University of California
Santa Barbara

Topics in Stochastic Stability, Optimal Control and Estimation Theory

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Mechanical Engineering

by

Maurice G. Filo

Committee in charge:

Professor Bassam Bamieh, Chair
Professor Francesco Bullo
Professor Igor Mezic
Professor Joao Hespanha

June 2018

The Dissertation of Maurice G. Filo is approved.

Professor Francesco Bullo

Professor Igor Mezic

Professor Joao Hespanha

Professor Bassam Bamieh, Committee Chair

June 2018

Topics in Stochastic Stability, Optimal Control and Estimation Theory

by

Maurice G. Filo

To Jennifer and my parents...

Acknowledgements

First and foremost, I would like to express my deep gratitude to my graduate advisor, Professor Bassam Bamieh. His unique style in mentoring and doing research captured me since day one when he was visiting the American University of Beirut. Every topic he taught seemed more interesting. I was captivated by his outstanding ability of making any concept, no matter how convoluted, as trivial as it can get. I am particularly grateful to his valuable time and long meetings where he shared his broad knowledge with me and made me grow academically and mentally. His involvement, flexibility and utmost understanding made my experience as a PhD student smooth and memorable. Bassam, I could never ask for a better advisor. Thank you for everything!

I would like to thank all the outstanding professors at UCSB for making me love what I do. I was very lucky to be around such distinguished professors in both the mechanical and electrical engineering departments. I am especially grateful and proud to be a fellow in the Center for Control, Dynamical Systems and Computation (CCDC) surrounded by unparalleled faculty, postdocs, students and visitors. Thanks to CCDC, I was always up-to-date with the ongoing research in my field of study.

I would particularly like to thank the committee members Professors Igor Mezic, Francesco Bullo and Joao Hespanha for their help during my PhD program and for taking the time to read my long dissertation! I also owe thanks to Professor Jean-Pierre Fouque for his valuable help in stochastic differential equations.

I would like to thank my parents for their love and support to pursue my dreams. I owe it all to you! Last, but absolutely not least, there are no words that are capable of expressing my gratitude to my wife Jennifer. Living far from you for the last five years was extremely difficult for both of us. But your unconditional love, understanding and support made it bearable and motivated me to do my best (and a little more).

Curriculum Vitæ

Maurice G. Filo

Education

2018 PhD in Mechanical Engineering. University of California, Santa Barbara
GPA: 4.0/4.0

2017 MS in Mechanical Engineering. University of California, Santa Barbara.
GPA: 4.0/4.0

2013 MS in Electrical & Computer Engineering. American University of Beirut.
GPA: 4.0/4.0

2010 Diploma in Electrical Engineering. Lebanese University. Graduated with First Class Honors.

Selected Publications

Journals

1. Maurice Filo & Bassam Bamieh, *An input-output approach to structured stochastic uncertainty in continuous time*, submitted to IEEE Transactions on Automatic Control, 2018.
2. Bassam Bamieh & Maurice Filo, *An input-output approach to structured stochastic uncertainty*, submitted to IEEE Transactions on Automatic Control, 2018.
3. Maurice Filo & Bassam Bamieh, *Stochastic models for cochlear instabilities*, submitted to Journal of Acoustical Society of America, 2018.
4. Maurice Filo, Fadi Karameh & Mariette Awad, *Order reduction and efficient implementation of nonlinear nonlocal cochlear response models*, Journal of Biological Cybernetics, 2016, 110(6), 435-454.
5. Nadine Hajj, Maurice Filo & Mariette Awad, *Automated composer recognition for multi-voice piano compositions using rhythmic features, n-grams and modified cortical algorithms*, 2017, Journal of Complex and Intelligent Systems, 1-11.

Refereed Proceedings

1. Maurice Filo & Bassam Bamieh, *A block diagram approach to stochastic calculus with application to multiplicative uncertainty analysis*, submitted to IEEE Conference on Decision and Control, 2018.
2. Maurice Filo & Bassam Bamieh, *A function space approach to gradient descent in optimal control*, to appear in Proceedings of the 2018 American Control Conference.
3. Maurice Filo & Bassam Bamieh, *Investigating cochlear instabilities using structured stochastic uncertainty*, in Proceedings of the 56th IEEE Conference on Decision and Control, pp. 1634-1640, 2017.

4. Maurice Filo & Bassam Bamieh, *Sensor motion for optimal estimation in distributed dynamic environments* in Proceedinngs of the 2017 American Control Conference, pp. 3263-3269, 2017.

Fellowship and Awards

Featured as the Researcher of the Term, Center for Control, Dynamical Systems & Computations, Spring 2016, https://www.ccdc.ucsb.edu/content/maurice-filo.

Best Teacher Assistant Award at the Mechanical Engineering Department at UCSB, Spring 2015.

Center for Control, Dynamical Systems and Computations (CCDC) Fellowship, Fall 2013.

Institute for Energy Efficiency Holbrook Foundation Fellowship, Fall 2013.

Abstract

Topics in Stochastic Stability, Optimal Control and Estimation Theory

by

Maurice G. Filo

This dissertation consists of four parts that revolve around structured stochastic uncertainty and optimal control/estimation theory.

In the first part, we consider the continuous-time setting of linear time-invariant (LTI) systems in feedback with multiplicative stochastic uncertainties. The objective is to characterize the conditions of Mean-Square Stability (MSS) using a purely input-output approach. This approach leads to uncovering new tools such as stochastic block diagrams. Various stochastic interpretations are considered, such as Itō and Stratonovich, and block diagram conversion schemes between different interpretations are devised. The MSS conditions are given in terms of the spectral radius of a matrix operator that takes different forms when different stochastic interpretations are considered.

The second part applies the developed theory to analyze the mean-square stability and performance of stochastic cochlear models. The analysis is carried out for a generalized class of biomechanical models of the cochlea, that is formulated as a stochastic spatially distributed system, by allowing stochastic spatio-temporal perturbations within the cochlear amplifier. The simulation-free analysis explains the underlying mechanisms that give rise to cochlear instabilities such as spontaneous otoacoustic emissions and/or tinnitus. Furthermore, nonlinear stochastic simulations are carried out to validate the predictions of the theoretical analysis.

The third part revisits the development of numerical methods to solve optimal control problems using a function-space approach. This approach has the advantage of unify-

ing the framework upon which the various (existing) numerical methods are based on. In fact, this approach motivates the definition of various system and projection operators that make the derivations conceptually transparent. Furthermore, the function-space approach builds useful geometric intuitions that inspire the development of new projection-based methods.

In the last part, we propose a methodology of optimal path design for sensors through a distributed environment. We consider time-limited scenarios where the sensors can only make a small number of measurements, but where some portion of a physics-based model is available for the field of interest (such as temperature). We consider both point-wise and tomographic sensors. The main idea is to recast the sensor path planning problem as a deterministic optimal control problem to minimize metrics related to the optimal estimation error covariance.

Contents

Chapter 1

Introduction

The dissertation spans a broad spectrum of topics under stochastic dynamics and optimal control/estimation theory. Although the first two parts are connected: the second is an application for the theory developed in the first, they can yet be read separately. In fact, each part has its own introduction and appendices and can be read with minimal reference to one another.

In the first part, we consider the continuous-time setting of linear time-invariant (LTI) systems in feedback with multiplicative stochastic uncertainties. The objective of this part of the dissertation is to characterize the conditions of Mean-Square Stability (MSS) using a purely input-output approach, i.e. without having to resort to state space realizations. This has the advantage of encompassing a wider class of models (such as infinite dimensional systems and systems with delays). The input-output approach leads to uncovering new tools such as stochastic block diagrams that have an intimate connection with the more general Stochastic Integral Equations (SIE), rather than Stochastic Differential Equations (SDE). Various stochastic interpretations are considered, such as Itō and Stratonovich, and block diagram conversion schemes between different interpretations are devised. The MSS conditions are given in terms of the spectral radius of a

matrix operator that takes different forms when different stochastic interpretations are considered.

Much effort has been made to make this exposition self-contained. This part is organized to first describe the problem statement and then immediately state the results. Thus, the reader can get the flavor of this work without having to dig into the technicalities. After stating the results, we provide the proofs and underlying analysis from which the results are based on.

The second part applies the developed theory to track the mean-square stability and performance of stochastic cochlear models. Instabilities that emerge due to random perturbations at the level of the cochlear amplifier are investigated. These perturbations are allowed to be time-and-location-varying to emulate the stochastic nature of the possible sources of biological disturbances. Various scenarios are considered to examine the effects of different types of disturbances on the instabilities. Particularly, it is shown that different types of disturbances (e.g. correlated, uncorrelated, localized) induce spontaneous vibrations at different locations on the cochlear partition. This leads to Spontaneous Otoacoustic Emissions (SOAEs) with different frequencies in the absence of any stimulus. Furthermore, it is believed that if these spontaneous vibrations are intense enough, they may be perceived as tinnitus.

The stability analysis is carried out on a generalized class of biomechanical models of the cochlea that is formulated in continuous space-time by defining relevant spatial operators. Furthermore, the analysis is simulation-free and is performed by borrowing notions from stochastic and robust control theory that is developed in the first part of the dissertation. Finally, nonlinear stochastic simulations are carried out to validate the predictions of the theoretical analysis. The simulations show that the nonlinearities saturate the spontaneous stochastic vibrations of the basilar membrane, but do not significantly deform its vibration modes (and thus the emitted frequencies).

The third part revisits the development of numerical methods to solve optimal control problems using a function-space approach. This approach has the advantage of unifying the framework upon which the various (existing) numerical methods are based on. In fact, this approach motivates the definition of various system and projection operators that make the derivations conceptually transparent. It also facilitates the classification of the various methods and uncovers the connections between them. Furthermore, the function-space approach builds useful geometric intuitions that inspire the development of new projection-based methods.

Particularly, this part develops a preconditioned constrained-gradient descent (PCGD) method which is based on projected gradient descent in infinite dimensional optimization problems. The key is to exploit the special structure of optimal control problems to precondition the state-control space, and thus achieve a higher convergence rate than the well known gradient descent method.

Finally, **in the last part**, we propose a methodology of optimal path design for sensors through a distributed environment represented by a field quantity. We consider time-limited scenarios where the sensors can only make a small number of measurements, but where some portion of a physics-based model is available for the field of interest such as fluid flows, temperatures or concentrations. Thus the highly underdetermined inverse problem can be augmented with dynamical models. We consider stochastic settings where the fields are subject to partially unknown disturbances and boundary conditions. The main idea is to recast the sensor path planning problem as a deterministic optimal control problem to minimize metrics related to the optimal estimation error covariance, thus converting the stochastic estimation problem to a deterministic operator-valued problem. In the specific case of linear field dynamics, the signal to be designed is the sensors paths which are inputs to the optimal error covariance Riccati equation, resulting in a deterministic, nonlinear, optimal control problem where the trace of the error covariance

operator is to be minimized. For sensing modalities, we consider point-wise sampling as well as the more unusual case of line-integral measurements. The latter is motivated by tomographic reconstruction scenarios with a small number of sensors.

Part I

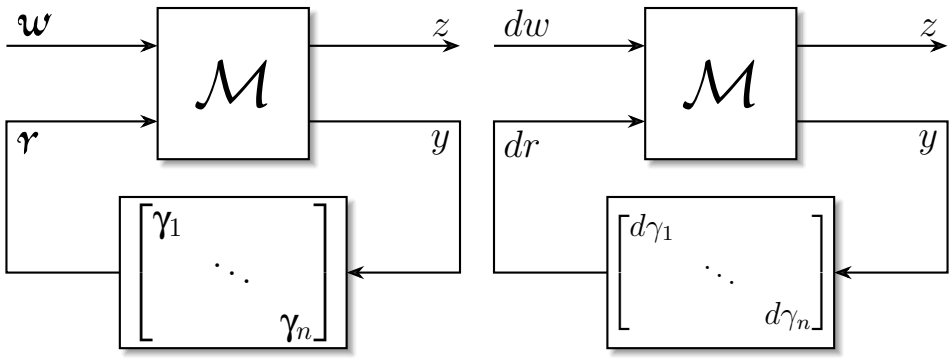
Structured Stochastic Uncertainty

Chapter 2

An Input-Output Approach to Structured Stochastic Uncertainty in Continuous Time

Linear Time-Invariant (LTI) systems with stochastic disturbances is a powerful modeling technique that is used to analyze and control a large class of physical systems. While additive disturbances are most commonly used to model process and measurement noise in a system, multiplicative disturbances are often necessary to model stochastic uncertainties in the system parameters (such as coefficients in dynamical equations). LTI systems driven by additive stochastic processes are more common in the literature; whereas simultaneous additive and multiplicative disturbances are relatively less addressed. This chapter develops a methodology to study the mean-square stability of continuous-time systems with both additive and multiplicative disturbances, while adopting different stochastic interpretations (such as Itō and Stratonovich).

The general setting we consider in this chpater is the continuous-time analog of that presented in [3] and is depicted in Figure 2.1(a). An LTI system is in feedback with



(a) White Process Representation (b) Wiener Process Representation

Figure 2.1: The general continuous-time setting of linear systems with both additive and multiplicative stochastic disturbances. Both block diagrams describe the same setting, given in (2.1) and (2.3), using white processes (to the left) and Wiener processes (to the right), respectively. The LTI system $\mathcal{M}$ is in feedback with multiplicative stochastic gains represented here as a diagonal matrix. In Figure (a), $\mathbf{w}$ is an additive stationary white process, while $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_n$ are multiplicative stationary white processes. In Figure (b), dw represents the differential of an additive Wiener process, while $d\gamma_1, \cdots, d\gamma_n$ represent the differentials of (possibly correlated) Wiener processes that enter the dynamics multiplicatively. The signal z represents an output whose variance quantifies a performance measure.

stochastic gains $\boldsymbol{\gamma}_1(t), ...\boldsymbol{\gamma}_n(t)$, that are assumed to be "white" in time (i.e. temporally independent) but possibly mutually correlated. Another set of stochastic disturbances are represented by the vector-valued signal $\mathbf{w}$ which is also assumed to be white but enters the dynamics additively. The signal z is an output whose variance quantifies a performance measure. The feedback term is then a diagonal matrix with the individual gains $\{\boldsymbol{\gamma}_i\}$ appearing on the diagonal. Such gains are commonly referred to as structured uncertainties. Note that if the gains are deterministic (but uncertain), we obtain the general setting considered in the robust control literature (e.g. [62]). The main objective of the present chapter is to derive the necessary conditions of Mean-Square Stability (MSS) for systems taking the form of Figure 2.1(a). The treatment is carried out using a purely input-output approach (i.e. without giving $\mathcal{M}$ a state space realization). This

has the advantage of encompassing a wider class of models $\mathcal{M}$ (e.g. infinite dimensional systems).

In a discrete-time setting, there is no ambiguity of defining white (i.e. temporally independent) signals. However, in a continuous-time setting, technical issues arise because white signals are not mathematically well defined when they enter the dynamics multiplicatively. Hence, the block diagram in Figure 2.1(a) is only used to pose the problem setup in an analogous fashion to the discrete-time setting in [3], but at the cost of abandoning mathematical rigor. In fact, the equations describing Figure 2.1 can be written using the white processes $\mathbf{w}$ and $\{\boldsymbol{\gamma}_i\}$ as

$$\begin{bmatrix} z \\ y \end{bmatrix} = \mathcal{M} \begin{bmatrix} \mathbf{w} \\ \mathbf{r} \end{bmatrix} \iff \begin{bmatrix} z(t) \\ y(t) \end{bmatrix} = \int_0^t M(t-\tau) \begin{bmatrix} \mathbf{w}(\tau) \\ \mathbf{r}(\tau) \end{bmatrix} d\tau$$

$$\mathbf{r}(t) = \mathcal{D}\big(\boldsymbol{\gamma}(t)\big) y(t), \tag{2.1}$$

where M is the impulse response of $\mathcal{M}$, and $\mathcal{D}\big(\boldsymbol{\gamma}(t)\big)$ is a diagonal matrix whose elements are equal to those of $\boldsymbol{\gamma}(t) := \begin{bmatrix} \boldsymbol{\gamma}_1(t) & \cdots & \boldsymbol{\gamma}_n(t) \end{bmatrix}^*$. To resort back to mathematical rigor, we think of the white processes $\mathbf{w}$ and $\{\boldsymbol{\gamma}_i\}$ as the formal derivatives of Wiener processes (or Brownian motion) that are mathematically well defined [50]. More precisely, define

$$\boldsymbol{\gamma}_i(t) := \frac{d\gamma_i(t)}{dt}; \quad \mathbf{w}(t) := \frac{dw(t)}{dt}; \quad \mathbf{r}(t) := \frac{dr(t)}{dt}, \tag{2.2}$$

such that $\boldsymbol{\gamma}(t) := \begin{bmatrix} \gamma_1(t) & \cdots & \gamma_n(t) \end{bmatrix}^*$ and $w(t)$ represent nonstandard, vector-valued Wiener processes (i.e. their covariances do not have to be the identity matrix). Furthermore, $r(t)$ will be shown (Section 2.6.1.3) to have temporally independent increments when $\mathcal{M}$ is causal and the Itō interpretation is adopted. Hence, the equations can be rewritten using differential forms as

$$\begin{bmatrix} z \\ y \end{bmatrix} = \mathcal{M} \begin{bmatrix} dw \\ dr \end{bmatrix} \iff \begin{bmatrix} z(t) \\ y(t) \end{bmatrix} = \int_0^t M(t-\tau) \begin{bmatrix} dw(\tau) \\ dr(\tau) \end{bmatrix}$$

$$dr(t) = \mathcal{D}\big(d\boldsymbol{\gamma}(t)\big)y(t). \tag{2.3}$$

These equations are now mathematically well defined when given some desired interpretation such as in the sense of Itō or Stratonovich. It will be shown in Section 2.3.2 that different interpretations produce different conditions of MSS.

We should note the other common and related models in the literature which are usually done in a state space setting and can be represented as Stochastic Differential Equations (SDEs). One such model is a linear system with a random "A matrix" such as

$$\dot{x}(t) = A(t)x(t) + B\mathbf{w}(t), \tag{2.4}$$

where $A(t)$ is a matrix-valued stochastic process independent of $\{x(\tau), \tau \leq t\}$. One can always rewrite $A(t)$ in terms of scalar-valued stochastic processes so that

$$\dot{x}(t) = \big(A_0 + \boldsymbol{\gamma}_1(t)A_1 + \cdots + \boldsymbol{\gamma}_n(t)A_n\big)x(t) + B\mathbf{w}(t).$$

If the matrices $A_1, \ldots, A_n$ are all of rank 1 (e.g. $A_i = b_i c_i$, for column and row vectors b_i, c_i respectively, $i = 1, \ldots, n$), then it is well-known [62] that the model (2.4) can always be reconfigured like the block diagram of Figure 2.1(a) by setting

$$\mathcal{M} \;=\; \left[\begin{array}{c|cc} A_0 & B & B_0 \\ \hline C & 0 & 0 \\ C_0 & 0 & 0 \end{array}\right],$$

where $B_0 := \begin{bmatrix} b_1 & \cdots & b_n \end{bmatrix}$ and $C_0 := \begin{bmatrix} c_1^* & \cdots & c_n^* \end{bmatrix}^*$. In the example above, we have chosen $z = Cx$. If the matrices $\{A_i\}_{i=1}^n$ are not rank one, it is still possible to reconfigure (2.4) into a diagram like Figure 2.1(a), but with the perturbation blocks being "repeated" [51].

When the processes $\{\boldsymbol{\gamma}_i\}$ and $\boldsymbol{w}$ are "white" in time, we resort to the configuration of Figure 2.1(b) to express the stochastic disturbances in terms of Wiener processes. Exploiting (2.2) yields

$$\mathcal{M}: \begin{cases} dx(t) = A_0 x(t)dt + B_0 dr(t) + Bdw(t) \\ y(t) = C_0 x(t) \\ z(t) = Cx(t) \end{cases} \tag{2.5}$$

$$dr(t) = \mathcal{D}\big(d\boldsymbol{\gamma}(t)\big)y(t). \tag{2.6}$$

Equations (2.5) and (2.6) describe the block diagram of Figure 2.1(b) when $\mathcal{M}$ is given as a state space realization. In fact, the impulse response can be easily calculated to be

$$M(t) := \begin{bmatrix} C \\ C_0 \end{bmatrix} e^{A_0 t} \begin{bmatrix} B & B_0 \end{bmatrix},$$

thus showing that models like those given in (2.4) are a special case of the purely input-output approach that we consider here. On a side note, observe that the underlying stochastic dynamics of the state x in (2.5) and (2.6) can be rewritten in a single SDE, that involves both additive and multiplicative disturbances, as

$$dx(t) = A_0 x(t)dt + B_0 \mathcal{D}\big(Cx(t)\big)d\boldsymbol{\gamma}(t) + Bdw(t). \tag{2.7}$$

Particularly, [17] studied SDEs having the form of (2.7) interpreted in the sense of Itō, where $B = 0$ (i.e. no additive noise) and $\boldsymbol{\gamma}$ is "spatially uncorrelated", i.e. $\mathbb{E}[\gamma_i \gamma_j] = 0, \forall i \neq j$.

Our goal in this chapter is to extend the machinery developed in [3] to provide a rather elementary, and purely input-output treatment and derivation of the necessary and sufficient conditions of MSS for systems like that of Figure 2.1. Furthermore, our treatment covers both Itō and Stratonovich interpretations. It is shown that the conditions of MSS can be stated in terms of the spectral radius of a finite dimensional linear

operator defined in Section 2.3.2. It is also shown that this operator takes different forms when different stochastic interpretations are prescribed (such as Itō or Stratonovich).

The chapter is organized as follows. First we provide some useful definitions and notation. Then, in Section 2.2, we give a precise formulation of the problem statement by setting up a general "stochastic block diagram" and describing the underlying assumptions. In Section 2.3, we present the main results of the chapter that can be divided into two parts. The first part shows a block diagram conversion scheme from Stratonovich to Itō interpretations, and the second part states the conditions of mean-square stability. The special cases of state space realizations are then treated in Section 2.4. Sections 2.5 and 2.6 provide the detailed derivations that explain the results. Finally, we conclude in Section 2.7.

2.1 Preliminaries and Notation

All the signals considered in this chapter are defined on the semi-infinite, continuous-time interval $\mathbb{R}^+ := [0, +\infty)$. The dynamical systems considered are maps between various signal spaces over the time interval $\mathbb{R}^+$. Unless stated otherwise, all stochastic processes considered here are random vector-valued functions of (continuous) time.

Notation Summary

2.1.1 Variance & Covariance Matrix of a Signal

If v is a stochastic signal, then its instantaneous variance and covariance matrix are denoted by the lowercase and uppercase bold letters respectively

$$\mathbf{v}(t) := \mathbb{E}\left[v^*(t)v(t)\right] \quad \text{and} \quad \mathbf{V}(t) := \mathbb{E}\left[v(t)v^*(t)\right],$$

where v^* denotes the transpose of v. The entries of $\mathbf{V}(t)$ are the mutual correlations of the vector $v(t)$, and are sometimes referred to as *spatial correlations*. Note that $\mathrm{tr}\,(\mathbf{V}(t)) = \mathbf{v}(t)$.

2.1.2 Variance & Covariance Matrix of a Differential Signal

If the differential du of a stochastic signal u appears in a stochastic block diagram (see Figure 2.2 for example), its instantaneous variance and covariance are represented as

$$\mathbb{E}\left[du^*(t)du(t)\right] := \mathbf{u}(t)dt \quad \text{and} \quad \mathbb{E}\left[du(t)du^*(t)\right] := \mathbf{U}(t)dt,$$

respectively. This is a compact (differential) notation for

$$\mathbb{E}\left[u^*(t)u(t)\right] := \int_0^t \mathbf{u}(\tau)d\tau; \quad \mathbb{E}\left[u(t)u^*(t)\right] := \int_0^t \mathbf{U}(\tau)d\tau.$$

2.1.3 Steady State Variance & Covariance Matrix

The asymptotic limits of the instantaneous variance and covariance matrix, when they exist, are denoted by an overbar, i.e.

$$\bar{\mathbf{u}} := \lim_{t\to\infty} \mathbf{u}(t) \quad \text{and} \quad \bar{\mathbf{U}} := \lim_{t\to\infty} \mathbf{U}(t).$$

2.1.4 Second Order Process

A process v is termed *second order* if the entries of its covariance matrix, $\mathbf{V}(t)$, are finite for each $t \in \mathbb{R}^+$.

2.1.5 Probability Space

Let $(\Omega, \mathcal{F}, p)$ be a complete probability space with Ω being the sample space, $\mathcal{F}$ the associated $\sigma-$algebra and p the probability measure. Let $L_2(p)$ denote the space of

vector-valued random variables with finite second order moments. Note that $L_2(p)$ is a Hilbert space.

2.1.6 Equalities & Limits in the Mean-Square Sense

Two stochastic processes x and y are said to be equal in the mean-square sense if $\mathbb{E}\left[||x-y||^2\right]=0$, where throughout this chapter, $||.||$ denotes the ℓ^2-norm for vectors and the spectral norm for matrices.

A sequence of second order stochastic processes, $\{x_N\}$, is said to converge to $\bar{x}\in L_2(p)$ in the mean-square sense iff $\lim_{N\to\infty}||x_N-\bar{x}||^2=0$.

2.1.7 White Process

A stochastic process $\boldsymbol{\gamma}$ is termed *white* if it is uncorrelated at any two distinct times, i.e. $\mathbb{E}\left[\boldsymbol{\gamma}(t)\boldsymbol{\gamma}^*(\tau)\right]=\boldsymbol{\Gamma}\delta(t-\tau)$, where δ is the Dirac delta function. Note that in the present context, a white process $\boldsymbol{\gamma}$ may still have spatial correlations, i.e. its instantaneous covariance matrix $\boldsymbol{\Gamma}$ need not be the identity.

2.1.8 Vector-Valued Wiener Process

In a continuous-time setting, calculus operations on a white process entering the dynamics multiplicatively are not mathematically well defined. Hence, it is useful to represent a white process as the formal derivative of a Wiener process, i.e. $\boldsymbol{\gamma}(t):=\frac{d\boldsymbol{\gamma}(t)}{dt}$, where $\boldsymbol{\gamma}$ is a zero-mean, vector-valued Wiener process with an instantaneous covariance matrix $\mathbb{E}\left[\boldsymbol{\gamma}(t)\boldsymbol{\gamma}^*(t)\right]=\boldsymbol{\Gamma}t$. This can be equivalently written in differential form as $\mathbb{E}\left[d\boldsymbol{\gamma}(t)d\boldsymbol{\gamma}^*(t)\right]=\boldsymbol{\Gamma}dt$. Note that $\boldsymbol{\gamma}$ is said to have temporally independent increments, i.e. its differentials $\big(d\boldsymbol{\gamma}(t),d\boldsymbol{\gamma}(\tau)\big)$ are independent when $t\neq\tau$.

2.1.9 Partitions of Time Intervals

Let $\mathcal{P}_N[0,t]$ denote an arbitrary partition of the time interval $[0,t]$ into N subintervals $[t_k, t_{k+1}]$ for $k = 0, 1, \cdots, N-1$, such that $0 = t_0 < t_1 < \cdots < t_N = t$. The partition step-size is denoted by $\Delta_k := t_{k+1} - t_k$ and the norm of the partition $\mathcal{P}_N[0,t]$ is denoted by the bold letter $\boldsymbol{\Delta}$ defined as $\boldsymbol{\Delta} := ||\mathcal{P}_N[0,t]|| = \sup_k \Delta_k$. Note that $\lim_{N\to\infty} \boldsymbol{\Delta} = 0$.

2.1.10 Notation for Signals and Increments on $\mathcal{P}_N[0,t]$

With slight abuse of notation, a continuous-time stochastic signal $\{u(\tau), 0 \le \tau \le t\}$ is represented at node t_k of the partition $\mathcal{P}_N[0,t]$ as $u_k := u(t_k)$ for $k = 0, 1, \cdots, N$. The increments of $\{u(\tau), 0 \le \tau \le t\}$ at t_k are denoted by $\tilde{u}_k := u(t_{k+1}) - u(t_k)$ for $k = 0, 1, \cdots, N-1$, and they represent a finite approximation of the differential form $\{du(\tau), 0 \le \tau \le t\}$.

A continuous-time stochastic process u is said to have *temporally independent increments* if $\big(du(t), du(\tau)\big)$ are independent whenever $t \neq \tau$. This implies that, on the partition $\mathcal{P}_N[0,t]$, $(\tilde{u}_k, \tilde{u}_l)$ are independent whenever $k \neq l$.

2.1.11 Stochastic Integrals

Calculus operations on a Wiener process are mathematically well defined when some stochastic interpretation is prescribed (such as Itō or Stratonovich). Particularly, we distinguish Itō and Stratonovich integrals using the symbols "$\diamond_I$" and "$\diamond_S$", respectively. More precisely, let v be a vector-valued second order stochastic process and $\boldsymbol{\gamma}$ be a vector-valued Wiener process. If $\Gamma(t) := \mathcal{D}\big(\boldsymbol{\gamma}(t)\big)$ is a diagonal matrix whose entries are equal to those of $\boldsymbol{\gamma}(t)$, then the integral "$\int_0^t d\Gamma(\tau) v(\tau)$" may be interpreted differently using

partial sums as

$$\int_0^t d\Gamma(\tau) \diamond_I v(\tau) := \lim_{N\to\infty} \sum_{k=0}^{N-1} \tilde{\Gamma}_k v_k \tag{2.8}$$

$$\int_0^t d\Gamma(\tau) \diamond_S v(\tau) := \lim_{N\to\infty} \sum_{k=0}^{N-1} \tilde{\Gamma}_k \frac{v_k + v_{k+1}}{2}. \tag{2.9}$$

The partial sums are constructed using a partition $\mathcal{P}_N[0,t]$ as described in Section 2.1.9 and by following the notation developed in Section 2.1.10 for signals and increments.

2.1.12 Quadratic Variation

The quadratic variation, at time t, of a stochastic process v is denoted by $\langle v\rangle(t)$ and is defined using a partition $\mathcal{P}_N[0,t]$ as

$$\langle v\rangle(t) := \lim_{N\to\infty} \sum_{k=0}^{N-1} ||\tilde{v}_k||^2 .$$

2.1.13 Hadamard Product and the Diagonal Operator

For any two matrices A and B of the same dimensions, their Hadamard (or element-by-element) product is denoted by $A \circ B$. For any vector v (resp. square matrix V), $\mathcal{D}(v)$ (resp. $\mathcal{D}(V)$) denotes a diagonal matrix whose diagonal elements are equal to v (resp. diagonal entries of V).

2.2 Problem Formulation

In this section, we first provide a precise definition for Mean-Square Stability (MSS) from a purely input/output approach. Then we present a "stochastic block diagram" formalism that can be given a desirable interpretation by prescribing a suitable stochastic calculus (Itō or Stratonovich).

2.2.1 Input-Output Formulation of MSS

Let $\mathcal{M}$ be a causal LTI (MIMO) system. It is defined as a linear operator that acts on the differential of a second order stochastic signal u, denoted by du. Its action is defined by the stochastic convolution integral

$$y(t) = \big(\mathcal{M}du\big)(t) \Longleftrightarrow y(t) = \int_0^t M(t-\tau)\ du(\tau), \tag{2.10}$$

where M is a deterministic matrix-valued function denoting the impulse response of $\mathcal{M}$. Without loss of generality, zero initial conditions are assumed throughout this chapter. When u is zero-mean and has independent increments such that $\mathbb{E}\left[du(t)du^*(\tau)\right] = 0\ \forall t \neq \tau$ and $\mathbb{E}\left[du(t)du^*(t)\right] = \mathbf{U}(t)dt$, a standard calculation relates the input and output instantaneous covariances as

$$\mathbf{Y}(t) = \int_0^t M(t-\tau)\ \mathbf{U}(\tau)\ M^*(t-\tau)d\tau. \tag{2.11}$$

Note that (2.11) holds for any stochastic interpretation (eg. Itō or Stratonovich) of the stochastic integral in (2.10) as shown in Appendix 2.A. Therefore, the action of $\mathcal{M}$ as described in (2.10) is not given a particular stochastic interpretation here. Unlike (2.10), this matrix convolution relationship is deterministic, and it is only valid when the input du is temporally independent (i.e. u has independent increments). Taking the trace of both sides of (2.11) yields

$$\begin{aligned}
\mathbf{y}(t) = \operatorname{tr}\left(\mathbf{Y}(t)\right) &= \int_0^t \operatorname{tr}\big(M(t-\tau)\mathbf{U}(\tau)M^*(t-\tau)\big)d\tau \\
&= \int_0^t \operatorname{tr}\big(M^*(t-\tau)M(t-\tau)\mathbf{U}(\tau)\big)d\tau \\
&\leq \int_0^t \operatorname{tr}\big(M^*(t-\tau)M(t-\tau)\big)\operatorname{tr}\big(\mathbf{U}(\tau)\big)d\tau \\
&\leq \int_0^\infty \operatorname{tr}\big(M^*(t-\tau)M(t-\tau)\big)d\tau \sup_{0\leq\tau\leq\infty} \mathbf{u}(\tau),
\end{aligned}$$

where the first inequality holds because for any two positive semidefinite matrices A and B, we have $\mathrm{tr}\,(AB) \leq \mathrm{tr}\,(A)\,\mathrm{tr}\,(B)$ [13, Thm 1]. The calculation above motivates the following definition for input/output MSS when the input is temporally independent.

Definition 1 *A causal LTI system $\mathcal{M}$ is Mean-Square Stable (MSS) if for each input du, representing the differential of a stochastic process with independent increments and uniformly bounded variance, the output process $y = \mathcal{M}du$ has a uniformly bounded variance, i.e. there exists a constant c such that $\mathbf{y}(t) \leq c\ \sup_\tau \mathbf{u}(\tau)$.*

It is easy to check that $\mathcal{M}$ is MSS in the sense of Definition 1 if and only if $\|\mathcal{M}\|_2$ is finite, where $||.||_2$ denotes the H^2-norm. When MSS holds, the output covariance has a finite steady-state limit $\bar{\mathbf{Y}}$ whenever the input covariance has a finite steady-state limit $\bar{\mathbf{U}}$. From (2.11), it is straight forward to see that the steady-state covariances (if they exist) are related as

$$\bar{\mathbf{Y}} = \int_0^\infty M(\tau)\bar{\mathbf{U}}M^*(\tau)d\tau. \tag{2.12}$$

2.2.2 Stochastic Feedback Interconnection

Consider the "stochastic block diagram" depicted in Figure 2.2 where the forward block represents a causal LTI system which is in feedback with multiplicative stochastic gains represented here as the differential of a diagonal matrix denoted by $d\Gamma(t)$ where

$$d\Gamma(t) := \mathcal{D}\big(d\boldsymbol{\gamma}(t)\big) \quad \text{and} \quad d\boldsymbol{\gamma}(t) := \begin{bmatrix} d\gamma_1(t) & \cdots & d\gamma_n(t) \end{bmatrix}^*. \tag{2.13}$$

Furthermore, a different type of stochastic disturbance enters the dynamics additively and is represented in Figure 2.2 as the differential of w.

The main objective of this chapter is to investigate the MSS of Figure 2.2 under the following assumptions

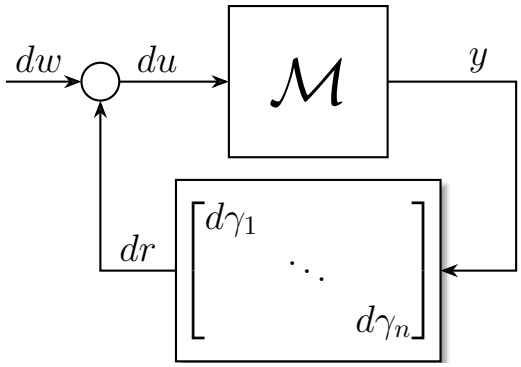


Figure 2.2: A continuous-time setting for a causal LTI system $\mathcal{M}$ in feedback with stochastic multiplicative gains $\{d\gamma_i\}$ that represent the differential forms of, possibly mutually correlated, Wiener processes. The equations describing the block diagram are given in (2.14).

- **Assumption 1** *$\mathcal{M}$ is a causal LTI (MIMO) system whose impulse response M belongs to the class $\mathcal{C}$ of deterministic, matrix-valued functions defined in Appendix 2.E. Note that for such M, $\exists$ a continuous scalar function c_M such that* $\sup_{0\leq\tau\leq t} ||M(\tau)|| = c_M(t)$.

- **Assumption 2** $\boldsymbol{\gamma}(t) := \begin{bmatrix}\gamma_1(t) & \cdots & \gamma_n(t)\end{bmatrix}^*$ *is a zero-mean, vector-valued Wiener process with an instantaneous covariance* $\mathbb{E}\left[\boldsymbol{\gamma}(t)\boldsymbol{\gamma}^*(t)\right] := \boldsymbol{\Gamma} t$ *which can be equivalently written as* $\mathbb{E}\left[d\boldsymbol{\gamma}(t)d\boldsymbol{\gamma}^*(t)\right] = \boldsymbol{\Gamma} dt$ *(refer to Section 2.1.8). Note that* $\boldsymbol{\Gamma}$ *is a constant positive semidefinite matrix.*

- **Assumption 3** *w is a zero-mean, vector-valued Wiener process with a (possibly) time-varying instantaneous covariance matrix, i.e.* $\mathbb{E}\left[dw(t)dw^*(t)\right] = \mathbf{W}(t)dt$, *where* $\mathbf{W}$ *is a positive semidefinite matrix whose entries remain bounded for all time. Furthermore,* $\mathbf{W}$ *is assumed to be monotone, i.e. if* $t_1 \leq t_2$ *then* $\mathbf{W}(t_1) \leq \mathbf{W}(t_2)$.

- **Assumption 4** *$\boldsymbol{\gamma}$ and w are uncorrelated for all time.*

Throughout this chapter, whenever the Stratonovich interpretation is adopted, a more

restrictive assumption on M is required for reasons that will become apparent in Section 2.5. Thus Assumption 1 is replaced by

- **Assumption 1′** *M is Lipschitz continuous.*

Note that the class of Lipschitz continuous functions is more restrictive than class $\mathcal{C}$ defined in Appendix 2.E. In fact, it is fairly straightforward to see that if M is Lipschitz continuous, then $M \in \mathcal{C}$.

The equations describing the block diagram in Figure 2.2 can be written as

$$\begin{cases} y(t) = (\mathcal{M}du)(t) \\ du(t) = dw(t) + dr(t) \\ dr(t) = d\Gamma(t)y(t). \end{cases} \tag{2.14}$$

Note that, without prescribing a stochastic interpretation for the calculus operations on the Wiener processes w and Γ, the set of equations in (2.14) are not sufficient to fully describe the underlying stochastic dynamics. We consider here the two most common interpretations named after Itō and Stratonovich; however, the analysis can be generalized to other interpretations as well. We encode the stochastic interpretations in (2.14) by rewriting them as

$$\begin{cases} y(t) = (\mathcal{M}du)(t) \\ du(t) = dw(t) + dr(t) \\ dr(t) = d\Gamma(t) \diamond y(t); \qquad \text{for} \quad \diamond = \{\diamond_I, \diamond_S\}, \end{cases} \tag{2.15}$$

where the last equation is the differential form of an integral equation that can be written as

$$r(t) = \int_0^t d\Gamma(\tau) \diamond y(\tau), \qquad \text{where} \quad \diamond = \{\diamond_I, \diamond_S\}.$$

Refer to Section 2.1.11 for an explanation of the different interpretations. Note that We close this section by giving a definition for MSS of the stochastic feedback system in Figure 2.2 by following the convention given in [16].

Definition 2 *Consider the stochastic feedback interconnection in Figure 2.2 satisfying assumptions 1-4. The overall feedback system is said to be MSS if all the signals in the loop, i.e. du, dr and y have uniformly bounded variances. More precisely, there exists a constant c such that*

$$max\{||\mathbf{u}||_\infty, ||\mathbf{r}||_\infty, ||\mathbf{y}||_\infty\} \leq c\,||\mathbf{w}||_\infty .$$

The next section characterizes the conditions of MSS for Figure 2.2 for different stochastic interpretations.

2.3 Main Results

Observe that the set of equations (2.15) can be rewritten as a single equation

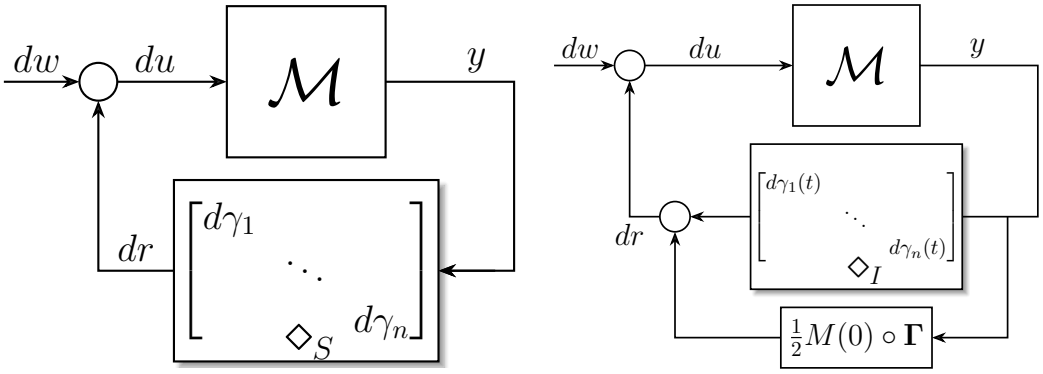
$$y(t) = \int_0^t M(t-\tau)dw(\tau) + \int_0^t M(t-\tau) \diamond d\Gamma(\tau)y(\tau); \quad \text{for } \diamond = \{\diamond_I, \diamond_S\}. \tag{2.16}$$

Equation (2.16) is a linear Stochastic Integral Equation (SIE) of Volterra type. The Itō version of (2.16) has been addressed in the literature ([34], [5], [6], [4]). For example, it is easy to check that (2.16), interpreted in the sense of Itō, has a unique solution [4, Thm 5A] under the assumption that M is finite over bounded intervals (Assumption 1). However, SIEs interpreted in the sense of Stratonovich are less common in the literature. In contrast, SDEs interpreted in the sense of Stratonovich [60] are analyzed by converting them to their equivalent Itō representation using the conversion formulas that were derived several decades ago (see e.g. [58]). In the present paper, the analysis is carried out from a purely input-output approach, and thus a more general conversion formula is required to convert an SIE interpreted in the sense of Stratonovich to its equivalent Itō counterpart. In this section, we first describe the conversion scheme, then state the MSS conditions of Figure 2.2 when different stochastic interpretations are adopted.

2.3.1 Block Diagram Conversion from Stratonovich to Itō Interpretations

Consider the block diagram in Figure 2.3(a) such that Assumptions 1′, 2, 3, and 4 are satisfied. As opposed to Figure 2.2, the multiplicative gains are now given a Stratonovich interpretation indicated by the symbol "$\diamond_S$" in the feedback block. Now we present a theorem that describes a conversion scheme of block diagrams from Stratonovich to Itō interpretations.

Theorem 1 *Under Assumptions 1′, 2, 3, and 4, the two block diagrams in Figures 2.3(a) and (b) are equivalent in the mean-square sense. That is, all the signals du, y, dw and dr in both block diagrams are equal in the mean-square sense.*



(a) Stratonovich Interpretation (b) Equivalent Itō Interpretation

Figure 2.3: (a) A continuous-time causal LTI system $\mathcal{M}$ in feedback with stochastic multiplicative gains $\{d\gamma_i\}$ that represent the differential forms of, possibly mutually correlated, Wiener processes. The diamond "$\diamond_S$" in the feedback block indicates a Stratonovich interpretation. (b) The equivalent Itō interpretation, in the mean-square sense, of the block diagram given in (a). The symbol "$\circ$" denotes the Hadamard (element-by-element) product and "$\diamond_I$" indicates an Itō interpretation of the multiplicative gains.

The proof of Theorem 1 is given in Section 2.5. A remark is worth noting here.

Remark 2.3.1 *If $M(0) = 0$, the block diagrams in Figures 2.3 (a) and (b) become identical. This means that there is no difference between Itō and Stratonovich interpretations if the impulse response is zero at initial time. This sort of reintroduces a notion of "strict causality" that forces the Stratonovich interpretation to behave in the same way as that of Itō. Therefore, LTI systems $\mathcal{M}$ with relative degrees [1] ≥ 2 have the same MSS conditions for both Itō and Stratonovich interpretations.*

2.3.2 MSS Conditions

The MSS setting considered here is given in Figure 2.2 and is repeated here in Figure 2.4 to explicitly show the adopted stochastic interpretation of the feedback block. In this section, MSS conditions are given in terms of a linear operator, denoted by $\mathbb{L}$, that acts on a positive semidefinite matrix to produce another positive semidefinite matrix. Its role is to propagate the steady-state covariance (if it exists) of du, denoted by $\bar{\mathbf{U}}$,

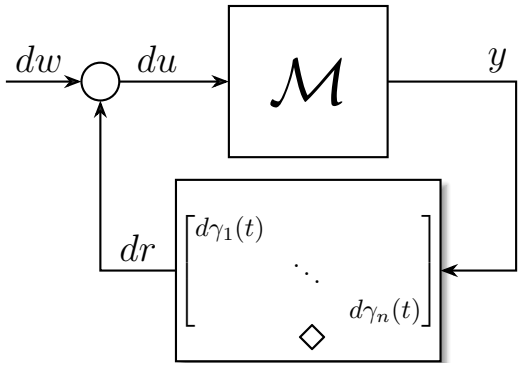


Figure 2.4: Mean-square stability setting. This figure is similar to the general setting given Figure 2.2. The only difference is that the stochastic interpretation of the feedback block is encoded by the symbol "$\diamond$" such that $\diamond = \diamond_I$ denotes an Itō interpretation, whereas $\diamond = \diamond_S$ denotes a Stratonovich interpretation.

through the loop to yield that of dr, denoted by $\bar{\mathbf{R}}$. This "Loop Gain Operator" (LGO)

[1] The relative degree of an LTI system with impulse response M is defined as the largest positive integer p such that $\lim_{s\to\infty} s^p M(s) < \infty$.

is the continuous-time counterpart of that defined in [3] for the discrete-time setting. For the Itō setting (i.e. $\diamond = \diamond_I$ in Figure 2.4), the LGO is denoted by $\mathbb{L}_I$ and is given by

$$\bar{\mathbf{R}} = \mathbb{L}_I\left(\bar{\mathbf{U}}\right) := \mathbf{\Gamma} \circ \left(\int_0^\infty M(\tau)\bar{\mathbf{U}}M^*(\tau)d\tau\right). \tag{2.17}$$

Refer to Section 2.6 for a detailed derivation of the LGO. A key step in the derivation of $\mathbb{L}_I$ is showing that du is temporally independent which is required to propagate $\bar{\mathbf{U}}$ in the forward block $\mathcal{M}$ using (2.12). As will be shown in Section 2.6.1, this temporal independence is a consequence of (1) the causality of $\mathcal{M}$, (2) the temporal independence of the stochastic multiplicative gains, and (3) the Itō interpretation. However, for the Stratonovich setting (i.e. $\diamond = \diamond_S$ in Figure 2.4), du is not temporally independent. This is a consequence of the nature of the Stratonovich integral in (2.9) that "looks into the future". In this case, (2.12) cannot be used to propagate the covariance in the forward block of Figure 2.3(a). Nonetheless, one can exploit the block diagram conversion scheme in Section 2.3.1 and rearrange the block diagram in Figure 2.3(b) so that it looks like the Itō setting as depicted in Figure 2.5. The equivalent forward block, now denoted by $\mathcal{H}$, is still a causal LTI system whose transfer function is

$$H(s) = \left(I - M(s)G\right)^{-1} M(s), \tag{2.18}$$

where $G := \frac{1}{2}M(0) \circ \mathbf{\Gamma}$ and $M(s)$ is the transfer function of $\mathcal{M}$. The input differential signal du_S in Figure 2.5 is now temporally independent and thus (2.12) can be exploited to propagate the steady state covariance through the equivalent forward block $\mathcal{H}$. Thus, the LGO for the Stratonovich setting propagates the steady-state covariance (if it exists) of du_S, denoted by $\bar{\mathbf{U}}_S$, through the loop of Figure 2.5 to yield that of dr_S, denoted by $\bar{\mathbf{R}}_S$. It is now denoted by $\mathbb{L}_S$ and is given by

$$\bar{\mathbf{R}}_S = \mathbb{L}_S\left(\bar{\mathbf{U}}_S\right) := \mathbf{\Gamma} \circ \left(\int_0^\infty H(\tau)\bar{\mathbf{U}}_S H^*(\tau)d\tau\right), \tag{2.19}$$

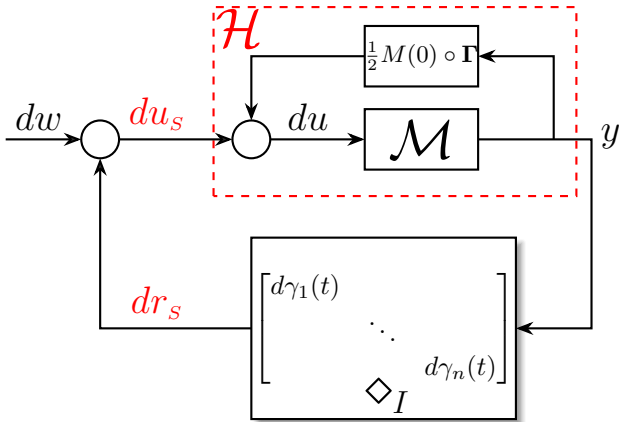


Figure 2.5: Rearrangement of the block diagram in Figure 2.3(b)

where H is given in (2.18). The spectral radius of $\mathbb{L}$ completely characterizes the MSS condition as will be seen next.

Theorem 2 *Consider the system in Figure 2.4 such that Assumptions 1-4 are satisfied. The feedback system is MSS if and only if the two conditions are satisfied*

1. *The equivalent forward block in Figure 2.4 has a finite H^2- norm.*

2. *The spectral radius of the loop gain operator is strictly less than 1, i.e. $\rho(\mathbb{L})<1$.*

where

- *For the Itō interpretation, the equivalent forward block is $\mathcal{M}$, and $\mathbb{L}$ is given in (2.17).*

- *For the Stratonovich interpretation, the equivalent forward block is $\mathcal{H}$, whose transfer function is given in (2.18), $\mathbb{L}$ is given in (2.19), and Assumption 1 is replaced by Assumption 1′.*

The proof of Theorem 2 is given in Section 2.6. Observe that, under the Itō interpretations, the covariance matrix $\mathbf{\Gamma}$ only plays a role in the second condition. However, under

the Stratonovich interpretation, $\mathbf{\Gamma}$ plays a role in both conditions since the equivalent forward block $\mathcal{H}$ now depends on $\mathbf{\Gamma}$ (Figure 2.5). Therefore, the conditions of MSS can be very different when different stochastic interpretations are adopted.

2.4 Application to State Space Realizations & SDEs

In this section, we consider the mean-square stability problems for both the Itō and Stratonovich settings given in Figure 2.4, but for the special case when $\mathcal{M}$ is given a state space realization. Thus, the underlying equations can be written as SDEs, i.e.

$$\begin{aligned} dx(t) &= Ax(t)dt + Bdu(t); \qquad y(t) = Cx(t) \\ du(t) &= dw(t) + dr(t) \\ dr(t) &= d\Gamma(t) \diamond y(t) \quad \text{for} \quad \diamond = \{\diamond_I, \diamond_S\}, \end{aligned} \tag{2.20}$$

where the last equation refers to either an Itō or Stratonovich interpretation. The impulse response of $\mathcal{M}$ can thus be written as $M(t) = Ce^{At}B$. Then, the realization of the loop gain operator, for each interpretation, can be calculated using (2.17) and (2.19). Starting with the Itō interpretation, we have

$$\begin{aligned} \bar{\mathbf{R}} = \mathbb{L}_I(\bar{\mathbf{U}}) &:= \mathbf{\Gamma} \circ \left(\int_0^\infty M(\tau) \bar{\mathbf{U}} M^*(\tau) d\tau \right) \\ &= \mathbf{\Gamma} \circ \left(C \int_0^\infty e^{A\tau} B \bar{\mathbf{U}} B^* e^{A^*\tau} d\tau \right) C \\ &= \mathbf{\Gamma} \circ \left(C \bar{\mathbf{X}} C \right), \end{aligned}$$

where $\bar{\mathbf{X}} := \int_0^\infty e^{A\tau} B \bar{\mathbf{U}} B^* e^{A^*\tau} d\tau$ which satisfies the algebraic Lyapunov equation given by

$$A\bar{\mathbf{X}} + \bar{\mathbf{X}} A^* + B\bar{\mathbf{U}} B^* = 0.$$

For the Stratonovich interpretation, we use Figure 2.5 to give the equivalent Itō representation. The impulse response of $\mathcal{H}$ in Figure 2.3(b) can be shown to be $H(t) = Ce^{A_S t}$

with $A_S = A + 1/2B\big((CB)\circ\mathbf{\Gamma}\big)C$ and the LGO can be similarly given a realization. To summarize, let $\mathbb{L}_I$ and $\mathbb{L}_S$ denote the loop gain operators for the Itō and Stratonovich interpretations as given in (2.17) and (2.19), respectively. Then their state space realizations are given by

$$\begin{aligned}&\bar{\mathbf{R}} = \mathbb{L}_k(\bar{\mathbf{U}})\\ &(k = I, S)\end{aligned} \quad \Longleftrightarrow \quad \begin{cases}\bar{\mathbf{R}} = \mathbf{\Gamma}\circ\big(C\bar{\mathbf{X}}C^*\big)\\ 0 = A_k\bar{\mathbf{X}} + \bar{\mathbf{X}}A_k^* + B\bar{\mathbf{U}}B^*;\end{cases} \tag{2.21}$$

where $A_I := A$ and $A_S := A + \frac{1}{2}B\big((CB)\circ\mathbf{\Gamma}\big)C$. Therefore, as a direct application of Theorem 2, the necessary and sufficient conditions of MSS are (1) A_k is Hurwitz and (2) $\rho(\mathbb{L}_k) < 1$ for $k = I, S$ for Itō and Stratonovich interpretations, respectively.

2.5 Stochastic Block Diagram Conversion Technique

In this section, we provide a proof for Theorem 1. Consider the Stratonovich setting in Figure 2.3(a) such that Assumptions 1′, 2, 3, and 4 are satisfied. The block diagram can be described by a single SIE given in (2.16) with $\diamond = \diamond_S$, and the goal of this section is to show that it is equivalent (in the mean-square sense) to

$$y(t) = \int_0^t M(t-\tau)dw(\tau) + \int_0^t M(t-\tau)\diamond_I d\Gamma(\tau)y(\tau) + \frac{1}{2}\int_0^t M(t-\tau)\big(M_0\circ\mathbf{\Gamma}\big)y(\tau)d\tau, \tag{2.22}$$

where $M(0)$ is denoted by M_0 for notational convenience. This can be shown by exploiting the following two propositions.

Proposition 1 *Consider the SIE given in (2.22) (or equivalently (2.16) with $\diamond = \diamond_S$) such that Assumptions 1′, 2, 3, and 4 are satisfied. Then the second moments of y and its quadratic variation (Section 2.1.12) are both finite over finite intervals. That is, there exist two scalar continuous functions c_y and c_q such that*

$$\sup_{0\le\tau\le t}\mathbb{E}\left[||y(\tau)||^2\right] = c_y(t);\ \sup_{0\le\tau\le t}\mathbb{E}\left[\langle y\rangle^2(\tau)\right] = c_q(t). \tag{2.23}$$

The proof of the boundedness of $\mathbb{E}\left[||y(\tau)||^2\right]$ *is given in [4, Thm 5A] while that of the quadratic variation is given in Section 2.F. These bounds will be useful to prove Proposition 2.*

Proposition 2 *Consider the Stratonovich integral*

$$S(t) := \int_0^t M(t-\tau) d\Gamma(\tau) \diamond_S y(\tau),$$

where M *satisfies Assumption 1,* $d\Gamma(t)$ *is defined in (2.13) such that* $\boldsymbol{\gamma}$ *satisfies Assumption 2, and* y *is a stochastic process that satisfies (2.16) with* $\diamond = \diamond_S$*. Then* $S(t) = I(t) + \frac{1}{2}R(t)$ *in the mean-square sense, where*

$$I(t) := \int_0^t M(t-\tau) \diamond_I d\Gamma(\tau) y(\tau) \qquad \text{and} \qquad R(t) := \int_0^t M(t-\tau)\big(M_0 \circ \mathbf{\Gamma}\big) y(\tau) d\tau$$

are Itō and Riemann integrals, respectively.

Proof: Start by using the definitions of the various integrals in Section 2.1.11 to construct the partial sums over a partition $\mathcal{P}_N[0,t]$ (2.1.9) as

$$\begin{aligned} S_N(t) &:= \frac{1}{2}\sum_{k=0}^{N-1}\Big(M(t-t_{k+1})\tilde{\Gamma}_k y_{k+1} + M(t-t_k)\tilde{\Gamma}_k y_k\Big) \\ I_N(t) &:= \sum_{k=0}^{N-1} M(t-t_k)\tilde{\Gamma}_k y_k \\ R_N(t) &:= \sum_{k=0}^{N-1} M(t-t_k)\big(M_0 \circ \mathbf{\Gamma}\big) y_k \Delta_k. \end{aligned} \tag{2.24}$$

The proof is carried out on the partition $\mathcal{P}_N[0,t]$ but can be passed to the limit in $L_2(p)$ (since it is a Hilbert space and all Cauchy sequences are convergent). More precisely, we are required to prove that $\lim_{N\to\infty}\mathbb{E}\left[D_N^2(t)\right] = 0 \ \forall t \geq 0$,

$$\text{where} \qquad D_N(t) = S_N(t) - \left(I_N(t) + \frac{1}{2}R_N(t)\right). \tag{2.25}$$

After carrying out a sequence of algebraic manipulations (Appendix 2.B), the expression of $D_N(t)$ can be rewritten as

$$\begin{aligned} D_N(t) = \frac{1}{2}\Big(\lambda_N(t) + J_N(t) + \nu_N(t) + \xi_N(t) + T_N^{\zeta}(t)\Big) \\ + \frac{1}{4}\Big(\theta_N(t) + \eta_N(t) + T_N^{\alpha}(t) + T_N^{\beta}(t)\Big), \end{aligned} \tag{2.26}$$

where

$$\begin{aligned} \lambda_N(t) &:= \sum_{k=0}^{N-1} M(t-t_k)\Big(\big(\tilde{\boldsymbol{\gamma}}_k\tilde{\boldsymbol{\gamma}}_k^* - \boldsymbol{\Gamma}\Delta_k\big)\circ M_0\Big)y_k \\ J_N(t) &:= \sum_{k=0}^{N-1}\Big(M(t-t_{k+1}) - M(t-t_k)\Big)\tilde{\Gamma}_k y_k \\ \nu_N(t) &:= \sum_{k=0}^{N-1}\Big(M(t-t_{k+1}) - M(t-t_k)\Big)\tilde{\Gamma}_k M_0\tilde{\Gamma}_k y_k \\ \theta_N(t) &:= \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k M_0\tilde{\Gamma}_k\tilde{y}_k \\ \eta_N(t) &:= \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k\Big(M(\Delta_k) - M_0\Big)\tilde{\Gamma}_k y_k \\ \chi_N(t) &:= \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k M(\Delta_k)\tilde{w}_k \\ T_N^x(t) &:= \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k x_k \quad \text{for} \quad x\in\{\alpha,\beta,\zeta\} \\ \alpha_k &:= \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_{l+1}) - M(t_k-t_{l+1})\Big)\tilde{\Gamma}_l\tilde{y}_l \\ \beta_k &:= \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_{l+1}) - M(t_k-t_{l+1}) \\ &\qquad + M(t_{k+1}-t_l) - M(t_k-t_l)\Big)\tilde{\Gamma}_l y_l \\ \zeta_k &:= \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_l) - M(t_k-t_l)\Big)\tilde{w}_l. \end{aligned} \tag{2.27}$$

The rest of the proof shows that the second moment of each term in (2.26) goes to zero in the limit as N goes to infinity. Note that there is no need to check the expectation of cross terms (Appendix 2.C).

2.5.0.1 Mean-Square Convergence of $\lambda_N(t)$

Recall that $\boldsymbol{\gamma}_k$ has independent increments that are also independent from present and past values of y_k. Furthermore, $\mathbb{E}[Z_k] = 0$ with $Z_k := \tilde{\boldsymbol{\gamma}}_k \tilde{\boldsymbol{\gamma}}_k^* - \boldsymbol{\Gamma}\Delta_k$. Then we invoke Lemma 2.D.6 to yield the following inequality

$$\begin{aligned}\mathbb{E}\left[||\lambda_N(t)||^2\right] &\leq \sum_{k=0}^{N-1} ||M(t-t_k)||^2 \, \mathbb{E}\left[\left|\left|\left(Z_k \circ M_0\right)\right|\right|^2\right] \mathbb{E}\left[||y_k||^2\right] \\ &\leq ||M_0||^2 \sum_{k=0}^{N-1} ||M(t-t_k)||^2 \, \mathbb{E}\left[||Z_k||^2\right] \mathbb{E}\left[||y_k||^2\right],\end{aligned}$$

where the second inequality follows from the sub-multiplicative property of the matrix spectral norm with respect to matrix and Hadamard products (see [33]). Knowing that $\tilde{\boldsymbol{\gamma}}_k \sim \mathcal{N}(0, \boldsymbol{\Gamma}\Delta_k)$, we can write $\tilde{\boldsymbol{\gamma}}_k = \boldsymbol{\Gamma}^{1/2}\boldsymbol{\xi}_k\sqrt{\Delta_k}$, where $\boldsymbol{\Gamma}^{1/2}$ denotes the Cholesky factorization of $\boldsymbol{\Gamma}$. The random vector $\boldsymbol{\xi}_k$ follows a standard multivariate normal distribution for all $k = 0, 1, ...N-1$ such that $\boldsymbol{\xi}_k$ and $\boldsymbol{\xi}_l$ are independent for $k \neq l$. To bound $\mathbb{E}\left[||Z_k||^2\right]$, we proceed as follows

$$\begin{aligned}\mathbb{E}\left[||Z_k||^2\right] &= \mathbb{E}\left[\left|\left|\boldsymbol{\Gamma}^{1/2}(\boldsymbol{\xi}_k\boldsymbol{\xi}_k^* - I)\boldsymbol{\Gamma}^{1/2}\right|\right|^2 \Delta_k^2\right] \\ &\leq \mathbb{E}\left[||\boldsymbol{\Gamma}||\,||\boldsymbol{\xi}_k\boldsymbol{\xi}_k^* - I||^2\,\Delta_k^2\right] \\ &\leq \mathbb{E}\left[||\boldsymbol{\Gamma}||\,||\boldsymbol{\xi}_k\boldsymbol{\xi}_k^* - I||_F^2\,\Delta_k^2\right] \\ &= \mathbb{E}\left[||\boldsymbol{\Gamma}||\,\mathrm{tr}\Big((\boldsymbol{\xi}_k\boldsymbol{\xi}_k^* - I)^*(\boldsymbol{\xi}_k\boldsymbol{\xi}_k^* - I)\Big)\Delta_k^2\right] \\ &= ||\boldsymbol{\Gamma}||\,\Delta_k^2\Big(\mathbb{E}\left[||\boldsymbol{\xi}_k||^4\right] - 2\mathbb{E}\left[||\boldsymbol{\xi}_k||^2\right] + n\Big) \\ &= ||\boldsymbol{\Gamma}||\,\Delta_k^2(n^2+n).\end{aligned}$$

where the second inequality follows from the fact that the Frobenius norm of a matrix is larger than its spectral norm. The last equality follows by using Lemma 2.D.2, where n is the number of gains γ_i. Finally, we obtain

$$\mathbb{E}\left[||\lambda_N(t)||^2\right] \leq ||M_0||^2\, c_M^2(t)\,||\boldsymbol{\Gamma}||\,(n^2+n)c_y(t)\sum_{k=0}^{N-1}\Delta_k^2 \underset{N\to\infty}{\longrightarrow} 0,$$

where Assumption 1 and (2.23) are exploited.

2.5.0.2 Mean-Square Convergence of $J_N(t)$

This partial sum is similar to that of $\lambda_N(t)$, and thus we define $F_k(t) := M(t-t_{k+1}) - M(t-t_k)$ and invoke Lemma 2.D.6 again to yield

$$\begin{aligned}\mathbb{E}\left[||J_N(t)||^2\right] &\leq \sum_{k=0}^{N-1} ||F_k(t)||^2 \, \mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^2\right] \mathbb{E}\left[||y_k||^2\right] \\ &\leq c_y(t)\mathrm{tr}\left(\mathbf{\Gamma}\right) \sum_{k=0}^{N-1} ||M(t-t_{k+1}) - M(t-t_k)||^2 \, \Delta_k \\ &\leq c_y(t)\mathrm{tr}\left(\mathbf{\Gamma}\right) \mathbf{\Delta}\mathcal{Q}\mathcal{V}_0^t\left(M\right) \underset{N\to\infty}{\longrightarrow} 0,\end{aligned}$$

where the second inequality follows from (2.23), Lemma 2.D.2 and the fact that $\left|\left|\tilde{\Gamma}_k\right|\right| \leq ||\tilde{\boldsymbol{\gamma}}_k||$ since $\tilde{\Gamma}_k = \mathcal{D}(\tilde{\boldsymbol{\gamma}}_k)$ so that

$$\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^2\right] \leq \mathrm{tr}\left(\mathbf{\Gamma}\right)\Delta_k. \tag{2.28}$$

The last inequality follows from the fact that the quadratic variation of M is finite (Lemma 2.E.1).

2.5.0.3 Mean-Square Convergence of $\nu_N(t)$

By using the same previous definition of $F_k(t)$, invoke Lemma 2.D.5 (with $X_k := \tilde{\Gamma}_k M_0 \tilde{\Gamma}_k$) to yield

$$\begin{aligned}\mathbb{E}\left[||\nu_N(t)||^2\right] &\leq \left(\sum_{k=0}^{N-1} ||F_k(t)|| \left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k M_0 \tilde{\Gamma}_k\right|\right|^2\right] \mathbb{E}\left[||y_k||^2\right]\right)^{\frac{1}{2}}\right)^2 \\ &\leq c_y(t)\,||M_0||^2 \left(\sum_{k=0}^{N-1} ||F_k(t)|| \left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^4\right]\right)^{\frac{1}{2}}\right)^2 \\ &\leq c_y(t)\,||M_0||^2\, c(2,n)\,||\mathbf{\Gamma}||^2 \left(\sum_{k=0}^{N-1} ||M(t-t_{k+1}) - M(t-t_k)||\,\Delta_k\right)^2\end{aligned}$$

$$\leq c_y(t)\,||M_0||^2\,c(2,n)\boldsymbol{\Delta}\Big(\mathcal{TV}_0^t\,(M)\Big)^2 \underset{N\to\infty}{\longrightarrow} 0,$$

where the second inequality follows from (2.23) and the sub-multiplicative property of the spectral norm. The third inequality follows from Lemma 2.D.2 where

$$\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^4\right] \leq c(2,n)\,||\boldsymbol{\Gamma}||^2\,\Delta_k^2, \tag{2.29}$$

and the last inequality follows from the fact that the total variation of M is finite (Lemma 2.E.1).

2.5.0.4 Mean-Square Convergence of $\eta_N(t)$

In a similar fashion to the previous calculation, define $G_k := M(\Delta_k) - M_0$ and invoke Lemma 2.D.5 (with $X_k := \tilde{\Gamma}_k G_k \tilde{\Gamma}_k$) to yield

$$\begin{aligned}
\mathbb{E}\left[||\eta_N(t)||^2\right] &\leq \left(\sum_{k=0}^{N-1} ||M(t-t_k)||\left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k G_k \tilde{\Gamma}_k\right|\right|^2\right]\mathbb{E}\left[||y_k||^2\right]\right)^{\frac{1}{2}}\right)^2 \\
&\leq c_y(t)c_M^2(t)\left(\sum_{k=0}^{N-1} ||M(\Delta_k)-M_0||\left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^4\right]\right)^{\frac{1}{2}}\right)^2 \\
&\leq c_y(t)c_M^2(t)c(4,n)\,||\boldsymbol{\Gamma}||^2\left(\sum_{k=0}^{N-1} ||M(\Delta_k)-M_0||\,\Delta_k\right)^2 \\
&\underset{N\to\infty}{\longrightarrow} 0,
\end{aligned}$$

where the second inequality follows from (2.23), Assumption 1, and the sub-multiplicative property of the spectral norm. Again, the last inequality follows from (2.29). The limit is zero because Assumption 1 guarantees that M is right-continuous at $t = 0$.

2.5.0.5 Mean-Square Convergence of $\chi_N(t)$

Since w and $\{\gamma_i\}$ are uncorrelated (Assumption 4), invoking Lemma 2.D.6 yields

$$\mathbb{E}\left[||\chi_N(t)||^2\right] \leq \sum_{k=0}^{N-1} ||M(t-t_{k+1})||^2\,\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^2\right]\mathbb{E}\left[||M(\Delta_k)\tilde{w}_k||^2\right]$$

$$\leq c_M^4(t)\mathrm{tr}\left(\mathbf{\Gamma}\right)\sum_{k=0}^{N-1}\Delta_k\mathrm{tr}\left(\mathbf{W}_k\right)\Delta_k$$
$$\leq c_M^4(t)\mathrm{tr}\left(\mathbf{\Gamma}\right)c_w(t)\sum_{k=0}^{N-1}\Delta_k^2 \underset{N\to\infty}{\longrightarrow} 0,$$

where the second inequality follows from assumptions 1 and 3 and (2.28). The last inequality follows because under Assumption 3, $\exists$ a continuous scalar function c_w such that

$$\sup_{0\leq\tau\leq t}\mathrm{tr}\left(\mathbf{W}(\tau)\right)=c_w(t). \tag{2.30}$$

2.5.0.6 Mean-Square Convergence of $\theta_N(t)$

By invoking Lemma 2.D.4, we obtain the following inequality

$$\mathbb{E}\left[||\theta_N(t)||^2\right]\leq\sum_{k=0}^{N-1}||M(t-t_{k+1})||^2\left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k M_0\tilde{\Gamma}_k\right|\right|^4\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\sum_{k=0}^{N-1}||\tilde{y}_k||^2\right)^2\right]\right)^{\frac{1}{2}},$$

where the second term converges to $\left(\mathbb{E}\left[\langle y\rangle^2(t)\right]\right)^{\frac{1}{2}}\leq\sqrt{c_q(t)}$ defined in (2.23). Now apply the submultiplicative property of the spectral norm to yield

$$\mathbb{E}\left[||\theta_N(t)||^2\right]\leq\sqrt{c_q(t)}\,||M_0||^2\sum_{k=0}^{N-1}||M(t-t_{k+1})||^2\left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^8\right]\right)^{\frac{1}{2}}$$
$$\leq\sqrt{c_q(t)}\sqrt{c(4,n)}\,||\mathbf{\Gamma}||^2\,c_M^2(t)\,||M_0||^2\sum_{k=0}^{N-1}\Delta_k^2\underset{N\to\infty}{\longrightarrow}0,$$

where the last inequality follows from Assumption 1 and Lemma 2.D.2 where $c(4,n)\,||\mathbf{\Gamma}||^4\,\Delta_k^4$ serves as an upper bound for the eighth moment $\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^8\right]$.

2.5.0.7 Mean-Square Convergence of $T_N^\alpha(t), T_N^\beta(t)$ and $T_N^\zeta(t)$

Observe using (2.27) that the pairs $(\tilde{\Gamma}_k,\alpha_k),(\tilde{\Gamma}_k,\beta_k)$ and $(\tilde{\Gamma}_k,\zeta_k)$ are independent for all $k=0,1,\cdots,N-1$. Then, for $x\in\{\alpha,\beta,\zeta\}$, invoking Lemma 2.D.6 yields

$$\mathbb{E}\left[||T_N^x(t)||^2\right]\leq\sum_{k=0}^{N-1}||M(t-t_{k+1})||^2\,\mathbb{E}\left[\left|\left|\tilde{\Gamma}_k\right|\right|^2\right]\mathbb{E}\left[||x_k||^2\right]$$

$$\leq c_M^2(t)\mathrm{tr}\left(\mathbf{\Gamma}\right)\sum_{k=0}^{N-1}\mathbb{E}\left[||x_k||^2\right]\Delta_k,$$

where the last inequality follows from Assumption 1 and (2.28). Now, we examine $\mathbb{E}\left[||\alpha_k||^2\right]$. Define $F_{k,l} := M(t_{k+1}-t_{l+1}) - M(t_k - t_{l+1})$ and invoke Lemma 2.D.4 to yield

$$\begin{aligned}\mathbb{E}\left[||\alpha_k||^2\right] &\leq \sum_{l=0}^{k-1}||F_{k,l}||^2\left(\mathbb{E}\left[\left|\left|\tilde{\Gamma}_l\right|\right|^4\right]\right)^{\frac{1}{2}}\left(\mathbb{E}\left[\sum_{l=0}^{k-1}||\tilde{y}_l||^2\right]\right)^{\frac{1}{2}}\\ &\leq \sqrt{c(2,n)}\,||\mathbf{\Gamma}||\sqrt{c_q(t)}\sum_{l=0}^{k-1}||F_{k,l}||^2\,\Delta_l\\ &\leq \sqrt{c(2,n)}\,||\mathbf{\Gamma}||\sqrt{c_q(t)}\mathbf{\Delta}\mathcal{QV}_0^t\left(M\right),\end{aligned}$$

where $\mathbf{\Delta} = \sup_l \Delta_l$. Note that the second inequality follows from (2.23) and (2.29), and the third inequality follows by observing that the sum converges to the quadratic variation of M on the interval $[0, t_k]$ (Appendix 2.E). The last equality exploits the fact that $\mathcal{QV}_0^t\left(M\right)$ is an increasing function in t. Substituting in $\mathbb{E}\left[||T_N^\alpha(t)||^2\right]$ yields

$$\begin{aligned}\mathbb{E}\left[||T_N^\alpha(t)||^2\right] &\leq c_M^2(t)\mathrm{tr}\left(\mathbf{\Gamma}\right)\sqrt{c(2,n)}\,||\mathbf{\Gamma}||\sqrt{c_q(t)}\mathbf{\Delta}\mathcal{QV}_0^t\left(M\right)\sum_{k=0}^{N-1}\Delta_k\\ &\leq c_M^2(t)\mathrm{tr}\left(\mathbf{\Gamma}\right)\sqrt{c(2,n)}\,||\mathbf{\Gamma}||\sqrt{c_q(t)}\mathbf{\Delta}\mathcal{QV}_0^t\left(M\right)t \underset{N\to\infty}{\longrightarrow} 0.\end{aligned}$$

Recalling from Appendix 2.C that there is no need to check the convergence of the cross terms, the same arguments used for $\mathbb{E}\left[||T_N^\alpha(t)||^2\right]$ can be used here to show that

$$\mathbb{E}\left[\left|\left|T_N^\beta(t)\right|\right|^2\right] \underset{N\to\infty}{\longrightarrow} 0 \qquad \text{and} \qquad \mathbb{E}\left[\left|\left|T_N^\zeta(t)\right|\right|^2\right] \underset{N\to\infty}{\longrightarrow} 0.$$

This completes the proof of Proposition 2. ■

A direct application of Proposition 2 to (2.16) with $\diamond = \diamond_S$ yields (2.22). This is exactly the result shown in Figure 2.3(b) and given in Theorem 1.

2.6 Loop Gain Operator & MSS Conditions

In this section, we give the mathematical derivations of the LGO (2.17) for the Itō setting. The same analysis can be carried out for the Stratonovich case by using the conversion scheme developed in Section 2.3.1. We first lay down the necessary framework to construct a deterministic block diagram that describes the continuous-time evolution of the covariance matrices of the various signals in the loop (see Figure 2.7). Once this deterministic setting is constructed, the MSS analysis from there onwards resembles that of the discrete-time counterpart in [3].

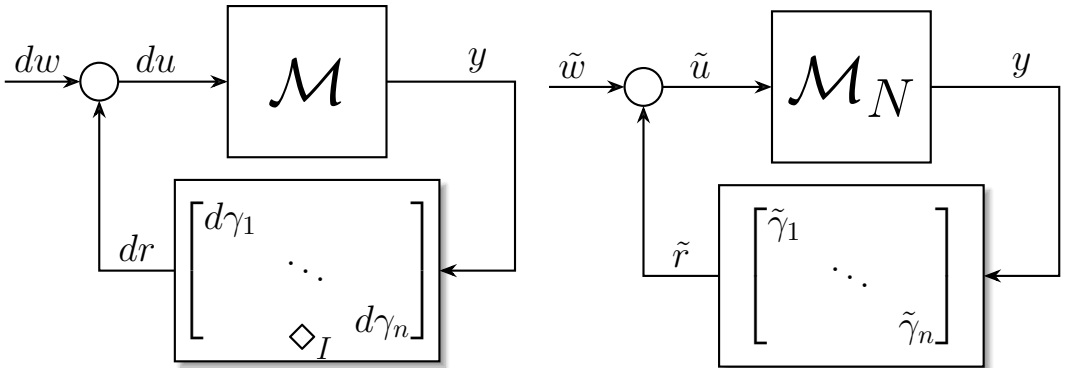
2.6.1 Stochastic Block Diagram Interpretation

Consider the stochastic continuous-time setting depicted in Figure 2.6(a) satisfying assumptions 1-4. It is the same as the general setting in Figure 2.2, but it also indicates an Itō interpretation of the stochastic multiplicative gains. By using the definition of Itō integrals in Section 2.1.11, we construct a discrete-time block diagram, depicted in Figure 2.6(b), which explicitly describes the Itō interpretation of Figure 2.6(a). In fact, it is constructed by using a partition $\mathcal{P}_N[0,t]$ of N subintervals on $[t_0,t_N] := [0,t]$ as described in Section 2.1.11. Therefore, Figure 2.6(a) can be interpreted as the limit of Figure 2.6(b) as $N \to \infty$. Note that $\mathcal{M}_N$ denotes a finite dimensional approximation of $\mathcal{M}$ on the partition $\mathcal{P}_N[0,t]$, i.e.

$$y = \mathcal{M}_N \tilde{u} \Longleftrightarrow y_N = \sum_{k=0}^{N-1} M(t_N - t_k)\tilde{u}_k,$$

where the "tilde" is used to denote the increments of a signal (refer to Section 2.1.11).

The equations describing the block diagrams in Figures 2.6(a) and (b) can be respectively written as



(a) Continuous-Time Setting (b) Discrete-Time Setting

Figure 2.6: A causal LTI system $\mathcal{M}$ in feedback with stochastic multiplicative gains $\{d\gamma_i\}$ that represent the differential forms of, possibly mutually correlated, Wiener processes. Figure (a) shows the continuous-time MSS setting when the Itō interpretation is adopted. Figure (b) explicitly describes the Itō interpretation of Figure (a) by using a partition $\mathcal{P}_N[0,t]$ of N subintervals as explained in 2.1.11. In fact, Figure (a) is interpreted as the limit of Figure (b) as $N \to \infty$.

$$\begin{cases} y(t) = (\mathcal{M}du)(t) \\ du(t) = dw(t) + dr(t) \\ dr(t) = d\Gamma(t) \diamond_I y(t) \end{cases} \qquad \begin{cases} y_N = (\mathcal{M}_N \tilde{u})_N \\ \tilde{u}_N = \tilde{w}_N + \tilde{r}_N \\ \tilde{r}_N = \tilde{\Gamma}_N y_N \end{cases}$$

(2.31a) (2.31b)

The rest of this subsection shows that by adopting the Itō interpretation (2.31b), the stochastic signal r will have independent increments. Furthermore, we will derive the expression that describes the propagation of the instantaneous covariance through the feedback block. The analysis is carried out using Figure 2.6(b) and then is passed to the limit as $N \to \infty$.

2.6.1.1 Disturbance-to-signals mapping

It is fairly straightforward to show that the disturbance $\tilde{w}$ is mapped to the various signals in the loop as

$$\begin{bmatrix} \tilde{u} \\ y \\ \tilde{r} \end{bmatrix} = \begin{bmatrix} (I - \tilde{\Gamma}\mathcal{M}_N)^{-1} \\ (I - \mathcal{M}_N\tilde{\Gamma})^{-1}\mathcal{M}_N \\ (I - \tilde{\Gamma}\mathcal{M}_N)^{-1}\tilde{\Gamma}\mathcal{M}_N \end{bmatrix} \tilde{w}. \tag{2.32}$$

2.6.1.2 Independence of $\big(d\Gamma(t), y(\tau)\big)$ for $\tau \leq t$

This can be shown by analyzing the second equation in (2.32). Examining the operator $(I - \mathcal{M}_N\tilde{\Gamma})^{-1}$ allows us to write it, over the time horizon of the partition $\mathcal{P}_N[0,t]$, as

$$\begin{bmatrix} I & & & \\ -M(t_1-t_0)\tilde{\Gamma}_0 & I & & \\ & \ddots & \ddots & \\ -M(t_N-t_0)\tilde{\Gamma}_0 & \cdots & -M(t_N-t_{N-1})\tilde{\Gamma}_{N-1} & I \end{bmatrix}^{-1} = \begin{bmatrix} I & & \\ & \ddots & \\ * & & I \end{bmatrix},$$

where $*$ denotes the blocks of matrices that are functions of $\tilde{\Gamma}_k$ for $k = 0, 1, ..., N-1$. Hence the second equation in (2.32) can be written as

$$\begin{bmatrix} y_0 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} I & & \\ & \ddots & \\ * & & I \end{bmatrix} \begin{bmatrix} I & & & \\ M(t_1-t_0) & I & & \\ & \ddots & \ddots & \\ M(t_N-t_0) & \cdots & M(t_N-t_{N-1}) & I \end{bmatrix} \begin{bmatrix} \tilde{w}_0 \\ \vdots \\ \tilde{w}_N \end{bmatrix}.$$

Clearly, y_N does not depend on $\tilde{\Gamma}_N$ for any positive integer N. Furthermore, by carrying out a similar reasoning, it is straightforward to see that $\tilde{\Gamma}_N$ is independent of the past values of all the signals in the loop (particularly y). This analysis shows that $(\tilde{\Gamma}_N, y_k)$ are independent for $k \leq N$. Finally, taking the limit as $N \to \infty$ completes the argument.

2.6.1.3 Temporal independence of the increments of r

The following calculation shows that r has independent increments. For $k < l$, we have

$$\mathbb{E}\left[\tilde{r}_k \tilde{r}_l^*\right] = \mathbb{E}\left[\tilde{\Gamma}_k y_k y_l^* \tilde{\Gamma}_l^*\right] = \mathbb{E}\left[\tilde{\Gamma}_k y_k y_l^*\right] \mathbb{E}\left[\tilde{\Gamma}_l^*\right] = 0,$$

where the third equality holds because $\tilde{\Gamma}$ has a zero-mean, and the second equality follows because Γ has independent increments (Wiener process) and also $\tilde{\Gamma}$ is independent of present and past values of y (Section 2.6.1.2).

The combination between the causality of $\mathcal{M}$ and the Itō interpretation introduces a sort of "strict causality" in continuous-time systems. Thus the multiplicative, temporally independent gains $\{d\gamma_i(t)\}$ has a "whitening" effect. In fact, although y has nonzero temporal correlations, the signal r is guaranteed to have independent increments dr, i.e. $\mathbb{E}\left[dr(t)dr^*(\tau)\right] = 0, \ \forall t \neq \tau$.

Finally, the instantaneous covariance of dr is calculated as

$$\begin{aligned}
\mathbb{E}\left[dr(t)dr(t)^*\right] &= \mathbb{E}\left[d\Gamma(t)y(t)y^*(t)d\Gamma^*(t)\right] \\
&= \mathbb{E}\Big[d\Gamma(t)\mathbb{E}\left[y(t)y^*(t)\right]d\Gamma^*(t)\Big] \\
&= \mathbf{\Gamma} \circ \mathbf{Y}(t)dt =: \mathbf{R}(t)dt,
\end{aligned}$$

where the second equality is a consequence of Lemma 2.D.1 since $d\Gamma(t)$ and $y(t)$ are independent (Section 2.6.1.2). The third equality is an immediate consequence of the fact that $d\Gamma(t) = \mathcal{D}\big(d\boldsymbol{\gamma}(t)\big)$. Finally, we have

$$\mathbf{R}(t) = \mathbf{\Gamma} \circ \mathbf{Y}(t). \tag{2.33}$$

2.6.2 Covariance Feedback System

The goal of this section is to construct a deterministic feedback system that describes the evolution of the instantaneous covariance matrices of the various signals in Figure 2.6 and finally derive the expression of the LGO given in (2.17).

In the previous section, we showed that r has temporally independent increments. As a result, it is straightforward to see that u also has temporally independent increments, because for $k < l$ we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{u}_k \tilde{u}_l^*\right] &= \mathbb{E}\left[(\tilde{w}_k + \tilde{r}_k)(\tilde{w}_l + \tilde{r}_l)^*\right] \\
&= \mathbb{E}\left[\tilde{w}_k \tilde{w}_l^*\right] + \mathbb{E}\left[\tilde{r}_k \tilde{r}_l^*\right] + \mathbb{E}\left[\tilde{r}_k \tilde{w}_l\right] + \mathbb{E}\left[\tilde{w}_k \tilde{r}_l^*\right] \\
&= 0 + 0 + 0 + \mathbb{E}\left[\tilde{w}_k y_l^* \tilde{\Gamma}_l^*\right] \\
&= \mathbb{E}\left[\tilde{w}_k y_l^*\right] \mathbb{E}\left[\tilde{\Gamma}_l^*\right] = 0,
\end{aligned}$$

where the third equality follows from the fact that w (Wiener process) and r (Section 2.6.1.3) both have independent increments and the fact that w is independent of past values of all the signals in the loop. The fourth equality follows from Section 2.6.1.2 and the assumption that w and Γ are independent. Finally, passing to the limit as $N \to \infty$ yields that du is temporally independent.

As for the instantaneous covariance of $\tilde{u}$, we have

$$\begin{aligned}
\mathbb{E}\left[\tilde{u}_k \tilde{u}_k^*\right] &= \mathbb{E}\left[\tilde{w}_k \tilde{w}_k^*\right] + \mathbb{E}\left[\tilde{r}_k \tilde{r}_k^*\right] + \mathbb{E}\left[\tilde{r}_k \tilde{w}_k^*\right] + \mathbb{E}\left[\tilde{w}_k \tilde{r}_k^*\right] \\
&= \mathbf{W}_k \Delta_k + \mathbf{R}_k \Delta_k + \mathbb{E}\left[\tilde{\Gamma}_k y_k \tilde{w}_k^*\right] + \mathbb{E}\left[\tilde{w}_k y_k^* \tilde{\Gamma}_k^*\right] \\
&= (\mathbf{W}_k + \mathbf{R}_k)\Delta_k + 0 + 0 =: \mathbf{U}_k \Delta_k.
\end{aligned}$$

Therefore, the addition junction in Figure 2.6 remains as an addition operation on the associated covariance matrices, i.e.

$$\mathbf{U}(t) = \mathbf{W}(t) + \mathbf{R}(t). \tag{2.34}$$

Furthermore, the propagation of the covariance through the forward block of Figure 2.6 is given by (2.11) which requires the input du to be temporally independent for its validity. Finally, the propagation of the covariance through the feedback block is given by (2.33). Therefore, (2.11), (2.33) and (2.34) can be used to construct the deterministic feedback block diagram depicted in Figure 2.7, where each signal is matrix-valued. The advantage

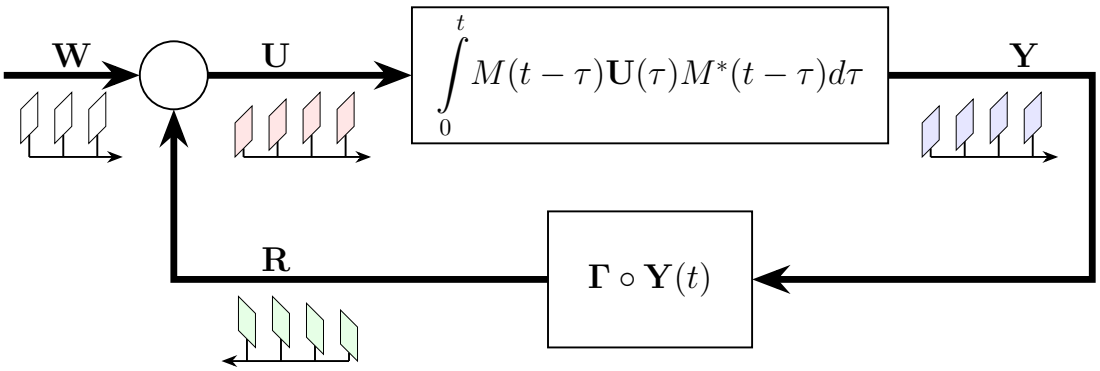


Figure 2.7: A deterministic block diagram describing the evolution of the covariance matrices of the various signals in the feedback loop of Figure 2.6(a). The forward block represents a convolution integral of matrices and the feedback block represents a Hadamard (element-by-element) product. Note that all the covariance matrices in the loop are positive semi-definite and non-decreasing in time when $\mathbf{W}$ is non-decreasing, i.e. for $t_2 \geq t_1$, $\mathbf{W}(t_2) - \mathbf{W}(t_1) \geq 0$ (refer to [3]).

of the covariance feedback system in Figure 2.7 is that it describes a deterministic dynamical system unlike its corresponding stochastic feedback system in Figure 2.6. Before we construct the loop gain operator, we give a remark.

Remark 2.6.1 *All the covariance signals in Figure 2.7 are monotone. Particularly, if* $t_1 \leq t_2$ *then* $\mathbf{U}(t_1) \leq \mathbf{U}(t_2)$*, where the matrix ordering is taken in the usual positive semidefinite sense. Refer to [3, Section II-E].*

2.6.3 Loop Gain Operator

We are now equipped with all the necessary tools to define the continuous-time counterpart of the LGO introduced in [3]. Over a finite time horizon $[0, t]$, the instantaneous covariance $\mathbf{R}(t)$ can be expressed in terms of $\{\mathbf{U}(\tau), 0 \leq \tau \leq t\}$ using (2.11) and (2.33) as

$$\begin{aligned}\mathbf{R}(t) &= \mathbf{\Gamma} \circ \mathbf{Y}(t) \\ &= \mathbf{\Gamma} \circ \left(\int_0^t M(t-s)\mathbf{U}(s)M(t-s)ds \right) \\ \mathbf{R}(t) &= \mathbf{\Gamma} \circ \left(\int_0^t M(\tau)\mathbf{U}(t-\tau)M^*(\tau)d\tau \right). \end{aligned} \tag{2.35}$$

The previous calculation motivates the definition of a finite dimensional linear operator over the infinite time horizon, i.e. as $t \to \infty$

$$\bar{\mathbf{R}} = \mathbb{L}\left(\bar{\mathbf{U}}\right) := \mathbf{\Gamma} \circ \left(\int_0^\infty M(\tau)\bar{\mathbf{U}}M^*(\tau)d\tau \right) \tag{2.36}$$

where $\bar{\mathbf{U}}$ and $\bar{\mathbf{R}}$ are the steady-state limits (if they exist) of the covariances. This linear operator acts on a matrix to produce another matrix, and it propagates the steady state covariance $\bar{\mathbf{U}}$ "once around the loop" to produce the steady state covariance $\bar{\mathbf{R}}$ (and thus the name *loop gain operator*, refer to Figure 2.7). Before moving to the next section, we define here a truncated version of the LGO as

$$\mathbb{L}_T\left(X\right) := \mathbf{\Gamma} \circ \left(\int_0^T M(\tau)XM^*(\tau)d\tau \right), \tag{2.37}$$

which will be useful when proving Theorem 2. Before stating the proof, we summarize some useful properties of the LGO in three remarks.

Remark 2.6.2 *The operator $\mathbb{L}_T$ defined in (2.37) is a monotone operator, i.e. if $0 \leq X \leq Y$, then $0 \leq \mathbb{L}_T(X) \leq \mathbb{L}_T(Y)$. The same property holds for $\mathbb{L}$ defined in (2.36) since $\mathbb{L} = \lim_{T\to\infty} \mathbb{L}_T$. Refer to [3, Section II-E] for details, noting that the same arguments also hold for integrals as well as summations.*

Remark 2.6.3 *The operator $\mathbb{L}_T$ is also monotone in time, i.e. if $T_1 \leq T_2$, then $0 \leq \mathbb{L}_{T_1}(X) \leq \mathbb{L}_{T_2}(X)$ for any $X \geq 0$. This is easy to validate by checking that $\mathbb{L}_{T_2}(X) - \mathbb{L}_{T_1}(X)$ is positive semidefinite. Consequently, for any $T > 0$ and $X \geq 0$, we have $0 \leq \mathbb{L}_T(X) \leq \mathbb{L}(X)$.*

Remark 2.6.4 *The spectral radius of $\mathbb{L}$ is its largest eigenvalue which is guaranteed to be a real number. Furthermore, the "eigen-matrix" associated with the largest eigenvalue is guaranteed to be positive semidefinite. That is, if $\rho(\mathbb{L})$ denotes the spectral radius of $\mathbb{L}$, then $\exists \hat{\mathbf{U}} \geq 0$ s.t. $\mathbb{L}(\hat{\mathbf{U}}) = \rho(\mathbb{L})\hat{\mathbf{U}}$. Note that $\hat{\mathbf{U}}$ is the matrix counterpart of the Perron-Frobenius vector for matrices with nonnegative entries. This is the covariance mode that has the fastest growth rate if MSS is violated, and therefore we refer to $\hat{\mathbf{U}}$ as the worst-case covariance. (Refer to [3, Thm 2.3] for more details.)*

2.6.4 MSS Conditions

Equipped with the LGO, we can now present the proof of Theorem 2. The proof is very similar to [3] and thus some of the details are omitted.

Proof:

"if": Using (2.34) and (2.35), $\mathbf{U}(t)$ can be written as

$$\begin{aligned}\mathbf{U}(t) &= \mathbf{\Gamma} \circ \left(\int_0^t M(\tau)\mathbf{U}(t-\tau)M^*(\tau)d\tau \right) + \mathbf{W}(t) \\ &\leq \mathbf{\Gamma} \circ \left(\int_0^t M(\tau)\mathbf{U}(t)M^*(\tau)d\tau \right) + \mathbf{W}(t) \\ &\leq \mathbb{L}\Big(\mathbf{U}(t)\Big) + \mathbf{W}(t),\end{aligned}$$

where the first inequality follows from Schur's theorem and the fact that $\mathbf{U}(t-\tau) \leq \mathbf{U}(t)$ for all $\tau \in [0, t]$ (Remark 2.6.1). The second inequality follows from Remark 2.6.3. To obtain an upper bound on $\mathbf{U}(t)$, we let $\mathbb{I}$ denote the identity operator and rearrange to

obtain

$$(\mathbb{I}-\mathbb{L})\mathbf{U}(t) \leq \mathbf{W}(t) \leq \bar{\mathbf{W}}$$
$$\mathbf{U}(t) \leq (\mathbb{I}-\mathbb{L})^{-1}\bar{\mathbf{W}},$$

where the second equality is obtained by replacing $\mathbf{W}(t)$ with its steady state value $\bar{\mathbf{W}}$ since it is assumed to be monotone (Assumption 3). The third inequality is obtained by applying [3, Thm 2.3] which guarantees that the operator $(\mathbb{I}-\mathbb{L})^{-1}$ exists and is monotone whenever $\mathbb{L}$ is monotone and $\rho(\mathbb{L}) < 1$. Finally the stability of $\mathcal{M}$ (finite H^2- norm) guarantees that all other covariance signals in the loop of Figure 2.7 are also uniformly bounded thus guaranteeing MSS.

"only if": First it is straightforward to show that MSS is lost if the H^2-norm of M is infinite (regardless of the value of $\rho(\mathbb{L})$). Using Figure 2.7, we can write the covariance $\mathbf{Y}(t)$ as

$$\begin{aligned}\mathbf{Y}(t) &= \int_0^t M(t-\tau)\mathbf{U}(\tau)M^*(t-\tau)d\tau \\ &= \int_0^t M(t-\tau)\Big(\mathbf{W}(\tau)+\mathbf{\Gamma}\circ\mathbf{Y}(\tau)\Big)M^*(t-\tau)d\tau \\ &\geq \int_0^t M(t-\tau)\mathbf{W}(\tau)M^*(t-\tau)d\tau,\end{aligned}$$

where the inequality follows from the fact that $\mathbf{\Gamma}\circ\mathbf{Y}(\tau)$ is positive semidefinite. Thus, clearly $\mathbf{Y}(t)$ grows unboundedly when M has an infinite H^2-norm (take $\mathbf{W}(t) = I$ for example).

Next, assume that M has a finite H^2-norm. We will show that if $\rho(\mathbb{L}) \geq 1$, then $\mathbf{U}(t)$ grows unboundedly in time. We do so by examining $\mathbf{U}(t)$ at the time samples $t_k := kT$, where k is a positive integer and $T > 0$. Using Figure 2.7, we obtain

$$\begin{aligned}\mathbf{U}(t_k) &= \mathbf{\Gamma}\circ\int_0^{t_k} M(t_k-\tau)\mathbf{U}(\tau)M^*(t_k-\tau)d\tau + \mathbf{W}(t_k) \\ &\geq \mathbf{\Gamma}\circ\int_{t_{k-1}}^{t_k} M(t_k-\tau)\mathbf{U}(\tau)M^*(t_k-\tau)d\tau + \mathbf{W}(t_k)\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbf{\Gamma} \circ \int_{t_{k-1}}^{t_k} M(t_k - \tau)\mathbf{U}(t_{k-1})M^*(t_k - \tau)d\tau + \mathbf{W}(t_k) \\
&\geq \mathbf{\Gamma} \circ \int_{0}^{T} M(s)\mathbf{U}(t_{k-1})M^*(s)ds + \mathbf{W}(t_k) \\
&= \mathbb{L}_T\Big(\mathbf{U}(t_{k-1})\Big) + \mathbf{W}(t_k) \\
\mathbf{U}(t_k) &\geq \mathbb{L}_T^k\Big(\mathbf{U}(0)\Big) + \sum_{r=0}^{k-1} \mathbb{L}_T^r\Big(\mathbf{W}(t_{k-r})\Big),
\end{aligned} \tag{2.38}
$$

where the first inequality follows from the fact that the integrand is positive semidefinite, the second inequality follows because $\mathbf{U}(\tau) \geq \mathbf{U}(t_{k-1})$ for $\tau \in [t_{k-1}, t_k]$, and the third inequality is a consequence of applying the change of variable $s := t_k - \tau$. The last inequality is a consequence of a simple induction argument that exploits the monotonicity of $\mathbb{L}_T$ (Remark 2.6.2). Establishing the inequality (2.38) allows us to use the same arguments in [3] (repeated here for completeness) to show that $\mathbf{U}(t_k)$ grows unboundedly.

Set the exogenous covariance $\mathbf{W}(t_k) = \hat{\mathbf{U}}$, where $\hat{\mathbf{U}}$ is the worst-case covariance described in Remark 2.6.4. Note that the initial covariance is $\mathbf{U}_0 = \hat{\mathbf{U}}$. Substituting in (2.38) yields

$$
\mathbf{U}(t_k) \geq \sum_{r=0}^{k} \mathbb{L}_T^r\left(\hat{\mathbf{U}}\right). \tag{2.39}
$$

Since $\lim_{T\to\infty} \mathbb{L}_T(\hat{\mathbf{U}}) = \mathbb{L}(\hat{\mathbf{U}}) = \rho(\mathbb{L})\hat{\mathbf{U}}$, then for any $\epsilon > 0$, $\exists\ T > 0$ such that $||\rho(\mathbb{L})\hat{\mathbf{U}} - \mathbb{L}_T(\hat{\mathbf{U}})|| \leq \epsilon||\hat{\mathbf{U}}||$. This inequality coupled with the fact that $0 \leq \mathbb{L}_T(\hat{\mathbf{U}}) \leq \rho(\mathbb{L})\hat{\mathbf{U}}$ allows us to invoke [3, Lemma A.3] to obtain

$$
\mathbb{L}_T(\hat{\mathbf{U}}) \ \geq\ (\rho(\mathbb{L}) - \epsilon c)\ \hat{\mathbf{U}} \ =:\ \alpha\ \hat{\mathbf{U}}, \tag{2.40}
$$

where c is a positive constant that only depends on $\hat{\mathbf{U}}$. Then, by (2.38), the one-step lower bound (2.40) becomes

$$
\mathbf{U}(t_k) \ \geq\ \left(\sum_{r=0}^{k} \alpha^r\right)\ \hat{\mathbf{U}} \ =\ \frac{\alpha^{k+1} - 1}{\alpha - 1}\ \hat{\mathbf{U}}. \tag{2.41}
$$

First consider the case when $\rho(\mathbb{L}) > 1$, then ϵ can be chosen small enough so that $\alpha > 1$ and therefore $\{\hat{\mathbf{U}}(t_k)\}$ is a geometrically growing sequence. As for the case where $\rho(\mathbb{L}) = 1$, we have $\alpha = 1 - \epsilon$. Then for $0 < \epsilon < 1$, we have

$$\bar{\mathbf{U}} = \lim_{k\to\infty} \mathbf{U}(t_k) \geq \frac{1}{\epsilon}\hat{\mathbf{U}}.$$

This proves that $\mathbf{U}(t)$ can grow arbitrarily large (although not necessarily geometrically) since ϵ can be chosen to be arbitrarily small. ■

2.7 Conclusion

This chapter examines the conditions of MSS for LTI systems in feedback with multiplicative stochastic gains. The analysis is carried out from a purely-input output approach as compared to (the more common) state space approach in the literature. The advantage of this approach is encompassing a wider range of models. It is shown that in the continuous-time setting, technical subtleties arise that require to exploit several tools from stochastic calculus. Different stochastic interpretations are considered for which different stochastic block diagram representations are constructed. Finally, it is shown that MSS analysis for state space realizations can be transparently carried out as a special case of our approach.

Appendix

2.A Interpretations of Stochastic Convolution

Consider the stochastic convolution in (2.10) satisfying Assumption 1. Exploiting the partition $\mathcal{P}_N[0,t]$ described in Section 2.1.9 and the notation developed in Section 2.1.10 yield

$$y(t) = \lim_{N\to\infty} \sum_{k=0}^{N-1} M(t-\bar{t}_k)\tilde{u}_k,$$

where $\bar{t}_k \in [t_k, t_{k+1}]$. The choice of $\bar{t}_k$ prescribes a particular stochastic interpretation of the integral, for example $\bar{t}_k = t_k$ corresponds to an Itō interpretation. The following calculation shows that the covariance of y does not depend on the choice of $\bar{t}_k$ when $M \in \mathcal{C}$ defined in Appendix 2.E.

$$\begin{aligned}
\mathbf{Y}(t) &:= \mathbb{E}\left[y(t)y^*(t)\right] \\
&= \lim_{N\to\infty} \sum_{k,l=0}^{N-1} M(t-\bar{t}_k)\mathbb{E}\left[\tilde{u}_k\tilde{u}_l^*\right] M^*(t-\bar{t}_l) \\
&= \lim_{N\to\infty} \sum_{k=0}^{N-1} M(t-\bar{t}_k)\mathbb{E}\left[\tilde{u}_k\tilde{u}_k^*\right] M^*(t-\bar{t}_k) \\
&= \lim_{N\to\infty} \sum_{k=0}^{N-1} M(t-\bar{t}_k)\mathbf{U}(t_k)\Delta_k M^*(t-\bar{t}_k) \\
&= \int_0^t M(t-\tau)\mathbf{U}(\tau)M^*(t-\tau)d\tau,
\end{aligned}$$

where the third equality follows from the temporal independence of u and the fourth equality follows from the definition of the covariance of du. The last equality is a consequence of Riemann integrability which guarantees convergence to a unique value when $M \in \mathcal{C}$. As a result, there is no need to prescribe a stochastic interpretation of (2.10) since different stochastic interpretations play the same role in the mean-square sense.

2.B Calculation of $D_N(t)$ in (2.25)

This appendix shows the required algebraic manipulations to arrive at the expression of $D_N(t)$ in (2.26). Start by adding and subtracting $M(t-t_k)\tilde{\Gamma}_k y_k$ in the partial sum of $S_N(t)$ in (2.24) to obtain

$$S_N(t) = I_N(t) + \frac{1}{2}\sum_{k=0}^{N-1}\Big(M(t-t_{k+1})\tilde{\Gamma}_k y_{k+1} - M(t-t_k)\tilde{\Gamma}_k y_k\Big),$$

where $I_N(t)$ is defined in (2.24). Adding and subtracting $M(t-t_{k+1})\tilde{\Gamma}_k y_k$ in the sum of the second term yields

$$S_N(t) = I_N(t) + \frac{1}{2}\big(Q_N(t) + J_N(t)\big), \tag{2.B.1}$$

where $J_N(t)$ is given in (2.27) and

$$Q_N(t) := \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k \tilde{y}_k \tag{2.B.2}$$

Observe that $Q_N(t)$ (2.B.2) is a cross quadratic-variation-like term whose limit is not obvious, so we examine the increments $\tilde{y}_k$ using (2.16) with $\diamond = \diamond_S$. We have

$$\begin{aligned}\tilde{y}_k &= E_{k+1}(t_{k+1}) - E_k(t_k) + S_{k+1}(t_{k+1}) - S_k(t_k)\\ \tilde{y}_k &=: \tilde{E}_k + \tilde{I}_k + \frac{1}{2}\left(\tilde{Q}_k + \tilde{J}_k\right).\end{aligned} \tag{2.B.3}$$

where $E_N(t) := \sum_{k=0}^{N-1} M(t-t_k)\tilde{w}_k$. Start by calculating $\tilde{E}_k$

$$\begin{aligned}\tilde{E}_k &= \sum_{l=0}^{k} M(t_{k+1}-t_l)\tilde{w}_l - \sum_{l=0}^{k-1} M(t_k-t_l)\tilde{w}_l \\ &= M(\Delta_k)\tilde{w}_k + \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_l) - M(t_k-t_l)\Big)\tilde{w}_l.\end{aligned}$$

Carrying out similar calculations for $\tilde{I}_k, \tilde{Q}_k$ and $\tilde{J}_k$ yields

$$\begin{aligned}\tilde{I}_k &= M(\Delta_k)\tilde{\Gamma}_k y_k + \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_l) - M(t_k-t_l)\Big)\tilde{\Gamma}_l y_l \\ \tilde{Q}_k &= M_0\tilde{\Gamma}_k\tilde{y}_k + \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_{l+1}) - M(t_k-t_{l+1})\Big)\tilde{\Gamma}_l\tilde{y}_l \\ \tilde{J}_k &= \Big(M_0 - M(\Delta_k)\Big)\tilde{\Gamma}_k y_k + \sum_{l=0}^{k-1}\Big(M(t_{k+1}-t_{l+1}) \\ &\qquad - M(t_k-t_{l+1}) + M(t_k-t_l) - M(t_{k+1}-t_l)\Big)\tilde{\Gamma}_l y_l,\end{aligned}$$

where M_0 denotes $M(0)$ for notational brevity. Substituting for the expression of $\tilde{y}_k$ (2.B.3) in $Q_N(t)$ (2.B.2) and collecting terms yield

$$\begin{aligned}Q_N(t) = \frac{1}{2}\Big(\theta_N(t) + \eta_N(t) + T_N^{\alpha}(t) + T_N^{\beta}(t)\Big) \\ + \chi_N(t) + T_N^{\zeta}(t) + \sum_{k=0}^{N-1} M(t-t_{k+1})\tilde{\Gamma}_k M_0\tilde{\Gamma}_k y_k,\end{aligned}$$

where $\theta_N(t), \eta_N(t), \chi_N(t), T_N^{\alpha}(t), T_N^{\beta}(t)$ and $T_N^{\zeta}(t)$ are all defined in (2.27). Adding and subtracting $M(t-t_k)\tilde{\Gamma}_k M_0\tilde{\Gamma}_k y_k$ in the partial sum of the last term yields

$$\begin{aligned}Q_N(t) = \frac{1}{2}\Big(\theta_N(t) + \eta_N(t) + T_N^{\alpha}(t) + T_N^{\beta}(t)\Big) + \nu_N(t) \\ + \chi_N(t) + T_N^{\zeta}(t) + \sum_{k=0}^{N-1} M(t-t_k)\tilde{\Gamma}_k M_0\tilde{\Gamma}_k y_k,\end{aligned} \tag{2.B.4}$$

where $\nu_N(t)$ is defined in (2.27). Finally, $D_N(t)$ is calculated as

$$D_N(t) := S_N(t) - \Big(I_N(t) + \frac{1}{2}R_N(t)\Big)$$

$$= \frac{1}{2}\Big(Q_N(t) - R_N(t) + J_N(t)\Big). \tag{2.B.5}$$

Substituting for $Q_N(t)$ from (2.B.4), $R_N(t)$ from (2.24), and $J_N(t)$ from (2.27), yields the expression of $D_N(t)$ given in (2.26) after exploiting the following equation

$$\tilde{\Gamma}_k M_0 \tilde{\Gamma}_k - (M_0 \circ \mathbf{\Gamma})\Delta_k = \Big(\tilde{\boldsymbol{\gamma}}_k \tilde{\boldsymbol{\gamma}}_k^* - \mathbf{\Gamma}\Delta_k\Big) \circ M_0,$$

where $\tilde{\boldsymbol{\gamma}}_k = \mathcal{D}(\Gamma_k)$ is the vector formed of the diagonal entries of Γ_k.

2.C Second Moments of Cross Terms

Let x and y be two vector-valued random variables. The subsequent calculation shows that to check if $\mathbb{E}\left[||x+y||^2\right]$ is zero, it suffices to check that $\mathbb{E}\left[||x||^2\right] = \mathbb{E}\left[||y||^2\right] = 0$.

$$\begin{aligned}
\mathbb{E}\left[||x+y||^2\right] &\leq \mathbb{E}\left[(||x|| + ||y||)^2\right] \\
&= \mathbb{E}\left[||x||^2 + ||y||^2 + 2\,||x||\,||y||\right] \\
&\leq \mathbb{E}\left[||x||^2\right] + \mathbb{E}\left[||y||^2\right] + 2\sqrt{\mathbb{E}\left[||x||^2\right]\mathbb{E}\left[||y||^2\right]},
\end{aligned}$$

where the first inequality is a consequence of applying the triangle inequality, and the last one follows from Cauchy-Schwarz inequality with respect to expectations. Observe that if $\mathbb{E}\left[||x||^2\right]$ or $\mathbb{E}\left[||y||^2\right]$ is zero, then the cross term is zero. Therefore, to prove that the variance of the sum of random variables is equal to zero, there is no need to calculate the expectation of cross terms.

2.D Useful Equalities & Inequalities

This appendix provides a sequence of lemmas that give some useful equalities and inequalities (upper bounds) that are used in the proofs throughout this chapter.

Lemma 2.D.1 *Let X and v be a matrix-valued and vector-valued random variables, respectively. If X and v are independent and $D_v := \mathcal{D}(v)$, then*

$$\mathbb{E}\left[D_v X D_v\right] = \mathbb{E}\left[vv^*\right] \circ \mathbb{E}\left[X\right].$$

Proof: Let X_{ij} denote the ij^{th} entry of the matrix X. Then

$$\begin{aligned}\mathbb{E}\left[D_v X D_v\right]_{ij} &= \mathbb{E}\left[v_i X_{ij} v_j\right] = \mathbb{E}\left[v_i v_j\right] \mathbb{E}\left[X_{ij}\right] \\ &= \mathbb{E}\left[vv^*\right]_{ij} \mathbb{E}\left[X\right]_{ij},\end{aligned}$$

where the first equality holds because $D_v := \mathcal{D}(v)$ is diagonal, and the second equality hold because X and v are independent. The proof is complete since the Hadamard product "$\circ$" is the element-by-element multiplication. ■

Lemma 2.D.2 *Let $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^*$ be a zero-mean random vector that follows a multivariate normal distribution with a covariance matrix $\mathbf{\Sigma} := \mathbb{E}\left[xx^*\right]$. Then*

$$\mathbb{E}\left[||x||^2\right] = tr(\mathbf{\Sigma}) \quad \textit{and} \quad \mathbb{E}\left[||x||^{2p}\right] \leq c(p,n)\,||\mathbf{\Sigma}||^p,$$

where p is any positive integer and c is a constant that depends on p and n. For example, one can check that $c(1,n) = n$ and $c(2,n) = n^2 + 2n$.

Proof: For the second moment, we have

$$\mathbb{E}\left[||x||^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[x_i^2\right] = \sum_{i=1}^{n} \mathbf{\Sigma}_{ii} = \operatorname{tr}(\mathbf{\Sigma}).$$

To calculate the fourth moment, let $\mathbf{\Sigma}^{1/2}$ denote the Cholesky factorization of $\mathbf{\Sigma}$ so that $x = \mathbf{\Sigma}^{1/2}\xi$ where ξ follows the standard multivariate normal distribution. Then

$$\begin{aligned}\mathbb{E}\left[||x||^{2p}\right] &= \mathbb{E}\left[\left|\left|\mathbf{\Sigma}^{\frac{1}{2}}\xi\right|\right|^{2p}\right] \leq ||\mathbf{\Sigma}||^p\, \mathbb{E}\left[||\xi||^{2p}\right] \\ &= ||\mathbf{\Sigma}||^p\, \mathbb{E}\left[\left(\sum_{i=1}^{n} \xi_i^2\right)^p\right]\end{aligned}$$

$$
\begin{aligned}
&= ||\mathbf{\Sigma}||^p \, \mathbb{E}\left[\sum_{k_1+k_2+\cdots+k_n=p} p! \prod_{i=1}^{n} \frac{\xi_i^{2k_i}}{k_i!}\right] \\
&= ||\mathbf{\Sigma}||^p \sum_{k_1+k_2+\cdots+k_n=p} p! \prod_{i=1}^{n} \frac{\mathbb{E}\left[\xi_i^{2k_i}\right]}{k_i!} \\
&= ||\mathbf{\Sigma}||^p \, p! \sum_{k_1+k_2+\cdots+k_n=p} \prod_{i=1}^{n} \frac{(2k_i-1)!!}{k_i!} \\
&=: c(p,n) \, ||\mathbf{\Sigma}||^p ,
\end{aligned}
$$

where "!!" is the double factorial operation. The inequality follows from the submultiplicative property of the norms, the third equality is a direct application of the multinomial theorem, and the fourth equality holds because $\{\xi_i\}$ are mutually independent. Finally, the fifth equality follows because the m^{th} moment of a standard normal random variable is $(m-1)!!$ when m is even. ■

Throughout Lemmas 2.D.3-2.D.6, let $\{X_k\}$ and $\{y_k\}$ be two sequences of square random matrices and random vectors, respectively, with bounded second moments. Furthermore, let $\{F_k\}$ be a sequence of deterministic matrices.

Lemma 2.D.3 *Exploiting the triangle inequality and the sub-multiplicative property of the norm yields*

$$
\mathbb{E}\left[\left|\left|\sum_{k=0}^{N-1} F_k X_k y_k\right|\right|^2\right] \leq \mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k|| \, ||X_k|| \, ||y_k||\right)^2\right].
$$

Lemma 2.D.4 *Suppose that (X_k, y_k) are in general dependent, but $\{X_k\}$ has independent increments, i.e. (X_k, X_l) are independent for $k \neq l$. Then*

$$
\mathbb{E}\left[\left|\left|\sum_{k=0}^{N-1} F_k X_k y_k\right|\right|^2\right] \leq \sum_{k=0}^{N-1} ||F_k||^2 \left(\mathbb{E}\left[||X_k||^4\right]\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||y_k||^2\right)^2\right]\right)^{\frac{1}{2}}.
$$

Proof:

$$\begin{aligned}\mathbb{E}\left[\left|\left|\sum_{k=0}^{N-1} F_k X_k y_k\right|\right|^2\right] &\leq \mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k||\,||X_k||\,||y_k||\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_{k=0}^{N-1} ||F_k||^2\,||X_k||^2 \sum_{k=0}^{N-1} ||y_k||^2\right] \\ &\leq \left(\mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k||^2\,||X_k||^2\right)^2\right]\mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||y_k||^2\right)^2\right]\right)^{\frac{1}{2}}\end{aligned}$$

where the first inequality follows from Lemma 2.D.3, the second follows by applying the Cauchy-Schwarz inequality, and the last one follows by applying again the Cauchy-Schwarz inequality but with respect to the expectation. To complete the proof, we find a bound on the first term of the last inequality. We have

$$\begin{aligned}&\mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k||^2\,||X_k||^2\right)^2\right] \\ &\quad= \sum_{k,l=0}^{N-1} ||F_k||^2\,||F_l||^2\,\mathbb{E}\left[||X_k||^2\,||X_l||^2\right] \\ &\quad\leq \sum_{k,l=0}^{N-1} ||F_k||^2\,||F_l||^2\left(\mathbb{E}\left[||X_k||^4\right]\mathbb{E}\left[||X_l||^4\right]\right)^{\frac{1}{2}} \\ &\quad\leq \left(\sum_{k=0}^{N-1} ||F_k||^2\left(\mathbb{E}\left[||X_k||^4\right]\right)^{\frac{1}{2}}\right)^2,\end{aligned}$$

where the first inequality is obtained by using the Cauchy-Schwarz inequality with respect to expectations. Finally, putting the results all together completes the proof. ■

Lemma 2.D.5 *Suppose that (X_k, y_k) are independent for $k = 0, 1, \cdots N-1$. Then*

$$\mathbb{E}\left[\left|\left|\sum_{k=0}^{N-1} F_k X_k y_k\right|\right|^2\right] \leq \left(\sum_{k=0}^{N-1} ||F_k||\left(\mathbb{E}\left[||X_k||^2\right]\mathbb{E}\left[||y_k||^2\right]\right)^{\frac{1}{2}}\right)^2.$$

Proof:

$$\mathbb{E}\left[\left|\left|\sum_{k=0}^{N-1} F_k X_k y_k\right|\right|^2\right] \leq \mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k||\,||X_k||\,||y_k||\right)^2\right]$$

$$
\begin{aligned}
&= \sum_{k,l=0}^{N-1} \mathbb{E}\left[\Big(||F_k||\,||X_k||\,||y_k|| \Big) \Big(||F_l||\,||X_l||\,|\,||y_l|| \Big) \right] \\
&\leq \sum_{k,l=0}^{N-1} ||F_k||\,||F_l|| \left(\mathbb{E}\left[||X_k||^2\,||y_k||^2 \right] \mathbb{E}\left[||X_l||^2\,|\,||y_l||^2 \right] \right)^{\frac{1}{2}} \\
&\leq \left(\sum_{k=0}^{N-1} ||F_k|| \left(\mathbb{E}\left[||X_k||^2 \right] \mathbb{E}\left[||y_k||^2 \right] \right)^{\frac{1}{2}} \right)^2 ,
\end{aligned}
$$

where the first inequality follows from Lemma 2.D.3, the second inequality follows from applying the Cauchy Schwarz inequality with respect to expectations, and the last one is a result of the mutual independence of (X_k, y_k). ■

Lemma 2.D.6 *Suppose that* $\mathbb{E}[X_k] = 0$, $\{X_k\}$ *has independent increments, i.e.* (X_k, X_l) *are independent for* $k \neq l$, *and* (X_k, y_l) *are independent for* $k \geq l$ *with* $k, l = 0, 1, \cdots N-1$. *Then*

$$
\mathbb{E}\left[\left|\left| \sum_{k=0}^{N-1} F_k X_k y_k \right|\right|^2 \right] \leq \sum_{k=0}^{N-1} ||F_k||^2\, \mathbb{E}\left[||X_k||^2 \right] \mathbb{E}\left[||y_k||^2 \right].
$$

Proof:

$$
\begin{aligned}
\mathbb{E}\left[\left|\left| \sum_{k=0}^{N-1} F_k X_k y_k \right|\right|^2 \right] &\leq \mathbb{E}\left[\left(\sum_{k=0}^{N-1} ||F_k||\,||X_k||\,||y_k|| \right)^2 \right] \\
&= \sum_{k=0}^{N-1} ||F_k||^2\, \mathbb{E}\left[||X_k||^2\,||y_k||^2 \right] \\
&\qquad + \sum_{\substack{k,l=0 \\ k<l}}^{N-1} ||F_k||\,||F_l||\, \mathbb{E}\Big[\, ||X_k||\,||y_k||\,||y_l|| \,\Big] \mathbb{E}\Big[\, ||X_l|| \,\Big] \\
&\qquad + \sum_{\substack{k,l=0 \\ k>l}}^{N-1} ||F_k||\,||F_l||\, \mathbb{E}\Big[\, ||y_k||\,||X_l||\,||y_l|| \,\Big] \mathbb{E}\Big[\, ||X_k|| \,\Big] \\
&= \sum_{k}^{N-1} ||F_k||^2\, \mathbb{E}\left[||X_k||^2\,||y_k||^2 \right] + 0 + 0 \\
&= \sum_{k}^{N-1} ||F_k||^2\, \mathbb{E}\left[||X_k||^2 \right] \mathbb{E}\left[||y_k||^2 \right],
\end{aligned}
$$

where the first inequality follows by applying Lemma 2.D.3, and the first equality follows from the independence of (X_k, y_l) when $k > l$ and the fact that X_k has independent increments. The second equality follows because X_k is zero-mean, and the last equality holds because the pair (X_k, y_k) are mutually independent. ■

2.E Total & Quadratic Variations of Deterministic Functions

Let $\mathcal{C}$ denote the class of deterministic, matrix-valued functions M that can be decomposed into two parts $M(t) = C(t) + D(t)$, where $C(t)$ is differentiable and $D(t)$ includes all the jumps (or discontinuities) of M, i.e.

$$M(t) = C(t) + D(t); \quad \text{s.t.} \quad D(t) = \sum_j A_j \mathbb{1}(t - \tau_j), \tag{2.E.1}$$

where $\{A_j\}$ are constant matrices that correspond to the jumps at $\{\tau_j\}$, and $\mathbb{1}(t)$ is the Heaviside step function centered at zero. Note that if M is a scalar function, $\mathcal{C}$ boils down to the class of functions with bounded absolute variations.

Define the total and quadratic variations of $M \in \mathcal{C}$ over the interval $[0, t]$ as

$$\mathcal{TV}_0^t(M) := \lim_{N\to\infty} \sum_{k=0}^{N-1} ||M(t_{k+1}) - M(t_k)||$$
$$\mathcal{QV}_0^t(M) := \lim_{N\to\infty} \sum_{k=0}^{N-1} ||M(t_{k+1}) - M(t_k)||^2,$$

respectively, where $\mathcal{P}_N[0, t]$ (Section 2.1.9) is used to partition the interval $[0, t]$.

Lemma 2.E.1 *If $M \in \mathcal{C}$, then $\mathcal{TV}_0^t(M)$ and $\mathcal{QV}_0^t(M)$ are finite for any finite time t.*

Proof: Since $M \in \mathcal{C}$, we exploit the decomposition in (2.E.1) to write the total variation of M as

$$\mathcal{TV}_0^t(M) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\tilde{C}_k + \tilde{D}_k\right|\right|$$

$$\leq \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\tilde{C}_k\right|\right| + \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\tilde{D}_k\right|\right|$$
$$= \mathcal{TV}_0^t(C) + \mathcal{TV}_0^t(D),$$

where the notation in Section 2.1.10 for the increments is used, i.e. $\tilde{C}_k := C(t_{k+1}) - C(t_k)$. $\mathcal{TV}_0^t(C)$ is shown to be finite by exploiting the fact that C is differentiable, i.e.

$$\mathcal{TV}_0^t(C) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\frac{\tilde{C}_k}{\Delta_k}\right|\right| \Delta_k = \int_0^t \left|\left|\dot{C}(\tau)\right|\right| d\tau.$$

The integral is finite, because C is differentiable and thus $\left|\left|\dot{C}(t)\right|\right|$ is finite for finite time. Furthermore, $\mathcal{TV}_0^t(D)$ is finite because

$$\mathcal{TV}_0^t(D) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\sum_j A_j \Big(\mathbb{1}(t_{k+1} - \tau_j) - \mathbb{1}(t_k - \tau_j)\Big)\right|\right|$$
$$= \sum_j ||A_j||,$$

where the second equality follows from the fact that the increments of the Heaviside step function are zeros everywhere except at the jumps $\{\tau_j\}$. Therefore, $\mathcal{TV}_0^t(M)$ is finite over any bounded interval $[0, t]$ with an upper bound given by

$$\mathcal{TV}_0^t(M) \leq \int_0^t \left|\left|\dot{C}(\tau)\right|\right| d\tau + \sum_j ||A_j||.$$

Similar reasoning can be carried out to show that $\mathcal{QV}_0^t(M)$ is also finite. In fact, using similar arguments we obtain

$$\mathcal{QV}_0^t(M) \leq \mathcal{QV}_0^t(C) + \mathcal{QV}_0^t(D) + 2 \lim_{N\to\infty} \sum_{k=0}^{N-1} \left|\left|\tilde{C}_k\right|\right| \left|\left|\tilde{D}_k\right|\right|$$
$$\leq 0 + \sum_j ||A_j||^2 + 0.$$

■

2.F Second Moment of Quadratic Variations

The goal of this appendix is to show that the second moment of the quadratic variation of the solutions of (2.22) is finite over finite time. For simplicity, we consider the scalar case with $w = 0$, $M_0 = 0$ and $\mathbf{\Gamma} = 1$; however the same analysis can be carried out for the general case. Over the partition $\mathcal{P}_N[0,t]$, (2.22) can be expressed as $y_k = \sum_{l=0}^{k-1} M(t_k - t_l) y_l \tilde{\gamma}_l$ and thus the increments can be written as

$$\tilde{y}_k = M(\Delta_k) y_k \tilde{\gamma}_k + \sum_{l=0}^{k-1} \big(M(t_{k+1} - t_l) - M(t_k - t_l) \big) y_l \tilde{\gamma}_l.$$

Using the inequality $(a+b)^4 \leq 8(a^4 + b^4)$ and the Cauchy Schwarz inequality, we obtain

$$\mathbb{E}\left[\tilde{y}_k^4\right] \leq 8M^4(\Delta_k)\mathbb{E}\left[y_k^4 \tilde{\gamma}_k^4\right] + 8L_M^4 \mathbf{\Delta}^2 t_k^2 \mathbb{E}\left[\left(\sum_{l=0}^{k-1} y_l^2 \tilde{\gamma}_l^2\right)^2\right],$$

where L_M is the Lipschitz constant of M and $\mathbf{\Delta}$ is defined in Section 2.1.9. Using Lemma 2.D.5, $\mathbb{E}\left[\tilde{\gamma}_k^4\right] = 3\Delta_k^2$, and Assumption 1 yield the upper bound $\mathbb{E}\left[\tilde{y}_k^4\right] \leq c(t)\mathbf{\Delta}^2$, where $c(t) = 24\left(c_M^4(t) + L_M^4 t^4\right) \sup_{\tau \leq t} \mathbb{E}\left[y^4(\tau)\right]$. Note that $\sup_{\tau \leq t} \mathbb{E}\left[y^4(\tau)\right]$ is shown to be finite in the corollary of [34, Thm 3.1]. Therefore, using the Cauchy-Schwarz inequality with respect to expectations, the second moment of the quadratic variation over $\mathcal{P}_N[0,t]$ can be bounded as follows

$$\mathbb{E}\left[\left(\sum_{k=0}^{N-1} \tilde{y}_k^2\right)^2\right] \leq \left(\sum_{k=0}^{N-1} \sqrt{\mathbb{E}\left[\tilde{y}_k^4\right]}\right)^2 \leq c(t)\left(\sum_{k=0}^{N-1} \mathbf{\Delta}\right)^2.$$

Finally, taking the limit as $N \to \infty$ shows that $\mathbb{E}\left[\langle y \rangle^2(t)\right]$ is bounded for finite time t.

Part II

Investigating Cochlear Instabilities Using Structured Stochastic Uncertainty

Chapter 3

Introduction & Brief Physiology

The cochlea is a highly sensitive device that is capable of sensing sound waves across a broad spectrum of frequencies ($20 - 20000$ Hz) and across a wide range of sound intensities ranging from 0 dB (threshold of hearing) up to 120 dB (sound of a jet engine). The cochlea was believed to be a passive device that acts like a Fourier analyzer: each frequency causes a vibration at a particular location on the basilar membrane (BM). This mechanism was discovered by the Nobel Prize winner George von Békésy who carried out his experiments on cochleae of human cadavers. However, in 1948, Thomas Gold hypothesized that the ear is rather an active device that has a component termed the cochlear amplifier. Although Gold's hypothesis was rejected by von Békésy, David Kemp validated it thirty years later by measuring emissions from the ear. These emissions, termed otoacoustic emissions (OAEs) are sound waves that are produced by the cochlea and can be measured in the ear canal.

It is widely accepted that the outer hair cells, anchored on the cochlear partition, are responsible for the active gain in the cochlea that produces these emissions. However, the underlying mechanism is still not well understood. For example, spontaneous otoacoustic emissions (SOAEs) – emissions generated in the absence of any stimulus – are studied

in [24] and [39]. The remarkable high sensitivity of the cochlea makes it vulnerable to stochastic perturbations that are believed to be the cause of these emissions. Particularly, in [39], the authors studied the instabilities that arise in a linear biomechanical cochlear model with spatially random active gain profiles that are static in time. In [24], similar analysis was carried out on simplified cochlear models comprised of coupled active nonlinear oscillators. The randomness, or disorder, was introduced via static variations of a bifurcation parameter. In these previous works, the analysis was carried out through Monte Carlo simulations by studying the stability of different randomly generated active gain (or bifurcation) profiles.

In this part of the dissertation, we carry out a *simulation-free* stability analysis of the linearized dynamics of a nonlinear model of the cochlea. Our analysis employs the structured stochastic uncertainty theory developed in the previous chapters rather than Monte Carlo simulations, where the active gain is stochastic in space and time and may have a spatially-varying expectation and/or covariance. It turns out that letting the active gain be a stochastic process puts the model in the standard setting of linear time-invariant (LTI) systems in feedback with a diagonal stochastic process that enters the dynamics multiplicatively (see Figure 4.2). This analysis allows us to predict the locations on the BM where the dynamics are more likely to destabilize due to the underlying uncertainties. It also provides a bound on the variance of the perturbations allowed such that stability is maintained.

The rest of this chapter gives a brief exposé of the physiology of the ear as an adaptive transduction device. For a more thorough reading on the physiology of the ear, we refer the reader to [54].

The primate ear is built to adapt for different sound intensity levels and across the entire audible frequency range (20Hz to 20 kHz). It is composed anatomically, of three principal parts: outer, middle and inner ear (refer to Figure 3.1). The outer ear is

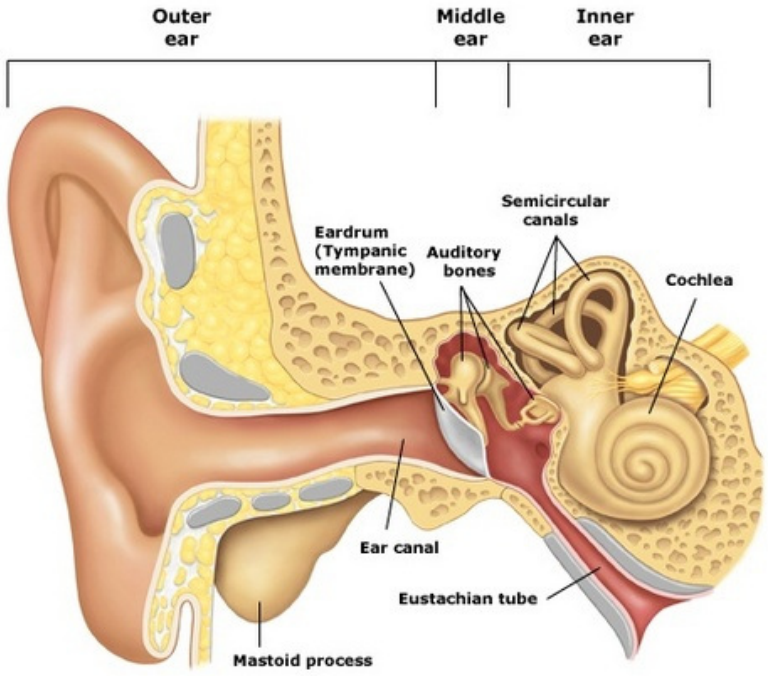


Figure 3.1: Ear Anatomy

mainly composed of the pinna and the external auditory canal. The pinna collects and transforms the sound waves and plays a role in sound source localization. The external auditory canal serves as a filter, which resonates and amplifies tones ranging between 3 and 4 kHz. The middle ear is mainly composed of the ear drum (tympanic membrane), the ossicles and the neighboring cavity. Sound pressure waves pass through the external ear canal and reach the eardrum causing it to vibrate. The neighboring cavity balances the pressure between the middle and outer ear thus preventing eardrum vibrations in the absence of sound waves. Induced eardrum vibrations are then transmitted to the inner ear via three bone structures (ossicles) that collectively act both as an amplifier of the vibration force and as an impedance matching device between the air medium (middle ear) and fluid medium (inner ear) thus preventing excessive energy loss as waves travel

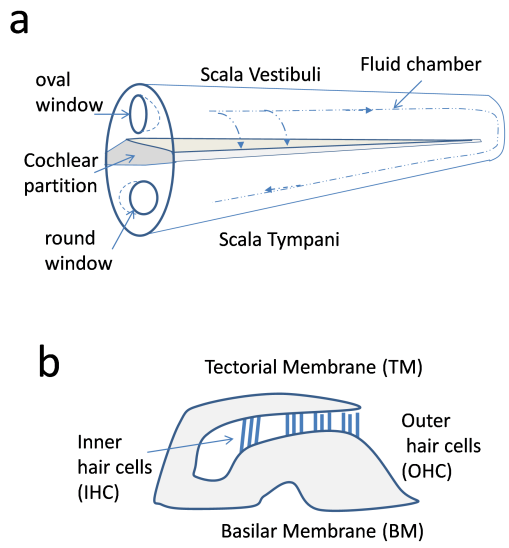


Figure 3.2: (a) Stretched Cochlea (b) Cochlear Partition

between the two different media. In the inner ear, the cochlea is the organ where the main nonlinear biomechanical processing takes place. It is a sensory organ where sound signals are transformed into electrical signals. The cochlea is divided into two chambers: Scala Vestibuli (SV) and Scala Tympani (ST) filled with incompressible fluid and are partly separated by the cochlear partition (refer to Figure). At one end of the SV, the oval window acts as an entry port where pressure waves arriving from the stapes of the middle ear enter the inner ear. These waves travel along the SV and enter the second chamber ST through a connection point (Helicotrema). Finally, a round window at the other end of the ST serves to release pressure traveling in the incompressible fluid. As the pressure waves travel along the two chambers, fluid pressure fluctuations permeate the first wall of the cochlear partition to cause vibrations in two connected wall structures termed the tectorial membrane (TM) and basilar membrane (BM). Anchored in the BM are rows of thin cells termed inner and outer hair cells which are moved as

the two membranes vibrate in different directions. The inner hair cells are the main nerve cells that transduce the mechanical vibrations to electrical impulses. Finally, the outer hair cells act to amplify vibrations specifically under low pressure fluctuations. The mechanical characteristics of the BM varies along its length from being narrow and stiff at the oval window (entry point) to being wide and compliant at the apex. This endows the cochlea with spatially-tuned resonances: lower frequencies cause slow vibrations closer to the apex while higher frequencies are closer to the oval window. Other factors that contribute to cochlear response include dynamics of the fluid and active feedback of the outer hair cells.

Chapter 4

Mean-Square Stability Analysis of the Cochlea

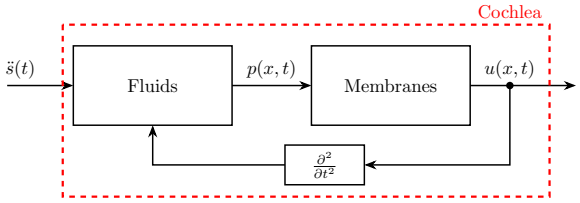
In this chapter, we show how the structured stochastic uncertainty theory developed in the previous chapters can be exploited to analyze the mean-square stability of the cochlea. This chapter is organized as follows: we start by providing a brief description of a class of biomechanical models of the cochlea in section 4.1.1. Then, in section 4.1.2, we recast this class of models in a descriptor state space (DSS) form using operator language (i.e. in continuous space-time). In section 4.2, we reformulate the DSS form in a standard setting that is particularly useful to carry out our stochastic uncertainty analysis. We also provide the conditions for mean-square stability (MSS). We conclude this chapter in section 4.3, where we present the numerical results of the possible instabilities caused by stochastic gain profiles with different statistic properties.

4.1 Biomechanical Model of the Cochlea

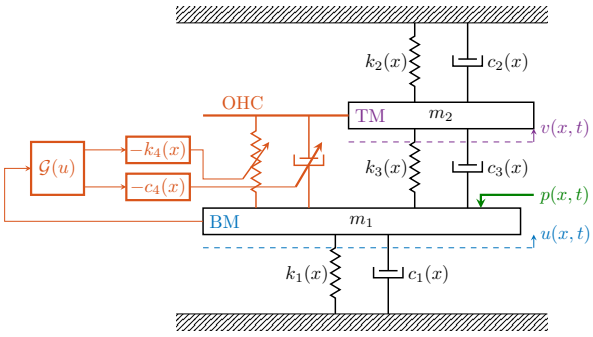
Throughout the literature, cochlear modeling attempts varied depending on two main factors. The first is concerned with the degree of biological realism of the mathematical model. This is realized by the incorporation of various biological structures ([25], [41], [47]) and the dimensionality of the fluid filling the cochlear chambers ([57], [26]). The second factor is concerned with the computational aspect of the models. Different numerical methods were devised to approach the spatio-temporal nature of the cochlea ([46], [20]). Particularly, [19] used a finite difference method developed in [46] to discretize space and formulate the model in state space form. Moreover, computationally efficient methods and model reduction techniques were developed for fast simulations of cochlear response ([7], [22]). This section starts by describing the mathematical model adopted in this dissertation. Then, we reformulate the latter in a continuous space-time descriptor state space form, using operator language. This form has two advantages: (a) it encompasses a wider class of cochlear models and (b) it makes the dynamics more transparent by treating the exact model and its finite dimensional approximation (i.e. discretizing space by some numerical method) separately [23].

4.1.1 Mathematical Model Description

The mathematical model can be divided into two main blocks as illustrated in Figure 4.1(a). For a detailed derivation of the governing mechanics, refer to [19] and [22] for a one and two dimensional modeling of the fluid stage, respectively. The fluid block, commonly referred to as the macro-mechanical stage, is linear and memoryless under the appropriate assumptions and approximations (refer to Appendix-5.A). This block introduces spatial coupling along the different locations on the BM. Its output is the pressure $p(x,t)$ acting on each location of the BM. The governing equation can be written as a



(a) Block Diagram of the Cochlea



(b) Detailed Schematic Representing the Membranes Block

Figure 4.1: (a) The cochlea processes the acceleration of the stapes $\ddot{s}(t)$, in two stages, to produce the vibrations at every location of the BM, $u(x,t)$. The first stage is governed by the fluid that is stimulated by both the stapes and BM accelerations to yield a pressure $p(x,t)$ acting on every location of the BM. The second stage is governed by the dynamics of the membranes. The two stages are in feedback through the BM acceleration. (b) This figure is a schematic of a cross section (at a location x) of the cochlear partition showing the membranes governing the dynamics of the micro-mechanical stage. The spatially varying parameters m_i, $c_i(x)$ and $k_i(x)$ are the mass, damping coefficient and stiffness of the BM and TM for $i = 1$ and 2, respectively. Furthermore, $c_3(x)$ and $k_3(x)$ are the mutual damping coefficient and stiffness, respectively; while $c_4(x)$ and $k_4(x)$ are the damping coefficient and stiffness associated with the active feedback gain from the outer hair cells (OHC) to the BM. The spring and damper between the BM and the OHC have variable negative values to capture the effect of the active force acting only on the BM without any direct effect on the TM. Their values depend on the the BM displacement u via the nonlinear gain $\mathcal{G}(u)$. Equation(4.2) describes the underlying dynamics.

general expression, regardless of the dimensionality of the fluid and the numerical method used, as

$$p(x,t) = -[\mathcal{M}_f \ddot{u}](x,t) - [\mathcal{M}_s \ddot{s}](t), \tag{4.1}$$

where¨represents the second time derivative operation, and $\mathcal{M}_f$ and $\mathcal{M}_s$ are linear spatial operators associated with the fluid and stapes mass, respectively. Refer to Appendix-5.A for a more detailed discussion of these mass operators and their finite dimensional approximations as matrices M_f and M_s, respectively. The second block, commonly referred to as the micro-mechanical stage, takes the distributed pressure $p(x,t)$ as an input to produce the BM vibrations $u(x,t)$ at every location according to the following differential equations

$$\begin{bmatrix} \frac{g}{b}m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} \ddot{u} \\ \ddot{v} \end{bmatrix} + \begin{bmatrix} \frac{g}{b}(c_1+c_3-\mathcal{G}(u)c_4) & \mathcal{G}(u)c_4-c_3 \\ -\frac{g}{b}c_3 & c_2+c_3 \end{bmatrix} \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} + \begin{bmatrix} \frac{g}{b}(k_1+k_3-\mathcal{G}(u)k_4) & \mathcal{G}(u)k_4-k_3 \\ -\frac{g}{b}k_3 & k_2+k_3 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} p \\ 0 \end{bmatrix}, \tag{4.2}$$

where $v(x,t)$ is the tectorial membrane (TM) vibration (refer to Figure 4.1(b)). Note that the space and time variables (x,t) are dropped where necessary for notational compactness. The constant b is the ratio of the average to maximum vibration along the width of the BM, and g is the BM to outer hair cells lever gain. Refer to [47] for a detailed explanation of the parameters. Finally, $\mathcal{G}$ is the nonlinear active gain operator that captures the active nature of the outer hair cells, commonly referred to as the cochlear amplifier. In the spirit of [41], the action of $\mathcal{G}$ on a distributed BM displacement profile u is given by

$$[\mathcal{G}(u)](x,t) = \frac{\gamma(x)}{1+\theta\left[\Phi_\eta\left(\frac{u^2}{R^2}\right)\right](x,t)}, \tag{4.3}$$

where the gain coefficient $\gamma(x)$ represents the gain at a location x, in the absence of any stimulus ($u(x,t)=0$). The constants θ and R are the nonlinear coupling coefficient and BM displacement normalization factor, respectively. The operator Φ_η is a normalized

Gaussian operator such that its action on u is defined as

$$[\Phi_\eta(u)](x,t) := \frac{\int_0^L \phi_\eta(x-\xi)u(\xi,t)d\xi}{\int_0^L \phi_\eta(x-\xi)d\xi}; \tag{4.4}$$

$$\phi_\eta(x) := \frac{1}{\eta\sqrt{2\pi}} e^{\frac{-x^2}{2\eta^2}}, \tag{4.5}$$

where L is the length of the BM and ϕ_η is the Gaussian kernel with a width η. Note that $\eta = 0.5345$ mm corresponds to the equivalent rectangular bandwidth on the BM (refer to Appendix-5.D for a detailed explanation). Observe that the spatial coupling in the micro-mechanical stage appears only in the nonlinear active gain (4.3).

4.1.2 Deterministic Descriptor State Space Formulation of the Linearized Dynamics in Continuous Space-Time

This section gives a Descriptor State Space (DSS) formulation of the cochlear model described in (4.1) and (4.2). The DSS form is given for the linearized dynamics around the only fixed point which is the origin.

It can be shown (Appendix-5.C) that the linearized dynamics can be achieved by simply replacing the nonlinear active gain $[\mathcal{G}(u)](x,t)$ in (4.2) by its gain coefficient $\gamma(x)$. First, define the state space variable $\psi(x,t)$ in continuous space-time as

$$\psi(x,t) := \begin{bmatrix} u(x,t) & v(x,t) & \dot{u}(x,t) & \dot{v}(x,t) \end{bmatrix}^T. \tag{4.6}$$

Then the DSS form of the linearized dynamics is

$$\begin{aligned} \mathcal{E}\frac{\partial}{\partial t}\psi(x,t) &= \mathcal{A}_\gamma \psi(x,t) + \mathcal{B}\ddot{s}(t) \\ u(x,t) &= \mathcal{C}\psi(x,t), \end{aligned} \tag{4.7}$$

where $\mathcal{E}$, $\mathcal{A}_\gamma$ and $\mathcal{B}$ are matrices of linear spatial operators defined as follows

$$\mathcal{E} := \begin{bmatrix} \mathcal{I} & 0 & 0 & 0 \\ 0 & \mathcal{I} & 0 & 0 \\ 0 & 0 & \frac{g}{b}m_1\mathcal{I} + \mathcal{M}_f & 0 \\ 0 & 0 & 0 & m_2\mathcal{I} \end{bmatrix}; \quad \mathcal{B} := \begin{bmatrix} 0 \\ 0 \\ -\mathcal{M}_s \\ 0 \end{bmatrix};$$

$$\mathcal{A}_\gamma := \mathcal{A}_0 + \mathcal{B}_0\gamma\mathcal{C}_0; \qquad \mathcal{C} := \begin{bmatrix} \mathcal{I} & 0 & 0 & 0 \end{bmatrix};$$

$$\mathcal{A}_0 := \begin{bmatrix} 0 & 0 & \mathcal{I} & 0 \\ 0 & 0 & 0 & \mathcal{I} \\ -\frac{g}{b}(k_1+k_3) & k_3 & -\frac{g}{b}(c_1+c_3) & c_3 \\ \frac{g}{b}k_3 & -(k_2+k_3) & \frac{g}{b}c_3 & -(c_2+c_3) \end{bmatrix};$$

$$\mathcal{B}_0 := \begin{bmatrix} 0 & 0 & \mathcal{I} & 0 \end{bmatrix}^T; \qquad \mathcal{C}_0 := \begin{bmatrix} \frac{g}{b}k_4 & -k_4 & \frac{g}{b}c_4 & -c_4 \end{bmatrix};$$

and $\mathcal{I}$ is the identity operator. The equations in (4.7) represent a deterministic evolution differential equation and an output equation that provides the distributed displacement of the BM $u(x,t)$. Other outputs can be selected, such as the TM displacement, by appropriately constructing the $\mathcal{C}$ operator. In the subsequent section, we slightly modify the dynamical equations to account for stochastic perturbations in the gain coefficient $\gamma(x)$.

4.2 Stochastic Uncertainties in the Active Gain

This section investigates the Mean Square Stability (MSS, which we will formally define in section 4.2.1) of the linearized cochlear dynamics when the gain coefficient is a *spatio-temporal stochastic process*. The stochastic gain coefficient, now denoted by $\gamma(x,t)$ to account for spatio-temporal perturbations, enters the dynamics (4.7) multiplicatively. We first reformulate the dynamics as an LTI system in feedback with a diagonal stochastic

gain which is a standard setting in robust control theory [62, Section 10.3]. Then we carry out our MSS analysis based on Chapter 2. By tracking the evolution of the instantaneous spatial covariances, MSS analysis allows us to predict the locations on the BM that are more likely to become unstable due to the underlying stochastic uncertainty. We conclude this section by defining and analyzing a linear operator, whose spectral radius provides a condition for MSS.

4.2.1 Stochastic Feedback Interconnection

The purpose of this section is to separate the stochastic portion of the gain coefficient in a feedback interconnection. We assume that $\gamma(x,t)$ is a *spatio-temporal stochastic process* that is white in time (but may be colored in space), and whose expectation and covariance are independent of time. More precisely, let $\bar{\gamma}(x)$ be the expectation of $\gamma(x,t)$ and $\tilde{\gamma}(x,t)$ be a temporally independent, zero mean stochastic perturbation, such that

$$\begin{aligned} &\gamma(x,t) = \bar{\gamma}(x) + \epsilon\tilde{\gamma}(x,t), \\ &\text{with } \begin{cases} \mathbb{E}[\gamma(x,t)] = \bar{\gamma}(x) \\ \mathbb{E}[\tilde{\gamma}(x,t)\tilde{\gamma}(\xi,\tau)] = \mathbf{\Gamma}(x,\xi)\delta(t-\tau) \end{cases} \forall t \geq 0, \end{aligned} \tag{4.8}$$

where $\mathbb{E}[.]$ denotes the expectation, ϵ is a perturbation parameter, $\delta(t)$ is the Dirac Delta function, and $\mathbf{\Gamma}(x,\xi)$ is a positive semi-definite covariance kernel. Substituting (4.8) in (4.7) yields

$$\begin{aligned} \mathcal{E}\frac{\partial}{\partial t}\psi(x,t) &= (\mathcal{A}_{\bar{\gamma}} + \epsilon\mathcal{B}_0\tilde{\gamma}\mathcal{C}_0)\psi(x,t) + \mathcal{B}\ddot{s}(t) \\ u(x,t) &= \mathcal{C}\psi(x,t). \end{aligned} \tag{4.9}$$

The evolution equation in (4.9) is a Stochastic Partial Differential Equation (SPDE) that is given an Itō interpretation in the time variable. For more details on Itō calculus, refer to [50].

Define a secondary output related to the difference in BM and TM displacements and velocities as

$$y(x,t) := \epsilon \mathcal{C}_0 \psi(x,t). \tag{4.10}$$

Furthermore, define the active feedback pressure resulting from the stochastic perturbations to be

$$p_a(x,t) := \tilde{\gamma}(x,t) y(x,t). \tag{4.11}$$

Therefore, using (4.9), (4.10) and (4.11), construct the feedback block diagram depicted in Figure 4.2. This is a standard setting(Chapter 2, [42], [18]) for structured stochas-

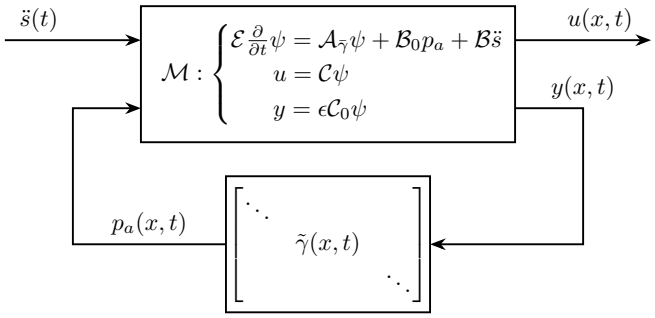


Figure 4.2: The linearized cochlear model in feedback with multiplicative stochastic gain. The block to the top represents the deterministic portion of the linearized cochlear dynamics casted in a descriptor state space form. The feedback block is a diagonal spatial operator that represents the multiplicative stochastic gain. $y(x,t)$ is the differential vibration and velocity between the BM and TM as given by (4.10). $p_a(x,t)$ is the active pressure that results from the stochastic component of the active gain.

tic uncertainty analysis, where the feedback gain is a diagonal spatial operator. This configuration is used to investigate the MSS of the cochlea which is formally defined next.

Definition: The feedback system in Figure 4.2 is MSS if, in the absence of an input (i.e. $\ddot{s}(t) = 0$), the state $\psi(x,t)$ and the active feedback pressure $p_a(x,t)$ have bounded variances for all time.

Therefore, to study MSS, we need to track the temporal evolution of the variances and look at their steady state limits as t goes to $+\infty$. This is the topic of the next subsection.

4.2.2 Temporal Evolution of the Covariance Operators

This section tracks the time evolution of the covariance operators in the absence of any input (i.e. we set $\ddot{s}(t) = 0$ for the rest of the dissertation). We use the term covariance "operators" rather than covariance matrices because the spatial variables x and ξ are continuous. After using some numerical method to discretize space, the covariance operators can be approximated by covariance matrices. With slight abuse of notation, we use the same symbol to denote the covariance operator and its associated kernel. Define the following instantaneous spatial covariance kernels

$$\begin{aligned}
\mathcal{X}(x,\xi;t) &:= \mathbb{E}[\psi(x,t)\psi(\xi,t)] \\
\mathcal{Y}(x,\xi;t) &:= \mathbb{E}[y(x,t)y(\xi,t)] \\
\mathcal{P}(x,\xi;t) &:= \mathbb{E}[p_a(x,t)p_a(\xi,t)] \\
\mathcal{U}(x,\xi;t) &:= \mathbb{E}[u(x,t)u(\xi,t)] \\
\mathbf{\Gamma}(x,\xi) &:= \mathbb{E}[\tilde{\gamma}(x,t)\tilde{\gamma}(\xi,t)] \qquad \forall t \geq 0.
\end{aligned} \tag{4.12}$$

Given that the stochastic perturbations $\tilde{\gamma}$ are temporally independent, it can be shown that the time evolution of the covariance operators are governed by the following operator-valued, *differential algebraic* equations

$$\begin{aligned}
\mathcal{E}\dot{\mathcal{X}}\mathcal{E}^* &= \mathcal{A}_{\tilde{\gamma}}\mathcal{X}\mathcal{E}^* + \mathcal{E}\mathcal{X}\mathcal{A}_{\tilde{\gamma}}^* + \mathcal{B}_0\mathcal{P}\mathcal{B}_0^* \\
\mathcal{Y} &= \epsilon^2\mathcal{C}_0\mathcal{X}\mathcal{C}_0^* \\
\mathcal{P} &= \mathbf{\Gamma} \circ \mathcal{Y},
\end{aligned} \tag{4.13}$$

where $*$ is the adjoint operation and $\circ$ is the Hadamard product; i.e. the element-by-element multiplication of the kernels $\mathcal{P}(x,\xi;t) = \mathbf{\Gamma}(x,\xi)\mathcal{Y}(x,\xi;t)$.

In order to study the MSS, we need to look at the steady state limit of the covariances. We denote by the asymptotic limit of a covariance operator, when it exists, by an overbar. That is

$$\bar{\mathcal{X}} := \lim_{t\to\infty} \mathcal{X}(t); \quad \bar{\mathcal{Y}} := \lim_{t\to\infty} \mathcal{Y}(t); \quad \bar{\mathcal{P}} := \lim_{t\to\infty} \mathcal{P}(t). \tag{4.14}$$

At the steady state, the covariances become constant in time and thus their time derivatives go to zero. Hence, the steady state covariances, if they exist, are governed by the following operator-valued *algebraic* equations:

$$\begin{aligned} &\mathcal{A}_{\bar{\gamma}}\bar{\mathcal{X}}\mathcal{E}^* + \mathcal{E}\bar{\mathcal{X}}\mathcal{A}_{\bar{\gamma}}^* + \mathcal{B}_0\bar{\mathcal{P}}\mathcal{B}_0^* = 0 \\ &\bar{\mathcal{Y}} = \epsilon^2\mathcal{C}_0\bar{\mathcal{X}}\mathcal{C}_0^* \\ &\bar{\mathcal{P}} = \mathbf{\Gamma} \circ \bar{\mathcal{Y}}. \end{aligned} \tag{4.15}$$

In the next section, we will use (4.15) to define a new operator as a tool to check the boundedness of the steady state covariances.

4.2.3 Loop Gain Operator & MSS

Using (4.15), define the loop gain operator $\mathbb{L}_{\mathbf{\Gamma}}$, parametrized by the perturbation covariance $\mathbf{\Gamma}$, as

$$\mathbb{L}_{\mathbf{\Gamma}}(\bar{\mathcal{P}}_{\text{in}}) = \bar{\mathcal{P}}_{\text{out}} \iff \begin{cases} \bar{\mathcal{P}}_{\text{out}} = \mathbf{\Gamma} \circ (\mathcal{C}_0\bar{\mathcal{X}}\mathcal{C}_0^*) \\ \mathcal{A}_{\bar{\gamma}}\bar{\mathcal{X}}\mathcal{E}^* + \mathcal{E}\bar{\mathcal{X}}\mathcal{A}_{\bar{\gamma}}^* + \mathcal{B}_0\bar{\mathcal{P}}_{\text{in}}\mathcal{B}_0^* = 0. \end{cases} \tag{4.16}$$

The MSS condition is given in terms of the spectral radius of the loop gain operator as explained next.

Theorem: *Consider the system in Figure 4.2 where $\tilde{\gamma}$ is a temporally independent multiplicative noise, interpreted in the sense of Itō, with instantaneous spatial covariance $\mathbf{\Gamma}$, and $\mathcal{M}$ is a stable causal LTI system. The feedback system is MSS if and only if the spectral radius of the loop gain operator is strictly less than one, i.e.*

$$\epsilon^2\rho(\mathbb{L}_{\mathbf{\Gamma}}) < 1, \tag{4.17}$$

where $\mathbb{L}_{\mathbf{\Gamma}}$ *is defined in (4.16) and* $\rho(\mathbb{L}_{\mathbf{\Gamma}})$ *is its spectral radius.*

The proof of this theorem is given in Chapter 2. This theorem will be used to find an upper bound on the perturbation constant ϵ above which MSS is violated.

4.2.4 Worst-Case Covariances

The loop gain operator maps a covariance operator $\bar{\mathcal{P}}_{\text{in}}$ into another covariance operator $\bar{\mathcal{P}}_{\text{out}}$. Hence, the eigenvectors of $\mathbb{L}_{\mathbf{\Gamma}}$ are themselves operators. When a finite dimensional approximation of $\mathbb{L}_{\mathbf{\Gamma}}$ is carried out using some numerical method, these eigenvectors can be approximated as matrices. We are particularly interested in the eigenvector (or eigen-operator) of $\mathbb{L}_{\mathbf{\Gamma}}$ associated with the largest eigenvalue because it has a significant meaning explained in this subsection.

First, since the loop gain operator is a monotone operator [3], it is guaranteed to have a real largest eigenvalue equal to $\rho(\mathbb{L}_{\mathbf{\Gamma}})$. It is also guaranteed that the eigen-operator associated with the largest eigenvalue is positive semidefinite, i.e. there exists a positive semidefinite covariance operator $\mathbf{P}$ such that

$$\mathbb{L}_{\mathbf{\Gamma}}(\mathbf{P}) = \rho(\mathbb{L}_{\mathbf{\Gamma}})\mathbf{P}. \tag{4.18}$$

Note that $\mathbf{P}$ is the operator counterpart of the Perron-Frobenius eigenvector for matrices with non-negative entries. Refer to [3, Thm 2.3] for a proof of the aforementioned guarantees. If the stability condition (4.17) is violated, $\mathbf{P}$ will be the covariance mode that has the highest growth rate, hence the name "worst-case" covariance. This provides information about the locations on the BM that are more likely to destabilize due to the stochastic perturbations of the gain. Particularly, since we are interested in the instabilities at the BM, the worst-case covariance of the BM vibrations, denoted by $\mathbf{U}$, can be computed by propagating the worst-case pressure covariance $\mathbf{P}$ through the cochlear

dynamics (at steady state) as follows

$$\begin{aligned} &\mathcal{A}_{\bar{\gamma}}\mathbf{X}\mathcal{E}^* + \mathcal{E}\mathbf{X}\mathcal{A}_{\bar{\gamma}}^* + \mathcal{B}_0\mathbf{P}\mathcal{B}_0^* = 0 \\ &\mathbf{U} = \mathcal{C}\mathbf{X}\mathcal{C}^*, \end{aligned} \tag{4.19}$$

where $\mathbf{X}$ denotes the worst-case covariance operator corresponding to the state space variable ψ.

4.3 Instabilities in Linearized Cochlear Dynamics

This section contains the main results on the effects of stochastic uncertainties on cochlear instabilities. The analysis is carried out for three different scenarios of the perturbation covariance $\mathbf{\Gamma}(x,\xi)$:

- $\mathbf{S}_1$: spatially uncorrelated uncertainties, i.e. $\mathbf{\Gamma}(x,\xi) = \delta(x-\xi)$
- $\mathbf{S}_2$: spatially correlated uncertainties with a correlation length λ, i.e. $\mathbf{\Gamma}(x,\xi) = \phi_\lambda(x-\xi)$
- $\mathbf{S}_3$: spatially localized and uncorrelated uncertainties, i.e. $\mathbf{\Gamma}(x,\xi) = \phi_\sigma(x-\mu)\delta(x-\xi)$,

where ϕ_λ and ϕ_σ are the Gaussian kernels defined in (4.5) such that λ is the spatial correlation length and σ is the spatial localization length. In the subsequent analysis, scenarios $\mathbf{S}_1$ and $\mathbf{S}_2$ are treated simultaneously because, in both cases, the perturbation covariance is a Toeplitz operator since $\mathbf{\Gamma}(x,\xi)$ depends solely on the difference $x-\xi$ rather than the absolute locations x and ξ. However, in scenario $\mathbf{S}_3$, the perturbation covariance is spatially localized and $\mathbf{\Gamma}(x,\xi)$ depends on the absolute locations, and thus it is treated separately in subsection 4.3.4. Recall that the linearized cochlear dynamics excludes micro-mechanical spatial coupling along different locations of the BM; whereas, scenario

$\mathbf{S}_2$ sort of reintroduces spatial coupling via the spatial correlations of the stochastic active gain.

The condition of MSS (4.17) can be rewritten as

$$\epsilon < \frac{1}{\sqrt{\rho(\mathbb{L}_{\mathbf{\Gamma}})}}, \tag{4.20}$$

for scenarios $\mathbf{S}_1, \mathbf{S}_2$ and $\mathbf{S}_3$. This bound is the maximum allowed perturbation in (4.9) such that MSS is maintained. In this section, we compute the upper bound on ϵ and the "worst-case" covariance $\mathbf{U}$ for the linearized cochlear dynamics.

4.3.1 Numerical Considerations

This section describes the numerical considerations of the model and the numerical method used to compute the spectral radius and worst-case covariance of $\mathbb{L}_{\mathbf{\Gamma}}$.

The numerical values of the parameters in this paper are taken from Table I in [39] for the linear cochlea. However, the expectation of the gain coefficient, $\bar{\gamma}(x)$, (which was considered to be spatially constant in [39]) is left as a spatially distributed parameter to be tuned. The fluids block in Figure 4.1(a) considered here is the one dimensional traveling wave as described in Appendix-5.A. A spatial discretization grid of step size $\Delta_x := L/N_x$, where $N_x = 400$, is used to give a finite dimensional approximation of the operators (as matrices) describing the dynamics in Figure 4.2 (refer to Appendix-5.B).

Special care has to be taken when dealing with spatially white continuous processes (Scenario $\mathbf{S}_1$). Let Γ denote a matrix approximation of the uncertainty covariance operator $\mathbf{\Gamma}$ and approximate the Dirac delta function as

$$\begin{aligned} \delta(x) &\approx \frac{1}{\Delta_x}\mathrm{rect}_{\Delta_x}(x) \\ \text{such that,} \quad \mathrm{rect}_{\Delta_x} &:= \begin{cases} 1, & \text{if } -\frac{\Delta_x}{2} \leq x \leq \frac{\Delta_x}{2} \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \tag{4.21}$$

Hence, the finite dimensional approximation of the perturbation covariance needs to be scaled with the discretization step Δ_x as follows

$$\Gamma = \frac{1}{\Delta_x} I, \tag{4.22}$$

where I is the identity matrix.

Furthermore, our analysis requires the computation of the largest eigenvalue of the loop gain operator and its associated eigenvector (or eigen-operator). The matrices that approximate the spatial operators have a size of $(4(N_x + 1) = 1604)$, and keeping track of the underlying sparsity of all the approximated operators is essential for carrying out the computations efficiently. Note that to maintain the sparsity of (4.16) for scenario $\mathbf{S}_2$, we use a truncated Gaussian kernel to approximate ϕ_λ given in (4.5), i.e. $\phi_\lambda(x - \xi) \approx 0$, for $|x - \xi| > d$, where d is a pre-specified constant that represents a compromise between computational accuracy and sparsity. Finally, the power iteration method is employed for eigenvalue and eigenmatrix computations as recommended by [53]. This requires solving the Lyapunov-like equation in (4.16) at each iteration.

4.3.2 Stochastic Gain Coefficient with a Spatially Constant Expectation

In this section, we set the expectation of the gain coefficient to one everywhere along the BM, i.e. $\bar{\gamma}(x) = 1$. To study the effects of the spatial correlations in the gain coefficient, we compare scenarios $\mathbf{S}_1$ and $\mathbf{S}_2$ by keeping in mind that $\mathbf{S}_1$ can be seen as a special case of $\mathbf{S}_2$ at the limit when λ goes to zero. First, we compute the upper bounds on ϵ in (4.20) such that MSS is maintained. Then we compute the worst-case covariance $\mathbf{U}$ in (4.19).

By applying the power iteration method on (4.18), we compute the spectral radii $\rho(\mathbb{L}_{\mathbf{\Gamma}})$ and their associated eigen-operators $\mathbf{P}$ for scenarios $\mathbf{S}_1$ and $\mathbf{S}_2$ with different cor-

relation lengths λ. Then, (4.20) yields the upper bounds on ϵ. The results are illustrated in Figure 4.3 showing the small upper bounds on ϵ. This reflects the high sensitivity of the model to such stochastic perturbations. As one would expect, a larger correlation length λ requires a larger perturbation to destabilize the linearized cochlea.

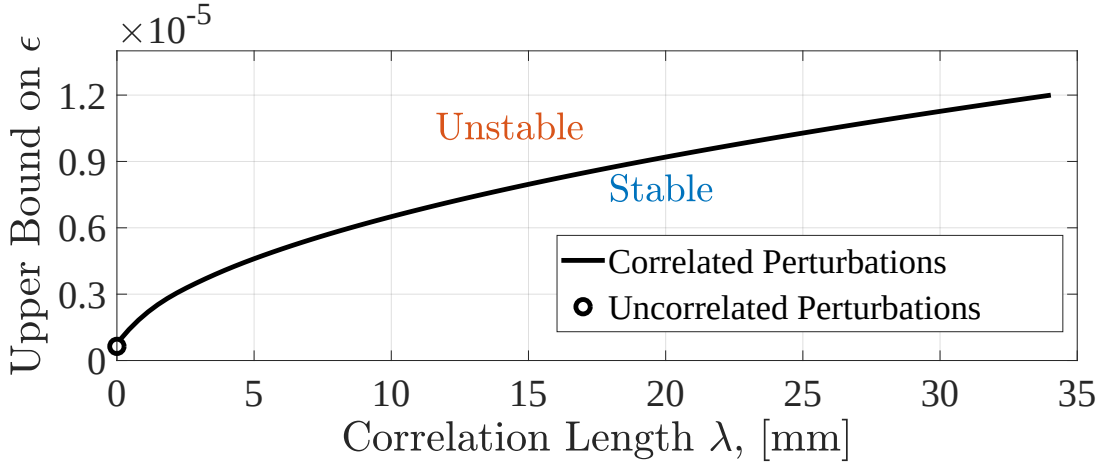


Figure 4.3: Mean Square Stability Curve: Upper bound on the perturbation parameter, ϵ, of the stochastic gain (4.8) whose expectation is $\bar{\gamma}(x) = 1$. The black dot corresponds to scenario $\mathbf{S}_1$ (uncorrelated gain perturbations) and the solid black line corresponds to scenario $\mathbf{S}_2$ (correlated gain perturbations) for different spatial correlation lengths λ. The figure shows that larger correlation lengths make the model more immune to stochastic perturbations.

The eigen-operator $\mathbf{P}$ computed by the power iteration method is the worst-case pressure covariance. The corresponding worst-case covariance of the BM displacement $\mathbf{U}$ is then computed using (4.19). Figure 4.4(a) shows $\mathbf{U}$ for scenario $\mathbf{S}_1$, zoomed in for $0 \leq x, \xi \leq L/10$. The intensity plot shows two sets of axes. The first axis represents the location on the BM and the second represents the corresponding characteristic frequency at each location, calculated using the Greenwood location-to-frequency mapping [27]. Observe that the covariance is band limited and the diagonal entries are dominant near the stapes ($x = 0$). This shows that instabilities essentially occur at high frequencies. Figure 4.4(b) plots the diagonal entries of $\mathbf{U}$ for scenarios $\mathbf{S}_1$ and $\mathbf{S}_2$ for different correla-

tion lengths λ. A smaller correlation length gives a slightly broader spectrum of unstable frequencies. However, for small ϵ, the effect of the correlation length on the shape of the unstable BM modes is negligible. This is illustrated in Figure 4.4(c), where the dominant eigenfunction of $\mathbf{U}$ is plotted for different cases.

4.3.3 Stochastic Gain Coefficient with a Spatially Varying Expectation

This section shows that the frequencies of instabilities (or, equivalently, the locations on the BM) can shift depending on the shape of the expectation of the gain coefficient $\bar{\gamma}(x)$. For illustration purposes, four different profiles of $\bar{\gamma}_0(x)$ are generated as

$$\bar{\gamma}_0(x) = \frac{\tanh(x/10) + \beta}{\tanh(L/10) + \beta}, \tag{4.23}$$

where x and L are expressed in mm and $\beta = 0, 2, 4$ and 6. First, we show the MSS curves, similar to Figure 4.3 for the four different profiles generated using (4.23). Figure 4.5(b) clearly shows that the shape of $\bar{\gamma}(x)$ affects the margin of MSS. Particularly, the larger the dip in the gain coefficient, the higher ϵ needs to be to destabilize the linearized dynamics in the MSS sense.

Since the correlation length for small values of ϵ has a negligible effect on the shape of the unstable modes as shown in Figure 4.4(c), we only present the worst-case covariances for scenario $\mathbf{S}_1$. In fact, the correlation length only affects the margin of stability as illustrated in Figure 4.5(b). Figure 4.5(c) depicts the dominant eigenfunctions of $\mathbf{U}$ for the four different profiles of $\bar{\gamma}(x)$. Clearly, the peaks of the unstable modes of the BM shift depending on the shape of $\bar{\gamma}(x)$. In fact, as the dip in $\bar{\gamma}(x)$ is increased, the peaks shift farther from the stapes resulting in instabilities of lower frequencies.

4.3.4 Stochastic Gain Coefficient with a Spatially Localized Covariance

We now treat the case where the gain coefficient $\gamma(x,t)$ in (4.9) has a spatially constant expectation, but spatially localized covariance given in scenario $\mathbf{S}_3$, i.e.

$$\bar{\gamma}(x) = 1 \qquad \text{and} \qquad \mathbf{\Gamma}(x,\xi) = \phi_\sigma(x-\mu)\delta(x-\xi),$$

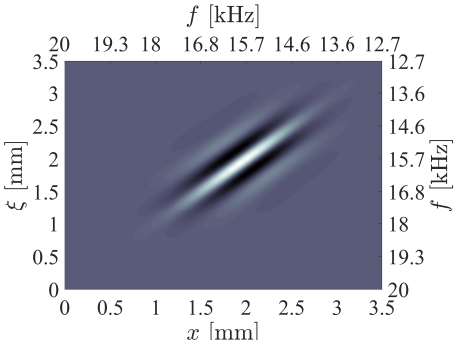
for different values of σ and μ. Observe that for this form of $\mathbf{\Gamma}(x,\xi)$, the covariance is localized around μ. Hence, this section investigates the cochlear instabilities that emerge as a result of stochastic perturbations localized around a particular location on the BM.

In particular, we are interested in tracking the unstable BM modes for different values of μ and σ, where μ is the location of the perturbation and σ represents the local spread of the perturbation in the neighborhood of μ. Following the same calculations of the previous sections, we compute the dominant eigenfunction of the worst-case covariance of the BM displacement $\mathbf{U}$ for different values of μ and σ. The results are depicted in Figure 4.6. Observe that localized perturbations of the active gain coefficient at some location μ of the BM causes instabilities in that neighborhood. Particularly, for relatively small spread $\sigma = L/100$, the instabilities emerge at the same locations of the perturbations as shown in Figure 4.6(a). However, as the spread of the uncertainty is increased up to $\sigma = L/30$ and $L/10$, the location of the instability shifts towards the stapes. In fact, the wider the spread the larger the shift is as illustrated in Figures 4.6(b) and (c).

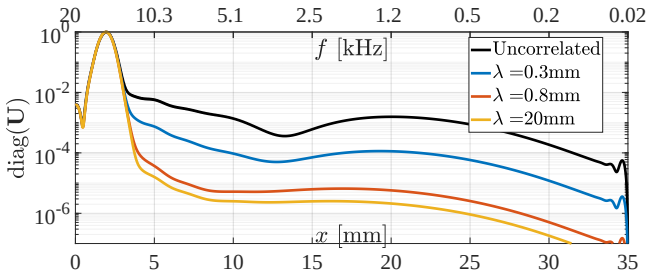
This "basal shifting" resembles the phenomenon of detuning observed in the cochlea. Acting as a frequency analyzer (or "inverse-piano"), each location on the BM vibrates in response to a sound stimulus at a particular frequency. Thus, the BM has a frequency-to-location map such that every stimulus frequency has a preferred place on the BM called Characteristic Place (CP). The detuning phenomenon is observed as the shifting

of the CP towards the stapes as the intensity of the stimulus (in dB) is increased. In this section, we showed that increasing the spread of the stochastic perturbations also shifts the BM vibrations towards the stapes. Nonlinear dynamics are necessary to model the detuning phenomenon. However, modeling this "detuning-like" phenomenon doesn't require nonlinearities, instead a locally perturbed active gain is sufficient to explain it.

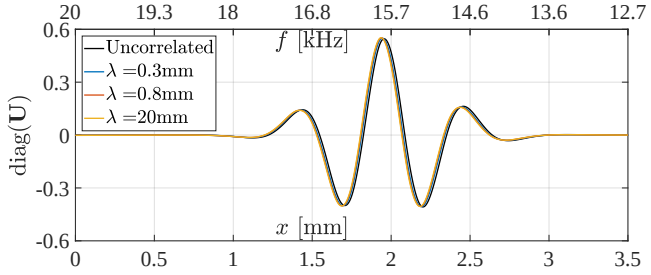
It is believed that these instabilities in the BM reflect back to the middle ear causing SOAEs [49]. It is also believed that if these BM vibrations are intense enough, they can be perceived as tinnitus. Our results suggest a mechanism that explains the frequencies that can be detected in the ear canal due to SOAEs and/or perceived as tinnitus. As a matter of fact, the shape of the statistics (expectation and covariance) of the gain coefficient is a factor that controls the bands of the frequencies that are emitted as SOAEs. These emissions arise due to (a) *spatially variant inhomogeneities* along the cochlear partition and (b) *temporal stochastic perturbations* that give rise to structured stochastic uncertainties.



(a) Worst-Case Covariance of BM Displacement $\mathbf{U}(x, \xi)$

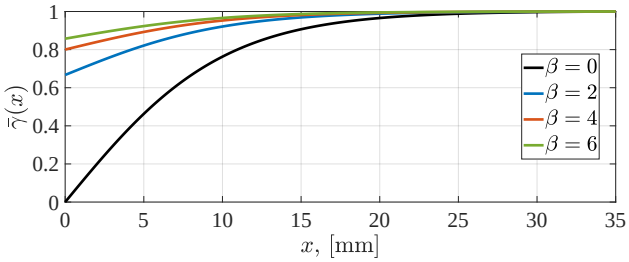


(b) Diagonal Entries of $\mathbf{U}$

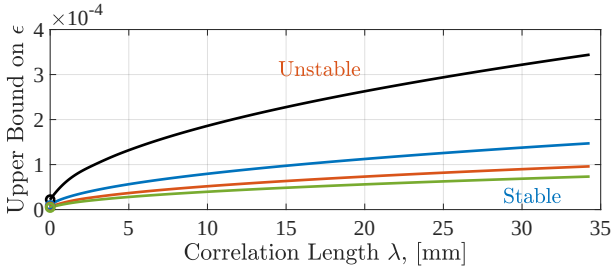


(c) Dominant Eigenfunction of $\mathbf{U}$

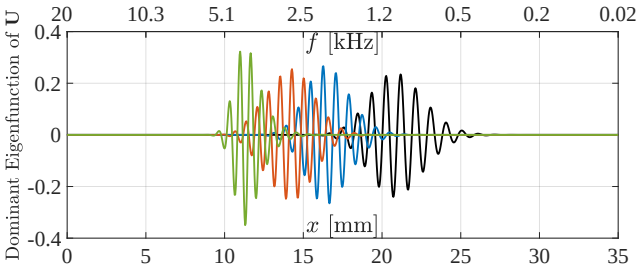
Figure 4.4: Figure (a) shows an intensity plot of the worst-case covariance $\mathbf{U}$ for scenario $\mathbf{S}_1$ (uncorrelated gain perturbation) zoomed in for $0 \leq x, \xi \leq 3.5$ mm. The axes correspond to the physical location x in mm on the BM and the corresponding characteristic frequency f in kHz. Figure (b) shows the diagonal entries of $\mathbf{U}$ for scenarios $\mathbf{S}_1$ and $\mathbf{S}_2$ for different correlation lengths λ. Figure (c) depicts the dominant eigenfunction of $\mathbf{U}$ for the different cases indicating the insignificant effect of λ on the shape of the dominant eigenfunctions.



(a) Gain Coefficient Expectation Profiles

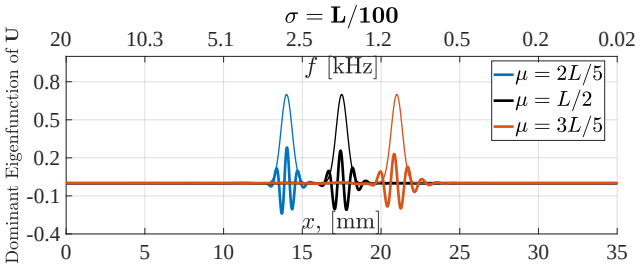


(b) Corresponding MSS Curves

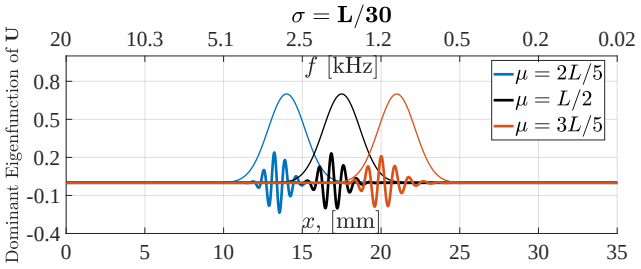


(c) Eigenfunctions for scenario $\mathbf{S}_1$

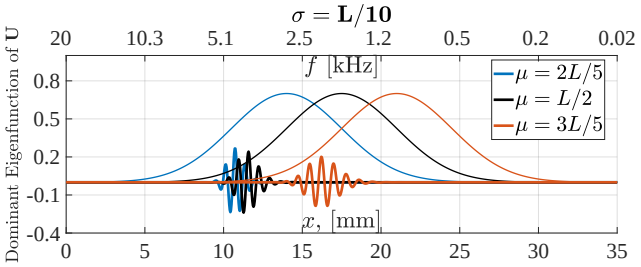
Figure 4.5: Mean Square Stability Curves for different gain coefficient expectation profiles: Figure(a) shows four different profiles of $\bar{\gamma}(x)$ generated as examples of spatially varying gain coefficients using (4.23). The same values of β are used in figures (b) and (c). Particularly, Figure(b) shows the upper bound on the perturbation parameter ϵ for the corresponding profiles of $\bar{\gamma}(x)$ in Figure(a). The circles correspond to scenario $\mathbf{S}_1$ (uncorrelated gain perturbations) and the solid lines correspond to scenario $\mathbf{S}_2$ (correlated gain perturbations) for different spatial correlation lengths λ. Figure (c) shows the eigenfunctions of the worst-case covariance operator $\mathbf{U}$ corresponding to the different profiles of $\bar{\gamma}(x)$. The peaks of the eigenfunctions shift consistently with the shape of the gain profiles.



(a)



(b)



(c)

Figure 4.6: Eigenfunctions of the worst-case covariance operator $\mathbf{U}$ for different localized gain coefficient perturbations. These figures show the dominant eigenfunctions of the worst-case covariance operators for three different values of μ and σ. Particularly, in each figure, we fix σ and vary μ. Each thin curve represents a particular uncertainty spread function $\phi_\sigma(x-\mu)$ (not drawn to scale in the vertical axis) and each thick curve (with the same color) represents the corresponding dominant eigenfunction of the worst-case covariance operator. This figure illustrates the "basal shifting" observation that resembles the phenomenon of detuning.

Chapter 5

Nonlinear Stochastic Simulation of the Cochlea

In the previous chapter, the MSS analysis is carried out on the linearized dynamics. In this chapter, we carry out stochastic simulations of the nonlinear model to validate the predictions of our analysis of the linearized dynamics.

5.1 Nonlinear Descriptor State Space Formulation in Continuous Space-Time

We first start by formulating the nonlinear dynamics in a DSS form similar to that given in section 4.1.2. Recall that, the nonlinear deterministic active gain is given by (4.3) with $\gamma(x)$ representing the gain coefficient. To include stochastic perturbations, we substitute (4.8) in (4.3) so that the nonlinear stochastic active gain can be written as

$$\begin{aligned}\left[\mathcal{G}(u)\right](x,t) &= \frac{\bar{\gamma}(x)+\epsilon\tilde{\gamma}(x,t)}{1+\theta\left[\Phi_\eta\left(\frac{u^2}{R^2}\right)\right](x,t)} \\ &=: \left(\bar{\gamma}(x)+\epsilon\tilde{\gamma}(x,t)\right)\left[\tilde{\mathcal{G}}(u)\right](x,t).\end{aligned} \tag{5.1}$$

Recall that Φ_η is the Gaussian spatial operator given by (4.4), $\theta = 0.5$, $R = 1$ nm and $\eta = 0.5345$ mm. By substituting (5.1) in (4.2), we can rewrite the nonlinear model in a nonlinear DSS form as

$$\mathcal{E}\frac{\partial}{\partial t}\psi(x,t) = \big(\mathcal{A}_{\bar{\gamma}}(u) + \epsilon\tilde{\mathcal{B}}_0(u)\tilde{\gamma}\mathcal{C}_0\big)\psi(x,t), \tag{5.2}$$

where $\mathcal{A}_{\bar{\gamma}}(u) := \mathcal{A}_0 + \tilde{\mathcal{B}}_\prime(u)\bar{\gamma}\mathcal{C}_0$ and $\tilde{\mathcal{B}}_0(u)\tilde{\gamma}\mathcal{C}_0$ are nonlinear spatial operators that represent the deterministic and stochastic portions of the dynamics, respectively. Note that $\mathcal{E}, \mathcal{A}_0$, and $\mathcal{C}_0$ are all defined in (4.7), and $\tilde{\mathcal{B}}_0(u) = \begin{bmatrix} 0 & 0 & \tilde{\mathcal{G}}(u) & 0 \end{bmatrix}^T$. Therefore, (5.2) represents the nonlinear stochastic dynamics given in a DSS operator form, where the spatial variable is continuous. This is really a Stochastic Partial Differential Equation (SPDE) that needs to be discretized in space and time in order to carry out our simulations.

5.2 Description of the Numerical Method for Simulations

In this section, we discretize (5.2) in space and time so that numerical simulations become fairly straightforward to implement. On a side note, if the stochastic perturbation $\tilde{\gamma} = 0$, (5.2) becomes a deterministic Partial Differential Equation (PDE). This can be easily integrated by discretizing space using a spatial grid, and then employ a time marching solver such as ODE45 in MATLAB. However, for an SPDE, one has to carefully treat the scaling of the covariances with the discretization steps.

Space and time are discretized as $x_i = i\Delta_x$ and $t_n = n\Delta_t$ with discretization steps $\Delta_x = L/N_x$ and $\Delta_t = t_f/N_t$ for $i = 0, 1, ..., N_x$ and $n = 0, 1, ..., N_t$, where t_f is the final time. Let the BM and TM displacements on the discretized space-time grid be denoted

by the vectors u_n and $v_n \in \mathbb{R}^{N_x+1}$, respectively such that

$$u_n := \begin{bmatrix} u(x_0, t_n) & \cdots & u(x_{N_x}, t_n) \end{bmatrix}^T$$
$$v_n := \begin{bmatrix} v(x_0, t_n) & \cdots & v(x_{N_x}, t_n) \end{bmatrix}^T.$$

Then the discretized state space variable can be expressed by $\psi_n \in \mathbb{R}^{4(N_x+1)}$ as

$$\psi_n := \begin{bmatrix} u_n^T & v_n^T & \dot{u}_n^T & \dot{v}_n^T \end{bmatrix}^T.$$

For scenarios $\mathbf{S}_1$ and $\mathbf{S}_3$, $\tilde{\gamma}(x,t)$ is a zero-mean white process in space and time. It can be approximated at the spatial grid points $\{x_i\}_{i=0,1,\ldots,N_x}$ and at time t_n as follows

$$\begin{bmatrix} \tilde{\gamma}(x_0, t_n) & \tilde{\gamma}(x_1, t_n) & \cdots & \tilde{\gamma}(x_{N_x}, t_n) \end{bmatrix}^T \approx \frac{1}{\sqrt{\Delta_x \Delta_t}} w_n,$$

where $w_n \in \mathbb{R}^{N_x+1}$ is a zero-mean Gaussian random vector with a covariance matrix $\mathbb{E}\left[w_n w_n^T\right] = I$ for $\mathbf{S}_1$ and $\mathbb{E}\left[w_n w_n^T\right] = \mathcal{D}\left(\begin{bmatrix} \phi_\sigma(x_0 - \mu) & \cdots & \phi_\sigma(x_{N_x} - \mu) \end{bmatrix}\right)$ for $\mathbf{S}_3$, where $\mathcal{D}$ is the diagonal operator such that $\mathcal{D}(w_n)$ is a diagonal matrix with w_n arranged on its diagonal entries.

For scenario $\mathbf{S}_2$, $\tilde{\gamma}(x,t)$ is a stochastic process that is white in time but "colored" in space with a spatial covariance $\mathbf{\Gamma}(x,\xi) = \epsilon^2 \phi_\lambda(x - \xi)$. In this scenario, the noise is smooth in space and there is no need to scale the covariance by the spatial discretization step. More precisely, $\tilde{\gamma}(x,t)$ can be approximated as

$$\begin{bmatrix} \tilde{\gamma}(x_0, t_n) & \tilde{\gamma}(x_1, t_n) & \cdots & \tilde{\gamma}(x_{N_x}, t_n) \end{bmatrix}^T \approx \frac{1}{\sqrt{\Delta_t}} w_n,$$

where $\mathbb{E}\left[w_n w_n^T\right]$ is now a symmetric matrix whose $(i,j)^{\text{th}}$ entry is given by $\phi_\lambda(x_i - x_j)$.

Therefore, a first order approximation of (5.2) can be carried out in the spirit of the Euler-Maruyama method [43] to obtain

$$E\psi_{n+1} = E\psi_n + \Delta_t A_{\tilde{\gamma}}(u_n)\psi_n + \alpha \tilde{B}_0(u_n)\mathcal{D}(w_n)C_0\psi_n \tag{5.3}$$

where $\alpha = \epsilon\sqrt{\Delta_t/\Delta_x}$ for $\mathbf{S}_1$ and $\mathbf{S}_3$; and $\alpha = \epsilon\sqrt{\Delta_t}$ for $\mathbf{S}_2$. The matrices $E, A_{\bar{\gamma}}(u_n), \tilde{B}_0(u_n)$ and C_0 are all finite dimensional approximations of the operators $\mathcal{E}, \mathcal{A}_{\bar{\gamma}}(u), \tilde{\mathcal{B}}_0(u)$ and $\mathcal{C}_0$, respectively (Appendix-5.B). Equation (5.3) represents the recursive numerical methods to solve (5.2) for all three scenarios with the right choice of α and $\mathbb{E}[w_n w_n^T]$.

5.3 Simulation of the Nonlinear Stochastic Model

To validate our MSS analysis of the linearized dynamics and evaluate how well it copes with the nonlinear dynamics, we carry out a simulation of (5.2). This section considers scenario $\mathbf{S}_1$. Hence, the numerical method used here is that given in (5.3) for $\alpha = \epsilon^2\sqrt{\Delta_t/\Delta_x}$ and $\mathbb{E}[w_n w_n^T] = I$.

The nonlinear stochastic simulation shown here is for $\bar{\gamma}(x)$ given in (4.23) with $\beta = 2$. All other scenarios are in agreement with our MSS analysis; however, this particular case study ($\beta = 2$) is chosen here to illustrate the effectiveness of our analysis. Observe using Figure 4.5(b) that for $\beta = 2$, the MSS condition is violated if $\epsilon \geq 9.1 \times 10^{-6}$. We choose $\epsilon = 1.1 \times 10^{-5}$ which slightly violates the MSS condition for the linearized dynamics and allows the nonlinearity to kick in and saturate the response. The spatio-temporal response of the BM is depicted in Figure 5.1(a) for $t \in [0, t_f]$ with $t_f = 200$ ms. The response is maximal in a band limited region 10 mm $< x <$ 20 mm which corresponds to a frequency range of roughly between 1 kHz and 5 kHz. To be more precise, we compute the empirical covariance $\mathbf{U}_{\text{Emp}}(x, \xi)$ as follows

$$\mathbf{U}_{\text{Emp}}(x, \xi) = \frac{1}{t_f}\int_0^{t_f} u(x, \tau)u(\xi, \tau)d\tau. \tag{5.4}$$

The time averaging replaces the expectation assuming ergodicity. Figures 5.1(b) and (c) compare the empirical covariance to the predicted worst-case covariance. By visual inspection, we observe that the empirical results are in good agreement with our theoretical

predictions. For a more precise comparison, we plot the first twenty dominant eigenvalues and first three dominant eigenfunctions of both the predicted and empirical covariances in Figure 5.1(d). This eigen-decomposition is referred to as the *Karhunen-Loève decomposition*. The eigenfunctions are the modes of BM vibrations that have the highest growth rate and are more likely to destabilize for small perturbations of the active gain. The plots doesn't show any significant difference between the empirical and theoretical results. In fact, although the nonlinear active gain slightly deforms the response, but its fundamental role (in the absence of a stimulus) is to saturate the linearized instabilities to form oscillations that remain bounded in time.

5.4 Discussion

The mechanisms underlying cochlear instabilities such as SOAEs and tinnitus are still controversial and not well understood. This work suggests a new possible source of cochlear instabilities: spatio-temporal stochastic perturbations of the active gain.

It is widely accepted that Outer Hair Cells (OHC) are responsible for the active gain in the cochlea. This work proposes a *simulation-free* control theory framework to analyze the effects of small stochastic perturbations that may occur on the level of the OHCs. These perturbations can have several physical origins such as noisy nearby neuronal activities, cellular activities, blood flow, etc...

Studying the effects of randomness in the active gain is not new [24], [39]. However, the previous studies on this matter considered random spatial perturbations that are time-invariant. This type of randomness is referred to as "frozen" or *quenched disorder* in the statistical physics community. In fact, [39] investigated the effects of the frozen spatial randomness by carrying out Monte Carlo simulations to study the statistics of the instabilities. However, to achieve a broad spectrum of unstable frequencies, the authors

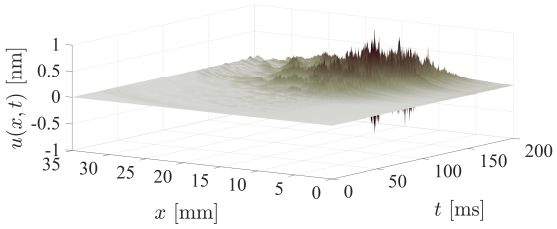
allowed severe perturbations of the active gain which is not realistic. Without these severe perturbations, the unstable frequencies would be limited to a band of high frequencies only (Section 4.3.2). This doesn't agree with the experimental observations where, for example, SOAEs are mainly found between 0.5 and 4.5 kHz.

A more realistic case is to treat the active gain as a stochastic process, where the randomness may occur in space and time, simultaneously. In addition to that, only small perturbations of the active gain are considered (three to four orders of magnitude less than [39]). A major advantage of our analysis is that it is *simulation-free* and no Monte Carlo simulations are required to study the statistics of the emerging instabilities. In our analysis, we also show that the band of unstable frequencies can be controlled by the tuning of the structural parameters of the cochlea such as the active gain coefficient. Hence, we show that even for very small perturbations, the unstable frequencies can be shifted dramatically. Furthermore, examining localized stochastic perturbations in the active gain allowed us to observe local instabilities that shift toward the stapes as the localization length or spread is larger. This observation resembles the detuning phenomenon present in the cochlea.

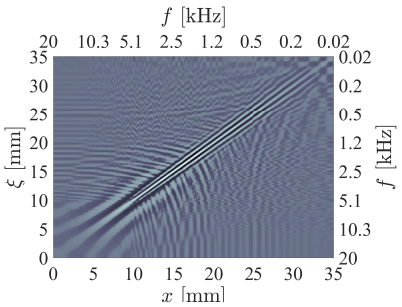
5.5 Conclusion and Future Work

This paper examines the instabilities that occur in the linearized dynamics due to spatio-temporal stochastic perturbations in the distributed structure of the cochlear partition. The simulation-free analysis is carried out through a structured stochastic uncertainty framework. It is shown that the spatial shape of the expectation and covariance of the gain coefficient affect the locations of the instabilities on the basilar membrane. These instabilities eventually saturate to form bounded oscillations due to the saturation nonlinearity of the active gain (4.3) producing spontaneous basilar membrane vibrations.

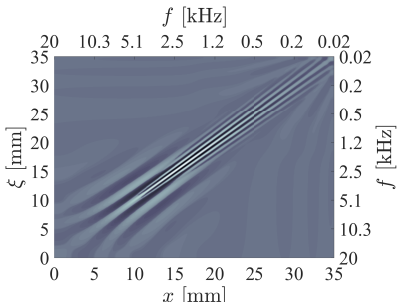
It is believed that these instabilities are reflected to the middle ear as spontaneous otoacoustic emissions (SOAEs) [49] with frequencies corresponding to the location of the instability on the basilar membrane. This analysis also suggests an explanation of one possible source of tinnitus, which is less addressed in the literature. Particularly, if the spontaneous BM vibrations were intense enough, they may be perceived as tinnitus. Future work will address instabilities that may occur due to stochastic uncertainties in structural parameters other than the active gain coefficient, such as the cochlear fluid density.



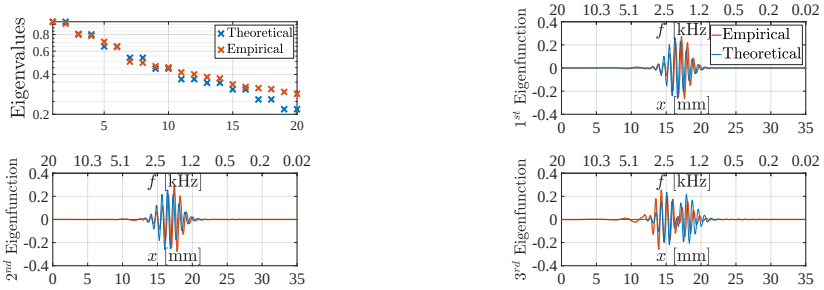
(a) Spatio-Temporal Stochastic Evolution of the BM



(b) Empirical Covariance $\mathbf{U}_{\text{Emp}}(x,\xi)$



(c) Predicted Worst-Case Covariance $\mathbf{U}(x,\xi)$



(d) Empirical and Theoretical Dominant Eigenvalues/functions

Figure 5.1: Nonlinear Stochastic Simulation. Figure (a) shows the BM response to spatially uncorrelated stochastic active gain (scenario $\mathbf{S}_1$) with an expectation given by (4.23) where $\beta = 2$ and a perturbation of $\epsilon = 1.1 \times 10^{-5}$. Figures (b) and (c) show a comparison between the empirical and predicted covariances. The predicted covariance is computed for the linearized dynamics via the power iteration method applied on the loop gain operator (4.16). The empirical covariance is computed using the data obtained from one nonlinear stochastic simulation using (5.3) and integrated in time using (5.4) assuming ergodicity. Figure (d) shows a comparison between the dominant eigenvalues/functions of the empirical and predicted covariances shown in Figures (b) and (c), respectively. This eigen-decomposition is referred to as the *KarhunenLoève* decomposition. Clearly the theoretical predictions match the empirical data, thus suggesting that the nonlinearities only saturate the response without significantly deforming the waveforms.

Appendix

5.A Mass Operators

The fluids block in Figure 4.1(a) can be modeled in 1D, 2D, or 3D. In two dimensions, Navier-Stokes equations boil down to the Laplace equation with the appropriate boundary conditions as shown in [41]. This simplification is valid under the assumptions of incompressible, inviscid fluid where the magnitude of the vibrations of the membranes are negligible relative to the dimensions of the cochlea. These assumptions make the fluid block in Figure 4.1 memoryless and amenable to be represented by the two linear spatial operators $\mathcal{M}_f$ and $\mathcal{M}_s$ in (4.1). In this paper, we give these operators for the 1D case only. Higher dimensions can be treated similarly. As in [19], the fluid block in 1D can be represented by the traveling wave equation as follows

$$\frac{\partial^2}{\partial x^2}p(x,t) = \frac{2\rho}{H}\ddot{u}(x,t); \begin{cases} \frac{\partial}{\partial x}p(0,t) = 2\rho\ddot{s}(t) \\ p(L,t) = 0, \end{cases} \tag{5.A.1}$$

where ρ is the density of the fluid, H is the height of the fluid chamber and L is the length of the BM. This is a linear system with two inputs: $\ddot{u}$ and $\ddot{s}$. It can be shown that

the solution of (5.A.1) is

$$\begin{aligned} p(x,t) &= -[\mathcal{M}_f \ddot{u}](x,t) - [\mathcal{M}_s]\ddot{s}(t) \\ [\mathcal{M}_f \ddot{u}](x,t) &:= -\frac{2\rho}{H}\sum_{n=0}^{\infty}\frac{1}{\lambda_n}\phi_n(x)\langle\phi_n, \ddot{u}(.,t)\rangle \\ [\mathcal{M}_s \ddot{s}](t) &:= 2\rho(L-x)\ddot{s}(t), \end{aligned} \tag{5.A.2}$$

where $\langle .,. \rangle$ denotes the inner product in the space of square integrable functions over $[0, L]$, and

$$\lambda_n = -(n+\frac{1}{2})^2\frac{\pi^2}{L^2} \longleftrightarrow \phi_n(x) = \sqrt{\frac{2}{L}}\cos\left[\left(n+\frac{1}{2}\right)\frac{\pi}{L}x\right],$$

for $n = 0, 1, 2, ...$ It is fairly straightforward to verify that (5.A.2) is indeed a solution by substituting in (5.A.1).

Finite dimensional approximations can be obtained by representing $\mathcal{M}_f$ and $\mathcal{M}_s$ by the matrix $M_f \in \mathbb{R}^{(N_x+1)\times(N_x+1)}$ and the vector $M_s \in \mathbb{R}^{N_x+1}$, respectively, where N_x+1 is the spatial grid size that discretizes the spatial variable x. This is done by truncating the sum and by using a quadrature rule to compute the inner product (or simply a trapezoidal rule). Note that finite difference methods, in the spirit of [19] and [46], can also be used to approximate the mass operators. However, the spectral method we presented here provides a better and more efficient approximation.

5.B Matrix Approximation of Spatial Operators

Let the matrices

$$F_\eta \in \mathbb{R}^{(N_x+1)\times(N_x+1)}; \qquad A_0 \in \mathbb{R}^{4(N_x+1)\times 4(N_x+1)};$$

$$B_0 \in \mathbb{R}^{4(N_x+1)\times(N_x+1)}; \qquad \tilde{B}_0(u_n) \in \mathbb{R}^{4(N_x+1)\times(N_x+1)};$$

$$C_0 \in \mathbb{R}^{(N_x+1)\times 4(N_x+1)}; \qquad E \in \mathbb{R}^{4(N_x+1)\times 4(N_x+1)};$$

$$A_{\tilde{\gamma}}(u_n) \in \mathbb{R}^{4(N_x+1)\times 4(N_x+1)},$$

be the finite dimensional approximations of the spatial operators $\Phi_\eta, \mathcal{A}_0, \mathcal{B}_0, \tilde{\mathcal{B}}_0(u), \mathcal{C}_0, \mathcal{E}$ and $\mathcal{A}_{\bar{\gamma}}(u_n)$, respectively. Using the trapezoidal integration rule on (4.4), we can construct the matrix F_η as

$$F_\eta = \mathcal{D}\left(\tilde{F}_\eta T\mathbf{1}\right)^{-1}\tilde{F}_\eta T,$$

where $\mathcal{D}$ is the diagonal operator, $\tilde{F}_\eta \in \mathbb{R}^{(N_x+1)\times(N_x+1)}$ and its $(i,j)^{th}$ entry is defined as $\left(\tilde{F}_\eta\right)_{ij} := e^{-(i-j)^2\frac{\Delta_x^2}{\eta^2}}$, $\mathbf{1} \in \mathbb{R}^{N_x+1}$ is a vector whose entries are all ones and $T \in \mathbb{R}^{(N_x+1)\times(N_x+1)}$ is a diagonal matrix defined as

$$T := \mathcal{D}\left(\begin{bmatrix}\frac{1}{2} & 1 & \cdots & 1 & \frac{1}{2}\end{bmatrix}\right).$$

Furthermore, define the following diagonal matrices $\in \mathbb{R}^{(N_x+1)\times(N_x+1)}$

$$\begin{aligned}
K_l &:= \mathcal{D}\left(\begin{bmatrix}k_l(x_0) & \cdots & k_l(x_{N_x})\end{bmatrix}\right), \quad l = 1,2,3,4;\\
C_l &:= \mathcal{D}\left(\begin{bmatrix}c_l(x_0) & \cdots & c_l(x_{N_x})\end{bmatrix}\right), \quad l = 1,2,3,4;\\
D_{\bar{\gamma}} &:= \mathcal{D}\left(\begin{bmatrix}\bar{\gamma}(x_0) & \cdots & \bar{\gamma}(x_{N_x})\end{bmatrix}\right);\\
\tilde{G}(u_n) &:= \mathcal{D}\left(1 + \frac{\theta}{R^2}F_\eta(u_n \circ u_n)\right)^{-1},
\end{aligned}$$

where $\circ$ is the Hadamard (element-by-element) product. Therefore

$$E := \begin{bmatrix} I & 0 & 0 & 0\\ 0 & I & 0 & 0\\ 0 & 0 & \frac{g}{b}m_1 I + M_f & 0\\ 0 & 0 & 0 & m_2 I\end{bmatrix}; \quad B := \begin{bmatrix}0\\ 0\\ -M_s\\ 0\end{bmatrix};$$

$$A_0 := \begin{bmatrix} 0 & 0 & I & 0\\ 0 & 0 & 0 & I\\ -\frac{g}{b}(K_1+K_3) & K_3 & -\frac{g}{b}(C_1+C_3) & C_3\\ \frac{g}{b}K_3 & -(K_2+K_3) & \frac{g}{b}C_3 & -(C_2+C_3)\end{bmatrix};$$

$$B_0 := \begin{bmatrix} 0 & 0 & I & 0 \end{bmatrix}^T; \qquad C_0 := \begin{bmatrix} \frac{g}{b}K_4 & -K_4 & \frac{g}{b}C_4 & -C_4 \end{bmatrix};$$
$$\tilde{B}_0(u_n) := \begin{bmatrix} 0 & 0 & \tilde{G}(u_n) & 0 \end{bmatrix}^T;$$
$$A_{\bar{\gamma}}(u_n) := A_0 + \tilde{B}_0(u_n) D_{\bar{\gamma}} C_0.$$

5.C System Linearization

The only nonlinear portion of the dynamics appears in the active gain given by (4.3). Thus, to linearize the dynamics around the origin, it suffices to linearize the active gain. Up to first order, the active gain can be expanded around some $\bar{u}$, by letting $u := \bar{u} + \epsilon\tilde{u}$. The expansion is given by

$$\mathcal{G}(u) = \mathcal{G}(\bar{u}) + \epsilon \left[\frac{\partial}{\partial u}\mathcal{G}(\bar{u})\right](\tilde{u}) + \mathcal{O}(\epsilon^2),$$

where $\left[\frac{\partial}{\partial u}\mathcal{G}(\bar{u})\right](\tilde{u})$ is the Fréchet derivative in the direction of $\tilde{u}$. It can be calculated as follows

$$\begin{aligned}\left[\frac{\partial}{\partial u}\mathcal{G}(\bar{u})\right](\tilde{u}) &:= \lim_{\epsilon\to 0} \frac{\mathcal{G}(\bar{u}+\epsilon\tilde{u}) - \mathcal{G}(\bar{u})}{\epsilon} \\ &= -\frac{2\theta}{R^2} \frac{\gamma\Phi_\eta(\bar{u}\tilde{u})}{\left(1+\theta\Phi_\eta\left(\frac{u^2}{R^2}\right)\right)^2}.\end{aligned}$$

To linearize around the origin, we set $\bar{u} = 0$. This yields

$$\mathcal{G}(0) = \gamma \qquad \text{and} \qquad \left[\frac{\partial}{\partial u}\mathcal{G}(0)\right](\tilde{u}) = 0.$$

Therefore, up to first order, the linearization around the fixed point of the active gain is $\mathcal{G}(u) = \gamma + \mathcal{O}(\epsilon^2)$.

5.D Equivalent Rectangular Bandwidth

The width, η, of the Gaussian kernel in (4.5) controls the spatial coupling length along the BM. The numerical value of η in this paper is chosen based on the critical bands in the cochlea. In psychoacoustics, the concept of critical bands was introduced by Harvey Fletcher in 1933. He described the bands of audio frequencies within which two tones interfere in the perception of each other, thus indicating the length of spatial coupling along the cochlea. This band, which is termed Equivalent Rectangular Bandwidth (ERB), is believed to be equivalent to 0.89 mm on the BM [45].

We model the spatial coupling along the BM using a Gaussian kernel as shown in (4.3-4.5). Hence, we require to calculate the width η of the Gaussian kernel that fits an ERB of 0.89 mm as shown in Figure 5.2. It is fairly straight forward to calculate η, by

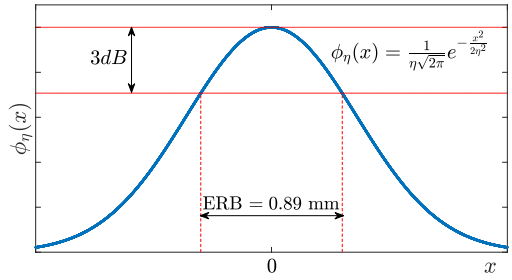


Figure 5.2: Equivalent Rectangular Bandwidth (ERB). The spatial coupling in the micro-mechanical stage is modeled using a Gaussian kernel whose width is chosen to respect the ERB in the cochlea.

setting $\phi_\eta\left(0.89/2\right) = \frac{\sqrt{2}}{2}\phi_\eta(0)$, we get $\eta = 0.5345$ mm.

Part III

Function Space Approach to Numerical Methods in Optimal Control

Chapter 6

Introduction, Notation & Preliminaries

The goal of optimal control is to design a control input for a given dynamical system, so that a cost functional is optimized. In most applications, optimal control problems (OCPs) cannot be solved analytically due to their mathematical complexity; instead, numerically methods are designed to obtain approximate solutions. The objective of this paper is to derive various numerical methods (first and second order) to solve OCPs using a function-space approach. Some of the numerical methods derived in this paper already exist in the literature [55]. However, our goal is to unify the framework upon which the various numerical methods are based on. In fact, the results are re-derived by (1) treating the OCP as an optimization problem in function space, and (2) exploiting the special structure (control-state dynamics) of the optimization problem. This approach gives rise to the definition of various system and projection operators that make the derivations conceptually transparent. It also facilitates the classification of the various methods and uncovers the connections between them. Furthermore, the function-space approach builds useful geometric intuitions that inspire the development of new projection-based methods.

Particularly, this paper develops a preconditioned constrained-gradient descent (PCGD) method which is based on projected gradient descent in infinite dimensional optimization problems [61]. The key is to exploit the special structure of OCPs to precondition the state-control space, and thus achieving a higher convergence rate than the well known gradient descent method [37].

A projection-based approach was proposed by Hauser [29], where the dynamically constrained optimization is transformed into an unconstrained one by using a particular trajectory-tracking, nonlinear, projection operator. A Newton method is then applied to solve the resulting unconstrained optimization problem. Although the projection operator adds more computational cost, the convergence is guaranteed to be quadratic in the vicinity of the solution [29]. Hauser's method approaches the optimal control problem by treating the dynamical system as a manifold in a Banach space as developed in [30]. In this paper, we extend Hauser's method to encompass a more general class of projection operators. Furthermore, we show that the PCGD method yields a particular algorithm that lies in the family of Quasi-Newton methods explained by Hauser. In fact, we carry the dynamical constraints throughout without the calculation of second derivatives of the dynamics (as Newton methods require). This allows us to give a geometric interpretation for the method as a constrained-gradient descent, after preconditioning of the cost functional.

6.1 Problem Statement, Notation & Preliminaries

This section is devoted to define some useful notation that is adopted throughout the paper. The notation is essentially introduced to pose the standard optimal control problem, using operator language, in function space. Let $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ denote

the state and control variables for $0 \leq t \leq T$, respectively. Consider the general OCP

$$\begin{aligned} \underset{x,u}{\text{minimize}} \quad & J(x,u) = \int_0^T L\big(x(t),u(t)\big)dt + \phi\big(x(T)\big) \\ \text{subject to} \quad & \dot{x}(t) = f\big(x(t),u(t)\big); \quad x(0) = \boldsymbol{x_0}, \end{aligned} \tag{6.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, and the terminal cost $\phi : \mathbb{R}^n \to \mathbb{R}$ are all twice differentiable functions, and $\boldsymbol{x_0} \in \mathbb{R}^n$ is a given initial condition. Furthermore, $\dot{x}(t)$ denotes the time derivative of $x(t)$. Note that the results in this paper are also applicable to the case where f and L explicitly depend on time.

To rewrite (6.1) using operator language in function space, we let $\mathbb{L}_n^2[0,T]$ denote the set of n-vector-valued, square-integrable functions over the time interval $[0,T]$. We use the letters $x \in \mathbb{L}_n^2[0,T]$ and $u \in \mathbb{L}_m^2[0,T]$, without the time argument, to represent the state and control variables as functions of time. Furthermore, let $z := (x,u) \equiv \begin{bmatrix} x^* & u^* \end{bmatrix}^*$ denote the state-control pair, where x^* denotes the transpose of x. Note that the parentheses and vector notation for pairing x and u (to form z) is interchangeably used throughout the paper for convenience. Define the time derivative operator $\mathcal{D} : \mathbb{X} \to \mathbb{L}_n^2[0,T]$ as $[\mathcal{D}x](t) := \dot{x}(t)$, where $\mathbb{X} \subset \mathbb{L}_n^2[0,T]$ is the domain of $\mathcal{D}$ defined as

$$\mathbb{X} := \{x \in \mathbb{L}_n^2[0,T] : \ \mathcal{D}x \in \mathbb{L}_n^2[0,T], \ x(0) = \boldsymbol{x_0}\}.$$

Note that $\mathcal{D}$ is a differential operator that is bounded on its domain (by construction), and it imposes a Dirichlet boundary condition on the dynamics. Let $\mathcal{C}$ denote the nonlinear dynamical constraints operator, that acts on z as

$$\begin{aligned} \mathcal{C} : \mathbb{X} \times \mathbb{L}_m^2[0,T] &\to \mathbb{L}_n^2[0,T] \\ \mathcal{C}(z) &:= f(z) - \mathcal{D}x. \end{aligned} \tag{6.2}$$

Therefore, the optimal control problem (6.1) can be rewritten as

$$\begin{aligned} \underset{z}{\text{minimize}} \quad & J(z) \\ \text{subject to} \quad & \mathcal{C}(z) = 0, \end{aligned} \tag{6.3}$$

where the nonlinear cost functional $J : \mathbb{L}^2_{n+m}[0,T] \to \mathbb{R}$ is twice differentiable. The rest of this section introduces more definitions and notation that are useful to derive the various numerical methods using a function-space approach.

6.1.1 Inner Product

Let $\langle .,. \rangle$ and $\langle .,. \rangle_{\mathbb{R}^n}$ denote the usual inner products in $\mathbb{L}^2_n[0,T]$ and $\mathbb{R}^n$, respectively. That is, for any $x, y \in \mathbb{L}^2_n[0,T]$ and $v, w \in \mathbb{R}^n$, we have

$$\langle x, y \rangle := \int_0^T x^*(t) y(t) dt, \qquad \langle v, w \rangle_{\mathbb{R}^n} := v^* w.$$

In this paper, we consider real function spaces and thus the order in the inner products is not significant.

6.1.2 Differential Operators

In addition to $\mathcal{D}$, define the following time derivative operators $\mathcal{D}_0 : \mathbb{X}_0 \to \mathbb{L}^2_n[0,T]$ and $\mathcal{D}_T : \mathbb{X}_T \to \mathbb{L}^2_n[0,T]$, where

$$\mathbb{X}_0 := \{x \in \mathbb{L}^2_n[0,T] : \ \mathcal{D}_0 x \in \mathbb{L}^2_n[0,T], \ x(0) = 0\}$$
$$\mathbb{X}_T := \{x \in \mathbb{L}^2_n[0,T] : \ \mathcal{D}_T x \in \mathbb{L}^2_n[0,T], \ x(T) = 0\},$$

such that they have the same action as $\mathcal{D}$, that is $[\mathcal{D}_0 x](t) = \dot{x}(t)$ and $[\mathcal{D}_T x](t) = \dot{x}(t)$, but their domains of definition are different. Observe that the domain of $\mathcal{D}$ is an affine subspace; whereas, the domains of $\mathcal{D}_0$ and $\mathcal{D}_T$ are linear subspaces. In fact, it is fairly straight forward to see that the three operators are related as

$$\partial_x \mathcal{D} = \mathcal{D}_0 \qquad \text{and} \qquad \mathcal{D}_0^* = -\mathcal{D}_T, \tag{6.4}$$

where $\partial_x \mathcal{D}$ and $\mathcal{D}_0^*$ denote the directional derivative of $\mathcal{D}$ and adjoint of $\mathcal{D}_0$, respectively. See Appendix 10.A for more details.

6.1.3 Evaluation Operator & the Delta (Generalized) Function

Let $\mathcal{S}_T$ denote the evaluation operator that evaluates a bounded continuous function in $\mathbb{X}_0$ at time $t = T$. That is, for any $x \in \mathbb{X}_0$, we have $\mathcal{S}_T x := x(T)$. Formally, we have

$$\mathcal{S}_T x = x(T) = \int_0^T \delta(t - T)x(t)dt,$$

where δ is the Dirac delta function. Thus, the adjoint $\mathcal{S}_T^*$ of $\mathcal{S}_T$ is given by

$$\mathcal{S}_T^*(t) = \delta(t - T), \tag{6.5}$$

and with slight abuse of notation, we write $\langle \mathcal{S}_T x, v\rangle_{\mathbb{R}^n} = \langle x, \mathcal{S}_T^* v\rangle$ for any $v \in \mathbb{R}^n$. Refer to 10.B for a rigorous treatment that justifies this abuse of notation.

6.1.4 Subscripts & Superscripts

Throughout the paper, the subscript k is used to denote the iteration number of the numerical methods. For example, $z_k(t)$ denotes a vector-valued function at the k^{th} iteration. However, when there is a need to index the entries of the vector, we switch notation to $z_i^{(k)}$ where now the subscript i denotes the i^{th} entry, and the superscript denotes the k^{th} iteration.

6.1.5 Partial Derivatives

Define the following partial derivatives, evaluated at a given $z_k(t) := \begin{bmatrix} x_k^*(t) & u_k^*(t) \end{bmatrix}^*$, as

$$L_k^x(t) := \partial_x L^*_{z_k(t)} \in \mathbb{R}^n, \qquad L_k^u(t) := \partial_u L^*_{z_k(t)} \in \mathbb{R}^m,$$

$$Q_k(t) := \partial_x^2 L_{z_k(t)} \in \mathbb{R}^{n\times n}, \quad R_k(t) := \partial_u^2 L_{z_k(t)} \in \mathbb{R}^{m\times m}, \quad N_k(t) := \partial_{xu} L_{z_k(t)} \in \mathbb{R}^{n\times m},$$

$$\phi_k^x := \left[\partial_x \phi_{x_k(T)}\right]^* \in \mathbb{R}^n, \qquad \phi_k^{xx} := \partial_x^2 \phi_{x_k(T)} \in \mathbb{R}^{n\times n},$$

where for example, $\partial_x L_{z_k(t)}$ means the partial derivative (with respect to x) of L, evaluated at $z_k(t) := \big(x_k(t), u_k(t)\big)$, and the star denotes the transpose. Furthermore, define

$$L_k(t) := \partial L^*_{z_k(t)} = \begin{bmatrix} \partial_x L_{z_k(t)} & \partial_u L_{z_k(t)} \end{bmatrix}^* = \begin{bmatrix} L^x_k(t) \\ L^u_k(t) \end{bmatrix},$$

$$H_k(t) := \partial^2 L^*_{z_k(t)} = \begin{bmatrix} \partial^2_x L_{z_k(t)} & \partial_{xu} L_{z_k(t)} \\ \partial_{ux} L_{z_k(t)} & \partial^2_u L_{z_k(t)} \end{bmatrix} = \begin{bmatrix} Q_k(t) & N_k(t) \\ N^*_k(t) & R_k(t) \end{bmatrix},$$

where the subscript of the partial derivative symbol "∂" is suppressed to indicate the total derivative (that is, the derivative with respect to all its arguments). Note that the Hessians $H_k(t), Q_k(t)$ and $R_k(t)$ are all symmetric matrices.

Now define the Jacobian of f evaluated at a given $z_k(t)$ as

$$\partial f_{z_k(t)} = \begin{bmatrix} \partial_x f_{z_k(t)} & \partial_u f_{z_k(t)} \end{bmatrix} =: \begin{bmatrix} A_k(t) & B_k(t) \end{bmatrix},$$

where $A_k(t) \in \mathbb{R}^{n\times n}$ and $B_k(t) \in \mathbb{R}^{n\times m}$. Furthermore, define the second derivative of f evaluated at z_k and acting on $\tilde{z}$ as

$$\partial^2 f_{z_k(t)}\big(\tilde{z}(t)\big) = \begin{bmatrix} \tilde{z}^*(t) F^{(k)}_1(t) \tilde{z}(t) \\ \tilde{z}^*(t) F^{(k)}_2(t) \tilde{z}(t) \\ \vdots \\ \tilde{z}^*(t) F^{(k)}_n(t) \tilde{z}(t) \end{bmatrix}, \qquad F^{(k)}_i(t) := \begin{bmatrix} \partial^2_x f_i(z_k(t)) & \partial_{xu} f_i(z_k(t)) \\ \partial_{ux} f_i(z_k(t)) & \partial^2_u f_i(z_k(t)) \end{bmatrix}, \quad (6.6)$$

where f_i denotes the i^{th} component of the vector-valued function f. Note that $F^{(k)}_i \in \mathbb{R}^{(n+m)\times(n+m)}$ is the Hessian of the i^{th} component of f evaluated at z_k and is symmetric. To obtain a more compact notation, define the second derivative operator of f at z_k, denote by $\mathcal{F}_k(t) := \{F^{(k)}_1(t), F^{(k)}_2(t), \cdots F^{(k)}_n(t)\}$, so that the second derivative of f can be written as

$$\partial^2 f_{z_k(t)}\big(\tilde{z}(t)\big) := \tilde{z}^*(t) \mathcal{F}_k(t) \tilde{z}(t) \in \mathbb{R}^n, \quad (6.7)$$

which is nothing but a short notation for (6.6). This compact notation is particularly useful to perform the following operation. For some $\lambda_k(t) \in \mathbb{R}^n$, we have

$$\left\langle \lambda_k, \partial^2 f_{z_k}\left(\tilde{z}\right)\right\rangle = \langle \lambda_k, \tilde{z}^* \mathcal{F}_k \tilde{z}\rangle = \left\langle \lambda_k, \begin{bmatrix} \tilde{z}^* F_1^{(k)} \\ \vdots \\ \tilde{z}^* F_n^{(k)} \end{bmatrix} \tilde{z} \right\rangle = \left\langle \begin{bmatrix} F_1^{(k)} \tilde{z} & \cdots & F_2^{(k)} \tilde{z} \end{bmatrix} \lambda_k, \tilde{z} \right\rangle$$
$$= \left\langle \left(\sum_{i=1}^{n} F_i^{(k)} \lambda_i^{(k)} \right) \tilde{z}, \tilde{z} \right\rangle =: \langle \mathcal{F}_k \lambda_k \tilde{z}, \tilde{z}\rangle ,$$

where the operation $\mathcal{F}_k \lambda_k$ is defined as

$$\mathcal{F}_k(t)\lambda_k(t) := \sum_{i=1}^{n} F_i^{(k)}(t)\lambda_i^{(k)}(t). \tag{6.8}$$

Finally, we obtain the compact equality

$$\langle \lambda_k, \tilde{z}^* \mathcal{F}_k \tilde{z}\rangle = \langle \mathcal{F}_k \lambda_k \tilde{z}, \tilde{z}\rangle . \tag{6.9}$$

6.1.6 Geometric Notation

In this section, we introduce some geometric notation that is useful to provide geometric interpretations of the relevant numerical methods developed in this paper.

- ***Dynamical Constraint Set:***

  Let $\mathcal{T}$ denote the set of trajectories satisfying the dynamics, that is

  $$\mathcal{T} = \Big\{ z = (x, u) \in \mathbb{X} \times L_m^2[0, T] : \mathcal{C}(z) = 0 \Big\}.$$

- ***Trajectory Projection Operator:***

  Let $\Pi_{\mathcal{T}}$ denote a nonlinear projection operator that acts on an arbitrary state-control pair $\hat{z} = (\hat{x}, \hat{u})$ to yield another pair $z = (x, u) \in \mathcal{T}$ that are trajectories of the dynamics as

  $$\Pi_{\mathcal{T}}(\hat{z}) = z \iff \begin{cases} u = \hat{u} \\ \mathcal{D}x = f(z). \end{cases}$$

- ***Dynamical Tangent Space:***

  Let $T_{z_k}\mathcal{T}$ denote the tangent space of $\mathcal{T}$ at z_k.

$$T_{z_k}\mathcal{T} := \{\tilde{z} \in \mathbb{X}_0 \times L^2_m[0,T] : \partial\mathcal{C}_{z_k}(\tilde{z}) = 0\}.$$

  Note that $T_{z_k}\mathcal{T}$ represents the linearized dynamics around z_k.

- ***Oblique Projection Operator:***

  Let $\Pi^H_{T_{z_k}\mathcal{T}}$ denote a linear oblique-projection operator parameterized by $H = H^* \geq 0$ that projects onto the tangent space $T_{z_k}\mathcal{T}$.

$$\Pi^H_{T_{z_k}\mathcal{T}}(z) := \underset{\tilde{z}}{\operatorname{argmin}} \frac{1}{2}\langle H(z-\tilde{z}), z-\tilde{z}\rangle$$
$$\text{s.t.} \quad \tilde{z} \in T_{z_k}\mathcal{T}.$$

  Note that the projection becomes orthogonal if H is equal to the identity matrix. In this case, the superscript $H = I$ is dropped.

6.2 Brief Tutorial on Optimization in Function-Space

We give a brief review of how an unconstrained optimization in function space can be (abstractly) solved using first and second order iterative methods. Consider a nonlinear functional $\mathcal{J} : \Psi \subset \mathbb{L}^2_n[0,T] \to \mathbb{R}$ on some dense subspace of $\mathbb{L}^2_n[0,T]$, that is if $\eta \in \Psi$, then $\mathcal{J}(\eta) \in \mathbb{R}$. The goal is to find a particular function $\boldsymbol{\eta}$ that minimizes the cost functional, that is

$$\boldsymbol{\eta} = \underset{\eta}{\operatorname{argmin}}\ \mathcal{J}(\eta). \tag{6.10}$$

An iterative method to solve (6.10) can be written, in general, as

$$\eta_{k+1} = \eta_k + \alpha_k \tilde{\eta}_k, \tag{6.11}$$

that is, given the current estimate of the minimum η_k, calculate an update direction $\tilde{\eta}_k$ (at the current iteration k) and "move" along that direction in a step size of α_k to obtain a new estimate η_{k+1}. This iteration is repeated until a desirable convergence measure is achieved. Therefore, the various numerical methods to solve (6.10) differ by the choice of the update direction $\tilde{\eta}_k$ at each iteration. We give two methods here: (a) a first order method and (b) a second order method. Note that the step size α_k can be chosen to be a constant throughout all iterations, or can be designed using various schemes that exist in the literature (e.g. [2]). Before we give a description of the two different methods, we provide a brief review on gradients and Hessians of functionals.

6.2.1 Gradients & Hessians of Nonlinear Functionals

The directional (Gâteaux) derivative of $\mathcal{J}$, evaluated at $\eta_k \in \Psi$, acting on the direction of some $\tilde{\eta}$ is defined as

$$\partial \mathcal{J}_{\eta_k}(\tilde{\eta}) := \lim_{\epsilon \to 0} \frac{\mathcal{J}(\eta_k + \epsilon\tilde{\eta}) - \mathcal{J}(\eta_k)}{\epsilon}.$$

Note that $\partial \mathcal{J}_{\eta_k}$ is the gradient of $\mathcal{J}$ at η_k. In fact, it is a linear functional whose action can be expressed using an inner product (more precisely a bilinear form, see 10.B) as

$$\partial \mathcal{J}_{\eta_k}(\tilde{\eta}) = \langle \partial \mathcal{J}_{\eta_k}, \tilde{\eta} \rangle .$$

Furthermore, the second directional derivative of $\mathcal{J}$, evaluated at η_k, acting on the direction of some $\tilde{\eta}$ is defined as

$$\partial^2 \mathcal{J}_{\eta_k}(\tilde{\eta}) := \lim_{\epsilon \to 0} \frac{\partial \mathcal{J}_{\eta_k + \epsilon\tilde{\eta}}(\tilde{\eta}) - \partial \mathcal{J}_{\eta_k}(\tilde{\eta})}{\epsilon}.$$

This can be seen as the directional derivative of the directional derivative of $\mathcal{J}$. Note that $\partial^2 \mathcal{J}_{\eta_k}$ is the Hessian of $\mathcal{J}$ evaluated at η_k. It defines a quadratic functional whose action can be expressed as

$$\partial^2 \mathcal{J}_{\eta_k}(\tilde{\eta}) = \left\langle \partial^2 \mathcal{J}_{\eta_k} \tilde{\eta}, \tilde{\eta} \right\rangle .$$

Equipped with the gradient and Hessian of $\mathcal{J}$, the (abstract) numerical methods to solve (6.10) can be developed as explained next. A first order method can be constructed by picking the steepest descent direction (negative of the gradient) at each iteration, that is

$$\tilde{\eta}_k := -\partial \mathcal{J}_{\eta_k}, \tag{6.12}$$

where $\mathcal{J}_{\eta_k}$ is the gradient of $\mathcal{J}$ evaluated at the current iteration η_k. In fact, a necessary condition of optimality is obtained by setting the gradient to zero, that is $\partial \mathcal{J}_{\boldsymbol{\eta}} = 0$.

A second order method can be constructed by choosing the update direction as

$$\tilde{\eta}_k := \underset{\tilde{\eta}}{\operatorname{argmin}} \quad \mathcal{J}(\eta_k) + \langle \partial \mathcal{J}_{\eta_k}, \tilde{\eta} \rangle + \frac{1}{2} \left\langle \partial^2 \mathcal{J}_{\eta_k} \tilde{\eta}, \tilde{\eta} \right\rangle, \tag{6.13}$$

where $\partial \mathcal{J}_{\eta_k}$ and $\partial^2 \mathcal{J}_{\eta_k}$ are the gradient and Hessian of $\mathcal{J}$ evaluated at the current iteration η_k, respectively. In words, instead of solving the nonlinear optimization (6.10), we approximate the nonlinear cost functional up to second order (linear-quadratic) and thus solve a simpler linear-quadratic optimization at each iteration. This is referred to as *Sequential Quadratic Programming* (SQP). In fact, the (abstract) solution of (6.13) can be easily obtained by setting the gradient (with respect to $\tilde{\eta}$) of the linear-quadratic cost functional in (6.13) to zero. This yields a linear equation to be solved for the update direction $\tilde{\eta}_k$

$$\partial^2 \mathcal{J}_{\eta_k}(\tilde{\eta}_k) = -\partial \mathcal{J}_{\eta_k}. \tag{6.14}$$

Note that this SQP is equivalent to solving the nonlinear equation giving the necessary condition of optimality, $\partial \mathcal{J}_{\boldsymbol{\eta}} = 0$, using a Newton iteration method (when $\alpha_k = 1$).

This section gives two numerical methods to solve unconstrained optimization problems. However, the optimal control problem (6.3) has dynamical constraints. In this paper, we show that the difference between various numerical methods in optimal control boils down to the technique of converting the constrained optimization given in (6.3)

to an unconstrained one. That is, they differ by the way $\mathcal{J}$ is constructed. Once we have an unconstrained optimization, the methods presented in this section can be directly applied.

6.3 Gradient and Hessian of J

The gradient and Hessian of the cost functional J in (6.1) (or equivalently in (6.3)) are given in this section, and will be used throughout the paper. The directional derivative of J, evaluated at $z_k = (x_k, u_k)$, acting on the direction of $\tilde{z}$ is calculated as

$$\partial J_{z_k}(\tilde{z}) = \int_0^T \partial L_{z_k(t)}\tilde{z}(t)dt + \partial_x \phi_{x_k(T)}\tilde{x}(T) = \langle L_k, \tilde{z}\rangle + \langle \phi_k^x, \mathcal{S}_T\tilde{x}\rangle_{\mathbb{R}^n} = \langle L_k, \tilde{z}\rangle + \langle \mathcal{S}_T^*\phi_k^x, \tilde{x}\rangle$$

$$\partial J_{z_k}(\tilde{z}) = \left\langle \begin{bmatrix} L_k^x + \mathcal{S}_T^*\phi_k^x \\ L_k^u \end{bmatrix}, \begin{bmatrix} \tilde{x} \\ \tilde{u} \end{bmatrix} \right\rangle =: \langle \partial J_{z_k}, \tilde{z}\rangle\,. \tag{6.15}$$

See Section 6.1.3 for details on $\mathcal{S}_T$ and its adjoint $\mathcal{S}_T^*$. Equation (6.15) characterizes the action of the gradient on $\tilde{z}$.

The second directional derivative of J, evaluated at z_k, acting on the direction of $\tilde{z}$ is calculated as

$$\begin{aligned}\partial^2 J_{z_k}(\tilde{z}) &= \int_0^T \tilde{z}^*(t)\partial^2 L_{z_k(t)}\tilde{z}(t)dt + \tilde{x}^*(T)\partial_x^2\phi_{x_k(T)}\tilde{x}(T) = \langle H_k\tilde{z}, \tilde{z}\rangle + \langle \phi_k^{xx}\mathcal{S}_T\tilde{x}, \mathcal{S}_T\tilde{x}\rangle_{\mathbb{R}^n} \\ &= \langle H_k\tilde{z}, \tilde{z}\rangle + \langle \mathcal{S}_T^*\phi_k^{xx}\mathcal{S}_T\tilde{z}, \tilde{z}\rangle \\ \partial^2 J_{z_k}(\tilde{z}) &= \left\langle \begin{bmatrix} Q_k + \mathcal{S}_T^*\phi_k^{xx}\mathcal{S}_T & N_k \\ N_k^* & R_k \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{u} \end{bmatrix}, \begin{bmatrix} \tilde{x} \\ \tilde{u} \end{bmatrix} \right\rangle =: \left\langle \partial^2 J_{z_k}\tilde{z}, \tilde{z}\right\rangle. \end{aligned} \tag{6.16}$$

Equation (6.16) characterizes the action of the Hessian on $\tilde{z}$.

Chapter 7

Lagrangian Approach

The goal of this approach is to transform the constrained optimization problem (6.3) into an unconstrained one using the Lagrangian, that is we let $\mathcal{J}$ be the Lagrangian. Then we apply the machinery explained in Section 6.2 to solve the resulting unconstrained optimization.

Define the Lagrangian as

$$\mathcal{J}(z, \lambda) := J(z) + \langle \mathcal{C}(z), \lambda \rangle, \tag{7.1}$$

where $\lambda(t) \in \mathbb{R}^n$ is a Lagrange multiplier. To develop numerical methods using this approach, we calculate the gradient and Hessian of the Lagrangian $\mathcal{J}$.

7.1 Gradient of the Lagrangian

The gradient of the Lagrangian $\mathcal{J}$, evaluated at (z_k, λ_k), is a linear functional. Its action on a given $(\tilde{z}, \tilde{\lambda})$, where $\tilde{z} := (\tilde{x}, \tilde{u})$, is calculated next. Starting from (7.1), we have

$$\partial \mathcal{J}_{(z_k, \lambda_k)}(\tilde{z}, \tilde{\lambda}) = \partial_z \mathcal{J}_{(z_k, \lambda_k)}(\tilde{z}) + \partial_\lambda \mathcal{J}_{(z_k, \lambda_k)}(\tilde{\lambda})$$

$$
\begin{aligned}
&= \partial J_{z_k}(\tilde{z}) + \langle \partial \mathcal{C}_{z_k}(\tilde{z}), \lambda_k \rangle + \left\langle \mathcal{C}(z_k), \tilde{\lambda} \right\rangle \qquad (7.2)\\
&= \partial J_{z_k}(\tilde{z}) + \langle \partial_z f_{z_k} \tilde{z} - \mathcal{D}_0 \tilde{x}, \lambda_k \rangle + \left\langle \mathcal{C}(z_k), \tilde{\lambda} \right\rangle \\
&= \partial J_{z_k}(\tilde{z}) + \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix} \tilde{z}, \lambda_k \right\rangle + \langle \tilde{x}, \mathcal{D}_T \lambda_k \rangle + \left\langle \mathcal{C}(z_k), \tilde{\lambda} \right\rangle \\
&= \left\langle L_k + \begin{bmatrix} A_k & B_k \end{bmatrix}^* \lambda_k, \tilde{z} \right\rangle + \langle \mathcal{D}_T \lambda_k + \mathcal{S}_T^* \phi_k^x, \tilde{x} \rangle + \left\langle \mathcal{C}(z_k), \tilde{\lambda} \right\rangle
\end{aligned}
$$

$$
\partial \mathcal{J}_{(x_k, u_k \lambda_k)}(\tilde{x}, \tilde{u}, \tilde{\lambda}) = \left\langle \begin{bmatrix} L_k^x + (\mathcal{D}_T + A_k^*)\lambda_k + \mathcal{S}_T^* \phi_k^x \\ L_k^u + B_k^* \lambda_k \\ f(x_k, u_k) - \mathcal{D} x_k \end{bmatrix}, \begin{bmatrix} \tilde{x} \\ \tilde{u} \\ \tilde{\lambda} \end{bmatrix} \right\rangle, \qquad (7.3)
$$

where the third and fourth equalities follow from (6.4), and the fifth equality follows from (6.15). Note that a necessary condition of optimality is obtained by setting the gradient to zero. That is, by invoking Appendix 10.E, the optimal variables $(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\lambda})$ satisfy

$$
\begin{cases}
\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t)); & \boldsymbol{x}(0) = \boldsymbol{x_0} \\
\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{A}^*(t)\boldsymbol{\lambda}(t) - \boldsymbol{L}^x(t); & \boldsymbol{\lambda}(T) = \boldsymbol{\phi}^x \\
\boldsymbol{L}^u(t) = -\boldsymbol{B}^*(t)\boldsymbol{\lambda}(t), &
\end{cases} \qquad (7.4)
$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{L}^x, \boldsymbol{L}^u$ and $\boldsymbol{\phi}^x$ are all evaluated at $(\boldsymbol{x}, \boldsymbol{u})$.

7.2 Hessian of the Lagrangian

The Hessian of the Lagrangian $\mathcal{J}$, evaluated at (z_k, λ_k) is a quadratic functional. Its action on a given $(\tilde{z}, \tilde{\lambda})$ is calculated next. Starting from (7.2), we have

$$
\begin{aligned}
\partial^2 \mathcal{J}_{(z_k, \lambda_k)}(\tilde{z}, \tilde{\lambda}) &= \partial^2 J_{z_k}(\tilde{z}) + \left\langle \partial^2 \mathcal{C}_{z_k}(\tilde{z}), \lambda_k \right\rangle + \left\langle \partial \mathcal{C}_{z_k}(\tilde{z}), \tilde{\lambda} \right\rangle + \left\langle \partial \mathcal{C}_{z_k}(\tilde{z}), \tilde{\lambda} \right\rangle \\
&= \partial^2 J_{z_k}(\tilde{z}) + \left\langle \partial^2 f_{z_k}(\tilde{z}), \lambda_k \right\rangle + 2 \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix} \tilde{z} - \mathcal{D}_0 \tilde{x}, \tilde{\lambda} \right\rangle \\
&= \partial^2 J_{z_k}(\tilde{z}) + \langle \tilde{z}^* \mathcal{F}_k \tilde{z}, \lambda_k \rangle + 2 \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix} \tilde{z}, \tilde{\lambda} \right\rangle - 2 \left\langle \mathcal{D}_0 \tilde{x}, \tilde{\lambda} \right\rangle \\
&= \partial^2 J_{z_k}(\tilde{z}) + \langle \tilde{z}, \mathcal{F}_k \lambda_k \tilde{z} \rangle + 2 \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix} \tilde{z}, \tilde{\lambda} \right\rangle - 2 \left\langle \mathcal{D}_0 \tilde{x}, \tilde{\lambda} \right\rangle,
\end{aligned}
$$

where $\mathcal{F}_k$ is defined in (6.6), and the last equality follows from (6.9). Note that $\mathcal{F}_k\lambda_k$ is a time-varying matrix that is defined in (6.8) and can be written as

$$\mathcal{F}_k(t)\lambda_k(t) = \sum_{i=1}^{n} T_i^{(k)}(t)\lambda_i^{(k)}(t) = \sum_{i=1}^{n} \begin{bmatrix} \partial_x^2 f_i(z_k(t))\lambda_i^{(k)}(t) & \partial_{xu} f_i(z_k(t))\lambda_i^{(k)}(t) \\ \partial_{ux} f_i(z_k(t))\lambda_i^{(k)}(t) & \partial_u^2 f_i(z_k(t))\lambda_i^{(k)}(t) \end{bmatrix}$$

$$\mathcal{F}_k(t)\lambda_k(t) =: W_k(t) =: \begin{bmatrix} W_k^{xx}(t) & W_k^{xu}(t) \\ W_k^{ux}(t) & W_k^{uu}(t) \end{bmatrix}. \tag{7.5}$$

Then, by using (6.16), the Hessian of the Lagrangian can be written as

$$\begin{aligned} \partial^2 \mathcal{J}_{(z_k,\lambda_k)}(\tilde{z},\tilde{\lambda}) = \langle (H_k + W_k)\,\tilde{z},\tilde{z}\rangle + \langle \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T \tilde{x},\tilde{x}\rangle + \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix} \tilde{z},\tilde{\lambda} \right\rangle \\ + \left\langle \begin{bmatrix} A_k & B_k \end{bmatrix}^* \tilde{\lambda},\tilde{z} \right\rangle + \left\langle \mathcal{D}_T\tilde{\lambda},\tilde{x}\right\rangle - \left\langle \mathcal{D}_0\tilde{x},\tilde{\lambda}\right\rangle \end{aligned}$$

$$\partial^2 \mathcal{J}_{(x_k,u_k\lambda_k)}(\tilde{x},\tilde{u},\tilde{\lambda}) = \left\langle \begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^*\phi_k^{xx}\mathcal{S}_T & N_k + W_k^{xu} & \mathcal{D}_T + A_k^* \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} & B_k^* \\ -\mathcal{D}_0 + A_k & B_k & 0 \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{u} \\ \tilde{\lambda} \end{bmatrix}, \begin{bmatrix} \tilde{x} \\ \tilde{u} \\ \tilde{\lambda} \end{bmatrix} \right\rangle. \tag{7.6}$$

It is worth to note that if the reader is familiar with gradients and Hessians in function space, the expression (7.6) can be immediately obtained by inspection of the gradient given in (7.3).

With the gradient and Hessian at hand, we construct a second order method to solve the OCP as follows: given the current iterate (x_k, u_k, λ_k), we calculate an update direction denoted by $(\tilde{x}_k, \tilde{u}_k, \tilde{\lambda}_k)$. Then we obtain the next iterate using some step size α_k as

$$\begin{bmatrix} x_{k+1} \\ u_{k+1} \\ \lambda_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ u_k \\ \lambda_k \end{bmatrix} + \alpha_k \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \\ \tilde{\lambda}_k \end{bmatrix}. \tag{7.7}$$

In this approach, we give a second order method only, because using a gradient descent (first order) method on the Lagrangian does not converge in general.

7.3 Second Order Method for the Lagrangian Approach

A second order method is obtained by choosing the update direction $\tilde{\eta}_k := (\tilde{x}_k, \tilde{u}_k, \tilde{\lambda}_k)$ according to (6.14). That is, by substituting the expressions of the gradient (7.3) and the Hessian (7.6), we obtain

$$\partial^2 \mathcal{J}_{(x_k,u_k,z_k)} \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \\ \tilde{\lambda}_k \end{bmatrix} = -\partial \mathcal{J}_{(x_k,u_k,\lambda_k)}$$

$$\begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} & \mathcal{D}_T + A_k^* \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} & B_k^* \\ -\mathcal{D}_0 + A_k & B_k & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \\ \tilde{\lambda}_k \end{bmatrix} = - \begin{bmatrix} L_k^x + (\mathcal{D}_T + A_k^*)\lambda_k + \mathcal{S}_T^* \phi_k^x \\ L_k^u + B_k^* \lambda_k \\ f(x_k, u_k) - \mathcal{D} x_k \end{bmatrix}.$$

This can be rearranged and rewritten as

$$\begin{aligned} \mathcal{D}_0 \tilde{x}_k &= A_k \tilde{x}_k + B_k \tilde{u}_k + f(x_k, u_k) - \mathcal{D} x_k \\ \mathcal{D}_T(\lambda_k + \tilde{\lambda}_k) + \mathcal{S}_T^*(\phi_k^{xx} \tilde{x}_k(T) + \phi_k^x) &= -\left(Q_k + W_k^{xx}\right) \tilde{x}_k - A_k^* \left(\tilde{\lambda}_k + \lambda_k\right) - (N_k + W_k^{xu}) \tilde{u}_k - L_k^x \\ \left(R_k + W_k^{uu}\right) \tilde{u}_k &= -(N_k^* + W_k^{ux}) \tilde{x}_k - B_k^* \left(\tilde{\lambda}_k + \lambda_k\right) - L_k^u. \end{aligned}$$

This is the operator form of the differential equations that govern the update direction $(\tilde{x}, \tilde{u}, \tilde{\lambda})$. By invoking Appendix 10.E, we can get rid of $\mathcal{S}_T^*$ and rewrite the result as a linear differential algebraic equation (DAE) where the differential equations take the

form a two point boundary value problem. We have

$$\frac{d}{dt}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} = \begin{bmatrix} A_k & 0 \\ -\left(Q_k + W_k^{xx}\right) & -A_k^* \end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix} B_k \\ -\left(N_k + W_k^{xu}\right)\end{bmatrix}\tilde{u}_k + \begin{bmatrix} f(x_k,u_k) - \mathcal{D}x_k \\ -L_k^x - \dot{\lambda}_k - A_k^*\lambda_k\end{bmatrix}$$

$$(R_k + W_k^{uu})\tilde{u}_k = -\begin{bmatrix} N_k^* + W_k^{ux} & B_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} - L_k^u - B_k^*\lambda_k$$

$$\text{such that} \qquad \tilde{x}_k(0) = 0 \qquad \text{and} \qquad \tilde{\lambda}_k(T) = \phi_k^x + \phi_k^{xx}\tilde{x}_k(T) - \lambda_k(T). \tag{7.8}$$

Finally, the algorithm for this second order method is summarized in Algorithm 1.

Algorithm 1 Second Order Method: Lagrangian Approach

1: Start with an initial guess (x_1, u_1, λ_1) and set $k = 1$.

2: Given (x_k, u_k, λ_k), compute :

$$
\begin{aligned}
A_k &= \partial_x f_{(x_k,u_k)}, & B_k &= \partial_u f_{(x_k,u_k)},\\
L_k^x &= \partial_x L^*_{(x_k,u_k)}, & L_k^u &= \partial_u L^*_{(x_k,u_k)},\\
Q_k &= \partial_x^2 L_{(x_k,u_k)}, & R_k &= \partial_u^2 L_{(x_k,u_k)},\\
N_k &= \partial_{xu} L_{(x_k,u_k)},\\
\phi_k^x &= \partial_x \phi^*_{x_k(T)}, & \phi_k^{xx} &= \partial_x^2 \phi_{x_k(T)},\\
W_k^{xx} &= \sum_{i=1}^{n} \partial_x^2 f_i(x_k,u_k)\lambda_i^{(k)}, & W_k^{xu} &= \sum_{i=1}^{n} \partial_{xu} f_i(x_k,u_k)\lambda_i^{(k)},\\
W_k^{ux} &= (W_k^{xu})^*, & W_k^{uu} &= \sum_{i=1}^{n} \partial_u^2 f_i(x_k,u_k)\lambda_i^{(k)}.
\end{aligned}
$$

3: Solve the following linear two point boundary value problem (with an algebraic constraint) to obtain $(\tilde{x}_k, \tilde{u}_k, \tilde{\lambda}_k)$:

$$
\frac{d}{dt}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} = \begin{bmatrix} A_k & 0\\ -(Q_k + W_k^{xx}) & -A_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix} B_k\\ -(N_k + W_k^{xu})\end{bmatrix}\tilde{u}_k + \begin{bmatrix} f(x_k,u_k) - \mathcal{D}x_k\\ -L_k^x - \dot{\lambda}_k - A_k^*\lambda_k\end{bmatrix}
$$

$$
(R_k + W_k^{uu})\tilde{u}_k = -\begin{bmatrix} N_k^* + W_k^{ux} & B_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} - L_k^u - B_k^*\lambda_k
$$

$$
\text{such that} \qquad \tilde{x}_k(0) = 0 \qquad \text{and} \qquad \tilde{\lambda}_k(T) = \phi_k^x + \phi_k^{xx}\tilde{x}_k(T) - \lambda_k(T).
$$

4: Update the state, control and Lagrange multiplier using a step size α_k:

$$
\begin{bmatrix} x_{k+1}\\ u_{k+1}\\ \lambda_{k+1}\end{bmatrix} = \begin{bmatrix} x_k\\ u_k\\ \lambda_k\end{bmatrix} + \alpha_k \begin{bmatrix}\tilde{x}_k\\ \tilde{u}_k\\ \tilde{\lambda}_k\end{bmatrix}.
$$

5: Set $k = k + 1$ and go back to step 2. Repeat until convergence.

Chapter 8

Substitution Approach

In this approach, we convert the constrained optimization (6.3) to an unconstrained one by *substituting* the dynamical constraints in the cost functional. This results in a new cost functional that depends only on the control input u.

Define the *system operator* $\mathcal{H}$ that acts on a control variable u to produce a state variable x as

$$\begin{aligned} \mathcal{H} : \mathbb{L}^2_m[0,T] &\to \mathbb{L}^2_n[0,T] \\ u &\mapsto x := \mathcal{H}(u) \quad \text{such that} \quad \mathcal{D}x = f(x,u), \end{aligned} \tag{8.1}$$

where $\mathcal{D}$ is the time derivative operator defined in Section 6.1. By substituting the expression of the state $x = \mathcal{H}(u)$ in the original cost functional J of (6.1), we obtain a new cost functional that only depends on the control input u, that is

$$\mathcal{J}(u) := J\left(\mathcal{H}(u), u\right) = \int_0^T L\Big([\mathcal{H}(u)](t), u(t)\Big)dt + \phi\Big(\mathcal{S}_T\mathcal{H}(u)\Big). \tag{8.2}$$

With the resulting unconstrained optimization problem at hand, we can develop first and second order numerical methods to approximate the solution as follows: given the current iterate u_k, we calculate the update direction $\tilde{u}_k$ and obtain the next iterate using some step size α_k as

$$u_{k+1} = u_k + \alpha_k \tilde{u}_k. \tag{8.3}$$

Table 8.1: System Operator Derivatives and Adjoint. This table shows the expressions of the first and second directional derivatives of the system operator $\mathcal{H}$ evaluated at a given u_k and acting on some $\tilde{u}$. Furthermore, the adjoint of the directional derivative evaluated at u_k is also shown. The expressions are given in their operator forms and their associated differential equations. Refer to Section 6.1 for an explanation of the time derivative operators $\mathcal{D}, \mathcal{D}_0$, and $\mathcal{D}_T$, and the second derivative operator $\mathcal{F}_k$.

	Notation	Operator Form	Differential Equations Form
System Operator	$x_k = \mathcal{H}(u_k)$	$\mathcal{D}x_k = f(x_k, u_k)$	$\dot{x}_k = f(x_k, u_k);\ x_k(0) = \boldsymbol{x_0}$
Derivative	$\tilde{x}_k = \partial\mathcal{H}_{u_k}(\tilde{u})$	$\tilde{x}_k = (\mathcal{D}_0 - A_k)^{-1} B_k \tilde{u}$	$\dot{\tilde{x}}_k = A_k\tilde{x}_k + B_k\tilde{u};\ \tilde{x}_k(0) = 0$
Adjoint	$\mu_k = \partial\mathcal{H}^*_{u_k}(\chi)$	$\mu_k = -B_k^*(\mathcal{D}_T + A_k^*)^{-1}\chi$	$\dot{\lambda}_k = -A_k^*\lambda_k - \chi;\ \lambda_k(T) = 0$ $\mu_k = B_k^*\lambda_k$
Second Derivative	$\bar{x}_k = \partial^2\mathcal{H}_{u_k}(\tilde{u})$	$\bar{x}_k = (\mathcal{D}_0 - A_k)^{-1}\tilde{z}_k^*\mathcal{F}_k\tilde{z}_k$ $\tilde{z}_k := \begin{bmatrix} \tilde{x}_k \\ \tilde{u} \end{bmatrix}$	$\dot{\bar{x}}_k = A_k\bar{x}_k + \tilde{z}_k^*\mathcal{F}_k\tilde{z}_k;\ \bar{x}_k(0) = 0$

The update direction depends on the gradient and/or Hessian of $\mathcal{J}$ which, naturally, depend on the first and second directional derivatives of $\mathcal{H}$. Table 8.1 summarizes the results of the calculations carried out in Appendix 10.C. It shows the formulas for the first directional derivative, its adjoint, and the second directional derivative of $\mathcal{H}$. The formulas are written in both operator form and differential equation form. We now proceed to calculate the gradient and Hessian of $\mathcal{J}$.

8.1 Gradient of $\mathcal{J}(u)$

Recall that the new cost functional $\mathcal{J}$ is a function of the control input u only. Using the chain rule, we calculate the directional derivative of $\mathcal{J}$ in terms of the original cost functional J. Starting from (8.2), we have

$$\begin{aligned}
\partial\mathcal{J}_{u_k}(\tilde{u}) &= \partial J_{(\mathcal{H}(u_k),u_k)}(\tilde{u}) = \partial_x J_{(\mathcal{H}(u_k),u_k)}\big(\partial\mathcal{H}_{u_k}(\tilde{u})\big) + \partial_u J_{(\mathcal{H}(u_k),u_k)}(\tilde{u}) \\
&= \left\langle \begin{bmatrix} \partial_x J_{(\mathcal{H}(u_k),u_k)} \\ \partial_u J_{(\mathcal{H}(u_k),u_k)} \end{bmatrix}, \begin{bmatrix} \partial\mathcal{H}_{u_k}(\tilde{u}) \\ \tilde{u} \end{bmatrix} \right\rangle = \left\langle \partial J_{z_k}, \begin{bmatrix} \partial\mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u} \right\rangle \\
\partial\mathcal{J}_{u_k}(\tilde{u}) &= \left\langle \begin{bmatrix} \partial\mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial J_{z_k}, \tilde{u} \right\rangle =: \langle \partial\mathcal{J}_{u_k}, \tilde{u} \rangle ,
\end{aligned} \tag{8.4}$$

where I_m is the identity matrix of size m, and $z_k := (x_k, u_k) = (\mathcal{H}(u_k), u_k)$. Note that $\mathcal{J}_{u_k}$ given in (8.4) represents the abstract form of the gradient of $\mathcal{J}$ at u_k. It is expressed explicity in terms of original cost functional J and the system operator $\mathcal{H}$. Substituting the expressions of $\partial\mathcal{H}_{u_k}^*$ from Table 8.1 and ∂J_{z_k} from (6.15) yields

$$\begin{aligned}
\partial\mathcal{J}_{u_k}(\tilde{u}) &= \left\langle - \begin{bmatrix} B_k^* \left(\mathcal{D}_T + A_k^*\right)^{-1} & I_m \end{bmatrix} \begin{bmatrix} L_k^x + \mathcal{S}_T^* \phi_k^x \\ L_k^u \end{bmatrix}, \tilde{u} \right\rangle \\
&= \left\langle -B_k^* \left(\mathcal{D}_T + A_k^*\right)^{-1} \left(L_k^x + \mathcal{S}_T^* \phi_k^x\right) + L_k^u, \tilde{u} \right\rangle \\
\partial\mathcal{J}_{u_k}(\tilde{u}) &= \langle B_k^* \lambda_k + L_k^u, \tilde{u} \rangle =: \langle \partial\mathcal{J}_{u_k}, \tilde{u} \rangle ,
\end{aligned} \tag{8.5}$$

where λ_k is an intermediate variable (which is referred to as the *costate* in the literature) and is defined as

$$\lambda_k := -\left(\mathcal{D}_T + A_k^*\right)^{-1} \left(L_k^x + \mathcal{S}_T^* \phi_k^x\right). \tag{8.6}$$

The differential equation associated with (8.6) can be obtained by acting on both sides by $\mathcal{D}_T + A_k^*$ and invoking Appendix 10.E to obtain

$$\dot{\lambda}_k(t) = -A_k^*(t)\lambda_k(t) - L_k^x(t); \quad \lambda_k(T) = \phi_k^x. \tag{8.7}$$

Note that the λ define in (8.7) is fundamentally different from the λ introduced in the Lagrangian approach, although they play similar roles. In the latter, λ is a Lagrange multiplier that doesn't have to satisfy any differential equation. However, the former is the *costate* variable that has to satisfy the differential equation in (8.7) at every iteration. Nonetheless, these two λ's become equal at the optimum.

8.2 Hessian of $\mathcal{J}(u)$

The Hessian of $\mathcal{J}$ at u_k acting on $\tilde{u}$ can be calculated using the chain rule. By calculating the directional derivative of the gradient from (8.4), we can express the Hessian in terms of the original cost functional J.

$$\begin{aligned}\partial^2 \mathcal{J}_{u_k}(\tilde{u}) &= \left\langle \begin{bmatrix} \partial_x^2 J_{z_k} & \partial_{xu} J_{z_k} \\ \partial_{ux} J_{z_k} & \partial_u^2 J_{z_k} \end{bmatrix} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u} \right\rangle + \left\langle \begin{bmatrix} \partial_x J_{z_k} \\ \partial_u J_{z_k} \end{bmatrix}, \begin{bmatrix} \partial^2 \mathcal{H}_{u_k}(\tilde{u}) \\ 0 \end{bmatrix} \right\rangle \\ &= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial^2 J_{z_k} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle + \left\langle \partial_x J_{z_k}, \partial^2 \mathcal{H}_{u_k}(\tilde{u}) \right\rangle .\end{aligned}$$

Substituting the expressions of $\partial^2 \mathcal{H}_{u_k}(\tilde{u})$ from Table 8.1 and $\partial_x J_{z_k}$ from (6.15) yields

$$\begin{aligned}\partial^2 \mathcal{J}_{u_k}(\tilde{u}) &= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial^2 J_{z_k} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle + \left\langle L_k^x + \mathcal{S}_T^* \phi_k^x, (\mathcal{D}_0 - A_k)^{-1} \tilde{z}_k^* \mathcal{F}_k \tilde{z}_k \right\rangle \\ &= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial^2 J_{z_k} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle + \left\langle -(\mathcal{D}_T + A_k^*)^{-1} \left(L_k^x + \mathcal{S}_T^* \phi_k^x \right), \tilde{z}_k^* \mathcal{F}_k \tilde{z}_k \right\rangle\end{aligned}$$

where $\tilde{z}_k := \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}$, $\mathcal{F}_k$ is defined in (6.6), and the second equality follows from (6.4).

By using the same definition of the intermediate variable λ_k in (8.6) and exploiting

(6.9), we obtain

$$\begin{aligned}
\partial^2 \mathcal{J}_{u_k}(\tilde{u}) &= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial^2 J_{z_k} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle + \langle \mathcal{F}_k \lambda_k \tilde{z}_k, \tilde{z}_k \rangle \\
&= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \partial^2 J_{z_k} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle + \left\langle \mathcal{F}_k \lambda_k \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u} \right\rangle \\
&= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \left(\partial^2 J_{z_k} + \mathcal{F}_k \lambda_k \right) \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle,
\end{aligned}$$

where $\mathcal{F}_k \lambda_k$ is the time-varying matrix given in (7.5). Finally, substituting for $\partial^2 J_{z_k}$ from (6.16) yields

$$\begin{aligned}
\partial^2 \mathcal{J}_{u_k}(\tilde{u}) &= \left\langle \begin{bmatrix} \partial \mathcal{H}_{u_k}^* & I_m \end{bmatrix} \begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}, \tilde{u} \right\rangle \\
&=: \left\langle \partial^2 \mathcal{J}_{u_k} \tilde{u}, \tilde{u} \right\rangle. \qquad (8.8)
\end{aligned}$$

Equation (8.8) explicitly shows, in operator form, the action of the Hessian of $\mathcal{J}$ on a given $\tilde{u}$. Note that the expressions of $\partial \mathcal{H}_{u_k}$ and its adjoint are give in Table 8.1. Equipped with the gradient and Hessian of $\mathcal{J}$, we can now calculate the update direction $\tilde{u}_k$ in (8.3).

8.3 First Order Method for the Substitution Approach

A first order method is obtained by simply choosing the update direction $\tilde{u}_k$ to be the negative of the gradient, that is $\tilde{u}_k := -(B_k^* \lambda_k + L_k^u)$, where λ_k is given in (8.7). The algorithm for this first order method is summarized in Algorithm 2.

Algorithm 2 Gradient Descent Method

1: Start with an initial guess u_1 and set $k = 1$.

2: Solve for the state x_k:

$$\dot{x}_k = f(x_k, u_k); \quad x_k(0) = \boldsymbol{x_0}.$$

3: Compute:

$$A_k = \partial_x f_{(x_k,u_k)}, \quad B_k = \partial_u f_{(x_k,u_k)},$$

$$L_k^x = \partial_x L^*_{(x_k,u_k)}, \quad L_k^u = \partial_u L^*_{(x_k,u_k)}, \quad \phi_k^x = \partial_x \phi^*_{x_k(T)}.$$

4: Solve for the costate λ_k:

$$\dot{\lambda}_k = -A_k^* \lambda_k - L_k^x; \quad \lambda_k(T) = \phi_k^x.$$

5: Update the control with a step size α_k:

$$u_{k+1} = u_k - \alpha_k (B_k^* \lambda_k + L_k^u).$$

6: Set $k = k + 1$ and go back to step 2. Repeat until convergence.

8.4 Second Order Method for the Substitution Approach

A second order method is obtained by choosing the update direction $\tilde{u}_k$ according to (6.14). That is, by substituting the expressions of the gradient (8.5) and the Hessian (8.8), we obtain

$$\partial^2 \mathcal{J}_{u_k}(\tilde{u}_k) = -\partial \mathcal{J}_{u_k}$$

$$\begin{bmatrix} \partial \mathcal{H}^*_{u_k} & I_m \end{bmatrix} \begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix} \begin{bmatrix} \partial \mathcal{H}_{u_k} \\ I_m \end{bmatrix} \tilde{u}_k = -(B_k^* \lambda_k + L_k^u).$$

Carrying out the matrix-vector multiplications yields

$$\partial \mathcal{H}^*_{u_k} \left(Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T\right) \partial \mathcal{H}_{u_k}(\tilde{u}_k) + (R_k + W_k^{uu})\tilde{u}_k$$
$$+ \partial \mathcal{H}^*_{u_k}(N_k + W_k^{xu})\tilde{u}_k + (N_k^* + W_k^{ux}) \partial \mathcal{H}_{u_k}(\tilde{u}_k) = -(B_k^* \lambda_k + L_k^u).$$

Now define $\tilde{x}_k := \partial \mathcal{H}_{u_k}(\tilde{u}_k)$, and substitute for the expression of $\partial \mathcal{H}^*_{u_k}$ from Table 8.1 to obtain

$$- B_k^* (\mathcal{D}_T + A_k^*)^{-1} \Big(\left(Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T\right) \tilde{x}_k + (N_k + W_k^{xu})\tilde{u}_k \Big)$$
$$+ (N_k^* + W_k^{ux})\tilde{x}_k + (R_k + W_k^{uu})\tilde{u}_k = -(B_k^* \lambda_k + L_k^u).$$

Finally, introduce a new intermediate variable

$$\tilde{\lambda}_k := -(\mathcal{D}_T + A_k^*)^{-1} \Big(\left(Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T\right) \tilde{x}_k + (N_k + W_k^{xu})\tilde{u}_k \Big), \tag{8.9}$$

and therefore, the update direction $\tilde{u}_k$ is given by the following algebraic equation

$$(R_k + W_k^{uu})\tilde{u}_k = -B_k^* (\lambda_k + \tilde{\lambda}_k) - (N_k^* + W_k^{ux})\tilde{x}_k - L_k^u, \tag{8.10}$$

where $\tilde{x}_k = \partial \mathcal{H}_{u_k}(\tilde{u}_k)$, and $\tilde{\lambda}_k$ solves (8.9) that can be rewritten as a differential equation by invoking Appendix 10.E to get rid of $\mathcal{S}_T^*$. We have

$$\dot{\tilde{\lambda}}_k = -A_k^* \tilde{\lambda}_k - (Q_k + W_k^{xx})\tilde{x}_k - (N_k + W_k^{xu})\tilde{u}_k; \quad \tilde{\lambda}_k(T) = \phi_k^{xx} \tilde{x}_k(T). \tag{8.11}$$

Therefore, using $\tilde{x}_k = \partial \mathcal{H}_{u_k}(\tilde{u}_k)$, (8.11) and (8.10), we obtain the update direction by solving the following linear two point boundary value problem with an algebraic constraint

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} &= \begin{bmatrix} A_k & 0 \\ -(Q_k + W_k^{xx}) & -A_k^* \end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix} B_k \\ -(N_k + W_k^{xu}) \end{bmatrix}\tilde{u}_k \\ (R_k + W_k^{uu})\tilde{u}_k &= -\begin{bmatrix} N_k^* + W_k^{ux} & B_k^* \end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} - L_k^u - B_k^*\lambda_k \\ &\text{such that} \qquad \tilde{x}_k(0) = 0 \qquad \text{and} \qquad \tilde{\lambda}_k(T) = \phi_k^{xx}\tilde{x}_k(T), \end{aligned} \tag{8.12}$$

where λ_k solves (8.7). The algorithm for this second order method is summarized in Algorithm 3.

Algorithm 3 Second Order Method: Substitution Approach

1: Start with an initial guess u_1 and set $k = 1$.

2: Solve for the state $x_k(t)$:

$$\dot{x}_k = f(x_k, u_k); \quad x_k(0) = \boldsymbol{x_0}.$$

3: Compute:

$$A_k = \partial_x f_{(x_k,u_k)}, \qquad B_k = \partial_u f_{(x_k,u_k)},$$

$$L_k^x = \partial_x L^*_{(x_k,u_k)}, \qquad L_k^u = \partial_u L^*_{(x_k,u_k)},$$

$$Q_k = \partial_x^2 L_{(x_k,u_k)}, \qquad R_k = \partial_u^2 L_{(x_k,u_k)},$$

$$N_k = \partial_{xu} L_{(x_k,u_k)},$$

$$\phi_k^x = \partial_x \phi^*_{x_k(T)}, \qquad \phi_k^{xx} = \partial_x^2 \phi_{x_k(T)}.$$

4: Solve for the costate $\lambda_k(t)$:

$$\dot{\lambda}_k = -A_k^* \lambda_k - L_k^x; \quad \lambda_k(T) = \phi_k^x.$$

5: Compute:

$$W_k^{xx} = \sum_{i=1}^{n} \partial_x^2 f_i(x_k, u_k)\lambda_i^{(k)}; \qquad W_k^{xu} = \sum_{i=1}^{n} \partial_{xu} f_i(x_k, u_k)\lambda_i^{(k)};$$

$$W_k^{ux} = (W_k^{xu})^*; \qquad W_k^{uu} = \sum_{i=1}^{n} \partial_u^2 f_i(x_k, u_k)\lambda_i^{(k)}.$$

6: Solve for $\tilde{u}_k$:

$$\frac{d}{dt}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} = \begin{bmatrix} A_k & 0 \\ -(Q_k + W_k^{xx}) & -A_k^* \end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix} B_k \\ -(N_k + W_k^{xu}) \end{bmatrix}\tilde{u}_k; \quad \begin{matrix} x_k(0) = 0 \\ \tilde{\lambda}_k(T) = \phi_k^{xx}\tilde{x}_k(T) \end{matrix}$$

$$(R_k + W_k^{uu})\tilde{u}_k = -\begin{bmatrix} N_k^* + W_k^{ux} & B_k^* \end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} - L_k^u - B_k^*\lambda_k.$$

7: Update the control with a step size α_k:

$$u_{k+1} = u_k + \alpha_k \tilde{u}_k.$$

8: Set $k = k + 1$ and go back to step 2. Repeat until convergence.

Chapter 9

Projection-Based Approach

In this section, we generalize the projection-based method developed in [29]. This method is particularly useful for optimal control problems where the dynamics are either unstable or sensitive. Observe that in the numerical methods developed using the substitution approach, a simulation of the dynamics $\dot{x}_k = f(x_k, u_k)$ is required at each iteration. However, for unstable or sensitive systems, such methods are not recommended, because small perturbations in the control u_k induce large perturbations in the corresponding state x_k. This causes the methods to behave poorly and thus leads to either divergence or extremely slow convergence. In [29], the constrained OCP (6.3) is converted to an unconstrained optimization problem by means of a nonlinear projection operator that adds a degree of freedom to be tuned (a linear feedback gain). The objective of the tuning (or design of the feedback gain) is to massage the dynamics to either stabilize them or reduce their sensitivity to perturbations of the control input. In this paper, we generalize this method by using a projection operator with a general nonlinear feedback gain. This is advantageous in scenarios where, for example, a nonlinear feedback gain is known to stabilize the dynamics of an unstable system. It is worth to note that our mathematical derivations are slightly different from those in [29]. In our approach, we

exploit the system operator $\mathcal{H}$ defined (and analyzed) in the previous section (see the chapter on Substitution Approach for more details). This allows us to use previous results throughout the derivations.

Similar to the other methods, we first show how the OCP (6.3) is converted to an unconstrained optimization problem. Afterwards, what remains is simply the calculation of the gradient and the Hessian of the resulting unconstrained cost functional. Although the calculations are tedious, they really boil down to the proper application of the chain rule. We start by defining and analyzing the projection operator similar to [29], but with a nonlinear feedback gain.

9.1 Projection Operator

Define the projection operator $\mathcal{P}$ that acts on a state-control pair $\hat{z} := (\hat{x}, \hat{u})$, which does not have to satisfy the system dynamics, to yield another state-control pair $z := (x, u)$ that satisfies the system dynamics

$$z = \mathcal{P}(\hat{z}) \Longleftrightarrow \begin{cases} x = \mathcal{H}(u) \\ \\ u = \hat{u} + g(\hat{x} - x), \end{cases} \tag{9.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ is a twice differentiable function such that $g(0) = 0$, and $\mathcal{H}$ is the system operator defined in (8.1). Note that the projection operator in [29] is obtained by setting $g(x - \hat{x}) = K(x - \hat{x})$, where K is a linear gain (that can be time-varying). Observe that all trajectories of the dynamics (that is, the state-control pair (x, u) that satisfies $x = \mathcal{H}(u)$) are fixed points of the mapping $\mathcal{P}$. In other words, if $\hat{x} = \mathcal{H}(\hat{u})$ then $\hat{z} = \mathcal{P}(\hat{z})$. Note also that $\mathcal{P}$ is a projection operator because $\mathcal{P} \circ \mathcal{P} = \mathcal{P}$. Therefore, the

OCP in (6.3) can be converted to an unconstrained optimization problem as follows

$$\begin{array}{ll}\underset{\hat{z}}{\text{minimize}} & J(\hat{z}) \\ \text{subject to} & \mathcal{C}(\hat{z}) = 0,\end{array} \iff \begin{array}{ll}\underset{\hat{z}}{\text{minimize}} & J(\hat{z}) \\ \text{subject to} & \hat{z} = \mathcal{P}(\hat{z}),\end{array} \iff \underset{\hat{z}}{\text{minimize}} \;\; J\big(\mathcal{P}(\hat{z})\big).$$

Therefore the new unconstrained cost functional is

$$\mathcal{J}(\hat{z}) = J\left(\mathcal{P}(\hat{z})\right). \tag{9.2}$$

With the resulting unconstrained optimization problem at hand, we can exploit the unconstrained optimization techniques explained in Section 6.2 to develop the numerical methods as follows: given the current iterate $\hat{z}_k$, we calculate an update direction, denoted here by $\underline{z}_k$, and obtain the next iterate using some step size α_k as

$$\hat{z}_{k+1} = \hat{z}_k + \alpha_k \underline{z}_k. \tag{9.3}$$

The update direction depends on the gradient and/or Hessian of $\mathcal{J}$ which, naturally, depend on the first and second directional derivatives of $\mathcal{P}$. Table 9.1 summarizes the results of the calculations carried out in 10.D. It shows the formulas for the first directional derivative, its adjoint, and the second directional derivative of $\mathcal{P}$. The formulas are written in both operator form and differential equations form. We now proceed to calculate the gradient and Hessian of $\mathcal{J}$.

9.2 Gradient of $\mathcal{J}$

The new cost functional is really the composition between the projection operator $\mathcal{P}$ and the original cost functional J. Using the chain rule, we calculate the directional derivative of $\mathcal{J}$, evaluated at $\hat{z}_k := (\hat{x}_k, \hat{u}_k)$ acting on $\underline{z} := (\underline{x}, \underline{u})$ as

$$\partial\mathcal{J}_{\hat{z}_k}(\underline{z}) = \partial J_{\mathcal{P}(\hat{z}_k)}\big(\partial\mathcal{P}_{\hat{z}_k}(\underline{z})\big) = \left\langle \partial J_{\mathcal{P}(\hat{z}_k)}, \partial\mathcal{P}_{\hat{z}_k}(\underline{z}) \right\rangle = \left\langle \partial\mathcal{P}^*_{\hat{z}_k}\left(\partial J_{z_k}\right), \underline{z} \right\rangle =: \left\langle \partial\mathcal{J}_{\hat{z}_k}, \underline{z} \right\rangle, \tag{9.4}$$

Table 9.1: Projection Operator Derivatives and Adjoint. This table shows the expressions of the first and second directional derivatives of the projection operator $\mathcal{P}$ evaluated at a given $\hat{z}_k := (\hat{x}_k, \hat{u}_k)$. Furthermore, the adjoint of the directional derivative evaluated at $\hat{z}_k$ is also shown. The expressions are given in their operator forms and their associated differential equations. Note that $\mathcal{H}, \partial\mathcal{H}, \partial\mathcal{H}^*$ and $\partial^2\mathcal{H}$ are given in Table 8.1. We use $\mathcal{F}_k$ and $\mathcal{G}_k$ to denote the second derivative operators of f at z_k and g at $\hat{x}_k - x_k$, respectively. Refer to Appendix 10.D for details.

	Notation	Operator Form	Differential Equations Form
Projection Operator	$(x_k, u_k) = \mathcal{P}(\hat{x}_k, \hat{u}_k)$ $\left(z_k = \mathcal{P}(\hat{z}_k)\right)$	$\begin{cases} x_k = \mathcal{H}(u_k) \\ u_k = \hat{u}_k + g(\hat{x}_k - x_k) \end{cases}$	$\begin{cases} \dot{x}_k = f(x_k, u_k);\ x_k(0) = \boldsymbol{x_0} \\ u_k = \hat{u}_k + g(\hat{x}_k - x_k) \end{cases}$
Derivative	$(\tilde{x}_k, \tilde{u}_k) = \partial\mathcal{P}_{\hat{z}_k}(\underline{x}, \underline{u})$ $\left(\tilde{z}_k = \partial\mathcal{P}_{\hat{z}_k}(\underline{z})\right)$	$\begin{cases} \tilde{x}_k = \partial\mathcal{H}_{u_k}(\tilde{u}_k) \\ \tilde{u}_k = \underline{u} + K_k(\underline{x} - \tilde{x}_k) \end{cases}$	$\begin{cases} \dot{\tilde{x}}_k = (A_k - B_k K_k)\tilde{x}_k + B_k(\underline{u}_k + K_k \underline{x}_k) \\ \tilde{u}_k = \underline{u} + K_k(\underline{x} - \tilde{x}_k) \end{cases}$ $\tilde{x}_k(0) = 0;$
Adjoint	$(\tilde{\chi}_k, \tilde{\mu}_k) = \partial\mathcal{P}^*_{\hat{z}_k}(\chi, \mu)$ $\left(\tilde{\zeta}_k = \partial\mathcal{P}^*_{\hat{z}_k}(\zeta)\right)$	$\begin{cases} \tilde{\chi}_k = K_k^* \tilde{\mu}_k \\ \tilde{\mu}_k = \mu + \partial\mathcal{H}^*_{u_k}(\chi - \tilde{\chi}_k) \end{cases}$	$\begin{cases} \tilde{\chi}_k = K_k^*(\mu + B_k^* \lambda_k) \\ \tilde{\mu}_k = \mu + B_k^* \lambda_k \\ \dot{\lambda}_k = -(A_k - B_k K_k)^* \lambda_k - (\chi - K_k^* \mu) \end{cases}$ $\lambda_k(T) = 0$
Second Derivative	$(\bar{x}_k, \bar{u}_k) = \partial^2\mathcal{P}_{\hat{z}_k}(\underline{x}, \underline{u})$ $\left(\bar{z}_k = \partial^2\mathcal{P}_{\hat{z}_k}(\underline{z})\right)$	$\begin{cases} \bar{x}_k = \partial\mathcal{H}_{u_k}(\bar{u}_k) + \partial^2\mathcal{H}_{u_k}(\tilde{u}_k) \\ \bar{u}_k = (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) - K_k \bar{x}_k \end{cases}$	$\begin{cases} \dot{\bar{x}}_k = (A_k - B_k K_k)\bar{x}_k + \tilde{z}_k^* \mathcal{F}_k \tilde{z}_k \\ \qquad\qquad + B_k(\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) \\ \bar{u}_k = (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) - K_k \bar{x}_k \end{cases}$ $\bar{x}_k(0) = 0$

where $z_k := (x_k, u_k) := \mathcal{P}(\hat{z}_k)$, ∂J_{z_k} is given in (6.15), and $\partial \mathcal{P}^*_{\hat{z}_k}$ is the adjoint of $\partial \mathcal{P}_{\hat{z}_k}$ and is given in Table 9.1. Equation (9.4) gives the expression of the gradient of $\mathcal{J}$ at $\hat{z}_k$ in operator form in terms of J and $\mathcal{P}$. Using (6.15) and Table 9.1, we can rewrite the gradient in differential equations form as

$$\partial \mathcal{J}_{\hat{z}_k}(\underline{z}) = \left\langle \partial \mathcal{P}^*_{\hat{z}_k} \left(L^x_k + \mathcal{S}^*_T \phi^x_k, L^u_k \right), \underline{z} \right\rangle = \left\langle \begin{bmatrix} K^*_k \\ I_m \end{bmatrix} \left(L^u_k + B^*_k \lambda_k \right), \underline{z} \right\rangle =: \left\langle \partial \mathcal{J}_{\hat{z}_k}, \underline{z} \right\rangle, \quad (9.5)$$

where $K_k := \partial g_{\hat{x}_k - x_k}$, and the intermediate variable λ_k (which is similar to the costate variable and Lagrange multiplier) is calculated by using Table 9.1

$$\dot{\lambda}_k = -(A_k - B_k K_k)^* \lambda_k - (L^x_k + \mathcal{S}^*_T \phi^x_k - K^*_k L^u_k); \quad \lambda_k(T) = 0,$$

which after invoking Appendix 10.E to get rid of $\mathcal{S}^*_T$ yields

$$\dot{\lambda}_k = -(A_k - B_k K_k)^* \lambda_k - (L^x_k - K^*_k L^u_k); \quad \lambda_k(T) = \phi^x_k. \quad (9.6)$$

Therefore, the gradient of $\mathcal{J}$ at $\hat{z}_k$ is given by (9.5) where λ_k is given by (9.6) and $z_k = \mathcal{P}(\hat{z}_k)$. The necessary conditions of optimality are obtained by setting the gradient to zero. It is easy to check that necessary conditions of optimality are the same as those given in (7.4).

9.3 Hessian of $\mathcal{J}$

The Hessian can be calculated by applying the chain rule on (9.4) to obtain

$$\begin{aligned} \partial^2 \mathcal{J}_{\hat{z}_k}(\underline{z}) &= \left\langle \partial^2 J_{\mathcal{P}(\hat{z}_k)} \partial \mathcal{P}_{\hat{z}_k} \underline{z}, \partial \mathcal{P}_{\hat{z}_k} \underline{z} \right\rangle + \left\langle \partial J_{\mathcal{P}(\hat{z}_k)}, \partial^2 \mathcal{P}_{\hat{z}_k}(\underline{z}) \right\rangle \\ &= \left\langle \partial \mathcal{P}^*_{\hat{z}_k} \partial^2 J_{z_k} \partial \mathcal{P}_{\hat{z}_k} \underline{z}, \underline{z} \right\rangle + \left\langle \partial J_{z_k}, \bar{z}_k \right\rangle, \end{aligned} \quad (9.7)$$

where $z_k := \mathcal{P}(\hat{z}_k)$ and $\bar{z}_k := \partial^2 \mathcal{P}_{\hat{z}_k}(\underline{z})$. Note that $\partial \mathcal{P}_{\hat{z}_k}, \partial \mathcal{P}^*_{\hat{z}_k}$ and $\partial^2 \mathcal{P}_{\hat{z}_k}$ are all given in Table 9.1. The rest of this section examines the second term in (9.7). In fact, the

differential equation form of $\partial^2\mathcal{P}_{\hat{z}_k}$ in Table 9.1 can be rewritten using the time derivative operator $\mathcal{D}_0$ as

$$\bar{z}_k = \begin{bmatrix} I_n \\ -K_k \end{bmatrix} \Big(\mathcal{D}_0 - (A_k - B_k K_k)\Big)^{-1} \Big(\tilde{z}_k^* \mathcal{F}_k \tilde{z}_k + B_k(\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k)\Big) + \begin{bmatrix} 0 \\ (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) \end{bmatrix}, \tag{9.8}$$

where $\tilde{z}_k := (\tilde{x}_k, \tilde{u}_k) = \partial\mathcal{P}_{\hat{z}_k}(\underline{z})$. Note that $\mathcal{F}_k$ and $\mathcal{G}_k$ are the second derivative operators of f at z_k and g at $\hat{x}_k - x_k$, respectively (refer to Section 6.1.5 for more details). To express $\langle \partial J_{z_k}, \bar{z}_k \rangle$ in terms of $\underline{z}$, we substitute for ∂J_{z_k} from (6.15) and $\bar{z}_k$ from (9.8) to obtain

$$\begin{aligned}
\langle \partial J_{z_k}, \bar{z}_k \rangle &= \left\langle -\Big(\mathcal{D}_T + (A_k - B_k K_k)^*\Big)^{-1} \begin{bmatrix} I_n & -K_k^* \end{bmatrix} \begin{bmatrix} L_k^x + \mathcal{S}_T^* \phi_k^x \\ L_k^u \end{bmatrix}, \tilde{z}_k^* \mathcal{F}_k \tilde{z}_k \right\rangle \\
&\quad + \left\langle -B_k^* \Big(\mathcal{D}_T + (A_k - B_k K_k)^*\Big)^{-1} \begin{bmatrix} I_n & -K_k^* \end{bmatrix} \begin{bmatrix} L_k^x + \mathcal{S}_T^* \phi_k^x \\ L_k^u \end{bmatrix}, (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) \right\rangle \\
&\quad + \left\langle \begin{bmatrix} L_k^x + \mathcal{S}_T^* \phi_k^x \\ L_k^u \end{bmatrix}, \begin{bmatrix} 0 \\ (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) \end{bmatrix} \right\rangle \\
&= \langle \lambda_k, \tilde{z}_k^* \mathcal{F}_k \tilde{z}_k \rangle + \langle B_k^* \lambda_k + L_k^u, (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) \rangle ,
\end{aligned}$$

where λ_k is the intermediate variable defined in (9.6). By letting $\theta_k := B_k^* \lambda_k + L_k^u$ and invoking the inner product property of the second derivative operators (6.9), we obtain

$$\begin{aligned}
\langle \partial J_{z_k}, \bar{z}_k \rangle &= \langle \mathcal{F}_k \lambda_k \tilde{z}_k, \tilde{z}_k \rangle + \langle \mathcal{G}_k \theta_k (\underline{x} - \tilde{x}_k), \underline{x} - \tilde{x}_k \rangle \\
&= \langle \mathcal{F}_k \lambda_k \partial\mathcal{P}_{\hat{z}_k} \underline{z}, \partial\mathcal{P}_{\hat{z}_k} \underline{z} \rangle + \left\langle \mathcal{G}_k \theta_k \begin{bmatrix} I_n & 0 \end{bmatrix} (\underline{z} - \tilde{z}_k), \begin{bmatrix} I_n & 0 \end{bmatrix} (\underline{z} - \tilde{z}_k) \right\rangle \\
&= \langle \partial\mathcal{P}_{\hat{z}_k}^* \mathcal{F}_k \lambda_k \partial\mathcal{P}_{\hat{z}_k} \underline{z}, \underline{z} \rangle + \left\langle \mathcal{G}_k \theta_k \begin{bmatrix} I_n & 0 \end{bmatrix} (I_{n+m} - \partial\mathcal{P}_{\hat{z}_k}) \underline{z}, \begin{bmatrix} I_n & 0 \end{bmatrix} (I_{n+m} - \partial\mathcal{P}_{\hat{z}_k}) \underline{z} \right\rangle \\
&= \left\langle \left(\partial\mathcal{P}_{\hat{z}_k}^* \mathcal{F}_k \lambda_k \partial\mathcal{P}_{\hat{z}_k} + (I_{n+m} - \partial\mathcal{P}_{\hat{z}_k}^*) \begin{bmatrix} I_n \\ 0 \end{bmatrix} \mathcal{G}_k \theta_k \begin{bmatrix} I_n & 0 \end{bmatrix} (I_{n+m} - \partial\mathcal{P}_{\hat{z}_k}) \right) \underline{z}, \underline{z} \right\rangle .
\end{aligned} \tag{9.9}$$

Note that $\mathcal{F}_k\lambda_k$ is a time-varying matrix defined in (7.5). Similarly, we define another time-varying matrix S_k as

$$\mathcal{G}_k(t)\theta_k(t) := \sum_{j=1}^{m} G_j^{(k)}(t)\theta_j^{(k)}(t) = \sum_{j=1}^{m} \partial^2 g_j\big(\hat{x}_k(t) - x_k(t)\big)\theta_j^{(k)}(t) =: S_k(t). \tag{9.10}$$

Thus, using the matrices W_k from (7.5) and S_k from (9.10), we can rewrite (9.9) as

$$\langle \partial J_{z_k}, \bar{z}_k \rangle = \left\langle \left(\partial \mathcal{P}^*_{\hat{z}_k} \left(\begin{bmatrix} W_k^{xx} + S_k & W_k^{xu} \\ W_k^{ux} & W_k^{uu} \end{bmatrix} \right) \partial \mathcal{P}_{\hat{z}_k} + \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} - \partial \mathcal{P}^*_{\hat{z}_k} \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \partial \mathcal{P}_{\hat{z}_k} \right) \underline{z}, \underline{z} \right\rangle$$

Finally, by substituting $\langle \partial J_{z_k}, \bar{z}_k \rangle$ in (9.7) and using the expression of $\partial^2 J_{z_k}$ from (6.16), we obtain the expression (in operator form) that describes the action of the Hessian, evaluated at $\hat{z}_k$, on $\underline{z}$

$$\partial^2 \mathcal{J}_{\hat{z}_k}(z) = \left\langle \left(\partial \mathcal{P}^*_{\hat{z}_k} \left(\begin{bmatrix} Q_k + W_k^{xx} + S_k + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix} \right) \partial \mathcal{P}_{\hat{z}_k} + \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} - \partial \mathcal{P}^*_{\hat{z}_k} \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \partial \mathcal{P}_{\hat{z}_k} \right) \underline{z}, \underline{z} \right\rangle \tag{9.11}$$

9.4 Second Order Method for the Projection Approach

A second order method is obtained by choosing the update direction $\underline{z}_k$ according to (6.14), that is $\partial^2 \mathcal{J}_{\hat{z}_k}(\underline{z}_k) = -\partial \mathcal{J}_{\hat{z}_k}$. By substituting the expressions of the gradient (8.5) and the Hessian (8.8), we obtain

$$\begin{aligned} \left(\partial \mathcal{P}^*_{\hat{z}_k} \left(\begin{bmatrix} Q_k + W_k^{xx} + S_k + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix} \right) \partial \mathcal{P}_{\hat{z}_k} + \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} - \partial \mathcal{P}^*_{\hat{z}_k} \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \right. \\ \left. - \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \partial \mathcal{P}_{\hat{z}_k} \right) \underline{z}_k = - \begin{bmatrix} K_k^* \\ I_m \end{bmatrix} \left(L_k^u + B_k^* \lambda_k \right). \end{aligned}$$

To obtain the differential equations that produce the update direction $\underline{z}_k$, we proceed as follows. Recall that $\theta_k := L_k^u + B_k^* \lambda_k$ and let $\tilde{z}_k := (\tilde{x}_k, \tilde{u}_k) = \partial \mathcal{P}_{\hat{z}_k}(\underline{z}_k)$, then

$$\partial \mathcal{P}^*_{\hat{z}_k} \left(\begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix} \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix} - \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} \underline{x}_k \\ \underline{u}_k \end{bmatrix} - \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix} \right) \right) + \begin{bmatrix} S_k & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} \underline{x}_k \\ \underline{u}_k \end{bmatrix} - \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix} \right) = - \begin{bmatrix} K_k^* \\ I_m \end{bmatrix} \theta_k. \tag{9.12}$$

For the sake of simplicity in the remaining mathematical manipulations, define the matrix C_k as

$$C_k := \begin{bmatrix} Q_k + W_k^{xx} + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k + W_k^{xu} \\ N_k^* + W_k^{ux} & R_k + W_k^{uu} \end{bmatrix}. \tag{9.13}$$

Then, by recalling that $\tilde{u}_k = \underline{u}_k + K_k(\underline{x}_k - \tilde{x}_k)$ (refer to Table 9.1), and by defining

$$d_k := \begin{bmatrix} K_k & I_m \end{bmatrix} \begin{bmatrix} \underline{x}_k \\ \underline{u}_k \end{bmatrix}, \tag{9.14}$$

we can rewrite (9.12) as

$$\partial \mathcal{P}^*_{\hat{z}_k} \left(C_k \begin{bmatrix} I_n \\ -K_k \end{bmatrix} \tilde{x}_k + C_k \begin{bmatrix} 0 \\ I_m \end{bmatrix} d_k - \begin{bmatrix} S_k \\ 0 \end{bmatrix} (\underline{x}_k - \tilde{x}_k) \right) + \begin{bmatrix} S_k \\ 0 \end{bmatrix} (\underline{x}_k - \tilde{x}_k) = - \begin{bmatrix} K_k^* \\ I_m \end{bmatrix} \theta_k,$$

or

$$\begin{bmatrix} \tilde{\chi}_k \\ \tilde{\mu}_k \end{bmatrix} + \begin{bmatrix} S_k \\ 0 \end{bmatrix} (\underline{x}_k - \tilde{x}_k) = - \begin{bmatrix} K_k^* \\ I_m \end{bmatrix} \theta_k, \tag{9.15}$$

where

$$\begin{bmatrix} \tilde{\chi}_k \\ \tilde{\mu}_k \end{bmatrix} := \partial \mathcal{P}^*_{\hat{z}_k} \left(\begin{bmatrix} \chi_k \\ \mu_k \end{bmatrix} \right); \qquad \begin{bmatrix} \chi_k \\ \mu_k \end{bmatrix} := C_k \begin{bmatrix} I_n \\ -K_k \end{bmatrix} \tilde{x}_k + C_k \begin{bmatrix} 0 \\ I_m \end{bmatrix} d_k - \begin{bmatrix} S_k \\ 0 \end{bmatrix} (\underline{x}_k - \tilde{x}_k). \tag{9.16}$$

Using the differential equation form of $\partial\mathcal{P}^*_{\hat{z}_k}$ in Table 9.1, (9.16) can be rewritten as

$$\begin{aligned}\begin{bmatrix}\tilde{\chi}_k\\ \tilde{\mu}_k\end{bmatrix} &= \begin{bmatrix}K_k^*\\ I_m\end{bmatrix}\left(B_k^*\tilde{\lambda}_k + \mu_k\right)\\ &= \begin{bmatrix}K_k^*\\ I_m\end{bmatrix}\left(B_k^*\tilde{\lambda}_k + \begin{bmatrix}0 & I_m\end{bmatrix} C_k \begin{bmatrix}I_n\\ -K_k\end{bmatrix}\tilde{x}_k + \begin{bmatrix}0 & I_m\end{bmatrix} C_k \begin{bmatrix}0\\ I_m\end{bmatrix} d_k\right), \end{aligned} \tag{9.17}$$

where the second intermediate variable $\tilde{\lambda}_k$ is defined as

$$\begin{aligned}\mathcal{D}_T\tilde{\lambda}_k &= -(A_k - B_kK_k)^*\tilde{\lambda}_k - \begin{bmatrix}I_n & -K_k^*\end{bmatrix}\begin{bmatrix}\chi_k\\ \mu_k\end{bmatrix}\\ &= -(A_k - B_kK_k)^*\tilde{\lambda}_k - \begin{bmatrix}I_n & -K_k^*\end{bmatrix}\left(C_k\begin{bmatrix}I_n\\ -K_k\end{bmatrix}\tilde{x}_k + C_k\begin{bmatrix}0\\ I_m\end{bmatrix}d_k - \begin{bmatrix}S_k\\ 0\end{bmatrix}(\underline{x}_k - \tilde{x}_k)\right)\\ \mathcal{D}_T\tilde{\lambda}_k &= -(A_k - B_kK_k)^*\tilde{\lambda}_k - \Big(Q_k + W_k^{xx} + \mathcal{S}_T^*\phi_k^{xx}\mathcal{S}_T + K_k^*(R_k + W_k^{uu})K_k\Big)\tilde{x}_k + S_k(\underline{x}_k - \tilde{x}_k)\\ &\quad + \Big(K_k^*(N_k^* + W_k^{ux}) + (N_k + W_k^{xu})K_k\Big)\tilde{x}_k - (N_k + W_k^{xu})d_k + K_k^*(R_k + W_k^{uu})d_k.\end{aligned} \tag{9.18}$$

Substituting (9.17) in (9.15) yields

$$\begin{bmatrix}K_k^*\\ I_m\end{bmatrix}\left(B_k^*\tilde{\lambda}_k + \begin{bmatrix}0 & I_m\end{bmatrix} C_k \begin{bmatrix}I_n\\ -K_k\end{bmatrix}\tilde{x}_k + \begin{bmatrix}0 & I_m\end{bmatrix} C_k \begin{bmatrix}0\\ I_m\end{bmatrix} d_k + \theta_k\right) + \begin{bmatrix}S_k\\ 0\end{bmatrix}(\underline{x}_k - \tilde{x}_k) = \begin{bmatrix}0\\ 0\end{bmatrix},$$

which, by substituting for C_k and θ_k, yields two equations

$$S_k(\underline{x}_k - \tilde{x}_k) = 0 \tag{9.19}$$

$$(R_k + W_k^{uu})d_k = -\Big(N_k^* + W_k^{ux} - (R_k + W_k^{uu})K_k\Big)\tilde{x}_k - B_k^*(\lambda_k + \tilde{\lambda}_k) - L_k^u. \tag{9.20}$$

By substituting for $S_k(\underline{x}_k - \tilde{x}_k)$ and $(R_k + W_k^{uu})d_k$ in (9.18), many terms cancel out to obtain

$$\mathcal{D}_T\tilde{\lambda}_k = -A_k^*\tilde{\lambda}_k - \Big(Q_k + W_k^{xx} + \mathcal{S}_T^*\phi_k^{xx}\mathcal{S}_T - (N_k + W_k^{xu})K_k\Big)\tilde{x}_k - (N_k + W_k^{xu})d_k - B_k^*\lambda_k - L_k^u. \tag{9.21}$$

Furthermore, since (9.19) has to be satisfied regardless of the choice of the nonlinear gain g (hence S_k), then $\underline{x}_k = \tilde{x}_k$ which also implies that $\underline{u}_k = \tilde{u}_k$. This further simplifies (9.20) and (9.21) by substituting $d_k = K_k\underline{x}_k + \underline{u}_k = K_k\tilde{x}_k + \tilde{u}_k$ to obtain

$$\begin{aligned}(R_k + W_k^{uu})\tilde{u}_k &= -(N_k^* + W_k^{ux})\tilde{x}_k - B_k^*(\lambda_k + \tilde{\lambda}_k) - L_k^u \\ \mathcal{D}_T\tilde{\lambda}_k &= -A_k^*\tilde{\lambda}_k - \Big(Q_k + W_k^{xx} + \mathcal{S}_T^*\phi_k^{xx}\Big)\tilde{x}_k - (N_k + W_k^{xu})\tilde{u}_k - B_k^*\lambda_k - L_k^u.\end{aligned} \tag{9.22}$$

Finally, by combining $\tilde{z}_k = \partial\mathcal{P}_{\hat{z}_k}(\underline{z}_k)$ and (9.22), and invoking Appendix 10.E, we obtain the following linear two-point boundary value problem coupled with an algebraic constraint

$$\begin{aligned}\frac{d}{dt}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} &= \begin{bmatrix}A_k & 0 \\ -\big(Q_k + W_k^{xx}\big) & -A_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix}B_k \\ -(N_k + W_k^{xu})\end{bmatrix}\tilde{u}_k + \begin{bmatrix}0 \\ -B_k^*\lambda_k - L_k^u\end{bmatrix} \\ (R_k + W_k^{uu})\tilde{u}_k &= -\begin{bmatrix}N_k^* + W_k^{ux} & B_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{\lambda}_k\end{bmatrix} - B_k^*\lambda_k - L_k^u \\ \text{such that} \quad \tilde{x}_k(0) &= 0, \qquad \tilde{\lambda}_k(T) = \phi_k^{xx}\tilde{x}_k(T).\end{aligned} \tag{9.23}$$

The algorithm for this second order method is summarized in Algorithm 4.

Algorithm 4 Second Order Method: Projection Operator Approach

1: Start with an initial guess $(\hat{u}_1, \hat{x}_1)$ and set $k = 1$.

2: Compute the projection $(x_k, u_k) := \mathcal{P}(\hat{x}_k, \hat{u}_k)$:

$$\begin{cases} x_k = f(x_k, u_k) \\ u_k = \hat{u}_k + g(\hat{x}_k - x_k), \end{cases}$$

3: Compute:

$$\begin{aligned} A_k &= \partial_x f_{(x_k,u_k)}, \qquad B_k = \partial_u f_{(x_k,u_k)}, \\ L_k^x &= \partial_x L^*_{(x_k,u_k)}, \qquad L_k^u = \partial_u L^*_{(x_k,u_k)}, \\ Q_k &= \partial_x^2 L_{(x_k,u_k)}, \qquad R_k = \partial_u^2 L_{(x_k,u_k)}, \\ N_k &= \partial_{xu} L_{(x_k,u_k)}, \quad K_k = \partial^2 g_{\hat{x}_k - x_k}, \\ \phi_k^x &= \partial_x \phi^*_{x_k(T)}, \qquad \phi_k^{xx} = \partial_x^2 \phi_{x_k(T)}, \end{aligned}$$

4: Solve for the costate $\lambda_k(t)$:

$$\dot{\lambda}_k = -(A_k - B_k K_k)^* \lambda_k - (L_k^x - K_k L_k^u); \quad \lambda_k(T) = \phi_k^x.$$

5: Compute:

$$\begin{aligned} W_k^{xx} &= \sum_{i=1}^n \partial_x^2 f_i(x_k, u_k)\lambda_i^{(k)}; & W_k^{xu} &= \sum_{i=1}^n \partial_{xu} f_i(x_k, u_k)\lambda_i^{(k)}; \\ W_k^{ux} &= (W_k^{xu})^*; & W_k^{uu} &= \sum_{i=1}^n \partial_u^2 f_i(x_k, u_k)\lambda_i^{(k)}. \end{aligned}$$

6: Solve for $(\tilde{x}_k, \tilde{u}_k)$:

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} = \begin{bmatrix} A_k & 0 \\ -\big(Q_k + W_k^{xx}\big) & -A_k^* \end{bmatrix}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} + \begin{bmatrix} B_k \\ -(N_k + W_k^{xu}) \end{bmatrix}\tilde{u}_k + \begin{bmatrix} 0 \\ -B_k^*\lambda_k - L_k^u \end{bmatrix}$$

$$(R_k + W_k^{uu})\tilde{u}_k = -\begin{bmatrix} N_k^* + W_k^{ux} & B_k^* \end{bmatrix}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} - B_k^*\lambda_k - L_k^u$$

$$\text{such that} \quad \tilde{x}_k(0) = 0, \qquad \tilde{\lambda}_k(T) = \phi_k^{xx}\tilde{x}_k(T).$$

7: Update the control with a step size α_k:

$$\begin{bmatrix} \hat{x}_{k+1} \\ \hat{u}_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{x}_k \\ \hat{u}_k \end{bmatrix} + \alpha_k \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix}$$

8: Set $k = k + 1$ and go back to step 2. Repeat until convergence.

Chapter 10

Preconditioned Constrained-Gradient Descent

This section develops a preconditioned constrained-gradient descent (PCGD) method as an iterative numerical algorithm to solve (6.3). The building blocks of the algorithm are based on projected gradient descent methods in infinite dimensional optimization problems (for example [61]) . By utilizing the special structure of optimal control problems and preconditioning the state-control space, we achieve higher convergence rates than the well known gradient descent method [37].

Projection based methods have been widely used to numerically solve optimal control problems with constraints ([8], [44], [9] among others). These methods treat the dynamical equality constraint as part of the cost functional. That is, the states are thought of as functions of the controls within the cost functional, leaving only inequality constraints. Typically, these methods project the cost functional gradient onto the feasible set defined by the inequality constraints. The PCGD method developed in this section, on the other hand, projects onto the dynamical equality constraint itself, thus treating the states and controls within the cost functional as two independent variables. This allows

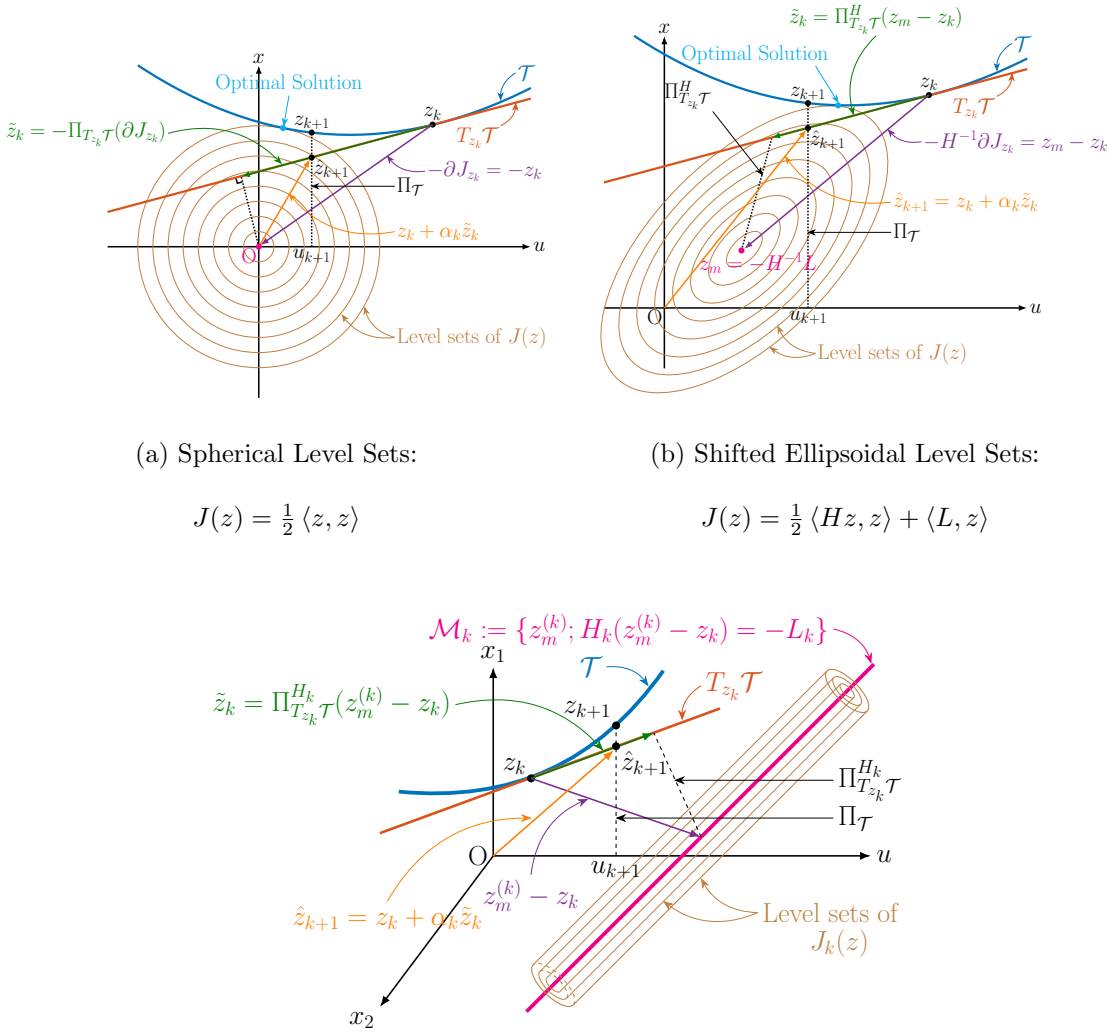
us to precondition the state-control space to boost the convergence rate of the method. This is possible because, generally, in optimal control problems, the complexity arises in the nonlinear dynamics; whereas, the cost functionals are typically simpler (such as quadratic functionals). We show that the PCGD method yields a particular algorithm that lies under the family of Quasi-Newton methods explained by [29] (which is generalized in the previous section). In fact, we carry the dynamical constraints throughout without the calculation of second derivatives of the dynamics (as second order methods require).

10.1 Geometric Description of the PCGD

We start by providing the geometric description of the PCGD algorithm. For clarity of exposure, we consider three different cases in increasing generality: (a) quadratic cost functional with spherical level sets, (b) quadratic cost functional with shifted ellipsoidal level sets and (c) general positive semi-definite cost functional. For simplicity, the geometric description is given, in the absence of a terminal cost, using a finite-dimensional geometric demonstration. It should be noted that the geometric demonstration is not meant to provide a rigorous proof but to build a geometric intuition.

10.1.1 Cost Functional with Spherical Level Sets

First, consider the simplest case where the cost functional has spherical level sets (centered around the origin). That is, we set the cost functional in (6.3) to $J(z) = \frac{1}{2}\langle z, z\rangle$ (see Figure 10.1-a). Observe that $\partial J_{z_k} = z_k$ and $\partial^2 J_{z_k} = I_{n+m}$, where I_{n+m} is the identity matrix of size $n + m$. The algorithm proceeds as follows: given the current iterate z_k,



(a) Spherical Level Sets:

$J(z) = \frac{1}{2}\langle z, z\rangle$

(b) Shifted Ellipsoidal Level Sets:

$J(z) = \frac{1}{2}\langle Hz, z\rangle + \langle L, z\rangle$

(c) General Positive Semi-definite Cost Functional

Figure 10.1: The three figures show the dynamical constraints manifold $\mathcal{T}$, the tangent space $T_{z_k}\mathcal{T}$ at the current iterate z_k and the level sets of three different cost functionals. The main idea is to project the vector anchored at the current iterate z_k and pointing towards the minimum of the unconstrained cost functional onto the tangent space. For (a), the unconstrained minimum is the origin and the projection is orthogonal since the level sets are spherical. For (b), the unconstrained minimum is shifted to z_m and the projection is oblique to respect the skewness of the ellipsoidal level sets governed by the positive definite Hessian H. For (c), the unconstrained minimum may degenerate into an affine space $\mathcal{M}_k$ due to the possible non-definiteness of the Hessian. This makes the level sets take an elliptic cylindrical shape (thus the need for a three dimensional representation). For all scenarios, an additional procedure is carried out, after taking a step in the tangent space, to force the dynamical constraints to be satisfied. This is achieved by the projection $\Pi_{\mathcal{T}}$ which takes the u-component and computes the corresponding state x.

compute the next iterate z_{k+1} using some step size α_k as

$$\begin{cases} \tilde{z}_k = -\Pi_{T_{z_k}\mathcal{T}}(\partial J_{z_k}) \\ \hat{z}_{k+1} = z_k + \alpha_k \tilde{z}_k \\ z_{k+1} = \Pi_{\mathcal{T}}(\hat{z}_{k+1}), \end{cases} \tag{10.1}$$

where $\Pi_{T_{z_k}\mathcal{T}}$ is an orthogonal projection operator that projects onto the tangent space $T_{z_k}\mathcal{T}$, and $\Pi_{\mathcal{T}}$ is a nonlinear projection operator that projects onto the dynamical constraint set $\mathcal{T}$ (see Section 6.1.6 for more details). In words, the gradient $\partial J_{z_k} = z_k$ is first projected onto the tangent space, which really corresponds to the linearized dynamics around z_k. The negative of the projected gradient constitutes the update direction along which we "move" using a step size α_k to obtain $\hat{z}_{k+1}$. Finally, we project $\hat{z}_{k+1}$ onto the constraint set using $\Pi_{\mathcal{T}}$ to force the next iterate z_{k+1} to be a trajectory of the nonlinear dynamics. Therefore, in this case, the PCGD is really the projected-gradient descent method followed by a projection $\Pi_{\mathcal{T}}$. Figure 10.1-a provides a geometric picture of (10.1). The spherical nature of the level sets gives rapid convergence properties of the projected-gradient descent algorithm. In fact, if the dynamical system were linear (that is, $\mathcal{T}$ in Figure 10.1-a is a straight line), it is easy to see geometrically that convergence is achieved in only one step.

10.1.2 Cost Functional with Shifted Ellipsoidal Level Sets

We generalize the previous method to a linear-quadratic cost functional $J(z) = \frac{1}{2}\langle Hz, z\rangle + \langle L, z\rangle$ where H is positive definite. Observe that $\partial J_{z_k} = Hz_k + L$, $\partial^2 J_{z_k} = H$, and $z_m := -H^{-1}L$ is the unconstrained minimum of the cost functional J. The level sets of J are now elliptical and are centered around the unconstrained minimum z_m (see Figure 10.1-b). Applying a projected-gradient descent method here is likely to result in slow convergence due to the skewness of the level sets. The main advantage of treating

the cost functional and dynamical constraints separately is that it allows to precondition the original state-control space (x, u) based on the cost functional only. More precisely, carrying out an affine transformation on the z-space defined as

$$z' = T(z) := H^{\frac{1}{2}}(z - z_m) \iff z = T^{-1}(z') = z_m + H^{-\frac{1}{2}}z', \tag{10.2}$$

yields a new cost functional $J'(z') = J(z)$ that has spherical level sets, because

$$\begin{aligned}
J'(z') = J(z) &= \frac{1}{2}\langle Hz, z\rangle + \langle L, z\rangle \\
&= \frac{1}{2}\left\langle Hz_m + H^{\frac{1}{2}}z', z_m + H^{-\frac{1}{2}}z'\right\rangle + \left\langle L, z_m + H^{-\frac{1}{2}}z'\right\rangle \\
&= \frac{1}{2}\left\langle -L + H^{\frac{1}{2}}z', -H^{-1}L + H^{-\frac{1}{2}}z'\right\rangle + \left\langle L, -H^{-1}L + H^{-\frac{1}{2}}z'\right\rangle \\
&= \frac{1}{2}\left\langle H^{\frac{1}{2}}z', H^{-\frac{1}{2}}z'\right\rangle - \frac{1}{2}\left\langle H^{\frac{1}{2}}z', H^{-1}L\right\rangle - \frac{1}{2}\left\langle L, H^{-\frac{1}{2}}z'\right\rangle + \frac{1}{2}\langle L, H^{-1}L\rangle \\
&\qquad + \left\langle L, H^{-\frac{1}{2}}z'\right\rangle - \langle L, H^{-1}L\rangle \\
J'(z') &= \frac{1}{2}\langle z', z'\rangle + \frac{1}{2}\langle L, z_m\rangle,
\end{aligned} \tag{10.3}$$

where the last equality follows by exploiting the fact that H is symmetric positive definite and thus $H, H^{-1}, H^{-\frac{1}{2}}$, and $H^{\frac{1}{2}}$ are all symmetric. Recall that $z_m := -H^{-1}L$, and thus the second term in (10.3) is a constant. Since the Hessian of J' is the identity matrix, the level sets in the new state-control space z' are spherical. Therefore, applying a projected-gradient descent in the transformed space yields a faster convergence rate. The PCGD method in this case can thus be written as follows: given the current iterate z_k, we obtain the next iterate z_{k+1} using some step size α_k as

$$\begin{cases}
z'_k = T(z_k) \\
\tilde{z}'_k = -\Pi_{T_{z'_k}\mathcal{T}'}(\partial J'_{z'_k}) \\
\hat{z}'_{k+1} = z'_k + \alpha_k \tilde{z}'_k \\
\hat{z}_{k+1} = T^{-1}(\hat{z}'_{k+1}) \\
z_{k+1} = \Pi_{\mathcal{T}}(\hat{z}_{k+1}),
\end{cases} \tag{10.4}$$

where $\partial J'_{z'_k} = z'_k$, and $\mathcal{T}'$ is the dynamical constraint set in the transformed z'-space. The key idea here is that the projection of the gradient onto the tangent space is carried out in the transformed z'-space where the level sets are spherical rather than elliptical. The rest of this section shows that the orthogonal projection in (10.4) is equivalent to an oblique projection (see Section 6.1.6 for details) in the original z-space. Using the definition of the orthogonal projection, $\tilde{z}'_k$ in (10.4) can be written as

$$\begin{aligned} \tilde{z}'_k \;=\; \Pi_{T_{z'_k}\mathcal{T}'}(-z'_k) \;&= \underset{\tilde{z}'}{\operatorname{argmin}} \; \frac{1}{2}\left\langle -z'_k - \tilde{z}', -z'_k - \tilde{z}' \right\rangle \;= \underset{\tilde{z}'}{\operatorname{argmin}} \; \frac{1}{2}\left\langle z'_k + \tilde{z}', z'_k + \tilde{z}' \right\rangle, \\ &\text{s.t.} \quad \tilde{z}' \in T_{z'_k}\mathcal{T}' \qquad\qquad \text{s.t.} \quad \partial \mathcal{C}'_{z'_k}(\tilde{z}') = 0 \end{aligned} \tag{10.5}$$

where $\mathcal{C}'(z') := \mathcal{C}(z)$ is the dynamical constraint operator in the transformed z'-space. Note that if $\tilde{z}'$ is in the tangent space, then it has to satisfy the linearized dynamics, that is $\partial \mathcal{C}'_{z'_k}(\tilde{z}') = 0$. Using the chain rule, we have $\partial \mathcal{C}_{z_k}(\tilde{z}) = \partial \mathcal{C}'_{z'_k}(H^{\frac{1}{2}}\tilde{z})$. By letting $\tilde{z}' := H^{\frac{1}{2}}\tilde{z}$, we obtain $\partial \mathcal{C}_{z_k}(\tilde{z}) = \partial \mathcal{C}'_{z'_k}(\tilde{z}')$. Then, (10.5) can be rewritten in the original z-space as

$$\begin{aligned} \tilde{z}'_k &= \underset{H^{\frac{1}{2}}\tilde{z}}{\operatorname{argmin}} \; \frac{1}{2}\left\langle H^{\frac{1}{2}}(z_k - z_m) + H^{\frac{1}{2}}\tilde{z}, H^{\frac{1}{2}}(z_k - z_m) + H^{\frac{1}{2}}\tilde{z} \right\rangle \\ &\qquad \text{s.t.} \quad \partial \mathcal{C}_{z_k}(\tilde{z}) = 0 \\ &= H^{\frac{1}{2}} \underset{\tilde{z}}{\operatorname{argmin}} \; \frac{1}{2}\left\langle H\Big((z_k - z_m) + \tilde{z}\Big), (z_k - z_m) + \tilde{z} \right\rangle \\ &\qquad \text{s.t.} \quad \partial \mathcal{C}_{z_k}(\tilde{z}) = 0 \\ \tilde{z}'_k &= -H^{\frac{1}{2}}\Pi^H_{T_{z_k}\mathcal{T}}(z_k - z_m). \end{aligned} \tag{10.6}$$

Finally, (10.4) can be rewritten in the original z-space as

$$\begin{cases} \tilde{z}_k = -\Pi^H_{T_{z_k}\mathcal{T}}(z_k - z_m) = -\Pi^H_{T_{z_k}\mathcal{T}}(H^{-1}\partial J_{z_k}) \\ \hat{z}_{k+1} = z_k + \alpha_k \tilde{z}_k \\ z_{k+1} = \Pi_{\mathcal{T}}\left(\hat{z}_{k+1}\right). \end{cases} \tag{10.7}$$

There are two key differences between (10.1) and (10.7). First, the projection is now oblique (rather than orthogonal) with a direction defined by the Hessian H. Second, $H^{-1}\partial J_{z_k}$ is now projected rather than the gradient ∂J_{z_k}. In fact, the gradient is first preconditioned before performing the projection. More precisely, the preconditioning maneuvers the gradient at z_k to point towards the minimum z_m of the unconstrained cost functional J as illustrated geometrically in Figure 10.1-b (since $H^{-1}\partial J_{z_k} = z_k - z_m$). Indeed, the preconditioning of the gradient together with the oblique projection have an effect of shifting the level sets in Figure 10.1-b to the origin and "de-skewing" them to become spherical as in Figure 10.1-a, and thus boosting the convergence of the PCGD method as compared to the traditional gradient descent.

10.1.3 General Positive Semi-Definite Cost Functional

Now we consider the general case for any cost functional J described in Section 6.3. First, consider the case where there is no terminal cost, that is $\phi = 0$. We assume that the Hessian at z_k, $H_k := \partial^2 J_k$, is generally positive semi-definite. The quadratic approximation of the unconstrained cost functional J around the current iterate z_k is denoted by J_k and can be written as

$$J_k(z) = \frac{1}{2}\langle H_k(z - z_k), z - z_k\rangle + \langle L_k, z - z_k\rangle + J(z_k), \tag{10.8}$$

where $L_k := \partial J_{z_k}$. In general, as opposed to the previous two scenarios, J_k doesn't have a unique minimum because, possibly, H_k might have a nontrivial nullspace. In fact, the unconstrained minimum of J_k at the k^{th} iteration, denoted by $z_m^{(k)}$, satisfies

$$H_k(z_m^{(k)} - z_k) = -L_k. \tag{10.9}$$

This is obtained by setting the gradient of J_k to zero. Clearly, when H_k is positive semi-definite, the solution to (10.9) degenerates to an affine subspace denoted by $\mathcal{M}_k$, that is

$$\mathcal{M}_k := \{z_m^{(k)}; H_k(z_m^{(k)} - z_k) = -L_k\}. \tag{10.10}$$

This is illustrated in Figure 10.1-c. The degeneration of the unconstrained minimum into an affine subspace causes the level sets to become elliptic-cylindrical (see Figure 10.1-c). The PCGD method in this case is similar to (10.7)

$$\begin{cases} \tilde{z}_k = -\Pi^{H_k}_{T_{z_k}\mathcal{T}}(z_k - z_m^{(k)}) \\ \hat{z}_{k+1} = z_k + \alpha_k \tilde{z}_k \\ z_{k+1} = \Pi_{\mathcal{T}}\left(\hat{z}_{k+1}\right). \end{cases} \tag{10.11}$$

The two differences (generalizations) are that the oblique projection direction, governed by H_k, changes every iteration, and the unconstrained minimum $z_m^{(k)} \in \mathcal{M}_k$ also changes every iteration and is not unique. However, we show next that the subsequent iterate z_{k+1} does not depend on a particular choice of $z_m^{(k)} \in \mathcal{M}_k$. Using the definition of the oblique projection (refer to Section 6.1.6), we have

$$\begin{aligned} \tilde{z}_k = \underset{\tilde{z}}{\operatorname{argmin}} \frac{1}{2} \left\langle H_k(z_m^{(k)} - z_k - \tilde{z}), z_m^{(k)} - z_k - \tilde{z} \right\rangle \\ \text{s.t.} \quad \partial \mathcal{C}_{z_k} \tilde{z} = 0. \end{aligned} \tag{10.12}$$

This is a linear-quadratic optimization problem whose necessary conditions of optimality can be derived by constructing the Lagrangian at the k^{th} iteration as

$$\mathcal{L}^{(k)}(\tilde{z}, \tilde{\lambda}) := \frac{1}{2} \left\langle H_k(z_m^{(k)} - z_k - \tilde{z}), z_m^{(k)} - z_k - \tilde{z} \right\rangle + \left\langle \tilde{\lambda}, \partial \mathcal{C}_{z_k} \tilde{z} \right\rangle.$$

Thus the necessary conditions of optimality are simply obtained by setting the gradient of $\mathcal{L}^{(k)}$ to zero. We have

$$\begin{aligned} \partial \mathcal{L}^{(k)}_{(\tilde{z}_k, \tilde{\lambda}_k)}(w, \xi) &= \left\langle H_k(\tilde{z}_k + z_k - z_m^{(k)}), w \right\rangle + \left\langle \partial \mathcal{C}^*_{z_k} \tilde{\lambda}_k, w \right\rangle + \langle \xi, \partial \mathcal{C}_{z_k} \tilde{z}_k \rangle \\ &= \left\langle \begin{bmatrix} H_k(\tilde{z}_k + z_k - z_m^{(k)}) + \partial \mathcal{C}^*_{z_k} \tilde{\lambda}_k \\ \partial \mathcal{C}_{z_k} \tilde{z} \end{bmatrix}, \begin{bmatrix} w \\ \xi \end{bmatrix} \right\rangle =: \left\langle \partial \mathcal{L}^{(k)}_{(\tilde{z}_k, \tilde{\lambda}_k)}, \begin{bmatrix} w \\ \xi \end{bmatrix} \right\rangle \end{aligned}$$

Setting the gradient to zero yields the necessary conditions of optimality

$$\begin{bmatrix} H_k & \partial\mathcal{C}_{z_k}^* \\ \partial\mathcal{C}_{z_k} & 0 \end{bmatrix} \begin{bmatrix} \tilde{z}_k \\ \tilde{\lambda}_k \end{bmatrix} = \begin{bmatrix} H_k(z_m^{(k)} - z_k) \\ 0 \end{bmatrix}. \tag{10.13}$$

However, observe that the right hand side of (10.13) doesn't depend on $z_m^{(k)}$ since for any $z_m^{(k)} \in \mathcal{M}_k$, we have $H_k(z_m^{(k)} - z_k) = -L_k$. This shows that the next iterate doesn't depend on $z_m^{(k)} \in \mathcal{M}_k$.

Finally, the same analysis can be carried out to include the terminal costs to obtain

$$\begin{bmatrix} \partial^2 J_{z_k} & \partial\mathcal{C}_{z_k}^* \\ \partial\mathcal{C}_{z_k} & 0 \end{bmatrix} \begin{bmatrix} \tilde{z}_k \\ \tilde{\lambda}_k \end{bmatrix} = \begin{bmatrix} -\partial J_{z_k} \\ 0 \end{bmatrix}. \tag{10.14}$$

Solving (10.14) gives the update direction $\tilde{z}_k$. Note that (10.14) is written in operator form. To obtain the underlying differential equations, we substitute the expressions of the Hessian and gradient of J from (6.16) and (6.15), respectively, to obtain

$$\begin{bmatrix} Q_k + \mathcal{S}_T^* \phi_k^{xx} \mathcal{S}_T & N_k & \mathcal{D}_T + A_k^* \\ N_k^* & R_k & B_k^* \\ -\mathcal{D}_0 + A_k & B_k & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \\ \tilde{\lambda}_k \end{bmatrix} = - \begin{bmatrix} L_k^x + \mathcal{S}_T^* \phi_k^x \\ L_k^u \\ 0 \end{bmatrix},$$

where $\partial\mathcal{C}_{z_k} = \begin{bmatrix} -\mathcal{D}_0 + A_k & B_k \end{bmatrix}$. This can be rearranged and rewritten as

$$\begin{aligned} \mathcal{D}_0 \tilde{x}_k &= A_k \tilde{x}_k + B_k \tilde{u}_k \\ \mathcal{D}_T \tilde{\lambda}_k + \mathcal{S}_T^*(\phi_k^{xx} \tilde{x}_k(T) + \phi_k^x) &= -Q_k \tilde{x}_k - A_k^* \tilde{\lambda}_k - N_k \tilde{u}_k - L_k^x \\ R_k \tilde{u}_k &= -N_k^* \tilde{x}_k - B_k^* \tilde{\lambda}_k - L_k^u. \end{aligned}$$

By invoking Appendix 10.E, we can get rid of $\mathcal{S}_T^*$ and rewrite the result as a linear differential algebraic equation (DAE) where the differential equations take the form a

two point boundary value problem. We have

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} &= \begin{bmatrix}A_k & 0\\ -Q_k & -A_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} + \begin{bmatrix}B_k\\ -N_k\end{bmatrix}\tilde{u}_k + \begin{bmatrix}0\\ -L_k^x\end{bmatrix} \\ R_k\tilde{u}_k &= -\begin{bmatrix}N_k^* & B_k^*\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ \tilde{\lambda}_k\end{bmatrix} - L_k^u \\ \text{such that} \qquad \tilde{x}_k(0) &= 0 \qquad \text{and} \qquad \tilde{\lambda}_k(T) = \phi_k^x + \phi_k^{xx}\tilde{x}_k(T). \end{aligned} \tag{10.15}$$

Finally, the algorithm for the PCGD method is summarized in Algorithm 5.

10.2 Connection with the General Projection Approach

It turns out that the PCGD method is a particular realization of the family of quasi-Newton methods developed by Hauser ([29], [31], [56]) that is generalized in the previous section. The Projection Operator based Newton Method for Trajectory Optimization (PRONTO) employs a stabilizing projection operator to project the whole state-control space onto the trajectory manifold. This transforms the constrained optimization problem into an unconstrained one to be solved using a Newton method. Subsequently, a family of quasi-Newton methods can be devised (example [29]) in order to obtain descent directions when the Newton method fails to do so.

The comparison with the projection approach can be done by examining Algorithms 4 and 5. Observe that by setting $g = 0$ (which makes $\Pi_{\mathcal{T}} = \mathcal{P}$), and neglecting the matrix W_k in Algorithm 4 yields Algorithm 5. In fact, neglecting W_k is exactly what makes PCGD a quasi-newton method under the (Newton) projection approach. This follows by examining (9.11) which shows that neglecting W_k means that we are simply approximating the Hessian, and thus constructing a quasi-newton method. Although

Algorithm 5 PCGD

1: Start with an initial guess $(\hat{x}_1, \hat{u}_1)$ and set $k = 1$.

2: Compute the projection $(x_k, u_k) := \Pi_{\mathcal{T}}(\hat{x}_k, \hat{u}_k)$:

$$\begin{cases} x_k = f(x_k, u_k) \\ u_k = \hat{u}_k, \end{cases}$$

3: Given (x_k, u_k), compute :

$$\begin{aligned} A_k &= \partial_x f_{(x_k,u_k)}, & B_k &= \partial_u f_{(x_k,u_k)}, \\ L_k^x &= \partial_x L^*_{(x_k,u_k)}, & L_k^u &= \partial_u L^*_{(x_k,u_k)}, \\ Q_k &= \partial_x^2 L_{(x_k,u_k)}, & R_k &= \partial_u^2 L_{(x_k,u_k)}, \\ N_k &= \partial_{xu} L_{(x_k,u_k)}, & & \\ \phi_k^x &= \partial_x \phi^*_{x_k(T)}, & \phi_k^{xx} &= \partial_x^2 \phi_{x_k(T)}. \end{aligned}$$

4: Solve the following linear two point boundary value problem (with an algebraic constraint) to obtain $(\tilde{x}_k, \tilde{u}_k)$:

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} = \begin{bmatrix} A_k & 0 \\ -Q_k & -A_k^* \end{bmatrix}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} + \begin{bmatrix} B_k \\ -N_k \end{bmatrix}\tilde{u}_k + \begin{bmatrix} 0 \\ -L_k^x \end{bmatrix}$$

$$R_k\tilde{u}_k = -\begin{bmatrix} N_k^* & B_k^* \end{bmatrix}\begin{bmatrix} \tilde{x}_k \\ \tilde{\lambda}_k \end{bmatrix} - L_k^u$$

$$\text{such that} \qquad \tilde{x}_k(0) = 0 \qquad \text{and} \qquad \tilde{\lambda}_k(T) = \phi_k^x + \phi_k^{xx}\tilde{x}_k(T).$$

5: Update the state, control and Lagrange multiplier using a step size α_k:

$$\begin{bmatrix} x_{k+1} \\ u_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \alpha_k \begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix}.$$

6: Set $k = k + 1$ and go back to step 2. Repeat until convergence.

this approximation may seem at first somehow heuristic, the derivations carried out to develop the PCGD method gives a geometrical meaning to this approximation.

In fact, the PCGD method begins by keeping the control and state variables separate as a constraint in function space that defines a manifold. A gradient descent algorithm is then used but with a "constrained-gradient", that is, a gradient that is projected onto the tangent space of the constraint manifold. This is geometrically compelling, and has the advantage of making the required preconditioning obvious since the objective is not mixed up with the dynamical mapping as in PRONTO. In our derivation, a second projection onto the actual manifold (to yield feasible trajectories) is done after the descent direction is projected onto the tangent space. While the PCGD method can be regarded as a particular realization of the family of quasi-Newton methods of PRONTO, the derivation is geometrically more transparent and clarifies why the preconditioning (which is critical) produces faster convergence. Finally, we note that the PCGD method can be easily generalized to $g \neq 0$ to treat unstable or sensitive systems. This boils down to setting $W_k = 0$ only in Algorithm 4.

10.3 Illustrative Numerical Examples

In this section, we present two numerical examples of nonlinear optimal control problems to compare the gradient descent method (Algorithm 2) to the PCGD method (Algorithm 5). Note that the Armijo rule [2] is employed to calculate the step size α_k in all the numerical examples.

10.3.1 A Continuous Stirred-Tank Chemical Reactor

The state equations for a continuous stirred-tank chemical reactor are given below [38, Example 6.2-2]

$$\begin{aligned}\frac{d}{dt}x_1 &= -2(x_1+0.25)+(x_2+0.5)e^{\frac{25x_1}{x_1+2}}-(x_1+0.25)u,\\ \frac{d}{dt}x_2 &= 0.5-x_2-(x_2+0.5)e^{\frac{25x_1}{x_1+2}},\end{aligned} \tag{10.16}$$

$$x_1(0)=0.05, \qquad x_2(0)=0.$$

This dynamical system represents the first order, irreversible exothermic reaction controlled by the flow of a coolant, u, through a coil inserted in the reactor. The deviation from the steady state temperature and concentration are expressed by x_1 and x_2, respectively. It is required to maintain the temperature and concentration close to their steady state values without expending large amounts of control effort. The cost functional is thus given by

$$J(x,u)=\frac{1}{2}\int_0^{0.78}\left[x(t)^TQx(t)+Ru^2(t)\right]dt, \tag{10.17}$$

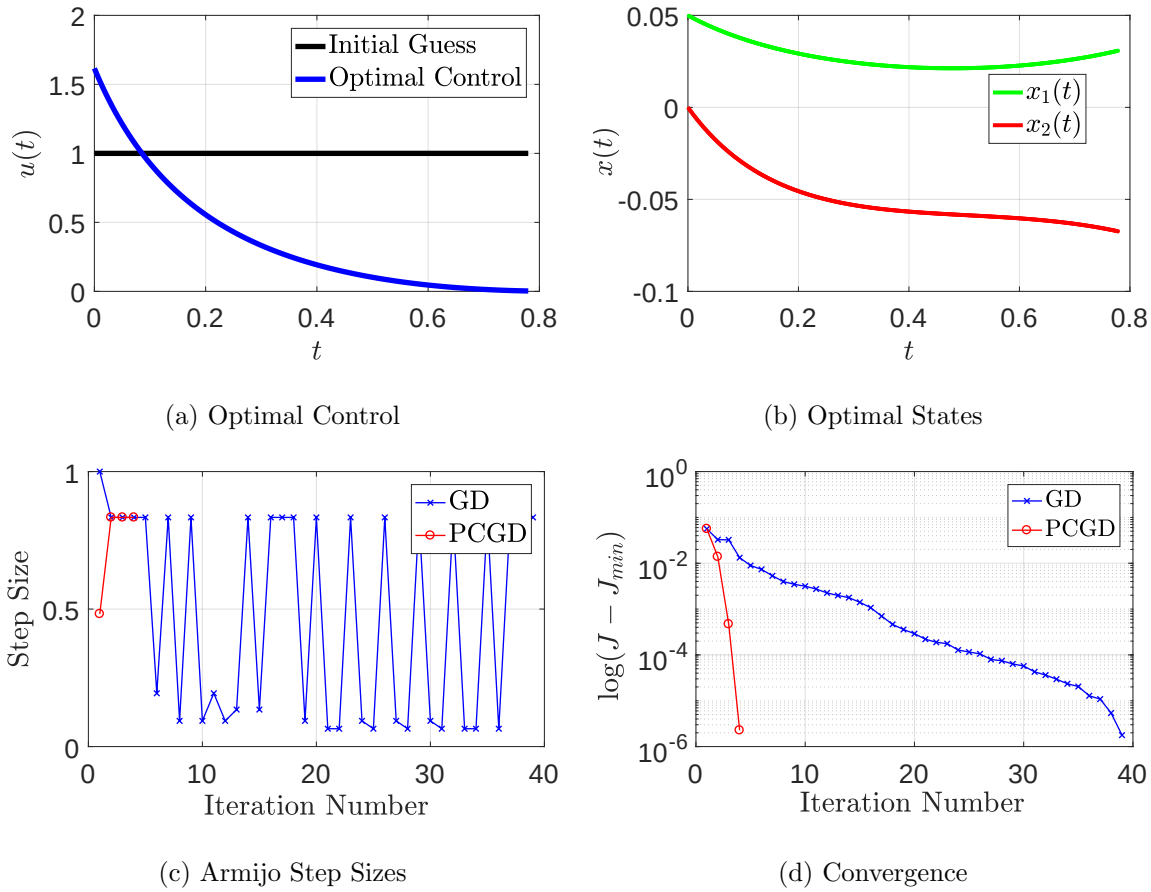
where $Q=2I$, $R=0.2$, $x=\begin{bmatrix}x_1 & x_2\end{bmatrix}^*$ and I is a 2×2 identity matrix. The results are shown in Figure 10.2. The gradient descent method takes 39 iterations to converge with considerable variations of the step size in each iteration . Whereas, the PCGD method converges in only 4 iterations with a consistent choice of the step size at each iteration.

10.3.2 A Bilinear Quantum System

A quantum system acted upon an external field is governed by the famous Schrödinger equation with a forcing term

$$i\hbar\frac{d}{dt}\psi(t)=[H_0+Vu(t)]\psi(t); \quad \psi(0)=\psi_0 \tag{10.18}$$

where ψ is the complex wave function, $i=\sqrt{-1}$ and $\hbar$ is the Planck constant divided by 2π, but is considered to be one here due to normalization. The time-independent system



(a) Optimal Control

(b) Optimal States

(c) Armijo Step Sizes

(d) Convergence

Figure 10.2: Optimal Control of a Continuous Stirred-Tank Chemical Reactor. (a) and (b) show the optimal control and states as calculated by the two methods with an initial guess of $u_1(t) = 1$. (c) shows the step sizes taken at each iteration for both methods. (d) compares the convergence rates. The gradient descent method takes 39 iterations to converge with considerable variations of the step size in each iteration . Whereas, the PCGD method converges in only 4 iterations with a consistent choice of the step size at each iteration.

Hamiltonian H_0 describes the internal dynamics of the system. V is referred to as the control Hamiltonian that describes the coupling of the system to the external field $u(t)$. In this section, we consider energy-optimal population transfers similar to that presented in [28]. That is, we aim at finding a control u that transfers the system to a desired final population and minimizes

$$J(x,u) = \frac{1}{2}\int_0^T \left[|\psi(t)^* \bar{Q}\psi(t)| + Ru^2(t)\right] dt, \tag{10.19}$$

where $|.|$ is the modulus of complex numbers and $\bar{Q} \geq 0$ is designed depending on the desired final population. By defining $x = \begin{bmatrix} \text{real}(\psi) & \text{imag}(\psi) \end{bmatrix}^*$, we transform the complex optimal control problem into a real optimal control problem of the following form

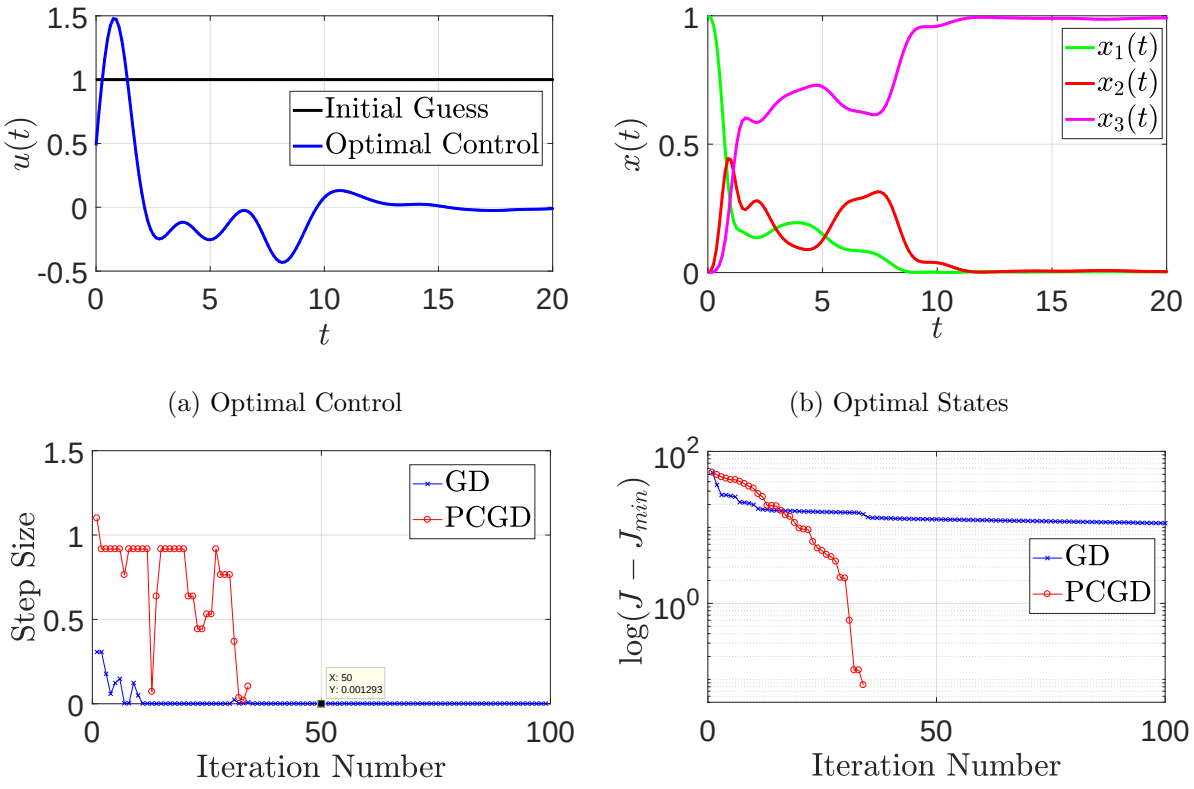
$$\begin{aligned} &\underset{z}{\text{minimize}} \quad J(z) = \frac{1}{2}\langle x, Qx\rangle + \frac{1}{2}\langle u, Ru\rangle \\ &\text{subject to} \quad \frac{d}{dt}x(t) = \left[A + Bu(t)\right]x(t); \quad x(0) = x_0 \end{aligned}$$

$$A = \begin{bmatrix} 0 & H_0 \\ -H_0 & 0 \end{bmatrix}; \quad B = \begin{bmatrix} 0 & V \\ -V & 0 \end{bmatrix}; \quad Q = \begin{bmatrix} \bar{Q} & 0 \\ 0 & \bar{Q} \end{bmatrix}.$$

In this example, we consider a three-state quantum system where it is required to carry out a population transfer : $\psi_0 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^* \mapsto \psi_d = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^*$ in $T = 20\pi$. Hence $\bar{Q}$ is designed as

$$\bar{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and $R = 1$. The results are shown in Figure 10.3. Clearly, the gradient descent method fails to achieve the optimum in 100 iterations, and it has to choose very small step sizes to proceed. On the other hand, the PCGD shows a rapid convergence near the solution.



(a) Optimal Control

(b) Optimal States

(c) Armijo Step Sizes

(d) Convergence

Figure 10.3: Optimal Control of a Three-State Quantum System. (a) and (b) show the optimal control and states as calculated by the PPGD method with a random initial guess of $u_1(t) = 1$. (c) shows the step sizes taken at each iteration of both methods. (d) shows the convergence rates on a log scale. The gradient descent method fails to converge in 100 iterations, and it has to choose very small step size α_k to proceed. Whereas, the PCGD method shows rapid convergence near the solution.

Appendix

10.A Directional Derivative & Adjoint

Let $\tilde{x} \in \mathbb{X}_0$ and $x \in \mathbb{X}$. The directional (Gâteaux) derivative of $\mathcal{D}x$ in the direction of $\tilde{x}$ is calculated as

$$\begin{aligned}\left[\partial_x \mathcal{D}x\right](\tilde{x}) &= \lim_{\epsilon \to 0} \frac{\mathcal{D}(x + \epsilon\tilde{x}) - \mathcal{D}x}{\epsilon} = \lim_{\epsilon \to 0} \frac{\dot{x} + \epsilon\dot{\tilde{x}} - \dot{x}}{\epsilon} = \dot{\tilde{x}} \\ &=: \mathcal{D}_0 \tilde{x},\end{aligned}$$

where the last equality holds because $\tilde{x} \in \mathbb{X}_0$. For short, we simply write $\partial_x \mathcal{D} = \mathcal{D}_0$. Furthermore, it can be shown that $\mathcal{D}_0^* = -\mathcal{D}_T$. Let $x \in \mathbb{X}_0$ and $y \in \mathbb{L}_n^2[0, T]$, then

$$\begin{aligned}\langle \mathcal{D}_0 x, y \rangle &= \int_0^T \dot{x}^*(t) y(t) dt \\ &= x^*(T)y(T) - x^*(0)y(0) - \int_0^T x^*(t)\dot{y}(t)dt \\ &= -\langle x, \mathcal{D}_T y \rangle ,\end{aligned}$$

where the second equality follows by applying integration by parts, and the third equality follows by recalling that $x(0) = 0$ (since $x \in \mathbb{X}_0$) and by requiring that $y \in \mathbb{X}_T$ which guarantees that $y(T) = 0$.

10.B Rigged Hilbert Space and Bilinear Forms

Let $\mathcal{S}_T : \Psi \subset \mathbb{L}_n^2[0,T] \to \mathbb{R}^n$ be the evaluation operator over the subspace Ψ. That is, the action of $\mathcal{S}_T$ on some $y \in \Psi$ is defined as

$$\mathcal{S}_T y := y(T).$$

Observe that $\mathcal{S}_T$ is an unbounded operator on $\mathbb{L}_n^2[0,T]$; however, it is bounded over the subspace $\Psi := C_n[0,T]$ (space of bounded continuous functions). The goal here is to give a rigorous explanation that justifies the formal mathematical statement that for any $y \in \Psi$ and $v \in \mathbb{R}^n$, we have

$$\langle \mathcal{S}_T y, v \rangle = \langle y, \mathcal{S}_T^* v \rangle \tag{10.B.1}$$

where the first inner product is understood in $\mathbb{R}^n$, that is $\langle \mathcal{S}_T y, v \rangle = y^*(T)v$, and $\mathcal{S}_T^*(t) := \delta(t-T)$. First, notice that the second "inner product" is not well posed, because $\mathcal{S}_T^* v \notin \mathbb{L}_n^2[0,T]$. However, with the suitable framework, this "inner product" is given a meaning and the technical issue here boils down to a slight abuse of notation.

Let $\Psi' \subset \Psi^*$ be a subset of the dual space Ψ^* of Ψ, containing all linear continuous functionals that map $\Psi \to \mathbb{R}$. Since Ψ is dense in $\mathbb{L}_n^2[0,T]$, then the triple $(\Psi, \mathbb{L}_n^2[0,T], \Psi')$ forms a *Rigged Hilbert Space* where $\Psi \subset \mathbb{L}_n^2[0,T] \subset \Psi'$. For any $v \in \mathbb{R}^n$, define the family of bounded linear functionals $\mathcal{S}_T^* v \in \Psi'$ that map any $y \in \Psi$ as

$$\mathcal{S}_T^* v(y) := y^*(T)v.$$

Furthermore, the action of the functional $\mathcal{S}_T^* v$ on $y \in \Psi$ can be represented using the canonical bilinear form [1] over $\Psi \times \Psi'$ as

$$\langle y, \mathcal{S}_T^* v \rangle_{(\Psi,\Psi')} := \mathcal{S}_T^* v(y).$$

Hence, for any $v \in \mathbb{R}^n$ and $y \in \Psi$, this bilinear form (unlike the inner product) is well

defined. Then, we have

$$\langle y, \mathcal{S}_T^* v \rangle_{(\Psi,\Psi')} := \mathcal{S}_T^* v(y) := y^*(T)v = \langle \mathcal{S}_T y, v \rangle.$$

Therefore, with a slight abuse of notation, we drop the subscript "(Ψ, Ψ')" of the bilinear form to yield (10.B.1).

This extends the inner products in $\mathbb{L}_n^2[0,T]$ to the more general notion of bilinear forms on $(\Psi \times \Psi')$ and motivates viewing $\mathcal{S}_T^* : \mathbb{R}^n \to \Psi^*$ as the adjoint operator of $\mathcal{S}_T$. It also justifies the common abuse of notation

$$y^*(T)v = \left(\text{“} \int_0^T \delta(t-T) y^*(t) dt \text{”} \right) v,$$

where the right hand side really means the bilinear form $\langle y, \mathcal{S}_T^* v \rangle_{(\Psi,\Psi')}$. Finally we note that we adopt this abuse of notation throughout the paper, for simplicity, since $\mathbb{X}_0 \subset \Psi$.

10.C Directional Derivatives & Adjoint of the System Operator

Let $\mathcal{H}$ be the system operator defined in (8.1). To calculate the first and second directional derivatives of $\mathcal{H}$, we carry out a perturbation analysis. More precisely, we perturb u_k by $\epsilon\tilde{u}$ to obtain $\mathcal{H}(u_k + \epsilon\tilde{u})$. A Taylor expansion in ϵ yields

$$\mathcal{H}(u_k + \epsilon\tilde{u}) = \mathcal{H}(u_k) + \epsilon \partial H_{u_k}(\tilde{u}) + \frac{1}{2}\epsilon^2 \partial^2 \mathcal{H}_{u_k}(\tilde{u}) + \mathcal{O}(\epsilon^3). \tag{10.C.1}$$

Define

$$\begin{aligned} x_\epsilon &:= \mathcal{H}(u_k + \epsilon\tilde{u}) \\ x_k &:= \mathcal{H}(u_k) \\ \tilde{x}_k &:= \partial \mathcal{H}_{u_k}(\tilde{u}) \end{aligned} \tag{10.C.2}$$

$$\bar{x}_k := \partial^2 \mathcal{H}_{u_k}(\tilde{u}),$$

so that (10.C.1) can be rewritten as

$$x_\epsilon = x_k + \epsilon \tilde{x}_k + \frac{1}{2}\epsilon^2 \bar{x}_k + \mathcal{O}(\epsilon^3). \tag{10.C.3}$$

Using the definition (8.1) of the system operator, (10.C.2) can be rewritten as

$$\dot{x}_\epsilon = f(x_\epsilon, u_k + \epsilon\tilde{u}); \quad x(0) = \boldsymbol{x_0}, \tag{10.C.4}$$

Substituting for the expression of x_ϵ given by (10.C.3) in (10.C.4) (while truncating higher orders of ϵ) yields

$$\dot{x}_k + \epsilon\dot{\tilde{x}}_k + \frac{1}{2}\epsilon^2\dot{\bar{x}}_k = f\left(x_k + \epsilon\left(\tilde{x}_k + \frac{1}{2}\epsilon\bar{x}_k\right), u_k + \epsilon\tilde{u}\right); \quad x_k(0) + \epsilon\tilde{x}_k(0) + \frac{1}{2}\epsilon^2\bar{x}_k(0) = \boldsymbol{x_0}.$$

Then $x_k(0) = \boldsymbol{x_0}$ and $\tilde{x}_k(0) = \bar{x}_k(0) = 0$. Using the appropriate time derivative operators to respect the domains (see Section 6.1 for details) and expanding f around $z_k := (x_k, u_k)$ up to second order in ϵ, we obtain

$$\mathcal{D}x_k + \epsilon\mathcal{D}_0\tilde{x}_k + \frac{1}{2}\epsilon^2\mathcal{D}_0\bar{x}_k = f(x_k, u_k) + \epsilon\partial f_{z_k}\begin{bmatrix}\tilde{x}_k + \frac{1}{2}\epsilon\bar{x}_k \\ \tilde{u}\end{bmatrix} + \frac{1}{2}\epsilon^2\partial^2 f_{z_k}(\tilde{x}_k, \tilde{u})$$

$$\mathcal{D}x_k + \epsilon\mathcal{D}_0\tilde{x}_k + \frac{1}{2}\epsilon^2\mathcal{D}_0\bar{x}_k = f(x_k, u_k) + \epsilon\begin{bmatrix}A_k & B_k\end{bmatrix}\begin{bmatrix}\tilde{x}_k \\ \tilde{u}\end{bmatrix} + \frac{1}{2}\epsilon^2\left(A_k\bar{x}_k + \partial^2 f_{z_k}(\tilde{x}_k, \tilde{u})\right)$$

Finally, to obtain the expressions of the directional derivatives of $\mathcal{H}$, we equate the same orders in ϵ. In fact, equating the zeroth orders in ϵ simply yields $x_k = \mathcal{H}(u_k)$, and equating the first orders in ϵ yields $\mathcal{D}_0\tilde{x}_k = A_k\tilde{x}_k + B_k\tilde{u}$, and thus

$$\tilde{x}_k = \partial\mathcal{H}_{u_k}(\tilde{u}) \iff \tilde{x}_k = (\mathcal{D}_0 - A_k)^{-1} B_k\tilde{u}.$$

Furthermore, equating the second orders in ϵ yields $\mathcal{D}_0\bar{x}_k = A_k\bar{x}_k + \partial^2 f_{z_k}(\tilde{x}_k, \tilde{u})$. By exploiting the notation for the second derivative in (6.7), we obtain

$$\bar{x}_k = \partial^2\mathcal{H}_{u_k}(\tilde{u}) \iff \bar{x}_k = (\mathcal{D}_0 - A_k)^{-1}\tilde{z}_k^*\mathcal{F}_k\tilde{z}_k,$$

where $\tilde{z}_k$ is defined as

$$\tilde{z}_k := \begin{bmatrix} \tilde{x}_k \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} \partial \mathcal{H}_{u_k}(\tilde{u}) \\ \tilde{u} \end{bmatrix}.$$

Finally, since $\partial \mathcal{H}_{u_k} = (\mathcal{D}_0 - A_k)^{-1} B_k$, then by using (6.4), we obtain $\partial \mathcal{H}^*_{u_k} = -B_k^* (\mathcal{D}_T + A_k^*)^{-1}$.

10.D Directional Derivatives & Adjoint of the Projection Operator

Let $\mathcal{P}$ be the projection operator defined in (9.1). To calculate the first and second directional derivatives of $\mathcal{P}$, we carry out a perturbation analysis. More precisely, we perturb $\hat{z}_k := (\hat{x}_k, \hat{u}_k)$ by $\epsilon \underline{z} := \epsilon(\underline{x}, \underline{u})$ to obtain $\mathcal{P}(\hat{z}_k + \epsilon \underline{z})$. A Taylor expansion in ϵ yields

$$\mathcal{P}(\hat{z}_k + \epsilon \underline{z}) = \mathcal{P}(\hat{z}_k) + \epsilon \partial P_{\hat{z}_k}(\underline{z}) + \frac{1}{2}\epsilon^2 \partial^2 P_{\hat{z}_k}(\underline{z}) + \mathcal{O}(\epsilon^3). \tag{10.D.1}$$

Define

$$\begin{aligned} z_\epsilon &:= (x_\epsilon, u_\epsilon) := \mathcal{P}(\hat{z}_k + \epsilon \underline{z}) \\ z_k &:= (x_k, u_k) := \mathcal{P}(\hat{z}_k) \\ \tilde{z}_k &:= (\tilde{x}_k, \tilde{u}_k) := \partial \mathcal{P}_{\hat{z}_k}(\underline{z}) \\ \bar{z}_k &:= (\bar{x}_k, \bar{u}_k) := \partial^2 \mathcal{P}_{\hat{z}_k}(\underline{z}), \end{aligned} \tag{10.D.2}$$

so that (10.D.1) can be rewritten as

$$z_\epsilon = z_k + \epsilon \tilde{z}_k + \frac{1}{2}\epsilon^2 \bar{z}_k + \mathcal{O}(\epsilon^3). \tag{10.D.3}$$

Using the definition (9.1) of the projection operator, (10.D.2) can be rewritten as

$$\begin{cases} x_\epsilon = \mathcal{H}(u_\epsilon) \\ u_\epsilon = \hat{u}_k + \epsilon \underline{u} + g(\hat{x}_k + \epsilon \underline{x} - x_\epsilon). \end{cases} \tag{10.D.4}$$

Substituting the expressions of (x_ϵ, u_ϵ), given by (10.D.3), in (10.D.4) (while truncating higher orders of ϵ) yields

$$\begin{cases} x_k + \epsilon\tilde{x}_k + \frac{1}{2}\epsilon^2\bar{x}_k = \mathcal{H}\left(u_k + \epsilon\left(\tilde{u}_k + \frac{1}{2}\epsilon\bar{u}_k\right)\right) \\ u_k + \epsilon\tilde{u}_k + \frac{1}{2}\epsilon^2\bar{u}_k = \hat{u}_k + \epsilon\underline{u} + g\left((\hat{x}_k - x_k) + \epsilon\left(\underline{x} - \tilde{x}_k - \frac{1}{2}\epsilon\bar{x}_k\right)\right). \end{cases}$$

Expanding $\mathcal{H}$ around u_k and g around $\hat{x}_k - x_k$, up to second order in ϵ, yields

$$\begin{cases} x_k + \epsilon\tilde{x}_k + \frac{1}{2}\epsilon^2\bar{x}_k = \mathcal{H}(u_k) + \epsilon\partial\mathcal{H}_{u_k}\left(\tilde{u}_k + \frac{1}{2}\epsilon\bar{u}_k\right) + \frac{1}{2}\epsilon^2\partial^2\mathcal{H}_{u_k}\left(\tilde{u}_k + \frac{1}{2}\epsilon\bar{u}_k\right) \\ u_k + \epsilon\tilde{u}_k + \frac{1}{2}\epsilon^2\bar{u}_k = \hat{u}_k + \epsilon\underline{u} + g(\hat{x}_k - x_k) + \epsilon\partial g_{\hat{x}_k - x_k}\left(\underline{x} - \tilde{x}_k - \frac{1}{2}\epsilon\bar{x}_k\right) \\ \qquad\qquad + \frac{1}{2}\epsilon^2\partial^2 g_{\hat{x}_k - x_k}\left(\underline{x} - \tilde{x}_k - \frac{1}{2}\epsilon\bar{x}_k\right), \end{cases}$$

where the directional derivatives of the system operator $\mathcal{H}$ are given in Table 8.1. Note that $\partial g_{\hat{x}_k - x_k}$ and $\partial^2 g_{\hat{x}_k - x_k}$ are the Jacobian and second derivative of g evaluated at $\hat{x}_k - x_k$, respectively. By exploiting the linearity of $\partial\mathcal{H}_{u_k}$ and truncating ϵ^3, we obtain

$$\begin{cases} x_k + \epsilon\tilde{x}_k + \frac{1}{2}\epsilon^2\bar{x}_k = \mathcal{H}(u_k) + \epsilon\partial\mathcal{H}_{u_k}(\tilde{u}_k) + \frac{1}{2}\epsilon^2\Big(\partial\mathcal{H}_{u_k}(\bar{u}_k) + \partial^2\mathcal{H}_{u_k}(\tilde{u}_k)\Big) \\ u_k + \epsilon\tilde{u}_k + \frac{1}{2}\epsilon^2\bar{u}_k = \hat{u}_k + g(\hat{x}_k - x_k) + \epsilon\Big(\underline{u} + \partial g_{\hat{x}_k - x_k}(\underline{x} - \tilde{x}_k)\Big) \\ \qquad\qquad + \frac{1}{2}\epsilon^2\Big(\partial^2 g_{\hat{x}_k - x_k}(\underline{x} - \tilde{x}_k) - \partial g_{\hat{x}_k - x_k}(\bar{x}_k)\Big). \end{cases}$$

Finally, to obtain the expressions of the directional derivatives of $\mathcal{P}$, we equate the same orders in ϵ. In fact, equating the zeroth orders in ϵ simply yields $(x_k, u_k) = \mathcal{P}(\hat{x}_k, \hat{u}_k)$, and equating the first orders in ϵ yields

$$(\tilde{x}_k, \tilde{u}_k) = \partial\mathcal{P}_{\hat{z}_k}(\underline{x}, \underline{u}) \iff \begin{cases} \tilde{x}_k = \partial\mathcal{H}_{u_k}(\tilde{u}_k) \\ \tilde{u}_k = \underline{u} + K_k(\underline{x} - \tilde{x}_k), \end{cases} \tag{10.D.5}$$

where $K_k := \partial g_{\hat{x}_k - x_k}$. Equation (10.D.5) describes the action of $\partial\mathcal{P}_{\hat{z}_k}$ in terms of the system operator $\mathcal{H}$. Using the expression of $\partial\mathcal{H}_{u_k}$ in Table 8.1, we obtain the differential equation form shown in Table 9.1.

To calculate the second derivative, equate the second orders in ϵ to obtain

$$(\bar{x}_k, \bar{u}_k) = \partial^2 P_{\hat{z}_k}(\underline{x}, \underline{u}) \Longleftrightarrow \begin{cases} \bar{x}_k = \partial \mathcal{H}_{u_k}(\bar{u}_k) + \partial^2 \mathcal{H}_{u_k}(\tilde{u}_k) \\ \bar{u}_k = (\underline{x} - \tilde{x}_k)^* \mathcal{G}_k(\underline{x} - \tilde{x}_k) - K_k \bar{x}_k, \end{cases} \tag{10.D.6}$$

where we exploit the notation developed in the section on Section 6.1.5 for the second derivative of a vector-valued function. Again, by substituting the expressions of $\partial \mathcal{H}_{u_k}$ and $\partial^2 \mathcal{H}_{u_k}$ from Table 8.1, we obtain the differential equation form shown in Table 9.1.

Now we calculate the adjoint of $\partial \mathcal{P}_{\hat{z}_k}$, denoted by $\partial \mathcal{P}^*_{\hat{z}_k}$. We first write $\partial \mathcal{P}_{\hat{z}_k}$ as an operator-valued matrix. From (10.D.5), we have

$$\tilde{x}_k = \partial \mathcal{H}_{u_k}\big(\underline{u} + K_k(\underline{x} - \tilde{x}_k)\big) = \partial \mathcal{H}_{u_k} \begin{bmatrix} K_k & \mathcal{I} \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{u} \end{bmatrix} - \partial \mathcal{H}_{u_k} K_k \tilde{x}_k$$

$$\tilde{x}_k = (\mathcal{I} + \partial \mathcal{H}_{u_k} K_k)^{-1} \partial \mathcal{H}_{u_k} \begin{bmatrix} K_k & \mathcal{I} \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{u} \end{bmatrix},$$

where $\mathcal{I}$ is the identity operator. Then the operation $(\tilde{x}_k, \tilde{u}_k) = \partial \mathcal{P}_{\hat{z}_k}(\underline{x}, \underline{u})$ in (10.D.5) can be rewritten as

$$\begin{bmatrix} \tilde{x}_k \\ \tilde{u}_k \end{bmatrix} := \partial \mathcal{P}_{\hat{z}_k} \begin{bmatrix} \underline{x} \\ \underline{u} \end{bmatrix} = \begin{bmatrix} (\mathcal{I} + \partial \mathcal{H}_{u_k} K_k)^{-1} \partial \mathcal{H}_{u_k} \\ \mathcal{I} - K_k (\mathcal{I} + \partial \mathcal{H}_{u_k} K_k)^{-1} \partial \mathcal{H}_{u_k} \end{bmatrix} \begin{bmatrix} K_k & \mathcal{I} \end{bmatrix} \begin{bmatrix} \underline{x} \\ \underline{u} \end{bmatrix}.$$

Then the adjoint is

$$\partial \mathcal{P}^*_{\hat{z}_k} = \begin{bmatrix} K_k^* \\ \mathcal{I} \end{bmatrix} \begin{bmatrix} \partial \mathcal{H}^*_{u_k} \left(\mathcal{I} + K_k^* \partial \mathcal{H}^*_{u_k}\right)^{-1} & \mathcal{I} - \partial \mathcal{H}^*_{u_k} (\mathcal{I} + K_k^* \partial \mathcal{H}^*_{u_k})^{-1} K_k^* \end{bmatrix},$$

where $\partial \mathcal{H}^*_{u_k}$ is given in Table 8.1. To compute the action of the adjoint in terms of $\partial \mathcal{H}^*_{u_k}$ without the inverse operations, we proceed as follows. Let $(\tilde{\chi}_k, \tilde{\mu}_k) = \partial \mathcal{P}^*_{\hat{z}_k}(\chi, \mu)$, then

$$\begin{cases} \tilde{\chi}_k = K_k^* \tilde{\mu}_k \\ \tilde{\mu}_k = \partial \mathcal{H}^*_{u_k} (\mathcal{I} + K_k^* \partial \mathcal{H}^*_{u_k})^{-1} (\chi - K_k^* \mu) + \mu =: \partial \mathcal{H}^*_{u_k} \alpha_k + \mu, \end{cases} \tag{10.D.7}$$

where the intermediate variable α_k is defined as

$$\alpha_k := (\mathcal{I} + K_k^* \partial \mathcal{H}_{u_k}^*)^{-1}(\chi - K_k^* \mu). \tag{10.D.8}$$

But $\tilde{\chi}_k$ in (10.D.7) can be rewritten in terms of α_k as

$$\tilde{\chi}_k = K_k^* \tilde{\mu}_k = K_k^* \partial \mathcal{H}_{u_k}^* \alpha_k + K_k^* \mu, \tag{10.D.9}$$

and α_k in (10.D.8) can be rewritten as

$$\chi - \alpha_k = K_k^* \partial \mathcal{H}_{u_k}^* \alpha_k + K_k^* \mu. \tag{10.D.10}$$

Comparing (10.D.9) and (10.D.10) yields $\tilde{\chi}_k = \chi - \alpha_k$ or $\alpha_k = \chi - \tilde{\chi}_k$. Finally, substituting for α_k in (10.D.7) yields

$$(\tilde{\chi}_k, \tilde{\mu}_k) := \partial \mathcal{P}_{\hat{z}_k}^*(\chi, \mu) \Longleftrightarrow \begin{cases} \tilde{\chi}_k = K_k^* \tilde{\mu}_k \\ \tilde{\mu}_k = \mu + \partial \mathcal{H}_{u_k}^*(\chi - \tilde{\chi}_k). \end{cases} \tag{10.D.11}$$

By substituting the expression of $\partial \mathcal{H}_{u_k}^*$ from Table 8.1, we obtain the differential equation form shown in Table 9.1.

10.E Replacing $\mathcal{S}_T^*$ with a Boundary Condition

In this appendix, we show that when $\mathcal{S}_T^*$ appears in a differential equation, we can remove it by a suitable modification of the boundary condition.

Consider the following differential equation

$$\dot{\lambda}(t) + f(t) + \mathcal{S}_T^* v = 0; \quad \lambda(T) = 0, \tag{10.E.1}$$

where f is some bounded function of time and v is a constant vector. Integrating both sides of the differential equation from $T - \epsilon$ to $T + \epsilon$ yields

$$\int_{T-\epsilon}^{T+\epsilon} \dot{\lambda}(t) dt + \int_{T-\epsilon}^{T+\epsilon} f(t) dt + \int_{T-\epsilon}^{T+\epsilon} \delta(t - T) v dt = 0.$$

When we take the limit as $\epsilon \to 0$, the second term goes to zero because f is finite to obtain

$$\lambda(T+\epsilon) - \lambda(T-\epsilon) + v = 0,$$

where we exploit the sifting property of the Dirac delta function. Given the original boundary condition $\lambda(T) = 0$, that is $\lambda(T+\epsilon) = 0$, we obtain

$$\lambda(T-\epsilon) = \lambda(T+\epsilon) + v = v.$$

Therefore, the new boundary condition becomes $\lambda(T) = v$, and hence (10.E.1) can be replaced by the following differential equation

$$\dot{\lambda}(t) + f(t) = 0; \quad \lambda(T) = v.$$

Part IV

Optimal Estimation & Tomographic Sensing in Distributed Environments

Chapter 11

Introduction

Model-based dynamic estimation is a widely used methodology for incorporating partial measurements of a process with some knowledge of its underlying dynamics. Most notably is the Kalman filter in the Linear Quadratic case, and its many nonlinear versions. In this paper, we are concerned with dynamical processes that are described by Partial Differential Equations (PDEs) in two or three spatial dimensions, particularly ones that describe fluid flow and advection of temperature and concentration fields. Incorporating measurements into such basic physics models is referred to as "data assimilation" in the Atmospheric Sciences literature [52].

The setting that motivates our current work is on a smaller scale such as local outdoor environments, or time-critical situations such as forest fires or hazardous plumes. In such situations mobile sensors collect limited measurements that need to be "assimilated" into fluid flow models. Since temperature, concentration and flow fields follow well-known physical laws, it is reasonable to expect that incorporating these PDEs can significantly enhance the reconstruction fidelity and spatial resolution of limited measurements.

While the foregoing is a standard dynamic estimation and data assimilation problem, our concern in the present work is how should mobile sensors move so as to optimize

estimation errors in the dynamic setting? This is especially important if sensors' maneuverability and time is limited, and this optimization of the sensors' paths become critical. This is essentially a mobile version of the classic "sensor placement" problem [14, 15, 21, 32, 40].

The approach we adopt is a sort of double optimization (a min-min problem), where the sensors' paths are chosen to minimize metrics related to the estimator of minimal error. For example, if the underlying dynamics are linear and the estimation error criterion is quadratic, then the optimal estimation error covariance (over a finite-time horizon) is given by the differential Riccati equation of the Kalman filter. The sensor's paths enter as a time-varying signal in the "C matrix" of the system's output equation, which enters quadratically in the Riccati equation. One can now think of a purely *deterministic optimal control* problem where the dynamics are given by the matrix (or operator)-valued Riccati differential equation, and the time-varying sensors' paths are the "control inputs" into this equation. The sensor paths can now be chosen to minimize criteria such as combinations of the trace of the error covariance and costs of sensor motion. In the nonlinear dynamics case, the error covariance is not given by a Riccati equation, but the min-min optimization problem formulation is still valid. The optimal path is chosen to minimize an error criterion (error variance, relative entropy or others) of the best estimator. This is similar in spirit to the approach adopted in [12, 32].

First we formulate the optimal distributed estimation problem where the sensors' locations are chosen a-priori. Then, we address the problems of sensor placement (for static sensors) and path planning (for mobile sensors) in the subsequent chapters.

The setting is the standard stochastic estimation with process disturbances and measurement noise. In addition, we do not assume that boundary conditions are known or fixed, but rather stochastic with some prior knowledge of the relative time scale of their variations (e.g. the daily cycle of the sun's radiation heating). We model two

different measurement operators: (1) point-wise measurements and (2) line integral measurements. The latter being relevant to acoustic tomography sensing (which will be addressed in details in the subsequent chapters).

Chapter 12

Acoustic Tomography & Estimation of Static Temperature Fields

In this chapter, we first give a brief description on the basic concepts of acoustic tomography and explain how it can be exploited in estimating temperature fields. Then a couple of case studies are considered to describe how we estimate unknown temperature fields in 2 dimensions using tomographic sensing techniques.

12.1 Acoustic Tomography for Static Temperature Fields

Acoustic Tomography [35], [36], [63] is a technique for reconstructing scalar fields (e.g. temperature) or vector fields (e.g. wind velocity) from the time of flight of ultrasonic sound signals between transmitters and receivers. The transceivers can be deployed outside the region to be mapped which might be advantageous in some scenarios (such as hazardous plumes, forest fires, etc.). In this section, we show how acoustic tomographic can be exploited to estimate unknown temperature fields.

In dry air, the equation that relates the temperature to the speed of sound is given by

$$c = \sqrt{\gamma\psi}, \tag{12.1}$$

where $\gamma = \frac{\gamma_0 R}{M}$, c is the speed of sound [m/s] at a temperature ψ [°K]. $\gamma_0 = 1.4$ is the adiabatic index, $R = 8.31451$ is the molar gas constant and $M = 0.0289645$ is the mean molar mass of dry air. The velocity of an acoustic ray is given by

$$\vec{v}_{ray} = c\vec{n} + \vec{v}, \tag{12.2}$$

where $\vec{n}$ is the unit wavefront normal and $\vec{v}$ is the wind speed. Denote by τ_{ij} to be the time of flight between transceivers i and j. Denote by $\vec{l}$ to be the unit direction vector of the line joining the two transceivers. Also, let Γ_{ij} be the straight path from transceiver i to j. Then, the time of flight τ_{ij} is

$$\tau_{ij} := \int_{\Gamma_{ij}} dt = \int_{\Gamma_{ij}} \frac{1}{\vec{v}_{ray}.\vec{l}} dl. \tag{12.3}$$

Hence the time of flight measurements are going to be some known nonlinear function of the temperature and velocity fields

$$\tau_{ij} = \mathcal{G}(\psi, \vec{v}). \tag{12.4}$$

The goal of acoustic tomography is to recover the temperature and velocity fields from available time of flight measurements. Naturally, some questions can be asked here: (a) is this inverse problem solvable? (b) if yes, how many transceivers are required for accurate recovery? (c) where to deploy the limited number of available transceivers? (d) how to deal with dynamical fields that are time varying? The Fourier Slice theorem [48] proves that using the Radon transform (which computes line integrals over the whole domain), one can recover only the temperature field. However, other set of measurements are needed in addition to (12.4) for a full reconstruction of the velocity vector field [10].

For example, [35] utilizes the angle of departure/arrival of the acoustic waves using the bent ray model. In this dissertation, we study the recovery of temperature fields in a stationary medium ($\vec{v} = 0$) with a straight ray model for acoustic signals ($\vec{n} = \vec{l}$).

12.2 Posing the Inverse Problem

Let $\psi(x, y)$ denote the temperature at location (x, y). Then the time of flight between transceivers i and j, denoted by τ_{ij}, is given by

$$\tau_{ij} = \int_{\Gamma_{ij}} \frac{dl}{\sqrt{\gamma\psi}}. \tag{12.5}$$

Linearizing around some operating point $\bar{\psi}$, we obtain

$$\tau_{ij} = \frac{1}{\sqrt{\gamma\bar{\psi}}} \left(\frac{3}{2} L_{ij} - \frac{1}{2\bar{\psi}} \int_{\Gamma_{ij}} \psi(x, y) dl \right). \tag{12.6}$$

with L_{ij} being the distance between transceivers i and j. We denote the line integral of the temperature field along Γ_{ij} by m_{ij} which can be expressed in terms of the time of flight τ_{ij} as

$$m_{ij} := 2\bar{\psi} \left(\frac{3}{2} L_{ij} - \sqrt{\gamma\bar{\psi}} \tau_{ij} \right). \tag{12.7}$$

For the rest of the dissertation, we assume that m_{ij} is directly available as measurements. Therefore the inverse problem for one line integral measurement is given by $m_{ij} = \mathcal{C}_{ij}(\psi)$, where $\mathcal{C}_{ij}$ is the line integral operator defined as follows:

$$\mathcal{C}_{ij}(\psi) := \int_{\Gamma_{ij}} \psi(x, y) dl. \tag{12.8}$$

Now, let m and $\mathcal{C}$ be two vectors that concatenate m_{ij} and $\mathcal{C}_{ij}$ for all possible lines, respectively. That is, $m := \{m_{ij}\}$ and $\mathcal{C} := \{\mathcal{C}_{ij}\}$. Let N_m be the number of lines used. Then the inverse problem is to find $\psi(x, y)$ that satisfies

$$m = \mathcal{C}(\psi). \tag{12.9}$$

Ultimately, since the scalar temperature field has an infinite number of unknowns and the measurements (line integrals) are finite, then the inverse problem is under-determined. A common method ([35] and the references therein) to overcome this issue is to parametrize the scalar field with a finite number of unknowns N_n using some numerical method such as finite elements or finite differences with $N_n < N_m$. Then a pseudo inverse is used to invert the line integral operator. In other words, the numerical method is deliberately designed to make the inverse problem solvable. However, for applications where the region of the unknown field is large, and the number of transceivers is limited, this method will force a coarse grid. This significantly degrades the reconstruction accuracy. A better scheme should not depend on the numerical method used to solve the infinite dimensional inverse problem.

12.3 Solution Schemes for the Inverse Problem

In this section, we show and compare three different methods to solve the inverse problem (12.9).

12.3.1 Norm Minimization of the Error

In this section we present the method that has been widely used by signal processing people and we thus call it "traditional method". This method, as mentioned before, parametrizes the scalar field with a number of unknowns less than the number of available measurements so that a pseudo-inverse method can be applied to minimize the error. Formally, this method solves the following problem

$$\psi_* = \underset{\psi}{\operatorname{argmin}} ||m - \mathcal{C}(\psi)|| \tag{12.10}$$

To solve this problem using finite elements, we cover the region with a tile of N_t triangles with a total of N_n nodes. Inside each triangle we have

$$\psi(x,y) = \sum_{n_l=1}^{n_l=3} \psi_{n_l}\phi_{n_l}(x,y) \tag{12.11}$$

where n_l is the local index of nodes for a particular triangle, ψ_{n_l} is the value of the scalar field at the node of a local index n_l, and ϕ_{n_l} is a linear basis function that is equal to 1 at the local node n_l and zero at all other nodes in the region. Thus, $\psi(x,y)$ is linearly interpolated between the nodes. Finally, the line integral operator $\mathcal{C}_{ij}$ can be numerically approximated as follows (refer to figure 12.1)

$$\begin{aligned}\mathcal{C}_{ij}(\psi) &\approx \sum_{n_t=1}^{n_t=N_t} \int_{\Gamma_{ij}\cap \text{Triangle}_{n_t}} dl \sum_{n_l=1}^{n_l=3} \psi_{n_l}\phi_{n_l}(x,y) \\ &= \sum_{n_t=1}^{n_t=N_t}\sum_{n_l=1}^{n_l=3} \psi_{n_l} \int_{\Gamma_{ij}\cap \text{Triangle}_{n_t}} \phi_{n_l}(x,y)dl\end{aligned} \tag{12.12}$$

Eventually, using this method, $\mathcal{C}$ will be realized as some matrix C. Moreover the

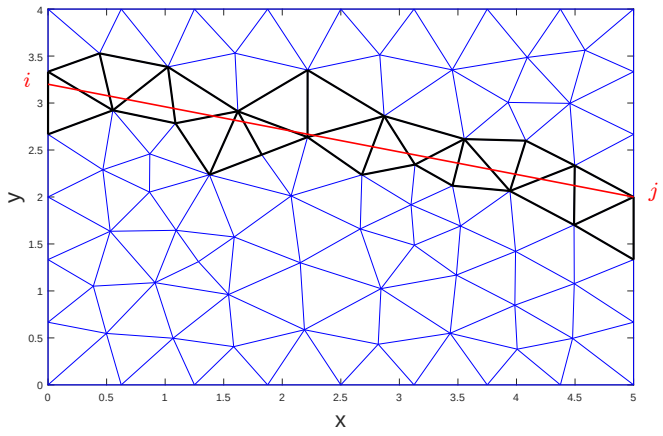


Figure 12.1: Line Integral on Finite Element Triangulation

unknowns can be formed as a vector

$$\text{vec}(\psi) := [\psi_1 \quad \psi_2 \quad ... \quad \psi_{N_n}]^T. \tag{12.13}$$

The problem can thus be expressed as

$$\text{vec}(\psi_*) = \underset{\psi}{\text{argmin}}||m - C\text{vec}(\psi)||. \tag{12.14}$$

This problem can be easily solved using the pseudo inverse. Hence the reconstructed temperature is

$$\text{vec}(\psi_*) = (C^T C)^{-1} C^T m. \tag{12.15}$$

For this method to give meaningful results, N_m needs to be greater or equal to N_n. In applications with large region Ω and limited number of transceivers, the mentioned condition means that the finite element grid needs to be coarse enough to reduce the number of unknowns. This will give accurate reconstruction only at the locations of the nodes. However, the accuracy will degrade significantly at the locations between nodes especially for fields with rapid spatial variations. This is illustrated in subsequent case studies.

12.3.2 Alternative Minimization

The actual inverse problem is different, in essence, from the approach presented in the previous section. The number of unknowns is actually infinite and in practice we can only obtain a finite number of measurements. This certainly fails to apply to the condition for the previous approach: $N_m \geq N_n$. In fact, the null space of the operator $\mathcal{C}$ is infinite dimensional. A better solution scheme shouldn't depend on the numerical method used. Attempting to overcome this limitation, we incorporate the physical laws that govern the scalar field under study. For the temperature problem, the field is governed by the heat equation

$$\begin{aligned} \mathcal{A}\psi(x,y) = 0 \quad & \text{for all}(x,y) \in \Omega \\ \text{with} \quad & \mathcal{A} := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \end{aligned} \tag{12.16}$$

Incorporating Physical Laws (Attempt 0): Instead of solving the inverse problem in equation (12.9), we solve the following inverse problem

$$\begin{bmatrix} 0 \\ m \end{bmatrix} = \mathcal{M}_0 \psi \qquad \text{where} \qquad \mathcal{M}_0 = \begin{bmatrix} \mathcal{A} \\ \mathcal{C} \end{bmatrix}. \tag{12.17}$$

Here, the extra piece of information added is the physical laws that govern the interior of the region Ω with no additional information on the boundaries. It turns out the null space of $\mathcal{M}_0$ is still non trivial. This will be illustrated in the numerical example later on.

Incorporating Physical Laws with a Minimization Criterion: The previous attempt didn't trivialize the null space of the operator to be inverted. Then there are infinite number of solutions from which we have to pick the best in some sense. One way to do that is by casting an alternative minimization problem. Two different minimization criteria were chosen: the spatial variations and the spatial curvature on the boundaries; that is, we minimize the tangential derivative $\partial_{\vec{t}}$ and tangential second derivative $\partial^2_{\vec{t}}$ on the boundaries, respectively.

Attempt 1:

$$\begin{aligned} \psi_* = \underset{\psi}{\text{argmin}} \quad & \frac{1}{2}\langle \partial_{\vec{t}}\psi, \partial_{\vec{t}}\psi \rangle \\ \text{subject to:} \quad & \mathcal{M}_0\psi = \begin{bmatrix} 0 \\ m \end{bmatrix} \end{aligned} \tag{12.18}$$

Attempt 2:

$$\begin{aligned} \psi_* = \underset{\psi}{\text{argmin}} \quad & \frac{1}{2}\langle \partial^2_{\vec{t}}\psi, \partial^2_{\vec{t}}\psi \rangle \\ \text{subject to:} \quad & \mathcal{M}_0\psi = \begin{bmatrix} 0 \\ m \end{bmatrix} \end{aligned} \tag{12.19}$$

Solution of the Minimization Problems: In this section, we present the solution procedure for attempts 1 and 2. In general, assume that the minimization problem is:

$$\begin{aligned} \psi_* = \underset{\psi}{\operatorname{argmin}} \quad & \frac{1}{2}\langle \mathcal{D}\psi, \mathcal{D}\psi \rangle \\ \text{subject to:} \quad & \mathcal{M}\psi = b. \end{aligned} \tag{12.20}$$

where $\mathcal{D}$ is a linear operator and b is some vector.

Using Lagrangian multipliers λ, we form:

$$\mathcal{L}(\psi, \lambda) := \frac{1}{2}\langle \mathcal{D}\psi, \mathcal{D}\psi \rangle + \lambda^*(\mathcal{M}\psi - b). \tag{12.21}$$

Hence to find the stationary points, we require the partial derivatives $\frac{\partial \mathcal{L}}{\partial \psi}$ and $\frac{\partial \mathcal{L}}{\partial \lambda}$ to be zero. Thus, we arrive at the necessary conditions:

$$\begin{bmatrix} \mathcal{D}\mathcal{D}^* & \mathcal{M}^* \\ \mathcal{M} & 0 \end{bmatrix} \begin{bmatrix} \psi \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}. \tag{12.22}$$

Equation (12.22) can be solved to recover the temperature ψ.

In the subsequent section, we give a case study where we estimate the temperature distribution using line integral measurements as described in this section.

12.4 Case Study 1: Estimating a Static Temperature Field on a Rectangle

Consider a two-dimensional rectangular region Ω of height $H = 4$ and length $L = 5$ with no heat sources/sinks within. The boundaries admit Dirichlet conditions on the top and the left ($T_1 = 20$ and $T_2 = 30$) and the other boundaries admit Neumann conditions (refer to figure 12.2). However, the boundary conditions (on $\partial\Omega$) are assumed to be unknown. The operators $\mathcal{A}$, $\mathcal{C}$ $\partial_{\bar{t}}$, and $\partial_{\bar{t}}^2$ are realized using a finite difference

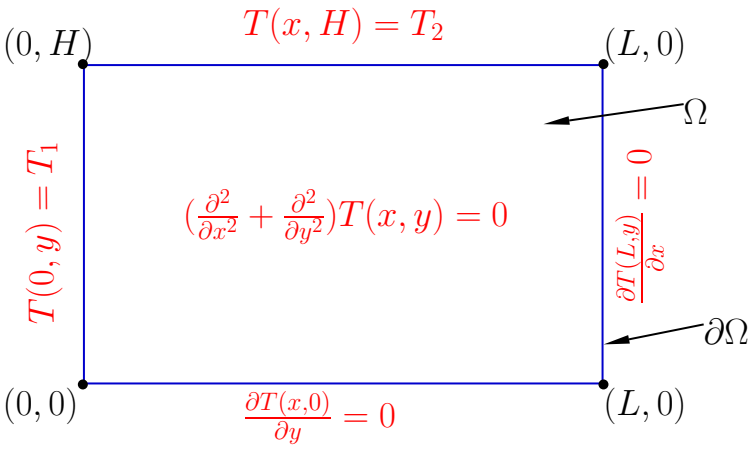


Figure 12.2: Heat Equation on a Rectangular Region

method with a grid size of 30 by 20 and the linearization was carried out around $\bar{T} = 25$. 14 transceivers are deployed on the boundary as shown in figure 12.3 and time of flights between transceivers along the boundaries are not allowed thus giving $N_m = 59$ measurements. With this spatial discretization, the number of unknowns is $N_n = 600$.

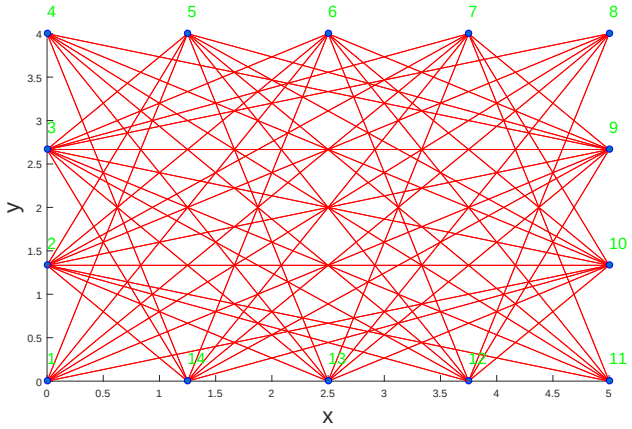


Figure 12.3: Location of the Deployed Transceivers

Clearly, the first method cannot be applied. In fact with this number of measurements available, it can only recover the temperature field for a grid with a maximum of 59 nodes (for example an 8 by 7 grid). This will certainly degrade the quality of the reconstructed

field. Attempt 0 still has nontrivial null space. For this method, there are an infinite number of solutions out of which that of a minimal norm is shown in figure 12.4. Indeed

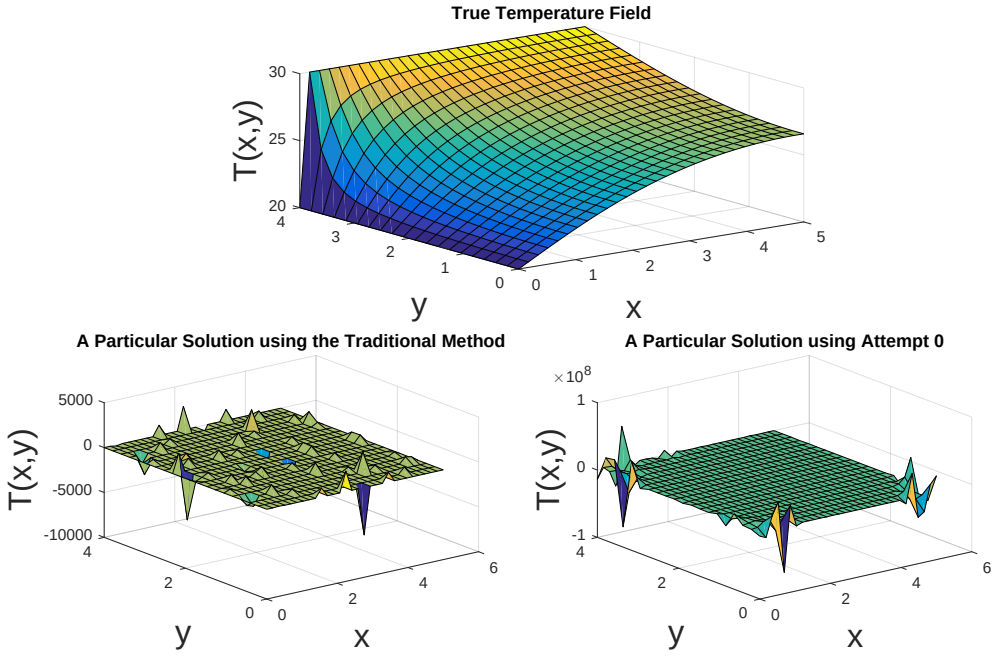


Figure 12.4: Failed Temperature Reconstruction using the Traditional Method and Attempt 0

figure 12.4 shows how the two methods fails to reconstruct the temperature. However, the large variations of the reconstructed field at the boundary in attempt 0 motivated us to look for solutions with some conditions on the boundaries. Thus attempts 1 and 2 were made. In fact, figure 12.5 shows the error percentage of the reconstructed temperature fields using attempts 1 and 2. Although attempt 1 shows better reconstruction inside the region, the reconstruction at the boundaries shows curvature variations. This motivated attempt 2 where the curvature along the boundaries is minimized thus reconstructing the temperature field with small errors.

The message taken from this case study is that when the number of transceivers

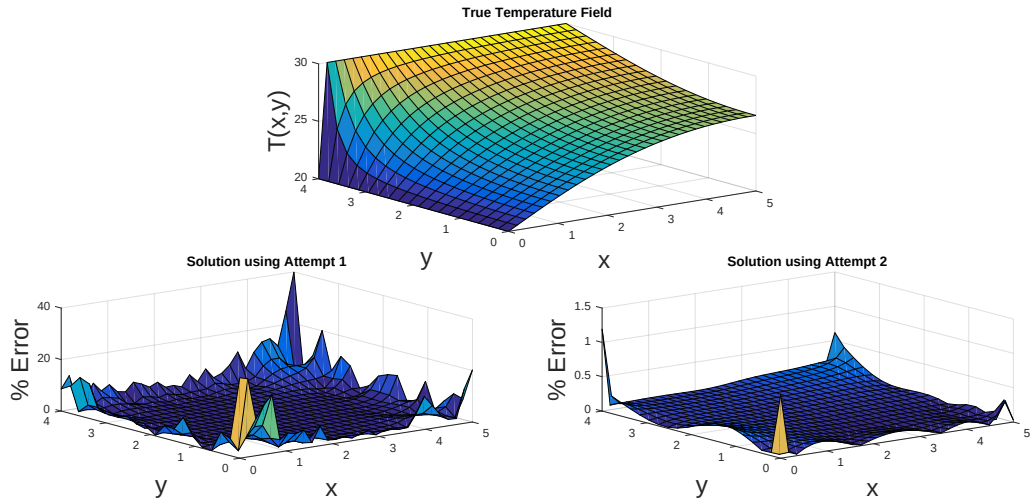


Figure 12.5: Temperature Reconstruction using Attempt 1 and 2

is limited, there is no way around incorporating our knowledge from the physics to reconstruct the unknown field accurately.

12.5 Case Study 2: Temperature Reconstruction on a Disk, Analytical Example

To get insights and intuition on the number of transceivers required to fully recover the temperature field, we study a particular example that can be analyzed analytically. In this example, we consider the exact reconstruction of unknown temperature fields on a disk using line integral measurements. The temperature field is assumed to be governed by the two dimensional static heat equation.

Let Ω be the region strictly inside a unit disk. Let (x, y) and (r, θ) be the Cartesian and polar coordinates respectively such that

$$x = r\cos(\theta) \qquad \text{and} \qquad y = r\sin(\theta). \tag{12.23}$$

Then, the region Ω along with its boundary $\partial\Omega$ can be represented as follows:

$$\begin{aligned} \Omega &= \{(x,y);\ 0 \le x^2+y^2 < 1\} &&= \{(r,\theta);\ 0 \le r < 1, -\pi \le \theta < \pi\} \\ \partial\Omega &= \{(x,y);\ x^2+y^2 = 1\} &&= \{(r,\theta); r = 1, -\pi \le \theta < \pi\}. \end{aligned} \tag{12.24}$$

Let $\psi(x,y)$ represent the temperature at location (x,y). Assume that the steady state heat equation (Laplace equation) governs Ω with a Dirichlet boundary condition on $\partial\Omega$.

$$\begin{aligned} \left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}\right)\psi(x,y) &= 0 && (x,y) \in \Omega \\ \psi(x,y) &= h(\theta) && (x,y) \in \partial\Omega. \end{aligned} \tag{12.25}$$

The Poisson Integral Formula [11] gives a closed form for boundary value problem (12.25) in polar coordinates:

$$\psi(r,\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} h(\phi)\frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)}d\phi. \tag{12.26}$$

Line Integrals over Diameters: First, we assume that the line integrals are calculated on lines passing through the origin as shown in figure 12.6.

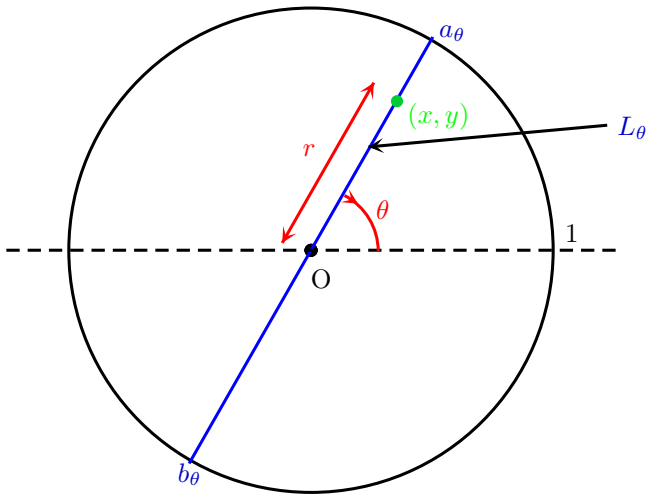


Figure 12.6: Unit Disk

Define $\mathcal{C}_\theta$ to be the line integral operator where the line of integration is parametrized by the angle θ.

$$\begin{aligned} \mathcal{C}_\theta \colon \mathbb{L}^2([0,1],[-\pi,\pi[) &\to \mathbb{R} \\ \psi &\mapsto \mathcal{C}_\theta(\psi) = \int_{a_\theta}^{b_\theta} \psi(r,\theta) dl. \end{aligned} \tag{12.27}$$

The line integral can be divided into two parts

$$\mathcal{C}_\theta(\psi) = \int_{a_\theta}^{0} \psi(r,\theta) dl + \int_{0}^{b_\theta} \psi(r,\theta) dl = \int_0^1 \psi(r,\theta) dr + \int_0^1 \psi(r,\theta+\pi) dr, \tag{12.28}$$

by exploiting the fact that $\psi(r,\theta)$ is periodic in θ.

Now, define

$$I(\theta) := \int_0^1 \psi(r,\theta) dr. \tag{12.29}$$

Hence

$$\mathcal{C}_\theta(\psi) = I(\theta) + I(\theta+\pi). \tag{12.30}$$

Substituting the expression $\psi(r,\theta)$ from (12.26) in (12.30), we get

$$\begin{aligned} I(\theta) &= \int_0^1 \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\phi) \frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)} d\phi dr \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\phi) \int_0^1 \frac{1-r^2}{1+r^2-2r\cos(\theta-\phi)} dr d\phi \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\phi) J(\theta-\phi) d\phi, \end{aligned} \tag{12.31}$$

where J is the integral of the Poisson kernel

$$J(\phi) := \int_0^1 \frac{1-r^2}{1+r^2-2r\cos(\phi)} dr. \tag{12.32}$$

Equation(12.31) suggests that the line integral $I(\theta)$ is the convolution of the boundary condition with the Poisson kernel J. As a matter of fact, the integral of the Poisson kernel can be shown to have a closed form. It is given by

$$J(\phi) = -1 - \cos(\phi)\log(2-2\cos(\phi)) + 2\pi|\sin(\phi)| \left[\mathrm{tri}_{2\pi}(\phi-\pi) + \frac{1}{2}\mathrm{tri}_{2\pi}(\phi) - \frac{1}{2}\right], \tag{12.33}$$

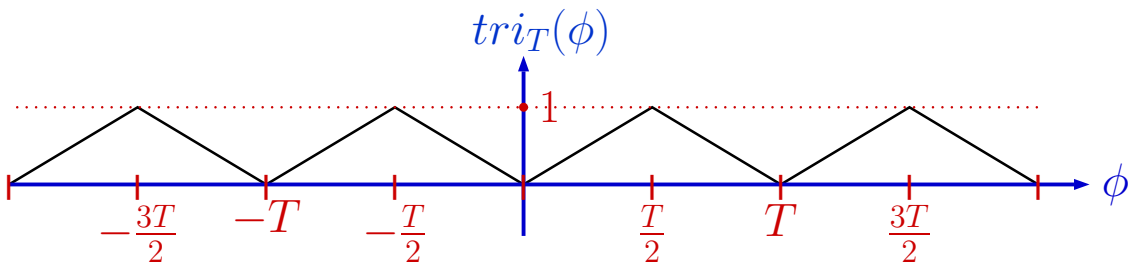


Figure 12.7: $tri_{2\pi}(\phi)$

where $\text{tri}_T(\phi)$ is a triangular periodic function as shown in figure 12.7. Furthermore, $J(\phi)$ is plotted in Figure 12.8. Since $h(\phi)$, $J(\phi)$ and $I(\phi)$ are periodic functions in their argument, then their Fourier Series can be calculated.

$$\begin{aligned} J(\phi) &= \sum_{k=-\infty}^{k=+\infty} \hat{J}[k] e^{ik\phi} \qquad & \hat{J}[k] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} J(\phi) e^{-ik\phi} d\phi \\ h(\phi) &= \sum_{k=-\infty}^{k=+\infty} \hat{h}[k] e^{ik\phi} \qquad & \hat{h}[k] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} h(\phi) e^{-ik\phi} d\phi \\ I(\phi) &= \sum_{k=-\infty}^{k=+\infty} \hat{I}[k] e^{ik\phi} \qquad & \hat{I}[k] &= \frac{1}{2\pi}\int_{-\pi}^{\pi} I(\phi) e^{-ik\phi} d\phi. \end{aligned} \tag{12.34}$$

$\hat{J}[k]$, $\hat{h}[k]$ and $\hat{I}[k]$ are the Fourier coefficients of $J(\phi)$, $h(\phi)$ and $I(\phi)$, respectively. In fact, the Fourier series for $J(\phi)$ can be explicitly calculated

$$\begin{aligned} \hat{J}[k] = \frac{1}{2\pi}\int_{-\pi}^{\pi} &(-1 - \cos(\phi)\log(2 - 2\cos(\phi)) \\ &+2\pi|\sin(\phi)|\left[tri_{2\pi}(\phi - \pi) + \frac{1}{2}tri_{2\pi}(\phi) - \frac{1}{2}\right]\Big) e^{-ik\phi} d\phi \\ = \frac{1}{|k|+1}. \end{aligned} \tag{12.35}$$

Figure 12.8 plots the Fourier coefficients $J[k]$ of $J(\phi)$ and some truncated partial sums of the Fourier series. Taking the Fourier Series of equation (12.30) yields

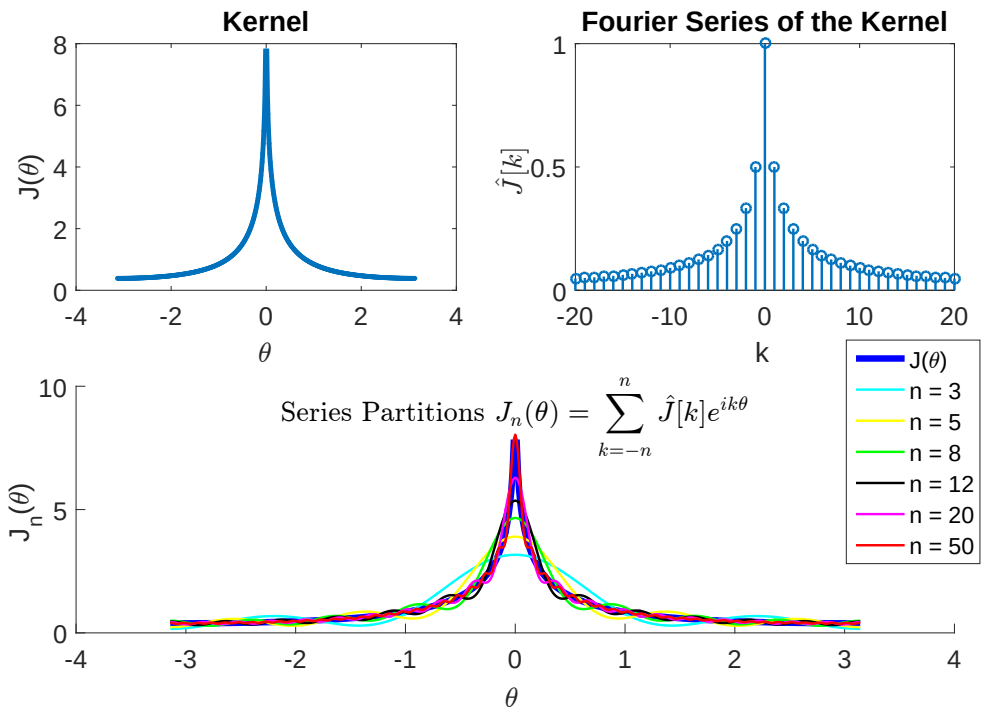


Figure 12.8: Integral of the Poisson Kernel and its Fourier Series

$$\begin{aligned}
\hat{\mathcal{C}}_k(\psi) = \hat{I}[k] + e^{ik\pi}\hat{I}[k] \qquad &= (1 + e^{ik\pi})\hat{I}[k] \\
= (1 + e^{ik\pi})\hat{h}[k]\hat{J}[k] \quad &= \frac{1 + (-1)^k}{|k| + 1}\hat{h}[k] \\
= \hat{F}[k]\hat{h}[k] \qquad & \qquad (12.36) \\
\hat{F}[k] = \begin{cases} 0, & \text{k odd} \\ \frac{2\hat{h}[k]}{|k|+1}, & \text{k even} \end{cases} &
\end{aligned}$$

where $\hat{\mathcal{C}}_k(\psi)$ is the Fourier Coefficient of the line integral operator.

$$\hat{\mathcal{C}}_k(\psi) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{C}_\theta(\psi)e^{-ik\theta}d\theta. \qquad (12.37)$$

Equation(12.36) shows that $\mathcal{C}_\theta(\psi) = 0$ for any angle θ if the boundary condition $h(\theta)$ has only odd spatial frequency contents. This result is intuitive since the scalar field with odd spatial frequency contents on the boundary forms an anti symmetric temperature

distribution within the disk, and thus the line integral will cancel out. We conclude that the operator $\mathcal{C}_\theta$ has an infinite dimensional nullspace and is not invertible: any scalar field with anti symmetric components (due to odd spatial frequency contents on the boundary) cannot be reconstructed from line integrals along the diameters.

As an illustrative example, we solve the forward problem using the following boundary condition:

$$h(\theta) = 20 + 5\cos(\theta) + 3\cos(2\theta) + 7\cos(3\theta). \tag{12.38}$$

The line integral for all possible θ is calculated by first calculating $\hat{\mathcal{C}}_k(\psi)$ using equation(12.36) and then calculate its inverse Fourier Series using:

$$\mathcal{C}_\theta(\psi) = \sum_{k=-\infty}^{k=+\infty} \hat{\mathcal{C}}_k(\psi) e^{ik\theta}. \tag{12.39}$$

Figure 12.9 shows the calculated line integral along with the Fourier Series and the imposed boundary condition.

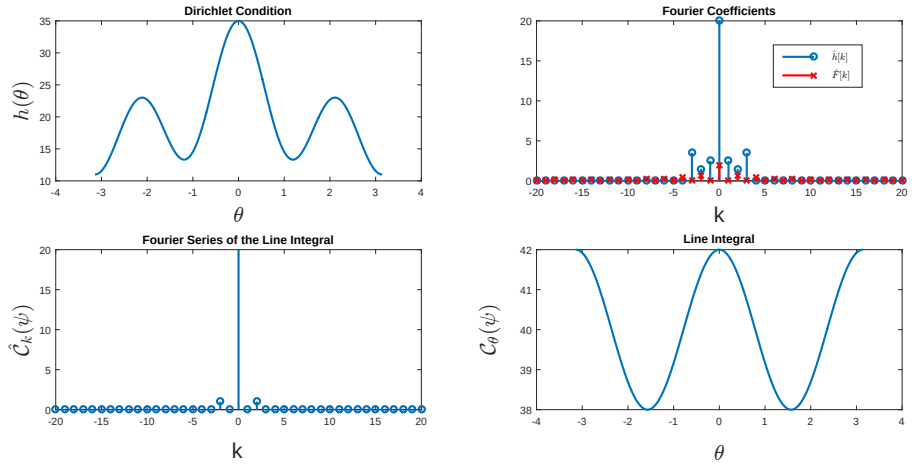


Figure 12.9: Line Integral along Diameters of the Disk

Now to solve the inverse problem, we assume that we are given the line integrals for all lines passing through the origin as shown in figure 12.9 and we are required to reconstruct the scalar field $\psi(r, \theta)$ using equations (12.34, 12.36 and 12.37). Figure 12.10 shows the

original temperature field along with the reconstructed field. The reconstructed field is not the same as the true field. In fact, it corresponds to the boundary condition with no odd spatial frequency contents. Moreover, the unreconstructed field corresponds to the field with only odd spatial frequency contents on the boundary.

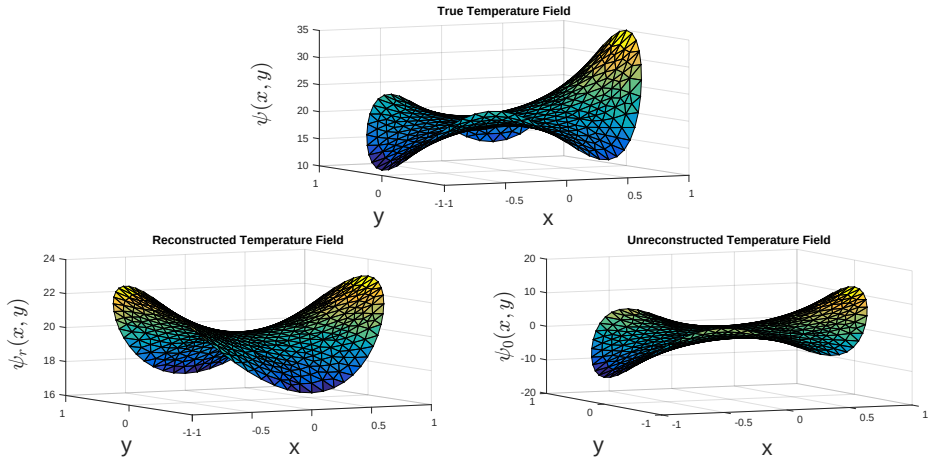


Figure 12.10: Reconstructed and Unreconstructed Temperature Fields

As a conclusion, this reconstruction procedure is capable of reconstructing only the symmetric component of the temperature field.

Line Integrals along Radii: The reason for the aforementioned null space is that the line integrals are along diameters. On the other hand, if sensors are allowed to be deployed inside the region, then one can deploy only one transmitter at the origin and spread receivers along the boundary. Then the line integral operator would be:

$$\mathcal{C}_\theta(\psi) = \int_0^1 \psi(r,\theta)dr = I(\theta). \tag{12.40}$$

Hence, its Fourier series would be

$$\hat{\mathcal{C}}_k(\psi) = \hat{J}[k]\hat{h}[k] = \frac{1}{|k|+1}\hat{h}[k]. \tag{12.41}$$

Equation(12.41) suggests that the new line integral operator will never nullify the scalar field unless the boundary condition is zero. That is, the null space of this operator is trivial and hence it is invertible. As a conclusion, this scheme is capable of completely reconstructing the temperature field from the line integrals. In practice, only finite number of receivers can be deployed. In fact, to fully recover the temperature field from a finite number of line integrals, the Nyquist-Shannon sampling theorem must be respected. That is, if the highest spatial frequency contained in $h(\theta)$ is N_h, then $2N_h + 1$ receivers are required to fully recover the temperature field.

Chapter 13

Estimation of Dynamic Distributed Fields Via Tomographic Sensing

In this chapter, we formulate the optimal distributed estimation problem where the sensors' locations are chosen a-priori. After developing the theoretical framework, we consider a case study where we estimate an unknown dynamically evolving temperature field in a two dimensional room.

13.1 Formulation of the Dynamic Distributed Estimation Problem

The setting considered here is the standard stochastic estimation with process disturbances and measurement noise. In addition, we do not assume that boundary conditions are known or fixed, but rather stochastic with some prior knowledge of the relative time scale of their variations (e.g. the daily cycle of the sun's radiation heating). We model two different measuring operators for the cases of point-wise sampling, and line integral measurements respectively. The latter being relevant to acoustic tomography sensing.

13.1.1 Physical Dynamics and the Measurement Models

Let ψ denote a dynamical field in a region Ω with its boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$; $\partial\Omega_D$ and $\partial\Omega_N$ are subsets of the boundary where Dirichlet and Neumann conditions are respectively imposed. The governing dynamics are transcribed by the following partial differential equation:

$$\begin{aligned} &\frac{\partial}{\partial t}\psi(\boldsymbol{x},t) = \mathcal{A}\psi(\boldsymbol{x},t) + w(\boldsymbol{x},t); \qquad \psi(0,\boldsymbol{x}) = \psi_0(\boldsymbol{x}) \\ &\psi(\boldsymbol{x},t)\Big|_{\partial\Omega_D} = \psi_D(\boldsymbol{x},t); \qquad \frac{\partial}{\partial\vec{n}}\psi(\boldsymbol{x},t)\Big|_{\partial\Omega_N} = \psi_N(\boldsymbol{x},t) \end{aligned} \tag{13.1}$$

where w is a zero-mean white Gaussian noise field with covariance $\mathcal{P}_w$. $\mathcal{A}$ is a spatial operator defined on the domain of fields satisfying the boundary conditions in (13.1), and $\vec{n}$ is a unit vector normal to $\partial\Omega_N$. ψ_D and ψ_N are unknown, possibly time varying, fields defined on the boundaries $\partial\Omega_D$ and $\partial\Omega_N$, respectively. For the rest of the paper, we drop the dependence of the fields on the spatial variable $\boldsymbol{x}$ whenever no confusion is caused.

The available sensors are assumed to be capable of taking either point-wise or line integral measurements (such as transceivers). For this purpose, define the line integral and sampling operators as follows:

$$\mathcal{C}_{ij}\psi := \int_{\Gamma_{ij}} \psi(\boldsymbol{x})d\boldsymbol{x} \quad \text{and} \qquad \mathcal{C}_k\psi := \psi(\boldsymbol{x}_k) \tag{13.2}$$

The linear spatial operator $\mathcal{C}_{ij}$ acts on a field, defined on Ω, to yield its line integral over the straight path Γ_{ij} connecting transceivers i and j. On the other hand, $\mathcal{C}_k$ samples the field at the location $\boldsymbol{x}_k$ of the point-wise sensor k. Taking all possible measurements by the available transceivers and point-wise sensors, we get the output equation:

$$m(t) = \mathcal{C}\psi(t) + v_c(t) \tag{13.3}$$

where $\mathcal{C}$ is the vector concatenation of $\mathcal{C}_{ij}$ and $\mathcal{C}_k$, and v_c is a zero-mean white Gaussian noise vector with covariance R_c.

The ultimate target of this paper is to design schemes to assimilate our knowledge of the model with the available measurements in order to optimally estimate the unknown fields ψ, ψ_D and ψ_N. In fact, we target two challenges facing this problem: (1) unknown boundary conditions and (2) where to place the sensors and how to control their movement.

13.1.2 Incorporating Unknown Boundary Conditions and their Dynamics

We now tackle the first challenge by modeling the dynamics of the boundary conditions based on our knowledge of the physical laws governing the fields. By absorbing our modeled stochastic dynamics of ψ_D and ψ_N, we formulate the estimation problem in a standard Kalman filter setting.

The boundary conditions ψ_D and ψ_N can be seen as inputs to the dynamical system defined in (13.1). Since they are unknown, we assume that they have dynamics of their own. We model their dynamics by the following evolution equations driven by white Gaussian noise:

$$\begin{aligned} \frac{\partial}{\partial t}\xi_D(t) &= \mathcal{A}_D\xi_D(t) + \mathcal{B}_D w_D(t); \quad \xi_D(0) = \xi_{D0} \\ \frac{\partial}{\partial t}\xi_N(t) &= \mathcal{A}_N\xi_N(t) + \mathcal{B}_N w_N(t); \quad \xi_N(0) = \xi_{N0} \end{aligned} \tag{13.4}$$

where ξ_D and ξ_N are the states of the dynamical systems modeling the Dirichlet and Neumann boundary conditions, respectively. For example, if the modeled dynamics are of first order, then $\xi_D = \psi_D$. In the case of second order dynamics, $\xi_D = \begin{bmatrix} \psi_D & \frac{\partial}{\partial t}\psi_D \end{bmatrix}^T$ and so on. Moreover, w_D and w_N are zero-mean white Gaussian noise fields with covariances $\mathcal{P}_D$ and $\mathcal{P}_N$, respectively. The operators $\mathcal{A}_D$, $\mathcal{A}_N$, $\mathcal{B}_D$ and $\mathcal{B}_N$ will shape the response of the modeled boundary dynamics and are designed depending on the application at hand.

By absorbing the modeled dynamics of the boundary conditions (13.4) in (13.1), we

get the augmented (continuous time) evolution equation with the corresponding output equation:

$$\begin{aligned} \psi_c(t) &= \mathcal{A}_c\psi_c(t) + w_c(t); \qquad \psi_c(0) = \psi_{c0} \\ m(t) &= \begin{bmatrix} \mathcal{C} & 0 \end{bmatrix} \psi_c(t) + v_c(t) \\ E\{w_c(\boldsymbol{x},t)w_c^*(\boldsymbol{\chi},\tau)\} &= \mathcal{Q}_c(\boldsymbol{x},\boldsymbol{\chi})\delta(t-\tau) \\ E\{v_c(t)v_c^T(\tau)\} &= R_c\delta(t-\tau) \end{aligned} \tag{13.5}$$

$$\psi_c := \begin{bmatrix} \psi \\ \xi_D \\ \xi_N \end{bmatrix} \quad \psi_{c0} := \begin{bmatrix} \psi_0 \\ \xi_{D0} \\ \xi_{N0} \end{bmatrix} \mathcal{A}_c := \begin{bmatrix} \mathcal{A} & 0 & 0 \\ 0 & \mathcal{A}_D & 0 \\ 0 & 0 & \mathcal{A}_N \end{bmatrix}$$

$$w_c := \begin{bmatrix} w \\ w_D \\ w_N \end{bmatrix} \quad \mathcal{Q}_c := \begin{bmatrix} \mathcal{P}_w & 0 & 0 \\ 0 & \mathcal{B}_D\mathcal{P}_D\mathcal{B}_D^* & 0 \\ 0 & 0 & \mathcal{B}_N\mathcal{P}_N\mathcal{B}_N^* \end{bmatrix}$$

where $\delta(t)$ is the Dirac delta function and "$*$" is the adjoint operator. Note that, we assume that there is no correlation between the different boundary conditions and the interior field, hence $\mathcal{Q}_c$ is block diagonal. If correlation is required in a particular application, off-diagonal terms can be added.

Assuming that the sensors are already deployed in fixed locations, the $\mathcal{C}$ operator is thus known and time invariant. Hence, the estimation problem can be optimally solved using Kalman filters. The design parameters are the modeled dynamics of the boundary conditions in (13.4) and $(\mathcal{Q}_c, R_c)$ in (13.5). The design completely depends on the application at hand. As a case study, the next section will illustrate an application on temperature fields and acoustic tomography.

13.2 Case Study: Dynamic Acoustic Tomography of Temperature Fields

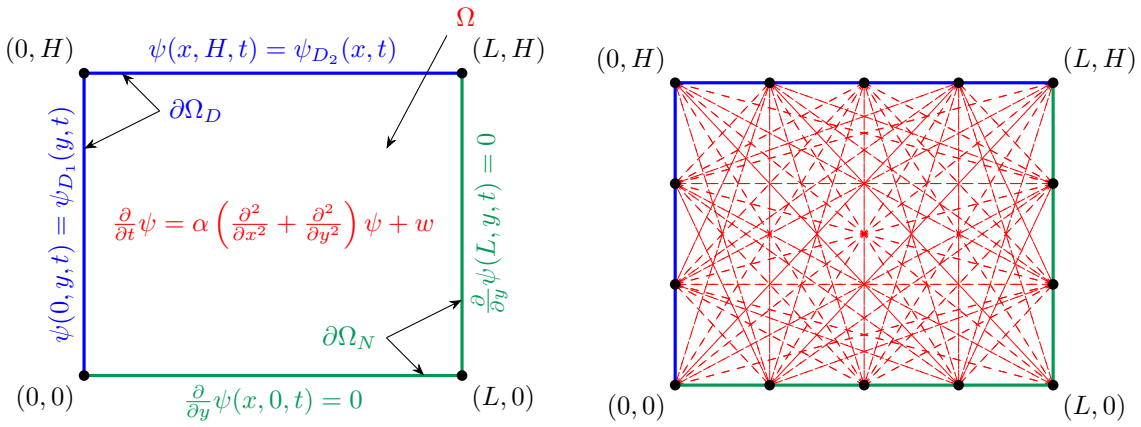
Acoustic tomography [35], [36], [63] is a technique for reconstructing scalar temperature fields and/or vector velocity fields from the time of flight of ultrasonic sound signals between transceivers. The transceivers can be deployed outside the region to be mapped which might be advantageous in some scenarios (such as hazardous plumes, forest fires, etc.). In this example, we utilize the technique developed in the previous section to estimate dynamic temperature fields in a static (zero velocity) medium. The measurement scheme employed is based on tomographic sensing, i.e. line integral measurements are taken.

13.2.1 Temperature Dynamics and Tomographic Sensing

Consider Ω to be a rectangular region to simulate a two dimensional room as shown in Fig. 13.1(a). Hence, the spatial variable $\boldsymbol{x} = \begin{bmatrix} x & y \end{bmatrix}^T$ is two dimensional. The temperature field is governed by the dynamical heat equation with a diffusion constant α. Unknown non-homogeneous Dirichlet conditions are imposed on the left ($\partial\Omega_1$) and top ($\partial\Omega_2$) boundaries to simulate heat sources/sinks (i.e. $\partial\Omega_D = \partial\Omega_1 \cup \partial\Omega_2$). For simplicity, the other boundary conditions (on $\partial\Omega_N$) are known to be insulated walls with homogeneous Neumann conditions imposed. Then the $\mathcal{A}$ operator in (13.1) is defined as follows:

$$\mathcal{A}\psi := \alpha \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) \psi$$
$$\text{with} \quad \psi_N(t) = 0 \quad \text{and} \quad \psi_D(t) = \begin{bmatrix} \psi_{D_1}(t) \\ \psi_{D_2}(t) \end{bmatrix} \tag{13.6}$$

In acoustic tomography, transceivers measure the time of flight of ultrasonic acoustic



(a) Room Dynamics (b) Deployed Transceivers

Figure 13.1: (a) The 2-dimensional dynamical heat equation is considered with a diffusion constant α. The unknown Dirichlet boundary conditions to the top and the left are allowed to be varying in space and time. The Neumann boundary conditions to the bottom and the right are assumed to be known and homogeneous, thus modeling insulated walls. (b) Transceivers are deployed on the boundaries to measure the time of flight of ultrasonic signals between them.

signals. It can be shown [35] that the time of flight depends on the temperature field along the path traveled. For simplicity, we assume that the transceivers are directly measuring the line integrals of the temperature field along the straight paths between them.

13.2.2 Modeling the Unknown Dynamics of the Boundary Conditions

Since the boundaries with Neumann conditions are known to be homogeneous, x_{N0}, $\mathcal{A}_N$ and $\mathcal{B}_N$ are all zeros in (13.4). In this example, we assume that the temporal frequency of the temperature variations along the boundaries are known to be less than a given frequency f_n. As a result, we model the boundaries with Dirichlet conditions to be the outputs of a second order low pass filter fed by a zero-mean white Gaussian noise field,

w_D. Let f_n and ζ denote the natural frequency and damping ratio of the second order low pass filter. Note that w_D can be divided into two separate sources, w_{D_1} and w_{D_2}, affecting each Dirichlet boundary condition separately. That is, $w_D(t) := \begin{bmatrix} w_{D_1}(t) & w_{D_2}(t) \end{bmatrix}^T$. Let $\mathcal{P}_{D_1}$ and $\mathcal{P}_{D_2}$ denote the covariance of w_{D_1} and w_{D_2}, respectively. Based on a physical intuition of the smoothness of temperature distributions, we assume that the temperature fields along the Dirichlet boundaries are spatially correlated with correlation lengths of σ_1 and σ_2. Hence one way to represent the covariance is by using a Gaussian kernel as follows:

$$\mathcal{P}_{D_1}(y,\xi) = a_1 e^{-\frac{(y-\xi)^2}{2\sigma_1^2}}; \quad \mathcal{P}_{D_2}(x,\chi) \quad = a_2 e^{-\frac{(x-\chi)^2}{2\sigma_2^2}} \tag{13.7}$$

Finally, by letting $\omega_n = 2\pi f_n$ and recalling the dynamics of a second order low pass filter, the parameters in (13.4) are summarized as follows:

$$\begin{aligned} \mathcal{A}_D &= \begin{bmatrix} 0 & 0 & \mathcal{I} & 0 \\ 0 & 0 & 0 & \mathcal{I} \\ -\omega_n^2\mathcal{I} & 0 & -2\zeta\omega_n\mathcal{I} & 0 \\ 0 & -\omega_n^2\mathcal{I} & 0 & -2\zeta\omega_n\mathcal{I} \end{bmatrix} \\ \mathcal{B}_D &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \omega_n^2\mathcal{I} & 0 \\ 0 & \omega_n^2\mathcal{I} \end{bmatrix} \quad \xi_D = \begin{bmatrix} \psi_{D1} \\ \psi_{D2} \\ \frac{\partial}{\partial t}\psi_{D_1} \\ \frac{\partial}{\partial t}\psi_{D_2} \end{bmatrix} \quad \mathcal{P}_D = \begin{bmatrix} \mathcal{P}_{D_1} & 0 \\ 0 & \mathcal{P}_{D_2} \end{bmatrix} \end{aligned} \tag{13.8}$$

where $\mathcal{I}$ is the identity operator.

13.2.3 Numerical Results for Temperature Field Estimation

For a numerical example, we use $L = 5$ m and $H =$ m with periodic Dirichlet boundary conditions (period = 1 day) as follows:

$$\psi(0, y, t) = 20 + 10 \sin\left(\frac{2\pi}{24 \times 60} t\right)$$
$$\psi(x, H, t) = 30 - 10 \sin\left(\frac{2\pi}{24 \times 60} t\right)$$

where t is expressed in minutes and ψ in °C. The initial temperature field is $\psi_0 = 10$°C. The true diffusion constant α is $0.05 m^2 s^{-1}$. To simulate an inaccessible interior region, we deploy 14 transceivers on the boundaries, as shown in Fig. 13.1(b), allowing us to take a total of 59 measurements at each instant of time. The operators $\mathcal{A}_c$, $\mathcal{C}$, and Q_c are realized using a finite difference method, by laying down a 35 by 40 two dimensional grid. In our simulations, we assume that our sensors and the model in (13.6) are accurate, so we let $R_c = 0.01I$ and $\mathcal{P}_w = 0.01\mathcal{I}$, where I and $\mathcal{I}$ are the identity matrix and identity operator, respectively. The design parameters are $a_1, a_2, \sigma_1, \sigma_2, \zeta$ and f_n. Their choice should be based on the available information (physical intuition) on the temperature field. For example, we predict that the temperature variations on the boundaries are around 40°C with a period more than 3 days. We also predict that the spatial temperature variations along the Dirichlet boundaries are small. Then we choose $a_1 = a_2 = 40^2$, $f_n = \frac{1}{5}$ days^{-1}, $\sigma_1 = 5H$ and $\sigma_2 = 5L$. The low pass filter is designed to be critically damped, that is $\zeta = 0.707$.

To test the robustness of our estimation scheme, we carry out two simulations with no prior knowledge of the initial temperature field in both cases. First, we assume that we have exact knowledge of the diffusion constant, that is we know that $\alpha = 0.05$. This is still challenging the Kalman filter since the Dirichlet boundary conditions are unknown. Fig. 13.2 shows two snapshots of the exact and estimated temperature fields at $t = 12$ min and $t = 2.5$ hrs. The figure clearly shows how accurate the estimation is. For the second

simulation, we assume that we don't have perfect knowledge of the diffusion constant; say we predict $\hat{\alpha} = 0.1$. This is considerably different from the actual diffusion constant. Fig. 13.3 shows the estimation error as a function of time for both scenarios: exact and perturbed diffusion constant. In fact, the figure shows how the estimation accuracy is degraded for the perturbed model; however, it is still doing a very good job given the severe perturbation of the diffusion constant.

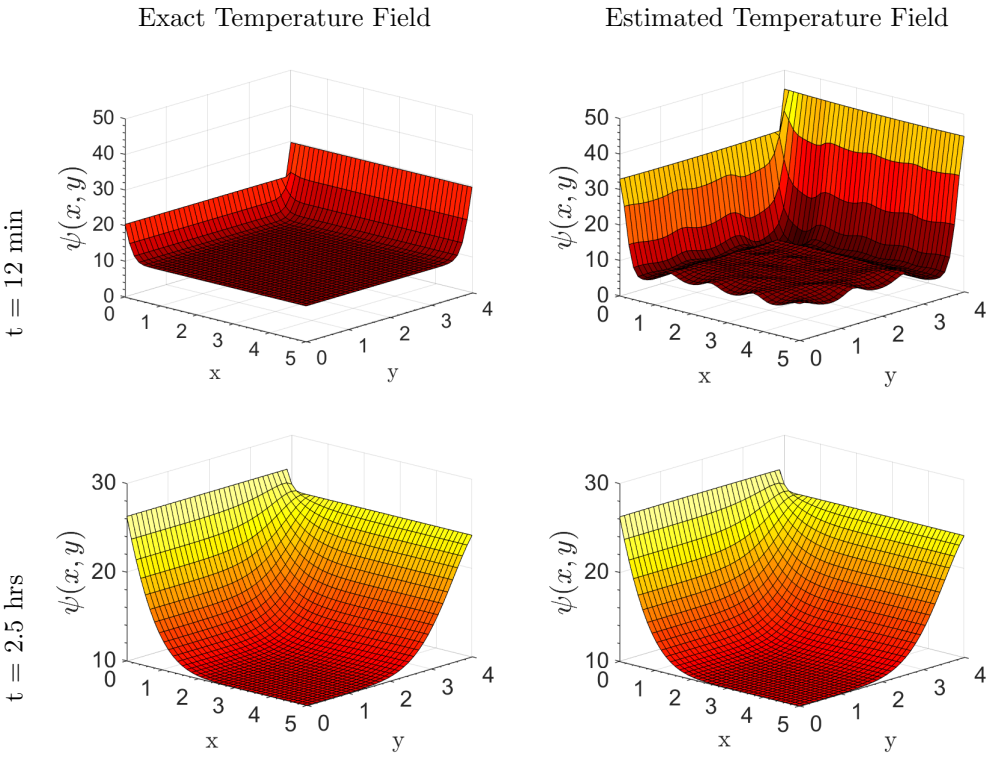


Figure 13.2: Performance of the Kalman Filter. This figure shows two snapshots of the exact temperature field and the estimated temperature field at $t = 12$ min and $t = 2.5$ hrs, respectively. The estimation process is carried out by deploying 14 transceivers to take line integral measurements. A Kalman filter algorithm that absorbed the dynamics of the unknown boundary conditions, as described by section 13.1, is employed to estimate the temperature field. Indeed, the figure shows the high accuracy of the estimation even if the initial estimate was considerably off from the actual temperature field.

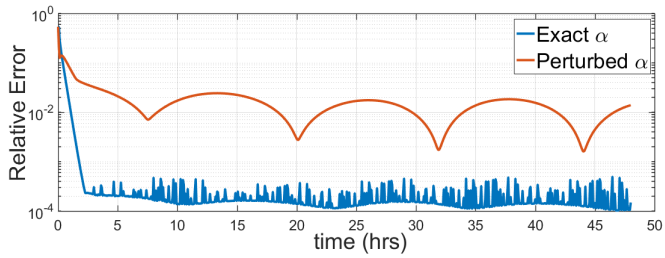


Figure 13.3: Comparison between the performance of the Kalman filter with an exact and a perturbed diffusion constant for the 2D heat equation. This figure shows the relative estimation error for two different scenarios. The first scenario assumes that the diffusion constant is predicted exactly. The second scenario assumes that the diffusion constant was predicted to be 0.1 when it is actually 0.05. The relative error was calculated by taking the norm of the estimation error relative to the norm of the true temperature field at each time instant. The red curve shows a drop in accuracy compared to the blue curve. However, taking into consideration the large perturbation of the model, the performance is still acceptable.

Chapter 14

Optimal Sensor Placement & Path Planning in Distributed Environments

After formulating the estimation problem in the previous chapter, we target the challenges of optimal sensor placement and path planning. This chapter deals with schemes to design the $\mathcal{C}$ operator in (13.5) using optimal control theory. For simplicity, we consider taking only one measurement (point-wise or line integral). Clearly, the optimization problem at hand requires a performance measure to quantify the estimation accuracy. [12] employs the mutual information between the estimated states and the measurements as a performance measure for point-wise sensor trajectory planning. In this paper, we intend to minimize the state estimation error variance which is the trace of the estimation error covariance. For this purpose, define the state estimate and estimation error covariance,

respectively as follows:

$$\begin{aligned} &\hat{\psi}(t) := E\{\psi(t)\} \\ &E\{[\psi(t) - \hat{\psi}(t)][\psi(\tau) - \hat{\psi}(\tau)]^*\} := \mathcal{X}(t)\delta(t-\tau). \end{aligned} \tag{14.1}$$

Note that, for notational convenience, the subscript of the state variable ψ_c in (13.5) is dropped for the rest of the paper. Let $\mathcal{C}_p$ denote the measurement operator in (13.3) parametrized by p. In the case of point-wise measurements, p is just the coordinates of the measurement location. Hence, $\mathcal{C}_p$ is a sampling operator that acts on a field ψ as follows:

$$\begin{aligned} &\mathcal{C}_p\psi := \psi(p) = \int_\Omega \delta(\boldsymbol{x} - p)\psi(\boldsymbol{x})d\boldsymbol{x} \\ &\mathcal{C}_p^*(\boldsymbol{x}) = \delta(\boldsymbol{x} - p) \end{aligned} \tag{14.2}$$

where $\mathcal{C}_p^*$ is the adjoint operator of $\mathcal{C}_p$. On the other hand, in the case of line integral measurements, p is the set of parameters of a line (for example polar coordinates in Fig. 14.1). Hence, $\mathcal{C}_p$ is a line integral operator that acts on a field ψ as follows:

$$\begin{aligned} &\mathcal{C}_p\psi := \int_{\Gamma_p} \psi(\boldsymbol{x})d\boldsymbol{x} = \int_\Omega \left(\int_{\Gamma_p} \delta(\boldsymbol{x} - \boldsymbol{x}_p(l))dl \right) \psi(\boldsymbol{x})d\boldsymbol{x} \\ &\mathcal{C}_p^*(\boldsymbol{x}) = \int_{\Gamma_p} \delta(\boldsymbol{x} - \boldsymbol{x}_p(l))dl \end{aligned} \tag{14.3}$$

where Γ_p is a line parametrized by p and $\boldsymbol{x}_p$ is a position vector spanning Γ_p (refer to Fig. 14.1). The propagation of the estimation error covariance $\mathcal{X}$ in the dynamics of (13.5) is governed by the continuous time Riccati equation. For notational convenience, we define the Riccati operator as follows:

$$\mathcal{R}(p, \mathcal{X}) := \mathcal{A}_c\mathcal{X} + \mathcal{X}\mathcal{A}_c^* + \mathcal{Q}_c - \frac{1}{R_c}\mathcal{X}\mathcal{C}_p^*\mathcal{C}_p\mathcal{X}. \tag{14.4}$$

Equipped with a performance metric, $tr(\mathcal{X})$, and a measurement operator $\mathcal{C}_p$, (14.2) and (14.3), we will show next how to find the optimal locations for fixed sensors, and the optimal trajectories for mobile sensors.

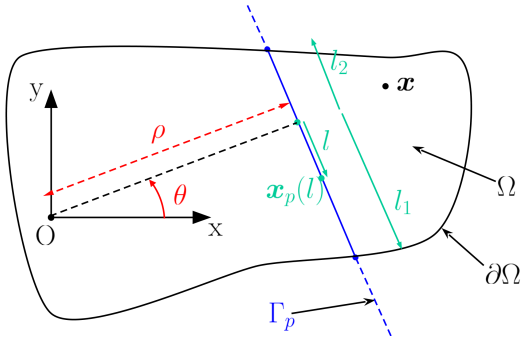


Figure 14.1: Parametrization of a line in a region Ω. Any line can be parametrized by its polar coordinates $p = (\rho, \theta)$. Let $x_p(l)$ denote a position vector that spans the line Γ_p as l varies between l_1 to l_2. Note that l_1 and l_2 specify the shape of the region Ω.

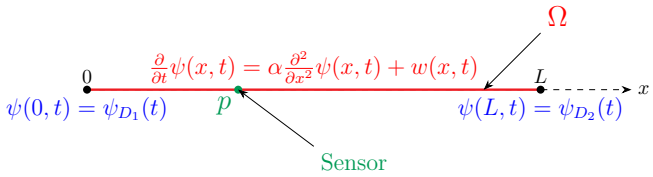


Figure 14.2: The one-dimensional dynamic heat equation is considered with a diffusion constant α. The two boundaries satisfy Dirichlet conditions. The sensor capable of taking point-wise measurements is located at $x = p$.

14.1 Optimal Static Sensor Placement

In this section, we consider another case study: the dynamical heat equation in one dimension with unknown periodic Dirichlet boundary conditions on both ends (refer to Fig. 14.2). Hence, in this example, $\psi_{D_1}(t)$ and $\psi_{D_2}(t)$ are unknown scalar periodic functions of time with an unknown frequency f. We apply the technique described in section 13.1 in a similar fashion to the case study in section 13.2. Thus, our modeled dynamics of the boundaries in (13.4) are given by (13.8), where $\mathcal{I}$, $\mathcal{P}_{D_1}$ and $\mathcal{P}_{D_2}$ are all scalars now. Our goal is to deploy one sensor, capable of taking point-wise measurements continuously in time, at an optimal fixed location $x = p$. The optimization objective is to

minimize the steady state [1] estimation error. For a given sensor location p_0, we know that the steady state estimation error covariance, $\mathcal{X}_{ss}$, solves the algebraic Riccati Equation: $\mathcal{R}(p_0, \mathcal{X}_{ss}) = 0$. Hence, we can pose the underlying optimization problem as follows:

$$P_0 : \begin{cases} \bar{p} = \underset{\{0 \le p \le L; \mathcal{X}_{ss}\}}{\operatorname{argmin}} & tr(\mathcal{X}_{ss}) \\ \text{s.t.} & \mathcal{R}(p, \mathcal{X}_{ss}) = 0 \end{cases} \tag{14.5}$$

To explore and characterize the problem, we do a brute force search by computing $tr(\mathcal{X}_{ss})$ for all possible values of p. Fig. 14.3 shows the cost function ($tr(\mathcal{X}_{ss})$) as a function of sensor location for different values of three design parameters: f_n, $\mathcal{P}_{D_2}$ and β, where we let $\mathcal{P}_w = \beta\mathcal{I}$ (recall that $\mathcal{P}_w$ is the covariance of the white Gaussian noise field in (13.1) and f_n is the natural frequency of the second order low pass filter introduced in section 13.2.2). The conclusions that can be drawn here are: (a) as the natural frequency f_n is increased compared to the actual frequency of the Dirichlet boundary conditions f, the filter realizes that the boundary conditions are varying rapidly and thus it would be more informative to measure towards the interior. (b) as β is increased, less trust is put into the interior of the model, thus the filter chooses to measure locations in the interior domain to compensate for the lack of trust. (c) as $\mathcal{P}_{D_2}$ is decreased relative to $\mathcal{P}_{D_1}$, more trust is put into the corresponding boundary and thus the filter chooses to measure the other boundary (Note that $\mathcal{P}_{D_1} = 20^2$).

14.2 Optimal Sensor Path Planning

In this section we allow the sensor to move around and take measurements (be it point-wise or line integral). Our goal here is to design an optimal path for the sensor. The optimization objective depends on the application requirement. In this section, we

[1] Steady state is achieved after the transient response (due to initial conditions) dies out. Hence, a steady state can be fixed or oscillatory.

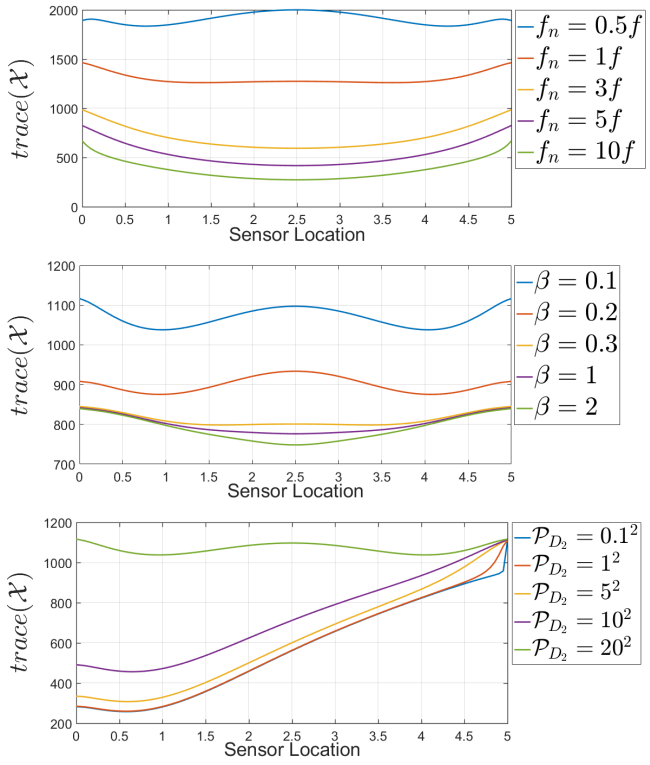


Figure 14.3: Effect of the design parameters on the optimal static sensor location for the 1D heat equation. This figure shows plots of the trace of the steady state estimation error covariance $\mathcal{X}_{ss}$ as a function of sensor location for point-wise measurements. The top plot to the left varies the natural frequency of the second order low pass filter f_n compared to the frequency f of the periodic Dirichlet boundary conditions. That to the right varies the intensity of the process noise inside the region of the heat equation. The plot at the bottom varies the covariance of the white Gaussian noise feeding the low pass filter of one of the two boundary conditions.

will consider two different objective functions.

This optimization can be thought of as an optimal control problem. We let the control to be the velocity of the sensor in order to penalize it in the objective function. Otherwise, the sensor would be allowed to move instantaneously from one location to another. Hence the states are the covariance operator $\mathcal{X}$ and the sensor location p. Two

different problems can be posed here:

$$P_1 : \begin{cases} \bar{p}(t) = \underset{\{p(t);\mathcal{X}(t)\}}{\operatorname{argmin}} & tr(\mathcal{X}(t_f)) + \dfrac{\mu}{2}\displaystyle\int_0^{t_f} (\dot{p}(t))^2 dt \\ \text{s.t.} & \dot{\mathcal{X}}(t) = \mathcal{R}(p(t), \mathcal{X}(t)); \quad \mathcal{X}(0) = \mathcal{X}_0 \\ & \dot{p}(t) = u(t); \qquad\qquad p(0) = p_0 \end{cases} \tag{14.6}$$

$$P_2 : \begin{cases} \bar{p}(t) = \underset{\{p(t);\mathcal{X}(t)\}}{\operatorname{argmin}} & \displaystyle\int_0^{t_f} \left(tr(\mathcal{X}(t)) + \frac{\mu}{2}(\dot{p}(t))^2 \right) dt \\ \text{s.t.} & \dot{\mathcal{X}}(t) = \mathcal{R}(p(t), \mathcal{X}(t)); \quad \mathcal{X}(0) = \mathcal{X}_0 \\ & \dot{p}(t) = u(t); \qquad\qquad p(0) = p_0 \end{cases} \tag{14.7}$$

where μ is the mobility penalty of the sensor. P_1 and P_2 are nonlinear optimal control problems. P_1 searches for the optimal control that gives the best estimate at the final time. So, given a time duration t_f, the sensors are allowed to move to give the best estimate at the end of the given time duration. On the other hand, P_2 searches for the optimal control that minimizes the estimation error as the sensor is moving. To solve the two optimal control problems, we form the Hamiltonian then develop the costate equations. To do so, we need to calculate the Frechét derivatives: $\left(\frac{\partial}{\partial p}\mathcal{R}(p,\mathcal{X})\right)(\delta p)$ and $\left(\frac{\partial}{\partial \mathcal{X}}\mathcal{R}(p,\mathcal{X})\right)(\delta \mathcal{X})$. These are the directional partial derivatives of $\mathcal{R}(p,\mathcal{X})$ in the directions of δp and $\delta\mathcal{X}$, respectively. It can be shown [2] that

$$\begin{aligned} \left(\frac{\partial}{\partial p}\mathcal{R}(p,\mathcal{X})\right)(\delta p) &= -\frac{1}{R_c}\mathcal{X}\mathcal{W}_p(\delta p)\mathcal{X} \\ \left(\frac{\partial}{\partial \mathcal{X}}\mathcal{R}(p,\mathcal{X})\right)(\delta \mathcal{X}) &= [\mathcal{A}_c - \mathcal{L}_p\mathcal{C}_p]\,\delta\mathcal{X} + \delta\mathcal{X}\,[\mathcal{A}_c - \mathcal{L}_p\mathcal{C}_p]^* \end{aligned} \tag{14.8}$$

where $\mathcal{L}_p := \mathcal{X}\mathcal{C}_p^* R_c^{-1}$ is the Kalman gain and $\mathcal{W}_p(\delta p) := \left(\frac{\partial}{\partial p}[\mathcal{C}_p^*\mathcal{C}_p]\right)(\delta p)$. The Hamiltonian functions for P_1 and P_2 are denoted by $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively.

$$\begin{aligned} \mathcal{H}_1 &:= \frac{\mu}{2}u^2(t) + \langle \Lambda(t), \mathcal{R}(p(t),\mathcal{X}(t))\rangle + \langle \lambda(t), u(t)\rangle \\ \mathcal{H}_2 &:= tr(\mathcal{X}(t)) + \mathcal{H}_1 \end{aligned} \tag{14.9}$$

[2] The derivations for a simpler (finite dimensional) setting is given in Appendix B.

The costates for P_1 can be shown to be:

$$\begin{aligned}
\dot{\Lambda}(t) &= -\Lambda(t)[\mathcal{A}_c - \mathcal{L}_{p(t)}(t)\mathcal{C}_{p(t)}] - [\mathcal{A}_c - \mathcal{L}_{p(t)}(t)\mathcal{C}_{p(t)}]^*\Lambda(t) \\
\dot{\lambda}(t) &= \frac{1}{R_c} tr(\Lambda^*(t)\mathcal{X}(t)\mathcal{W}_{p(t)}\mathcal{X}(t)) \\
\Lambda(t_f) &= -\mathcal{X}^2(t_f); \qquad \lambda(t_f) = 0
\end{aligned} \tag{14.10}$$

The costates for P_2 can be shown to be:

$$\begin{aligned}
\dot{\Lambda}(t) &= -\mathcal{I} - \Lambda(t)[\mathcal{A}_c - \mathcal{L}_{p(t)}(t)\mathcal{C}_{p(t)}] - \\
&\qquad\qquad [\mathcal{A}_c - \mathcal{L}_{p(t)}(t)\mathcal{C}_{p(t)}]^*\Lambda(t) \\
\dot{\lambda}(t) &= \frac{1}{R_c} tr(\Lambda^*(t)\mathcal{X}(t)\mathcal{W}_{p(t)}\mathcal{X}(t)) \\
\Lambda(t_f) &= 0; \qquad \lambda(t_f) = 0
\end{aligned} \tag{14.11}$$

The state equations for both P_1 and P_2 are the same:

$$\begin{cases}
\dot{\mathcal{X}}(t) = \mathcal{R}(p(t), \mathcal{X}(t)); & \mathcal{X}(0) = \mathcal{X}_0 \\
\dot{p}(t) = u(t); & p(0) = p_0 \\
u(t) = -\dfrac{\lambda(t)}{\mu}
\end{cases} \tag{14.12}$$

To solve the optimal control problems P_1 and P_2, one needs to solve the costate and state equations in (14.10) through (14.12). This is, numerically, a very large scale problem since the states are typically large covariance matrices. Efficient numerical schemes to tackle these problems are currently under investigation.

14.2.1 Sub-optimal Path Planning in Discrete Time

We present a sub-optimal algorithm for the path planning problem in discrete time. First, we discretize (13.5) in time to get:

$$\begin{cases}
\psi_{k+1} = \mathcal{A}_d\psi_k + w_k \\
m_k = \mathcal{C}_{p_k}\psi_k + v_k
\end{cases} \tag{14.13}$$

with

$$\mathcal{A}_d := \exp(\mathcal{A}_c \Delta t); \quad \mathcal{Q}_d := \int_0^{\Delta t} \exp(\mathcal{A}_d \tau) \mathcal{Q}_c \exp(\mathcal{A}_d^* \tau) d\tau$$
$$R_d := \frac{R_c}{\Delta t}; \quad E[w_k w_s^*] := \mathcal{Q}_d \delta_{ks}; \quad E[v_k v_s^*] = R_d \delta_{ks}$$

where δ_{ks} is the Kronecker delta and Δt is the discretization time step. Note that $\mathcal{Q}_d$ can be computed using Van Loan's algorithm [59] for example. The propagation of the covariance in discrete time is dictated by the discrete time Riccati equation:

$$\begin{aligned} \mathcal{Y}_k &= \mathcal{A}_d \mathcal{X}_{k-1} \mathcal{A}_d^* + \mathcal{Q}_d \\ \mathcal{X}_k &= \left[\mathcal{Y}_k^{-1} + \mathcal{C}_{p_k}^* R_d^{-1} \mathcal{C}_{p_k}\right]^{-1} \end{aligned} \tag{14.14}$$

Hence, at each time step we have an optimization problem to be solved as follows. Given p_{k-1} and $\mathcal{X}_{k-1}$, select p_k that minimizes the estimation error at time step k. For the case study explained in section 14.1 but with a moving point-wise sensor, the optimization problem can be written as follows:

$$\begin{aligned} \bar{p}_k = \underset{\{0 \le p_k \le L; \mathcal{X}_k\}}{\text{argmin}} \quad & tr(\mathcal{X}_k) + \frac{\mu}{2\Delta t^2}(p_k - p_{k-1})^2 \\ \text{s.t.} \quad & \mathcal{Y}_k = \mathcal{A}_d \mathcal{X}_{k-1} \mathcal{A}_d^* + \mathcal{Q}_d \\ & \mathcal{X}_k = \left[\mathcal{Y}_k^{-1} + \mathcal{C}_{p_k}^* R_d^{-1} \mathcal{C}_{p_k}\right]^{-1} \end{aligned} \tag{14.15}$$

Fig. 14.4 plots the optimal trajectory for different set of design parameters. In fact, for the set of design parameters used, the optimal trajectory turned out to be periodic. The period and the shape of the trajectory depend on the design parameters used. Note that the typical numerical values used for the design parameters are as follows: $\mu = 0.1$, $f_n = 3f$, $\beta = 5^2$, and $\mathcal{P}_{D_1} = \mathcal{P}_{D_2} = 40^2$, where $f = 1$ reflects a time scale of heat transfer on the rod (refer to Appendix A). Fig. 14.4 shows only the values of the modified design parameters. As a matter of fact, for higher mobility penalty, the sensor tends to move less and spends more time on the boundaries. Moreover, when $\mathcal{P}_{D_2}$ is decreased, the sensor visits the second boundary for smaller duration of time to reflect higher trust in

the latter. On the other hand, higher values of f_n indicates that the boundaries are allowed to vary faster and thus the sensor stays at the boundaries for longer periods of time. Finally, smaller values of β indicates that we trust the interior of the model more, thus the sensor visits the interior less frequently.

14.3 Conclusion and Future Work

This paper approaches the optimal estimation problem in distributed dynamic environments, where measurements taken by available mobile sensors and physics-based models are assimilated to enhance the estimation accuracy. The optimal sensor path planning was then cast as a continuous time-space optimal control problem. The necessary conditions of optimality were derived to yield operator valued state and costate differential equations. Efficient numerical methods to solve this, generally, large scale optimal control problem are currently under investigation. It is believed that the optimal control has a special structure (such as periodic) as shown in the discrete time version solution of the one dimensional heat equation example. Solving the optimal control problem will give us insights on the structure of the optimal sensor path. On the other hand, other applications such as flow estimation will be considered. In this application, the physics-based model to be employed is the nonlinear Navier-Stokes equation. Acoustic tomography sensing techniques can also be used, using the framework developed in section 13.1, to design sensor trajectories that optimally estimate the flow fields.

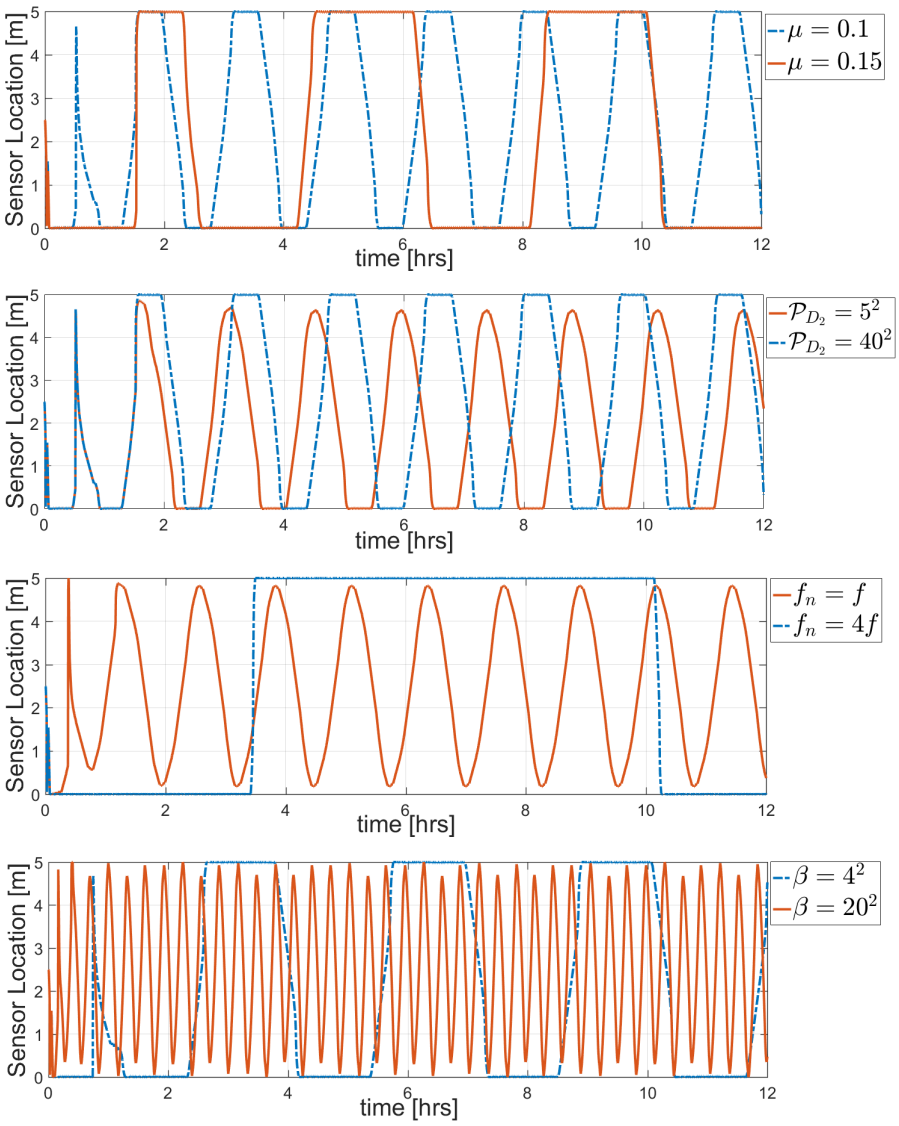


Figure 14.4: Optimal paths for a mobile sensor capable of taking point-wise measurements using the discrete time Kalman filter algorithm for the 1D heat equation. The four plots show the calculated optimal paths for two different values of the mobility penalty μ, the error covariance $\mathcal{P}_{D_2}$ of the white Gaussian noise feeding one of the Dirichlet boundaries, the natural frequency f_n of the low pass filter and the intensity of the process noise β inside the region of the heat equation.

Appendix

14.A Heat Equation on a Rod: Transfer Functions

Consider the one-dimensional simplified heat transfer problem on a rod that is dynamically heated on the left boundary. We are interested in studying how heat propagates in the rod in response to the dynamically heated end. We will analyze two problems. In the first, we consider an semi-infinite rod and calculate a transfer function from $x = 0$ to $x = L$. Then, we consider a finite rod of length L that is thermally isolated on the right boundary. Mathematically, the dynamics for the two problems are give by

$$\mathcal{M}_\infty : \begin{cases} \partial_t \psi(x,t) = \alpha \partial_x^2 \psi(x,t); \quad x \in [0, +\infty) \\ \psi(0,t) = u(t) \\ \psi(x,0) = 0 \end{cases}$$

$$\mathcal{M}_L : \begin{cases} \partial_t \psi(x,t) = \alpha \partial_x^2 \psi(x,t); \quad x \in [0, L] \\ \psi(0,t) = u(t) \\ \partial_x \psi(L,t) = 0 \\ \psi(x,0) = 0. \end{cases} \tag{14.A.1}$$

Our goal is find a transfer function between the input heater $u(t)$ and the output $y(t) = \psi(L,t)$ for both problems $\mathcal{M}_\infty$ and $\mathcal{M}_L$.

14.A.1 Heat Equation on the Semi-infinite Line

In this section, we consider $\mathcal{M}_\infty$. This is approached by taking the Laplace transform in time. Taking Laplace transforms in time of $\mathcal{M}_\infty$, we get:

$$s\hat{\psi}(x,s) = \alpha\partial_x^2\hat{\psi}(x,s)$$
$$\hat{\psi}(0,s) = \hat{u}(s).$$

This really can be posed as follows:

$$\partial_x \begin{bmatrix} \hat{\psi}(x,s) \\ \partial_x\hat{\psi}(x,s) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{s}{\alpha} & 0 \end{bmatrix} \begin{bmatrix} \hat{\psi}(x,s) \\ \partial_x\hat{\psi}(x,s) \end{bmatrix}; \qquad \begin{bmatrix} \hat{\psi}(0,s) \\ \partial_x\hat{\psi}(0,s) \end{bmatrix} = \begin{bmatrix} \hat{u}(s) \\ \hat{v}(s) \end{bmatrix},$$

where $\hat{v}(s)$ is unknown and needs to be determined from the extra condition that

$$\lim_{x\to+\infty} \psi(x,t) = \lim_{x\to+\infty} \hat{\psi}(x,s) = 0.$$

The solution can be written using the variation of constants formula as

$$\begin{aligned} \begin{bmatrix} \hat{\psi}(x,s) \\ \partial_x\hat{\psi}(x,s) \end{bmatrix} &= e^{\left(\begin{bmatrix} 0 & 1 \\ \frac{s}{\alpha} & 0 \end{bmatrix} x\right)} \begin{bmatrix} \hat{u}(s) \\ \hat{v}(s) \end{bmatrix} \\ &= \begin{bmatrix} \cosh\left(\sqrt{\frac{s}{\alpha}}x\right) & \sqrt{\frac{\alpha}{s}}\sinh\left(\sqrt{\frac{s}{\alpha}}x\right) \\ \sqrt{\frac{s}{\alpha}}\sinh\left(\sqrt{\frac{s}{\alpha}}x\right) & \cosh\left(\sqrt{\frac{s}{\alpha}}x\right) \end{bmatrix} \begin{bmatrix} \hat{u}(s) \\ \hat{v}(s) \end{bmatrix} \end{aligned}$$

Writing the first equation of the matrix equation, we have

$$\begin{aligned} \hat{\psi}(x,s) &= \cosh\left(\sqrt{\frac{s}{\alpha}}x\right)\hat{u}(s) + \sqrt{\frac{\alpha}{s}}\sinh\left(\sqrt{\frac{s}{\alpha}}x\right)\hat{v}(s) \\ &= \left(e^{\sqrt{\frac{s}{\alpha}}x} + e^{-\sqrt{\frac{s}{\alpha}}x}\right)\frac{\hat{u}(s)}{2} + \left(e^{\sqrt{\frac{s}{\alpha}}x} - e^{-\sqrt{\frac{s}{\alpha}}x}\right)\sqrt{\frac{\alpha}{s}}\frac{\hat{v}(s)}{2} \\ &= \frac{e^{\sqrt{\frac{s}{\alpha}}x}}{2}\left(\hat{u}(s) + \sqrt{\frac{\alpha}{s}}\hat{v}(s)\right) + \frac{e^{-\sqrt{\frac{s}{\alpha}}x}}{2}\left(\hat{u}(s) - \sqrt{\frac{\alpha}{s}}\hat{v}(s)\right) \end{aligned}$$

Observe that for the limit of $\hat{\psi}(x,s)$ as $x \to \infty$ to vanish, we have $\hat{u}(s) + \sqrt{\frac{\alpha}{s}}\hat{v}(s) = 0$. Then

$$\hat{\psi}(x,s) = e^{-\sqrt{\frac{s}{\alpha}}x}.$$

Finally, the output is given by $y(t) = \psi(L,t)$, then

$$\frac{\hat{y}(s)}{\hat{u}(s)} = e^{-\sqrt{\frac{s}{\alpha}}L}. \tag{14.A.2}$$

The frequency response is then given by

$$\begin{aligned}\frac{\hat{y}(j2\pi f)}{\hat{u}(j2\pi f)} &= e^{-\sqrt{\frac{j2\pi f}{\alpha}}L} = e^{-\sqrt{\frac{2\pi f}{\alpha}}L\sqrt{j}} = e^{-\sqrt{\frac{2\pi f}{\alpha}}L\frac{1}{\sqrt{2}}(1+j)} \\ &= e^{-\sqrt{\frac{L^2\pi f}{\alpha}}}e^{-j\sqrt{\frac{L^2\pi f}{\alpha}}}\end{aligned}$$

Therefore the magnitude and phase of the transfer function are given by

$$\left\|\frac{\hat{y}(j2\pi f)}{\hat{u}(j2\pi f)}\right\| = e^{-\sqrt{\frac{L^2\pi f}{\alpha}}} \qquad \text{Phase}\left(\frac{\hat{y}(j2\pi f)}{\hat{u}(j2\pi f)}\right) = -\sqrt{\frac{L^2\pi f}{\alpha}}. \tag{14.A.3}$$

For $L = 5m$, $\alpha = 0.05m^2/s$, the frequency response is depicted in Figure 14.A.1.

14.A.2 Solution of the Heat Equation with Inhomogeneous Dirichlet and Homogeneous Neumann Boundary Conditions, Method 1

To transform $\mathcal{M}_L$ in (14.A.1) so that the boundary conditions become homogeneous, we define a new state space variable $\Psi(x,t) := \psi(x,t) - u(t)$. Hence, the dynamics in the new state space variable Ψ can be expressed as

$$\begin{aligned}\partial_t\Psi(x,t) &= \alpha\partial_x^2\Psi(x,t) - \dot{u}(t) \\ \Psi(0,t) &= 0 \\ \partial_x\Psi(L,t) &= 0 \\ \Psi(x,0) &= g(x) - u(0).\end{aligned} \tag{14.A.4}$$

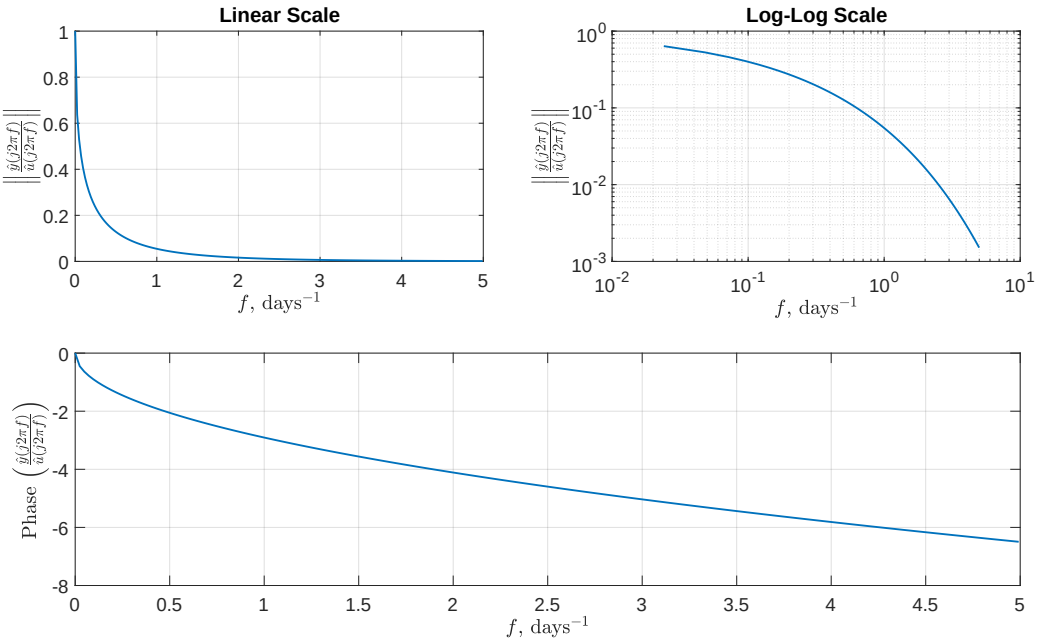


Figure 14.A.1: Frequency Response at a location $x = L$ of a semi-infinite rod heated at $x = 0$.

Now, we will solve the easier problem (14.A.4) since it has homogeneous boundary conditions. Define the following operator

$$\nabla\Psi(x) = \partial_x^2\Psi(x); \text{ with domain } \mathcal{D}(\nabla) := \{\Psi \in L_2[0,L]; \partial_x^2 \in L_2[0,L]; \Psi(0) = \partial_x\Psi(L) = 0\}.$$

It can be shown that this operator is self-adjoint and it has a full set of orthonormal eigenfunctions given by:

$$\lambda_n = -\left(n + \frac{1}{2}\right)^2 \frac{\pi^2}{L^2} \qquad \longleftrightarrow \qquad \phi_n(x) = \sqrt{\frac{2}{L}}\sin(\sqrt{-\lambda_n}x) \qquad n \in \mathbb{N}.$$

Knowing that the eigenfunctions for $n \in \mathbb{N}$ form an orthonormal basis, we can expand any $\Psi \in \mathcal{D}(\nabla)$ as

$$\Psi(x,t) = \sum_{n\in\mathbb{N}} \hat{\Psi}_n(t)\phi_n(x) \qquad \longleftrightarrow \qquad \hat{\Psi}_n(t) = \langle \phi_n, \Psi(.,t)\rangle,$$

where $\langle ., . \rangle$ is the $L_2[0, L]$ inner product defined as $\langle f, g \rangle = \int_0^L f(x)g(x)dx$. We substitute the basis expansion in (14.A.4) and proceed as

$$\sum_{n\in\mathbb{N}} \frac{d}{dt}\hat{\Psi}_n(t)\phi_n(x) = \alpha \sum_{n\in\mathbb{N}} \hat{\Psi}_n(t)\lambda_n\phi_n(x) - \dot{u}(t)$$

$$\implies \langle \phi_k, \sum_{n\in\mathbb{N}} \frac{d}{dt}\hat{\Psi}_n(t)\phi_n \rangle = \langle \phi_k, \sum_{n\in\mathbb{N}} \alpha\hat{\Psi}_n(t)\lambda_n\phi_n - \dot{u}(t) \rangle \qquad \text{(Project on } \phi_k\text{)}$$

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) - \int_0^L \dot{u}(t)\phi_k(x)dx$$

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) - \dot{u}(t)\sqrt{\frac{2}{L}}\int_0^L \sin(\sqrt{-\lambda_k}x)dx$$

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) + \dot{u}(t)\sqrt{\frac{2}{L}}\frac{1}{\sqrt{-\lambda_k}}\cos(\sqrt{-\lambda_k}x)\Big|_0^L$$

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) + \dot{u}(t)\sqrt{\frac{2}{L}}\frac{L}{(n+\frac{1}{2})\pi}\Big(\cos(\sqrt{-\lambda_k}L) - 1\Big)$$

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) + \mu_k\dot{u}(t) \qquad \Big(\mu_k := -\frac{\sqrt{2L}}{(k+\frac{1}{2})\pi}\Big).$$

The initial condition can be calculated as

$$\hat{\Psi}_k(0) = \langle \phi_k, \Psi(., 0) \rangle = \langle \phi_k, g - u(0) \rangle = \langle \phi_k, g \rangle - u(0)\sqrt{\frac{2}{L}}\int_0^L \sin(\sqrt{-\lambda_k}x)dx$$

$$= \hat{g} + \mu_k u(0)$$

Therefore, the coefficients of $\Psi(x, t)$ in the basis $\{\phi_k\}_{k\in\mathbb{N}}$ evolves according to the following set of decoupled ordinary differential equations

$$\frac{d}{dt}\hat{\Psi}_k(t) = \alpha\lambda_k\hat{\Psi}_k(t) + \mu_k\dot{u}(t); \qquad \hat{\Psi}_k(0) = \hat{g} + \mu_k u(0),$$

where $\mu_k := -\frac{\sqrt{2L}}{(k+\frac{1}{2})\pi}$. The solution of these decoupled ODEs is easily obtained using the variation of constants formula

$$\hat{\Psi}_n(t) = e^{\alpha\lambda_n t}\hat{\Psi}_n(0) + \int_0^t e^{\alpha\lambda_n(t-\tau)}\mu_n\dot{u}(\tau)d\tau; \qquad \hat{\Psi}_n(0) = \hat{g} + \mu_k u(0).$$

We, now use integration by parts to express $\hat{\Psi}_n(t)$ in terms of $u(t)$ rather than $\dot{u}(t)$.

$$\hat{\Psi}_n(t) = e^{\alpha\lambda_n t}\hat{\Psi}_n(0) + \mu_n e^{\alpha\lambda_n t}\int_0^t e^{-\alpha\lambda_n\tau}\dot{u}(\tau)d\tau$$

$$
\begin{aligned}
&= e^{\alpha\lambda_n t}\hat{\Psi}_n(0) + \mu_n e^{\alpha\lambda_n t}\left(e^{-\alpha\lambda_n\tau}u(\tau)\Big|_0^t + \alpha\lambda_n\int_0^t e^{-\alpha\lambda_n\tau}u(\tau)d\tau\right)\\
&= e^{\alpha\lambda_n t}\hat{\Psi}_n(0) + \mu_n e^{\alpha\lambda_n t}\left(e^{-\alpha\lambda_n t}u(t) - u(0) + \alpha\lambda_n\int_0^t e^{-\alpha\lambda_n\tau}u(\tau)d\tau\right)\\
&= e^{\alpha\lambda_n t}\left(\hat{g} + \mu_n u(0) + \mu_n\left(e^{-\alpha\lambda_n t}u(t) - u(0) + \alpha\lambda_n\int_0^t e^{-\alpha\lambda_n\tau}u(\tau)d\tau\right)\right)\\
&= e^{\alpha\lambda_n t}\hat{g} + \mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau.
\end{aligned}
$$

The solution of (14.A.4) can thus be written as

$$
\Psi(x,t) = \sum_{n\in\mathbb{N}}\left(e^{\alpha\lambda_n t}\hat{g} + \mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\right)\phi_n(x)
$$

Therefore, the solution in the original state space is given by

$$
\psi(x,t) = u(t) + \sum_{n\in\mathbb{N}}\left(e^{\alpha\lambda_n t}\hat{g} + \mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\right)\phi_n(x)
$$

Transfer Function: In this section, we assume that $g(x) = 0$, then the output $y(t)$ can be calculated as

$$
\begin{aligned}
y(t) = \psi(L,t) &= u(t) + \sum_{n\in\mathbb{N}}\left(\mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\right)\phi_n(L)\\
&= u(t) + \sum_{n\in\mathbb{N}}\left(\mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\right)\sqrt{\frac{2}{L}}\sin(\sqrt{-\lambda_n}L)\\
&= u(t) + \sqrt{\frac{2}{L}}\sum_{n\in\mathbb{N}}\left(\mu_n u(t) + \alpha\lambda_n\mu_n\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\right)(-1)^n\\
&= u(t) - \sqrt{\frac{2}{L}}\sum_{n\in\mathbb{N}}(-1)^n\frac{\sqrt{2L}}{(n+\frac{1}{2})\pi}u(t)\\
&\qquad + \alpha\sqrt{\frac{2}{L}}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)^2\frac{\pi^2}{L^2}\frac{\sqrt{2L}}{(n+\frac{1}{2})\pi}\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\\
&= u(t) - \frac{2}{\pi}u(t)\sum_{n\in\mathbb{N}}\frac{(-1)^n}{n+\frac{1}{2}} + \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau\\
&= u(t) - \frac{4}{\pi}u(t)\sum_{n\in\mathbb{N}}\frac{(-1)^n}{2n+1} + \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau
\end{aligned}
$$

$$
\begin{aligned}
&= u(t) - \frac{4}{\pi}u(t)\tan^{-1}(1) + \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau \\
&= u(t) - \frac{4}{\pi}u(t)\frac{\pi}{4} + \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau \\
\implies\ y(t) &= \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}}(-1)^n\left(n+\frac{1}{2}\right)\int_0^t e^{\alpha\lambda_n(t-\tau)}u(\tau)d\tau
\end{aligned}
$$

This is an input-output dynamical system that can be realized in state space as explained next. First define a family of state space variables as follows:

$$
z_n(t) := \int_0^t e^{\alpha\lambda_n(t-\tau)}(-1)^n\left(n+\frac{1}{2}\right)u(\tau)d\tau, \qquad (n\in\mathbb{N}).
$$

Then the output can be written as

$$
y(t) = \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}} z_n(t).
$$

The state space realization can thus be expressed as an infinite dimensional system as follows:

$$
\begin{cases}
\dfrac{d}{dt}\begin{bmatrix} z_0(t) \\ z_1(t) \\ \vdots \\ z_n(t) \\ \vdots \end{bmatrix} = \begin{bmatrix} \alpha\lambda_0 & & & & \\ & \alpha\lambda_1 & & & \\ & & \ddots & & \\ & & & \alpha\lambda_n & \\ & & & & \ddots \end{bmatrix}\begin{bmatrix} z_0(t) \\ z_1(t) \\ \vdots \\ z_n(t) \\ \vdots \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ -\frac{3}{2} \\ \vdots \\ (-1)^n\left(n+\frac{1}{2}\right) \\ \vdots \end{bmatrix} u(t); \quad \begin{bmatrix} z_0(0) \\ z_1(0) \\ \vdots \\ z_n(0) \\ \vdots \end{bmatrix} = 0 \\
\qquad y(t) = \begin{bmatrix} \alpha\frac{2\pi}{L^2} & \alpha\frac{2\pi}{L^2} & \cdots & \alpha\frac{2\pi}{L^2} & \cdots \end{bmatrix}\begin{bmatrix} z_0(t) \\ z_1(t) \\ \vdots \\ z_n(t) \\ \vdots \end{bmatrix}
\end{cases}
\tag{14.A.5}
$$

The transfer function is thus obtained by taking Laplace transforms:

$$\hat{y}(s) = \begin{bmatrix} \alpha\frac{2\pi}{L^2} & \alpha\frac{2\pi}{L^2} & \cdots & \alpha\frac{2\pi}{L^2} & \cdots \end{bmatrix} \begin{bmatrix} \frac{1}{s-\alpha\lambda_0} & & & & \\ & \frac{1}{s-\alpha\lambda_1} & & & \\ & & \ddots & & \\ & & & \frac{1}{s-\alpha\lambda_n} & \\ & & & & \ddots \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ -\frac{3}{2} \\ \vdots \\ (-1)^n\left(n+\frac{1}{2}\right) \\ \vdots \end{bmatrix} \hat{u}(s)$$

Therefore

$$\hat{H}(s) := \frac{\hat{y}(s)}{\hat{u}(s)} = \alpha\frac{2\pi}{L^2}\sum_{n\in\mathbb{N}} \frac{(-1)^n\left(n+\frac{1}{2}\right)}{s+\left(n+\frac{1}{2}\right)^2\alpha\frac{\pi^2}{L^2}}. \tag{14.A.6}$$

By defining $c_n := \alpha\frac{\pi^2}{L^2}\left(n+\frac{1}{2}\right)^2$, we can rewrite $\hat{H}(s)$ as

$$\hat{H}(s) = \frac{2\sqrt{\alpha}}{L}\sum_{n\in\mathbb{N}}(-1)^n\frac{\sqrt{c_n}}{s+c_n} \qquad \longleftrightarrow \qquad H(t) = \frac{2\sqrt{\alpha}}{L}\sum_{n\in\mathbb{N}}(-1)^n\sqrt{c_n}e^{-c_n t}.$$

14.A.3 Solution of the Heat Equation with Inhomogeneous Dirichlet and Homogeneous Neumann Boundary Conditions, Method 2

By taking the Laplace transform of $\mathcal{M}_L$ in (14.A.1), we obtain a two point Boundary Value Problem in space, where we treat the Laplace variable s as a parameter. The BVP can be written as

$$\begin{aligned} s\hat{\psi}(x,s) &= \alpha\partial_x^2\hat{\psi}(x,s) \\ \hat{\psi}(0,s) &= \hat{u}(s) \\ \partial_x\hat{\psi}(L,s) &= 0 \end{aligned} \tag{14.A.7}$$

The solution to the BVP can be easily calculated to be

$$\hat{\psi}(x,s) = \left(\cosh\left(\sqrt{\frac{s}{\alpha}}x\right) - \sinh\left(\sqrt{\frac{s}{\alpha}}x\right)\tanh\left(\sqrt{\frac{s}{\alpha}}L\right)\right)\hat{u}(s).$$

Define a transfer function at location x as $\hat{H}(x,s) := \frac{\hat{\psi}(x,s)}{\hat{u}(s)}$. This can be rewritten as

$$\begin{aligned}\hat{H}(x,s) &= \cosh\left(\sqrt{\frac{s}{\alpha}}x\right) - \sinh\left(\sqrt{\frac{s}{\alpha}}x\right)\tanh\left(\sqrt{\frac{s}{\alpha}}L\right)\\ &= \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}x\right)\cosh\left(\sqrt{\frac{s}{\alpha}}L\right) - \sinh\left(\sqrt{\frac{s}{\alpha}}x\right)\sinh\left(\sqrt{\frac{s}{\alpha}}L\right)}{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}\\ &= \frac{\cosh\left(\sqrt{\frac{s}{\alpha}}(x-L)\right)}{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}.\end{aligned}$$

Finally, we have

$$\hat{H}(L,s) = \frac{1}{\cosh\left(\sqrt{\frac{s}{\alpha}}L\right)}. \tag{14.A.8}$$

Note that this derivation, by comparing with (14.A.6), allows us to compute the following sum (for $\alpha = 1$ and $L = \pi$)

$$\frac{2}{\pi}\sum_{n\in\mathbb{N}}(-1)^n\frac{n+\frac{1}{2}}{s+\left(n+\frac{1}{2}\right)^2} = \frac{1}{\cosh\left(\pi\sqrt{s}\right)}.$$

For $L = 5m$, $\alpha = 0.05m^2/s$, the frequency response is depicted in Figure 14.A.2. Observe that the ga

14.B Derivation of Sufficient Conditions of Optimality: A Finite Dimensional Example

In this appendix, we derive the necessary conditions of optimality of the sensor motion problem of Chapter 14. The setting considered here is the finite dimensional setting for simplicity.

14.B.1 System Dynamics

Consider the following linear, finite dimensional dynamical system

$$\begin{aligned}\dot{\psi}(t) &= A\psi(t) + w(t); \qquad \psi(0) = \psi_0\\ y(t) &= C(p(t))\psi(t) + v(t);\end{aligned} \tag{14.B.1}$$

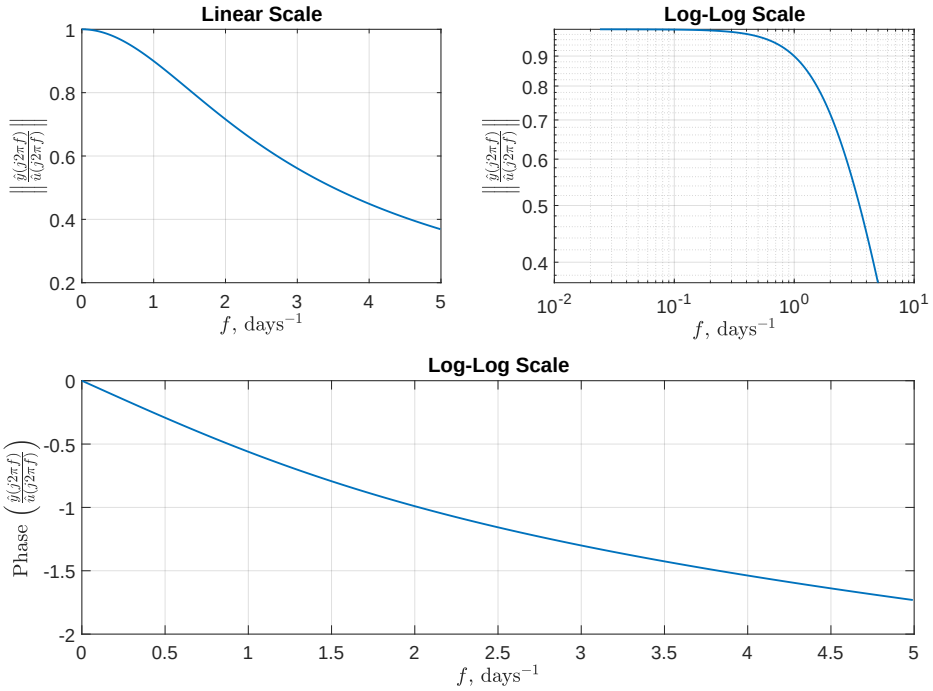


Figure 14.A.2: Frequency Response at a location $x = L$ of a Finite rod heated at $x = 0$.

where $w(t)$ and $v(t)$ are zero-mean Gaussian White Noise, such that

$$E\{w(t)w^T(t)\} := Q\delta(t-\tau)$$

$$E\{v(t)v^T(t)\} := R\delta(t-\tau)$$

and $p(t) \in \mathbb{R}$ is a scalar variable that parameterizes one sensor. The dimensions are given below:

$$\psi(t), w(t) \in \mathbb{R}^N; \qquad y(t), v(t), p(t) \in \mathbb{R}$$

$$A \in \mathbb{R}^{N\times N}; \qquad C(p(t)) \in \mathbb{R}^{1\times N}$$

$$Q = Q^T \geq 0 \in \mathbb{R}^{N\times N}; \qquad R > 0 \in \mathbb{R}$$

14.B.2 Sensor Dynamics

Let the sensor dynamics be governed by the state space equation

$$\begin{aligned} \dot{z}(t) &= A_s z(t) + B_s u(t); \qquad z(0) = z_0 \\ p(t) &= C_s z(t); \end{aligned} \tag{14.B.2}$$

where $z(t)$ and $u(t)$ are the state space variable and control of the sensor dynamics, respectively. The dimensions are given below:

$$\begin{aligned} z(t) &\in \mathbb{R}^{N_z}; \qquad u(t) \in \mathbb{R}^{N_u}; \qquad p(t) \in \mathbb{R} \\ A_s &\in \mathbb{R}^{N_z \times R_z}; \qquad B_s \in \mathbb{R}^{N_z \times N_u}; \qquad C_s \in \mathbb{R}^{1 \times N_z} \end{aligned}$$

14.B.3 Optimal Control Problem: Path Planning for Optimal State Estimation

Let the state estimate and estimation error be denoted by $\hat{\psi}(t)$ and $e(t)$, respectively such that

$$\begin{aligned} \hat{\psi}(t) &:= E\{\psi(t)\} \\ e(t) &:= \psi(t) - \hat{\psi}(t) \end{aligned}$$

Furthermore, let the estimation error covariance be denoted by the matrix $X(t) = X^T(t) \geq 0 \in \mathbb{R}^{N \times N}$ such that

$$E\{e(t)e^T(\tau)\} := X(t)\delta(t-\tau)$$

Let $P = P^T \geq 0 \in \mathbb{R}^{N\times N}$ be some positive semidefinite matrix. Observe that

$$\begin{aligned}
\mathrm{tr}(PX(t)) &= \mathrm{tr}(PE\{e(t)e^T(t)\}) \\
&= E\{\mathrm{tr}(Pe(t)e^T(t))\} \\
&= E\{\mathrm{tr}(e^T(t)Pe(t))\} \quad \text{(circular property of the trace)} \\
&= E\{(e^T(t)Pe(t)\} \qquad (e^T(t)Pe(t) \text{ is a scalar}) \\
&= E\{||e(t)||_P^2\} \qquad (P\text{-Weighed } L_2\text{-Norm})
\end{aligned}$$

This is a Weighed Mean Square Error which will be used as a cost functional to be minimized for optimal estimation.

14.B.4 Without a Terminal Cost

The optimal path $p(t)$ to estimate $\psi(t)$ can be calculated by solving the following optimal control problem

$$\begin{aligned}
&\min_{\{p(t);X(t)\}} \quad \int_0^{t_f} \left(\mathrm{tr}(PX(t)) + \frac{1}{2}z^T(t)Q_s z(t) + \frac{1}{2}u^T(t)R_s u(t)\right) dt \\
&\text{s.t.} \qquad \begin{cases} \dot{X}(t) = AX + XA^T + Q - \dfrac{1}{R}XC(p)^T C(p)X; & X(0) = X_0 \\ \dot{z}(t) = A_s z(t) + B_s u(t); & z(0) = z_0 \\ p(t) = C_s z(t) \end{cases}
\end{aligned} \tag{14.B.3}$$

where $P = P^T \geq 0 \in \mathbb{R}^{N\times N}$, $Q_s = Q_s^T \geq 0 \in \mathbb{R}^{N_z\times N_z}$ and $R_s = R_s^T > 0 \in \mathbb{R}^{N_u\times N_u}$ are penalization terms.

14.B.4.1 Necessary conditions of optimality: Summary

In this section, we state the necessary conditions of optimality leaving the derivations for the subsequent sections.

Necessary Conditions of Optimality: State and Costate Equations

$$\dot{X} = AX + XA^T + Q - \frac{1}{R}XC(p)^TC(p)X; \qquad X(0) = X_0$$

$$\dot{z} = A_s z + B_s u; \qquad z(0) = z_0$$

$$-\dot{\Lambda} = \Big(A - K(X,p)C(p)\Big)^T \Lambda + \Lambda\Big(A - K(X,p)C(p)\Big) + P; \qquad \Lambda(T) = 0$$

$$-\dot{\lambda} = A_s^T\lambda + Q_s z - \frac{1}{R}C_s^T \mathrm{tr}\Big(XW(p)X\Lambda\Big); \qquad \lambda(T) = 0$$

$$u = -R_s^{-1}B_s^T\lambda$$

$$p = C_s z$$

where

$$K(X,p) := XC^T(p)R^{-1}$$

$$W(\bar{p}) := \frac{d}{dp}C^T(\bar{p})C(\bar{p}) + C^T(\bar{p})\frac{d}{dp}C(\bar{p})$$

.

14.B.4.2 Definitions of Riccati operator and time-differentiation operators

First, we define some useful operators. Define the Riccati operator as follows

$$\mathcal{R}(X,p) := AX + XA^T + Q - \frac{1}{R}XC(p)^TC(p)X \tag{14.B.4}$$

Define the affine time-differentiation operator $\mathcal{D}$ that acts on matrix and vector functions of time as follows

$$\mathcal{D} : [\mathcal{D}X](t) = \dot{X}(t); \quad \mathrm{dom}(\mathcal{D}) = \{X; X(t) \in \mathbb{R}^{N\times N} \quad \forall t \in [0,T] \text{ and } X(0) = X_0\}$$

$$\mathcal{D} : [\mathcal{D}z](t) = \dot{z}(t); \quad \mathrm{dom}(\mathcal{D}) = \{z; z(t) \in \mathbb{R}^{N_z} \quad \forall t \in [0,T] \text{ and } z(0) = z_0\}$$

Moreover, define the linear time-differentiation operators $\mathcal{D}_0$ and $\mathcal{D}_T$ that act on matrix and vector functions of time as follows

$$\mathcal{D}_0 : [\mathcal{D}X](t) = \dot{X}(t); \quad \text{dom}(\mathcal{D}_0) = \{X; X(t) \in \mathbb{R}^{N\times N} \quad \forall t \in [0,T] \text{ and } X(0) = 0\}$$

$$\mathcal{D}_0 : [\mathcal{D}z](t) = \dot{z}(t); \quad \text{dom}(\mathcal{D}_0) = \{z; z(t) \in \mathbb{R}^{N_z} \quad \forall t \in [0,T] \text{ and } z(0) = 0\}$$

$$\mathcal{D}_T : [\mathcal{D}\Lambda](t) = \dot{\Lambda}(t); \quad \text{dom}(\mathcal{D}_T) = \{\Lambda; \Lambda(t) \in \mathbb{R}^{N\times N} \quad \forall t \in [0,T] \text{ and } \Lambda(T) = 0\}$$

$$\mathcal{D}_T : [\mathcal{D}\lambda](t) = \dot{\lambda}(t); \quad \text{dom}(\mathcal{D}_T) = \{\lambda; \lambda(t) \in \mathbb{R}^{N_z} \quad \forall t \in [0,T] \text{ and } \lambda(T) = 0\}$$

It is easy to derive the following relationships between the various time-differentiation operators

$$\frac{\partial}{\partial X}\mathcal{D} = \mathcal{D}_0, \qquad \frac{\partial}{\partial z}\mathcal{D} = \mathcal{D}_0 \qquad \text{and} \qquad \mathcal{D}_0^* = -\mathcal{D}_T$$

where * is the adjoint operator.

14.B.4.3 Lagrangian Analysis

Let $\Lambda \in \mathbb{R}^{N\times N}$ and $\lambda \in \mathbb{R}^{N_z}$ be the Lagrange multipliers associated with the optimization problem (14.B.3). Using appropriate inner products, the Lagrangian L can be written as follows

$$L(X,z,\Lambda,\lambda,u) := \langle P, X\rangle + \frac{1}{2}\langle Q_s z, z\rangle + \frac{1}{2}\langle R_s u, u\rangle + \langle \Lambda, \mathcal{R}(X, C_s z) - \mathcal{D}X\rangle + \langle \lambda, A_s z + B_s u - \mathcal{D}z\rangle$$

where the operators $\mathcal{R}$ and $\mathcal{D}$ are defined in section 14.B.4.2, and the inner products are defined as follows

$$\text{If } X_1(t), X_2(t) \in \mathbb{R}^{n\times n} \quad \forall t \in [0,T], \quad \text{then } \langle X_1, X_2\rangle := \int_0^T \text{tr}\left(X_1^T(t) X_2(t)\right) dt$$

$$\text{If } z_1(t), z_2(t) \in \mathbb{R}^{n} \quad \forall t \in [0,T], \quad \text{then } \langle z_1, z_2\rangle := \int_0^T z_1^T(t) z_2(t) dt$$

The necessary conditions of optimality are obtained by setting the Fréchet derivatives to zero as follows

$$\left[\partial_\eta L(\bar{X}, \bar{z}, \bar{\Lambda}, \bar{\lambda}, \bar{u})\right](\tilde{\eta}) = 0 \qquad \text{for } (\eta, \tilde{\eta}) \in \{(X, \tilde{X}); (z, \tilde{z}); (\Lambda, \tilde{\Lambda}); (\lambda, \tilde{\lambda}); (u, \tilde{u})\}$$

Before we start computing the Fréchet derivatives of the Lagrangian, lets calculate the Fréchet derivative of the Riccati operator defined in (14.B.4).

$$\begin{aligned}\left[\partial_X \mathcal{R}(\bar{X},\bar{p})\right](\tilde{X}) &= A\tilde{X} + \tilde{X}A^T - \frac{1}{R}\left(\bar{X}C^T(\bar{p})C(\bar{p})\tilde{X} + \tilde{X}C^T(\bar{p})C(\bar{p})\bar{X}\right)\\ \left[\partial_p \mathcal{R}(\bar{X},\bar{p})\right](\tilde{p}) &= -\frac{1}{R}\bar{X}\left(\frac{d}{dp}C^T(\bar{p})C(\bar{p}) + C^T(\bar{p})\frac{d}{dp}C(\bar{p})\right)\bar{X}\tilde{p}\end{aligned} \tag{14.B.5}$$

where $\frac{d}{dp}C(p)$ is defined as follows:

$$\begin{aligned}C(p) &= \begin{bmatrix} C_1(p) & C_2(p) & \cdots & C_{N_z}(p)\end{bmatrix}\\ \frac{d}{dp}C(p) &= \begin{bmatrix} \frac{d}{dp}C_1(p) & \frac{d}{dp}C_2(p) & \cdots & \frac{d}{dp}C_{N_z}(p)\end{bmatrix}\end{aligned}$$

Now, we are ready to compute the various Fréchet derivatives:

1. Setting $\left[\partial_\Lambda L(\bar{X},\bar{z},\bar{\Lambda},\bar{\lambda},\bar{u})\right](\tilde{\Lambda}) = 0$, we get the Differential Riccati Equation.

$$\dot{\bar{X}}(t) = \mathcal{R}(\bar{X}(t),\bar{p}(t)); \qquad \bar{X}(0) = X_0 \tag{14.B.6}$$

2. Setting $\left[\partial_\lambda L(\bar{X},\bar{z},\bar{\Lambda},\bar{\lambda},\bar{u})\right](\tilde{\lambda}) = 0$, we get the state space equation of the sensor dynamics.

$$\dot{\bar{z}}(t) = A_s\bar{z}(t) + B_s\bar{u}(t); \qquad \bar{z}(0) = z_0 \tag{14.B.7}$$

3. Let's calculate L_X such that $\left[\partial_X L(\bar{X},\bar{z},\bar{\Lambda},\bar{\lambda},\bar{u})\right](\tilde{X}) := \langle L_X, \tilde{X}\rangle$.

$$\begin{aligned}\langle L_X,\tilde{X}\rangle &= \langle P,\tilde{X}\rangle + \langle\bar{\Lambda},\left[\partial_X\mathcal{R}(\bar{X},\bar{p})\right](\tilde{X}) - \mathcal{D}_0\tilde{X}\rangle\\ &= \langle P,\tilde{X}\rangle + \langle\bar{\Lambda}, A\tilde{X} + \tilde{X}A^T - \frac{1}{R}\left(\bar{X}C^T(\bar{p})C(\bar{p})\tilde{X} + \tilde{X}C^T(\bar{p})C(\bar{p})\bar{X}\right) - \mathcal{D}_0\tilde{X}\rangle\\ &= \langle P + A^T\bar{\Lambda} + \bar{\Lambda}A + \mathcal{D}_T\bar{\Lambda}, \tilde{X}\rangle - \frac{1}{R}\langle C^T(\bar{p})C(\bar{p})\bar{X}\bar{\Lambda} + \bar{\Lambda}\bar{X}C^T(\bar{p})C(\bar{p}), \tilde{X}\rangle\\ &= \langle P + A^T\bar{\Lambda} + \bar{\Lambda}A + \mathcal{D}_T\bar{\Lambda} - \frac{1}{R}\left(C^T(\bar{p})C(\bar{p})\bar{X}\bar{\Lambda} + \bar{\Lambda}\bar{X}C^T(\bar{p})C(\bar{p})\right), \tilde{X}\rangle\\ \implies L_X &= P + A^T\bar{\Lambda} + \bar{\Lambda}A + \mathcal{D}_T\bar{\Lambda} - \frac{1}{R}\left(C^T(\bar{p})C(\bar{p})\bar{X}\bar{\Lambda} + \bar{\Lambda}\bar{X}C^T(\bar{p})C(\bar{p})\right)\end{aligned}$$

Setting $L_X = 0$, we get the following costate differential equation:

$$-\dot{\bar{\Lambda}} = \left(A - K(\bar{X},\bar{p})C(\bar{p})\right)^T\bar{\Lambda} + \bar{\Lambda}\left(A - K(\bar{X},\bar{p})C(\bar{p})\right) + P; \qquad \bar{\Lambda}(T) = 0 \tag{14.B.8}$$

where $K(X,p) := XC^T(p)R^{-1}$ is the Kalman gain.

4. Let's calculate L_z such that $\left[\partial_z L(\bar{X},\bar{z},\bar{\Lambda},\bar{\lambda},\bar{u})\right](\tilde{p}) := \langle L_z, \tilde{z}\rangle$. Define $\tilde{p} := C_s z$, and employ the chain rule to calculate the Fréchet derivative of $\mathcal{R}$ with respect to z.

$$\begin{aligned}\langle L_z,\tilde{z}\rangle &= \langle Q_s\bar{z},\tilde{z}\rangle + \langle\bar{\Lambda},\left[\partial_p\mathcal{R}(\bar{X},\bar{p})\right]\Big([\partial_z C_s\bar{z}](\tilde{z})\Big)\rangle + \langle\bar{\lambda},A_s\tilde{z}-\mathcal{D}_0\tilde{z}\rangle\\ &= \langle Q_s\bar{z}+A_s^T\bar{\lambda}+\mathcal{D}_T\bar{\lambda},\tilde{z}\rangle - \langle\bar{\Lambda},\frac{1}{R}\bar{X}\left(\frac{d}{dp}C^T(\bar{p})C(\bar{p})+C^T(\bar{p})\frac{d}{dp}C(\bar{p})\right)\bar{X}C_s\tilde{z}\rangle\end{aligned}$$

Define

$$W(\bar{p}) := \frac{d}{dp}C^T(\bar{p})C(\bar{p}) + C^T(\bar{p})\frac{d}{dp}C(\bar{p})$$

Observe that $W(\bar{p}) = W^T(\bar{p}) \in \mathbb{R}^{N\times N}$. Then

$$\langle L_z,\tilde{z}\rangle = \langle Q_s\bar{z}+A_s^T\bar{\lambda}+\mathcal{D}_T\bar{\lambda},\tilde{z}\rangle - \langle\bar{\Lambda},\frac{1}{R}\bar{X}W(\bar{p})\bar{X}C_s\tilde{z}\rangle$$

Let's calculate the second term while keeping in mind that $C_s\tilde{z}$ is a scaler function of time.

$$\begin{aligned}\langle\bar{\Lambda},\frac{1}{R}\bar{X}W(\bar{p})\bar{X}C_s\tilde{z}\rangle &= \frac{1}{R}\int_0^T \mathrm{tr}\Big(\bar{\Lambda}^T\bar{X}W(\bar{p})\bar{X}C_s\tilde{z}\Big)dt\\ &= \frac{1}{R}\int_0^T \mathrm{tr}\Big(\bar{\Lambda}^T\bar{X}W(\bar{p})\bar{X}\Big)C_s\tilde{z}dt\\ &= \frac{1}{R}\int_0^T \mathrm{tr}\Big(\bar{X}W(\bar{p})\bar{X}\bar{\Lambda}\Big)C_s\tilde{z}dt\\ &= \frac{1}{R}\langle\mathrm{tr}\Big(\bar{X}W(\bar{p})\bar{X}\bar{\Lambda}\Big),C_s\tilde{z}\rangle\\ &= \frac{1}{R}\langle C_s^T\mathrm{tr}\Big(\bar{X}W(\bar{p})\bar{X}\bar{\Lambda}\Big),\tilde{z}\rangle\end{aligned}$$

Hence,

$$\langle L_z,\tilde{z}\rangle = \langle Q_s\bar{z}+A_s^T\bar{\lambda}+\mathcal{D}_T\bar{\lambda}-\frac{1}{R}C_s^T\mathrm{tr}\Big(\bar{X}W(\bar{p})\bar{X}\bar{\Lambda}\Big),\tilde{z}\rangle$$

Setting $L_z = 0$, we get the second costate differential equation:

$$-\dot{\bar{\lambda}} = A_s^T\bar{\lambda}+Q_s\bar{z}-\frac{1}{R}C_s^T\mathrm{tr}\Big(\bar{X}W(\bar{p})\bar{X}\bar{\Lambda}\Big);\qquad \bar{\lambda}(T)=0 \tag{14.B.9}$$

5. Setting $\left[\partial_u L(\bar{X}, \bar{z}, \bar{\Lambda}, \bar{\lambda}, \bar{u})\right](\tilde{u}) = 0$, we get the simple relationship that links $\bar{\lambda}$ with $\bar{u}$.

$$\begin{aligned}\left[\partial_u L(\bar{X}, \bar{z}, \bar{\Lambda}, \bar{\lambda}, \bar{u})\right](\tilde{u}) &= \langle R_s \bar{u}, \tilde{u}\rangle + \langle \bar{\lambda}, B_s \tilde{u}\rangle \\ &= \langle R_s \bar{u} + B_s^T \bar{\lambda}, \tilde{u}\rangle\end{aligned}$$

Therefore,

$$\bar{u} = -R_s^{-1} B_s^T \bar{\lambda} \tag{14.B.10}$$

Equations (14.B.6), (14.B.7), (14.B.8), (14.B.9) and (14.B.10) form the set of necessary conditions of optimality.

Bibliography

[1] J Antoine. Rigged hilbert spaces in quantum field theory: A lesson drawn from charge operators. Technical report, Univ., Louvain, Belg. Univ., Geneva, 1972.

[2] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 16(1):1–3, 1966.

[3] Bassam Bamieh and Maurice Filo. An input-output approach to structured stochastic uncertainty. *Submitted to IEEE Transactions on Automatic Control*, 2018.

[4] Marc A Berger and Victor J Mizel. Theorems of fubini type for iterated stochastic integrals. *Transactions of the American Mathematical Society*, 252:249–274, 1979.

[5] Marc A Berger and Victor J Mizel. Volterra equations with itô integrals I. *The Journal of Integral Equations*, pages 187–245, 1980.

[6] Marc A Berger and Victor J Mizel. Volterra equations with itô integrals II. *The Journal of Integral Equations*, pages 319–337, 1980.

[7] Daniele Bertaccini and Renata Sisto. Fast numerical solution of nonlinear nonlocal cochlear models. *Journal of Computational Physics*, 230(7):2575–2587, 2011.

[8] Dimitri Bertsekas. On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control*, 21(2):174–184, 1976.

[9] Dimitri P Bertsekas. Projected newton methods for optimization problems with simple constraints. *SIAM Journal on control and Optimization*, 20(2):221–246, 1982.

[10] Hans Braun and Axel Hauck. Tomographic reconstruction of vector fields. *Signal Processing, IEEE Transactions on*, 39(2):464–471, 1991.

[11] JT Chen and CS Wu. Alternative derivations for the poisson integral formula. *International Journal of Mathematical Education in Science and Technology*, 37(2):165–185, 2006.

[12] Han-Lim Choi and Jonathan P How. Continuous trajectory planning of mobile sensors for informative forecasting. *Automatica*, 46(8):1266–1275, 2010.

[13] ID Coope. On matrix trace inequalities and related topics for products of hermitian matrices. *Journal of mathematical analysis and applications*, 188(3):999–1001, 1994.

[14] Neda Darivandi, Kirsten Morris, and Amir Khajepour. An algorithm for lq optimal actuator location. *Smart materials and structures*, 22(3):035001, 2013.

[15] Michael A Demetriou. Guidance of mobile actuator-plus-sensor networks for improved control and estimation of distributed parameter systems. *IEEE Transactions on Automatic Control*, 55(7):1570–1584, 2010.

[16] Charles A Desoer and Mathukumalli Vidyasagar. *Feedback systems: input-output properties*, volume 55. Siam, 1975.

[17] A El Bouhtouri and AJ Pritchard. Stability radii of linear systems with respect to stochastic perturbations. *Systems & control letters*, 19(1):29–33, 1992.

[18] Nicola Elia. Remote stabilization over fading channels. *Systems & Control Letters*, 54(3):237–249, 2005.

[19] Stephen J Elliott, Emery M Ku, and Ben Lineton. A state space model for cochlear mechanics. *The Journal of the Acoustical Society of America*, 122(5):2759–2771, 2007.

[20] Stephen J Elliott, Guangjian Ni, Brian R Mace, and Ben Lineton. A wave finite element analysis of the passive cochlea. *The Journal of the Acoustical Society of America*, 133(3):1535–1545, 2013.

[21] Fariba Fahroo and Michael A Demetriou. Optimal actuator/sensor location for active noise regulator and tracking control problems. *Journal of Computational and Applied Mathematics*, 114(1):137–158, 2000.

[22] Maurice Filo, Fadi Karameh, and Mariette Awad. Order reduction and efficient implementation of nonlinear nonlocal cochlear response models. *Biological cybernetics*, 110(6):435–454, 2016.

[23] Maurice G Filo. *Topics in Modeling of Cochlear Dynamics: Computation, Response and Stability Analysis*. PhD thesis, University of California, Santa Barbara, 2017.

[24] Florian Fruth, Frank Jülicher, and Benjamin Lindner. An active oscillator model describes the statistics of spontaneous otoacoustic emissions. *Biophysical journal*, 107(4):815–824, 2014.

[25] C Daniel Geisler and Chunning Sang. A cochlear model using feed-forward outer-hair-cell forces. *Hearing research*, 86(1):132–146, 1995.

[26] Edward Givelberg and Julian Bunn. A comprehensive three-dimensional model of the cochlea. *Journal of Computational Physics*, 191(2):377–391, 2003.

[27] Donald D Greenwood. A cochlear frequency-position function for several species29 years later. *The Journal of the Acoustical Society of America*, 87(6):2592–2605, 1990.

[28] Symeon Grivopoulos. *Optimal control of quantum systems*. University of California, Santa Barbara, 2005.

[29] John Hauser. A projection operator approach to the optimization of trajectory functionals. *IFAC Proceedings Volumes*, 35(1):377–382, 2002.

[30] John Hauser and David G Meyer. The trajectory manifold of a nonlinear control system. In *Decision and Control, 1998. Proceedings of the 37th IEEE Conference on*, volume 1, pages 1034–1039. IEEE, 1998.

[31] Andreas J Häusler, Alessandro Saccon, António Pedro Aguiar, John Hauser, and António M Pascoal. Energy-optimal motion planning for multiple robotic vehicles with collision avoidance. *IEEE Transactions on Control Systems Technology*, 24(3):867–883, 2016.

[32] K Hiramoto, H Doki, and G Obinata. Optimal sensor/actuator placement for active vibration control using explicit solution of algebraic riccati equation. *Journal of Sound and Vibration*, 229(5):1057–1075, 2000.

[33] Roger A Horn and Roy Mathias. An analog of the cauchy–schwarz inequality for hadamard products and unitarily invariant norms. *SIAM Journal on Matrix Analysis and Applications*, 11(4):481–498, 1990.

[34] Ichiro Ito. On the existence and uniqueness of solutions of stochastic integral equations of the volterra type. *Kodai Mathematical Journal*, 2(2):158–170, 1979.

[35] Ivana Jovanovic, Luciano Sbaiz, and Martin Vetterli. Acoustic tomography for scalar and vector fields: theory and application to temperature and wind estimation. *Journal of Atmospheric and Oceanic Technology*, 26(8):1475–1492, 2009.

[36] D Keith Wilson and Dennis W Thomson. Acoustic tomographic monitoring of the atmospheric surface layer. *Journal of Atmospheric and Oceanic Technology*, 11(3):751–769, 1994.

[37] Henry J Kelley. Gradient theory of optimal flight paths. *Ars Journal*, 30(10):947–954, 1960.

[38] Donald E Kirk. *Optimal control theory: an introduction*. Courier Corporation, 2012.

[39] Emery M Ku, Stephen J Elliott, and Ben Lineton. Statistics of instabilities in a state space model of the human cochlea. *The Journal of the Acoustical Society of America*, 124(2):1068–1079, 2008.

[40] CS Kubrusly and H Malebranche. Sensors and controllers location in distributed systems—a survey. *Automatica*, 21(2):117–128, 1985.

[41] M Drew LaMar, Jack Xin, and Yingyong Qi. Signal processing of acoustic signals in the time domain with an active nonlinear nonlocal cochlear model. *Signal processing*, 86(2):360–374, 2006.

[42] Jianbo Lu and Robert E Skelton. Mean-square small gain theorem for stochastic control: discrete-time case. *IEEE Transactions on Automatic Control*, 47(3):490–494, 2002.

[43] Gisiro Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1):48, 1955.

[44] Raman Mehra and R Davis. A generalized gradient method for optimal control problems with inequality constraints and singular arcs. *IEEE Transactions on Automatic Control*, 17(1):69–79, 1972.

[45] BCJ Moore. Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. 1986.

[46] Stephen T Neely. Finite difference solution of a two-dimensional mathematical model of the cochlea. *The Journal of the Acoustical Society of America*, 69(5):1386–1393, 1981.

[47] Stephen T Neely and DO Kim. A model for active elements in cochlear biomechanics. *The Journal of the Acoustical Society of America*, 79(5):1472–1480, 1986.

[48] Stephen J Norton. Tomographic reconstruction of 2-d vector fields: application to flow imaging. *Geophysical Journal International*, 97(1):161–168, 1989.

[49] AL Nuttall, K Grosh, J Zheng, E De Boer, Y Zou, and Tianying Ren. Spontaneous basilar membrane oscillation and otoacoustic emission at 15 khz in a guinea pig. *Journal of the Association for Research in Otolaryngology*, 5(4):337–348, 2004.

[50] Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.

[51] Andy Packard and John Doyle. Structured singular value with repeated scalar blocks. 1988.

[52] Seon Ki Park and Liang Xu. *Data assimilation for atmospheric, oceanic and hydrologic applications*, volume 2. Springer Science & Business Media, 2013.

[53] Pablo A Parrilo and Sven Khatri. On cone-invariant linear matrix inequalities. *IEEE Transactions on Automatic Control*, 45(8):1558–1563, 2000.

[54] James O Pickles. *An introduction to the physiology of hearing*, volume 2. Academic press London, 1988.

[55] Anil V Rao. A survey of numerical methods for optimal control. *Advances in the Astronautical Sciences*, 135(1):497–528, 2009.

[56] Alessandro Saccon, John Hauser, and Alessandro Beghi. Trajectory exploration of a rigid motorcycle model. *IEEE Transactions on Control Systems Technology*, 20(2):424–437, 2012.

[57] Charles R Steele and Larry A Taber. Comparison of wkb and finite difference calculations for a two-dimensional cochlear model. *The Journal of the Acoustical Society of America*, 65(4):1001–1006, 1979.

[58] RL Stratonovich. A new representation for stochastic integrals and equations. *SIAM Journal on Control*, 4(2):362–371, 1966.

[59] Charles Van Loan. Computing integrals involving the matrix exponential. Technical report, Cornell University, 1977.

[60] JL Willems. Mean square stability criteria for stochastic feedback systems. *International Journal of Systems Science*, 4(4):545–564, 1973.

[61] Hong-Kun Xu. Averaged mappings and the gradient-projection algorithm. *Journal of Optimization Theory and Applications*, 150(2):360–378, 2011.

[62] Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

[63] A Ziemann, K Arnold, and A Raabe. Acoustic tomography in the atmospheric surface layer. In *Annales Geophysicae*, volume 17, pages 139–148. Springer, 1998.

Chapter 15

Conclusion & Future Directions

The **first part** of the dissertation presented a framework to track mean-square stability and performance of linear time-invariant systems in feedback with multiplicative stochastic uncertainties. The analysis was carried out using a purely input/output approach which uncovered new tools that are borrowed from stochastic calculus. The main assumption of our analysis is that the multiplicative uncertainties are white (uncorrelated) in time. A future direction in this line of research is to consider colored (temporally correlated) uncertainties as well. Furthermore, although our analysis encompasses infinite dimensional systems, but the number of uncertain parameters considered are finite. Future work would include extending our results to spatially distributed uncertainties that obey certain symmetries (for example spatially invariant or circulant systems).

The **second part** applied the theory developed for structured stochastic uncertainty to stochastic cochlear models. Stochastic disturbances were assumed to infiltrate the cochlea within the cochlear amplifier, and thus mean-square stability and performance analysis was carried out. Future work includes studying the effect of stochastic uncertainties in different structural parameters in the cochlea, such as the fluid density.

The **third part** suggested an alternative derivations for existing numerical methods to solve optimal control problems using a function-space approach. This approach gave rise to geometrical interpretations and insights that lead to the development of two new numerical methods as well. Future work in this line of research would involve adding inequality constraints to the optimal control problems that are considered and following the same function-space approach.

Finally, the **fourth part** lay down a theoretical framework to design optimal trajectories for mobile sensors whose goal is to minimize the estimation error of an unknown field. The mobile sensors are assumed to move in stochastic and distributed environments to collect measurements of an unknown field whose underlying physical laws are available. Although the underlying dynamics are stochastic, we were able to pose the problem as a deterministic optimal control problem whose states are operator valued. Future work would involve using efficient numerical methods to solve such large scale optimal control problems.