Department of Mathematics and Systems Analysis

Infinite Dimensional Systems: Passivity and Kalman Filter Discretization

Atte Aalto

Aalto University

DOCTORAL DISSERTATIONS

Atte Aalto Infinite Dimensional Systems: Passivity and Kalman Filter Discretization Aalto University

Aalto University publication series
DOCTORAL DISSERTATIONS 160/2014

Infinite Dimensional Systems: Passivity and Kalman Filter Discretization

Atte Aalto

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall M1 of the school on 28 November 2014 at 12.

Aalto University
School of Science
Department of Mathematics and Systems Analysis

Supervising professor
Prof. Rolf Stenberg

Thesis advisor
Dr. Jarmo Malinen

Preliminary examiners
Prof. Alessandro Macchelli, University of Bologna, Italy
Dr. Philippe Moireau, Inria Saclay, France

Opponent
Prof. Giorgio Picci, University of Padova, Italy

Aalto University publication series
DOCTORAL DISSERTATIONS 160/2014

ISBN 978-952-60-5909-9 (printed)
ISBN 978-952-60-5910-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5910-5

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Atte Aalto

Name of the doctoral dissertation
Infinite Dimensional Systems: Passivity and Kalman Filter Discretization

Publisher School of Science

Unit Department of Mathematics and Systems Analysis

Series Aalto University publication series DOCTORAL DISSERTATIONS 160/2014

Field of research Mechanics

Manuscript submitted 19 August 2014 **Date of the defence** 28 November 2014

Permission to publish granted (date) 7 October 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

The results of this thesis can be divided into two categories, well-posedness and passivity of boundary control systems and Kalman filter discretization. It is shown that a composition of internally well-posed, impedance passive boundary control systems through Kirchhoff type couplings is also an internally well-posed, impedance passive boundary control system. The concept of a passive majorant is defined and it is shown that boundary control systems that possess a passive majorant are internally well-posed, passive boundary control systems.

The effect of both temporal and spatial discretization to Kalman filtering is studied. Firstly, convergence speed rates are derived for the convergence of the discrete time Kalman filter estimate to the continuous time estimate as the temporal discretization is refined. This result is established for various types of linear systems. Secondly, we derive the optimal one-step state estimate that takes values in a given finite dimensional subspace of the system's state space for a linear discrete-time system with Gaussian input and output noise. An upper bound is given for the error due to the spatial discretization.

Keywords Infinite dimensional systems, boundary control systems, passive systems, well-posedness, state estimation, Kalman filter, spatial discretization, temporal discretization

ISBN (printed) 978-952-60-5909-9 **ISBN (pdf)** 978-952-60-5910-5

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 157 **urn** http://urn.fi/URN:ISBN:978-952-60-5910-5

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Atte Aalto

Väitöskirjan nimi
Ääretönulotteiset systeemit: passiivisuus ja Kalman-suotimen diskretointi

Julkaisija Perustieteiden korkeakoulu

Yksikkö Matematiikan ja systeemianalyysin laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 160/2014

Tutkimusala Mekaniikka

Käsikirjoituksen pvm 19.08.2014 **Väitöspäivä** 28.11.2014

Julkaisuluvan myöntämispäivä 07.10.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Väitöskirjan tulokset voidaan jakaa kahteen luokkaan, reunakontrollisysteemien hyvinasetettuus ja passiivisuus sekä Kalman-suotimen diskretointi. Työssä osoitetaan, että hyvinasetettuja impedanssipassiivisia reunakontrollisysteemejä Kirchhoffin lakien kaltaisilla ehdoilla kytkemällä aikaansaatu kompositiosysteemi on myös hyvinasetettu impedanssipassiivinen reunakontrollisysteemi. Työssä määritellään myös passiivisen majorantin käsite ja näytetään, että reunakontrollisysteemi, jolla on passiivinen majorantti on hyvinasetettu ja passiivinen.

Sekä aika- että paikkadiskretoinnin vaikutusta Kalman-suodatukseen tarkastellaan. Ensin johdetaan suppenemisnopeusestimaatteja diskreettiaikaisen Kalman-suotimen antamalle tilaestimaatille, joka konvergoi jatkuva-aikaiseen tilaestimaattiin, kun aika-askellusta tihennetään. Tämä tulos johdetaan useille erilaisille lineaarisille systeemeille. Toiseksi johdetaan optimaalinen yksiaskeltilaestimaatti annetussa tila-avaruuden äärellisulotteisessa aliavaruudessa lineaariselle diskreettiaikaiselle systeemille, johon vaikuttaa Gaussinen kohinaprosessi tilaan ja systeemin ulostuloon. Työssä johdetaan myös yläraja paikkadiskretoinnista johtuvalle virheelle tilaestimaatissa.

Avainsanat ääretönulotteiset systeemit, reunakontrollisysteemit, passiiviset systeemit, hyvinasetettuus, tilaestimointi, Kalman-suodin, paikkadiskretointi, aikadiskretointi

ISBN (painettu) 978-952-60-5909-9 **ISBN (pdf)** 978-952-60-5910-5

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 157 **urn** http://urn.fi/URN:ISBN:978-952-60-5910-5

Preface

In the summer of 2007, I was working on a project work on Loewner and Bézout matrices. At some point during the summer I thought that maybe I should know where such matrices actually appear. I found out that they arise from a field called *control theory*. It sounded interesting and after reading a bit more about it I was convinced that I want it to be my field of study (although in my studies I never again heard of those matrices). After a master's-thesis-sized detour to the world of biomechanical modeling, I got to tackle the more theoretical studies, and now — five years later — it gives me great pleasure to finally announce my contribution to mathematical systems theory.

I want to take the opportunity to thank the people who have influenced the thesis work in one way or another. Firstly, I want to acknowledge the Finnish Graduate School in Engineering Mechanics for funding during 2010–2013. I want to thank my advisor, Dr. Jarmo Malinen for patient guidance along the way, and my supervisor, Prof. Rolf Stenberg for valuable help especially towards the end of the process. I also thank the preliminary examiners, Prof. Alessandro Macchelli and Dr. Philippe Moireau for reviewing the thesis manuscript and Prof. Giorgio Picci for agreeing to be my opponent. I thank Peter Seenan for English proofreading. For neverending support, I thank my family, Liisa, Jorma, Saija, and Henni, and my friends, of whom I want to mention Lauri Viitasaari, Jaakko Kurttila, current and former members of the "coffee room gang", Matti, Janne, Eerno, Linda, Mikko,... and close friends from m/aux Inga-Lill. Finally, I want to thank Helle for everything!

Espoo, October 16, 2014,

Atte Aalto

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I A. Aalto and J. Malinen. Compositions of Passive Boundary Control Systems. *Mathematical Control and Related Fields*, **3**, 1–19, March 2013.

II A. Aalto. Convergence of discrete time Kalman filter estimate to continuous time estimate. *http://arxiv.org/abs/1408.1275*, 21 pages, August 2014.

III A. Aalto. Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems. *http://arxiv.org/abs/1406.7160*, 19 pages, October 2014.

IV A. Aalto and T. Lukkari and J. Malinen. Acoustic wave guides as infinite-dimensional dynamical systems. Accepted for publication in *ESAIM: Control, Optimization and Calculus of Variations*, 35 pages, June 2013.

Author's Contribution

Publication I: "Compositions of Passive Boundary Control Systems"

The author is responsible for most of the research work and writing the manuscript. The problem originates from J. Malinen.

Publication II: "Convergence of discrete time Kalman filter estimate to continuous time estimate"

This is an individual work of the author.

Publication III: "Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems"

This is an individual work of the author.

Publication IV: "Acoustic wave guides as infinite-dimensional dynamical systems"

The author has participated in formulating the results of Section 3.

1. Introduction

Systems theory is a field of mathematics and engineering studying phenomena that can be controlled and observed through particular external signals. The underlying physical system is often called *a plant*. The signal affecting the system is called *input* and the observed signal is called *output*. This is illustrated on the left in Figure 1.1. At the end of Section 1.1, we list different types of problems that are typically addressed in mathematical systems theory. Let us take here a more historical perspective and present the example that led to the emergence of mathematical systems theory. In the example, the plant being controlled is a steam engine with varying load. The input u is the opening of the valve controlling the steam flow to the engine. The output y is the rotational speed of the engine. Of course if the valve is not adjusted when the engine load increases, the engine will slow down. To compensate the variations in the load, one can design a controller that somehow converts the output to an input signal in such a manner that reducing the rotational speed makes the valve open and vice versa. This principle is called *feedback control* and it is illustrated on the right in Figure 1.1, with K denoting the controller. James Watt designed a feedback controller for the steam engine, called a centrifugal governor. In his design, there are two masses attached to rods which, in turn, are attached to a central axle by a hinge mechanism. The rotation of the axle causes a centrifugal force pushing the two masses away from the axle, and the hinge mechanism converts this movement into a control of the valve.

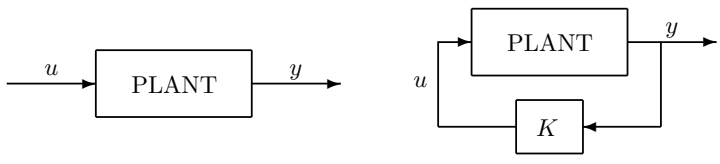

Figure 1.1. Left: A system with input u and output y. Right: The principle of feedback control.

Watt's controller was by no means the first feedback control mechanism ever developed, but its occasional instability prompted James Maxwell to do research on the matter. In his article [36] from 1868, titled "On governors", he noted that the motion of the controlled system consists of a steady motion and an additive perturbation. He divided these perturbations into four categories: increasing, diminishing, oscillation with increasing amplitude, and oscillation with diminishing amplitude. In short, he then derives differential equations and the corresponding characteristic polynomials for the coupled mechanical systems and concludes that for the system to be stable, the real parts of the roots of the characteristic polynomial must be negative. Maxwell's article is usually regarded as the starting point of mathematical systems theory.

This example also illuminates the methodology of mathematical systems theory. The first task in control problems is to develop a mathematical model for the plant. This modeling can be based on physical considerations, as in Maxwell's case, or it can be a so-called *black box* model, which is constructed by feeding some input signals into the system, and measuring the corresponding output. A model with some pre-defined structure is then fitted to the data. The mathematical model is then used for solving the problem at hand. One widely used representation for mathematical models is the *state space representation*. It is also used in this thesis and it is introduced in the next section.

1.1 Linear state space approach

In the state space representation it is assumed that all the essential information on the state of the plant can be represented as a vector called the *state* of the system. The vector space where the state takes values is called the state space and it can be either finite or infinite dimensional. The state is assumed to have some kind of dynamics in discrete or continuous time. These dynamics equations can be linear or nonlinear. The results of this thesis are exclusively concerned with linear state space models whose dynamics are formally governed by differential equations of the form

$$\begin{cases} \frac{d}{dt}x(t) = Ax(t) + Bu(t), & x(0) = x_0, \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{1.1}$$

or, in the discrete time setting, by difference equations (2.10), see Section 2.2 below. The state of the system is x, and u and y are the input and the output,

respectively. It is assumed that $x \in \mathcal{X}$, $u \in \mathcal{U}$, and $y \in \mathcal{Y}$ where $\mathcal{X}$ is the *state space*, $\mathcal{U}$ the *input space*, and $\mathcal{Y}$ the *output space* and they are all assumed to be separable Hilbert spaces. Thus the linear system can be represented as a block operator and the corresponding spaces,

$$S := \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right] : \mathcal{X} \times \mathcal{U} \to \mathcal{X} \times \mathcal{Y}. \tag{1.2}$$

The operator A is called the *main operator*, B is the *input* or *control operator*, C the *output* or *observation operator*, and D the *feedthrough operator*.

In the case when $\mathcal{X}$ and $\mathcal{U}$ are finite dimensional, the solution to (1.1) is given by the matrix exponential and the so-called variation of parameters formula,

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s)\,ds, \tag{1.3}$$

assuming $u \in L^2(\mathbb{R}^+;\mathbb{R}^m)$. Even this regularity assumption can be relaxed if the integral is understood in a more general sense; for example if u is white noise, then (1.3) has to be replaced by a Wiener integral.

To give a hint of what kind of problems are addressed in classical systems theory, let us give a non-exhaustive list, together with some classical examples and both historical and state-of-the-art references.

- **Well-posedness:** In (1.3) we already provided the solution to (1.1) if our system is finite dimensional. It is also rather easy to check that this solution is differentiable one time more than the input signal u. However, when the system is infinite dimensional then establishing the solvability of the dynamics equations and the smoothness of solutions can be far from trivial. These kinds of problems are typically known as well-posedness problems. In particular, when the system dynamics are governed by partial differential equations with control action through time-dependent boundary conditions, the control operator B is unbounded. The well-posedness of these *boundary control systems* is the subject of publications I and IV and so a more thorough introduction is given in Section 2.1.

- **Stability and stabilization:** The stability of a steam engine controlled by a governor system was already the topic of Maxwell's paper [36]. The different stability concepts and related results for infinite dimensional systems are discussed in [40] by Pritchard and Zabczyk and [49, Chapter 8] by Staffans. To give some intuition, we note that a finite dimensional system is stable if the eigenvalues of the matrix A have negative real parts. If $u = 0$ then for

any x_0, the solution of (1.1) converges to zero. Also for any $u \in L^2(\mathbb{R}^+;\mathcal{U})$, the solution $x(t)$ remains bounded.

If the feedback controller in Fig. 1.1 is linear, then plugging $u(t) = Ky(t)$ to (1.1) gives $\frac{d}{dt}x(t) = (A + KC)x(t)$. It is possible that A has eigenvalues with positive real parts but $A + KC$ does not. Then K is a stabilizing feedback controller for the system. For example, the case $B = C^*$ in (1.1) is called colocated control/observation. Then the feedback $u = -\kappa y$ with $\kappa > 0$ leads to $\frac{d}{dt}x(t) = (A - \kappa C^*C)x(t)$ and further, $\frac{d}{dt}\left(\frac{1}{2}\|x(t)\|_{\mathcal{X}}^2\right) = \langle x(t), Ax(t)\rangle_{\mathcal{X}} - \kappa\|Cx(t)\|_{\mathcal{Y}}^2$. Clearly such feedback has a stabilizing effect on the system, see [10] by Curtain and Weiss.

- **Controllability and observability of systems:** A fundamental question related to a system is whether for any vectors $x_0 \in \mathcal{X}$ and $x_1 \in \mathcal{X}$ there exists a control signal u so that $x(T) = x_1$ for some T. This property is called *exact controllability at time* T. In particular, in infinite dimensions, it is a rather strong property, and other, weaker notions exist, see [52, Chapter 11].

  With linear systems, the dual concept of controllability is observability. The observability at time T can be defined so that any initial state can be distinguished from the corresponding output on time interval $[0, T]$ (if $u = 0$). However, in literature, the characterization

$$\int_0^T \|CT(t)x_0\|_{\mathcal{Y}}^2 \, dt \geq k_T \|x_0\|_{\mathcal{X}}^2$$

  is often taken as the definition of *exact observability at time* T. This is equivalent to the existence of a bounded operator $K \in \mathcal{L}(L^2([0,T];\mathcal{Y}),\mathcal{X})$, such that $x_0 = Ky$, see [52, Remark 6.1.5].

  In finite dimensions ($\dim(\mathcal{X}) = n$), the exact controllability is equivalent to the *Kalman rank condition*, that is, $\operatorname{rank}\big([B|AB|A^2B|...|A^{n-1}B]\big) = n$. The exact observability is equivalent to $\operatorname{rank}\left(\begin{bmatrix} C \\ \vdots \\ CA^{n-1} \end{bmatrix}\right) = n$.

  For recent results on controllability and observability, see for example [27] by Li et al. for results on systems governed by partial differential equations, and [57] by Weiss and Zhao for results on coupled systems.

- **Optimal control:** One typical control problem is how to choose the control signal u so that some cost functional is minimized. This field is so wide that we only mention the classical problem with quadratic cost function

$$J = \langle x(T), P_T x(T)\rangle_{\mathcal{X}} + \int_0^T \big(\langle x(t), Qx(t)\rangle_{\mathcal{X}} + \langle u(t), Ru(t)\rangle_{\mathcal{U}}\big)dt$$

where $P_T, Q \in \mathcal{L}(\mathcal{X})$, and $R \in \mathcal{L}(\mathcal{U})$ are positive and self-adjoint operators. It is well known that this is a dual problem to Kalman–Bucy filtering, discussed in Chapter 3. Under sufficient assumptions, the solution to this optimal control problem is given by the feedback $u(t) = K(t)y(t)$ where $K(t)$ corresponds to the Kalman gain in the dual problem (see Section 3.2).

An interesting problem type is optimal control of systems with stochastic inputs. For recent progress, see [14] by Duncan et al. studying the linear quadratic control problem with fractional Brownian motion input and [37] by Muradore and Picci studying control strategies that are robust under stochastic disturbances.

- **State estimation:** In state estimation problems, the task is to estimate the state variable $x(t)$ when we are given the output (possibly corrupted by noise). Often also the input u might be partially or wholly unknown to us, thereby making the state estimation more difficult.

  The case with input and output corrupted by additive white noise is somewhat classical. The solution minimizing the estimation error variance is given by the *Kalman filter*, derived in 1960 in [23] by Kalman for discrete time systems and by Kalman and Bucy in 1961 in [24] in the continuous time setting. The timing of the results was perfect — the *space race* was booming and the method's potential in spaceflight trajectory estimation was quickly discovered. Even today, the Kalman filter is advocated for this renowned application, see [17] by Grewal and Andrews for the whole story. The infinite dimensional Kalman filter is the subject of publications II and III and so it will be presented in more detail in Chapter 3.

  Another well-known class of state estimation methods are the H^∞-techniques that are — loosely speaking — based on minimizing the "gain" from noise to estimation error. For an introduction, see [46, Chapter 11] by Simon, and for a recent study on infinite dimensional systems, see [8] by Chapelle et al.

  In the case the observations are not corrupted by noise, the state estimators are typically called observers. Perhaps the best-known class of observers are the Luenberger observers, see [28], that are based on updating the state estimate $\hat{x}(t)$ proportionally to the measurement discrepancy $y(t) - C\hat{x}(t)$. For recent development, see [42] by Ramdani et al. studying observers when the output operator C is not necessarily bounded, and [19] by Haine discussing observers in case the system is not exactly observable.

1.2 On the thesis

The results in the articles of this thesis can be divided into two categories — the well-posedness and passivity of boundary control systems is studied in publications I and IV and the effect of temporal and spatial discretization to Kalman filtering is studied in publications II and III, respectively. In publication I, it is shown that a composition of passive boundary control systems through Kirchhoff couplings is also a passive boundary control system. The results of publication IV essentially say that adding either boundary or state dissipation to a boundary control system preserves the system's well-posedness. Publication II treats the discrete time Kalman filter state estimate's convergence to the continuous time estimate as the temporal discretization is refined. Spatial discretization error in Kalman filtering is the subject of publication III. An optimal one step reduced-order state estimate is derived together with a bound for the discretization error. The main results of the publications are further discussed in Chapter 4.

Basic background on infinite dimensional linear systems is presented in Chapter 2, first for continuous time setting in Section 2.1 and then shortly for discrete time setting in Section 2.2. The background for the treatment of boundary control systems is given in Sections 2.1.2 and 2.1.3 — emphasis being on the well-posedness of systems. The Kalman filter is presented in Chapter 3. In Section 3.2, we derive the Kalman filter equations when the state space $\mathcal{X}$ is infinite dimensional but the output space $\mathcal{Y}$ is finite dimensional. The required background on Gaussian random variables is given in Section 3.1.

The reader is assumed to have knowledge on elementary functional analysis and stochastics (including treatment of random variables in Hilbert spaces). For introductory representations on these subjects, we refer to [25], and [53] or [11], respectively. For more comprehensive background on infinite dimensional linear systems, see [49].

Notation

Denote by $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ the space of bounded linear operators from normed space $\mathcal{H}_1$ to $\mathcal{H}_2$. Also denote $\mathcal{L}(\mathcal{H}) = \mathcal{L}(\mathcal{H}, \mathcal{H})$. The domain of an operator is denoted by $\mathcal{D}(\cdot)$, the null space by $\mathcal{N}(\cdot)$, and the range by $\mathcal{R}(\cdot)$. The resolvent set of A is denoted by $\rho(A)$ and the resolvent is $R(\lambda, A) = (\lambda - A)^{-1}$. The spectrum of an operator is denoted by $\sigma(\cdot)$ and the point spectrum by $\sigma_p(\cdot)$.

2. Infinite dimensional linear systems

The results of the thesis are all related to infinite dimensional linear systems, which are introduced in this chapter. The concept of well-posedness of systems will be discussed and the notion of semigroup will be introduced in Section 2.1.1. The results of publication II are more or less based on the semigroup approach and it is also needed in the further development of the system node concept and finally, boundary control systems. Publications I and IV treat well-posedness of boundary control systems and so emphasis will be given do the description of boundary control systems and well-posedness of infinite dimensional systems. In particular, the results of publications I and IV rely heavily on the results of [34] by Malinen and Staffans and so those results are reviewed in Section 2.1.3.

Finally, as discrete time systems are studied in publication III, some background will be given in Section 2.2. There we also go through some stability concepts that of course have their continuous time counterparts; but as they are not needed in the thesis, we only present the discrete time versions.

2.1 Continuous time systems

2.1.1 Semigroups and well-posedness

When the state space is infinite dimensional, the operator A in the formal equations (1.1) is often not bounded. Typically this is the case if the system dynamics are governed by partial differential equations when A is some kind of differential operator. Then, unlike in the finite dimensional setting, even the simple, homogeneous equation

$$\frac{d}{dt}x(t) = Ax(t), \qquad x(0) = x_0 \tag{2.1}$$

gives rise to numerous problems, starting from the unique existence and smoothness of the solution. Loosely speaking, these are known as well-posedness problems.

Firstly, a classical solution is defined as a function satisfying (2.1), such that $x \in C^1(\mathbb{R}^+; \mathcal{X})$ and $x(t) \in \mathcal{D}(A)$ for all $t \geq 0$. However, it is often desirable to formally study equation (2.1) when x_0 is not necessarily in $\mathcal{D}(A)$. To this end, we define a mild solution of (2.1) to be a function $x \in C(\mathbb{R}^+; \mathcal{X})$ satisfying

$$\int_0^t x(s)\, ds \in \mathcal{D}(A) \quad \text{and} \quad x(t) - x_0 = A \int_0^t x(s)\, ds$$

for all $t \geq 0$ where the integrals are Bochner integrals, see e.g., [1, Section 1.1].

The definition of well-posedness of a system varies depending on what we are interested in. Typically it is somehow related to the unique existence and smoothness of solutions. For the homogeneous time evolution problem, we adopt the following definition, due to [11, Section A.1]:

Definition 2.1.1. *The time evolution problem* (2.1)*, also known as* Cauchy problem*, is said to be* well-posed *if:*

(i) for any $x_0 \in \mathcal{D}(A)$, there exists a unique strongly differentiable (in $\mathcal{X}$) function $x(t, x_0)$ satisfying (2.1) *for all $t \geq 0$;*

(ii) for $\{x_n\} \subset \mathcal{D}(A)$ with $x_n \to 0$ strongly in $\mathcal{X}$ it holds that $x(t, x_n) \to 0$ strongly in $\mathcal{X}$ for all $t \geq 0$.

This definition gives rise to the notion of the semigroup generated by the operator A.

Definition 2.1.2. *If the problem* (2.1) *is well-posed, define the* semigroup generated by A *as the operator-valued function $T(t)$, that satisfies*

$$T(t)x_0 := x(t, x_0), \qquad t \geq 0$$

for $x_0 \in \mathcal{D}(A)$ where $x(t, x_0)$ is defined in part (i) of Definition 2.1.1.

The fact that $T(t)$ actually defines a linear operator in $\mathcal{D}(A)$ is easy to see by the linearity of differentiation. The semigroup $T(t)$ was defined in $\mathcal{D}(A)$ but by property *(ii)* in Definition 2.1.1, it can be uniquely extended to a bounded linear operator in the whole space $\mathcal{X}$. Henceforth $T(t)$ stands for this extension. This operator-valued function has the following well-known

properties:

- $T(0) = I$, (2.2)
- $T(t+s) = T(t)T(s)$ for $t, s \geq 0$, (2.3)
- for any $x \in \mathcal{X}$, the function $T(t)x$ is strongly continuous in $\mathcal{X}$. (2.4)

Note that strong differentiability in $\mathcal{X}$ holds only for $x \in \mathcal{D}(A)$.

Here we started with the formal equation (2.1) and ended up with a definition of a semigroup. However, we could also define a C_0-semigroup as an $\mathcal{L}(\mathcal{X})$-valued function satisfying the three conditions (2.2)–(2.4). If we are given such a function then the infinitesimal generator of the semigroup can be defined as follows (see [11, (A.7)]):

Definition 2.1.3. *Let $T(t)$ be an operator-valued function satisfying conditions* (2.2)–(2.4). *Define the domain of the* infinitesimal generator *A of the semigroup $T(t)$ as*

$$\mathcal{D}(A) := \left\{ x \in \mathcal{X} \ : \ \lim_{h \to 0} \frac{T(h)x - x}{h} \text{ exists (in strong sense) in } \mathcal{X} \right\}$$

and in $\mathcal{D}(A)$ define A as the limit, that is,

$$Ax := \lim_{h \to 0} \frac{T(h)x - x}{h}.$$

The infinitesimal generator given by the above equation is an extension of the original A in (2.1) but here we don't make the distinction between them.

Thus the well-posedness of the problem (2.1), as defined in Definition 2.1.1, means that the time evolution operator A is the generator of a C_0-semigroup. Perhaps the best-known characterization for C_0-semigroup generators is given by the Hille–Yosida theorem [1, Thm. 3.3.4]:

Theorem 2.1.1. Hille–Yosida. *Let A be a closed, densely defined operator on $\mathcal{X}$. Then it is the generator of a C_0-semigroup if and only if there exists $\omega \in \mathbb{R}$ and $M > 0$ such that*

$$\left\| (\lambda I - A)^{-n} \right\|_{\mathcal{L}(\mathcal{X})} \leq \frac{M}{(\lambda - \omega)^n} \qquad \text{for all } n \in \mathbb{N} \text{ and } \lambda > \omega.$$

A C_0-semigroup is called contractive if $\|T(t)\|_{\mathcal{L}(\mathcal{X})} \leq 1$ for all $t \geq 0$. Contractivity is related to the stability of the system and so it is a somewhat fundamental property. It is also a standing assumption in publication II that the system dynamics are governed by a contractive semigroup. A characterization for generators of contractive semigroups is given by the Lumer–Phillips

theorem, originally presented in [31] but it can also be found for example in [1, Thm. 3.4.5]:

Theorem 2.1.2. Lumer–Phillips. *Let A be a closed, densely defined operator on $\mathcal{X}$. Then it is the generator of a contractive C_0-semigroup iff*

(i) A is dissipative, meaning that for all $\lambda > 0$ and $x \in \mathcal{D}(A)$,

$$\|(\lambda I - A)x\|_{\mathcal{X}} \geq \lambda \|x\|_{\mathcal{X}}\,; \text{ and}$$

(ii) A is maximal in the sense that $\lambda_0 I - A$ is surjective for some $\lambda_0 > 0$.

One widely studied class of systems are such that the main operator A generates an analytic semigroup, that is, a semigroup that can be extended to a sector $t \in \{\lambda \in \mathbb{C} : |\arg(\lambda)| \leq \theta\}$ for some $\theta < \pi/2$ in such a way that conditions (2.2)–(2.4) hold in the whole sector. Analytic semigroups are also studied in Section 3.4 of publication II and so we give here their definition following [9, Definition 2.27], and present some of their properties.

Definition 2.1.4. *A C_0-semigroup $T(t)$ is* analytic *if*

(i) $T(t)$ can be continued analytically to a sector $\{\lambda \in \mathbb{C} : |\arg(\lambda)| \leq \theta\}$ for some $\theta < \pi/2$;

(ii) for all $t \in \{\lambda \in \mathbb{C} : |\arg(\lambda)| \leq \theta\}$, and $t \neq 0$, it holds that $AT(t) \in \mathcal{L}(\mathcal{X})$, and for any $x \in \mathcal{X}$,

$$\frac{d}{dt}T(t)x = AT(t)x;$$

(iii) $\|T(t)\|_{\mathcal{L}(\mathcal{X})}$ is uniformly bounded and $\|AT(t)\|_{\mathcal{L}(\mathcal{X})} \leq \frac{M}{|t|}$ for all $t \in \{\lambda \in \mathbb{C} : |\arg(\lambda)| \leq \theta\}$ for some $M > 0$.

Proposition 2.1.1. *Let A be the infinitesimal generator of an analytic semigroup $T(t)$. Then*

(i) the semigroup is given by $T(0) = I$ and

$$T(t) = \frac{1}{2\pi i}\int_{\gamma} e^{\lambda t}(\lambda - A)^{-1}d\lambda, \qquad t > 0$$

where $\gamma(\cdot)$ *is the path defined by parametrization* $\gamma(s) = \begin{cases} -se^{-i\theta} & \text{for } s < 0, \\ se^{i\theta} & \text{for } s \geq 0 \end{cases}$ *where* $\theta \in (\pi/2, \theta_0)$.

(ii) for any $t > 0$ *and* $x \in \mathcal{X}$, $T(t)x \in \mathcal{D}(A^k)$ *for all* $k \in \mathbb{N}$, *and for each* k *there exists a constant* $c(k)$, *such that*

$$\left\| A^k T(t) \right\|_{\mathcal{L}(\mathcal{X})} \leq \frac{c(k)}{t^k}, \qquad \text{for } t > 0;$$

(iii) if, in addition, $-A$ *is sectorial (see [1, Section 3.8]) then the above bound holds also for non-integer* k, *if* A^k *is replaced by* $(-A)^k$.

For a proof of part *(i)*, see [1, (3.46)]. For parts *(ii)* and *(iii)*, see [51, Thms. 3.3.1 & 3.3.3].

Let us finish this section by discussing the full system (1.1) under the assumption that A is the generator of a C_0-semigroup $T(\cdot) : \mathbb{R}^+ \to \mathcal{L}(\mathcal{X})$. In the case $B \in \mathcal{L}(\mathcal{U}, \mathcal{X})$, nothing is really changed compared to the finite dimensional case, and the solution to (1.1) is given by (1.3) with e^{At} replaced by the general semigroup $T(t)$, see, for example [1, Chapter 3]. However, if for example the system under consideration is governed by a partial differential equation with control action inflicting through the boundary conditions, then the input operator B is not bounded. To be able to study such systems, we proceed to introduce a more general framework of system nodes.

2.1.2 Operator and system nodes

Above we worked with the system's state space $\mathcal{X}$ and the domain of the main operator, $\mathcal{D}(A)$. In this section we define the *rigged spaces* $\mathcal{X}_j$ for $j \in \mathbb{Z}$, following [49, Section 3.6] and present the system node realization following [49, Section 4.7]. Let us also mention [44] and [45] by Salamon and [56] by Weiss as historical references on realization theory on Hilbert spaces. For more references, see the discussion sections 3.15 and 4.11 of [49].

If A is closed — as is usually assumed — then also $\mathcal{D}(A)$ can be made a Hilbert space if it is equipped with the graph norm $\|x\|^2_{\mathcal{D}(A)} := \|x\|^2_{\mathcal{X}} + \|Ax\|^2_{\mathcal{X}}$ or, assuming the resolvent set $\rho(A)$ is nonempty, with norm $\|x\|_{\mathcal{D}(A)} = \|(\alpha - A)x\|_{\mathcal{X}}$ with some $\alpha \in \rho(A)$. Note that different selection of α gives an equivalent norm to $\mathcal{D}(A)$. Let us denote $\mathcal{X}_1 := \mathcal{D}(A)$ and use there the latter norm. Then $(\alpha - A)^{-1}$ maps $\mathcal{X}$ isometrically to $\mathcal{X}_1$. Following [35, Proposition 2.1], define also the space $\mathcal{X}_{-1}$ as the completion of $\mathcal{X}$ with respect

to the norm $\|x\|_{\mathcal{X}_{-1}} := \|(\alpha - A)^{-1}x\|_{\mathcal{X}}$. By iteration of this construction, we can define spaces $\mathcal{X}_j$ for any $j \in \mathbb{Z}$ with $\mathcal{X}_j \subset \mathcal{X}_k$ if $j \leq k$ with a dense inclusion. Also it is possible to uniquely extend (or restrict) A and the corresponding semigroup $T(t)$ to $A_j \in \mathcal{L}(\mathcal{X}_{j+1}, \mathcal{X}_j)$ and $T_j(t) \in \mathcal{L}(\mathcal{X}_j)$, respectively.

After these preparations, we are now ready to extend the block notion (1.2) to cases where the input and output operators are not necessarily bounded.

Definition 2.1.5. *Let $\mathcal{X}$, $\mathcal{U}$, and $\mathcal{Y}$ be Hilbert spaces. A block operator*

$$S = \left[\begin{smallmatrix} A\&B \\ C\&D \end{smallmatrix}\right] : \mathcal{X} \times \mathcal{U} \to \mathcal{X} \times \mathcal{Y}$$

is called an operator node *on $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ if it has the following structure:*

(i) A is a closed, densely defined operator on $\mathcal{X}$ with a nonempty resolvent set.

(ii) $B \in \mathcal{L}(\mathcal{U}, \mathcal{X}_{-1})$.

(iii) $\mathcal{D}(S) := \{\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in \mathcal{X} \times \mathcal{U} \ : \ A_{-1}x + Bu \in \mathcal{X}\}$ where A_{-1} is the extension of A as described above. $\mathcal{D}(S)$ is equipped with the graph norm

$$\|\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right]\|^2_{\mathcal{D}(S)} := \|A_{-1}x + Bu\|^2_{\mathcal{X}} + \|x\|^2_{\mathcal{X}} + \|u\|^2_{\mathcal{U}}\,.$$

(iv) $C\&D \in \mathcal{L}(\mathcal{D}(S), \mathcal{Y})$.

If, in addition, A generates a C_0-semigroup on $\mathcal{X}$, then S is called a system node.

If S is a system node on $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ then for each $x_0 \in \mathcal{X}$ and $u \in C^2(\mathbb{R}^+; \mathcal{U})$ with $\left[\begin{smallmatrix} x_0 \\ u(0) \end{smallmatrix}\right] \in \mathcal{D}(S)$ the formal equations (1.1) have a unique solution $x \in C^1(\mathbb{R}^+; \mathcal{X})$ such that $\left[\begin{smallmatrix} x \\ u \end{smallmatrix}\right] \in C(\mathbb{R}^+; \mathcal{D}(S))$. This result can be found for example in [33, Lemma 2.2] but for a proof they refer to [49, Lemma 4.7.8].

Many systems satisfy different types of conservation laws that can be utilized when determining the solvability of a given system. An important conservation law is energy preservation:

Definition 2.1.6. *A system node is* scattering passive *if for all x_0 and u satisfying the conditions in the paragraph above, and for all $t \geq 0$, the solutions of* (1.1) *satisfy*

$$\|x(t)\|^2_{\mathcal{X}} - \|x_0\|^2_{\mathcal{X}} \leq \|u\|^2_{L^2((0,t);\mathcal{U})} - \|y\|^2_{L^2((0,t);\mathcal{Y})}\,. \tag{2.5}$$

A system node is scattering energy preserving *if this holds as an equality.*

Many characterizations for energy preserving systems an be found in [35, Section 3]. It is clear from the definition that the semigroup corresponding to a scattering passive system node is contractive.

An alternative framework for the presented system node setting is provided by the so-called port-Hamiltonian systems, that has been a very active field of research during the last fifteen years. Port-Hamiltonian systems form a unified approach for treating linear and nonlinear, and finite and infinite dimensional systems (including boundary control systems). The key idea is to utilize the systems' inherent conservation laws and to break the system at hand into components representing (Hamiltonian) "energy storages" and power conserving interconnections (through ports) between these storages. For an introduction, see the doctoral theses [32] by Macchelli or [54] by Villegas.

2.1.3 Boundary control systems

Boundary control systems are typically systems whose dynamics are governed by partial differential equations and the control action to them is inflicted through time-dependent boundary conditions. In principle, the system node framework allows treatment of such systems but these systems do not naturally adopt the form (1.1). So let us introduce slightly different looking dynamics equations:

$$\begin{cases} \frac{d}{dt}z(t) = Lz(t), & t \geq 0, \\ Gz(t) = u(t), \\ y(t) = Kz(t). \end{cases} \tag{2.6}$$

That is, the dynamics are not entirely governed by the first equation but an additional requirement $Gz(t) = u(t)$ has to be imposed for unique solvability. In a typical example, the operator G is a trace operator, and so this additional requirement consists of the boundary conditions for a partial differential equation. In this formalism, the operator L is called the *interior operator*, G the *input boundary operator*, and K the *output boundary operator*. This theoretical framework originates from [15] by Fattorini and [44] by Salamon. Our presentation is close to that of Malinen and Staffans in [33] and [34]. Related to equations of the form (2.6), we make the following definition.

Definition 2.1.7. *A triple of linear mappings (G, L, K) on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ with the same domain $\mathcal{Z} \subset \mathcal{X}$ is called a* colligation. *A colligation is* strong *if L is closed with $\mathcal{D}(L) = \mathcal{Z}$, and G and K are continuous with respect to the graph norm of L on $\mathcal{Z}$. The space $\mathcal{Z}$ is called the* solution space.

A colligation is a boundary node *if it has the following structure:*

(i) The block operator $\left[\begin{smallmatrix} G \\ L \\ K \end{smallmatrix}\right] : \mathcal{X} \to \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ *is closed;*

(ii) G *is surjective and its null space* $\mathcal{N}(G)$ *is dense in* $\mathcal{X}$*;*

(iii) The operator $A := L|_{\mathcal{N}(G)}$ *has a nonempty resolvent set* $\rho(A)$*;*

The boundary node is internally well-posed *if in addition,* A *generates a* C_0*-semigroup.*

Theorems 2.3 and 2.4 of [33] imply that every boundary node induces an operator node that is "of boundary control type", meaning that $\mathcal{R}(B) \cap \mathcal{X} = \{0\}$ and vice versa — every operator node that is of boundary control type induces a boundary node. The boundary node is internally well-posed if and only if the corresponding operator node is a system node. When this is the case, the solutions to respective equations (1.1) and (2.6) coincide.

From Definition 2.1.7 it is evident that $\left[\begin{smallmatrix} G \\ \alpha - L \end{smallmatrix}\right]$ is surjective for $\alpha \in \rho(A)$. Now regard α as fixed. Then there exists a right inverse for G, such that $LG_{right}^{-1} = \alpha G_{right}^{-1}$. In fact, by the proof of [33, Thm. 2.3], this inverse is given by $G_{right}^{-1} = (\alpha - A_{-1})^{-1} B$. So the solution space can be decomposed into a direct sum

$$\mathcal{Z} = \mathcal{X}_1 \oplus G_{right}^{-1} \mathcal{U},$$

that is, into components $\mathcal{X}_1 = \mathcal{N}(G)$ and another part taking care of the boundary conditions. We also have a bijective mapping and its inverse between $\mathcal{Z}$ and its decomposition:

$$\begin{bmatrix} I - G_{right}^{-1} G \\ G \end{bmatrix} : \mathcal{Z} \to \mathcal{X}_1 \times \mathcal{U} \qquad \text{and} \qquad \begin{bmatrix} I & G_{right}^{-1} \end{bmatrix} : \mathcal{X}_1 \times \mathcal{U} \to \mathcal{Z}.$$

The Cauchy problem associated with the boundary control system (2.6) can now be taken from the space $\mathcal{Z}$ to the decomposed space $\mathcal{X}_1 \times \mathcal{U}$. It can be solved there and the obtained solution can be taken back to $\mathcal{Z}$. This method is not used in the thesis but here it is presented. The interior operator can be split according to this decomposition,

$$Lz = L\left(I - G_{right}^{-1} G\right) z + L G_{right}^{-1} G z = A\left(I - G_{right}^{-1} G\right) z + \alpha G_{right}^{-1} G z,$$

and following this splitting, we write the time derivative of the $\mathcal{X}_1$-component

in the space $\mathcal{X}$:

$$\begin{aligned}\frac{d}{dt}\left(I - G_{right}^{-1}G\right)z(t) &= \frac{d}{dt}z(t) - G_{right}^{-1}\frac{d}{dt}u(t) = Lz(t) - G_{right}^{-1}\dot{u}(t)\\ &= A\left(I - G_{right}^{-1}G\right)z(t) + G_{right}^{-1}\big(\alpha u(t) - \dot{u}(t)\big).\end{aligned}$$

Consider now the Cauchy problem in the decomposed space $\mathcal{X}\times\mathcal{U}$. Equations (2.6) can be formulated in the decomposed space

$$\begin{cases}\frac{d}{dt}\left[\begin{smallmatrix}x\\u\end{smallmatrix}\right](t) = \left[\begin{smallmatrix}A & \alpha G_{right}^{-1}\\ 0 & 0\end{smallmatrix}\right]\left[\begin{smallmatrix}x\\u\end{smallmatrix}\right](t) + \left[\begin{smallmatrix}-G_{right}^{-1}\\ I\end{smallmatrix}\right]\dot{u}(t)\\ \left[\begin{smallmatrix}x\\u\end{smallmatrix}\right](0) = \left[\begin{smallmatrix}I - G_{right}^{-1}G\\ G\end{smallmatrix}\right]z_0.\end{cases} \tag{2.7}$$

This formulation resembles (1.1). The new control operator $\left[\begin{smallmatrix}-G_{right}^{-1}\\ I\end{smallmatrix}\right]$ is bounded from $\mathcal{U}$ to $\mathcal{X}\times\mathcal{U}$ but that is obtained at the cost of one temporal derivative in the input signal u.

Theorem 2.1.3. *The operator* $\tilde{A} := \left[\begin{smallmatrix}A & \alpha G_{right}^{-1}\\ 0 & 0\end{smallmatrix}\right] : \mathcal{X}\times\mathcal{U} \to \mathcal{X}\times\mathcal{U}$ *with domain* $\mathcal{X}_1\times\mathcal{U}$ *generates a* C_0*-semigroup* $\tilde{T}(t)$ *on* $\mathcal{X}\times\mathcal{U}$.

Proof. We use the Hille-Yosida theorem 2.1.1. The resolvent of $\tilde{A}$ is $R(\lambda,\tilde{A}) = \left[\begin{smallmatrix}R(\lambda,A) & \frac{\alpha}{\lambda}R(\lambda,A)G_{right}^{-1}\\ 0 & \lambda^{-1}\end{smallmatrix}\right]$. For some $\omega > 0$ we have $\|(\lambda-\omega)^n R(\lambda,A)\|_{\mathcal{L}(\mathcal{X})} < M$ for all $\lambda > \omega$ and $n \in \mathbb{N}$ and we need to find a similar uniform bound for the resolvent of $\tilde{A}$. For that we have

$$(\lambda-\omega)^n R(\lambda,\tilde{A})^n = \begin{bmatrix}(\lambda-\omega)^n R(\lambda,A)^n & \frac{\alpha}{\lambda}\sum_{j=1}^{n}\left(\frac{\lambda-\omega}{\lambda}\right)^{n-j}(\lambda-\omega)^j R(\lambda,A)^j G_{right}^{-1}\\ 0 & \left(\frac{\lambda-\omega}{\lambda}\right)^n\end{bmatrix}.$$

The only nontrivial element is the one in the upper right corner and for that we have a uniform bound

$$\begin{aligned}&\left\|\frac{\alpha}{\lambda}\sum_{j=1}^{n}\left(\frac{\lambda-\omega}{\lambda}\right)^{n-j}(\lambda-\omega)^j R(\lambda,A)^j G_{right}^{-1}\right\|_{\mathcal{L}(\mathcal{U},\mathcal{X})}\\ \le\ &\frac{\alpha}{\lambda}M\left\|G_{right}^{-1}\right\|_{\mathcal{L}(\mathcal{U},\mathcal{X})}\sum_{j=1}^{n}\left(\frac{\lambda-\omega}{\lambda}\right)^{n-j} \le \frac{\alpha}{\omega}M\left\|G_{right}^{-1}\right\|_{\mathcal{L}(\mathcal{U},\mathcal{X})}.\end{aligned}$$

□

The semigroup generated by $\tilde{A}$ is given by $\tilde{T}(t) = \left[\begin{smallmatrix}T(t) & \alpha\int_0^t T(u)G_{right}^{-1}\,du\\ 0 & I\end{smallmatrix}\right]$ where the integral is a Bochner integral computed in $\mathcal{X}$ but with value in $\mathcal{X}_1$ and $T(t)$ is the semigroup generated by A. The solution to (2.6) is then given by $z(t) = T_a(t)z_0 + \int_0^t T_b(t-s)\dot{u}(s)\,ds$ where

$$T_a(t) = T(t)\left(I - G_{right}^{-1}G\right) + \left(\alpha\int_0^t T(u)\,du + I\right)G_{right}^{-1}G$$

and $$T_b(t-s) = (I - T(t-s))\, G_{right}^{-1} + \alpha \int_0^{t-s} T(u)\, du\, G_{right}^{-1}.$$

We remark that the definition of energy preservation/passivity (Def. 2.1.6) above did not have any references to the system operators and so the same definition is directly extended to internally well-posed boundary nodes. Related to energy preservation, let us also define conservativity following [33]:

Definition 2.1.8. *The* time-flow inverse *of a given colligation* $\Xi = (G, L, K)$ *on spaces* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ *with domain* $\mathcal{Z}$ *is given by* $(K, -L, G)$ *on* $(\mathcal{Y}, \mathcal{X}, \mathcal{U})$ *with the same domain* $\mathcal{Z}$.

The boundary node is scattering conservative *if both* Ξ *and its time-flow inverse are scattering energy preserving.*

The reason we have used the term "scattering" when talking about energy preservation is that the energy inequality (2.5) is not the only naturally arising alternative. So opposed to scattering type systems, let us introduce *impedance* type systems assuming $\mathcal{U}$ and $\mathcal{Y}$ are a dual pair. In the impedance formulation, if the system equations have a solution $z(t)$ then, instead of (2.5), the energy passivity is characterized by the inequality

$$\frac{d}{dt}\left(\frac{1}{2}\left\|z(t)\right\|_{\mathcal{X}}^2\right) \leq \langle y(t), u(t)\rangle_{(\mathcal{Y},\mathcal{U})}\,. \tag{2.8}$$

The expression $\frac{1}{2}\left\|z(t)\right\|_{\mathcal{X}}^2$ is interpreted as the energy stored in the system and the right hand side is the instantaneous power inflicted. For example, in electric circuits the input u might be some control voltage and the output y the corresponding current — the inflicted power is then their product (recall the well-known formula $P = UI$). Other examples are acoustics (see the example in Section 5 of article I), where the input and output variables in impedance form would be the pressure and flow, and in mechanical systems, the inflicted force and velocity.

In article [34], it is noted that impedance type systems are obtained from scattering type systems by the *external Cayley transform*. They also define impedance passivity (and conservativity) through the Cayley transform. However, a more straightforward definition often serves the purpose better, and so we adopt the following definition, due to [34, Theorem 3.4]:

Definition 2.1.9. *Let* $\Xi = (G, L, K)$ *be a colligation on Hilbert spaces* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$.

(i) Ξ *is* impedance passive *if the following conditions hold:*

(a) $\begin{bmatrix} \beta G + K \\ \alpha - L \end{bmatrix}$ *is surjective for some* $\alpha, \beta \in \mathbb{C}^+$;

(b) For all $z \in \mathcal{D}(\Xi)$ *we have the* Green–Lagrange inequality

$$\Re\langle z, Lz\rangle_{\mathcal{X}} \leq \Re\langle Kz, Gz\rangle_{(\mathcal{Y},\mathcal{U})}. \tag{2.9}$$

(ii) Impedance passive Ξ *is* impedance conservative *if* (2.9) *holds as an equality, and (a) holds also for some* $\alpha, \beta \in \mathbb{C}^-$.

Note that the concept of impedance passivity does not require internal well-posedness. If Ξ is internally well-posed, then (2.9) is equivalent to (2.8). It is evident by (2.8) that the semigroup of an impedance passive boundary control system is contractive. Impedance passivity and also the Green–Lagrange inequality alone can be used for confirming the internal well-posedness using the following results, due to [34, Theorems 4.3 and 4.7]:

Theorem 2.1.4. *Let* $\Xi = (G, L, K)$ *be a strong colligation on spaces* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ *with domain* $\mathcal{Z}$ *where* $\mathcal{U}$ *and* $\mathcal{Y}$ *are a dual pair.*

(i) Assume that (2.9) *holds for all* $z \in \mathcal{Z}$. *If* $\left[\begin{smallmatrix} G \\ \alpha - L \end{smallmatrix}\right]$ *is surjective for some* $\alpha \in \mathbb{C}$ *with* $\Re(\alpha) \geq 0$ *then* Ξ *is an internally well-posed, impedance passive boundary node.*

(ii) Assume Ξ *is impedance passive. Then it is internally well-posed if and only if* G *is surjective.*

2.2 Discrete time systems

The dynamics of a discrete time system are governed by difference equations

$$\begin{cases} x_k = Ax_{k-1} + Bu_k \\ y_k = Cx_k + Du_k. \end{cases} \tag{2.10}$$

In some sense the theory of infinite dimensional discrete time systems is not as rich as that of continuous time systems. The equations are always solvable and there are no problems caused by unbounded operators.

The solution to the state evolution equation (2.10) is given by

$$x_k = A^k x_0 + \sum_{j=0}^{k-1} A^j B u_{k-j}.$$

If the input signal u is extended so that $u_k = 0$ for $k \leq 0$, the second term can be written as $\sum_{j=0}^{\infty} A^j B u_{k-j}$ which motivates us to define the *input map* $\mathcal{B} : l^2(\mathbb{Z}^-; \mathcal{U}) \to \mathcal{X}$ by $\{u_k\}_{k\in\mathbb{Z}^-} \mapsto \sum_{j=0}^{\infty} A^j B u_{-j}$ where $\mathbb{Z}^- = \{0, -1, -2, ...\}$. Define then the relevant stability concepts.

Definition 2.2.1. *The discrete time system composed of the operator quadruple* $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$ *and the dynamics equation* (2.10) *is*

(i) exponentially stable *if for* $u_k = 0$ *for all* k*, it holds that* $\sum_{k=1}^{\infty} \|x_k\|_{\mathcal{X}}^2 < \infty$ *for any initial state* $x_0 \in \mathcal{X}$*;*

(ii) asymptotically stable *if for* $u_k = 0$ *for all* k*, it holds that* $\|x_k\|_{\mathcal{X}} \to 0$ *as* $k \to \infty$ *for any initial state* $x_0 \in \mathcal{X}$*;*

(iii) output stable *if for* $u_k = 0$ *for all* k*, it holds that* $y \in l^2(\mathcal{Y})$ *for any initial state* $x_0 \in \mathcal{X}$*;*

(iv) input stable *if its dual system, composed of* $\left[\begin{smallmatrix} A^* & C^* \\ B^* & D^* \end{smallmatrix}\right]$*, is output stable.*

Characterizations for different stability concepts can be found in Opmeer's doctoral thesis [38, Chapter 3]. The connection between different stability concepts and solvability of the Lyapunov equation

$$S = ASA^* + W \tag{2.11}$$

with bounded, self-adjoint load $W \in \mathcal{L}(\mathcal{X})$ was studied by Przyluski in his classic article [41]. Here we present some results on the stability concepts which are essential considering this thesis, while other results are presented just to give some insight on the subject.

Theorem 2.2.1. Exponential and asymptotical stability. *The following statements are equivalent:*

(i) The discrete time system (2.10) *is exponentially stable.*

(ii) The spectral radius of A *is smaller than one, that is,* $\sigma(A) \subset D_1$ *where* D_1 *denotes the open unit disc in the complex plane (recall that as* A *is bounded,* $\sigma(A)$ *is closed).*

(iii) The Lyapunov equation (2.11) *with load* $W = I$ *has a nonnegative, self-*

adjoint solution $S \in \mathcal{L}(\mathcal{X})$.

In addition, exponential stability implies asymptotical stability and input stability. Asymptotical stability implies $\sigma_p(A) \subset D_1$ *and* $\sigma(A) \subset \overline{D_1}$.

Theorem 2.2.2. Input stability. *The following statements are equivalent:*

(i) The discrete time system (2.10) *is input stable.*

(ii) The input map satisfies $\mathcal{B} \in \mathcal{L}(l^2(\mathbb{Z}^-;\mathcal{U}), \mathcal{X})$.

(iii) The Lyapunov equation (2.11) *with load* $W = BB^*$ *has a nonnegative, self-adjoint solution* $S \in \mathcal{L}(\mathcal{X})$.

In addition, input stability implies that $\sigma_p(A) \subset \overline{D(0,1)}$.

2.2.1 Discretizing continuous time systems

Sometimes the considered real-life system has continuous time dynamics but for technical reasons we can only observe the output and control the input with discrete time intervals. Then the system can be transformed to a discrete time model. Consider the solution (1.3) in the case discussed in the end of Section 2.1.1, that is, A is the generator of a C_0-semigroup $T(\cdot)$ and $B \in \mathcal{L}(\mathcal{U}, \mathcal{X})$. Denoting $x_k := x(k\Delta t)$, the solution can be written as

$$x_k = T(\Delta t)x_{k-1} + \int_{(k-1)\Delta t}^{k\Delta t} T(t-s)Bu(s)\,ds.$$

If we then assume that $u(s)$ is constant u_k on the interval $s \in [(k-1)\Delta t, k\Delta t)$ then the solution can be written in discrete time form

$$x_k = A_d x_{k-1} + B_d u_k$$

where $A_d = T(\Delta t)$ and $B_d = \int_0^{\Delta t} T(s)B\,ds$. In the general system node setting with $B \in \mathcal{L}(\mathcal{U}, \mathcal{X}_{-1})$ it was required that $u \in C^2(\mathbb{R}^+;\mathcal{U})$ for the classical solution to exist. Thus, the piecewise constant u is not smooth enough. However, the integrated semigroup operator $\int_0^{\Delta t} T(s)\,ds$ has a smoothing effect, that is, $\int_0^{\Delta t} T_j(s)\,ds \in \mathcal{L}(\mathcal{X}_j, \mathcal{X}_{j+1})$ where the subindex j refers to the rigged spaces

discussed in the beginning of Section 2.1.2. In fact, it holds that

$$\left\| \int_0^{\Delta t} T_j(s)\, ds \right\|_{\mathcal{L}(\mathcal{X}_j, \mathcal{X}_{j+1})} = \left\| (\alpha - A_j) \int_0^{\Delta t} T_j(s)\, ds \right\|_{\mathcal{L}(\mathcal{X}_j)}$$
$$\leq |\alpha| \Delta t \sup_{s \in [0, \Delta t]} \|T_j(s)\|_{\mathcal{L}(\mathcal{X}_j)} + \|T_j(\Delta t) - I\|_{\mathcal{L}(\mathcal{X}_j)}\,.$$

So even $B \in \mathcal{L}(\mathcal{U}, \mathcal{X}_{-1})$ yields a bounded discrete time input operator $B_d \in \mathcal{L}(\mathcal{U}, \mathcal{X})$ with this so-called "zero-order-hold" discretization. Note that care must be taken when choosing the output of the discretized system. If also the output is a boundary observation, then $C \in \mathcal{L}(\mathcal{X}_1, \mathcal{Y})$ and then Cx_k is not well defined. However, integrating the state $x(s)$ from $(k-1)\Delta t$ to $k\Delta t$ gives a vector in $\mathcal{X}_1$ and so the discrete output y_k can be defined as the average of $y(s)$ on this interval, that is,

$$C_d x_{k-1} + D_d u_k := \frac{C}{\Delta t} \int_0^{\Delta t} \left(T(u) x_{k-1} + \int_{(k-1)\Delta t}^{(k-1)\Delta t + u} T(t-s) B u_k\, ds \right) du + D u_k.$$

The discretization given above is accurate, given that the input actually is piecewise constant. However, actually computing $T(\Delta t)$ might be impossible and one typically needs to rely on approximative schemes. A widely used method for approximating the discrete operators is given by the Cayley transform where $A_d = (\sigma + A)(\sigma - A)^{-1}$ and $B_d = \sqrt{2\sigma}(\sigma - A_{-1})^{-1} B$ with $\sigma = 2/\Delta t$. This method is studied in [6] by Besseling and in [20] by Havu and Malinen from the point of view of mathematical systems theory.

3. Infinite dimensional Kalman filter

In this chapter we introduce the discrete time Kalman filter, originally derived in [23] in the finite dimensional setting. The infinite dimensional generalization can be found, for example in [21] by Horowitz and [18] by Hager and Horowitz. It is the subject of publications II and III. Even though we also define the continuous time state estimate in II, an explicit representation is not needed. The proofs there make use of the discrete time Kalman filter with non-constant output operator. For the sake of notational simplicity, we here only treat the case where the operators do not depend on time. The continuous time variant is known as the Kalman–Bucy filter which was originally derived in [24]. The infinite dimensional Kalman–Bucy filter is presented, for example, in [3] by Bensoussan and in [9, Chapter 6] by Curtain and Pritchard.

The Kalman filter was originally developed for discrete time systems with noisy input and output:

$$\begin{cases} x_k = Ax_{k-1} + Bu_k \\ y_k = Cx_k + w_k. \end{cases} \tag{3.1}$$

where the input u_k and the output noise w_k are Gaussian random variables with values in $\mathcal{U}$ and $\mathcal{Y}$, respectively. They are assumed to have mean zero and covariance operators Q and R, respectively. Also the initial state is an $\mathcal{X}$-valued Gaussian random variable, $x_0 \sim N(m, P_0)$. It is assumed that u, w, and x_0 are mutually independent, and also w_k and u_k are independent of w_j and u_j, respectively, when $k \neq j$.

In this chapter, we first introduce Gaussian random variables in Section 3.1. In Section 3.2, we derive the Kalman filter equations assuming that the state space $\mathcal{X}$ is a separable Hilbert space and the output space $\mathcal{Y}$ is finite dimensional. Finally, in Section 3.3, we present some results on the Kalman filter and the corresponding Riccati equations that are needed in publication III.

3.1 Gaussian random variables

Definition 3.1.1. *A random variable v taking values in the Hilbert space $\mathcal{X}$ is said to be* Gaussian *if $\langle v, h\rangle_{\mathcal{X}}$ is normally distributed for all $h \in \mathcal{X}$.*

Gaussian random variables are extensively used when modeling uncertainty and external noise in dynamical systems — partly because they truly are somewhat fundamental (recall the central limit theorem), but also because they have so many nice properties making them easy to work with.

Proposition 3.1.1. *Let $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right]$ be a Gaussian random variable in $\mathcal{H}_x \times \mathcal{H}_z$ where $\mathcal{H}_x$ and $\mathcal{H}_z$ are separable Hilbert spaces. Then the following assertions hold:*

(i) **Fernique theorem.** *There exists $\lambda > 0$ such that $\mathbb{E}\Big(e^{\lambda \|x\|^2_{\mathcal{H}_x}}\Big) < \infty$. As a corollary, note that $\mathbb{E}\big(\|x\|^n_{\mathcal{H}_x}\big) < \infty$ for all $n \geq 1$.*

(ii) **Mean and covariance.** *There exists a vector $\left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right] \in \mathcal{H}_x \times \mathcal{H}_z$ and a symmetric, nonnegative trace class operator $P = \left[\begin{smallmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{smallmatrix}\right]$ such that*

$$\mathbb{E}\Big(\Big\langle \left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right], \begin{bmatrix} h_x \\ h_z \end{bmatrix} \Big\rangle\Big) = \Big\langle \left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right], \begin{bmatrix} h_x \\ h_z \end{bmatrix} \Big\rangle$$

for all $\begin{bmatrix} h_x \\ h_z \end{bmatrix} \in \mathcal{H}_x \times \mathcal{H}_z$ and

$$\begin{aligned} &\mathbb{E}\Big(\Big\langle \begin{bmatrix} h_{x1} \\ h_{z1} \end{bmatrix}, \left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right] \Big\rangle \Big\langle \begin{bmatrix} h_{x2} \\ h_{z2} \end{bmatrix}, \left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right] \Big\rangle\Big) - \Big\langle \left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right], \begin{bmatrix} h_{x1} \\ h_{z1} \end{bmatrix} \Big\rangle \Big\langle \left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right], \begin{bmatrix} h_{x2} \\ h_{z2} \end{bmatrix} \Big\rangle \\ &= \Big\langle \left[\begin{smallmatrix} P_{xx} & P_{xz} \\ P_{zx} & P_{zz} \end{smallmatrix}\right] \begin{bmatrix} h_{x1} \\ h_{z1} \end{bmatrix}, \begin{bmatrix} h_{x2} \\ h_{z2} \end{bmatrix} \Big\rangle \end{aligned}$$

for all $\begin{bmatrix} h_{x1} \\ h_{z1} \end{bmatrix}, \begin{bmatrix} h_{x2} \\ h_{z2} \end{bmatrix} \in \mathcal{H}_x \times \mathcal{H}_z$. Here $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{H}_x \times \mathcal{H}_z}$.

It holds that $\mathbb{E}\Big(\|x - m_x\|^2_{\mathcal{H}_x}\Big) = \operatorname{tr}(P_{xx})$. Also, the properties of a Gaussian random variable are completely comprised in its mean and covariance. Thus, it is meaningful to write $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right] \sim N\left(\left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right], P\right)$ meaning that $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right]$ is a Gaussian random variable with mean $\left[\begin{smallmatrix} m_x \\ m_z \end{smallmatrix}\right]$ and covariance P.

(iii) **Independence.** *x and z are independent if and only if $P_{xz} = 0$. Also if $\tilde{x}$ and $\tilde{z}$ are independent Gaussian random variables then $\left[\begin{smallmatrix} \tilde{x} \\ \tilde{z} \end{smallmatrix}\right]$ is a Gaussian random variable.*

(iv) **Conditional expectation.** *Assume $\dim(\mathcal{H}_z) < \infty$. The conditional expectation of x, given z, is given by*

$$\mathbb{E}(x|z) = m_x + P_{xz}P_{zz}^{-1}(z - m_z). \tag{3.2}$$

If P_{zz} is not invertible then P_{zz}^{-1} is replaced by pseudoinverse. The error covariance is

$$\operatorname{Cov}\left[x - \mathbb{E}(x|z)\,, x - \mathbb{E}(x|z)\right] = P_{xx} - P_{xz}P_{zz}^{-1}P_{zx}. \tag{3.3}$$

The conditional expectation minimizes $\mathbb{E}\Big(\|x - m_x - K(z - m_z))\|_{\mathcal{H}_x}^2\Big)$ over $K \in \mathcal{L}(\mathcal{H}_z, \mathcal{H}_x)$.

(v) **Linear combinations.** *If $A \in \mathcal{L}(\mathcal{H}_x, \mathcal{H})$ and $B \in \mathcal{L}(\mathcal{H}_z, \mathcal{H})$ then $Ax + Bz \sim N(Am_x + Bm_z, AP_{xx}A^* + AP_{xz}B^* + BP_{zx}A^* + BP_{zz}B^*)$.*

(vi) **Estimation.** *The best linear estimate is the best global estimate, that is, (3.2) minimizes $\mathbb{E}\Big(\|x - f(z)\|_{\mathcal{H}_x}^2\Big)$ over all measurable functions $f : \mathcal{H}_z \to \mathcal{H}_x$.*

For proofs, for part *(i)*, see [11, Theorem 2.6] (also a more general formulation is presented there). For part *(ii)*, see Lemma 2.14 and Proposition 2.15 in [11] and the discussion related to those results. Part *(iii)* follows by studying the characteristic function of $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right]$. A proof for the first claim can be found in [53, Proposition 4.10]. The second claim follows by writing the characteristic function for $\left[\begin{smallmatrix} \tilde{x} \\ \tilde{z} \end{smallmatrix}\right]$ and by independence noting that it corresponds to the characteristic function of a Gaussian random variable with mean $\left[\begin{smallmatrix} \mathbb{E}(\tilde{x}) \\ \mathbb{E}(\tilde{z}) \end{smallmatrix}\right]$ and covariance $\left[\begin{smallmatrix} \operatorname{Cov}[\tilde{x},\tilde{x}] & 0 \\ 0 & \operatorname{Cov}[\tilde{z},\tilde{z}] \end{smallmatrix}\right]$. Part *(v)* is easy to see directly from part *(ii)*, Definition 3.1.1, and properties of Bochner integral ($\mathbb{E}(\cdot)$ can be defined as a Bochner integral in the probability space, see [11, Section 1.1]). Part *(vi)* is proved in [9, Lemma 5.13]. Note that the condition (5.12) there is equivalent to $\mathcal{N}(P_{zz}) \subset \mathcal{N}(P_{xz})$ which is easy to confirm if $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right]$ is Gaussian.

A simple proof for *(iv)* (in the desired case when $\mathcal{H}_x$ is not necessarily finite dimensional) seems to be hard to find in the literature, so let us present steps leading to the proof. Firstly, $\mathbb{E}(x|z)$ is the unique element that is measurable with respect to the sigma algebra generated by z, for which $x - \mathbb{E}(x|z)$ and z are independent. Clearly $m_x + P_{xz}P_{zz}^{-1}(z - m_z)$ is measurable with respect to the sigma algebra generated by z. Now $\left[\begin{smallmatrix} x - (m_x + P_{xz}P_{zz}^{-1}(z - m_z)) \\ z \end{smallmatrix}\right]$ is also Gaussian so that independence of z and $m_x + P_{xz}P_{zz}^{-1}(z - m_z)$ can be verified by *(iii)*:

$$\operatorname{Cov}\left[x - (m_x + P_{xz}P_{zz}^{-1}(z - m_z)), z\right] = \operatorname{Cov}\left[x, z\right] - \operatorname{Cov}\left[P_{xz}P_{zz}^{-1}z, z\right] = 0.$$

In case P_{zz} is not invertible and pseudoinverse is used, the last term above becomes $P_{xz}P_{zz}^{+}P_{zz}$ where $P_{zz}^{+}P_{zz}$ is an orthogonal projection to the range

of P_{zz}. As $\mathcal{R}(P_{zz})^{\perp} \subset \mathcal{N}(P_{xz})$, the above covariance is still zero.

The minimization property in *(iv)* can be checked by directly solving the minimization problem which leads to expression (3.2), as in the proof of [9, Lemma 5.12].

It is noteworthy that by (3.2), $\operatorname{Cov}[\mathbb{E}(x|z), \mathbb{E}(x|z)] = P_{xz}P_{zz}^{-1}P_{zx}$ so that from (3.3) we see that

$$\operatorname{Cov}[x - \mathbb{E}(x|z), x - \mathbb{E}(x|z)] = \operatorname{Cov}[x, x] - \operatorname{Cov}[\mathbb{E}(x|z), \mathbb{E}(x|z)].$$

By computing the trace of both sides, we get the sort of Pythagorean identity

$$\mathbb{E}\Big(\|x - m_x\|_{\mathcal{X}}^2\Big) = \mathbb{E}\Big(\|\mathbb{E}(x|z) - m_x\|_{\mathcal{X}}^2\Big) + \mathbb{E}\Big(\|x - \mathbb{E}(x|z)\|_{\mathcal{X}}^2\Big).$$

Also it holds that $\operatorname{Cov}[x, x] \geq \operatorname{Cov}[\mathbb{E}(x|z), \mathbb{E}(x|z)]$ meaning that $\operatorname{Cov}[x, x] - \operatorname{Cov}[\mathbb{E}(x|z), \mathbb{E}(x|z)]$ is positive (semi)definite. These simple facts are used in publication III.

From linearity of the dynamics equations (3.1) and parts *(iii)* and *(v)* of Proposition 3.1.1, it follows that the state x_k is an $\mathcal{X}$-valued Gaussian random variable for all $k \geq 0$. The mean is $\mathbb{E}(x_k) = A^k m$ and covariance $\operatorname{Cov}[x_k, x_k] =: S_k$ is given by the recursive equation

$$S_k = AS_{k-1}A^* + BQB^*, \qquad S_0 = P_0. \tag{3.4}$$

Further, $[x_0, ..., x_k, y_1, ..., y_k]$ is a Gaussian random variable in $\mathcal{X}^{k+1} \times \mathcal{Y}^k$ for all $k \geq 0$. Let us conclude the section with the following result.

Theorem 3.1.1. *Let x_k be given by* (3.1) *with $P_0 = 0$ and assume that the system is input stable. Then the covariance $S_k = \operatorname{Cov}[x_k, x_k]$ given by* (3.4) *converges strongly to $S \in \mathcal{L}(\mathcal{X})$ which is the solution of the Lyapunov equation $S = ASA^* + BQB^*$. If, in addition, the system is asymptotically stable, then S_k converges strongly to S starting from any symmetric $S_0 = P_0$.*

Note that the limit S is not a trace class operator in general. If the system is even exponentially stable then the limit is a trace class operator and the convergence is in operator norm.

Proof. Recall that input stability is equivalent to $\hat{S} = A\hat{S}A^* + BB^*$ having a nonnegative solution. Consider first the case $S_0 = 0$. Clearly the solution to the covariance equation (3.4) is $S_k = \sum_{j=0}^{k-1} A^j BQB^*(A^*)^j$, from which it is easy to see that $S_{k+1} \geq S_k$. Assuming $S_{k-1} \leq \|Q\|_{\mathcal{L}(\mathcal{U})}\, \hat{S}$ for some k, then

$$S_k = AS_{k-1}A^* + BQB^* \leq \|Q\|_{\mathcal{L}(\mathcal{U})}\, A\hat{S}A^* + \|Q\|_{\mathcal{L}(\mathcal{U})}\, BB^* = \|Q\|_{\mathcal{L}(\mathcal{U})}\, \hat{S}.$$

So S_k is increasing and uniformly bounded implying strong convergence to some operator by [43, p. 249]. Letting $k \to \infty$ in (3.4) yields that the limit is S.

Then assume asymptotical stability and consider $S_0 \neq 0$. Then $S_k = A^k S_0 (A^*)^k + \sum_{j=0}^{k-1} A^j B Q B^* (A^*)^j$. Asymptotical stability means that $A^k \to 0$ strongly and so also $A^k S_0 (A^*)^k \to 0$ strongly. □

3.2 Kalman filter derivation

Assume now that $\dim(\mathcal{Y}) < \infty$. Define $Y_k := [y_1, ..., y_k]^T$ and consider $\mathbb{E}(x_k|Y_k)$. Also $[x_k, Y_k]$ is a Gaussian random variable in $\mathcal{X} \times \mathcal{Y}^k$ and so the conditional expectation is given by (3.2):

$$\hat{x}_k := \mathbb{E}(x_k|Y_k) = \mathbb{E}(x_k) + \mathrm{Cov}\,[x_k, Y_k]\,\mathrm{Cov}\,[Y_k, Y_k]^{-1}\,(Y_k - \mathbb{E}(Y_k)). \tag{3.5}$$

Note that $\mathrm{Cov}\,[Y_k, Y_k]$ is invertible because it is the sum of a positive definite block diagonal matrix (with R:s on the diagonal), and a positive semidefinite matrix.

Now decompose $Y_k = \left[\begin{smallmatrix} Y_{k-1} \\ y_k \end{smallmatrix}\right]$ in (3.5), and write the covariances in corresponding block form. Firstly,

$$\begin{aligned} &\mathrm{Cov}\,[x_k, Y_k] = \mathrm{Cov}\left[Ax_{k-1} + Bu_k, \left[\begin{smallmatrix} Y_{k-1} \\ y_k \end{smallmatrix}\right]\right] \\ &= A\mathrm{Cov}\left[x_{k-1}, \left[\begin{smallmatrix} Y_{k-1} \\ 0 \end{smallmatrix}\right]\right] + A\mathrm{Cov}\left[x_{k-1}, \left[\begin{smallmatrix} 0 \\ CAx_{k-1} \end{smallmatrix}\right]\right] + B\mathrm{Cov}\left[u_k, \left[\begin{smallmatrix} 0 \\ CBu_k \end{smallmatrix}\right]\right] \end{aligned} \tag{3.6}$$

where in the second equality we have used $y_k = CAx_{k-1} + CBu_k + w_k$ and the independence of u_k, w_k, and x_{k-1}. Then recall the block matrix inversion formula for symmetric matrices

$$\begin{bmatrix} F & G \\ G^T & H \end{bmatrix}^{-1} = \begin{bmatrix} F^{-1} + F^{-1}G(H - G^T F^{-1} G)^{-1} G^T F^{-1} & -F^{-1}G(H - G^T F^{-1} G)^{-1} \\ -(H - G^T F^{-1} G)^{-1} G^T F^{-1} & (H - G^T F^{-1} G)^{-1} \end{bmatrix}$$

and apply that to

$$\mathrm{Cov}\,[Y_k, Y_k] = \begin{bmatrix} \mathrm{Cov}\,[Y_{k-1}, Y_{k-1}] & \mathrm{Cov}\,[Y_{k-1}, y_k] \\ \mathrm{Cov}\,[y_k, Y_{k-1}] & \mathrm{Cov}\,[y_k, y_k] \end{bmatrix}.$$

Then we collect terms of (3.5). First, from (3.6) the term multiplying the first row of $\mathrm{Cov}\,[Y_k, Y_k]^{-1}$ is $A\mathrm{Cov}\,[x_{k-1}, Y_{k-1}]$. By picking only the term $F^{-1} = \mathrm{Cov}\,[Y_{k-1}, Y_{k-1}]^{-1}$ from the upper left corner of the inverse formula,

and $\mathbb{E}(x_k) = A\mathbb{E}(x_{k-1}) + B\mathbb{E}(u_k) = A\mathbb{E}(x_{k-1})$ from (3.5), we get

$$\mathbb{E}(x_k) + A\mathrm{Cov}\left[x_{k-1}, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\left(Y_{k-1} - \mathbb{E}(Y_{k-1})\right) = A\hat{x}_{k-1}.$$

Then observe that the remaining terms in the inverse formula can be factorized so that (3.5) becomes

$$\begin{aligned}
&\hat{x}_k - A\hat{x}_{k-1} \\
&= \mathrm{Cov}\left[x_k, Y_k\right]\left[\begin{smallmatrix} -F^{-1}G \\ I \end{smallmatrix}\right](H - G^TF^{-1}G)^{-1}\left[-G^TF^{-1}\ I\right]\left[\begin{smallmatrix} Y_{k-1}-\mathbb{E}(Y_{k-1}) \\ y_k-\mathbb{E}(y_k) \end{smallmatrix}\right]. \quad (3.7)
\end{aligned}$$

Now $-G^TF^{-1} = -\mathrm{Cov}\left[y_k, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}$ so the last product is

$$\begin{aligned}
&\left[-G^TF^{-1}\ I\right]\left[\begin{smallmatrix} Y_{k-1}-\mathbb{E}(Y_{k-1}) \\ y_k-\mathbb{E}(y_k) \end{smallmatrix}\right] \\
&= -\mathrm{Cov}\left[y_k, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\left(Y_{k-1} - \mathbb{E}(Y_{k-1})\right) + y_k - \mathbb{E}(y_k) \\
&= y_k - \mathbb{E}(y_k|Y_{k-1})
\end{aligned}$$

where the second equality holds by (3.2). Now it holds that $\mathbb{E}(y_k|Y_{k-1}) = \mathbb{E}(CAx_{k-1} + CBu_k + w_k|Y_{k-1}) = CA\hat{x}_{k-1}$ because u_k and w_k are independent of Y_{k-1}, and $\mathbb{E}(u_k) = \mathbb{E}(w_k) = 0$. Further, the inverse in (3.7) is

$$\begin{aligned}
&H - G^TF^{-1}G \\
&= \mathrm{Cov}\left[y_k, y_k\right] - \mathrm{Cov}\left[y_k, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\mathrm{Cov}\left[Y_{k-1}, y_k\right] \\
&= \mathrm{Cov}\left[y_k - \mathbb{E}(y_k|Y_{k-1}), y_k - \mathbb{E}(y_k|Y_{k-1})\right] \\
&= \mathrm{Cov}\left[CAx_{k-1} + CBu_k + w_k - CA\hat{x}_{k-1}, CAx_{k-1} + CBu_k + w_k - CA\hat{x}_{k-1}\right] \\
&= CA\mathrm{Cov}\left[x_{k-1} - \hat{x}_{k-1}, x_{k-1} - \hat{x}_{k-1}\right]A^*C^* + CBQB^*C^* + R
\end{aligned}$$

where the second equality holds by (3.3) and the last because u_k and w_k are independent of x_{k-1} and $\hat{x}_{k-1}$. Finally, using (3.6) for $\mathrm{Cov}\left[x_k, Y_{k-1}\right]$, and $-F^{-1}G = -\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\mathrm{Cov}\left[Y_{k-1}, y_k\right]$, the first product in (3.7) is

$$\begin{aligned}
&\mathrm{Cov}\left[x_k, Y_k\right]\left[\begin{smallmatrix} -F^{-1}G \\ I \end{smallmatrix}\right] \\
&= -A\mathrm{Cov}\left[x_{k-1}, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\mathrm{Cov}\left[Y_{k-1}, y_k\right] \\
&\qquad + A\mathrm{Cov}\left[x_{k-1}, CAx_{k-1}\right] + B\mathrm{Cov}\left[u_k, CBu_k\right] \\
&= A\Big(\mathrm{Cov}\left[x_{k-1}, x_{k-1}\right] - \mathrm{Cov}\left[x_{k-1}, Y_{k-1}\right]\mathrm{Cov}\left[Y_{k-1}, Y_{k-1}\right]^{-1}\mathrm{Cov}\left[Y_{k-1}, x_{k-1}\right]\Big)A^*C^* \\
&\qquad + BQB^*C^* \\
&= A\mathrm{Cov}\left[x_{k-1} - \hat{x}_{k-1}, x_{k-1} - \hat{x}_{k-1}\right]A^*C^* + BQB^*C^*.
\end{aligned}$$

In the second equality, $\mathrm{Cov}\left[Y_{k-1}, y_k\right]$ was treated as above, and the last equal-

ity follows from (3.3). Now we have all of the terms in (3.7) for computing $\hat{x}_k$. Performing the same decomposition and term gathering for the estimation error covariance $P_k := \mathrm{Cov}\,[x_k - \hat{x}_k, x_k - \hat{x}_k]$ given by (3.3) leads to the recursive Kalman filter equations that are typically written in the following form, known as *Riccati difference equations*,

$$\begin{cases} \tilde{P}_k = AP_{k-1}A^* + BQB^*, \\ P_k = \tilde{P}_k - \tilde{P}_k C^*(C\tilde{P}_k C^* + R)^{-1} C\tilde{P}_k \end{cases} \tag{3.8}$$

with P_0 being the initial state covariance, and

$$\hat{x}_k = A\hat{x}_{k-1} + K_k(y_k - CA\hat{x}_{k-1}) \tag{3.9}$$

where $K_k := \tilde{P}_k C^*(C\tilde{P}_k C^* + R)^{-1}$ is known as the *Kalman gain*.

3.3 Discussion and auxiliary results

One of the reasons why Kalman filter has been very popular in practical applications is its computational lightness. The error covariances given by (3.8) and the Kalman gains K_k do not depend on observations and thus they can be computed offline beforehand, leaving only (3.9) to be solved online.

It is easy to show that for any quadratically integrable random variable $\left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right] \in \mathcal{H}_x \times \mathcal{H}_z$, that is, $\mathbb{E}\Big(\|x\|^2_{\mathcal{H}_x} + \|z\|^2_{\mathcal{H}_z}\Big) < \infty$, the solution to the minimization problem

$$\min_{K \in \mathcal{L}(\mathcal{H}_z, \mathcal{H}_x)} \mathbb{E}\Big(\|x - m_x - K(z - m_z)\|^2_{\mathcal{H}_x}\Big) \tag{3.10}$$

is given by (3.2) and the error covariance by (3.3). Recall that our derivation of the Kalman filter was based solely on these equations. Thus the Kalman filter provides the optimal (in terms of error measure (3.10)) linear filter for systems of the form (3.1), even when the noise processes u and w and initial state x_0 are uncorrelated and quadratically integrable, but not necessarily Gaussian. Of course, better nonlinear filters might exist in this case.

Let us end the chapter by presenting some results on Kalman filter and the corresponding Riccati difference equations. Some of these results are used in publication III while others are just "nice-to-know".

Theorem 3.3.1. *Let P_k and $P_k^{(j)}$ for $j = 1, 2$ be the solutions of equations (3.8) with the load term BQB^* replaced by self-adjoint, positive trace class operators W and $W^{(j)}$, $j = 1, 2$, respectively. The following assertions hold.*

(i) If $W^{(2)} \geq W^{(1)}$ *and* $P_0^{(2)} \geq P_0^{(1)}$ *then* $P_k^{(2)} \geq P_k^{(1)}$ *for all* k.

(ii) If $P_k \geq P_{k-1}$ *for some* k *then* $P_{k+1} \geq P_k$.

(iii) If $2P_k \geq P_{k-1} + P_{k+1}$ *for some* k *then* $2P_{k+1} \geq P_k + P_{k+2}$.

The first assertion follows from [12, Lemma 3.1]. The proof is presented in the finite dimensional case, but it holds also for infinite dimensional systems assuming $\dim(\mathcal{Y}) < \infty$. It is also presented in Lemma 3.2 of III with a simple proof. The other two assertions are not needed in the thesis, but here they are given just to illuminate the properties of Riccati difference equations and the state estimation problem. Part *(ii)* follows directly from *(i)*. Part *(iii)* is proven in [12, Lemma 3.2].

Theorem 3.3.2. *Let* P_k *be the solution of* (3.8)*. The following assertions hold.*

(i) If the underlying system is input stable and $P_0 = 0$ *then* P_k *converges strongly to* P *as* $k \to \infty$ *where* P *is a solution of the* discrete algebraic Riccati equation *(DARE)*

$$\begin{cases} \tilde{P} = APA^* + BQB^*, \\ P = \tilde{P} - \tilde{P}C^*(C\tilde{P}C^* + R)^{-1}C\tilde{P}. \end{cases} \tag{3.11}$$

(ii) If the asymptotic filter is exponentially stable, that is, $r(A - KCA) < 1$ *where* $K = \tilde{P}C^*(C\tilde{P}C^* + R)^{-1}$ *then* P_k *converges to* P*, starting from any self-adjoint trace class operator* $P_0 \in \mathcal{L}(\mathcal{X})$*. Also,* P *is the unique nonnegative solution of* (3.11)*.*

The proofs of *(i)* and *(ii)* can be found in [18, Theorem 1] and [18, Theorem 3], respectively. The first proof is based on showing that P_k is an increasing sequence (see part *(ii)* of Theorem 3.3.1). It is also bounded by S which is the limit of (3.4), see Theorem 3.1.1. The proof of *(ii)* is rather similar. The sufficient stability assumption for part *(ii)* is actually *uniform asymptotical stability at large* which is implied by exponential stability. In publication III the exponential stability of the Kalman filter is needed elsewhere and therefore it is taken as an assumption here as well.

4. Summaries of the articles

I: Compositions of passive boundary control systems

Recall the formulation of impedance type boundary control systems in Section 2.1.3, and in particular, the energy inequality (2.8). Consider then an electric circuit. The well-known Kirchhoff laws say that in any vertex of the circuit, the voltage is the same for all leads connected in the vertex and the electrical currents must sum up to zero. These coupling conditions are natural also for the other mentioned example cases.

In publication I, the coupling conditions are formulated in terms of the input and output operators of the subsystems, whose dynamics are governed by colligations $(G^{(j)}, L^{(j)}, K^{(j)})$ on Hilbert spaces $(\mathcal{U}^{(j)}, \mathcal{X}^{(j)}, \mathcal{Y}^{(j)})$ where the index j refers to the subsystem. Assume that the input and output spaces can be split into two parts, that is, $\mathcal{U}^{(j)} = \mathcal{U}_1^{(j)} \oplus \mathcal{U}_2^{(j)}$ and $\mathcal{Y}^{(j)} = \mathcal{Y}_1^{(j)} \oplus \mathcal{Y}_2^{(j)}$, each representing a part of the boundary where the control action takes place — consider, for example, the two ends of a transmission line. Then, for example, the Kirchhoff coupling conditions for three systems ($j = 1, 2, 3$) coupled through the first parts of the input and output are

$$\begin{cases} G_1^{(1)} z_1(t) = G_1^{(2)} z_2(t) = G_1^{(3)} z_3(t), \\ K_1^{(1)} z_1(t) + K_1^{(2)} z_2(t) + K_1^{(3)} z_3(t) = 0, \end{cases} \tag{4.1}$$

assuming that the corresponding spaces are compatible, that is, $\mathcal{U}_1^{(1)} = \mathcal{U}_1^{(2)} = \mathcal{U}_1^{(3)}$ and $\mathcal{Y}_1^{(1)} = \mathcal{Y}_1^{(2)} = \mathcal{Y}_1^{(3)}$. This is a slightly simplified example. In publication I, the spaces $\mathcal{U}^{(j)}$ and $\mathcal{Y}^{(j)}$ can be split into more than two parts.

The main result of this article is that if internally well-posed, impedance passive (or conservative) boundary control systems (see Definitions 2.1.7 and 2.1.9) are interconnected through Kirchhoff type coupling conditions (4.1), then also

the resulting composed system (called *a transmission graph*, see Definitions 3.1 and 3.2 in I) is an internally well-posed, impedance passive (or conservative) boundary control system.

Compositions of port-Hamiltonian systems (see the end of Section 2.1.2) are studied in [7] by Cervera et al. and in [26] by Kurula et al. The presented formalism does not allow connecting finite dimensional subsystems to boundary control systems. To do that, one would need to work with the system node setting. This would require some further investigation. Such ideas can be found for example in [57] by Weiss and Zhao.

II: Convergence of discrete time Kalman filter estimate to continuous time estimate

In publication II we study systems of the form

$$\begin{cases} \frac{d}{dt}z(t) = Az(t), \\ z(0) = x \sim N(m, P_0), \\ y(t) = \int_0^t Cz(s)\,ds + w(t) \end{cases}$$

where w is Brownian motion with incremental covariance R. We define the discrete and continuous time state estimates as

$$\hat{x}_{T,n} := \mathbb{E}\Big(x \,\Big|\, \big\{y\left(\tfrac{iT}{n}\right)\big\}_{i=1}^{n}\Big) \quad \text{and} \quad \hat{x}(T) := \mathbb{E}\big(x \,\big|\, \{y(s), s \le T\}\big)$$

respectively. These estimates are given by the Kalman(–Bucy) filter — given that the continuous time Kalman–Bucy filter equations are solvable. By the Martingale convergence theorem, when the temporal discretization is refined then the discrete time estimate converges to the continuous time estimate. The purpose of publication II is to establish convergence speed estimates for $\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big)$ in various cases under different assumptions. First the result is established assuming $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and either $P_0 \in \mathcal{L}(\mathcal{X}, \mathcal{D}(A))$ or $x \in \mathcal{D}(A)$ almost surely. The latter covers the case $\dim(\mathcal{X}) < \infty$. Then the case $C \in \mathcal{L}(\mathcal{D}(A), \mathcal{Y})$ is treated assuming $x \in \mathcal{D}(A)$ almost surely and that A is diagonalizable and its point spectrum satisfies the asymptotic condition *(ii)* in Theorem 3.5 and C satisfies the regularity assumption *(iii)* in Theorem 3.5, or that the system is scattering passive. Then an estimate is shown when A generates an analytic semigroup. The proofs are based on applying the discrete time Kalman filter starting from $\hat{x}_{T,n}$ and taking into account more

and more measurements from a dense, numerable set in $[0, T]$.

To the author's knowledge, such results have not been published before. The articles [2] by Axelsson and Gustafsson and [55] by Wahlström et al. study the effect of using different numerical schemes for approximating the matrix exponential $e^{A\Delta t}$ on the solution of the Lyapunov equation and the Kalman filtering problem. Further effort would be required to obtain similar convergence results when for example the Cayley transformation (introduced in Section 2.2.1) would be used for obtaining the discretized system.

III: Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems

Publication III deals with state estimation problem for infinite dimensional discrete time systems. A practical implementation of the Kalman filter cannot be done in infinite dimensions. The system dynamics can be approximated by projecting equations (3.1) by an orthogonal projection $\Pi_s : \mathcal{X} \to \mathcal{X}$. The finite dimensional subspace $\Pi_s\mathcal{X}$ can be for example a finite element space (see the example in Section 5 of III) or a truncated eigenspace, see [48]. If the Kalman filter is directly implemented to the discretized system, the result is biased and hence not optimal. In Section 2 of III, an optimal one-step state estimate is derived that takes values in the finite dimensional subspace. One-step estimate means here that the k^{th} state estimate depends only on the previous estimate and the k^{th} measurement — recall the remarkable property of the Kalman filter, $\mathbb{E}(x_k|Y_k) = \mathbb{E}(x_k|\hat{x}_{k-1}, y_k)$. In Section 3, a Riccati difference equation is derived for the estimation error. The main results of the article are presented in Section 4, namely estimates for the discrepancy between the full state Kalman filter estimate $\hat{x}_k$ and the presented reduced-order estimate $\tilde{x}_k$. It is shown that if $\sup_k \mathbb{E}\Big(\|x_k\|^2_{\mathcal{X}_1}\Big) < \infty$, the system is input stable, and the full state Kalman filter is exponentially stable, then as $\|I - \Pi^*\Pi\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}$ becomes small, then

$$\limsup_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|^2_{\mathcal{X}}\Big) = \mathcal{O}\left(\|I - \Pi^*\Pi\|^2_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}\right)$$

where Q_k is a certain post-processing operator that is obtained when computing the Kalman gains for the reduced-order method. The proof is based on applying perturbation theory for algebraic Riccati equations, developed by Sun in [50], to the corresponding DAREs.

Another state estimator that takes the discretization error into account is developed by Pikkarainen in [39] and implemented numerically by Huttunen

and Pikkarainen in [22]. Their method is based on keeping track of the discretization error and then ignoring the correlation of the discretization error for different time steps in order to obtain a one step estimate. The direct implementation of the finite dimensional Kalman filter to the discretized system is studied by Bensoussan in [3, Chapter 9] and by Germani et al. in [16]. The latter includes a convergence result for the finite dimensional state estimate and the corresponding error covariance. In a very recent manuscript [13], Dihlmann and Haasdonk propose a reduced-basis Kalman filter for PDEs with possibly non-constant (in time) parameters.

Our approach is closely related to the reduced-order filtering methods. Let us mention articles [4] and [5] by Bernstein and Hyland and [47] by Simon because they had some influence on the results of this article — even though their results are not explicitly used.

IV: Acoustic wave guides as infinite-dimensional dynamical systems

This publication is a part of a trilogy containing also articles [29] and [30] by Lukkari and Malinen. The author's contribution is restricted to Section 3, titled "Conservative majorants", and so only that part is discussed here.

A passive boundary control system described by a colligation (G, L, K) on $\mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ with domain $\mathcal{Z}$ (see Definitions 2.1.8 and 2.1.9) can often be "split" into a sum of a conservative part and a dissipative perturbation (see (12) in the example in Section 5 of I). Alternatively, at some part of the boundary of an otherwise energy preserving system, there is a resistive boundary condition (see the second and third boundary conditions in (14) in I). These cases can be formulated as follows:

Definition. *Let* $\left(\begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, L, \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \right)$ *on Hilbert spaces* $(\mathcal{U}_1 \times \mathcal{U}_2, \mathcal{X}, \mathcal{Y}_1 \times \mathcal{Y}_2)$ *with domain* $\mathcal{Z}$ *be a scattering passive (or conservative) boundary node. It is called a* passive (or conservative) majorant *of colligations of the form* $(G_1, L+H, K_1)$ *on Hilbert spaces* $(\mathcal{U}_1, \mathcal{X}, \mathcal{Y}_1)$ *with domain* $\mathcal{Z} \cap \mathcal{N}(G_2)$ *where* $\mathcal{Z} \cap \mathcal{N}(G_2) \subset \mathcal{D}(H)$ *and* $\langle z, Hz \rangle_{\mathcal{X}} \leq 0$ *for all* $z \in \mathcal{Z} \cap \mathcal{N}(G_2)$ *and* H *is dominated by* L*, meaning that it satisfies one (or both) of the conditions* (i) *or* (ii) *of Theorem 3.2 in IV.*

The results of Section 3 of IV then say that if a colligation has a passive majorant, then also the system itself is a scattering passive boundary node. For example the internal well-posedness in the example in Section 5 of I is shown using such argument in the simple special case $H \in \mathcal{L}(\mathcal{X})$. For similar ideas in the port-Hamiltonian context, see [54, Chapter 6] by Villegas.

Bibliography

[1] W. Arendt, C.J.K. Batty, M. Hieber, and F. Neubrander. *Vector-valued Laplace Transforms and Cauchy Problems*. Birkhäuser, 2001.

[2] P. Axelsson and F. Gustafsson. Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering. Technical report, Linkopings Universitetet, 9 pages, December 2012.

[3] A. Bensoussan. *Filtrage Optimal des Systèmes linéaires*. Dunod, Paris, 1971.

[4] D.S. Bernstein and D.C. Hyland. The optimal projection equations for reduced-order state estimation. *Transactions on Automatic control*, 30(6):583–585, 1985.

[5] D.S. Bernstein and D.C. Hyland. The optimal projection equations for finite-dimensional fixed-order dynamic compensation of infinite-dimensional systems. *SIAM Journal on Control and Optimization*, 24:122–151, 1986.

[6] N. Besseling. *Stability analysis in continuous and discrete time*. Ph.D. thesis, University of Twente, 2011.

[7] J. Cervera, A.J. van der Schaft, and A. Baños. Interconnection of port-Hamiltonian systems and composition of Dirac structures. *Automatica J. of IFAC*, 43:212–225, 2007.

[8] D. Chapelle, P. Moireau, and P. Le Tallec. Robust filtering for joint state-parameter estimation in distributed mechanical systems. *Discrete and Continuous Dynamical Systems - Series A*, 23:65–84, 2009.

[9] R. Curtain and A.J. Pritchard. *Infinite Dimensional Linear Systems Theory*. Springer-Verlag, 1979.

[10] R. Curtain and G. Weiss. Exponential stabilization of well-posed systems by colocated feedback. *SIAM Journal on Control and Optimization*, 45(1):273–297, 2006.

[11] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications, **44**, Cambridge University Press, 1979.

[12] C.E. De Souza. On stabilizing properties of solutions of the Riccati difference equation. *Transactions on Automatic control*, 34(12):1313–1316, 1989.

[13] M. Dihlmann and B. Haasdonk. A reduced basis Kalman filter for parametrized partial differential equations. Technical report, Universität Stuttgart, 44 pages, August 2014.

[14] T. Duncan, B. Maslowski, and B. Pasik-Duncan. Linear-quadratic control for stochastic equations in a Hilbert space with fractional Brownian motions. *SIAM Journal on Control and Optimization*, 50(1):507–531, 2012.

[15] H.O. Fattorini. Boundary control systems. *SIAM Journal on Control*, 6(3):349–385, 1968.

[16] A. Germani, L. Jetto, and M. Piccioni. Galerkin approximation for optimal linear filtering of infinite-dimensional linear systems. *SIAM Journal on Control and Optimization*, 26(6):1287–1305, 1988.

[17] M. Grewal and A. Andrews. Applications of Kalman filtering in aerospace 1960 to the present. *IEEE Control Systems Magazine*, 30(3):69–78, 2010.

[18] W. Hager and L. Horowitz. Convergence and stability properties of the discrete Riccati operator equation and the associated optimal control and filtering problems. *SIAM Journal on Control and Optimization*, 14(2):295–312, 1976.

[19] G. Haine. Recovering the observable part of the initial data of an infinite-dimensional linear system with skew-adjoint generator. *To appear in: Mathematics of Control, Signals, and Systems*, 2014.

[20] V. Havu and J. Malinen. The Cayley transform as a time discretization scheme. *Numerical Functional Analysis and Optimization*, 28(7–8):825–851, 2007.

[21] L. Horowitz. *Optimal Filtering of Gyroscopic Noise*. Ph.D. thesis, Massachusetts Institute of Technology, 1974.

[22] J.M.J. Huttunen and H. Pikkarainen. Discretization error in dynamical inverse problems: one-dimensional model case. *Journal of Inverse and Ill-posed Problems*, 15(4):365–386, 2007.

[23] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[24] R.E. Kalman and R.S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83:95–107, 1961.

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

[26] M. Kurula, H. Zwart, A. van der Schaft, and J. Behrndt. Dirac structures and their composition on Hilbert spaces. *Journal of Mathematical Analysis and Applications*, 372:402–422, 2010.

[27] H. Li, Q. Lü, and X. Zhang. Recent progress on controllability/observability for systems governed by partial differential equations. *Journal of Systems Science and Complexity*, 23:527–545, 2010.

[28] D. Luenberger. Observing the state of a linear system. *Transactions on Military Electronics*, 8:74–80, 1964.

[29] T. Lukkari and J. Malinen. Webster's equation with curvature and dissipation. *Manuscript, 30 pages, arXiv:1204:4075*, 2013.

[30] T. Lukkari and J. Malinen. A posteriori error estimates for Webster's equation in wave propagation. *Manuscript*, 2014.

[31] G. Lumer and R.S. Phillips. Dissipative operators in a Banach space. *Pacific Journal of Mathematics*, 11:679–698, 1961.

[32] A. Macchelli. *Port Hamiltonian Systems. A unified approach for modeling and control finite and infinite dimensional physical systems.* Ph.D. thesis, University of Bologna, 2003.

[33] J. Malinen and O.J. Staffans. Conservative boundary control systems. *Journal of Differential Equations*, 231(1):290–312, 2006.

[34] J. Malinen and O.J. Staffans. Impedance passive and conservative boundary control systems. *Complex Analysis and Operator Theory*, 1:279–300, 2007.

[35] J. Malinen, O.J. Staffans, and G. Weiss. When is a linear system conservative? *Quarterly of Applied Mathematics*, 64:61–91, 2006.

[36] J. Maxwell. On governors. *Proceedings of the Royal Society of London*, 16:270–283, 1868.

[37] R. Muradore and G. Picci. Mixed H_2/H_∞ control: the discrete-time case. *Systems and Control Letters*, 54:1–13, 2005.

[38] M. Opmeer. *Model reduction for controller design for infinite-dimensional systems.* Ph.D. thesis, University of Groningen, 2006.

[39] H. Pikkarainen. State estimation approach to nonstationary inverse problems: discretization error and filtering problem. *Inverse Problems*, 22:365–379, 2006.

[40] A. Pritchard and J. Zabczyk. Stability and stabilizability of infinite dimensional systems. *SIAM Review*, 23(1):25–52, 1981.

[41] K.M. Przyluski. The Lyapunov equation and the problem of stability for linear bounded discrete-time systems in Hilbert space. *Applied Mathematics and Optimization*, 6:97–112, 1980.

[42] K. Ramdani, M. Tucsnak, and G. Weiss. Recovering the initial state of an infinite-dimensional system using observers. *Automatica*, 46:1616–1625, 2010.

[43] F. Riesz and B. Sz.-Nagy. *Vorlesungen über Funktionalanalysis.* Hochschulbücher für Mathematik, 27, 1968.

[44] D. Salamon. Infinite dimensional linear systems with unbounded control and observation: A functional analytic approach. *Transactions of the American Mathematical Society*, 300(2):383–431, 1987.

[45] D. Salamon. Realization theory in hilbert space. *Mathematical Systems Theory*, 21(1):147–164, 1989.

[46] D. Simon. *Optimal State Estimation — Kalman, H_∞, and Nonlinear Approaches.* John Wiley & Sons, 2006.

[47] D. Simon. Reduced order kalman filtering without model reduction. *Control and Intelligent Systems*, 35(2):169–174, 2007.

[48] A. Solin and S. Särkkä. Infinite-dimensional Bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data. *Physical Review E*, 88(5):052909, 2013.

[49] O.J. Staffans. *Well-posed Linear Systems*. Encyclopedia of Mathematics and its Applications **103**, Cambridge University Press, 2005.

[50] J. Sun. Sensitivity analysis of the discrete-time algebraic Riccati equation. *Linear Algebra and its Applications*, 275–276:595–615, 1998.

[51] H. Tanabe. *Equations of Evolution*. Pitman, London, 1979.

[52] M. Tucsnak and G. Weiss. *Observation and Control for Operator Semigroups*. Birkhäuser-Verlag, Basel, Switzerland, 2009.

[53] J. van Neerven. *Stochastic Evolution Equations*. ISEM Lecture Notes, Delft University of Technology, 2007/2008.

[54] J. Villegas. *A Port-Hamiltonian Approach to Distributed Parameter Systems*. Ph.D. thesis, University of Twente, 2007.

[55] N. Wahlström, P. Axelsson, and F. Gustafsson. Discretizing stochastic dynamical systems using Lyapunov equations. arXiv:1402.1358, February 2014.

[56] G. Weiss. Admissibility of unbounded control operators. *SIAM Journal on Control and Optimization*, 27(3):527–545, 1989.

[57] G. Weiss and X. Zhao. Well-posedness and controllability of a class of coupled linear systems. *SIAM Journal on Control and Optimization*, 48:2719–2750, 2009.

Errata

Publication I

On page 4, the right hand side of (6) should be $\operatorname{Re}\langle Kz, Gz\rangle_{\mathcal{U}}$.

Publication I

A. Aalto and J. Malinen. Compositions of Passive Boundary Control Systems. Mathematical Control and Related Fields, 3, 1–19, March 2013.

MATHEMATICAL CONTROL AND
RELATED FIELDS
Volume **3**, Number **1**, March **2013**

doi:10.3934/mcrf.2013.3.1

pp. **1–19**

COMPOSITIONS OF PASSIVE BOUNDARY CONTROL SYSTEMS

Atte Aalto and Jarmo Malinen

Department of Mathematics and Systems Analysis
Aalto University School of Science
PB 11100, 00076-Aalto, Finland

(Communicated by Olof Staffans)

Abstract. We show under mild assumptions that a composition of internally well-posed, impedance passive (or conservative) boundary control systems through Kirchhoff type connections is also an internally well-posed, impedance passive (resp., conservative) boundary control system. The proof is based on results of Malinen and Staffans [21]. We also present an example of such composition involving Webster's equation on a Y-shaped graph.

1. **Introduction.** We treat the solvability (forward in time) of dynamical boundary control systems that are composed by interconnecting a finite number of more simple boundary control *subsystems* that are already known to be solvable forward in time. The interconnections are given in terms of algebraic equations involving the boundary control/observation operators of the subsystems. The aggregate formed by the subsystems and their interconnections is called *a transmission graph* (see Definition 3.1), and it can be seen as a generalisation of mathematical transmission lines and networks. We assume throughout this work that all the subsystems are passive or conservative as described in, *e.g.*, Gorbachuk and Gorbachuk [9], Livšic [17], Malinen and Staffans [20, 21], Salamon [24, 25], and Staffans [26], and they are represented by equations of the form (5) below involving *strong boundary nodes*. Moreover, the interconnections respect passivity in the sense that they do not create energy. In Theorem 3.3 — the main result of this paper — we give conditions for checking the solvability (*i.e.*, internal well-posedness) and passivity of the transmission graph in terms of simple conditions on the subsystems and interconnections.

To illuminate the purpose of this paper, let us consider the following example from acoustic wave propagation. Given the interconnection graph in Fig. 1, the longitudinal wave propagation on its edges (*i.e.*, wave guides) is governed by

$$\frac{\partial^2 \psi^{(j)}}{\partial t^2}(x,t) = c^2 \frac{\partial^2 \psi^{(j)}}{\partial x^2}(x,t), \qquad x \in [0, l_j], \text{ and } t \in \mathbb{R}^+. \tag{1}$$

Here the index $j = A, ..., D$ refers to the index of the edge, and the arrows in Fig. 1 show the positive direction of the parametrisation $x \in [0, l_j]$. To the vertices

2010 *Mathematics Subject Classification.* Primary: 47A48; Secondary: 35R02, 47N70, 35L65.
Key words and phrases. Boundary control, passive system, distributed parameter system, well-posedness, composition, Cauchy problem.
The first author is supported by the Finnish graduate school in engineering mechanics
The results of this paper were presented in IFAC World Congress 2011 ([1]).

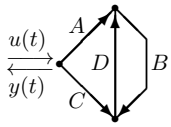

FIGURE 1. The example graph

ABD and BCD we impose Kirchhoff law type coupling (boundary) conditions (take vertex ABD for example):

$$\begin{cases} \frac{\partial\psi^{(A)}}{\partial t}(l_A,t) = \frac{\partial\psi^{(B)}}{\partial t}(0,t) = \frac{\partial\psi^{(D)}}{\partial t}(l_D,t), \\ A_A\frac{\partial\psi^{(A)}}{\partial x}(l_A,t) - A_B\frac{\partial\psi^{(B)}}{\partial x}(0,t) + A_D\frac{\partial\psi^{(D)}}{\partial x}(l_D,t) = 0. \end{cases} \tag{2}$$

We remark that in acoustics applications the state $\psi^{(j)}$ is chosen to be a velocity potential; then $p^{(j)} = \rho\frac{\partial\psi^{(j)}}{\partial t}$ gives the perturbation pressure and $v^{(j)} = -\frac{\partial\psi^{(j)}}{\partial x}$ gives the perturbation velocity for each edge. Thus, the first equation in (2) says that the pressure is continuous, and the second equation is a flux conservation law (the weights A_j can be understood as the cross-sectional areas of the wave guides).

We want to control the pressure at the vertex AC and observe the perturbation flux to the wave guides A and C. Defining the input and output

$$\begin{cases} u(t) := \frac{\partial\psi^{(A)}}{\partial t}(0,t) = \frac{\partial\psi^{(C)}}{\partial t}(0,t), \\ y(t) := -A_A\frac{\partial\psi^{(A)}}{\partial x}(0,t) - A_C\frac{\partial\psi^{(C)}}{\partial x}(0,t), \end{cases} \tag{3}$$

respectively, then equations (1) for $j = A, ..., D$ and (2) define a dynamical system whose solvability and energy conservation we wish to verify using Theorem 3.3.

We must consider first the solvability of the subsystems, that is, equations (1) on the edges with boundary conditions

$$\begin{bmatrix} \frac{\partial\psi^{(j)}}{\partial t}(0,t) \\ \frac{\partial\psi^{(j)}}{\partial t}(l_j,t) \end{bmatrix} = \begin{bmatrix} u_1^{(j)}(t) \\ u_2^{(j)}(t) \end{bmatrix} =: u^{(j)}(t). \tag{4}$$

After reducing (1) to a first order equation of form $\dot{z} = Lz$ with $z = \begin{bmatrix} \psi^{(j)} \\ p^{(j)} \end{bmatrix}$, defining operator G by (4), that is, by $Gz(t) = u^{(j)}(t)$, and K in a similar manner, we obtain *an internally well-posed boundary node* $\Xi^{(j)} = (G, L, K)$ that is *impedance conservative*, see Definitions 2.2 and 2.3. As explained after Definition 2.2, the initial value problem

$$\begin{array}{rcll} u(t) & = & Gz(t), & \\ \dot{z}(t) & = & Lz(t), & \\ y(t) & = & Kz(t), & t \in \mathbb{R}^+, \\ z(0) & = & z_0 & \end{array} \tag{5}$$

has a solution such that $\psi^{(j)}$ in equation (1) satisfies $\psi^{(j)} \in C^1(\mathbb{R}^+; L^2(0,l_j)) \cap C(\mathbb{R}^+; H^1[0,l_j])$ for all inputs $u^{(j)} \in C^2(\mathbb{R}^+;\mathbb{C}^2)$ and for all initial states z_0 that satisfy the boundary condition $Gz_0 = u(0)$. For technical details, see the (more general) example of Webster's equation presented in Section 5.

Now we have boundary nodes $\Xi^{(j)}$, $j = A, ..., D$ and coupling conditions of the form (2) for all vertices except the one that defines the external input and output through (3). They form a transmission graph as defined in Definition 3.1. Since

the components $\Xi^{(j)}$ are solvable and conservative, then by Theorem 3.3, also the resulting composed system is solvable forward in time and conservative in a similar way as any of its components.

Let us review the most relevant literature on compositions of (boundary control) systems. The feedback theory for (regular) well-posed linear systems is treated by Staffans in [26, Chapter 7] and by Weiss in [28] whose concept of admissibility of the feedback loops is related to the (internal) well-posedness of the composed system, but the theory can be used only when well-posedness of the components is verified by other means.

Transport equation on graphs is studied by Engel *et al.* in [7] by using semigroup techniques. For a study on the non-linear Saint-Venant equations on a star-shaped graph, see Gugat *et al.* [6]. A control algorithm for a string network is developed by Hundhammer and Leugering in [12] using a domain decomposition method. Further practical examples of compositions of PDEs with 1D spatial domains include semiconductor strips and lattice structures constructed of Timoshenko beams. Such systems have also been studied from the spectral point of view: asymptotic spectral properties of the Laplacian are studied by Kuchment and Zeng in [13] and by Rubinstein and Schatzman in [23] when its "graph-like" 3D spatial domain collapses to a graph with 1D edges. See also Latushkin and Pivovarchik [16] for a study on the spectral properties of the Sturm-Liouville equation on a Y-shaped graph.

Compositions of PDEs on 1D spatial domains are treated by Villegas in [27] and by Zwart *et al.* in [30] in terms of port-Hamiltonian framework. Compositions of more general systems are studied in, *e.g.*, Cervera *et al.* in [3] and Kurula *et al.* in [15] who treat systems that give raise to Dirac structures on their state spaces (see also Derkach *et al.* [5]). These contain impedance conservative, strong boundary control systems (as characterised in Definitions 2.2 and 2.3) as a special case. However, our approach is based on results of Malinen and Staffans [20, 21] that are reviewed in Section 2, and we are able to treat couplings of both passive and conservative systems at once.

In Section 5 we present a concrete example of a transmission graph, namely the human vocal tract with nasal cavity, modelled by Webster's equation on a Y-shaped graph. For more concrete examples, we refer to Malinen [19].

2. **Background.**

In this work we treat linear *boundary control systems* described by operator differential equations of the form (5) involving linear mappings G, L, and K:

Definition 2.1. Let $\Xi := (G, L, K)$ be a triple of linear mappings.

(i) Ξ is a *colligation* on the Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ if G, L, and K have the same domain $\mathcal{Z} = \mathrm{dom}(\Xi) \subset \mathcal{X}$ and values in $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$, respectively;

(ii) A colligation Ξ is *strong* if $\begin{bmatrix} G \\ L \\ K \end{bmatrix}$ is closed as an operator $\mathcal{X} \to \begin{bmatrix} \mathcal{U} \\ \mathcal{X} \\ \mathcal{Y} \end{bmatrix}$ with domain $\mathcal{Z}$, and L is closed with $\mathrm{dom}(L) = \mathcal{Z}$.

We call L the *interior operator*, G the *input (boundary) operator*, and K the *output (boundary) operator*. The space $\mathcal{Z}$ we call the *solution space*, $\mathcal{X}$ the *state space*, and $\mathcal{U}$ and $\mathcal{Y}$ the *input* and *output spaces*, respectively. In $\mathcal{Z}$ we use the graph norm $\|z\|_{\mathcal{Z}}^2 = \|z\|_{\mathcal{X}}^2 + \|Gz\|_{\mathcal{U}}^2 + \|Lz\|_{\mathcal{X}}^2 + \|Kz\|_{\mathcal{Y}}^2$.

In this paper we use the notations $\begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$ and $\bigoplus$ to represent orthogonal direct sum of (sub)spaces. See also Remark 3 for a discussion on the terms input and output.

The definition of strongness coincides with [21, Definition 4.4]. By [21, Lemma 4.5], Ξ is strong if and only if L is closed with $\mathrm{dom}(\Xi)$ and G and K are bounded with respect to the graph norm of L on $\mathrm{dom}(\Xi)$. We shall later make use of this fact.

Many dynamical systems defined by boundary controlled partial differential equations naturally adopt the form (5) associated with some colligation (G, L, K) on properly chosen spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$, see the example in Section 5. Equations (5) are solvable forward in time (at least) if Ξ satisfies somewhat stronger assumptions:

Definition 2.2. A strong colligation $\Xi = (G, L, K)$ is a *boundary node* on the Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ if the following conditions are satisfied:

(i) G is surjective and $\mathcal{N}(G)$ is dense in $\mathcal{X}$;
(ii) The operator $L|_{\mathcal{N}(G)}$ (interpreted as an operator in $\mathcal{X}$ with domain $\mathcal{N}(G)$) has a nonempty resolvent set.

This boundary node is *internally well-posed* (in the forward time direction) if, in addition,

(iii) $L|_{\mathcal{N}(G)}$ generates a C_0 semigroup.

This definition coincides with [20, Definition 1.1] for strong colligations. There are, in fact, well-posed boundary nodes that are not strong (see [21, Proposition 6.3]) but we do not consider such nodes in this paper[1]. We remark that also [8], [9], and [15] treat strong colligations (with different names), see [21, Theorem 5.2] and [15, Remark 4.4].

Note that "boundary node" does not refer to the vertices of the underlying graph structure. In fact, boundary nodes are related to the *edges* of the graph. Therefore, we always talk about *vertices* when referring to the graph structure.

If $\Xi = (G, L, K)$ is an internally well-posed boundary node, then (5) has a unique solution for sufficiently smooth input functions u and initial states z_0 compatible with $u(0)$. More precisely, as shown in [20, Lemma 2.6], for all $z_0 \in \mathcal{Z}$ and $u \in C^2(\mathbb{R}^+; \mathcal{U})$ with $Gz_0 = u(0)$ the first, second, and fourth of the equations in (5) have a unique solution $z \in C^1(\mathbb{R}^+; \mathcal{X}) \cap C(\mathbb{R}^+; \mathcal{Z})$, and hence we can define $y \in C(\mathbb{R}^+; \mathcal{Y})$ by the third equation in (5). In the rest of this article, when we say "a smooth solution of (5) on $\mathbb{R}^+$" we mean a solution with the above properties.

In a practical application, checking the solvability of (5), that is, verifying the conditions of Definition 2.2 may be difficult. However, in many cases this is not necessary because the system satisfies energy (in)equalities that can be verified using the Green–Lagrange inequality without an *a priori* knowledge of the well-posedness. Such energy laws make it easier to check the solvability, see Proposition 1 below. First we shall define impedance passivity/conservativity. To keep the notation simple, we assume that $\mathcal{U} = \mathcal{Y}$ even though it would be enough to assume that $\mathcal{U}$ and $\mathcal{Y}$ are a dual pair of Hilbert spaces with duality pairing $\langle \cdot, \cdot \rangle_{(\mathcal{Y}, \mathcal{U})}$; see [21, Definition 3.6] and the discussion preceding it.

Definition 2.3. Let $\Xi = (G, L, K)$ be a colligation on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$.

(i) Ξ is *impedance passive* if the following conditions hold:
  (a) $\begin{bmatrix} \beta G + K \\ \alpha - L \end{bmatrix}$ is surjective for some $\alpha, \beta \in \mathbb{C}^+$;
  (b) For all $z \in \mathrm{dom}(\Xi)$ we have the *Green–Lagrange inequality*

$$\mathrm{Re}\langle z, Lz \rangle_{\mathcal{X}} \leq \langle Kz, Gz \rangle_{\mathcal{U}}. \tag{6}$$

[1]To avoid confusion, we shall use the term strong boundary node below.

(ii) Impedance passive Ξ is *impedance conservative* if (6) holds as an equality, and (a) holds also for some $\alpha, \beta \in \mathbb{C}^-$.

Impedance passivity/conservativity is defined in [21, Definition 3.2] using the external Cayley transform of scattering passivity/conservativity (see also the discussion there). These definitions are equivalent by [21, Theorem 3.4]. We further remark that [21, Theorem 3.4] also states that for an impedance passive Ξ, condition (a) holds for all $\alpha, \beta \in \mathbb{C}^+$, and for an impedance conservative Ξ, condition (a) holds also for all $\alpha, \beta \in \mathbb{C}^-$.

Suppose now that Ξ is an internally well-posed, impedance passive boundary node and z a smooth solution of (5). Then (6) means plainly the energy inequality

$$\frac{d}{dt}\left(\frac{1}{2}\|z(t)\|_{\mathcal{X}}^2\right) \leq \langle y(t), u(t)\rangle_{\mathcal{U}} \quad \text{for all } t \in \mathbb{R}^+$$

where the right hand side stands for the instantaneous power entering the system, and the norm of $\mathcal{X}$ measures the energy stored in the state.

The following proposition utilising the energy balance laws is needed for checking internal well-posedness and impedance passivity/conservativity.

Proposition 1. *Let $\Xi = (G, L, K)$ be a strong colligation on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{U})$.*

(i) Suppose that (6) holds for all $z \in \operatorname{dom}(\Xi)$, and that $\left[\begin{smallmatrix} G \\ \alpha - L \end{smallmatrix}\right]$ is surjective for some $\alpha \in \mathbb{C}$ with $\operatorname{Re}(\alpha) \geq 0$. Then Ξ is an internally well-posed, impedance passive boundary node. If, in addition, (6) holds as an equality and $\left[\begin{smallmatrix} G \\ \alpha - L \end{smallmatrix}\right]$ is surjective also for some $\operatorname{Re}(\alpha) \leq 0$, then the internally well-posed boundary node Ξ is impedance conservative.

(ii) If Ξ is impedance passive, then it is an internally well-posed boundary node if and only if its input operator G is surjective.

For a proof, see [21, Theorem 4.3 and Remark 4.6] for part (i) and [21, Theorem 4.7] for part (ii).

Internally well-posed boundary nodes can always be written in terms of more general and complicated *system nodes* (see [20], [22], and [26]) but they are excluded from *state linear systems* studied in [4]. A functional analytic setting of boundary control systems, that is independent of the system node setting, was formulated by Fattorini in [8] and significant progress was made by Salamon in [24, 25]. See also Greiner [10] for a similar presentation.

3. **Transmission graphs as colligations.** Assume that we have colligations $\Xi^{(j)} = \left(G^{(j)}, L^{(j)}, K^{(j)}\right)$ on Hilbert spaces $\left(\mathcal{U}^{(j)}, \mathcal{X}^{(j)}, \mathcal{Y}^{(j)}\right)$ with solution spaces $\mathcal{Z}^{(j)}$, $j = 1, ..., m$, where

$$\begin{aligned} G^{(j)} &= \begin{bmatrix} G_1^{(j)} \\ \vdots \\ G_{k_j}^{(j)} \end{bmatrix} : \operatorname{dom}(\Xi^{(j)}) \to \mathcal{U}^{(j)} = \begin{bmatrix} \mathcal{U}_1^{(j)} \\ \vdots \\ \mathcal{U}_{k_j}^{(j)} \end{bmatrix} \quad \text{and} \\ K^{(j)} &= \begin{bmatrix} K_1^{(j)} \\ \vdots \\ K_{k_j}^{(j)} \end{bmatrix} : \operatorname{dom}(\Xi^{(j)}) \to \mathcal{Y}^{(j)} = \begin{bmatrix} \mathcal{Y}_1^{(j)} \\ \vdots \\ \mathcal{Y}_{k_j}^{(j)} \end{bmatrix}. \end{aligned} \tag{7}$$

That is, the Hilbert spaces $\mathcal{U}^{(j)}$ and $\mathcal{Y}^{(j)}$ are represented by an orthogonal direct sum of k_j subspaces each, and the corresponding input and output operators are split accordingly.

In order to define the topological structure of the transmission graph, we define *control vertices* $\mathcal{I}^1, ..., \mathcal{I}^N$ (where $N \neq 0$) and *closed vertices* $\mathcal{J}^1, ..., \mathcal{J}^M$ as pairwise disjoint sets of index pairs (j, i) where j refers to the subsystem $\Xi^{(j)}$ and $i \in \{1, ..., k_j\}$ refers to the i^{th} component in the splitting (7). We assume that every pair (j, i) for $j = 1, ..., m;\ i = 1, ..., k_j$ belongs to some vertex. This is not a restriction since uncoupled input/output pairs can be included as singleton vertices, as in our example in Section 5.

Each vertex defines a coupling between the subsystems in such a way that all inputs $u_i^{(j)}$ whose index pairs (j, i) belong to the same vertex are equal, and the corresponding outputs are summed up. In addition, for closed vertices we require that the outputs sum up to zero. For such coupling to be possible, it is required that the compatibility conditions

$$\mathcal{U}_i^{(j)} = \mathcal{U}_q^{(p)} \quad \text{and} \quad \mathcal{Y}_i^{(j)} = \mathcal{Y}_q^{(p)} \tag{8}$$

hold for all $(j, i), (p, q) \in \mathcal{I}^k,\ k = 1, ..., N$ and for all $(j, i), (p, q) \in \mathcal{J}^l,\ l = 1, ..., M$. The couplings are written in terms of input and output operators as follows:

(i) for all control and closed vertices, *the continuity equations*

$$G_i^{(j)} z^{(j)} = G_q^{(p)} z^{(p)} \qquad \text{for } z^{(j)} \in \mathcal{Z}^{(j)} \text{ and } z^{(p)} \in \mathcal{Z}^{(p)} \tag{9}$$

hold, *i.e.*, (9) holds for all $(j, i), (p, q) \in \mathcal{I}^k,\ k = 1, ..., N$ and for all $(j, i), (p, q) \in \mathcal{J}^l,\ l = 1, ..., M$; and

(ii) for closed vertices, also *the balance equations*

$$\sum_{(j,i) \in \mathcal{J}^l} K_i^{(j)} z^{(j)} = 0 \qquad \text{for } z^{(j)} \in \mathcal{Z}^{(j)} \text{ and } l = 1, ..., M \tag{10}$$

hold.

Control vertices are exactly those couplings where external signals are applied. If the transfer function (see [20, Section 2]) of each $\Xi^{(j)}$ represents electrical admittance, then the physical dimensions of $\mathcal{U}^{(j)}$ and $\mathcal{Y}^{(j)}$ are the voltage and current, respectively. Equations (9) and (10) are the classical Kirchhoff laws, namely, the continuity of voltage and the conservation of charge.

Definition 3.1. Assume that $\Xi^{(j)}$ are colligations with splittings as described above in (7). Suppose that sets $\mathcal{I}^1, ..., \mathcal{I}^N$ and $\mathcal{J}^1, ..., \mathcal{J}^M$ are defined consistently with this splitting so that the compatibility conditions (8) hold. The ordered triple

$$\Gamma := \left(\left\{ \Xi^{(j)} \right\}_{j=1}^{m}, \left\{ \mathcal{I}^k \right\}_{k=1}^{N}, \left\{ \mathcal{J}^l \right\}_{l=1}^{M} \right)$$

is called a *transmission graph*.

A transmission graph is a notion that contains the building blocks and the "assembly instructions" of the composition. Together with coupling conditions (9) and (10), it gives rise to a dynamical system as follows:

Definition 3.2. Let Γ be a transmission graph as in Definition 3.1. Using the same notation, we define the *colligation of the transmission graph* as the triple

$\Xi_\Gamma = (G, L, K)$ on the Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ where[2]

$$\mathcal{X} := \bigoplus_{j=1}^{m} \mathcal{X}^{(j)}, \qquad \mathcal{U} := \bigoplus_{\substack{(j,i)\in\mathcal{I}^k \\ k=1,\dots,M}} \mathcal{U}_i^{(j)}, \qquad \mathcal{Y} := \bigoplus_{\substack{(j,i)\in\mathcal{I}^k \\ k=1,\dots,M}} \mathcal{Y}_i^{(j)},$$

$$\mathrm{dom}(\Xi_\Gamma) := \left\{ \bigoplus_{j=1}^{m} \mathcal{Z}^{(j)} \;\middle|\; (9) \text{ and } (10) \text{ hold} \right\},$$

$$G := [G_{k,j}]_{\substack{k=1,\dots,N \\ j=1,\dots,m}}, \quad L := \begin{bmatrix} L^{(1)} & & \\ & \ddots & \\ & & L^{(m)} \end{bmatrix}, \text{ and } K := [K_{k,j}]_{\substack{k=1,\dots,N \\ j=1,\dots,m}}$$

where

$$G_{k,j} := \begin{cases} G_k^{(j)}/|\mathcal{I}^k|, & \text{if } (j,k) \in \mathcal{I}^k, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad K_{k,j} := \begin{cases} K_k^{(j)}, & \text{if } (j,k) \in \mathcal{I}^k, \\ 0, & \text{otherwise.} \end{cases}$$

In order to make the preceding definitions more intuitive, let us return to the example on the wave equation on the graph of Fig. 1, presented in the introduction. We have four boundary nodes $\Xi^{(j)}$, $j = A, ..., D$ whose input and output spaces are split into two parts, see equation (4). In the graph, there is one control vertex $\mathcal{I}^1 = \{(A,1),(C,1)\}$ and two closed vertices $\mathcal{J}^1 = \{(A,2),(B,1),(D,2)\}$ and $\mathcal{J}^2 = \{(B,2),(C,2),(D,1)\}$.

The dynamical system given by (1), (2), and (3) corresponds to the colligation of the transmission graph $\Gamma := \left(\left\{\Xi^{(j)}\right\}_{j=A}^{D}, \left\{\mathcal{I}^1\right\}, \left\{\mathcal{J}^1, \mathcal{J}^2\right\}\right)$. More precisely, equations in (2) are equivalent with (9) and (10) and the input and output operators given in Definition 3.2 yield the input/output of equation (3).

The main result of this paper is the following:

Theorem 3.3. *Assume that the transmission graph Γ is composed of internally well-posed, impedance passive (or conservative), strong boundary nodes $\Xi^{(j)} = \left(G^{(j)}, L^{(j)}, K^{(j)}\right)$ with the following property:*

$$\text{all of the operators } \begin{bmatrix} G^{(j)} \\ K^{(j)} \end{bmatrix} \text{ are surjective.} \tag{11}$$

Then the colligation of Γ is an impedance passive (respectively, conservative), internally well-posed, strong boundary node.

This is proved in three steps (Lemmas 4.1, 4.2, and 4.3) presented in the following section. The assumption (11) can be relaxed (see Remark 1) but this condition appears to hold in many applications (as in our example in Section 5).

4. **Proof of Theorem 3.3.** Suppose we are given a transmission graph Γ. We reconstruct this graph by a finite number of three different kinds of steps, starting from its components $\Xi^{(j)}$. In step 1, we form a partial parallel connection between two compatible colligations to obtain a new colligation, see Fig. 2a. We remark that such parallel connections are treated in [26, Examples 2.3.13 and 5.1.17] for system nodes. In step 2, we form loops by joining two signals of a single colligation to obtain a new colligation, see Fig. 2b. Both the control vertices and the closed vertices are treated similarly at this stage: all the vertices are left "open" so that (9) is satisfied but (10) is not. After constructing the full coupling graph structure by taking a

[2] In sums of $\mathcal{U}$ and $\mathcal{Y}$, pick *one* pair $(j,i) \in \mathcal{I}^k$ for each k.

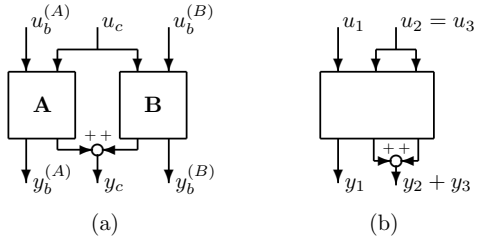

FIGURE 2. (a) The partial parallel coupling; (b) The loop coupling

finite number of steps 1 and 2 in some order, the final step 3 is taken to close those vertices that are not used for control/observation; then condition (10) is satisfied as well. The transmission graph Γ and its colligation have now been reconstructed, and the remaining (open) vertices are exactly the control vertices of Γ.

By this procedure, it is possible to synthesise any transmission graph. In Lemmas 4.1, 4.2, and 4.3, we show that if we start from internally well-posed, impedance passive/conservative strong boundary nodes, then the resulting colligations after steps 1, 2, and 3 (respectively) are internally well-posed, impedance passive/conservative, strong boundary nodes as well. This is required for iterated application of these steps in order to prove Theorem 3.3. The reconstruction procedure is demonstrated in Section 4.4 by using the graph of Fig. 1.

4.1. **Step 1: Partial parallel coupling.** Assume that we have two colligations $\Xi^{(A)} = \left(\begin{bmatrix} G_b^{(A)} \\ G_c^{(A)} \end{bmatrix}, L^{(A)}, \begin{bmatrix} K_b^{(A)} \\ K_c^{(A)} \end{bmatrix}\right)$ and $\Xi^{(B)} = \left(\begin{bmatrix} G_b^{(B)} \\ G_c^{(B)} \end{bmatrix}, L^{(B)}, \begin{bmatrix} K_b^{(B)} \\ K_c^{(B)} \end{bmatrix}\right)$ on Hilbert spaces $\left(\begin{bmatrix} \mathcal{U}_b^{(A)} \\ \mathcal{U}_c \end{bmatrix}, \mathcal{X}^{(A)}, \begin{bmatrix} \mathcal{Y}_b^{(A)} \\ \mathcal{Y}_c \end{bmatrix}\right)$ and $\left(\begin{bmatrix} \mathcal{U}_b^{(B)} \\ \mathcal{U}_c \end{bmatrix}, \mathcal{X}^{(B)}, \begin{bmatrix} \mathcal{Y}_b^{(B)} \\ \mathcal{Y}_c \end{bmatrix}\right)$ with solution spaces $\mathcal{Z}^{(A)}$ and $\mathcal{Z}^{(B)}$, respectively.

Now define the composed colligation $\Xi^{(AB)} := \left(G^{(AB)}, L^{(AB)}, K^{(AB)}\right)$ on the Hilbert spaces

$$\mathcal{X}^{(AB)} := \begin{bmatrix} \mathcal{X}^{(A)} \\ \mathcal{X}^{(B)} \end{bmatrix}, \quad \mathcal{U}^{(AB)} := \begin{bmatrix} \mathcal{U}_b^{(A)} \\ \mathcal{U}_c \\ \mathcal{U}_b^{(B)} \end{bmatrix}, \quad \text{and} \quad \mathcal{Y}^{(AB)} := \begin{bmatrix} \mathcal{Y}_b^{(A)} \\ \mathcal{Y}_c \\ \mathcal{Y}_b^{(B)} \end{bmatrix}$$

by

$$L^{(AB)} := \begin{bmatrix} L^{(A)} & 0 \\ 0 & L^{(B)} \end{bmatrix},$$

$$G^{(AB)} := \begin{bmatrix} G_b^{(A)} & 0 \\ G_c^{(A)} & 0 \\ 0 & G_b^{(B)} \end{bmatrix}, \quad \text{and} \quad K^{(AB)} := \begin{bmatrix} K_b^{(A)} & 0 \\ K_c^{(A)} & K_c^{(B)} \\ 0 & K_b^{(B)} \end{bmatrix}.$$

The domain of the colligation is

$$\operatorname{dom}(\Xi^{(AB)}) := \left\{ \begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \begin{bmatrix} \operatorname{dom}(\Xi^{(A)}) \\ \operatorname{dom}(\Xi^{(B)}) \end{bmatrix} \;\middle|\; G_c^{(A)} z^{(A)} = G_c^{(B)} z^{(B)} \right\}.$$

Such partial parallel coupling is illustrated in Fig. 2a. We now show that such coupling of two boundary nodes is also a boundary node and the coupling preserves internal well-posedness and passivity/conservativity.

Lemma 4.1. *Let $\Xi^{(A)}$, $\Xi^{(B)}$, and $\Xi^{(AB)}$ be as defined above. If the colligations $\Xi^{(A)}$ and $\Xi^{(B)}$ are internally well-posed, impedance passive (conservative), strong boundary nodes such that both $\begin{bmatrix} G^{(A)} \\ K^{(A)} \end{bmatrix}$ and $\begin{bmatrix} G^{(B)} \\ K^{(B)} \end{bmatrix}$ are surjective, then the composed colligation $\Xi^{(AB)}$ is an internally well-posed, impedance passive (resp., conservative), strong boundary node with the property that $\begin{bmatrix} G^{(AB)} \\ K^{(AB)} \end{bmatrix}$ is surjective.*

Proof. We start by showing that $\Xi^{(AB)}$ is a strong colligation. First, we show that $\Xi^{(AB)}$ is closed. Assume that $\operatorname{dom}(\Xi^{(AB)}) \ni \begin{bmatrix} z_n^{(A)} \\ z_n^{(B)} \end{bmatrix} \to \begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix}$ and

$$\begin{bmatrix} G_b^{(A)} & 0 \\ G_c^{(A)} & 0 \\ 0 & G_b^{(B)} \end{bmatrix} \begin{bmatrix} z_n^{(A)} \\ z_n^{(B)} \end{bmatrix} \to \begin{bmatrix} u_b^{(A)} \\ u_c \\ u_b^{(B)} \end{bmatrix}, \quad \begin{bmatrix} L^{(A)} & 0 \\ 0 & L^{(B)} \end{bmatrix} \begin{bmatrix} z_n^{(A)} \\ z_n^{(B)} \end{bmatrix} \to \begin{bmatrix} x^{(A)} \\ x^{(B)} \end{bmatrix},$$

$$\text{and} \quad \begin{bmatrix} K_b^{(A)} & 0 \\ K_c^{(A)} & K_c^{(B)} \\ 0 & K_b^{(B)} \end{bmatrix} \begin{bmatrix} z_n^{(A)} \\ z_n^{(B)} \end{bmatrix} \to \begin{bmatrix} y_b^{(A)} \\ y_c \\ y_b^{(B)} \end{bmatrix}.$$

Since colligations $\Xi^{(A)}$ and $\Xi^{(B)}$ are strong, the operators $L^{(A)}$ and $L^{(B)}$ are closed, $\begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \begin{bmatrix} \operatorname{dom}(\Xi^{(A)}) \\ \operatorname{dom}(\Xi^{(B)}) \end{bmatrix}$, and also $L^{(A)} z^{(A)} = x^{(A)}$ and $L^{(B)} z^{(B)} = x^{(B)}$. To show that $\begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \operatorname{dom}(\Xi^{(AB)})$, we need to use the strongness of $\Xi^{(A)}$ and $\Xi^{(B)}$ which means that $G_c^{(A)}$ and $G_c^{(B)}$ are continuous with respect to the graph norms of $L^{(A)}$ and $L^{(B)}$, respectively (see the comment after Definition 2.1). Hence

$$\begin{aligned} &\|G_c^{(A)} z^{(A)} - G_c^{(B)} z^{(B)}\|_{\mathcal{U}_c} \le \|G_c^{(A)}(z^{(A)} - z_n^{(A)})\|_{\mathcal{U}_c} + \|G_c^{(B)}(z^{(B)} - z_n^{(B)})\|_{\mathcal{U}_c} \\ &\le M_A \left(\|z^{(A)} - z_n^{(A)}\|_{\mathcal{X}^{(A)}} + \|L^{(A)}(z^{(A)} - z_n^{(A)})\|_{\mathcal{X}^{(A)}} \right) + \\ &\quad + M_B \left(\|z^{(B)} - z_n^{(B)}\|_{\mathcal{X}^{(B)}} + \|L^{(B)}(z^{(B)} - z_n^{(B)})\|_{\mathcal{X}^{(B)}} \right) \to 0 \text{ when } n \to \infty \end{aligned}$$

where we have used the fact $G_c^{(A)} z_n^{(A)} = G_c^{(B)} z_n^{(B)}$. This implies $G_c^{(A)} z^{(A)} = G_c^{(B)} z^{(B)}$ meaning that $\begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \operatorname{dom}(\Xi^{(AB)})$. By a similar computation we can verify

$$\begin{bmatrix} G_b^{(A)} & 0 \\ G_c^{(A)} & 0 \\ 0 & G_b^{(B)} \end{bmatrix} \begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} = \begin{bmatrix} u_b^{(A)} \\ u_c \\ u_b^{(B)} \end{bmatrix} \text{ and } \begin{bmatrix} K_b^{(A)} & 0 \\ K_c^{(A)} & K_c^{(B)} \\ 0 & K_b^{(B)} \end{bmatrix} \begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} = \begin{bmatrix} y_b^{(A)} \\ y_c \\ y_b^{(B)} \end{bmatrix}.$$

Closedness of $L^{(AB)}$ with domain $\operatorname{dom}(L^{(AB)}) = \operatorname{dom}(\Xi^{(AB)})$ is shown similarly. Thus, $\Xi^{(AB)}$ is strong colligation. Note that in the preceding computation, we did not need $G_c^{(A)} z_n^{(A)} \to u_c$ to show $\begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \operatorname{dom}(\Xi^{(AB)})$, *i.e.*, $G_c^{(A)} z^{(A)} = G_c^{(B)} z^{(B)}$.

We proceed to show that $\Xi^{(AB)}$ is an internally well-posed, impedance passive boundary node with the help of Proposition 1. Surjectivity of $\begin{bmatrix} G^{(AB)} \\ \alpha - L^{(AB)} \end{bmatrix}$ (with domain $\operatorname{dom}(\Xi^{(AB)})$) for some $\alpha \in \mathbb{C}$ with $\operatorname{Re} \alpha \ge 0$ follows from the fact that $\begin{bmatrix} G^{(A)} \\ \alpha - L^{(A)} \end{bmatrix}$ and $\begin{bmatrix} G^{(B)} \\ \alpha - L^{(B)} \end{bmatrix}$ are surjective for the same α. All that is left is to show

that the Green–Lagrange identity (6) holds:

$$\begin{aligned}\operatorname{Re}\langle z, L^{(AB)}z\rangle_{\mathcal{X}^{(AB)}} &= \operatorname{Re}\Big(\langle z^{(A)}, L^{(A)}z^{(A)}\rangle_{\mathcal{X}^{(A)}} + \langle z^{(B)}, L^{(B)}z^{(B)}\rangle_{\mathcal{X}^{(B)}}\Big)\\ &\le \operatorname{Re}\Big(\langle K_b^{(A)}z^{(A)}, G_b^{(A)}z^{(A)}\rangle_{\mathcal{U}_b^{(A)}} + \langle K_c^{(A)}z^{(A)}, G_c^{(A)}z^{(A)}\rangle_{\mathcal{U}_c} +\\ &\quad + \langle K_b^{(B)}z^{(B)}, G_b^{(B)}z^{(B)}\rangle_{\mathcal{U}_b^{(B)}} + \langle K_c^{(B)}z^{(B)}, G_c^{(B)}z^{(B)}\rangle_{\mathcal{U}_c}\Big)\\ &= \operatorname{Re}\langle K^{(AB)}z, G^{(AB)}z\rangle_{\mathcal{U}^{(AB)}}\end{aligned}$$

where the last equation follows from $G_c^{(A)}z^{(A)} = G_c^{(B)}z^{(B)}$ and definitions of $G^{(AB)}$ and $K^{(AB)}$. Surjectivity of $\left[\begin{smallmatrix} G^{(AB)} \\ K^{(AB)}\end{smallmatrix}\right]$ follows from surjectivity of $\left[\begin{smallmatrix} G^{(A)} \\ K^{(A)}\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} G^{(B)} \\ K^{(B)}\end{smallmatrix}\right]$.

The conservativity is verified by repeating the latter part of the proof with $-\alpha$ in place of α and replacing the inequality in Green–Lagrange identity by equality. □

4.2. **Step 2: Loop coupling.** Now assume that we have a colligation $\Xi = (G, L, K)$ on the Hilbert spaces $\left(\left[\begin{smallmatrix}\mathcal{U}_1\\ \mathcal{U}_c\\ \mathcal{U}_c\end{smallmatrix}\right], \mathcal{X}, \left[\begin{smallmatrix}\mathcal{Y}_1\\ \mathcal{Y}_c\\ \mathcal{Y}_c\end{smallmatrix}\right]\right)$ where $G = \left[\begin{smallmatrix} G_1\\ G_2\\ G_3\end{smallmatrix}\right]$ and $K = \left[\begin{smallmatrix} K_1\\ K_2\\ K_3\end{smallmatrix}\right]$, *i.e.*, the input and output operators and spaces can be split into (at least) three parts. We "glue" two of these parts together to form another colligation $\widehat{\Xi} := \big(\widehat{G}, \widehat{L}, \widehat{K}\big)$ on the Hilbert spaces $\left(\left[\begin{smallmatrix}\mathcal{U}_1\\ \mathcal{U}_c\end{smallmatrix}\right], \mathcal{X}, \left[\begin{smallmatrix}\mathcal{Y}_1\\ \mathcal{Y}_c\end{smallmatrix}\right]\right)$ with $\operatorname{dom}(\widehat{\Xi}) := \big\{z \in \operatorname{dom}(\Xi) \;\big|\; G_2 z = G_3 z\big\}$, $\widehat{L} := L|_{\operatorname{dom}(\widehat{\Xi})}$, $\widehat{G} := \left[\begin{smallmatrix} G_1\\ G_2\end{smallmatrix}\right]$, and $\widehat{K} := \left[\begin{smallmatrix} K_1\\ K_2+K_3\end{smallmatrix}\right]$.

The block diagram of such coupling is shown in Fig. 2b. As in step 1, we show that if the original colligation Ξ is an internally well-posed, impedance passive (conservative), strong boundary node, then $\widehat{\Xi}$ is one as well.

Lemma 4.2. *Let Ξ and $\widehat{\Xi}$ be as defined above. If the colligation Ξ is an internally well-posed, impedance passive (conservative), strong boundary node such that $\left[\begin{smallmatrix} G\\ K\end{smallmatrix}\right]$ is surjective, then also $\widehat{\Xi}$ is an internally well-posed, impedance passive (resp., conservative), strong boundary node with the property that $\left[\begin{smallmatrix} \widehat{G}\\ \widehat{K}\end{smallmatrix}\right]$ is surjective.*

Proof. Strongness of $\widehat{\Xi}$ is shown as before in Lemma 4.1.

Surjectivity of $\left[\begin{smallmatrix} \widehat{G}\\ \alpha - \widehat{L}\end{smallmatrix}\right]$ for some $\alpha \in \mathbb{C}$ with $\operatorname{Re}\alpha \ge 0$ is easy to see, and also Green–Lagrange identity holds in $\operatorname{dom}(\widehat{\Xi})$:

$$\begin{aligned}\operatorname{Re}\langle z, \widehat{L}z\rangle_{\widehat{\mathcal{X}}} &\le \operatorname{Re}\langle K_1 z, G_1 z\rangle_{\mathcal{U}_1} + \operatorname{Re}\langle K_2 z, G_2 z\rangle_{\mathcal{U}_c} + \operatorname{Re}\langle K_3 z, G_3 z\rangle_{\mathcal{U}_c}\\ &= \operatorname{Re}\langle K_1 z, G_1 z\rangle_{\mathcal{U}_1} + \operatorname{Re}\langle (K_2+K_3) z, G_2 z\rangle_{\mathcal{U}_c}\\ &= \operatorname{Re}\langle \widehat{K}z, \widehat{G}z\rangle_{\widehat{\mathcal{U}}}\end{aligned}$$

where the second equality follows from $G_2 z = G_3 z$ and the last from the definitions of $\widehat{G}$ and $\widehat{K}$. Surjectivity of $\left[\begin{smallmatrix} \widehat{G}\\ \widehat{K}\end{smallmatrix}\right]$ follows from surjectivity of $\left[\begin{smallmatrix} G\\ K\end{smallmatrix}\right]$.

If Ξ is conservative, then to show conservativity of $\widehat{\Xi}$, just repeat the proof with $-\alpha$ in place of α and replace the inequality in the Green–Lagrange identity with equality. □

4.3. **Step 3: Closing the vertices.** In this step, we single out some vertices as control/observation vertices and permanently "close" all others with respect to additional external signals. Note that after steps 1 and 2, under the assumptions of Lemmas 4.1 and 4.2, the resulting colligation is an internally well-posed boundary node, such that (9) is satisfied. This closing means that we require also (10) to be

satisfied, and we now show that this can be done without sacrificing the internal well-posedness or passivity/conservativity.

So let $\Xi = (G, L, K)$ be a colligation on the Hilbert spaces $\left(\left[\begin{smallmatrix}\mathcal{U}_1\\ \mathcal{U}_2\end{smallmatrix}\right], \mathcal{X}, \left[\begin{smallmatrix}\mathcal{Y}_1\\ \mathcal{Y}_2\end{smallmatrix}\right]\right)$ with splittings $G = \left[\begin{smallmatrix}G_1\\ G_2\end{smallmatrix}\right]$ and $K = \left[\begin{smallmatrix}K_1\\ K_2\end{smallmatrix}\right]$ where G_2 and K_2 correspond to vertices we want to close. Define the new colligation by $\widehat{\Xi} := \left(G_1, \widehat{L}, K_1\right)$ on the Hilbert spaces $(\mathcal{U}_1, \mathcal{X}, \mathcal{Y}_1)$ with $\operatorname{dom}(\widehat{\Xi}) := \operatorname{dom}(\Xi) \cap \mathcal{N}(K_2)$ and $\widehat{L} := L|_{\operatorname{dom}(\widehat{\Xi})}$.

Lemma 4.3. *Let Ξ and $\widehat{\Xi}$ be as defined above. If Ξ is an internally well-posed, impedance passive (conservative), strong boundary node with the property that $\left[\begin{smallmatrix}G\\ K\end{smallmatrix}\right]$ is surjective, then also $\widehat{\Xi}$ is an internally well-posed, impedance passive (resp., conservative), strong boundary node.*

Proof. We carry out a partial flow inversion and interchange the roles of G_2 and K_2. More precisely, we shall prove that $\widetilde{\Xi} := \left(\widetilde{G}, L, \widetilde{K}\right)$ on Hilbert spaces $\left(\left[\begin{smallmatrix}\mathcal{U}_1\\ \mathcal{Y}_2\end{smallmatrix}\right], \mathcal{X}, \left[\begin{smallmatrix}\mathcal{Y}_1\\ \mathcal{U}_2\end{smallmatrix}\right]\right)$ where $\widetilde{G} := \left[\begin{smallmatrix}G_1\\ K_2\end{smallmatrix}\right]$, $\widetilde{K} := \left[\begin{smallmatrix}K_1\\ G_2\end{smallmatrix}\right]$, and $\operatorname{dom}(\widetilde{\Xi}) := \operatorname{dom}(\Xi)$, is an internally well-posed, impedance passive (conservative), strong boundary node. Colligation $\widehat{\Xi}$ is then obtained from $\widetilde{\Xi}$ by restricting the solution space to $\mathcal{N}(K_2)$, and it clearly has all the properties as claimed, see Definition 2.2 and the comment after Definition 2.1 concerning the strongness of $\widehat{\Xi}$.

It is trivial that $\widetilde{\Xi}$ is a strong colligation. One way to see the interchangeability of G_2 and K_2 is directly by Definition 2.3 with $\beta = 1$:

$$\begin{bmatrix}\widetilde{G}+\widetilde{K}\\ \alpha - L\end{bmatrix} = \begin{bmatrix}\left[\begin{smallmatrix}G_1\\ K_2\end{smallmatrix}\right] + \left[\begin{smallmatrix}K_1\\ G_2\end{smallmatrix}\right]\\ \alpha - L\end{bmatrix} = \begin{bmatrix}\left[\begin{smallmatrix}G_1\\ G_2\end{smallmatrix}\right] + \left[\begin{smallmatrix}K_1\\ K_2\end{smallmatrix}\right]\\ \alpha - L\end{bmatrix} = \begin{bmatrix}G+K\\ \alpha - L\end{bmatrix}.$$

The surjectivity of this operator follows from impedance passivity of Ξ. Similarly for the conservative system we also need the operator

$$\begin{bmatrix}\widetilde{G}-\widetilde{K}\\ \alpha - L\end{bmatrix} = \left[\begin{array}{cc|c} I & 0 & 0\\ 0 & -I & 0\\ \hline 0 & 0 & I\end{array}\right] \begin{bmatrix}G-K\\ \alpha - L\end{bmatrix}$$

to be surjective which holds by the conservativity of Ξ, see Definition 2.3 with $\beta = -1$. The Green–Lagrange (in)equality trivially holds, and it follows that $\widetilde{\Xi}$ is an impedance passive (conservative), strong colligation.

Finally, by Proposition 1, the surjectivity of $\left[\begin{smallmatrix}G_1\\ K_2\end{smallmatrix}\right]$ implies that $\widetilde{\Xi}$ is an internally well-posed boundary node. □

Remark 1. Assumption (11) is actually stronger than what was needed in Theorem 3.3. Indeed, it was only used in the last lines of the proof of Lemma 4.3. However, the minimal sufficient conditions are impossible to formulate in terms of the control/observation operators of the subsystems. Instead, we would have to consider the whole composed system. The requirement is that the control operator of the composed system has to remain surjective despite the couplings in the closed vertices.

Remark 2. The partial parallel coupling could be constructed by first forming a cross product of systems $\Xi^{(A)}$ and $\Xi^{(B)}$, see [26, Example 2.3.10]. It is easy to see that the cross product preserves all the desired properties of the colligations. The partial parallel coupling can then be formed by applying a loop coupling to the product system. This means that Lemma 4.1 actually follows from Lemma 4.2.

Remark 3. Using the words input and output for Gz and Kz is somewhat misleading. In fact, since our coupling equations (9) and (10) include conditions involving both Gz and Kz, we have to assume that also the flow-inverted system is solvable; that is, solvable if G and K are interchanged. For many systems this is not a serious restriction and, in fact, the whole concept of *abstract boundary spaces* (introduced in [9]) is based on the existence of such interchangeable pair of possible boundary conditions. See also Derkach *et al.* [5] for a study of compositions of systems using such abstract boundary spaces and Kurula [14] for an introduction of *state/signal systems* that are based on equal treatment of inputs and outputs.

4.4. **Example on constructing the composition.** Let us once more return to the example of the introduction. We reconstruct the interconnection graph in four phases which are illustrated in Fig. 3. We start with four boundary nodes labelled with A, B, C, and D. The input and output operators and spaces of each system are split into two parts, *i.e.*, $k_j = 2$. The vertices are labelled with 1 and 2 and the arrows in Fig. 3 point from 1 to 2.

• *Phase 1.* We start with colligations $\Xi^{(j)} = \left(\begin{bmatrix} G_1^{(j)} \\ G_2^{(j)} \end{bmatrix}, L^{(j)}, \begin{bmatrix} K_1^{(j)} \\ K_2^{(j)} \end{bmatrix} \right)$ on the Hilbert spaces $\left(\begin{bmatrix} \mathcal{U}_1^{(j)} \\ \mathcal{U}_2^{(j)} \end{bmatrix}, \mathcal{X}^{(j)}, \begin{bmatrix} \mathcal{Y}_1^{(j)} \\ \mathcal{Y}_2^{(j)} \end{bmatrix} \right)$, $j = A, B, C, D$.

• *Phase 2.* The system A is connected to B, and C to D, by a partial parallel coupling so that we obtain two colligations $\Xi^{(AB)}$ and $\Xi^{(CD)}$ with

$$G^{(AB)} = \begin{bmatrix} G_1^{(A)} & 0 \\ G_2^{(A)} & 0 \\ 0 & G_2^{(B)} \end{bmatrix}, \quad K^{(AB)} = \begin{bmatrix} K_1^{(A)} & 0 \\ K_2^{(A)} & K_1^{(B)} \\ 0 & K_2^{(B)} \end{bmatrix},$$

$$\text{and} \quad \operatorname{dom}(\Xi^{(AB)}) = \left\{ \begin{bmatrix} z^{(A)} \\ z^{(B)} \end{bmatrix} \in \begin{bmatrix} \operatorname{dom}(\Xi^{(A)}) \\ \operatorname{dom}(\Xi^{(B)}) \end{bmatrix} \;\middle|\; G_2^{(A)} z^{(A)} = G_1^{(B)} z^{(B)} \right\}$$

and similarly $G^{(CD)}$, $K^{(CD)}$, and $\operatorname{dom}(\Xi^{(CD)})$.

Note that these colligations are induced by transmission graphs; for example the colligation of $\Gamma^{(AB)} := \left(\{\Xi^{(A)}, \Xi^{(B)}\}, \{\{(A,1)\}, \{(A,2),(B,1)\}, \{(B,2)\}\}, \emptyset \right)$ is exactly $\Xi^{(AB)}$.

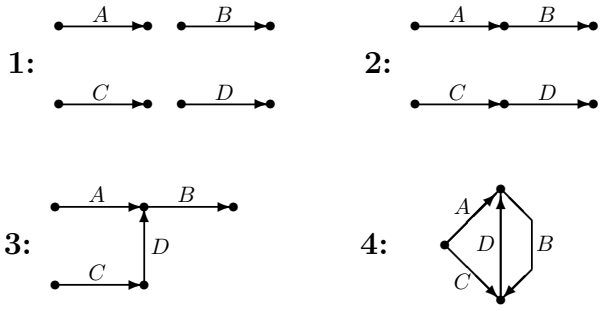

FIGURE 3. Composing a transmission graph

• *Phase 3.* Now $\Xi^{(AB)}$ is connected to $\Xi^{(CD)}$ by a partial parallel coupling. The part of the operator $G^{(AB)}$ which is not involved in the connection is $G_b^{(AB)} = \begin{bmatrix} G_1^{(A)} & 0 \\ 0 & G_2^{(B)} \end{bmatrix}$ and the part that is, is $G_c^{(AB)} = \begin{bmatrix} G_2^{(A)} & 0 \end{bmatrix}$. Correspondingly $K_b^{(AB)} = \begin{bmatrix} K_1^{(A)} & 0 \\ 0 & K_2^{(B)} \end{bmatrix}$ and $K_c^{(AB)} = \begin{bmatrix} K_2^{(A)} & K_1^{(B)} \end{bmatrix}$. The system $\Xi^{(CD)}$ is connected by its free vertex $\{(D,2)\}$ to the common vertex $\{(A,2),(B,1)\}$ of $\Xi^{(AB)}$ so the CD-splitting is done differently, namely $G_b^{(CD)} = \begin{bmatrix} G_1^{(C)} & 0 \\ 0 & G_1^{(D)} \end{bmatrix}$, $G_c^{(CD)} = \begin{bmatrix} 0 & G_2^{(D)} \end{bmatrix}$, $K_b^{(CD)} = \begin{bmatrix} K_1^{(C)} & 0 \\ K_2^{(C)} & K_1^{(D)} \end{bmatrix}$, and $K_c^{(CD)} = \begin{bmatrix} 0 & K_2^{(D)} \end{bmatrix}$.

Thus, as described in Section 4.1, we obtain a system with

$$G = \left[\begin{array}{cc|cc} G_1^{(A)} & 0 & 0 & 0 \\ 0 & G_2^{(B)} & 0 & 0 \\ \hline G_2^{(A)} & 0 & 0 & 0 \\ \hline 0 & 0 & G_1^{(C)} & 0 \\ 0 & 0 & 0 & G_1^{(D)} \end{array}\right], \quad K = \left[\begin{array}{cc|cc} K_1^{(A)} & 0 & 0 & 0 \\ 0 & K_2^{(B)} & 0 & 0 \\ \hline K_2^{(A)} & K_1^{(B)} & 0 & K_2^{(D)} \\ \hline 0 & 0 & K_1^{(C)} & 0 \\ 0 & 0 & K_2^{(C)} & K_1^{(D)} \end{array}\right],$$

and

$$\mathrm{dom}(\Xi) = \left\{ z^{(j)} \in \mathrm{dom}(\Xi^{(j)}),\ j = A,B,C,D \;\middle|\; G_2^{(A)} z^{(A)} = G_1^{(B)} z^{(B)} = G_2^{(D)} z^{(D)},\ G_2^{(C)} z^{(C)} = G_1^{(D)} z^{(D)} \right\}.$$

Again, the colligation Ξ is induced by a transmission graph $\Gamma := \left(\{\Xi^{(j)}\}_{j=A}^{D}, \{\mathcal{I}_l\}_{l=1}^{5}, \emptyset\right)$ where $\mathcal{I}_1 = \{(A,1)\}$, $\mathcal{I}_2 = \{(A,2),(B,1),(D,2)\}$, $\mathcal{I}_3 = \{(B,2)\}$, $\mathcal{I}_4 = \{(C,1)\}$, and $\mathcal{I}_5 = \{(C,2),(D,1)\}$.

• *Phase 4.* In the last phase, the vertex $\{(B,2)\}$ is connected to $\{(C,2),(D,1)\}$, and $\{(A,1)\}$ to $\{(C,1)\}$, by a loop coupling. The parts of input and output that are not involved in the connection are $G_1 = [G_2^{(A)}\ 0\ 0\ 0]$ and $K_1 = [K_2^{(A)}\ K_1^{(B)}\ 0\ K_2^{(D)}]$. The operators that are involved are $G_2 = \begin{bmatrix} G_1^{(A)} & 0 & 0\ 0 \\ 0 & G_2^{(B)} & 0\ 0 \end{bmatrix}$, $K_2 = \begin{bmatrix} K_1^{(A)} & 0 & 0\ 0 \\ 0 & K_2^{(B)} & 0\ 0 \end{bmatrix}$, $G_3 = \begin{bmatrix} 0\ 0 & G_1^{(C)} & 0 \\ 0\ 0 & 0 & G_1^{(D)} \end{bmatrix}$, and $K_3 = \begin{bmatrix} 0\ 0 & K_1^{(C)} & 0 \\ 0\ 0 & K_2^{(C)} & K_1^{(D)} \end{bmatrix}$. As described in Section 4.2, the new input and output operators are $G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}$ and $K = \begin{bmatrix} K_1 \\ K_2 + K_3 \end{bmatrix}$. To $\mathrm{dom}(\Xi)$ we impose the additional condition $G_2 z_2 = G_3 z_3$. In terms of the original blocks, this means $G_1^{(A)} z^{(A)} = G_1^{(C)} z^{(C)}$ and $G_2^{(B)} z^{(B)} = G_1^{(D)} z^{(D)}$.

In block operators G and K, before closing any vertices, each column corresponds to one system (an edge of the graph) and each row corresponds to a coupling (a vertex of the graph). Thus, in phase 2, the block operators $G^{(AB)}$, $K^{(AB)}$, $G^{(CD)}$, and $K^{(CD)}$ have three rows and two columns. In phase 3, G and K have five rows and four columns. And finally, when connecting vertex $\{(B,2)\}$ to $\{(C,2),(D,1)\}$ and $\{(A,1)\}$ to $\{(C,1)\}$, two rows are lost.

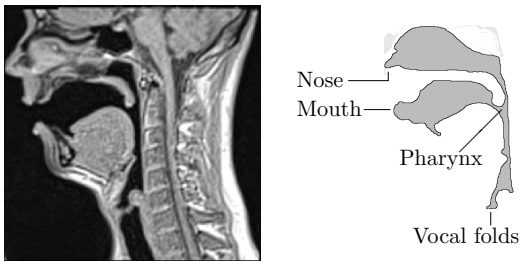

FIGURE 4. The human vocal tract and nasal cavity

5. **Webster's equation with dissipation on a graph.** An MR-image of the human vocal tract is shown in Fig. 4. The vocal tract can be considered as a Y-shaped graph whose three free vertices are at the vocal folds, mouth, and nose (in Fig. 4, the nasal cavity is only partially visible). The closed vertex with three outgoing edges is located in the pharynx. Wave propagation in such domain can be computed by Webster's equation up to frequencies of about 4 kHz where the effect of the transversal resonances becomes significant, see [11, Section 5 and Fig. 1].

The generalised Webster's equation is derived in [18], and it is given by

$$\psi_{tt}(x,t) + \frac{2\pi\theta S(x)c(x)^2}{A(x)}\psi_t(x,t) - \frac{c(x)^2}{A(x)}\frac{\partial}{\partial x}\left(A(x)\frac{\partial\psi}{\partial x}(x,t)\right) = 0. \tag{12}$$

The solution ψ is *Webster's velocity potential* that approximates the wave equation velocity potential when averaged over a transversal cross-section at distance $x \in [0,l]$ from the tube end. Functions $A(\cdot)$, $S(\cdot)$, and $c(\cdot)$ are the cross-sectional area of the tube, the surface area factor, and the corrected sound velocity, respectively. The coefficient $\theta \geq 0$ regulates the dissipation at the tube walls. The classical Webster's equation is obtained by setting $\theta = 0$ and $c(\cdot) = c$.

As explained above, the model for the vocal tract is divided into three parts. In each of these parts we have velocity potentials $\psi^{(j)} : [0,l_j] \times \mathbb{R}_+ \to \mathbb{C},\ j = A,B,C$ that satisfy (12) with respective functions $A_j \in C^1[0,l_j]$ such that $A_j(x) > \epsilon > 0$, $S_j \in L^2(0,l_j)$ such that $S_j(x) \geq 0$, and c_j such that $\infty > c_j(x) > \epsilon > 0$ and $c_j^{-2}(x) \in L^2(0,l_j)$. The potentials are connected through Kirchhoff conditions

$$\begin{cases} \frac{\partial\psi^{(A)}}{\partial t}(0,t) = \frac{\partial\psi^{(B)}}{\partial t}(0,t) = \frac{\partial\psi^{(C)}}{\partial t}(0,t), \\ A_A(0)\frac{\partial\psi^{(A)}}{\partial x}(0,t) + A_B(0)\frac{\partial\psi^{(B)}}{\partial x}(0,t) + A_C(0)\frac{\partial\psi^{(C)}}{\partial t}(0,t) = 0. \end{cases} \tag{13}$$

The system is controlled by the flow u through the vocal folds, and there is an acoustic resistance at the mouth and nose openings:

$$\begin{cases} \frac{\partial\psi^{(A)}}{\partial x}(l_A,t) = u(t) & \text{at vocal folds,} \\ \frac{\partial\psi^{(B)}}{\partial t}(l_B,t) + \theta_B c_B(l_B)\frac{\partial\psi^{(B)}}{\partial x}(l_B,t) = 0 & \text{at mouth, and} \\ \frac{\partial\psi^{(C)}}{\partial t}(l_C,t) + \theta_C c_C(l_C)\frac{\partial\psi^{(C)}}{\partial x}(l_C,t) = 0 & \text{at nose} \end{cases} \tag{14}$$

where θ_B and θ_C are the dimensionless normalised acoustic resistances.

We proceed to formulate this model as a transmission graph. First, we write Webster's equation as a first order system by choosing the state vector as $z = \left[\begin{smallmatrix} \psi \\ \psi_t \end{smallmatrix}\right]$. The state and solution spaces are

$$\mathcal{X}^{(j)} := h^1[0,l_j] \times L^2(0,l_j) \qquad \text{and} \qquad \mathcal{Z}^{(j)} := h^2[0,l_j] \times H^1[0,l_j]$$

respectively, where $h^1[0,l_j] = H^1[0,l_j]\big/\sim$ and $h^2[0,l_j] = H^2[0,l_j]\big/\sim$ where the equivalence relation $z \sim v$ holds if $z - v$ is constant Lebesgue almost everywhere in $(0,l_j)$. We equip $h^1[0,l_j]$ with the norm $\|\psi\|_{h^1[0,l_j]} := \left\|\frac{\partial \psi}{\partial x}\right\|_{L^2(0,l_j)}$, and the state spaces with inner products

$$\left\langle z, v \right\rangle_{\mathcal{X}^{(j)}} := \rho \left(\int_0^{l_j} \frac{\partial z_1}{\partial x}(x) \overline{\frac{\partial v_1}{\partial x}}(x)\ A_j(x) dx + \int_0^{l_j} z_2(x) \overline{v_2(x)}\ \frac{A_j(x)}{c_j(x)^2} dx \right)$$

where ρ is the fluid density. The induced $\mathcal{X}^{(j)}$-norm corresponds to the physical energy — the first term gives the kinetic energy of the fluid and the second term gives the potential energy (recall that acoustic pressure is obtained from the velocity potential through $p(x,t) = \rho \psi_t(x,t)$). In the solution spaces we use norms

$$\|z\|^2_{\mathcal{Z}^{(j)}} := \|z_1\|^2_{h^1[0,l_j]} + \left\| \frac{\partial^2 z_1}{\partial x^2} \right\|^2_{L^2(0,l_j)} + \|z_2\|^2_{H^1[0,l_j]}.$$

The input and output spaces are $\mathcal{U}^{(j)} = \mathcal{Y}^{(j)} = \mathbb{C}^2$ with the Euclidian norm. The interior operators are defined by

$$L^{(j)} := W^{(j)} + D^{(j)} :\ \mathcal{Z}^{(j)} \to \mathcal{X}^{(j)}$$

where

$$W^{(j)} := \begin{bmatrix} 0 & 1 \\ \frac{c_j(x)^2}{A_j(x)} \frac{\partial}{\partial x} \left(A_j(x) \frac{\partial}{\partial x} \right) & 0 \end{bmatrix} \quad \text{and} \quad D^{(j)} := \begin{bmatrix} 0 & 0 \\ 0 & -\frac{2\pi\theta S_j(x) c_j(x)^2}{A_j(x)} \end{bmatrix};$$

the dissipative part $D^{(j)}$ acts as a bounded perturbation (in $\mathcal{X}^{(j)}$) to the classical Webster-related part $W^{(j)}$. The input and output operators are defined by

$$G^{(j)} z^{(j)} := \begin{bmatrix} \rho z_2^{(j)}(0,t) \\ \rho z_2^{(j)}(l_j,t) \end{bmatrix} \quad \text{and} \quad K^{(j)} z^{(j)} := \begin{bmatrix} -A_j(0) \frac{\partial z_1^{(A)}}{\partial x}(0,t) \\ A_j(l_j) \frac{\partial z_1^{(j)}}{\partial x}(l_j,t) \end{bmatrix}.$$

The pressure controlled, velocity observed Webster's equation can finally be written in the form

$$\begin{cases} u^{(j)}(t) &= G^{(j)} z^{(j)}(t), \\ \dot{z}^{(j)}(t) &= L^{(j)} z^{(j)}(t), \\ y^{(j)}(t) &= K^{(j)} z^{(j)}(t), \quad t \in \mathbb{R}^+, \end{cases}$$

and it remains to show that each $\Xi^{(j)} = (G^{(j)}, L^{(j)}, K^{(j)})$ satisfies the conditions of Definitions 2.2 and 2.3.

Theorem 5.1. *Each colligation $\Xi^{(j)} = (G^{(j)}, L^{(j)}, K^{(j)})$ on spaces $\left(\mathbb{C}^2, \mathcal{X}^{(j)}, \mathbb{C}^2\right)$ defined above is an impedance passive (even conservative if $\theta = 0$), internally well-posed, strong boundary node.*

Proof. Here we drop the index j, and begin by showing the claim in the special impedance conservative case $\widehat{\Xi} = (G, W, K)$ on $\left(\mathbb{C}^2, \mathcal{X}, \mathbb{C}^2\right)$.

It is easy to see that $\widehat{\Xi}$ is a strong colligation, and that G is surjective. Thus, to show surjectivity of $\left[\begin{smallmatrix} G \\ \alpha - W \end{smallmatrix}\right]$ it is sufficient to show $(\alpha - W)|_{\mathcal{N}(G)}$ to be bijective.

Fix $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right] \in \mathcal{X}$ (in the following we treat $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right]$ as a representative from the equivalence class) and $\alpha \neq 0$. We wish to find $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \mathcal{N}(G)$, s.t.

$$\begin{bmatrix} \alpha & -1 \\ -\frac{c(x)^2}{A(x)} \frac{\partial}{\partial x}\left(A(x)\frac{\partial}{\partial x}\right) & \alpha \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}. \tag{15}$$

The first row implies $\alpha z_1 - z_2 = f$ (in $H^1[0,l]$). The condition $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \mathcal{N}(G)$ is equivalent to $z_2(0) = z_2(l) = 0$ so that $z_1(0) = \frac{f(0)}{\alpha}$ and $z_1(l) = \frac{f(l)}{\alpha}$. Multiplying the first row in (15) with α and adding it to the second row gives

$$\alpha^2 z_1(x) - \frac{c(x)^2}{A(x)} \frac{\partial}{\partial x}\left(A(x)\frac{\partial z_1}{\partial x}(x)\right) = \alpha f(x) + g(x) \quad \left(\in L^2(0,l)\right).$$

This equation with the aforementioned boundary conditions has a unique variational solution $z_1 \in H^2[0,l]$ that satisfies $\left[\begin{smallmatrix} z_1 \\ \alpha z_1 - f \end{smallmatrix}\right] \in \mathcal{N}(G)$. If we solve (15) for a different representative of the same equivalence class, that is, with right hand side $\left[\begin{smallmatrix} f+C \\ g \end{smallmatrix}\right]$ where $C \in \mathbb{C}$, then we get for (15) the respective solution $\left[\begin{smallmatrix} z_1 + C/\alpha \\ \alpha(z_1 + C/\alpha) - f - C \end{smallmatrix}\right] = \left[\begin{smallmatrix} z_1 + C/\alpha \\ \alpha z_1 - f \end{smallmatrix}\right]$ which is in the same equivalence class with $\left[\begin{smallmatrix} z_1 \\ \alpha z_1 - f \end{smallmatrix}\right]$. Hence, equation (15) has a unique solution in $\mathcal{Z}$ for all $\left[\begin{smallmatrix} f \\ g \end{smallmatrix}\right] \in \mathcal{X}$. The Green–Lagrange identity (6) for $\widehat{\Xi}$ as an equality can be shown by partial integration. The claim is now proved for $\widehat{\Xi}$ by Proposition 1.

Since $D : \mathcal{X} \to \mathcal{X}$ is bounded, also $L|_{\mathcal{N}(G)} = (W + D)|_{\mathcal{N}(G)}$ generates a C_0-semigroup by [2, Corollary 3.5.6]. Because $S(x) \geq 0$ and $\theta \geq 0$, it follows

$$\left\langle z, Dz \right\rangle_{\mathcal{X}} = -2\pi\theta\rho \int_0^l S(x) z_2(x)^2 \, dx \leq 0$$

which means that Green–Lagrange identity for Ξ holds as an inequality. Because bounded perturbations of closed operators are closed, nodes Ξ and $\widehat{\Xi}$ are simultaneously strong. □

The boundary conditions (13) in the pharynx correspond to conditions (9) and (10). Thus, after noting that operators $\left[\begin{smallmatrix} G^{(j)} \\ K^{(j)} \end{smallmatrix}\right]$ are surjective (try polynomial functions in $\mathcal{Z}$), Theorems 3.3 and 5.1 yield:

Theorem 5.2. *Define the transmission graph* Γ *with three control vertices and one closed vertex by*

$$\Gamma = \left(\left\{\Xi^{(A)}, \Xi^{(B)}, \Xi^{(C)}\right\}, \{\{(A,2)\}, \{(B,2)\}, \{(C,2)\}\}, \{\{(A,1),(B,1),(C,1)\}\}\right).$$

The colligation induced by Γ *is* $\Xi = (G, L, K)$ *on spaces* $\left(\mathbb{C}^3, \mathcal{X}, \mathbb{C}^3\right)$ *where*

$$G \begin{bmatrix} z^{(A)} \\ z^{(B)} \\ z^{(C)} \end{bmatrix} := \begin{bmatrix} \rho z_2^{(A)}(l_A, t) \\ \rho z_2^{(B)}(l_B, t) \\ \rho z_2^{(C)}(l_C, t) \end{bmatrix}, \qquad L := \begin{bmatrix} L^{(A)} & 0 & 0 \\ 0 & L^{(B)} & 0 \\ 0 & 0 & L^{(C)} \end{bmatrix}, \quad \textit{and}$$

$$K \begin{bmatrix} z^{(A)} \\ z^{(B)} \\ z^{(C)} \end{bmatrix} := \begin{bmatrix} A_A(l_A) \frac{\partial z_1^{(A)}}{\partial x}(l_A, t) \\ A_B(l_B) \frac{\partial z_1^{(B)}}{\partial x}(l_B, t) \\ A_C(l_C) \frac{\partial z_1^{(C)}}{\partial x}(l_C, t) \end{bmatrix}, \ \textit{with} \quad \mathcal{X} := \mathcal{X}^{(A)} \oplus \mathcal{X}^{(B)} \oplus \mathcal{X}^{(C)} \ \textit{and}$$

$$\operatorname{dom}(\Xi) := \left\{ \begin{bmatrix} z^{(A)} \\ z^{(B)} \\ z^{(C)} \end{bmatrix} \in \begin{bmatrix} \mathcal{Z}^{(A)} \\ \mathcal{Z}^{(B)} \\ \mathcal{Z}^{(C)} \end{bmatrix} \;\middle|\; z_1^{(A)}(0,t) = z_2^{(B)}(0,t) = z_2^{(C)}(0,t), \right.$$
$$\left. A_A(0)\tfrac{\partial z_1^{(A)}}{\partial x}(0,t) + A_B(0)\tfrac{\partial z_1^{(B)}}{\partial x}(0,t) + A_C(0)\tfrac{\partial z_1^{(C)}}{\partial x}(0,t) = 0 \right\}.$$

Then Ξ is an impedance passive, internally well-posed, strong boundary node. The node Ξ is conservative if and only if $\theta = 0$.

Here also the vertices corresponding to the mouth and nose are also chosen to be control vertices which does not correspond to boundary conditions (14). It can be shown that an impedance passive internally well-posed system remains as one with such resistive termination but we do not do it here.

6. **Remarks and conclusions.** Many kinds of passive boundary control systems can be interconnected with each other so that the composed system is also a passive and internally well-posed boundary control system. The presented Kirchhoff couplings are natural when connecting impedance passive systems. We remark that it is also possible to form partial couplings using the presented techniques. This is needed, *e.g.*, when beams are connected to each other by a hinge that does not transmit all the degrees of freedom between the subsystems. This can be done by splitting the input and output spaces using orthogonal projections and then treating these as independent inputs and outputs.

However, if the junctions themselves have (finite-dimensional) dynamics then these methods are not (directly) applicable — consider, for example, a hinge junction between two beams with a spring or a damper. In such case the resulting system is not necessarily of boundary control form, and instead, these systems should be treated in the more general system node setting. See the work of Weiss and Zhao [29] for this kind of ideas.

All results in this paper require the colligations to be strong in the sense of Definition 2.1. As mentioned before, there are internally well-posed boundary nodes (in the sense of [21, Definition 2.2]) that are even impedance conservative and satisfy $\mathcal{U} = \mathcal{Y}$ but are not strong. One such example is given in [21, Proposition 6.3] in terms of the boundary controlled wave equation on $\Omega \subset \mathbb{R}^n$ with smooth boundary $\partial\Omega$. However, the same PDE with the same boundary control can be written as a strong node at the cost of $\mathcal{U} \neq \mathcal{Y}$; these spaces are still a dual pair. Note that Theorem 3.3 can be applied also in this case even though the smoothness assumption on $\partial\Omega$ seriously restricts the possible couplings of this kind of systems.

Acknowledgments. We thank the anonymous reviewer for pointing out the possible simplification of the proof of our main theorem (see Remark 2).

REFERENCES

[1] A. Aalto and J. Malinen, *Wave propagation in networks: A system theoretic approach*, in "Proceedings of the 18th IFAC World Congress" (eds. S. Bittanti, A. Cenedese and S. Zampieri), (2011), 8854–8859.

[2] W. Arendt, C. Batty, M. Hieber and F. Neubrander, "Vector-valued Laplace Transforms and Cauchy Problems," Monographs in Mathematics, **96**, Birkhäuser Verlag, Basel, 2001.

[3] J. Cervera, A. J. van der Schaft and A. Baños, *Interconnection of port-Hamiltonian systems and composition of Dirac structures,* Automatica J. of IFAC, **43** (2007), 212–225.

[4] R. F. Curtain and H. Zwart, "An Introduction to Infinite-Dimensional Linear Systems Theory," Texts in Applied Mathematics, **21**, Springer-Verlag, New York, 1995.

[5] V. Derkach, S. Hassi, M. Malamud and H. de Snoo, *Boundary relations and their Weyl families,* Transactions of the American Mathematical Society, **358** (2006), 5351–5400.

[6] M. Gugat, G. Leugering, K. Schittkowski and E. J. P. Georg Schmidt, *Modelling, stabilization, and control of flow in networks of open channels,* in "Online Optimization of Large Scale Systems," Springer, Berlin, (2001), 251–270.

[7] K.-J. Engel, M. Kramar Fijavž, R. Nagel and E. Sikolya, *Vertex control of flows in networks,* Networks and Heterogeneous Media, **3** (2008), 709–722.

[8] H. Fattorini, *Boundary control systems,* SIAM Journal of Control, **6** (1968), 349–385.

[9] V. I. Gorbachuk and M. L. Gorbachuk, "Boundary Value Problems for Operator Differential Equations," Mathematics and its Applications (Soviet Series), **48**, Kluwer Academic Publishers Group, Dordrecht, 1991.

[10] G. Greiner, *Perturbing the boundary conditions of a generator,* Houston Journal of Mathematics, **13** (1987), 213–229.

[11] A. Hannukainen, T. Lukkari, J. Malinen and P. Palo, *Vowel formants from the wave equation,* Journal of Acoustical Society of America Express Letters, **122** (2007).

[12] R. Hundhammer and G. Leugering, *Instantaneous control of vibrating string networks,* in "Online Optimization of Large Scale Systems," Springer, Berlin, (2001), 229–249.

[13] P. Kuchment and H. Zeng, *Convergence of spectra of mesoscopic systems collapsing onto a graph,* Journal of Mathematical Analysis and Applications, **258** (2001), 671–700.

[14] Mikael Kurula, "Towards Input/Output-Free Modelling of Linear Infinite-Dimensional Systems in Continuous Time," Ph.D thesis, Åbo Akademi, 2010.

[15] M. Kurula, H. Zwart, A. van der Schaft and J. Behrndt, *Dirac structures and their composition on Hilbert spaces,* Journal of Mathematical Analysis and Applications, **372** (2010), 402–422.

[16] Y. Latushkin and V. Pivovarchik, *Scattering in a forked-shaped waveguide,* Integral Equations and Operator Theory, **61** (2008), 365–399.

[17] M. S. Livšic, "Operators, Oscillations, Waves (Open Systems)," Translations of Mathematical Monographs, Vol. 34, American Mathematical Society, Providence, Rhode Island, 1973.

[18] T. Lukkari and J. Malinen, *Webster's equation with curvature and dissipation,* preprint, `arXiv:1204.4075`, 2012.

[19] J. Malinen, *Conservativivity of time-flow invertible and boundary control systems,* Helsinki University of Technology Institute of Mathematics Research Reports, A479, (2004).

[20] J. Malinen and O. Staffans, *Conservative boundary control systems,* Journal of Differential Equations, **231** (2006), 290–312.

[21] J. Malinen and O. Staffans, *Impedance passive and conservative boundary control systems,* Complex Analysis and Operator Theory, **1** (2007), 279–300.

[22] J. Malinen, O. Staffans and G. Weiss, *When is a linear system conservative,* Quarterly of Applied Mathematics, **64** (2006), 61–91.

[23] J. Rubinstein and M. Schatzman, *Variational problems on multiply connected thin strips. I. Basic estimates and convergence of the Laplacian spectrum,* Archive for Rational Mechanics and Analysis, **160** (2001), 271–308.

[24] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach,* Transactions of the American Mathematical Society, **300** (1987), 383–431.

[25] D. Salamon, *Realization theory in Hilbert space,* Mathematical Systems Theory, **21** (1989), 147–164.

[26] O. Staffans, "Well-Posed Linear Systems," Encyclopedia of Mathematics and its Applications, **103**, Cambridge University Press, Cambridge, 2005.

[27] Javier Villegas, "A Port-Hamiltonian Approach to Distributed Parameter Systems," Ph.D thesis, University of Twente, 2007.

[28] G. Weiss, *Regular linear systems with feedback,* Mathematics of Control, Signals, and Systems, **7** (1994), 23–57.

[29] G. Weiss and X. Zhao, *Well-posedness and controllability of a class of coupled linear systems,* SIAM Journal of Control and Optimization, **48** (2009), 2719–2750.

[30] H. Zwart, Y. Le Gorrec, B. Maschke and J. Villegas, *Well-posedness and regularity of hyperbolic boundary control systems on a one-dimensional spatial domain,* ESAIM: Control, Optimisation and Calculus of Variations, **16** (2010), 1077–1093.

Received December 2011; revised May 2012.

E-mail address: atte.aalto@aalto.fi
E-mail address: jarmo.malinen@aalto.fi

Publication II

A. Aalto. Convergence of discrete time Kalman filter estimate to continuous time estimate. http://arxiv.org/abs/1408.1275, 21 pages, August 2014.

CONVERGENCE OF DISCRETE TIME KALMAN FILTER ESTIMATE TO CONTINUOUS TIME ESTIMATE

ATTE AALTO

Department of Mathematics and Systems Analysis,
Aalto University School of Science

Abstract. This article is concerned with the convergence of the state estimate obtained from the discrete time Kalman filter to the continuous time estimate as the temporal discretization is refined. We derive convergence rate estimates for different systems, first finite dimensional and then infinite dimensional with bounded or unbounded observation operators. Finally, we derive the convergence rate in the case where the system dynamics is governed by an analytic semigroup. The proofs are based on applying the discrete time Kalman filter on a dense numerable subset of a certain time interval $[0, T]$.

1. Introduction

It is well known that Kalman filter gives the optimal solution to the state estimation problem for discrete time linear systems with Gaussian initial state, and Gaussian input and output noise processes. The continuous time estimator is generally known as the Kalman–Bucy filter. These filters have proven to be very robust and so they have been widely used in practical applications since their introduction in the 1960s. The implementation is straightforward since especially the discrete time filter is readily formulated in an algorithmic manner. Thus, it may often be tempting to use the discrete time filter on the temporally discretized continuous time system. The purpose of this article is to study the convergence of a state estimate from discrete time Kalman filter to the continuous time state estimate as the temporal discretization is refined. In particular, we show convergence speed estimates for the quadratic error between the discrete time and continuous time estimate first for finite dimensional systems, then for infinite dimensional systems with a bounded observation operator, and finally, for systems with unbounded observation operator.

The class of systems studied here is described by a pair of mappings $(A, C) : \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ and the corresponding dynamics equations

$$\begin{cases} \frac{d}{dt} z(t) = Az(t), & t \in \mathbb{R}^+, \\ z(0) = x, \\ dy(t) = Cz(t)\, dt + dw(t). \end{cases} \tag{1}$$

2010 *Mathematics Subject Classification.* 93E11, 93C57.

Key words and phrases. Kalman filter, infinite dimensional systems, temporal discretization, sampled data.

Here $\mathcal{X}$ is called the *state space* and $\mathcal{Y} = \mathbb{R}^q$ is the *output space*. The mapping A is the generator of a contractive C_0-semigroup e^{At} on $\mathcal{X}$ with domain $\mathcal{D}(A)$ and $C : \mathcal{X} \to \mathbb{R}^q$ is called the *observation operator*. The observation operator can be bounded or not but it always maps to a finite dimensional space in this article. The process y is called the *output process*. The *output noise process* w is assumed to be q-dimensional Brownian motion with incremental covariance matrix $R > 0$ and the *initial state* x is assumed to be an $\mathcal{X}$-valued Gaussian random variable.

The discrete and continuous time state estimates are defined by

$$\hat{x}_{T,n} := \mathbb{E}\Big(x \,\Big|\, \big\{y\big(\tfrac{iT}{n}\big)\big\}_{i=1}^{n}\Big) \quad \text{and} \quad \hat{x}(T) := \mathbb{E}\big(x \,\big|\big\{y(s), s \le T\big\}\big), \tag{2}$$

respectively. That is, we are estimating the initial state of the system (1). In the absence of system input (or input noise — deterministic input can be removed by the usual techniques) it holds that $\mathbb{E}(z(t)|\sigma) = e^{At}\mathbb{E}(x|\sigma)$. These estimates are given by the discrete and continuous time Kalman filter, respectively — given that the continuous time Kalman filter equations are solvable. The purpose of this article is to study the convergence $\hat{x}_{T,n} \to \hat{x}(T)$ as $n \to \infty$.

In Section 2, we cover the necessary background concerning stochastics and the Kalman filter. In particular, in Section 2.1, it is shown that $\hat{x}_{T,n} \to \hat{x}(T)$ strongly in $\mathcal{X}$ almost surely. Gaussian random variables and the Kalman filter are introduced in Section 2.2. In Section 2.3 we show how to take into account an intermediate measurement in Kalman filtering — an important tool in the article. Section 3 contains the main results of this article, namely estimates of the convergence speed of $\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big)$ when n is increased first for finite dimensional systems (Thm. 3.1) and then for infinite dimensional systems with bounded (Thms. 3.3 and 3.4) and then with unbounded observation operator C (Thm. 3.5) when A is diagonalizable. In this case we have to make an assumption on the spectral asymptotics of A and pose a slight restriction on how badly behaving C can be. The case where A is not diagonalizable but satisfies a well-posedness condition is treated in Thm. 3.7. Finally, we show two convergence rate results if A is the generator of an analytic semigroup (Thms. 3.8 and 3.9).

The Kalman filter performance has been widely studied in literature. Even though it was originally derived for state estimation for finite dimensional linear systems with Gaussian input and output noise processes it has proven to be very robust and thus applicable to a variety of other scenarios. Variants for non-linear systems have been developed, such as the extended Kalman filter and the unscented Kalman filter, see the book [19] by Simon. Kalman filter sensitivity to modelling errors has been studied by for example Sun in [21] and Gelb in [8: Chapter 7]. See also the recent work [13] by Lee *et al.* for a study on the effect of modelling errors in an infinite dimensional example case, namely the one dimensional wave equation. The effect of state space discretization to Kalman filtering has been studied in, *e.g.*, [9] by Germani *et al.* and in [1] by Aalto.

However, the error that stems from using the discrete time filter on the temporally discretized continuous time system has not received much attention. Two recent articles, [3] by Axelsson and Gustafsson and [23] by Wahlström *et al.*, have studied different numerical methods for approximating the matrix exponential $e^{A\Delta t}$ and the effect of this approximation on the solution of the corresponding Lyapunov equations and Kalman filtering. A convergence result of the discrete time Kalman

filter estimate in finite dimensional setting is shown by Salgado *et al.* in [18] without convergence rate estimate. They use similar techniques that can also be used to (formally) obtain the Kalman-Bucy filter as a limit of the discrete time Kalman filter, as is done for example in [19: Section 8.2] and [8: Section 4.3].

Notation and standing assumptions.

- The space of bounded operators from a Hilbert space $\mathcal{H}_1$ to another Hilbert space $\mathcal{H}_2$ is denoted by $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$, and $\mathcal{L}(\mathcal{H}_1) = \mathcal{L}(\mathcal{H}_1, \mathcal{H}_1)$.
- We assume that the state space $\mathcal{X}$ is a separable Hilbert space. Denote by $\{e_k\}_{k=1}^{p/\infty} \subset \mathcal{X}$ an orthonormal basis for the p/∞-dimensional state space.
- A is the generator of a contractive C_0-semigroup on $\mathcal{X}$. The semigroup is denoted by e^{At} even though A is not bounded in general.
- The space $\mathcal{D}(A)$ is equipped with the graph norm $\|x\|^2_{\mathcal{D}(A)} = \|x\|^2_{\mathcal{X}} + \|Ax\|^2_{\mathcal{X}}$ which makes $\mathcal{D}(A)$ a Hilbert space since A is closed.
- $C \in \mathcal{L}(\mathcal{D}(A), \mathcal{Y})$. This is a minimal assumption, and sometimes we assume more. The output space is always finite dimensional, $\mathcal{Y} = \mathbb{R}^q$.
- Ω is a probability space and $L^2(\Omega; \mathcal{X})$ is the space of $\mathcal{X}$-valued random variables x satisfying $\mathbb{E}\left(\|x\|^2_{\mathcal{X}}\right) < \infty$.
- The sigma algebra generated by a random variable h is denoted by $\sigma\{h\}$.
- To improve readability, we use index n only when referring to the discretization level in the state estimate $\hat{x}_{T,n}$ defined in (2), index k only to denote different dimensions of the state space, and index j only when referring to the martingale $\tilde{x}_j$ defined below in Section 2.1.

2. Background and preliminary results

As mentioned above, the proofs of this article are based on applying the discrete time Kalman filter on a dense, numerable subset on the interval $[0, T]$ — starting from the discrete time state estimate $\hat{x}_{T,n}$ — and computing an upper bound for the change in the estimate. In section 2.1, we establish that the limit thus obtained is indeed $\hat{x}(T)$. Gaussian random variables and the Kalman filter are discussed in Section 2.2. In Section 2.3 it is shown how an intermediate observation is taken into account in the state estimate.

2.1. **Stochastics.** In the cases where the state space $\mathcal{X}$ is infinite dimensional it is always assumed either that $x \in \mathcal{D}(A)$ almost surely or that $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. This guarantees that the stochastic process y given by (1) has almost surely continuous sample paths. Let $\{t_i\}_{i=1}^{\infty}$ be a dense subset of the interval $[0, T]$ and denote $\mathrm{T}_j := \{t_i\}_{i=1}^{j}$. Now let x be an integrable $\mathcal{X}$-valued random variable and y a stochastic process with almost surely continuous sample paths. Then $[x]_k := \langle x, e_k \rangle_{\mathcal{X}}$ is an integrable $\mathbb{R}$-valued random variable for each k. Define the martingales $[\tilde{x}_j]_k := \mathbb{E}(\langle x, e_k \rangle_{\mathcal{X}} | \mathcal{F}_j)$ where $\mathcal{F}_j$ is the sigma algebra generated by $\{y(t), t \in \mathrm{T}_j\}$, that is, $\mathcal{F}_j = \sigma\{y(t), t \in \mathrm{T}_j\}$. It holds that $\mathbb{E}(|[\tilde{x}_j]_k|) \leq \mathbb{E}(|\langle x, e_k \rangle_{\mathcal{X}}|)$ for all j and thus by Doob's Martingale convergence theorem (see [16: Appendix C], in particular, Theorem C.6 and Corollary C.9), $[\tilde{x}_j]_k \to [\tilde{x}_\infty]_k$ almost surely. As y has continuous sample paths, it holds that $[\tilde{x}_\infty]_k = \mathbb{E}(\langle x, e_k \rangle_{\mathcal{X}} | \{y(s), s \leq T\})$ almost surely. Using this componentwise implies that $\tilde{x}_j := \mathbb{E}(x|\mathcal{F}_j) = \sum_{k=1}^{\infty} [\tilde{x}_j]_k e_k$ converges strongly (in $\mathcal{X}$) almost surely to $\tilde{x}_\infty = \sum_{k=1}^{\infty} [\tilde{x}_\infty]_k e_k$.

Below we sometimes need the assumption that $x \in \mathcal{D}(A)$ almost surely. With Gaussian random variables this means that x is actually a $\mathcal{D}(A)$-valued random variable.

Proposition 2.1. *Let z be an $\mathcal{X}$-valued Gaussian random variable s.t. $z \in \mathcal{X}_1$ almost surely where $\mathcal{X}_1 \subset \mathcal{X}$ is another Hilbert space with continuous and dense embedding. Then z is an $\mathcal{X}_1$-valued Gaussian random variable.*

Proof. Pick $h \in \mathcal{X}_1$. We intend to show that $\langle z, h\rangle_{\mathcal{X}_1}$ is a real-valued Gaussian random variable. For $h \in \mathcal{X}_1$ there exists $x \in \mathcal{X}_1'$, the dual space of $\mathcal{X}_1$, s.t. $\langle z, h\rangle_{\mathcal{X}_1} = \langle z, x\rangle_{(\mathcal{X}_1, \mathcal{X}_1')}$ and further, there exists a sequence $\{x_i\}_{i=1}^{\infty} \subset \mathcal{X}$ such that $\langle z, x\rangle_{(\mathcal{X}_1, \mathcal{X}_1')} = \lim_{i\to\infty} \langle z, x_i\rangle_{\mathcal{X}}$. Now $\langle z, x_i\rangle_{\mathcal{X}}$ is a pointwise converging sequence of Gaussian random variables and so the limit is also Gaussian. □

Fernique's theorem [6: Theorem 2.6] can be applied to note that if x is an $\mathcal{X}_1$-valued Gaussian random variable then $x \in L^p(\Omega; \mathcal{X}_1)$ for any $p > 0$. In particular, $\mathbb{E}\Big(\|x\|_{\mathcal{X}_1}^2\Big) < \infty$ and if $C \in \mathcal{L}(\mathcal{X}_1, \mathcal{Y})$ then Cx is a $\mathcal{Y}$-valued Gaussian random variable.

2.2. **Kalman filter.** The discrete time Kalman filter was originally presented in [11]. The continuous time filter is known as the Kalman–Bucy filter, and it was presented in [12]. We also refer to the book [8] by Gelb for a thorough introduction to both discrete and continuous time Kalman filters as well as the usual techniques needed in different scenarios. Of course, the original presentations are in finite dimensional setting. The infinite dimensional generalization of the discrete time Kalman filter is rather straightforward, and it can be found for example in [10] by Horowitz. The infinite dimensional Kalman–Bucy filter is considered by Curtain and Pritchard in [4: Chapter 6]. For what comes to the continuous time filter in infinite dimensions, care must be taken to make sure that the crucial operator-valued error covariance equation is solvable. This problem is considered for example by Flandoli in [7] and Da Prato and Ichikawa in [5] in the case of an analytic semigroup with unbounded control and observation operators. In our proofs, we do not even need to be concerned with the solvability of the continuous time equations. Our approach is based on using the discrete time Kalman filter on a numerable set $\{t_j\}$ that is dense on an interval $[0, T]$, and showing that this state estimate converges. In this section we thus review the discrete time Kalman filter equations.

The Kalman filter is based on the fact that with linear systems with Gaussian initial state and input and output noise processes, the state vector remains a Gaussian stochastic process. Also, the conditional expectation of the state with respect to the measurements is a Gaussian process. The statistical properties of the Gaussian $\mathcal{X}$-valued random variable x are completely characterized by the mean $m = \mathbb{E}(x) \in \mathcal{X}$ and the covariance operator $P = \operatorname{Cov}[x, x] \in \mathcal{L}(\mathcal{X})$, defined for $h \in \mathcal{X}$ by $\operatorname{Cov}[x, x]\, h := \mathbb{E}((x - m)\langle x - m, h\rangle_{\mathcal{X}})$. Thus it is meaningful to write $x \sim N(m, P)$ meaning that x is a Gaussian random variable with mean m and covariance P. The covariance operator is symmetric and nonnegative and, in addition, it is a trace class operator with $\operatorname{tr}(P) = \mathbb{E}\Big(\|x - m\|_{\mathcal{X}}^2\Big)$, see [6: Lemma 2.14 & Proposition 2.15]. In fact, by Fernique's theorem, Gaussian random variables are p-integrable for every $p > 0$.

For square integrable random variables, the conditional expectation with respect to a random variable h is a projection onto the subspace generated by h. With

jointly Gaussian random variables $h_1 \in \mathcal{X}$ and finite dimensional h_2, this projection has an easy representation. That is, if $h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \sim N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^* & P_{22} \end{bmatrix} \right)$ then

$$\mathbb{E}(h_1|h_2) = m_1 + P_{12}P_{22}^+(h_2 - m_2)$$

where P_{22}^+ denotes the (Moore-Penrose) pseudoinverse of P_{22}. The error covariance is

$$\operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right] = P_{11} - P_{12}P_{22}^+P_{12}^*.$$

Now applying the above equations to a Gaussian random variable $[h_1, h_2, h_3]$ where h_2 and h_3 are finite dimensional, and the 2-by-2 blockwise matrix inversion formula to $\operatorname{Cov}\left[\begin{bmatrix} h_2 \\ h_3 \end{bmatrix}, \begin{bmatrix} h_2 \\ h_3 \end{bmatrix}\right]$ leads directly to

$$\begin{aligned} \mathbb{E}(h_1|[h_2, h_3]) =& \mathbb{E}(h_1|h_2) + \operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right] \times \\ & \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right]^+ \left(h_3 - \mathbb{E}(h_3|h_2)\right) \end{aligned} \tag{3}$$

and

$$\begin{aligned} &\operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|[h_2, h_3])\,, h_1 - \mathbb{E}(h_1|[h_2, h_3])\right] \\ &= \operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right] - \operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right] \\ &\qquad \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right]^+ \\ &\qquad \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right]. \end{aligned} \tag{4}$$

These equations make it possible to update the state estimate (here $\mathbb{E}(h_1|h_2)$) recursively when a new measurement (here h_3) is obtained from the system.

From (3) we get the covariance for the increment $\mathbb{E}(h_1|[h_2, h_3]) - \mathbb{E}(h_1|h_2)$,

$$\begin{aligned} &\operatorname{Cov}\left[\mathbb{E}(h_1|[h_2, h_3]) - \mathbb{E}(h_1|h_2)\,, \mathbb{E}(h_1|[h_2, h_3]) - \mathbb{E}(h_1|h_2)\right] \\ &= \operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right] \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right]^+ \\ &\qquad \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right], \end{aligned}$$

and further,

$$\begin{aligned} &\mathbb{E}\Big(\|\mathbb{E}(h_1|[h_2, h_3]) - \mathbb{E}(h_1|h_2)\|_{\mathcal{X}}^2\Big) \\ &= \operatorname{tr}\Big(\operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right] \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right]^+ \\ &\qquad \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right]\Big), \end{aligned}$$

that is, the squared $L^2(\Omega; \mathcal{X})$-norm of the change in the state estimate is the trace of the change in the error covariance. This fact will be used multiple times in the proofs below.

The familiar discrete time Kalman filter equations follow directly from (3) and (4) if h_1 is chosen to be the current state x_i that is to be estimated, h_2 consists of the old outputs $[y_1, \ldots, y_{i-1}]$, and h_3 is the new output y_i.

2.3. **Intermediate observations.** The convergence rate estimates are based on computing how much $\hat{x}_{T,n}$ can change at most (measured with the $L^2(\Omega; \mathcal{X})$-norm) when more and more output values $y(t)$ are taken into account from the intervals $t \in ((i-1)T/n, iT/n)$ for $i = 1, \ldots, n$. In this section, it is shown how an intermediate

measurement is taken into account. Consider the output of the system (1), $dy(t) = Ce^{At}x\,dt + dw(t)$, which is a shortened notation for

$$y(t) = C\int_0^t e^{As}x\,ds + w(t) \tag{5}$$

where A and C are (possibly unbounded) operators from $\mathcal{X}$ to $\mathcal{X}$ and $\mathcal{Y} = \mathbb{R}^q$, respectively, and w is a Brownian motion with incremental covariance matrix R.

Assume we have a state estimate $\tilde{x}_j := \mathbb{E}(x|\{y(t_1), y(t_2), \ldots, y(t_j)\})$ for the initial state x, and the corresponding error covariance $P_j := \mathrm{Cov}\,[x - \tilde{x}_j, x - \tilde{x}_j]$. Now the next measurement to be taken into account in state estimation is $y(t_{j+1})$. Say $t_{j+1} \in (t_a, t_b)$ for some $a, b \in \{1, \ldots, j\}$ and that this interval does not contain any earlier included measurements, that is $t_i \notin (t_a, t_b)$ for $i = 1, \ldots, j$. The new state estimate $\tilde{x}_{j+1}$ and the corresponding error covariance $P_{j+1} := \mathrm{Cov}\,[x - \tilde{x}_{j+1}, x - \tilde{x}_{j+1}]$ are given by (3) and (4), respectively, if we set $h_1 = x$, $h_2 = [y(t_1), y(t_2), \ldots, y(t_j)]$, and $h_3 = y(t_{j+1})$.

To get a simple representation for the covariances in (3) and (4), define a new output

$$\tilde{y} := y(t_{j+1}) - \frac{t_b - t_{j+1}}{t_b - t_a}y(t_a) - \frac{t_{j+1} - t_a}{t_b - t_a}y(t_b).$$

That is, $\tilde{y}$ is $y(t_{j+1})$ from which the linear interpolant between $y(t_a)$ and $y(t_b)$ has been removed. By plugging (5) here, this can be written in the form $\tilde{y} = \tilde{C}x + \tilde{w}$ where

$$\begin{aligned}\tilde{C} &= C\int_0^{t_{j+1}} e^{As}\,ds - C\frac{t_b - t_{j+1}}{t_b - t_a}\int_0^{t_a} e^{As}\,ds - \frac{t_{j+1} - t_a}{t_b - t_a}\int_0^{t_b} e^{As}\,ds \\ &= C\left(\frac{t_b - t_{j+1}}{t_b - t_a}\int_{t_a}^{t_{j+1}} e^{As}\,ds - \frac{t_{j+1} - t_a}{t_b - t_a}\int_{t_{j+1}}^{t_b} e^{As}\,ds\right)\end{aligned}$$

and

$$\tilde{w} = w(t_{j+1}) - \frac{t_b - t_{j+1}}{t_b - t_a}w(t_a) - \frac{t_{j+1} - t_a}{t_b - t_a}w(t_b).$$

Since w is Brownian motion, it holds that $\tilde{w} \sim N\left(0, \frac{(t_{j+1}-t_a)(t_b-t_{j+1})}{t_b-t_a}R\right)$ and $\tilde{w}$ is independent of the already included measurements (that is, of h_2) and hence of $\tilde{x}_j$, as well. Thus $\mathbb{E}(\tilde{y}|h_2) = \tilde{C}\tilde{x}_j$,

$$\mathrm{Cov}\left[x - \tilde{x}_j, \tilde{y} - \tilde{C}\tilde{x}_j\right] = P\tilde{C}^*,$$

and

$$\mathrm{Cov}\left[\tilde{y} - \tilde{C}\tilde{x}_j, \tilde{y} - \tilde{C}\tilde{x}_j\right] = \tilde{C}P\tilde{C}^* + \frac{(t_{j+1} - t_a)(t_b - t_{j+1})}{t_b - t_a}R.$$

By (3), the new estimate $\tilde{x}_{j+1} := \mathbb{E}(x|\{y(t_1), y(t_2), \ldots, y(t_{j+1})\})$ is given by

$$\tilde{x}_{j+1} = \tilde{x}_j + P_j\tilde{C}^*\left(\tilde{C}P_j\tilde{C}^* + \frac{(t_{j+1} - t_a)(t_b - t_{j+1})}{t_b - t_a}R\right)^{-1}\left(\tilde{y} - \tilde{C}\tilde{x}_j\right) \tag{6}$$

and by (4), the new error covariance $P_{j+1} := \mathrm{Cov}\,[x - \tilde{x}_{j+1}, x - \tilde{x}_{j+1}]$ by

$$P_{j+1} = P_j - P_j\tilde{C}^*\left(\tilde{C}P_j\tilde{C}^* + \frac{(t_{j+1} - t_a)(t_b - t_{j+1})}{t_b - t_a}R\right)^{-1}\tilde{C}P_j. \tag{7}$$

This will be used with $t_b - t_{j+1} = t_{j+1} - t_a = h$, and we define

$$C_h(t)x := \frac{C}{2}\left(\int_{t-h}^{t} e^{As}x\,ds - \int_{t}^{t+h} e^{As}x\,ds\right), \qquad \text{for } t \geq h > 0. \tag{8}$$

Lemma 2.2. *If $C \in \mathcal{L}(\mathcal{D}(A), \mathcal{Y})$ then $C_h(t) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. If $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ then it holds that*

(i) $\|C_h(t)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \leq h\,\|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$ *and*

(ii) $\|C_h(t)\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})} \leq \dfrac{h^2}{2}\,\|A\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{X})}\,\|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$.

In finite dimensional case $\|A\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{X})}$ means plainly the matrix norm of A. In infinite dimensional case $\|A\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{X})} = 1$ because $\mathcal{D}(A)$ is equipped with the graph norm of A.

This could also be shown for more general $\tilde{C}$ with $t_b - t_a$ replacing h in *(i)* and $\frac{(t_{j+1}-t_a)^2}{2} + \frac{(t_b-t_{j+1})^2}{2}$ replacing h^2 in *(ii)* but that is not needed. Also, part *(ii)* can be made a bit better. In fact, $\|C_h(t)x\|_{\mathcal{Y}} \leq \frac{h^2}{2}\,\|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\,\|Ax\|_{\mathcal{X}}$.

Proof. First assume just $C \in \mathcal{L}(\mathcal{D}(A), \mathcal{Y})$. If $x \in \mathcal{X}$ then $\int_t^{t+h} e^{As}x\,ds \in \mathcal{D}(A)$ since

$$\begin{aligned}\left\|\int_t^{t+h} e^{As}x\,ds\right\|_{\mathcal{D}(A)}^2 &= \left\|\int_t^{t+h} e^{As}x\,ds\right\|_{\mathcal{X}}^2 + \left\|A\int_t^{t+h} e^{As}x\,ds\right\|_{\mathcal{X}}^2 \\ &\leq h^2\,\|x\|_{\mathcal{X}}^2 + \left\|\left(e^{A(t+h)} - e^{At}\right)x\right\|_{\mathcal{X}}^2 \leq \left(h^2+4\right)\|x\|_{\mathcal{X}}^2.\end{aligned}$$

Then $\|C_h(t)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \leq \sqrt{h^2+4}\,\|C\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})}$.

Then assume $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Part *(i)* of the Lemma is clear from the definition (8) since e^{At} is contractive. For part *(ii)*, note that $Ce^{At}x \in C^1(\mathbb{R}^+; \mathcal{Y})$ with $\frac{d}{dt}Ce^{At}x = CAe^{At}x$ and $\left\|CAe^{At}x\right\|_{\mathcal{Y}} \leq \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\,\|A\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{X})}\,\|x\|_{\mathcal{D}(A)}$. Then by Bochner integral properties, C can be taken inside the integral and thus

$$\begin{aligned}&\int_{t-h}^{t} Ce^{As}x\,ds - \int_t^{t+h} Ce^{As}x\,ds \\ &= \int_{t-h}^{t}\left(Ce^{At}x - \int_s^t CAe^{Ar}x\,dr\right)ds - \int_t^{t+h}\left(Ce^{At}x + \int_t^s CAe^{Ar}x\,dr\right)ds \\ &= -\int_{t-h}^{t}\int_s^t CAe^{Ar}x\,dr\,ds - \int_t^{t+h}\int_t^s CAe^{Ar}x\,dr\,ds.\end{aligned}$$

This together with the bound for $\left\|CAe^{At}x\right\|_{\mathcal{Y}}$ imply *(ii)*. □

3. Convergence results

3.1. **Finite dimensional systems.** We begin by showing a convergence rate estimate in the case of a finite dimensional system. This result could be obtained as a special case of Thm. 3.4 below since $x \in \mathcal{D}(A)$ holds trivially in the finite dimensional state space. However, the proofs of all cases follow the same outline and in order to convey the idea of the proofs as clearly as possible, we give a complete proof of the simplest, finite dimensional case.

Theorem 3.1. *Let now $\mathcal{X} = \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$ and $C \in \mathbb{R}^{q\times p}$ (with $q \leq p$) and let $\hat{x}_{T,n}$ and $\hat{x}(T)$ be as defined above in* (2)*. Then*

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^3}{n^2}$$

where $M = \frac{\mathrm{tr}(P_n)^2\|C\|^2\|A\|^2}{6\min(\mathrm{eig}(R))}$ and $P_n = \mathrm{Cov}\,[x - \hat{x}_{T,n}, x - \hat{x}_{T,n}]$.

The constant M depends on n through P_n which is the error covariance of the discrete time state estimate $\hat{x}_{T,n}$. It holds that $P_n \leq P_0$ and so $\mathrm{tr}(P_n) \leq \mathrm{tr}(P_0)$. So a strict *a priori* result is obtained if P_n is replaced by P_0 in M.

Proof. The outline of the proof is as follows. First, we define the martingale $\tilde{x}_j$ as in Section 2.1. That is, $\tilde{x}_j = \mathbb{E}(x|\mathcal{F}_j)$, where $\mathcal{F}_j = \sigma\{y(t), t \in \mathrm{T}_j\}$ and $\mathrm{T}_j = \{t_i\}_{i=1}^j$. The martingale is Gaussian and hence square integrable, and so we have the following telescope identity for $L, N \in \mathbb{N}$ with $L \geq N$:

$$\mathbb{E}\Big(\|\tilde{x}_L - \tilde{x}_N\|_{\mathcal{X}}^2\Big) = \sum_{j=N}^{L-1} \mathbb{E}\Big(\|\tilde{x}_{j+1} - \tilde{x}_j\|_{\mathcal{X}}^2\Big). \tag{9}$$

Second, we find an upper bound for $\mathbb{E}\Big(\|\tilde{x}_{j+1} - \tilde{x}_j\|_{\mathcal{X}}^2\Big)$ using the results of Section 2.3. Third, we prove that the sum in (9) converges as $L \to \infty$ and thus $\tilde{x}_j$ is a Cauchy sequence in $L^2(\Omega;\mathcal{X})$. It has a limit in this space by completeness and the limit must be $\hat{x}(T)$ by the considerations in Section 2.1. Also, setting $N = n$ (we have $\tilde{x}_n = \hat{x}_{T,n}$) and letting $L \to \infty$ in (9) gives $\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big)$.

(I) Martingale $\tilde{x}_j$: Let $t_i = iT/n$ for $i = 1, \ldots, n$. Then $\tilde{x}_j$ for $j = 1, \ldots, n$ are the state estimates from the discrete time Kalman filter and, in particular, $\tilde{x}_n = \hat{x}_{T,n}$ defined in (2). The idea is to then halve the intervals $((i-1)T/n, iT/n)$ for $i = 1, \ldots, n$ between the already included measurements. That is, we include n measurements $y\left(\frac{(i-1/2)T}{n}\right)$ for $i = 1, \ldots, n$. Then we halve the new intervals $((i-1)T/2n, iT/2n)$ for $i = 1, \ldots, 2n$ by including $2n$ measurements $y\left(\frac{(i-1/2)T}{2n}\right)$ for $i = 1, \ldots, 2n$ and so on.

(II) Increment $\tilde{x}_{j+1} - \tilde{x}_j$: Assume that the current state estimate is $\tilde{x}_j$ with $j \geq n$, the corresponding error covariance matrix is P_j, and the next measurement being included is $y\left(iT/n - \frac{2l-1}{2^K}T/n\right)$ with some $i \in \{1, \ldots, n\}$, $K \in \mathbb{N}$, and $l \in \{1, \ldots, 2^{K-1}\}$. The new state estimate $\tilde{x}_{j+1}$ is then given by (6) with $\tilde{C} = C_h\left(iT/n - \frac{2l-1}{2^K}T/n\right)$ — denoted below simply by C_h — and $h = \frac{T}{2^K n}$. We are only interested in the $L^2(\Omega;\mathcal{X})$-norm of the increment, and as discussed in Section 2.2, it is obtained from the covariance increment given in (7):

$$\mathbb{E}\Big(\|\tilde{x}_{j+1} - \tilde{x}_j\|_{\mathcal{X}}^2\Big) = \mathrm{tr}\left(P_j C_h^*(C_h P_j C_h^* + h/4\, R)^{-1} C_h P_j\right).$$

Now we wish to establish a bound for this trace. To this end, recall that the norm of the inverse of a positive definite matrix is $\left\|Q^{-1}\right\| = \frac{1}{\min(\mathrm{eig}(Q))}$, and thus,

$$\left\|\left(C_h P_j C_h^* + \frac{h}{4}R\right)^{-1}\right\| \leq \frac{4}{h\min(\mathrm{eig}(R))} =: \frac{C_R}{h}. \tag{10}$$

Using this and part *(ii)* of Lemma 2.2 gives

$$\begin{aligned}
\operatorname{tr}\left(P_jC_h^*\left(C_hP_jC_h^*+\frac{h}{4}R\right)^{-1}C_hP_j\right) &= \sum_{k=1}^{p}\left\langle C_hP_je_k,\left(C_hP_jC_h^*+\frac{h}{4}R\right)^{-1}C_hP_je_k\right\rangle \\
&\le \frac{C_R}{h}\sum_{k=1}^{p}\|C_hP_je_k\|_{\mathcal{Y}}^2 \\
&= \frac{C_R}{h}\sum_{k=1}^{p}\left\|\mathbb{E}\big(C_h(\tilde{x}_j-x)\,\langle\tilde{x}_j-x,e_k\rangle_{\mathcal{X}}\big)\right\|_{\mathcal{Y}}^2 \\
&\le \frac{C_R}{h}\mathbb{E}\Big(\|C_h(\tilde{x}_j-x)\|_{\mathcal{Y}}^2\Big)\sum_{k=1}^{p}\mathbb{E}\Big(\langle\tilde{x}_j-x,e_k\rangle_{\mathcal{X}}^2\Big) \\
&\le \frac{C_R}{h}\operatorname{tr}(C_hP_jC_h^*)\operatorname{tr}(P_j) \qquad (11)\\
&\le \frac{h^3}{\min(\operatorname{eig}(R))}\|C\|^2\|A\|^2\operatorname{tr}(P_j)^2. \qquad (12)
\end{aligned}$$

(III) Convergence: It holds that $\operatorname{tr}(P_j)\le\operatorname{tr}(P_n)$. In part (II) of the proof we had $h=2^{-K}T/n$ and that bound is used for all $2^{K-1}n$ new measurements corresponding to this h. Finally, setting $N=n$ and $L\to\infty$ in (9) and using (12) to bound the terms of the sum yields

$$\mathbb{E}\Big(\|\hat{x}_{T,n}-\hat{x}(T)\|_{\mathcal{X}}^2\Big)\le\sum_{K=1}^{\infty}2^{K-1}n\left(\frac{T}{2^Kn}\right)^3\frac{\operatorname{tr}(P_n)^2\|C\|^2\|A\|^2}{\min(\operatorname{eig}(R))}=\frac{\operatorname{tr}(P_n)^2\|C\|^2\|A\|^2T^3}{6\min(\operatorname{eig}(R))n^2}$$

completing the proof. □

3.2. **Infinite dimensional systems with bounded C.** We move on to infinite dimensional state space $\mathcal{X}$. Compared to the finite dimensional case, the main difficulty arises from that the bound for C_h in part *(ii)* of Lemma 2.2 utilizes the differentiability of $Ce^{At}x$ and thus it holds for $x\in\mathcal{D}(A)$. A natural assumption that would make it possible to use this bound is that x is a $\mathcal{D}(A)$-valued random variable. This is exactly what is done in Theorem 3.4. Before that, in Theorem 3.3 we shall see, however, that a reasonable convergence estimate can be obtained with slightly less smooth initial state x. Before tackling this problem, we present an example illuminating the necessity of some additional assumptions.

Example 3.2. This example shows that there is a system with $C\in\mathcal{L}(\mathcal{X},\mathbb{R})$ such that $\mathbb{E}\Big(\|\hat{x}_{T,n}-\hat{x}(T)\|_{\mathcal{X}}^2\Big)$ converges arbitrarily slowly where $\hat{x}_{T,n}$ and $\hat{x}(T)$ are defined in (2). Consider the one-dimensional wave equation with augmented state vector,

$$\begin{cases}\dfrac{d}{dt}\begin{bmatrix}z(s,t)\\v(s,t)\end{bmatrix}=\begin{bmatrix}0 & I\\ \frac{\partial^2}{\partial s^2} & 0\end{bmatrix}\begin{bmatrix}z(s,t)\\v(s,t)\end{bmatrix}, \quad s\in[0,1],\ t\in\mathbb{R}^+,\\ z(s,0)=0,\ v(s,0)=x(s),\\ dy(t)=Cz(t)\,dt+dw(t)\end{cases} \qquad (13)$$

in state space $\mathcal{X}=H_0^1[0,1]\times L^2(0,1)$ and $\mathcal{D}(A)=\big(H^2[0,1]\cap H_0^1[0,1]\big)\times H_0^1[0,1]$. The output operator $C\in\mathcal{L}(\mathcal{X},\mathbb{R})$ is given by $Cz=\int_0^1c(s)z(s)\,ds$ where $c(s)=\sum_{k=1}^{\infty}c_ke_k(s)$ with some $\{c_k\}\in l^2$ and $\{e_k\}$ is the orthonormal basis in $L^2(0,1)$

formed by the sine functions, that is $e_k(s) = \frac{1}{\sqrt{2}}\sin(k\pi s)$. The initial velocity is $x = \sum_{k=1}^{\infty} a_k e_{2^k}$ where $a_k \sim N(0,\sigma_k^2)$ and $a_k \perp a_i$ for $k \neq i$. It holds that $\mathbb{E}\Big(\|x\|_{\mathcal{X}}^2\Big) = \sum_{k=1}^{\infty}\sigma_k^2$ and thus this sum is assumed to converge. Then the solution to (13) and the corresponding output are

$$\begin{cases} z(s,t) = \frac{1}{\sqrt{2}}\sum_{k=1}^{\infty} a_k \sin(2^k\pi s)\sin(2^k\pi t), \\ v(s,t) = \frac{1}{\sqrt{2}}\sum_{k=1}^{\infty} a_k \sin(2^k\pi s)\cos(2^k\pi t), \\ dy(t) = \frac{1}{\sqrt{2}}\sum_{k=1}^{\infty} a_k c_{2^k}\sin(2^k\pi t)\,dt + dw(t). \end{cases}$$

Now set $T = 1$ and consider the subsequence $\hat{x}_{T,2^l}$ of the discrete time estimates, defined in (2). As noted in the proof of Thm. 3.1, it holds that $\mathbb{E}\Big(\big\|\hat{x}_{T,2^l} - \hat{x}(T)\big\|_{\mathcal{X}}^2\Big) = \sum_{i=l}^{\infty}\mathbb{E}\Big(\big\|\hat{x}_{T,2^{i+1}} - \hat{x}_{T,2^i}\big\|_{\mathcal{X}}^2\Big)$. The estimate $\hat{x}_{T,2^{l+1}}$ is obtained from the previous estimate $\hat{x}_{T,2^l}$ by including measurements $y\left(\frac{2i-1}{2^{l+1}}\right)$ for $i = 1,\dots,2^l$ as described in Section 2.3. In order to obtain a lower bound for $\mathbb{E}\Big(\big\|\hat{x}_{T,2^{l+1}} - \hat{x}_{T,2^l}\big\|_{\mathcal{X}}^2\Big)$, define $\widehat{C} := [C_h(h), C_h(3h),\dots,C_h(1-h)]^T : \mathcal{X} \to \mathbb{R}^{2^l}$ where $h = \frac{1}{2^{l+1}}$. That is, $\widehat{C}$ gives the whole batch of the measurements needed for the update. Then denoting $P_l = \mathrm{Cov}\left[\hat{x}_{T,2^l} - x, \hat{x}_{T,2^l} - x\right]$, it holds that

$$\mathbb{E}\Big(\big\|\hat{x}_{T,2^{l+1}} - \hat{x}_{T,2^l}\big\|_{\mathcal{X}}^2\Big) = \mathrm{tr}\left(P_l\widehat{C}^*\left(\widehat{C}P_l\widehat{C}^* + \frac{h}{4}RI\right)^{-1}\widehat{C}P_l\right)$$

$$\geq \left\langle \widehat{C}P_l e_{2^{l+1}}, \left(\widehat{C}P_l\widehat{C}^* + \frac{h}{4}RI\right)^{-1}\widehat{C}P_l e_{2^{l+1}}\right\rangle_{\mathbb{R}^{2^l}} \geq \frac{\left\|\widehat{C}P_l e_{2^{l+1}}\right\|_{\mathbb{R}^{2^l}}^2}{\max\Big(\mathrm{eig}\Big(\widehat{C}P_l\widehat{C}^* + \frac{h}{4}RI\Big)\Big)}.$$

For $h = 2^{-l}$ it holds that $C_h(ih)e_{2^k} = 0$ when $l < k$ and $i = 1,\dots,2^l - 1$ because when computing $C_h(ih)e_{2^k}$ by (8), the integrals are always over full periods of the sine function $\sin(2^k\pi t)$. When $l = k$ it holds that $C_h(ih)e_{2^k} = \frac{\sqrt{2}h}{\pi}c_{2^k}$ for every $i = 1,3,\dots,2^k - 1$. So, loosely speaking, the already included output values $y\left(\frac{2i-1}{2^l}\right)$ do not carry any information on a_k for $k > l$. Thus $P_l e_{2^{l+1}} = \sigma_{l+1}^2 e_{2^{l+1}}$ and $\left\|\widehat{C}P_l e_{2^{l+1}}\right\|_{\mathbb{R}^{2^l}}^2 = 2^l\sigma_{l+1}^2\left(\frac{\sqrt{2}h}{\pi}c_{2^{l+1}}\right)^2$. For the denominator it holds by part *(i)* of Lemma 2.2 that

$$\max\Big(\mathrm{eig}\Big(\widehat{C}P_l\widehat{C}^* + \frac{h}{4}RI\Big)\Big) \leq \frac{h}{4}R + \mathbb{E}\left(\left\|\widehat{C}x\right\|_{\mathbb{R}^{2^l}}^2\right) \leq \frac{h}{4}R + 2^l h^2\|C\|_{\mathcal{L}(\mathcal{X},\mathbb{R})}^2\,\mathrm{tr}(P_0).$$

Recalling $h = \frac{1}{2^{l+1}}$, we finally get $\mathbb{E}\Big(\big\|\hat{x}_{T,2^{l+1}} - \hat{x}_{T,2^l}\big\|_{\mathcal{X}}^2\Big) \geq \frac{4\sigma_{l+1}^2 c_{2^{l+1}}^2}{\pi^2 R + 2\pi^2\|C\|_{\mathcal{L}(\mathcal{X},\mathbb{R})}^2\,\mathrm{tr}(P_0)}$ and further

$$\mathbb{E}\Big(\big\|\hat{x}_{T,2^l} - \hat{x}(T)\big\|_{\mathcal{X}}^2\Big) \geq \frac{4\sum_{i=l+1}^{\infty}\sigma_i^2 c_{2^i}^2}{\pi^2 R + 2\pi^2\|C\|_{\mathcal{L}(\mathcal{X},\mathbb{R})}^2\,\mathrm{tr}(P_0)}$$

where there is no h-dependence and the variances $\{\sigma_k^2\}$ can be chosen so that the convergence is arbitrarily slow, concluding the example.

Clearly some additional assumptions are needed for getting any convergence rate estimates. In the following theorem, the initial state is assumed to be so smooth that the covariance operator satisfies $P_0 \in \mathcal{L}(\mathcal{X},\mathcal{D}(A))$. The problem here is that $\|P_j\|_{\mathcal{L}(\mathcal{X},\mathcal{D}(A))}$ is not necessarily decreasing as more measurements are taken

into account. Thus the convergence speed estimate has to be based on the initial covariance P_0.

Theorem 3.3. *Let $\hat{x}_{T,n}$ and $\hat{x}(T)$ be as defined in (2) and assume $C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$. Assume $x \sim N(m, P_0)$ where the covariance operator satisfies $P_0 \in \mathcal{L}(\mathcal{X}, \mathcal{D}(A))$. Then*

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^2}{n}$$

where $M = \frac{q\,\mathrm{tr}(P_n)\|P_0\|_{\mathcal{L}(\mathcal{X},\mathcal{D}(A))}\|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}}{\min(\mathrm{eig}(R))}$ and $P_n = \mathrm{Cov}\,[x - \hat{x}_{T,n}, x - \hat{x}_{T,n}]$.

Proof. The main idea of the proof is the same as in the proof of Theorem 3.1 and we note that every step taken until equation (11) in that proof can be taken in the infinite dimensional setting as well — p just has to be replaced by ∞ in the sums but this does not cause any problems.

So we pick up from (11) and note first that

$$\begin{aligned}\mathrm{tr}\,(C_h P_j C_h^*) &\leq q\,\|C_h P_j C_h^*\|_{\mathcal{L}(\mathcal{Y})} = q \sup_{\|y\|_{\mathcal{Y}}=1} \langle y, C_h P_j C_h^* y\rangle_{\mathcal{Y}} \\ =& q \sup_{\|y\|_{\mathcal{Y}}=1} \langle C_h^* y, P_j C_h^* y\rangle_{\mathcal{X}} \leq q \sup_{\|y\|_{\mathcal{Y}}=1} \langle C_h^* y, P_0 C_h^* y\rangle_{\mathcal{X}} = q\,\|C_h P_0 C_h^*\|_{\mathcal{L}(\mathcal{Y})}\end{aligned}$$

where $q = \dim(\mathcal{Y})$. The inequality $P_j \leq P_0$ was used in $\mathcal{X}$, but now the $\mathcal{L}(\mathcal{X}, \mathcal{D}(A))$-norm can be used for P_0. Then using both parts *(i)* and *(ii)* of Lemma 2.2 gives

$$\|C_h P_0 C_h^*\|_{\mathcal{L}(\mathcal{Y})} \leq \frac{h^3}{2}\,\|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\,\|P_0\|_{\mathcal{L}(\mathcal{X},\mathcal{D}(A))}\,.$$

As before, this leads to an estimate

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{q\,\mathrm{tr}(P_n)\,\|P_0\|_{\mathcal{L}(\mathcal{X},\mathcal{D}(A))}\,\|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\,T^2}{\min(\mathrm{eig}(R))\,n} =: \frac{MT^2}{n}$$

completing the proof. □

Checking the assumption $P_0 \in \mathcal{L}(\mathcal{X}, \mathcal{D}(A))$ might be difficult. Under the stronger smoothness assumption $x \in \mathcal{D}(A)$ almost surely, we get the same convergence rate as in the finite dimensional case:

Theorem 3.4. *Make the same assumptions as in Theorem 3.3. Assume, in addition, that $x \in \mathcal{D}(A)$ almost surely. Then*

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^3}{n^2}$$

where $M = \frac{\mathrm{tr}(P_n)\mathrm{tr}(AP_nA^*)\|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}}{6\min(\mathrm{eig}(R))}$ and $P_n = \mathrm{Cov}\,[x - \hat{x}_{T,n}, x - \hat{x}_{T,n}]$.

Proof. The proof is the same as that of Theorem 3.1 but from Eq. (11) we proceed differently. It holds that

$$\begin{aligned}\mathrm{tr}\,(C_h P_j C_h^*) \leq \mathrm{tr}\,(C_h P_n C_h^*) &= \mathbb{E}\Big(\|C_h(x - \hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big) \\ &\leq \frac{h^4}{4}\,\|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\,\mathbb{E}\Big(\|A(x - \hat{x}_{T,n})\|_{\mathcal{X}}^2\Big)\end{aligned}$$

where the last inequality holds by part *(ii)* of Lemma 2.2. The term is finite by Proposition 2.1 and Fernique's theorem. Further, it holds that $\mathbb{E}\Big(\|A(x - \hat{x}_{T,n})\|_{\mathcal{X}}^2\Big) = \mathrm{tr}(AP_nA^*)$. Now the result follows as above. □

3.3. **Unbounded observation operator C.** We proceed to prove a similar result for systems with unbounded observation operator C — provided that A is (unitarily) diagonalizable. The proof is quite similar to that of Theorem 3.3. Again the main difference is how we proceed from (11). To get a useful bound for $\operatorname{tr}(C_h P_j C_h^*)$, some assumptions on C and the spectral asymptotics of A are required.

Theorem 3.5. *Let $\hat{x}_{T,n}$ and $\hat{x}(T)$ be as defined above in (2). Denote by $\{\mu_k + i\lambda_k\}_{k=1}^{\infty}$ the spectrum of A ordered so that $|\mu_k + i\lambda_k|$ is non-decreasing and let $\{e_k\}_{k=1}^{\infty} \subset \mathcal{D}(A)$ be the corresponding set of eigenvectors that give an orthonormal basis for $\mathcal{X}$. Make the following assumptions on x, A, and C:*

(i) *$x \in \mathcal{D}(A)$ almost surely;*

(ii) *$\mu_k \leq 0$ for all k, and there exists $\delta > 1/2$ such that*

$$\lim_{k\to\infty} \frac{|\mu_k + i\lambda_k|}{k^\beta} = \begin{cases} 0 & \text{when } \beta > \delta, \\ \infty & \text{when } \beta < \delta; \end{cases}$$

(iii) *There exists $\gamma \in [0, 1)$ such that $2\gamma + 1/\delta < 2$ and*

$$\sup_k \frac{\|Ce_k\|_{\mathcal{Y}}}{|\mu_k + i\lambda_k|^\gamma} < \infty.$$

Then the following holds:

- *If $\lim_{k\to\infty} \frac{|\mu_k+i\lambda_k|}{k^\delta} = \Gamma \in (0, \infty)$, then*

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^{3-2\gamma-1/\delta}}{n^{2-2\gamma-1/\delta}}$$

  where the constant M is given below in (15).

- *If either this limit does not exist, or it is 0 or ∞, then for all $\epsilon \in \left(0, \delta - \frac{1}{2-\gamma}\right)$*

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{M_\epsilon T^{3-2\gamma-1/(\delta+\epsilon)}}{n^{2-2\gamma-1/(\delta-\epsilon)}}$$

  where the ϵ-dependent constant M_ϵ is given below also in (15) but with different, ϵ-dependent parameters (see the last paragraph of the proof).

For example, 1D wave equation on interval $[0, L]$ with Dirichlet boundary conditions in the natural state space where some pointwise value of the state is observed, satisfies the assumptions of the above theorem with $\delta = 1$ and $\gamma = 0$. The limit of $\frac{|\mu_k+i\lambda_k|}{k}$ as $k \to \infty$ exists and it is $\Gamma = \frac{\pi}{2L}$. This would imply convergence rate $\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^2}{n}$.

Proof. Assume first that $\lim_{k\to\infty} \frac{|\mu_k+i\lambda_k|}{k^\delta} = \Gamma \in (0, \infty)$. Note that assumption *(i)* with Proposition 2.1 and Fernique's theorem imply that $\mathbb{E}\Big(\|Ax\|_{\mathcal{X}}^2\Big) < \infty$. Denoting $x = \hat{x}_{T,n} + \sum_{k=1}^{\infty} \alpha_k e_k$, this condition can be expressed as $\mathbb{E}\Big(\|A\hat{x}_{T,n}\|_{\mathcal{X}}^2\Big) + \mathbb{E}\big(\sum_{k=1}^{\infty} |\mu_k + i\lambda_k|^2 \alpha_k^2\big) < \infty$. Again the proof proceeds exactly as the proof of Theorem 3.3 until Equation (11).

As in the proof of Theorem 3.4, note that $\operatorname{tr}(C_h P_j C_h^*) \leq \operatorname{tr}(C_h P_n C_h(t)^*) = \mathbb{E}\Big(\|C_h(x - \hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big)$. Then

$$C_h(x - \hat{x}_{T,n}) = \sum_{k=1}^{\infty} \frac{\alpha_k}{2} \left(\int_{t-h}^{t} e^{(\mu_k+i\lambda_k)s} ds - \int_{t}^{t+h} e^{(\mu_k+i\lambda_k)s} ds \right) Ce_k. \tag{14}$$

For the term inside parentheses, we have

$$\left|\int_{t-h}^{t} e^{(\mu_k+i\lambda_k)s}ds - \int_{t}^{t+h} e^{(\mu_k+i\lambda_k)s}ds\right| \leq h^2 \sup_{s\geq 0}\left|\frac{d}{ds}e^{(\mu_k+i\lambda_k)s}\right| = h^2|\mu_k+i\lambda_k|,$$

since $\mu_k \leq 0$. On the other hand, computing the integrals yields

$$\left|\int_{t-h}^{t} e^{(\mu_k+i\lambda_k)s}ds - \int_{t}^{t+h} e^{(\mu_k+i\lambda_k)s}ds\right| \leq \frac{4}{|\mu_k+i\lambda_k|}.$$

Now the idea is to bound the sum in (14) by using the first bound for small k and the latter for large k. Define the index $n(h) := \lceil h^{-1/\delta} \rceil$ for splitting the sum to get

$$\begin{aligned}\|C_h(x-\hat{x}_{T,n})\|_{\mathcal{Y}} &\leq \sum_{k=1}^{n(h)} \frac{|\alpha_k|}{2}\|Ce_k\|_{\mathcal{Y}}\,|\mu_k+i\lambda_k|h^2 + \sum_{k=n(h)+1}^{\infty}|\alpha_k|\,\|Ce_k\|_{\mathcal{Y}}\,\frac{2}{|\mu_k+i\lambda_k|}\\ &=: (I)+(II).\end{aligned}$$

We then proceed to find upper bounds for the two parts. Using Cauchy-Schwartz inequality and denoting $\hat{\Gamma} := \sup_k \frac{|\mu_k+i\lambda_k|}{k^\delta}$ gives

$$\begin{aligned}(I) &\leq \frac{h^2}{2}\left(\sum_{k=1}^{n(h)}\alpha_k^2\|Ce_k\|_{\mathcal{Y}}^2\,|\mu_k+i\lambda_k|^{2-2\gamma}\right)^{1/2}\left(\sum_{k=1}^{n(h)}|\mu_k+i\lambda_k|^{2\gamma}\right)^{1/2}\\ &\leq \frac{h^2\hat{\Gamma}^\gamma}{2}M_I\left(\sum_{k=1}^{n(h)}k^{2\gamma\delta}\right)^{1/2}\end{aligned}$$

where $M_I = \left(\sum_{k=1}^{n(h)}\alpha_k^2\|Ce_k\|_{\mathcal{Y}}^2\,|\mu_k+i\lambda_k|^{2-2\gamma}\right)^{1/2}$. The sum inside the parentheses can be bounded from above by the integral $\int_0^{n(h)+1}x^{2\gamma\delta}dx$ to get

$$\begin{aligned}(I) &\leq \frac{h^2\hat{\Gamma}^\gamma}{2\sqrt{2\gamma\delta+1}}M_I\sqrt{(n(h)+1)^{2\gamma\delta+1}}\\ &\leq \frac{3^\delta\hat{\Gamma}^\gamma}{2\sqrt{2\gamma\delta+1}}M_I h^{2-\gamma-\frac{1}{2\delta}} \leq \frac{3^\delta\hat{\Gamma}^\gamma}{2}M_I h^{2-\gamma-\frac{1}{2\delta}}\end{aligned}$$

where the last row follows from the facts that

$$\sqrt{(n(h)+1)^{2\gamma\delta+1}} \leq \sqrt{(h^{-1/\delta}+2)^{2\gamma\delta+1}} = (1+2h^{1/\delta})^{\gamma\delta+\frac{1}{2}}h^{-\gamma-\frac{1}{2\delta}} \leq 3^\delta h^{-\gamma-\frac{1}{2\delta}}$$

if $h \leq 1$, and that $2\gamma\delta+1 > 1$.

For the second part, assume $|\mu_k+i\lambda_k| \geq \check{\Gamma}k^\delta$ for $k \geq n(h)+1$ where $\check{\Gamma} = 0.9\Gamma$ for example. Again, using Cauchy-Schwartz inequality yields

$$\begin{aligned}(II) &\leq 2\left(\sum_{k=n(h)+1}^{\infty}\alpha_k^2\|Ce_k\|_{\mathcal{Y}}^2\,|\mu_k+i\lambda_k|^{2-2\gamma}\right)^{1/2}\left(\sum_{k=n(h)+1}^{\infty}\frac{1}{|\mu_k+i\lambda_k|^{4-2\gamma}}\right)^{1/2}\\ &\leq \frac{2}{\check{\Gamma}^{2-\gamma}}M_{II}\left(\sum_{k=n(h)+1}^{\infty}\frac{1}{k^{(4-2\gamma)\delta}}\right)^{1/2}\end{aligned}$$

where $M_{II} = \left(\sum_{k=n(h)+1}^{\infty} \alpha_k^2 \left\|Ce_k\right\|_{\mathcal{Y}}^2 |\mu_k + i\lambda_k|^{2-2\gamma}\right)^{1/2}$. Now the sum inside the parentheses can be bounded from above by the integral $\int_{n(h)}^{\infty} \frac{1}{x^{(4-2\gamma)\delta}} dx$. Note that our assumptions on γ and δ imply $(4-2\gamma)\delta > 2$. So we get

$$(II) \leq \frac{2M_{II}}{\check{\Gamma}^{2-\gamma}\sqrt{(4-2\gamma)\delta - 1}} \left(\frac{1}{n(h)^{(4-2\gamma)\delta-1}}\right)^{1/2} \leq \frac{2}{\check{\Gamma}^{2-\gamma}} M_{II} h^{2-\gamma-\frac{1}{2\delta}}$$

where in the last row we have used $n(h) \geq h^{-1/\delta}$.

Combining the bounds gives

$$\begin{aligned}
&\mathbb{E}\Big(\left\|C_h(x - \hat{x}_{T,n})\right\|_{\mathcal{Y}}^2\Big) \leq 2\mathbb{E}\big((I)^2 + (II)^2\big) \\
&\leq 2\left(M_I^2 + M_{II}^2\right) \max\left(\frac{9^\delta \hat{\Gamma}^{2\gamma}}{4}, \frac{4}{\check{\Gamma}^{4-2\gamma}}\right) h^{4-2\gamma-1/\delta} \\
&\leq 2\mathbb{E}\Big(\left\|A(x - \hat{x}_{T,n})\right\|_{\mathcal{X}}^2\Big) \sup_k \frac{\left\|Ce_k\right\|_{\mathcal{Y}}^2}{|\mu_k + i\lambda_k|^{2\gamma}} \max\left(\frac{9^\delta \hat{\Gamma}^{2\gamma}}{4}, \frac{4}{\check{\Gamma}^{4-2\gamma}}\right) h^{4-2\gamma-1/\delta} \\
&=: 2M_0\mathbb{E}\Big(\left\|A(x - \hat{x}_{T,n})\right\|_{\mathcal{X}}^2\Big) \sup_k \frac{\left\|Ce_k\right\|_{\mathcal{Y}}^2}{|\mu_k + i\lambda_k|^{2\gamma}} h^{4-2\gamma-1/\delta}
\end{aligned}$$

where we have used

$$M_I^2 + M_{II}^2 = \sum_{k=1}^{\infty} \alpha_k^2 \left\|Ce_k\right\|_{\mathcal{Y}}^2 |\mu_k + i\lambda_k|^{2-2\gamma} \leq \sum_{k=1}^{\infty} |\mu_k + i\lambda_k|^2 \alpha_k^2 \sup_j \frac{\left\|Ce_j\right\|_{\mathcal{Y}}^2}{|\mu_j + i\lambda_j|^{2\gamma}}.$$

Note that we assumed that we could choose for example $\check{\Gamma} = 0.9\Gamma$. In some sense this is not our choice but we need to make sure that the "original" $h = \frac{T}{2n}$ is small enough so that $n(T/(2n)) = \left(\frac{2n}{T}\right)^{1/\delta}$ is such that there exists $\check{\Gamma} > 0$ for which $\frac{|\mu_k + i\lambda_k|}{k^\delta} \geq \check{\Gamma}$ for $k \geq n(T/(2n))$.

To finish the proof, we continue as in the proof of Theorem 3.3; that is, we conclude

$$\begin{aligned}
&\operatorname{tr}\left(P_j C_h^* \left(C_h P_j C_h^* + \frac{h}{4} R\right)^{-1} C_h P_j\right) \\
&\leq 2C_R M_0 \operatorname{tr}(P_n) \operatorname{tr}(A P_n A^*) \sup_k \frac{\left\|Ce_k\right\|_{\mathcal{Y}}^2}{|\mu_k + i\lambda_k|^{2\gamma}} h^{3-2\gamma-1/\delta}
\end{aligned}$$

where $P_n = \operatorname{Cov}[x - \hat{x}_{T,n}, x - \hat{x}_{T,n}]$. Now doing the same summation as before in part (III) of the proof of Theorem 3.1, it follows that

$$\mathbb{E}\Big(\left\|\hat{x}_{T,n} - \hat{x}(T)\right\|_{\mathcal{X}}^2\Big) \leq \frac{nC_R M_0}{2^{2-2\gamma-1/\delta} - 1} \operatorname{tr}(P_n) \operatorname{tr}(A P_n A^*) \sup_k \frac{\left\|Ce_k\right\|_{\mathcal{Y}}^2}{|\mu_k + i\lambda_k|^{2\gamma}} \left(\frac{T}{n}\right)^{3-2\gamma-1/\delta}$$

completing the proof in the first case with constant

$$M = \frac{C_R \operatorname{tr}(P_n) \operatorname{tr}(A P_n A^*)}{2^{2-2\gamma-1/\delta} - 1} \sup_k \frac{\left\|Ce_k\right\|_{\mathcal{Y}}^2}{|\mu_k + i\lambda_k|^{2\gamma}} \max\left(\frac{9^\delta \hat{\Gamma}^{2\gamma}}{4}, \frac{4}{\check{\Gamma}^{4-2\gamma}}\right) \tag{15}$$

where $C_R = 4/\min(\operatorname{eig}(R))$ is defined in (10).

In the case that $\lim_{k\to\infty} \frac{|\mu_k + i\lambda_k|}{k^\delta}$ is 0, ∞, or it does not exist, some modifications are required to the bounds of (I) and (II). In the bound for (I), δ needs to be

replaced by $\delta+\epsilon$ and then $\hat{\Gamma}_\epsilon = \sup_k \frac{|\mu_k+i\lambda_k|}{k^{\delta+\epsilon}} < \infty$. In the bound for (II), δ needs to be replaced by $\delta-\epsilon$ and then $\check{\Gamma}_\epsilon = \inf_{k\geq n(h)+1} \frac{|\mu_k+i\lambda_k|}{k^{\delta-\epsilon}} > 0$. □

The assumption *(iii)* in the theorem differs from our minimal assumption $C \in \mathcal{L}(\mathcal{D}(A),\mathcal{Y})$ which is equivalent to $\left\{\frac{\|Ce_k\|_{\mathcal{Y}}}{|\mu_k+i\lambda_k|}\right\} \in l^2$ for unitarily diagonalizable A. It is possible to construct a system for which $C \in \mathcal{L}(\mathcal{D}(A),\mathcal{Y})$ but *(iii)* does not hold.

Remark 3.6. Theorem 3.5 can be extended to $\gamma < 0$. In that case, when determining the bounds for (I) and (II), the computations are carried out as if γ were zero. This eventually leads to a bound $\mathbb{E}\Big(\|\hat{x}_{T,n}-\hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^{3-1/\delta}}{n^{2-1/\delta}}$. Note that if assumption *(iii)* holds for $\gamma < -\frac{1}{2\delta}$ then C is actually bounded.

There is no assumption on the diagonalizability of A in the following theorem. Unfortunately, the obtained convergence rate is not very impressive.

Theorem 3.7. *Let $\hat{x}_{T,n}$ and $\hat{x}(T)$ be as defined above in* (2)*. Make the following assumptions:*

(i) *$x \in \mathcal{D}(A)$ almost surely;*
(ii) *The orthonormal basis $\{e_k\} \subset \mathcal{X}$ is such that $e_k \in \mathcal{D}(A^2)$ for every $k \in \mathbb{N}$ and there exists $\delta > 1/2$ such that for $x = \sum_{k=1}^{\infty} \alpha_k e_k$ the norm given by $\sqrt{\sum_{k=1}^{\infty} k^{2\delta}\alpha_k^2}$ is equivalent to the $\mathcal{D}(A)$-norm and $\sqrt{\sum_{k=1}^{\infty} k^{4\delta}\alpha_k^2}$ is equivalent to the $\mathcal{D}(A^2)$-norm;*
(iii) *The system is well-posed in the sense that $\left\|Ce^{A(\cdot)}x\right\|_{L^2((0,T);\mathcal{Y})} \leq H_T \|x\|_{\mathcal{X}}$ for some $H_T > 0$.*

Then

$$\mathbb{E}\Big(\|\hat{x}_{T,n}-\hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{M(T)\,T^{2-1/2\delta}}{n^{1-1/2\delta}}$$

with $M(T) = \frac{C_R \mathrm{tr}(P_n)\mathrm{tr}(AP_nA^*)}{2^{1-1/2\delta}-1} \max\left(\frac{3^{2\delta+1}T\|C\|^2_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})}}{8\delta+4}, \frac{H_T}{2\delta-1}\right)$ *where* $C_R = \frac{4}{\min(\mathrm{eig}(R))}$ *is defined in* (10) *and* $P_n = \mathrm{Cov}\,[x-\hat{x}_{T,n}, x-\hat{x}_{T,n}]$.

Proof. In this proof, the aforementioned norms are used in $\mathcal{D}(A)$ and $\mathcal{D}(A^2)$. We need to utilize the global output bound $\left\|Ce^{A(\cdot)}x\right\|_{L^2((0,T);\mathcal{Y})} \leq H_T\|x\|_{\mathcal{X}}$. To this end, define a stacked operator $\widehat{C}_h := [C_h(h), C_h(3h), \ldots, C_h(T-h)]^T$ for $h = 2^{-K}\frac{T}{n}$ mapping to a product space $\mathcal{Y}^{2^{K-1}n}$. This operator is used to add a whole batch of intermediate measurements at once as was done in Example 3.2. Below $\left[a_i\right]_{i=1}^N$ is used to denote an augmented vector with components a_i.

Otherwise the proof proceeds similarly as the proof of Thm. 3.5 but the sum in (14) is split using the index $n(h) = \lceil h^{-1/2\delta}\rceil$ to get $\left\|\widehat{C}_h(x-\hat{x}_{T,n})\right\|_{\mathcal{Y}^{2^{K-1}n}} \leq (I)+(II)$ where

$$(I) = \left\|\sum_{k=1}^{n(h)} \frac{\alpha_k}{2}\left[C\int_{(2j-2)h}^{(2j-1)h} e^{As}e_k\,ds - C\int_{(2j-1)h}^{2jh} e^{As}e_k\,ds\right]_{i=1}^{2^{K-1}n}\right\|_{\mathcal{Y}^{2^{K-1}n}}$$

$$\leq \sum_{k=1}^{n(h)} \frac{|\alpha_k|}{2}\sqrt{\frac{T}{2h}}h^2\|C\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})}\,k^{2\delta} \leq \sqrt{\frac{Th^3}{8}}\|C\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})}\left(\sum_{k=1}^{n(h)} k^{2\delta}\alpha_k^2\right)^{1/2}\left(\sum_{k=1}^{n(h)} k^{2\delta}\right)^{1/2}$$

where the first inequality is obtained by bounding the derivative of $Ce^{At}e_k$ by

$$\left\| \frac{d}{dt} Ce^{At}e_k \right\|_{\mathcal{Y}} \leq \|C\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})} \|e_k\|_{\mathcal{D}(A^2)} = \|C\|_{\mathcal{L}(\mathcal{D}(A),\mathcal{Y})} k^{2\delta}$$

and using the same argument as in the proof of part *(ii)* of Lemma 2.2 and noting that $2^{K-1}n = \frac{T}{2h}$. For the remaining part it holds that

$$(II) = \left\| \sum_{k=n(h)+1}^{\infty} \frac{\alpha_k}{2} \left[C \int_{(2i-2)h}^{(2i-1)h} e^{As}e_k \, ds - C \int_{(2i-1)h}^{2ih} e^{As}e_k \, ds \right]_{i=1}^{2^{K-1}n} \right\|_{\mathcal{Y}^{2^{K-1}n}}$$

$$\leq \sum_{k=n(h)+1}^{\infty} \frac{|\alpha_k|}{2} \sqrt{2h} H_T \leq \sqrt{\frac{h}{2}} H_T \left(\sum_{k=n(h)+1}^{\infty} k^{2\delta} \alpha_k^2 \right)^{1/2} \left(\sum_{k=n(h)+1}^{\infty} k^{-2\delta} \right)^{1/2}$$

since it holds that

$$\left\| \left[C \int_{(2i-2)h}^{(2i-1)h} e^{As}e_k \, ds - C \int_{(2i-1)h}^{2ih} e^{As}e_k \, ds \right]_{i=1}^{2^{K-1}n} \right\|_{\mathcal{Y}^{2^{K-1}n}}$$

$$\leq \left(\sum_{i=1}^{2^{K-1}n} \left(\int_{(2i-2)h}^{2ih} \left\| Ce^{As}e_k \right\|_{\mathcal{Y}} ds \right)^2 \right)^{1/2} \leq \sqrt{2h} \left\| Ce^{As}e_k \right\|_{L^2((0,T);\mathcal{Y})}$$

where the last inequality follows from Cauchy-Schwartz inequality. Finally, the result is obtained by proceeding as in the proof of Thm. 3.5 and doing the summation over $K = 1, 2, \ldots$. □

The theory of well-posed systems has been extensively studied. A comprehensive treatment can be found in the book [20] by Staffans. One good example of systems that satisfy assumption (iii) is provided by scattering passive boundary control systems, see the article [15] by Malinen and Staffans. This condition is also known as admissibility of the output operator C, introduced in [24] by Weiss.

3.4. **Analytic semigroup e^{At}.** In this section we show the convergence estimate when A is the generator of an analytic semigroup. One result is first shown without additional assumptions for bounded and unbounded observation operator C. Then we assume further that $-A$ is a sectorial operator in $\mathcal{X}$ which enables us to treat non-integer powers $(-A)^\eta$ for $\eta \geq 0$. An example of such case is provided by heat equation treated below in Example 3.10.

An important tool here is that for analytic semigroups it holds that

$$\left\| A^\kappa e^{At} \right\|_{\mathcal{L}(\mathcal{X})} \leq \frac{c(\kappa)}{t^\kappa}, \qquad t > 0, \ \kappa \in \mathbb{N} \tag{16}$$

(see [22: Theorem 3.3.1]). This gives

$$\|C_h(t)x\|_{\mathcal{Y}} \leq \frac{c(1)\|x\|_{\mathcal{X}/\mathcal{D}(A)}}{t-h} \frac{\|C\|_{\mathcal{L}(\mathcal{X}/\mathcal{D}(A),\mathcal{Y})}}{2} h^2, \qquad t > h. \tag{17}$$

For $t = h$, we can use part *(i)* of Lemma 2.2.

Theorem 3.8. *Let $\hat{x}_{T,n}$ and $\hat{x}(T)$ be as defined above in* (2)*. Assume A is the generator of an analytic and contractive C_0-semigroup and assume either*

(i) *$C \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, or*

(ii) $C \in \mathcal{L}(\mathcal{D}(A), \mathcal{Y})$ *and* $x \in \mathcal{D}(A)$ *almost surely.*

Then

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT}{n}$$

where $M = C_R \mathrm{tr}(P_n)\, \|C\|_{\mathcal{L}(\mathcal{X}/\mathcal{D}(A),\mathcal{Y})}^2\, \mathbb{E}\Big(\|x - \hat{x}_{T,n}\|_{\mathcal{X}/\mathcal{D}(A)}^2\Big)\, \Big(1 + \frac{c(1)^2\pi^2}{96}\Big)$ *and* $P_n = \mathrm{Cov}\,[x - \hat{x}_{T,n}, x - \hat{x}_{T,n}]$.

Proof. The proofs for the two cases are identical so only the case *(i)* is presented. In the second case just replace $\mathcal{X}$ by $\mathcal{D}(A)$ in $\|x\|_{\mathcal{X}}$ and $\|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}$.

As many times before, the proof is based on finding an upper bound for $\mathbb{E}\Big(\|C_h(t)(x - \hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big)$. The difference to earlier proofs is that here we use (17) and the t-dependence of the bound has to be utilized. Because of this, it is not possible to just multiply a bound found for certain $h = \frac{T}{2^K n}$ by $2^{K-1}n$ as has been done above but instead, we need to calculate and add up all bounds separately, that is, compute

$$\frac{C_R \mathrm{tr}(P_n)}{h} \sum_{l=1}^{2^{K-1}n} \mathbb{E}\Big(\|C_h(t_l)(x - \hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big), \qquad t_l = (2l-1)h,\ h = \frac{T}{2^K n} \tag{18}$$

and sum these up for $K = 1, 2, \ldots$. For $l = 1$, we use $\|C_h(h)(x - \hat{x}_{T,n})\|_{\mathcal{Y}} \leq h\, \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\, \|x - \hat{x}_{T,n}\|_{\mathcal{X}}$ from part *(i)* of Lemma 2.2. For $l > 1$, we use (17) to obtain

$$\begin{aligned}
&\frac{C_R \mathrm{tr}(P_n)}{h} \sum_{l=1}^{2^{K-1}n} \mathbb{E}\Big(\|C_h(t_l)(x - \hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big) \\
&\quad \leq C_R \mathrm{tr}(P_n)\, \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}^2\, \mathbb{E}\Big(\|x - \hat{x}_{T,n}\|_{\mathcal{X}}^2\Big) \left(1 + \frac{c(1)^2}{16} \sum_{j=1}^{2^{K-1}n-1} 1/j^2\right) h \\
&\quad \leq C_R \mathrm{tr}(P_n)^2\, \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}^2 \left(1 + \frac{c(1)^2\pi^2}{96}\right) h.
\end{aligned}$$

Now summing up over $K = 1, 2, \ldots$ completes the proof. □

Then one more case is treated where A is as before and, in addition, $-A$ is a sectorial operator, see [2: Section 3.8] for definitions. Then it is possible to define non-integer powers $(-A)^\eta$ where $\eta \in \mathbb{R}$ and spaces $\mathcal{D}((-A)^\eta)$ equipped with the corresponding graph norm. Also (16) holds then for non-integer $\kappa \geq 0$ if A is replaced by $-A$, see [22: Thm. 3.3.3]. In particular, if A is strictly negative definite, then it is sectorial. This type of systems are also studied in [5] and [7].

Theorem 3.9. *Let* $\hat{x}_{T,n}$ *and* $\hat{x}(T)$ *be as defined above in in* (2)*. Assume* A *is the generator of an analytic and contractive* C_0*-semigroup and, in addition,* $-A$ *is a sectorial operator. Then assume* $C \in \mathcal{L}(\mathcal{D}((-A)^\nu), \mathcal{Y})$ *and* $x \in \mathcal{D}((-A)^\eta)$ *almost surely where* $\nu \in \mathbb{R}$ *and* $\eta \in \mathbb{R}$ *are such that* $|\eta - \nu| < 1/2$*. Then*[1]

$$\mathbb{E}\Big(\|\hat{x}_{T,n} - \hat{x}(T)\|_{\mathcal{X}}^2\Big) \leq \frac{MT^{1+2(\eta-\nu)}}{n^{1+2(\eta-\nu)}}$$

where M *is given below in* (20).

[1]This result extends to $\eta - \nu = 1/2$ in which case the convergence rate is $\mathcal{O}(T^2 n^{-2} \ln n)$.

Proof. This is done exactly as the proof of Theorem 3.8 above. Just the bounds for $\|C_h(t_l)(x-\hat{x}_{T,n})\|_{\mathcal{Y}}$ in the summation (18) are computed differently. To begin with, we note that we get from (16) with non-integer $\kappa = 1-\eta+\nu$,

$$\left\|CAe^{At}(x-\hat{x}_{T,n})\right\|_{\mathcal{Y}} \le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\,\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\frac{c(1-\eta+\nu)}{t^{1-\eta+\nu}}. \quad (19)$$

When treating the term with $l=1$ in (18), the cases $\nu \ge \eta$ and $\nu < \eta$ have to be considered separately. First for $\nu \ge \eta$,

$$\begin{aligned}\left\|Ce^{At}(x-\hat{x}_{T,n})\right\|_{\mathcal{Y}} &= \left\|C(-A)^{-\nu}(-A)^{\nu-\eta}e^{At}(-A)^{\eta}(x-\hat{x}_{T,n})\right\|_{\mathcal{Y}}\\ &\le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\,\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\,\frac{c(\nu-\eta)}{t^{\nu-\eta}}.\end{aligned}$$

Then for $\eta \le \nu < 1+\eta$,

$$\begin{aligned}\|C_h(h)(x-\hat{x}_{T,n})\|_{\mathcal{Y}} &\le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\,c(\nu-\eta)\int_0^{2h}\frac{1}{s^{\nu-\eta}}ds\\ &\le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\,\frac{c(\nu-\eta)}{1+\eta-\nu}(2h)^{1+\eta-\nu}.\end{aligned}$$

For $\nu < \eta < 1+\nu$, one can show a similar bound by the same technique that was used in the proof of part *(ii)* of Lemma 2.2. Instead of bounding the derivative norm $\left\|CAe^{At}x\right\|_{\mathcal{Y}}$ by a constant, using (19) gives a bound

$$\begin{aligned}&\|C_h(h)(x-\hat{x}_{T,n})\|_{\mathcal{Y}}\\ &\le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\,\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\,\frac{c(1-\eta+\nu)}{\eta-\nu}\frac{2^{1+\eta-\nu}-2}{1+\eta-\nu}h^{1+\eta-\nu}\\ &\le \|C\|_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\,\|x-\hat{x}_{T,n}\|_{\mathcal{D}((-A)^\eta)}\,\frac{4\ln 2\,c(1-\eta+\nu)}{1+\eta-\nu}h^{1+\eta-\nu}.\end{aligned}$$

To cover $l>1$ in (18), we use (19) to get

$$\frac{C_R\mathrm{tr}(P_n)}{h}\sum_{l=1}^{2^{K-1}n}\mathbb{E}\Big(\|C_h(t_l)(x-\hat{x}_{T,n})\|_{\mathcal{Y}}^2\Big) \le M_0 h^{1+2(\eta-\nu)}$$

where M_0 is gathered from the used inequalities. Finally summing over $K=1,2,\ldots$ yields the result with

$$\begin{aligned}M = C_R\mathrm{tr}(P_n)\,\|C\|^2_{\mathcal{L}(\mathcal{D}((-A)^\nu),\mathcal{Y})}\,\frac{\mathbb{E}\Big(\|x-\hat{x}_{T,n}\|^2_{\mathcal{D}((-A)^\eta)}\Big)}{2^{1+2(\eta-\nu)}-1}\times\\ \times\left(M_{\nu,\eta}+c(1-\eta+\nu)^2\frac{2-2(\eta-\nu)}{1-2(\eta-\nu)}\right)\end{aligned} \quad (20)$$

where we have used $\sum_{j=1}^{\infty}\frac{1}{j^{2+2(\nu-\eta)}} \le \frac{2-2(\eta-\nu)}{1-2(\eta-\nu)}$ and the term with $l=1$ gives

$$M_{\nu,\eta} = \begin{cases}\frac{16(\ln 2)^2c(1-\eta+\nu)^2}{(1+\eta-\nu)^2} & \text{if } \eta > \nu,\\ \frac{2^{2+2(\eta-\nu)}c(\nu-\eta)^2}{(1+\eta-\nu)^2} & \text{if } \eta \le \nu\end{cases}$$

and $P_n = \mathrm{Cov}\,[x-\hat{x}_{T,n}, x-\hat{x}_{T,n}]$. □

Example 3.10. Consider the 1D heat equation

$$\begin{cases} \frac{\partial}{\partial t} z(x,t) = \frac{\partial^2}{\partial x^2} z(x,t), \quad x \in [0,1], \\ z(0,t) = z(1,t) = 0, \\ z(x,0) = z_0, \\ dy(t) = \frac{\partial}{\partial x} z(0,t)\, dt + dw(t) \end{cases}$$

with state space $\mathcal{X} = L^2(0,1)$ and $\mathcal{D}(A) = H_0^2[0,1]$. Assume $z_0 \in \mathcal{D}(A)$ almost surely. Now the spectrum of A is $\{-\pi^2 k^2\}$ and the corresponding eigenvectors are $e_k = \sin(\pi k x)$. Then it is easy to see that the assumptions of Theorem 3.5 are satisfied with $\delta = 2$ and $\gamma = 1/2$ and thus the theorem implies convergence rate $\mathcal{O}(n^{-1/2})$ for $\hat{x}_{T,n}$. Clearly Theorem 3.8 implies convergence rate $\mathcal{O}(n^{-1})$ but we can do better.

Denoting $z = \sum_{k=1}^{\infty} \alpha_k e_k$ we have $\|z\|^2_{\mathcal{D}((-A)^\nu)} = \sum_{k=1}^{\infty} k^{4\nu} \alpha_k^2$. For the output it holds that

$$|Cz|^2 = \left| \sum_{k=1}^{\infty} \pi k \alpha_k \right|^2 \leq \pi \sum_{k=1}^{\infty} \frac{1}{k^{1+\epsilon}} \sum_{k=1}^{\infty} k^{3+\epsilon} \alpha_k^2$$

from which it can be deduced that $C \in \mathcal{L}(\mathcal{D}((-A)^\nu), \mathcal{Y}))$ for $\nu > 3/4$. Now Theorem 3.9 implies convergence rate $\mathcal{O}(n^{-3/2+\epsilon})$ for $\hat{x}_{T,n}$ with $\epsilon > 0$ — of course, with a multiplicative constant that tends to infinity as $\epsilon \to 0$.

4. Discussion

Since the implementation of the discrete time Kalman filter is straightforward, it is a tempting choice for state estimation for discretized continuous time systems. As the temporal discretization is refined, the discrete time state estimate converges pointwise to the continuous time estimate in $L^2(\Omega; \mathcal{X})$. In this article, we derived convergence speed estimates with various assumptions on the system. With infinite dimensional systems even with bounded observation operator, some smoothness assumption on the initial state is needed for obtaining any convergence speed estimates. This was demonstrated in Example 3.2. Possible additional assumptions are (i): for the initial state covariance it holds that $P_0 \in \mathcal{L}(\mathcal{X}, \mathcal{D}(A))$; or (ii): for the initial state it holds that $x \in \mathcal{D}(A)$ almost surely. In the latter case we obtained the same convergence speed estimate as for finite dimensional systems.

In the case of unbounded output operator, some additional assumptions were needed, including a slightly nonstandard assumption on the output operator (assumption *(iii)* in Thm. 3.5). In the problems arising from PDEs on one dimensional spatial domains, this is not a big problem but unfortunately with more complicated systems, finding a suitable γ might be close to a mission impossible. The spectral asymptotics, on the other hand, is an extensively studied field — so much so that it has even been a subject of a few books, such as [14] by Levendorskiĭ and [17] by Safarov and Vassiliev.

Some of the major topics that would require further work are adding input noise to the system and accepting infinite dimensional output space. With input noise, one should at least establish the sufficient "smoothness of the state". Also, the technique used here, based on taking into account more and more intermediate output process values, would become significantly more complicated. For what comes to the dimension of the output space in the results of this article, the

output space dimension q does not appear explicitly in the convergence speed estimates, except for Thm. 3.3. However, in the proofs we need an upper bound for $\left\| \left(C_h P_j C_h^* + \frac{h}{4} R \right)^{-1} \right\|_{\mathcal{L}(\mathcal{Y})}$ and thus, in order to obtain (10), we made a coercivity assumption $R \geq \epsilon I > 0$ which excludes infinite dimensional output space since R is required to be a trace class operator.

Two more topics that are not covered by this article are the long time behaviour as $T \to \infty$, and using some approximate time integration scheme for taking the time step. When T grows, the error covariance converges under some assumptions on the observability of the system. When there is no input noise, the limit is 0. Of course, the observability of the continuous time system does not imply the observability of the discretized system. In the case where there is input noise affecting the system, the error covariance limits are obtained as the solutions P_d and P_c of the corresponding discrete or continuous time algebraic Riccati equations, respectively. Then it holds that $\lim_{n\to\infty} \mathbb{E}\Big(\|\hat{x}_{n\Delta t,n} - \hat{x}(n\Delta t)\|_{\mathcal{X}}^2 \Big) = \operatorname{tr}(P_d - P_c)$ where $\hat{x}_{n\Delta t,n}$ and $\hat{x}(n\Delta t)$ are defined in (2). Finally, further research would be needed to study the error caused to the state estimate if some numerical time integration scheme is used for computing the discrete time update, that is, $e^{A\Delta t}$ is not computed accurately. A similar problem is addressed in [3] and [23], but they are mainly concerned with the stability of the resulting filter.

Acknowledgements. The author was financially supported by the Finnish Graduate School in Engineering Mechanics. The author thanks Dr. Jarmo Malinen for valuable comments on the manuscript.

References

[1] A. Aalto, Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems, ArXiv:1406:7160 (2014), 19 pages.
[2] W. Arendt, C. Batty, M. Hieber, and F. Neubrander, "Vector-valued Laplace Transforms and Cauchy Problems," Birkhäuser, 2001.
[3] P. Axelsson and F. Gustafsson, Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering, Technical report, Linkopings Universitetet (2012), 9 pages.
[4] R. Curtain and A. Pritchard, "Infinite Dimensional Linear Systems Theory," Springer-Verlag, 1979.
[5] G. Da Prato and A. Ichikawa, Riccati equations with unbounded coefficients, Ann. Mat. Pura Appl., 140 (1985), 209–221.
[6] G. Da Prato and J. Zabczyk, "Stochastic Equations in Infinite Dimensions," Encyclopedia of Mathematics and its Applications 44, Cambridge University Press, 1979.
[7] F. Flandoli, Direct solution of a Riccati equation arising in a stochastic control problem with control and observation on the boundary, Appl. Math. Optim., 14 (1986), 107–129.
[8] A. Gelb, "Applied Optimal Estimation," MIT Press, Cambridge, MA, 1974.
[9] A. Germani, L. Jetto, and M. Piccioni, Galerkin approximation for optimal linear filtering of infinite-dimensional linear systems, SIAM J. Control Optim., 26 (1988), 1287–1305.
[10] L. L. Horowitz, "Optimal Filtering of Gyroscopic Noise," PhD. thesis, Massachusetts Institute of Technology, 1974.

[11] R. Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering, 82 (1960), 35–45.
[12] R. Kalman and R. Bucy, New results in linear filtering and prediction theory, Journal of Basic Engineering, 83 (1961), 95–107.
[13] W. Lee, D. McDougall, and A. Stuart, Kalman filtering and smoothing for linear wave equations with model error, Inverse Problems, 27 (2011).
[14] S. Levendorskiĭ, "Asymptotic Distribution of Eigenvalues of Differential Operators," Mathematics and its Applications (Soviet Series), Kluwer Academic Publishers, 1990.
[15] J. Malinen and O. Staffans, Conservative boundary control systems, J. Differential Equations, 231 (2006), 290–312.
[16] B. O. Øksendal, "Stochastic Differential Equations: An Introduction with Applications," Springer-Verlag, 1998.
[17] Y. Safarov and D. Vassiliev, "The Asymptotic Distribution of Eigenvalues of Partial Differential Operators," Translations of mathematical monographs, American Mathematical Society, 1997.
[18] M. Salgado, R. Middleton, and G. Goodwin, Connection between continuous and discrete Riccati equations with applications to Kalman filtering, IEE Proceedings, 135 (1988), 28–34.
[19] D. Simon, "Optimal State Estimation — Kalman, H_∞, and Nonlinear Approaches," John Wiley & Sons, 2006.
[20] O. Staffans, "Well-posed Linear Systems," Encyclopedia of Mathematics and its Applications 103, Cambridge University Press, 2005.
[21] J. Sun, Sensitivity analysis of the discrete-time algebraic Riccati equation, Linear Algebra Appl., 275–276 (1998), 595–615.
[22] H. Tanabe, "Equations of Evolution," Pitman, London, 1979.
[23] N. Wahlström, P. Axelsson, and F. Gustafsson, Discretizing stochastic dynamical systems using Lyapunov equations, arXiv:1402.1358 (2014), 17 pages.
[24] G. Weiss, Admissible observation operators for linear semigroups, Israel J. Math., 65 (1989), 17–43.

E-mail address: atte.aalto@aalto.fi

Publication III

A. Aalto. Spatial discretization error in Kalman filtering for discrete-time infinite dimensional systems. http://arxiv.org/abs/1406.7160, 19 pages, October 2014.

SPATIAL DISCRETIZATION ERROR IN KALMAN FILTERING FOR DISCRETE-TIME INFINITE DIMENSIONAL SYSTEMS

ATTE AALTO

Abstract. We derive a reduced-order state estimator for discrete-time infinite dimensional linear systems with finite dimensional Gaussian input and output noise. This state estimator is the optimal one-step estimate that takes values in a fixed finite dimensional subspace of the system's state space — consider, for example, a Finite Element space. We then derive a Riccati difference equation for the error covariance and use sensitivity analysis to obtain a bound for the error of the state estimate due to the state space discretization.

1. Introduction

In this paper, we consider the state estimation problem for infinite dimensional discrete time linear systems with finite dimensional Gaussian input and output noise. The objective is to find the optimal one-step state estimate from a given subspace of the original state space (for example a Finite Element space). We shall also find a bound for the error due to the spatial discretization to the state estimate at the infinite time limit.

The dynamics of the system under consideration is given by

$$\begin{cases} x_k = Ax_{k-1} + Bu_k, \\ y_k = Cx_k + w_k, \\ x_0 \sim N(m, S_0) \end{cases} \tag{1}$$

where $x_k \in \mathcal{X}$, $A \in \mathcal{L}(\mathcal{X})$, $B \in \mathcal{L}(\mathbb{C}^q, \mathcal{X})$, and $C \in \mathcal{L}(\mathcal{X}, \mathbb{C}^m)$. The state space $\mathcal{X}$ is a separable Hilbert space. The noise processes are assumed to be Gaussian, $u_k \sim N(0, U)$ and $w_k \sim N(0, R)$ where $U \in \mathbb{R}^{q\times q}$ and $R \in \mathbb{R}^{m\times m}$ are positive-definite and symmetric. It is also assumed that u, w, and x_0 are mutually independent, and the noises at different times are independent.

When measurements y_j for $j = 1, ..., k$ are known, the state estimate $\hat{x}_k$ minimizing the conditional expectation $\mathbb{E}\Big(\|\hat{x}_k - x_k\|_{\mathcal{X}}^2 \,\Big|\, \{y_j,\, j = 1, ..., k\}\Big)$ is given by $\hat{x}_k = \mathbb{E}(x_k | \{y_j,\, j \le k\})$. In the presented Gaussian case, the conditional expectation $\hat{x}_k$ can be computed recursively from $\hat{x}_{k-1}$ and y_k. This recursive scheme is known as the Kalman filter, originally presented in [12] in the finite dimensional setting. For infinite dimensional systems, the generalization is straightforward and it can be done, for example, using the presentation by Bogachev [4: Section 3.10] or the more explicit presentation [14] by Krug. Let us present a short introduction. It is well known that linear combinations of Gaussian random variables are

2010 *Mathematics Subject Classification.* 93E11, 93E25.

Key words and phrases. Kalman filter, infinite dimensional systems, reduced-order filtering, spatial discretization, optimal estimation, Riccati equation.

also Gaussian random variables. Further, if $\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \sim N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^* & P_{22} \end{bmatrix} \right)$ where $h_1 \in \mathcal{X}$ and h_2 is finite dimensional, then

$$\mathbb{E}(h_1|h_2) = m_1 + P_{12}P_{22}^+(h_2 - m_2) \tag{2}$$

and

$$\operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right] = P_{11} - P_{12}P_{22}^+P_{12}^*. \tag{3}$$

Remark that $\operatorname{Cov}\left[\mathbb{E}(h_1|h_2)\,, \mathbb{E}(h_1|h_2)\right] = P_{12}P_{22}^+P_{12}^*$ so that in fact,

$$\begin{aligned} &\operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_1 - \mathbb{E}(h_1|h_2)\right] \\ &\qquad = \operatorname{Cov}\left[h_1, h_1\right] - \operatorname{Cov}\left[\mathbb{E}(h_1|h_2)\,, \mathbb{E}(h_1|h_2)\right]. \end{aligned} \tag{4}$$

Applying (2) and (3) to the jointly Gaussian random variable $[x_k, y_1, ..., y_k]$ and the block matrix inversion formula

$$\begin{bmatrix} F & G \\ G^T & H \end{bmatrix}^{-1} = \begin{bmatrix} F^{-1}+F^{-1}G(H-G^TF^{-1}G)^{-1}G^TF^{-1} & -F^{-1}G(H-G^TF^{-1}G)^{-1} \\ -(H-G^TF^{-1}G)^{-1}G^TF^{-1} & (H-G^TF^{-1}G)^{-1} \end{bmatrix} \tag{5}$$

to $P_{22} \mathrel{\widehat{=}} \operatorname{Cov}\left[[y_1, ..., y_k], [y_1, ..., y_k]\right]$ eventually leads to the *full state Kalman filter* equations

$$\hat{x}_k = A\hat{x}_{k-1} + K_k^{(F)}(y_k - CA\hat{x}_{k-1}) \tag{6}$$

where $K_k^{(F)}$ for $k = 1, 2, ...$ are called *Kalman gains*, and they are given by $K_k^{(F)} = \tilde{P}_k^{(F)}C^*(C\tilde{P}_k^{(F)}C^* + R)^{-1}$, and the *Riccati difference equation* (RDE)

$$\begin{cases} \tilde{P}_k^{(F)} = AP_{k-1}^{(F)}A^* + BUB^*, \\ P_k^{(F)} = \tilde{P}_k^{(F)} - \tilde{P}_k^{(F)}C^*(C\tilde{P}_k^{(F)}C^* + R)^{-1}C\tilde{P}_k^{(F)}. \end{cases} \tag{7}$$

Here $P_k^{(F)} = \operatorname{Cov}\left[x_k - \hat{x}_k, x_k - \hat{x}_k\right]$ is the *(estimation) error covariance* and $\tilde{P}_k^{(F)} = \operatorname{Cov}\left[x_k - \mathbb{E}(x_k|[y_1, ..., y_{k-1}])\,, x_k - \mathbb{E}(x_k|[y_1, ..., y_{k-1}])\right]$ is the *prediction error covariance*. The initial values are $\hat{x}_0 = m$ and $P_0^{(F)} = S_0$. The superscript (F) refers to full Kalman filter estimate and it is used for later purposes.

Numerical implementation of the Kalman filter to infinite dimensional systems requires discretization of the state space. If the implementation is then carried out directly to the discretized system, the result is not optimal. In particular, if the state estimation is performed online, the restrictions in computing power might prevent using a very fine mesh for the simulations. In such cases it is beneficial to take the discretization error into account in the state estimation. The purpose of this paper is to derive the optimal one-step state estimate that takes values in the discretized state space, and to analyze the discrepancy between the proposed state estimate and the full state Kalman filter estimate.

We tackle this task in Section 2 by first fixing the structure of the filter in (8). In the spirit of Kalman filtering, we require that the k^{th} estimate depends only on the previous estimate and the current measured output y_k. We then find the expression for a filter with such structure. The rest of the paper is organized as follows: In Section 3, we derive a Riccati difference equation for the estimation error covariance for the proposed method. Compared to (7), this equation contains an additional term due to the discretization. In Section 4, we use sensitivity analysis for algebraic Riccati equations — developed by Sun in [23] — to determine a bound for the error due to the discretization at the infinite time limit. In short, it is shown that when the approximation properties of the subspace improve at some rate as the

spatial discretization is refined, then the finite dimensional state estimate converges to the full state Kalman filter estimate at least with the same convergence rate. In Section 5, the proposed method is implemented to one dimensional wave equation with damping, and the result is compared with the Kalman filter that does not take into account the spatial discretization error.

The "engineer's approach", *i.e.*, the direct Kalman filter implementation to the discretized system is studied in [1] by Bensoussan and in [7] by Germani *et al.* The latter contains a convergence result for the finite dimensional state estimate (in continuous time) with a convergence rate estimate. They also show convergence of the solutions of the corresponding Riccati differential equations in the space of continuous Hilbert-Schmidt operator-valued functions. A method where the discretization error is taken into account is proposed by Pikkarainen in [17]. Their approach is based on keeping track of the discretization error mean and covariance. Then with certain approximations on the error distributions, they too end up with a one-step method that is numerically implemented in [11] by Huttunen and Pikkarainen.

Our method is very closely related to the reduced-order filtering methods that have been studied since the introduction of the Kalman filter itself; see *e.g.*, [2; 3; 19; 20; 22]. The articles by Bernstein and Hyland, [2; 3] yield a state estimator similar to ours for continuous time. They obtain algebraic optimality equations for the error covariance and Kalman gain limits as the time index $k \to \infty$, in terms of "optimal projections". Our solution is somewhat more straightforward, and we obtain the error covariances and Kalman gains for all time steps. A similar method is developed by Simon in [19] with a more restrictive assumption on the filter structure. For a more thorough introduction and review on the earliest results on reduced-order filtering techniques, we refer to [22] by Stubberud and Wismer and to [20] by Sims.

Infinite dimensional Kalman filter has numerous applications. The practical application that motivated the paper [17] is the electrical impedance process tomography, studied by Seppänen *et al.* in [18]. Infinite dimensional Kalman filter implementation to optical tomography problem can be found in [9] by Hiltunen *et al.* Quasiperiodic phenomena is studied by Solin and Särkkä in [21] using the infinite dimensional Kalman filter. They use a weather prediction model and fMRI brain imaging as example cases. The numerical treatment is done using truncated eigenbasis approach instead of using FEM as in the example of this article.

Notation. We denote by $\mathcal{L}(\mathcal{X}_1, \mathcal{X}_2)$ the space of bounded linear operators from $\mathcal{X}_1$ to $\mathcal{X}_2$, and $\mathcal{L}(\mathcal{X}) = \mathcal{L}(\mathcal{X}, \mathcal{X})$. The subspace of self-adjoint operators in $\mathcal{X}$ is denoted by $\mathcal{L}^*(\mathcal{X})$. The spectrum of an operator is denoted by $\sigma(\cdot)$. The sigma algebra generated by a random variable (or random variables) is denoted by $\mathcal{S}(\cdot)$. The Moore-Penrose pseudoinverse of a matrix T is denoted by T^+.

The covariance of square integrable random variables $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ is the operator in $\mathcal{L}(\mathcal{X}_2, \mathcal{X}_1)$ defined for $h \in \mathcal{X}_2$ by
$\operatorname{Cov}[x_1, x_2]\, h := \mathbb{E}\big((x_1 - \mathbb{E}(x_1))\, \langle x_2 - \mathbb{E}(x_2)\,, h\rangle_{\mathcal{X}_2}\big)$.

2. The reduced-order state estimate

Let $\Pi_s : \mathcal{X} \to \mathcal{X}$ be an orthogonal projection from the state space $\mathcal{X}$ (a separable, complex Hilbert space) to an n-dimensional subspace of $\mathcal{X}$ (*e.g.*, a finite element space). Assume we have a coordinate system in $\mathbb{C}^n$ associated to this subspace,

such that the inner product is preserved, and denote by $\Pi : \mathcal{X} \to \mathbb{C}^n$ the representation of the projection Π_s in this coordinate system. That is, $\langle \Pi_s x_1, \Pi_s x_2 \rangle_{\mathcal{X}} = \langle \Pi x_1, \Pi x_2 \rangle_{\mathbb{C}^n}$ for $x_1, x_2 \in \mathcal{X}$. Then it holds that $\Pi\Pi^* = I \in \mathbb{C}^{n\times n}$ and $\Pi^*\Pi = \Pi_s$.

Finding an exact solution to the estimation problem of the finite dimensional Πx_k would require solving the full state Kalman filtering problem and then projecting the estimate by Π. This, of course, doesn't make much practical sense. As mentioned above, we want to find the optimal state estimate $\tilde{x}_k$ in $\Pi_s\mathcal{X}$ that can be computed from the previous state estimate $\tilde{x}_{k-1}$ and the current measurement y_k. More precisely, we want to obtain $\tilde{x}_k$'s satisfying

$$\begin{cases} \tilde{x}_0 = \Pi m, \\ \tilde{x}_k = \Pi \mathbb{E}(x_k | \tilde{x}_{k-1}, y_k)\,, \qquad k \geq 1, \end{cases} \tag{8}$$

where x_k satisfy (1). One thing to notice here is that in contrast to the full state filtering, the conditioning is not done over a filtration, because — loosely speaking — we lose some information when we only take into account the last measurement and the last estimate of the state projection. Without loss of generality, we may assume that $m = 0$ (see Remark 2.1). Note that this also implies $\mathbb{E}(x_k) = 0$ and further, $\mathbb{E}(\tilde{x}_k) = 0$ and $\mathbb{E}(y_k) = 0$ for all $k \geq 1$.

We then proceed to find a concrete representation for $\tilde{x}_k$. From (8) it can be inductively deduced that $[x_{k-1}, \tilde{x}_{k-1}]$ is Gaussian and from (1), also $[x_k, \tilde{x}_{k-1}, y_k]$ is Gaussian. The reasoning leading to the full state Kalman filter equations utilizing equations (2) and (3) together with the block matrix inversion formula (5) can be generalized for any Gaussian random variable $[h_1, h_2, h_3]$ with $h_1 \in \mathcal{X}$, and h_2 and h_3 finite dimensional, to obtain

$$\begin{aligned} \mathbb{E}(h_1|h_2, h_3) =& \mathbb{E}(h_1|h_2) + \operatorname{Cov}\left[h_1 - \mathbb{E}(h_1|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right] \times \\ & \times \operatorname{Cov}\left[h_3 - \mathbb{E}(h_3|h_2)\,, h_3 - \mathbb{E}(h_3|h_2)\right]^{-1} (h_3 - \mathbb{E}(h_3|h_2)). \end{aligned} \tag{9}$$

The corresponding equation can be obtained for the covariance operator. The full state Kalman filter equations (6) and (7) are obtained by applying (9) to $h_1 = x_k$, $h_2 = [y_1, ..., y_{k-1}]$, and $h_3 = y_k$. In what follows, we obtain $\tilde{x}_k$ by applying (9) to $h_1 = x_k$, $h_2 = \tilde{x}_{k-1}$, and $h_3 = y_k$.

Since $m = 0$, there exists an operator $Q_{k-1} \in \mathcal{L}(\mathbb{R}^n, \mathcal{X})$ such that

$$\mathbb{E}(x_{k-1}|\tilde{x}_{k-1}) = Q_{k-1}\tilde{x}_{k-1} \tag{10}$$

and the *(estimation) error covariance*

$$P_{k-1} := \operatorname{Cov}\left[x_{k-1} - Q_{k-1}\tilde{x}_{k-1}, x_{k-1} - Q_{k-1}\tilde{x}_{k-1}\right]. \tag{11}$$

Using these we can make an orthogonal decomposition of the state

$$x_{k-1} = \mathbb{E}(x_{k-1}|\tilde{x}_{k-1}) + (x_{k-1} - \mathbb{E}(x_{k-1}|\tilde{x}_{k-1})) =: Q_{k-1}\tilde{x}_{k-1} + v_{k-1}$$

where $v_{k-1} \sim N\big(0, P_{k-1}\big)$ and it is independent of the estimate $\tilde{x}_{k-1}$. Together with (1), this gives decompositions for the state x_k and output y_k:

$$\begin{cases} x_k = Ax_{k-1} + Bu_k = A(Q_{k-1}\tilde{x}_{k-1} + v_{k-1}) + Bu_k, \\ y_k = Cx_k + w_k = C(A(Q_{k-1}\tilde{x}_{k-1} + v_{k-1}) + Bu_k) + w_k \end{cases} \tag{12}$$

from which one can deduce $\mathbb{E}(x_k|\tilde{x}_{k-1}) = AQ_{k-1}\tilde{x}_{k-1}$ and $\mathbb{E}(y_k|\tilde{x}_{k-1}) = CAQ_{k-1}\tilde{x}_{k-1}$.

Then we need the two covariances in (9). To this end, define the *prediction error covariance* for which we get a representation from (12),

$$\tilde{P}_k := \mathrm{Cov}\left[x_k - \mathbb{E}(x_k|\tilde{x}_{k-1})\,, x_k - \mathbb{E}(x_k|\tilde{x}_{k-1})\right] = AP_{k-1}A^* + BUB^*. \tag{13}$$

Using the two equations in (12), we get

$$\mathrm{Cov}\left[x_k - \mathbb{E}(x_k|\tilde{x}_{k-1})\,, y_k - \mathbb{E}(y_k|\tilde{x}_{k-1})\right] = \tilde{P}_k C^*$$

and the covariance of output prediction error from the second equation in (12)

$$\mathrm{Cov}\left[y_k - \mathbb{E}(y_k|\tilde{x}_{k-1})\,, y_k - \mathbb{E}(y_k|\tilde{x}_{k-1})\right] = C\tilde{P}_k C^* + R.$$

Now we have all the components for obtaining $\tilde{x}_k$ by (9),

$$\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k) = AQ_{k-1}\tilde{x}_{k-1} + \underbrace{\tilde{P}_k C^*\big(C\tilde{P}_k C^* + R\big)^{-1}}_{=:K_k}(y_k - CAQ_{k-1}\tilde{x}_{k-1}). \tag{14}$$

It remains to compute the error covariance P_k defined in (11), and the operator Q_k defined through (10). By (4), P_k is given by

$$P_k = S_k - Q_k\tilde{S}_k Q_k^*$$

where $S_k = \mathrm{Cov}\left[x_k, x_k\right]$ is the *state covariance* and $\tilde{S}_k = \mathrm{Cov}\left[\tilde{x}_k, \tilde{x}_k\right]$ is the *state estimate covariance*. The state x_k is a linear combination of mutually independent Gaussian random variables x_{k-1} and u_k and so S_k can be obtained from the Lyapunov difference equation

$$S_k = AS_{k-1}A^* + BUB^* \tag{15}$$

and the first one, S_0, is the initial state covariance in (1). Also, by (12),

$$y_k - CAQ_{k-1}\tilde{x}_{k-1} = CAv_k + CBu_k + w_k \sim N\big(0, C\tilde{P}_k C^* + R\big) \tag{16}$$

where v_k, u_k, and w_k are mutually independent and also independent with the state estimate $\tilde{x}_{k-1}$. Thus, by (14), also $\tilde{S}_k$ is obtained from a Lyapunov difference equation,

$$\tilde{S}_k = \Pi AQ_{k-1}\tilde{S}_{k-1}Q_{k-1}^*A^*\Pi^* + \Pi K_k\big(C\tilde{P}_k C^* + R\big)(\Pi K_k)^T \tag{17}$$

with $\tilde{S}_0 = 0$.

By (2), Q_k is given by

$$Q_k = \mathrm{Cov}\left[x_k, \tilde{x}_k\right]\tilde{S}_k^{-1}. \tag{18}$$

The case when $\tilde{S}_k$ is not invertible is discussed in Remark 2.2. The cross covariance operator $V_k := \mathrm{Cov}\left[x_k, \tilde{x}_k\right]$ in (18) can be computed by "anchoring" x_k and $\tilde{x}_k$ to $\tilde{x}_{k-1}$ using equations (12) and (14) and the fact that $Av_{k-1} + Bu_k \sim N\big(0, \tilde{P}_k\big)$,

$$\mathrm{Cov}\left[x_k, \tilde{x}_k\right] = AQ_{k-1}\tilde{S}_{k-1}Q_{k-1}^*A^*\Pi^* + \tilde{P}_k C^*\big(C\tilde{P}_k C^* + R\big)^{-1}C\tilde{P}_k\Pi^*.$$

It is worth noting here that $\tilde{S}_k = \Pi\mathrm{Cov}\left[x_k, \tilde{x}_k\right]$ implying the intuitive fact, $\Pi Q_k = I$ in the case that $\tilde{S}_k$ is invertible.

Let us conclude by presenting some remarks concerning the derivation of the reduced-order state estimate and then collecting the relevant equations to an algorithm.

Remark 2.1. The assumption $m = 0$ does not restrict generality, since we can always always add $\Pi A^k m$ to $\tilde{x}_k$ and subtract $CA^k m$ from y_k in (12). However, this is how to make the derivation accurate. In practical implementation, it is reasonable to just start the state estimate from $\tilde{x}_0 = \Pi m$ and then proceed as described.

Remark 2.2. If $\tilde{S}_k$ is not invertible, it means that $\mathcal{R}(\tilde{S}_k)$, the range of $\tilde{S}_k$, does not cover the whole space $\mathbb{C}^n$. The estimate $\tilde{x}_k$ lies on $\mathcal{R}(\tilde{S}_k)$ almost surely. Thus Q_k is not determined uniquely in this case. By imposing additional requirements $\Pi Q_k = I$ and $(I - \Pi_s)Q_k\big|_{\mathcal{R}(\tilde{S}_k)^\perp} = 0$ then Q_k is uniquely determined and it is given by $Q_k = \tilde{Q}_k + \Pi^*(I - \Pi\tilde{Q}_k) = \Pi^* + (I - \Pi_s)\tilde{Q}_k$ where $\tilde{Q}_k = \operatorname{Cov}[x_k, \tilde{x}_k]\,\tilde{S}_k^+$.

Algorithm 2.3. As with the full state Kalman filter, the following operator-valued equations can be computed beforehand (offline):

$$\begin{aligned}
S_k &= AS_{k-1}A^* + BUB^*,\\
\tilde{P}_k &= AP_{k-1}A^* + BUB^*,\\
K_k &= \tilde{P}_k C^*\big(C\tilde{P}_k C^* + R\big)^{-1},\\
V_k &= AQ_{k-1}\tilde{S}_{k-1}Q_{k-1}^*A^*\Pi^* + K_k\big(C\tilde{P}_k C^* + R\big)(\Pi K_k)^T,\\
\tilde{S}_k &= \Pi V_k,\\
Q_k &= \Pi^* + (I - \Pi_s)\tilde{S}_k^+ V_k,\\
P_k &= S_k - Q_k\tilde{S}_k Q_k^*.
\end{aligned}$$

The initial values are S_0 (given in (1)), $P_0 = S_0$, $\tilde{S}_0 = 0$, and $Q_0 = \Pi^*$. The state estimate is given by

$$\begin{aligned}
\tilde{x}_0 &= \Pi m,\\
\tilde{x}_k &= \Pi AQ_{k-1}\tilde{x}_{k-1} + \Pi K_k(y_k - CAQ_{k-1}\tilde{x}_{k-1}).
\end{aligned}$$

Practical implementation of the proposed method is discussed in Section 6.1. An alternative equation for P_k is derived in the following section.

3. The error covariance equation

Motivated by the main theorem of [2], we next seek for a Riccati difference equation satisfied by the error covariance P_k. This equation will be needed later for determining a bound for the error in the state estimate due to the spatial discretization. To this end, define the augmented state $\bar{x}_k := \begin{bmatrix} x_k \\ \tilde{x}_k \end{bmatrix}$ for which we have dynamic equations

$$\begin{aligned}
\begin{bmatrix} x_k \\ \tilde{x}_k \end{bmatrix} &= \begin{bmatrix} A & 0 \\ \Pi K_k CA & \Pi(A - K_k CA)Q_{k-1} \end{bmatrix}\begin{bmatrix} x_{k-1} \\ \tilde{x}_{k-1} \end{bmatrix} + \begin{bmatrix} B & 0 \\ \Pi K_k CB & \Pi K_k \end{bmatrix}\begin{bmatrix} u_k \\ w_k \end{bmatrix}\\
&=: \bar{A}_k\bar{x}_{k-1} + \bar{B}_k\bar{u}_k.
\end{aligned}$$

The augmented state covariance satisfies the Lyapunov difference equation

$$\bar{S}_k = \bar{A}_k\bar{S}_{k-1}\bar{A}_k^* + \bar{B}_k\bar{U}\bar{B}_k^* \tag{19}$$

where $\bar{U} = \begin{bmatrix} U & 0 \\ 0 & R \end{bmatrix}$. This covariance can be written as a block operator by $\bar{S}_k = \begin{bmatrix} S_k & V_k \\ V_k^* & \tilde{S}_k \end{bmatrix}$ where S_k and $\tilde{S}_k$ are the state and state estimate covariances, given in

(15) and (17), respectively. Now it holds that $Q_k = V_k\tilde{S}_k^{-1}$ (or $Q_k = V_k\tilde{S}_k^{+} + \Pi^*(I - \Pi V_k\tilde{S}_k^{+})$ if $\tilde{S}_k$ is not invertible) and thus for the reduced-order error covariance defined in (11), it holds that $P_k = S_k - V_k\tilde{S}_k^{+}V_k^*$. Also, for the prediction error covariance we have $\tilde{P}_k = A(S_k - V_k\tilde{S}_k^{+}V_k^*)A^* + BUB^*$ by (13). Using these notations we get from (19)

$$\begin{aligned} V_k &= AS_{k-1}A^*C^*K^*\Pi^* + AV_{k-1}\tilde{S}_{k-1}^{+}V_{k-1}^*A^*(\Pi - \Pi K_kC)^* + BUB^*C^*K_k^*\Pi^* \\ &= \tilde{P}_kC^*(C\tilde{P}_kC^* + R)^{-1}C\tilde{P}_k\Pi^* + AV_{k-1}\tilde{S}_{k-1}^{+}V_{k-1}^*A^*\Pi^*, \end{aligned}$$

and similarly $\tilde{S}_k = \Pi V_k = V_k^*\Pi^*$. Using the state covariance Lyapunov equation (15) and the equations above and noting that $V_k\tilde{S}_k^{+}V_k^* = Q_kV_k^* = V_kQ_k^* = Q_k\tilde{S}_kQ_k^*$, we see that the error covariance P_k satisfies the *Riccati difference equation* (RDE)

$$(20)\quad \begin{cases} \tilde{P}_k = AP_{k-1}A^* + BUB^*, \\ P_k = \tilde{P}_k - \tilde{P}_kC^*(C\tilde{P}_kC^* + R)^{-1}C\tilde{P}_k + \\ \quad +(I - Q_k\Pi)(AV_{k-1}\tilde{S}_{k-1}^{+}V_{k-1}^*A^* + \tilde{P}_kC^*(C\tilde{P}_kC^* + R)^{-1}C\tilde{P}_k)(I - Q_k\Pi)^*. \end{cases}$$

This equation is posed in $\mathcal{L}(\mathcal{X})$. Note that this is not a complete set of equations, but the last equation in Algorithm 2.3 can be replaced by the second equation in (20). Compared to the RDE (7) for the full state Kalman filter, this equation contains the additional load term in the last line of (20). In the next section we find an upper bound for the effect of this additional term to the solution at the infinite time limit but first we need to go through some auxiliary results.

Proposition 3.1. *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be sigma algebras, such that $\mathcal{S}_1 \subset \mathcal{S}_2$ and x an integrable random variable. Then $\mathbb{E}(x|\mathcal{S}_1) = \mathbb{E}(\mathbb{E}(x|\mathcal{S}_2)\,|\mathcal{S}_1)$.*

If x is quadratically integrable then

$$\operatorname{Cov}\left[\mathbb{E}(x|\mathcal{S}_1)\,, \mathbb{E}(x|\mathcal{S}_1)\right] \le \operatorname{Cov}\left[\mathbb{E}(x|\mathcal{S}_2)\,, \mathbb{E}(x|\mathcal{S}_2)\right] \le \operatorname{Cov}\left[x, x\right].$$

Lemma 3.2. *Assume that the state covariance S_k defined in* (15) *satisfies $S_k \le S$ for all k for some trace class operator $S \in \mathcal{L}^*(\mathcal{X})$. For the discretization error term in the RDE* (20)*, it holds that*

$$\begin{aligned} M_k &:= (I - Q_k\Pi)(AV_{k-1}\tilde{S}_{k-1}^{+}V_{k-1}^*A^* + \tilde{P}_kC^*(C\tilde{P}_kC^* + R)^{-1}C\tilde{P}_k)(I - Q_k\Pi)^* \\ &\le (I - \Pi_s)S(I - \Pi_s)^* =: M. \end{aligned} \tag{21}$$

Proof. Note that $V_{k-1}\tilde{S}_{k-1}^{+}V_{k-1}^* = Q_{k-1}\tilde{S}_{k-1}Q_{k-1}^*$. Then by (14) and (16) it can be seen that

$$M_k = \operatorname{Cov}\left[(I - Q_k\Pi)\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\,, (I - Q_k\Pi)\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\right].$$

It holds that

$$Q_k\Pi\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k) = Q_k\tilde{x}_k = \mathbb{E}(x_k|\tilde{x}_k) = \mathbb{E}(\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\,|\tilde{x}_k)$$

where the first equality follows by (8), the second by the definition of Q_k, (10), and the third by Proposition 3.1 and $\mathcal{S}(\tilde{x}_k) \subset \mathcal{S}(\tilde{x}_{k-1}, y_k)$ which, in turn, can be seen from (8).

Thus Q_k minimizes

$$\begin{aligned} &\mathbb{E}\Big(\langle e, \mathbb{E}(x_k|\tilde{x}_{k-1}, y_k) - Z\tilde{x}_k\rangle_{\mathcal{X}}^2\Big) = \mathbb{E}\Big(\langle e, (I - Z\Pi)\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\rangle_{\mathcal{X}}^2\Big) \\ &= \langle e, (I - Z\Pi)\operatorname{Cov}\left[\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\,, \mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\right](I - Z\Pi)^*e\rangle_{\mathcal{X}} \end{aligned}$$

over $Z \in \mathcal{L}(\mathbb{C}^n, \mathcal{X})$ for all $e \in \mathcal{X}$. Since $\Pi_s = \Pi^*\Pi$, it holds that

$$\begin{aligned} M_k &\leq (I - \Pi_s)\mathrm{Cov}\left[\mathbb{E}(x_k|\tilde{x}_{k-1}, y_k), \mathbb{E}(x_k|\tilde{x}_{k-1}, y_k)\right](I - \Pi_s)^* \\ &\leq (I - \Pi_s)\mathrm{Cov}\left[x_k, x_k\right](I - \Pi_s)^* \leq M \end{aligned}$$

where the middle inequality holds by Proposition 3.1. □

Lemma 3.3. *Let $P_k^{(j)}$ for $j = 1, 2$, be the solutions of the RDEs*

$$\begin{cases} \tilde{P}_k^{(j)} = AP_{k-1}^{(j)}A^* + W_k^{(j)}, \\ P_k^{(j)} = \tilde{P}_k^{(j)} - \tilde{P}_k^{(j)}C^*(C\tilde{P}_k^{(j)}C^* + R)^{-1}C\tilde{P}_k^{(j)} \end{cases} \tag{22}$$

where $P_0^{(2)} \geq P_0^{(1)} \geq 0$ and $W_k^{(2)} \geq W_k^{(1)} \geq 0$. Then $P_k^{(2)} \geq P_k^{(1)}$ for all $k \geq 0$.

This follows from [6: Lemma 3.1] by de Souza in the finite dimensional setting. The proof is just algebraic manipulation and it holds also in the infinite dimensional setting (if the output is finite dimensional). However, we shall present a straightforward proof.

Proof. We show $P_1^{(2)} \geq P_1^{(1)}$. For larger k the result follows by induction. Define the block diagonal covariances in $\mathcal{L}^*(\mathcal{X}^3)$

$$\tilde{P}_B^{(1)} = \begin{bmatrix} AP_0^{(1)}A^* & & \\ & W_1^{(1)} & \\ & & 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{P}_B^{(2)} = \begin{bmatrix} AP_0^{(2)}A^* & & \\ & W_1^{(1)} & \\ & & W_1^{(2)} - W_1^{(1)} \end{bmatrix}$$

and $C_B := [C\ C\ C]$. Then define

$$\begin{aligned} P_B^{(j)} &= \tilde{P}_B^{(j)} - \tilde{P}_B^{(j)}C_B^*(C_B\tilde{P}_B^{(j)}C_B^* + R)^{-1}C_B\tilde{P}_B^{(j)} \qquad \text{for } j = 1, 2 \\ P_B^{(\times)} &= \tilde{P}_B^{(2)} - \tilde{P}_B^{(2)}C_B^*(C_B\tilde{P}_B^{(1)}C_B^* + R)^{-1}C_B\tilde{P}_B^{(2)}. \end{aligned}$$

Now $\tilde{P}_B^{(2)} \geq \tilde{P}_B^{(1)}$ implies $P_B^{(2)} \geq P_B^{(\times)}$. Then $P_B^{(1)} = \begin{bmatrix} I & & \\ & I & 0 \end{bmatrix} P_B^{(\times)} \begin{bmatrix} I & & \\ & I & \\ 0 & & \end{bmatrix}$ and so $P_B^{(\times)} \geq P_B^{(1)}$. Now $P_1^{(j)} = [I\ I\ I]P_B^{(j)} \begin{bmatrix} I \\ I \\ I \end{bmatrix}$ and so $P_1^{(2)} \geq P_1^{(1)}$. □

The following lemma is due to Hager and Horowitz, [8]:

Lemma 3.4. *Assume that $S_k \leq S$ for all k for some trace class operator $S \in \mathcal{L}^*(\mathcal{X})$ where S_k is defined in (15). Let $P_k^{(F)}$ be the solution of (7) and $P_k^{(b)}$ be the solution of (22) with $W_k^{(b)} = W^{(b)} = BUB^* + AMA^*$ where M is defined in (21). Assuming $P_0^{(b)} = P_0^{(F)} = 0$, then $P_k^{(b/F)} \to P^{(b/F)}$ strongly as $k \to \infty$. Also, the limit operators $P^{(b/F)} \geq 0$ are the unique nonnegative solutions of the discrete time algebraic Riccati equation (DARE)*

$$\begin{cases} \tilde{P}^{(b/F)} = AP^{(b/F)}A^* + W^{(b/F)}, \\ P^{(b/F)} = \tilde{P}^{(b/F)} - \tilde{P}C^*(C\tilde{P}^{(b/F)}C^* + R)^{-1}C\tilde{P}^{(b/F)} \end{cases} \tag{23}$$

where $W^{(F)} = BUB^*$.

If $\sigma(A - K^{(F)}CA) \subset B(0, \rho)$ with $\rho < 1$ where $K^{(F)}$ is the limit of the full state Kalman gain, that is

$$K^{(F)} = \tilde{P}^{(F)}C^*(C\tilde{P}^{(F)}C^* + R)^{-1}, \tag{24}$$

then $P_k^{(F)} \to P^{(F)}$ strongly, starting from any $P_0^{(F)} \geq 0$.

The first part follows from [8: Theorem 1] because $P_k^{(j)} \leq S$, and the second part from [8: Theorem 3].

Even the weak convergence would suffice for the dominated convergence of trace class operators:

Lemma 3.5. *If P, S, and P_k for $k = 0, 1, \ldots$ are trace class operators in $\mathcal{L}^*(\mathcal{X})$, $P_k \leq S$ for all k, and $P_k \xrightarrow{w} P$, then $\operatorname{tr}(P_k) \to \operatorname{tr}(P)$.*

The proof is rather straightforward after noting that $\langle e_j, P_k e_j \rangle_{\mathcal{X}} \to \langle e_j, P e_j \rangle_{\mathcal{X}}$ as $k \to \infty$, for all $j \in \mathbb{N}$ where $\{e_j\}_{j\in\mathbb{N}}$ is an orthonormal basis for $\mathcal{X}$.

4. Error analysis

Next we use sensitivity analysis for DAREs and the results of the preceding section to show a bound for the discrepancy $\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big)$ of the full and reduced-order state estimates, defined in (6) and (8), respectively. The results of this section are based on bounding the effect of the perturbation M_k in (21) caused by the spatial discretization. Such bound is possible if we have additional information about the smoothness of the state x_k. That is, it is assumed that x_k lies in a subspace $\mathcal{X}_1$ of $\mathcal{X}$ — which is a Hilbert space itself — and that the projection Π_s approximates well the vectors in that subspace, meaning that the norm $\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}$ becomes small as the spatial discretization is refined.

We show two theorems — first (Thm. 4.1) is an *a priori* type estimate on the convergence rate of $\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big)$, and the second (Thm. 4.2) is an *a posteriori* estimate of the error $\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big)$.

Theorem 4.1. *Consider the system (1) and the reduced order state estimator $Q_k\tilde{x}_k$ derived in Sections 2 and 3. Make the following assumptions:*

(i) *$x_k \in \mathcal{X}_1$ a.s. for all k where $\mathcal{X}_1$ is a Hilbert space that is a vector subspace of $\mathcal{X}$ and $\sup_k \mathbb{E}\Big(\|x_k\|_{\mathcal{X}_1}^2\Big) < \infty$.*

(ii) *The state covariance S_k defined in (15) converges to the solution of the Lyapunov equation $S = ASA^* + BUB^*$, that is, $S = \sum_{j=0}^{\infty} A^j BUB^*(A^*)^j$ and $S_k \leq S$ for all $k \geq 0$. Use this S in the definition of M in (21).*

(iii) *The converged full state Kalman filter is exponentially stable, meaning $\sigma(A - K^{(F)}CA) \subset B(0,\rho)$ for some $\rho < 1$ where $K^{(F)}$ is the Kalman gain of the converged full state Kalman filter, introduced in (24).*

If $\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}$ is small enough, it holds that

$$\limsup_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) \leq C\,\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}^2 + \mathcal{O}\Big(\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}^4\Big)$$

where $C = \left(1 + L\left\|A - K^{(F)}CA\right\|_{\mathcal{L}(\mathcal{X})}^2\right) \sup_k \mathbb{E}\Big(\|x_k\|_{\mathcal{X}_1}^2\Big)$ and L is defined in Lemma A.1.

Proof. Assume first that the initial state is completely known, that is, $S_0 = 0$. Let P_k be the error covariance of the reduced order method, satisfying the RDE (20) and M_k be defined in (21). It is easy to confirm that the *shifted covariance* $P_k^{(a)} := P_k - M_k$ satisfies the RDE

$$\begin{cases} \tilde{P}_k^{(a)} = AP_{k-1}^{(a)}A^* + BUB^* + AM_kA^*, \\ P_k^{(a)} = \tilde{P}_k^{(a)} - \tilde{P}_k^{(a)}C^*\big(C\tilde{P}_k^{(a)}C^* + R\big)^{-1}C\tilde{P}_k^{(a)}. \end{cases}$$

Then denote by $P_k^{(b)}$ and $\tilde{P}_k^{(b)}$ the solution of a similar RDE but with the term AM_kA^* replaced by AMA^* where M is the upper bound for M_k, defined in (21). Finally, let $P_k^{(F)}$ be the error covariance of the full Kalman filter estimate, given in (7) and $\hat{x}_k = \mathbb{E}(x_k|\{y_j, j \le k\})$ is given in (6).

By computing the trace of both sides of (4), we see that for a Gaussian random variable $[h_1, h_2]$ it holds that

$$\mathbb{E}\Big(\|h_1\|_{\mathcal{X}}^2\Big) = \mathbb{E}\Big(\|\mathbb{E}(h_1|h_2)\|_{\mathcal{X}}^2\Big) + \mathbb{E}\Big(\|h_1 - \mathbb{E}(h_1|h_2)\|_{\mathcal{X}}^2\Big).$$

Now $\tilde{x}_k$ depends linearly on $[y_1, ..., y_k]$ and thus clearly $\mathcal{S}(\tilde{x}_k) \subset \mathcal{S}(y_1, ..., y_k)$. By Proposition 3.1, it holds that $Q_k\tilde{x}_k = \mathbb{E}(x_k|\tilde{x}_k) = \mathbb{E}(\hat{x}_k|\tilde{x}_k)$. Thus it holds that

$$\begin{aligned}
\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) &= \mathbb{E}\Big(\|\hat{x}_k\|_{\mathcal{X}}^2\Big) - \mathbb{E}\Big(\|Q_k\tilde{x}_k\|_{\mathcal{X}}^2\Big) \\
&= \mathbb{E}\Big(\|x_k\|_{\mathcal{X}}^2\Big) - \mathbb{E}\Big(\|Q_k\tilde{x}_k\|_{\mathcal{X}}^2\Big) - \Big(\mathbb{E}\Big(\|x_k\|_{\mathcal{X}}^2\Big) - \mathbb{E}\Big(\|\hat{x}_k\|_{\mathcal{X}}^2\Big)\Big) \\
&= \mathbb{E}\Big(\|x_k - Q_k\tilde{x}_k\|_{\mathcal{X}}^2\Big) - \mathbb{E}\Big(\|x_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) = \operatorname{tr}\big(P_k^{(a)}\big) + \operatorname{tr}(M_k) - \operatorname{tr}\big(P_k^{(F)}\big).
\end{aligned}$$

By Lemmas 3.2 and 3.3, $P_k^{(F)} \le P_k^{(a)} \le P_k^{(b)}$ and thus $\operatorname{tr}\big(P_k^{(a)}\big) - \operatorname{tr}\big(P_k^{(F)}\big) \le \operatorname{tr}\big(P_k^{(b)}\big) - \operatorname{tr}\big(P_k^{(F)}\big)$. By Lemma 3.4, $P_k^{(b)} \to P^{(b)}$ and $P_k^{(F)} \to P^{(F)}$ strongly (recall $S_0 = 0$) where $P^{(b)}$ and $P^{(F)}$ are the solutions of the corresponding DAREs, that is, equation (23) with $W^{(b)} = BUB^* + AMA^*$ and $W^{(F)} = BUB^*$. Also, by Lemma 3.5, $\operatorname{tr}\big(P_k^{(b)}\big) \to \operatorname{tr}\big(P^{(b)}\big)$ and $\operatorname{tr}\big(P_k^{(F)}\big) \to \operatorname{tr}\big(P^{(F)}\big)$. Denote $\Delta P := P^{(b)} - P^{(F)}$ and note that $\Delta P \in \mathcal{L}^*(\mathcal{X})$ is a positive (semi-)definite trace class operator. Then an upper bound for the discrepancy is given by

$$\limsup_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) \le \operatorname{tr}(\Delta P) + \operatorname{tr}(M). \tag{25}$$

Equation (30) in Lemma A.2 gives a representation for ΔP. The next step is to use this equation to find a bound for $\operatorname{tr}(\Delta P)$. Because the full Kalman filter is assumed to be exponentially stable, by Lemmas A.1 and A.2, we have

$$\operatorname{tr}(\Delta P) \le \operatorname{tr}\big(\mathbf{L}^{-1}(E_1 + E_2 + h_1(\Delta P))\big)$$

where $\mathbf{L} \in \mathcal{L}(\mathcal{L}^*(\mathcal{X}))$ is defined in Lemma A.1 and E_1, E_2, and $h_1(\Delta P)$ are defined in Lemma A.2. The term $h_2(\Delta P)$ in (30) is excluded here because it is negative definite (see the discussion after Lemma A.1).

Now we have $E_1 \ge 0$ and so by Lemma A.1,

$$\operatorname{tr}(\mathbf{L}^{-1}E_1) \le L\operatorname{tr}(E_1) \le L\left\|A - K^{(F)}CA\right\|_{\mathcal{L}(\mathcal{X})}^2 \operatorname{tr}(M)$$

where L is defined in Lemma A.1. From E_2 the negative definite part can be omitted and thus

$$\begin{aligned}
\operatorname{tr}(\mathbf{L}^{-1}E_2) &\le L\left\|K^{(F)}C\right\|_{\mathcal{L}(\mathcal{X})}^2 \operatorname{tr}\left(AMA^*C^*\left(C(\tilde{P}^{(F)} + AMA^*)C^* + R\right)^{-1}CAMA^*\right) \\
&\le L\left\|K^{(F)}C\right\|_{\mathcal{L}(\mathcal{X})}^2 \|A\|_{\mathcal{L}(\mathcal{X})}^4 \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}^2 \operatorname{tr}\left(\left(C(\tilde{P}^{(F)} + AMA^*)C^* + R\right)^{-1}\right) \operatorname{tr}(M)^2 \\
&\le L\left\|K^{(F)}C\right\|_{\mathcal{L}(\mathcal{X})}^2 \|A\|_{\mathcal{L}(\mathcal{X})}^4 \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}^2 \operatorname{tr}\left(\left(C\tilde{P}^{(F)}C^* + R\right)^{-1}\right) \operatorname{tr}(M)^2.
\end{aligned}$$

To get a bound for $\operatorname{tr}\left(\mathbf{L}^{-1}h_1(\Delta P)\right)$, recall the following properties of the operator trace and the Hilbert-Schmidt norm:

$$\|AB\|_{HS} \le \|A\|_{\mathcal{L}(\mathcal{X})} \|B\|_{HS}, \quad \|A\|_{HS} \le \operatorname{tr}(A), \text{ for } A \in \mathcal{L}^*(\mathcal{X}),\ A \ge 0,$$

$$\text{and} \quad \operatorname{tr}(AB) \le \|A\|_{HS} \|B\|_{HS}.$$

Using these and (29) yields $\operatorname{tr}\left(\mathbf{L}^{-1}h_1(\Delta P)\right)$

$$\le L_0 \left(2 \left\| A - K^{(F)}CA \right\|_{\mathcal{L}(\mathcal{X})} \|CA\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \|\Delta K\|_{HS} + \|CA\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \|\Delta K\|^2_{HS} \right) \operatorname{tr}(\Delta P)$$

where L_0 is defined in Lemma A.1, $\Delta K = K^{(F)} - K^{(b)}$, and $K^{(b)} = \tilde{P}^{(b)}C^*(C\tilde{P}^{(b)}C^* + R)^{-1}$. By the last part of Lemma A.2, we have

$$\|\Delta K\|_{HS} \le \left(\hat{c}_1 + \hat{c}_2 \operatorname{tr}(M)\right) \operatorname{tr}(M) \tag{26}$$

where

$$\hat{c}_1 = \left(1 + \left\| \tilde{P}^{(F)} \right\|_{\mathcal{L}(\mathcal{X})} \|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \left\| \left(C\tilde{P}^{(F)}C^* + R \right)^{-1} \right\|_{\mathcal{L}(\mathcal{Y})} \right) \times$$
$$\times \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \|A\|^2_{\mathcal{L}(\mathcal{X})} \left\| \left(C\tilde{P}^{(F)}C^* + R \right)^{-1} \right\|_{\mathcal{L}(\mathcal{Y})}$$

and
$$\hat{c}_2 = \|A\|^4_{\mathcal{L}(\mathcal{X})} \|C\|^3_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \left\| \left(C\tilde{P}^{(F)}C^* + R \right)^{-1} \right\|^2_{\mathcal{L}(\mathcal{Y})}.$$

Collecting these inequalities we finally get

$$\operatorname{tr}(\Delta P) \le \frac{a\operatorname{tr}(M) + b\operatorname{tr}(M)^2}{1 - \left(c_1\operatorname{tr}(M) + c_2\operatorname{tr}(M)^2 + c_3\operatorname{tr}(M)^3 + c_4\operatorname{tr}(M)^4\right)} \tag{27}$$

where

$$\begin{aligned}
a &= L \left\| A - K^{(F)}CA \right\|^2_{\mathcal{L}(\mathcal{X})}, \\
b &= L \left\| K^{(F)}C \right\|^2_{\mathcal{L}(\mathcal{X})} \|A\|^4_{\mathcal{L}(X)} \|C\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \operatorname{tr}\left(\left(C\tilde{P}^{(F)}C^* + R \right)^{-1} \right), \\
c_1 &= 2L_0 \left\| A - K^{(F)}CA \right\|_{\mathcal{L}(\mathcal{X})} \|CA\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \hat{c}_1, \\
c_2 &= 2L_0 \left\| A - K^{(F)}CA \right\|_{\mathcal{L}(\mathcal{X})} \|CA\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \hat{c}_2 + L_0 \|CA\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \hat{c}_1^2, \\
c_3 &= 2L_0 \|CA\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \hat{c}_1\hat{c}_2, \\
c_4 &= L_0 \|CA\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})} \hat{c}_2^2.
\end{aligned}$$

To complete the proof under the assumption $S_0 = 0$, use (25), (27), and note that by the definition of M in (21) and S in assumption *(ii)*,

$$\operatorname{tr}(M) = \sup_k \mathbb{E}\left(\|(I - \Pi_s)x_k\|^2_{\mathcal{X}} \right) \le \|I - \Pi_s\|^2_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})} \sup_k \mathbb{E}\left(\|x_k\|^2_{\mathcal{X}_1} \right).$$

In case $S_0 > 0$, the convergence $P_k^{(b)} \to P^{(b)}$ has to be established. Denote $\Phi = A - K^{(F)}CA$ and $\Delta\Phi = \Delta K CA$. Pick $\lambda \in \mathbb{C}$ from the resolvent set of Φ. Then

using the Woodbury formula, we get

$$\begin{aligned}\Big(\lambda - (A - K^{(b)}CA)\Big)^{-1} &= (\lambda - \Phi - \Delta\Phi)^{-1} \\ &= (\lambda-\Phi)^{-1} + (\lambda-\Phi)^{-1}\Delta\Phi\left(I - (\lambda-\Phi)^{-1}\Delta\Phi\right)^{-1}(\lambda-\Phi)^{-1}\end{aligned}$$

and

$$\left\|(\lambda-\Phi)^{-1}\Delta\Phi\right\|_{\mathcal{L}(\mathcal{X})} \leq \frac{\|\Delta\Phi\|_{\mathcal{L}(\mathcal{X})}}{|\lambda|-\rho}$$

where $\rho < 1$ is the spectral radius of Φ. The invertibility of $\lambda - (A - K^{(b)}CA)$ is then guaranteed if $\|\Delta KCA\|_{\mathcal{L}(\mathcal{X})} < |\lambda| - \rho$ which implies that the spectral radius of $A - K^{(b)}CA$ is at most $\rho + \|\Delta KCA\|_{\mathcal{L}(\mathcal{X})}$. So when $\mathrm{tr}(M)$ is small enough, then also $A - K^{(b)}CA$ is exponentially stable and $P_k^{(b)} \to P^{(b)}$ strongly. □

The assumption *(iii)* in Theorem 4.1 is very difficult to check. Also, it is hard to say what it means that "$\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}$ *is small enough*" which is related to the denominator in Eq. (27) and the exponential stability of $A - K^{(b)}CA$. Consequently, this theorem should be considered as an *a priori* convergence speed estimate when the discretization is refined, that is, when $\|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})} \to 0$.

However, if one has already computed the operators Q_k and K_k and they have converged to Q_∞ and K_∞ and it has turned out that $\sigma(A - K_\infty CA) \subset B(0,\rho)$ for some $\rho < 1$, then by the same argument as in Theorem 4.1 we get the following improved error estimate:

Theorem 4.2. *Make the assumptions (i) and (ii) in Theorem 4.1. Assume also that the operators K_k, Q_k, and M_k related to the reduced order filter have converged to K_∞, Q_∞ and M_∞, respectively, and $\sigma(A - K_\infty CA) \subset B(0,\rho)$ for some $\rho < 1$. Then*

$$\limsup_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) \leq C_1 \|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}^2 + C_2 \|I - \Pi_s\|_{\mathcal{L}(\mathcal{X}_1,\mathcal{X})}^4$$

where $C_1 = \left(1 + \tilde{L}\left\|A - K^{(F)}CA\right\|_{\mathcal{L}(\mathcal{X})}^2\right)\sup_k \mathbb{E}\Big(\|x_k\|_{\mathcal{X}_1}^2\Big)$,

$$C_2 = \tilde{L}\left\|K^{(F)}C\right\|_{\mathcal{L}(\mathcal{X})}^2 \|A\|_{\mathcal{L}(\mathcal{X})}^4 \|C\|_{\mathcal{L}(\mathcal{X},\mathcal{Y})}^2 \,\mathrm{tr}\left(\left(C\tilde{P}^{(F)}C^* + R\right)^{-1}\right)\left(\sup_k \mathbb{E}\Big(\|x_k\|_{\mathcal{X}_1}^2\Big)^2\right),$$

and $\tilde{L}$ is defined in Lemma A.1.

Proof. The covariances $P_k^{(a)}$ and $\tilde{P}_k^{(a)}$ defined in the proof of Theorem 4.1 converge to $P^{(a)}$ and $\tilde{P}^{(a)}$ that are the solution of the DARE

$$\begin{cases}\tilde{P}^{(a)} = AP^{(a)}A^* + BUB^* + AM_\infty A^*, \\ P^{(a)} = \tilde{P}^{(a)} - \tilde{P}^{(a)}C^*\big(C\tilde{P}^{(a)}C^* + R\big)^{-1}C\tilde{P}^{(a)}.\end{cases}$$

Now bounding $\Delta P := P^{(a)} - P^{(F)}$ by using the alternative expression (31) for ΔP given in Lemma A.2 and otherwise proceeding as in the proof of Theorem 4.1 leads to the result. Note that

$$K_\infty = \lim_{k\to\infty} \tilde{P}_k C^*(C\tilde{P}_k C^* + R)^{-1}$$

but since $\tilde{P}_k^{(a)} = \tilde{P}_k$ for all k, it holds that $K_\infty = \tilde{P}^{(a)}C^*(C\tilde{P}^{(a)}C^* + R)^{-1}$. □

Remark 4.3. The coefficients C_1 and C_2 in the above theorem depend on $K^{(F)}$ and $\tilde{P}^{(F)}$ which is not desirable. It is possible to bound these coefficients from above without computing them. Firstly, we have

$$\left\|A - K^{(F)}CA\right\|^2_{\mathcal{L}(\mathcal{X})} \leq 2\left\|A - K_\infty CA\right\|^2_{\mathcal{L}(\mathcal{X})} + 2\left\|CA\right\|^2_{\mathcal{L}(\mathcal{X},\mathcal{Y})}\left\|\Delta K\right\|^2_{\mathcal{L}(\mathcal{Y},\mathcal{X})}.$$

Now $\left\|\Delta K\right\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \leq \left\|\Delta K\right\|_{HS}$ for which we have (26), $\left\|\tilde{P}^{(F)}\right\|_{\mathcal{L}(\mathcal{X})} \leq \left\|\tilde{P}^{(a)}\right\|_{\mathcal{L}(\mathcal{X})}$, $\left\|\left(C\tilde{P}^{(F)}C^* + R\right)^{-1}\right\|_{\mathcal{L}(\mathcal{Y})} \leq \frac{1}{\min(\mathrm{eig}(R))}$, and $\mathrm{tr}\left(\left(C\tilde{P}^{(F)}C^* + R\right)^{-1}\right) \leq \mathrm{tr}(R^{-1})$.

5. Numerical example

In this section, Algorithm 2.3 is implemented to the temporally discretized 1D wave equation with damping,

$$\begin{cases} \frac{\partial^2}{\partial t^2}z(x,t) = -\epsilon\frac{\partial}{\partial t}z(x,t) + \frac{\partial^2}{\partial x^2}z(x,t) + Bu(t), & x \in [0,1], \\ z(0,t) = z(1,t) = 0, \\ y(t) = Cz(x,t) + w(t), \\ z(x,0) = z_0 \end{cases} \tag{28}$$

where $u \in \mathbb{R}^3$ and $w \in \mathbb{R}^2$ are the formal derivatives of Brownian motions with incremental covariances U and R, respectively. The initial state is a Gaussian random variable $z_0 \sim N(0, P_0)$, and u, w, and z_0 are mutually independent. The input operator B is a multiplication operator but we define its structure only on the discrete-time level. The output operator $C \in \mathcal{L}(\mathcal{X}, \mathbb{R}^2)$ is given by $Cz = \left[\langle c_1, z\rangle_{L^2(0,1)}, \langle c_2, z\rangle_{L^2(0,1)}\right]^T$ where $c_1(x) = \frac{1.4}{(x+1)\cdot 7}$ and $c_2(x) = \frac{1}{(2-x)\cdot 3}$.

The equation is transformed to a first order differential equation with respect to the time variable by introducing the augmented state $\begin{bmatrix} z \\ v \end{bmatrix}$ where $v = \frac{\partial}{\partial t}z$ is the velocity variable. The natural augmented state space is $\mathcal{X} = H^1_0[0,1] \times L^2(0,1)$. In $H^1_0[0,1] := \{z \in H^1[0,1] \,|\, z(0) = z(1) = 0\}$ we use the norm $\|z\|^2_{H^1_0[0,1]} := \int_0^1 z'(x)^2 dx$. The equation is then temporally discretized using the implicit Euler method with time step Δt. The state space discretization is carried out by Finite Element Method using piecewise linear elements on two meshes on the interval $[0,1]$. The first one is a finer mesh with N_f equispaced discretization points. The fine mesh solution is regarded as the true solution. The second, coarse mesh consists of N_c discretization points, also equally spaced with discretization intervals of length $h_c = 1/(N_c+1)$. It is required that the function space consisting of the piecewise linear elements on the coarse mesh is a subspace of the fine mesh space. This is satisfied when $N_f + 1 = k(N_c + 1)$ for some integer k. The coarse mesh space is the range of Π. In the augmented state of the discretized system, the input operator is $B_d = \begin{bmatrix} 0 & 0 & 0 \\ b_1(x) & b_2(x) & b_3(x) \end{bmatrix}$ where $b_1(x) = (1-x)\sin(\pi x)$, $b_2(x) = 7x^2(1-x)$, and $b_3(x) = \sin(6\pi x)^2/x$. The input noise covariance for the discrete time system is $U_d = \Delta t U$.

The solution of (28) actually has additional smoothness, namely $[z\ v]^T \in \mathcal{X}_1 = \left(H^1_0[0,1] \cap H^2[0,1]\right) \times H^1_0[0,1]$ almost surely — note that $B_d \in \mathcal{L}(\mathbb{R}^3, \mathcal{X}_1)$. It is well known that the piecewise linear elements approximate H^2-functions in one

TABLE 1. Left: Simulation parameters. Right: Squared error averages over 500 simulations.

Symbol	value
Δt	.01
U	diag(1 , 1 , .25)
R	diag(.3 , .15)
N_f	65
N_c	5
ϵ	.4

Method	F	A	C
Position	.6122	.6126	.6352
Velocity	.8150	.8154	.9294

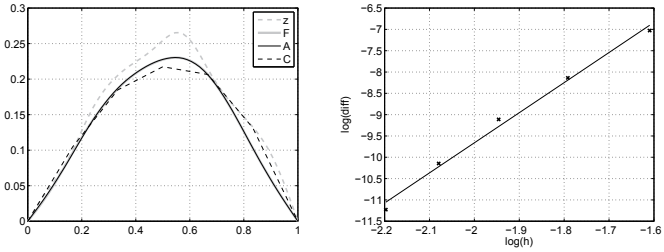

FIGURE 1. Left: The true solution and estimates given by the three filtering methods. Right: The convergence of $\lim_{k\to\infty} \mathbb{E}\Big(\|Q\tilde{x}_k - \hat{x}_k\|^2_{\mathcal{X}}\Big)$ as $h_c \to 0$ is shown with the x-markers. The solid line is a fitted regression line. The plot is in logarithmic scale.

dimension so that $\|z - \Pi_s z\|_{H^1[0,1]} \leq C_2 h_c \|z\|_{H^2[0,1]}$ and H^1-functions so that $\|v - \Pi_s v\|_{L^2(0,1)} \leq C_1 h_c \|v\|_{H^1[0,1]}$, see for example [15: Section 5.1].

Fig. 1 (left) shows the state $z(x,t)$ together with the three different state estimates in one simulation. The full state Kalman filter estimate (F) and the reduced-order state estimate (A) cannot be distinguished from each other. The third state estimate (C) is computed in the coarse mesh without taking the discretization error into account. The simulation parameters are shown in Table 1 (left). The spectral radius was .996 for both the full state Kalman filter and the reduced-order filter. We are interested in the stationary Kalman filter and so the simulations were first run 2000 steps to get rid of initial transitions. The expected (squared) errors of the different methods are shown in Table 1 (right) separately for the position variable z and the velocity variable v.

As $h_c \to 0$, the expected squared difference between the reduced-order estimate and full state Kalman filter estimate, $\lim_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|^2_{\mathcal{X}}\Big)$, tends to zero. Fig. 1 (right) illustrates this convergence in the example case. Regression analysis gives $\lim_{k\to\infty} \mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|^2_{\mathcal{X}}\Big) \approx 86.8h^{7.06}$ whereas Theorem 4.1 gives $\mathcal{O}(h^2)$ convergence rate.

6. Conclusions and remarks

When the system at hand is infinite dimensional (or its dimension is very large), one needs to make some finite (or lower) dimensional approximation of the system in order to be able to actually compute something. For what comes to the Gaussian state estimation problem, the spatial discretization introduces a bias in the Kalman filter but the result can be improved by taking that error into account when determining the Kalman gain.

In this paper, we derived the optimal one-step state estimator $\tilde{x}_k$ for an infinite dimensional system that takes values in a pre-defined finite dimensional subspace $\Pi_s\mathcal{X}$ of the system's state space $\mathcal{X}$. The presented method also gives an operator Q_k that gives $\mathbb{E}(x_k|\tilde{x}_k) = Q_k\tilde{x}_k$. This operator can be used as a sort of post-processor of the obtained state estimate.

Sections 3 and 4 were devoted to finding a bound for the error caused by the discretization. The error measure is the $L^2(\Omega,\mathcal{X})$-distance between the reduced-order state estimate $Q_k\tilde{x}_k$ and the full state Kalman filter estimate $\hat{x}_k$, that is, $\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big)$. It was found that this distance converges to zero as the approximation abilities of the projection Π_s improve.

A numerical example on temporally discretized 1D wave equation was presented in Section 5. It was noted that the presented method worked well even with fairly low level of discretization. The spatial discretization was done using piecewise linear hat functions whose approximating properties were noted to converge with rate $\mathcal{O}(h)$ when the discretization is refined. By Theorem 4.1 this would imply convergence rate $\mathbb{E}\Big(\|Q_k\tilde{x}_k - \hat{x}_k\|_{\mathcal{X}}^2\Big) = \mathcal{O}(h^2)$ for the reduced-order state estimate. However, numerical simulations showed that this convergence was actually of order $\mathcal{O}(h^7)$ in the example case.

6.1. **On practical implementation.**

Even though all the computations needed for the update of the state estimate are carried out in the finite dimensional subspace $\Pi_s\mathcal{X}$ in the presented method, the offline computations needed for determining the Kalman gains K_k and the operators Q_k are still formally carried out in the infinite dimensional $\mathcal{X}$. In practice, there are very few cases where this can be done analytically, and even then it is hardly worth the effort. A practical approach is proposed in the example, namely introducing two computational meshes for the problem at hand — a fine mesh and a coarse mesh. The fine mesh discretization is then regarded as the true system and K_k and Q_k are computed using this discretization. This mesh should be as fine as reasonably possible. The online state estimation is then carried out in the coarse mesh. Of course, the criterion for this mesh is that the time evolution of the state estimator has to be solvable with the available computing power in time before the next measurement arrives.

In practical implementation of the presented method, one weak point is the computation of Q_k which in theory requires computation of the (pseudo)inverse of the $n \times n$ matrix $\tilde{S}_k$, see (18). As noted in Remark 2.2, when $\tilde{S}_k$ is not invertible then $Q_k = \Pi^* + (I - \Pi_s)V_k\tilde{S}_k^+$. This equation for Q_k could also be used if the pseudoinverse is not computed accurately, but by using some approximative or regularizing scheme. Then the part that Q_k maps to $\Pi_s\mathcal{X}$ is readily taken care of and from $V_k\tilde{S}_k^+$ one can compute an approximation to a couple of the most important dimensions in the null space of Π.

We also remark that there is no guarantee that Q_k and K_k would converge. Further, even if they do converge, there are no algebraic equations for obtaining the limits directly. Thus, the only way to obtain them is to iterate the recursive equation sufficiently many times. However, consider the case that we are given Πx_k and we want to recover x_k. Then (assuming $\mathbb{E}(x_k) = 0$) the optimal solution is given by $\mathbb{E}(x_k|\Pi x_k) =: \widehat{Q}_k \Pi x_k$ where

$$\widehat{Q}_k = \Pi^* + (I - \Pi_s) S_k \Pi^* (\Pi S_k \Pi^*)^+$$

where $S_k = \operatorname{Cov}[x_k, x_k]$ is given by (15). Then we have $x_k = \widehat{Q}_k \Pi x_k + v_k$ where $v_k \sim N(0, \widehat{V}_k)$ where

$$\widehat{V}_k = (I - \Pi_s) S_k (I - \Pi_s)^* - (I - \Pi_s) S_k \Pi^* (\Pi S_k \Pi^*)^+ \Pi S_k (I - \Pi_s)^*.$$

Now S_k converges and the limit S_∞ can be obtained as the solution of the Lyapunov equation $S_\infty = A S_\infty A^* + B U B^*$. Of course, the error v_k is correlated but making the (false) assumption that it is not, leads to an approximate reduced order error covariance (in converged form)

$$\begin{cases} \tilde{P} = \Pi A \widehat{Q}_\infty P \widehat{Q}_\infty^* A^* \Pi^* + \Pi B U B^* \Pi^* + \Pi A \widehat{V}_\infty A^* \Pi^*, \\ P = \tilde{P} - \tilde{P} \widehat{Q}_\infty^* C^* (C \widehat{Q}_\infty \tilde{P} \widehat{Q}_\infty^* C^* + R)^{-1} C \widehat{Q}_\infty \tilde{P}. \end{cases}$$

It was found that using this approximative state estimate worked reasonably well in the presented example. With the parameters on the left in Table 1, the error $\left\| \widehat{Q}_\infty \tilde{x}_k - x_k \right\|_{\mathcal{X}}^2$ was in average over 500 simulations .6148 for the position variable and .8179 for the velocity variable (cf. the right panel of Table 1).

6.2. **Further work.** Let us end the paper by briefly discussing topics that would require further work. An immediate question is whether a similar result can be obtained for the Kalman–Bucy filter, that is, for continuous time systems. Here the discrete time systems were studied for technical convenience but, in principle, there should not be any reasons why it couldn't be done. For example the results of [2], [3] and [7] were obtained in the continuous time setting. In particular [7] might give useful tools for treating this problem.

The dual problem to the Gaussian state estimation problem is the optimal control problem for linear systems with quadratic cost functions. A natural question is whether the results of this paper can be translated to that problem. For example Mohammadi *et al.* use truncated eigenbasis approach to approximately solve the algebraic Riccati equation arising from optimal control of a diffusion-convection-reaction in [16].

One topic that was not given much attention in this paper is the optimality of the assumptions on the system. It is well known that the classical Kalman filter might work just fine even though the underlying system is not stable. We, on the other hand, used many times the input stability of the system, *i.e.*, the state covariance is uniformly bounded by some trace class operator $S_k \leq S$. Also, we had to state as an assumption that the full state Kalman filter is exponentially stable, that is, $\sigma(A - \hat{K} C A) \subset B(0, \rho)$ for some $\rho < 1$. Relaxing this assumption would be desirable since for example strong (that is, asymptotical) stability of the full state filter is proved in [10: Theorem 4.2] — although under a controllability assumption that would exclude finite dimensional control.

Acknowledgements. The author has been supported by the Finnish Graduate School in Engineering Mechanics. The author thanks Dr. Jarmo Malinen for valuable comments on the manuscript.

Appendix A. Auxiliary results

Lemma A.1. *Define the operator* $\mathbf{L} \in \mathcal{L}(\mathcal{L}^*(\mathcal{X}))$ *by*

$$\mathbf{L}W := W - (A - K^{(F)}CA)W(A - K^{(F)}CA)^*$$

where $\sigma(A - K^{(F)}CA) \subset B(0,\rho)$ *for* $\rho < 1$. *This operator has the following properties:*

(i) $\mathbf{L}$ *is boundedly invertible.*
(ii) *If* $\mathbf{L}W = X$, *then* $X \geq 0$ *implies* $W \geq 0$.
(iii) *There exists a constant* $L > 0$ *s.t.* $\operatorname{tr}(\mathbf{L}^{-1}X) \leq L\operatorname{tr}(X)$ *for all positive definite trace class operators* $X \in \mathcal{L}^*(\mathcal{X})$. *Denote by* L *the smallest possible constant. Denote* $L_0 := \sum_{j=0}^{\infty} \left\|(A - K^{(F)}CA)^{2j}\right\|_{\mathcal{L}(\mathcal{X})}$. *We have* $L \leq L_0 < \infty$.

Define also $\tilde{\mathbf{L}}W := W - (A - K_\infty CA)W(A - K_\infty CA)^*$ *where* K_∞ *is the converged gain of the reduced order filter (if it converges) and denote by* $\tilde{L}$ *the corresponding trace bound for* $\tilde{\mathbf{L}}^{-1}$.

Proof. (i): The inverse of $\mathbf{L}$ is given by

$$\mathbf{L}^{-1}X = \sum_{j=0}^{\infty}(A - K^{(F)}CA)^j X((A - K^{(F)}CA)^*)^j. \tag{29}$$

By Gelfand's formula (see [13: Theorem 7.5-5]), the sum converges in operator topology because $\sigma(A - K^{(F)}CA) \subset B(0,\rho)$ for some $\rho < 1$.

(ii): Assume that $X \in \mathcal{L}^*(\mathcal{X})$ is positive semidefinite. From (29) it is easy to see that $\mathbf{L}^{-1}X$ is positive semidefinite. Clearly also if X is negative semidefinite then W is negative semidefinite.

(iii): If $X \in \mathcal{L}^*(\mathcal{X})$ is a positive definite trace class operator and $T \in \mathcal{L}(\mathcal{X})$ then $\operatorname{tr}(TXT^*) \leq \|T\|^2_{\mathcal{L}(\mathcal{X})} \operatorname{tr}(X)$. This together with (29) imply *(iii)*. □

If $\mathbf{L}W = X_+ - X_-$ where $X_+, X_- \geq 0$ then $W = W_+ - W_-$ where $\mathbf{L}W_\pm = X_\pm$ and $W_+, W_- \geq 0$. Of course $\operatorname{tr}(W) \leq \operatorname{tr}(W_+) \leq L\operatorname{tr}(X_+)$. Thus, if the right hand side can be represented as a sum of a positive definite and a negative definite part, then only the positive definite part needs to be taken into account when computing an upper bound for the trace of the solution.

Lemma A.2. *The perturbation* ΔP *in the proof of Theorem 4.1 satisfies*

$$\begin{aligned}\Delta P &= (A - K^{(F)}CA)\Delta P(A - K^{(F)}CA)^* + E_1 + E_2 + h_1(\Delta P) + h_2(\Delta P)\\ &= \mathbf{L}^{-1}\left(E_1 + E_2 + h_1(\Delta P) + h_2(\Delta P)\right) \qquad (30)\end{aligned}$$

where

$$E_1 = (I - K^{(F)}C)AMA^*(I - K^{(F)}C)^*,$$

$$\begin{aligned}E_2 = &-(I - K^{(F)}C)AMA^*C^*\left(C(\tilde{P}^{(F)} + AMA^*)C^* + R\right)^{-1}CAMA^*(I - K^{(F)}C)^*\\ &+ K^{(F)}CAMA^*C^*\left(C(\tilde{P}^{(F)} + AMA^*)C^* + R\right)^{-1}CAMA^*C^*K^{(F)*},\end{aligned}$$

$h_1(\Delta P) = \Delta K C A \Delta P (A - K^{(F)}CA)^* + (A - K^{(F)}CA)\Delta P (\Delta K C A)^* + \Delta K C A \Delta P (\Delta K C A)^*$
where $\Delta K = K^{(F)} - K^{(b)}$, *and*

$$h_2(\Delta P) = -(A - K^{(b)}CA)\Delta P A^* C^* \Big(C(\tilde{P}^{(F)} + AMA^* + A\Delta P A^*)C^* + R \Big)^{-1} \times \\ \times CA\Delta P (A - K^{(b)}CA)^*.$$

Alternatively, the equation (30) *can be written as*

$$\Delta P = (A - K^{(b)}CA)\Delta P (A - K^{(b)}CA)^* + E_1 + E_2 + h_2(\Delta P). \tag{31}$$

The perturbation of the Kalman gain is given by

$$\Delta K = \Big((\tilde{P}^{(F)} + AMA^*)C^* \Big(C(\tilde{P}^{(F)} + AMA^*)C^* + R \Big)^{-1} C - I \Big) \times \\ \times AMA^* C^* (C\tilde{P}^{(F)} C^* + R)^{-1}.$$

For a proof, see [23: Lemma 2.1]. There everything is finite-dimensional but the proof of this Lemma is based on just algebraic manipulation and it holds also in the infinite-dimensional setting. Note that the matrix $C(\tilde{P}^{(F)} + M + A\Delta P A^*)C^* + R$ is invertible because $C(\tilde{P}^{(F)} + M + A\Delta P A^*)C^* \geq 0$ and $R > 0$. In the proof of [23: Lemma 2.1], some additional assumptions on the perturbations is needed to guarantee the invertibility of the corresponding matrix (denoted by $\tilde{C}$ there). To get (31), note that

$$h_1(\Delta P) = (A - K^{(b)}CA)\Delta P (A - K^{(b)}CA)^* - (A - K^{(F)}CA)\Delta P (A - K^{(F)}CA)^*.$$

For the last part, see in particular [23: Eq. (A.8)].

References

[1] Bensoussan, A. (**1971**). Filtrage Optimal des Systèmes linéaires, Dunod, Paris.

[2] Bernstein, D. and Hyland, D. (**1985**). "The optimal projection equations for reduced-order state estimation," Transactions on Automatic control **30**, 583–585.

[3] Bernstein, D. and Hyland, D. (**1986**). "The optimal projection equations for finite-dimensional fixed-order dynamic compensation of infinite-dimensional systems," SIAM Journal on Control and Optimization **24**, 122–151.

[4] Bogachev, V. (**1998**). Gaussian Measures, American Mathematical Society, Mathematical Surveys and Monographs, **62**.

[5] Da Prato, G. and Zabczyk, J. (**1979**). Stochastic Equations in Infinite Dimensions, Encyclopedia of Mathematics and its Applications, **44**, Cambridge University Press.

[6] De Souza, C. (**1989**). "On stabilizing properties of solutions of the Riccati difference equation," Transactions on Automatic control **34**, 1313–1316.

[7] Germani, A., Jetto, L., and Piccioni, M. (**1988**). "Galerkin approximation for optimal linear filtering of infinite-dimensional linear systems," SIAM Journal on Control and Optimization **26**, 1287–1305.

[8] Hager, W. W. and Horowitz, L. L. (**1976**). "Convergence and stability properties of the discrete Riccati operator equation and the associated optimal control and filtering problems," SIAM Journal on Control and Optimization **14**, 295–312.

[9] Hiltunen, P., Särkkä, S., Nissilä, I., Lajunen, A., and Lampinen, J. (**2011**). "State space regularization in the nonstationary inverse problem for diffuse optical tomography," Inverse Problems **27**.

[10] Horowitz, L. L. (**1974**). Optimal Filtering of Gyroscopic Noise, PhD. thesis, Massachusetts Institute of Technology.

[11] Huttunen, J. and Pikkarainen, H. (**2007**). "Discretization error in dynamical inverse problems: one-dimensional model case," Journal of Inverse and Ill-posed Problems **15**, 365–386.

[12] Kalman, R. (**1960**). "A new approach to linear filtering and prediction problems," Journal of Basic Engineering **82**, 35–45.

[13] Kreyszig, E. (**1989**). Introductory Functional Analysis with Applications, Wiley & Sons.

[14] Krug, P. (**1991**). "The conditional expectation as estimator of normally distributed random variables with values in infinitely dimensional Banach spaces," Journal of Multivariate Analysis **38**, 1–14.

[15] Larsson, S. and Thomée, V. (**2005**). Partial Differential Equations with Numerical Methods, Texts in Applied Mathematics 45, Springer-Verlag.

[16] Mohammadi, L., Aksikas, I., Dubljevic, S., and Forbes, J. (**2012**). "LQ-boundary control of a diffusion-convection-reaction system," International Journal of Control **2**, 171–181.

[17] Pikkarainen, H. (**2006**). "State estimation approach to nonstationary inverse problems: discretization error and filtering problem," Inverse Problems **22**, 365–379.

[18] Seppänen, A., Vauhkonen, M., Somersalo, E., and Kaipio, J. (**2001**). "State space models in process tomography — approximation of state noise covariance," Inverse Problems in Engineering **9**, 561–585.

[19] Simon, D. (**2007**). "Reduced order kalman filtering without model reduction," Control and Intelligent Systems **35**, 169–174.

[20] Sims, C. (**1982**). "Reduced-order modelling and filtering," Control and Dynamic Systems **18**, 55–103.

[21] Solin, A. and Särkkä, S. (**2013**). "Infinite-dimensional Bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data," Physical Review E **88**, 052909.

[22] Stubberud, A. and Wismer, D. (**1970**). "Suboptimal Kalman filter techniques," in C. Leondes, ed., "Theory and Applications of Kalman Filtering," Advisory Group for Aerospace Research and Development, 105–117.

[23] Sun, J. (**1998**). "Sensitivity analysis of the discrete-time algebraic Riccati equation," Linear Algebra and its Applications **275–276**, 595–615.

Publication IV

A. Aalto and T. Lukkari and J. Malinen. Acoustic wave guides as infinite-dimensional dynamical systems. Accepted for publication in ESAIM: Control, Optimization and Calculus of Variations, 35 pages, June 2013.

Acoustic wave guides as infinite-dimensional dynamical systems

Atte Aalto, Teemu Lukkari, and Jarmo Malinen

June 3, 2013

Abstract

We prove the unique solvability, passivity/conservativity and some regularity results of two mathematical models for acoustic wave propagation in curved, variable diameter tubular structures of finite length. The first of the models is the generalised Webster's model that includes dissipation and curvature of the 1D waveguide. The second model is the scattering passive, boundary controlled wave equation on 3D waveguides. The two models are treated in an unified fashion so that the results on the wave equation reduce to the corresponding results of approximating Webster's model at the limit of vanishing waveguide intersection.

Keywords. Wave propagation, tubular domain, wave equation, Webster's horn model, passivity, regularity.

AMS classification. Primary 35L05, secondary 35L20, 93C20, 47N70.

1 Introduction

This is the second part of the three part mathematical study on acoustic wave propagation in a narrow, tubular 3D domain $\Omega \subset \mathbb{R}^3$. The other parts of the work are [25, 26]. Our current interest in wave guide dynamics stems from modelling of acoustics of speech production; see, e.g., [1, 3, 13] and the references therein.

The main purpose of the present paper is to give a rigorous treatment of solvability and energy passivity/conservativity questions of the two models for wave propagations that are discussed in detail in [26]: these are (i) the boundary controlled wave equation on a tubular domain, and (ii) the generalised Webster's horn model that approximates the wave equation in low frequencies. The *a posteriori* error estimate for the Webster's model is ultimately given in [25], and it is in an essential part based on Theorems 4.1 and 5.1 below.

The secondary purpose of this paper is to introduce the new notion of *conservative majoration* for passive boundary control systems. The underlying systems theory idea is simple and easy to explain: it is to be expected on engineering and physical grounds that adding energy dissipation to a forward time solvable (i.e., internally well-posed, typically even conservative) system cannot make the system ill-posed, e.g., unsolvable in forward time direction. Thus, it should be enough to treat mathematically only the lossless conservative case that "majorates" all models where dissipation is included as far as we are not reversing the arrow of time. That this intuition holds true for many types of energy dissipation is proved in Theorem 3.1 for boundary dissipation and in Theorem 3.2 for a class of dissipation terms for PDE's. These theorems are given in the general context of *boundary nodes* that have been discussed in, e.g., [29, 30, 42].

Early work concerning Webster's equation can be found in [5, 40, 41, 47]. Webster's original work [47] was published in 1919, but the model itself has a longer history spanning over 200 years and starting from the works of D. Bernoulli, Euler, and Lagrange. More modern approaches is provided by [20, 21, 31, 32, 34, 33]. Webster's horn model is a special case of the wave equation in a non-homogenous medium in $\Omega \subset \mathbb{R}^n$, $n \geq 1$, which has been treated with various boundary and interior point control actions in, e.g., [9, Appendix 2], [18, Section 2], [22], [37, Section 6], and, in particular, [19, Section 7] containing also historical remarks. There exists a rich literature on the damped wave equation in 1D spatial domain, and instead of trying to give here a comprehensive account we refer to the numerous references given [10].

The boundary of $\Omega \subset \mathbb{R}^n$, $n \geq 2$, is smooth or C^2 in the works cited above, which excludes polygons (for $n = 2$) or their higher dimensional counterparts such as the tubular structures discussed here. From systems theory point of view, this is a serious restriction since it is obviously impossible to connect finitely many, disjoint, smooth domains seamlessly to each other without leaving holes whose interior is non-empty. The generality of this article makes it possible to interconnect 3D wave equation systems on geometrically compatible elements $\Omega_j \subset \mathbb{R}^3$ to form aggregated systems on $\cup_j \Omega_j$ in the same way as described in [2, Section 5] for Webster's horn model.

Theorems 4.1 and 5.1 treat the questions of unique solvability, passivity, and regularity of the two wave propagation models in the exactly same form as these results are required in companion papers [25, 26]. The strict passivity (i.e., the case $\alpha > 0$) in Theorems 4.1 and 5.1 could be proved without resorting to Theorems 3.1 and 3.2 as they both concern single PDE's with simple dissipation models. However, the direct approach becomes technically quite cumbersome if we have more complicated aggregated systems to treat (not all of which need be defined by PDE's), and combinations of various dissipation models are involved. An example of such systems is pro-

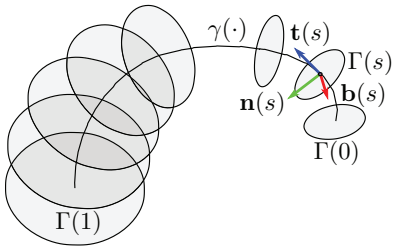

Figure 1: The Frenet frame of the planar centreline for a tubular domain Ω, represented by some of its intersection surfaces $\Gamma(s)$ for $s \in [0,1]$. The wall $\Gamma \subset \partial\Omega$ is not shown, and the global coordinate system is detailed in [26, Section 2].

vided by *transmission graphs* as introduced in [2] where the general passive case is treated by reducing it to the conservative case and arguing as in Theorem 3.2. In the context of transmission graphs, see also the literature on port-Hamiltonian systems [4, 16, 46]. That the conservative majoration method cannot be used for all possible dissipation terms is shown in Section 6 by an example involving Kelvin–Voigt structural damping.

Let us return to wave propagation models on a tubular domain Ω referring to Fig. 1. The cross sections $\Gamma(s)$ of Ω are normal to the planar curve $\gamma = \gamma(s)$ that serves as the centreline of Ω as shown in Fig. 1. We denote by $R(s)$ and $A(s) := \pi R(s)^2$ the radius and the area of $\Gamma(s)$, respectively. We call Γ the *wall*, and the circular plates $\Gamma(0)$, $\Gamma(1)$ the *ends* of the tube Ω. The boundary of Ω satisfies $\partial\Omega = \overline{\Gamma} \cup \Gamma(0) \cup \Gamma(1)$. Without loss of generality, the parameter $s \geq 0$ can be regarded as the arc length of γ, measured from the control/observation surface $\Gamma(0)$ of the tube.

As is well known, acoustic wave propagation in Ω can be modelled by the wave equation for the velocity potential ϕ as

$$\begin{cases} \phi_{tt}(\mathbf{r},t) = c^2 \Delta\phi(\mathbf{r},t) \quad \text{for } \mathbf{r} \in \Omega \text{ and } t \in \mathbb{R}^+, \\ c\frac{\partial\phi}{\partial\nu}(\mathbf{r},t) + \phi_t(\mathbf{r},t) = 2\sqrt{\frac{c}{\rho A(0)}}\, u(\mathbf{r},t) \quad \text{for } \mathbf{r} \in \Gamma(0) \text{ and } t \in \mathbb{R}^+, \\ \phi(\mathbf{r},t) = 0 \quad \text{for } \mathbf{r} \in \Gamma(1) \text{ and } t \in \mathbb{R}^+, \\ \alpha\frac{\partial\phi}{\partial t}(\mathbf{r},t) + \frac{\partial\phi}{\partial\nu}(\mathbf{r},t) = 0 \quad \text{for } \mathbf{r} \in \Gamma, \text{ and } t \in \mathbb{R}^+, \text{ and} \\ \phi(\mathbf{r},0) = \phi_0(\mathbf{r}), \quad \rho\phi_t(\mathbf{r},0) = p_0(\mathbf{r}) \quad \text{for } \mathbf{r} \in \Omega \end{cases} \tag{1.1}$$

with the observation defined by

$$c\frac{\partial\phi}{\partial\nu}(\mathbf{r},t) - \phi_t(\mathbf{r},t) = 2\sqrt{\frac{c}{\rho A(0)}}\, y(\mathbf{r},t) \quad \text{for } \mathbf{r} \in \Gamma(0) \text{ and } t \in \mathbb{R}^+, \tag{1.2}$$

where ν denotes the unit normal vector on $\partial\Omega$, c is the sound speed, ρ is the density of the medium, and $\alpha \geq 0$ is a parameter associated to boundary

dissipation. The functions u and y are control and observation signals in *scattering form*, and the normalisation constant $2\sqrt{\frac{c}{\rho A(0)}}$ takes care of their physical dimension which is power per area. Solvability, stability, and energy questions for the wave equation in various geometrical domains $\Omega \subset \mathbb{R}^n$ have a huge literature, and it is not possible to give a historically accurate review here. The wave equation is a prototypal example of a linear hyperbolic PDE whose classical mathematical treatment can be found, e.g., in [23, Chapter 5], and the underlying physics is explained well in [8, Chapter 9]. In the operator and mathematical system theory context, it has been given as an example (in various variations) in [27, 30, 43, 44, 48] and elsewhere. For applications in speech research, see, e.g., [3, 13, 26] and the references therein.

One computationally and analytically simpler wave propagation model is the *generalised Webster's horn model* for the same tubular domain Ω that is now represented by the *area function* $A(\cdot)$ introduced above. To review this model in its generalised form, let us recall some notions from [26]. To take into account the curvature $\kappa(s)$ of the centreline $\gamma(\cdot)$ of Ω, we adjust the sound speed c in (1.1) by defining $c(s) := c\Sigma(s)$ where $\Sigma(s) := \left(1 + \frac{1}{4}\eta(s)^2\right)^{-1/2}$ is the *sound speed correction factor*, and $\eta(s) := R(s)\kappa(s)$ is the *curvature ratio* at $s \in [0,1]$. We also need take into consideration the deformation of the outer wall Γ by defining the *stretching factor* $W(s) := R(s)\sqrt{R'(s)^2 + (\eta(s)-1)^2}$; see [26, Eq. (2.8)]. It is a standing assumption that $\eta(s) < 1$ to prevent the tube Ω from folding on itself locally.

Following [26], the generalised Webster's horn model for the velocity potential $\psi = \psi(s,t)$ is now given by

$$\begin{cases} \psi_{tt} = \frac{c(s)^2}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial \psi}{\partial s}\right) - \frac{2\pi\alpha W(s)c(s)^2}{A(s)}\frac{\partial \psi}{\partial t} \\ \qquad\qquad\qquad\qquad \text{for } s \in (0,1) \text{ and } t \in \mathbb{R}^+, \\ -c\psi_s(0,t) + \psi_t(0,t) = 2\sqrt{\frac{c}{\rho A(0)}}\,\tilde{u}(t) \quad \text{for } t \in \mathbb{R}^+, \\ \psi(1,t) = 0 \quad \text{for } t \in \mathbb{R}^+, \quad \text{and} \\ \psi(s,0) = \psi_0(s), \quad \rho\psi_t(s,0) = \pi_0(s) \quad \text{for } s \in (0,1), \end{cases} \tag{1.3}$$

and the observation $\tilde{y}$ is defined by

$$-c\psi_s(0,t) - \psi_t(0,t) = 2\sqrt{\frac{c}{\rho A(0)}}\,\tilde{y}(t) \quad \text{for } t \in \mathbb{R}^+. \tag{1.4}$$

The constants c, ρ, α are same as in (1.1). The input and output signals $\tilde{u}$ and $\tilde{y}$ of (1.3)–(1.4) correspond to u and y in (1.1)–(1.2) by spatial averaging over the control surface $\Gamma(0)$. Hence, their physical dimension is power per area as well. Based on [25, 26], the solution ψ of (1.3) approximates the averages

$$\bar{\phi}(s,t) := \frac{1}{A(s)}\int_{\Gamma(s)} \phi \,\mathrm{d}A \quad \text{for} \quad s \in (0,1) \quad \text{and} \quad t \geq 0 \tag{1.5}$$

of ϕ in (1.1) when ϕ is regular enough. Note that the dissipative boundary condition $\alpha\frac{\partial\phi}{\partial\nu}(\mathbf{r},t)+\frac{\partial\phi}{\partial\nu}(\mathbf{r},t)=0$ in (1.1) has been replaced by the dissipation term $2\pi\alpha W(s)A(s)^{-1}c(s)^2\frac{\partial\psi}{\partial t}$ (with the same parameter α) in (1.3). For classical work on Webster's horn model, see [20, 31, 40] and in particular [33] where numerous references can be found.

We show in Theorem 5.1 that the wave equation model (1.1)–(1.2) is uniquely solvable in both directions of time, and the solution satisfies an energy inequality if $\alpha > 0$. By Corollary 5.2, the model has the same properties for $\alpha = 0$ but then the energy inequality is replaced by an equality, and the model is even time-flow invertible. In all cases, the solution ϕ is observed to have the regularity required for the treatment given in [26] if the input u is twice continuously differentiable. The generalised Webster's horn model (1.3)–(1.4) is treated in a similar manner in Theorem 4.1.

This paper is organised as follows: Background on boundary control systems is given in Section 2. Conservative majoration of passive boundary control systems is treated in Section 3. The Webster's horn model and the wave equation are treated in Sections 4 and 5 respectively. Some immediate extensions of these results are given in Section 6. Because of the lack of accessible, complete, and sufficiently general references, the paper is completed by a self-contained appendix on Sobolev spaces, boundary trace operators, Green's identity, and Poincaré inequality for special Lipschitz domains that are required in the rigorous analysis of typical wave guide geometries.

2 On infinite dimensional systems

Linear boundary control systems such as (1.1) and (1.3) are treated as dynamical systems that can be described by operator differential equations of the form

$$u(t) = Gz(t), \quad \dot{z}(t) = Lz(t), \qquad \text{with the initial condition} \qquad z(0) = z_0 \tag{2.1}$$

and the observation equation

$$y(t) = Kz(t), \tag{2.2}$$

where $t \in \mathbb{R}^+$ denotes time. The signals in (2.1), (2.2) are as follows: u is the input, y is the output, and the state trajectory is z.

Cauchy problems

To make (2.1) properly solvable for all twice differentiable u and compatible initial states z_0, the axioms of an *internally well-posed boundary node* should be satisfied:

Definition 2.1. *A triple of operators $\Xi = (G, L, K)$ is an* internally well-posed boundary node *on the Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ if the following conditions are satisfied:*

(i) *G, L, and K are linear operators with the same domain $\mathcal{Z} \subset \mathcal{X}$;*

(ii) $\left[\begin{smallmatrix} G \\ L \\ K \end{smallmatrix}\right]$ *is a closed operator from $\mathcal{X}$ into $\mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ with domain $\mathcal{Z}$;*

(iii) *G is surjective, and $\ker(G)$ is dense in $\mathcal{X}$; and*

(iv) $L\big|_{\ker(G)}$ *(understood as an unbounded operator in $\mathcal{X}$ with domain $\ker(G)$) generates a strongly continuous semigroup on $\mathcal{X}$.*

If, in addition, L is a closed operator on $\mathcal{X}$ with domain $\mathcal{Z}$, we say that the boundary node Ξ is strong.

The history of abstract boundary control system dates back to [7, 38, 39]. The phrase "internally well-posed" refers to condition (iv) of Definition 2.1, and it is a much weaker property than well-posedness of systems in the sense of [42]. It plainly means that the boundary node defines an evolution equation that is uniquely solvable in forward time direction. Boundary nodes that are not necessarily internally well-posed are characterised by the weaker requirement in place of (iv): $\alpha - L\big|_{\ker(G)}$ is a bijection from $\ker(G)$ onto $\mathcal{X}$ for some $\alpha \in \mathbb{C}$.

We call $\mathcal{U}$ the *input space*, $\mathcal{X}$ the *state space*, $\mathcal{Y}$ the *output space*, $\mathcal{Z}$ the *solution space*, G the *input boundary operator*, L the *interior operator*, and K the *output boundary operator*. The operator $A := L\big|_{\ker(G)}$ is called the *semigroup generator* if Ξ is internally well-posed, and otherwise it is known as the *main operator* of Ξ. Because $\begin{bmatrix} G & L & K \end{bmatrix}^T$ is a closed operator, we can give its domain the Hilbert space structure by the *graph norm*

$$\|z\|_{\mathcal{Z}}^2 = \|z\|_{\mathcal{X}}^2 + \|Lz\|_{\mathcal{X}}^2 + \|Gz\|_{\mathcal{U}}^2 + \|Kz\|_{\mathcal{Y}}^2. \tag{2.3}$$

If the node is strong, we have an equivalent norm for $\mathcal{Z}$ given by omitting the last two terms in (2.3). If $\Xi = (G, L, K)$ is an internally well-posed boundary node, then (2.1) has a unique "smooth" solution:

Proposition 2.2. *Assume that $\Xi = (G, L, K)$ is an internally well-posed boundary node. For all $z_0 \in \mathcal{X}$ and $u \in C^2(\mathbb{R}^+; \mathcal{U})$ with $Gz_0 = u(0)$ the equations* (2.1) *have a unique solution $z \in C^1(\mathbb{R}^+; \mathcal{X}) \cap C(\mathbb{R}^+; \mathcal{Z})$. Hence, the output $y \in C(\mathbb{R}^+; \mathcal{Y})$ is well defined by the equation* (2.2)*.*

Indeed, this is [29, Lemma 2.6].

Energy balances

Now that we have treated the solvability of the dynamical equations, it remains to consider energy notions. We say that the internally well-posed boundary node $\Xi = (G, L, K)$ is *(scattering) passive* if all smooth solutions of (2.1) satisfy

$$\frac{d}{dt}\|z(t)\|_{\mathcal{X}}^2 + \|y(t)\|_{\mathcal{Y}}^2 \leq \|u(t)\|_{\mathcal{U}}^2 \quad \text{for all} \quad t \in \mathbb{R}^+ \tag{2.4}$$

with y given by (2.2). All such systems are well-posed in the sense of [42]; see also [45]. We say that Ξ is *(scattering) energy preserving* if (2.4) holds as an equality.

Many boundary nodes arising from hyperbolic PDE's (such as (1.1)–(1.2) and (1.3)–(1.4)) have the property that they remain boundary nodes if we (i) change the sign of L (i.e., reverse the direction of time); and (ii) interchange the roles of K and G (i.e., reverse the flow direction). Such boundary nodes are called *time-flow invertible*, and we write $\Xi^{\leftarrow} = (K, -L, G)$ for the time-flow inverse of Ξ. There are many equivalent definitions of *conservativity* in the literature, and we choose here the following:

Definition 2.3. *An internally well-posed boundary node Ξ is* (scattering) conservative *if it is time-flow invertible, and both Ξ itself and the time-flow inverse $\Xi^{\leftarrow}$ are (scattering) energy preserving.*[1]

For system nodes that have been introduced in [42, 28], an equivalent definition for conservativity is to require that both S and its *dual node* S^d are energy preserving. This is the straightforward generalisation from the finite-dimensional theory but it is not very practical when dealing with boundary control. For conservative systems, the time-flow inverse and the dual system coincide, and we have then, in particular, $A^* = -L\big|_{\ker(K)}$ if $A = L\big|_{\ker(G)}$. For details, see [29, Theorems 1.7 and 1.9].

It is possible to check economically, without directly using Definition 2.1, that the triple $\Xi = (G, L, K)$ is a dissipative/conservative boundary node:

Proposition 2.4. *Let $\Xi = (G, L, K)$ be a triple of linear operators with a common domain $\mathcal{Z} \subset \mathcal{X}$, and ranges in the Hilbert spaces $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$, respectively. Then Ξ is a passive boundary node on $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ if and only if the following conditions hold:*

(i) *We have the* Green–Lagrange inequality

$$2\mathrm{Re}\,\langle z, Lz\rangle_{\mathcal{X}} + \|Kz\|_{\mathcal{Y}}^2 \leq \|Gz\|_{\mathcal{U}}^2 \quad \textit{for all} \quad z \in \mathcal{Z}; \tag{2.5}$$

[1] The words "energy preserving" can be replaced by "passive" without changing the class of systems one obtains.

(ii) $G\mathcal{Z} = \mathcal{U}$ *and* $(\beta - L)\ker(G) = \mathcal{X}$ *for some* $\beta \in \mathbb{C}^+$ *(hence, for all* $\beta \in \mathbb{C}^+$*).*

Similarly, Ξ *is a conservative boundary node on* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ *if and only if* (ii) *above holds together with the additional conditions:*

(iii) *We have the* Green–Lagrange identity

$$2\mathrm{Re}\,\langle z, Lz\rangle_{\mathcal{X}} + \|Kz\|_{\mathcal{Y}}^2 = \|Gz\|_{\mathcal{U}}^2 \quad \text{for all} \quad z \in \mathcal{Z}. \tag{2.6}$$

(iv) $K\mathcal{Z} = \mathcal{Y}$ *and* $(\gamma + L)\ker(K) = \mathcal{X}$ *for some* $\gamma \in \mathbb{C}^+$ *(hence, for all* $\gamma \in \mathbb{C}^+$*).*

This is a slight modification of [30, Theorem 2.5]. See also [29, Proposition 2.5]. The abstract boundary spaces as discussed in [11] are essentially (impedance) conservative strong nodes as explained in [30, Section 5].

3 Conservative majorants

In some applications, the dissipative character of a linear dynamical system is often due to a distinct part of the model such as a term or a boundary condition imposed on the defining PDE. If this part is completely removed from the model, the resulting more simple system is conservative and, in particular, internally well-posed. We call it a *conservative majorant* of the original dissipative system.

Intuition from engineering and physics hints that increasing dissipation should make the system "better behaved" and not spoil the internal well-posedness.[2] The following Theorems 3.1 and 3.2 apply to many boundary control systems. However, they are written for *passive* majorants since the proofs remain the same, and this way the results can be applied successively to systems having both boundary dissipation and dissipative terms.

Theorem 3.1. *Let* $\widetilde{\Xi} = (\left[\begin{smallmatrix} G \\ \tilde{G} \end{smallmatrix}\right], L, \left[\begin{smallmatrix} K \\ \tilde{K} \end{smallmatrix}\right])$ *be a scattering passive boundary node on Hilbert spaces* $(\mathcal{U} \oplus \tilde{\mathcal{U}}, \mathcal{X}, \mathcal{Y} \oplus \tilde{\mathcal{Y}})$ *with solution space* $\tilde{\mathcal{Z}}$*. Then* $\Xi := (G\big|_{\mathcal{Z}}, L\big|_{\mathcal{Z}}, K\big|_{\mathcal{Z}})$ *is a scattering passive boundary node on* $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ *with the solution space* $\mathcal{Z} := \ker\left(\tilde{G}\right)$*. Both* $\widetilde{\Xi}$ *and* Ξ *have the same semigroup generators, equalling* $L\big|_{\ker(G)\cap\ker\left(\tilde{G}\right)}$*. If* $\widetilde{\Xi}$ *is a strong node, so is* Ξ*.*

Proof. The Green–Lagrange inequality holds for Ξ since for $z \in \ker\left(\tilde{G}\right)$ we have $\|Gz\|_{\mathcal{U}} = \|\left[\begin{smallmatrix} G \\ \tilde{G} \end{smallmatrix}\right] z\|_{\mathcal{U}\oplus\tilde{\mathcal{U}}}$, and hence we get by the passivity of $\widetilde{\Xi}$

$$2\mathrm{Re}\,\langle z, Lz\rangle_{\mathcal{X}} - \|Gz\|_{\mathcal{U}}^2 \le -\|\left[\begin{smallmatrix} Kz \\ \tilde{K}z \end{smallmatrix}\right]\|_{\mathcal{Y}\oplus\tilde{\mathcal{Y}}}^2 \le -\|Kz\|_{\mathcal{Y}}^2.$$

[2] The dissipativity or even the internal well-posedness of the time-flow inverted system is, if course, destroyed since adding dissipation creates the "arrow of time".

The surjectivity $G\mathcal{Z} = \mathcal{U}$ follows from $\mathcal{U} \oplus \{0\} \subset \mathcal{U} \oplus \tilde{\mathcal{U}} = \left[\begin{smallmatrix} G \\ \tilde{G} \end{smallmatrix}\right] \mathcal{Z}$ and $\mathcal{Z} = \ker\left(\tilde{G}\right)$. Since $(\beta - L)\ker\left(G\big|_{\mathcal{Z}}\right) = (\beta - L)\big|_{\ker(\tilde{G})}\ker(G) = (\beta - L)\left(\ker(G) \cap \ker\left(\tilde{G}\right)\right) = (\beta - L)\ker\left(\left[\begin{smallmatrix} G \\ \tilde{G} \end{smallmatrix}\right]\right) = \mathcal{X}$, the passivity of Ξ follows by Proposition 2.4.

Suppose that L is closed (i.e., $\widetilde{\Xi}$ is strong) and that $\tilde{\mathcal{Z}} \supset \mathcal{Z} \ni z_j \to z$ in $\mathcal{X}$ is such that $Lz_j \to x$ in $\mathcal{X}$ as $j \to \infty$. Because L is closed, $z \in \operatorname{dom}(L) = \tilde{\mathcal{Z}}$ and $Lz = x$. Thus, $\|z_j - z\|_{\mathcal{Z}}^2 := \|z_j - z\|_{\mathcal{X}}^2 + \|L(z_j - z)\|_{\mathcal{X}}^2 \to 0$. Because $\tilde{G} \in \mathcal{L}(\mathcal{Z};\tilde{\mathcal{U}})$ by applying (2.3) on $\widetilde{\Xi}$, the space $\mathcal{Z} = \ker\left(\tilde{G}\right)$ is closed in $\tilde{\mathcal{Z}}$ and thus $z \in \mathcal{Z}$. We have now shown that $L\big|_{\mathcal{Z}}$ is closed with $\operatorname{dom}\left(L\big|_{\mathcal{Z}}\right) = \mathcal{Z}$. □

The restriction of the original solution space to $\ker\left(\tilde{G}\right)$ in Theorem 3.1 is a functional analytic description of boundary dissipation of a particular kind. If the original scattering passive $\widetilde{\Xi}$ is translated to an impedance passive boundary node by the external Cayley-transform (see [30, Definition 3.1]), then the abstract boundary condition by restriction to $\ker\left(\tilde{G}\right)$ can be understood as a termination to an ideally resistive element as depicted in [30, Fig. 1].

Theorem 3.2. *Let $\Xi = (G, L, K)$ be a scattering passive boundary node on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ with solution space $\mathcal{Z}$ and $\mathcal{X}_1 = \ker(G)$ with the norm $\|z\|_{\mathcal{X}_1} = \|(1 - L)z\|_{\mathcal{X}}$. Let H be a dissipative operator on $\mathcal{X}$ with $\mathcal{Z} \subset \operatorname{dom}(H)$.*[3] *Denote the two assumptions as follows:*

(i) *There is $a > 0$ and $0 \le b < 1$ such that $\|Hz\|_{\mathcal{X}} \le a\|z\|_{\mathcal{X}} + b\|Lz\|_{\mathcal{X}}$ for all $z \in \ker(G)$.*

(ii) *There is a Hilbert space $\tilde{\mathcal{X}}$ such that $\mathcal{X}_1 \subset \tilde{\mathcal{X}} \subset \operatorname{dom}(H)$, the inclusion $\mathcal{X}_1 \subset \tilde{\mathcal{X}}$ is compact and $H\big|_{\tilde{\mathcal{X}}} \in \mathcal{L}(\tilde{\mathcal{X}}; \mathcal{X})$.*

If either (i) *or* (ii) *holds, then $\Xi_H := (G, L + H, K)$ is a scattering passive boundary node. We have $\operatorname{dom}(A) = \operatorname{dom}(A_H)$ where $A = L\big|_{\ker(G)}$ and $A_H = (L+H)\big|_{\ker(G)}$ are the semigroup generators of Ξ and Ξ_H, respectively. If the node Ξ is strong and $H \in \mathcal{L}(\mathcal{X})$ (i.e., $b = 0$ in assumption* (i)*), then Ξ_H is a strong boundary node as well.*

Both the assumptions (i) and (ii) hold if $H \in \mathcal{L}(\mathcal{X})$ and $\mathcal{X}_1 \subset \mathcal{X}$ with a compact inclusion. This is the case in [2, Section 5] in the context of an impedance passive system. The compactness property is typically a consequence of the Rellich–Kondrachov theorem [6, Theorem 1, p. 144] for boundary nodes defined by PDE's on bounded domains. In many applications such as Theorem 4.1 below, the operator H is even self-adjoint. We

[3] This means that $H : \operatorname{dom}(H) \subset \mathcal{X} \to \mathcal{X}$ is an operator satisfying $\mathcal{Z} \subset \operatorname{dom}(H)$ and $\operatorname{Re}\langle z, Hz\rangle_{\mathcal{X}} \le 0$ for all $z \in \mathcal{Z}$.

give an example of the 1D wave equation with Kelvin–Voigt damping in Section 6 where Theorem 3.2 cannot be applied.

Proof. By using assumption (i): This argument is motivated by [14, Theorem 2.7 on p. 501]. Let us first show that $A_H := A + H\big|_{\ker(G)}$ with $\operatorname{dom}(A_H) = \ker(G)$ generates a contraction semigroup on $\mathcal{X}$ where $A = L\big|_{\ker(G)}$ generates the contraction semigroup of Ξ as usual. As a first step, we establish the inequality $\|H(s-A)^{-1}\|_{\mathcal{L}(\mathcal{X})} < 1$ for all real s large enough.

Let $\beta > 0$ be arbitrary. For all $s > \beta$ and $z \in \mathcal{X}$ we have

$$\begin{aligned}\|H(s-A)^{-1}z\|_{\mathcal{X}} \leq & a\|(s-A)^{-1}z\|_{\mathcal{X}} + b\|A(s-A)^{-1}z\|_{\mathcal{X}} \\ \leq & (a+\beta b)\|(s-A)^{-1}z\|_{\mathcal{X}} \\ & + \frac{b}{s-\beta}\left\|\left(\frac{1}{s-\beta} - (A-\beta)^{-1}\right)^{-1} z\right\|_{\mathcal{X}}\end{aligned} \tag{3.1}$$

since

$$-A(s-A)^{-1} = \frac{1}{s-\beta}\left(\frac{1}{s-\beta} - (A-\beta)^{-1}\right)^{-1} - \beta(s-A)^{-1}.$$

Since A is a maximally dissipative operator on $\mathcal{X}$, we have for all $z = (A-\beta)x \in \mathcal{X}$ with $x \in \operatorname{dom}(A)$

$$\begin{aligned}\operatorname{Re}\left\langle (A-\beta)^{-1}z, z\right\rangle_{\mathcal{X}} = & \operatorname{Re}\left\langle (A-\beta)^{-1}(A-\beta)x, (A-\beta)x\right\rangle_{\mathcal{X}} \\ = & \operatorname{Re}\left\langle x, (A-\beta)x\right\rangle_{\mathcal{X}} \\ = & \operatorname{Re}\left\langle x, Ax\right\rangle_{\mathcal{X}} - \beta\|x\|_{\mathcal{X}}^2 \leq 0.\end{aligned}$$

Thus, the operator $(A-\beta)^{-1}$ is dissipative, and it is maximally so because $(A-\beta)^{-1} \in \mathcal{L}(\mathcal{X})$.

Because $(A-\beta)^{-1}$ generates a C_0 contraction semigroup on X, the Hille–Yoshida generator theorem gives the resolvent estimate

$$\frac{1}{s-\beta}\left\|\left(\frac{1}{s-\beta} - (A-\beta)^{-1}\right)^{-1}\right\|_{\mathcal{L}(\mathcal{X})} \leq 1$$

for $s > \beta > 0$. Similarly, $\|(s-A)^{-1}\|_{\mathcal{L}(\mathcal{X})} \leq 1/s$ for $s > 0$. These together with (3.1) give

$$\frac{\|H(s-A)^{-1}z\|_{\mathcal{X}}}{\|z\|_{\mathcal{X}}} \leq \frac{a+\beta b}{s} + b < 1 \text{ for all } s > \frac{a+\beta b}{1-b}.$$

Because $\beta > 0$ was arbitrary, we get $\|H(s-A)^{-1}\|_{\mathcal{L}(\mathcal{X})} < 1$ for all $s > \frac{a}{1-b}$. We conclude that $(a/(1-b), \infty) \subset \rho(A_H)$ and

$$(s-A_H)^{-1} = (s-A)^{-1}(I - H(s-A)^{-1})^{-1} \tag{3.2}$$

where $\mathrm{dom}\,(A_H) = \mathrm{dom}\,(A) = \ker(G)$. In particular, we have shown that $(2a/(1-b) - L - H)\ker(G) = \mathcal{X}$ (that $G\mathcal{Z} = \mathcal{U}$ holds, follows because Ξ itself is a boundary node with the same input boundary operator G). Since the Green–Lagrange inequality (2.5) holds by the passivity of Ξ and $\mathrm{Re}\,\langle z, Hz\rangle_{\mathcal{X}} \leq 0$ by assumption, we conclude that (2.5) holds with $L + H$ in place of L, too. Thus Ξ_H is a scattering passive boundary node by Proposition 2.4.

By using assumption (ii): As in the first part of this proof, it is enough to prove that $\rho(A_H) \cap \mathbb{C}_+ \neq \emptyset$ by verifying (3.2). Because $(s-A)^{-1} \in \mathcal{L}(\mathcal{X}; \mathcal{X}_1)$, $\mathcal{X}_1 \subset \tilde{\mathcal{X}}$ is compact, and $H\big|_{\tilde{\mathcal{X}}} \in \mathcal{L}(\tilde{\mathcal{X}}; \mathcal{X})$, we conclude that $H(s-A)^{-1} \in \mathcal{L}(\mathcal{X})$ is a compact operator for all $s \in \mathbb{C}_+$. If there is a $s > 0$ such that $1 \notin \sigma(H(s-A)^{-1}) \subset \sigma_p(H(s-A)^{-1}) \cup \{0\}$, then (3.2) holds, $s \in \rho(A_H)$, and Ξ_H is a passive boundary node as argued in the first part of the proof. For contradiction, assume that $1 \in \sigma_p(H(s_0 - A)^{-1})$ for some $s_0 > 0$. This implies $A_H x_0 = s_0 x_0$ for some $x_0 \in \mathrm{dom}\,(A_H)$, and hence

$$\mathrm{Re}\,\langle A_H x_0, x_0\rangle_{\mathcal{X}} = s_0 \|x_0\|_{\mathcal{X}}^2 > 0$$

which contradicts the dissipativity of $A_H = A + H\big|_{\ker(G)}$. Thus (3.2) holds and $\mathrm{dom}\,(A) = \mathrm{dom}\,(A_H)$. The final claim about strongness of Ξ_H holds because perturbations of closed operators by bounded operators are closed. □

The perturbation H in Theorem 3.2 is a densely defined dissipative operator on $\mathcal{X}$. As such, it has a maximally dissipative (closed) extension $\widetilde{H} : \mathrm{dom}\left(\widetilde{H}\right) \subset \mathcal{X} \to \mathcal{X}$ satisfying $\widetilde{H}^* \subset H^*$, and the adjoint $\widetilde{H}^*$ is maximally dissipative as well. Without loss of generality we may assume that $H = \widetilde{H}$ in Theorem 3.2. Furthermore, it is possible to use $\tilde{\mathcal{X}} = \mathrm{dom}\left(\widetilde{H}\right)$ equipped with the graph norm $\|z\|^2_{\mathrm{dom}\left(\widetilde{H}\right)} = \|z\|^2_{\mathcal{X}} + \|\widetilde{H}z\|^2_{\mathcal{X}}$ in assumption (ii), and it only remains to check whether $\mathcal{X}_1 \subset \mathrm{dom}\left(\widetilde{H}\right)$ compactly.

Let us consider the adjoint semigroup of the passive boundary node $\Xi_H = (G, L+H, K)$, majorated by the conservative node $\Xi = (G, L, K)$. The adjoint semigroup is generated by the maximally dissipative operator A_H^* where $A_H = (L+H)\big|_{\ker(G)}$ is maximally dissipative under the assumptions of Theorem 3.2.

Proposition 3.3. *Let $\Xi = (G, L, K)$ be a scattering conservative boundary node on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{Y})$ with solution space $\mathcal{Z}$. Let H be a dissipative operator on $\mathcal{X}$ with $\mathcal{Z} \subset \mathrm{dom}\,(H)$. Assume that either of the assumptions* (i) *or* (ii) *of Theorem 3.2 holds, and let the extension $\widetilde{H}$ be defined as above.*

(i) *If $\ker(K) \subset \mathrm{dom}\left(\widetilde{H}^*\right)$, then $(-L + \widetilde{H}^*)\big|_{\ker(K)} \subset A_H^*$.*

(ii) *If* Ξ *is time-flow invertible and* $\mathcal{Z} \subset \operatorname{dom}\left(\widetilde{H}^*\right)$*, then* $\Xi^{\leftarrow}_{\widetilde{H}^*} := (K, -L+\widetilde{H}^*, G)$ *is an internally well-posed boundary node if and only if* $(-L+\widetilde{H}^*)\big|_{\ker(K)} = A_H^*$.

(iii) *If* Ξ *is conservative and* $\mathcal{Z} \subset \operatorname{dom}\left(\widetilde{H}^*\right)$*, then* $\Xi^{\leftarrow}_{\widetilde{H}^*}$ *is a passive boundary node if and only if* $(-L+\widetilde{H}^*)\big|_{\ker(K)} = A_H^*$.

If $\Xi = (G, L, K)$ is conservative, so is its time-flow inverse $\Xi^{\leftarrow} = (K, -L, G)$ by Definition 2.3. In this case, it may be possible to use Theorem 3.2 to conclude that $\Xi^{\leftarrow}_{\widetilde{H}^*}$ is a passive boundary node as well. If both Ξ_H and $\Xi^{\leftarrow}_{\widetilde{H}^*}$ are passive, then they cannot be time-flow inverses of each other unless both nodes are, in fact, conservative; i.e., $H = \widetilde{H}^* = 0$ on $\mathcal{Z}$.

Proof. It is easy to see that $A^*+T^* \subset (A+T)^*$ holds for operators A, T on $\mathcal{X}$ with $\operatorname{dom}(A) \cap \operatorname{dom}(T)$ dense in $\mathcal{X}$. Applying this on $A = L\big|_{\ker(G)}$ and $T := \widetilde{H}\big|_{\ker(G)}$ we get on $\ker(K)$ the inclusion $-L\big|_{\ker(K)} + \left(\widetilde{H}\big|_{\ker(G)}\right)^* \subset A_H^*$. Here we used $A^* = -L\big|_{\ker(K)}$ which holds because $\Xi = (G, L, K)$ is a conservative boundary node whose dual system (with semigroup generator A^*) coincides with the time-flow inverse $\Xi^{\leftarrow} = (K, -L, G)$. Since $\ker(K) \subset \operatorname{dom}\left(\widetilde{H}^*\right)$ has been assumed, it follows that $\left(\widetilde{H}\big|_{\ker(G)}\right)^* z = \widetilde{H}^* z$ for all $z \in \ker(K)$, and claim (i) now follows.

The "only if" part of claims (ii) and (iii): By the internal well-posedness of $\Xi^{\leftarrow}_{\widetilde{H}^*}$, its main operator $(-L+\widetilde{H}^*)\big|_{\ker(K)}$ generates a C_0 semigroup, and its resolvent set contains some right half plane by the Hille–Yoshida theorem. By claim (i) and the fact that A_H^* is (even maximally) dissipative, it follows that $(-L+\widetilde{H}^*)\big|_{\ker(K)}$ is dissipative. But then $(-L+\widetilde{H}^*)\big|_{\ker(K)}$ is maximally dissipative, and the converse inclusion $A_H^* \subset (-L+\widetilde{H}^*)\big|_{\ker(K)}$ follows.

The "if" part of claim (ii): The operator $(-L+\widetilde{H}^*)\big|_{\ker(K)}$ generates a contraction semigroup on $\mathcal{X}$ because it equals by assumption A_H^* where A_H itself is a generator of a contraction semigroup by Theorem 3.2.

Equip the Hilbert space $\operatorname{dom}\left(\widetilde{H}^*\right)$ with the graph norm of the closed operator $\widetilde{H}^*$. Since $\mathcal{Z} \subset \operatorname{dom}\left(\widetilde{H}^*\right)$ has been assumed, and both $\mathcal{Z}$ and $\operatorname{dom}\left(\widetilde{H}^*\right)$ are continuously embedded in $\mathcal{X}$, the inclusion $\mathcal{Z} \subset \operatorname{dom}\left(\widetilde{H}^*\right)$ is continuous, too. Now $\widetilde{H}^*\big|_{\mathcal{Z}} \in \mathcal{L}(\mathcal{Z}; \mathcal{X})$ follows from $\widetilde{H}^* \in \mathcal{L}(\operatorname{dom}\left(\widetilde{H}^*\right); \mathcal{X})$. Since now $-L+\widetilde{H}^* \in \mathcal{L}(\mathcal{Z}; \mathcal{X})$, it follows that $\Xi^{\leftarrow}_{\widetilde{H}^*}$ is an internally well-posed boundary node by [29, Proposition 2.5]. (You could also argue by verifying Definition 2.1(ii) directly.)

The "if" part of claim (iii): The "if" part of claim (ii) gives the internal well-posedness of $\Xi^{\leftarrow}_{\widetilde{H}^*}$. To show passivity, only the Green–Lagrange

inequality $2\mathrm{Re}\,\langle z, (-L+\widetilde{H}^*)z\rangle_{\mathcal{X}} \leq \|Kz\|_{\mathcal{Y}}^2 - \|Gz\|_{\mathcal{U}}^2$ is needed. This follows from (2.6) (by the conservativity of $\Xi^{\leftarrow}$) and the dissipativity of $\widetilde{H}^*$ with $\mathcal{Z} \subset \mathrm{dom}\left(\widetilde{H}^*\right)$ (since $\widetilde{H}$ is *maximally* dissipative). □

4 Generalised Webster's model for wave guides

As proved in [26], we arrive (under some mild technical assumptions on Ω as explained in [26, Section 3]) to the following equations for the approximate spatial averages of solutions of (5.1):

$$\begin{cases} \psi_{tt} = \frac{c(s)^2}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial \psi}{\partial s}\right) - \frac{2\pi\alpha W(s)c(s)^2}{A(s)}\frac{\partial \psi}{\partial t} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } s \in (0,1) \text{ and } t \in \mathbb{R}^+, \\ -c(0)\psi_s(0,t) + \psi_t(0,t) = 2\sqrt{\frac{c(0)}{\rho A(0)}}\,\tilde{u}(t) \quad \text{for } t \in \mathbb{R}^+, \\ \psi(1,t) = 0 \quad \text{for } t \in \mathbb{R}^+, \quad \text{and} \\ \psi(s,0) = \psi_0(s), \quad \rho\psi_t(s,0) = \pi_0(s) \quad \text{for } s \in (0,1), \end{cases} \tag{4.1}$$

and the observation equation averages to

$$-c(0)\psi_s(0,t) - \psi_t(0,t) = 2\sqrt{\frac{c(0)}{\rho A(0)}}\,\tilde{y}(t) \quad \text{for } t \in \mathbb{R}^+. \tag{4.2}$$

The notation has been introduced in Section 1. Analogously with the wave equation, the solution ψ is called *Webster's velocity potential*. In [25, Section 3] we add a load function $f(s,t)$ to obtain the PDE $\psi_{tt} = \frac{c(s)^2}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial \psi}{\partial s}\right) - \frac{2\pi\alpha W(s)c(s)^2}{A(s)}\frac{\partial \psi}{\partial t} + f(s,t)$ because the argument there is based on the feed-forward connection detailed in [26, Fig. 1]. Only the boundary control input is considered here, and it can be treated using boundary nodes.

We assume that the sound speed correction factor $\Sigma(s)$ and the area function $A(s)$ are continuously differentiable for $s \in [0,1]$, and that the estimates

$$0 < \min_{s\in[0,1]} A(s) \leq \max_{s\in[0,1]} A(s) < \infty \text{ and } 0 < \min_{s\in[0,1]} c(s) \leq \max_{s\in[0,1]} c(s) < \infty \tag{4.3}$$

hold. These are natural assumptions recalling the geometry of the tubular domain Ω. Define the operators

$$W := \frac{1}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial}{\partial s}\right) \text{ and } D := -\frac{2\pi W(s)}{A(s)}. \tag{4.4}$$

The operator D should be understood as a multiplication operator on $L^2(0,1)$ by the strictly negative function $-2\pi W(\cdot)A(\cdot)^{-1}$. Then the first of the equations in (4.1) can be cast into first order form by using the rule

$$\psi_{tt} = c(s)^2\left(W\psi + \alpha D\psi_t\right) \quad \hat{=} \quad \frac{d}{dt}\begin{bmatrix}\psi \\ \pi\end{bmatrix} = \begin{bmatrix} 0 & \rho^{-1} \\ \rho c(s)^2 W & \alpha c(s)^2 D\end{bmatrix}\begin{bmatrix}\psi \\ \pi\end{bmatrix}.$$

Henceforth, let

$$L_W := \begin{bmatrix} 0 & \rho^{-1} \\ \rho c(s)^2 W & 0 \end{bmatrix} : \mathcal{Z}_W \to \mathcal{X}_W \text{ and } H_W := \begin{bmatrix} 0 & 0 \\ 0 & c(s)^2 D \end{bmatrix} : \mathcal{X}_W \to \mathcal{X}_W$$

where the Hilbert spaces are given by

$$\mathcal{Z}_W := \Big(H^1_{\{1\}}(0,1) \cap H^2(0,1) \Big) \times H^1_{\{1\}}(0,1), \quad \mathcal{X}_W := H^1_{\{1\}}(0,1) \times L^2(0,1)$$
$$\text{where} \quad H^1_{\{1\}}(0,1) := \big\{ f \in H^1(0,1) : f(1) = 0 \big\}.$$

Clearly we have $H_W \in \mathcal{L}(\mathcal{X}_W)$, $H_W^* = H_W$, and this operator is negative in the sense that $\langle H_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right], \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \rangle_{\mathcal{X}_W} = -2\pi \int_0^1 |z_2(s)|^2 W(s) c(s)^2 A(s)^{-1} \, ds \leq 0$. So, the operator αH_W for $\alpha > 0$ satisfies assumption (i) of Theorem 3.2 with $b = 0$ and also assumption (ii) of the same theorem with $\tilde{\mathcal{X}} = \mathcal{X}$.

The Hilbert spaces $\mathcal{Z}_W$ and $\mathcal{X}_W$ are equipped with the norms

$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{\mathcal{Z}_W} := \|z_1\|^2_{H^2(0,1)} + \|z_2\|^2_{H^1(0,1)} \quad \text{and}$$
$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{H^1(0,1) \times L^2(0,1)} := \|z_1\|^2_{H^1(0,1)} + \|z_2\|^2_{L^2(0,1)},$$

respectively. We will use the *energy norm* on $\mathcal{X}_W$, which for any $\rho > 0$ is defined by

$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{\mathcal{X}_W} := \frac{1}{2} \left(\rho \int_0^1 \left| z_1'(s) \right|^2 A(s) \, ds + \frac{1}{\rho c^2} \int_0^1 |z_2(s)|^2 \, A(s) \Sigma(s)^{-2} \, ds \right). \tag{4.5}$$

This is an equivalent norm for $\mathcal{X}_W$ because the conditions (4.3) hold and $\sqrt{2}\|z_1\|_{L^2(0,1)} \leq \|z_1'\|_{L^2(0,1)}$ for all $z_1 \in H^1_{\{1\}}(0,1)$. To see that the Poincaré inequality holds in $H^1_{\{1\}}(0,1)$, note that for smooth functions z with $z(1) = 0$, one has from the fundamental theorem of calculus that

$$|z(s)| = \left| \int_s^1 z'(t) \, dt \right| \leq (1-s)^{1/2} \|z'\|_{L^2(0,1)}.$$

From this, we proceed by squaring and integrating with respect to s, and then passing to general Sobolev functions by approximation.

We define $\mathcal{U}_W := \mathbb{C}$ with the absolute value norm $\|u_0\|_{\mathcal{U}_W} := |u_0|$. The endpoint control and observation functionals $G_W : \mathcal{Z}_W \to \mathcal{U}_W$ and $K_W : \mathcal{Z}_W \to \mathcal{U}_W$ are defined by

$$G_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] := \frac{1}{2} \sqrt{\frac{A(0)}{\rho c(0)}} \left(-\rho c(0) z_1'(0) + z_2(0) \right) \quad \text{and}$$
$$K_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] := \frac{1}{2} \sqrt{\frac{A(0)}{\rho c(0)}} \left(-\rho c(0) z_1'(0) - z_2(0) \right).$$

Now the generalised Webster's horn model (4.1)–(4.2) for the state $z(t) = \left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right]$ takes the form

$$\begin{cases} \frac{d}{dt}\left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right] = (L_W + \alpha H_W)\left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right], \\ \tilde{u}(t) = G_W\left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right], \end{cases} \tag{4.6}$$

and

$$\tilde{y}(t) = K_W\left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right] \tag{4.7}$$

for all $t \in \mathbb{R}^+$. The initial conditions are $\left[\begin{smallmatrix}\psi(0)\\ \pi(0)\end{smallmatrix}\right] = \left[\begin{smallmatrix}\psi_0\\ \pi_0\end{smallmatrix}\right]$. The state variable $\pi = \rho\psi_t$ has the dimension of pressure, as for the wave equation.

The impedance passive version of the following Theorem 4.1 is given in [2, Theorem 5.1], and it would be possible to deduce parts of Theorem 4.1 from that result using the external Cayley transform [30, Definition 3.1]. Here we give a direct proof instead.

Theorem 4.1. *Let the operators L_W, H_W, G_W, K_W, and spaces $\mathcal{Z}_W$, $\mathcal{X}_W$, $\mathcal{U}_W$ be defined as above. Let $\left[\begin{smallmatrix}\psi_0\\ \pi_0\end{smallmatrix}\right] \in \mathcal{Z}_W$ and $\tilde{u} \in C^2(\mathbb{R}^+;\mathbb{C})$ such that the compatibility condition $G_W\left[\begin{smallmatrix}\psi_0\\ \pi_0\end{smallmatrix}\right] = \tilde{u}(0)$ holds. Then for all $\alpha \geq 0$ the following holds:*

(i) *The triple $\Xi_\alpha^{(W)} := (G_W, L_W + \alpha H_W, K_W)$ is a scattering passive, strong boundary node on Hilbert spaces $(\mathcal{U}_W, \mathcal{X}_W, \mathcal{U}_W)$.*

The semigroup generator $A_{W,\alpha} = (L_W + \alpha H_W)\big|_{\ker(G_W)}$ of $\Xi_\alpha^{(W)}$ satisfies $A_{W,\alpha}^* = (-L_W + \alpha H_W)\big|_{\ker(K_W)}$ and $0 \in \rho(A_{W,\alpha}) \cap \rho(A_{W,\alpha}^*)$.

(ii) *The equations in (4.6) have a unique solution $\left[\begin{smallmatrix}\psi\\ \pi\end{smallmatrix}\right] \in C^1(\mathbb{R}^+;\mathcal{X}_W) \cap C(\mathbb{R}^+;\mathcal{Z}_W)$. Hence we can define $\tilde{y} \in C(\mathbb{R}^+;\mathbb{C})$ by equation (4.7).*

(iii) *The solution of (4.6) satisfies the energy dissipation inequality*

$$\frac{d}{dt}\|\left[\begin{smallmatrix}\psi(t)\\ \pi(t)\end{smallmatrix}\right]\|_{\mathcal{X}_W}^2 \leq |\tilde{u}(t)|^2 - |\tilde{y}(t)|^2, \quad t \in \mathbb{R}^+. \tag{4.8}$$

Moreover, $\Xi_0^{(W)}$ is a conservative boundary node, and (4.8) holds then as an equality.

Under the assumptions of this proposition, we have $\psi \in C(\mathbb{R}^+;H^2(0,1)) \cap C^1(\mathbb{R}^+;H^1(0,1)) \cap C^2(\mathbb{R}^+;L^2(0,1))$.

Proof. Claim (i): By Theorem 3.2, it is enough to show the conservative case $\alpha = 0$. Let us first verify the that the Green–Lagrange identity

$$2\mathrm{Re}\,\langle\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], L_W\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\rangle_{\mathcal{X}_W} + |K_W\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]|^2 = |G_W\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]|^2 \tag{4.9}$$

holds for all $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \mathcal{Z}_W$. By partial integration, we get

$$2\mathrm{Re}\,\langle \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right], L_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\rangle_{\mathcal{X}_W} = -A(0)\mathrm{Re}\left(\overline{z_1'(0)}z_2(0)\right).$$

Now (4.9) follows since $|G_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]|^2 - |K_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]|^2 = -A(0)\mathrm{Re}\left(\overline{z_1'(0)}z_2(0)\right)$ just as in equations (5.14) – (5.15).

It is trivial that $G_W\mathcal{Z}_W = K_W\mathcal{Z}_W = \mathcal{U}_W$ since $\dim \mathcal{U}_W = 1$ and neither of the operators G_W and K_W vanishes. We prove next that L_W maps $\ker(G_W)$ *bijectively* onto $\mathcal{X}_W$. Now, $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \ker(G_W)$ and $\left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right] \in \mathcal{X}_W$ satisfy $L_W \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right]$ if and only if $z_2 = \rho w_1$ and

$$\frac{\partial}{\partial s}\left(A(\cdot)\frac{\partial z_1}{\partial s}\right) = \frac{A(\cdot)w_2}{\rho c(\cdot)^2}, \quad z_1(1) = 0, \quad z_1'(0) = \frac{w_1(0)}{c(0)}.$$

Since this equation has always a unique solution $z_1 \in H^2(0,1)$ for any $w_1 \in H^1_{\{1\}}(0,1)$ and $w_2 \in L^2(0,1)$, it follows that $L_W\ker(G_W) = \mathcal{X}_W$ and $0 \in \rho(A_{W,0})$ where $A_{W,0} = L_W\big|_{\ker(G_W)}$ is the semigroup generator of $\Xi_0^{(W)}$. We conclude by Proposition 2.4 that $\Xi_0^{(W)}$ is a conservative boundary node as claimed. That $\Xi_\alpha^{(W)}$ is passive for $\alpha > 0$ with semigroup generator $A_{W,\alpha} = (L_W + \alpha H_W)\big|_{\ker(G_W)}$ follows by Theorem 3.2.

Because $H_W^* = H_W \in \mathcal{L}(\mathcal{X})$ is dissipative, we may apply Theorem 3.2 again to the time-flow inverted, conservative node $\left(\Xi_0^{(W)}\right)^{\leftarrow} = (K_W, -L_W, G_W)$ to conclude that the boundary node $(K_W, -L_W + \alpha H_W^*, G_W)$ is passive as well. Claim (iii) of Proposition 3.3 implies that $A_{W,\alpha}^* = (-L_W + \alpha H_W)\big|_{\ker(K_W)}$.

Let us argue next that $0 \in \rho(A_{W,\alpha}) \cap \rho(A_{W,\alpha}^*)$ for $\alpha > 0$. Because $A_{W,\alpha}$ is a compact resolvent operator, it is enough to exclude $0 \in \sigma_p(A_{W,\alpha})$. Suppose $A_{W,\alpha}z_0 = 0$, giving $\mathrm{Re}\,\langle A_{W,0}z_0, z_0\rangle_{\mathcal{X}} + \mathrm{Re}\,\langle \alpha H_W z_0, z_0\rangle_{\mathcal{X}} = \mathrm{Re}\,\langle A_{W,\alpha}z_0, z_0\rangle_{\mathcal{X}} = 0$. Thus

$$\mathrm{Re}\,\langle A_{W,0}z_0, z_0\rangle_{\mathcal{X}} = \alpha\mathrm{Re}\,\langle -H_W z_0, z_0\rangle_{\mathcal{X}} = \alpha\|(-H_W)^{1/2}z_0\|_{\mathcal{X}}^2 = 0$$

by the dissipativity of both $A_{W,0}$ and H_W, and the fact that $-H_W$ is a self-adjoint nonnegative operator. Thus $z_0 \in \ker(H_W)$ and hence $A_{W,0}z_0 = (A_{W,0} + \alpha H_W)z_0 = A_{W,\alpha}z_0 = 0$. Because $0 \in \rho(A_{W,0})$ has already been shown, we conclude that $z_0 = 0$.

The node $\Xi_0^{(W)}$ is strong (i.e., L_W is closed with $\mathrm{dom}\,(L_W) = \mathcal{Z}_W$) since $L_W = L_W^{**}$ and $L_W^* = -L_W\big|_{\mathrm{dom}(L_W^*)}$ where

$$\mathrm{dom}\,(L_W^*) = \left\{ \left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right] \in H^1_{\{1\}}(0,1) \cap H^2(0,1) \times H^1_0(0,1) \,:\, \tfrac{\partial w_1}{\partial s}(0) = 0 \right\}$$

which is dense in $\mathcal{X}_W$ and satisfies $\mathrm{dom}\,(L_W^*) \subset \mathrm{dom}\,(L_W)$. That $\Xi_\alpha^{(W)}$ is strong for $\alpha > 0$ follows from $H_W \in \mathcal{L}(\mathcal{X})$ as explained in Theorem 3.2.

Claims (ii) and (iii) follow from Proposition 2.2 and Eq. (2.4). □

5 Passive wave equation on wave guides

Define the tubular domain $\Omega \subset \mathbb{R}^3$ and its boundary components Γ, $\Gamma(0)$, and $\Gamma(1)$ as in Section 1. Each of the sets Γ, $\Gamma(0)$, and $\Gamma(1)$ are smooth manifolds but $\partial\Omega = \overline{\Gamma} \cup \Gamma(0) \cup \Gamma(1)$ is only Lipschitz. Other relevant properties of Ω and $\partial\Omega$ are listed in (i) – (iii) of Appendix A where we also make rigorous sense of the Sobolev spaces, boundary trace mappings, Poincaré inequality, and the Green's identity for such domains.

Following [26, Section 3], we consider the linear dynamical system described by

$$\begin{cases} \phi_{tt}(\mathbf{r},t) = c^2\Delta\phi(\mathbf{r},t) \quad \text{for } \mathbf{r}\in\Omega \text{ and } t\in\mathbb{R}^+, \\ c\frac{\partial\phi}{\partial\nu}(\mathbf{r},t) + \phi_t(\mathbf{r},t) = 2\sqrt{\frac{c}{\rho A(0)}}\,u(\mathbf{r},t) \quad \text{for } \mathbf{r}\in\Gamma(0) \text{ and } t\in\mathbb{R}^+, \\ \phi(\mathbf{r},t) = 0 \quad \text{for } \mathbf{r}\in\Gamma(1) \text{ and } t\in\mathbb{R}^+, \\ \frac{\partial\phi}{\partial\nu}(\mathbf{r},t) + \alpha\phi_t(\mathbf{r},t) = 0 \quad \text{for } \mathbf{r}\in\Gamma, \text{ and } t\in\mathbb{R}^+, \text{ and} \\ \phi(\mathbf{r},0) = \phi_0(\mathbf{r}), \quad \rho\phi_t(\mathbf{r},0) = p_0(\mathbf{r}) \quad \text{for } \mathbf{r}\in\Omega, \end{cases} \tag{5.1}$$

together with the observation y defined by

$$c\frac{\partial\phi}{\partial\nu}(\mathbf{r},t) - \phi_t(\mathbf{r},t) = 2\sqrt{\frac{c}{\rho A(0)}}\,y(\mathbf{r},t) \quad \text{for } \mathbf{r}\in\Gamma(0) \text{ and } t\in\mathbb{R}^+. \tag{5.2}$$

This model describes acoustics of a cavity Ω that has an open end at $\Gamma(1)$ and an energy dissipating wall Γ. The solution ϕ is the *velocity potential* as its gradient is the perturbation velocity field of the acoustic waves. The boundary control and observation on surface $\Gamma(0)$ (whose area is $A(0)$) are both of scattering type. The speed of sound is denoted by $c > 0$. The constants $\alpha \geq 0$ and $\rho > 0$ have physical meaning but we refer to [26] for details. Note that if $\alpha = 0$, we have the Neumann boundary condition modelling a hard, sound reflecting boundary on Γ. Our purpose is to show that (5.1)–(5.2) defines a passive boundary node (conservative, if $\alpha = 0$ by a slightly different argument in Corollary 5.2) by using Theorem 3.1 with the aid of the additional signals $\tilde{u} := \frac{1}{\sqrt{\alpha}}\frac{\partial\phi}{\partial\nu} + \sqrt{\alpha}\phi_t$ (that will be grounded) and $\tilde{y} := \frac{1}{\sqrt{\alpha}}\frac{\partial\phi}{\partial\nu} - \sqrt{\alpha}\phi_t$ (that will be disregarded) on the wall Γ.

The boundedness of the Dirichlet trace implies that the space

$$H^1_{\Gamma(1)}(\Omega) := \left\{ f \in H^1(\Omega) : f\big|_{\Gamma(1)} = 0 \right\}. \tag{5.3}$$

is a closed subspace of $H^1(\Omega)$. Define

$$\tilde{\mathcal{Z}}' := \{ f \in H^1_{\Gamma(1)}(\Omega) : \Delta f \in L^2(\Omega), \frac{\partial f}{\partial\nu}\big|_{\Gamma(0)\cup\Gamma} \in L^2(\Gamma(0)\cup\Gamma) \} \tag{5.4}$$

with the norm $\|f\|^2_{\tilde{\mathcal{Z}}'} = \|f\|^2_{H^1(\Omega)} + \|\Delta f\|^2_{L^2(\Omega)} + \|\frac{\partial f}{\partial\nu}\big|_{\Gamma(0)\cup\Gamma}\|^2_{L^2(\Gamma(0)\cup\Gamma)}$. Then

the operator

$$\frac{\partial}{\partial\nu}\big|_{\Gamma'} : f \mapsto \frac{\partial f}{\partial\nu}\big|_{\Gamma'} \quad \text{lies in} \quad \mathcal{L}(\tilde{\mathcal{Z}}'; L^2(\Gamma')) \quad \text{for} \quad \Gamma' \in \{\Gamma(0), \Gamma, \Gamma(0)\cup\Gamma\}. \tag{5.5}$$

The spaces $\tilde{\mathcal{Z}}$, $\mathcal{X}$, and the interior operator L are defined by

$$\begin{aligned} L &:= \left[\begin{smallmatrix} 0 & \rho^{-1} \\ \rho c^2\Delta & 0 \end{smallmatrix}\right] : \tilde{\mathcal{Z}} \to \mathcal{X} \quad \text{with} \\ \tilde{\mathcal{Z}} &:= \tilde{\mathcal{Z}}' \times H^1_{\Gamma(1)}(\Omega) \quad \text{and} \quad \mathcal{X} := H^1_{\Gamma(1)}(\Omega) \times L^2(\Omega) \end{aligned} \tag{5.6}$$

where $H^1_{\Gamma(1)}(\Omega)$ and $\tilde{\mathcal{Z}}'$ are given by (5.3) and (5.4). For the space $\mathcal{X}$, we use the *energy norm*

$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{\mathcal{X}} := \frac{1}{2}\left(\rho \| |\nabla z_1| \|^2_{L^2(\Omega)} + \frac{1}{\rho c^2}\|z_2\|^2_{L^2(\Omega)}\right). \tag{5.7}$$

The Poincaré inequality $\|z_1\|_{L^2(\Omega)} \le M_\Omega \|\nabla z_1\|_{L^2(\Omega)}$ holds for $z_1 \in H^1_{\Gamma(1)}(\Omega)$ as given in Theorem A.4 in Appendix A. Therefore (5.7) defines a norm on $\mathcal{X}$, equivalent to the Cartesian product norm

$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{H^1(\Omega)\times L^2(\Omega)} := \|z_1\|^2_{L^2(\Omega)} + \|\nabla z_1\|^2_{L^2(\Omega)} + \|z_2\|^2_{L^2(\Omega)}$$

so that $\tilde{\mathcal{Z}} \subset \mathcal{X}$ with a continuous embedding, and $L \in \mathcal{L}(\tilde{\mathcal{Z}}; \mathcal{X})$ with respect to the $\tilde{\mathcal{Z}}$-norm

$$\| \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \|^2_{\tilde{\mathcal{Z}}} := \|z_1\|^2_{\tilde{\mathcal{Z}}'} + \|z_2\|^2_{L^2(\Omega)} + \|\nabla z_2\|^2_{L^2(\Omega)}.$$

Defining $\mathcal{U} := L^2(\Gamma(0))$ and $\tilde{\mathcal{U}} := L^2(\Gamma)$ with the norms

$$\|u_0\|^2_{\mathcal{U}} = A(0)^{-1}\|u_0\|^2_{L^2(\Gamma(0))} \text{ and } \|\tilde{u}_0\|_{\tilde{\mathcal{U}}} = \|\tilde{u}_0\|_{L^2(\Gamma)}, \tag{5.8}$$

we get $\mathcal{U} \oplus \tilde{\mathcal{U}} = L^2(\Gamma(0)\cup\Gamma)$ where we use the Cartesian product norm of $\mathcal{U}$ and $\tilde{\mathcal{U}}$.

The boundedness of the Dirichlet trace and the property (5.5) of the Neumann trace imply that $\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right] \in \mathcal{L}(\tilde{\mathcal{Z}}; \mathcal{U}\oplus\tilde{\mathcal{U}})$ and $\left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right] \in \mathcal{L}(\tilde{\mathcal{Z}}; \mathcal{U}\oplus\tilde{\mathcal{U}})$ where

$$\begin{aligned} \begin{bmatrix} G \\ G_\alpha \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} &:= \frac{1}{2}\begin{bmatrix} \sqrt{\frac{A(0)}{\rho c}}\left(\rho c \frac{\partial z_1}{\partial\nu}\big|_{\Gamma(0)} + z_2\big|_{\Gamma(0)}\right) \\ \frac{\sqrt{\rho}}{\sqrt{\alpha}}\frac{\partial z_1}{\partial\nu}\big|_{\Gamma} + \frac{\sqrt{\alpha}}{\sqrt{\rho}} z_2\big|_{\Gamma} \end{bmatrix} \quad \text{and} \\ \begin{bmatrix} K \\ K_\alpha \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} &:= \frac{1}{2}\begin{bmatrix} \sqrt{\frac{A(0)}{\rho c}}\left(\rho c \frac{\partial z_1}{\partial\nu}\big|_{\Gamma(0)} - z_2\big|_{\Gamma(0)}\right) \\ \frac{\sqrt{\rho}}{\sqrt{\alpha}}\frac{\partial z_1}{\partial\nu}\big|_{\Gamma} - \frac{\sqrt{\alpha}}{\sqrt{\rho}} z_2\big|_{\Gamma} \end{bmatrix}. \end{aligned} \tag{5.9}$$

The reason for defining the triple $\widetilde{\Xi}_\alpha := (\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right], L, \left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right])$ is to obtain first order equations from (5.1), using the equivalence of $\phi_{tt} = c^2\Delta\phi$ and

$\frac{d}{dt}\left[\begin{smallmatrix}\phi\\ p\end{smallmatrix}\right] = \left[\begin{smallmatrix}0 & \rho^{-1}\\ \rho c^2\Delta & 0\end{smallmatrix}\right]\left[\begin{smallmatrix}\phi\\ p\end{smallmatrix}\right]$ where $p = \rho\phi_t$ is the sound pressure. More precisely, equations (5.1)–(5.2) are (at least formally) equivalent with

$$\begin{cases} \frac{d}{dt}\begin{bmatrix}\phi(t)\\ p(t)\end{bmatrix} = L\begin{bmatrix}\phi(t)\\ p(t)\end{bmatrix}, \\ \begin{bmatrix}u(t)\\ 0\end{bmatrix} = \begin{bmatrix}G\\ G_\alpha\end{bmatrix}\begin{bmatrix}\phi(t)\\ p(t)\end{bmatrix}, \end{cases} \tag{5.10}$$

and

$$\begin{bmatrix}y(t)\\ \tilde{y}(t)\end{bmatrix} = \begin{bmatrix}K\\ K_\alpha\end{bmatrix}\begin{bmatrix}\phi(t)\\ p(t)\end{bmatrix} \tag{5.11}$$

for $t \in \mathbb{R}^+$, with the initial conditions $\left[\begin{smallmatrix}\phi(0)\\ p(0)\end{smallmatrix}\right] = \left[\begin{smallmatrix}\phi_0\\ p_0\end{smallmatrix}\right]$. The *Green–Lagrange identity*

$$2\mathrm{Re}\,\langle \left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], L\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\rangle_{\mathcal{X}} + \|\left[\begin{smallmatrix}K\\ K_\alpha\end{smallmatrix}\right]\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\|^2_{\mathcal{U}\oplus\tilde{\mathcal{U}}} = \|\left[\begin{smallmatrix}G\\ G_\alpha\end{smallmatrix}\right]\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\|^2_{\mathcal{U}\oplus\tilde{\mathcal{U}}} \text{ for all } \left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right] \in \tilde{\mathcal{Z}} \tag{5.12}$$

is a key fact for proving the conservativity of $\widetilde{\Xi}_\alpha$, and we verify it next. Green's identity (Theorem A.3 in Appendix A) gives

$$\begin{aligned} &2\mathrm{Re}\,\langle \left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], L\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\rangle_{\mathcal{X}} = 2\mathrm{Re}\,\left\langle \left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], \left[\begin{smallmatrix}\rho^{-1}z_2\\ \rho c^2\Delta z_1\end{smallmatrix}\right]\right\rangle_{\mathcal{X}} \\ &= 2\mathrm{Re}\,\frac{1}{2}\left(\rho\int_\Omega \nabla\overline{z_1}\cdot\nabla(z_2/\rho)\,\mathrm{d}V + \frac{1}{\rho c^2}\left\langle \rho c^2\Delta\overline{z_1}, z_2\right\rangle_{L^2(\Omega)}\right) \\ &= \mathrm{Re}\left(\int_{\Gamma(0)\cup\Gamma\cup\Gamma(1)} \frac{\partial\overline{z_1}}{\partial\nu} z_2\,\mathrm{d}A\right) \\ &= \mathrm{Re}\left\langle \frac{\partial z_1}{\partial\nu}\Big|_{\Gamma(0)}, z_2\Big|_{\Gamma(0)}\right\rangle_{L^2(\Gamma(0))} + \mathrm{Re}\left\langle \frac{\partial z_1}{\partial\nu}\Big|_{\Gamma}, z_2\Big|_{\Gamma}\right\rangle_{L^2(\Gamma)} \end{aligned} \tag{5.13}$$

because $z_2\big|_{\Gamma(1)} = 0$ by (5.6). On the other hand, we obtain

$$\begin{aligned} &\|G\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\|^2_{\mathcal{U}} = A(0)^{-1}\left\langle G\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], G\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\right\rangle_{L^2(\Gamma(0))} \\ &= \frac{1}{4\rho c}\left(\rho^2c^2\left\|\frac{\partial z_1}{\partial\nu}\Big|_{\Gamma(0)}\right\|^2_{L^2(\Gamma(0))} + 2\rho c\,\mathrm{Re}\left\langle \frac{\partial z_1}{\partial\nu}\Big|_{\Gamma(0)}, z_2\Big|_{\Gamma(0)}\right\rangle_{L^2(\Gamma(0))} + \left\|z_2\big|_{\Gamma(0)}\right\|^2_{L^2(\Gamma(0))}\right) \end{aligned} \tag{5.14}$$

and also

$$\begin{aligned} &\|K\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\|^2_{\mathcal{U}} = A(0)^{-1}\left\langle K\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right], K\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]\right\rangle_{L^2(\Gamma(0))} \\ &= \frac{1}{4\rho c}\left(\rho^2c^2\left\|\frac{\partial z_1}{\partial\nu}\Big|_{\Gamma(0)}\right\|^2_{L^2(\Gamma(0))} - 2\rho c\,\mathrm{Re}\left\langle \frac{\partial z_1}{\partial\nu}\Big|_{\Gamma(0)}, z_2\Big|_{\Gamma(0)}\right\rangle_{L^2(\Gamma(0))} + \left\|z_2\big|_{\Gamma(0)}\right\|^2_{L^2(\Gamma(0))}\right), \end{aligned} \tag{5.15}$$

where $G\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]$ and $K\left[\begin{smallmatrix}z_1\\ z_2\end{smallmatrix}\right]$ are the first components in (5.9) respectively.

Similarly, we compute the two terms needed in

$$\|G_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\|_{\tilde{\mathcal{U}}}^2 - \|K_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\|_{\tilde{\mathcal{U}}}^2 \tag{5.16}$$
$$= \langle G_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right], G_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\rangle_{L^2(\Gamma)} - \langle K_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right], K_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\rangle_{L^2(\Gamma)} = \mathrm{Re} \left\langle \frac{\partial z_1}{\partial \nu}\Big|_\Gamma, z_2\Big|_\Gamma \right\rangle_{L^2(\Gamma)},$$

where $G_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ and $K_\alpha \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ are the second components in (5.9) respectively. Now (5.13) – (5.16) implies (5.12) as required.

We proceed to show that the the triple $\Xi_\alpha := (G\big|_{\mathcal{Z}_\alpha}, L\big|_{\mathcal{Z}_\alpha}, K\big|_{\mathcal{Z}_\alpha})$ for all $\alpha > 0$ is a scattering passive boundary node on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{U})$ with the solution space

$$\mathcal{Z}_\alpha := \left\{ \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \in \tilde{\mathcal{Z}}' \times H^1_{\Gamma(1)}(\Omega) : \frac{\partial z_1}{\partial \nu}\Big|_\Gamma + \frac{\alpha}{\rho} z_2\Big|_\Gamma = 0 \right\}. \tag{5.17}$$

Note that $\mathcal{Z}_\alpha$ is a closed subspace of $\tilde{\mathcal{Z}}$ because $G_\alpha \in \mathcal{L}(\tilde{\mathcal{Z}}; \tilde{\mathcal{U}})$ and $\mathcal{Z}_\alpha = \ker(G_\alpha)$. Therefore, we can use the norm of $\tilde{\mathcal{Z}}$ on $\mathcal{Z}_\alpha$. The conservative case $\alpha = 0$ is slightly different, and it is treated separately in Corollary 5.2.

Theorem 5.1. *Take $\alpha > 0$ and let the operators L, G, K, and Hilbert spaces $\mathcal{X}$, $\mathcal{U}$, and $\mathcal{Z}_\alpha$ be defined as above. Let $\left[\begin{smallmatrix} \phi_0 \\ p_0 \end{smallmatrix}\right] \in \mathcal{Z}_\alpha$ and $u \in C^2(\mathbb{R}^+; \mathcal{U})$ such that the compatibility condition $G \left[\begin{smallmatrix} \phi_0 \\ p_0 \end{smallmatrix}\right] = u(0)$ holds. Then the following holds:*

(i) *The triple $\Xi_\alpha := (G\big|_{\mathcal{Z}_\alpha}, L\big|_{\mathcal{Z}_\alpha}, K\big|_{\mathcal{Z}_\alpha})$ is a scattering passive boundary node on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{U})$ with solution space $\mathcal{Z}_\alpha$. The semigroup generator $A_\alpha = L\big|_{\ker(G)\cap\ker(G_\alpha)}$ of Ξ_α satisfies $A_\alpha^* = -L\big|_{\ker(K)\cap\ker(K_\alpha)}$ and $0 \in \rho(A_\alpha) \cap \rho(A_\alpha^*)$.*

(ii) *The equations[4] in (5.10) have a unique solution $\left[\begin{smallmatrix} \phi \\ p \end{smallmatrix}\right] \in C^1(\mathbb{R}^+; \mathcal{X}) \cap C(\mathbb{R}^+; \mathcal{Z}_\alpha)$. Hence we can define $y \in C(\mathbb{R}^+; \mathcal{U})$ by equation (5.11).*

(iii) *The solution of (5.10) satisfies the energy dissipation inequality*

$$\frac{d}{dt} \| \left[\begin{smallmatrix} \phi(t) \\ p(t) \end{smallmatrix}\right] \|_{\mathcal{X}}^2 \le \|u(t)\|_{\mathcal{U}}^2 - \|y(t)\|_{\mathcal{U}}^2, \quad t \in \mathbb{R}^+. \tag{5.18}$$

It follows from claim (ii) and the definition of the norms of $\mathcal{Z}_\alpha$ and $\mathcal{X}$ that $\phi \in C^1(\mathbb{R}^+; H^1(\Omega)) \cap C^2(\mathbb{R}^+; L^2(\Omega))$, $\nabla\phi \in C^1(\mathbb{R}^+; L^2(\Omega; \mathbb{R}^3))$, and $\Delta\phi \in C(\mathbb{R}^+; L^2(\Omega))$. These are the same smoothness properties that have been used in [26, see, in particular, Eq. (1.4)] for deriving the generalised Webster's equation in (1.3) from the wave equation.

[4]Note that (2.1) is equivalent with (5.1) and (5.10) in the context of this theorem.

Proof. Claim (i): By Theorem 3.1 and the discussion preceding this theorem, it is enough to show that $\widetilde{\Xi}_\alpha = (\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right], L, \left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right])$ introduced above is a conservative boundary node which is easiest done by using Proposition 2.4. Since the Green–Lagrange identity (2.6) has already been established, it remains to prove conditions (ii) (with $\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right]$ in place of G) and (iv) (with $\left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right]$ in place of K) of Proposition 2.4 with $\beta = \gamma = 0$. It is enough to consider only $\beta = \gamma = 0$ because the resolvent sets of $L\big|_{\ker(G)}$ and $-L\big|_{\ker(K)}$ in Proposition 2.4 are open, and then the same conditions hold for some $\beta, \gamma > 0$ as well.

For an arbitrary $g \in L^2(\Gamma(0) \cup \Gamma)$ there exists a unique variational[5] solution $z_1 \in H^1_{\Gamma(1)}(\Omega)$ of the problem

$$\Delta z_1 = 0, \quad z_1\big|_{\Gamma(1)} = 0, \quad \frac{\partial z_1}{\partial \nu}\Big|_{\Gamma(0)\cup\Gamma} = g. \tag{5.19}$$

Since $z_1 \in \tilde{\mathcal{Z}}'$, we have $\frac{\partial}{\partial \nu}\big|_{\Gamma(0)\cup\Gamma}\tilde{\mathcal{Z}}' = L^2(\Gamma(0)\cup\Gamma)$ which obviously gives both $\frac{\partial}{\partial \nu}\big|_{\Gamma(0)}\tilde{\mathcal{Z}}' = L^2(\Gamma(0))$ and $\frac{\partial}{\partial \nu}\big|_{\Gamma}\tilde{\mathcal{Z}}' = L^2(\Gamma)$. Clearly $\tilde{\mathcal{Z}}' \oplus \{0\} \subset \tilde{\mathcal{Z}}$ and the surjectivity of $\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right]$ follows from

$$\begin{bmatrix} G \\ G_\alpha \end{bmatrix}\begin{bmatrix} z_1 \\ 0 \end{bmatrix} := \frac{1}{2}\begin{bmatrix} \sqrt{A(0)\rho c}\frac{\partial}{\partial \nu}\big|_{\Gamma(0)} \\ \frac{\sqrt{\rho}}{\sqrt{\alpha}}\frac{\partial}{\partial \nu}\big|_{\Gamma} \end{bmatrix} z_1.$$

To see this, for a given $h \in L^2(\Gamma(0)\cup\Gamma)$, we choose

$$g = \begin{cases} 2\frac{1}{\sqrt{A(0)\rho c}}h, & \text{on } \Gamma(0), \\ 2\frac{\sqrt{\alpha}}{\sqrt{\rho}}h, & \text{on } \Gamma \end{cases}$$

in (5.19) to find a function z_1 so that $\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right]\left[\begin{smallmatrix} z_1 \\ 0 \end{smallmatrix}\right] = h$. The surjectivity of $\left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right]$ is proved similarly.

To show that $L\ker\left(\left[\begin{smallmatrix} G \\ G_\alpha \end{smallmatrix}\right]\right) = L\left(\ker(G) \cap \ker(G_\alpha)\right) = \mathcal{X}$, let $\left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right] \in \mathcal{X}$ be arbitrary. Then $\left[\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right] = L\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} \rho^{-1}z_2 \\ \rho c^2 \Delta z_1 \end{smallmatrix}\right]$ for $\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \ker(G) \cap \ker(G_\alpha)$ if and only if $z_2 = \rho w_1$ and the variational solution $z_1 \in H^1_{\Gamma(1)}(\Omega)$ of the problem

$$\rho c^2 \Delta z_1 = w_2, \quad z_1\big|_{\Gamma(1)} = 0, \quad \frac{\partial z_1}{\partial \nu}\Big|_{\Gamma} = -\alpha\rho w_1\big|_{\Gamma}, \quad c\frac{\partial z_1}{\partial \nu}\Big|_{\Gamma(0)} = -w_1\big|_{\Gamma(0)}$$

exists and belongs to the space $\mathcal{Z}'$. Now, this condition can be verified by standard variational techniques because $w_2 \in L^2(\Omega)$ and $w_1 \in H^1_{\Gamma(1)}(\Omega)$ which implies $w_1\big|_{\Gamma(0)\cup\Gamma} \in H^{1/2}(\Gamma(0)\cup\Gamma) \subset L^2(\Gamma(0)\cup\Gamma)$. That $L\ker\left(\left[\begin{smallmatrix} K \\ K_\alpha \end{smallmatrix}\right]\right) =$

[5] We leave it to the interested reader to derive the variational form using Green's identity (A.9) and then carry out the usual argument by the Lax–Milgram theorem; see, e.g., [12, Lemma 2.2.1.1].

$\mathcal{X}$ is proved similarly. All the conditions of Proposition 2.4 are now satisfied with $\beta = \gamma = 0$, and thus $\tilde{\Xi}_\alpha$ is a conservative boundary node. It now follows from Theorem 3.1 that Ξ_α is a passive boundary node which has the common semigroup generator $A_\alpha = L\big|_{\ker(G)\cap\ker(G_\alpha)}$ with the original conservative boundary node $\tilde{\Xi}_\alpha$. By [29, Theorem 1.9 and Proposition 4.3], the dual system of $\widetilde{\Xi}_\alpha$ is of boundary control type, and it coincides with the time-flow inverted boundary node $\widetilde{\Xi}_\alpha^{\leftarrow}$. Now, the unbounded adjoint A_α^* is the semigroup generator of the dual system $\widetilde{\Xi}_\alpha^{\leftarrow}$, and hence $A_\alpha^* = -L\big|_{\ker(K)\cap\ker(K_\alpha)}$ as claimed.

It remains to show that $0 \notin \sigma(A_\alpha)$. We have already shown above that $A_\alpha \mathrm{dom}\,(A_\alpha) = \mathcal{X}$ with $\mathrm{dom}\,(A_\alpha) = \ker(G) \cap \ker(G_\alpha)$, and the remaining injectivity part follows if we show that $\ker(L) \cap \ker(G) \cap \ker(G_\alpha) = \{0\}$. This follows because the variational solution in $H^1(\Omega)$ of the homogenous problem

$$\Delta z_1 = 0, \quad z_1\big|_{\Gamma(1)} = 0, \quad \frac{\partial z_1}{\partial \nu}\Big|_{\Gamma(0)\cup\Gamma} = 0$$

is unique. That $0 \notin \sigma(A_\alpha^*)$ follows similarly by considering the time-flow inverted system $\widetilde{\Xi}_\alpha^{\leftarrow}$ instead.

Claims (ii) and (iii): Since scattering passive boundary nodes are internally well-posed, it follows from, e.g., [29, Lemma 2.6] that equations (2.1) are solvable as has been explained in Section 2. □

Corollary 5.2. *Use the same notation and make the same assumptions as in Theorem 5.1. If $\alpha = 0$, then claims* (i) — (iii) *of Theorem 5.1 hold in the stronger form:* (i') *the triple $\Xi_0 := (G\big|_{\mathcal{Z}_0}, L\big|_{\mathcal{Z}_0}, K\big|_{\mathcal{Z}_0})$ is a scattering conservative boundary node on Hilbert spaces $(\mathcal{U}, \mathcal{X}, \mathcal{U})$ with the solution space $\mathcal{Z}_0 := \tilde{\mathcal{Z}}_0' \times H^1_{\Gamma(1)}(\Omega)$ where*

$$\tilde{\mathcal{Z}}_0' := \{f \in H^1_{\Gamma(1)}(\Omega) : \Delta f \in L^2(\Omega), \frac{\partial f}{\partial \nu}\Big|_{\Gamma(0)} \in L^2(\Gamma(0)), \frac{\partial f}{\partial \nu}\Big|_{\Gamma} = 0\}; \quad (5.20)$$

and (iii') *the energy inequality* (5.18) *holds as an equality.*

Claim (ii) of Theorem 5.1 remains true without change. Thus, the solution ϕ has the same regularity properties as listed right after Theorem 5.1.

Proof. Because the operators G_α and K_α refer to $1/\sqrt{\alpha}$, we cannot simply set $\alpha = 0$ in the proof. This problem could be resolved by making the norm of $\tilde{\mathcal{U}}$ dependent on α which we want to avoid. A direct argument can be given without ever defining $\widetilde{\Xi}_\alpha$. To prove the Green–Lagrange identity

$$2\mathrm{Re}\,\langle \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right], L\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\rangle_{\mathcal{X}} + \|K\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\|_{\mathcal{U}}^2 = \|G\left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]\|_{\mathcal{U}}^2 \text{ for all } \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right] \in \tilde{\mathcal{Z}}_0 \qquad (5.21)$$

for Ξ_0, one simply omits the last term on the right hand side of (5.13) by using the Neumann condition $\frac{\partial z_1}{\partial \nu}\big|_\Gamma = 0$ from (5.20). Then (5.21) follows

from (5.13)—(5.15), leading ultimately to (5.18) with an equality. The remaining parts of claim (i') follow by the argument given in the proof of Theorem 5.1. □

This result generalises the reflecting mirror example in [29, Section 5], and further generalisations are given in Section 6.

6 Conclusions and generalisations

We have given a unified treatment of a 3D wave equation model on tubular structures and the corresponding Webster's horn model in the form it is derived and used in [25, 26]. Both the forward time solvability and the energy inequalities have been treated rigorously, and the necessary but hard-to-find Sobolev space apparatus was presented in App. A. The strictly dissipative case was reduced to the conservative case using auxiliary Theorems 3.1 and 3.2 that have independent interest.

Theorem 5.1 can be extended and generalised significantly using only the techniques presented in this work. Firstly, a dissipation term, analogous with the one appearing in Webster's equation (4.1), can be added to the wave equation part of (5.1) while keeping rest of the model the same:

Corollary 6.1. *Theorem 5.1 remains true if the wave equation* $\phi_{tt} = c^2\Delta\phi$ *in* (5.1) *is replaced by* $\phi_{tt} = c^2\Delta\phi + g(\cdot)\phi_t$ *where* g *is a smooth function satisfying* $g(\mathbf{r}) \leq 0$ *for all* $\mathbf{r} \in \Omega$.

Indeed, this follows by using Theorem 3.2 on the result of Theorem 5.1 in the same way as has been done in Section 4. Even now the resulting negative perturbation H on the original interior operator L in (5.6) satisfies $H \in \mathcal{L}(\mathcal{X})$. The same dissipation term can, of course, be added to Corollary 5.2 (where $\alpha = 0$) as well but then the resulting boundary node is only passive unless $g \equiv 0$.

Theorem 5.1 can be generalised to cover much more complicated geometries $\Omega \subset \mathbb{R}^3$ than tube segments with circular cross-sections. Inspecting the construction of the boundary node Ξ_α and the accompanying Hilbert spaces in Section 5, it becomes clear that much more can be proved at the cost of more complicated notation but nothing more:

Corollary 6.2. *Let* $\Omega \subset \mathbb{R}^3$ *be a bounded Lipschitz domain satisfying standing assumptions* (i) – (iv) *in App. A. Denote the smooth boundary components of* Ω *by* Γ_j *where* $j \in J \subset \mathbb{N}$ *satisfying* $\partial\Omega = \cup_{j\in J}\overline{\Gamma_j}$. *Let* $J = J_1 \cup J_2 \cup J_3$ *where the sets are pairwise disjoint, and at least* J_1 *and* J_3 *are nonempty. Define the open Lipschitz surfaces* $\Gamma(0), \Gamma, \Gamma(1) \subset \partial\Omega$ *through their closures* $\overline{\Gamma(0)} = \cup_{j\in J_1}\overline{\Gamma_j}$, $\overline{\Gamma} = \cup_{j\in J_2}\overline{\Gamma_j}$, *and* $\overline{\Gamma(1)} = \cup_{j\in J_3}\overline{\Gamma_j}$, *respectively. Let* $\alpha = \{\alpha_j\}_{j\in J_2} \subset (-\infty, 0]$ *be a vector of dissipation parameters.*

Then the wave equation model (5.1) *with equations*

$$\alpha_j \frac{\partial \phi}{\partial t}(\mathbf{r}, t) + \frac{\partial \phi}{\partial \nu}(\mathbf{r}, t) = 0 \quad \text{for all } \mathbf{r} \in \Gamma_j,\ t \geq 0,\ \text{and } j \in J_2$$

in place of the fourth equation in (5.1) *defines the boundary node* Ξ_α *and the Hilbert spaces* $\mathcal{X}$, $\mathcal{U}$, *and* $\mathcal{Z}_\alpha$ *in a same way as presented in Section 5. Moreover, Theorem 5.1 and Corollary 5.2 (where* $\alpha_j = 0$ *for all* $j \in J_2$*) hold without change.*

In particular, the set Ω may be an union of a finite number of tubular domains described in Section 1. Even loops are possible and the interior domain dissipation can be added just like in Corollary 6.1. This configuration can be found in the study of the spectral limit behaviour of Neumann–Laplacian on graph-like structures in [15, 35].

Comments on the proof. The argument in Section 5 defines Ξ_α, the Hilbert spaces $\mathcal{X}$, $\mathcal{U}$, and $\mathcal{Z}_\alpha$, and the Green–Lagrange identity by splitting $\partial\Omega$ into three smooth components and patching things up using the results of App. A. The same can be done on any finite number of components since the results of App. A are sufficiently general to allow it. The solvability of the variational problems in the proof of Theorem 5.1 do not depend on the number of such boundary components either. □

There is nothing in Section 5 that would exclude the further generalisation to $\Omega \subset \mathbb{R}^n$ for any $n \geq 2$ if standing assumptions (i) – (iv) in App. A remain true. If $n = 2$ and Ω is a curvilinear polygon (i.e., it is simply connected), the necessary PDE toolkit can be found in [12, Section 1].

Also Theorem 4.1 has extensions but not as many as Theorem 5.1. Firstly, the nonnegative constant α can be replaced by a nonnegative function $\alpha(\cdot) \in C[0, 1]$ since the s-dependency is already present in the operator D in (4.4). Secondly, strong boundary nodes described by Theorem 4.1 can be scaled to different interval lengths and coupled to finite *transmission graphs* as explained in [2] for impedance passive component systems. The full treatment of a simple transmission graph, consisting of three Webster's horn models in Y-configuration, has been given in [2, Theorem 5.2]. More general finite configurations can be treated similarly, and the resulting impedance passive system can be translated to a scattering passive system by the external Cayley transform [30, Section 3], thus producing a generalisation of Theorem 4.1. We note that there is not much point in trying to derive the transmission graph directly from scattering passive systems since the continuity equation (for the pressure) and Kirchhoff's law (for the conservation of flow) at each node is easiest described by impedance notions.

That Theorem 3.2 cannot be used for all possible dissipation terms is seen by considering the wave equation with Kelvin–Voigt structural damping

term

$$\psi_{tt} = c^2\psi_{ss} + \frac{\partial}{\partial s}\left(\beta(s)\frac{\partial}{\partial s}\psi_t\right) \quad \text{where} \quad \beta(s) \geq 0. \tag{6.1}$$

For details of this dissipation model, see, e.g., [24]. To obtain the full dynamical system analogous to the one associated with Webster's equation, the same boundary and initial conditions can be used as in (1.3) for $\beta \in C^\infty[0,1]$ compactly supported $(0,1)$. Thus the operators G_W and K_W do not change. Following Section 4 we use the velocity potential and the pressure as state variables $\left[\begin{smallmatrix}\psi\\ \pi\end{smallmatrix}\right]$. We define the Hilbert spaces $\mathcal{Z}_W$ and $\mathcal{X}_W$ similarly as well as the operators

$$L_W := \begin{bmatrix} 0 & \rho^{-1} \\ \rho c^2 \frac{\partial^2}{\partial s^2} & 0 \end{bmatrix} : \mathcal{Z}_W \to \mathcal{X}_W \text{ and}$$

$$\widetilde{H} := \begin{bmatrix} 0 & 0 \\ 0 & \frac{\partial}{\partial s}\left(\beta(s)\frac{\partial}{\partial s}\right) \end{bmatrix} : \operatorname{dom}\left(\widetilde{H}\right) \subset \mathcal{X}_W \to \mathcal{X}_W$$

where $\operatorname{dom}\left(\widetilde{H}\right) := H^1_{\{1\}}(0,1) \times \{f \in L^2(0,1) : \beta(s)\frac{\partial f}{\partial s} \in H^1(0,1)\}$. The physical energy norm for $\mathcal{X}_W$ is given by (4.5) with $A(s) = \Sigma(s) \equiv 1$ representing a constant diameter straight tube. If the parameter $\beta \equiv 0$, the colligation (G_W, L_W, K_W) is a special case of the conservative system $\Xi_0^{(W)}$ described in Theorem 4.1. Clearly, the domain of $\widetilde{H}$ cannot be further extended without violating the range inclusion in $\mathcal{X}_W$. On the other hand, the inclusion $\mathcal{Z} \subset \operatorname{dom}\left(\widetilde{H}\right)$ required by Theorem 3.2 is not satisfied.

Acknowledgment

The authors have received support from the Finnish Graduate School on Engineering Mechanics, the Norwegian Research Council, and Aalto Starting Grant (grant no. 915587). The authors wish to thank the anonymous referees for many valuable comments.

References

[1] A. Aalto, D. Aalto, J. Malinen and M. Vainio, Modal locking between vocal fold and vocal tract oscillations, arXiv:1211.4788 (2012), submitted.

[2] A. Aalto and J. Malinen, Composition of passive boundary control systems. *Math. Control Relat. Fields* **3** (2013) 1–19.

[3] D. Aalto, A. Huhtala, A. Kivelä, J. Malinen, P. Palo, J. Saunavaara and M. Vainio, How far are vowel formants from computed vocal tract resonances? arXiv:1208.5963 (2012).

[4] J. Cervera, A. J. van der Schaft and A. Baños, Interconnection of port-Hamiltonian systems and composition of Dirac structures. *Automatica* **43** (2007) 212–225.

[5] E. Eisner, Complete solutions of the "Webster" horn equation. *J. Acoust. Soc. Am.* **41** (1967) 1126–1146.

[6] L. Evans and R. Gariepy, *Measure Theory and the Fine Properties of Functions.*, CRC Press (1992).

[7] H. Fattorini, Boundary control systems. *SIAM J. Control* **6** (1968) 349–385.

[8] A. Fetter and J. Walecka, *Theoretical mechanics of particles and continua*, Dover (2003).

[9] F. Flandoli, I. Lasiecka and R. Triggiani, Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernoulli boundary control problems. *Annali Mathematica Pura Applicata* **CLIII** (1988) 307–382.

[10] F. Gesztesy and H. Holden, The damped string problem revisited. *J. Differential Equations* **251** (2011) 1086–1127.

[11] V. Gorbachuk and M. Gorbachuk, *Boundary value problems for operator differential equations*, volume 48 of *Mathematics and its Applications (Soviet Series)*, Kluwer Academic Publishers Group, Dordrecht (1991).

[12] P. Grisvard, *Elliptic problems in non-smooth domains*, Pitman (1985).

[13] A. Hannukainen, T. Lukkari, J. Malinen and P. Palo, Vowel formants from the wave equation. *J. Acoust. Soc. Am. Express Letters* **122** (2007) EL1–EL7.

[14] T. Kato, *Perturbation theory for linear operators*, volume 132 of *Grundlehren der mathematischen Wissenschaften*, Springer Verlag (1980).

[15] P. Kuchment and H. Zeng, Convergence of spectra of mesoscopic systems collapsing onto a graph. *J. Math. Anal. Appl.* **258** (2001) 671–700.

[16] M. Kurula, H. Zwart, A. J. van der Schaft and J. Behrndt, Dirac structures and their composition on Hilbert spaces. *J. Math. Anal. Appl.* **372** (2010) 402–422.

[17] J. Lagnese, Decay of solutions of wave equations in a bounded region with boundary dissipation. *J. Differential Equations* **50** (1983) 163–182.

[18] I. Lasiecka, J. L. Lions and R. Triggiani, Nonhomogenous boundary value problems for second order hyperbolic operations. *J. Math. Pures Appl.* **65** (1986) 149–192.

[19] I. Lasiecka and R. Triggiani, *Control theory for partial differential equations: continuous and approximation theories. II*, volume 75 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press, Cambridge (2000), abstract hyperbolic-like systems over a finite time horizon.

[20] M. Lesser and J. Lewis, Applications of matched asymptotic expansion methods to acoustics. I. The Webster horn equation and the stepped duct. *J. Acoust. Soc. Am.* **51** (1971) 1664–1669.

[21] M. Lesser and J. Lewis, Applications of matched asymptotic expansion methods to acoustics. II. The open-ended duct. *J. Acoust. Soc. Am.* **52** (1972) 1406–1410.

[22] J. L. Lions, Exact controllability, stabilization and perturbations for distributed systems. *SIAM review* **30** (1988) 1–68.

[23] J. L. Lions and E. Magenes, *Non-homogenous boundary value problems and applications II*, volume 182 of *Die Grundlehren der mathematischen Wissenchaften*, Springer Verlag, Berlin (1972).

[24] K. Liu and Z. Liu, Exponential decay of energy of vibrating strings with local viscoelasticity. *Z. Angew. Math. Phys.* **53** (2002) 265–280.

[25] T. Lukkari and J. Malinen, A posteriori error estimates for Webster's equation in wave propagation (2011), manuscript.

[26] T. Lukkari and J. Malinen, Webster's equation with curvature and dissipation, arXiv:1204.4075 (2011), submitted.

[27] J. Malinen, Conservativity of time-flow invertible and boundary control systems, Technical Report A479, Helsinki University of Technology Institute of Mathematics (2004).

[28] J. Malinen, O. Staffans and G. Weiss, When is a linear system conservative? *Quart. Appl. Math.* **64** (2006) 61–91.

[29] J. Malinen and O. Staffans, Conservative boundary control systems. *J. Differential Equations* **231** (2006) 290–312.

[30] J. Malinen and O. Staffans, Impedance passive and conservative boundary control systems. *Complex Anal. Oper. Theory* **2** (2007) 279–300.

[31] A. Nayfeh and D. Telionis, Acoustic propagation in ducts with varying cross sections. *J. Acoust. Soc. Am.* **54** (1973) 1654–1661.

[32] S. Rienstra, Sound transmission in slowly varying circular and annular lined ducts with flow. *J. Fluid Mech.* **380** (1999) 279–296.

[33] S. Rienstra, Webster's horn equation revisited. *SIAM J. Appl. Math.* **65** (2005) 1981–2004.

[34] S. Rienstra and W. Eversman, A numerical comparison between the multiple-scales and finite-element solution for sound propagation in lined flow ducts. *J. Fluid Mech.* **437** (2001) 367–384.

[35] J. Rubinstein and M. Schatzman, Variational problems on multiply connected thin strips I: Basic estimates and convergence of the Laplacian spectrum. *Arch. Ration. Mech. Anal.* **160** (2001) 271–308.

[36] W. Rudin, *Real and Complex Analysis*, McGraw-Hill Book Company, New York, 3 edition (1986).

[37] D. Russell, Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions. *SIAM Review* **20**.

[38] D. Salamon, Infinite dimensional linear systems with unbounded control and observation: a functional analytic approach. *Trans. Amer. Math. Soc.* **300** (1987) 383–431.

[39] D. Salamon, Realization theory in Hilbert spaces. *Math. Systems Theory* **21** (1989) 147–164.

[40] V. Salmon, Generalized plane wave horn theory. *J. Acoust. Soc. Am.* **17** (1946) 199–211.

[41] V. Salmon, A new family of horns. *J. Acoust. Soc. Am* **17** (1946) 212–218.

[42] O. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge (2004).

[43] R. Triggiani, Wave equation on a bounded domain with boundary dissipation: An operator approach. *J. Math. Anal. Appl.* **137** (1989) 438–461.

[44] M. Tucsnak and G. Weiss, How to get a conservative well-posed linear system out of thin air. II. Controllability and stability. *SIAM J. Control Optim.* **42** (2003) 907–935.

[45] M. Tucsnak and G. Weiss, *Observation and control for operator semigroups*, Birkhäuser Verlag, Basel (2009).

[46] J. Villegas, *A Port-Hamiltonian Approach to Distributed Parameter Systems*, Ph.D. thesis, University of Twente (2007).

[47] A. Webster, Acoustic impedance, and the theory of horns and of the phonograph. *Proc. Natl. Acad. Sci. USA* **5** (1919) 275–282.

[48] G. Weiss and M. Tucsnak, How to get a conservative well-posed linear system out of thin air. I. Well-posedness and energy balance. *ESAIM Control Optim. Calc. Var.* **9** (2003) 247–274.

A Sobolev spaces and Green's identity

We prove a sufficiently general form of Green's identity that holds in a tubular domain Ω (that has a Lipschitz boundary) with minimal assumptions on any functions involved. We make the following standing assumptions on Ω:

(i) Ω is a bounded Lipschitz domain so that Ω locally on one side of is boundary $\partial\Omega$;

(ii) there is a finite number of smooth, open, connected, and disjoint $(n-1)$-dimensional surfaces Γ_j with the following property: the boundary $\partial\Omega$ is a union of all Γ_j's and parts of their common boundaries $\overline{\Gamma}_j \cap \overline{\Gamma}_k$ for $j \neq k$;

(iii) $\mathcal{H}^{n-2}(\overline{\Gamma}_j \cap \overline{\Gamma}_k) < \infty$ for all $j \neq k$ where $\mathcal{H}^m(M)$ is the m-dimensional Hausdorff measure for $1 \leq m \leq n$ of $M \subset \mathbb{R}^n$; and

(iv) for each j, there is a C^∞ vector field ν_j defined in a neighbourhood of $\overline{\Omega}$ such that $\nu_j(\mathbf{r})$ is the exterior unit normal to Γ_j at $\mathbf{r} \in \Gamma_j$.

That $\Gamma_j \subset \mathbb{R}^n$ is an open, bounded, and smooth $(n-1)$-dimensional surface means plainly the following: there is an open and bounded $\tilde{\Gamma}_j \subset \mathbb{R}^{n-1}$ and a C^∞-diffeomorphism ϕ_j from $\tilde{\Gamma}_j$ onto Γ_j. The pair $(\phi_j, \tilde{\Gamma}_j)$ is a global coordinate representation of Γ_j.

The boundary conditions in Section 5 involve Dirichlet conditions on some parts of the boundary $\partial\Omega$ and Neumann type conditions on other parts of the same *connected component* of $\partial\Omega$. All this is in contrast with the inconvenient technical assumption on $\partial\Omega$ in, e.g., [17, 29, 43] that must be avoided in the verification of the Green–Lagrange identity in Section 5 and elsewhere. We need a version of Green's identity suitable for this situation. This is in Theorem A.3 below. The key fact ensuring the validity of this identity is that the interfaces where we switch between different boundary conditions are so small that Sobolev functions do not see them. That this is the case is a consequence of the assumption (iii) above, and it is expressed rigorously in the following auxiliary result.

Lemma A.1. *Let Ω be a bounded domain with a Lipschitz boundary, and let $E \subset \mathbb{R}^n$ be a compact set of zero capacity; i.e.,*

$$C(E) := \inf_{u \in S(E)} \int_{\mathbb{R}^n} \left(|u|^2 + |\nabla u|^2 \right) \, \mathrm{d}V = 0 \tag{A.1}$$

where

$$S(E) := \{u \in C^\infty(\mathbb{R}^n) : 0 \le u \le 1 \text{ in } \mathbb{R}^n \text{ and } u = 1 \text{ in } N, \text{ where } N \text{ is open and } E \subset N\}.$$

Then

(i) *the set $\mathcal{D}_E(\mathbb{R}^n)$ is dense in $H^1(\mathbb{R}^n)$ where*

$$\mathcal{D}_E(\mathbb{R}^n) := \{u \in \mathcal{D}(\mathbb{R}^n) : u \text{ vanishes in an open neighbourhood of } E\}; \text{and} \tag{A.2}$$

(ii) *the set*

$$\mathcal{D}_E(\overline{\Omega}) := \{u\big|_\Omega : u \in \mathcal{D}_E(\mathbb{R}^n)\}$$

is dense in $H^1(\Omega)$.

Proof. Claim (i): Let $u \in H^1(\mathbb{R}^n)$ and $\varepsilon > 0$. Then by [12, Theorem 1.4.2.1] there is $v \in \mathcal{D}(\mathbb{R}^n)$ such that $\|u - v\|_{H^1(\mathbb{R}^n)} < \varepsilon/2$.

By the vanishing capacity assumption (A.1), there is a sequence $\{\varphi_j\}_{j=1,2,\ldots} \subset C^\infty(\mathbb{R}^n)$ such that $\varphi_j\big|_{N_j} = 1$ for some neighbourhoods N_j of E, and also

$$\lim_{j \to \infty} \int_{\mathbb{R}^n} \left(|\varphi_j|^2 + |\nabla \varphi_j|^2 \right) \, \mathrm{d}V = 0. \tag{A.3}$$

Defining $v_j(\mathbf{r}) := v(\mathbf{r})(1 - \varphi_j(\mathbf{r}))$ we see that each of these functions satisfies $v_j \in \mathcal{D}_E(\mathbb{R}^n)$. It remains to prove that $\|v_j - v\|_{H^1(\mathbb{R}^n)} < \varepsilon/2$ for all j large enough, since then

$$\|v_j - u\|_{H^1(\mathbb{R}^n)} \le \|v_j - v\|_{H^1(\mathbb{R}^n)} + \|u - v\|_{H^1(\mathbb{R}^n)} < \varepsilon.$$

By possibly replacing $\{\varphi_j\}_{j=1,2,\ldots}$ by its subsequence, we may assume that $\varphi_j \to 0$ pointwise almost everywhere; see [36, Theorem 3.12]. Because $|v_j(\mathbf{r})| \le |v(\mathbf{r})|$ for all $\mathbf{r} \in \mathbb{R}^n$ and $j = 1, 2, \ldots$, we have $v_j \to v$ in $L^2(\mathbb{R}^n)$ by the Lebesgue dominated convergence theorem. For the gradients, we note that $\nabla(v_j - v) = -\varphi_j \nabla v - v \nabla \varphi_j$. Thus $|\nabla(v_j - v)| \to 0$ in $L^2(\mathbb{R}^n)$, since both φ_j and $|\nabla \varphi_j|$ tend to zero in $L^2(\mathbb{R}^n)$ by (A.3).

Claim (ii): Let $u \in H^1(\Omega)$ and take $\varepsilon > 0$. Since Ω has a Lipschitz boundary, there is an extension operator $T \in \mathcal{L}(H^1(\Omega); H^1(\mathbb{R}^n))$ such that $(Tu)\big|_\Omega = u$; see [12, Theorem 1.4.3.1]. By claim (i), there is a function $v \in \mathcal{D}_E(\mathbb{R}^n)$ such that

$$\|u - v\big|_\Omega\|_{H^1(\Omega)} \le \|Tu - v\|_{H^1(\mathbb{R}^n)} < \varepsilon$$

which completes the proof. □

Let us review the Sobolev spaces and the boundary trace mappings on Ω and $\partial\Omega$ when the standing assumptions (i) – (iv) above hold. The boundary Sobolev spaces $H^s(\partial\Omega)$ and $H^s(\Gamma_j)$ for $s \in [-1,1]$ are defined as in [12, Definitions 1.2.1.1 and 1.3.3.2]. The zero extension Sobolev spaces on Γ_j are defined by

$$\tilde{H}^s(\Gamma_j) := \{u \in H^s(\Gamma_j) : \tilde{u} \in H^s(\partial\Omega)\}$$

for $s \in (0,1]$ where

$$\tilde{u}(\mathbf{r}) := \begin{cases} u(\mathbf{r}) & \text{if } \mathbf{r} \in \Gamma_j \\ 0 & \text{if } \mathbf{r} \in \partial\Omega \setminus \Gamma_j. \end{cases} \tag{A.4}$$

We use the Hilbert space norms $\|u\|_{\tilde{H}^s(\Gamma_j)} := \|\tilde{u}\|_{H^s(\partial\Omega)}$. The space $\tilde{H}^s(\Gamma_j)$ is closed in this norm since restriction to Γ_j from $\partial\Omega$ is a bounded operator from $H^s(\partial\Omega)$ to $H^s(\Gamma_j)$ for $0 \le s \le 1$. This boundedness follows trivially by restriction using the Gagliardo seminorm, see [12, Eq. (1,3,3,3) on p. 20]. Then $H^s(\partial\Omega) \subset L^2(\partial\Omega)$ and $\tilde{H}^s(\Gamma_j) \subset H^s(\Gamma_j) \subset L^2(\Gamma_j)$ with bounded inclusions.

The *Dirichlet trace operator* γ is first defined for functions $f \in \mathcal{D}(\overline{\Omega})$ simply by restriction $\gamma f := f\big|_{\partial\Omega}$. This operator has a unique extension to a bounded operator $\gamma \in \mathcal{L}(H^1(\Omega); H^{1/2}(\partial\Omega))$; see [12, Theorem 1.5.1.3] and Lemma A.1. All this holds for any Lipschitz domain Ω.

We define the *Neumann trace operator* separately on each surface Γ_j using the vector fields ν_j. Such an operator $\gamma_j \frac{\partial}{\partial \nu_j}$ is first defined on $\mathcal{D}(\overline{\Omega})$ (with values in $L^2(\partial\Omega)$) by setting $\left(\gamma_j \frac{\partial}{\partial \nu_j} f\right)(\mathbf{r}) := \nu_j(\mathbf{r}) \cdot \nabla f(\mathbf{r})$ for all $\mathbf{r} \in \Gamma_j$; here $\gamma_j f := f\big|_{\Gamma_j}$ and $\frac{\partial}{\partial \nu_j} := \nu_j \cdot \nabla$. It is easy to see that $\frac{\partial f}{\partial \nu_j} \in H^1(\Omega)$ and hence $\gamma_j \frac{\partial}{\partial \nu_j}$ has an extension to an operator in $\mathcal{L}(H^2(\Omega); H^{1/2}(\Gamma_j))$ by [12, Theorem 1.5.1.3]. We then define the full Neumann trace operator $\gamma \frac{\partial}{\partial \nu}$ on $\cup_j \Gamma_j$ by

$$\gamma \frac{\partial f}{\partial \nu}(\mathbf{r}) := \gamma_j \frac{\partial f}{\partial \nu_j}(\mathbf{r}) \quad \text{for all } f \in H^2(\Omega) \quad \text{and (almost) all} \quad \mathbf{r} \in \Gamma_j.$$

Note that the function $\gamma \frac{\partial f}{\partial \nu}$ is not defined at all on the exceptional set of capacity zero

$$E := \cup_{j \ne k} (\overline{\Gamma}_j \cap \overline{\Gamma}_k) \tag{A.5}$$

of the non-smooth part of $\partial\Omega$. That $C(E) = 0$ follows from the standing assumption (iii) by [6, Theorem 3, p. 154].

We need to extend each $\gamma_j \frac{\partial}{\partial \nu_j}$ to the Hilbert space

$$E(\Delta; L^2(\Omega)) := \{f \in H^1(\Omega) : \Delta f \in L^2(\Omega)\}$$

that is equipped with the norm defined by $\|f\|^2_{E(\Delta;L^2(\Omega))} = \|f\|^2_{H^1(\Omega)} + \|\Delta f\|^2_{L^2(\Omega)}$.

We use an appropriate L^2 space as the pivot space for Sobolev spaces and their duals.

Proposition A.2. *Let the domain $\Omega \subset \mathbb{R}^n$ satisfy the standing assumptions* (i) – (iv).

(i) *Then each Neumann trace operator $\gamma_j \frac{\partial}{\partial \nu_j}$ (originally defined on $\mathcal{D}(\overline{\Omega})$) has a unique extension (also denoted by $\gamma_j \frac{\partial}{\partial \nu_j}$) that is bounded from $E(\Delta; L^2(\Omega))$ into the dual space of $\tilde{H}^{1/2}(\Gamma_j)$.*

(ii) *We have*

$$\int_\Omega (\Delta u)\, v \,\mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}V = \sum_j \left\langle \gamma_j \frac{\partial u}{\partial \nu}, \gamma_j v \right\rangle_{[\tilde{H}^{1/2}(\Gamma_j)]^d, \tilde{H}^{1/2}(\Gamma_j)}$$

for all $u \in E(\Delta; L^2(\Omega))$ and $v \in H^1(\Omega)$ such that $\gamma_j v \in \tilde{H}^{1/2}(\Gamma_j)$ for all j.

Proof. The classical Green's identity for $u \in \mathcal{D}(\overline{\Omega})$ and $v \in \mathcal{D}_E(\overline{\Omega})$ is

$$\int_\Omega (\Delta u)\, v \,\mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}V = \sum_j \int_{\Gamma_j} \gamma_j \frac{\partial u}{\partial \nu_j} \gamma_j v \,\mathrm{d}A, \tag{A.6}$$

where E is the exceptional set in (A.5). Indeed, since v vanishes near the interfaces $\overline{\Gamma}_j \cap \overline{\Gamma}_k$ for $j \neq k$, we may initially apply Green's identity just like (A.6) but over a subdomain of Ω that has been obtained from Ω by rounding slightly at all $\partial\Gamma_j$'s but preserving essentially all of $\partial\Omega$. Then we get (A.6) by rewriting the result as integrals over the original Ω and the original boundary pieces Γ_j, noting that on additional points the integrands vanish because $v \in \mathcal{D}_E(\overline{\Omega})$.

It follows from (A.6) that we have for $u \in \mathcal{D}(\overline{\Omega})$ and $v \in \mathcal{D}_E(\overline{\Omega})$ the estimate

$$\left| \sum_j \left\langle \gamma_j \frac{\partial u}{\partial \nu_j}, \gamma_j v \right\rangle_{L^2(\Gamma_j)} \right| \leq \|u\|_{E(\Delta;L^2(\Omega))} \cdot 4 \|v\|_{H^1(\Omega)}. \tag{A.7}$$

Because $\mathcal{D}_E(\overline{\Omega})$ is dense in $H^1(\Omega)$ by Lemma A.1 and $\gamma \in \mathcal{L}(H^1(\Omega); H^{1/2}(\partial\Omega))$ by the trace theorem [12, Theorem 1.5.1.3], we conclude that (A.7) holds for all $u \in \mathcal{D}(\overline{\Omega})$ and $v \in H^1(\Omega)$.

Fix now j and $g \in \tilde{H}^{1/2}(\Gamma_j)$, and define $\tilde{g} \in H^{1/2}(\partial\Omega)$ by (A.4). Because the Dirichlet trace $\gamma : H^1(\Omega) \to H^{1/2}(\partial\Omega)$ is bounded and surjective, it has a continuous right inverse $P \in \mathcal{L}(H^{1/2}(\partial\Omega); H^1(\Omega))$, see [12, Theorem

1.5.1.3]. Thus there exists $v \in H^1(\Omega)$ such that $\gamma_j v = \tilde{g}\big|_{\Gamma_j} = g$ and $\gamma_k v = 0$ for $k \neq j$; we may choose $v = P\tilde{g}$. From this, we have the estimate $4\|v\|_{H^1(\Omega)} \leq K\|\tilde{g}\|_{H^{1/2}(\partial\Omega)} = K\|g\|_{\tilde{H}^{1/2}(\Gamma_j)}$.

It follows from all this and (A.7) that we have

$$|\Phi_g(u)| \leq K\|u\|_{E(\Delta;L^2(\Omega))} \cdot \|g\|_{\tilde{H}^{1/2}(\Gamma_j)} \tag{A.8}$$

for all $g \in \tilde{H}^{1/2}(\Gamma_j)$ where $\Phi_g(u) := \langle \gamma \frac{\partial u}{\partial \nu}, \tilde{g} \rangle_{L^2(\partial\Omega)} = \left\langle \gamma_j \frac{\partial u}{\partial \nu_j}, g \right\rangle_{L^2(\Gamma_j)}$ for $u \in \mathcal{D}(\overline{\Omega})$. Since $\mathcal{D}(\overline{\Omega})$ is dense in $E(\Delta; L^2(\Omega))$ by [12, Lemma 1.5.3.9], we may extend Φ_g, $g \in \tilde{H}^{1/2}(\Gamma_j)$, by continuity to a continuous linear functional on $E(\Delta; L^2(\Omega))$ satisfying estimate (A.8), too.

For each fixed $u \in E(\Delta; L^2(\Omega))$, the mapping $g \mapsto \Phi_g(u)$ is a continuous linear functional on $\tilde{H}^{1/2}(\Gamma_j)$ by (A.8). Hence, there is a representing vector – denoted by $\gamma_j \frac{\partial u}{\partial \nu_j}$ – in the dual space $[\tilde{H}^{1/2}(\Gamma_j)]^d$ such that $\Phi_g(u) = \left\langle \gamma_j \frac{\partial u}{\partial \nu_j}, g \right\rangle_{[\tilde{H}^{1/2}(\Gamma_j)]^d, \tilde{H}^{1/2}(\Gamma_j)}$. This proves claim (i). Claim (ii) follows by a density argument using claim (i) and (A.8). □

Theorem A.3 (Green's identity)**.** *Let the domain $\Omega \subset \mathbb{R}^n$ satisfy the standing assumptions* (i) – (iv) *above. Assume that $u \in H^1(\Omega)$ is such that $\Delta u \in L^2(\Omega)$ and satisfies $\frac{\partial u}{\partial \nu} \in L^2(\cup_{j=1}^k \Gamma_j)$ for some $1 \leq k \leq n$. Then the Green's identity*

$$\int_\Omega (\Delta u)\, v \,\mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}V = \sum_{j=1}^{k} \int_{\Gamma_j} \frac{\partial u}{\partial \nu} v \,\mathrm{d}A + \sum_{j=k+1}^{n} \left\langle \gamma_j \frac{\partial u}{\partial \nu_j}, \gamma_j v \right\rangle_{[\tilde{H}^{1/2}(\Gamma_j)]^d, \tilde{H}^{1/2}(\Gamma_j)} \tag{A.9}$$

holds for functions $v \in H^1(\Omega)$ such that $\gamma_j v \in \tilde{H}^{1/2}(\Gamma_j)$ for $k+1 \leq j \leq n$.

For $n = 2$, this is a generalisation of [12, Theorem 1.5.3.11]. See also [12, discussion on p. 62] for domains with $C^{1,1}$-boundaries. The assumption $\frac{\partial u}{\partial \nu} \in L^2(\cup_{j=1}^k \Gamma_j)$ simply means that $\gamma_j \frac{\partial u}{\partial \nu_j} \in L^2(\Gamma_j)$ for all $j = 1, 2, \ldots, k$ where $\gamma_j \frac{\partial u}{\partial \nu_j}$ is understood as an element of $[\tilde{H}^{1/2}(\Gamma_j)]^d$ which space includes $L^2(\Gamma_j)$; see Proposition A.2.

Proof. As explained above, we have $\gamma_j v, \gamma_j \frac{\partial u}{\partial \nu_j} \in L^2(\Gamma_j)$ for all $j = 1, \ldots, k$. Then (A.9) follows from claim (ii) of Proposition A.2 under the additional assumption that $\gamma_j v \in \tilde{H}^{1/2}(\Gamma_j)$ for all j. The functions in $\mathcal{D}_E(\overline{\Omega})$ clearly satisfy this additional assumption, and they are dense in $H^1(\Omega)$. This proves the claim. □

An alternative to the above piecewise construction is to start with the global Neumann trace $\gamma \frac{\partial}{\partial \nu} u$ defined for $u \in E(\Delta; L^2(\Omega))$ with values in $H^{-1/2}(\partial\Omega)$, see, e.g., [45, Theorem 13.6.9]. The global Neumann trace $\gamma \frac{\partial}{\partial \nu} u$

can be restricted to the spaces $\tilde{H}^{1/2}(\Gamma_j)$, and claim (ii) of Proposition A.2 follows from a global Green's identity in a general Lipschitz domain. However, one still needs Lemma A.1 to prove Theorem A.3.

It remains to prove the Poincaré inequality that is used to show that the expression (5.7) is a valid Hilbert space norm for the state space. Let Γ_j be one of the boundary components of $\partial\Omega$ as described above. By the standing assumptions (i) and (ii) given in the beginning of this appendix, the set Γ_j has a finite, positive area $A_j = \int_{\Gamma_j} \mathrm{d}A$. Thus, we can define the mean value operator $M_j : H^1(\Omega) \to \mathbb{C}$ on Γ_j by

$$M_j u = \frac{1}{A_j} \int_{\Gamma_j} \gamma_j u \, \mathrm{d}A,$$

It is clear that M_j is a bounded linear functional on $H^1(\Omega)$, and we may regard it as an element of $\mathcal{L}(H^1(\Omega))$ safistying $M_j^2 = M_j$ by considering $M_j u$ as a constant function on Ω.

Theorem A.4 (Poincaré inequality)**.** *Let the domain $\Omega \subset \mathbb{R}^n$ satisfy the standing assumptions* (i) – (iv) *above, and let Γ_j be one of the boundary components of $\partial\Omega$. There is a constant $C < \infty$ such that*

$$\|u - M_j u\|_{L^2(\Omega)} \leq C \|\nabla u\|_{L^2(\Omega)} \tag{A.10}$$

for all $u \in H^1(\Omega)$. Thus, we have $\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}$ for $u \in H^1(\Omega) \cap \ker(\gamma_j)$.

Proof. The argument is a standard argument by contradiction using the Rellich–Kondrachov compactness theorem, see e.g. [6, Theorem 1, p. 144]). For a contradiction against (A.10), assume that there exist functions $u_k \in H^1(\Omega)$ such that there is the strict inequality

$$\|u_k - M_j u_k\|_{L^2(\Omega)} > k \|\nabla u_k\|_{L^2(\Omega)} \quad \text{for} \quad k = 1, 2, \ldots.$$

None of the functions u_k are constant functions since for such functions (A.10) holds for any $C \geq 0$. So, we can define the functions

$$v_k := \frac{u_k - M_j u_k}{\|u_k - M_j u_k\|_{L^2(\Omega)}}$$

satisfying for all k the normalisation $\|v_k\|_{L^2(\Omega)} = 1$ and also $M_j v_k = 0$ by using $M_j^2 = M_j$. Since

$$\|\nabla v_k\|^2 = \frac{\|\nabla u_k\|^2_{L^2(\Omega)}}{\|u_k - M_j u_k\|^2_{L^2(\Omega)}} < \frac{1}{k^2}$$

by the counter assumption, we get

$$\|v_k\|^2_{H^1(\Omega)} = \|v_k\|^2_{L^2(\Omega)} + \|\nabla v_k\|^2_{L^2(\Omega)} \leq 1 + \frac{1}{k^2} \leq 2.$$

Since the embedding $H^1(\Omega) \subset L^2(\Omega)$ is compact (by the boundedness of Ω and the Rellich–Kondrachov compactness theorem, see e.g. [6, Theorem 1, p. 144]), we have a function v such that $v_k \to v$ in $L^2(\Omega)$ by possibly replacing $\{v_k\}$ by its subsequence. Moreover, $\|v\|_{L^2(\Omega)} = 1$ since $\|v_k\|_{L^2(\Omega)} = 1$ for all k.

Since $\|\nabla v_k\|_{L^2(\Omega)} \leq 1/k$, we see that $v_k \to v$ in $H^1(\Omega)$ and hence $\nabla v = 0$. Thus v is a constant function. Because $M_j v = \lim_{k\to\infty} M_j v_k = 0$, we conclude that $v = 0$ which contradicts the fact that $\|v\|_{L^2(\Omega)} = 1$. This proves (A.10), and the Poincaré equality follows trivially from this. $\square$

Aalto-DD 160/2014

ISBN 978-952-60-5909-9 (printed)
ISBN 978-952-60-5910-5 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Mathematics and Systems Analysis
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS