Numerical approximation of the data-rate limit for state estimation under communication constraints

Sigurdur Hafstein* and Christoph Kawan†

Abstract

In networked control, a fundamental problem is to determine the smallest capacity of a communication channel between a dynamical system and a controller above which a prescribed control objective can be achieved. Often, a preliminary task of the controller, before selecting the control input, is to estimate the state with a sufficient accuracy. For time-invariant systems, it has been shown that the smallest channel capacity C_0 above which the state can be estimated with an arbitrarily small error, depending on the precise formulation of the estimation objective, is given by the topological entropy or a quantity named restoration entropy, respectively. In this paper, we propose an algorithm that computes rigorous upper bounds of C_0, based on previous analytical estimates.

Keywords: State estimation; communication constraints; nonlinear systems; topological entropy; restoration entropy; Lyapunov-type functions; numerical computation

AMS Classification: 37B40, 93C10, 93C41

1 Introduction

Networked control systems are spatially distributed systems, in which the communication between sensors, controllers and actuators is accomplished through a shared digital communication network. Examples can be found, for instance, in vehicle tracking, underwater communications for remotely controlled surveillance and rescue submarines, remote surgery, space exploration and aircraft design. Another large field of applications can be found in modern industrial systems, where industrial production is combined with information and communication technology ('Industry 4.0'). A fundamental problem in networked control is to determine the minimal requirements on the communication network for a specified control objective to be achieved.

In the simplest model case, a sensor measures the states of a dynamical system at discrete sampling times and transmits the encoded state measurements through a finite-capacity channel to a controller at a remote location. In this framework, various works characterize or estimate the smallest channel capacity above which a given control objective (usually, stabilization of some sort) can be achieved. An even more fundamental problem is to determine the smallest capacity above which the controller is able to compute an estimate of the state with a given precision. This problem has been studied under various assumptions on the system and the channel. Notably, Savkin [20] characterized the critical capacity by a quantity, which turns out to be infinite if the system is genuinely affected by noise, and otherwise reduces to the topological entropy of the system. A more recent contribution is Matveev and Pogromsky [16], which discusses three estimation objectives of increasing strength and provides constructive methods to obtain upper and lower bounds for the associated critical channel capacities. The contribution of the paper at

*Faculty of Physical Sciences, University of Iceland, Dunhagi 5, IS-107 Reykjavik, Iceland; e-mail: shafstein@hi.is

†Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany; e-mail: christoph.kawan@uni-passau.de

hand consists in a numerical scheme to compute the upper bounds proposed in [15, 16]. Further studies about state estimation under communication constraints include [11, 13].

The systems studied in [16] are of the form $x_{t+1} = \phi(x_t)$ with a C^1-map $\phi : \mathbb{R}^n \to \mathbb{R}^n$. The aim is to generate an accurate estimate $\hat{x}_t$ of the state x_t at a remote location for initial states x_0 confined to a compact set $K \subset \mathbb{R}^n$. The only way to transmit information to the estimator is via a noiseless discrete channel. At each time instant t, a coder encodes x_t by a symbol e_t from a finite coding alphabet $\mathcal{M}$. This process can be described by maps $\mathcal{C}_t$ so that

$$e_t = \mathcal{C}_t(x_0, x_1, \ldots, x_t; \hat{x}_0, \delta), \quad \mathcal{C}_t : (\mathbb{R}^n)^{t+1} \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathcal{M},$$

where $\hat{x}_0$ is an initial estimate satisfying $\|x_0 - \hat{x}_0\| \leq \delta$ for some $\delta > 0$, depending on the aspired exactness the estimate. The estimation process similarly can be described by maps $\mathcal{E}_t$ so that

$$\hat{x}_t = \mathcal{E}_t(e_0, e_1, \ldots, e_t; \hat{x}_0, \delta), \quad \mathcal{E}_t : \mathcal{M}^{t+1} \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n.$$

To allow a certain flexibility in the transmission of information, the number of bits that can be transmitted in any time interval of length r is not fixed, but confined between two numbers $b_-(r) \leq b_+(r)$, satisfying

$$C := \lim_{r\to\infty} \frac{b_-(r)}{r} = \lim_{r\to\infty} \frac{b_+(r)}{r},$$

where C by definition is the capacity of the channel. A desirable objective is to obtain an estimate of the form $\|x_t - \hat{x}_t\| \leq \varepsilon$ for all $t \geq 0$, whenever $x_0, \hat{x}_0 \in K$ and $\|x_0 - \hat{x}_0\| \leq \delta$, where $\delta = \delta(\varepsilon)$. Writing C_0 for the smallest capacity C above which this can be achieved for any $\varepsilon > 0$, it was shown in [16] that $C_0 \geq h_{\mathrm{top}}(\phi; K)$ and $C_0 = h_{\mathrm{top}}(\phi; K)$ if K is forward-invariant, where $h_{\mathrm{top}}(\phi; K)$ is the topological entropy of ϕ on K.

One problem with the estimation objective addressed above is that the gap between the initial error δ and the final error ε may be very large. Another problem is that a coding and estimation policy based on topological entropy is likely to suffer from a severe non-robustness, since topological entropy is highly discontinuous with respect to the dynamical system under consideration. To avoid a drastic degradation of accuracy and at the same time obtain a coding and estimation scheme that is more robust with respect to perturbations, one may require instead that $\|x_t - \hat{x}_t\| \leq G\delta$ for all $t \geq 0$ with a constant $G > 0$. The smallest channel capacity C_0 above which this objective can be achieved can be described in terms of a different entropy notion, introduced in [17] under the name *restoration entropy*. A closed-form expression for restoration entropy can be formulated in terms of the singular values of the linearized system. This expression, which at the same time is an upper bound on $h_{\mathrm{top}}(\phi; K)$, has been derived earlier by the authors of [16] in their papers [15, 19], both for discrete- and continuous-time systems and also for time-varying systems.

In this paper, we consider a flow $(\phi_t)_{t\in\mathbb{R}}$ generated by an ODE $\dot{x} = f(x)$ with a sufficiently smooth vector field f on $\mathbb{R}^n$. Our analysis focuses on the dynamics of (ϕ_t) on a compact forward-invariant set K. Essentially following an approach used before for the computation of Lyapunov functions [1, 6, 7, 14], we numerically compute a piecewise affine Riemannian metric on the simplices of a triangulation of K, which is then used to produce an upper estimate on C_0 in terms of the eigenvalues of the symmetrized derivative of the vector field f, computed with respect to that metric. Our algorithm works in two steps. The first one produces a piecewise affine function P on the given triangulation with values in the positive definite $(n \times n)$ symmetric matrices, designed in such a way to minimize the maximum of the largest generalized eigenvalue. The second step produces a Lyapunov-like function, which is used to scale the metric P in order to make the largest generalized eigenvalue even smaller. It needs to be mentioned that the original estimate in [16, 19] involves not the largest generalized eigenvalue only, but the sum of the k largest generalized eigenvalues, where $1 \leq k \leq n$ is chosen to maximize this sum. Hence, we can only expect good results in low dimensions, where typically $k = 1$.

We apply our algorithm to the well-known Lorenz system with standard parameters on a region containing the attractor. Using a simplified algorithm which works with a constant P, we are already able to improve former entropy estimates obtained in [19] by analytical methods.

The paper is organized as follows. In Section 2, we recall the main result of [16, 19], yielding upper estimates on the topological entropy and the critical channel capacity. Section 3 explains the relevance of our algorithm for the problem of state estimation under communications constraints. A detailed description of the algorithm is presented in Section 4. In Section 5, the example of the Lorenz system is discussed. Finally, Section 6 contains some concluding remarks.

2 Preliminaries

Notation. We denote by $\mathbb{Z}$ the set of integers and write $\mathbb{Z}_+ = \{n \in \mathbb{Z} : n \geq 0\}$. We write $\mathcal{S}_n$ for the space of $(n \times n)$ real symmetric matrices and $\mathcal{S}_n^+ \subset \mathcal{S}_n$ for the space of all positive definite elements of $\mathcal{S}_n$. If $\phi(t,x)$ denotes the (local) flow of an ODE $\dot{x} = f(x)$ in $\mathbb{R}^n$ and $v : \mathbb{R}^n \to E$ is a C^1-function into a Euclidean space E, we write $\dot{v}(x)$ for the *orbital derivative* of v at x, i.e.

$$\dot{v}(x) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} v(\phi(t,x)) \in L(\mathbb{R}, E) \cong E.$$

By I we denote the $(n \times n)$ identity matrix for any $n \in \mathbb{N}$. By writing $A \succeq B$ for $A, B \in \mathcal{S}_n$, we mean that $A - B$ is positive semi-definite. Furthermore, we write $B_\varepsilon(x) = \{y \in \mathbb{R}^n : \|x - y\| < \varepsilon\}$. Finally, we use the notation $i = 1 : n$ as a short-cut for $i \in \{1, \ldots, n\}$.

Upper bounds for topological entropy. In the following, we recall the main result of [19], providing upper bounds on topological entropy and critical channel capacity. In [19], the result is formulated for nonautonomous ODEs. However, we only use the following autonomous version.

Consider an ODE of the form

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n \tag{1}$$

with a C^1-vector field $f : \mathbb{R}^n \to \mathbb{R}^n$. Since we only consider solutions that evolve within a compact set, we may assume that all solutions are defined on the whole time domain. We write $\phi(t, x_0)$ for the unique solution satisfying the initial condition $x(0) = x_0$. For a fixed $t \in \mathbb{R}$, we also write $\phi_t : \mathbb{R}^n \to \mathbb{R}^n$ for the diffeomorphism $x \mapsto \phi(t,x)$. We further assume the existence of a compact forward-invariant set $K \subset \mathbb{R}^n$, i.e. $\phi_t(K) \subset K$ for all $t \geq 0$.

The topological entropy of ϕ on K, denoted by $h_{\mathrm{top}}(\phi; K)$, can be defined as follows. For $\tau, \varepsilon > 0$, a subset $E \subset \mathbb{R}^n$ (τ,ε)-spans K if for every $x \in K$ there is $y \in E$ with

$$\max_{0 \leq t \leq \tau} \|\phi(t,x) - \phi(t,y)\| \leq \varepsilon.$$

Writing $r(\tau, \varepsilon, \phi, K)$ for the minimal cardinality of any (τ,ε)-spanning set for K,

$$h_{\mathrm{top}}(\phi; K) := \lim_{\varepsilon \downarrow 0} \limsup_{\tau \to \infty} \frac{1}{\tau} \log_2 r(\tau, \varepsilon, \phi, K).$$

2.1 Theorem: *Let $P : K \to \mathcal{S}_n^+$ and $v_d : K \to \mathbb{R}$, $1 \leq d \leq n$, be C^1-functions and let $\lambda_1(x) \geq \ldots \geq \lambda_n(x)$ denote the solutions of the algebraic equation*

$$\det\Big[\mathrm{D}f(x)^\top P(x) + P(x)\mathrm{D}f(x) + \dot{P}(x) - \lambda P(x)\Big] = 0. \tag{2}$$

Let $\Lambda_d \geq 0$, $1 \leq d \leq n$, be constants so that

$$\sum_{i=1}^{d} \lambda_i(x) + \dot{v}_d(x) \leq \Lambda_d \quad \textit{for all } x \in K. \tag{3}$$

Then for $\Lambda := \max_{1 \leq d \leq n} \Lambda_d$, the topological entropy of ϕ on K satisfies

$$h_{\mathrm{top}}(\phi; K) \leq \frac{\Lambda}{2 \ln 2}.$$

Some remarks about the formulation of the theorem are in order.

2.2 Remark: The functions P and v_d in [19] depend on three variables, i.e. $P = P(t,s,x_0)$ and $v_d = v_d(t,s,x_0)$, where $t \geq s$ are time variables. Such functions can be obtained from the above formulation by putting

$$\tilde{P}(t,s,x_0) := P(\phi(t-s,x_0)), \quad \tilde{v}_d(t,s,x_0) := v_d(\phi(t-s,x_0)),$$

and it is easy to verify that the so-defined functions satisfy the requirements of [19, Thm. 3.2].

2.3 Remark: The function P can be interpreted as a Riemannian metric on K, defined by

$$\langle v,w\rangle_x := \langle P(x)v,w\rangle \quad \text{for all } x \in K.$$

Indeed, let $X(\cdot)$ denote the solution to the following initial value problem corresponding to the variational equation of (1):

$$\dot{Y}(t) = \mathrm{D}f(\phi_t(x_0))Y(t), \quad Y(0) = I.$$

Let $\alpha_1(t) \geq \ldots \geq \alpha_n(t)$ denote the singular values of $X(t)$ w.r.t. the metric $\langle\cdot,\cdot\rangle_{(\cdot)}$, i.e. the eigenvalues of the self-adjoint operator $\sqrt{X(t)^*X(t)}$, where $X(t)^*$ is defined by $\langle X(t)v,w\rangle_{\phi_t(x_0)} \equiv \langle v, X(t)^*w\rangle_{x_0}$. Then, according to [19, Prop. 8.6],

$$\alpha_1(t)\alpha_2(t)\cdots\alpha_d(t) \leq \exp\left(\frac{1}{2}\int_0^t [\lambda_1(\phi(s,x_0)) + \cdots + \lambda_d(\phi(s,x_0))]\mathrm{d}s\right), \quad 1 \leq d \leq n. \tag{4}$$

We expect that the number d, where the maximum $\max_{1\leq d\leq n}\Lambda_d$ is attained, is more or less fixed for a given system under any reasonable choice of the functions $v_1,\ldots,v_n$ (it is something like the number of positive Lyapunov exponents). If this is the case, we can also incorporate the function v_d into the metric by putting

$$\langle v,w\rangle_x := \langle \mathrm{e}^{v_d(x)/d}P(x)v,w\rangle \quad \text{for all } x \in K.$$

Then (2) is equivalent to

$$\det\left[\mathrm{D}f(x)^\top P(x) + P(x)\mathrm{D}f(x) + \dot{P}(x) - \left(\lambda - \frac{1}{d}\dot{v}_d(x)\right)P(x)\right] = 0,$$

and thus, the sum $\sum_{i=1}^d \lambda_i(x)$ with the solutions of (2) becomes $\sum_{i=1}^d \lambda_i(x) + \dot{v}_d(x)$.

The functions v_i in Theorem 2.1 have some similarity with Lyapunov functions. Instead of $\dot{v}_i < 0$ we have the inequalities (3). In the rest of the paper, we call such functions *Lyapunov-type functions.*

3 The state estimation problem

In this section, we show how Theorem 2.1 is related to the problem of state estimation over a digital channel.

Consider a dynamical system given by an ODE of the form (1). Suppose that a sensor, fully observing the state x_t of the system, sends its data to an encoder. At discrete sampling times, the encoder sends a signal e_t through a noisefree discrete channel to a decoder (without transmission delay). The decoder acts as an observer of the system, trying to reconstruct the state from the received data. For simplicity, we assume that the times of transmissions are $t = 0,1,2,\ldots$. We write x_t for the state at time t and $\hat{x}_t$ for its estimate generated by the observer. Moreover, we assume that $x_0, \hat{x}_0 \in K$ for a compact and forward-invariant set $K \subset \mathbb{R}^n$. The encoder and the observer are described by mappings

$$e_t = \mathcal{C}_t(x_0, x_1, \ldots, x_t; \hat{x}_0, \delta), \quad \mathcal{C}_t : (\mathbb{R}^n)^{t+1} \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathcal{M},$$

and

$$\hat{x}_t = \mathcal{E}_t(e_0, e_1, \ldots, e_t; \hat{x}_0, \delta), \quad \mathcal{E}_t : \mathcal{M}^{t+1} \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n.$$

The argument δ corresponds to the initial error at time zero, i.e. $\|x_0 - \hat{x}_0\| \leq \delta$. In particular, we assume that both the encoder and the observer are given the data $\hat{x}_0$ and δ.

We assume that the channel can transmit at least $b_-(r)$ and at most $b_+(r)$ bits in any time interval of length r. The *capacity* of the channel is then defined by

$$C := \lim_{r \to \infty} \frac{b_-(r)}{r} = \lim_{r \to \infty} \frac{b_+(r)}{r},$$

assuming that these limits exist and coincide.

We consider the following two observation objectives:

(O1) The observer *observes* the system with exactness $\varepsilon > 0$ if there exists $\delta = \delta(\varepsilon, K)$ so that $x_0, \hat{x}_0 \in K$ with $\|x_0 - \hat{x}_0\| \leq \delta$ implies

$$\sup_{t \geq 0} \|x_t - \hat{x}_t\| \leq \varepsilon.$$

(O2) The observer *regularly observes* the system if there exist $G, \delta_* > 0$ so that for all $\delta \in (0, \delta_*)$ and $x_0, \hat{x}_0 \in K$ with $\|x_0 - \hat{x}_0\| \leq \delta$,

$$\sup_{t \geq 0} \|x_t - \hat{x}_t\| \leq G\delta.$$

We say that the system is

- *observable on* K over a channel of capacity C if for every $\varepsilon > 0$ an observer exists which observes the system with exactness ε over this channel;
- *regularly observable on* K over a channel of capacity C if there exists an observer which regularly observes the system over this channel.

For objective (O1) we have the following result, cf. [16]:

3.1 Theorem: *The smallest channel capacity C_0, so that system* (1) *is observable on K over every channel of capacity $C > C_0$ is given by*

$$C_0 = h_{\text{top}}(\phi; K).$$

Due to the problems related to estimation policies based on topological entropy, and the gap between the initial error δ and the final exactness ε, both explained in the introduction, [17] introduces another entropy notion tailored to characterize C_0 for objective (O2).

For $t > 0$, $x \in K$ and $\delta > 0$ let $p(t, x, \delta)$ denote the minimal number of δ-balls needed to cover the image $\phi_t(B_\delta(x) \cap K)$. The *restoration entropy* of ϕ on K is given by

$$h_{\text{res}}(\phi; K) := \lim_{t \to \infty} \frac{1}{t} \limsup_{\delta \downarrow 0} \sup_{x \in K} \log_2 p(t, x, \delta).$$

The limit in t exists due to subadditivity, and the following data-rate theorem holds, cf. [17].

3.2 Theorem: *The smallest channel capacity C_0, so that system (1) is regularly observable on K over every channel of capacity $C > C_0$ is given by*

$$C_0 = h_{\mathrm{res}}(\phi; K).$$

Now, $h_{\mathrm{res}}(\phi; K)$ is a quantity that is much better behaved than $h_{\mathrm{top}}(\phi; K)$ in several respects. A first manifestation of this is the following characterization of $h_{\mathrm{res}}(\phi; K)$ in terms of the singular values of the derivative $\mathrm{D}\phi_t(x)$, cf. [17, Thm. 11]:

3.3 Theorem: *Assume that the closure of K equals the closure of its interior. Then*

$$h_{\mathrm{res}}(\phi; K) = \lim_{t\to\infty} \frac{1}{t} \max_{x\in K} \sum_{i=1}^{n} \max\{0, \log_2 \alpha_i(t,x)\}, \tag{5}$$

where $\alpha_1(t,x) \geq \ldots \geq \alpha_n(t,x)$ are the singular values of $\mathrm{D}\phi_t(x)$.

The existence of the limit in (5) follows again from subadditivity. Hence, the limit can be replaced by the infimum over all $t > 0$. From this fact, one easily sees that h_{res} depends upper semicontinuously on the system under consideration (in the C^1-topology).

We claim that Theorem 2.1 also holds with $h_{\mathrm{res}}(\phi; K)$ in place of $h_{\mathrm{top}}(\phi; K)$. A heuristic argument, neglecting the functions $v_1, \ldots, v_n$, proceeds as follows. First, one shows that formula (5) also holds if we compute the singular values of $\mathrm{D}\phi_t(x)$ with respect to some Riemannian metric on K, described by a C^1-function $P : K \to \mathcal{S}_n^+$. The adjoint of $\mathrm{D}\phi_t(x)$ is then given by

$$\mathrm{D}\phi_t(x)^* = P(x)^{-1} \mathrm{D}\phi_t(x)^\top P(\phi_t(x)),$$

hence the singular value equation can be written as

$$\det\left[\mathrm{D}\phi_t(x)^\top P(\phi_t(x)) \mathrm{D}\phi_t(x) - \lambda P(x)\right] = 0. \tag{6}$$

Thus, due to subadditivity, for every $t > 0$ we have

$$h_{\mathrm{res}}(\phi; K) \leq \frac{1}{t} \max_{x\in K} \sum_{i=1}^{n} \max\{0, \log_2 \alpha_i^P(t,x)\}, \tag{7}$$

where $\alpha_1^P(t,x) \geq \ldots \geq \alpha_n^P(t,x)$ are the square-roots of the solutions to (6). Assuming the existence of differentiable curves $\lambda : [0,\varepsilon) \to \mathbb{R}$ and $v : [0,\varepsilon) \to \mathbb{R}^n$ with $\|v(t)\| \equiv 1$ so that

$$\mathrm{D}\phi_t(x)^\top P(\phi_t(x)) \mathrm{D}\phi_t(x) v(t) = \lambda(t) P(x) v(t) \quad \text{for all } t \in [0,\varepsilon), \tag{8}$$

differentiation with respect to t at $t = 0$ yields

$$\big(\mathrm{D}f(x)^\top P(x) + P(x)\mathrm{D}f(x) + \dot{P}(x)\big) v(0) + P(x)\dot{v}(0) = \dot{\lambda}(0) P(x) v(0) + \lambda(0) P(x) \dot{v}(0).$$

For $t = 0$, equation (8) reduces to $P(x)v(0) = \lambda(0)P(x)v(0)$, hence $\lambda(0) = 1$. Consequently, the above equation is equivalent to

$$\big(\mathrm{D}f(x)^\top P(x) + P(x)\mathrm{D}f(x) + \dot{P}(x) - \dot{\lambda}(0)P(x)\big) v(0) = 0.$$

Letting $t \to 0$ in the right-hand side of (7) and comparing with Theorem 2.1 then yields the claim. For a precise formulation and proof, we refer to [17, Thm. 14].

Hence, we can conclude that Theorem 2.1, and thus our algorithm yields upper bounds for $h_{\mathrm{res}}(\phi; K)$, i.e. for the smallest channel capacity above which the estimation objective (O2) can be achieved. Moreover, the output of our algorithm can be used to implement a coding and estimation policy which leads to regular observation, as is shown in [16, 17].

4 Description of the algorithm

In this section, we describe the algorithm for computing the upper bounds provided by Theorem 2.1, which is split into two optimization problems. Before we go into details, we provide a short outline: Our algorithm aims at the computation of optimal functions P and v_i by solving two optimization problems. Starting with a triangulation $\mathcal{T}$ of the compact forward-invariant set K (or some larger set), the first optimization problem delivers a piecewise affine function P on K, affine on each simplex of the triangulation $\mathcal{T}$, with values in $\mathcal{S}_n^+$. This is accomplished by solving a semidefinite feasibility problem with linear matrix inequality constraints at each vertex and extension to the whole domain by affine interpolation of the values obtained at the vertices. The decisive quantity in this problem is a positive parameter μ, so that the solution P (if it exists) satisfies $\lambda_{\max}(x) \leq \mu$ for all $x \in K$, where $\lambda_{\max}(x)$ denotes the largest generalized eigenvalue of the pair $(A(x), P(x))$ with

$$A(x) := P(x)\mathrm{D}f(x) + \mathrm{D}f(x)^\top P(x) + (w_{ij}^\nu \cdot f(x))_{i,j=1:n},$$

where w_{ij}^ν stands for the gradient of the (i,j)-th entry of P on the simplex $\mathfrak{S}_\nu$ satisfying $x \in \mathfrak{S}_\nu$. If the algorithm yields a feasible solution for one parameter μ_1, it can be run again for a smaller parameter $\mu_2 < \mu_1$ to check if there is still a feasible solution. Repeating this procedure, the maximum of the largest generalized eigenvalues over K can be minimized.

The second optimization problems takes as an input a feasible solution P of the first problem and an upper bound $\widetilde{m}$ on the number of positive generalized eigenvalues of the matrix pairs $(A(x), P(x))$. It delivers a piecewise affine real-valued function V, affine on each simplex of a triangulation $\mathcal{T}^*$, which is a refinement of $\mathcal{T}$, and another real-valued function μ. This is done by solving a semidefinite problem with linear matrix inequality constraints at each vertex of $\mathcal{T}^*$ and extending again by affine interpolation. The optimization minimizes Q so that for all $x \in K$,

$$A(x) - \mu(x)P(x) \preceq 0 \quad \text{and} \quad \dot{V}(x) + \widetilde{m}\mu(x) \leq Q.$$

A detailed description of these two steps is given in the following subsections. The main results are Theorem 4.10 and Theorem 4.12, which show that solutions to the optimization problems, computed at the vertices of the triangulation, extend to solutions on the whole domain of interest by affine interpolation.

4.1 The semidefinite optimization problem

Given vectors $x_0, x_1, \ldots, x_n \in \mathbb{R}^n$ that are affinely independent, i.e. the vectors $x_1 - x_0, x_2 - x_0, \ldots, x_n - x_0$ are linearly independent, the convex hull

$$\mathfrak{S} = \mathrm{co}(x_0, x_1, \ldots, x_n) := \left\{ \sum_{k=0}^{n} \lambda_k x_k \, : \, \lambda_k \in [0,1] \text{ and } \sum_{k=0}^{n} \lambda_k = 1 \right\}$$

is called an n-simplex or simply a simplex. A set

$$\mathrm{co}(x_{k_0}, x_{k_1}, \ldots, x_{k_j}) := \left\{ \sum_{i=0}^{j} \lambda_{k_i} x_{k_i} \, : \, \lambda_{k_i} \in [0,1] \text{ and } \sum_{i=0}^{j} \lambda_{k_i} = 1 \right\}$$

with $0 \leq k_0 < k_1 < \ldots < k_j \leq n$ and $0 \leq j < n$ is called a j-face of the simplex $\mathfrak{S}$.

4.1 Definition: (Triangulation) *We call a finite set $\mathcal{T} = \{\mathfrak{S}_\nu\}_\nu$ of n-simplices $\mathfrak{S}_\nu$ a triangulation in $\mathbb{R}^n$ if two simplices $\mathfrak{S}_\nu, \mathfrak{S}_\mu \in \mathcal{T}$, $\mu \neq \nu$, intersect in a common face or not at all and the interior $\mathcal{D}_\mathcal{T}^\circ$ of $\mathcal{D}_\mathcal{T} := \bigcup_{\mathfrak{S}_\nu \in \mathcal{T}} \mathfrak{S}_\nu$ is connected.*

4.1 Optimization Problem *Given is a system $\dot{x} = f(x)$, $f \in C^3(\mathbb{R}^n; \mathbb{R}^n)$, a triangulation $\mathcal{T}$ in $\mathbb{R}^n$, and a parameter $\mu \geq 0$. The optimization problem is a semidefinite feasibility problem with linear matrix inequality constraints.*

Constants: *The constants used in this problem are*

1. $\epsilon_0 > 0$ – *lower bound on the matrix $P(x_k)$*

2. *The diameter h_ν of each simplex $\mathfrak{S}_\nu \in \mathcal{T}$:*

$$h_\nu := \operatorname{diam}(\mathfrak{S}_\nu) = \max_{x,y \in \mathfrak{S}_\nu} \|x - y\|_2$$

3. *Upper bounds B_ν on the second-order derivatives of the components f_k of f on each simplex $\mathfrak{S}_\nu \in \mathcal{T}$:*

$$B_\nu \geq \max_{\substack{x \in \mathfrak{S}_\nu \\ i,j,k=1:n}} \left| \frac{\partial^2 f_k(x)}{\partial x_i \partial x_j} \right| \tag{9}$$

4. *Upper bounds $B_{3,\nu}$ on the third-order derivatives of the components f_k of f on each simplex $\mathfrak{S}_\nu \in \mathcal{T}$:*

$$B_{3,\nu} \geq \max_{\substack{x \in \mathfrak{S}_\nu \\ i,j,k,l=1:n}} \left| \frac{\partial^3 f_l(x)}{\partial x_i \partial x_j \partial x_k} \right|$$

Variables: *The variables of the problem are*

1. *$P_{ij}(x_k) \in \mathbb{R}$ for all $1 \leq i \leq j \leq n$ and all vertices x_k of all simplices $\mathfrak{S}_\nu = \operatorname{co}(x_0, \ldots, x_n) \in \mathcal{T}$. For $1 \leq i \leq j \leq n$ the variable $P_{ij}(x_k)$ is the (i,j)-th entry of the $(n \times n)$ matrix $P(x_k)$. The matrix $P(x_k)$ is assumed to be symmetric and therefore these components determine it.*

2. *$C_\nu \in \mathbb{R}_0^+$ for all simplices $\mathfrak{S}_\nu \in \mathcal{T}$ – upper bound on P in $\mathfrak{S}_\nu$*

3. *$D_\nu \in \mathbb{R}_0^+$ for all simplices $\mathfrak{S}_\nu \in \mathcal{T}$ – upper bound on the derivative of P_{ij} in $\mathfrak{S}_\nu$*

Objective: *The objective function of the optimization problem is not needed because it is a feasibility problem, but one can, e.g., minimize $\max_{\mathfrak{S}_\nu \in \mathcal{T}} C_\nu$.*

Constraints:

1. **Positive definiteness of P**

   For each simplex $\mathfrak{S}_\nu = \operatorname{co}(x_0, \ldots, x_n) \in \mathcal{T}$ and each vertex x_k of $\mathfrak{S}_\nu$:

$$P(x_k) \succeq \epsilon_0 I$$

2. **Upper bound on P**

   For each simplex $\mathfrak{S}_\nu = \operatorname{co}(x_0, \ldots, x_n) \in \mathcal{T}$ and each vertex x_k of $\mathfrak{S}_\nu$:

$$P(x_k) \preceq C_\nu I$$

3. **Bound on the derivative of P**

   For each simplex $\mathfrak{S}_\nu \in \mathcal{T}$ and all $1 \leq i \leq j \leq n$:

$$\|w_{ij}^\nu\|_1 \leq D_\nu$$

   Here $w_{ij}^\nu = \nabla P_{ij}\big|_{\mathfrak{S}_\nu}(x)$ for all $x \in \mathfrak{S}_\nu$. See Remark 4.2 for details.

4. **Bounds on the largest generalized eigenvalue**
For each simplex $\mathfrak{S}_\nu = \mathrm{co}(x_0, \ldots, x_n) \in \mathcal{T}$ and each vertex x_k of $\mathfrak{S}_\nu$:

$$0 \succeq A(x_k) - \mu P(x_k) + h_\nu^2 E_\nu I$$

Here

$$A(x_k) := P(x_k)\mathrm{D}f(x_k) + \mathrm{D}f(x_k)^\top P(x_k) + (w_{ij}^\nu \cdot f(x_k))_{i,j=1:n},$$

where $\mathrm{D}f(x_k)$ is the Jacobian matrix of f at x_k, $(w_{ij}^\nu \cdot f(x_k))_{i,j=1:n}$ denotes the symmetric $(n \times n)$-matrix with entries $w_{ij}^\nu \cdot f(x_k)$ and w_{ij}^ν is defined as in (10) *and is the same vector for all vertices in one simplex. Further,*

$$E_\nu := n^2[(1 + 4\sqrt{n})B_\nu D_\nu + 2nB_{3,\nu}C_\nu].$$

4.2 Remark: In Constraints 3 and 4 in Optimization Problem 4.1, the gradient w_{ij}^ν of the affine function $P_{ij}\big|_{\mathfrak{S}_\nu}$ on the simplex $\mathfrak{S}_\nu = \mathrm{co}(x_0, \ldots, x_n)$, i.e. $\nabla P_{ij}\big|_{\mathfrak{S}_\nu} = w_{ij}^\nu$, is given by the expression

$$w_{ij}^\nu := X_\nu^{-1} \begin{pmatrix} P_{ij}(x_1) - P_{ij}(x_0) \\ \vdots \\ P_{ij}(x_n) - P_{ij}(x_0) \end{pmatrix} \in \mathbb{R}^n, \tag{10}$$

where $X_\nu = (x_1 - x_0, x_2 - x_0, \ldots, x_n - x_0)^\top \in \mathbb{R}^{n \times n}$ is the so-called shape-matrix of the simplex $\mathfrak{S}_\nu$. For a proof of this fact and, moreover, that the definition is independent of the choice of the vertex x_0, see [7, Rem. 2.9].

The Constraints 3 are indeed linear and can be implemented using the auxiliary variables D_ν^k and the constraints

$$-D_\nu^k \leq [w_{ij}^\nu]_k \leq D_\nu^k \quad \text{for } k = 1 : n,$$

where $[w_{ij}^\nu]_k$ is the k-th component of the vector w_{ij}^ν, and setting $D_\nu = \sum_{k=1}^n D_\nu^k$. Similarly, the constraints $\|\nabla \mu_\xi\|_\infty \leq D_\xi^\mu$ in Optimization Problem 4.2 can be implemented as

$$-D_\xi^\mu \leq [\nabla \mu_\xi]_k \leq D_\xi^\mu \quad \text{for } k = 1 : n.$$

4.3 Remark: The constraints above are easily transferred into the standard form $\sum_{i=1}^m F_i y_i - F_0 \succeq 0$, $F_0, F_1, \ldots, F_m \in \mathbb{R}^{n \times n}$ constant matrices and $y_1, y_2, \ldots, y_m \in \mathbb{R}$ the variables, for semidefinite programming (SDP) with linear matrix inequality (LMI) constraints. See, e.g., [6, Rem. 4.10] for a similar transfer.

4.4 Remark: The Optimization Problem 4.1 always has a feasible solution if the parameter μ is chosen large enough. Indeed, even for a fixed $P \succeq \epsilon_0 I$ the constraints are fulfilled for a large enough μ.

4.2 Feasible solution to Optimization Problem 4.1

A feasible solution of the Optimization Problem 4.1 returns a matrix $P(x_k) = (P_{ij}(x_k))_{i,j=1:n}$ at each vertex x_k of the triangulation $\mathcal{T}$ and values C_ν and D_ν for each simplex $\mathfrak{S}_\nu \in \mathcal{T}$. From these we can easily obtain $A(x_k)$ and E_ν as in Constraints 4 at each vertex x_k and for each simplex $\mathfrak{S}_\nu$, respectively.

We define the CPA (*continuous piecewise affine*) metric M by affine interpolation on each simplex.

4.5 Definition: (CPA interpolation) *Let $\mathcal{T}$ be a triangulation in $\mathbb{R}^n$ with $\mathcal{D}_\mathcal{T} = \bigcup_{\mathfrak{S}_\nu \in \mathcal{T}} \mathfrak{S}_\nu$. Let $P_{ij}(x_k)$ be fixed by a feasible solution to the Optimization Problem 4.1. An $x \in \mathfrak{S}_\nu =$*

$\mathrm{co}(x_0,\ldots,x_n)\in\mathcal{T}$ can be written uniquely as $x=\sum_{k=0}^n\lambda_k x_k$ with $\lambda_k\in[0,1]$ and $\sum_{k=0}^n\lambda_k=1$ and we define

$$P_{ij}(x):=\sum_{k=0}^n\lambda_k P_{ij}(x_k)$$

and

$$P(x):=\begin{pmatrix}P_{11}(x) & P_{12}(x) & \cdots & P_{1n}(x)\\ P_{21}(x) & P_{22}(x) & \cdots & P_{2n}(x)\\ \vdots & \vdots & \ddots & \vdots\\ P_{n1}(x) & P_{n2}(x) & \cdots & P_{nn}(x)\end{pmatrix}. \tag{11}$$

We refer to the functions P_{ij} and P as the CPA interpolations of the values $P_{ij}(x_k)$ and $P(x_k)$, respectively, where the x_k are the vertices of the simplices in $\mathcal{T}$. Furthermore, we write CPA$[\mathcal{T}]$ *for the space of all piecewise affine functions on $\mathcal{D}_\mathcal{T}$ defined in this way by interpolation of the values on the vertices of $\mathcal{T}$.*

The following lemma can be proved exactly as [6, Lem. 4.13].

4.6 Lemma: *The matrix $P(x)$ in (11) is symmetric and positive definite for all $x\in\mathcal{D}_\mathcal{T}$.*

4.7 Definition: (Orbital derivative) *Let $P(x)$ be as in Definition 4.5 and fix a point $x\in\mathcal{D}_\mathcal{T}^\circ$. As shown in the proof of [6, Lem. 4.7], there exists a $\mathfrak{S}_\nu=\mathrm{co}(x_0,\ldots,x_n)\in\mathcal{T}$ and a number $\theta^*>0$ such that $x+\theta f(x)\in\mathfrak{S}_\nu$ for all $\theta\in[0,\theta^*]$. Then $x=\sum_{k=0}^n\lambda_k x_k$ with $\lambda_k\in[0,1]$, $\sum_{k=0}^n\lambda_k=1$, and we define the orbital derivative $\dot{P}_{ij}(x)$ of P_{ij} at x as*

$$\dot{P}_{ij}(x):=w_{ij}^\nu\cdot f(x).$$

Our definition of the orbital derivative is natural, because with $t\mapsto\phi(t,x)$ as the solution to $\dot{x}=f(x)$ crossing x at time $t=0$ and for any locally Lipschitz function $g:\mathbb{R}^n\to\mathbb{R}$, we have (cf. [14, Thm. 1.17])

$$\limsup_{h\to 0+}\frac{g(\phi(h,x))-g(x)}{h}=\limsup_{h\to 0+}\frac{g(x+hf(x))-g(x)}{h}$$

and with $\mathfrak{S}_\nu$ chosen for x as in Definition 4.7, we have

$$\limsup_{h\to 0+}\frac{P_{ij}(x+hf(x))-P_{ij}(x)}{h}=w_{ij}^\nu\cdot f(x).$$

Note that $P_{ij}\big|_{\mathfrak{S}_\nu}$ is an affine function and its gradient w_{ij}^ν was defined in Constraints 3 of Optimization Problem 4.1 and is the same vector for all points $x\in\mathfrak{S}_\nu$.

Before proceeding to prove the implications of Optimization Problem 4.1, let us first recall a few elementary relations about matrix norms. For an $A\in\mathbb{R}^{n\times n}$ we define

$$\|A\|_{\max}:=\max_{i,j=1:n}|a_{ij}|\quad\text{and}\quad\|A\|_p:=\max_{\|x\|_p=1}\|Ax\|_p\ \text{ for } p=1,2,\infty.$$

The following relations hold:

$$\|A\|_{\max}\le\|A\|_2\le n\|A\|_{\max},\ \|A\|_2\le\sqrt{n}\|A\|_1,\ \text{and}\ \|A\|_2\le\sqrt{n}\|A\|_\infty.$$

For a symmetric and positive definite A, the largest singular value $\lambda_{\max}$ of A, which equals $\|A\|_2$ and is the largest of its eigenvalues, is the smallest number such that $A\preceq\lambda_{\max}I$. Further,

$$\|A\|_1=\max_{j=1:n}\sum_{i=1}^n|a_{ij}|=\|A^\top\|_\infty.$$

We will now relate $\dot{P}(x)$ to $\dot{P}(x_k)$, as well as $P(x)\mathrm{D}f(x)$ to $P(x_k)\mathrm{D}f(x_k)$. For the proof we will need the following auxiliary result, see [1, Prop. 4.1 and Cor. 4.3]. The notation is as in Optimization Problem 4.1.

4.8 Lemma: *Denoting the Hessian of f by $H(x) := \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{ij}$, we have*

$$\left\| f(x) - \sum_{k=0}^{n} \lambda_k f(x_k) \right\|_\infty \leq \max_{x \in \mathfrak{S}_\nu} \|H(x)\|_2 h_\nu^2 \leq n B_\nu h_\nu^2.$$

4.9 Lemma: *Consider a feasible solution to Optimization Problem 4.1 and let P be defined as in Definition 4.5. Fix a point $x \in \mathcal{D}_{\mathcal{T}}^{\circ}$ and a corresponding simplex $\mathfrak{S}_\nu = \mathrm{co}(x_0, x_1, \ldots, x_n) \in \mathcal{T}$ as in Definition 4.7. Set*

$$A(y) := P(y)\mathrm{D}f(y) + \mathrm{D}f(y)^\top P(y) + (w_{ij}^\nu \cdot f(y))_{i,j=1:n}$$

for all $y \in \mathfrak{S}_\nu$. Then we have the following estimate:

$$\left\| A(x) - \sum_{k=0}^{n} \lambda_k A(x_k) \right\|_2 \leq h_\nu^2 E_\nu. \tag{12}$$

Proof: We show this in several steps:

Step 1: Entry-wise bounds on $\dot{\mathbf{P}}(\mathbf{x})$

The estimate

$$\left| w_{ij}^\nu \cdot f(x) - \sum_{k=0}^{n} \lambda_k w_{ij}^\nu \cdot f(x_k) \right| \leq n B_\nu D_\nu h_\nu^2 \tag{13}$$

follows by Hölder's inequality, Constraints 3, and Lemma 4.8:

$$\left| w_{ij}^\nu \cdot \left(f(x) - \sum_{k=0}^{n} \lambda_k f(x_k) \right) \right| \leq \|w_{ij}^\nu\|_1 \left\| f(x) - \sum_{k=0}^{n} \lambda_k f(x_k) \right\|_\infty \leq D_\nu n B_\nu h_\nu^2.$$

Step 2: Entry-wise bounds on $\mathbf{P}(\mathbf{x})\mathrm{D}\mathbf{f}(\mathbf{x})$ and $\mathrm{D}\mathbf{f}(\mathbf{x})^\top \mathbf{P}(\mathbf{x})$

We show that

$$\left| [P(x)\mathrm{D}f(x)]_{ij} - \sum_{k=0}^{n} \lambda_k [P(x_k)\mathrm{D}f(x_k)]_{ij} \right| \leq n h_\nu^2 (2\sqrt{n} B_\nu D_\nu + n B_{3,\nu} C_\nu). \tag{14}$$

Consider two scalar-valued functions $g, h \in C^2(\mathfrak{S}_\nu)$. We apply Lemma 4.8 to gh, yielding

$$\left| g(x)h(x) - \sum_{k=0}^{n} \lambda_k g(x_k) h(x_k) \right| \leq \max_{y \in \mathfrak{S}_\nu} \|H(y)\|_2 h_\nu^2, \tag{15}$$

where the matrix $H(y)$ is defined by $[H(y)]_{rs} := \frac{\partial^2 (gh)(y)}{\partial x_r \partial x_s}$. Set $g(y) := P_{il}(y)$. Since $P_{il}(y) = w_{il}^\nu \cdot (y - x_0) + P_{il}(x_0)$, we get $\frac{\partial g}{\partial x_s}(y) = [w_{il}^\nu]_s$ and $\frac{\partial^2 g}{\partial x_r \partial x_s}(y) = 0$ for all $y \in \mathfrak{S}_\nu$. Hence,

$$\frac{\partial}{\partial x_s} gh = \frac{\partial g}{\partial x_s} h + g \frac{\partial h}{\partial x_s} = [w_{il}^\nu]_s h + g \frac{\partial h}{\partial x_s}$$

and then

$$\frac{\partial^2}{\partial x_r \partial x_s} gh = [w_{il}^\nu]_s \frac{\partial h}{\partial x_r} + \frac{\partial g}{\partial x_r} \frac{\partial h}{\partial x_s} + g \frac{\partial^2 h}{\partial x_r \partial x_s} = [w_{il}^\nu]_s \frac{\partial h}{\partial x_r} + [w_{il}^\nu]_r \frac{\partial h}{\partial x_s} + P_{il} \frac{\partial^2 h}{\partial x_r \partial x_s}.$$

Now set $h(y) := [\mathrm{D}f(y)]_{lj}$. Then $\frac{\partial h}{\partial x_r} = \frac{\partial^2 f_l}{\partial x_r \partial x_j}$ and $\frac{\partial^2 h}{\partial x_r \partial x_s} = \frac{\partial^3 f_l}{\partial x_r \partial x_s \partial x_j}$ and thus

$$|[H(y)]_{rs}| = \left| \frac{\partial^2 (gh)(y)}{\partial x_r \partial x_s} \right| \leq |[w_{il}^\nu]_s| B_\nu + |[w_{il}^\nu]_r| B_\nu + |P_{il}(y)| B_{3,\nu}.$$

Using in succession for any $H_1, H_2, H_3 \in \mathbb{R}^{n \times n}$ that

$$\|H_1 + H_2 + H_3\|_2 \leq \|H_1\|_2 + \|H_2\|_2 + \|H_3\|_2$$

and

$$\|H_1\|_2 \leq \sqrt{n}\|H_1\|_\infty, \ \|H_2\|_2 \leq \sqrt{n}\|H_2\|_1, \text{ and } \|H_3\|_2 \leq n\|H_3\|_{\max},$$

this delivers

$$\begin{aligned} \|H(y)\|_2 &\leq \sqrt{n}\|w_{il}^\nu\|_1 B_\nu + \sqrt{n}\|w_{il}^\nu\|_1 B_\nu + nB_{3,\nu} \max_{x \in \mathfrak{S}_\nu} \max_{1 \leq i \leq l \leq n} |P_{il}(x)| \\ &\leq 2\sqrt{n} B_\nu D_\nu + nB_{3,\nu} C_\nu, \end{aligned} \tag{16}$$

because we have $|P_{il}(y)| \leq \|P(y)\|_2 \leq C_\nu$ by Constraints 2.

Hence, (15) and (16) establish

$$\begin{aligned} &\left| [P(x)\mathrm{D}f(x)]_{ij} - \sum_{k=0}^n \lambda_k [P(x_k)\mathrm{D}f(x_k)]_{ij} \right| = \left| \sum_{l=1}^n P_{il}(x)[\mathrm{D}f(x)]_{lj} - \sum_{l=1}^n \sum_{k=0}^n \lambda_k P_{il}(x_k)[\mathrm{D}f(x_k)]_{lj} \right| \\ &\leq \sum_{l=1}^n \left| P_{il}(x)[\mathrm{D}f(x)]_{lj} - \sum_{k=0}^n \lambda_k P_{il}(x_k)[\mathrm{D}f(x_k)]_{lj} \right| \leq n \cdot (2\sqrt{n} B_\nu D_\nu + nB_{3,\nu} C_\nu) \cdot h_\nu^2. \end{aligned}$$

Step 3: Bounds on matrices
From the definition of $A(y)$ we get

$$\begin{aligned} \left\| A(x) - \sum_{k=0}^n \lambda_k A(x_k) \right\|_2 &\leq \left\| P(x)\mathrm{D}f(x) - \sum_{k=0}^n \lambda_k P(x_k)\mathrm{D}f(x_k) \right\|_2 \\ &+ \left\| \mathrm{D}f(x)^\top P(x) - \sum_{k=0}^n \lambda_k \mathrm{D}f(x_k)^\top P(x_k) \right\|_2 \\ &+ \left\| (w_{ij}^\nu \cdot f(x))_{i,j=1:n} - \sum_{k=0}^n \lambda_k (w_{ij}^\nu \cdot f(x_k))_{i,j=1:n} \right\|_2 . \end{aligned}$$

The first two norms on the right-hand side are equal because P is symmetric and therefore the matrices in the norms are conjugate. The entry-wise bounds (13) and (14) together with $\|H\|_2 \leq n\|H\|_{\max}$ for any $H \in \mathbb{R}^{n \times n}$ now deliver (12). □

Using Lemma 4.9 we can now establish, that the parameter μ in Optimization Problem 4.1 is an upper bound on the generalized eigenvalues of the matrix pair $(A(x), P(x))$ for all $x \in \mathcal{D}_\mathcal{T}^\circ$. For completeness, we first give a short description of generalized eigenvalues as needed here.

The generalized eigenvalue problem for two symmetric matrices $A, B \in \mathbb{R}^{n \times n}$, $B \succ 0$, is to find values $\lambda_i \in \mathbb{R}$ and corresponding nonzero vectors $x_i \in \mathbb{R}^n$ such that $Ax_i = \lambda_i B x_i$ for $i = 1 : n$. Since $B \succ 0$, we can define $B^{\frac{1}{2}} := O^\top D^{\frac{1}{2}} O$, where $B = O^\top D O$ is the spectral decomposition of B, i.e. O is orthogonal and D is a diagonal matrix with strictly positive entries on the diagonal. $D^{\frac{1}{2}}$ is then canonically defined as the diagonal matrix with the square-roots of the entries of D on the diagonal. Further, $B^{-\frac{1}{2}} := (B^{\frac{1}{2}})^{-1}$. The matrix $C = B^{-\frac{1}{2}} A B^{-\frac{1}{2}}$ is then symmetric and if $\lambda \in \mathbb{R}$ is an eigenvalue of C with corresponding eigenvector $y \in \mathbb{R}^n$, then $B^{-\frac{1}{2}} A B^{-\frac{1}{2}} y = \lambda y$. From this $AB^{-\frac{1}{2}} y = \lambda B^{\frac{1}{2}} y = BB^{-\frac{1}{2}} y$ or $Ax = \lambda Bx$ with $x = B^{-\frac{1}{2}} y$, i.e. λ is a generalized eigenvalue for the matrix pair A and B and x is a corresponding generalized eigenvector. With $\{y_i\}$ as an orthonormal set of eigenvectors of C, the generalized eigenvectors $x_i = B^{-\frac{1}{2}}y_i$ are thus a basis of $\mathbb{R}^n$ and $x_i^\top Bx_j = y_i^\top B^{-\frac{1}{2}}BB^{-\frac{1}{2}}y_j = y_i^\top y_j = \delta_{ij}$. It follows from an easy calculation that if $\lambda_{\max}$ is the largest generalized eigenvalue of the pair (A,B), then

$$0 \succeq A - \mu B \text{ if and only if } \mu \geq \lambda_{\max}.$$

Also note that $0 \succeq A - \mu B + \alpha I$ for an $\alpha \geq 0$ clearly implies $0 \succeq A - \mu B$. Since $B \succ 0$, the smallest eigenvalue of B is given by $\|B^{-1}\|_2^{-1}$ and then $\|B^{-1}\|_2^{-1}I \preceq B$ and with $\alpha \geq 0$ we get

$$A - \mu B + \alpha I \preceq A - \left(\mu - \alpha\|B^{-1}\|_2\right) B \preceq 0 \text{ if } \mu - \alpha\|B^{-1}\|_2 \geq \lambda_{\max}. \tag{17}$$

From this discussion on generalized eigenvalues and Lemma 4.9 we can draw the following conclusion:

4.10 Theorem: *Assume that we have a feasible solution to Optimization Problem 4.1 with parameter $\mu \geq 0$ and let $P(x)$ be defined from the feasible solution as in Definition 4.5 and define for every $x \in \mathcal{D}_\mathcal{T}^\circ$ the matrix*

$$A(x) := P(x)\mathrm{D}f(x) + \mathrm{D}f(x)^\top P(x) + (w_{ij}^\nu \cdot f(x))_{i,j=1:n}.$$

Denote for every $x \in \mathcal{D}_\mathcal{T}^\circ$ by $\lambda_{\max}(x)$ the largest generalized eigenvalue of the matrix pair $(A(x), P(x))$. Then $\mu \geq \lambda_{\max}(x)$ for every $x \in \mathcal{D}_\mathcal{T}^\circ$.

Proof: Let $x \in \mathcal{D}_\mathcal{T}^\circ$ be arbitrary and fix an $\mathfrak{S}_\nu \in \mathcal{T}$ as in Definition 4.7 such that $x \in \mathfrak{S}_\nu$. By Lemma 4.9 we have

$$\left\| A(x) - \sum_{k=0}^n \lambda_k A(x_k) \right\|_2 \leq h_\nu^2 E_\nu,$$

from which

$$A(x) - \sum_{k=0}^n \lambda_k A(x_k) \preceq \left\| A(x) - \sum_{k=0}^n \lambda_k A(x_k) \right\|_2 I \preceq h_\nu^2 E_\nu I$$

follows. But then

$$\begin{aligned} A(x) - \mu P(x) &\preceq \left\| A(x) - \sum_{k=0}^n \lambda_k A(x) \right\|_2 I + \sum_{k=0}^n \lambda_k A(x_k) - \mu \sum_{k=0}^n \lambda_k P(x_k) \\ &\preceq \sum_{k=0}^n \lambda_k \left[A(x_k) - \mu P(x_k) + h_\nu^2 E_\nu I \right] \preceq 0 \end{aligned}$$

by Constraints 4. The assertion of the lemma now follows by the discussion on generalized eigenvalues above. □

After we have found a parameter $\mu \geq 0$ and a triangulation $\mathcal{T}$ such that Optimization Problem 4.1 has a feasible solution, we can use this solution as the input to another optimization problem to get bounds on $h_{\text{top}}(\phi; K)$ as in Theorem 2.1. We now use $P(x)$ computed by Optimization Problem 4.1 to compute two functions, $\mu, V : \mathcal{D}_\mathcal{T} \to \mathbb{R}$. The function μ, satisfying $0 \leq \lambda_{\max}(x) \leq \mu(x) \leq \mu$, is a local upper bound on the largest generalized eigenvalue $\lambda_{\max}(x)$ of $(A(x), P(x))$ and V is a Lyapunov-type function as used in Theorem 2.1.

4.2 Optimization Problem *Given is a feasible solution to Optimization Problem 4.1 for a system $\dot{x} = f(x)$, $f \in C^3(\mathbb{R}^n; \mathbb{R}^n)$, with a triangulation $\mathcal{T}$, a parameter $\mu \geq 0$, and an upper bound $\widetilde{m}$ on the number of positive generalized eigenvalues of the matrix pairs $(A(x), P(x))$ from the feasible solution. Further, a refined triangulation $\mathcal{T}^*$ of $\mathcal{T}$ is given, i.e. $\mathcal{T}^*$ is a triangulation as in Definition 4.1, $\mathcal{D}_{\mathcal{T}^*} = \mathcal{D}_\mathcal{T}$, and each simplex $\mathfrak{S}_\nu \in \mathcal{T}$ is the union of simplices $\mathfrak{S}_\xi$ in $\mathcal{T}^*$.*

The optimization problem is a semidefinite problem with linear matrix inequality constraints.

Constants *The constants used in problem are:*

1. *The diameter* h_ξ *of each simplex* $\mathfrak{S}_\xi \in \mathcal{T}^*$ *:*

$$h_\xi := \operatorname{diam}(\mathfrak{S}_\xi) = \max_{x,y\in\mathfrak{S}_\xi} \|x-y\|_2$$

2. D_ν *and* E_ν *for each simplex* $\mathfrak{S}_\nu \in \mathcal{T}$ *– delivered by the feasible solution to Optimization Problem 4.1*

3. *Upper bounds* B^*_ξ *on the second-order derivatives of the components of* f *on each simplex* $\mathfrak{S}_\xi \in \mathcal{T}^*$*, just as in (9) but for the simplices in* $\mathcal{T}^*$*. For example, one can set* $B^*_\xi := B_\nu$ *for every* $\mathfrak{S}_\xi \in \mathcal{T}^*$ *fulfilling* $\mathfrak{S}_\xi \subset \mathfrak{S}_\nu \in \mathcal{T}$*.*

We additionally use the functions $A, P : \mathcal{D}^*_{\mathcal{T}} \to \mathbb{R}^{n\times n}$ *from the feasible solution to Optimization Problem 4.1.*

Variables *The variables of the semidefinite feasibility problem are:*

1. $\mu(x_k) \in \mathbb{R}$ *for all vertices* x_k *of all simplices* $\mathfrak{S}_\nu = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *– upper bound on the largest generalized eigenvalue*

2. $D^\mu_\xi \in \mathbb{R}^+_0$ *for all simplices* $\mathfrak{S}_\xi \in \mathcal{T}^*$ *– upper bound on the gradient of* μ

3. $V(x_k) \in \mathbb{R}$ *for all vertices* x_k *of all simplices* $\mathfrak{S}_\nu = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *– value of the Lyapunov-type function at* x_k

4. $D^V_\xi \in \mathbb{R}$ *for all simplices* $\mathfrak{S}_\xi \in \mathcal{T}^*$ *– upper bound on the gradient of* V

5. $Q \in \mathbb{R}^+_0$ *– the quantity to be minimized*

Objective

$$\text{minimize } Q$$

Constraints

1. **Bound on the gradient of** μ
For each simplex $\mathfrak{S}_\xi = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *:*

$$\|\nabla\mu_\xi\|_\infty \le D^\mu_\xi$$

See Remark 4.2 for details.

2. $\mu(\mathbf{x})$ **an upper bound on the generalized eigenvalues**
For each simplex $\mathfrak{S}_\xi = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *and each vertex* x_k *of* $\mathfrak{S}_\xi$ *:*

$$A(x_k) - \mu(x_k)P(x_k) + h_\xi^2(E_\nu + 2n\sqrt{n}D_\nu D^\mu_\xi)I \preceq 0 \tag{18}$$

E_ν *and* D_ν *correspond to the simplex* $\mathfrak{S}_\nu \in \mathcal{T}$ *such that* $\mathfrak{S}_\xi \subset \mathfrak{S}_\nu$*.*

3. **Bound on the gradient of V**
For each simplex $\mathfrak{S}_\xi = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *:*

$$\|\nabla V_\xi\|_1 \le D^V_\xi$$

Here $\nabla V_\xi := \nabla V\big|_{\mathfrak{S}_\xi}(x)$ *for all* $x \in \mathfrak{S}_\xi$*. It is constructed exactly as the vectors* w^ν_{ij} *in Optimization Problem 4.1, see also Remark 4.2.*

4. **Upper bound on the sum of generalized eigenvalues**
For each simplex $\mathfrak{S}_\xi = \operatorname{co}(x_0,\dots,x_n) \in \mathcal{T}^*$ *and each vertex* x_k *of* $\mathfrak{S}_\xi$ *:*

$$\nabla V_\xi \cdot f(x_k) + h_\xi^2 \cdot nB^*_\xi D^V_\xi + \widetilde{m}\mu(x_k) \le Q \tag{19}$$

The orbital derivative of V is defined as in Definition 4.7, but with the triangulation $\mathcal{T}^*$ of course.

4.11 Remark: The Optimization Problem 4.2 clearly has a solution $Q \geq \tilde{m}\mu$, where μ is the parameter from Optimization Problem 4.1 chosen such that it has a feasible solution. Indeed, just set $\mu(x_k) = \mu$ and $V(x_k) = 0$ for all vertices x_k of $\mathcal{T}^*$.

4.12 Theorem: *Consider a feasible solution to Optimization Problem 4.2 and let $\mu(x)$ and $V(x)$ be constructed as in Definition 4.5. Then for every $x \in \mathcal{D}^\circ_{\mathcal{T}^*}$ we have*

$$A(x) - \mu(x)P(x) \preceq 0 \quad \textit{and} \quad V'(x) + \tilde{m}\mu(x) \leq Q.$$

Proof: Fix a point $x \in \mathcal{D}^\circ_{\mathcal{T}^*}$ and a corresponding simplex $\mathfrak{S}_\xi = \mathrm{co}(x_0, x_1, \ldots, x_n) \in \mathcal{T}^*$ as in Definition 4.7. Further, denote by $\mathfrak{S}_\nu$ the simplex in $\mathcal{T}$ such that $\mathfrak{S}_\xi \subset \mathfrak{S}_\nu$. Now

$$\left\| A(x) - \mu(x)P(x) - \sum_{k=0}^{n} \lambda_k \left[A(x_k) - \mu(x_k)P(x_k) \right] \right\|_2$$
$$\leq \left\| A(x) - \sum_{k=0}^{n} \lambda_k A(x_k) \right\|_2 + \left\| \mu(x)P(x) - \sum_{k=0}^{n} \lambda_k \mu(x_k)P(x_k) \right\|_2 . \tag{20}$$

Just as in the proof of Lemma 4.9, we can show that

$$\left\| A(x) - \sum_{k=0}^{n} \lambda_k A(x_k) \right\|_2 \leq h_\xi^2 E_\nu. \tag{21}$$

For the second norm on the right-hand side of (20) consider two scalar-valued functions $g, h : \mathfrak{S}_\xi \to \mathbb{R}$, where $g, h \in C^2$. We apply Lemma 4.8 to gh, yielding

$$\left| g(x)h(x) - \sum_{k=0}^{n} \lambda_k g(x_k)h(x_k) \right| \leq \max_{y \in \mathfrak{S}_\xi} \|H(y)\|_2 h_\xi^2,$$

where the matrix $H(y)$ is defined by $[H(y)]_{rs} := \frac{\partial^2 (gh)(y)}{\partial x_r \partial x_s}$. Set $g(y) := P_{il}(y)$ and $h(y) := \mu(y)$. Since $\mathfrak{S}_\xi \subset \mathfrak{S}_\nu$, we have $P_{il}(y) = w_{il}^\nu \cdot (y - x_0) + P_{il}(x_0)$ and because $\mu(y)$ is defined as a CPA interpolation, we have $\mu(y) = [\nabla \mu_\xi] \cdot (y - x_0) + \mu(x_0)$. Thus, we have $\frac{\partial g}{\partial x_s}(y) = [w_{il}^\nu]_s$, $\frac{\partial^2 g}{\partial x_r \partial x_s}(y) = 0$, $\frac{\partial h}{\partial x_s}(y) = [\nabla \mu_\xi]_s$, and $\frac{\partial^2 h}{\partial x_r \partial x_s}(y) = 0$ for all $y \in \mathfrak{S}_\nu$. Hence,

$$\frac{\partial}{\partial x_s} gh = \frac{\partial g}{\partial x_s} h + g \frac{\partial h}{\partial x_s} = [w_{il}^\nu]_s h + g [\nabla \mu_\xi]_s$$

and

$$\frac{\partial^2}{\partial x_r \partial x_s} gh = [w_{il}^\nu]_s \frac{\partial h}{\partial x_r} + \frac{\partial g}{\partial x_r} [\nabla \mu_\xi]_s = [w_{il}^\nu]_s [\nabla \mu_\xi]_r + [w_{il}^\nu]_r [\nabla \mu_\xi]_s.$$

Thus

$$|[H(y)]_{rs}| = |[w_{il}^\nu]_s [\nabla \mu_\xi]_r + [w_{il}^\nu]_r [\nabla \mu_\xi]_s| \leq |[w_{il}^\nu]_s [\nabla \mu_\xi]_r| + |[w_{il}^\nu]_r [\nabla \mu_\xi]_s| .$$

Using that for any $H_1, H_2 \in \mathbb{R}^{n \times n}$ we have

$$\|H_1 + H_2\|_2 \leq \|H_1\|_2 + \|H_2\|_2 \leq \sqrt{n}\|H_1\|_1 + \sqrt{n}\|H_2\|_\infty,$$

we get

$$\|H(y)\|_2 \leq 2\sqrt{n} \|w_{il}^\nu\|_1 \|\nabla \mu_\xi\|_\infty \leq 2\sqrt{n} D_\nu D_\xi^\mu,$$

because $\|w_{il}^\nu\|_1 \leq D_\nu$ by Constraints 3 in Optimization Problem 4.1 and $\|\nabla \mu_\xi\|_\infty \leq D_\xi^\mu$ by Constraints 1 in Optimization Problem 4.2.

We have shown that

$$\left\|\mu(x)P(x)-\sum_{k=0}^{n}\lambda_k\mu(x_k)P(x_k)\right\|_{\max}\le h_\xi^2\cdot 2\sqrt{n}D_\nu D_\xi^\mu$$

and it follows that

$$\left\|\mu(x)P(x)-\sum_{k=0}^{n}\lambda_k\mu(x_k)P(x_k)\right\|_{2}\le h_\xi^2\cdot 2n\sqrt{n}D_\nu D_\xi^\mu. \tag{22}$$

Thus, we have by (20), (21), (22), and Constraints 2 of Optimization Problem 4.2

$$\begin{aligned}A(x)-\mu(x)P(x)&\preceq\sum_{k=0}^{n}\lambda_k[A(x_k)-\mu(x_k)P(x_k)]+\left\|A(x)-\sum_{k=0}^{n}\lambda_kA(x_k)\right\|_2 I\\&\quad+\left\|\mu(x)P(x)-\sum_{k=0}^{n}\lambda_k\mu(x_k)P(x_k)\right\|_2 I\\&\preceq\sum_{k=0}^{n}\lambda_k[A(x_k)-\mu(x_k)P(x_k)+h_\xi^2(E_\nu+2n\sqrt{n}D_\nu D_\xi^\mu)I]\preceq 0\end{aligned}$$

and the estimate (23) follows and, as before, it also follows that $\mu(x)$ is an upper bound on the largest generalized eigenvalue of the matrix pair $(A(x),P(x))$ for every $x\in\mathcal{D}_{\mathcal{T}}^\circ$.

Let $x\in\mathcal{D}_{\mathcal{T}^*}$ and $\mathfrak{S}_\xi$ be as above. We now show the implications of Constraints 4. By Hölder's inequality and Lemma 4.8 we get

$$\begin{aligned}\nabla V_\xi\cdot f(x)&\le\sum_{k=0}^{n}\lambda_k\nabla V_\xi\cdot f(x_k)+\left|\nabla V_\xi\cdot f(x)-\sum_{k=0}^{n}\lambda_k\nabla V_\xi\cdot f(x_k)\right|\\&\le\sum_{k=0}^{n}\lambda_k\nabla V_\xi\cdot f(x_k)+\|\nabla V_\xi\|_1\left\|f(x)-\sum_{k=1}^{n}\lambda_kf(x_k)\right\|_\infty\le\sum_{k=0}^{n}\lambda_k\nabla V_\xi\cdot f(x_k)+h_\xi^2\cdot nB_\xi^*D_\xi^V\end{aligned}$$

and then by Constraints 4 of Optimization Problem 4.2

$$\nabla V_\xi\cdot f(x)+\widetilde{m}\mu(x)\le\sum_{k=0}^{n}\lambda_k\left[\nabla V_\xi\cdot f(x_k)+h_\xi^2\cdot nB_\xi^*D_\xi^V+\widetilde{m}\mu(x_k)\right]\le\sum_{k=0}^{n}\lambda_kQ=Q,$$

i.e. the estimate (24). □

The following lemma shows that the piecewise affine functions P, V and μ computed by our algorithm can be approximated by smooth functions asymptotically satisfying the same inequalities. Hence, we do not get into trouble because of the differentiability assumptions in Theorem 2.1.

For any subset $\mathcal{D}\subset\mathbb{R}^n$ and $\varepsilon>0$ define $\mathcal{D}_{-\varepsilon}:=\{x\in\mathcal{D}\,:\,B_\varepsilon(x)\subset\mathcal{D}\}$. Define $\phi:\mathbb{R}^n\to\mathbb{R}_+$, $\phi(x):=C\exp(-1/(1-\|x\|_2))$ for $\|x\|_2<1$ and $\phi(x):=0$ otherwise and choose the constant C such that $\int_{\mathbb{R}^n}\phi(y)\,\mathrm{d}y=1$. For an $\varepsilon>0$ define

$$\widetilde{\phi}_\varepsilon(x):=\frac{\phi(x/\varepsilon)}{\varepsilon^n}.$$

For a locally integrable $g:\mathcal{D}\to\mathbb{R}$, $\mathcal{D}\subset\mathbb{R}^n$, define the function $g_\varepsilon:=g*\widetilde{\phi}_\varepsilon$, i.e. $g_\varepsilon(x)=\int_{\mathcal{D}}g(y)\widetilde{\phi}_\varepsilon(y-x)\mathrm{d}y$. It is well-known that $g_\varepsilon,\widetilde{\phi}_\varepsilon\in C^\infty(\mathbb{R}^n)$ and if g is continuous on $\mathcal{D}\subset\mathbb{R}^n$ and $\mathcal{K}\subset\mathcal{D}^\circ$ is compact, then the functions g_ε approximate g uniformly on $\mathcal{K}$, i.e. $\max_{x\in\mathcal{K}}|g_\varepsilon(x)-g(x)|\to 0$ as $\varepsilon\to 0+$.

4.13 Lemma: *Assume $g \in \mathrm{CPA}[\mathcal{T}] \to \mathbb{R}$ (cf. Definition 4.5), $\varepsilon > 0$, and denote by w_ν the gradient of g on $\mathfrak{S}_\nu$. That is, $g(x) = w_\nu \cdot x + b_\nu$ on $\mathfrak{S}_\nu$. Then for every $x \in (\mathcal{D}_\mathcal{T})_{-\varepsilon}$ we have*

$$\nabla g_\varepsilon(x) = \sum_\nu \alpha_\nu^{x,\varepsilon} w_\nu, \quad \textit{where} \;\; \alpha_\nu^{x,\varepsilon} := \int_{\mathfrak{S}_\nu \cap B_\varepsilon(x)} \widetilde{\phi}_\varepsilon(x-y)\,\mathrm{d}y.$$

Especially, the nonnegative numbers $\alpha_\nu^{x,\varepsilon}$ only depend on x and $\varepsilon > 0$ and not on the function g and they sum to one.

Proof: This follows from the following calculation, using integration by parts:

$$\begin{aligned}\nabla g_\varepsilon(x) &= \int \nabla_x \widetilde{\phi}_\varepsilon(x-y) g(y)\,\mathrm{d}y = \sum_\nu \int_{\mathfrak{S}_\nu \cap B_\varepsilon(x)} -\nabla_y \widetilde{\phi}_\varepsilon(x-y) g(y)\,\mathrm{d}y \\ &= \sum_\nu \int_{\mathfrak{S}_\nu \cap B_\varepsilon(x)} \widetilde{\phi}_\varepsilon(x-y) \nabla_y g(y)\,\mathrm{d}y = \sum_\nu \int_{\mathfrak{S}_\nu \cap B_\varepsilon(x)} \widetilde{\phi}_\varepsilon(x-y) w_\nu\,\mathrm{d}y = \sum_\nu \alpha_\nu^{x,\varepsilon} w_\nu.\end{aligned}$$

□

4.14 Lemma: *Given the same assumption as in Theorem 4.12, let $\delta > 0$. Then there exist smooth $P_\varepsilon : (\mathcal{D}_\mathcal{T})_{-\delta} \to \mathbb{R}^{n\times n}$ and $V_\varepsilon, \mu_\varepsilon : (\mathcal{D}_\mathcal{T})_{-\delta} \to \mathbb{R}$, such that for every $x \in (\mathcal{D}_\mathcal{T})_{-\delta}$ we have*

$$A_\varepsilon(x) - \mu_\varepsilon(x) P_\varepsilon(x) \preceq \delta I \tag{23}$$

and

$$\dot{V}_\varepsilon(x) + \widetilde{m}\mu_\varepsilon(x) \le Q + \delta, \tag{24}$$

where

$$A_\varepsilon(x) := P_\varepsilon(x)\mathrm{D}f(x) + \mathrm{D}f(x)^\top P_\varepsilon(x) + \dot{P}_\varepsilon(x)$$

and Q is the same constant as in Theorem 4.12.

Proof: Let P, V, and μ be defined as in Theorem 4.12 and set

$$G := \max\left\{ \max_{\substack{\mathfrak{S}_\nu \in \mathcal{T} \\ i,j=1:n}} \|w_{ij}^\nu\|_2, \max_{\xi \in \mathcal{T}^*} \|\nabla V_\xi\|_2 \right\}.$$

Fix $0 < \varepsilon < \delta$ so small that for all $x \in (\mathcal{D}_\mathcal{T})_{-\delta}$ and all y satisfying $\|x-y\|_2 < \varepsilon$ we have

$$\begin{gathered}|(B_\varepsilon)_{ij}(x) - B_{ij}(x)| < \frac{\delta}{3n}, \quad |B_{ij}(x) - B_{ij}(y)| < \frac{\delta}{3n}, \\ \|f(x) - f(y)\|_2 < \frac{\delta}{3nG}, \quad |V(x) - V(y)| < \frac{\delta}{2}, \\ |\mu(x) - \mu(y)| < \frac{\delta}{3\widetilde{m}}, \quad |\mu_\varepsilon(x) - \mu(x)| < \frac{\delta}{3},\end{gathered}$$

where the mollified functions with ε in the subscript are defined as in Lemma 4.13,

$$B(x) := P(x)\mathrm{D}f(x) + \mathrm{D}f(x)^\top P(x) - \mu(x)P(x)$$

and

$$B_\varepsilon(x) := P_\varepsilon(x)\mathrm{D}f(x) + \mathrm{D}f(x)^\top P_\varepsilon(x) - \mu_\varepsilon(x)P_\varepsilon(x).$$

Fix $x \in (\mathcal{D}_\mathcal{T})_{-\delta}$. For each $\alpha_\nu^{x,\varepsilon} > 0$, cf. Lemma 4.13, for a $\mathfrak{S}_\nu \in \mathcal{T}$ select an x_ν in the interior of $\mathfrak{S}_\nu \cap B_\varepsilon(x)$ and for each $\alpha_\xi^{x,\varepsilon} > 0$ for a $\mathfrak{S}_\xi \in \mathcal{T}^*$ select an x_ξ in the interior of $\mathfrak{S}_\xi \cap B_\varepsilon(x)$.

Now for all $i, j = 1 : n$ we have by the estimates above, Lemma 4.13, and the Cauchy-Schwarz inequality

$$\begin{aligned}(B_\varepsilon)_{ij}(x) + \nabla(P_\varepsilon)_{ij}(x)\cdot f(x) &< B_{ij}(x) + \frac{\delta}{3n} + \sum_\nu \alpha_\nu^{x,\varepsilon} w_{ij}^\nu \cdot f(x)\\ &< \sum_\nu \alpha_\nu^{x,\varepsilon}\left(B_{ij}(x_\nu) + \frac{\delta}{3n} + w_{ij}^\nu \cdot f(x_\nu) + \|w_{ij}^\nu\|_2 \frac{\delta}{3nG}\right) + \frac{\delta}{3n}\\ &< \sum_\nu \alpha_\nu^{x,\varepsilon}\left(B_{ij}(x_\nu) + w_{ij}^\nu \cdot f(x_\nu)\right) + \frac{\delta}{n}.\end{aligned}$$

Hence,

$$\begin{aligned}A_\varepsilon(x) - \mu_\varepsilon(x)P_\varepsilon(x) = B_\varepsilon(x) + \dot{P}_\varepsilon(x) &\preceq \sum_\nu \alpha_\nu^{x,\varepsilon}\left(B(x_\nu) + \dot{P}(x_\nu)\right) + \frac{\delta}{n}(1)_{ij}\\ &= \sum_\nu \alpha_\nu^{x,\varepsilon}\left(A(x_\nu) - \mu(x_\nu)P(x_\nu)\right) + \frac{\delta}{n}(1)_{ij} \preceq \delta I.\end{aligned}$$

Similarly,

$$\begin{aligned}\dot{V}_\varepsilon(x) + \widetilde{m}\mu_\varepsilon(x) &= \sum_\xi \alpha_\xi^{x,\varepsilon}\left(\nabla V_\xi \cdot f(x) + \widetilde{m}\mu(x)\right) + \frac{\delta}{3}\\ &= \sum_\xi \alpha_\xi^{x,\varepsilon}\left(\nabla V_\xi \cdot f(x_\xi) + \widetilde{m}\mu(x_\xi)\right) + \frac{\delta}{3n} + \frac{\delta}{3} + \frac{\delta}{3} \le Q + \delta.\end{aligned}$$

□

Note that the Constraints 1 and 2 in Optimization Problem 4.2 are not very strongly coupled to the Constraints 3 and 4. Constraints 2 balance the values $\mu(x_k)$ and the gradient D_ξ^μ, whereas Constraints 4 do not have to take the gradient of μ into account. Since the gradient is multiplied by h_ξ^2, which is small for small simplices $\mathfrak{S}_\xi \in \mathcal{T}^*$, the gradient can be rendered less important in Constraints 2 by using smaller simplices. It is thus tempting to split Optimization Problem 4.2 into two optimization problems, the first with Constraints 1 and 2 and some objective function that makes the collection of the $\mu(x_k)$ small in some sense, and then consecutively run an optimization problem with Constraints 3 and 4, where the $\mu(x_k)$ from a solution to the first optimization problem are constants. This is especially tempting, because SDP solvers have not reached the maturity of linear programming solvers and are sometimes not able to deliver solutions to moderately sized feasible problems or worse, deliver solutions that are quite far from being feasible.

Further, if we use Optimization Problem 4.1 to find a constant matrix $P(x)$, the optimization problem is much smaller and easier to solve. From such a solution the Optimization Problem 4.2 can be naturally split into two optimization problems as described above without any disadvantage, because the coupling between Constraints 1 and 2 on the one hand and Constraints 3 and 4 on the other hand vanishes completely. Even better, since $P(x)$ is constant, its gradient is zero and therefore $D_\nu = 0$. Thus, the gradient of μ plays no role, because its upper bound D_ξ^μ is multiplied by D_ν in Constraints 2. We can thus drop Constraints 1 and compute the optimal $\mu(x_k)$ directly.

First, for each vertex x_k we find the minimum $\mu(x_k)$ such that (18) is fulfilled for every ν such that x_k is a vertex of $\mathfrak{S}_\nu$. Then we minimize Q under the linear constraints (19). This is described in more detail in the next section.

4.3 Simplified procedure

In the simplified procedure we restrict our search for a matrix $P(x)$ in Optimization Problem 4.1 to a constant matrix and then split Optimization Problem 4.2 into two simpler problems. In detail:

In Optimization Problem 4.1 we set $P_{ij}(x_k) := P_{ij}$ and then $P(x_k) := (P_{ij})$ for all vertices x_k of all simplices of $\mathcal{T}$. Then clearly we can set $D_\nu := 0$ and $C := C_\nu$ for all $\mathfrak{S}_\nu \in \mathcal{T}$ and the constraints simplify to:

$$\epsilon_0 I \preceq P \preceq CI$$

and for each simplex $\mathfrak{S}_\nu = \mathrm{co}(x_0, \ldots, x_n) \in \mathcal{T}$ and each vertex x_k of $\mathfrak{S}_\nu$:

$$0 \succeq A(x_k) - \mu P + h_\nu^2 \cdot 2n^3 B_{3,\nu} CI,$$

where

$$A(x_k) = P\mathrm{D}f(x_k) + \mathrm{D}f(x_k)^\top P,$$

because w_{ij}^ν, the gradient of P, is now the zero vector.

Let us now consider Optimization Problem 4.2. The Constraints 2 become: For each simplex $\mathfrak{S}_\xi = \mathrm{co}(x_0, \ldots, x_n) \in \mathcal{T}^*$ and each vertex x_k of $\mathfrak{S}_\nu$:

$$0 \succeq A(x_k) - \mu(x_k)P + h_\xi^2 \cdot 2n^3 B_{3,\nu} CI, \tag{25}$$

because $D_\nu = 0$. The variables D_ξ^μ are thus redundant and we can eliminate Constraints 1. An even farther reaching consequence is that we do not even have to combine the Constraints (25) with Constraints 3 and 4. We can compute the optimal $\mu(x_k)$ locally for each vertex x_k by solving: For each vertex x_k of the triangulation $\mathcal{T}^*$ maximize the value $\mu(x_k)$ under the constraints

$$0 \succeq A(x_k) - \mu(x_k)P + h_\xi^2 \cdot 2n^3 B_{3,\nu} CI$$

for every ν such that $x_k \in \mathfrak{S}_\nu \in \mathcal{T}$.

An even simpler version of this optimization problem is obtained by defining

$$B_3^y := \max_{\mathfrak{S}_\nu \in \mathcal{T}} B_{3,\nu}$$

and solving: For each vertex x_k of the triangulation $\mathcal{T}^*$ maximize the value $\mu(x_k)$ under the constraints

$$0 \succeq A(x_k) - \mu(x_k)P + h_\xi^2 \cdot 2n^3 B_3^{x_k} CI.$$

For small $h_\xi > 0$ good estimates on these optimal $\mu(x_k)$ for the Optimization Problem 4.2 can be directly computed using (17). Just set

$$\mu(x_k) := \lambda_{\max}(x_k) + h_\xi^2 \cdot 2n^3 B_3^{x_k} C \|P^{-1}\|_2,$$

where $\lambda_{\max}(x_k)$ is the largest generalized eigenvalue of the matrix pair $(A(x_k), P)$. Since $C \geq \|P\|_2$, this formula can be further simplified to

$$\mu(x_k) := \lambda_{\max}(x_k) + h_\xi^2 \cdot 2n^3 B_3^{x_k} \kappa_2(P),$$

using the condition number $\kappa_2(P) := \|P\|_2 \|P^{-1}\|_2$ of P.

After this being done, we can minimize Q under the Constraints 3 and 4 of Optimization Problem 4.2, and this is a linear programming problem, for which much more mature solvers exist.

5 An example: the Lorenz system

We consider the Lorenz system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\sigma x + \sigma y \\ rx - y - xz \\ -bz + xy \end{pmatrix} =: g(x, y, z). \tag{26}$$

For our approach it is advantageous to scale the system such that its attractors are contained in a smaller set. For this purpose, define $S := \mathrm{diag}(s_x, s_y, s_z)$ for constants $s_x, s_y, s_z > 0$ and consider the system $\dot{\mathbf{x}} = f(\mathbf{x})$ with $f(\mathbf{x}) = S^{-1}g(S\mathbf{x})$, i.e. the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\sigma x + \sigma \frac{s_y}{s_x} y \\ r\frac{s_x}{s_y}x - y - \frac{s_x s_z}{s_y} xz \\ -bz + \frac{s_x s_y}{s_z} xy \end{pmatrix}. \tag{27}$$

Clearly one can take

$$B_\nu = s_x \cdot \max\left\{\frac{s_y}{s_z}, \frac{s_z}{s_y}\right\}$$

in the optimization problems for all $\mathfrak{S}_\nu$, independent of the triangulation $\mathcal{T}$, because the right-hand side is a global bound on the second-order derivatives of f, and we can set $B_{3,\nu} := 0$ for all $\mathfrak{S}_\nu$, because the components of f are second-order polynomials. We use the scaling parameters $s_x = 24.5$ and $s_y = s_z = 100$ and thus $B_\nu = 24.5$ in what follows.

In [2, §2.2] it is shown that if $\sigma \geq 1$ and $b \geq 2$ in (26), then the system is dissipative in the sense of Levinson and the region $\mathcal{D}$ of dissipation fulfills

$$\mathcal{D} \subset \left\{(x, y, z) \in \mathbb{R}^3 \,:\, x^2 + y^2 + (z - [\sigma + r])^2 \leq \frac{b}{2}(\sigma + r)^2\right\}, \tag{28}$$

$$\mathcal{D} \subset \left\{(x, y, z) \in \mathbb{R}^3 \,:\, 2x^2 + y^2 + (z - [\sigma + r])^2 \leq \left[1 + \frac{(b-2)^2}{4(b-1)}\right](\sigma + r)^2\right\}, \tag{29}$$

$$\mathcal{D} \subset \left\{(x, y, z) \in \mathbb{R}^3 \,:\, y^2 + (z - r)^2 \leq \frac{b^2 r^2}{4(b-1)}\right\}, \tag{30}$$

$$\mathcal{D} \subset \left\{(x, y, z) \in \mathbb{R}^3 \,:\, z \geq 0\right\}. \tag{31}$$

All attractors of (26) are inside $\mathcal{D}$. For the common parameters $\sigma = 10$, $r = 28$, and $b = 8/3$ in (26), it suffices to compute our metric P and Lyapunov-type function V on the set

$$\mathcal{K} := [-1, 1] \times [-0.29, 0.29] \times [0, 0.57],$$

because with these parameters (30) implies

$$|y| \leq \frac{4}{\sqrt{15}} r \leq 29 = s_y \cdot 0.29$$

and (30) and (31) imply

$$0 \leq z \leq \left(1 + \frac{4}{\sqrt{15}}\right) r \leq 57 = s_z \cdot 0.57,$$

which in turn with (29) implies that

$$|x| \leq \sqrt{\frac{1}{2}\left(\frac{16}{15}(\sigma + r)^2 - \left(\frac{4}{\sqrt{15}} r - \sigma\right)^2\right)} \leq 24.5 = s_x \cdot 1.$$

Thus, all the attractors of (27) are inside of $\mathcal{K}$.

For our example we used the simplified procedure from Section 4.3 and the computer we used has an i9-7900X CPU.

5.1 Remark: The implementation of our algorithm, even the simplified one, is not simple, and a detailed discussion of it is beyond the scope of this paper. We refer the reader to [9, 10] for some implementation details on triangulations for the computation of Lyapunov functions. We used similar methods, adapted to our problem.

First we compute a constant metric $P(x) = P$ on the set $\mathcal{K}$ using a triangulation with the vertices

$$\left(1.0 \cdot \frac{i_x}{12}, 0.29 \cdot \frac{i_y}{6}, 0.57 \cdot \frac{i_z}{10}\right) \quad \text{for } i_x = -12 : 12,\ i_y = -6 : 6, \text{ and } i_z = 0 : 10.$$

The triangulation we used is a so-called *standard-triangulation* as in [8, §4], but with different scaling along the different axes, i.e. ρ in [8, Def. 4.8] is $\rho_x = 1/12$ along the x-axis, $\rho_y = 0.29/6$ along the y-axis, and $\rho_z = 0.57/10$ along the z-axis. We set $\epsilon_0 := 0.1$ in Optimization Problem 4.1 and the constraints become

$$0.1I \preceq P \preceq CI$$

for each simplex $\mathfrak{S}_\nu \in \mathcal{T}$ and for each vertex x_k of $\mathfrak{S}_\nu$:

$$0 \succeq P\mathrm{D}f(x_k) + \mathrm{D}f(x_k)^\top P - \mu P.$$

The optimization problem is especially simple because $B_{3,\nu} = 0$ for all $\mathfrak{S}_\nu$, and therefore we do not even have to repeat the second constraints for all $\mathfrak{S}_\nu$, i.e. we can replace *for each simplex* $\mathfrak{S}_\nu \in \mathcal{T}$ *and for each vertex* x_k *of* $\mathfrak{S}_\nu$ with *for each vertex* x_k *of a simplex in* $\mathcal{T}$.

By trying out a few different μs, we obtained a feasible solution with $\mu = 27$. Investigation gave us that there is only one positive generalized eigenvalue for all x, so we can take $\widetilde{m} = 1$. Writing the problem took a few seconds with our software and solving the optimization problem took 140 sec. using the solver PENSDP 2.2 [12]. The problem has 7 variables and 69,122 matrix constraints. The matrix computed is

$$P = \begin{pmatrix} 0.1008469737786 & -0.01415360101927 & 0 \\ -0.01415360101927 & 0.3361537095909 & 0 \\ 0 & 0 & 0.3139832543019 \end{pmatrix}$$

which has $\|P^{-1}\|_2 = 0.1$ as the smallest and $C = \|P\|_2 = 0.3370007906512$ as the largest eigenvalue. If we only used these results in the formula in [19, Thm. 3.2] for the upper bound on the topological/restoration entropy, i.e. set $V(x)$ =const., this P delivers $\lambda/(2\ln(2)) \approx 19.4764$ as an upper bound.

The formula (13) in [19] delivers the upper bound

$$\frac{1}{2\ln(2)}\left(\sqrt{(\sigma-1)^2 + 4r\sigma} - (\sigma+1)\right) \approx 17.0638 \tag{32}$$

for our parameters.

For computing the Lyapunov-type function we used the triangulation $\mathcal{T}^*$, which is constructed exactly as the triangulation $\mathcal{T}$ above, but with the vertices

$$\left(1.0 \cdot \frac{i_x}{N_x}, 0.29 \cdot \frac{i_y}{N_y}, 0.57 \cdot \frac{i_z}{N_z}\right) \quad \text{for } i_x = -N_x : N_x,\ i_y = -N_y : N_y, \text{ and } i_z = -1 : N_z, \tag{33}$$

for some $N_x, N_y, N_z \in \mathbb{N}$. We tried a few different sets of parameter values, expecting lower upper bounds on the topological/restoration entropy for larger values of N_x, N_y, and N_z. We used the state of the art solver GUROBI, which is free for academic use, to solve the LP problems using the barrier method.

The first set of parameters was $N_x = 30$, $N_y = 14$, and $N_z = 28$ and with those the LP problem with 930,032 variables and 3,800,122 constraints was written in 11 sec. with our software and solved in 291 sec. with the optimal value of $Q = 23.9094$ which delivers the upper bound 17.247, which is slightly worse than in (32).

The second set of parameters was $N_x = 42$, $N_y = 14$, and $N_z = 28$ and with those the LP problem with 1,301,696 variables and 5,320,176 constraints was written in 28 sec. with our software and

N_x	N_y	N_z	time [s]	impr. bounds	Q	u.b.
30	14	28	302	No	23.909	17.247
30	14	28	448	Yes	22.540	16.260
42	14	28	917	No	23.525	16.970
42	14	28	536	Yes	22.094	15.937
50	18	32	901	Yes	21.701	15.654
70	22	40	5862	Yes	21.311	15.373

Table 1: The results of our computations. N_x, N_y, N_z are the parameters for the grid in (33), 'time' is the total time in seconds needed to write and solve the problem, 'impr. bounds' states whether the improved bounds discussed in the text are used (Yes) or not (No), Q is the objective that is minimized in Optimization Problem 4.2, and 'u.b.' is the associated upper bound on the topological/restoration entropy. For reference, the upper bound 17.064 is computed in [19].

solved in 889 sec. with the optimal value of $Q = 23.5254$ which delivers the upper bound 16.970, which is slightly better than in (32).

Because we are using such a simple axially parallel triangulation, one can use somewhat less conservative bounds the LP problems. That is, the term $nh_\xi^2 B_\xi^*$ in Constraints 4 in Optimization Problem 4.2 can be replaced with a smaller number and Theorem 4.12 still holds true. For these less conservative bounds we refer to [14, Lem. 4.16]. Using these less conservative bounds in the LP problems gave notably better results. Using the first set of parameters, the LP problem was solved in 437 sec. with the optimal value of $Q = 22.5403$, which delivers the upper bound 16.260, and using the second set of parameters the LP problem was solved in 508 sec. with the optimal value of $Q = 22.094$, which delivers the upper bound 15.937.

For the third set of parameters we took $N_x = 50$, $N_y = 18$, and $N_z = 32$ and only used the less conservative bounds on the second-order derivatives of f. The LP problem had 2,265,460 variables and 9,266,354 constraints, was written in 24 sec. with our software and solved in 877 sec. with the optimal value of $Q = 21.701$ which delivers the upper bound 15.654.

The fourth and final set of parameters was $N_x = 70$, $N_y = 22$, and $N_z = 40$ and only used the less conservative bounds on the second-order derivatives of f. The LP problem had 4,812,572 variables and 19,699,632 constraints, was written in 61 sec. with our software and solved in 5,801 sec. with the optimal value of $Q = 21.311$ which delivers the upper bound 15.373.

Thus, the best estimate we got with our method was the upper bound 15.373 on the topological/restoration entropy, which is considerably better than 17.064 given by formula (13) in [19].The results of the computations are summarized in Table 1.

Running the full Optimization algorithms 4.1 and 4.2 and thus computing a nonconstant matrix P would be very interesting, but is hardly possible for examples of interest with today's SDP problems solvers. It remains interesting to see if these solvers mature enough in the near future for this to change and how much the upper bound decreases using our fully fledged method.

6 Concluding remarks

In this paper, we proposed an algorithm for computing upper bounds on the critical channel capacity for state estimation over a finite-capacity channel, a typical problem studied in networked control. The upper bounds computed by our algorithm are, at the same time, upper bounds on the topological entropy of the dynamical system under consideration. Moreover, the output of the algorithm can be used to implement a coding and estimation policy which operates over a channel of the corresponding capacity.

It is not hard to see that topological entropy, in general, cannot be approximated very well by our algorithm, since the computed values are upper bounds on restoration entropy h_{res}, as shown in Section 3, and the strict inequality $h_{\mathrm{top}} < h_{\mathrm{res}}$ holds for most dynamical systems. Hence, we do not claim that our paper contributes to the problem of numerical computation of topological entropy. Some standard references on this quite intricate subject (for multi-dimensional systems) are [3, 4, 5, 18].

At this point, it is not clear whether restoration entropy can be approximated (and not only upper-bounded) by the estimates of Theorem 2.1. We believe, however, that this is the case and hope to deliver a proof in a future work.

Our full algorithm, using both Optimization Problem 4.1 and Optimization Problem 4.2 for a non-constant matrix P, overstrains currently even the best semidefinite-programming solvers in problems of interest. It will be interesting to see if this situation changes in the near future. Therefore, we derived a simplified algorithm in Section 4.3, which computes an then uses a constant P. Using this simplified algorithm allowed us to study the Lorenz system with our method and we got superior results to [19], where an analytical bound is derived. It remains an interesting question how much the full algorithm can improve these bounds.

References

[1] R. Baier, L. Grüne, S. Hafstein. *Linear programming based Lyapunov function computation for differential inclusions.* Discrete Contin. Dyn. Syst. Ser. B 17(1) (2012), 33–56.

[2] V. Boichenko, G. Leonov, V. Reitmann. *Dimension Theory for Ordinary Differential Equations.* Teubner, 2005.

[3] Q. Chen, E. Ott, L. Hurd. *Calculating topological entropies of chaotic dynamical systems.* Phys. Lett. A 156 (1991), no. 1–2, 48–52.

[4] G. D'Alessandro, P. Grassberger, S. Isola, A. Politi. *On the topology of the Hénon map.* J. Phys. A 23 (1990), no. 22, 5285–5294.

[5] G. Froyland, O. Junge, G. Ochs. *Rigorous computation of topological entropy with respect to a finite partition.* Phys. D 154 (2001), no. 1–2, 68–84.

[6] P. Giesl, S. Hafstein. *Construction of a CPA contraction metric for periodic orbits using semidefinite optimization.* Nonlinear Anal. 86 (2013), 114–134.

[7] P. Giesl, S. Hafstein. *Revised CPA method to compute Lyapunov functions for nonlinear systems.* J. Math. Anal. Appl. 410 (2014), 292–306.

[8] P. Giesl, S. Hafstein. *Computation and Verification of Lyapunov Functions.* SIAM J. Appl. Math. 14 (4) (2015), 1663–1698.

[9] S. Hafstein. *Implementation of Simplicial Complexes for CPA functions in C++11 using the Armadillo Linear Algebra Library.* In Proceedings of 4th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), Reykjavik, Iceland, 2013, 49–57.

[10] S. Hafstein. *Efficient Algorithms for Simplicial Complexes Used in the Computation of Lyapunov Functions for Nonlinear Systems.* In Proceedings of 7th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), Madrid, Spain, 2017, 398–409.

[11] C. Kawan, S. Yüksel. *On optimal coding of non-linear dynamical systems.* IEEE Trans. Inform. Theory 64 (2018), no. 10, 6816–6829.

[12] M. Kocvara, M. Stingl. *PENNON: Software for Linear and Nonlinear Matrix Inequalities. In: Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer 2012, 755–794.

[13] D. Liberzon, S. Mitra. *Entropy and minimal bit rates for state estimation and model detection.* IEEE Trans. Automat. Control 63 (2018), no. 10, 3330–3344.

[14] S. Marinósson. *Stability Analysis of Nonlinear Systems with Linear Programming: A Lyapunov Functions Based Approach.* PhD thesis: Gerhard-Mercator-University Duisburg, Duisburg, Germany, 2002.

[15] A. Matveev, A. Pogromsky. *A topological entropy approach for observation via channels with limited data rate.* IFAC Proceedings Volumes 44.1 (2011), 14416–14421.

[16] A. Matveev, A. Pogromsky. *Observation of nonlinear systems via finite capacity channels: constructive data rate limits.* Automatica J. IFAC 70 (2016), 217–229.

[17] A. Matveev, A. Pogromsky. *Observation of nonlinear systems via finite capacity channels. Part II: Restoration entropy and its estimates.* Submitted, 2017.

[18] S. Newhouse, T. Pignataro. *On the estimation of topological entropy.* J. Statist. Phys. 72 (1993), no. 5–6, 1331–1351.2016).

[19] A. Pogromsky, A. Matveev. *Estimation of topological entropy via the direct Lyapunov method.* Nonlinearity 24 (2011), no. 7, 1937–1959.

[20] A. V. Savkin. *Analysis and synthesis of networked control systems: topological entropy, observability, robustness and optimal control.* Automatica J. IFAC 42 (2006), no. 1, 51–62.