = ||A_1 - A_2||_F^2$ for all $t \in [0,1]$. Thus, this yields

$$\begin{aligned}||\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)||_F &\leq \int_0^1 \max\left\{1, \sqrt{2}\left(\frac{1}{4}\frac{|\lambda_1(t)+\lambda_2(t)|}{\lambda_3(t)-\lambda_2(t)}\right)\right\} \mathrm{d}t||A_1 - A_2||_F \\ &\leq \sup_{t\in[0,1]}\left\{1, \sqrt{2}\left(\frac{1}{4}\frac{|\lambda_1(t)+\lambda_2(t)|}{\lambda_3(t)-\lambda_2(t)}\right)\right\}||A_1 - A_2||_F.\end{aligned}$$

The supremum is bounded since $\lambda_2(t) < \lambda_3(t)$ for all $t \in [0,1]$, and is attained since $[0,1]$ is closed. □

*Lemma* 6. For arbitrary matrices $A_1$, $A_2$, such that

$$\text{symm}(A_1)t + (1-t)\text{symm}(A_2)$$

has eigenvalues $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$ with $\lambda_1(t) + \lambda_2(t) < 0$ for all $t \in [0, 1]$, it holds that

$$\|\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)\|_F \leq L_{\mathcal{M}} \|A_1 - A_2\|_F , \tag{2.80}$$

with

$$L_{\mathcal{M}} := \max_{t\in[0,1]} \left\{ 1, \frac{\sqrt{2}}{4} \frac{|\lambda_1(t) + \lambda_2(t)|}{\lambda_3(t) - \lambda_2(t)} \right\}. \tag{2.81}$$

*Proof.* From the analytic solution of $\text{prox}_{\mathcal{M}}$, Lemma 5, and the fact that $L_{\mathcal{M}} \geq 1$, it can be inferred that

$$\begin{aligned} ||\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)||_F &= ||\text{skew}(A_1) - \text{skew}(A_2) + \\ &\qquad \text{prox}_{\mathcal{M}}(\text{symm}(A_1)) - \text{prox}_{\mathcal{M}}(\text{symm}(A_2))||_F \\ &\leq L_{\mathcal{M}}||\text{symm}(A_1) - \text{symm}(A_2)||_F + ||\text{skew}(A_1) - \text{skew}(A_2)||_F \\ &\leq L_{\mathcal{M}} \left(||\text{symm}(A_1 - A_2)||_F + ||\text{skew}(A_1 - A_2)||_F\right) \\ &\leq L_{\mathcal{M}}||A_1 - A_2||_F. \end{aligned}$$

□

From the previous result it follows that

**Proposition 4.** *Let $A^* \in \mathcal{M}$ be such that* $\text{symm}(A^*)$ *has eigenvalues $\lambda_1 = \lambda_2 < \lambda_3 = 0$. Then, for any $\varepsilon > 0$ there exists a $\delta > 0$, such that for all $A_1, A_2 \in \mathbb{R}^{3\times3}$, with $||A_1 - A^*||_F < \delta$, $||A_1 - A_2||_F < \delta$ implies*

$$||\text{prox}_{\mathcal{M}}(A_1) - \text{prox}_{\mathcal{M}}(A_2)||_F \leq (1 + \varepsilon)||A_1 - A_2||_F.$$

*Proof.* Note that $||A_1 - A_2||_F < \delta$ and $||A_1 - A^*||_F < \delta$ implies $||A_2 - A^*||_F < 2\delta$. The eigenvalues are continuous functions of the matrix elements, [21, p. 26], and therefore there exists a $\delta > 0$ such that for all $||A_1 - A_2||_F < \delta$ it holds that the eigenvalues of

$$\text{symm}(A_1)(t) + (1 - t)\text{symm}(A_2)$$

fulfill $\lambda_1(t) \leq \lambda_2(t) < \lambda_3(t)$ and $\lambda_1(t) + \lambda_2(t) < 0$ for all $t \in [0, 1]$, and

$$L_{\mathcal{M}} = \max_{t\in[0,1]} \left\{ 1, \frac{\sqrt{2}}{4} \frac{|\lambda_1(t) + \lambda_2(t)|}{\lambda_3(t) - \lambda_2(t)} \right\} \leq 1 + \varepsilon.$$

Note that for $\delta \to 0$ it follows that $A_1 \to A^*$, $A_2 \to A^*$ and therefore $L_{\mathcal{M}} \to 1$. Therefore by Lemma 6 the result follows. □

## B. Bound on the Lipschitz constant of $\text{prox}_{S^2}$

Similar to the derivation of the Lipschitz constant of $\text{prox}_{\mathcal{M}}$, the rotational invariance of the two norm can be exploited to derive bounds on the Lipschitz constant of $\text{prox}_{S^2}$. Without loss of generality, we consider the case where $x_0 = (0, 0, r)$, $r > 0$. It can be shown that

$$\left.\frac{\partial \text{prox}_{S^2}}{\partial x_1}\right|_{x_0} = \begin{pmatrix} \frac{g_0}{r} \\ 0 \\ 0 \end{pmatrix}, \quad \left.\frac{\partial \text{prox}_{S^2}}{\partial x_2}\right|_{x_0} = \begin{pmatrix} 0 \\ \frac{g_0}{r} \\ 0 \end{pmatrix}, \quad \left.\frac{\partial \text{prox}_{S^2}}{\partial x_3}\right|_{x_0} = 0. \tag{2.82}$$

Using a similar argument to the proof of Lemma 4 it follows that $||\text{prox}_{S^2}(g_1) - \text{prox}_{S^2}(g_2)||_2$ can be bounded by

$$\int_0^1 \left|\left| \left.\frac{\partial \text{prox}_{S^2}}{\partial x_i}\right|_{g_1 t + (1-t)g_2} (g_1 - g_2)_i \right|\right|_2 \mathrm{d}t \leq \frac{g_0}{\min_{t\in[0,1]} ||g_1 t + (1-t)g_2||_2} ||g_1 - g_2||_2. \tag{2.83}$$

This leads to

**Proposition 5.** *Let the vector $a^*$ be such that $a^* \in S^2$. Then, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $a_1, a_2 \in \mathbb{R}^3$ with $||a_1 - a^*||_2 < \delta$, $||a_1 - a_2||_2 < \delta$ implies that*

$$||\text{prox}_{S^2}(a_1) - \text{prox}_{S^2}(a_2)||_2 \leq (1 + \varepsilon)||a_1 - a_2||_2.$$

*Proof.* We have that $\min_{t\in[0,1]} ||a_1 t + (1-t)a_2||_2 > g_0 - 2\delta$, since $||a_1 - a^*||_2 < \delta$ and $||a_2 - a^*||_2 < 2\delta$. Therefore

$$||\text{prox}_{S^2}(a_1) - \text{prox}_{S^2}(a_2)||_2 \leq \frac{g_0}{g_0 - 2\delta} ||a_1 - a_2||_2. \tag{2.84}$$

Choosing $\delta = g_0\varepsilon/(2(1+\varepsilon))$ implies $g_0/(g_0 - 2\delta) = (1 + \varepsilon)$ and the result follows. □

## References

[1] S. Trimpe and R. D'Andrea, "The Balancing Cube", *IEEE Control Systems Magazine*, pp. 48–75, 2012.

[2] M. Muehlebach and R. D'Andrea, "Nonlinear analysis and control of a recation-wheel-based 3-D inverted pendulum", *IEEE Transactions on Control Systems Technology*, vol. 25, no. 1, pp. 235–246, 2017.

[3] J. M. Hilkert, "Inertially stabilized platform technology", *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 26–46, 2008.

[4] J. E. Kain and C. Yates, "Airborne imaging system using global positioning system (GPS) and inertial measurement unit (IMU) data", US5894323 A, 1999.

[5] M. K. Masten, "Inertially stabilized platforms for optical imaging systems", *IEEE Control Systems Magazine*, vol. 28, no. 1, pp. 47–64, 2008.

[6] R. G. Brown, *Introduction to Random Signal Analysis and Kalman Filtering*. John Wiley & Sons, 1983.

[7] R. Mahony, T. Hamel, and J.-M. Pflimlin, "Nonlinear complementary filters on the special orthogonal group", *IEEE Transactions on Automatic Control*, vol. 53, no. 5, pp. 1203–1218, 2008.

[8] H. Rehbinder and X. Hu, "Drift-free attitude estimation for accelerated rigid bodies", *Automatica*, vol. 40, no. 4, pp. 653–659, 2004.

[9] H. F. Grip, T. I. Fossen, T. A. Johansen, and A. Saberi, "Atttitude estimation using biased gyro and vector measurements with time-varying reference vectors", *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1332–1338, 2012.

[10] M.-D. Hua, K. Rudin, G. Ducard, T. Hamel, and R. Mahony, "Nonlinear attitude estimation with measurement decoupling and anti-windup gyro-bias compensation", *Proceedings of the 18th IFAC World Congress*, pp. 2972–2978, 2011.

[11] B. Barshan and H. F. Durrant-Whyte, "Inertial navigation systems for mobile robots", *IEEE Transactions on Robotics and Automation*, vol. 11, no. 3, pp. 328–342, 1995.

[12] E. J. Lefferts, F. L. Markley, and M. D. Shuster, "Kalman filtering for spacecraft attitude estimation", *Journal of Guidance, Control, and Dynamics*, vol. 5, no. 5, pp. 417–429, 1982.

[13] R. C. Leishman, J. C. Macdonald Jr., R. W. Beard, and T. W. Mclain, "Quadrotors and accelerometers", *IEEE Control Systems Magazine*, vol. 34, no. 1, pp. 28–41, 2014.

[14] J. L. Crassidis and F. L. Markley, "Unscented filtering for spacecraft attitude estimation", *Journal of Guidance, Control, and Dynamics*, vol. 26, no. 4, pp. 536–542, 2003.

[15] S. Bonnabel and J.-J. Slotine, "A contraction theory-based analysis of the stability of the deterministic extended Kalman filter", *IEEE Transactions on Automatic Control*, vol. 60, no. 2, pp. 565–569, 2015.

[16] S. Trimpe and R. D'Andrea, "Accelerometer-based tilt estimation of a rigid body with only rotational degrees of freedom", *International Conference on Robotics and Automation*, pp. 2630–2636, 2010.

[17] M. Gajamohan, M. Muehlebach, T. Widmer, and R. D'Andrea, "The Cubli: A reaction wheel based 3D inverted pendulum", *European Control Conference*, 2013.

[18] R. W. Cottle, J.-S. Pang, and R. E. Stone, *The Linear Complementarity Problem*. SIAM, 1992.

[19] J. Kopp, "Efficient numerical diagonalization of Hermitian $3 \times 3$ matrices", *International Journal of Modern Physics C*, vol. 19, no. 3, pp. 523–548, 2008.

[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd. Springer, 2006.

[21] E. E. Tyrtyshnikov, *A brief introduction to numerical analysis*. Springer Science & Business Media, 1997.

[22] K. P. Murphy, *Machine Learning - A Probabilistic Perspective*. MIT Press, 2012.

[23] M. Vallisneri, "Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects", *Physical Review D*, vol. 77, no. 4, 2008.

[24] F. E. Udwadia, "Methodology for optimum sensor locations for parameter identification in dynamic systems", *Journal of Engineering Mechanics*, vol. 120, no. 2, pp. 368–390, 1994.

[25] D. S. Bernstein, *Matrix Mathematics*. Princeton University Press, 2005.

[26] P. Aravind, "A comment on the moment of inertia of symmetrical solids", *American Journal of Physics*, vol. 60, no. 8, pp. 754–755, 1992.

[27] F. A. Cotton, *Chemical Applications of Group Theory*, 3rd. John Wiley & Sons, 1990.

# Part B

## A flying vehicle actuated by ducted fans

# Paper P3

# The Flying Platform - A testbed for ducted fan actuation and control design

Michael Muehlebach and Raffaello D'Andrea

**Abstract**

This article discusses the design of an unmanned aerial vehicle whose purpose is to study the use of electric ducted fans as control and propulsion system. Thrust vectoring is essential for stabilizing the vehicle. We present measurement results characterizing the thrust vectoring capabilities of the propulsion system (both statically and dynamically), discuss a first-principle model describing the behavior of the flying machine, and analyze and quantify the controllability about hover. The first-principle model is subsequently used for a cascaded control design, which is shown to work reliably in practice. Furthermore, system identification results are discussed and used to extend the model. The resulting augmented model is shown to match the measured frequency response function.

Published in *Mechatronics*.

## 1. Introduction

The design and control of unmanned aerial vehicles has been an active field of research in the past years, not least because of the numerous applications including surveillance, data acquisition, aerial photography, construction, transportation, and entertainment. Often, flying vehicles combining efficient forward flight, high maneuverability, and vertical take-off and landing capabilities are highly desirable. This article aims therefore at studying the properties of electric ducted fans as control and propulsion system for flying machines, where size is limited, but high static thrusts are required. This includes, for example, tailsitters, hovercrafts or even actuated wingsuit flight, [1].

To that extent, the Flying Platform, a flying vehicle actuated by three electric ducted fans is introduced, see Fig. 3.1.[14] In addition to their aerodynamic efficiency, [2, p. 322], resulting in high thrusts at moderate sizes, ducted fans have the advantage that the moving parts are shielded, protecting the propeller blades from undesired contacts with the environment. Moreover, the high exit velocities can be exploited for thrust vectoring. Thus, each ducted fan of the Flying Platform is augmented with an exit nozzle and control flaps to direct the airflow. The thrust vectoring is essential for stabilizing the vehicle.

The article includes experimental results characterizing the static maps from flap angles to thrusts, as well as the transfer functions from fan and servo commands to thrusts, thereby quantifying the available actuation bandwidth. For control and analysis purposes a low-complexity model is introduced. The mechanical design of the Flying Platform is optimized for maximum control authority; a closed-form expression for the determinant of the controllability Gramian is derived, providing a means to quantify and optimize the controllability of the vehicle by trading off the total inertia with the lever arm of the thrust vectoring system. The low-complexity model is used to derive a cascaded control law, stabilizing the vehicle about hover. The parameters of the control law are related to time constants of the closed-loop dynamics, which enables an intuitive tuning. The controller is shown to work reliably in flight experiments. A frequency domain system identification is presented, showing the limitations of the low-complexity model at frequencies below 1 Hz. We extend the model by including gyroscopic and aerodynamic effects, such as momentum drag (due to the redirection of the airflow by the ducted fans) yielding an augmented model that roughly matches the measured frequency response function.

*Related work:* Previous work, see e.g. [3], [4], [5], [6], focused on aspects related to the modeling, the design and the control laws of flying vehicles with a single duct. The authors of [3] present a controller based on dynamic inversion of a low-complexity model in combination with a neural network for capturing the unmodeled dynamics. The control design is shown to work reliably in real world experiments. In [4], nonlinear control techniques are applied for simultaneous force and position tracking by a ducted-fan vehicle. The authors emphasize the unstable zero dynamics of the open-loop system, which is

[14] A video showing the Flying Platform can be found under `https://www.youtube.com/watch?v=NYY9q-vs4Nw`.

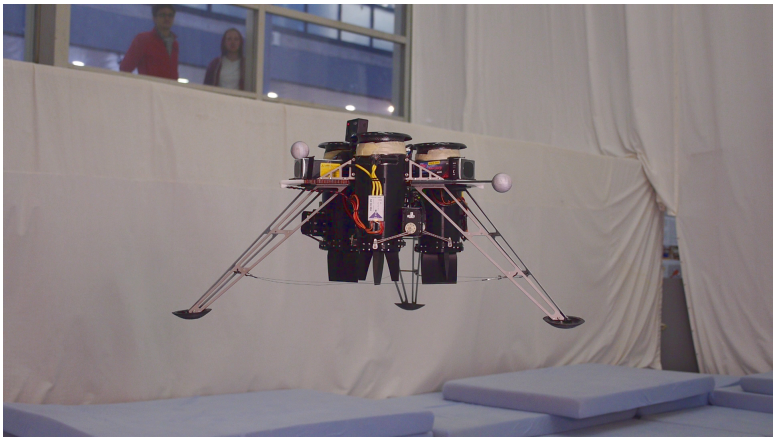

**Figure 3.1.** The Flying Platform hovering in the Flying Machine Arena.

attributed to the fact that the thrust vectoring acts below the center of gravity, see also [7]. Other nonlinear control approaches include a sliding mode controller, [8], and nonlinear receding horizon control accounting for actuator saturations in [9]. In contrast, the authors in [10] present a linear cascaded control design and a linear estimator design for a ducted-fan vehicle with two counter-rotating rotors. The authors emphasize the benefits of the cascaded control design with regards to a practical implementation. In [5] and [6], the effects of crosswinds on the aerodynamics of a ducted fan vehicle are discussed. It is pointed out that the redirection of crosswinds by the propeller and the duct results in a drag force, linearly dependent on the forward velocity of the flying vehicle. This force induces a pitching moment on the center of gravity leading to an unstable open-loop system. This effect is further investigated in [11] by means of computational fluid dynamics and wind tunnel testing (see also [12] for further experimental results). The authors of [13] use a planar particle image velocimeter system to investigate the velocity profile in ducted fans. Both experimental data and computational predictions based on the Navier-Stokes equation are shown to agree at hover, as well as for horizontal movements. The results confirm that a horizontal movement redirects, respectively distorts the incoming airflow.

In [14] and [15], 4 ducted fans are assembled in a quadrotor configuration and the resulting flight performance is analyzed. Thereby two ducted fans are counter rotating for stabilizing yaw. In a more recent work, the authors of [16] compare and implement several extensions to a standard quadrotor configuration: 1) a quadrotor that can tilt its rotors, 2) a quadrotor that is extended with two ducted fans, both of which can vector the thrust, 3) four ducted fans aligned in a quadrotor configuration, all of which can vector the thrust. Thrust vectoring is achieved by moving the whole exit nozzle. The designs are motivated by the fact that these vehicles can perform position set point changes or compensate cross winds without requiring the vehicle to tilt. In all the designs, thrust vectoring is enhancing the maneuverability of the vehicle, but is not crucial for stability.

Compared to the ducted fan vehicles presented in the literature, the Flying Platform

is significantly different. Instead of a relatively large shroud, covering a single or two counter-rotating propellers with rotary speeds of roughly 10000 rpm, see e.g. [3], three electric ducted fans, each of which can vector the thrust, are used to actuate the Flying Platform. Thurst vectoring is essential for stabilizing the vehicle. The electric ducted fans have a diameter of 90 mm, which is small compared to the overall dimension of the Flying Platform (about 1 m). Nevertheless, they provide a total thrust of roughly 45 N each, at around 30000 rpm, and achieve exit velocities up to 90 m/s.[15] The high exit velocities enable efficient thrust vectoring; compared to the single-duct vehicle presented in [3],[16] the thrust generated by the control surfaces of the Flying Platform is roughly 5-10 times larger. Moreover, the high rotation speeds of the ducted fans lead to gyroscopic torques, which are quantified by analyzing real flight data.

The characterization and modeling of flying vehicles with system identification techniques has a long history, [17]. In the past years, various models for different types of unmanned aerial vehicles have been identified. Helicopters are for example considered in [18], a fixed-wing aircraft in [19], and multirotors in [20]. A survey and categorization of these identification results can be found in [21]. We will present a non-parametric frequency domain-based system identification of the Flying Platform, which provides a means to asses the accuracy of two first-principle models with various degrees of complexity.

*Outline:* The hardware design is covered in Sec. 2, where the properties of a single actuation unit, comprising an electric ducted fan, an exit nozzle, and control flaps for thrust vectoring, are investigated. Both static and dynamic thrust measurements are presented. The section concludes by discussing how the actuation units are combined in the Flying Platform design. In Sec. 3, a low-complexity model describing the dynamics of the Flying Platform is presented. The dynamic model is used for optimizing the control authority of the thrust vectoring by the mechanical design, leading to a systematic trade-off between the lever arm and the total inertia of the vehicle. The model is also used for a cascaded control design as presented in Sec. 4. Flight tests show the effectiveness of the proposed control design. Sec. 5 discusses the results of a frequency domain system identification. It is shown that the low-complexity model explains the frequencies above 1 Hz well, but has limited predictive power at lower frequencies. It is argued that the model mismatch is possibly due to unmodelled aerodynamic effects, which are inherent to the ducted fan actuation. Therefore an augmented model is derived providing a better explanation of the measured data. The article concludes with final remarks in Sec. 6.

## 2. Hardware design

This section describes the hardware design of the Flying Platform. We start by presenting the design of a single actuation unit, before explaining how these are combined to actuate

[15] These values are taken from the datasheet of the fans.

[16] Other previously presented vehicles seem to be similar; [4]-[6] do not provide measurement results explicitly quantifying the thrust vectoring.

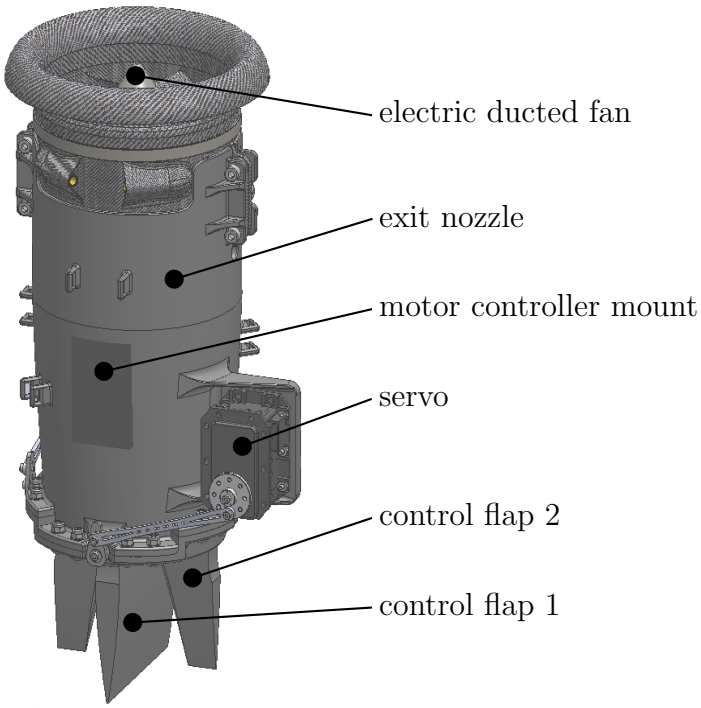


**Figure 3.2.** The different components of a single actuation unit.

the Flying Platform.

### 2.1 Actuation unit

An actuation unit consists of an electric ducted fan, an outlet nozzle, and two control flaps for thrust vectoring, see Fig. 3.2.

The thrust is generated by the Schübeler DS-51-DIA HST electric ducted fan driven by the brushless DC motor DSM4640-950. According to the datasheet of the manufacturer the ducted fan is optimized for high static thrust, yielding exit velocities up to 90 m/s. This makes thrust vectoring particularly interesting, since the force resulting from a redirection of the airflow is proportional to the square of the airflow velocity. The electric ducted fan is embedded in a convergent exit nozzle, which has an inlet area of 6940 $\mathrm{mm}^2$ and an outlet area of 5540 $\mathrm{mm}^2$. Thus, the cross section is reduced by around 20% causing the airflow to accelerate through the nozzle. The motor controller (YGE 90HV) is mounted to the exit nozzle and is cooled by the airflow. More precisely, the fins that are attached to the motor controller are inserted in the airflow through the hole in the exit nozzle, as shown in Fig. 3.2. The hole in the exit nozzle is designed such that the motor controller holds in place (press fit). In addition, the outlet nozzle has a mount for the two servos (Dynamixel RX-24F), where each of them actuate a control flap. Both the outlet nozzle and the control flaps are 3D-printed in ABS-M30. The roughness average characterizing the surface roughness of the flaps and the exit nozzle is estimated to be around $R_a = 3.2\,\mu\mathrm{m}$. In [22], the impact of roughness on the lift characteristics of a NACA0015

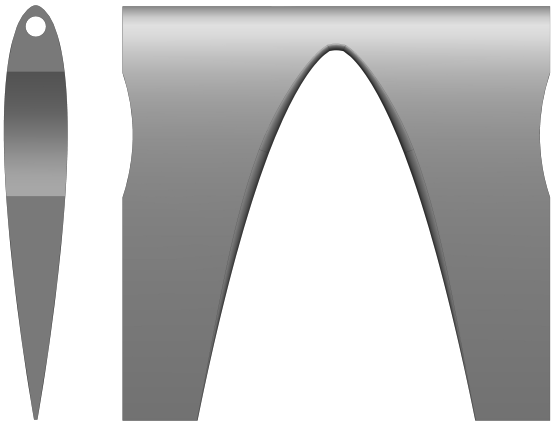

**Figure 3.3.** Side and top view of control flap 2. The flap has 80mm length (chord length) and 83mm width. Compared to the control flap 1, which has the same dimensions and the same airfoil (NACA0015), a triangular part of control flap 2 is cut out.

airfoil is characterized. It is found that at a Reynolds number of 220000, which is comparable to our set-up, an increased roughness would reduce the produced lift up to 40%. However, an increased roughness also delays the airfoil's stall to higher angles of attacks (up to a factor of two). In our case, the flaps, which have likewise a NACA0015 profile, operate at relatively high angles of attack, and therefore the roughness of the airfoil is not necessarily a disadvantage, as it might prevent stall. Note that the effect of roughness seems strongly influenced by the Reynolds number and the specific airfoil. For example, the reduction of lift due to roughness reported in [23], where windtunnel tests with the DU300-mod airfoil at Reynolds numbers above $3.6 \cdot 10^6$ are presented, are less drastic.

The two control flaps are aligned orthogonally to simplify the mechanical design of the actuation mechanism. To achieve an actuation radius of $\pm 18^\circ$ for both flaps, a triangular part of control flap 2 is cut out, thereby reducing the maximum thrust deviation achieved by control flap 2 approximately by a factor of two. A chord length of 80 mm is chosen for both flaps. The choice of the airfoil (NACA0015) is based on an optimization of the stall angle with the XFoil software package[17] at a Reynolds number of 350000, corresponding to a typical airflow velocity of 70 m/s.[18]

*Characterization of a single actuation unit:* The available thrust, and the ability of the control flaps to vector thrust is characterized using force measurements with the transducer ATI Mini-40 using the SI-20-1 calibration. This results in a sensing range of $\pm 60$ N in the vertical direction and $\pm 20$ N in the horizontal direction, with a resolution of 0.01 N. The experimental results are presented in the following.

[17] See `http://web.mit.edu/drela/Public/web/xfoil/`.

[18] The airflow velocity estimate is based on momentum theory, [2, p. 322, equation (6.41)]. For the calculation of the Reynolds number a temperature of 30° C is assumed, which is motivated by the heat loss of the motor controller and the electrical motor. The parameter Ncrit that describes the transition criterion in XFoil is set to 7.

Static thrust measurements are shown in Fig. 3.4. A single actuation unit is attached to the load cell. The motorcontroller, the servos, and the load cell are interfaced using the PX4 flight management unit, [24]. The fan is run at a constant pulse-width modulation (PWM) rate, resulting in a constant thrust of 26.4 N when both control flaps point straight down (this corresponds roughly to the hover condition of the Flying Platform). Measurements are taken at 7 different flap angles, which is found to be enough for guaranteeing that the 68% confidence interval of a resulting linear fit is below 0.024 N/° (slope) and 0.27 N (offset). The flap angle is set by the servo, which has a resolution of 0.29°. For each flap angle, 500 measurement points are taken at a sampling frequency of 50 Hz. The standard deviation obtained at each measurement point is indicated by the bars shown in Fig. 3.4. The thrust measurements display a relatively large standard deviation, which is possibly due to the turbulent flow in the exit nozzle induced by the high Reynolds number, the roughness of the 3D print, and the motor controller mount, but also due to a slight play in the connection of the control flaps with the servos. Summarizing, a maximum horizontal thrust of 3 N can be generated by control flap 1, whereas control flap 2 generates a maximum horizontal thrust of 1.5 N. This is not unexpected, since compared to control flap 1, control flap 2 has roughly half the area available for deviating the thrust. Moreover, if control flap is fully inclined, the total thrust magnitude is reduced by around 2 N, as shown in Fig. 3.4 (bottom). The decrease in total thrust is not entirely symmetric. This might be caused by the motor controller mount that destroys the symmetry of the airflow through the exit nozzle.

Similar experiments are carried out for characterizing the total thrust as a function of the PWM rate given to the motorcontroller, see Fig. 3.5. The plateau that is visible above a duty cycle of 0.8 is most likely due to limitations of the motor controller. The rotational speed of the ducted fan is found to be roughly constant at a fixed PWM rate, as can be inferred from current measurements of a single motor phase. A linear fit through the data points neighboring the PWM rate of 0.6 is performed, and will be used later. The 65% confidence interval of the fit is 0.127 N/% for the linear part and 0.7 N for the offset.

Dynamic measurements reveal that the flap angle to thrust maps can be approximated by second-order systems, with a natural frequency of around 80 rad/s and a damping of roughly 0.4. The map from PWM rate to total thrust (in case the control flaps are pointing straight down) behaves as a first-order system with a time constant of 0.01 s. The dynamic measurements were carried out using a similar procedure as presented in Sec. 5. The control flaps and the ducted fan are excited using multisine signals containing a flat frequency spectrum up to 20 Hz, respectively 10 Hz. The excitation signals have an amplitude below 10° for the flaps and an amplitude below 0.05 for the PWM rate controlling the fan. The sampling frequency is set to 100 Hz for the horizontal thrusts and 50 Hz for the vertical thrust, which is due to the limited update rate of the motor controller. The transfer function estimates are based on data collected over 62 periods, where the first two periods are discarded for eliminating transients. The experimental results are shown in Fig. 3.6 (control flap 1, control flap 2 is similar) and Fig. 3.7 (total thrust). The sharp resonance peak at 100 rad/s visible in Fig. 3.6 is attributed to the

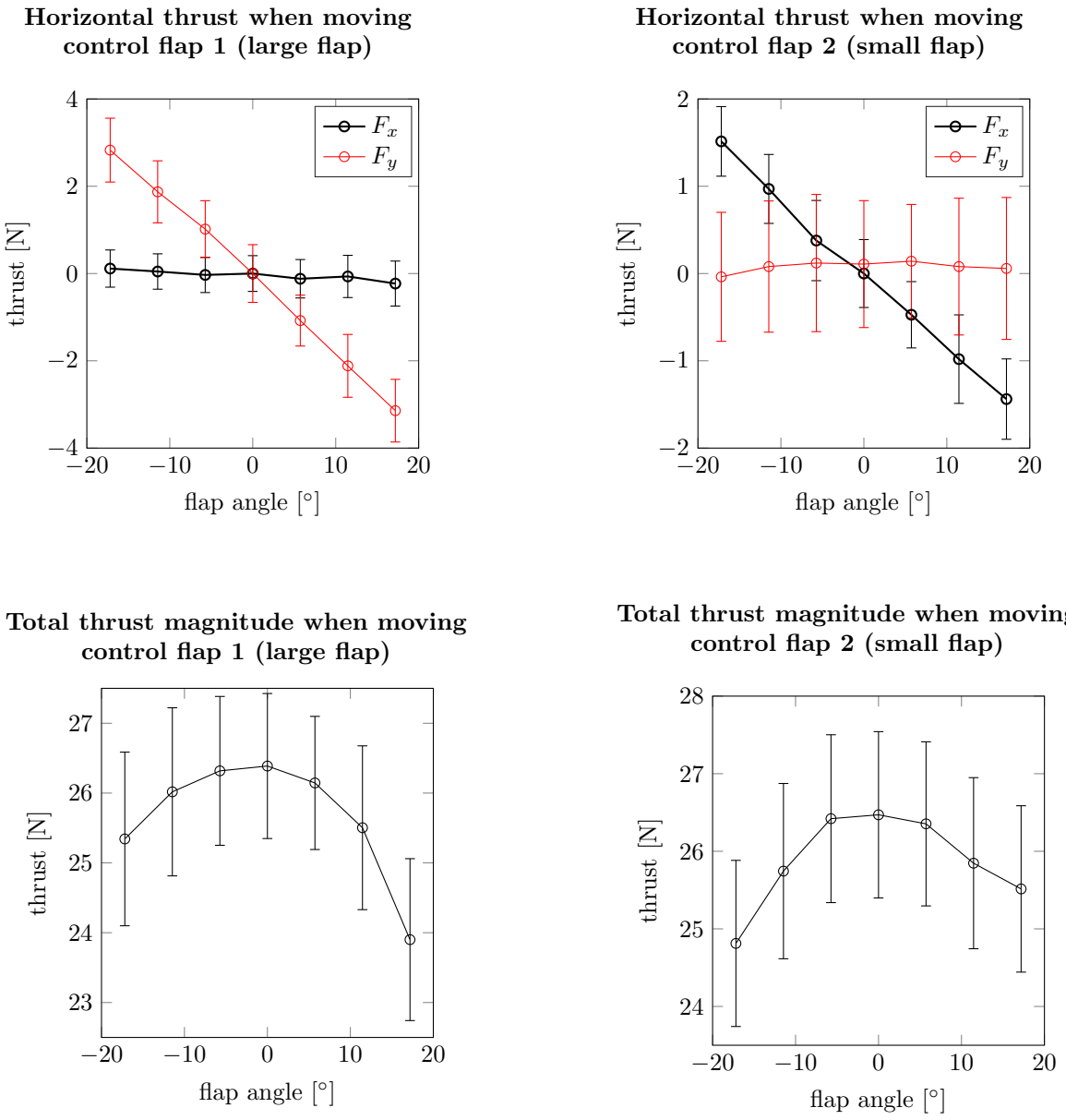


**Figure 3.4.** Shown are the x and y-components of the thrust (top row) and the total thrust magnitude (bottom row), when moving control flap 1 (left row), respectively control flap 2 (right row). The x-axis is aligned with control flap 1, the y-axis with control flap 2.

measurement setup. More precisely, it corresponds to the first eigenmode of the beam holding the load cell and the actuation unit. The parametric fit is obtained by minimizing a weighted residual, similar to Sec. 5.

## 2.2 Flying Platform

The Flying Platform design combines three actuation units, which are aligned with the corners of an equilateral triangle of 20 cm side length, as shown in Fig. 3.8. The actuation units are oriented such that the axis of the larger flap points to the center of the

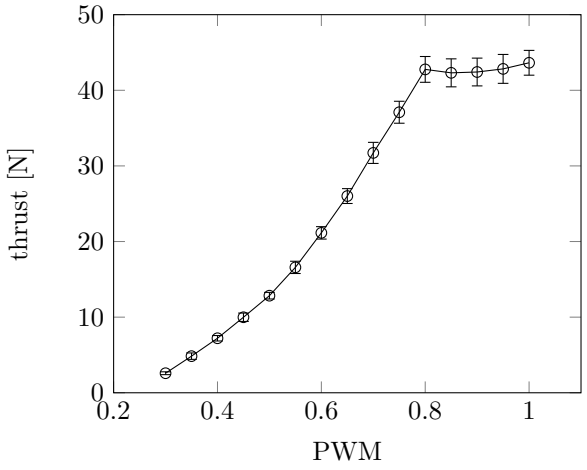


**Figure 3.5.** Total thrust generated by the actuation unit (in the vertical direction), with both control flaps pointing straight down.

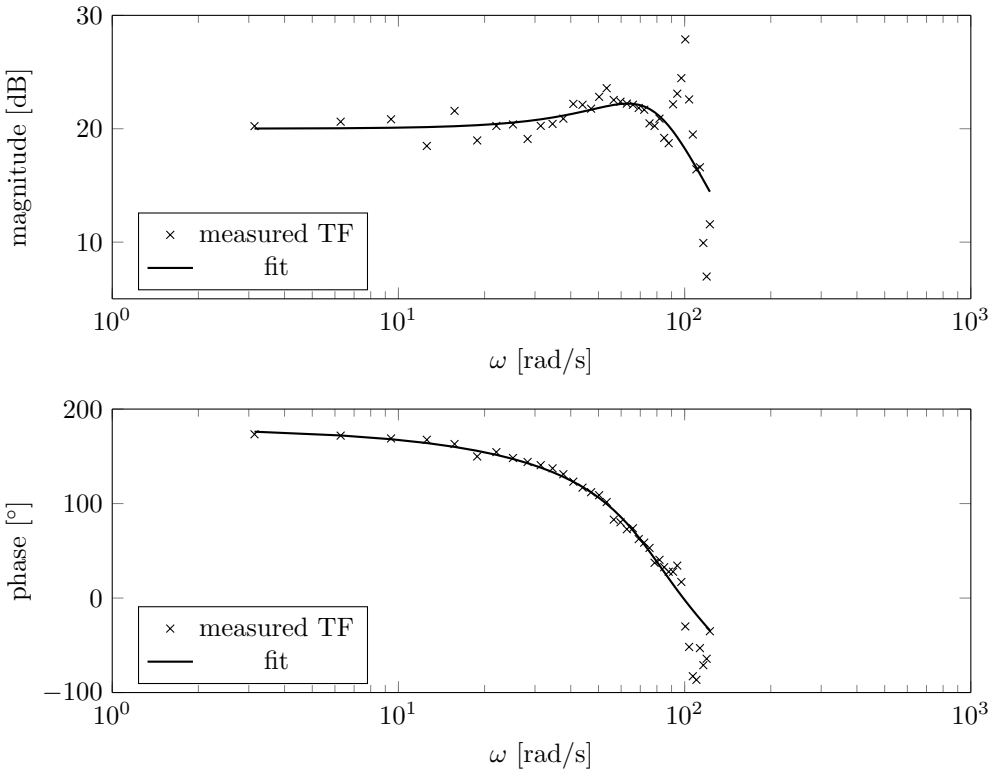


**Figure 3.6.** Transfer function from the flap angle to the horizontal thrust.

equilateral triangle. The fan units are mounted on a honeycomb carbon fibre sandwich structure. Three legs support the weight of the Flying Platform when it is on the ground. The electronics are located close to the estimated center of gravity. Tab. 3 in App. B summarizes the mechanical specification of the Flying Platform.

The PX4 flight management unit, [24] is used to run the control algorithms. The

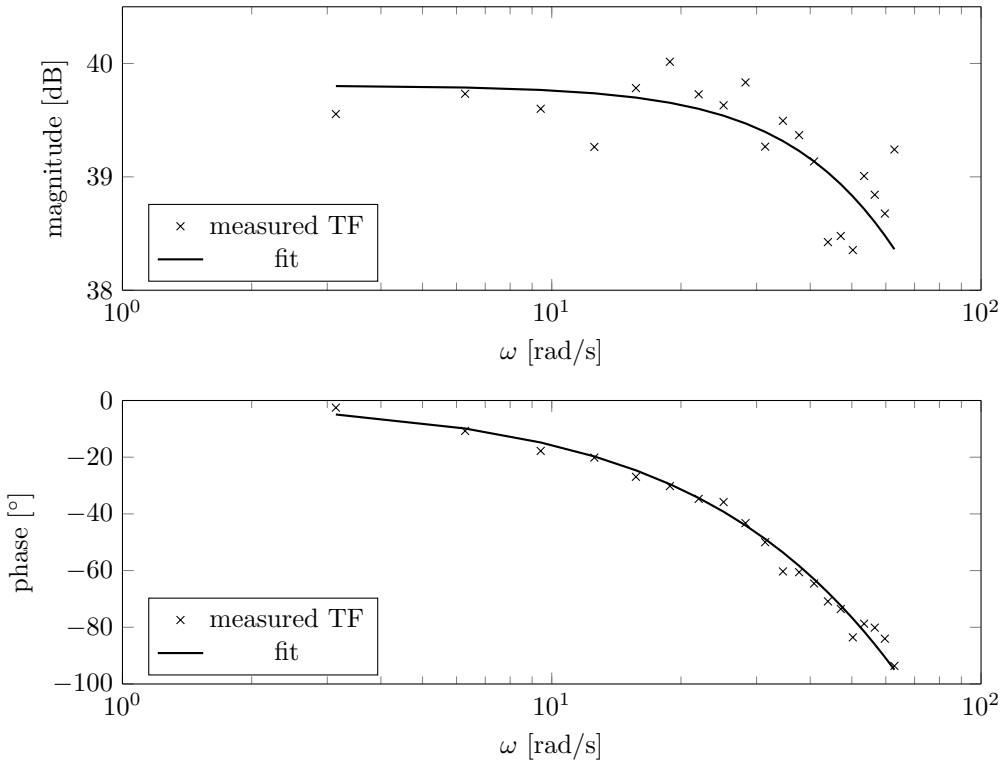


**Figure 3.7.** Transfer function from the PWM rate to the vertical thrust.

motor controllers of the electric ducted fans are interfaced via PWM. Servo commands for actuating the control flaps are sent to the servos via a serial RS485 bus. Power is delivered by three 4-cell Thunderpower Magma batteries with 6600 mAh each. The power consumption at hover is around 6kW resulting in a flight time of around 3 min. The batteries weigh 680 g each, leading to a total weight of the Flying Platform of 8.0 kg.

## 3. Dynamics

This section presents a low-complexity model of the Flying Platform. The nonlinear equations of motion are linearized about hover for control and analysis purposes. We will optimize the determinant of the controllability Gramian as a function of the actuator placements and thereby maximize the controllability about hover.

*Notation:* We introduce an inertial coordinate system {I}, a body-fixed coordinate system {B}, and local body-fixed coordinate systems $\{i\}$ oriented along the control flaps of the actuation units, see Fig. 3.9. The projection of a tensor onto a particular coordinate frame is denoted by a preceding superscript, i.e. ${}^{\mathrm{B}}\Theta \in \mathbb{R}^{3\times3}$, ${}^{\mathrm{B}}F \in \mathbb{R}^3$. The arrow notation, e.g. in Fig. 3.9, is used to emphasize that a vector (and tensor) should be a priori thought of as a linear object in a normed vector space detached from its coordinate representation in a particular coordinate frame. The transformation matrix $R_{\mathrm{IB}} \in \mathrm{SO}(3)$

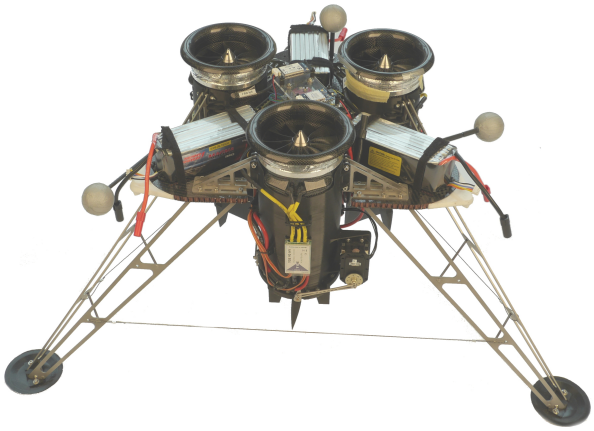

**Figure 3.8.** The Flying Platform.

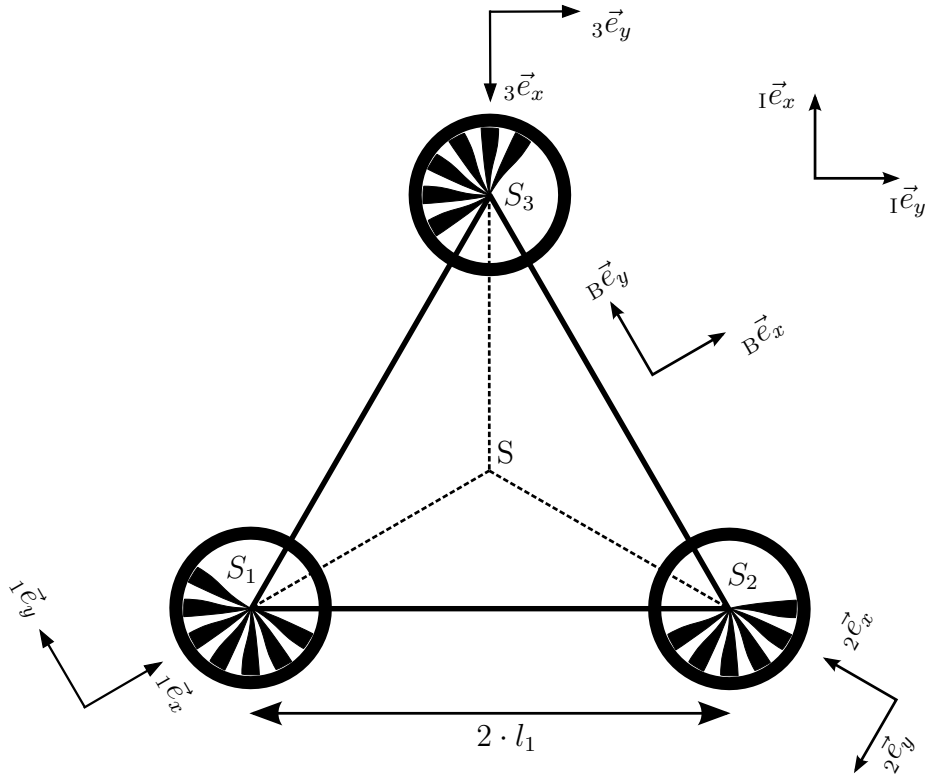


**Figure 3.9.** Schematic outline of the Flying Platform showing the coordinate frames $\{I\}$, $\{i\}$, $i = 1, 2, 3$, and $\{B\}$ (courtesy of Tobias Meier).

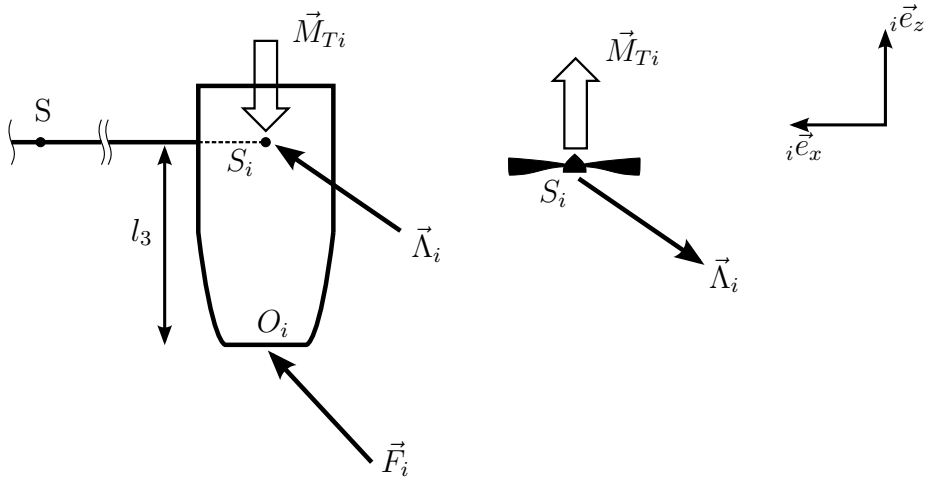


**Figure 3.10.** Free body diagram of a single actuation unit (courtesy of Tobias Meier). The motor torque $\vec{M}_{Ti}$ is aligned with the z-axis of the local coordinate frame $\{i\}$. The vertical thrust, as well as the horizontal thrust generated by the two the control flaps are combined in the force $\vec{F}_i$.

relates vectors from the body-fixed frame to their representation in the inertial frame, that is ${}^{\mathrm{I}}v = R_{\mathrm{IB}}{}^{\mathrm{B}}v$, for all vectors ${}^{\mathrm{B}}v \in \mathbb{R}^3$. Moreover, the skew symmetric matrix corresponding to a vector $a \in \mathbb{R}^3$, denoted by $\widetilde{a}$, is defined as $a \times b = \widetilde{a}b$, for all $b \in \mathbb{R}^3$, where $a \times b$ refers to the cross product of the two vectors $a$ and $b$. Since the body-fixed coordinate frame $\{\mathrm{B}\}$ is the most commonly projected coordinate frame, its preceding superscript is usually removed for ease of notation, that is, ${}^{\mathrm{B}}m = m$, ${}^{\mathrm{B}}\Theta_0 = \Theta_0$, etc. The standard unit vectors in $\mathbb{R}^3$ are denoted by $e_x$, $e_y$, and $e_z$. Vectors are expressed as $n$-tuples $(x_1, x_2, \ldots, x_n)$ with dimension and stacking clear from the context.

*Dynamics:* The equations of motion can be derived, for example, by using the principle of virtual power, [25, Ch. 3]. To that extent, the moving parts of the $i$'th actuation unit (turbine blades and shaft of the electrical motor) are separated from the remaining structure by introducing the constraint forces $\vec{\Lambda}_i$ and the motor torques $\vec{M}_{Ti}$, see Fig. 3.10. Requiring the virtual power to vanish for all virtual velocities (translational and rotational) yields the following characterization of the dynamic equilibrium,

$$\Theta\dot{\omega} + \sum_{i=1}^{3} \Theta_i \dot{\omega}_i = -\widetilde{\omega}\left(\Theta\omega + \sum_{i=1}^{3} \Theta_i \omega_i\right) + \sum_{i=1}^{3} \widetilde{r}_i F_i, \tag{3.1}$$

$$m^{\mathrm{I}}\dot{v} = m^{\mathrm{I}}g + \sum_{i=1}^{3} R_{\mathrm{IB}} F_i, \tag{3.2}$$

$$Ce_z^{\mathsf{T}}(\dot{\omega} + \dot{\omega}_i) = M_i, \quad i = 1, 2, 3, \tag{3.3}$$

where $\Theta$ denotes the total inertia of the Flying Platform referred to its center of gravity $S$, $\Theta_i$ the inertia of the moving parts of the $i$'th ducted fan referred to its center of rotation, and $m$ the total mass. The velocity of the center of mass of the vehicle is denoted by $v$, whereas $\omega$ refers to its angular velocity, i.e. the angular velocity of the frame $\{B\}$ with respect to frame $\{I\}$. The thrust generated by the $i$'th actuation unit, that is, the vertical thrust from the electric ducted fan, vectored by the two control flaps, is denoted by $F_i$. The vector from the center of gravity to the point of origin of the force $F_i$ is denoted by $r_i$. Aerodynamic effects except the forces generated by the control flaps and the thrust of the fans are neglected. These will be included in an augmented model as presented in Sec. 5. The scalar $M_i$ and the vector $\omega_i$ denote the torque of the electrical motor, respectively the angular rate (relative to the body-fixed frame $\{B\}$) of the $i$'th ducted fan. The rotating parts (turbine blades and electrical motor) of the actuation units are assumed to be symmetric and rotate about their respective center of gravity resulting in[19]

$$\Theta_i =: \operatorname{diag}(\bar{C}, \bar{C}, C). \tag{3.4}$$

The angular velocity vector $\omega_i$ is assumed to have only a component along the z-axis of the body-fixed frame. Therefore its rate of change $\dot{\omega}_i$ appearing in (3.1) can be eliminated with (3.3) resulting in

$$\hat{\Theta}\dot{\omega} = -\widetilde{\omega}\left(\Theta\omega + \sum_{i=1}^{3}\Theta_i\omega_i\right) + \sum_{i=1}^{3}(\widetilde{r}_i F_i - e_z M_i), \tag{3.5}$$

where

$$\hat{\Theta} := \Theta - 3\ C e_z e_z^{\mathsf{T}}. \tag{3.6}$$

We will consider the thrusts generated by the actuation units and expressed in their local coordinate frames $\{i\}$ to be the inputs to the system. The servo and PWM-commands for the electric ducted fans are then calculated by inverting the linearization of the static maps presented in Sec. 2.1. The total thrust and the resulting torque are linear in the thrusts generated by the actuation units (the inputs), more precisely,

$$\sum_{i=1}^{3} F_i = T_1 u, \quad \sum_{i=3}^{3} \widetilde{r}_i F_i = T_2 u, \tag{3.7}$$

where $u := ({}^1F_1, {}^2F_2, {}^3F_3)$,

$$T_1 := \begin{pmatrix} T_{11} \\ T_{12} \end{pmatrix}, \quad T_2 := \begin{pmatrix} T_{21} \\ T_{22} \end{pmatrix}, \tag{3.8}$$

[19] In fact, the expression remains unchanged if the inertia is expressed in the local frame $\{i\}$ or in a frame attached to the moving parts of fan $i$.

with

$$T_{11} := \begin{pmatrix} 1 & 0 & 0 & -1/2 & -\sqrt{3}/2 & 0 & -1/2 & \sqrt{3}/2 & 0 \\ 0 & 1 & 0 & \sqrt{3}/2 & -1/2 & 0 & -\sqrt{3}/2 & -1/2 & 0 \end{pmatrix}, \tag{3.9}$$

$$T_{12} := \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}, \tag{3.10}$$

$$T_{21} := l_3 J T_{11} + 2\sqrt{3}/3 l_1 V_1, \tag{3.11}$$

$$T_{22} := -2\sqrt{3}/3 l_1 \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{3.12}$$

$$V_1 := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -\sqrt{3}/2 & 0 & 0 & \sqrt{3}/2 \\ 0 & 0 & 1 & 0 & 0 & -1/2 & 0 & 0 & -1/2 \end{pmatrix}, \tag{3.13}$$

$$J := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.14}$$

As a result, the evolution of the center of gravity and the evolution of the angular velocity are given by

$$m^{\mathrm{I}} \dot{v} = m^{\mathrm{I}} g + R_{\mathrm{IB}} T_1 u, \tag{3.15}$$

$$\hat{\Theta}\dot{\omega} = -\widetilde{\omega}\left(\Theta\omega + \sum_{i=1}^{3} \Theta_i \omega_i\right) + T_2 u - e_z \sum_{i=1}^{3} M_i. \tag{3.16}$$

*Linearization:* For control and analysis purposes the dynamics are linearized about hover. The three ducted fans are assumed to be identical and to rotate in the same direction. Thus, at hover, the torques $M_i$ have the same values, that is, $M_i = M$, $i = 1, 2, 3$. Moreover, the torques $M$ and the weight of the vehicle must be balanced by the thrust generated by the ducted fans and deviated by the control flaps, which is achieved by the thrust command

$$\bar{u} := (0,\ -M\sqrt{3}/(2l_1),\ mg_0/3,\quad 0,\ -M\sqrt{3}/(2l_1),\ mg_0/3,\quad 0,\ -M\sqrt{3}/(2l_1),\ mg_0/3),$$

where $g_0 := 9.81\ \mathrm{m/s^2}$ denotes the gravitational acceleration. For better readability the components of the vector $\bar{u}$ in the above equation are grouped according to the different actuation units, that is, the first line contains the x, y, and z-components of the thrust assigned to the first fan unit, the second line contains the thrust assigned to the second fan unit, etc. We further introduce Euler angles $(\alpha, \beta, \gamma)$ (roll, pitch, yaw) to parametrize the rotation matrix $R_{\mathrm{IB}}$. Using the matrix exponential, the rotation matrix $R_{\mathrm{IB}}$ can be expressed as

$$R_{\mathrm{IB}} = e^{\widetilde{e}_z \gamma} e^{\widetilde{e}_y \beta} e^{\widetilde{e}_x \alpha}. \tag{3.17}$$

For control purposes it will be convenient to obtain a linearization that is invariant to yaw. Therefore the position and velocity of the center of gravity will be expressed in a separate coordinate system $\{J\}$ obtained by rotating the inertial system $\{I\}$ about ${}_{\mathrm{I}}\vec{e}_z$

by the angle $\gamma$. Hence, the rotation matrix $R_{\mathrm{IB}}$ is decomposed according to

$$R_{\mathrm{IB}} = R_{\mathrm{IJ}} R_{\mathrm{JB}}, \quad R_{\mathrm{IJ}} = e^{\widetilde{e}_z \gamma}, \quad R_{\mathrm{JB}} = e^{\widetilde{e}_y \beta} e^{\widetilde{e}_x \alpha}, \tag{3.18}$$

and (3.15) is reformulated as

$$m^{\mathrm{J}} \dot{v} = -m\dot{\gamma} e_z \times {}^{\mathrm{J}}v + m^{\mathrm{J}} g + R_{\mathrm{JB}} T_1 u, \tag{3.19}$$

where the convective derivative enters due to the fact that the frame $\{J\}$ is non-inertial. Linearizing the translational dynamics around hover, i.e. ${}^{\mathrm{J}}\bar{v} = 0$, $\bar{R}_{\mathrm{JB}} = I$, $\bar{\omega} = 0$, yields

$$\begin{aligned} {}^{\mathrm{J}}\dot{v} &\approx -\alpha \widetilde{e}_x {}^{\mathrm{J}}g - \beta \widetilde{e}_y {}^{\mathrm{J}}g + \frac{1}{m} T_1 (u - \bar{u}) && (3.20)\\ &= g_0 \left(\alpha \widetilde{e}_x e_z + \beta \widetilde{e}_y e_z\right) + \frac{1}{m} T_1 (u - \bar{u}) && (3.21)\\ &= g_0 \left(-e_y \alpha + e_x \beta\right) + \frac{1}{m} T_1 (u - \bar{u}), && (3.22) \end{aligned}$$

which holds independent of the angle $\gamma$. Similarly, linearizing the rotational dynamics (3.16) around $\bar{\omega} = 0$, and neglecting the gyroscopic term $C\hat{\Theta}^{-1}\widetilde{\omega}\omega_i$ results in

$$\dot{\omega} \approx \hat{\Theta}^{-1} T_2 u - \hat{\Theta}^{-1} e_z \sum_{i=1}^{3} M_i. \tag{3.23}$$

From (3.22) and (3.23) it can be inferred that the poles of the open-loop system all lie at 0, and that the height and yaw dynamics are decoupled from the x, y, and roll and pitch dynamics.

Assuming further that the mass distribution of the Flying Platform has a three-fold rotational symmetry about its figure axis ${}_{\mathrm{B}}\vec{e}_z$ simplifies the inertia tensor $\hat{\Theta}$ to

$$\hat{\Theta} =: \mathrm{diag}(I_1, I_1, I_3). \tag{3.24}$$

This is a reasonable assumption due to the symmetric placement of both the actuation units and the batteries, and the symmetry of the frame, which together constitute the main mass of the Flying Platform. Thus, the x, y, and roll and pitch dynamics can be rewritten as

$$\begin{pmatrix} \dot{v}_x \\ \dot{v}_y \end{pmatrix} \approx g_0 J \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \frac{1}{m} T_{11} (u - \bar{u}), \qquad \begin{pmatrix} \ddot{\alpha} \\ \ddot{\beta} \end{pmatrix} \approx \frac{1}{I_1} T_{21} (u - \bar{u}), \tag{3.25}$$

whereas the vertical and the yaw dynamics are given by

$$\dot{v}_z \approx \frac{1}{m} T_{12}(u - \bar{u}), \quad \ddot{\gamma} \approx \frac{1}{I_3} T_{22} u - \frac{1}{I_3} \sum_{i=1}^{3} M_i. \tag{3.26}$$

*Controllability analysis:* We determine the overall dimensions of the Flying Platform, that is the lengths $l_1$ and $l_3$ by maximizing the determinant of the controllability Gramian subject to the dynamics (3.25). This amounts to maximizing the volume of the state space from which the Flying Platform can be steered to zero within a fixed time $T$ and with unit energy (assuming linear dynamics, i.e. near hover conditions), [26, Ch. 8]. We focus entirely on the actuation via thrust vectoring, and therefore the differential thrust is set to zero. As we will show in the remainder, this leads to a simple closed-form expression of the determinant of the controllability Gramian, which enables a physical interpretation, and leads to a straightforward optimization of the mechanical design.

By defining the state vector to be

$$x := (v_x, v_y, \alpha, \beta, \dot{\alpha}, \dot{\beta}), \tag{3.27}$$

the linearized system dynamics (3.25) can be rewritten in the standard form

$$\dot{x} = Ax + B(u - \bar{u}), \tag{3.28}$$

for which the controllability Gramian, [26, p. 227], is defined as

$$W_c(T) := \int_0^T e^{-At} B B^{\mathsf{T}} e^{-A^{\mathsf{T}} t} \mathrm{d}t. \tag{3.29}$$

The values of the matrices $A$ and $B$ are given in (3.59) in App. A. Given that we have unit energy at our disposal, the system can be steered within the time $T$ to the origin from any initial condition within the ellipsoid

$$\mathcal{W}(T) := \{z \in \mathbb{R}^6 \mid z^{\mathsf{T}} W_c(T)^{-1} z \leq 1\}. \tag{3.30}$$

The area of $\mathcal{W}(T)$ is proportional to the square root of the determinant of $W_c(T)$. For the given dynamics, the determinant of $W_c(T)$ can be calculated in closed form, see App. A, leading to

$$\det(W_c(T)) = \frac{g_0^4 T^{18}}{102400} \left(\frac{l_3}{I_1}\right)^{12}. \tag{3.31}$$

The following observations can be made:

1) For any $I_1 \neq 0, l_3 \neq 0$ the Flying Platform can be steered from any initial condition

to the origin, provided that $T$ is sufficiently large. This is not surprising, as the system's poles all lie at 0.

2) The area of $\mathcal{W}(T)$ only depends on the inertia $I_1$ and the length $l_3$. The total mass, for example, enters the expression only through the inertia $I_1$.

3) For a fixed, but arbitrary $T$, the area of $\mathcal{W}(T)$ attains its maximum if the ratio $l_3/I_1$ is maximized.

Hence we chose the dimensions of the Flying Platform, $l_1$ and $l_3$, such that the ratio $l_3/I_1$ is as large as possible. Clearly, $I_1$ is implicitly dependent on $l_3$, as the actuation units have substantial mass. This dependence is captured by approximating the inertia $I_1$ as

$$I_1 \approx I_0 + 2(l_1^2 + l_3^2\delta^2)m_t, \tag{3.32}$$

where $m_t$ refers to the mass of a single actuation unit, whose center of gravity lies at a height of $\delta l_3$ below the center of gravity of the vehicle, and $I_0$ refers to the remaining inertia, which is independent of $l_3$. As a result, we seek to maximize the ratio

$$\frac{l_3}{I_0 + 2l_1^2 m_t + 2m_t\delta^2 l_3^2}, \tag{3.33}$$

which is achieved by decreasing $I_0$ and $l_1$ as much as possible. Moreover, for a fixed inertia $I_0$, length $l_1$, and mass $m_t$, the previous expression is maximized for

$$l_{3,\max} = \sqrt{\frac{\frac{I_0}{2m_t} + l_1^2}{\delta^2}}. \tag{3.34}$$

By assuming that the weight of the Flying Platform is mainly given by the actuation units and the weight of the batteries, which are located at a horizontal distance $l_1$ from the ${}_{\mathrm{B}}\vec{e}_x$, respectively the ${}_{\mathrm{B}}\vec{e}_y$ axis, we obtain $I_0 \approx 1.4\,\mathrm{kg}\; l_1^2$. Together with $m_t \approx 1.2\,\mathrm{kg}$, and $\delta \approx 0.75$ this yields $l_{3,\max} \approx 1.7 l_1$. In the design the length $l_1$ was bounded from below to $l_1 = 10\,\mathrm{cm}$ for ease of assembly, and therefore $l_3$ was chosen to be roughly 17 cm.

The optimization over $l_3$ can be viewed as a trade-off between the total inertia of the flying vehicle and the lever arm of the thrust vectoring; the further away the actuation units are placed, the larger the lever arm, and the higher the torque generated by the thrust vectoring, but at the same time the inertia is increased. The lever arm grows linearly with $l_3$, whereas the inertia grows quadratically leading to the optimum captured by (3.34).

Moreover, the formula (3.31) is valid irrespective of the sign of $l_3$. Thus, the above derivation remains valid even in case the thrust vectoring is placed above the center of gravity. Note that having the thrust vectoring below the center of gravity leads to a nonminimum phase zero in the transfer function from the horizontal thrust to the horizontal velocity, as for example noted in [4]. It stems from the fact that the thrust

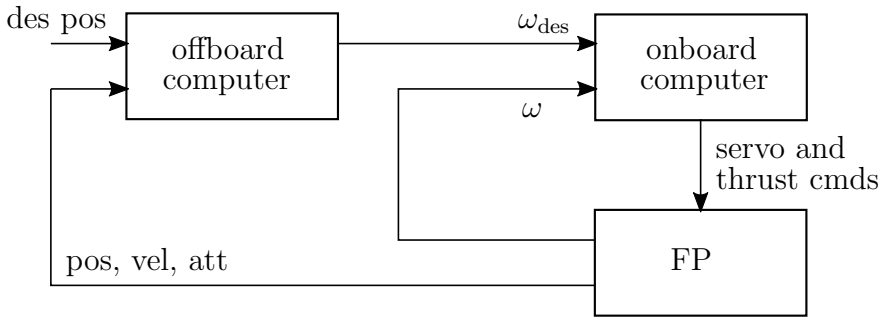


**Figure 3.11.** Overview of the control architecture. FP stands for Flying Platform. The angular rates $\omega$ are measured with an onboard gyroscope. The position, velocity and attitude of the vehicle are obtained from a motion capture system.

vectoring generates lateral forces, which induce a torque with respect to the center of gravity, causing the vehicle to accelerate horizontally and tilt in the opposite direction at the same time. In case the thrust vectoring is placed above the center of gravity two complex conjugated, pure imaginary zeros are obtained instead. For ease of construction we decided to choose $l_3 > 0$.

## 4. Control Design

We present a linear control design for stabilizing hover. The controller has a cascaded structure, with a part running onboard at 50 Hz, accessing onboard sensor measurements and controlling the angular rates of the vehicle, and a part running offboard, controlling the position and attitude, see Fig. 3.11.

*Control system overview:* The position, velocity, and attitude of the vehicle is estimated using a motion capture system, [27]. The system estimates position with a precision of roughly 0.3 mm, and attitude with a precision of roughly 0.3° ($2\sigma$-bounds, sampled at 200 Hz). The velocity is obtained by low-pass filtering and numerical differentiation of the position estimate. The data from the motion capture system is sent to an offboard computer, which implements a user interface and calculates the desired angular rates for the flying vehicle. The offboard computer runs at a sampling rate of 50 Hz. The desired angular rates are sent to the vehicle via a low-latency protocol, and are then tracked by the flying vehicle using the gyroscope included on the PX4 flight computer. The onboard control algorithm runs at 50 Hz. Telemetry data from the flying vehicle is sent out via a separate wireless radio.

*Onboard control:* The onboard controller tracks the desired angular rates $\omega_{\text{des}}$, which are obtained from the offboard computer. About hover, the rotational dynamics can be approximated by, c.f. (3.23),

$$\dot{\omega} = \hat{\Theta}^{-1} T_2 (u - \bar{u}), \tag{3.35}$$

where the torques $M_i$ are approximated as constants, compensated by the steady-state

control input $\bar{u}$. A linear quadratic regulator, with state weight $512 \cdot I$ and input weight

$$\text{diag}( \underbrace{1,2,2,}_{\substack{\text{1st actuation unit}\\ \text{x,y,z-components}}} \underbrace{1,2,2,}_{\substack{\text{2nd actuation unit}\\ \text{x,y,z-components}}} \underbrace{1,2,2}_{\substack{\text{3rd actuation unit}\\ \text{x,y,z-components}}} ) \tag{3.36}$$

is used to compute a constant feedback gain $K$, rendering (3.35) asymptotically stable with

$$u = \bar{u} - K(\omega - \omega_{\text{des}}) + (0,0,1,\ \ 0,0,1,\ \ 0,0,1)F_z, \tag{3.37}$$

where $F_z$ denotes the collective thrust of the three electric ducted fans. The collective thrust does not affect the angular rates and will be used in a later stage to control the height of the flying vehicle. The obtained feedback gain $K$ results in closed-loop poles at 42 rad/s (for $\omega_x$), 42 rad/s (for $\omega_y$), and 25 rad/s (for $\omega_z$).

*Offboard control:* Under the assumption that the inner control loop has a substantially faster time constant, we consider $\omega_{\text{des}}$ to be the control input of the outer control loop, controlling the position, attitude, and velocity of the flying vehicle. As a result, (3.22) simplifies to

$$\dot{v}_x \approx \beta g_0, \quad \dot{v}_y \approx -\alpha g_0, \quad \dot{v}_z = \frac{3}{m} F_z, \tag{3.38}$$

where ${}^{\text{J}}v =: (v_x, v_y, v_z)$. Differentiating the first two equations with respect to time yields

$$\ddot{v}_x = \omega_{\text{des},y} g_0, \quad \ddot{v}_y = -\omega_{\text{des},x} g_0. \tag{3.39}$$

Thus we choose

$$\omega_{\text{des},x} = \frac{1}{g_0}\left(-(2d_y w_y + p_y)g_0\alpha + (w_y^2 + 2d_y w_y p_y)v_y + p_y w_y^2(y - y_{\text{des}})\right), \tag{3.40}$$

$$\omega_{\text{des},y} = \frac{1}{g_0}\left(-(2d_x w_x + p_x)g_0\beta - (w_x^2 + 2d_x w_x p_x)v_x - p_x w_x^2(x - x_{\text{des}})\right), \tag{3.41}$$

$$\omega_{\text{des},z} = -\frac{1}{g_0} p_z(\gamma - \gamma_{\text{des}}), \tag{3.42}$$

$$F_z = \frac{m}{3}(-2d_z w_z v_z - w_z^2(z - z_{\text{des}})), \tag{3.43}$$

where $d_i, w_i, p_i$ with $i = x, y, z$ are constants, $x, y, z$ and $x_{\text{des}}, y_{\text{des}}, z_{\text{des}}$ denotes the actual and desired position of the vehicle expressed in the $\{J\}$ frame, and $\gamma_{\text{des}}$ the desired yaw angle. The constants $d_i, w_i, p_i$ with $i = x, y$ are chosen such that the translational closed-loop dynamics in the $\{J\}$ frame result in two decoupled third-order systems with one pole located at $-p_x$ (respectively $-p_y$) and a remaining second-order system with damping $d_x$ (respectively $d_y$) and natural frequency $w_x$ (respectively $w_y$). The constant $p_z$ determines

| | rms error (x,y,z component) | | |
|---|---|---|---|
| ${}^{\mathrm{I}}r$ | 0.013 m | 0.029 m | 0.005 m |
| $\phi$ | 0.007° | 0.004° | 0.008° |
| $\omega$ | 0.029 rad/s | 0.022 rad/s | 0.011 rad/s |

**Table 1.** Root-mean-squared errors when hovering in steady state.

the time-constant of the yaw dynamics, whereas the closed-loop dynamics for the height result in a second-order system with damping $d_z$ and natural frequency $w_z$. The constants are set to the following values

$$\begin{aligned} &d_x = d_y = d_z = 1, \\ &\omega_x = \omega_y = 3\,\mathrm{rad/s}, \quad \omega_z = 2\,\mathrm{rad/s}, \\ &p_x = p_y = 1\,\mathrm{rad/s}, \quad p_z = 2\,\mathrm{rad/s}, \end{aligned}$$

leading to a clear separation of the time constants associated with the inner and the outer control loop. This results in a symmetric behavior in the x and y-directions, whereas the height is controlled in a slightly less aggressive manner ($\omega_z < \omega_x, \omega_y$). The damping is set to 1, leading to critically damped systems.

Flight experiments are carried out in the Flying Machine Arena, [27]. Tab. 1 shows the root-mean-squared errors when hovering in steady state. It follows that the vehicle maintains its position within a few centimeters. Disturbance rejection measurements are shown in Fig. 3.12. The disturbance is generated by commanding a constant angular rate in $y$-direction, $\omega_y = 0.3\,\mathrm{rad/s}$ for 0.18 s, leading to a pitch of approximately 4° from which the vehicle is able to recover.

## 5. System Identification

The following section describes a frequency domain-based approach for identifying the parameters of the Flying Platform. Specifically, the aim is to quantify the model quality and identify the matrices $T_1/m$ and $\hat{\Theta}^{-1}T_2$, essentially determining the rotational and translational dynamics, (3.15) and (3.16). This is done by exciting the system while hovering with periodic, sinusoidal inputs, and measuring its reaction. Due to the fact that the system has nine inputs defined as the thrust commands of each actuation unit, at least nine different experiments are used to measure the corresponding frequency response function. In order to reduce the noise influence we performed in total 18 different experiments, which are based on two different excitation signals (for increasing robustness against nonlinearities, [28, Ch. 3]). The experiments, which are referred to by the subscript $e$, $e \in \{1, 2, \ldots, 18\}$, can be grouped in three parts: Part 1) ($e \in \{1, 4, 7, 10, 13, 16\}$): excitation of the control flaps 1 of each actuation unit; Part 2) ($e \in \{2, 5, 8, 11, 14, 17\}$):

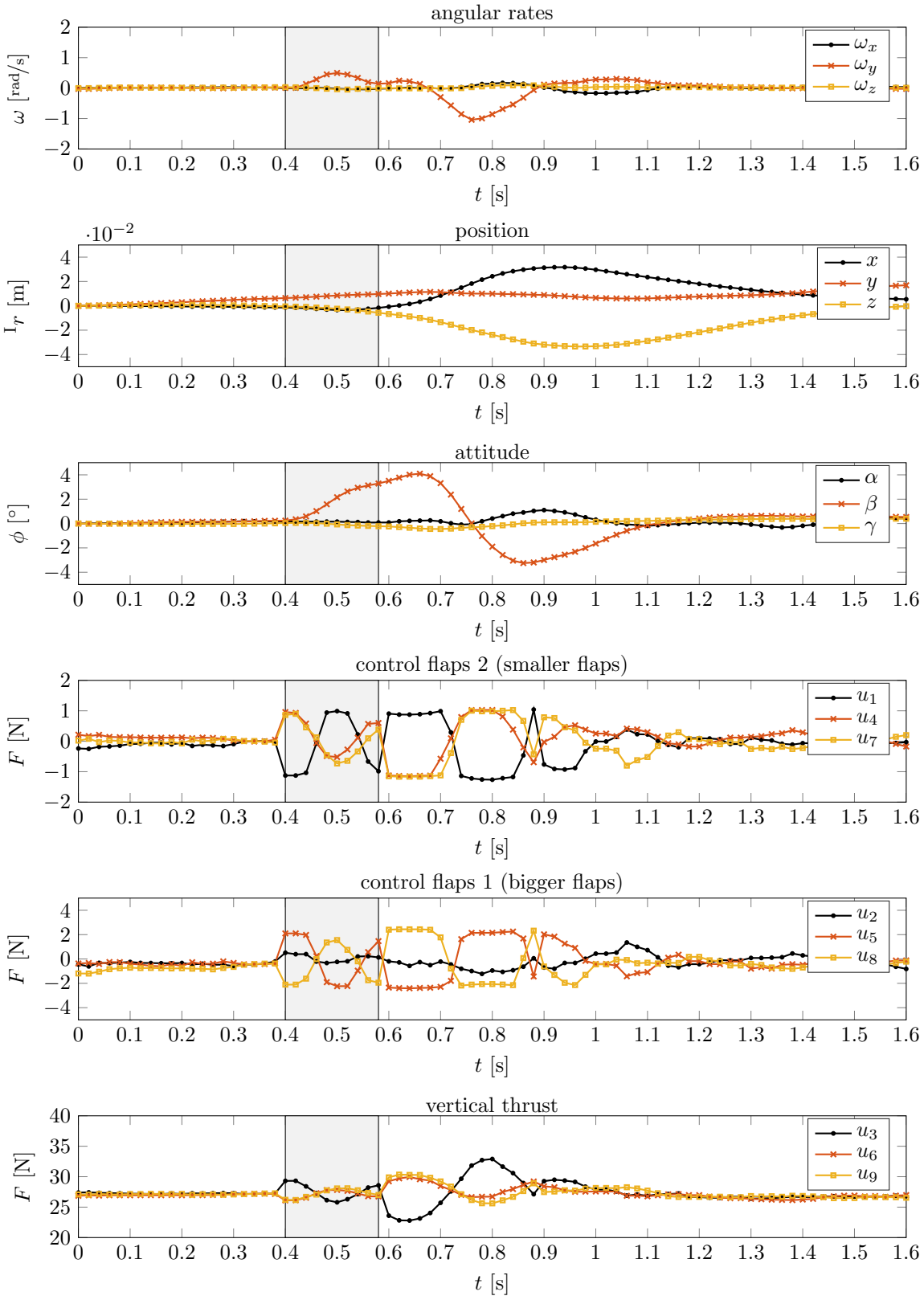


**Figure 3.12.** Disturbance rejection. At time $t = 0.4\,\mathrm{s}$ the disturbance is injected, by commanding angular rates of $(0, 0.3\,\mathrm{rad/s}, 0)$ for $0.18\,\mathrm{s}$. The time instances at which the disturbance is active are highlighted. The position and attitude (yaw) is shifted to zero at time 0.

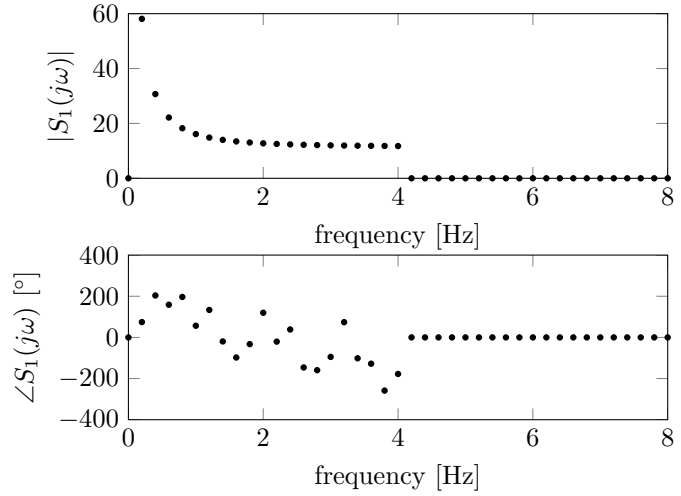


**Figure 3.13.** Excitation signal $S_1(j\omega)$. The low frequencies have a larger magnitude to compensate the fact that the signal to noise ratio is worse at low frequencies. The excitation signal $S_2(j\omega)$ has the same magnitude, but a different phase realization.

excitation of the control flaps 2 of each actuation unit; Part 3) ($e \in \{3, 6, 9, 12, 15, 18\}$): excitation of the vertical thrusts of each actuation unit. The different excitation signals are obtained by multiplying two scalar random phase multisine signals $S_1(j\omega)$ and $S_2(j\omega)$ (to be made precise below) with the 3-point discrete Fourier transform matrix $V(j\omega) \in \mathbb{C}^{3\times 3}$, resulting in

$$R(j\omega) = \begin{pmatrix} (V(j\omega) \otimes \mathrm{diag}(\lambda))\, S_1(j\omega) \\ (V(j\omega) \otimes \mathrm{diag}(\lambda))\, S_2(j\omega) \end{pmatrix}, \quad R(j\omega) \in \mathbb{C}^{18\times 9}, \tag{3.44}$$

where $\lambda \in \mathbb{R}^3, \lambda > 0$ represents a positive gain for scaling the excitation, and $\otimes$ refers to the Kronecker product. Multiplying the scalar multisine signals with the 3-point discrete Fourier transform matrix leads to an improved condition number of the pseudo-inverse needed to calculate the frequency response function, [28, p. 66]. The matrix $R(j\omega)$ contains the excitation signals for the different inputs as rows. Hence, for example in the first experiment of Part 1), the excitation signals $\lambda_1 V_{11}(j\omega)S_1(j\omega)$, $\lambda_1 V_{12}(j\omega)S_1(j\omega)$, $\lambda_1 V_{13}(j\omega)S_1(j\omega)$ are used to excite the control flaps 1 of each actuation unit (the remaining control flaps and the vertical thrusts are not excited). The multisine signals $S_1(j\omega)$ and $S_2(j\omega)$ have a random phase uniformly distributed in $[0, 2\pi)$, are sampled with 50 Hz, and have a period of 250 samples. The Crest-factor, [28, p. 153] is reduced by optimizing over 1000 different phase-realizations. The resulting signal $S_1(j\omega)$ used for the identification is shown in Fig. 3.13, the signal $S_2(j\omega)$ has the same magnitude, but a different phase realization. Note that due to their periodicity, the random phase multisine signals prevent spectral leakage.

Thus, while hovering, the Flying Platform is excited with the signal $R_e(j\omega)$, where $R_e(j\omega)$ denotes the $e$'th row of $R(j\omega)$. The setup is illustrated in Fig. 3.14. The perio-

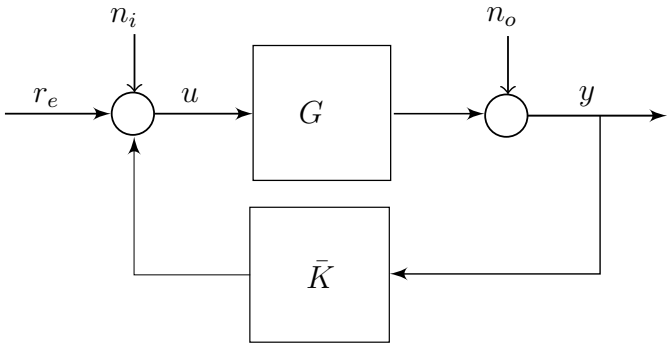


**Figure 3.14.** The block diagram of the system identification procedure. The Flying Platform ($G$) is controlled by the nominal linear feedback controller $\bar{K}$, as presented in Sec. 4, and is excited by the random phase multisine signal $r_e$, that is, the time-domain representation of the signal $R_e(j\omega)$. The noise on the input and on the output is denoted by $n_i$, respectively $n_0$.

dic excitation leads naturally to a periodic input $U_e(j\omega)$ and a periodic output $Y_e(j\omega)$ (assuming the system is linear). The input is given by the thrust commands to each actuation unit, $u := ({}^1F_1, {}^2F_2, {}^3F_3)$ and the output is taken to be the angular velocity and the velocity of the center of mass, $y := ({}^{\mathrm{J}}v, \omega)$. By averaging over multiple periods the impact of the noise can be reduced, leading to

$$Y_e(j\omega) = \frac{1}{P}\sum_{p=1}^{P} Y_{ep}(j\omega), \quad Y_e(j\omega) \in \mathbb{C}^{12}, \tag{3.45}$$

$$U_e(j\omega) = \frac{1}{P}\sum_{p=1}^{P} U_{ep}(j\omega), \quad U_e(j\omega) \in \mathbb{C}^{9}, \tag{3.46}$$

where $P = 10$ refers to the number of periods, and $Y_{ep}(j\omega)$ refers to the Fourier transform of the output of the $e$'th experiment and the $p$'th period. In order to reduce the effect of transients the first 200 samples are discarded. In a similar way, the sample covariances are given by

$$\hat{\sigma}^2_{XZe}(j\omega) = \frac{1}{P(P-1)}\sum_{p=1}^{P}(X_{ep}(j\omega) - X_e(j\omega)(Z_{ep}(j\omega) - Z_e(j\omega))^*, \tag{3.47}$$

where $X = U, Y$, and $Z = U, Y$.

An estimate of the transfer function $G(j\omega)$ is obtained by combining the inputs and outputs of all experiments, i.e. $Y(j\omega) = (Y_1(j\omega), Y_2(j\omega), \ldots, Y_{18}(j\omega))$, $U(j\omega) = (U_1(j\omega), U_2(j\omega), \ldots, U_{18}(j\omega))$, and evaluating

$$G(j\omega) = Y(j\omega)U(j\omega)^\dagger, \tag{3.48}$$

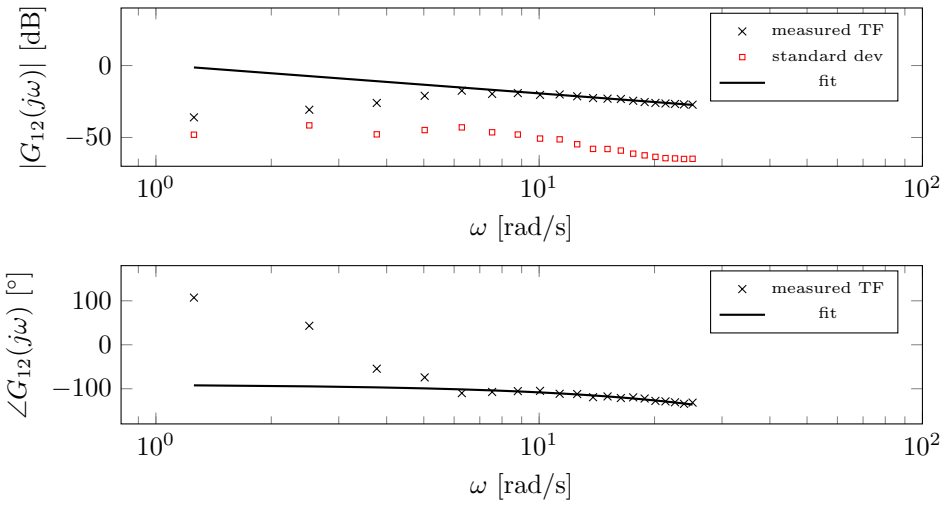


**Figure 3.15.** Estimated transfer function (black crosses) from the control flap 1 (bigger flap) of the first actuation unit to the angular rates $\omega_x$. The resulting fit of the simplified model is shown in black (solid line) and the estimate of the standard deviation is depicted in red (squares).

where † denotes the pseudo-inverse. Due to the fact that the input and output noise is correlated, the pseudo-inverse leads to small biases in the estimate of $G(j\omega)$ (dependent on the signal to noise ratios). However, even for a moderate signal to noise ratio of 6dB these biases are on the order of few percents (relative to the true $G(j\omega)$), [28, p. 46].

The resulting transfer functions from the inputs to the angular velocity and the velocity of the center of mass are depicted in Fig. 3.15 and Fig. 3.16 (blue dots). The variance of the transfer function is estimated via

$$\hat{\sigma}_G^2(j\omega) = \frac{1}{E(E-1)} \sum_{e=1}^{E} (U_e(j\omega)U_e(j\omega)^*)^{-1} \otimes (\hat{\sigma}_{YYe}^2(j\omega) - G(j\omega)\hat{\sigma}_{YUe}^2(j\omega)^* - \hat{\sigma}_{YUe}^2(j\omega)G(j\omega)^* + G(j\omega)\hat{\sigma}_{UUe}^2(j\omega)G(j\omega)^*), \quad (3.49)$$

where $E = 18$ refers to the number of experiments. Note that the variance $\hat{\sigma}_G(j\omega)$ has size $54 \times 54$ and refers to the variance of the vector $\mathrm{vec}(G(j\omega))$, where vec denotes vectorization.

### 5.1 Low-complexity model

We fit the parameters of the low-complexity model as derived in Sec. 3 to the measured frequency response. The parameters, denoted by $\theta$, are given by the matrices $T_1$, $T_2$, and the inertia $I_1$ and $I_3$. We denote the parametric transfer function corresponding to the dynamics (3.25) by $G_\theta(j\omega)$. In addition, the parametric transfer function is augmented with a delay modeling the sample and hold. The parameters $\theta$ are obtained by optimizing

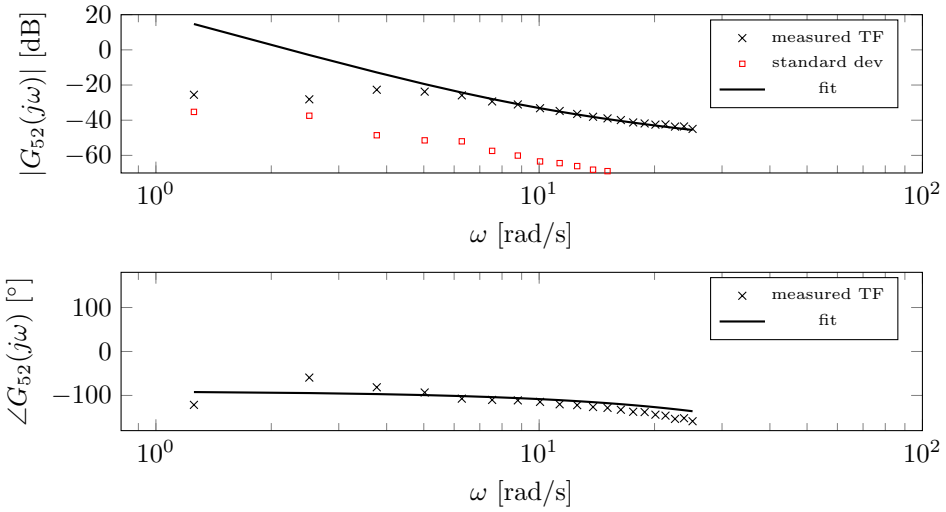


**Figure 3.16.** Estimated transfer function (black crosses) from the control flap 1 (bigger flap) of the first actuation unit to the velocity $v_y$. The resulting fit of the simplified model is shown in black (solid line) and the estimate of the standard deviation is depicted in red (squares).

the cost function

$$V(\theta) := \sum_{\omega \in \Omega} \mathrm{vec}(G_\theta(j\omega) - G(j\omega))^* (\hat{\sigma}_G^2(j\omega))^{-1} \mathrm{vec}(G_\theta(j\omega) - G(j\omega)), \tag{3.50}$$

where the set $\Omega$ is given by all frequencies that are excited by the excitation signal, that is, $\Omega := 2\pi \{0.2, 0.4, \ldots, 4\}$. Note that $V$ is formed by the squared distance of the matrix elements of $G_\theta$ from $G$, weighted with the variance $\hat{\sigma}_G^2$. If $G_\theta(j\omega)$ is assumed to be circularly-symmetric complex normally distributed with variance $\hat{\sigma}_G^2(j\omega)$, then (3.50) corresponds to the maximum likelihood cost function.

The cost is optimized using a quasi-Newton method, where the Jacobian and Hessian are obtained via numerical differentiation. An absolute tolerance of $10^{-8}$ of the optimizer $\theta$ is used as a stopping criterion. The resulting fit is exemplarily shown for the angular velocity $\omega_x$ and the linear velocity $v_y$ in Fig. 3.15, respectively Fig. 3.16. It can be concluded that the model explains well the frequencies above 1 Hz, but is not able to represent the lower frequencies accurately, which is most likely due to lack of aerodynamic effects in the model, as will be discussed in the following.

## 5.2 Augmented model

In order to explain the frequencies below 1 Hz the model is augmented to account for the following two effects, which were found to be dominant:

1) gyroscopic torques due to the fact that the ducted fans are all rotating in the same direction,

2) the redirection of a horizontal inlet airflow due to forward motion by the ducted

fan leading to so-called momentum drag, [1].

According to (3.16), the gyroscopic torques are given by the term

$$-\widetilde{\omega}\left(\Theta\omega + \sum_{i=1}^{3}\Theta_i\omega_i\right), \tag{3.51}$$

whose linearization about hover yields

$$3C\omega_{T_0}\widetilde{e}_z\omega, \tag{3.52}$$

with $\omega_{T_0}$ the angular velocity of a single ducted fan at hover.

The second effect stems from the fact that the airflow is redirected by the electric ducted fan and the outlet nozzle, leading to drag-like forces acting on the Flying Platform, see Fig. 3.17. This force is modeled to be proportional to the velocity at a certain point $P_i$ (to be determined by the measured data), [5],

$$F_{M_i} = -C_\alpha(v + \omega \times r_{Pi}), \quad i = 1, 2, 3, \tag{3.53}$$

$$C_\alpha := \mathrm{diag}(c_{\alpha_1}, c_{\alpha_1}, c_{\alpha_2}), \tag{3.54}$$

where $r_{Pi}$ denotes the vector from the center of gravity to the point $P_i$ and $c_{\alpha_1} > 0$, $c_{\alpha_2} > 0$ are two constants. The different constants $c_{\alpha_1}$ and $c_{\alpha_2}$ aim at modeling a potentially different behavior for horizontal and vertical motions. In addition, these forces induce a torque with respect to the center of gravity. As a result, due to the three-fold rotational symmetry of the fan configuration, the total force is modeled as

$$F_M := \sum_{i=1}^{3} F_{M_i} = -3C_\alpha v + 3l_{\alpha_1}C_\alpha\widetilde{e}_z\omega, \tag{3.55}$$

and the total torque (with respect to the center of gravity) is modeled as

$$M_M := -3l_{\alpha_2}C_\alpha\widetilde{e}_z v - 3L_\alpha C_\alpha\omega, \tag{3.56}$$

where

$$L_\alpha := \mathrm{diag}(l_{\alpha_3}, l_{\alpha_3}, l_{\alpha_4}), \tag{3.57}$$

and $c_\alpha$, $l_{\alpha_1}$, $l_{\alpha_2}$, $l_{\alpha_3}$, and $l_{\alpha_4}$ refer to different lengths describing the points $P_i$ and the lever arms of the forces $F_{M_i}$. Note that the constants $c_{\alpha_1}$ and $c_{\alpha_2}$ have units Ns/m, $l_{\alpha_1}$ and $l_{\alpha_2}$ have units m, and $l_{\alpha_3}$ and $l_{\alpha_4}$ have units $\mathrm{m}^2$.

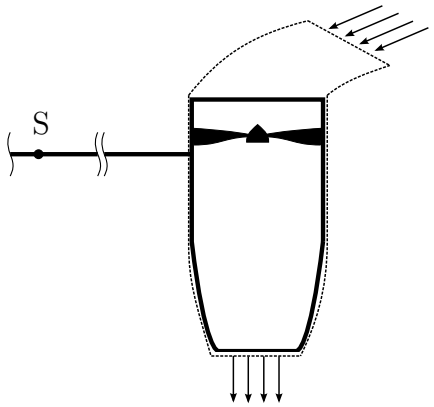


**Figure 3.17.** A single actuation unit with a body-fixed control volume (dashed line). The arrows refer to the inlet, respectively the outlet flow. An incoming airflow having a lateral component, which might stem from a translational motion is redirected by the electric ducted fan and the outlet nozzle, leading to a drag-like force acting on the Flying Platform (as can be seen from a momentum balance over the control volume).

Combining these three effects yields the augmented linear model

$$\dot{x}_a = A_a x_a + B_a(u - \bar{u}), \tag{3.58}$$

where $x_a := ({}^{\mathrm{J}}v, \alpha, \beta, \gamma, \omega)$, and

$$A_a := \begin{pmatrix} -3\frac{c_{\alpha_1}}{m} & 0 & 0 & 0 & g_0 & 0 & 0 & -3\frac{l_{\alpha_1}c_{\alpha_1}}{m} & 0 \\ 0 & -3\frac{c_{\alpha_1}}{m} & 0 & -g_0 & 0 & 0 & 3\frac{l_{\alpha_1}c_{\alpha_1}}{m} & 0 & 0 \\ 0 & 0 & -3\frac{c_{\alpha_2}}{m} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 3\frac{l_{\alpha_2}c_{\alpha_1}}{I_1} & 0 & 0 & 0 & 0 & -3\frac{l_{\alpha_3}c_{\alpha_1}}{I_1} & -3\frac{C\omega_{T_0}}{I_1} & 0 \\ -3\frac{l_{\alpha_2}c_{\alpha_1}}{I_1} & 0 & 0 & 0 & 0 & 0 & 3\frac{C\omega_{T_0}}{I_1} & -3\frac{l_{\alpha_3}c_{\alpha_1}}{I_1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3\frac{l_{\alpha_4}c_{\alpha_2}}{I_3} \end{pmatrix},$$

$$B_a := \begin{pmatrix} \frac{1}{m}T_{11} \\ \frac{1}{m}T_{12} \\ 0_{3\times 9} \\ \frac{1}{I_1}(l_3 J T_{11} + 2\sqrt{3}/3 l_1 V_1) \\ \frac{1}{I_3}T_{22} \end{pmatrix}.$$

The parameters $\theta_a$, describing the augmented parametric transfer function are given by $c_{\alpha_1}, c_{\alpha_2}, l_{\alpha_1}, l_{\alpha_2}, l_{\alpha_3}, l_{\alpha_4}, T_1, T_2, V_1$, and are found by optimizing (3.50) (with respect to the augmented model). The remaining parameters $m, l_1$, and $I_1$ are fixed to $m = 8\,\mathrm{kg}$, $l_1 = 10\,\mathrm{cm}$, and $I_1 = 0.07\,\mathrm{kg\,m^2}$ (a rough estimate from the CAD-model) to eliminate redundancies, and a delay accounting for the sample-and-hold is included. The resulting

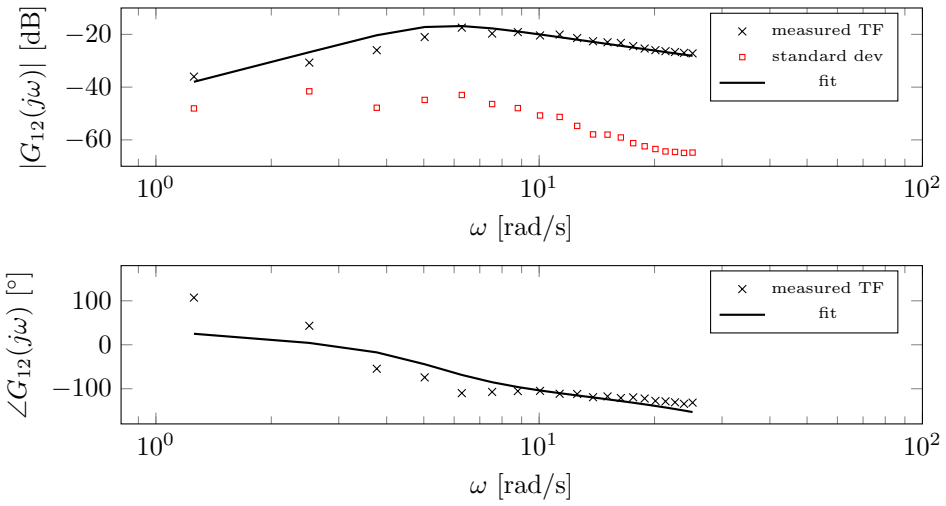


**Figure 3.18.** Estimated transfer function (black crosses) from the control flap 1 (larger flap) of actuation unit 1 to the angular rate $\omega_x$. The fit resulting from the augmented model is shown in black (solid line) and the standard deviation is indicated in red (squares).

fit is exemplarily shown for the angular velocity $\omega_x$ and the linear velocity $v_y$ in Fig. 3.18 and Fig. 3.19. Compared to the low-complexity model, the augmented model captures the behavior at frequencies below 1 Hz substantially better. By introducing the augmented model, the cost function $V(\theta)$ is decreased by roughly two orders of magnitude, which corresponds to a reduction of 99%. Most of the decrease can be attributed to introduction of the momentum drag, as the introduction of the gyroscopic effects leads to a decrease of the cost of roughly 1.3%.

We further investigated the sensitivity of the cost function with respect to shifts in the center of gravity, variations of the inertia, and misalignment of coordinate systems used for measuring ${}^{\mathrm{J}}v$ and $\omega$. To that extent, we analyzed the standard deviation of the cost function when sampling these parameter variations uniformly. The results are reported in Tab. 2. The cost is most sensitive to shifts in the center of gravity. However, even in case all effects are included, the cost alters by less than 3%, which is small especially when considering the number of additional degrees of freedom that these variations introduce. Thus, although a higher-order model might explain the data even better, we believe that the augmented model we presented yields a reasonable trade-off between model complexity and accuracy. The resulting numerical parameter values are listed in App. B. The full fit of the augmented model to the experimental data can be found on the first author's homepage[20].

We validated the model on a different dataset. The resulting time-domain fit of the angular velocities $\omega_x$ and $\omega_y$ is exemplarily depicted in Fig. 3.20.

[20] `http://www.idsc.ethz.ch/research-dandrea/people/person-detail.html?persid=156097`

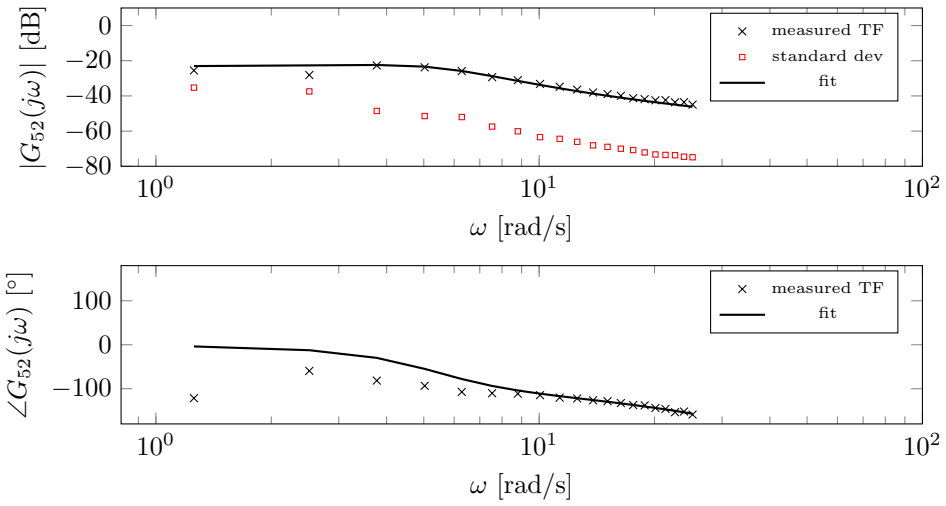


**Figure 3.19.** Estimated transfer function (black crosses) from the control flap 1 (larger flap) of actuation unit 1 to the velocity $v_y$. The fit resulting from the augmented model is shown in black (solid line) and the standard deviation is indicated in red (squares).

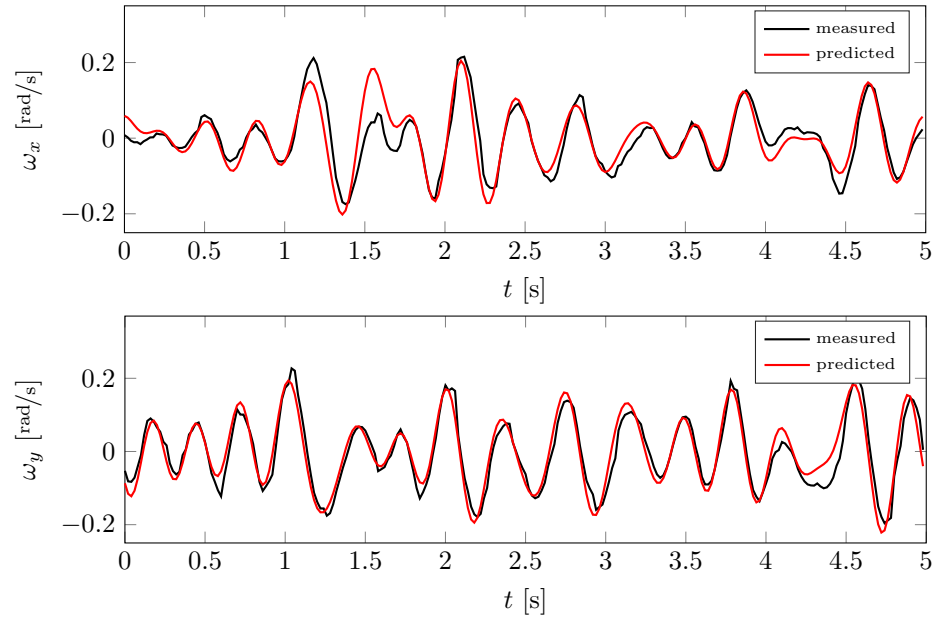


**Figure 3.20.** Validation of the augmented model. The Flying Platform is excited by a random phase multisine signal acting on the control flaps 1. The measurements are averaged over 8 periods to reduce the noise influence. The estimated standard deviation of the measurements is on the order of few percent and is therefore not shown.

## 6. Conclusion

This article presented the mechatronic design of the Flying Platform, an aerial vehicle whose purpose is to study ducted fan actuation. We discussed the mechanical design of a single actuation unit, including the control flap design to vector the thrust. The resulting

| parameter var. | std$[V(\theta)]$/E$[V(\theta)]$ | number samples |
|---|---|---|
| COG shift | 0.027 | $10^4$ |
| inertia | 0.0045 | $10^5$ |
| misalignment | 0.0085 | $10^5$ |
| all | 0.029 | $10^7$ |

**Table 2.** Sensitivity of the cost function $V(\theta)$ estimated from monte-carlo sampling. The parameter variations, that is, a shift in the center of gravity (COG shift), variations in the inertia (inertia), and a misalignment of coordinate systems (misalignment) are uniformly sampled, and the corresponding variation of the cost is quantified by the ratio between its standard deviation and its expected value. The shift in the center of gravity is restricted to a radius of 2 cm and the variations of the inertia are obtained by varying the diagonal elements of diag$(I_1, I_1, I_3)$ by 5% and rotating the resulting matrix along a uniformly sampled direction by an angle of less than 2° (also uniformly sampled). The misalignment of coordinate systems is characterized by rotations comprising a uniformly sampled direction and a uniformly sampled rotation angle of less than 2°.

thrust vectoring capabilities were characterized by static and dynamic measurements. A low-complexity rigid body model was introduced for control and analysis purposes. In particular, it was shown that the determinant of the controllability Gramian is a function of the ratio between lever arm and inertia. As a result, the mechanical design of the Flying Platform was chosen to maximize controllability. A linear control design was presented subsequently, which was shown to work reliably in practice. The quality of the model was assessed via a frequency domain system identification. It was shown that the low-complexity model captures roughly the frequencies above 1 Hz, but is unable to explain the lower frequencies. As a result, the model was extended to incorporate gyroscopic and aerodynamic effects, while keeping the model order fixed. The augmented model was found to roughly explain the measured transfer function from vectored thrusts to angular and linear velocities even at frequencies below 1 Hz.

It is hoped that the modeling and the measurement results presented throughout this article are useful for future aerial vehicle designs, and/or feasibility studies of aerial vehicles propelled by ducted fans.

Possible future work includes performing more aggressive maneuvers and evaluating advanced control algorithms, that account, for example, for input and state constraints, or incorporate the nonlinearities in the attitude dynamics. The ducted fan actuation, as presented in this paper, could be used for controlling aerial vehicles with lifting surfaces, thereby enabling efficient forward flight combined with high maneuverability, and vertical take-off and landing capabilities.

## A. Determinant of controllability Gramian

The system matrices in (3.28) are given by

$$A := \begin{pmatrix} 0 & g_0 J & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix}, \quad B := \begin{pmatrix} \frac{1}{m} T_{11} \\ 0 \\ \frac{l_3}{I_1} J T_{11} \end{pmatrix}. \tag{3.59}$$

The matrix exponential $e^{-At}$ yields therefore

$$e^{-At} = \begin{pmatrix} I & -g_0 t J & \frac{g_0 t^2}{2} J \\ 0 & I & -tI \\ 0 & 0 & I \end{pmatrix}, \tag{3.60}$$

leading to

$$e^{-At} B = \frac{l_3}{I_1} \begin{pmatrix} (\frac{I_1}{l_3 m} - \frac{g_0 t^2}{2}) T_{11} \\ -tJT_{11} \\ JT_{11} \end{pmatrix}, \tag{3.61}$$

and

$$e^{-At} B B^{\mathsf{T}} e^{-At} = \left(\frac{l_3}{I_1}\right)^2 \operatorname{diag}(I, J, J) \begin{pmatrix} \zeta^2 & -\zeta t & \zeta \\ -\zeta t & t^2 & -t \\ \zeta & -t & 1 \end{pmatrix} \otimes T_{11} T_{11}^{\mathsf{T}} \operatorname{diag}(I, J, J)^{\mathsf{T}}, \tag{3.62}$$

where $\zeta := I_1/(l_3 m) - g_0 t^2/2$. This yields

$$\det(W_c(T)) = \left(\frac{l_3}{I_1}\right)^{12} \det\left( \int_0^T \begin{pmatrix} \zeta^2 & -\zeta t & \zeta \\ -\zeta t & t^2 & -t \\ \zeta & -t & 1 \end{pmatrix} \mathrm{d}t \right)^2 \det(T_{11} T_{11}^{\mathsf{T}})^3, \tag{3.63}$$

where the fact that $\det(\operatorname{diag}(I, J, J)) = 1$ and the property $\det(X \otimes Y) = \det(X)^m \det(Y)^n$ of the Kronecker product has been used (where $X \in \mathbb{R}^{n \times n}$ and $Y \in \mathbb{R}^{m \times m}$). Moreover, $T_{11} T_{11}^{\mathsf{T}}$ simplifies to $3I$, and therefore, we obtain

$$\det(W_c(T)) = 9^3 \left(\frac{l_3}{I_1}\right)^{12} \det\left( \int_0^T \begin{pmatrix} \zeta^2 & -\zeta t & \zeta \\ -\zeta t & t^2 & -t \\ \zeta & -t & 1 \end{pmatrix} \mathrm{d}t \right)^2. \tag{3.64}$$

Using straightforward manipulations it can be shown that

$$\det\left(\int_0^T \begin{pmatrix} \zeta^2 & -\zeta t & \zeta \\ -\zeta t & t^2 & -t \\ \zeta & -t & 1 \end{pmatrix} \mathrm{d}t\right) = \frac{g_0^2}{8640} T^9, \tag{3.65}$$

which results in

$$\det(W_c(T)) = \frac{g_0^4 T^{18}}{102400} \left(\frac{l_3}{I_1}\right)^{12}. \tag{3.66}$$

## B. Parameter Values

The parameter values of the augmented model are listed below.

| | value | comment |
|---|---|---|
| $m$ | $8\,\mathrm{kg}$ | mass |
| $l_1$ | $0.1\,\mathrm{m}$ | lever arm of actuation |
| $l_3$ | $0.079\,\mathrm{m}$ | lever arm of actuation |
| $I_1$ | $0.07\ \mathrm{kg\,m^2}$ | inertia (roll, pitch, estimate) |
| $I_3$ | $0.11\ \mathrm{kg\,m^2}$ | inertia (yaw) |
| $C$ | $7 \cdot 10^{-6}\ \mathrm{kg\,m^2}$ | inertia of moving parts (motor and fan) |
| $c_{\alpha_1}$ | $2.388\,\mathrm{Ns/m}$ | drag force |
| $c_{\alpha_2}$ | $4.939\,\mathrm{Ns/m}$ | |
| $l_{\alpha_1}$ | $0.242\,\mathrm{m}$ | |
| $l_{\alpha_2}$ | $-0.138\,\mathrm{m}$ | drag force - lever arm |
| $l_{\alpha_3}$ | $-0.0084\,\mathrm{m^2}$ | |
| $l_{\alpha_4}/I_3$ | $0.362\,\mathrm{kg^{-1}}$ | |
| $C\omega_{T_0}$ | $0.018\,\mathrm{kg\,m^2/s}$ | gyroscopic effects |
| $T_d$ | $0.045\,\mathrm{s}$ | time delay |

**Table 3.** Scalar parameters of the augmented model. Note that the values of $I_1$ and $I_3$ are estimated using a CAD model. The value of $C$ is obtained by dividing $C\omega_{T_0}$ by a rough estimate of the fan velocity at hover (obtained from the datasheet of the manufacturer of the ducted fan).

The matrices $T_1$ and $T_2$ are given by

$$T_{11} = \begin{pmatrix} 0.6966 & -0.0311 & 0 & -0.3643 & -0.7165 & 0 & -0.3867 & 0.7286 & 0 \\ 0.0330 & 0.8656 & 0 & 0.6447 & -0.3826 & 0 & -0.6196 & -0.3775 & 0 \end{pmatrix},$$

$$T_{12} = \begin{pmatrix} 0 & 0 & 0.0967 & 0 & 0 & 0.0896 & 0 & 0 & 0.0670 \end{pmatrix},$$

$$V_1 = \begin{pmatrix} 0 & 0 & 0.0156 & 0 & 0 & -0.3433 & 0 & 0 & 0.3873 \\ 0 & 0 & 0.4450 & 0 & 0 & -0.2501 & 0 & 0 & -0.2068 \end{pmatrix},$$

$$\frac{1}{I_3} T_{22} = \begin{pmatrix} 0 & -0.7062 & -0.1536 & 0 & -0.6859 & -0.1030 & 0 & -0.7037 & -0.1076 \end{pmatrix}.$$

This leads to the following system matrices

$$A_a = \begin{pmatrix} -0.896 & 0 & 0 & 0 & 9.810 & 0 & 0 & -0.217 & 0 \\ 0 & -0.896 & 0 & -9.810 & 0 & 0 & 0.217 & 0 & 0 \\ 0 & 0 & -1.852 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -14.136 & 0 & 0 & 0 & 0 & 0.856 & -0.751 & 0 \\ 14.136 & 0 & 0 & 0 & 0 & 0 & 0.751 & 0.856 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5.366 \end{pmatrix}, \tag{3.67}$$

$$B_a = \begin{pmatrix} 0.087 & -0.004 & 0 & -0.046 & -0.090 & 0 & -0.048 & 0.091 & 0 \\ 0.004 & 0.108 & 0 & 0.081 & -0.048 & 0 & -0.077 & -0.047 & 0 \\ 0 & 0 & 0.097 & 0 & 0 & 0.090 & 0 & 0 & 0.067 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.037 & 0.980 & 0.026 & 0.730 & -0.433 & -0.566 & -0.702 & -0.427 & 0.639 \\ -0.789 & 0.035 & 0.734 & 0.413 & 0.811 & -0.413 & 0.438 & -0.825 & -0.341 \\ 0 & -0.706 & -0.154 & 0 & -0.686 & -0.103 & 0 & -0.704 & -0.108 \end{pmatrix}. \tag{3.68}$$

## Acknowledgement

This research was supported by the ETH-Grant ETH-48 15-1. We would like to thank Tobias Meier and Yeo Yih Tang for numerous contributions, for example to the flap design, the characterization of a single actuation unit, and the implementation of low-level hardware interfaces. We would like to express our gratitude towards Michael Egli and Marc-Andrè Corzillius for their contribution to the mechanical and electrical design. We would like to thank the Flying Machine Arena members for sharing the software and hardware infrastructure. A list of past and present participants is available at `http://flyingmachinearena.org/people/`.

## References

[1] G. Robson and R. D'Andrea, "Longitudinal stability of a jet-powered wingsuit", *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, 2010.

[2] J. G. Leishman, *Principles of Helicopter Aerodynamics*, second. Cambridge University Press, 2006.

[3] E. N. Johnson and M. A. Turbe, "Modeling, control, and flight testing of a small ducted-fan aircraft", *Journal of Guidance, Control, and Dynamics*, vol. 29, no. 4, pp. 769–779, 2006.

[4] L. Marconi and R. Naldi, "Control of aerial robots", *Control System Magazine*, vol. 32, no. 4, pp. 43–65, 2012.

[5] J.-M. Pfimlin, P. Binetti, P. Souères, T. Hamel, and D. Trouchet, "Modeling and attitude control analysis of a ducted-fan micro aerial vehicle", *Control Engineering Practice*, vol. 18, no. 3, pp. 209–218, 2010.

[6] J. M. Pflimlin, P. Souères, and T. Hamel, "Hovering flight stabilization in wind gusts for ducted fan UAV", *Proceedings of the 43rd Conference on Decision on Control*, pp. 3491–3496, 2004.

[7] R. Olfati-Saber, "Global configuration stabilization for the VTOL aircraft with strong input coupling", *IEEE Transactions on Automatic Control*, vol. 47, no. 11, pp. 1949–1952, 2002.

[8] R. A. Hess and M. Bakhtiari-Nejad, "Sliding mode control of a nonlinear ducted-fan UAV model", *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, pp. 748–762, 2006.

[9] R. Franz, M. Milam, and J. Hauser, "Applied receding horizon control of the caltech ducted fan", *Proceedings of the American Control Conference*, pp. 3735–3740, 2002.

[10] I. K. Peddle, T. Jones, and J. Treurnicht, "Practical near hover flight control of a ducted fan (SLADe)", *Control Engineering Practice*, vol. 17, no. 1, pp. 48–58, 2009.

[11] J. Fleming, T. Jones, W. Ng, P. Gelhausen, and D. Enns, "Improving control system effectiveness for ducted fan VTOL UAVs operating in crosswinds", *Proceedings of the 2nd AIAA "Unmanned Unlimited" System, Technologies and Operations-Aerospace Conference*, 2003.

[12] J. L. Pereira, "Hover and wind-tunnel testing of shrouded rotors for improved micro air vehicle design", PhD thesis, University of Maryland, 2008.

[13] A. Akturk and C. Camci, "Experimental and computational assessment of a ducted-fan rotor flow model", *Journal of Aircraft*, vol. 49, no. 3, pp. 885–897, 2012.

[14] V. Hrishikeshavan, J. Black, and I. Chopra, "Development of a quad shrouded rotor mircro air vehicle and performance evaluation in edgewise flow", *Proceedings of the American Helicopter Society Forum*, 2012.

[15] M. Miwa, S. Uemura, Y. Ishihara, A. Imamura, J.-h. Shim, and K. Ioi, "Evaluation of quad ducted-fan helicopter", *International Journal of Intelligient Unmanned Systems*, vol. 1, no. 2, pp. 187–198, 2013.

[16] A. Imamura, M. Miwa, and J. Hino, "Flight characteristics of quad rotor helicopter with thrust vectoring equipment", *Journal of Robotics and Mechatronics*, vol. 28, no. 3, pp. 334–342, 2016.

[17] P. G. Hamel and R. V. Jategaonkar, "Evolution of flight vehicle system identification", *Journal of Aircraft*, vol. 33, no. 1, 1996.

[18] B. Mettler, M. B. Tischler, and T. Kanade, "System identification of small-size unmanned helicopter dynamics", *Proceedings of the American Helicopter Society Forum*, 1999.

[19] A. Dorobantu, A. M. Murch, B. Mettler, and G. J. Balas, "Frequency domain system identification for a small, low-cost, fixed-wing UAV", *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, 2011.

[20] L. Derafa, T. Madani, and A. Benallegue, "Dynamic modelling and experimental identification of four rotors helicopter parameters", *Proceedings of the International Conference on Industrial Technology*, pp. 1834–1839, 2006.

[21] N. V. Hoffer, C. Coopmans, A. M. Jensen, and Y. Chen, "A survey and categorization of small low-cost unmanned aerial vehicle system identification", *Journal on Intelligent and Robotic Systems*, vol. 74, no. 1, pp. 129–145, 2014.

[22] K. W. Lewis, "The cumulative effects of roughness and reynolds number on NACA 0015 airfoil section characteristics", *Master's thesis in mechanical engineering, Texas Tech University*, 1984.

[23] K. Freudenreich, K. Kaiser, A. Schaffarczyk, H. Winkler, and B. Stahl, "Reynolds number and roughness effects on thick airfoils for wind turbines", *Wind Engineering*, vol. 28, no. 5, pp. 529–546, 2004.

[24] *PX4 flight management unit*, `https://pixhawk.org/modules/px4fmu`, Accessed: July 2016.

[25] F. Pfeiffer and C. Glocker, *Multibody Dynamics with Unilateral Contacts*. Wiley-VCH, 2004.

[26] F. M. Callier and C. A. Desoer, *Linear System Theory*. Springer Science + Business Media, 1991.

[27] S. Lupashin, M. Hehn, M. W. Mueller, A. P. Schoellig, M. Sherback, and R. D'Andrea, "A platform for aerial robotics research and demonstration: The Flying Machine Arena", *Mechatronics*, vol. 24, no. 1, pp. 41–54, 2014.

[28] R. Pintelon and J. Schoukens, *System Identification: A Frequency Domain Approach*, second. John Wiley & Sons, 2012.

# Part C

## Approximations of the constrained linear quadratic regulator problem

# Paper P4

# On the Approximation of Constrained Linear Quadratic Regulator Problems and their Application to Model Predictive Control

Michael Muehlebach and Raffaello D'Andrea

**Abstract**

This article is concerned with the approximation of constrained continuous-time linear quadratic regulator problems, which are, for example, encountered in model predictive control. By representing input and state trajectories using basis functions, the underlying infinite-dimensional optimal control problems are reduced to convex finite-dimensional optimization problems that can be solved efficiently. The article quantifies the suboptimality and establishes convergence of the obtained approximations. The results are applied in the context of model predictive control. In particular, it will be shown that the truncation of the prediction horizon can be avoided, leading to recursive feasibility and closed-loop stability guarantees. The resulting finite-dimensional convex optimization problems typically include semi-infinite constraints. Several strategies to handle these constraints are discussed. The approach is shown to be numerically efficient. It is shown to outperform state-of-the-art model predictive control algorithms on a quadruple integrator system, without necessarily degrading closed-loop performance.

Submitted to *Automatica*, September 2017.

## 1. Introduction

Model predictive control (MPC) has become a well-known and widely used control strategy for solving challenging control problems. Unlike many other approaches, MPC addresses input and state constraints in a systematic way. It is based on repeatedly solving an optimal control problem, including the actual state as an initial condition and a prediction of the system's evolution. This leads naturally to an implicit feedback law, providing robustness against modeling errors and disturbances, [1].

Due to the fact that an optimal control problem has to be solved at each sampling interval, online MPC, where the optimization is solved online, is computationally demanding. Thus, in order to render MPC computationally tractable, the underlying optimal control problem is simplified, typically by discretizing the dynamics and truncating the prediction horizon.

Herein, we propose an alternative approach that relies on a parametrization of input and state trajectories using basis functions. A Galerkin method is used to formulate the dynamics as an equality constraint relating the parameter vectors describing the input and the state. We will show on an example that this parametrized approach leads to different trade-offs between computational effort and achieved closed-loop cost. Moreover, the basis functions can be used to encode a priori knowledge of the system's dynamics (e.g. different time scales), and even provide a means to retain an infinite prediction horizon. We will show that this leads to an MPC algorithm with inherent recursive feasibility and closed-loop stability guarantees. This contrasts the discrete-time approach where stability is often imposed indirectly using a combination of a terminal cost and a terminal set constraint. The proposed parametrized MPC approach is benchmarked against state-of-the-art discrete-time MPC strategies for underlining the numerical efficiency of the parametrized approach. Simulations of a quadruple integrator system are presented, where it is shown that the average execution time for achieving a given closed-loop cost can be reduced by roughly one order of magnitude. This is particularly interesting for embedded systems with fast dynamics, where fast sampling is required but where the available computational power is limited.

*Outline of the paper:* The paper is divided into three parts. The first part is concerned with the approximation of constrained continuous-time linear quadratic regulator problems. The discussion is not restricted to infinite-horizon problems, which are often encountered in MPC, but also includes finite-horizon problems, with terminal costs and/or terminal state constraints. By approximating input and state trajectories using basis functions and exploiting duality, two different finite-dimensional optimization problems are derived, whose optimal costs yield upper and lower bounds on the cost of the underlying infinite-dimensional problem. By increasing the basis functions' complexity (which might correspond to the polynomial order, for example), the resulting optimal costs are found to yield monotonic sequences approximating the underlying optimal control problem from above and below. This makes it possible to quantify the suboptimality of the obtained approximation, and, as we will show in the remainder, also bounds the

$L^2$-distance from the optimal trajectories. Moreover, conditions guaranteeing convergence to the underlying optimal control problem will be established, thereby providing a theoretical justification of the proposed approach.

The second part deals with the application of the proposed approach to MPC. By choosing exponentially decaying basis functions, a truncation of the prediction horizon can be avoided, and as a result, closed-loop stability and recursive feasibility are shown to be inherent to the resulting parametrized MPC formulation.

The third part is concerned with numerical solution routines for solving the resulting finite-dimensional optimization problems. Due to the fact that the parametrization is done in continuous time, input and state constraints, which are enforced over a certain time interval and not only at a finite number of time instants, lead to semi-infinite constraints. These constraints describe convex sets that are not necessarily polytopic. To that extent, several strategies for handling these semi-infinite constraints are presented. In particular, a dedicated active-set approach is proposed, which is shown to perform well in the numerical experiments that are conducted subsequently. Moreover, we show that this active-set method indeed converges and derive an upper bound on the number of iterations the method takes to achieve a given tolerance.

*Related work:* For our analysis of the approximations to constrained continuous-time linear quadratic regulator problems we adopt a similar point of view than presented in [2], where (weighted) Sobolev spaces are introduced as state space and (weighted) Lebesgue spaces are introduced as control spaces. By doing so, the author establishes a Pontryagin type of Maximum Principle for linear infinite-horizon optimal control problems. These problems have proven to be difficult to analyze as the standard transversality conditions cannot be extended directly to the infinite-horizon case, see for example [3] or [4, Ch. 3.7, Ch. 6.5].

In [5], polynomials are used for approximating continuous linear programs.[21] Duality is exploited for constructing approximations yielding upper and lower bounds on the underlying continuous linear program. The resulting semi-infinite constraints are reformulated using sum-of-squares techniques yielding semidefinite programs. Our approach is similar in the sense that the lower bounds are also derived using duality. However, the optimal control problem that we consider cannot be cast as a continuous linear program, and as a result, our approach for constructing the lower bounds differs significantly. Moreover, we do not restrict ourselves to polynomial basis functions, and treat equality constraints in the form of linear ordinary differential equations by means of a Galerkin approach.

The optimal control problems that are discussed in the following can also be approximated by "standard" numerical optimization approaches such as shooting or collocation methods, see for example [6] or [7] and references therein. However, these approaches are typically tailored to nonlinear problems and as such, do not yield guarantees on the approximation quality in general. Moreover, these approaches tend to be computationally

[21] Continuous linear programs are related to the constrained linear quadratic regulator problems considered herein by the fact that the constraints occurring in continuous linear programs could be used to encode linear dynamics.

expensive.

Constrained linear quadratic infinite-horizon problems are often encountered in MPC, which is the main motivation for our work. In contrast to our formulation, the "standard" MPC approach relies mostly on a discrete-time finite-horizon formulation, [1]. In order to guarantee closed-loop stability, terminal cost and terminal state constraints are often needed, [8]. Moreover, as remarked in [9], truncating the prediction horizon leads to a discrepancy between the closed-loop performance objective and the finite-horizon open-loop performance objective that is minimized at every time step. An alternative approach is proposed in [10, Ch. 3, Ch. 6] and [11], where the finite differences (in the discrete-time setting), respectively the time-derivatives (in the continuous-time setting) of the control inputs are described with so-called Laguerre or Kauz basis functions. An analytical expression for the corresponding state trajectory as a function of the parametrized inputs is derived, eliminating thereby the state variables in the resulting optimization problem. Still, a finite prediction horizon is retained. In contrast, by choosing exponentially decaying basis function for parametrizing input and state trajectories, our MPC formulation avoids the truncation of the prediction horizon. As a consequence, we will show that recursive feasibility and closed-loop stability are guaranteed, provided that the resulting optimization problem is feasible at time $t = 0$. If the basis functions are well-chosen, only few basis functions are needed for obtaining a relatively good approximation. This leads to optimization problems with relatively few optimization variables that can be solved efficiently. Compared to state-of-the-art MPC solvers [12] and [13], the average execution time can be reduced by roughly one order of magnitude for the quadruple integrator system presented in Sec. 5, without degrading closed-loop performance. Compared to the approach presented in [11], we parametrize the control inputs directly, which avoids lifting the system, and consequently reduces the number of variables. In our approach, the dynamics are represented by linear equality constraints, which may or may not be eliminated. In our experience, the numerical optimization routines tend to be more effective if the equality constraints are not eliminated.[22]

The suboptimality of the "standard" MPC without terminal constraints and terminal cost with respect to the underlying infinite-horizon problem is discussed and quantified in [16, Ch. 6], [17], [18], and [19]. These approaches are based on approximate dynamic programming and typically involve finding so-called control Lyapunov functions. The approach presented in the following is constructive in the sense that the suboptimality can be quantified by solving two finite-dimensional convex optimization problems.

Due to the fact that an infinite prediction horizon can be retained, the optimal open-loop cost of our approach always acts as an upper bound on the achieved closed-loop cost. Similarly, it is shown in [20] that the infinite-horizon closed-loop cost can be upper bounded by a corresponding discrete-time receding-horizon scheme including a terminal cost and a terminal state constraint. This leads naturally to stability guarantees and constraint satisfaction for all times (even between the sampling instants).

[22]This observation is well-known in the literature, see e.g. [14, p. 522], [15, p. 455].

The authors from [21] and [22] propose to solve the discrete-time infinite-horizon linear quadratic regulator problem directly, by means of an operator splitting technique in [21] or by successively extending the prediction horizon of a finite-horizon approximation in [22]. These schemes require successive solutions of the discrete-time finite-horizon problem with varying prediction horizons. This contrasts the proposed approach, where the number of optimization variables is fixed and can be adjusted for trading off the computational complexity with the approximation quality.

Exploiting a parametrization of the input for reducing the computational complexity of MPC has already been explored by previous work, see for example [23], [24], [25], and [26]. In [23], the implications regarding closed-loop stability and recursive feasibility of an input parametrization are investigated in the context of nonlinear MPC. The input parametrization is required to be invariant to time shifts, which parallels the approach presented in the following. The formulation is based on a finite prediction horizon, and a terminal equality constraint (if the prediction horizon remains fixed) or a contraction property (if the prediction horizon enters the optimization) is required for guaranteeing closed-loop stability. In case the prediction horizon is fixed, the input parametrization is assumed to be translatable, see [23, Def. 1.5], which results either in a standard sample-and-hold parametrization, a nonlinear input parametrization, or requires additional assumptions compared to the parametrization presented in the following. In case the contraction property in combination with a varying prediction horizon is used for guaranteeing closed-loop stability, a nonlinear and in general non-convex optimization problem is obtained. We show that our parametrization evolves naturally from a time-shift requirement related to closed-loop stability and the fact that the open-loop trajectories should achieve a finite cost. In contrast to [23], the infinite-horizon formulation avoids the use of additional equality constraints for guaranteeing closed-loop stability and leads to a finite-dimensional convex optimization problem with a quadratic cost and linear constraints.[23]

In [27], multiresolution analysis is used for parametrizing the input trajectory. However, the approach is mainly applicable to open-loop stable systems, where the impulse response is assumed to be negligible after a certain time horizon. For dealing with unstable systems, the proposed approach would require additional terminal constraints on the unstable modes. Similarly, the authors from [28] apply the wavelet transformation for simplifying the control laws obtained with explicit model predictive control. They show that the resulting simplified control law is everywhere feasible and quantify the suboptimality.

The problem of imposing semi-infinite constraints (in our case due to the fact that we require input and state constraints to be fulfilled for a certain time interval) has been extensively studied in the literature. In [29] and [30], a stochastic constraint sampling

[23] In the following, a continuous-time point of view is adopted, resulting in semi-infinite constraints due to the fact that the constraints are imposed over compact time intervals. These semi-infinite constraints can be avoided in a discrete-time setting. Moreover, we will present a computationally efficient approach to deal with semi-infinite constraints.

approach is presented. The authors provide bounds on the probability that constraint violations occur, when solving the problem with sampled constraints. Alternative approaches include relaxation techniques, for example based on sum-of-squares programming, as presented in [31]. The exactness of certain relaxations is established in [32]. We will discuss several possibilities for handling these semi-infinite constraints and introduce a dedicated active-set method. Unlike stochastic constraint sampling strategies, our approach is deterministic, and does not require the solution of semidefinite programming problems. We will establish convergence of our approach and derive an explicit bound on the number of iterations needed to achieve a given tolerance.

Preliminary results to the ones presented herein appeared in the conference papers [33] and [34]. In [33], the application of our strategy to MPC and the implications regarding closed-loop stability and recursive feasibility are discussed in detail. In [34], the approximation quality with respect to the underlying optimal control problem is discussed. The results from [34] are extended herein by deriving a bound on the approximation error of the resulting optimal input and state trajectories in the $L^2$-norm. We state conditions guaranteeing that our approximations will actually converge to the solutions of the underlying optimal control problem, when increasing the basis functions complexity. Moreover, compared to [34], the results presented herein are derived using a different approach (for example not relying on conjugate functions), which we think is more accessible. Compared to earlier work, we also discuss in detail how to solve the resulting optimization problems that typically include semi-infinite constraints. We propose and analyze a dedicated active-set method, which we benchmark against a state-of-the-art MPC solver.

## 2. Part I: Theoretical Foundation

### 2.1 Problem Formulation

In a first step, we present and analyze approximations to the following optimal control problem

$$\begin{aligned} J_\infty := \min \; & \frac{1}{2}\left(||x||^2 + ||u||^2\right) + \psi(x_T) \\ \text{s.t.} \; & \dot{x} = Ax + Bu, x(0) = x_0, x(T) = x_T, \\ & C_{\mathrm{x}} x + C_{\mathrm{u}} u \leq b, x_T \in \mathcal{X}, \\ & x \in L_n^2, u \in L_m^2, \end{aligned} \tag{4.1}$$

where the space of square integrable functions mapping from the interval $I := (0, T)$ to $\mathbb{R}^q$ is denoted by $L_q^2$, where $q$ is a positive integer; and the $L_q^2$-norm is defined as the map

$L_q^2 \to \mathbb{R}$,

$$x \to ||x||, \quad ||x||^2 := \int_I x^\mathsf{T} x \, \mathrm{d}t, \tag{4.2}$$

with $\mathrm{d}t$ the Lebesgue measure. The function $\psi : \mathbb{R}^n \to \mathbb{R}$ is assumed to be positive definite and strongly convex, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C_\mathrm{x} \in \mathbb{R}^{n_c\times n}$, $C_\mathrm{u} \in \mathbb{R}^{n_c\times m}$, and $b \in \mathbb{R}^{n_c}$ are constant, and the set $\mathcal{X}$ is closed and convex. The dynamics as well as the stage constraints are assumed to be fulfilled almost everywhere. Thus we simply write

$$f = g, \quad f \leq g, \tag{4.3}$$

when we mean $f(t) = g(t)$, respectively $f(t) \leq g(t)$ for all $t \in I$ almost everywhere, with $f, g \in L_q^2$, or equivalently,

$$\int_I \delta p^\mathsf{T}(f - g)\mathrm{d}t = 0, \quad \int_I \delta \hat{p}^\mathsf{T}(f - g)\mathrm{d}t \leq 0, \tag{4.4}$$

for all smooth compactly supported test functions $\delta p$ and $\delta \hat{p}$, with $\delta \hat{p}(t) \geq 0$ for all $t \in I$. The weak derivative of $x$ is denoted by $\dot{x}$.[24] To simplify notation we abbreviate the domain of the objective function by

$$X := L_n^2 \times L_m^2 \times \mathbb{R}^n. \tag{4.5}$$

We assume throughout the article that the constraints in (4.1) are nonempty, i.e. there exist trajectories $x$ and $u$, fulfilling the dynamics, the initial condition, the constraints, and thus achieve a finite cost.

The main motivation for studying problem (4.1) comes from the fact that (4.1) often serves as a starting point for MPC.

*Discussion of the assumptions:* The assumption of linear time-invariant dynamics will be important in the following, as it leads to approximate solutions of (4.1) that fulfill the equations of motion exactly. The assumption of a quadratic cost could be relaxed to include strongly convex running costs; we will comment on such extensions in due course. However, these extensions will generally increase the computational complexity needed for obtaining (approximate) numerical solutions. From a practical point of view, a quadratic running cost often represents a good compromise between generality and computational tractability.

The interval $I$ is not restricted to have finite measure. The subsequent analysis remains valid even if $T \to \infty$, with $\psi = 0$, $\mathcal{X} = \{0\}$, and $x_T = \lim_{t\to\infty} x(t) = 0$. Furthermore, the more general cost

$$\frac{1}{2}\int_I x^\mathsf{T} Q x + u^\mathsf{T} R u \, \mathrm{d}t \tag{4.6}$$

[24] The equations of motion imply that $\dot{x} \in L_n^2$, which can be used to conclude that $x$ has a unique absolutely continuous representative defined on the closure of $I$ (a classical solution of the equations of motion). With $x(t)$ we refer to the value this unique absolutely continuous representative takes at time $t \in [0, T]$.

can be cast into (4.1) by means of a linear coordinate transformation, provided that the matrices $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are positive definite. In addition, the $L_q^2$ space can be replaced by the weighted Lebesgue space with squared norm

$$\int_I x^\mathsf{T} x \; e^{-\alpha t} \; \mathrm{d}t, \tag{4.7}$$

where $\alpha > 0$ is constant, without changing the subsequent analysis. The assumption that $b$ is constant can be relaxed, for example by requiring $b$ to be square integrable.

The strong convexity of the running cost and the terminal cost $\psi$ is important for guaranteeing uniqueness of the corresponding minimizer. This will be established in Sec. 2.5, where we also argue that the minimum in (4.1) is indeed attained.

### 2.2 Motivation

In the remainder we will construct two series of finite-dimensional approximations to (4.1) yielding upper and lower bounds on $J_\infty$. We will do this in three steps; 1) parametrization of input and state trajectories using basis functions, 2) approximation of the dynamics, 3) approximation of the constraints. We will show that the upper and lower bounds will get tighter and tighter as more and more basis functions are included in the approximation. Furthermore, we provide conditions guaranteeing convergence to $J_\infty$ as the number of basis functions tends to infinity. As the following derivations and proofs are fairly technical, we would like to convey and motivate the underlying concepts using the following simple example

$$\min_{z\in\mathcal{Z}} |z|^2, \tag{4.8}$$

where $\mathcal{Z}$ is taken as the closed convex set

$$\mathcal{Z} := \{z \in \mathbb{R}^2 \mid z_1 \geq 1, z_2 \geq 1\}, \tag{4.9}$$

and $|\cdot|$ denotes the Euclidean norm. Clearly, an upper bound on (4.8) is obtained by restricting the variable $z$ to a subspace of $\mathbb{R}^2$, for instance $z_1 + z_2 = 3$. This results in

$$\min_{z\in\mathcal{Z}_\mathrm{U}} |z|^2 \geq \min_{z\in\mathcal{Z}} |z|^2, \tag{4.10}$$

with

$$\mathcal{Z}_\mathrm{U} := \{z \in \mathbb{R}^2 \mid 1 \leq z_1 \leq 2, z_1 + z_2 = 3\} \subset \mathcal{Z}. \tag{4.11}$$

We will follow exactly the same strategy to obtain upper bounds on the cost $J_\infty$ in (4.1), that is, we simply restrict the trajectories $x$ and $u$ to be spanned by a fixed (and finite) number of basis functions. We therefore restrict the trajectories $x$ and $u$ to a linear

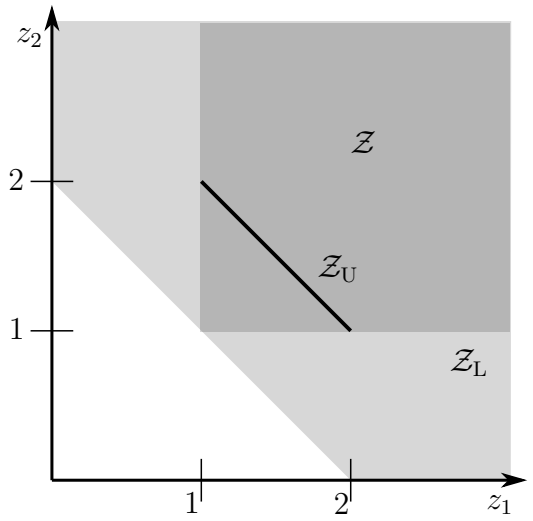


**Figure 4.1.** Illustration of the sets $\mathcal{Z}$, $\mathcal{Z}_\mathrm{U}$, and $\mathcal{Z}_\mathrm{L}$. The sets $\mathcal{Z}$, $\mathcal{Z}_\mathrm{U}$, and $\mathcal{Z}_\mathrm{L}$ are all closed and convex.

subspace. If this subspace is dense, the optimal cost of the underlying problem might be recovered.

Moreover, the constraints in (4.8) form a closed convex set and can be rewritten as the intersection of all closed half-spaces that pass through the point $(1, 1)$ and contain $\mathcal{Z}$. More precisely, a vector $z$ is an element of $\mathcal{Z}$ if and only if

$$\delta z_1(z_1 - 1) + \delta z_2(z_2 - 1) \geq 0, \quad \forall \delta z_1 \geq 0, \delta z_2 \geq 0 \tag{4.12}$$

holds. Thus, by restricting the variations $\delta z_1$ and $\delta z_2$ to lie in a subspace, for instance $\delta z_1 = \delta z_2$, we obtain an approximation to the set $\mathcal{Z}$, that is,

$$\mathcal{Z}_\mathrm{L} := \{z \in \mathbb{R}^2 \mid \delta z_1(z_1 - 1) + \delta z_2(z_2 - 1) \geq 0, \forall \delta z \in \mathbb{R}^2 : \delta z \geq 0, \delta z_1 = \delta z_2\}. \tag{4.13}$$

As a result, $\mathcal{Z}$ is a subset of $\mathcal{Z}_\mathrm{L}$, simply because compared to (4.12), fewer variations are allowed in (4.13). This leads naturally to a lower bound on (4.8), which is obtained by optimizing over $\mathcal{Z}_\mathrm{L}$ instead of $\mathcal{Z}$. Again, if the subspace to which the variations are restricted is dense, the original problem might be recovered. Moreover, (4.13) is in general still an intersection of closed half-spaces and is therefore closed. The sets $\mathcal{Z}$, $\mathcal{Z}_\mathrm{U}$, and $\mathcal{Z}_\mathrm{L}$ are illustrated in Fig. 4.1. The same strategy is applied in the following: The feasible set of problem (4.1) will be described using a variational formulation, that is, as the intersection of half-spaces in a Hilbert space. Restricting the variations will lead to a finite-dimensional optimization problem, approximating (4.1) from below.

### 2.3 Finite-dimensional approximations of (4.1)

*1) Parametrization with basis functions* We will parametrize input and state trajectories using basis functions, that is,

$$\tilde{x}(t) = (I_n \otimes \tau^s(t))^\mathsf{T} \eta_x, \quad \tilde{u}(t) = (I_m \otimes \tau^s(t))^\mathsf{T} \eta_u,$$

where $\eta_x \in \mathbb{R}^{ns}$ and $\eta_u \in \mathbb{R}^{ms}$ are the parameter vectors, $\tau^s(t) := (\tau_1(t), \tau_2(t), \ldots, \tau_s(t)) \in \mathbb{R}^s$ contains the first $s$ basis functions, $\otimes$ denotes the Kronecker product, and $I_q \in \mathbb{R}^{q\times q}$ refers to the identity matrix for any integer $q > 0$. The superscript $s$ refers to the number of basis functions used for the approximation. For ease of notation the superscript $s$ will be dropped, whenever it is clear from context, and we will indicate vectors as $n$-tuples, where the dimension and stacking can be inferred from context. The basis functions are required to satisfy the following assumptions:

A1) The basis functions $\tau_i \in L_1^2$, $i = 1, 2, \ldots, s$ are linearly independent and orthonormal with respect to the standard $L_1^2$-scalar product.

A2) The basis functions fulfill $\dot{\tau}^s(t) = M_s \tau^s(t)$ for all $t \in I$, for some $M_s \in \mathbb{R}^{s\times s}$.

Note that in case $I$ has infinite measure, $M_s$ is required to have strictly negative eigenvalues. This is a natural requirement, since a feasible state trajectory $x$ in (4.1) is guaranteed to decay due to the fact that it is required to be square integrable and to have a weak derivative in $L_n^2$, see [35, Cor. 8.9].

The assumption of linearly independent basis functions is necessary for the approximations to be unique. The assumption of orthonormal basis functions is without loss of generality, since orthonormal basis functions can be constructed from linearly independent ones via the Gram-Schmidt procedure. Assumption A2 is more restrictive. Well-known examples fulfilling Assumption A2 are sinusoids or polynomials. Assumption A2 implies, however, that the basis functions are able to capture an arbitrary time-shift, and can be used to conclude that the equations of motion are (depending on the formulation) fulfilled exactly by the parametrized input and state trajectories, [33]. If the basis functions are assumed to be continuously differentiable, the converse is also true, as we illustrate next. Thus, we set forth that the basis functions should be able to capture time-shifts, that is, for every vector $\eta \in \mathbb{R}^s$ and every time-shift $T_\mathrm{s} \geq 0$ there exists a vector $\hat{\eta}(\eta, T_\mathrm{s})$ such that

$$\tau(t - T_\mathrm{s})^\mathsf{T} \eta = \tau(t)^\mathsf{T} \hat{\eta}(\eta, T_\mathrm{s}), \quad \forall t \in (T_\mathrm{s}, T). \tag{4.14}$$

We will now show that this implies that the basis functions must fulfill Assumption A2. In order to do so, we take the derivative with respect to $T_\mathrm{s}$ and evaluate the resulting expression for $T_\mathrm{s} \to 0$, leading to

$$-\dot{\tau}(t)^\mathsf{T} \eta = \tau(t)^\mathsf{T} \left.\frac{\partial \hat{\eta}}{\partial T_\mathrm{s}}\right|_{T_\mathrm{s}\downarrow 0}, \quad \forall t \in I. \tag{4.15}$$

We may choose the canonical unit vectors for $\eta$, which readily implies that $\dot{\tau}$ must be a linear combination of the vector $\tau$. This concludes that any set of continuously differentiable basis functions that can capture an arbitrary time-shift must fulfill Assumption A2. In the context of MPC, Assumption A2 is used to guarantee recursive feasibility and closed loop stability and will therefore be of paramount importance. In a discrete-time finite-horizon setting, the importance of the time-shift property regarding closed-loop stability has already been emphasized in [23], resulting in similar requirements on the basis functions, see for example [23, Def. 1.8].

In the infinite-horizon case, that is for $T \to \infty$, examples fulfilling Assumptions A1 and A2 are given by exponentially decaying polynomials, or exponentially decaying sinusoids. In the case of polynomials, this leads to so-called Laguerre functions, which are given by

$$\tau_i(t) = \sqrt{2\nu} L_i(2\nu t) e^{-\nu t}, \tag{4.16}$$

where $L_i$ denotes the $i$th Laguerre polynomial, $i = 1, 2, \ldots, s$, and $\nu > 0$ is the rate of the exponential decay. The corresponding matrix $M_s$ has then the form

$$M_s = \begin{pmatrix} -\nu & 0 & 0 & \ldots \\ -2\nu & -\nu & 0 & \ldots \\ -2\nu & -2\nu & -\nu & \ldots \\ \vdots & & & \ddots \end{pmatrix}. \tag{4.17}$$

These basis functions will be used to approximate infinite-horizon problems arising in the context of MPC, see Sec. 3 and Sec. 5.

We will denote the finite-dimensional subspace spanned by the first $s$ basis functions as $X^s$,

$$X^s := \{(x, u, x_T) \in X \mid \eta_x \in \mathbb{R}^{ns}, \eta_u \in \mathbb{R}^{ms}, \\ x = (I_n \otimes \tau^s)^\mathsf{T} \eta_x, u = (I_m \otimes \tau^s)^\mathsf{T} \eta_u\}. \tag{4.18}$$

The fact that $X^s$ is finite-dimensional can be used to conclude that $X^s$ is complete, i.e. that every Cauchy sequence in $X^s$ converges and has its limit in $X^s$. As a result, it follows that $X^s$ is a closed subspace of $X$.[25] This will become important for arguing that the minima of the resulting optimization problems are indeed attained. In addition, the following straightforward, but important relation

$$X^s \subset X^{s+1} \tag{4.19}$$

holds for all integers $s > 0$.

[25] In a metric space a set is closed if and only if it is sequentially closed.

We can think of an element in $X^s$ not only as an element in $X$ (i.e. a tuple of a finite-dimensional vector and two square integrable functions), but also as a finite-dimensional vector given by the corresponding parameter vectors $\eta_x$ and $\eta_u$. To make this distinction explicit, we introduce the map $\pi^{qs} : L_q^2 \to \mathbb{R}^{qs}$, defined as

$$x \to \int_I (I_q \otimes \tau^s) x \mathrm{d}t, \tag{4.20}$$

which maps an arbitrary element $x \in L_n^2$ to its first $s$ basis function coefficients. Similarly, we define $\pi^s : X \to \mathbb{R}^{ns} \times \mathbb{R}^{ms} \times \mathbb{R}^n$ as

$$(x, u, x_T) \to (\pi^{ns}(x), \pi^{ms}(u), x_T). \tag{4.21}$$

As a consequence, we write $\pi^s(x)$ for describing the finite dimensional representation of $x \in X^s$, that is, its representation in terms of the parameter vectors $\eta_x$ and $\eta_u$. The adjoint map $(\pi^{qs})^* : \mathbb{R}^{qs} \to L_q^2$ is given by

$$\eta \to (I_n \otimes \tau)^{\mathsf{T}} \eta, \tag{4.22}$$

and is used to obtain the trajectory corresponding to the vector $\eta \in \mathbb{R}^{qs}$, containing the first $s$ basis function coefficients. Similarly, we define $(\pi^s)^* : \mathbb{R}^{ns} \times \mathbb{R}^{ms} \times \mathbb{R}^n \to X$ as

$$(\eta_x, \eta_u, x_T) \to ((\pi^{ns})^*(\eta_x), (\pi^{ms})^*(\eta_u), x_T). \tag{4.23}$$

The composition $(\pi^s)^* \pi^s : X \to X$ yields the projection of an element $x \in X$ onto the subspace $X^s \subset X$.

*2) Approximation of the constraints* In the following we seek to approximate the constraint

$$\mathcal{C} := \{(x, u, x_T) \in X \mid C_{\mathrm{x}} x + C_{\mathrm{u}} u \leq b, x_T \in \mathcal{X}\}, \tag{4.24}$$

characterizing all square integrable functions $x$ and $u$, satisfying the inequality constraints. The first approximation is obtained by restricting the trajectories $x$ and $u$ to be spanned by the first $s$ basis functions, i.e.

$$\mathcal{C}_{\mathrm{U}}^s := \{(\tilde{x}, \tilde{u}, x_T) \in X^s \mid C_{\mathrm{x}} \tilde{x} + C_{\mathrm{u}} \tilde{u} \leq b, x_T \in \mathcal{X}\}.$$

In other words, $\mathcal{C}_{\mathrm{U}}^s$ is defined as the intersection of $\mathcal{C}$ with $X^s$.

The set (4.24) can be reformulated using a variational formulation, leading to

$$\mathcal{C} = \{(x,u,x_T) \in X \mid \int_I \delta p^\mathsf{T}(-C_\mathrm{x}x - C_\mathrm{u}u + b)\mathrm{d}t \geq 0 \\ \forall \delta p \in L^2_{n_c} : \delta p \geq 0, \Big| \int_I \delta p^\mathsf{T} b \mathrm{d}t \Big| < \infty; x_T \in \mathcal{X}\}, \quad (4.25)$$

where in case $I$ has finite measure the Lebesgue integral of $\delta p^\mathsf{T} b$ over $I$ is guaranteed to be finite. In the light of (4.25), a second approximation is thus naturally obtained by restricting the test functions $\delta p$ to be spanned by the first $s$ basis functions, that is,

$$\mathcal{C}^s_\mathrm{L} := \{(x,u,x_T) \in X \mid \int_I \delta \tilde{p}^\mathsf{T}(-C_\mathrm{x}x - C_\mathrm{u}u + b)\mathrm{d}t \geq 0 \\ \forall \delta \tilde{p} = (I_{n_c} \otimes \tau)^\mathsf{T} \delta \eta_p : \delta \tilde{p} \geq 0, \delta \eta_p \in \mathbb{R}^{n_c s}; x_T \in \mathcal{X}\}. \quad (4.26)$$

The sets $\mathcal{C}^s_\mathrm{U}$, $\mathcal{C}^s_\mathrm{L}$, and $\mathcal{C}$ have the following properties:

B1) the sets $\mathcal{C}^s_\mathrm{U}$, $\mathcal{C}^s_\mathrm{L}$, and $\mathcal{C}$ are closed and convex.

B2) the sets $\pi^s(\mathcal{C}^s_\mathrm{U})$, $\pi^s(\mathcal{C}^s_\mathrm{L})$ are closed and convex.

B3) $\mathcal{C}^s_\mathrm{U} \subset \mathcal{C}^{s+1}_\mathrm{U} \subset \mathcal{C}$.

B4) $\mathcal{C}^s_\mathrm{L} \supset \mathcal{C}^{s+1}_\mathrm{L} \supset \mathcal{C}$.

B5) $(\pi^s)^*\pi^s(\mathcal{C}^s_\mathrm{U}) \subset \mathcal{C}^s_\mathrm{U}$, $(\pi^s)^*\pi^s(\mathcal{C}^s_\mathrm{L}) \subset \mathcal{C}^s_\mathrm{L}$.

Property B1 ensures that the resulting optimizations over $\mathcal{C}^s_\mathrm{U}$, $\mathcal{C}^s_\mathrm{L}$, and $\mathcal{C}$ will be convex, and that the corresponding minima will be attained. The sets $\mathcal{C}^s_\mathrm{U}$ and $\mathcal{C}^s_\mathrm{L}$ are represented as subsets of the Euclidean space through the map $\pi^s$. As a result, Property B2 ensures that the corresponding (finite-dimensional) optimization problems will be convex and that the corresponding minima will be attained. Property B3 is used to obtain a monotonically decreasing sequence bounding $J_\infty$ from above, whereas Property B4 is used to construct a montonically increasing sequence bounding $J_\infty$ from below. Property B5 will guarantee consistency of the corresponding finite-dimensional optimization problems.

The proof of Properties B1-B5 can be found in App. A. A schematic illustration of the sets $\mathcal{C}^s_\mathrm{L}$, $\mathcal{C}$, and $\mathcal{C}^s_\mathrm{U}$ is shown in Fig. 4.2.

It follows from the property of the set $\mathcal{C}^s_\mathrm{U}$ being contained in $\mathcal{C}^{s+1}_\mathrm{U}$, and $\mathcal{C}^{s+1}_\mathrm{L}$ being contained in $\mathcal{C}^s_\mathrm{L}$, that the limits

$$\lim_{s\to\infty} \mathcal{C}^s_\mathrm{U} = \bigcup_{s=1}^{\infty} \mathcal{C}^s_\mathrm{U}, \quad \lim_{s\to\infty} \mathcal{C}^s_\mathrm{L} = \bigcap_{s=1}^{\infty} \mathcal{C}^s_\mathrm{L} \quad (4.27)$$

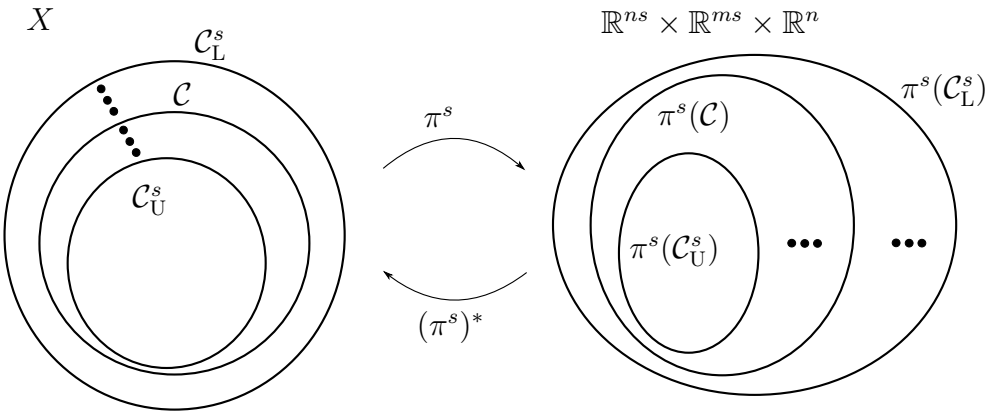


**Figure 4.2.** Schematic illustration of the Properties B1-B4 and the maps $\pi^s$ and $(\pi^s)^*$.

exist, [36, p. 18, p. 21].[26] Next, we will provide conditions asserting that the two limits agree. To that extent we introduce the following definition: The set $\mathcal{A}$ is an algebra, if it is closed under addition, scalar multiplication, and multiplication, that is,

$$f + g \in \mathcal{A}, \quad fg \in \mathcal{A}, \quad cf \in \mathcal{A}, \tag{4.28}$$

for any $f, g \in \mathcal{A}$, $c \in \mathbb{R}$, [37, p. 161].

*Lemma* 7. Given that the basis functions form an algebra and that the basis functions are dense in the set of smooth functions with compact support in $I$, it holds that

$$\lim_{s\to\infty} \mathcal{C}_{\mathrm{U}}^s = \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s.$$

*Proof.* We consider the case where $I$ has finite measure. The proof for the case where $I$ has infinite measure is given in App. C. We claim that $\lim_{s\to\infty} \mathcal{C}_{\mathrm{U}}^s \supset \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$. We prove the claim by contradiction. Let $(x, u, x_T) \in \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$ be such that there exists an open set $U$ with

$$C_{\mathrm{x}k}x(t) + C_{\mathrm{u}k}u(t) > b_k, \quad \forall t \in U \quad \text{a.e.}, \tag{4.29}$$

for some $k \in \{1, 2, \ldots, n_c\}$, where $C_{\mathrm{x}k}$ and $C_{\mathrm{u}k}$ denotes the $k$th row of $C_{\mathrm{x}}$, respectively $C_{\mathrm{u}}$. Thus, it holds that

$$\int_I \delta v(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t < 0 \tag{4.30}$$

for all smooth test functions $\delta v : I \to \mathbb{R}$, vanishing outside $U$, with $\delta v(t) > 0\ \forall t \in U$. We pick one of these test functions and denote it by $\delta p$. Due to the fact that the basis functions are dense in the set of smooth functions with compact support, there exists a sequence $\sqrt{\delta \tilde{p}_i}$, which converges uniformly to $\sqrt{\delta p}$, that is, for any given $\epsilon > 0$ there

[26] Moreover, $\lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$ is closed, as it can be written as the intersection of the closed sets $\mathcal{C}_{\mathrm{L}}^s$.

exists an integer $N > 0$ such that

$$||\sqrt{\delta\tilde{p}_i} - \sqrt{\delta p}||_\infty < \epsilon, \quad \forall i > N, \tag{4.31}$$

where $||\cdot||_\infty$ denotes the supremum norm. From the assumption that the basis functions form an algebra that is closed under multiplication, we can infer that $\delta\tilde{p}_i$ lies likewise in the span of the basis functions. Moreover, we have that

$$\begin{aligned} ||\delta\tilde{p}_i - \delta p||_\infty &= ||(\sqrt{\delta\tilde{p}_i} - \sqrt{\delta p})(\sqrt{\delta\tilde{p}_i} + \sqrt{\delta p})||_\infty && (4.32)\\ &\leq ||\sqrt{\delta\tilde{p}_i} - \sqrt{\delta p}||_\infty ||\sqrt{\delta\tilde{p}_i} + \sqrt{\delta p}||_\infty && (4.33)\\ &< \epsilon||\sqrt{\delta\tilde{p}_i} - \sqrt{\delta p} + 2\sqrt{\delta p}||_\infty && (4.34)\\ &< \epsilon(\epsilon + 2||\sqrt{\delta p}||_\infty) \leq C_1\epsilon, && (4.35) \end{aligned}$$

for all integers $i > N$, where $C_1 > 0$ is constant (for $\epsilon$ sufficiently small). By assumption, $(x, u, x_T) \in \mathcal{C}^s_{\mathrm{L}}$, for all integers $s > 0$, and therefore

$$\int_I \delta\tilde{p}_i(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b)\mathrm{d}t \geq 0, \tag{4.36}$$

for all integers $i > 0$. However, the above integral can be rewritten as

$$\begin{aligned} 0 \leq &\int_I \delta p(-C_{\mathrm{x}k}x - C_{\mathrm{u}k} + b)\mathrm{d}t && (4.37)\\ &+ \int_I (\delta\tilde{p}_i - \delta p)^{\mathsf{T}}(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b)\mathrm{d}t, && (4.38) \end{aligned}$$

where the last term can be bounded by (using Hölder's inequality twice)

$$\epsilon C_1 \int_I |C_{\mathrm{x}k}x + C_{\mathrm{u}k}u - b|\mathrm{d}t \leq \epsilon C_1 ||C_{\mathrm{x}k}x + C_{\mathrm{u}k}u - b||_2\sqrt{T}, \tag{4.39}$$

for all integers $i > N$. The fact that the expression (4.38) converges to zero as $i \to \infty$ leads to a contradiction with (4.30). It follows therefore that $\lim_{s\to\infty} \mathcal{C}^s_{\mathrm{U}} \supset \lim_{s\to\infty} \mathcal{C}^s_{\mathrm{L}}$, which, combined with $\mathcal{C}^s_{\mathrm{U}} \subset \mathcal{C}^s_{\mathrm{L}}$ for all integers $s > 0$, leads to the desired conclusion. □

### 2.4 Approximation of the dynamics

We define the the set

$$\mathcal{D} := \{(x, u, x_T) \in X | \dot{x} = Ax + Bu, x(0) = x_0, x(T) = x_T\}, \tag{4.40}$$

containing all trajectories $x$ and $u$ fulfilling the equations of motion in a weak sense. We

obtain a first approximation to the set $\mathcal{D}$ by restricting $x$ and $u$ to be spanned by the first $s$ basis functions, that is,

$$\mathcal{D}_{\mathrm{U}}^{s} := \mathcal{D} \bigcap X^{s}. \tag{4.41}$$

It was shown in [33] that the linearity of the dynamics implies that the set $\mathcal{D}_{\mathrm{U}}^{s}$ can be rewritten as the elements $(\tilde{x}, \tilde{u}, x_T) \in X^s$, satisfying $\tilde{x}(0) = x_0$, $\tilde{x}(T) = x_T$, and

$$\int_I \delta\tilde{p}^{\mathsf{T}}(\dot{\tilde{x}} - A\tilde{x} - B\tilde{u})\mathrm{d}t = 0, \tag{4.42}$$

for all variations $\delta\tilde{p}$ that are spanned by the first $s$ basis functions. In particular, (4.42) reduces to a linear equation in the coefficient vectors $\eta_x$ and $\eta_u$ defining the trajectories $\tilde{x}$ and $\tilde{u}$ compatible with the equations of motion.

In order to obtain a second approximation, we reformulate the dynamics in terms of the variational equality

$$\int_I \delta p^{\mathsf{T}}(\dot{x} - Ax - Bu)\mathrm{d}t + \delta p(0)^{\mathsf{T}}(x(0) - x_0) + \delta p(T)^{\mathsf{T}}(x_T - x(T)) = 0, \tag{4.43}$$

for all test functions $\delta p \in H_n$, where $H_n$ denotes the set of functions in $L_n^2$ having a weak derivative in $L_n^2$. As remarked earlier, $\delta p$ has therefore a unique absolutely continuous representative to which we refer when writing $\delta p(0)$, $\delta p(T)$. In case $I$ has infinite measure, the above equation reduces naturally to

$$\int_I \delta p^{\mathsf{T}}(\dot{x} - Ax - Bu)\mathrm{d}t + \delta p(0)^{\mathsf{T}}(x(0) - x_0) = 0,$$

for all test functions $\delta p \in H_n$. The formulation (4.43) is equivalent to the one in (4.40), which is implied by the fundamental lemma of the calculus of variations [38, p. 18]. Applying integration by parts results in

$$-\int_I \delta\dot{p}^{\mathsf{T}} x\mathrm{d}t - \int_I \delta p^{\mathsf{T}}(Ax + Bu)\mathrm{d}t - \delta p(0)^{\mathsf{T}}x_0 + \delta p(T)^{\mathsf{T}}x_T = 0, \quad \forall \delta p \in H_n, \tag{4.44}$$

and is often referred to as weak formulation of the dynamics. The above equation is well-defined for all $x \in L_n^2$ (and equivalent to (4.40)). Therefore, by restricting the variations in (4.44) to be spanned by the first $s$ basis functions, we obtain

$$\begin{aligned}\mathcal{D}_{\mathrm{L}}^{s} := \{&(x, u, x_T) \in X \mid \delta\tilde{p} = (I_n \otimes \tau)^{\mathsf{T}}\delta\eta_p, \\ &-\int_I \delta\dot{\tilde{p}}^{\mathsf{T}} x\mathrm{d}t - \int_I \delta\tilde{p}^{\mathsf{T}}(Ax + Bu)\mathrm{d}t - \delta\tilde{p}(0)^{\mathsf{T}}x_0 + \delta\tilde{p}(T)^{\mathsf{T}}x_T = 0, \forall \delta\eta_p \in \mathbb{R}^{ns}\},\end{aligned} \tag{4.45}$$

as an approximation to $\mathcal{D}$. Note that while the formulation (4.40) (and likewise (4.44)) implies $x \in H_n$, the set $\mathcal{D}_{\mathrm{L}}^s$ contains also elements $x \in L_n^2$ that do not necessarily have a weak derivative in $L_n^2$. The use of the weak formulation for the definition of $\mathcal{D}_{\mathrm{L}}^s$ is motivated by the fact that the resulting set $\mathcal{D}_{\mathrm{L}}^s$ is closed (see below), which is important to ensure that the minimum is attained, when optimizing over $\mathcal{D}_{\mathrm{L}}^s$.

The sets $\mathcal{D}_{\mathrm{U}}^s$, $\mathcal{D}_{\mathrm{L}}^s$, and $\mathcal{D}$ have the following properties:

C1) the sets $\mathcal{D}_{\mathrm{U}}^s$, $\mathcal{D}_{\mathrm{L}}^s$, and $\mathcal{D}$ are closed and convex.

C2) the sets $\pi^s(\mathcal{D}_{\mathrm{U}}^s)$ and $\pi^s(\mathcal{D}_{\mathrm{L}}^s)$ are closed and convex.

C3) $\mathcal{D}_{\mathrm{U}}^s \subset \mathcal{D}_{\mathrm{U}}^{s+1} \subset \mathcal{D}$.

C4) $\mathcal{D}_{\mathrm{L}}^s \supset \mathcal{D}_{\mathrm{L}}^{s+1} \supset \mathcal{D}$.

C5) $(\pi^s)^*\pi^s(\mathcal{D}_{\mathrm{U}}^s) \subset \mathcal{D}_{\mathrm{U}}^s$, $(\pi^s)^*\pi^s(\mathcal{D}_{\mathrm{L}}^s) \subset \mathcal{D}_{\mathrm{L}}^s$.

The above properties are in complete analogy to the previous section, and will be used to draw analogous conclusions. In particular, Properties C1 and C2 ensure that the resulting optimization problems will have unique well-defined minimizers. Properties C3 and C4 will lead to a monotonically decreasing sequence bounding $J_\infty$ above, respectively to a monotonically increasing sequence bounding $J_\infty$ below. Property C5 will guarantee consistency. A proof of Properties C1-C5 can be found in App. B.

### 2.5 Resulting optimization problems

Using the definitions of the previous section we can rewrite (4.1) as

$$\begin{aligned} J_\infty = &\min ||x||^2 + ||u||^2 + \psi(x_T) \\ &\text{s.t. } (x, u, x_T) \in \mathcal{C} \cap \mathcal{D}. \end{aligned} \tag{4.46}$$

By assumption, there exists a feasible trajectory satisfying the dynamics and the constraints. Therefore the set $\mathcal{C} \cap \mathcal{D}$ is nonempty and the objective is bounded above. As a consequence, (4.46) reduces to an optimization over a closed convex and bounded set in the Banach space $X$ (the set $\mathcal{C} \cap \mathcal{D}$ is closed). Thus, the minimum in (4.46) is attained and due to the strong convexity of the objective function the corresponding minimizer $(x, u, x_T) \in X$ is unique, [39, Thm. 26, p. 93].

By combining the approximation of the constraints in Sec. 2 and the approximation of the dynamics in Sec. 2.4 we obtain the two auxiliary problems

$$\begin{aligned} J_s := &\inf ||\tilde{x}||^2 + ||\tilde{u}||^2 + \psi(x_T) \\ &\text{s.t. } (\tilde{x}, \tilde{u}, x_T) \in \mathcal{C}_{\mathrm{U}}^s \cap \mathcal{D}_{\mathrm{U}}^s, \end{aligned} \tag{4.47}$$

and

$$\begin{aligned}\tilde{J}_s := \min\, &||\tilde{x}||^2 + ||\tilde{u}||^2 + \psi(x_T) \\ \text{s.t. } &(\tilde{x}, \tilde{u}, x_T) \in \mathcal{C}_{\mathrm{L}}^s \cap \mathcal{D}_{\mathrm{L}}^s.\end{aligned} \tag{4.48}$$

We make the assumption that there exists an integer $s_0 > 0$ large enough such that $\mathcal{C}_{\mathrm{U}}^{s_0} \cap \mathcal{D}_{\mathrm{U}}^{s_0}$ is nonempty. The closedness and convexity of the sets $\mathcal{C}_{\mathrm{U}}^s$, $\mathcal{D}_{\mathrm{U}}^s$, $\mathcal{C}_{\mathrm{L}}^s$, and $\mathcal{D}_{\mathrm{L}}^s$ leads to the conclusion that the infimum in (4.47) is attained and that the corresponding minimizer is unique for $s = s_0$. From the fact that $\mathcal{D}_{\mathrm{U}}^s$ is contained in $\mathcal{D}_{\mathrm{U}}^{s+1}$ and $\mathcal{C}_{\mathrm{U}}^s$ is contained in $\mathcal{C}_{\mathrm{U}}^{s+1}$ it follows that $J_{s+1} \leq J_s$ for all integers $s \geq s_0$. This implies that the infimium in (4.47) is attained and that the corresponding minimizers are unique for all integers $s \geq s_0$. Moreover, the fact that $\mathcal{D}_{\mathrm{U}}^s \subset \mathcal{D}$ and $\mathcal{C}_{\mathrm{U}}^s \subset \mathcal{C}$ implies that $J_s \geq J_\infty$ for all integers $s \geq s_0$. Similar arguments show that $\tilde{J}_s \leq \tilde{J}_{s+1} \leq J_\infty$ for all integers $s > 0$, that the minimum in (4.48) is indeed attained, and that the corresponding minimizers are unique for all integers $s > 0$. The results are summarized with the following lemma.

*Lemma* 8. Let the sets $\mathcal{C} \cap \mathcal{D}$ and $\mathcal{C}_{\mathrm{U}}^{s_0} \cap \mathcal{D}_{\mathrm{U}}^{s_0}$ be nonempty for some integer $s_0 > 0$. Then the optimization problems (4.46), (4.47), and (4.48) are well defined and the corresponding minima are attained and are unique. Moreover, the costs $J_s$ form a monotonically decreasing sequence bounding $J_\infty$ above for all integers $s \geq s_0$, whereas the costs $\tilde{J}_s$ form a monotonically increasing sequence bounding $J_\infty$ below for all integers $s > 0$.

By definition of the constraints $\mathcal{C}_{\mathrm{U}}^s$ and $\mathcal{D}_{\mathrm{U}}^s$, the minimizer of (4.47) (for $s \geq s_0$) is required to be an element of $X^s$. Consequently, the problem (4.47) is equivalent to

$$\begin{aligned}J_s = \inf\, &|\eta_x|^2 + |\eta_u|^2 + \psi(x_T) \\ \text{s.t. } &(\eta_x, \eta_u, x_T) \in \pi^s(\mathcal{C}_{\mathrm{U}}^s) \cap \pi^s(\mathcal{D}_{\mathrm{U}}^s),\end{aligned} \tag{4.49}$$

which corresponds to a convex finite-dimensional optimization problem. Note that the orthonormality of the basis functions can be used to conclude $|\eta_x|^2 = ||\tilde{x}||^2$ and likewise $|\eta_u|^2 = ||\tilde{u}||^2$. Similarly, the minimizer $(x, u, x_T) \in X$ of (4.48) is guaranteed to lie in $X^s$. This can be shown by contradiction: We assume therefore $(x, u, x_T) \in X \setminus X^s$. We construct $\tilde{x} := (\pi^{ns})^* \pi^{ns}(x)$, and $\tilde{u} := (\pi^{ms})^* \pi^{ms}(u)$. As a consequence of Properties B5 and C5 it follows that $(\tilde{x}, \tilde{u}, x_T) \in \mathcal{C}_{\mathrm{L}}^s \cap \mathcal{D}_{\mathrm{L}}^s$. Moreover, by orthonormality of the basis functions it holds that

$$\int_I \tilde{x}^{\mathsf{T}}(x - \tilde{x})\mathrm{d}t = 0, \quad \int_I \tilde{u}^{\mathsf{T}}(u - \tilde{u})\mathrm{d}t = 0, \tag{4.50}$$

which leads to

$$||x||^2 = ||x - \tilde{x}||^2 + ||\tilde{x}||^2, \quad ||u||^2 = ||u - \tilde{u}||^2 + ||\tilde{u}||^2. \tag{4.51}$$

This implies that $\tilde{x}$, $\tilde{u}$, and $x_T$ achieve a cost that is below $\tilde{J}_s$, contradicting the fact that $(x, u, x_T)$ are the minimizer of (4.48).

This shows that the convex finite-dimensional problem

$$\begin{aligned}\tilde{J}_s = &\min |\eta_x|^2 + |\eta_u|^2 + \psi(x_T) \\ &\text{s.t. } (\eta_x, \eta_u, x_T) \in \pi^s(\mathcal{C}_{\mathrm{L}}^s) \cap \pi^s(\mathcal{D}_{\mathrm{L}}^s),\end{aligned} \tag{4.52}$$

is equivalent to (4.48) in the sense that its (unique) minimizer $(\eta_x, \eta_u, \bar{x}_T)$ is related to the minimizer $(x, u, x_T) \in X$ of (4.48) by $x = (\pi^{ns})^*(\eta_x)$, $u = (\pi^{ms})^*(\eta_u)$, $\bar{x}_T = x_T$, and achieves the same cost. By virtue of Lemma 8 we therefore conclude

**Theorem 9.** *Let the sets $\mathcal{C} \cap \mathcal{D}$ and $\mathcal{C}_U^{s_0} \cap \mathcal{D}_U^{s_0}$ be nonempty for some integer $s_0 > 0$. Then, the optimization problems* (4.46), (4.49), *and* (4.52) *are well-defined and the corresponding minima are attained and are unique. Moreover, the costs $J_s$ form a monotonically decreasing sequence bounding $J_\infty$ above for all integers $s \geq s_0$, whereas the costs $\tilde{J}_s$ form a monotonically increasing sequence bounding $J_\infty$ below for all integers $s > 0$.*

From the fact that $J_s$ and $\tilde{J}_s$ form monotonically increasing, respectively monotonically decreasing sequences it follows at once that

$$\lim_{s\to\infty} J_s, \quad \lim_{s\to\infty} \tilde{J}_s \tag{4.53}$$

exist. We will argue that not only the optimal cost, but also the optimal trajectories converge (strongly). In addition, this will provide a means to quantify the $L^2$-error of the input and state trajectories with respect to the trajectories corresponding to (4.1).

**Proposition 6.** *Let the assumptions of Thm. 9 be fulfilled. Let the optimal trajectories of* (4.47), *respectively* (4.49) *be denoted by $\tilde{x}^s$, $\tilde{u}^s$, $x_{Ts}$, and the optimal trajectories of* (4.1) *by $x$, $u$, $x_T$. It is further assumed that $\psi$ is $\mu$-strongly convex. Then, $\tilde{x}^s$, $\tilde{u}^s$, and $x_{Ts}$ converge strongly, and for all integers $s \geq s_0$ it holds that*

$$||\tilde{x}^s - x||^2 + ||\tilde{u}^s - u||^2 + \frac{\mu}{2}|x_{Ts} - x_T|^2 \leq 2(J_s - J_\infty). \tag{4.54}$$

*Proof.* We first note that the strong convexity of $\psi$ implies that for any real numbers $x_1$ and $x_2$ it holds that

$$\psi(\frac{1}{2}(x_1 + x_2)) \leq \frac{1}{2}(\psi(x_1) + \psi(x_2)) - \frac{1}{4}\frac{\mu}{2}|x_1 - x_2|^2, \tag{4.55}$$

where $\mu$ is a strictly positive constant. We construct from the two optimizers $(\tilde{x}^s, \tilde{u}^s, x_{Ts})$

and $(\tilde{x}^{s+1}, \tilde{u}^{s+1}, x_{T_{s+1}})$ the feasible candidate

$$(\frac{1}{2}(\tilde{x}^s + \tilde{x}^{s+1}), \frac{1}{2}(\tilde{u}^s + \tilde{u}^{s+1}), \frac{1}{2}(x_{Ts} + x_{T_{s+1}})) \in \mathcal{C}^{s+1} \cap \mathcal{D}^{s+1}. \tag{4.56}$$

Feasibility is guaranteed due to the convexity of both $\mathcal{C}^{s+1}$ and $\mathcal{D}^{s+1}$. This results in

$$J_{s+1} \quad \leq \quad ||\frac{1}{2}(\tilde{x}^s + \tilde{x}^{s+1})||^2 + ||\frac{1}{2}(\tilde{u}^s + \tilde{u}^{s+1})||^2 + \psi(\frac{1}{2}(x_{Ts} + x_{T_{s+1}})). \tag{4.57}$$

Applying the relation (4.55), which holds with $\mu = 2$ in case of the $L^2$-norm, yields

$$J_{s+1} \leq \frac{1}{2}(J_s + J_{s+1}) - \frac{1}{4}(||\tilde{x}^s - \tilde{x}^{s+1}||^2 + ||\tilde{u}^s - \tilde{u}^{s+1}||^2) - \frac{1}{4}\frac{\mu}{2}|x_{Ts} - x_{T_{s+1}}|^2. \tag{4.58}$$

This implies

$$||\tilde{x}^s - \tilde{x}^{s+1}||^2 + ||\tilde{u}^s - \tilde{u}^{s+1}||^2 + \frac{\mu}{2}|x_{Ts} - x_{T_{s+1}}|^2 \quad \leq \quad 2(J_s - J_{s+1}), \tag{4.59}$$

and from the convergence of $J_s$ it follows therefore that the elements $(\tilde{x}^s, \tilde{u}^s, x_{Ts}) \in X^s$ form a Cauchy sequence, which, due to the completeness of $X^s$, converges (strongly). The above argument can be repeated by replacing $\tilde{x}^{s+1}$, $\tilde{u}^{s+1}$ and $x_{T_{s+1}}$ by the optimizer of (4.1), resulting in the inequality (4.54). □

The inequality (4.54) is particularly interesting, as it can be used to determine the quality of the approximation of (4.47) (or likewise (4.49)) with respect to (4.1). This is because the suboptimality in the cost can be bounded by

$$J_s - J_\infty \leq J_s - \tilde{J}_s, \tag{4.60}$$

which stems from the fact that $\tilde{J}_s \leq J_\infty$ for all integers $s > 0$. As a result, by solving the two finite-dimensional problems (4.47) and (4.48) (or likewise (4.49) and (4.52)) not only the suboptimality of the cost $J_s$ compared to $J_\infty$ can be quantified, but also the $L^2$-distance of the corresponding optimal input and state trajectories.

The same reasoning can be applied to the trajectories obtained by solving (4.48) or (4.52).

**Proposition 7.** *Let the assumptions of Thm. 9 be fulfilled. Let the optimal trajectories to* (4.48)*, respectively* (4.52) *be denoted by* $\tilde{x}^s$*,* $\tilde{u}^s$*,* $x_{Ts}$*, and the optimal trajectories to* (4.1) *by* $x$*,* $u$*,* $x_T$*. It is further assumed that* $\psi$ *is* $\mu$*-strongly convex. Then,* $\tilde{x}^s$*,* $\tilde{u}^s$*, and*

*$x_{Ts}$ converge strongly, and for all integers $s \geq 0$ it holds that*

$$||\tilde{x}^s - x||^2 + ||\tilde{u}^s - u||^2 + \frac{\mu}{2}|x_{Ts} - x_T|^2 \leq 2(J_\infty - \tilde{J}_s). \tag{4.61}$$

In case the assumption of $\psi$ being strongly convex is dropped, strong convergence of $\tilde{x}^s$ and $\tilde{u}^s$ can still be established. Moreover, in the absence of the terminal cost $\psi$, for instance in the infinite-horizon case, the bounds (4.54) and (4.61) continue to hold, and it remains true that $\tilde{x}^s$ and $\tilde{u}^s$ converge strongly.

We now provide conditions under which $J_s$ and $\tilde{J}_s$ converge both to $J_\infty$.

**Theorem 10.** *Given that the basis functions form an algebra and are dense in the set of continuous functions with compact support in $I$, it holds that*

$$\lim_{s\to\infty} J_s = \lim_{s\to\infty} \tilde{J}_s = J_\infty. \tag{4.62}$$

*Proof.* From Thm. 9 it can be inferred that $J_s$ and $\tilde{J}_s$ converge, and that the corresponding sequence of optimal input and state trajectories is bounded. Furthermore, by virtue of Lemma 7 we have that $\lim_{s\to\infty} \mathcal{C}^s_{\mathrm{U}} = \lim_{s\to\infty} \mathcal{C}^s_{\mathrm{L}}$, and as a consequence we can apply Prop. 3.6 in [34] (the proposition extends naturally to the finite-horizon case), which leads to the desired result. □

By combining Thm. 10 with Prop. 6 or Prop. 7 it follows that not only the optimal value function but also the corresponding sequence of optimal trajectories converges strongly to the optimal trajectories of (4.1).

### 2.6 Remarks

In the previous section the basis functions were assumed to be continuous throughout the time interval $I$. It is, however, straightforward to extend the previous results in case the interval is split up, for example in $I_1$, $I_2$, ..., $I_N$, with $\cup_{i=1}^N I_i = I$, and piecewise continuous basis functions defined over the intervals $I_i$, $i = 1, 2, \ldots, N$ are used. The basis functions are then required to fulfill Assumptions A1 and A2 over the intervals $I_i$ separately, $i = 1, 2, \ldots, N$. This includes for example the case where polynomials are used as basis functions on $(0, 1)$ and exponentially decaying polynomials on $(1, \infty)$. Thereby, the basis functions complexity is in general increased, which potentially improves the approximation quality and leads to tighter upper and lower bounds on $J_\infty$. However, in the context of MPC, closed-loop stability and recursive feasibility is in general lost when splitting the interval $I$, due to the fact that the obtained (potentially discontinuous) solutions cannot be shifted in time (see Sec. 3).

Moreover, Thm. 9, Prop. 6, Prop. 7, and Thm. 10 can be generalized to a strongly convex running cost instead of a quadratic one. In practice, a quadratic running cost has the advantage of yielding a quadratic objective function, facilitating the numerical solution of the resulting optimization problem.

### 2.7 A remark on the discrete-time formulation

The main results presented earlier can be translated to the discrete-time case. Due to the high similarity, we will restrict the discussion to the following few remarks.

In the discrete-time case the trajectories $x$, $u$ corresponding to the discrete-time counterpart of (4.1) are approximated via

$$\tilde{x}(k) = (I_n \otimes \tau(k))^{\mathsf{T}} \eta_x, \quad \tilde{u}(k) = (I_m \otimes \tau(k))^{\mathsf{T}} \eta_u, \tag{4.63}$$

for all $k \in I$, where $I$ is a subset of all non-negative integers. We will slightly abuse notation and denote both discrete-time and continuous-time trajectories with the same variables, that is, $\tilde{x}$, $\tau$, etc. The discrete-time analogue of Assumption A2 is given by

A2D) The basis functions fulfill $\tau^s(k+1) = M_{\mathrm{d}s}\tau^s(k)$ for all $k \in I$.

The subscript 'd' highlights that the matrix $M_{\mathrm{d}s}$ and $M_s$ are a priori unrelated. However, we may choose $M_{\mathrm{d}s} = e^{M_s T_{\mathrm{d}}}$, for a fixed time $T_{\mathrm{d}} > 0$, in which case the discrete-time basis function $\tau(k)$ matches the corresponding continuous-time basis function $\tau(t)$ at time $t = kT_{\mathrm{d}}$ for all $k \in I$. In complete analogy to the continuous-time case, Assumption A2D leads to an invariance of the basis functions with respect to shifts in the index $k$. Hence, in the context of MPC, closed-loop stability guarantees can be shown by the same arguments as in the continuous-time setting (see Sec. 3).

The major difference compared to the continuous-time formulation is that the inequality constraints in the discrete-time version of (4.1) are at most enforced at a countable number of time indices. No matter whether $I$ has finite or infinite cardinality, this always leads to a finite number of inequality constraints that need to be enforced in the corresponding approximations.[27] Hence, in the discrete-time setting the resulting approximations can always be reduced to standard quadratic programs.

## 3. Part II: Model Predictive Control

The proposed approximations can be applied in the context of MPC. We will show that by repeatedly solving the infinite-horizon optimal control problem (4.49) (with $I = (0, \infty)$, $\psi = 0$, $x_T = \lim_{t\to\infty} x(t) = 0$), recursive feasibility and closed-loop stability are inherent to the resulting MPC algorithm. Due to the fact that the basis functions are decaying the constraint $\lim_{t\to\infty} \tilde{x}(t) = 0$ is satisfied by construction and does not lead to an additional

[27] In case $I$ has not a finite cardinality, the basis functions are required to be exponentially decaying, due to Assumption A2D and the fact that they are square summable. Hence, as will be shown in the following (the results translate to the discrete setting), it is enough to check the inequality constraints at a finite number of points.

terminal constraint. For the sake of completeness, (4.49) is written out as

$$\begin{aligned} \min\ & \eta_x^{\mathsf{T}}(Q \otimes I_s)\eta_x + \eta_u^{\mathsf{T}}(R \otimes I_s)\eta_u \\ \text{s.t. } & (I_n \otimes M_s^{\mathsf{T}} - A \otimes I_s)\eta_x - (B \otimes I_s)\eta_u = 0, \\ & (I_n \otimes \tau(0))^{\mathsf{T}}\eta_x = x_0, \\ & (C_{\mathrm{x}} \otimes \tau(t))^{\mathsf{T}}\eta_x + (C_{\mathrm{u}} \otimes \tau(t))^{\mathsf{T}}\eta_u \leq b, \forall t \in [0, \infty), \end{aligned} \tag{4.64}$$

where the input and state costs are weighted with the matrices $Q > 0$ and $R > 0$, which is common in MPC.

The MPC algorithm consists of two steps: In a first step, input and state trajectories $\tilde{x}$ and $\tilde{u}$ are obtained by solving (4.64) subject to the current state as initial condition $x_0$. In a second step, the first portion of the input $\tilde{u}$ is applied to the system, and the procedure is repeated in the next sampling interval. As a consequence, feedback control is achieved.

We recall that Assumption A2 implies that the basis functions can capture arbitrary time-shifts in the sense that for every time-shift $T_{\mathrm{s}}$ and any given trajectory $f(t) = \tau^s(t)^{\mathsf{T}}\eta$, where $\eta \in \mathbb{R}^s$ is the parameter vector, there exists a different set of parameters $\hat{\eta}$, such that $f(t - T_{\mathrm{s}}) = \tau^s(t)^{\mathsf{T}}\hat{\eta}$ for all times $t$. This result can be easily established by rewriting the basis functions in terms of a matrix exponential as done in [33].

We now discuss the stability properties of the proposed control strategy. Without loss of generality we set $t = 0$. According to the first step of the MPC algorithm we solve (4.64) to obtain the optimal trajectories $\tilde{x}$ and $\tilde{u}$. The input $\tilde{u}$ is then applied to the system over the time span $[0, T_{\mathrm{d}})$, where $T_{\mathrm{d}}$ is the sampling time. In the absence of modeling errors, the system will evolve along the predicted trajectory $\tilde{x}$, which is due to the fact that the predictions $\tilde{x}$ and $\tilde{u}$ are exact (as shown in [33]). Due to the time-shift property of the basis functions, there exist parameters $\hat{\eta}_x$ and $\hat{\eta}_u$ for expressing the shifted trajectories $\tilde{x}(t + T_{\mathrm{d}})$, $\tilde{u}(t + T_{\mathrm{d}})$ as a linear combination of the same basis functions, that is

$$\tilde{x}(t + T_{\mathrm{d}}) = (I_n \otimes \tau(t))^{\mathsf{T}}\hat{\eta}_x, \tilde{u}(t + T_{\mathrm{d}}) = (I_m \otimes \tau(t))^{\mathsf{T}}\hat{\eta}_u, \tag{4.65}$$

for all times $t \in [0, \infty)$. The trajectories $\tilde{x}$ and $\tilde{u}$ are guaranteed to satisfy the equations of motion and the stage constraints for all times and therefore the parameters $\hat{\eta}_x$ and $\hat{\eta}_u$ are feasible candidates for the optimization (4.64) with $x_0 = \tilde{x}(T_{\mathrm{d}})$ at the time instant $t = T_{\mathrm{d}}$. Recursive feasibility follows then by induction. Moreover, the resulting optimal cost (obtained by solving (4.64) at time $t = T_{\mathrm{d}}$) is certainly lower than the cost achieved by the feasible candidates $\hat{\eta}_x$ and $\hat{\eta}_u$ corresponding to the trajectories $\tilde{x}(t + T_{\mathrm{d}})$ and $\tilde{u}(t + T_{\mathrm{d}})$. The optimal cost at time $t = T_{\mathrm{d}}$ is therefore bounded by the difference of the optimal cost at time $t = 0$ with the integral of the running cost over the interval $[0, T_{\mathrm{d}})$. As a consequence, the cost is guaranteed to decrease over time, acts therefore as a Lyapunov function, and can be used to conclude closed-loop stability. The previous

argument is summarized with the following proposition.

**Proposition 8.** *Provided that the optimization* (4.64) *is feasible at time* $t = 0$*, it remains feasible for all times* $t > 0$*, and the resulting closed-loop system is guaranteed to be asymptotically stable.*

*Proof.* A formal proof can be found in [33] and is included in App. D. □

The previous result continues to hold even if (4.64) is not solved to full optimality (which is often not practicable), provided that the numerical solution algorithm is monotonic in the cost. Given that the numerical solution algorithm is initialized with feasible trajectories at time $t = 0$, a single iteration of the solver at each time-step is enough for the stability guarantee to hold, as follows from the above argument.

### 3.1 Implementation of the semi-infinite constraint

The optimization (4.64) is a convex finite-dimensional optimization problem. However, it is not a quadratic program, as it includes the semi-infinite constraint

$$(C_{\mathrm{x}} \otimes \tau(t))^{\mathsf{T}} \eta_x + (C_{\mathrm{u}} \otimes \tau(t))^{\mathsf{T}} \eta_u \leq b, \forall t \in [0, \infty). \tag{4.66}$$

As a result, (4.64) cannot be solved by a standard quadratic programming solver. Three different approaches to deal with the semi-infinite constraint are immediate:

1) global polyhedral approximation

2) sum-of-squares approximation

3) local polyhedral approximation (active-set approach).

The first is based on a fixed polyhedral approximation, leading to a quadratic program. The second is based on exploiting polynomial basis functions for obtaining a characterization using linear matrix inequalities, whereas the third is based on an iterative constraint sampling scheme, resulting in a local polyhedral approximation. All approaches aim at leaving Prop. 8, Prop. 6, and Prop. 7 intact. In the following subsections we will focus on approach 1 and 2. We will describe an efficient solution algorithm based on approach 3 in detail in Sec. 4.

It turns out that it is enough to check the constraint (4.66) over a compact time interval, instead of the unbounded interval $t \in [0, \infty)$. This is because the basis functions are assumed to be linearly independent and exponentially decaying according to Assumption A1 and A2. A formal proof of this claim can be found in App. E. The compact time interval for which the constraint (4.66) has to be checked is denoted by $[0, T_{\mathrm{c}}]$ and depends on the choice of basis functions, on the order $s$, and in some cases also on the bound $b$ (see App. E).

In case the presented approach is used for situations with a changing set-point, the semi-infinite constraint must be adapted accordingly. This leads to changes in the right-hand side of (4.66), resulting in a translation of the finite-dimensional convex set described by (4.66).

*1) Global polyhedral approximation* In order to construct a global polyhedral approximation of the set $\mathcal{C}_{\mathrm{U}}^s$ we assume that an upper bound on the achievable cost $J_s$, denoted by $\bar{J}_s$, is available. In order to simplify the discussion, we assume further that for now $C_{\mathrm{x}}$ and $C_{\mathrm{u}}$ are row vectors, and that $b$ is a scalar. The proposed scheme can be readily extended to the case where $C_{\mathrm{x}}$ and $C_{\mathrm{u}}$ are matrices, and $b$ is a vector. The approximation is based on constraint sampling, where we tighten the constraint slightly to

$$C_{\mathrm{x}}\tilde{x}(t_i) + C_{\mathrm{u}}\tilde{u}(t_i) \leq (1-\epsilon)b, \tag{4.67}$$

with $\epsilon > 0$, and where $t_i$, denotes the constraint sampling instances, $i = 1, 2, \ldots, n_i$, which are to be determined. The algorithm is based on the following two steps:

1) Compute

$$\begin{aligned} h(t) := \max\ & C_{\mathrm{x}}\tilde{x}(t) + C_{\mathrm{u}}\tilde{u}(t) - b \\ \text{s.t. } & C_{\mathrm{x}}\tilde{x}(t_i) + C_{\mathrm{u}}\tilde{u}(t_i) \leq (1-\epsilon)b,\ i = 1, \ldots, n_i, \\ & (\tilde{x}, \tilde{u}, \lim_{t_{\mathrm{e}}\to\infty} \tilde{x}(t_{\mathrm{e}})) \in \mathcal{D}_{\mathrm{U}}^s, \\ & ||\tilde{x}||^2 + ||\tilde{u}||^2 \leq \bar{J}_s, \end{aligned}$$

   for all times $t \in I_s$, where $I_s$ contains a finite number of sampling instances (to be made precise below).

2) Find the local peaks of $h(t)$, denoted by $t_i^*$. Add each $t_i^*$ to the constraint sampling points if $h(t_i^*) > -b\epsilon/2$. Repeat the procedure until $h(t) \leq -b\epsilon/2$ for all $t \in I_s$.

Note that the function $h(t)$ is again only evaluated at the discrete time points $t \in I_s$. The index set $I_s$ has to be chosen such that $h(t) \leq -b\epsilon/2$ for all $t \in I_s$ implies that $h(t) \leq 0$ for all $t \in [0, \infty)$. As remarked earlier, due to the fact that the basis functions are exponentially decaying and linearly independent it is enough to check $h(t) \leq 0$ for all $t \in [0, T_{\mathrm{c}}]$, for a fixed time $T_{\mathrm{c}}$, as $h(t) \leq 0$ for all $t \in (T_{\mathrm{c}}, \infty)$ will be fulfilled automatically. Moreover, a Lipschitz constant of

$$C_{\mathrm{x}}\tilde{x}(t) + C_{\mathrm{u}}\tilde{u}(t) - b \tag{4.68}$$

can be found by using an upper bound on its time-derivative, that is, for example,

$$|\tau(t)^{\mathsf{T}} M^{\mathsf{T}}((I_s \otimes C_{\mathrm{x}})\eta_x + (I_s \otimes C_{\mathrm{u}})\eta_u)| \leq |\tau(t)||M^{\mathsf{T}}((I_s \otimes C_{\mathrm{x}})\eta_x + (I_s \otimes C_{\mathrm{u}})\eta_u)|, \tag{4.69}$$

where the first term can be bounded for all $t \in [0, \infty)$ due to the fact that the basis functions are exponentially decaying and the second term can be bounded using the fact that the cost $J_s$ is below $\bar{J}_s$. We therefore choose the index set $I_s$ as

$$I_s = \{t_k < T_c \mid t_k = k\frac{2L}{b\epsilon}, k = 0, 1, 2, \dots\}, \tag{4.70}$$

where $L$ denotes a Lipschitz constant of (4.68).

It is important to note that the optimization in step 1 imposes the dynamics and the upper bound $\bar{J}_s$ on the cost. Both constraints tend to reduce the number of constraint sampling points $t_i$ greatly. The initial condition $x_0$ enters the optimization as an optimization variable. The optimization in step 1 represents a quadratically constrained linear program for each time instant $t \in I_s$, and as such, it can be solved using standard software packages. The whole procedure for determining the constraint sampling points is done offline. Once these time instances are found, the optimization problem that is solved online reduces to a quadratic program. The number of constraint sampling points $t_i$ is upper bounded by the cardinality of the index set $I_s$, and thus guaranteed to be finite. Due to the fact that the above procedure is greedy, it will not necessarily lead to the smallest number of constraint sampling points.

In case the presented approach is used for situations with a changing set-point, the upper bound $\bar{J}_s$ might have to be adapted, requiring a re-computation of the constraint sampling points $t_i$. If the upper bound $\bar{J}_s$ is still valid, the constraint sampling points $t_i$ do not have to be recomputed and it suffices to adapt the right-hand-side of (4.67).

*2) Sum-of-squares approximation* In case exponentially decaying polynomials are used as basis functions, c.f. (4.16), sum-of-squares techniques can be applied. In particular, it is shown in [32] that the set

$$\{\eta \in \mathbb{R}^s \mid \eta^\mathsf{T}(1, t, \dots, t^{s-1}) \geq 0, \forall t \in [0, \infty)\} \tag{4.71}$$

can be expressed using matrix inequalities that are linear in the coefficients $\eta$. In the case of exponentially decaying polynomials it is therefore enough to approximate the exponential decay by a polynomial upper bound (for example by appropriately truncating a Taylor series expansion at 0), in order to approximate the constraint (4.66) in a slightly conservative manner. As a result, by applying the results from [32], the optimization problem (4.64) is approximated by a semidefinite program that can be solved using standard optimization routines.

## 4. Part III: An efficient optimization routine

In the following section we present an efficient optimization routine for solving (4.49)

(and likewise (4.64)). The method is an extension of traditional active set methods and generalizes to optimization problems with a linear quadratic cost function, linear equality constraints, and linear semi-infinite inequality constraints, i.e.

$$\hat{J}(I_c) := \min z^{\mathsf{T}} H z \tag{4.72}$$

$$\text{s.t. } A_{\text{eq}} z = b_{\text{eq}}, \tag{4.73}$$

$$l_{\text{b}} \leq (I_{n_c} \otimes \tau(t))^{\mathsf{T}} C_{\text{z}} z \leq l_{\text{u}}, \quad \forall t \in I_c, \tag{4.74}$$

where $I_c$ is any subset of the non-negative real line. In case of (4.64), the interval $I_c$ is taken to be $[0, \infty)$. Note that an optimization problem, whose objective function has a linear part, can be brought to the form (4.72) by completing the squares. It is assumed that the optimization problem (4.72) is feasible, that $l_{\text{b}} \leq 0$ and $l_{\text{u}} \geq 0$, and that the Hessian $H$ is positive definite, which guarantees existence and uniqueness of the corresponding minimizer[28].

The method is based on the observation that if the set $I_c$ consists merely of a collection of time instants (constraint sampling instances) $t_i$, (4.72) reduces to a quadratic program that can be solved efficiently. Moreover, due to the fact that the basis functions fulfill Assumptions A1 and A2, a trajectory parametrized with the basis functions has a finite number of maxima and minima, as is shown in App. F. Consequently, (4.72) has only a finite number of active constraints. The collection of the time instants corresponding to these active constraints will be denoted by $I_c^*$. If this finite collection of constraint sampling instants is known ahead of time, one could simply solve (4.72) with respect to $I_c^*$ instead of $I_c$, resulting in $\hat{J}(I_c^*) = \hat{J}(I_c)$. In addition, for any subset $I_c^k$ of $I_c$ it holds that $\hat{J}(I_c) \geq \hat{J}(I_c^k)$, and likewise, if $I_c^k$ is a subset of $I_c^{k+1}$ we have that $\hat{J}(I_c^{k+1}) \geq \hat{J}(I_c^k)$. Hence, a monotonically increasing sequence $J(I_c^k)$, bounded above by $\hat{J}(I_c)$ can be constructed using any sequence of sets $I_c^k$ that fulfill $I_c^k \subset I_c^{k+1} \subset \cdots \subset I_c$. In particular, such sets are obtained by starting with an arbitrary initial guess $I_c^0$ containing a finite number of constraint sampling points (or even the empty set), and by adding at least one constraint violation point at each iteration. Moreover, at each iteration the inactive constraints contained in the set $I_c^k$ can be removed, as this will not alter the optimizer nor the optimal value $\hat{J}(I_c^k)$. In that way, the number of constraint sampling instances contained in $I_c^k$ remains finite. This motivates Alg. 2, which solves (4.72) up to a given tolerance, by constructing an approximation to the set of active constraints $I_c^*$.

**Proposition 9.** *Alg. 2 converges, that is,* $\lim_{k\to\infty} \hat{J}(I_c^k) = \hat{J}(I_c^*) = \hat{J}(I_c)$. *In order to achieve constraint violations smaller than* $\epsilon$ *at most*

$$\frac{4c_\tau(\hat{J}(I_c) - \hat{J}(I_c^0))}{\sigma\epsilon^2} \tag{4.75}$$

[28]This is due to the fact that the constraints describe a closed convex set and due to the strong convexity of the objective function.

**Procedure 2** Iterative constraint sampling

**Initialize:** initial guess for the constraint sampling points: $I_c^0 = \{t_1, t_2, \ldots, t_N\}$; maximum number of iterations: MAXITER; constraint satisfaction tolerance: $\epsilon$;

```
k = 0
for k <MAXITER do
    solve (4.72) for I_c^k → z^k, Ĵ(I_c^k)
    if infeasible then
        abort
    else if z^k fulfills (4.74) for all t ∈ I (with tol. ε) then
        algorithm converged
        return z^k
    end if
    find at least one constraint violation instant → t_c
    remove inactive time instants in I_c^k
    I_c^{k+1} = I_c^k ∪ {t_c}, k = k + 1
end for
```

*steps are required, where $c_\tau$ is defined as*

$$c_\tau := \sup_{t \in I_c} |\tau(t)|^2, \tag{4.76}$$

*and $\sigma$ denotes the smallest eigenvalue of the Hessian $H$.*

*Proof.* From the above arguments it can be concluded that $\hat{J}(I_c^k)$ is monotonically increasing whenever $\hat{J}(I_c^k) < \hat{J}(I_c)$. Therefore the sequence $\hat{J}(I_c^k)$ converges. It remains to show that $\hat{J}(I_c^k)$ converges to $\hat{J}(I_c)$. Similar to Prop. 6, the strong convexity of the objective function can be used to establish

$$|z^{k+1} - z^k|^2 \leq 4\sigma^{-1}(\hat{J}(I_c^{k+1}) - \hat{J}(I_c^k)), \tag{4.77}$$

where $z^{k+1}$ and $z^k$ are the minimizer corresponding to $\hat{J}(I_c^k)$ and $\hat{J}(I_c^{k+1})$. As a result we can conclude that $z^k$ converges, and that $\lim_{k\to\infty} z^k$ is well defined and satisfies the constraint (4.74). It is therefore a feasible candidate for (4.72), implying that $\lim_{k\to\infty} \hat{J}(I_c^k) \geq \hat{J}(I_c)$, which, combined with $\hat{J}(I_c^k) \leq \hat{J}(I_c)$ for all integers $k$, leads to $\lim_{k\to\infty} \hat{J}(I_c^k) = \hat{J}(I_c)$.

It remains to show that (4.75) is fulfilled. To that extent, let $\epsilon > 0$ denote the smallest constraint violation that occurs within the first $N$ steps. For all $k \leq N-1$, there exists the constraint violation point $t_j^k$, which will be added to $I_c^k$, and therefore

$$\epsilon \leq |(e(t_j^k) \otimes \tau(t_j^k))^\mathsf{T}(z^k - z^{k+1})|, \tag{4.78}$$

where $e(t_j^k)$ is a canonical unit vector. Combining the Cauchy-Schwarz inequality and the

above bound on $|z^{k+1} - z^k|$ results in

$$\epsilon^2 \leq 4c_\tau \sigma^{-1}(\hat{J}(I_c^{k+1}) - \hat{J}(I_c^k)). \tag{4.79}$$

By summing over the first $N$ steps we arrive at

$$N\epsilon^2 \leq 4\sigma^{-1}c_\tau(\hat{J}(I_c^N) - \hat{J}(I_c^0)) \leq 4\sigma^{-1}c_\tau(\hat{J}(I_c^*) - \hat{J}(I_c^0)), \tag{4.80}$$

since the sequence $\hat{J}(I_c^k)$ is strictly increasing and bounded above by $\hat{J}(I_c^*)$. Dividing by $\epsilon^2$ on both sides concludes the proof. □

### 4.1 Implementation details

Alg. 2 can be naturally embedded in an active-set method. An introduction to active-set methods for solving quadratic programs can be found for example in [40, Ch. 10]. Starting from an initial guess of the active constraint sampling instants, which is denoted by $I_c^0$, the quadratic program with optimal cost $\hat{J}(I_c^0)$ is solved: This is done by initially assuming that all constraints in the set $I_c^0$ are active. The resulting optimization problem reduces to an equality constrained quadratic program, whose solution can be calculated by solving a linear system of equations. The Lagrange multipliers corresponding to (4.74), which are denoted by $\mu(t)$, $t \in I_c^0$, are evaluated subsequently. If all the constraints are indeed active, then the optimizer to the quadratic program with cost $\hat{J}(I_c^0)$ has been found. If, however, not all the constraints are found to be active, that is, if there are some Lagrange multipliers that are zero, the standard active-set procedure, see [40, Ch. 10] is used to find the subset of active constraints $I_a \subset I_c^0$. Provided that the active set $I_a \subset I_c^0$ and the optimizer to the quadratic program with cost $\hat{J}(I_c^0)$ has been found, the constraint (4.74) is then checked for all $t \in I_c$. If no constraint violations occur, the solution to (4.72) has been found. If constraint violations occur, the time instant $t_c$ for which a violation occurs, is added to the set of active constraints $I_a$ resulting in $I_c^1 = I_a \cup \{t_c\}$. The above procedure is then repeated until convergence.

Each iteration requires solving equality constrained quadratic programs of the type

$$\min z^\mathsf{T} H z \tag{4.81}$$

$$\text{s.t. } A_{\text{eq}} z = b_{\text{eq}}, \quad (c(t) \otimes \tau(t))^\mathsf{T} C_\text{z} z = l_\text{a}(t), \quad \forall t \in I_a, \tag{4.82}$$

where $c(t) \in \mathbb{R}^{n_c}$, $l_\text{a}(t) \in \mathbb{R}$, $t \in I_a$, and $I_a \subset I_c^k$ describe the active constraints corresponding to (4.74). Due to the fact that very few constraints are expected to be active, we use a range space approach, [40, p. 238]. To that extent, the equality constraint is eliminated and the optimizer $z^*$ corresponding to (4.81) is rewritten as

$$z^* = \hat{b} + \hat{H} C_\text{z}^\mathsf{T} \sum_{t \in I_a} (c(t) \otimes \tau(t))^\mathsf{T} \mu(t), \tag{4.83}$$

where the dual variable $\mu(t) \in \mathbb{R}$, defined for $t \in I_c^k$, satisfies

$$(c(t_j) \otimes \tau(t_j))^{\mathsf{T}} C_{\mathrm{z}} \hat{H} C_{\mathrm{z}}^{\mathsf{T}} \sum_{t \in I_a} (c(t) \otimes \tau(t)) \mu(t) \;=\; l_{\mathrm{a}}(t_j) - (c(t_j) \otimes \tau(t_j))^{\mathsf{T}} C_{\mathrm{z}} \hat{b}, \quad (4.84)$$

for all $t_j \in I_a \subset I_c^k$, and $\mu(t) = 0$ for all $t \in I_c^k \setminus I_a$, with

$$\hat{H} := H^{-1} A_{\mathrm{eq}}^{\mathsf{T}} (A_{\mathrm{eq}} H^{-1} A_{\mathrm{eq}}^{\mathsf{T}})^{-1} A_{\mathrm{eq}} H^{-1} - H^{-1}, \quad (4.85)$$

$$\hat{b} := H^{-1} A_{\mathrm{eq}}^{\mathsf{T}} (A_{\mathrm{eq}} H^{-1} A_{\mathrm{eq}}^{\mathsf{T}})^{-1} b. \quad (4.86)$$

The dual variable $\mu(t)$ is therefore obtained by solving (4.84), and the optimizer $z^*$ is then determined via (4.83). At each iteration, a single constraint is either added or removed. Therefore, in order to efficiently find a solution to (4.84), a Cholesky factorization (more precisely a $LDL^{\mathsf{T}}$-decomposition) of the matrix

$$\{(c(t_j) \otimes \tau(t_j))^{\mathsf{T}} C_{\mathrm{z}} \hat{H} C_{\mathrm{z}}^{\mathsf{T}} (c(t_i) \otimes \tau(t_i))\}_{(t_j, t_i) \in I_a \times I_a} \quad (4.87)$$

is computed and adapted at each step by performing rank-1 updates. The matrix $\hat{H}$ and the vector $\hat{b}$ are precomputed. For additional details regarding the regularity of (4.87), and issues related to cycling and stalling we refer to [40, Ch. 10] and [15, p. 467].

### 4.2 Constraint check

It remains to explain how to efficiently check whether the constraint (4.74) is fulfilled for a given solution candidate $z$. We assume that the interval $I_c$ has the form $I_c = [0, T_c]$. As it has been explained in Sec. 3.1, the constraint check over the interval $[0, \infty)$ reduces to the check over a compact interval, provided that the basis functions fulfill Assumptions A1 and A2.

A straightforward approach would be to exploit the specific structure of the basis functions. For example, if the basis functions consist of exponentially decaying polynomials having a degree of at most 4, determining the stationary points of (4.74) amounts to solving a quartic equation, which can be done analytically. As a result, it would be enough to check the constraints at these stationary points in order to determine if the constraint is satisfied or not.

We propose a more general approach that is based on local Taylor approximations, and thus valid for arbitrary basis functions compatible with Assumptions A1 and A2. In order to simplify the discussion, we consider the special case of (4.74), where $C_{\mathrm{z}}$ is the identity and $n_c = 1$. The resulting algorithm extends naturally to the more general case. According to Taylor's theorem we obtain the following identity

$$\tau(t)^{\mathsf{T}} z = \tau(0)^{\mathsf{T}} z + \dot{\tau}(0)^{\mathsf{T}} z t + \ddot{\tau}(0)^{\mathsf{T}} z \frac{t^2}{2} + \tau^{(3)}(\bar{t})^{\mathsf{T}} z \frac{t^3}{6}, \quad (4.88)$$

where $\bar{t} \in [0, t]$. As will become clear in the following, a third-order Taylor expansion represents a good compromise between approximation quality and computational effort. The last term of the previous equation can be bounded by the Cauchy-Schwarz inequality leading to

$$|\tau^{(3)}(s)^\mathsf{T} z| \leq \sup_{\bar{t} \in I_c} |\tau(\bar{t})||(M^\mathsf{T})^3 z| =: R(z), \quad \forall s \in [0, t]. \tag{4.89}$$

As a consequence, the following upper and lower bounds are obtained

$$b_\mathrm{l}(t) \leq \tau(t)^\mathsf{T} z \leq b_\mathrm{u}(t), \tag{4.90}$$

for all $t \in I_c$, with

$$b_\mathrm{l}(t) := \tau(0)^\mathsf{T} z + \dot{\tau}(0)^\mathsf{T} z t + \ddot{\tau}(0)^\mathsf{T} z \frac{t^2}{2} - R(z)\frac{t^3}{6}, \tag{4.91}$$

$$b_\mathrm{u}(t) := \tau(0)^\mathsf{T} z + \dot{\tau}(0)^\mathsf{T} z t + \ddot{\tau}(0)^\mathsf{T} z \frac{t^2}{2} + R(z)\frac{t^3}{6}. \tag{4.92}$$

The situation is exemplarily depicted in Fig. 4.3. Given that $\dot{\tau}(0)^\mathsf{T} z \geq 0$ the lower bound attains its maximum at time

$$t_\mathrm{u} := \frac{\ddot{\tau}(0)^\mathsf{T} z + \sqrt{(\ddot{\tau}(0)^\mathsf{T} z)^2 + 2R(z)\dot{\tau}(0)^\mathsf{T} z}}{R(z)} > 0, \tag{4.93}$$

whereas if $\dot{\tau}(0)^\mathsf{T} z < 0$ the upper bound attains its minimum at time

$$t_\mathrm{l} := \frac{-\ddot{\tau}(0)^\mathsf{T} z + \sqrt{(\ddot{\tau}(0)^\mathsf{T} z)^2 - 2R(z)\dot{\tau}(0)^\mathsf{T} z}}{R(z)} > 0. \tag{4.94}$$

Thus, if the lower bound exceeds $l_\mathrm{u}$ or the upper bound drops below $l_\mathrm{l}$, that is, if

$$b_\mathrm{l}(t_\mathrm{u}) > l_\mathrm{u}, \quad \text{or} \quad b_\mathrm{u}(t_\mathrm{l}) < l_\mathrm{l}, \tag{4.95}$$

the constraint (4.74) is guaranteed to be violated at time $t_\mathrm{u}$, respectively at time $t_\mathrm{l}$. If this is not the case, we are guaranteed that the constraint (4.74) is satisfied in the interval $[0, t_\mathrm{s}]$, with

$$t_\mathrm{s} := \min\{t_1, t_2 \mid b_\mathrm{u}(t_1) = l_\mathrm{u}, b_\mathrm{l}(t_2) = l_\mathrm{l}\}. \tag{4.96}$$

Thus, finding the value $t_\mathrm{s}$ requires the solution of two cubic equations, which stems from the fact that a third order Taylor approximation was used as a starting point. By shifting the parameter vector in time by $t_\mathrm{s}$ and repeating the above procedure, the constraint is

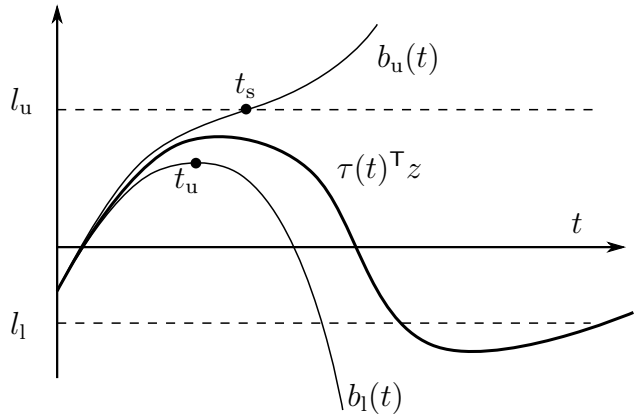


**Figure 4.3.** Illustration of the upper and lower bounds $b_{\mathrm{u}}(t)$ and $b_{\mathrm{l}}(t)$ obtained from the Taylor expansion of $\tau(t)^{\mathsf{T}}z$. In that case the test (4.95) is indecisive, and constraint satisfaction over the interval $[0, t_{\mathrm{s}}]$ can be guaranteed.

either found to be satisfied for all $t \in I_c$ or a constraint violation is detected. Shifting the parameter vector by $t_{\mathrm{s}}$ amounts in multiplying $z$ with the matrix exponential

$$e^{M^{\mathsf{T}}t_{\mathrm{s}}}z \to z. \tag{4.97}$$

The procedure is illustrated by the flow chart depicted in Fig. 4.4.

The efficiency of the proposed strategy can be improved via the following observation. A constraint violation occurring close to $t = 0$, is often found within the first few iterations. However, if no constraint violation occurs, the whole interval $I_c$ needs to be traversed, which tends to increase computation. The computational effort may be reduced by including additional conservative constraint satisfaction checks. For example, upper and lower bounds can be tightened by a factor $\gamma \in (0, 1)$ such that the satisfaction of

$$\gamma l_{\mathrm{l}} \leq \tau(t_i)^{\mathsf{T}}z \leq \gamma l_{\mathrm{u}} \tag{4.98}$$

for certain time instances $t_i$ implies (4.74) (for all $t \in I_c$). As a result, at each iteration, the above inequality is checked. If it is found to be fulfilled, then constraint satisfaction can be guaranteed.

## 5. Simulation example

The proposed approach is illustrated on an quadruple integrator system, that is,

$$x_{\mathrm{qi}}^{(4)} = u, \tag{4.99}$$

where $x_{\mathrm{qi}}$ corresponds to the quadruple integrator state and $u$ to the input. The example is

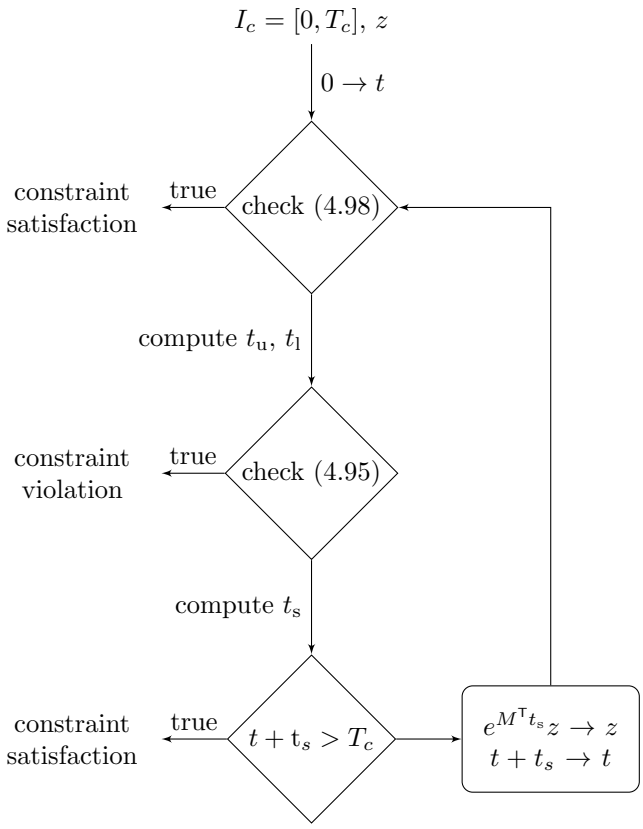


**Figure 4.4.** Flow chart that illustrates the proposed constraint satisfaction check.

used to highlight the potential of the proposed MPC approach. However, we do not claim that our approach is superior in general, but believe that it leads to a different trade-off (basis function complexity vs. computation) compared to "standard" MPC, which might be beneficial for some applications. We define the state vector as $x := (x_\mathrm{qi}, \dot{x}_\mathrm{qi}, \ddot{x}_\mathrm{qi}, \dddot{x}_\mathrm{qi})$ and consider the task of driving the system from $x(0) = x_0$ back to the origin. We penalize input and state deviations with the following cost

$$\int_0^\infty \frac{1}{2} x^\mathsf{T} x + \frac{1}{2} u^2 \, \mathrm{d}t, \tag{4.100}$$

and constrain the input $u$ to lie in $[-0.5, 0.5]$ and the state $x_\mathrm{qi}$ to be non-negative, that is, $x_\mathrm{qi} \geq 0$. The basis functions $\tau$ are designed to be orthonormal and spanned by

$$\tau \in \exp(-\nu t) \, \mathrm{span}(1, t, t^2, \ldots, t^{s-1}), \tag{4.101}$$

where $\nu$ is set to $0.7\,\mathrm{s}^{-1}$ (this corresponds approximately to the closed-loop poles of an LQR design). This leads to so-called Laguerre functions, c.f. (4.16) that fulfill Assumptions A1 and A2. Note that the theorem of Stone-Weierstrass, [41, p. 147], states that the basis functions given by (4.101) are dense in the set of continuous functions vanishing

at infinity. The set of smooth compactly supported functions is contained in the set of continuous functions vanishing at infinity, [42, p. 70] and as a result, the assumptions of Lemma 7 are fulfilled.[29]

The open-loop state and input trajectories resulting from solving (4.49) with $x_0 = (0.3, 0.3, 0.3, 0.3)$ and $s = 12$ are shown in Fig. 4.5. The resulting cost amounts to $J_{12} = 13.95$. By solving the problem (4.52), we obtain the lower bound $\tilde{J}_{12} = 11.0$. Thus we can conclude that the cost corresponding to (4.1) is at most 20% below $J_{12}$.

Fig. 4.5 compares open-loop and closed-loop trajectories. The closed-loop trajectories are obtained when resolving the optimization problem (4.49) every $T_\mathrm{d} = 20$ms, and applying the obtained input trajectory $\tilde{u}(t)$ in between. In practice this could be realized with two different processes running at different frequencies, one solving (4.49) at a slower rate and one applying the input $\tilde{u}(t)$ at a higher rate.[30] Closed-loop and open-loop trajectories are significantly different, which is due to the fact that a high polynomial order is required to approximate the bang-bang behavior accurately. The achieved closed-loop cost is $J_{12\mathrm{cl}} = 13.06$, lying between $\tilde{J}_{12}$ and $J_{12}$.

Next, the proposed parametrized MPC approach is benchmarked against the discrete-time MPC approach used by FORCES, [12], and qpOASES, [13]. The MPC solver FORCES implements an interior point method that exploits the so-called multistage formulation obtained from the discrete-time MPC formulation. The quadratic programming solver qpOASES implements an active set method tailored to MPC. A terminal cost that matches an LQR design is included in the discrete-time formulation. No terminal set constraint is added, hence closed-loop stability is not guaranteed in the discrete-time approach. The evolution of the system, starting from $x_0$ is simulated over 20s and is used to compute the closed-loop cost according to

$$\sum_{k=1}^{1000} \left( \frac{1}{2} x(kT_\mathrm{d})^\mathsf{T} x(kT_\mathrm{d}) + \frac{1}{2} u(kT_\mathrm{d})^2 \right) T_\mathrm{d}. \tag{4.102}$$

The time horizon (in the discrete-time approach) is increased from 93 to 130 samples with a step length of two samples. A time horizon of 92 samples was found to yield an unstable closed-loop trajectory for the given initial condition. The optimization routine FORCES was run with the standard settings, including an absolute tolerance of $10^{-6}$ for the duality gap and a constraint satisfaction tolerance of $10^{-6}$. The standardized MPC-settings were used for qpOASES. For the parametrized approach the number of basis functions is increased from 8 to 12. The semi-infinite constraint is handled via the active-set method proposed in Sec. 4, where an absolute tolerance of $10^{-6}$ is used for

[29] The exponential decay can be expressed using polynomials, which shows that the basis functions form an algebra.

[30] We applied the proposed parametrized approach in practice, see [43], [44], by using zero-order hold. Although the stability guarantees are lost when applying zero-order hold, we did not experience any issues due to the robustness of MPC. By relying on a discrete-time formulation of our parametrized approach closed-loop stability can be guaranteed with zero-order hold.

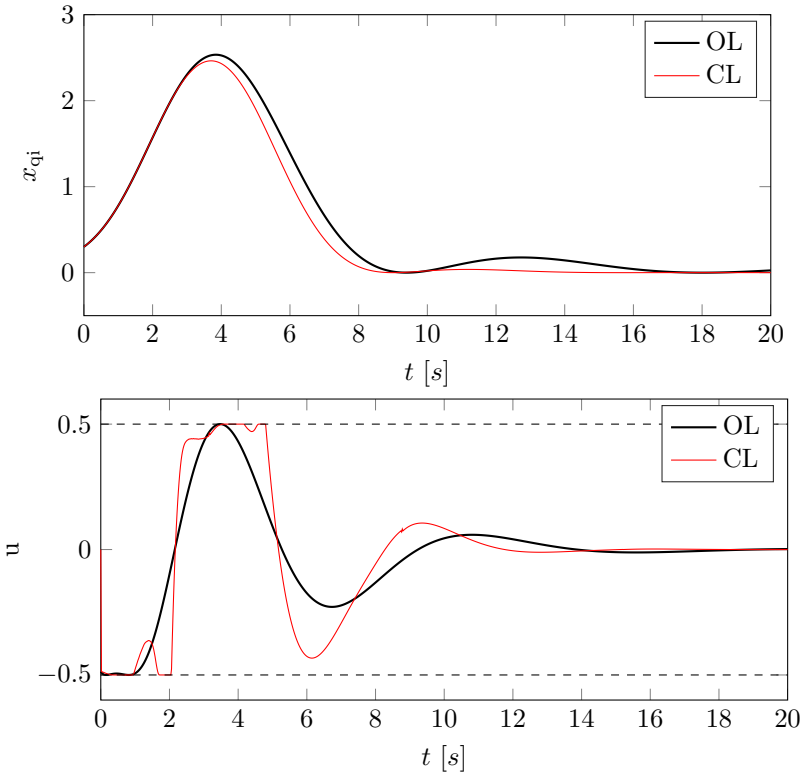


**Figure 4.5.** Open-loop (black, thick) and closed-loop (red,thin) trajectories for $s = 12$. The quadruple integrator state $x_{\mathrm{qi}}$ is depicted in the first graph (top), and the input is shown in the second graph (bottom).

constraint satisfaction. The results are displayed in Fig. 4.6, where the average execution time (averaged over the 20 s simulation) is plotted as a function of the achieved closed-loop cost. The parametrized MPC approaches is shown to outperform FORCES and qpOASES by up to one order of magnitude in terms of the average execution time. The achieved closed-loop cost with parametrized approach increases for $s = 11, 12$ compared to $s = 10$. This can most probably be attributed to the relative large discrepancy between closed-loop and open-loop trajectories.

The benchmark is repeated for 100 random initial conditions, that are uniformly distributed in $[0, 0.2]^4$. The corresponding results are shown in Fig. 4.7. Note that the initial conditions are guaranteed to be stabilizable, and were indeed stabilized by the parametrized MPC approach. In the discrete-time case, a time horizon below 36 was found to yield unstable closed-loop trajectories. In addition, Fig. 4.7 displays the sensitivity of the execution time of the parametrized approach with respect to the chosen constraint satisfaction tolerance. The constraint satisfaction tolerance was also varied for FORCES, but found to influence the execution time only insignificantly. In the result shown the constraint satisfaction tolerance for FORCES was set to $10^{-6}$.

In order to demonstrate the scalability of the parametrized approach with the number

of states, a chain of $n$ integrators is considered,

$$x_{ni}^{(n)} = u, \tag{4.103}$$

where $x_{ni}$ corresponds to the first integrator state and $u$ to the input. We define the state vector to be $x := (x_{ni}, \dot{x}_{ni}, \ldots, x_{ni}^{(n-1)})$, and penalize input and state deviations with the cost function (4.100). The basis function are chosen to be orthonormal and spanned according to (4.101) with $\nu = 0.7\,\mathrm{s}^{-1}$ and $s = 12$. The input constraint $u \in [-0.5, 0.5]$ and the state constraint $x_{ni} \geq 0$ is included. The execution time required to compute a single solution of (4.49) subject to the initial condition $x_0 = (0.1, 0.1, \ldots, 0.1)$ is shown in Fig. 4.8. No feasible solution was found with the given basis functions for values of $n$ larger than 7. For $n = 7$ a time horizon of approximately 450 samples is required to achieve closed-loop stability with the discrete-time formulation, which amounts to 3607 optimization variables. In contrast, for $s = 12$ and $n = 7$ the optimization problem resulting from the parametrized approach includes 91 optimization variables.

Summarizing, we can conclude that the parametrized approach might be promising, in particular for systems with marginally stable or unstable dynamics that require high sampling frequencies. On the example of the quadruple integrator system, the parametrized approach outperformed the standard discrete-time approach in terms of execution time, without necessarily degrading performance. We believe that the computational advantages stem from a reduction in the number of optimization variables and the fact that only very few constraints are typically active (consider for example the open-loop trajectory shown in Fig. 4.5), which is exploited by our active-set approach. Hence, it is conjectured that the computational benefits are even higher for systems with higher state and/or input dimensions.

## 6. Conclusion

The article discussed approximations to the constrained linear quadratic regulator problem, which are based on representing input and state trajectories with basis functions. In particular, a sequence of lower and upper bounds on the cost of the underlying optimal control problem is derived. The approximations are shown to converge. The proposed framework is applied to MPC, where it is shown that an infinite prediction horizon can be retained, leading to recursive feasibility and closed-loop stability. Efficient solution methods are presented to solve the resulting finite-dimensional convex optimization problems. The results are illustrated on a quadruple integrator system. The proposed approach leads to different computational trade-offs compared to "standard" MPC, which might be beneficial for some applications. In case of the quadruple integrator system, it is shown to outperform the state-of-the-art discrete-time solvers in terms of execution time, without necessarily degrading performance.

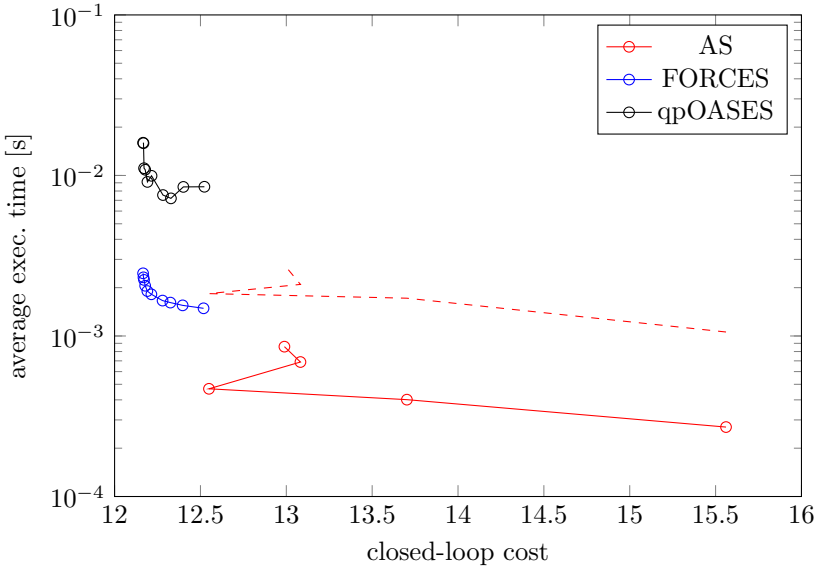


**Figure 4.6.** Shown is the average execution time as a function of the closed-loop cost obtained from a 20s simulation starting at $x_0$. The dashed line indicates the 68% confidence interval of the execution time (plus one standard deviation) for the parametrized approach. The execution time was found to vary only insignificantly for FORCES and qpOASES, and hence, the standard deviation is not shown. To trade-off the execution time with the closed-loop cost, the prediction horizon is changed from 93 to 130 (first in steps of two, then in steps of five) in the discrete-time formulation (used by FORCES and qpOASES), whereas the number of basis functions is increased from 8 to 12 in the parametrized formulation.

## A. Properties B1-B5

We will sketch the proofs of Properties B1-B5 in the following. It follows from the linearity of the stage constraint, and the fact that the terminal constraint $\mathcal{X}$ is convex, that the sets $\mathcal{C}$, $\mathcal{C}^s_{\mathrm{U}}$, and $\mathcal{C}^s_{\mathrm{L}}$ are likewise convex. We will sketch the proof that the set $\mathcal{C}^s_{\mathrm{L}}$ is closed. The argument can be translated to the set $\mathcal{C}$ by using the formulation according to (4.25). We will argue indirectly, i.e. that the complement of $\mathcal{C}^s_{\mathrm{L}}$ is open. To that extent we choose $(x, u, x_T) \in X \setminus \mathcal{C}^s_{\mathrm{L}}$. As a result, there exists a test function $\delta\tilde{p}$, with $\delta\tilde{p} \geq 0$, which is spanned by the first $s$ basis functions and is such that

$$\int_I \delta\tilde{p}^{\mathsf{T}}(-C_{\mathrm{x}}x - C_{\mathrm{u}}u + b)\mathrm{d}t < 0. \tag{4.104}$$

For any $\hat{x} \in L^2_n$, $\hat{u} \in L^2_m$, and $\hat{x}_T \in \mathbb{R}^n$, with $||\hat{x} - x|| < \epsilon$, $||\hat{u} - u|| < \epsilon$, and $|\hat{x}_T - x_T| < \epsilon$, it follows that

$$\begin{aligned}\int_I \delta\tilde{p}^{\mathsf{T}}(-C_{\mathrm{x}}\hat{x} - C_{\mathrm{u}}\hat{u} + b)\mathrm{d}t = \int_I \delta\tilde{p}^{\mathsf{T}}(-C_{\mathrm{x}}x - C_{\mathrm{u}}u + b)\mathrm{d}t \\ + \int_I \delta\tilde{p}^{\mathsf{T}}(C_{\mathrm{x}}(x - \hat{x}) + C_{\mathrm{u}}(u - \hat{u}))\mathrm{d}t,\end{aligned} \tag{4.105}$$

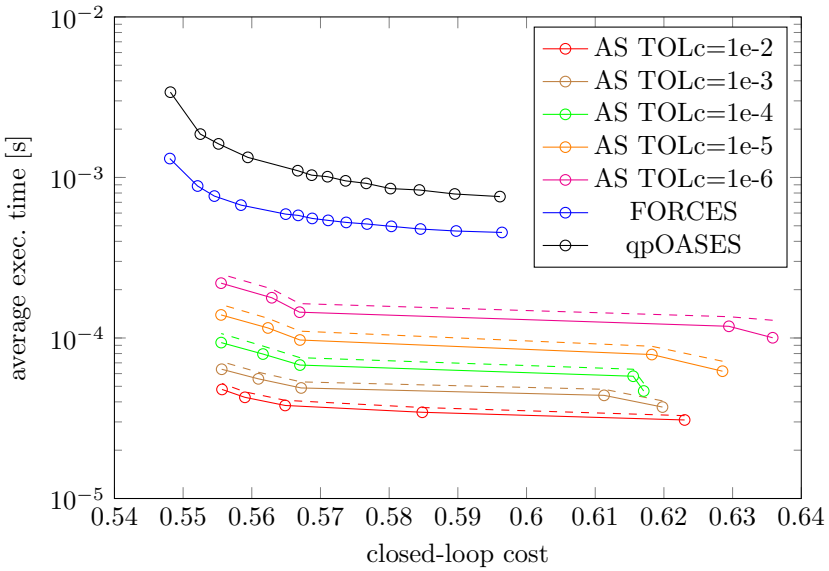


**Figure 4.7.** Simulation of 100 random initial conditions. Depicted is the average execution time as a function of the average closed-loop cost. The dashed line indicates the 68% confidence interval of the average execution time for the parametrized approach. The execution time was found to vary only insignificantly for FORCES and qpOASES, and therefore the corresponding standard deviation is not shown. In order to trade-off the execution time with the closed-loop cost, the prediction horizon is changed from 36 to 60 (36,38,40,42,44,45,50,55,60) in the discrete-time formulation (used by FORCES and qpOASES), whereas the number of basis functions is increased from 8 to 12 in the parametrized formulation.

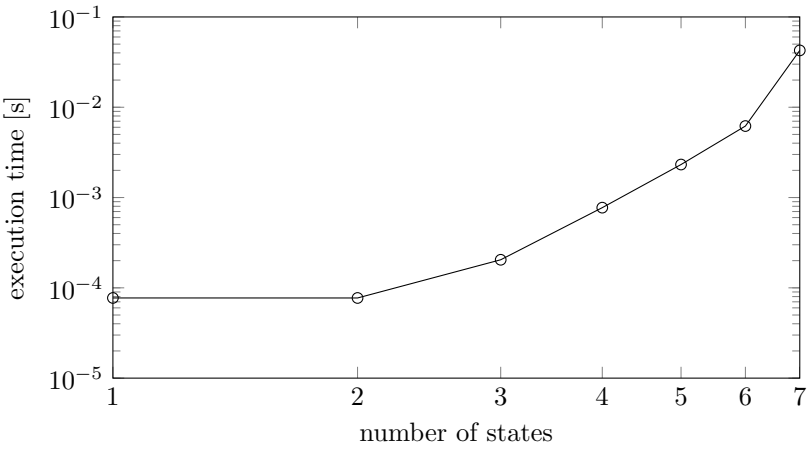


**Figure 4.8.** Execution time required for solving (4.49) subject to the initial condition $x_0 = (0.1, 0.1, \ldots, 0.1)$ for the $n$th order integrator system.

where the last integral can be bounded by (using the Cauchy-Schwarz inequality)

$$||\delta\tilde{p}||(|C_{\mathrm{x}}|||x - \hat{x}|| + |C_{\mathrm{u}}|||u - \hat{u}||) < ||\delta\tilde{p}||(|C_{\mathrm{x}}| + |C_{\mathrm{u}}|)\epsilon. \tag{4.106}$$

We can infer from $\mathcal{X}$ being closed, that there exists an open ball centered at $x_T$, which

does not intersect $\mathcal{X}$. As a result, by choosing $\epsilon$ small enough, it can be concluded that

$$\int_I \delta\tilde{p}^{\mathsf{T}}(-C_{\mathrm{x}}\hat{x} - C_{\mathrm{u}}\hat{u} + b)\mathrm{d}t < 0, \quad \hat{x}_T \notin \mathcal{X}, \tag{4.107}$$

and therefore $(\hat{x}, \hat{u}, x_T) \in X \setminus \mathcal{C}_{\mathrm{L}}^s$, for all $\hat{x} \in L_n^2$, $\hat{u} \in L_m^2$, $\hat{x}_T \in \mathbb{R}^n$ with $||x - \hat{x}|| < \epsilon$, $||u - \hat{u}|| < \epsilon$, and $|\hat{x}_T - x_T| < \epsilon$. Hence, the complement of $\mathcal{C}_{\mathrm{L}}^s$ is open, and therefore $\mathcal{C}_{\mathrm{L}}^s$ is closed.

Note that the set $\mathcal{C}_{\mathrm{U}}^s$ is given by the intersection of the set $\mathcal{C}$ with the linear subspace $X^s$ spanned by the first $s$ basis functions. Both of these sets are closed[31] implying that $\mathcal{C}_{\mathrm{U}}^s$ is closed as well.

The projection $\pi^s$ is linear, which asserts the convexity of the sets $\pi^s(\mathcal{C}_{\mathrm{U}}^s)$ and $\pi^s(\mathcal{C}_{\mathrm{L}}^s)$. Moreover, it is surjective, and hence, by the open mapping theorem, it follows directly from $\mathcal{C}_{\mathrm{L}}^s$ and $\mathcal{C}_{\mathrm{U}}^s$ being closed that $\pi^s(\mathcal{C}_{\mathrm{L}}^s)$ and $\pi^s(\mathcal{C}_{\mathrm{U}}^s)$ are closed as well.

The inclusion $\mathcal{C}_{\mathrm{U}}^s \subset \mathcal{C}_{\mathrm{U}}^{s+1}$ follows directly from $\mathcal{C}_{\mathrm{U}}^s = \mathcal{C} \cap X^s$ and the inclusion $X^s \subset X^{s+1}$. In other words, given $(x, u, x_T) \in \mathcal{C}_{\mathrm{U}}^s$, the parameter vectors $\eta_x$ and $\eta_u$, corresponding to the state and input trajectories $x$ and $u$, can be extended with zeros resulting in trajectories $\hat{x}$, $\hat{u}$ spanned by $s+1$ basis functions. But $\hat{x} = x$ and $\hat{u} = u$ and therefore $(x, u, x_T) \in \mathcal{C}_{\mathrm{U}}^{s+1}$. The inclusion $\mathcal{C}_{\mathrm{L}}^s \supset \mathcal{C}_{\mathrm{L}}^{s+1}$ follows from the fact that the dimension of the subspace to which the test functions $\delta\tilde{p}$ are constrained increases with $s$.

The claim $(\pi^s)^*\pi^s(\mathcal{C}_{\mathrm{U}}^s) \subset \mathcal{C}_{\mathrm{U}}^s$ follows from the fact that $(\pi^s)^*\pi^s$ is a projection from $X$ onto $X^s \subset X$ and that $\mathcal{C}_{\mathrm{U}}^s \subset X^s$. The claim that $(\pi^s)^*\pi^s(\mathcal{C}_{\mathrm{L}}^s) \subset \mathcal{C}_{\mathrm{L}}^s$ follows from the linearity of the stage constraints. More precisely, it follows by noting that for any $\delta\tilde{p} = (I_{n_c} \otimes \tau)^{\mathsf{T}}\delta\eta_p$ and any $x \in L_n^2$,

$$\begin{aligned}\int_I \delta\tilde{p}^{\mathsf{T}} C_{\mathrm{x}} x \mathrm{d}t &= \delta\eta_p^{\mathsf{T}} C_{\mathrm{x}} \pi^{ns}(x) \\ &= \int_I \delta\tilde{p}^{\mathsf{T}} C_{\mathrm{x}}\ (\pi^{ns})^*\pi^{ns}(x)\mathrm{d}t\end{aligned} \tag{4.108}$$

holds.

## B. Properties C1-C5

We will sketch the proof of Properties C1-C5 below.

The convexity of the sets $\mathcal{D}_{\mathrm{U}}^s$, $\mathcal{D}_{\mathrm{L}}^s$, and $\mathcal{D}$ follows directly from the linearity of the dynamics.

The fact that the set $\mathcal{D}_{\mathrm{L}}^s$ is closed can be seen by a similar argument used for showing closedness of $\tilde{\mathcal{C}}^s$ in App. A, that is, showing that $X \setminus \mathcal{D}_{\mathrm{L}}^s$ is open. The set $\mathcal{D}$ can be

[31] $X^s$ is finite-dimensional, thus complete, and hence also closed.

rewritten using the variational equality (4.44), and thus, again a similar argument can be applied to show that $\mathcal{D}$ is closed. It follows that $\mathcal{D}_{\mathrm{U}}^s$ is closed, since $\mathcal{D}_{\mathrm{U}}^s$ is defined as the intersection of $\mathcal{D}$ with the closed set $X^s$.

The linearity and surjectivity of the map $\pi^s$ implies that $\pi^s(\mathcal{D}_{\mathrm{U}}^s)$ and $\pi^s(\mathcal{D}_{\mathrm{L}}^s)$ are indeed closed (by the open mapping theorem) and convex.

The inclusion $\mathcal{D}_{\mathrm{U}}^s \subset \mathcal{D}_{\mathrm{U}}^{s+1} \subset \mathcal{D}$ for all $s$ follows directly from the fact that $X^s \subset X^{s+1} \subset X$ and $\mathcal{D}_{\mathrm{U}}^s = \mathcal{D} \cap X^s$. The inclusion $\mathcal{D}_{\mathrm{L}}^{s+1} \subset \mathcal{D}_{\mathrm{L}}^s$ for all $s$ can be seen by noting that the variational equality in (4.45) has to hold for variations spanned by more and more basis functions as $s$ increases. The claim that $\mathcal{D}$ is contained in $\mathcal{D}_{\mathrm{L}}^s$ follows from the equivalence of (4.44) with the formulation in (4.40). The properties $(\pi^s)^*\pi^s(\mathcal{D}_{\mathrm{U}}^s) \subset \mathcal{D}_{\mathrm{U}}^s$ and $(\pi^s)^*\pi^s(\mathcal{D}_{\mathrm{L}}^s) \subset \mathcal{D}_{\mathrm{L}}^s$ can be shown using the same arguments as in App. A, where the latter relies on the linearity of the dynamics.

## C. Proof of Lemma 7 (infinite measure case)

We prove Lemma 7 for the case where $I$ has infinite measure.

*Proof.* The idea of the proof is the same as in the finite measure case: We claim that $\lim_{s\to\infty} \mathcal{C}_{\mathrm{U}}^s \supset \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$. We assume that the claim is incorrect and show that this leads to a contradiction. Thus, we choose $(x, u, x_T) \in \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$, such that there exists an open set $U$ (bounded) and a $k \in \{1, 2, \ldots, n_c\}$, for which

$$\int_I \delta v(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t < 0 \tag{4.109}$$

holds for all smooth test functions $\delta v : I \to \mathbb{R}$, $\delta v \geq 0$, with support in $U$, and $\delta v(t_0) > 0$ for some $t_0 \in U$. Due to the smoothness of the test functions, $\delta v(t_0) > 0$ readily implies that there is an open neighborhood of $t_0$, denoted by $\mathcal{N}(t_0)$, such that $\delta v(t) > 0$, $\forall t \in \mathcal{N}(t_0)$. The above integral exists, since $\delta v$ is bounded, has compact support and $x \in L_n^2$, $u \in L_m^2$. We fix $t_0 \in U$ and pick one of these variations that is positive, strictly positive at time $t_0$, and has support in $U$, which we name $\delta p$. Due to the fact that the basis functions are dense in the set of smooth functions with compact support, there exists a sequence $\sqrt{\delta \tilde{p}_i}$ that converges uniformly to $\sqrt{\delta p}$. It was shown in the proof of Lemma 7 that $\delta \tilde{p}_i$ lies likewise in the span of the basis functions and that for a given $\epsilon > 0$ (small enough) there exits an integer $N > 0$ such that

$$||\delta \tilde{p}_i - \delta p||_\infty < C_1 \epsilon$$

holds for all integers $i > N$, where $C_1 > 0$ is constant.

We claim that there is an integer $p > 0$ such that the basis function $\tau_p : I \to \mathbb{R}$ is nonzero for $t_0$. This can be shown by a contradiction argument: If the claim was

not true, then all basis functions $\tau^p = (\tau_1, \tau_2, \ldots, \tau_p)$ would be zero at time $t_0$. The basis functions fulfill Assumption A2, and hence the first order differential equation $\dot{\tau}^p = M_p \tau^p$, which is guaranteed to have unique solutions. From $\tau^p(t_0) = 0$, $\dot{\tau}^p(t_0) = 0$ we can infer that $\tau^p(t) = 0$ for all $t \in (0, \infty)$ is the (unique) solution to $\dot{\tau}^p = M_p \tau^p$, contradicting Assumption A1.

Thus, we we can choose the basis function $\tau_p$ that is nonzero for $t_0$ and hence also nonzero in a neighborhood around $t_0$, due to the smoothness of the basis functions. The same applies to the function $\tau_p^2$, which is likewise contained in the set of basis functions (the basis functions form an algebra that is closed under multiplication). Moreover, due to the fact that the basis functions are orthonormal, it follows that

$$\int_I \tau_p^2 \mathrm{d}t = 1. \tag{4.110}$$

In the same way, the function $\delta\tilde{p}_i(t)\tau_p^2(t)$ is also contained in the set of basis functions, is non-negative for all $t \in I$, and is bounded and integrable for all integers $i > 0$ ($|\tau^p|$ is bounded, see App. F). Thus, it follows that the integral

$$\int_I \tau_p^2 \delta\tilde{p}_i(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t \tag{4.111}$$

exists. By assumption $(x, u, x_T) \in \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$, and therefore

$$0 \leq \int_I \tau_p^2 \delta\tilde{p}_i(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t \tag{4.112}$$

$$= \int_I \tau_p^2 \delta p(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t \tag{4.113}$$

$$+ \int_I \tau_p^2 (\delta\tilde{p}_i - \delta p)(-C_{\mathrm{x}k}x - C_{\mathrm{u}k}u + b_k)\mathrm{d}t, \tag{4.114}$$

where the last term can be bounded by

$$\epsilon C_1 \int_I \tau_p^2 |C_{\mathrm{x}k}x + C_{\mathrm{u}k}u - b_k|\mathrm{d}t. \tag{4.115}$$

The above integral is bounded due to the fact that $\tau_p^2$ is bounded and integrable, $x \in L_n^2$, $u \in L_m^2$, and that $b_k$ is bounded, and therefore (4.114) can be made arbitrarily small by sufficiently increasing $i$. However, this leads to a contradiction, since (4.113) is strictly negative, according to (4.109). This proves that $\lim_{s\to\infty} \mathcal{C}_{\mathrm{U}}^s \supset \lim_{s\to\infty} \mathcal{C}_{\mathrm{L}}^s$. The desired result is then established due to the fact that $\mathcal{C}_{\mathrm{U}}^s \subset \mathcal{C}_{\mathrm{L}}^s$ holds for all integers $s > 0$. □

## D. Proof of Prop. 8

*Proof.* The proof is taken from [33] and included for completeness. The following notation is introduced: The closed-loop state and input trajectories are denoted by $x(t)$ and $u(t)$. The predicted trajectories are referred to as $\tilde{x}(t_\mathrm{p}|t)$, $\tilde{u}(t_\mathrm{p}|t)$, where $t_\mathrm{p} > 0$ denotes the prediction horizon. For $t_\mathrm{p} = 0$, the prediction matches the true trajectory, that is $\tilde{x}(0|t) = x(t)$, $\tilde{u}(0|t) = u(t)$ for all $t \in [0, \infty)$. The predictions $\tilde{x}(t_\mathrm{p}|t)$, $\tilde{u}(t_\mathrm{p}|t)$ are obtained by solving (4.64) subject to the initial condition $x_0 = x(t)$, which yields the parameters $\eta_x$ and $\eta_u$ defining $\tilde{x}(t_\mathrm{p}|t)$ and $\tilde{u}(t_\mathrm{p}|t)$ by

$$\tilde{x}(t_\mathrm{p}|t) = (I_n \otimes \tau(t_\mathrm{p}))^\mathsf{T}\eta_x, \quad \tilde{u}(t_\mathrm{p}|t) = (I_m \otimes \tau(t_\mathrm{p}))^\mathsf{T}\eta_u. \tag{4.116}$$

In order to highlight the dependence on the initial condition, the resulting optimal cost of (4.64) is denoted by $J^\mathrm{MPC}(x(t))$.

By assumption, (4.64) is feasible at time $t = 0$. The resulting trajectories $\tilde{x}(t_\mathrm{p}|0)$, $\tilde{u}(t_\mathrm{p}|0)$ fulfill the equations of motion, the initial condition $\tilde{x}(0|0) = x(0)$, and the constraints and hence, the system evolves according to $x(t) = \tilde{x}(t|0)$, $u(t) = \tilde{u}(t|0)$, $\forall t \in [0, T_\mathrm{d})$. Due to the time-shift property of the basis functions implied by Assumption A2, the feasible candidates

$$\begin{aligned} \tilde{x}(t_\mathrm{p} + T_\mathrm{d}|0) &= (I_n \otimes \tau(t_\mathrm{p})^\mathsf{T})(I_n \otimes \exp(M_s T_\mathrm{d})^\mathsf{T})\eta_x, \\ \tilde{u}(t_\mathrm{p} + T_\mathrm{d}|0) &= (I_m \otimes \tau(t_\mathrm{p})^\mathsf{T})(I_m \otimes \exp(M_s T_\mathrm{d})^\mathsf{T})\eta_u \end{aligned} \tag{4.117}$$

for the optimization at time $T_\mathrm{d}$ can be constructed from the optimizer $\eta_x$ and $\eta_u$ at time 0. As a result, recursive feasibility of (4.64) follows by induction.

We will show that the optimal cost $J^\mathrm{MPC}$ acts as a Lyapunov function. The function $J^\mathrm{MPC}$ is a valid Lyapunov candidate since $J^\mathrm{MPC}(x) > 0$ for all $x \neq 0$ and $J^\mathrm{MPC}(x) = 0$ if and only if $x = 0$. Due to the fact that the shifted trajectories $\tilde{x}(t_\mathrm{p}+T_\mathrm{d}|0)$ and $\tilde{u}(t_\mathrm{p}+T_\mathrm{d}|0)$ (as defined in (4.117)) are feasible for the optimization at time $T_\mathrm{d}$, the following upper bound on $J^\mathrm{MPC}(x(T_\mathrm{d}))$ can be established

$$J^\mathrm{MPC}(x(T_\mathrm{d})) \leq \int_{T_\mathrm{d}}^{\infty} \tilde{x}(t_\mathrm{p}|0)^\mathsf{T} Q \tilde{x}(t_\mathrm{p}|0) + \tilde{u}(t_\mathrm{p}|0)^\mathsf{T} R \tilde{u}(t_\mathrm{p}|0) \mathrm{d}t_\mathrm{p}. \tag{4.118}$$

The right-hand side can be rewritten as

$$J^\mathrm{MPC}(x(0)) - \int_{0}^{T_\mathrm{d}} \tilde{x}(t_\mathrm{p}|0)^\mathsf{T} Q \tilde{x}(t_\mathrm{p}|0) + \tilde{u}(t_\mathrm{p}|0)^\mathsf{T} R \tilde{u}(t_\mathrm{p}|0) \mathrm{d}t_\mathrm{p}, \tag{4.119}$$

resulting in

$$J^{\mathrm{MPC}}(x(T_{\mathrm{d}})) - J^{\mathrm{MPC}}(x(0)) \leq \\ -\int_0^{T_{\mathrm{d}}} \tilde{x}(t_{\mathrm{p}}|0)^{\mathsf{T}} Q \tilde{x}(t_{\mathrm{p}}|0) + \tilde{u}(t_{\mathrm{p}}|0)^{\mathsf{T}} R \tilde{u}(t_{\mathrm{p}}|0) \mathrm{d}t_{\mathrm{p}}. \tag{4.120}$$

The right-hand side of (4.120) is guaranteed to be strictly negative, except for $x(0) = 0$, and thus, by induction, $J^{\mathrm{MPC}}(x(kT_{\mathrm{d}}))$ is strictly decreasing, which concludes the proof. □

## E. Reduction of the semi-infinite constraint

The following section discusses the reduction of the semi-infinite constraint

$$a \leq \tau(t)^{\mathsf{T}} \eta \leq b \tag{4.121}$$

over the interval $t \in [0, \infty)$, with $a, b \in \mathbb{R}$, $a < 0$, $b > 0$ to a compact interval. Thereby we consider the symmetric case, where $|a| = |b|$ first, before discussing the asymmetric case $|a| \neq |b|$. It is shown that the length of this compact time interval depends only on the properties of the basis functions and on the ratio between $|a|$ and $|b|$. Both $a$ and $b$ are assumed to be finite.

### E.1 The symmetric case

**Proposition 10.** *Provided that the basis functions fulfill Assumptions A1 and A2 for all $t \in [0, \infty)$ there exists a positive real number $T_c$ such that*

$$\sup_{t \in [0,\infty)} |\tau(t)^{\mathsf{T}} \eta| = \max_{t \in [0, T_c]} |\tau(t)^{\mathsf{T}} \eta| \tag{4.122}$$

*holds for all parameter vectors $\eta \in \mathbb{R}^s$.*

*Proof.* Without loss of generality we restrict the parameter vectors to have unit magnitude, that is, $|\eta| = 1$.[32]

We prove the claim in 4 steps. We first derive an exponentially decaying upper bound on the Euclidean norm of the basis function vector $\tau$. We then use this bound to argue that the basis functions are linearly independent over the interval $[0, T_{\mathrm{i}}]$ (the scalar $T_{\mathrm{i}}$ will be determined). The third step consists of constructing a lower bound on

$$\max_{t \in [0, T_{\mathrm{i}}]} |\tau(t)^{\mathsf{T}} \eta| \tag{4.123}$$

[32] The claim holds trivially for $\eta = 0$; in case $\eta \neq 0$ we can always normalize $\eta$.

that holds for all parameter vectors $\eta$ with $|\eta| = 1$. Linear independence of the basis functions on $[0, T_\mathrm{i}]$ will be used to argue that this lower bound is strictly positive. In the last step, we show that if $t$ is sufficiently large, $|\tau(t)^\mathsf{T}\eta|$ will be below this lower bound, which concludes the proof.

Step 1): The fact that $M_s$ is asymptotically stable implies that there exists a quadratic Lyapunov function that decays exponentially. This provides a means to establish the following bound

$$|\tau(t)|^2 \leq C_2 e^{-c_2 t}, \quad \forall t \in [0, \infty), \tag{4.124}$$

where $C_2 > 0$, $c_2 > 0$ are constant.

Step 2): We use the Gram matrix to argue that the basis functions are linearly independent over the interval $[0, T_\mathrm{i}]$. According to [45, p. 2, Thm. 3] it holds that the basis functions are linearly independent in the set $[0, T_\mathrm{i}]$ if and only if the matrix

$$\int_0^{T_\mathrm{i}} \tau\tau^\mathsf{T} \mathrm{d}t \tag{4.125}$$

has full rank. This is the case if the bilinear form

$$v^\mathsf{T} \int_0^{T_\mathrm{i}} \tau\tau^\mathsf{T} \mathrm{d}t \; v \tag{4.126}$$

is strictly positive for all $v \in \mathbb{R}^s$ with $|v| = 1$. Combining the fact that the basis functions are orthonormal with the Cauchy-Schwarz inequality, leads to the following lower bound of the above bilinear form,

$$1 - \int_{T_\mathrm{i}}^{\infty} (v^\mathsf{T}\tau)^2 \mathrm{d}t \geq 1 - \int_{T_\mathrm{i}}^{\infty} |\tau|^2 \mathrm{d}t. \tag{4.127}$$

Using the upper bound (4.124) we therefore obtain

$$v^\mathsf{T} \int_0^{T_\mathrm{i}} \tau\tau^\mathsf{T} \mathrm{d}t v \geq 1 - \int_{T_\mathrm{i}}^{\infty} C_2 e^{-c_2 t} \mathrm{d}t = 1 - \frac{C_2}{c_2} e^{-c_2 T_\mathrm{i}}, \tag{4.128}$$

for all $v \in \mathbb{R}^s$ with $|v| = 1$. Thus, we fix $T_\mathrm{i} > 0$, such that

$$1 > \frac{C_2}{c_2} e^{-c_2 T_\mathrm{i}}, \tag{4.129}$$

implying that the matrix (4.125) is positive definite and has therefore full rank. Consequently, the basis functions are guaranteed to be linearly independent on the interval $[0, T_\mathrm{i}]$.

Step 3): We claim that

$$c^* := \inf_{|\eta|=1} \max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}\eta| \tag{4.130}$$

is well-defined and strictly positive. To that extent we first prove that the function $g : \mathbb{R}^s \to [0,\infty)$,

$$g(\eta) := \max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}\eta| \tag{4.131}$$

is continuous (in fact Lipschitz-continuous). Therefore we consider two parameter vectors $\eta_1$ and $\eta_2$ with $g(\eta_1) \leq g(\eta_2)$ (without loss of generality). From the fact that $g(\eta_2)$ can be rewritten as $g(\eta_1 + (\eta_2 - \eta_1))$ and by invoking the triangle inequality it can be inferred that

$$g(\eta_1 + (\eta_2 - \eta_1)) \leq \max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}\eta_1| + |\tau(t)^\mathsf{T}(\eta_2 - \eta_1)|. \tag{4.132}$$

By noting that the maximum of the above sum is smaller than the sum of the summand's maxima it can be concluded that

$$g(\eta_2) \leq g(\eta_1) + \max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}(\eta_2 - \eta_1)|. \tag{4.133}$$

Moreover, by combining the Cauchy-Schwarz inequality and the bound (4.124) we obtain

$$|g(\eta_2) - g(\eta_1)| \leq \sqrt{C_2}|\eta_2 - \eta_1|, \tag{4.134}$$

showing that the function $g$ is indeed (Lipschitz) continuous. As a result, the Bolzano-Weierstrass theorem asserts that the infimum in (4.130), is attained and well-defined. It remains to argue that $c^* > 0$. For the sake of contradiction, we assume $c^* = 0$. This implies the existence of the minimizer $\eta^*$, with $|\eta^*| = 1$, which fulfills

$$\max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}\eta^*| = 0. \tag{4.135}$$

As a result, it follows that $\tau(t)^\mathsf{T}\eta^*$ is zero for all $t \in [0, T_\mathrm{i}]$, which contradicts the fact that the basis functions $\tau$ are linearly independent on $[0, T_\mathrm{i}]$.

Step 4): From the upper bound (4.124) and the Cauchy-Schwarz inequality it follows that

$$|\tau(t)^\mathsf{T}\eta| \leq \sqrt{C_2}e^{-\frac{c_2 t}{2}}, \quad \forall t \in [0,\infty), \tag{4.136}$$

and for all $\eta \in \mathbb{R}^s$ with $|\eta| = 1$. Clearly, $c^* \leq \sqrt{C_2}$ and therefore we can choose the time $T_\mathrm{c}$ such that

$$c^* = \sqrt{C_2}e^{-\frac{c_2 T_\mathrm{c}}{2}}, \tag{4.137}$$

implying

$$\sup_{t\in(T_\mathrm{c},\infty)} |\tau(t)^\mathsf{T}\eta| < c^* \leq \max_{t\in[0,T_\mathrm{i}]} |\tau(t)^\mathsf{T}\eta|, \tag{4.138}$$

for all $\eta \in \mathbb{R}^s$ with $|\eta| = 1$. This proves the claim. □

### E.2 The asymmetric case

**Proposition 11.** *Provided that the basis functions fulfill Assumptions A1 and A2 for all $t \in [0, \infty)$ there exists a positive real number $T_c$ such that*

$$a \leq \tau(t)^\mathsf{T}\eta \leq b, \quad \forall t \in [0, T_c] \tag{4.139}$$

*implies*

$$a \leq \tau(t)^\mathsf{T}\eta \leq b, \quad \forall t \in [0, \infty), \tag{4.140}$$

*for any $\eta \in \mathbb{R}^s$, where $a, b \in \mathbb{R}$, $a < 0, b > 0$.*

*Proof.* We define $\tilde{f} := \tau^\mathsf{T}\eta$ and establish upper and lower bounds on $\tilde{f}$. Without loss of generality we assume $\eta \neq 0$. Combining Assumption A2 with the Caley-Hamilton theorem leads to

$$\begin{aligned} \tilde{f}^{(s)}(t) + a_1\tilde{f}^{(s-1)}(t) + \cdots + a_s\tilde{f}(t) &= \eta^\mathsf{T}\tau^{(s)}(t) + a_1\eta^\mathsf{T}\tau^{(s-1)}(t) + \cdots + a_s\eta^\mathsf{T}\tau(t) \quad (4.141) \\ &= \eta^\mathsf{T}\left(M^s + a_1M^{s-1} + \cdots + a_sM\right)\tau(t) = 0, \quad (4.142) \end{aligned}$$

where $a_1, a_2, \ldots a_s$ are the coefficients of the characteristic polynomial of the matrix $M$. Thus, the trajectory $\tilde{f}$ and its time derivatives fulfill the following set of differential equations

$$\dot{f} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ -a_s & -a_{s-1} & -a_{s-2} & \ldots & -a_1 \end{pmatrix}}_{:=\hat{M}} f, \quad f(0) = \underbrace{\begin{pmatrix} \tau(0)^\mathsf{T} \\ \tau(0)^\mathsf{T}M^\mathsf{T} \\ \vdots \\ \tau(0)^\mathsf{T}(M^{(s-1)})^\mathsf{T} \end{pmatrix}}_{:=H}\eta, \tag{4.143}$$

where $f := (\tilde{f}, \tilde{f}^{(1)}, \ldots, \tilde{f}^{(s-1)})$. The matrix $\hat{M}$ is Hurwitz and therefore, due to the Lyapunov theorem, there exists a symmetric matrix $P > 0$, $P \in \mathbb{R}^{s\times s}$ that satisfies

$$P\hat{M} + \hat{M}^\mathsf{T}P + Q = 0 \tag{4.144}$$

for any symmetric matrix $Q > 0$, $Q \in \mathbb{R}^{s\times s}$. We fix the positive definite matrix $Q$ and consider the quadratic Lyapunov function $V(t) = f(t)^\mathsf{T}Pf(t)$, where $P$ satisfies (4.144).

The time derivative of $V$ can be upper bounded by

$$\dot{V} = -f^\mathsf{T} Q f \leq -\lambda_Q |f|^2 \leq -\frac{\lambda_Q}{\lambda^P} f^\mathsf{T} P f \tag{4.145}$$

$$\leq -\frac{\lambda_Q}{\lambda^P} V, \tag{4.146}$$

where the minimum eigenvalue of $Q$ is denoted by $\lambda_Q$ and the maximum eigenvalue of $P$ is denoted by $\lambda^P$. As a result, this yields the upper bound

$$V(t) \leq V(0) e^{-\frac{\lambda_Q}{\lambda^P} t} \leq \lambda^P \lambda^{H^\mathsf{T} H} |\eta|^2 e^{-\frac{\lambda_Q}{\lambda^P} t}, \tag{4.147}$$

where $\lambda^{H^\mathsf{T} H}$ denotes the maximum eigenvalue of the matrix $H^\mathsf{T} H$. According to the proof of Prop. 10, we may choose $T_\mathrm{i}$ such that

$$1 > \frac{C_2}{c_2} e^{-c_2 T_\mathrm{i}}, \tag{4.148}$$

implying that the basis functions are linearly independent on the interval $[0, T_\mathrm{i}]$ (see (4.129) and Prop. 10 for the definition of the constants $c_2$ and $C_2$). Linear independence can be used to establish the following lower bound, c.f. Prop. 10:

$$\int_0^{T_\mathrm{i}} |\tilde{f}|^2 \mathrm{d}t = \eta^\mathsf{T} \int_0^{T_\mathrm{i}} \tau \tau^\mathsf{T} \mathrm{d}t \; \eta \geq c_3 |\eta|^2, \tag{4.149}$$

where the constant $c_3 > 0$ denotes the minimum eigenvalue of the matrix

$$\int_0^{T_\mathrm{i}} \tau \tau^\mathsf{T} \mathrm{d}t. \tag{4.150}$$

Without loss of generality it is assumed that $|a| \leq |b|$. Choosing the real number $T_\mathrm{c} > T_\mathrm{i}$ implies that the constraint is imposed over the interval $[0, T_\mathrm{i}]$ and therefore

$$\int_0^{T_\mathrm{i}} |\tilde{f}|^2 \mathrm{d}t \leq T_\mathrm{i} |b|^2. \tag{4.151}$$

Combined with the above upper bound on $|\eta|^2$ it follows that

$$|\eta|^2 \leq \frac{T_\mathrm{i} |b|^2}{c_3}. \tag{4.152}$$

This can be used to upper bound the squared magnitude of $\tilde{f}(t)$, that is,

$$|\tilde{f}(t)|^2 \leq \frac{1}{\lambda_P} V(t) \leq \frac{\lambda^P \lambda^{H^\mathsf{T} H} |b|^2 T_\mathrm{i}}{\lambda_P c_3} e^{-\frac{\lambda_Q}{\lambda^P} t}, \tag{4.153}$$

where $\lambda_P$ denotes the minimum eigenvalue of the matrix $P$. As a result, we may choose $T_\mathrm{c}$ such that the above upper bound is below $|a|^2$. This may be achieved by choosing $T_\mathrm{c} > \max\{\hat{t}, T_\mathrm{i}\}$, where

$$\hat{t} := -\frac{\lambda^P}{\lambda_Q} \left( 2 \ln \left( \frac{|b|}{|a|} \right) + \ln \left( \frac{c_3 \lambda_P}{T_\mathrm{i} \lambda^P \lambda^{H^\mathsf{T} H}} \right) \right) \tag{4.154}$$

and ln refers to the natural logarithm. □

## F. Additional properties

In the following we will discuss some of the properties of the basis functions that fulfill Assumptions A1 and A2.

### F.1 Conditions on $M_s$

The fact that the basis functions are orthonormal on the interval $I = (0, T)$ implies

$$\tau(T)\tau(T)^\mathsf{T} - \tau(0)\tau(0)^\mathsf{T} = \int_I \frac{\mathrm{d}}{\mathrm{d}t}(\tau\tau^\mathsf{T})\mathrm{d}t \tag{4.155}$$

$$= M_s + M_s^\mathsf{T}. \tag{4.156}$$

In case the interval $I$ has infinite measure, that is, $I = (0, \infty)$, the above formula reduces naturally to

$$-\tau(0)\tau(0)^\mathsf{T} = M_s + M_s^\mathsf{T}. \tag{4.157}$$

### F.2 Bounds on the Euclidean norm

In case the interval $I$ has infinite measure, that is, $I = (0, \infty)$, it holds that

$$|\tau(t)| \leq |\tau(0)|, \quad \forall t \in [0, \infty). \tag{4.158}$$

This results from

$$\frac{\mathrm{d}}{\mathrm{d}t}(\tau(t)^\mathsf{T}\tau(t)) = \tau(t)^\mathsf{T}(M_s + M_s^\mathsf{T})\tau(t) \tag{4.159}$$

$$= -|\tau(0)^\mathsf{T}\tau(t)|^2 \leq 0, \tag{4.160}$$

for all $t \in [0, \infty)$, which follows from (4.157).

### F.3 Finite number of minima and maxima

In the following we will argue that a function that is not everywhere zero and spanned by basis functions fulfilling Assumptions A1 and A2, has a finite number of minima and maxima. The function will be denoted by $\tilde{f} := \tau^{\mathsf{T}}\eta$, $\eta \in \mathbb{R}^s, \eta \neq 0$, where the function and the basis functions are defined over the interval $t \in I = (0, T)$. If $I$ has infinite measure, it follows from App. E that $\tilde{f}$ takes is maxima and minima within a compact interval, and hence, without loss of generality, we assume in the following that $I$ has finite measure. Moreover, due to the Caley-Hamilton theorem it holds that

$$\tilde{f}^{(s)}(t) + a_1\tilde{f}^{(s-1)}(t) + \cdots + a_s\tilde{f}(t) = 0, \tag{4.161}$$

for all $t \in I$, where the $a_k$, $k = 1, 2, \ldots, s$ are the coefficients of the characteristic polynomial of the matrix $M_s$ (see Prop. 11). According to [46], the time derivative of $\tilde{f}$ has at most $s - 1$ zeros on any subinterval of length $l$, where

$$\sum_{k=1}^{s} \frac{a_k l^k}{k!} < 1. \tag{4.162}$$

This proves readily that $\tilde{f}$ has a finite number of minima and maxima in the interval $I$.

## References


[1] M. Morari and J. H. Lee, "Model predictive control: Past, present and future", *Computers and Chemical Engineering*, vol. 23, no. 4, pp. 667–682, 1999.

[2] S. Pickenhain, "Infinite horizon optimal control problems in the light of convex analysis in Hilbert spaces", *Set-Valued and Variational Analysis*, vol. 23, no. 1, pp. 169–189, 2015.

[3] H. Halkin, "Necessary conditions for optimal control problems with infinite horizons", *Econometrica*, vol. 42, no. 2, pp. 267–272, 1974.

[4] A. Seierstad and K. Sydsæter, *Optimal Control Theory with Economic Applications*. Elsevier, 1987.

[5] D. Bampou and D. Kuhn, "Polynomial approximations for continuous linear programs", *Journal on Optimization*, vol. 22, no. 2, pp. 628–648, 2012.

[6] M. Diehl, H. G. Bock, H. Diedam, and P.-B. Wieber, "Fast direct multiple shooting algorithms for optimal robot control", in *Fast Motions in Biomechanics and Robotics*, M. Diehl and K. Mombaur, Eds., Springer, 2006, ch. 4, pp. 65–93.

[7] A. Barclay, P. E. Gill, and J. B. Rosen, "SQP methods and their application to numerical optimal control", in *Variational Calculus, Optimal Control and Applications*, W. H. Schmidt, K. Heier, L. Bittner, and R. Bulirsch, Eds., Springer, 1996, ch. 3, pp. 207–222.

[8] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality", *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[9] J. A. Primbs and V. Nevistic, "Constrained finite receding horizon linear quadratic control", *Proceedings of the IEEE Conference on Decision and Control*, pp. 3196–3201, 1997.

[10] L. Wang, *Model Predictive Control System Design and Implementation using MATLAB*. Springer, 2009.

[11] L. Wang, "Continuous time model predictive control design using orthogonal functions", *International Journal of Control*, vol. 74, no. 16, pp. 1588–1600, 2001.

[12] A. Domahidi and J. Jerez, *FORCES Professional*, embotech GmbH (`http://embotech.com/FORCES-Pro`), Jul. 2014.

[13] H. J. Ferreau, C. Kirches, A. Potschka, H. G. Bock, and M. Diehl, "qpOASES: A parametric active-set algorithm for quadratic programming", *Mathematical Programming Computation*, vol. 6, no. 4, pp. 327–363, 2014.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[15] J. Nocedal and S. J. Wright, *Numerical Optimization*, second. Springer, 2006.

[16] L. Grüne and J. Pannek, *Nonlinear Model Predictive Control*, second. Springer, 2017.

[17] L. Grüne and A. Rantzer, "On the infinite horizon performance of receding horizon controllers", *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2100–2111, 2008.

[18] J. Pannek, "Receding horizon control", PhD thesis, Universität Bayreuth, 2009.

[19] M. Reble, "Model predictive control for nonlinear continuous-time systems with and without time-delays", PhD thesis, Universität Stuttgart, 2013.

[20] F. Blanchini, S. Miani, and F. A. Pellegrino, "Suboptimal receding horizon control for continuous-time systems", *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 1081–1086, 2003.

[21] G. Stathopoulos, M. Korda, and C. N. Jones, "Solving the infinite-horizon constrained LQR problem using accelerated dual proximal methods", *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1752–1767, 2017.

[22] P. Scokaert and J. Rawlings, "Constrained linear quadratic regulation", *IEEE Transactions on Automatic Control*, vol. 43, no. 8, pp. 1163–1169, 1998.

[23] M. Alamir, *Stabilization of Nonlinear Systems Using Receding-horizon Control Schemes*. Springer, 2006.

[24] M. Alamir, "A framework for real-time implementation of low-dimensional parametrized NMPC", *Automatica*, vol. 48, no. 1, pp. 198–204, 2012.

[25] D. DeHaan and M. Guay, "A real-time framework for model-predictive control of continuous-time nonlinear systems", *IEEE Transactions on Automatic Control*, vol. 52, no. 11, pp. 2047–2057, 2007.

[26] E. T. van Donkelaar, O. H. Bosgra, and P. M. V. den Hof, "Model predictive control with generalized input parametrization", *Proceedings of the European Control Conference*, pp. 1693–1698, 1999.

[27] J. H. Lee, Y. Chikkula, Z. Yu, and J. C. Kantor, "Improving computational efficiency of model predictive control algorithm using wavelet transformation", *International Journal of Control*, vol. 61, no. 4, pp. 859–883, 1995.

[28] S. Summers, C. N. Jones, J. Lygeros, and M. Morari, "A multiresolution approximation method for fast explicit model predictive control", *IEEE Transactions on Automatic Control*, vol. 56, no. 11, pp. 2530–2541, 2011.

[29] G. Calafiore and M. Campi, "Uncertain convex programs: Randomized solutions and confidence levels", *Mathematical Programming*, vol. 102, no. 1, pp. 25–46, 2005.

[30] D. P. de Farias and B. V. Roy, "On constraint sampling in the linear programming approach to approximate dynamic programming", *Mathematics of Operations Research*, vol. 29, no. 3, pp. 462–478, 2004.

[31] P. A. Parilo, "Polynomial optimization, sums of squares, and applications", in *Semidefinite Optimization and Convex Algebraic Geometry*, G. Blekherman, P. A. Parrilo, and R. R. Thomas, Eds., MOS-SIAM, 2013, ch. 3, pp. 47–157.

[32] Y. Nesterov, "Squared functional systems and optimization problems", in *High Performance Optimization*, H. Frenk, K. Roos, T. Terlaky, and S. Zhang, Eds., Springer, 2000, ch. 17, pp. 405–440.

[33] M. Muehlebach and R. D'Andrea, "Parametrized infinite-horizon model predictive control for linear time-invariant systems with input and state constraints", *Proceedings of the American Control Conference*, pp. 2669–2674, 2016.

[34] M. Muehlebach and R. D'Andrea, "Approximation of continuous-time infinite-horizon optimal control problems arising in model predictive control", *Proceedings of the IEEE Conference on Decision and Control*, pp. 1464–1470, 2016.

[35] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

[36] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*. Birkhäuser, 2009.

[37] W. Rudin, *Principles of Mathematical Analysis*, third. McGraw-Hill, 1976.

[38] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Second. AMS Chelsea Publishing, 1980.

[39] A. Göpfert, *Mathematische Optimierung in allgemeinen Vektorräumen*. BSB B. G. Teubner Verlagsgesellschaft, 1973.

[40] R. Fletcher, *Practical Methods of Optimization*, second. John Wiley & Sons, 1987.

[41] J. B. Conway, *A Course in Functional Analysis*, Second. Springer, 1990.

[42] W. Rudin, *Real and Complex Analysis*, third. McGraw-Hill, 1987.

[43] M. Muehlebach, C. Sferrazza, and R. D'Andrea, "Implementation of a parametrized infinite-horizon model predictive control scheme with stability guarantees", *Proceedings of the International Conference on Robotics and Automation*, pp. 2723–2730, 2017.

[44] M. Hofer, M. Muehlebach, and R. D'Andrea, "Application of an approximate model predictive control scheme on an unmanned aerial vehicle", *Proceedings of the International Conference on Robotics and Automation*, pp. 2952–2957, 2016.

[45] G. Sansone, *Orthogonal Functions*. Interscience Publishers, 1959.

[46] S. Yakovenko, "On functions and curves defined by ordinary differential equations", *Fields Institute Communications*, vol. 24, pp. 497–525, 1999.

# Part D

## A state estimation algorithm for distributed networked systems

# Paper P5

# Distributed Event-Based State Estimation for Networked Systems: An LMI-Approach

Michael Muehlebach and Sebastian Trimpe

**Abstract**

In this work, a dynamic system is controlled by multiple sensor-actuator agents, each of them commanding and observing parts of the system's input and output. The different agents sporadically exchange data with each other via a common bus network according to local event-triggering protocols. From these data, each agent estimates the complete dynamic state of the system and uses its estimate for feedback control. We propose a synthesis procedure for designing the agents' state estimators and the event triggering thresholds. The resulting distributed and event-based control system is guaranteed to be stable and to satisfy a predefined estimation performance criterion. The approach is applied to the control of a vehicle platoon, where the method's trade-off between performance and communication, and the scalability in the number of agents is demonstrated.

Accepted for publication in *IEEE Transactions on Automatic Control*, 2017, to appear.

## 1. Introduction

The majority of today's control systems are implemented on digital hardware with a periodic exchange of data between the various system's components, e.g. reading sensor values, providing actuation commands, etc. While periodic information exchange simplifies the analysis of the resulting control systems, it is fundamentally limited: system resources such as computation and communication are used at predetermined time instants irrespective of the current state of the system, or the information content of the data to be passed between the components. This is not the case with event-based strategies, where information is exchanged or processed only when certain events indicate that an update would be favorable, for instance, to improve the control or estimation performance. System resources are therefore only used when necessary. As a consequence, event-based communication for control, estimation, and optimization is an active and growing area of research, see e.g. [1]–[6] and references therein.

In this work, we consider event-based communication for a distributed control system, where multiple sensor and actuator agents observe and control a dynamic system and exchange data via a common bus network, as shown in Fig. 5.1. In previous work [7], [8], an architecture for distributed state estimation with event-based communication between the agents was proposed. Each agent consists of three main components: the controller computes actuation commands based on the information obtained from the state estimator; the event generator (EG) decides whether local measurements are transmitted over the common bus network and shared with all agents; and the state estimator reconstructs the system's state based on the measurements communicated over the bus network. The event generator compares the current measurement to the prediction of the measurement by the state estimator for making effective transmit decisions. The architecture is distributed due to the fact that transmit decisions, state estimates, and control inputs are computed locally. The common bus is a key element of the proposed architecture as it facilitates information sharing between all components. Bus systems as assumed herein are common in industry automation [9], and have recently also been proposed for multi-hop low-power wireless networks [10].

The approach in [7], [8] has been shown to be effective for reducing measurement communication in experiments on the Balancing Cube test bed [11], which has a network architecture as in Fig. 5.1. The method in [7], [8] relies on a distributed and event-based implementation that emulates a given centralized observer and controller design. In [7], closed-loop stability is shown in an ideal scenario with perfect communication (no delay or packet loss) and identically initialized state estimates. To guarantee closed-loop stability also for the case where state estimates may differ (e.g. due to packet losses), additional periodic estimator resets are introduced in [8]. Both approaches require periodic communication of the inputs.

In this work, a modified design is proposed, which further reduces network load by avoiding the communication of the control inputs altogether and under favorable circumstances (to be made precise) also the periodic estimator resets. In contrast to [7], [8],

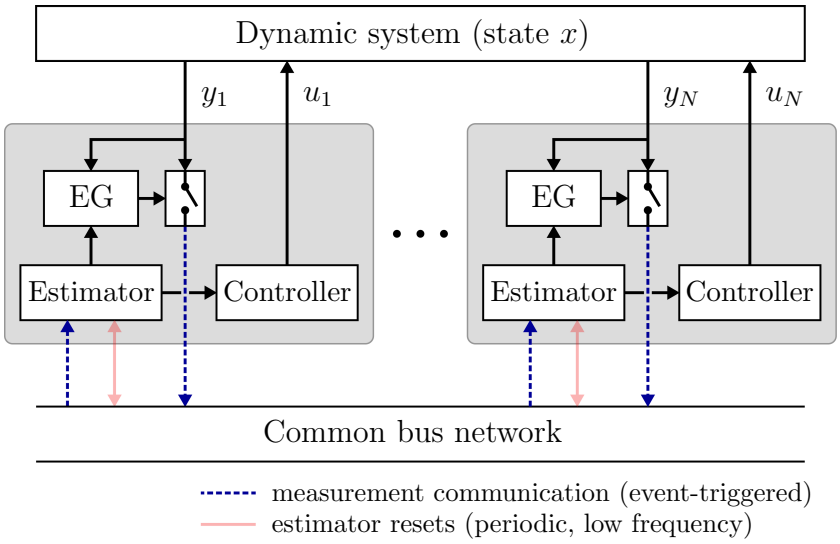


**Figure 5.1.** Networked control system considered in this paper. Each agent observes part of the system state $x$ through local sensors $y_i$ and sends commands $u_i$ to its local actuator. Event-triggered communication is indicated by dashed arrows, while periodic communication is shown by solid ones. The periodic estimator resets can be avoided under certain conditions (to be made precise later). The common-bus architecture is motivated by commonly used field-bus systems [9], such as CAN on the Balancing Cube [11], as well as recent wireless systems [10].

which obtain the estimator gains from a centralized Luenberger observer design and the event triggering thresholds by manual tuning, we synthesize observer gains *and* triggering thresholds specifically for the distributed and event-based estimation problem. A flexible performance objective is derived, such that the state estimator design can be formulated as an optimization problem. The optimization is augmented with linear matrix inequalities (LMIs) imposing closed-loop stability. As a result, both, the state estimator and the event generator, are designed by solving convex optimization problems, [12].

Preliminary results of those herein were presented in the conference papers [13] and [14], which focused on stability and performance, respectively. The main extensions of this article include a less conservative stability condition for the inter-agent error; a relaxation of the LMI-design that scales linearly instead of exponentially in the number of agents; new simulation examples; and the unified presentation of previous stability and performance results.

*Related Work:* Distributed event-based state estimation designs based on LMI formulations are also proposed in [15]–[18], whose relation to this work is discussed next. For a general overview and references on event-based state estimation, the reader is referred to the reviews in [5], [19], [20].

While herein filtering performance is considered in terms of an $\mathcal{H}_2$ index (e.g. like in the steady-state Kalman filter), [15] considers $\mathcal{H}_\infty$ performance and proposes an LMI-based sufficiency condition for filter design. Similarly, [16] proposes a synthesis procedure guaranteeing closed-loop stability and dissipativity for a type of event-based output feedback systems. In [17], the problem of distributed state estimation in a sensor network

described by a directed graph with communication only between neighbors is considered. As in [15] and [16], the transmit decision is based on the difference between the actual measurement and the last measurement, which was transmitted. In contrast, the transmit decision presented herein uses model-based predictions of the output and compares it with the actual measurement, which typically yields more effective triggering decisions (see [21], [22]).

In [18], local observers combining a Luenberger observer and consensus-like correction are proposed. An LMI-based design is used to synthesize the observer gains according to the periodic-update (full communication) scenario, and, only in a second step, the event-based mechanism is introduced. While a similar Luenberger-type observer structure is used herein, the closed-loop stability conditions are not based on the periodic communication scenario, but respect the event-based nature of the control system.

Most of the mentioned references treat the state estimation problem only, while we simultaneously address stability and performance of the state estimation, and stability of the distributed event-based control system that results when local estimates are used for feedback control. The developed results generalize to the pure estimation problem; it suffices to set the state feedback gain $F$ (to be made precise below) to zero.

*Outline:* The distributed event-based estimation and control architecture is presented in Sec. 2, and the problem formulation is made precise in Sec. 3. The closed-loop dynamics are derived in Sec. 4 and are then used to obtain conditions guaranteeing closed-loop stability in Sec. 5. The proposed synthesis procedure is introduced in Sec. 6 and illustrated in simulation examples in Sec. 7. The article concludes with remarks in Sec. 8.

## 2. Architecture

The following section introduces the distributed event-based control system, which is analyzed subsequently. The architecture is similar to [7] and [8].

### 2.1 Networked Control System

The following discrete-time linear system is considered

$$\begin{aligned} x(k) &= Ax(k-1) + Bu(k-1) + v(k-1) \\ y(k) &= Cx(k) + w(k), \end{aligned} \tag{5.1}$$

where $k$ denotes the time index, $x(k) \in \mathbb{R}^n$ the state at time $k$, $u(k) \in \mathbb{R}^{n_u}$ the input at time $k$, and $y(k) \in \mathbb{R}^p$ the output at time $k$. The disturbances $v$ and $w$ are bounded (but not necessarily deterministic), $(A, B)$ is assumed to be stabilizable, and $(A, C)$ is assumed to be detectable.

The inputs and outputs of the system are measured by independent sensor-actuator agents. Therefore the input $u$ and output $y$ is split up according to

$$B\,u(k-1) = \begin{bmatrix} B_1 & B_2 & \dots & B_N \end{bmatrix} \begin{bmatrix} u_1(k-1) \\ \vdots \\ u_N(k-1) \end{bmatrix} \tag{5.2}$$

$$y(k) = \begin{bmatrix} y_1(k) \\ \vdots \\ y_N(k) \end{bmatrix} = \begin{bmatrix} C_1 \\ \vdots \\ C_N \end{bmatrix} x(k) + \begin{bmatrix} w_1(k) \\ \vdots \\ w_N(k) \end{bmatrix}, \tag{5.3}$$

where $u_i(k) \in \mathbb{R}^{q_i}$ is agent $i$'s input and $y_i(k) \in \mathbb{R}^{p_i}$ its measurement. The agents can be heterogeneous, thus the dimensions $q_i$ and $p_i$ may differ, including the cases $q_i = 0$ and $p_i = 0$. It is *not* assumed that the system is detectable or stabilizable by a single agent, i.e. $(A, B_i)$ is not necessarily stabilizable and $(A, C_i)$ not necessarily detectable.

The agents can exchange sensor data $y_i(k)$ with each other over a broadcast network; that is, if one agent communicates, all other agents will receive the data. The communication is assumed to be instantaneous and the agents are synchronized in time. The event-based mechanism determining when sensor data is exchanged will be made precise in the next subsection. It is assumed that the network bandwidth is sufficient to support such communication, and contention among the agents is resolved by low-level protocols. In the Controller Area Network (CAN) on the Balancing Cube [11], for example, contention is resolved through fixed priorities, and the network bandwidth is sufficient to support communication of several agents in one time step. In contrast to [7], [8] the agents do not share input data $u_i(k)$ among each other.

We assume that a static state-feedback controller $u(k) = Fx(k)$ is given, rendering $A + BF$ asymptotically stable (all eigenvalues lie strictly within the unit circle). The existence of such a feedback gain is guaranteed since $(A, B)$ is stabilizable. The controller can be designed using standard methods, see e.g. [23].

### 2.2 Distributed Event-Based State Estimation

Each agent implements an *event generator* that makes the transmit decision for the local measurement, and a *state estimator* that computes a local state estimate.

*1) Event Generator* The event generator triggers the communication of a local measurement $y_i(k)$ of agent $i$ to all other agents. The transmit decision is made according to

$$\text{transmit } y_i(k) \Leftrightarrow |\Delta_i^{-1}\,(y_i(k) - C_i\hat{x}_i(k|k-1))\,| \geq 1, \tag{5.4}$$

where $\Delta_i \in \mathbb{R}^{p_i \times p_i}$ is symmetric and positive definite, $\hat{x}_i(k|k-1)$ is agent $i$'s prediction of the state $x(k)$ based on measurements until time $k-1$ (to be made precise below), $C_i\hat{x}_i(k|k-1)$ is agent $i$'s prediction of its measurement $y_i(k)$, and the Euclidean norm is denoted by $|\cdot|$. The communication thresholds $\Delta_i$ will enter the design process as decision variables.

The underlying idea of the trigger (5.4) is that a communication should happen whenever the predicted output does not match the actual measurement $y_i(k)$. Such triggers

have been considered under the terms *measurement-based trigger*, *innovation-based trigger*, or *predictive sampling* in [21], [22], [24], [25], for example.

To simplify notation, the index set of all agents transmitting their measurements at time $k$ is denoted by

$$I(k) := \left\{ i \in \mathbb{N} \mid 1 \leq i \leq N, |\Delta_i^{-1} \left( y_i(k) - C_i \hat{x}_i(k|k-1) \right)| \geq 1 \right\}, \tag{5.5}$$

where $\mathbb{N}$ denotes the set of natural numbers.

*2) State Estimator* Each agent estimates the full state $x$. Let $\hat{x}_i(k) = \hat{x}_i(k|k)$ denote agent $i$'s estimate of the state at time $k$ given measurement data up to time $k$, which is computed by

$$\hat{x}_i(k|k-1) = A\hat{x}_i(k-1|k-1) + B\hat{u}^i(k-1) \tag{5.6}$$

$$\hat{x}_i(k) = \hat{x}_i(k|k-1) + \sum_{j \in I(k)} L_j \big( y_j(k) - C_j \hat{x}_i(k|k-1) \big) + d_i(k), \tag{5.7}$$

where $\hat{u}^i(k)$ is agent $i$'s belief of the input $u(k)$, $L_j$ are observer gains to be designed, and $d_i$ represents a disturbance, which is assumed to be bounded. The disturbance $d_i$ models[33] mismatches between the estimates of the individual agents, which may stem from unequal initialization, different computation accuracy, or imperfect communication. For example, if the communication from agent $m$ to agent $i$ fails at time $k$, the disturbance $d_i(k)$ takes the value

$$d_i(k) = -L_m(y_m(k) - C_m \hat{x}_i(k|k-1)). \tag{5.8}$$

In Sec. 7, random packet drops are simulated in this way. While $d_i$ cannot be bounded for random drops in general, the simulation results demonstrate that the design is effective also in this case. In App. D, we discuss a packet drop model where the assumption of bounded disturbances is valid provided that packet drops are sufficiently rare.

The disturbance signal $d_i$ in (5.7), which may cause the agents' estimates to differ, plays a crucial role with regards to stability. While closed-loop stability is shown in [7] for $d_i = 0$, it was found in [8] that stability can be lost in case $d_i \neq 0$. To recover stability even in case of nonzero disturbances $d_i$, periodic estimator resets were introduced in [8]. By incorporating the event-based and distributed nature of the control system in the observer design herein, the communication of inputs and (under favorable circumstances) the periodic estimator resets are avoided, while still guaranteeing closed-loop stability for $d_i \neq 0$.

The communication protocol (5.4) implies that a measurement is either transmitted to all agents (and thus included in all state estimates (5.7)), or it is discarded. In App. C, the case where each agent updates its state estimate with its local measurements $y_i$ at

[33] We emphasize that $d_i$ is introduced as a generic disturbance signal for the purpose of stability analysis. When implementing the event-based estimator (5.6), (5.7), $d_i(k)$ is omitted.

every time step is discussed. It is shown that stability is still preserved, while at the same time the estimation performance might be improved.

*3) Distributed Control* Given agent $i$'s state estimate, its local input $u_i$ is obtained by

$$u_i(k) = F_i \hat{x}_i(k), \tag{5.9}$$

where $F^\mathsf{T} = [F_1^\mathsf{T}, F_2^\mathsf{T}, \ldots, F_N^\mathsf{T}]$ is the decomposition of the feedback gain $F$ according to the dimensions of $u_1(k), u_2(k), \ldots, u_N(k)$. Agent $i$'s belief $\hat{u}^i(k)$ of the complete input $u(k)$ is defined as

$$\hat{u}^i(k) := F \hat{x}_i(k), \tag{5.10}$$

and is used in the state estimator update (5.6). This contrasts earlier work, [7], [8], where it was assumed that each agent has access to the true input $u(k)$. Hence, we do not require the communication of the inputs $u_i(k)$ in this work, which reduces the communication load.

## 3. Problem Formulation

The objective of this article is to present a synthesis procedure for both the estimator gains $L_i$ and the communication thresholds $\Delta_i$. The estimators are designed to guarantee i) closed-loop stability (stable dynamics (5.1), (5.6), (5.7), (5.9), and (5.10) for bounded disturbances $v$, $w_i$, and $d_i$), and ii) achieve a predefined $\mathcal{H}_2$ performance incorporating estimation and communication objectives.

## 4. Closed-loop Dynamics

In this section, the closed-loop dynamics are expressed in terms of the system state, local estimation errors, and inter-agent estimation errors, which forms the basis for deriving the stability conditions in Sec. 5. This decomposes the closed-loop dynamics into a series of subsystems connected in feedforward, which facilitates the subsequent analysis. We obtain

$$x(k) = (A+BF)x(k-1) - \sum_{i=1}^{N} B_i F_i e_i(k-1) + v(k-1), \tag{5.11}$$

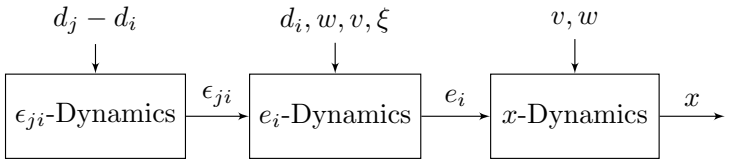


**Figure 5.2.** Simplified block diagram representing the closed-loop system as a feedforward connection of subsystems. The disturbances $d_i$, $w$, $v$, and $\xi$ are bounded (either by assumption or by the event-triggering rule (5.4)).

where $e_i$ is the estimation error of agent $i$ defined by $e_i := x - \hat{x}_i$,

$$\begin{aligned} e_i(k) &= (I - LC)Ae_i(k-1) + (I - LC)v(k-1) \\ &\quad + (I - LC)\sum_{j=1}^{N} B_j F_j \epsilon_{ji}(k-1) + \xi(k) - d_i(k) \\ &\quad + \sum_{j \in I^{\mathsf{c}}(k)} L_j C_j (A + BF)\epsilon_{ji}(k-1) - \sum_{j=1}^{N} L_j w_j(k) \end{aligned} \tag{5.12}$$

with

$$\xi(k) := \sum_{j \in I^{\mathsf{c}}(k)} L_j (y_j(k) - C_j \hat{x}_j(k|k-1)), \tag{5.13}$$

where $I^{\mathsf{c}}(k)$ denotes the complement of $I(k)$ and $\epsilon_{ji} := \hat{x}_j - \hat{x}_i$ refers to the inter-agent error, and

$$\epsilon_{ji}(k) = A_{\mathrm{cl}}(I(k))\epsilon_{ji}(k-1) + d_{ji}(k), \tag{5.14}$$

with $d_{ji}$ defined as $d_{ji} := d_j - d_i$, and

$$A_{\mathrm{cl}}(I(k)) := (I - \sum_{m \in I(k)} L_m C_m)(A + BF). \tag{5.15}$$

## 5. Stability Analysis

Next, conditions on the observer gains $L_i$ are derived to guarantee stability of the closed-loop system. These conditions are expressed as LMIs and can be used for the synthesis of stabilizing observer gains $L_i$ as presented in Sec. 6.

Stability is discussed using the concept of input-to-state stability (ISS) as defined in [26, Def. 3.1]. A feedforward connection of systems is ISS if each system is ISS by itself [27, Cor. 1]. Since this applies to the closed-loop dynamics (see Fig. 5.2), conditions guaranteeing ISS for each subsystem (i.e. the inter-agent dynamics (5.14), the agent error (5.12), and the system state (5.11)) are derived first to subsequently conclude stability for the entire system.

### 5.1 Stability of the Inter-Agent Error

For the subsequent analysis, the inter-agent error (5.14) is regarded as a switched linear system under arbitrary switching. While the event-based design will not typically lead to arbitrary switching, it is difficult to determine all possible communication patterns a-priori, without additional restrictions on the system's structure and the disturbances. However, the consideration of arbitrary switching provides a means to derive general stability conditions that can be expressed as LMIs. The following theorem establishes stability of the inter-agent error dynamics (5.14) by means of a switched quadratic Lyapunov function. This result extends the one in previous work [13], which employed a common Lyapunov function leading to a more conservative condition.

**Theorem 11.** *Let the matrix inequalities*

$$\begin{aligned} A_{\mathrm{cl}}^{\mathsf{T}}(\Pi_i)P_1A_{\mathrm{cl}}(\Pi_i) - P_1 < 0, \quad A_{\mathrm{cl}}^{\mathsf{T}}(\Pi_i)P_1A_{\mathrm{cl}}(\Pi_i) - P_2 < 0, \\ A_{\mathrm{cl}}^{\mathsf{T}}(\emptyset)P_2A_{\mathrm{cl}}(\emptyset) - P_2 < 0, \quad A_{\mathrm{cl}}^{\mathsf{T}}(\emptyset)P_2A_{\mathrm{cl}}(\emptyset) - P_1 < 0, \end{aligned} \tag{5.16}$$

*be fulfilled for symmetric positive definite matrices $P_1, P_2 \in \mathbb{R}^{n\times n}$, and for all $\Pi_i \in \Pi \setminus \emptyset$, where $\emptyset$ denotes the empty set and $\Pi$ the power set of $\{1, 2, \ldots, N\}$. Then the inter-agent error* (5.14) *is ISS.*

*Proof.* Consider a trajectory $\epsilon_{ji}(k)$, $k = 1, 2, \ldots$, subjected to (5.14) and starting at $\epsilon_{ji}(0)$. Let the trajectory $V$ be defined as

$$V(k) = \begin{cases} \epsilon_{ji}^{\mathsf{T}}(k)P_1\epsilon_{ji}(k) & I(k) \neq \emptyset \\ \epsilon_{ji}^{\mathsf{T}}(k)P_2\epsilon_{ji}(k) & I(k) = \emptyset, \end{cases} \tag{5.17}$$

$k = 0, 1, \ldots$. Note that $V(k) \geq 0$ for all $k$, where equality holds only if $\epsilon_{ji}(k)$ vanishes. Moreover, $V$ can be bounded by

$$0 \leq \underline{\sigma}|\epsilon_{ji}(k)|^2 \leq V(k) \leq \bar{\sigma}|\epsilon_{ji}(k)|^2, \tag{5.18}$$

where $\underline{\sigma} := \min\{\sigma_{\min}(P_1), \sigma_{\min}(P_2)\}$ and $\bar{\sigma} := \max\{\sigma_{\max}(P_1), \sigma_{\max}(P_2)\}$, and $\sigma_{\min}(P)$, $\sigma_{\max}(P)$ denote the minimum and maximum singular values of a matrix $P$. The time evolution of $V$ is given by

$$\begin{aligned} V(k) - V(k-1) = {} & 2d_{ji}^{\mathsf{T}}(k)P_mA_{\mathrm{cl}}(I(k))\epsilon_{ji}(k-1) \\ & + \epsilon_{ji}^{\mathsf{T}}(k-1)\left(A_{\mathrm{cl}}^{\mathsf{T}}(I(k))P_mA_{\mathrm{cl}}(I(k)) - P_l\right)\epsilon_{ji}(k-1) \\ & + d_{ji}^{\mathsf{T}}(k)P_md_{ji}(k), \end{aligned}$$

where $m \in \{1, 2\}$, $l \in \{1, 2\}$, depending on $I(k)$ and $I(k-1)$. Denoting the maximum

eigenvalue of $A_{\mathrm{cl}}^{\mathsf{T}}(\Pi_i)P_m A_{\mathrm{cl}}(\Pi_i) - P_l$ over all $\Pi_i \in \Pi$ by $\bar{\lambda}$, yields the bound

$$V(k) - V(k-1) \leq 2|d_{ji}(k)||P_m A_{\mathrm{cl}}(I(k))||\epsilon_{ji}(k-1)| + \bar{\lambda}|\epsilon_{ji}(k-1)|^2 + |P_m||d_{ji}(k)|^2.$$

Completing the squares with an $\alpha > 0$ results in

$$V(k) - V(k-1) \leq (\bar{\lambda} + \alpha)|\epsilon_{ji}(k-1)|^2 - \left(\sqrt{\alpha}|\epsilon_{ji}(k-1)| - \frac{|P_m A_{\mathrm{cl}}(I(k))|}{\sqrt{\alpha}}|d_{ji}(k)|\right)^2 + \left(\frac{|P_m A_{\mathrm{cl}}(I(k))|^2}{\alpha} + |P_m|\right)|d_{ji}(k)|^2.$$

Therefore

$$V(k) - V(k-1) \leq (\bar{\lambda} + \alpha)|\epsilon_{ji}(k-1)|^2 + \left(\frac{|P_m A_{\mathrm{cl}}(I(k))|^2}{\alpha} + |P_m|\right)|d_{ji}(k)|^2$$

and consequently

$$V(k) \leq aV(k-1) + b|d_{ji}(k)|^2, \tag{5.19}$$

where

$$b := \max_{\Pi_i \in \Pi, m \in \{1,2\}} \left(\frac{|P_m A_{\mathrm{cl}}(\Pi_i)|^2}{\alpha} + |P_m|\right), a := \frac{\bar{\lambda} + \alpha}{\underline{\sigma}} + 1.$$

By assumption, c.f. (5.16), $\bar{\lambda}$ is negative and therefore an $\alpha > 0$ can be chosen such that $0 < a < 1$. As a consequence, (5.19) implies that $V(k)$ remains bounded for all $k$. In particular, it follows that

$$V(k) \leq a^k V(0) + b\sum_{l=0}^{k-1} a^l |d_{ji}(k-l)|^2, \tag{5.20}$$

and therefore

$$|\epsilon_{ji}(k)|^2 \leq a^k \frac{\bar{\sigma}}{\underline{\sigma}}|\epsilon_{ji}(0)|^2 + \frac{b}{\underline{\sigma}}\sum_{l=0}^{k-1} a^l |d_{ji}(k-l)|^2. \tag{5.21}$$

The constants $\underline{\sigma}, \bar{\sigma}, a, b$ are all positive, which results in

$$|\epsilon_{ji}(k)| \leq a^{\frac{k}{2}}\sqrt{\frac{\bar{\sigma}}{\underline{\sigma}}}|\epsilon_{ji}(0)| + \sqrt{\frac{b}{\underline{\sigma}}}\sum_{l=0}^{k-1} a^{\frac{l}{2}}|d_{ji}(k-l)|, \tag{5.22}$$

and proves that the inter-agent error is ISS. □

In Thm. 11, the Lyapunov function is switched depending on whether there is communication or not. Using the Schur complement, the conditions (5.16) can be rewritten as

$$\begin{pmatrix} P_k & P_k(I - \sum_{m\in\Pi_i} L_m C_m)(A + BF) \\ * & P_l \end{pmatrix} > 0, \tag{5.23}$$

for all $\Pi_i \in \Pi$ with $k = 1$ if $\emptyset \notin \Pi_i$, $k = 2$ if $\Pi_i = \{\emptyset\}$ and $l = 1, 2$, where the placeholder $*$ is implied by symmetry of the matrix. Thus, using the change of variables $U_m = P_1 L_m$, the previous set of matrix inequalities is linear in $U_m$, $P_1$, and $P_2$ for all $m = 1, 2, \ldots, N$ and can therefore be used as auxiliary condition for the synthesis of the observer gains $L_m$, as done in Sec. 6.

By introducing a Lyapunov function that switches for each communication pattern (i.e. distinct $P_i$'s for each $\Pi_i \in \Pi$), and not only between the case of communication or no communication, the conservativeness of Thm. 11 could be reduced further. However, in that case the resulting stability conditions are not suitable for synthesis, as they are no longer linear in the decision variables. In addition, such an extension would result in a significant increase in the number of LMIs (number of LMIs of the order $2^{2N}$).

*1) Relaxation of the LMI conditions in Thm. 11* In the following, we aim to reduce the number of LMI conditions required to guarantee inter-agent error stability. We first note that the result from [13] follows from Thm. 11 as a corollary,

*Corollary* 12. Let the matrix inequality

$$A_{\mathrm{cl}}^{\mathsf{T}}(\Pi_i) P A_{\mathrm{cl}}(\Pi_i) - P < 0, \tag{5.24}$$

be satisfied for a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$ and for all $\Pi_i \in \Pi$, where $\Pi$ denotes the power set of $\{1, 2, \ldots, N\}$. Then the inter-agent error is ISS.

*Proof.* Set $P_1 = P_2$ in Thm. 11. □

The power set $\Pi$ has cardinality $2^N$, which leads to a rapid growth in the number of LMIs used to ensure inter-agent stability even in Cor. 12. For a large number of agents, the corresponding synthesis problem may become intractable. Therefore the conditions from Cor. 12 are further relaxed, such that the number of LMIs scales linearly with the number of agents. This comes at the price of more conservative conditions.

*Corollary* 13. Let the matrix inequalities

$$H \geq \begin{pmatrix} P & P(A + BF) \\ (A + BF)^{\mathsf{T}} P & P \end{pmatrix} > 0, \tag{5.25}$$

$$\begin{pmatrix} P & P(I - L_m C_m)(A + BF) \\ * & P \end{pmatrix} > \frac{N-1}{N} H \tag{5.26}$$

be satisfied for a symmetric positive definite matrix $H \in \mathbb{R}^{2n\times 2n}$, a symmetric positive definite matrix $P \in \mathbb{R}^{n\times n}$, and for all $m \in \{1, 2, \ldots, N\}$. Then the inter-agent error is ISS.

*Proof.* Applying the Schur complement to (5.24) results in

$$\begin{pmatrix} P & PA_{\mathrm{cl}}(\Pi_i) \\ A_{\mathrm{cl}}(\Pi_i)^{\mathsf{T}}P & P \end{pmatrix} > 0, \tag{5.27}$$

for all $\Pi_i \in \Pi$ and therefore (5.25) implies (5.24) for $\Pi_i = \emptyset$. Note that the sum in the expression of $A_{\mathrm{cl}}(\Pi_i)$ can be rearranged to

$$A_{\mathrm{cl}}(\Pi_i) = -(|\Pi_i|-1)(A+BF) + \sum_{m\in\Pi_i}(I-L_mC_m)(A+BF),$$

such that the LMI (5.27) can be reformulated as

$$\sum_{m\in\Pi_i}\begin{pmatrix} P & P(I-L_mC_m)(A+BF) \\ * & P \end{pmatrix} > (|\Pi_i|-1)\begin{pmatrix} P & P(A+BF) \\ * & P \end{pmatrix}, \tag{5.28}$$

for all $\Pi_i \in \Pi$. In contrast, combining (5.25) and (5.26) leads to

$$\sum_{m\in\Pi_i}\begin{pmatrix} P & P(I-L_mC_m)(A+BF) \\ * & P \end{pmatrix} > \frac{N-1}{N}|\Pi_i|\begin{pmatrix} P & P(A+BF) \\ (A+BF)^{\mathsf{T}}P & P \end{pmatrix}$$

for all $\Pi_i \in \Pi\setminus\emptyset$. It holds that $|\Pi_i|(N-1)/N \geq (|\Pi_i|-1)$, and therefore (5.25) and (5.26) imply (5.28) (and thereby also (5.24)) for all $\Pi_i \in \Pi \setminus \emptyset$, which concludes the proof. □

In Sec. 7, the different stability conditions are compared by means of simulation examples.

*Remark.* In case the open-loop system is unstable, it is essential for guaranteeing inter-agent error stability that each agent reconstructs the input $u$ based on its current state estimate $\hat{x}_i$, as opposed to the case where all agents have access to the true input $u$ (proposed in [7], [8]). This seems counterintuitive, as providing the agents with more information should potentially improve the closed-loop performance. The mechanism leading to a destabilization is further discussed and illustrated on a simple example in App. F.

### 5.2 Stability of the Agent Error

Stability of the agent error (5.12) follows directly from the agent-error dynamics (5.12), the inter-agent error being bounded, and the communication protocol, which bounds the disturbance $\xi$.

*Lemma* 14. Let the inter-agent errors $\epsilon_{ji}$, $j = 1, 2, \ldots, N$ be bounded. Then the agent error $e_i$ is ISS if and only if the eigenvalues of $(I - LC)A$ have magnitude strictly less than one.

*Proof.* See App. A. □

We remark that $(I - LC)A$ corresponds to the error dynamics for the estimator (5.6), (5.7) with full communication; that is, stability of $(I - LC)A$ is a natural requirement for the estimator design. Due to the detectability of $(A, C)$, the existence of such estimator gains $L$ is guaranteed.

### 5.3 Stability of the Closed-loop System

By combining the previous results, conditions for the closed-loop dynamics to be ISS can be established. Provided that the agent error is bounded, it follows from (5.11) that the state $x$ is ISS, since, by assumption, $A + BF$ has all eigenvalues strictly within the unit circle. This leads to the following conclusion:

**Theorem 15.** *Let the eigenvalues of $A + BF$ have magnitude strictly less than one. The closed-loop system is ISS if both, the agent error* (5.12) *and the inter-agent error* (5.14) *are ISS.*

## 6. Performance Analysis and Synthesis

In this section, a general $\mathcal{H}_2$ performance measure is introduced that can capture both estimation performance and communication requirements. LMI-conditions will be established guaranteeing a worst-case performance. Moreover, a unified synthesis procedure for the distributed and event-based estimator (5.6), (5.7) is presented that combines stability requirements (from Sec. 5) and performance criteria.

We will focus on the design of the estimator gains $L_i$ and the communication thresholds $\Delta_i$. However, a similar approach could be used to synthesize the feedback gain $F$ subject to the stability conditions provided by Thm. 11, Cor. 12, or Cor. 13. Likewise, $\mathcal{H}_2$ or $\mathcal{H}_\infty$ performance measures could be included in the design. The resulting synthesis procedures are very similar to the ones presented herein and thus not discussed in detail.

### 6.1 Performance Measure

To simplify the derivation of the performance metric, we assume that the disturbances $d_i$ are absent and that all agents are initialized with the same state estimate. According to (5.14), this implies $\epsilon_{ji}(k) = 0$ for all $k$ (i.e., all agents' estimates are identical), and as a result, we formulate a performance metric based on the estimation error $e_i$ of a single agent. We emphasize that this simplification only serves to obtain a tractable performance criterion; the final synthesis procedure is then augmented with conditions

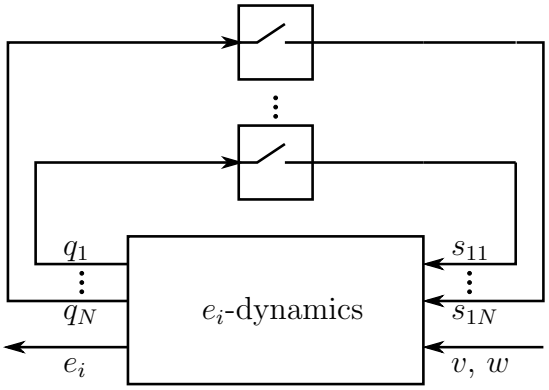


**Figure 5.3.** Block diagram of the simplified agent error dynamics (5.29). The error $e_i$ is driven by the external disturbances $v$ and $w$. The switches and the signals $q_i$ and $s_{1i}$ are used to model the event-based communication. Based on the magnitude of the signal $q_i(k)$ at time instant $k$, the $i$th switch is either closed (no communication in case $|q_i(k)| < 1$) implying $q_i(k) = s_{1i}(k)$, or opened (communication in case $|q_i(k)| \geq 1$) implying $s_{1i}(k) := 0$.

from the previous section ensuring ISS of the closed-loop system and account for the general case of nonzero disturbances $d_i$.

With the above assumptions, the estimation error (5.12) simplifies to

$$e_i(k) = (I - LC)Ae_i(k-1) + (I - LC)v(k-1) - Lw(k) + \xi(k). \tag{5.29}$$

The disturbance $\xi(k)$, as defined in (5.13), can be reformulated as $\xi(k) = L\Delta s_1(k)$, where

$$\Delta := \operatorname{diag}(\Delta_1, \Delta_2, \ldots, \Delta_N) \in \mathbb{R}^{p\times p}, \tag{5.30}$$
$$s_1(k) := (s_{11}^{\mathsf{T}}(k), s_{12}^{\mathsf{T}}(k), \ldots, s_{1N}^{\mathsf{T}}(k))^{\mathsf{T}} \in \mathbb{R}^{p}, \tag{5.31}$$
$$s_{1i}(k) := \chi_{i\in I^{\mathsf{c}}}(k) q_i(k) \in \mathbb{R}^{p_i}, \tag{5.32}$$
$$q_i(k) := \Delta_i^{-1}(y_i(k) - C_i\hat{x}_i(k|k-1)) \in \mathbb{R}^{p_i}, \tag{5.33}$$

and $\chi_{i\in I^{\mathsf{c}}}(k)$ denotes the indicator function, that is, $\chi_{i\in I^{\mathsf{c}}}(k) = 1$ if $i \in I^{\mathsf{c}}(k)$ and 0 otherwise, for $k \in \mathbb{N}$ and $i = 1, 2, \ldots, N$. Note that the signal $q$ is directly related to the communication since a transmission is triggered if $|q_i(k)| > 1$. Furthermore, the communication protocol guarantees that $|s_{1i}(k)|$ is strictly less than one. The agent error dynamics (5.29) can be represented by the block diagram shown in Fig. 5.3.

The communication protocol results in nonlinear feedback terms because of the switching behavior of the event triggers. Therefore, the direct minimization of the performance criterion (to be made precise below) is difficult. Instead, we minimize an upper bound, which is obtained by considering the worst-case performance with respect to all perturbations $s_{1i}(k)$ with Euclidean norm less than one. This leads to a robust control problem and, as a consequence, the resulting synthesis procedure can be formulated as a convex optimization problem.

The power semi-norm [28, p. 816] is used as performance objective:

$$||z||_{\mathcal{P}} := \lim_{K\to\infty} \sqrt{\frac{1}{K}\sum_{k=1}^{K} z^{\mathsf{T}}(k)z(k)}, \tag{5.34}$$

where

$$z(k) := \hat{C}e_i(k-1) + \hat{D}_{21}w(k) + \hat{D}_{22}v(k-1), \tag{5.35}$$

with $\hat{C}$, $\hat{D}_{21}$, $\hat{D}_{22}$ arbitrary matrices of appropriate dimensions. In particular, (5.35) allows for the choices $z(k) = (q_1^{\mathsf{T}}(k), \ldots, q_N^{\mathsf{T}}(k))^{\mathsf{T}}$ and $z(k) = e_i(k)$, which can be used to reduce, respectively, average communication and estimation error, as shall be demonstrated later.

In the following, a synthesis procedure for the observer gains $L_i$ and the communication thresholds $\Delta_i$ is developed, which seeks to minimize $||z||_{\mathcal{P}}$. However, for the reasons stated above, we do not minimize $||z||_{\mathcal{P}}$ directly, but an upper bound, which is formulated in terms of $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms. Expressing the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms using LMIs, see e.g. [29], leads to the following result:

**Theorem 16.** *Let the disturbances $v(k)$ and $w_i(k)$ be bounded, zero mean, independent and identically distributed for all $k$ with covariances $V$ and $W_i$, respectively, $i = 1, 2, \ldots, N$. Define*

$$\hat{A} := (I - LC)A, \qquad \hat{B}_2 := \begin{bmatrix} -LW^{\frac{1}{2}} & (I-LC)V^{\frac{1}{2}} \end{bmatrix},$$
$$\hat{D}_2 := \begin{bmatrix} \hat{D}_{21}W^{\frac{1}{2}} & \hat{D}_{22}V^{\frac{1}{2}} \end{bmatrix}, W := \operatorname{diag}(W_1, W_2, \ldots, W_N),$$

*and let the matrix inequalities*

$$\begin{pmatrix} I & 0 & \hat{C} \\ 0 & P & P\hat{A} \\ \hat{C}^{\mathsf{T}} & \hat{A}^{\mathsf{T}}P & P \end{pmatrix} > 0, \begin{pmatrix} I & 0 & \hat{D}_2 \\ 0 & P & P\hat{B}_2 \\ \hat{D}_2^{\mathsf{T}} & \hat{B}_2^{\mathsf{T}}P & X \end{pmatrix} > 0, \tag{5.36}$$

$$\begin{pmatrix} Q & \hat{A}Q & L\Delta & 0 \\ Q\hat{A}^{\mathsf{T}} & Q & 0 & Q\hat{C}^{\mathrm{T}} \\ \Delta L^{\mathsf{T}} & 0 & I & 0 \\ 0 & \hat{C}Q & 0 & \gamma I \end{pmatrix} > 0, \tag{5.37}$$

*be fulfilled for symmetric matrices $P \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n}$, $X \in \mathbb{R}^{(n+p)\times(n+p)}$, and a scalar $\gamma \in \mathbb{R}$. Then it holds that*

$$||z||_{\mathcal{P}} < \sqrt{N\gamma} + \sqrt{\operatorname{tr}(X)}. \tag{5.38}$$

*Proof.* See App. B. □

The bound (5.38) consists of two terms: The expression $\sqrt{\operatorname{tr}(X)}$ captures the $\mathcal{H}_2$ gain from the disturbances $v$, $w$ to the signal $z$, whereas the expression $\sqrt{N\gamma}$ captures the $\mathcal{H}_\infty$ gain from the signal $s_1$ to the signal $z$, and bounds as such the effect of the nonlinear

feedback due to the event-based communication, see Fig. 5.3. In the full communication scenario it holds that $s_{1i} = 0$, and therefore the agent error reduces to a linear system excited by the disturbances $v$ and $w$, which implies $||z||_{\mathcal{P}} < \sqrt{\text{tr}(X)}$. Hence, the term $\sqrt{\text{tr}(X)}$ corresponds to the performance in the full communication case and represents a lower bound on the achievable performance in the event-based scenario, which is attained for $\Delta_i \to 0$. The term $\sqrt{N\gamma}$ bounds the effect of the disturbance $s_1$ due to the event-based communication.

### 6.2 Synthesis

We first discuss the synthesis of the estimator gains $L_i$ and the thresholds $\Delta_i$ for the relevant special case where the performance measure is the estimation error, which corresponds to the steady-state Kalman filter objective. We then comment on a synthesis procedure for a general performance measure.

*1) Kalman Filter Objective* In case the performance measure is chosen as $z(k) = e_i(k-1)$, that is $\hat{C} = I$, $\hat{D}_2 = 0$, and $\hat{D}_3 = 0$ in (5.35), it follows that (5.36) does not depend on the communication thresholds $\Delta_i$. We therefore propose to design the observer gains $L_i$ in a first step by minimizing $\sqrt{\text{tr}(X)}$ subject to (5.36) and to the conditions ensuring closed-loop stability. For example, if the stability conditions provided by Cor. 12 are used, we synthesize the observer gains according to

$$\begin{aligned} &\inf_{X,P,L} \text{tr}(X) \quad \text{subject to } P = P^{\mathsf{T}} \\ &\begin{pmatrix} I & 0 & I \\ 0 & P & P\hat{A} \\ I & \hat{A}^{\mathsf{T}}P & P \end{pmatrix} > 0, \begin{pmatrix} P & P\hat{B}_2 \\ \hat{B}_2^{\mathsf{T}}P & X \end{pmatrix} > 0, \\ &\begin{pmatrix} P & PA_{\text{cl}}(\Pi_i) \\ A_{\text{cl}}(\Pi_i)^{\mathsf{T}}P & P \end{pmatrix} > 0, \ \forall \Pi_i \in \Pi, \end{aligned} \tag{5.39}$$

where (5.24) has been rewritten using the Schur complement. In the absence of the stability conditions obtained from Cor. 12, this optimization would yield a centralized steady-state Kalman filter. Note that the first condition in (5.36) ensures that $(I - LC)A$ will have all eigenvalues strictly within the unit circle, which implies ISS of the closed-loop system according to Thm. 15.

As a result, the contribution $c^* = \sqrt{\text{tr}(X)}$ to the upper bound given by (5.38) can be calculated, and captures the $\mathcal{H}_2$ gain from the signals $v$ and $w$ to the output $e_i$ in the full communication case.

In a second step, the communication thresholds $\Delta_i$ are synthesized such that an a priori specified worst-case performance $J_{\max}$ is guaranteed (i.e. $||e_i||_{\mathcal{P}} < J_{\max}$). This is

achieved by solving

$$\sup_{Q,\Delta,\gamma} \operatorname{tr}(\Delta) \quad \text{subject to } Q = Q^{\mathsf{T}} \text{ and} \tag{5.40}$$
$$\begin{pmatrix} Q & \hat{A}Q & L\Delta & 0 \\ Q\hat{A}^{\mathsf{T}} & Q & 0 & Q^{\mathrm{T}} \\ \Delta L^{\mathsf{T}} & 0 & I & 0 \\ 0 & Q & 0 & \gamma I \end{pmatrix} > 0, \gamma < \frac{1}{N}(J_{\max} - c^*)^2,$$

while keeping the estimator gains $L_i$ fixed. Therefore this two-step procedure has the following interpretation: In the first step, a lower bound on the achievable cost $||e_i||_{\mathcal{P}}$ is obtained based on the full communication scenario (i.e. $s_1 = 0$), while respecting the stability conditions for the inter-agent error. In the second step, the communication thresholds $\Delta_i$ are designed such that the a priori specified worst-case performance $J_{\max}$ is guaranteed. Hence, the second step can be interpreted as performance versus communication trade-off: increasing $J_{\max}$ will generally downgrade estimation performance by giving the optimization more flexibility to find larger $\Delta_i$, which tends to reduce communication.

In general, feasibility of (5.39) cannot be guaranteed. However, the optimization (5.40) is guaranteed to be feasible provided that $J_{\max} > c^*$. Details regarding feasibility and extensions in case (5.39) is not feasible are discussed in App. E and [13].

*2) General Case* This two-step procedure can also be applied in case of a more general performance objective given by (5.35). The difference is that (5.36) might depend on the communication thresholds $\Delta_i$. As a result, we propose to keep the communication thresholds $\Delta_i$ fixed in the first step, yielding the observer gains $L_i$. In the second step, the observer gains $L_i$ are kept fixed and the communication thresholds are updated by solving an optimization similar to (5.40). The procedure is then repeated until convergence or satisfactory performance.

## 7. Simulation Example

The presented framework for event-based estimation and control is applied in a simulation example that is based on a simplified model for vehicle platooning. Thereby, the communication versus performance trade-off of the proposed approach is discussed, as well as the scalability with respect to a larger number of agents. Additional simulation studies can be found in App. G, [13], and [14].

The problem of vehicle platooning has been studied extensively in the literature, see e.g. [30], [31], and references therein. In [32], it is shown that the linear quadratic regulator problem is ill-posed as the number of vehicles tends to infinity. Moreover, [33] shows that string instability occurs for any local linear feedback law, where the input of the $i$th vehicle depends linearly on the relative distance to its two neighbors. This motivates the

use of a common network, where the different vehicles can exchange information across the platoon.

Similar to [31], we consider a chain of $M$ vehicles (agents), where each vehicle is modeled as a unit point mass. The aim is to control the velocity and the position of each vehicle relative to its neighbors. The following continuous-time model is introduced, c.f. [31],

$$x_i(t) := \begin{pmatrix} p_i(t) \\ r_i(t) - r_{i+1}(t) \end{pmatrix}, \dot{x}_i(t) = \begin{pmatrix} u_i(t) \\ p_i(t) - p_{i+1}(t) \end{pmatrix}, \tag{5.41}$$

$i = 1, 2, \ldots, M-1$, and $x_M(t) := p_M(t)$, $\dot{x}_M(t) = u_M(t)$ with $t \in [0, \infty)$, where $r_i$ and $p_i$ denote the position and velocity of the $i$th vehicle and $u_i$ the normalized force generated by the motor of the $i$th vehicle. The model is discretized with a sampling time of 20 ms leading to the model (5.1).[34]

Each vehicle measures the distance to the previous vehicle, except for the first vehicle, which measures its velocity. The measurements are corrupted by independent, uniformly distributed noise, with $[-0.1\,\mathrm{m}, 0.1\,\mathrm{m}]$ (distance measurements), $[-0.1\,\mathrm{m/s}, 0.1\,\mathrm{m/s}]$ (velocity measurements). Likewise, the inputs $u_i(k)$ are corrupted by independent, uniformly distributed noise $[-0.01\,\mathrm{m/s^2}, 0.01\,\mathrm{m/s^2}]$ . The system is controllable and observable, but neither controllable nor observable for each agent on its own.

A stabilizing feedback controller $F$ is obtained by solving the linear quadratic regulator problem with the identity $I \in \mathbb{R}^{(2M-1)\times(2M-1)}$ and the scaled identity $100\ I \in \mathbb{R}^{M\times M}$ for weighting the state and input costs.

*1) 3 Vehicles* We consider first the case of three vehicles ($M = 3$). As performance objective, the power of the estimation error, $||e_i||_{\mathcal{P}}$, is used, and the observer gains $L_i$ and the communication thresholds $\Delta_i$ are designed according to Sec. 6.2. The optimizations are solved up to a tolerance of $10^{-8}$ using SDPT-3, [34], interfaced through Yalmip, [35]. The different stability conditions, that is, the conditions given by Thm. 11, Cor. 12, and Cor. 13, lead in this case to a very similar design of the observer gains. We will therefore focus on the results obtained by Cor. 12. However, this does not necessarily need to be the case, as shown in App. G. For the synthesis of the communication thresholds, $J_{\max}$ is chosen to be roughly 35 times the power $||e_i||_{\mathcal{P}}$ corresponding to the full communication case, i.e. $J_{\max} = 0.38$, yielding $\Delta_1 = 0.107$, $\Delta_2 = 0.092$, and $\Delta_3 = 0.106$.

The resulting closed-loop system is studied in simulations, where the first car is initialized with a surplus velocity of $5\,\mathrm{m/s}$. The state estimates of the different agents are initialized with zero. In addition, a communication loss rate of 10% is introduced (independent Bernoulli-distributed). The simulation results indicate that the approach is robust also to non-deterministic and potentially unbounded disturbances $d_i$. Fig. 5.4 shows the evolution of the distances between the three vehicles. In steady state, the distance error between the vehicles is kept below $\pm 0.1\,\mathrm{m}$. The communication rates (smoothed with

[34] The same notation is used for continuous and discrete-time signals, e.g. $x(t)$ refers to the continuous-time state trajectory, $x(k)$ to the discrete-time state trajectory.

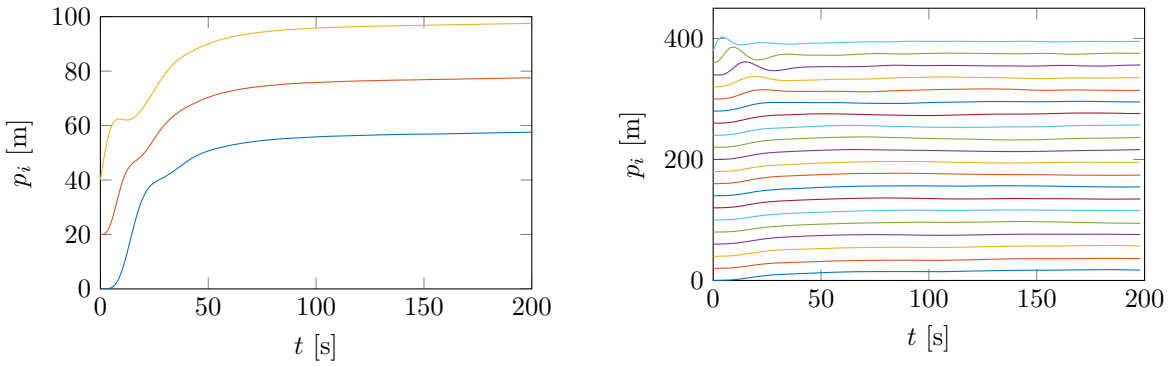


**Figure 5.4.** Left: Platoon with three vehicles, where the evolution of the absolute positions of vehicle 1 (yellow), vehicle 2 (red), and vehicle 3 (blue) is shown. Right: Platoon with 20 vehicles. The vehicles are initialized with an inter-vehicle distance of 20 m.

a moving average filter of length 200) of the different vehicles are depicted in Fig. 5.5. The communication rate is normalized such that a rate of 1.0 corresponds to all agents transmitting their measurements at every time step. In steady state, the second vehicle communicates its measurement in around 8% of the time, whereas the first and last vehicle communicate at a rate below 4%.

The trade-off between estimation performance and communication is obtained by varying $J_{\max}$. The corresponding steady-state performance $||e_i||_{\mathcal{P}}$ and the communication rates of the different designs are evaluated in simulations. Their values were estimated using 20 independent simulations (with different noise realizations) over 1000 s. The variability among the different noise realizations was found to be negligible and a time horizon of 1000 s sufficiently long for transients to be insignificant. The communication versus performance graph, as depicted in Fig. 5.5 is compared to a centralized discrete-time design with reduced sampling rates.[35] This reveals that a better trade-off is achieved by the event-based design as opposed to the centralized design with reduced periodic sampling rates.

*2) 20 Vehicles* The design procedure is repeated for the case $M = 20$, which results in an optimization including 1973 variables. For this example, the inter-agent error stability conditions provided by Cor. 12 would lead to a numerically intractable problem (this would amount to $2^{20}$ LMIs).

The resulting closed-loop performance is evaluated in simulations, where the leading car is initialized with a surplus velocity of 5 m/s, the state estimates of the different vehicles are initialized with zero, and again a packet loss rate of 10% is introduced. The absolute positions of all vehicles are shown in Fig. 5.4. In steady state, the distance error remains below 0.2 m for all 20 vehicles. The communication rates are found to be higher for the

[35] The centralized design is obtained by re-sampling the discrete-time system (5.1) at increasingly lower rates, and then performing a centralized steady-state Kalman filter design based on the performance objective $||e_i||_{\mathcal{P}}$. The fact that the inputs are also communicated is not accounted for in the corresponding communication rates shown in Fig. 5.5.

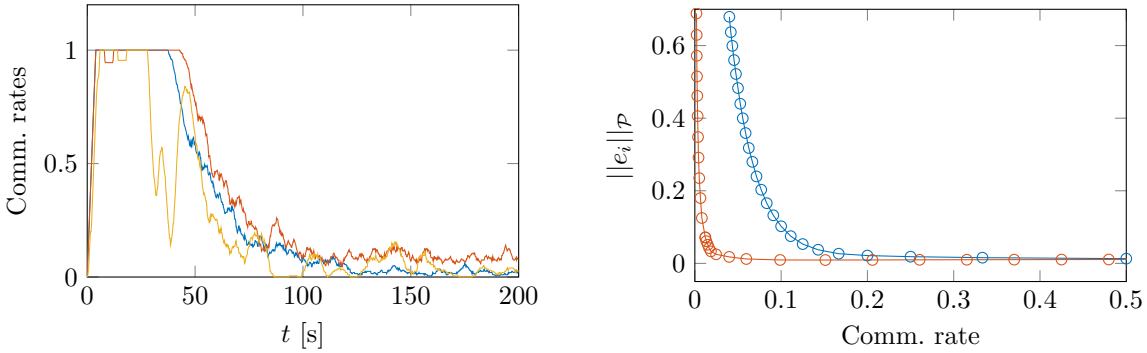


**Figure 5.5.** Left: Communication rates corresponding to the three vehicles in Fig. 5.4. Right: Performance versus communication plot for the event-based design (red) and the centralized design with reduced sampling rates (blue). The graph focuses on communication rates below 0.5, as the achieved performance $||e_i||_{\mathcal{P}}$ changes only insignificantly for rates above 0.5.

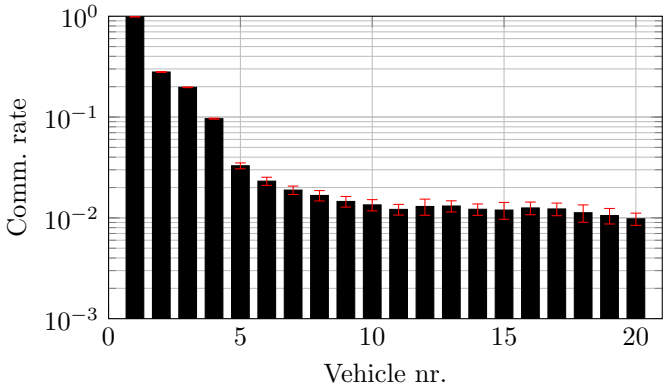


**Figure 5.6.** Communication at steady state for all 20 vehicles. The communication rates are determined by 20 independent simulations (different noise realizations) of the system over a time horizon of 1000 s, which was found to be sufficiently long for transients to die out. The error bars indicate the standard deviation over the different noise realizations.

leading vehicles, see Fig. 5.6, which can be explained by the fact that the actions of the leading vehicles influence all remaining vehicles.

## A. Proof of Lemma 14

*Proof. Sufficiency:* Let the matrix $(I - LC)A$ have eigenvalues with magnitude strictly less than one. According to (5.12) it is enough to show that $\xi$ is bounded, since $\epsilon_{ji}$, $w_i$, $\xi$, and $d_i$ are bounded by assumption. From the triangle inequality, the submultiplicativity of the two-norm, and the communication protocol (5.5), it follows that $|\xi(k)|$ is bounded by

$$\sum_{i \in I^c(k)} |L_i \Delta_i| |\Delta_i^{-1}(y_i(k) - C_i \hat{x}_i(k|k-1))| \leq \sum_{i=1}^{N} |L_i \Delta_i|.$$

*Necessity:* The argument is based on contradiction. Thus we assume the system to be ISS and the matrix $(I - LC)A$ to have at least one eigenvalue of magnitude greater or equal than one. Choosing disturbances $d_i$ parallel to an eigenvector of $(I - LC)A$ with corresponding eigenvalue having magnitude greater or equal than one contradicts the assumption that the agent error is ISS. □

## B. Proof of Thm. 16

*Proof.* The dynamics of the performance objective $z$, as defined in (5.35), can be written as

$$\begin{aligned} e_i(k) &= \hat{A}e_i(k-1) + L\Delta s_1(k) + \hat{B}_2 s_2(k) \\ z(k) &= \hat{C}e_i(k-1) + \hat{D}_2 s_2(k), \end{aligned} \tag{5.42}$$

where

$$s_2(k) := \begin{bmatrix} W^{-\frac{1}{2}} w(k) \\ V^{-\frac{1}{2}} v(k-1) \end{bmatrix}.$$

The communication protocol guarantees that $|s_{1i}(k)|$ is strictly less than one and therefore $|s_1(k)| < \sqrt{N}$.

Let the impulse response from $s_1$ to $z$ be denoted by $g_1$ and the impulse response from $s_2$ to $z$ by $g_2$. Both are well defined, since the matrix $\hat{A}$ has eigenvalues strictly within the unit circle, which is implied by the first matrix inequality in (5.36). Using the fact that $||\cdot||_{\mathcal{P}}$ is a semi-norm yields

$$||z||_{\mathcal{P}} \leq ||g_1 * s_1||_{\mathcal{P}} + ||g_2 * s_2||_{\mathcal{P}}, \tag{5.43}$$

where $*$ denotes the convolution operator. The first term can be upper bounded by, [36, p. 107][36]

$$||g_1 * s_1||_{\mathcal{P}} \leq ||G_1||_\infty ||s_1||_{\mathcal{P}} \leq ||G_1||_\infty \sqrt{N}, \tag{5.44}$$

whereas the second term yields $||g_2 * s_2||_{\mathcal{P}} = ||G_2||_2$, by the statistical properties of $s_2$, [36, p. 108]. Note that $G_1$ and $G_2$ represent the Z-transforms of $g_1$, respectively $g_2$, $||G_1||_\infty$ the $\mathcal{H}_\infty$ norm of $G_1$, and $||G_2||_2$ the $\mathcal{H}_2$ norm of $G_2$, see e.g. [36, pp. 97-100]. Thus, combining (5.43) and (5.44) yields

$$||z||_{\mathcal{P}} \leq ||G_1||_\infty \sqrt{N} + ||G_2||_2. \tag{5.45}$$

According to [29, Lemma 2], it holds that $||G_1||_\infty < \sqrt{\gamma}$, where $\gamma \in \mathbb{R}$ satisfies (5.37),

[36] A continuous-time derivation is presented in [36]. The discrete-time case used herein is analogous.

and according to [14, Theorem A.2 (Appendix)], $||G_2||_2 < \sqrt{\text{tr}(X)}$ holds, where $X = X^\mathsf{T}$ satisfies (5.36). □

## C. Continuous Local Measurement Update

According to (5.4), (5.5), the measurement $y_i(k)$ is used in the estimator update (5.7) only if the condition $|\Delta_i^{-1}(y_i(k) - C_i\hat{x}_i(k|k-1))| \geq 1$ is satisfied. However, each agent could include its local measurements $y_i$ in the update (5.7) continuously (irrespective of the event trigger) without requiring additional communication. The implications of this alternative scheme regarding closed-loop stability are analyzed next.

For each agent $i$, let the indicator function $\chi_{i\in I^\mathsf{c}}(k)$ be defined as $\chi_{i\in I^\mathsf{c}}(k) = 1$ if $i \in I^\mathsf{c}(k)$ and 0 otherwise, for $k \in \mathbb{N}$. In case each agent continuously updates its state estimate with local measurements, the estimation update (5.7) is replaced by

$$\begin{aligned}\hat{x}_i(k) = \hat{x}_i(k|k-1) &+ \sum_{j\in I(k)} L_j(y_j(k) - C_j\hat{x}_i(k|k-1)) \\ &+ \underbrace{\chi_{i\in I^\mathsf{c}}(k)L_i(y_i(k) - C_i\hat{x}_i(k|k-1)) + d_i(k)}_{:=\bar{d}_i(k)},\end{aligned} \tag{5.46}$$

where the additional term can be regarded as a disturbance and forms, together with $d_i(k)$, the disturbance $\bar{d}_i(k)$. In fact, $|\bar{d}_i|$ is bounded by

$$|d_i(k)| + \chi_{i\in I^\mathsf{c}}(k)|L_i||y_i(k) - C_i\hat{x}_i(k|k-1)| < |d_i(k)| + |L_i|\sigma_{\max}(\Delta_i), \tag{5.47}$$

since $\chi_{i\in I^\mathsf{c}}(k) = 1$ implies

$$|y_i(k) - C_i\hat{x}_i(k|k-1)| < \sigma_{\max}(\Delta_i). \tag{5.48}$$

Hence, the conditions ensuring ISS of the closed-loop system established previously remain valid even in case each agent continuously updates his state estimate with local measurements. While causing no additional communication, such a scheme potentially improves the estimation performance since each agent exploits all locally available measurements.

## D. Modeling Packet Drops

If the communication from agent $m$ to agent $i$ fails at time $k$, the disturbance $d_i(k)$ takes the value

$$d_i(k) = -L_m(y_m(k) - C_m\hat{x}_i(k|k-1)). \tag{5.49}$$

As shown below, (5.49) is a function of the agent errors $e_i$, the process noise $v$, and the measurement noise $w_m$. Hence, if $d_i$ is used to model packet drops, it is implicitly dependent on the agent error $e_i$, and boundedness of $d_i$ cannot be guaranteed a priori. However, we will argue that the $d_i$'s are indeed bounded if packet drops are sufficiently rare, and the conditions given by Thm. 11, Cor. 12, or Cor. 13 are fulfilled. We provide a qualitative argument, which can be turned into a quantitative statement about the allowed frequency of packet drops so as to still guarantee boundedness of the disturbances $d_i$. Although these statements tend to be conservative, the simulation examples presented in Sec. 7 indicate that relatively frequent packet drops can be tolerated (e.g. packet loss probability of 10%).

We assume $d_i(1)$ arbitrary and $d_i(k) = 0$ for all agents $i$ and for all $2 \leq k \leq k_0$, where $k_0$ is a positive integer, describing the earliest time instant at which the next packet drop can occur. We therefore model the packet drops as being sufficiently rare, that is, the number of time instants between two consecutive packet drops is greater or equal than $k_0$. We assume further that the conditions of Thm. 11 are fulfilled (the argument is analogous in case the conditions of Cor. 12 or Cor. 13 are satisfied). From (5.22) it follows that the inter-agent error decays exponentially due to the fact that $d_i(k) = 0$ for all $2 \leq k \leq k_0$. The agent-error can be regarded as a linear time-invariant system with system matrix $(I - LC)A$, which is Schur stable. Thus, an exponentially decaying input will lead to an exponentially decaying output. As a consequence, the agent-error $|e_i(k)|$ can be bounded by

$$a_1^k b_1 \sum_{j=1}^{N} |d_j(1)| + b_2, \tag{5.50}$$

where $a_1 < 1$ is the decay rate and $b_2$ is a constant depending on the bounds for $\xi$, $v$, $w$, and $|e_i(0)|$.

Provided that the communication from agent $m$ to agent $i$ fails at time $k$, the measurement equation in (5.1) can be used to rewrite (5.49) as

$$d_i(k) = -L_m C_m (x(k) - \hat{x}_i(k|k-1)) - L_m w_m(k), \tag{5.51}$$

which leads, according to (5.6), (5.10), and (5.11), to

$$d_i(k) = -L_m C_m [(A + BF)e_i(k-1) - \sum_{j=1}^{N} B_j F_j e_j(k-1) + v(k-1)] - L_m w_m(k). \tag{5.52}$$

Given that packet drops happen at times $mk_0 + 1$, $m \in \mathbb{N}$ (or less frequent), we bound $|d_i(mk_0 + 1)|$ for all agents $i$ using a worst case upper bound over all possible communi-

cation failures; that is,

$$|d_i(mk_0+1)| \leq a_1^{k_0} b_3 \sum_{j=1}^{N} |d_j((m-1)k_0+1)| + b_4, \tag{5.53}$$

where $b_3 > 0$ and $b_4 > 0$ are constants. For large enough $k_0$, it follows that $a_1^{k_0} b_3 < 1/N$ and therefore

$$\sum_{i=1}^{N} |d_i(mk_0+1)| < \sum_{i=1}^{N} |d_i((m-1)k_0+1)| + Nb_4, \tag{5.54}$$

for all $m \in \mathbb{N}$. Thus, if packet drops are sufficiently rare, the assumption that the disturbances $d_i$ are bounded is indeed valid.

## E. Feasibility

The stability conditions given by Thm. 11, Cor. 12, and Cor. 13 might be too restrictive, resulting in an infeasible synthesis problem (in Step 1). In this case, the inter-agent error is not guaranteed to be ISS. In [8], a reset strategy was introduced to periodically reset the inter-agent error using additional communication. In the following, an extension to this approach is provided ensuring input-to-state stability of the inter-agent error, even in case the corresponding LMI conditions are infeasible. We will use the conditions in Thm. 11 as starting point. The procedure is analogous if the conditions provided by Cor. 12, and Cor. 13 are used to guarantee inter-agent error stability.

In a first step, the conditions given by (5.16) are relaxed to

$$A_{\mathrm{cl}}^{\mathsf{T}}(\Pi_i) P_k A_{\mathrm{cl}}(\Pi_i) - P_l < \bar{\lambda} I, \tag{5.55}$$

for all $\Pi_i \in \Pi$ with $k = 1$ if $\emptyset \notin \Pi_i$, $k = 2$ if $\Pi_i = \{\emptyset\}$ and $l = 1, 2$. Note that $\bar{\lambda} \geq 0$ is either fixed, or can be included in the optimization problem as decision variable, see [13].

From the proof of Thm. 15, it follows that the function $V$ in (5.17) can be bounded by (c.f. (5.19))

$$V(k) \leq \left( \frac{\bar{\lambda} + \alpha}{\underline{\sigma}} + 1 \right) V(k-1) + \left( \frac{\bar{\gamma}^2}{\alpha} + \bar{\delta} \right) D^2, \tag{5.56}$$

where $D$ is an upper bound to the disturbances $d_{ji}(k)$, i.e. $|d_{ji}(k)| \leq D$ for all $k$, $\bar{\delta} := \max_{m \in \{1,2\}} |P_m|$, and $\bar{\gamma} := \max_{\Pi_i \in \Pi, m \in \{1,2\}} |P_m A_{\mathrm{cl}}(\Pi_i)|$. Therefore, an estimate $\hat{V}(k)$ with $\hat{V}(k) \geq V(k)$ is given by

$$\hat{V}(k) = \left( \frac{\bar{\lambda} + \alpha}{\underline{\sigma}} + 1 \right) \hat{V}(k-1) + \left( \frac{\bar{\gamma}^2}{\alpha} + \bar{\delta} \right) D^2, \tag{5.57}$$

for $k \in \mathbb{N}$, $\hat{V}(0) = 0$ (provided that all agents are initialized with the same state estimate). Note that in order to tighten the bound, the right hand side of (5.56) can be minimized with respect to $\alpha > 0$, as done in [13].

As soon as $\hat{V}$ exceeds the predefined threshold $V_{\max}$, i.e. $\hat{V}(k) \geq V_{\max}$, a communication is triggered and the different agents' state estimates are set to a common value, which resets the inter-agent errors $\epsilon_{ji}(k) = 0$ and implies $\hat{V}(k) = 0$. There are many different reset strategies that can be used, such as a majority vote, the mean, etc. Such resets bound the inter-agent error since $V_{\max} \geq V(k) \geq \underline{\sigma}|\epsilon_{ji}(k)|^2$ for all $k$. By the strict feedforward structure of the closed-loop dynamics, this implies ISS of the state $x$ and the agents' estimation errors $e_i$, $i = 1, 2, \ldots, N$. The time instants $k_{\text{reset}_i}$, where $\hat{V}(k_{\text{reset}_i})$ exceeds $V_{\max}$ for $i = 1, 2, \ldots, N$ can be precalculated, since the evolution of $\hat{V}(k)$ is not explicitly dependent on time. This amounts to periodic resets and extends the procedure presented in [8] by providing a method for choosing the reset period.

The synthesis of the communication thresholds $\Delta_i$ in Step 2 is guaranteed to be feasible. This is because the full communication scenario can be recovered by making the thresholds $\Delta_i$ arbitrarily small; that is, $\gamma \to 0$ in Thm. 16 (we refer to [14] for further details).

## F. Communication of the Inputs

In case of an unstable open-loop system, it is essential for guaranteeing inter-agent stability that each agent reconstructs the input $u$ based on its current state estimate $\hat{x}_i$, as opposed to the case where all agents have access to the true input $u$ (proposed in [7], [8]).

The mechanism leading to a destabilization in case the inputs are communicated can be illustrated by a simple two-agent system having an unstable mode, which is only controllable by agent 1, and only observable by agent 2. Roughly speaking, in case the agents cannot access the true inputs, the inter-agent error tends to decay (e.g. in case there is no communication according to the stable closed-loop dynamics $A+BF$) resulting in communication by agent 2 if the predicted and actual measurements are too far apart, thereby stabilizing the system. In case the agents have access to the true inputs, agent 2 might observe the unstable mode perfectly (but cannot control it), and might thus never share a measurement with agent 1 (who cannot observe the unstable mode at all, but would be able to control it).

Specifically, this mechanism can be illustrated on the system with matrices

$$A = \begin{pmatrix} 0.5 & 0 \\ 0 & 2 \end{pmatrix}, \; B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.58}$$

$$F = \begin{pmatrix} 0 & -2 \\ 0.1 & 0 \end{pmatrix}, \; L = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{5.59}$$

where the first agent measures the first component of $y$ and controls the first component of $u$, and the second agent measures the second component of $y$ and controls the second component of $u$. Clearly, the matrices

$$A + BF = \begin{pmatrix} 0.6 & 0 \\ 0 & 0 \end{pmatrix}, \quad (I - LC)A = 0, \tag{5.60}$$

are stable. The initial condition are chosen as

$$\hat{x}_1(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \hat{x}_2(0) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \quad x(0) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \tag{5.61}$$

and for simplicity, it is assumed that there is neither process noise nor measurement noise, and that the communication thresholds $\Delta_1$ and $\Delta_2$ are set to 1.

In case the input $u$ is communicated, the following sequences of inputs, states, and estimates is obtained

$$\begin{aligned}
\text{Step 1:} \quad & u(0) = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \ x(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \ y(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \\
& \hat{x}_1(1|0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \hat{x}_2(1|0) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \\
& \xrightarrow{\text{no comm.}} \hat{x}_1(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \hat{x}_2(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \\
\text{Step 2:} \quad & u(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ x(2) = \begin{pmatrix} 0 \\ 4 \end{pmatrix}, \ y(2) = \begin{pmatrix} 0 \\ 4 \end{pmatrix}, \\
& \hat{x}_1(2|1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \hat{x}_2(2|1) = \begin{pmatrix} 0 \\ 4 \end{pmatrix} \\
& \xrightarrow{\text{no comm.}} \hat{x}_1(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \hat{x}_2(2) = \begin{pmatrix} 0 \\ 4 \end{pmatrix},
\end{aligned}$$

leading to $u(n) = \hat{x}_1(n) = 0$ and

$$x(n) = y(n) = \hat{x}_1(n) = \hat{x}_2(n) = \begin{pmatrix} 0 \\ 2^n \end{pmatrix}, \tag{5.62}$$

for all $n > 0$. Agent 2, which can observe the unstable mode $x_2$, tracks the state perfectly, and as a result, will never communicate its local measurements $y_2$. In contrast, agent 1, which could control the unstable mode, obtains no information about $x_2$. Thus, in the above example, the state estimate $\hat{x}_1$ will stay at zero for all times, whereas $\hat{x}_2$ tracks $x$ perfectly. Overall an unstable closed-loop system is obtained, unless a periodic estimator

reset (as proposed in [8]) is introduced. Such a reset strategy will periodically set the agents' state estimates to a common average, thereby providing agent 1 with information about $x_2$, resulting in a stabilization of the closed-loop system, as shown in [8].

In case the input is not communicated, the following evolution of the closed-loop system is obtained

$$\begin{aligned}
\text{Step 1:}\quad & u(0) = \begin{pmatrix} -2 \\ 0 \end{pmatrix},\ x(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix},\ y(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \\
& \hat{x}_1(1|0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \hat{x}_2(1|0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
& \xrightarrow{\text{agent 2 comm}} \hat{x}_1(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix},\ \hat{x}_2(1) = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \\
\text{Step 2:}\quad & u(1) = \begin{pmatrix} 0 \\ -4 \end{pmatrix},\ x(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ y(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
& \hat{x}_1(2|1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \hat{x}_2(2|1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\
& \xrightarrow{\text{no comm.}} \hat{x}_1(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \hat{x}_2(2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},
\end{aligned}$$

leading to $u(n) = x(n) = \hat{x}_1(n) = \hat{x}_2(n) = 0$ for all $n > 1$. In that case, both agents track the state perfectly, because agent 2 communicates its measurement $y_2(1)$ and thus shares its information about the unstable mode with agent 1 who is able to drive the system to 0. Thus, by not sharing the inputs, a stable closed-loop system is obtained. The conditions from Cor. V.2 are clearly fulfilled, as the Lyapunov matrix $P$ can, for example, be chosen to be the identity. Thus, according to Thm. V.5 the closed-loop system is guaranteed to be stable.

## G. Inverted Pendulum System

The example is taken from [8], where it was proposed as an abstraction of the Balancing Cube [11], which was the experimental test bed for the distributed and event-based methods in [7] and [24]. The pendulum system is parametrized by the inclination angle $\theta$, the angle $\varphi_1$ of the lower arm (called Agent 1), and the angle $\varphi_2$ of the upper arm (Agent 2), see Fig. 5.7. A state-space model (5.1) is obtained through discretization of the continuous dynamics with a sampling time of 10 ms. The state is given by $x^{\mathrm{T}} = (\theta, \dot{\theta}, \varphi_1, \dot{\varphi}_1, \varphi_2, \dot{\varphi}_2)$, and the inputs are the desired angular rates for the arms, $u = (\dot{\varphi}_{1\mathrm{des}}, \dot{\varphi}_{2\mathrm{des}})$. We refer to [13] for details of the modeling and the numerical values of the state-space matrices.

Agent 1 measures $\varphi_1 + w_{\varphi_1}$, $\dot{\varphi}_1 + w_{\dot{\varphi}_1}$, and $\dot{\theta} + w_{\dot{\theta}}$; and controls $u_1 = \dot{\varphi}_{1\mathrm{des}} + v_{u_1}$. Agent 2 measures $\varphi_2 + n_{\varphi_2}$ and $\dot{\varphi}_2 + v_{\dot{\varphi}_2}$; and controls $u_2 = \dot{\varphi}_{2\mathrm{des}} + v_{u_2}$. The signals $v_{\varphi_1}$, $v_{\varphi_2}$,

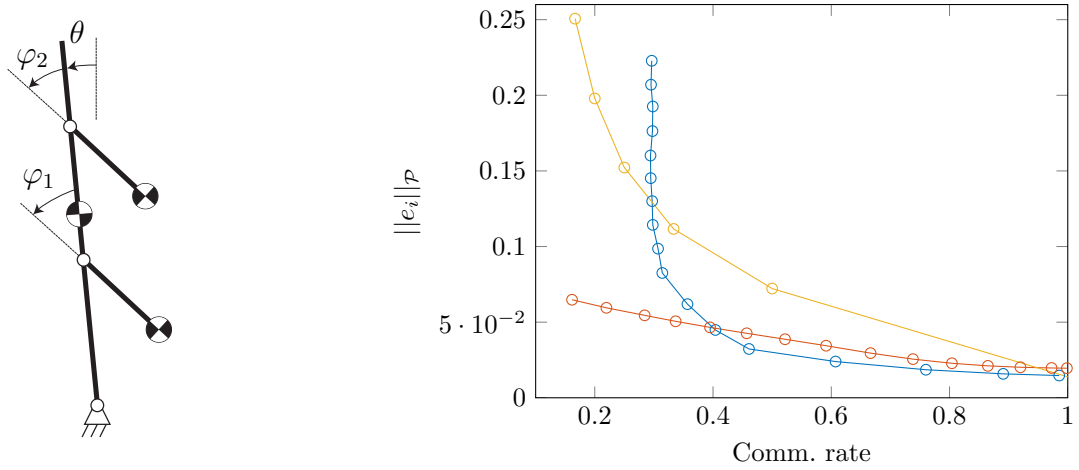


**Figure 5.7.** Inverted pendulum balanced by two independently controlled arms (left), and resulting performance versus communication plots for different event-based estimator designs (right). Blue: event-based design with the stability conditions of Cor. 12; Red: event-based design with the less conservative conditions of Thm. 11; Yellow: centralized design with reduced sampling rates.

$v_{\dot{\varphi}_1}$, $v_{\dot{\varphi}_2}$, $v_{\dot{\theta}}$, $w_{u_1}$, and $w_{u_2}$ are assumed to be independent, uniformly distributed with zero mean and variances $\sigma^2_{\varphi_i} = (0.05\,^\circ)^2$, $\sigma^2_{\dot{\varphi}_i} = (0.1\,^\circ/\mathrm{s})^2$, $\sigma^2_{\dot{\theta}} = (0.24\,^\circ/\mathrm{s})^2$, $\sigma^2_{u_i} = (1.73\,^\circ/\mathrm{s})^2$, $i = 1, 2$. Note that both measurement noise and input noise are introduced. A packet loss probability of 10% is assumed (independent Bernoulli-distributed). The simulation results indicate that the approach is robust also to non-deterministic and potentially unbounded disturbances $d_i$.

The system is controllable and observable, but neither controllable nor observable for each agent on its own. In order to stabilize the upright equilibrium, communication between the agents is indispensable.

A stabilizing state feedback controller $F$ is obtained via a linear quadratic regulator approach, whose values can be found in [13].

As performance measure, the power of the agent-error $e_i$ is used. Observer gains and communication thresholds are synthesized according to Sec. 6.2. The optimizations are solved up to a tolerance of $10^{-8}$ using SDPT-3, [34], interfaced through Yalmip, [35].

For the disturbance rejection properties of an event-based design based on Cor. 12, and a design primarily aimed at reducing communication, we refer to [13], respectively [14]. Herein, we focus on the trade-off between estimation performance and communication, which is obtained by varying $J_{\max}$. The steady-state performance $||e_i||_{\mathcal{P}}$ and the communication rates of the different designs (obtained by successively increasing $J_{\max}$) are evaluated in simulations. Their values were estimated using 20 independent simulations (with different noise realizations) over 150 s. The variability among the different noise realizations was found to be negligible and a time horizon of 150 s sufficiently long for transients to be insignificant. The communication versus performance graphs, resulting from the different designs, i.e. stability conditions according to Thm. 11 and Cor. 12, are depicted in Fig. 5.7 (right), which also includes the graph for a centralized discrete-time

design with reduced sampling rates for comparison. As in Sec. 7, the centralized design is obtained by re-sampling the discrete-time system (5.1) at increasingly lower rates, and then performing a centralized steady-state Kalman filter design based on the performance objective $||e_i||_{\mathcal{P}}$. The fact that the inputs are also communicated is not accounted for in the corresponding communication rates shown in Fig. 5.7. The communication rate is normalized such that a rate of 1.0 corresponds to both agents transmitting their measurements at every time step.

The comparison in Fig. 5.7 reveals that for communication rates above 40% the design based on the stability conditions given by Cor. 12 is superior. In case the communication is further reduced, but kept above 15%, the design based on the stability conditions given by Thm. 11 achieves a lower cost. If $J_{\max}$ is increased further, the communication rate is found to increase again, which is possibly due to nonlinear effects. Compared to the centralized design with reduced periodic sampling rates a better trade-off is achieved by the event-based design.

## References

[1] M. Lemmon, "Event-triggered feedback in control, estimation, and optimization", in *Networked Control Systems*, Springer, 2010, pp. 293–358.

[2] W. Heemels, K. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control", *Proc. of the 51st IEEE Conference on Decision and Control*, pp. 3270–3285, 2012.

[3] L. Grüne, S. Hirche, *et al.*, "Event-based control", in *Control Theory of Digitally Networked Dynamic Systems*, Springer, 2014, pp. 169–261.

[4] C. G. Cassandras, "The event-driven paradigm for control, communication and optimization", *Journal of Control and Decision*, vol. 1, no. 1, pp. 3–17, 2014.

[5] M. Miskowicz (editor), *Event-Based Control and Signal Processing*. CRC Press, 2016.

[6] G. S. Seyboth, D. V. Dimarogonas, and K. H. Johansson, "Event-based broadcasting for mulit-agent average consensus", *Automatica*, vol. 49, no. 1, pp. 245–252, 2013.

[7] S. Trimpe, "Event-based state estimation with switching static-gain observers", *Proc. of the 3rd IFAC Workshop on Distributed Estimation and Control in Networked Systems*, pp. 91–96, 2012.

[8] S. Trimpe, "Stability analysis of distributed event-based state estimation", *Proc. of the 53rd IEEE Conference on Decision and Control*, pp. 2013–2019, 2014.

[9] J.-P. Thomesse, "Fieldbus technology in industrial automation", *Proceedings of the IEEE*, vol. 93, no. 6, pp. 1073–1101, 2005.

[10] F. Ferrari, M. Zimmerling, L. Mottola, and L. Thiele, "Low-power wireless bus", *Proc. of the 10th ACM Conference on Embedded Network Sensor Systems*, pp. 1–14, 2012.

[11] S. Trimpe and R. D'Andrea, "The balancing cube: A dynamic sculpture as test bed for distributed estimation and control", *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 48–75, 2012.

[12] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. SIAM, 1994.

[13] M. Muehlebach and S. Trimpe, "LMI-based synthesis for distributed event-based state estimation", *Proc. of the American Control Conference*, pp. 4060–4067, 2015.

[14] M. Muehlebach and S. Trimpe, "Guaranteed $\mathcal{H}_2$ performance in distributed event-based state estimation", *Proc. of the International Conference on Event-based Control, Communication, and Signal Processing*, 2015.

[15] X.-M. Zhang and Q.-L. Han, "Event-based $H_\infty$ filtering for sampled-data systems", *Automatica*, vol. 51, pp. 55–69, 2015.

[16] X.-M. Zhang and Q.-L. Han, "A decentralized event-triggered dissipative control scheme for systems with multiple sensors to sample the system outputs", *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 2745–2757, 2016.

[17] L. Yan, X. Zhang, Z. Zhang, and Y. Yang, "Distributed state estimation in sensor networks with event-triggered communication", *Nonlinear Dynamics*, vol. 76, pp. 169–181, 2014.

[18] C. Fischione, D. Dimarogonas, F. Rubio, K. Johansson, P. Millan, and U. Tiberi, "Distributed event-based observers for LTI networked systems", *Proc. of the Portuguese Conference on Automatic Control*, 2012.

[19] D. Shi, L. Shi, and T. Chen, *Event-Based State Estimation*. Springer, 2016.

[20] S. Trimpe, "Event-based state estimation: An emulation-based approach", *IET Control Theory & Applications*, vol. 11, no. 11, pp. 1684–1693, 2017.

[21] S. Trimpe and M. C. Campi, "On the choice of the event trigger in event-based estimation", *Proc. of the International Conference on Event-based Control, Communication, and Signal Processing*, 2015.

[22] J. Sijs, L. Kester, and B. Noack, "A study on event triggering criteria for estimation", *Proc. of the 17th International Conference on Information Fusion*, 2014.

[23] K. Åström and B. Wittenmark, *Computer-controlled systems: theory and design*. Prentice Hall, 1997.

[24] S. Trimpe and R. D'Andrea, "An experimental demonstration of a distributed and event-based state estimation algorithm", *Proc. of the 18th IFAC World Congress*, pp. 8811–8818, 2011.

[25] J. Wu, Q.-S. Jia, K. Johansson, and L. Shi, "Event-based sensor data scheduling: Trade-off between communication rate and estimation quality", *IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 1041–1046, 2013.

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

[27] D. Nešić and A. R. Teel, "Changing supply functions in input to state stable systems: The discrete-time case", *IEEE Transactions on Automatic Control*, vol. 46, no. 6, pp. 960–962, 2001.

[28] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Prentice Hall, 1999.

[29] M. C. De Oliveira, J. C. Geromel, and J. Bernussou, "Extended $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norm characterizations and controller parametrizations for discrete-time systems", *International Journal of Control*, vol. 75, no. 9, pp. 666–679, 2002.

[30] A. Alam, J. Mårtensson, and K. H. Johansson, "Experimental evaluation of decentralized cooperative cruise control for heavy-duty vehicle platooning", *Control Engineering Practice*, vol. 38, pp. 11–25, 2015.

[31] W. Levine and M. Athans, "On the optimal error regulation of a string of moving vehicles", *IEEE Transactions on Automatic Control*, vol. 11, no. 3, pp. 355–361, 1966.

[32] M. R. Jovanovic and B. Bamieh, "On the ill-posedness of certain vehicular platoon control problems", *IEEE Transactions on Automatic Control*, vol. 50, no. 9, pp. 1307–1321, 2005.

[33] P. Seiler, A. Pant, and K. Hedrick, "Disturbance propagation in vehicle strings", *IEEE Transactions on Automatic Control*, vol. 49, no. 10, pp. 1835–1842, 2004.

[34] K. C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3 – a MATLAB software package for semidefinite programming, version 1.3", *Optimization Methods and Software*, vol. 11, pp. 545–581, 1999.

[35] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB", *Proc. of the IEEE International Symposium on Computer Aided Control Systems Design*, pp. 284–289, 2004.

[36] K. Zhou, J. Doyle, K. Glover, *et al.*, *Robust and optimal control*. Prentice Hall New Jersey, 1996.