MODEL-BASED ROBUST AND STOCHASTIC CONTROL, AND STATISTICAL INFERENCE FOR UNCERTAIN DYNAMICAL SYSTEMS

BY

KWANG KI KIM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Aerospace Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Associate Professor Cedric Langbort, Chair
Professor Richard D. Braatz (MIT), Director of Research
Professor Petros G. Voulgaris
Professor Geir E. Dullerud

Abstract

This thesis develops various methods for the robust and stochastic model-based control of uncertain dynamical systems. Several different types of uncertainties are considered, as well as different mathematical formalisms for quantification of the effects of uncertainties in dynamical systems.

For deterministic uncertain models and robust control, uncertainties are described as sets of unknowns and every element from a set is presumed to be realizable. Stability and performance characteristics and controlled system behaviors are required to be satisfied for any element in the set of uncertain models. This thesis extends and expands robust control theory to tackle control problems for specific classes of structured uncertain linear and nonlinear systems that include cone-invariant systems, descriptor systems, and Wiener systems. The resultant analysis and control methods are proposed in terms of conic programming that includes linear programming and semidefinite programming (SDP).

For stochastic uncertain models and stochastic control, uncertainties are described in terms of probability distribution functions. Stability and performance characteristics and controlled system behaviors are required to be satisfied with a desired probabilistic confidence. This thesis develops analysis and control schemes based on a spectral methods known as generalized polynomial chaos that can be used to approximate the propagation of uncertainties through dynamical systems. The proposed analysis and design methods are shown to be computationally efficient and accurate alternatives to sampling-based methods, especially when the methods are incorporated into model-based real-time control such as model predictive control.

In addition to accounting for uncertainties and disturbances, the occurrence of a system component fault or failure can significantly degrade the ability of the control system to satisfy the desired stability and performance criteria. This thesis presents an application of the robust control formalism to passively ensure the reliability of a closed-loop system. For an active and intelligent control under presumed fault scenarios, this thesis considers Bayesian inference and information theory that are suited for real-time model checking and selection. To maximize the performance of statistical inference decision-making in the presence of stochastic uncertainties, design methods for optimal probing input signals are proposed in terms of the solutions of mathematical programs. Due to excitation nature of probing inputs, the resultant mathematical programs

are nonconvex and a sequential SDP and convex relaxation methods are proposed to cope with such computational challenges. For real-time model checking and selection of complex distributed systems, this thesis develops methods of distributed hypothesis testing that are based on belief propagation and optimization in graphical models. The proposed methods are scalable and guarantee a consensus of distributed statistical inference decision-making.

To my parents, for their endless love and support.

Acknowledgments

My deepest gratitude goes to my advisor, Richard D. Braatz, for being a good natured advisor, available anytime and full of resources, knowledge, and willingness to help. My journey towards this dissertation has not been a straight path, but Braatz has kept his trust in me. This thesis would not have been possible without him and I enjoyed stress-free life last couple years at the University of Illinois and the Massachusetts Institute of Technology.

I am very grateful to the members of my thesis committee – Cedric Langbort, Petros G. Voulgaris, and Geir E. Dullerud. I also would like to thank my former advisors, Sang-Young Park at the Yonsei University and Naira Hovakimyan, who helped me stay motivated and inspired. My special thanks go to Rolf Findeisen, Davide M. Raimondo, Lixian Zhang, Bhushan Gopaluni for helpful discussions and enjoyable conversations on research problems during their visit to Braatz's group at the Massachusetts Institute of Technology. I also thank Braatz's and Hovakimyan's group members, my dear colleagues Takashi Tanaka, Hong Jang, Mo Jiang, and Evgeny Kharisov, and everybody who was important to complete this thesis work.

At a more fundamental level, this thesis has been built on the foundations I received from my family. My father Kim Jin-Bong, my mother Kwon Jung-Ja, and my three sisters, Hye-Jung, Kyeong-Ae, and Sung-Ook, have helped me grow up in an environment of caring and support. They are the reasons for my being.

Kwang-Ki K. Kim

Urbana-Champaign, Illinois

March 2013

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Uncertain Systems

Uncertainties are ubiquitous in mathematical models of complex systems and can be represented as certain classes of perturbations and disturbances. A general description of uncertainty can be categorizes into two classes: the plant perturbations and exogenous inputs. The description of plant perturbations including parametric and nonparametric perturbations is an outcome of modeling or system identification errors reflecting the discrepancy between the mathematical model and the actual system in operation. Typical sources of plant-model mismatches are neglected nonlinearities, unmodelled dynamics, model reduction, system parameter variations due to operation circumstance changes, and physical changes in system components. Uncertain exogenous inputs include disturbances and measurement noise. There are two main models for exogenous inputs: stochastic inputs with joint probability distributions and deterministic inputs from a set. Stochastic modeling of inputs is the traditional method for which the inputs are assumed to be random processes with known or estimated joint probability distributions. However, due to difficulties in fully characterizing the resultant probability distributions of solutions of the stochastic dynamical systems, it is typical to discuss/model only up to second-order statistics. Deterministic modeling of inputs describe the class of input signals as unknown but bounded in some mathematical sense. A typical method for representing the input signals is as being bounded by finite- or infinite-dimensional convex constraints.

An *uncertain system* is a set of mathematical models that is assumed to contain the real system as an element, in which the uncertainties can be represented either deterministically or stochastically. For both types of uncertainties, this thesis develops tools for the analysis and control for uncertain systems, for which the physical characteristics of actual systems are consistently kept in their associated mathematical models.

1.2 Robust and Stochastic Analysis and Control

Robust and stochastic analysis/control are two major directions in assessing and achieving satisfactory stability and performance characteristics in the presence of any conceivable uncertainties. Robust analysis and control deals with deterministic modeling of perturbations and disturbances, whereas stochastic analysis and control use stochastic models of perturbations and disturbances, in the mathematical models of systems.

Robust control has been one of the most active research areas of system and control theory since the 1980s. A conjunction of different disciplines of mathematics such as functional analysis, convex optimization, algebraic and differential geometry, and abstract algebra has led advances in developing theories of robust control as well as their applications to real engineering problems. In modern robust control, common and popular approaches use the theory of topological separation to characterize the well-posedness of the interconnected systems and their robust stability and performance, for which convex analysis is used to develop robustness criteria and convex optimization is applied for assessment of such criteria, equipped with algorithmic numerical computations. Exact computations for robust analysis and control are known to be NP-hard for general systems, whereas the associated conservative convex criteria can be evaluated in polynomial-time, except for some special cases such as copositive programs. This thesis presents some extensions of existing robust analysis and control methods, and applies those extended results to certain structured uncertain linear and nonlinear systems. An overview of the computational complexity of robust analysis and control problems is provided afterward.

Stochastic control is an area of control theory that considers stochastic disturbances and perturbations, as well as measurement noise, in the mathematical models of systems.[1] With regards to stochastic analysis, the major task is to compute the probability distributions of solution trajectories of stochastic dynamical systems for given stochastic disturbances and perturbations. With regards to stochastic control, the control objective is to achieve desired probability distributions of solution trajectories of stochastic dynamical systems by applying allowable control inputs. The main difficulties are that the resultant stochastic analysis and control problems are to find solutions of infinite-dimensional dynamical system equations, in particular, partial differential equations such as Fokker-Plank and Hamilton-Jackob-Bellman equations. These computational difficulties necessitate stochastic approximation methods and this thesis investigates the use of spectral methods as an efficient alternative to naive sampling-based approaches.

On a different requirement of control systems (in both of robust and stochastic control), constraining the controlled system behavior in the presence of uncertainties in the mathematical models is another major challenge. To achieve desired constrained controlled trajectories, this thesis considers the model predictive control technique equipped with the receding horizon scheme.

[1] In the literature, the terminology "stochastic control" has been used to refer to control theory that deals with stochastic uncertainty either in observations of the data or in processes that produce data. Throughout this thesis,"stochastic control" refers to control theory that deals with both stochastic data (signals) and stochastic model uncertainty in the systems of consideration. Alternative terms for such theory that have been used in the literature include *probabilistic robust control* or *stochastic robust control*.

1.3 Uncertainty Quantification

The methodology of *uncertainty quantification* is to characterize the effects of uncertainties on simulation or theoretical models of actual systems. Aforementioned sources of uncertainty include parametric model perturbations, lack of physical fidelity of models, and uncertain circumstances in system operation. The principal objectives of uncertainty quantification and propagation include [167]:

1. *Model Checking*: In model-based analysis and control, models must be validated or invalidated by assessing their consistency with measurements/observations that are available from the actual system. Physical measurements are inherently corrupted by uncertainties (e.g., measurement noise, sensor bias), and understanding the sources of uncertainties and modeling imperfections are indispensable to the application of a robust control and estimation scheme.

2. *Variance Analysis*: The simplest way of quantifying uncertainty propagation is to compute the variance of the system response around its mean value (or expectation). This variance analysis can provide important information and perspectives for robust design and optimization and can be used to characterize the robustness of the prediction, the reachability and controllability of the system, and compute confidence levels of associated predictions and estimations.

3. *Risk Analysis*: Apart from variance analysis, determining probabilities that certain system characteristics exceed critical values or operation safety thresholds has significant importance in risk and reliability assessment.

4. *Uncertainty Management*: In the presence of multiple sources of uncertainty, efficient robust control and estimation requires analyses of their relative impacts on certain system performances and behaviors. Isolating and reducing dominant sources of uncertainty are key steps for robust estimation.

Much research effort has been devoted to developing optimal and scalable uncertainty quantification methods, including polynomial chaos, stochastic response surface, and dynamic sampling methods (e.g., Markov Chain Monte Carlo).

For the purpose of stochastic uncertainty quantification in dynamical uncertain systems, this thesis focuses on the use of polynomial chaos and its generalizations with intrusive projection methods, called Galerkin projections, to identify the associated surrogate models. These methods can be considered as special types of spectral methods to construct finite-dimensional approximations in infinite-dimensional probability measure spaces. Surrogate models are built based on generalized polynomial chaos and used for analyzing and quantifying uncertainty propagation and for developing stochastic control methods.

1.4 Reliable Control

An inevitable consequence of practical operation of control systems is that actuators and sensors can become faulty or fail, which motivates the development of methods to evaluate the reliability of the closed-loop system to such imperfect operations. A feedback-controlled system is said to be *reliable* if it is guaranteed to retain desired closed-loop system properties while tolerating faults or failures of actuators and sensors. Maximizing the reliability of a system concerns minimizing its potential performance degradation while retaining closed-loop stability when a fault or failure occurs in a control and measurement channel.

This thesis provides methodologies for the design of robustly reliable decentralized control systems, for which the structured singular value theory is applied to cope with the concurrent presence of faults, failures, and uncertainties.

1.5 Fault Detection and Diagnosis

The complexity of devices and processes implies that faults are inevitable, and the tight interactions between instrumentation and other components of the overall system can result in cascading effects with significant economic, environmental, and human damages. Large amounts of data are collected in the operation of control systems and the data can be analyzed to determine whether or not a fault has occurred in the system, where a fault is defined as abnormal system behavior whether associated with equipment failure, equipment wear, or extreme process disturbances. This task of determining whether a fault has occurred is called *fault detection*, whereas *fault diagnosis* is the task of determining which fault has occurred. To properly and safely operate the facilities and devices in real-time while preventing any unallowable behaviors of the system, reliable FDD algorithms are needed that monitor the inputs and outputs of the system and determines whether a fault occurs and to point to the location of the fault (aka fault diagnosis). In addition, without an optimal integration between the monitoring and control systems, the response to faults can reduce reliability and profit or can be overly conservative, for example, by initiating an unnecessary automated shutdown of the facility due to false alarm. The design of FDD procedures are challenged by the presence of disturbances, noise, and model uncertainties that can make the symptoms of faults/failures indiscernible.

The most common and popular approaches to FDD problems are to use Bayesian hypothesis tests, for which certain system output data are monitored and collected to determine which hypothesized model is the most probable. Performance of those statistical methods of decision-making relies on quality of monitoring data for given exogenous and control inputs. To maximize performance of statistical FDD methods, this thesis considers the design of probing inputs that excite or perturb the actual system so that the resultant monitoring data are expected to have a larger discrimination between conceived hypotheses.

1.6 Distributed Decision Making

Distributed decision making has become of increasing importance in quantitative and quantitative decision theory and applications for complex large-scale and distributed systems. Control of complex systems is essentially characterized by distributed decision making and neccessitate a share of perspectives and cognitions among distributed decision makers. Major problems in distributed decision making are to coordinate the local decision makers called *agents* and to develop efficient protocols for information exchange and perspective sharing. This thesis particularly uses graphical models that are appropriate for representing cognitive maps of perspective sharing and information/data fusion and can be incorporated with distributed statistical inference problems that are special classes of distributed decision making problems.

This thesis develops methods of distributed Bayesian hypothesis tests based on belief propagation and optimization in graphical models, for which local evidences or measurements are exchanged among distributed agents, denoted as the nodes, via communication links, denoted as the edges, and marginal a posteriori probability distributions called beliefs are computed.

1.7 Chapter-by-Chapter Description

This thesis consists of twelve chapters, and compromises three main parts. Part I (Chapters 2 – 6) presents methods of robust control and absolute stability for structured uncertain systems that are further characterized, compared to traditional uncertain systems, and for which newly developed theories and computational methods are incorporated with extensions of existing analysis and control schemes. Part II (Chapters 7 and 8) discusses the use of spectral methods called *polynomial chaos* for stochastic uncertainty propagation and quantification and their applications to probabilistic analysis and control problems for stochastic uncertain dynamical systems. Part III (Chapters 9 – 11) considers statistical inference problems for fault detection and diagnosis based on Bayesian theory. In particular, optimum active probing input design for maximizing the performance of statistical inference is discussed and methods of distributed Bayesian hypothesis tests based on belief propagation and optimization are developed for graphical models of distributed uncertain systems.

Chapter 2 provides necessary and sufficient conditions for the stability of continuous- and discrete-time general cone-invariant LTI systems. The proposed stability criteria are conic Lyapunov stability conditions that are geometric algebraic conditions for the stability of an equilibrium state and established from using the concepts of dual and polar cones. Namely, the existence of a dual variable in the interior of the dual cone such that the adjoint operator maps the dual variable into the interior of the polar cone is a necessary and sufficient condition for the stability of the cone-invariant LTI system, where the cone can be an arbitrary proper cone in the state space. Another necessary and sufficient condition for stability of such a system is the existence of a quadratic Lyapunov solution for the associated copositive Lyapunov inequality. It is shown that the feasible solutions of the stability conditions with conic inequalities can be used to characterize the

extreme rays of the set of solutions for copositive Lyapunov inequalities.

Chapter 3 studies the impulsive behavior and robust stability and performance of continuous-time uncertain linear descriptor systems, which are described by a combination of differential and algebraic equations. Necessary and sufficient conditions are presented for robust stability and several dissipation performance indices of uncertain linear descriptor systems represented as generalized linear fractional transformations. For developing unified methods of robust analysis of uncertain linear descriptor systems, the full-block S-procedure is employed and extended for such further characterized uncertain systems. The conditions are written as coupling of linear matrix inequalities and equalities whose feasibility can be checked in polynomial-time by using interior-point methods. Unified and generalizable convex conditions are provided for the analysis of robust stability and performance for linear descriptor systems with structured uncertainty. A necessary and sufficient condition for robust impulse-free dynamics for structured uncertain systems is derived from using structured singular value theory and incorporated into associated robust stability conditions.

Chapter 4 presents necessary and sufficient conditions for several forms of controlled system reliability. For comparison purposes, past results on the reliability analysis of controlled systems are reviewed and it is shown that several of the past results are either conservative or have exponential complexity. For systems with real and complex uncertainties, conditions for robust reliable stability and performance are derived in terms of the structured singular values of certain transfer functions. The conditions are necessary and sufficient for the controller to stabilize the closed-loop system while retaining a desirable level of the closed-loop performance in the presence of system component faults and/or failures, as well as modeling uncertainties and external disturbances. The resulting conditions based on the structured singular value are applied to the decentralized control for a high-purity distillation column and singular value decomposition-based optimal control for a parallel reactor with combined precooling.

Chapter 5 considers certain classes of uncertain Wiener models that are interconnected systems of stable linear systems followed by a static nonlinearity. A nonlinear control design procedure is presented that provides robustness to uncertainties while being applicable to systems with unstable zero dynamics, unmeasured states, disturbances, and measurement noise. The design procedure combines nonlinear internal model control with linear matrix inequality feasibility or optimization problems, such that all robust stability and performance criteria are computable in polynomial-time using readily available software. The approach is applied to a case study involving the control of pH in which the Wiener model is identified from experimental data. This pH neutralization case study demonstrates the importance of taking uncertainty into account during the design of controllers for Wiener systems. The approach is generalizable to Hammerstein and sandwich systems, whether well- or poorly-conditioned, and to systems with actuator constraints.

Chapter 6 provides a comprehensive overview of research related to the computational complexity of robustness margin calculations. Followed by the pioneering papers on the structured singular value (SSV), there have been numerous efforts to develop efficient algorithms for computing for purely real, mixed real and complex, and purely complex uncertainties. Results on the NP-hardness of the exact computation of

SSV motivated interest on the computational complexity of and potential conservatism in the approximation of SSV. This chapter collects together many results that are not well known in the literature, including that the cost of SSV calculation scales by the rank of the M matrix, and that in worst case the widely used upper bound for SSV can be arbitrarily far off. The chapter also describes approaches for the extension of past results. The role of probabilistic randomized algorithms is also discussed, including their favorable scaling with problem size. Polynomial chaos expansion-based methods are described as a computationally efficient alternative for sampling-based stochastic robustness analysis and controller synthesis.

Chapter 7 considers the incorporation of generalized polynomial chaos expansions for uncertainty propagation and quantification into robust control design. Generalized polynomial chaos expansions are more computationally efficient than Monte Carlo simulation for quantifying the influence of stochastic parametric uncertainties on the states and outputs. Approximate surrogate models based on generalized polynomial chaos expansions are applied to design optimal controllers by solving stochastic optimizations in which the control laws are suitably parameterized, and the cost functions and probabilistic (chance) constraints are approximated by spectral representations. The approximation error is shown to converge to zero as the number of terms in the generalized polynomial chaos expansions increases. Several proposed approximate stochastic optimization problem formulations are demonstrated for a probabilistic robust optimal IMC control problem.

Chapter 8 considers the model predictive control of dynamic systems subject to stochastic uncertainties due to parametric uncertainties and exogenous disturbance. The effects of uncertainties are quantified using generalized polynomial chaos expansions with an additive Gaussian random process as the exogenous disturbance. With Gaussian approximation of the resulting solution trajectory of a stochastic differential equation using generalized polynomial chaos expansion, convex finite-horizon model predictive control problems are solved that are amenable to online computation of a stochastically robust control policy over the time horizon. Using generalized polynomial chaos expansions combined with convex relaxation methods, the probabilistic constraints are replaced by convex deterministic constraints that approximate the probabilistic violations. This approach to chance-constrained model predictive control provides an explicit way to handle a stochastic system model in the presence of both model uncertainty and exogenous disturbances.

Chapter 9 provides a concise introduction to methods of Bayesian hypothesis testing and generalizes some of existing work in the literature. The main applications of consideration are change detection and fault detection and diagnosis in stochastic dynamical systems, for which statistical inference problems based on Bayesian theory of statistics are formulated as mathematical programs.

Chapter 10 considers optimal/suboptimal *active* input design problems for fault detection and diagnosis (FDD). The design problems are formulated as optimizations in which an optimal sequence of inputs within a prediction horizon is computed for maximizing the statistical discrimination of different models of fault scenarios. The optimality criteria are information theoretic measures of the statistical distance between probability distributions and constraints on the predicted controlled output trajectory are imposed for ensuring operational safety as well as the input constraints that correspond to hardware limitations. Two

different approaches to such constrained optimal input design problems are presented. The first scheme is to compute an optima input sequence for maximizing discrimination between system models of fault scenarios in a statistical sense. Two different measures quantifying the degree of distinguishability between two stochastic LTI system models are considered, and their geometric properties are investigated. The presented constrained open- and closed-loop feedback input design problems are shown to be concave programs and an iteration algorithm to solve these special families of nonlinear programs is presented, in which semidefinite programs are sequentially solved and a local optimum can be achieved. The second scheme is semidefinite programming (SDP) relaxation in which three different measures for the degree of statistical discrimination between two hypothesized stochastic dynamical system models are considered and their mathematical properties that are related to Bayesian hypothesis tests are studied. The resulting input design problems are non-convex and associated convex relaxation methods are proposed that can be solved in polynomial time using interior point methods. Receding horizon method is used to implement the computed inputs for both approaches for constrained optimal probing input design. Numerical simulations with an aircraft model are provided to illustrate and demonstrate the presented methods of optimal input design for FDD.

Chapter 11 develops distributed Bayesian hypothesis tests for fault detection and diagnosis that are based on belief propagation and optimization in graphical models. The main challenges in developing distributed statistical estimation algorithms are (i) difficulties in ensuring convergence and consensus for the solutions of distributed inference problems, (ii) increasing computational costs due to lack of scalability, and (iii) communication constraints for networked multi-agent systems. To cope with those challenges, we consider (i) belief propagation and optimization in graphical models of complex distributed systems, (ii) decomposition methods of optimization for parallel and iterative computations, and (iii) distributed decision making protocols.

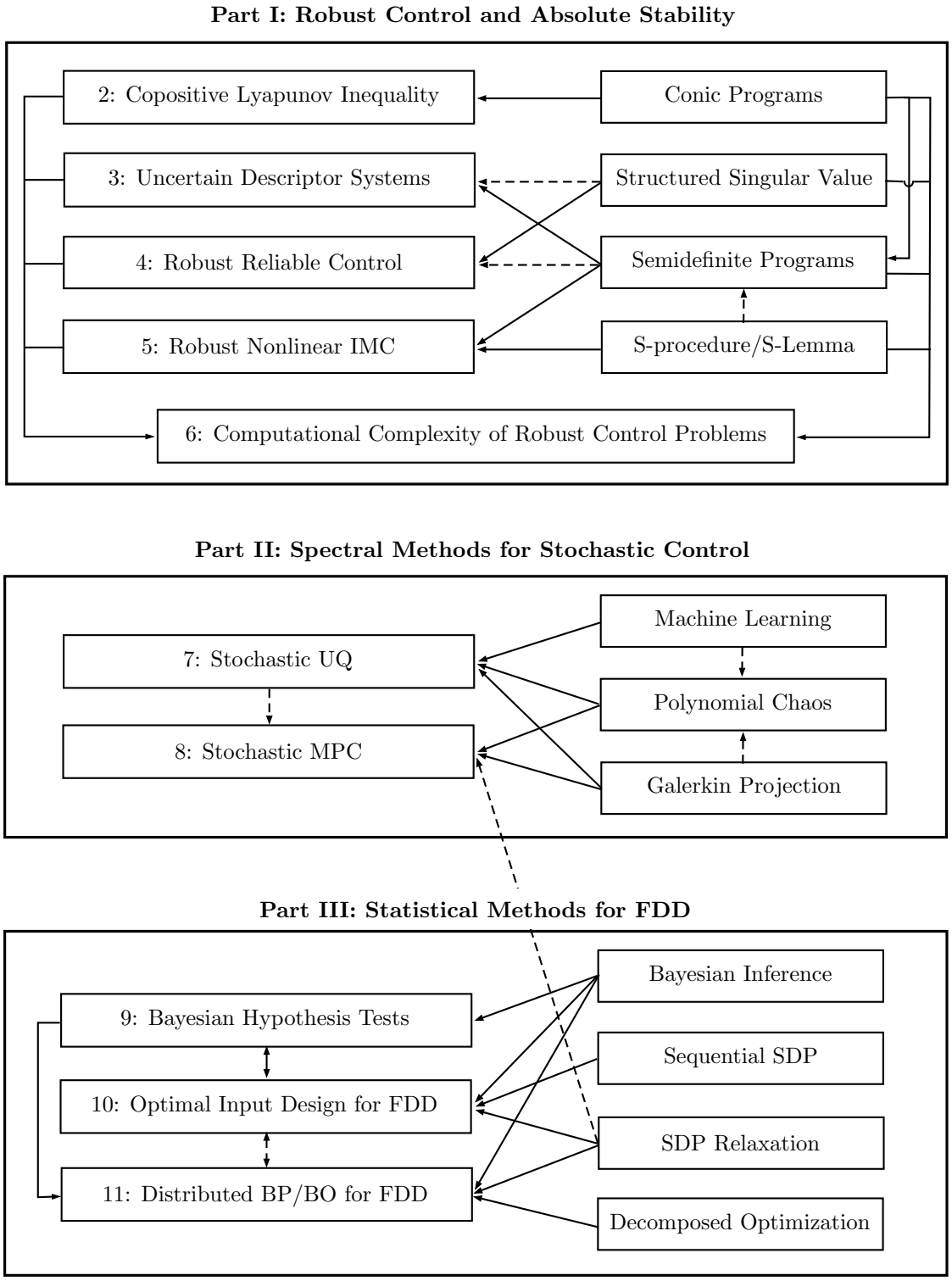


Figure 1.1: The main interactions and flows between the chapters and mathematical methods. A solid arrow connecting two boxes indicates that one chapter or mathematical method depends on the other and a dotted arrow indicates that only the main ideas of the parent chapter or mathematical method are used in the child chapter or mathematical method.

Part I

ROBUST CONTROL AND ABSOLUTE STABILITY

Chapter 2

A Characterization of Solutions for General Copositive Quadratic Lyapunov Inequalities

Abstract This chapter provides an answer to an open question raised in [48] with regard to checking existence of a solution for copositive Lyapunov inequalities. This chapter considers homogeneous LTI systems that preserve a proper cone $\mathcal{C}$.[1] A necessary and sufficient condition for stability of such a system is the existence of a quadratic Lyapunov solution for the associated copositive Lyapunov inequality. This chapter provides another necessary and sufficient condition for stability of the cone-invariant LTI system, in which geometric algebraic conditions for the stability of an equilibrium state are established from the concepts of dual and polar cones. The conditions are polynomial-time verifiable, provided $\mathcal{C}$ is a proper cone in $\mathbb{R}^n$ and has a polynomial-time evaluable self-concordant barrier function. This chapter shows that the feasible solutions of those conditions can be used to characterize the extreme rays of the set of solutions for copositive Lyapunov inequalities.

2.1 Introduction

Mathematical models of dynamical systems in which the state trajectories remain in a cone are prevalent in many systems and control problems [25, 80, 163] and their application to switched positive systems has become a popular research topic [148,170]. For a continuous-time linear time-invariant (LTI) system $\dot{x} = Ax$ with an initial condition $x(0) = x_0$, it is well-known that an equilibrium state is globally asymptotically stable (g.a.s)[2] if and only if there exists a positive definite solution $P \succ 0$ of the Lyapunov inequality $PA + A^{\mathrm{T}}P \prec 0$. This stability condition can be generalized for a cone-invariant LTI system, namely, the

[1] The cones considered in [48] were polyhedral cones, while our consideration is more general, to cover any proper cones $\mathcal{C}$ such that checking the condition $x \in \mathcal{C}$ (or $x \notin \mathcal{C}$) can performed from a polynomial-time evaluable self-concordant barrier function; see [197] for details on polynomial-time interior point methods using self-concordant barrier functions.

[2] The term "stable" means "g.a.s." in many places of this chapter.

state transition map corresponding to the system $\dot{x} = Ax$ ensures that $x(0) \in \mathcal{C}$ implies $x(t) \in \mathcal{C}$ for all $t \geq 0$, where $\mathcal{C} \in \mathbb{R}^n$ is a cone. It is straightforward to show that the aforementioned cone-invariant LTI system is g.a.s. if and only if there exists $P \succ_{\mathcal{C}} 0$ such that $PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0$, where $P \succ_{\mathcal{C}} 0$ is equivalent to $x^{\mathrm{T}}Px > 0$ for all $x \in \mathcal{C} \setminus \{0\}$ and $P \prec_{\mathcal{C}} 0$ is equivalent to $-P \succ_{\mathcal{C}} 0$. This stability condition for the $\mathcal{C}$-invariant LTI system is called the *copositive Lyapunov inequality*.

For the copositive Lyapunov inequality and function of a polyhedral cone-invariant LTI system, an open question was raised in [48]:

Problem 2.1. Consider a matrix $A \in \mathbb{R}^{n\times n}$ and a cone $\mathcal{C} \triangleq \{x \in \mathbb{R}^n : Cx \geq 0\}$. Determine necessary and sufficient conditions for the existence of a symmetric matrix $P \succ_{\mathcal{C}} 0$ such that $PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0$.

This chapter provides an answer to the stability of general cone-invariant LTI systems, which implicitly generalizes the answer to Problem 2.1 for which there exists a solution P satisfying the associated copositive Lyapunov inequality if and only if the corresponding LTI system is stable over the cone $\mathcal{C}$. Namely, we establish necessary and sufficient conditions for stability of continuous- and discrete-time cone-invariant LTI systems. Our conditions are polynomial-time verifiable, provided that there exists a polynomial-time evaluable self-concordant barrier function to check if $x \in \mathcal{C}$. Notice that copositive programs are convex, but NP-hard [189]. This is because the cone $\mathcal{K} \triangleq \{X \in \mathbb{R}^{n\times n} : X = X^{\mathrm{T}}, X \succeq_{\mathcal{C}} 0\}$ nor its dual allow self-concordant barrier functions that can be evaluated in polynomial-time, even for the case when $\mathcal{C}$ is the positive orthant, i.e., $C \equiv \mathrm{I}$ (see [197]). The stability conditions presented in this chapter are to check feasibility of constraints over the cones $\mathcal{C}$ (not $\mathcal{K}$), and its dual and polar cones. There has been some misunderstanding about the computational complexity of checking feasibility of the copositive Lyapunov inequality; for example, in [45] it was misstated that the feasibility problem of the copositive Lyapunov inequality is NP-hard so that there was little hope for there to exist an efficient solution method for large-scale systems.

A main contribution of this chapter is to provide polynomial-time checkable convex conditions for the stability of general cone-invariant LTI systems and correct some misunderstanding of computational complexity of the copositive Lyapunov inequality, for which we fully characterize quadratic Lyapunov solutions for copositive Lyapunov inequalities. A notable discrimination is that, since stability needs to be assessed only over a cone $\mathcal{C}$, it is natural to assume that the system $\dot{x} = Ax$ has a state-transition map that preserves the cone $\mathcal{C}$ and this cone-preserving property characterized in terms of the system matrix A can be exploited to derive polynomial-time solvable mathematical programs. The contribution of this chapter is to characterize solutions of copositive Lyapunov inequalities as a convex hull of rank-one matrices that are dyadic products of two vectors from a cone of the co-state satisfying semi-algebraic conditions with respect to the cone $\mathcal{C}$. We further assume that there exists a polynomial-time verifiable barrier function for the condition $x \in \mathcal{C}$, i.e., checking if $x \in \mathcal{C}$ can be performed in polynomial-time, e.g., polyhedral, second-order (aka Lorentz), and semidefinite cones are such types of proper cones.

2.2 Preliminaries

This section provides a concise overview of the theory of cones and consistency of linear inequalities. This theory is used to establish a main result of this chapter, namely, necessary and sufficient conditions for the cone-invariance and stability of homogeneous LTI systems.

2.2.1 Theory of Cones and Cone-preserving Matrices, and Consistency of Convex Inequalities

There are certain classes of matrices that are related to cone-preserving operators.

Definition 2.1 (*$\mathcal{Z}$-matrix [110]*). The set $\mathcal{Z}_n \subset \mathbb{R}^{n\times n}$ is defined by $\mathcal{Z}_n \triangleq \{A \in \mathbb{R}^{n\times n} : A_{ij} \leq 0 \text{ for } i \neq j\}$.

The set $-\mathcal{Z}_n$ defines the set of *Metzler* matrices, which have non-negative off-diagonal entries.

Definition 2.2 (*$\mathcal{M}$-matrix [110]*). A matrix A is called an *$\mathcal{M}$-matrix* if $A \in \mathcal{Z}_n$ and A is positive stable, i.e., $\sigma(A) \subset \mathbb{C}_+$ where $\mathbb{C}_+$ refers to the open right-half plane in complex variable domain.

Definition 2.3 (*Nonnegative matrix [109]*). A matrix A is called a *nonnegative matrix* if it has nonnegative entries only.

Remark 2.1. The sets of $\mathcal{Z}$-, Metzler, $\mathcal{M}$-, and nonnegative matrices define cones in $\mathbb{R}^{n\times n}$.

Some background on cones is provided below.

Definition 2.4 (*Cone [221]*). A set $\mathcal{C} \in \mathbb{R}^n$ is called a *cone* if it is closed under positive scalar multiplication, i.e., $\lambda x \in \mathcal{C}$ when $x \in \mathcal{C}$ and $\lambda \geq 0$. This set is a union of half-lines emanating from the origin.

A cone $\mathcal{C}$ is *pointed* if $\mathcal{C} \cap (-\mathcal{C}) = \{0\}$ and *solid* if the interior of $\mathcal{C}$ is not empty. A cone that is convex, closed, pointed, and solid is called a *proper* cone.

Definition 2.5 (*Dual and polar cone [161]*). The *dual* of a nonempty set $\mathcal{C} \in \mathbb{R}^n$ is $\mathcal{C}^\bullet \triangleq \{z \in \mathbb{R}^n : \langle z, x\rangle \geq 0 \ \forall x \in \mathcal{C}\}$. Similarly, the *polar* of a nonempty set $\mathcal{C}$ is $\mathcal{C}^\circ \triangleq \{z \in \mathbb{R}^n : \langle z, x\rangle \leq 0 \ \forall x \in \mathcal{C}\}$.

We now come to the basic concepts for the study of topological properties of sets in the vector spaces.

Definition 2.6 (*Interior of a set [161]*). Let $\mathcal{S}$ be a subset of a normed space $\mathcal{X}$. The point $s \in \mathcal{S}$ is said to be an *interior point* of $\mathcal{S}$ if there is an $\epsilon > 0$ such that all vectors $x \in \mathcal{B}(s, \epsilon) \subset \mathcal{S}$. The collection of interior points of $\mathcal{S}$ is called the *interior* of $\mathcal{S}$ and is denoted by $\odot\mathcal{S}$.

Definition 2.7 (*Closure of a set [161]*). Let $\mathcal{S}$ be a subset of a normed space $\mathcal{X}$. The point $s \in \mathcal{S}$ is said to a *closure point* of $\mathcal{S}$ if for a given $\epsilon > 0$ there exists a point $x \in \mathcal{B}(s, \epsilon)$. The collection of all closure points of $\mathcal{S}$ is called the *closure* of $\mathcal{S}$ and is denoted by $\circledast\mathcal{S}$.

The dual $\mathcal{C}^\bullet$ and polar $\mathcal{C}^\circ$ are always closed convex cones. Duality reverses inclusion, $\mathcal{C}_1 \subseteq \mathcal{C}_2 \Rightarrow \mathcal{C}_2^\bullet \subseteq \mathcal{C}_1^\bullet$. If $\mathcal{C}$ is a closed convex cone, then $\mathcal{C} = \mathcal{C}^{\bullet\bullet}$. Otherwise, $\mathcal{C}^{\bullet\bullet}$ is the closure of the smallest convex cone that contains $\mathcal{C}$.

One of the most common matrix operators that preserve a cone is the set of nonnegative matrices. Below is a summary of the appropriate definitions and spectral properties of nonnegative matrices.

Theorem 2.1. *(Perron-Frobenius [109])* Let $A \in M_n$ and suppose that A is irreducible and nonnegative. Then

(a) $\rho(A) > 0$,

(b) $\rho(A)$ is an eigenvalue of A,

(c) There is a positive vector x such that $Ax = \rho(A)x$,

(d) $\rho(A)$ is an algebraically (and hence geometrically) simple eigenvalue of A.

Corollary 2.1. *(Perron-Frobenius [109])* Suppose that $A \in M_n(\mathbb{R})$ is a Metzler matrix. Then

(a) $\rho(A)$ is an eigenvalue of A and there exists a nonnegative eigenvalue $x \geq 0$, $x \neq 0$ such that $Ax = \rho(A)x$,

(b) If $\lambda \in \sigma(A)$ and $|\lambda| = \rho(A)$ then the algebraic multiplicity of λ is not greater than the algebraic multiplicity of the eigenvalue of $\rho(A)$.

Krein and Rutman [150] generalized the results of the Perron-Frobenius theorem. Here, we only consider the linear operators on a finite-dimensional real vector space, although similar results can be extended to infinite-dimensional Banach spaces.

Theorem 2.2. *(Krein-Rutman Theorem for Linear Operators on Finite-dimensional Spaces [150])* Let $\mathcal{C}$ be a proper cone in a finite-dimensional real vector space $\mathcal{V}$, and $\mathcal{L}$ be a linear operator on $\mathcal{V}$ for which $\mathcal{L}(\mathcal{C}) \subseteq \mathcal{C}$. Then the spectral radius $\rho(\mathcal{L})$ is an eigenvalue and there exists a non-zero $x \in \mathcal{C}$ such that $\mathcal{L}(x) = \rho(\mathcal{L})x$. Furthermore, there exists a non-zero $y \in \mathcal{C}^\bullet$ such that $\mathcal{L}^*(y) = \rho(\mathcal{L})y$.

For convex functions defined on convex sets with non-empty interiors, there is a fundamental existence theorem represented as the form of two mutually exclusive alternatives for the associated convex level sets. Below are the results for a special case of affine functions and inequalities.

Theorem 2.3. *(Theorem of Alternatives [221])* Assume that the system $\langle a_i, x\rangle \leq \alpha_i$ for $i = 1, \ldots, m$ is consistent. An inequality $\langle a_0, x\rangle \leq \alpha_0$ is then a consequence of this system if and only if there exist non-negative real numbers $\lambda_i \geq 0$, $i = 1, \ldots, m$, such that $\sum_{i=1}^m \lambda_i a_i = a_0$ and $\sum_{i=1}^m \lambda_i \alpha_i \leq \alpha_0$.

For linear functions and inequalities, the result is known as the Farkas Lemma.

Lemma 2.1. *(Farkas Lemma [221])* An inequality $\langle a_0, x\rangle \leq 0$ is a consequence of the system $\langle a_i, x\rangle \leq 0$ for $i = 1, \ldots, m$ if and only if there exist non-negative real numbers $\lambda_i \geq 0$, $i = 1, \ldots, m$, such that $\sum_{i=1}^m \lambda_i a_i = a_0$.

2.2.2 Positive Invariance of LTI Systems

Some results related to cone invariant LTI systems are now presented.

Lemma 2.2. Consider the system $\dot{x}(t) = Ax(t)$ and a proper cone $\mathcal{C}$. Then $\mathcal{C}$ is a positively invariant set of the solution $x(t)$ if and only if $\Phi(t, t_0) \triangleq \mathrm{e}^{A(t-t_0)}$ has property of leaving positively invariant the proper cone, i.e., $\Phi(t, t_0)\mathcal{C} \subseteq \mathcal{C}$ for all $t \geq t_0$.

Corollary 2.2. Consider the inhomogeneous system $\dot{x}(t) = Ax(t) + v(t)$ and a proper cone $\mathcal{C}$. Then $\mathcal{C}$ is a positively invariant set of the solution $x(t)$ if and only if $\Phi(t, t_0)\mathcal{C} \subseteq \mathcal{C}$ for all $t \geq t_0$.

A special and common case of cone invariant LTI systems is a polyhedral cone that can be generated by a finite number of real vectors from $\mathbb{R}^n$.

Theorem 2.4. *(Polyhedron Invariant Continuous-time LTI Homogeneous System [234, 260, 261])* Consider the system $\dot{x}(t) = Ax(t)$ and a polyhedral cone $\mathcal{C}_{\mathrm{p}}(R) \triangleq \{x \in \mathbb{R}^n : Rx \geq 0\}$ where $R \in \mathbb{R}^{m\times n}$. Then $\mathcal{C}_{\mathrm{p}}(R)$ is a positively invariant set of the solution $x(t)$ if and only if there exists a $\mathcal{Z}$-matrix $Z \in \mathbb{R}^{m\times m}$ such that

$$AR + RZ = 0. \tag{2.1}$$

The above result includes the positive system that preserves the polyhedral cone $\mathbb{R}^n_+$.

Corollary 2.3. Consider the system $\dot{x}(t) = Ax(t)$ and the closed positive orthant $\circledast\mathbb{R}^n_+$. Then $\circledast\mathbb{R}^n_+$ is a positively invariant set of the solution $x(t)$ if and only if A is a Metzler matrix.

Similar results can be extended to discrete-time LTI systems, which directly follow from conditions for consistency of systems of linear inequalities.

Theorem 2.5. *(Polyhedron Invariant Discrete-time LTI Homogeneous System)* Consider the system $x_{k+1} = Ax_k$ and a polyhedral cone $\mathcal{C}_{\mathrm{p}}(R) \triangleq \{x \in \mathbb{R}^n : Rx \geq 0\}$ where $R \in \mathbb{R}^{m\times n}$. Then $\mathcal{C}_{\mathrm{p}}(R)$ is a positively invariant set of the solution x_k if and only if there exists a non-negative matrix $P \in \mathbb{R}^{m\times m}$ such that

$$RA = PR. \tag{2.2}$$

Proof. The Farkas lemma indicates that $Rx \geq 0$ implies $RAx \geq 0$ if and only if for each $i = 1, \ldots, m$, there exists $\lambda_i \in \mathbb{R}^m_+$ such that $R^{\mathrm{T}}\lambda_i = (RA)^{\mathrm{T}}_i$ where $(RA)_i$ refers to the ith row of the matrix RA. This equivalent condition for the dual variables λ_i, $i = 1, \ldots, m$, can be rewritten as the matrix form $R^{\mathrm{T}}\Lambda = (RA)^{\mathrm{T}}$ where the ith column of Λ is λ_i. Selecting $P = \Lambda^{\mathrm{T}}$ proves the result. QED

Corollary 2.4. Consider the system $x_{k+1} = Ax_k$ and the closed positive orthant $\circledast\mathbb{R}^n_+$. Then $\circledast\mathbb{R}^n_+$ is a positively invariant set of the solution x_k if and only if A is a non-negative matrix.

2.3 Necessary and Sufficient Conditions for Stability of Cone invariant LTI systems

Here, we show that the stability of a cone invariant LTI system can be tested by the existence of a dual variable in the interior of the associated dual cone that is mapped by the adjoint operator into the interior of the associated polar cone.

2.3.1 Conic Lyapunov Stability: Linear Functional

For the stability analysis of cone invariant LTI systems, we propose to use a conic linear Lyapunov functional as an alternative to copositive quadratic Lyapunov functions,[3] without introducing any conservatism while requiring less computational demand as the associated problems are convex, for which polynomial-time self-concordant barrier functions exist [197].

2.3.1.1 Continuous-time LTI Homogeneous Systems

The next theorem uses the concept of dual cones to establish a necessary and sufficient condition for the stability of cone invariant LTI systems. We also provide an explicit form of dual variables that define a conic linear Lyapunov functional proving stability.

Theorem 2.6. *(A Necessary and Sufficient Condition for Stability of Cone Invariant LTI Homogeneous Systems)* Consider the system $\dot{x}(t) = Ax(t)$. Suppose that this linear homogeneous system is $\mathcal{C}$-invariant, i.e, $x(t) \in \mathcal{C}$ implies $x(s) \in \mathcal{C}$ for all $s \geq t$, where $\mathcal{C}$ is a proper cone. Then the system is asymptotically stable on $\mathcal{C}$ if and only if there exists $p \in \odot\mathcal{C}^\bullet$ such that $A^{\mathrm{T}}p \in \odot\mathcal{C}^\circ$ where $\mathcal{C}^\bullet$ and $\mathcal{C}^\circ$ refer to the dual and polar cones of $\mathcal{C}$, respectively, and $\odot\mathcal{C}$ denotes the interior of $\mathcal{C}$, i.e., $\odot\mathcal{C} \triangleq \mathcal{C} \setminus \partial\mathcal{C}$.

Proof. ($\Rightarrow$): Consider the linear Lyapunov functional $V(x;p) \triangleq \langle p, x\rangle$ where $p \in \odot\mathcal{C}^\bullet$ such that $V(x;p) \geq 0$ for all $x \in \mathcal{C}$ and $V(x;p) = 0$ only for $x = 0$. The time derivative of this linear Lyapunov functional is

$$\begin{aligned} \frac{d}{dt}V(x;p) &= \frac{d}{dt}\langle p, x\rangle, \\ &= \langle p, Ax\rangle, \\ &= \langle A^{\mathrm{T}}p, x\rangle, \\ &\leq 0 \ (\because A^{\mathrm{T}}p \in \odot\mathcal{C}^\circ \subseteq \mathcal{C}^\circ), \end{aligned} \tag{2.3}$$

where the equality holds if and only if $x = 0$. From Lyapunov theory, this inequality implies that the origin

[3] Some open problems on copositive Lyapunov functions were posed in [48]. One open problem was the derivation of necessary and sufficient conditions for the existence of a Lyapunov function for a linear system that preserves a cone. There have been some efforts to answer this question afterwards (e.g., [45]) and some new results on computing copositive Lyapunov functions based on the conditions presented in this section will be provided in Section 2.4.

is asymptotically stable. ($\Leftarrow$): Suppose that A is Hurwitz stable. Define the vector

$$p := \int_0^\infty \mathrm{e}^{A^{\mathrm{T}}t} q dt$$

for some $q \in \odot\mathcal{C}^\bullet$. Then $\mathrm{e}^{At}\mathcal{C} \subseteq \mathcal{C}$ implies that $\mathrm{e}^{A^{\mathrm{T}}t}\mathcal{C}^\bullet \subseteq \mathcal{C}^\bullet$ for all $t \geq 0$, which also implies that $p \in \odot\mathcal{C}^\bullet$ since $q \in \odot\mathcal{C}^\bullet$. The expression

$$A^{\mathrm{T}}p = \int_0^\infty A^{\mathrm{T}}\mathrm{e}^{A^{\mathrm{T}}t} q dt = -q \in \odot\mathcal{C}^\circ, \tag{2.4}$$

shows that asymptotic stability on $\mathcal{C}$ ensures the existence of a well-defined vector $p \in \mathcal{C}^\bullet$ satisfying $A^{\mathrm{T}}p \in \odot\mathcal{C}^\circ$. QED

A special case of the previous results is a condition for the stability of polyhedral cone invariant LTI systems.

Lemma 2.3. *(A Necessary and Sufficient Stability Condition for Polyhedron Invariant Linear Homogeneous Systems)* Consider the dynamical system $\dot{x}(t) = Ax(t)$ and the polyhedron $\mathcal{C}_{\mathrm{p}}(R) \triangleq \{x \in \mathbb{R}^n : Rx \geq 0\}$. The system is $\mathcal{C}_{\mathrm{p}}(R)$-invariant and the origin is asymptotically stable on $\mathcal{C}_{\mathrm{p}}(R)$ if and only if there exist a $\mathcal{Z}$-matrix $Z \in \mathbb{R}^{m\times m}$ and a vector $p \in \mathcal{C}_{\mathrm{p}}(R)^\bullet$ such that

$$AR + RZ = 0 \quad \text{and} \quad (AG)^{\mathrm{T}}p < 0, \tag{2.5}$$

where $G \in \mathbb{R}^{n\times m}$ satisfies $RG = \mathrm{I}$, i.e., the columns of G define the extreme rays of the cone $\mathcal{C}_{\mathrm{p}}(R)$.

Proof. The set equivalences

$$\begin{aligned}
\odot\mathcal{C}_{\mathrm{p}}(R)^\circ &= \{x \in \mathbb{R}^n : \langle x, y\rangle < 0,\ \forall y \in \mathcal{C}_{\mathrm{p}}(R)\},\\
&= \{x \in \mathbb{R}^n : \langle x, G\lambda\rangle < 0,\ \forall \lambda \in \mathbb{R}^m_+\},\\
&= \{x \in \mathbb{R}^n : \langle G^{\mathrm{T}}x, \lambda\rangle < 0,\ \forall \lambda \in \mathbb{R}^m_+\},\\
&= \{x \in \mathbb{R}^n : G^{\mathrm{T}}x < 0\}
\end{aligned} \tag{2.6}$$

imply that the condition $A^{\mathrm{T}}p \in \odot\mathcal{C}_{\mathrm{p}}(R)^\circ$ is equivalent to the inequality $G^{\mathrm{T}}A^{\mathrm{T}}p < 0$. QED

A special case of the aforementioned results is for linear positive LTI systems.

Corollary 2.5 (*See [163]*). Consider the system $\dot{x}(t) = Ax(t)$ and the closed positive orthant $\circledast\mathbb{R}^n_+$. Then $\circledast\mathbb{R}^n_+$ is a positively invariant set of the solution $x(t)$ and the origin is asymptotically stable on $\circledast\mathbb{R}^n_+$ if and only if A is a Metzler matrix and there exists a vector $p > 0$ such that $A^{\mathrm{T}}p < 0$.

Remark 2.2. Any simplicial cones can be linear transformed into the positive orthant $\mathbb{R}^n_+$ by an invertible matrix in $\mathbb{R}^{n\times n}$, which relates Lemma 2.3 and Corollary 2.5.

2.3.1.2 Discrete-time LTI Homogeneous Systems

Similar results as the continuous-time cases can be obtained for discrete-time cone invariant LTI systems.

Theorem 2.7. *(A Necessary and Sufficient Condition for Stability of Cone Invariant LTI Homogeneous Systems)* Consider the system $x_{k+1} = Ax_k$. Suppose that this linear homogeneous system is $\mathcal{C}$-invariant, i.e, $x_k \in \mathcal{C}$ implies $x_s \in \mathcal{C}$ for all $s \geq t$, where $\mathcal{C}$ is a proper cone. Then the system is asymptotically stable on $\mathcal{C}$ if and only if there exists $p \in \odot\mathcal{C}^\bullet$ such that $(A - \mathrm{I})^{\mathrm{T}} p \in \odot\mathcal{C}^\circ$ where $\mathcal{C}^\bullet$ and $\mathcal{C}^\circ$ refer to the dual and polar cones of $\mathcal{C}$, respectively, and $\odot\mathcal{C}$ denotes the interior of $\mathcal{C}$, i.e., $\odot\mathcal{C} \triangleq \mathcal{C} \setminus \partial\mathcal{C}$.

Proof. ($\Rightarrow$): Consider the linear Lyapunov functional $V(x;p) \triangleq \langle p, x\rangle$ where $p \in \odot\mathcal{C}^\bullet$ such that $V(x;p) \geq 0$ for all $x \in \mathcal{C}$ and $V(x;p) = 0$ only for $x = 0$. The time difference of this linear Lyapunov functional is

$$\begin{aligned} \Delta V(x_k;p) &= V(x_{k+1};p) - V(x_k;p), \\ &= \langle p, (A-\mathrm{I})x_k\rangle, \\ &= \langle (A-\mathrm{I})^{\mathrm{T}} p, x_k\rangle, \\ &\leq 0 \ (\because (A-\mathrm{I})^{\mathrm{T}} p \in \odot\mathcal{C}^\circ \subseteq \mathcal{C}^\circ), \end{aligned} \tag{2.7}$$

where the equality holds if and only if $x = 0$. From Lyapunov theory, this inequality implies that the origin is asymptotically stable.

($\Leftarrow$): Suppose that A is Schur stable. Define the vector

$$p := \sum_{k=0}^{\infty} (A^{\mathrm{T}})^k q$$

for some $q \in \odot\mathcal{C}^\bullet$. Then $A\mathcal{C} \subseteq \mathcal{C}$ implies that $A^{\mathrm{T}}\mathcal{C}^\bullet \subseteq \mathcal{C}^\bullet$, which also implies that $p \in \odot\mathcal{C}^\bullet$ since $q \in \odot\mathcal{C}^\bullet$. The expression

$$(A-\mathrm{I})^{\mathrm{T}} p = (A-\mathrm{I})^{\mathrm{T}} \sum_{k=0}^{\infty} (A^{\mathrm{T}})^k q = -q \in \odot\mathcal{C}^\circ, \tag{2.8}$$

shows that asymptotic stability on $\mathcal{C}$ ensures the existence of a well-defined vector $p \in \mathcal{C}^\bullet$ satisfying $(A - \mathrm{I})^{\mathrm{T}} p \in \odot\mathcal{C}^\circ$. QED

Lemma 2.4. *(A Necessary and Sufficient Stability Condition for Polyhedron Invariant Linear Homogeneous Systems)* Consider the dynamical system $x_{k+1} = Ax_k$ and the polyhedron $\mathcal{C}_{\mathrm{p}}(R) \triangleq \{x \in \mathbb{R}^n : Rx \geq 0\}$. The system is $\mathcal{C}_{\mathrm{p}}(R)$-invariant and the origin is asymptotically stable on $\mathcal{C}_{\mathrm{p}}(R)$ if and only if there exist a non-negative matrix $P \in \mathbb{R}^{m\times m}$ and a vector $p \in \mathcal{C}_{\mathrm{p}}(R)^\bullet$ such that

$$AR = RP \quad \text{and} \quad ((A-\mathrm{I})G)^{\mathrm{T}} p < 0, \tag{2.9}$$

where $G \in \mathbb{R}^{n\times m}$ satisfies $RG = \mathrm{I}$, i.e., the columns of G define the extreme rays of the cone $\mathcal{C}_{\mathrm{p}}(R)$.

Proof. The set equivalences

$$\begin{aligned}\odot\mathcal{C}_{\mathrm{p}}(R)^{\circ} &= \{x \in \mathbb{R}^n : \langle x, y\rangle < 0,\ \forall y \in \mathcal{C}_{\mathrm{p}}(R)\},\\ &= \{x \in \mathbb{R}^n : \langle x, G\lambda\rangle < 0,\ \forall \lambda \in \mathbb{R}^m_+\},\\ &= \{x \in \mathbb{R}^n : \langle G^{\mathrm{T}}x, \lambda\rangle < 0,\ \forall \lambda \in \mathbb{R}^m_+\},\\ &= \{x \in \mathbb{R}^n : G^{\mathrm{T}}x < 0\}\end{aligned} \tag{2.10}$$

imply that the condition $(A-\mathrm{I})^{\mathrm{T}}p \in \odot\mathcal{C}_{\mathrm{p}}(R)^{\circ}$ is equivalent to the inequality $G^{\mathrm{T}}(A-\mathrm{I})^{\mathrm{T}}p < 0$. QED

Stability of the $\mathbb{R}^n_+$-invariant system (see [163]) is a special case of the results of Lemma 2.4.

Corollary 2.6. Consider the system $x_{k+1} = Ax_k$ and the closed positive orthant $\circledast\mathbb{R}^n_+$. Then $\circledast\mathbb{R}^n_+$ is a positively invariant set of the solution x_k and the origin is asymptotically stable on $\circledast\mathbb{R}^n_+$ if and only if A is a non-negative matrix and there exists a vector $p > 0$ such that $(A-\mathrm{I})^{\mathrm{T}}p < 0$.

2.4 Solutions for General Copositive Lyapunov Inequalities

This section provides an answer to Problem 2.1 and fully characterizes the associated quadratic Lyapunov solutions. Consider the copositivity condition

$$P \succ_{\mathcal{C}} 0,\ PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0 \tag{2.11}$$

and suppose that the system matrix A defines the $\mathcal{C}$-invariant system $\dot{x} = Ax$. The associated cone of linear operators (matrices) defined by

$$\mathcal{K}_{\mathrm{lyap}}(A|\mathcal{C}) \triangleq \left\{P \in \mathbb{S}^n : P \succ_{\mathcal{C}} 0,\ PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0\right\} \tag{2.12}$$

is a cone in $\mathbb{R}^{n\times n}$. The next lemma computes the dual cone of $\mathcal{K}_{\mathrm{lyap}}(A|\mathcal{C})$.

Lemma 2.5. The dual cone of $\mathcal{K}_{\mathrm{lyap}}(A|\mathcal{C})$ is the closure of

$$\mathcal{K}_{\mathrm{lyap}}(A^{\mathrm{T}}|\mathcal{C}^{\bullet}) \triangleq \left\{X : \mathbb{S}^n : X \succ_{\mathcal{C}^{\bullet}} 0,\ AX + XA^{\mathrm{T}} \prec_{\mathcal{C}^{\bullet}} 0\right\}, \tag{2.13}$$

i.e., $\mathcal{K}_{\mathrm{lyap}}(A|\mathcal{C})^{\bullet} = \circledast\mathcal{K}_{\mathrm{lyap}}(A^{\mathrm{T}}|\mathcal{C}^{\bullet})$.

Proof. From the definition of dual cone, $X \in \mathcal{K}_{\mathrm{lyap}}(A|\mathcal{C})^{\bullet}$ if and only if $\langle X, P\rangle \geq 0$ for all $P \succ_{\mathcal{C}} 0$ satisfying $PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0$. It is trivial to see that $\langle X, P\rangle \geq 0$ for all $P \succ_{\mathcal{C}} 0$ if and only if $X \succeq_{\mathcal{C}^{\bullet}} 0$. Furthermore, if $X \succeq_{\mathcal{C}^{\bullet}} 0$ then $\langle X, PA + A^{\mathrm{T}}P\rangle \leq 0$ for all P satisfying $PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0$. Since $\langle X, PA + A^{\mathrm{T}}P\rangle =$

$\langle P, AX + XA^{\mathrm{T}}\rangle$ and $P \succ_{\mathcal{C}} 0$, $\langle X, PA + A^{\mathrm{T}}P\rangle \leq 0$ for all P satisfying $PA + A^{\mathrm{T}}P \prec_{\mathcal{C}} 0$ if and only if $AX + XA^{\mathrm{T}} \preceq_{\mathcal{C}^\bullet} 0$. QED

The next lemma considers a generalized cone of $\mathcal{C}$-copositive matrices and computes its dual cone.

Lemma 2.6. Consider the cone of $\mathcal{C}$-copositive matrices defined by $\mathcal{K}_+(\mathcal{C}) \triangleq \{M \in \mathbb{S}^n : M \succeq_{\mathcal{C}} 0\}$. Then

$$\begin{aligned}\mathcal{P}_+(\mathcal{C}) &\triangleq \left\{X \in \mathbb{S}^n : X = \sum_{i=1}^n x_i x_i^{\mathrm{T}},\ x_i \in \mathcal{C},\ \forall i\right\} \\ &= \mathcal{K}_+(\mathcal{C})^\bullet.\end{aligned} \tag{2.14}$$

Proof. Consider $X \in \mathcal{P}_+(\mathcal{C})$. Then

$$\begin{aligned}\langle X, M\rangle &= \left\langle \sum_{i=1}^n x_i x_i^{\mathrm{T}}, M\right\rangle, \\ &= \sum_{i=1}^n \langle x_i x_i^{\mathrm{T}}, M\rangle, \\ &= \sum_{i=1}^n x_i^{\mathrm{T}} M x_i, \\ &\geq 0 \ \ \forall M \in \mathcal{K}_+(\mathcal{C})\ (\because x_i \in \mathcal{C}),\end{aligned} \tag{2.15}$$

which implies $\mathcal{P}_+(\mathcal{C}) \subseteq \mathcal{K}_+(\mathcal{C})^\bullet$. To show the reverse inclusion, suppose that $M \notin \mathcal{K}_+(\mathcal{C})$. Then there exists $x \in \mathcal{C}$ such that $x^{\mathrm{T}}Mx < 0$, but $xx^{\mathrm{T}} \in \mathcal{P}_+(\mathcal{C})$. This implies that if $M \notin \mathcal{K}_+(\mathcal{C})$ then $M \notin \mathcal{P}_+(\mathcal{C})^\bullet$. Thus, we have $\mathcal{P}_+(\mathcal{C})^\bullet \subseteq \mathcal{K}_+(\mathcal{C})$, which implies the relation of inclusion $\mathcal{K}_+(\mathcal{C})^\bullet \subseteq \mathcal{P}_+(\mathcal{C})^{\bullet\bullet}$. Since $\mathcal{P}_+(\mathcal{C})$ is a closed and convex cone, $\mathcal{P}_+(\mathcal{C})^{\bullet\bullet} = \mathcal{P}_+(\mathcal{C})$ so that $\mathcal{K}_+(\mathcal{C})^\bullet \subseteq \mathcal{P}_+(\mathcal{C})$. QED

It is straightforward to see the relations

$$\begin{aligned}\mathcal{K}_+(\mathcal{C})^\bullet &= \mathcal{K}_+(\mathcal{C}^\bullet) \\ \mathcal{K}_+(\mathcal{C})^{\bullet\bullet} &= \left\{X \in \mathbb{S}^n : X = \sum_{i=1}^n x_i x_i^{\mathrm{T}},\ x_i \in \mathcal{C}^\bullet,\ \forall i\right\} \\ &= \mathcal{K}_+(\mathcal{C}),\end{aligned} \tag{2.16}$$

which follows from $\mathcal{K}_+(\mathcal{C})^{\bullet\bullet} = \mathcal{P}_+(\mathcal{C})^\bullet$.

Definition 2.8. An extreme ray of a cone $\mathcal{C}$ is a subset of $\mathcal{C} \cup \{0\}$ of the form $\{\alpha E : \alpha \geq 0\}$ where $0 \neq E \in \mathcal{C}$ is such that $E = E_1 + E_2$, $E_1, E_2 \in \mathcal{C}$ implies $E_i = \alpha_i E$ for some $\alpha_i \geq 0$, $i = 1, 2$.

The next theorem shows that the extreme rays of solutions for the copositive Lyapunov inequalities can be represented as the set of dyadic products of vectors that satisfy the conditions in Theorem 2.6.

Theorem 2.8. Consider the closed convex cone

$$\circledast\mathcal{K}_{\text{lyap}}(A|\mathcal{C}) \triangleq \left\{P \in \mathbb{S}^n : P \succeq_{\mathcal{C}} 0,\ PA + A^{\mathrm{T}}P \preceq_{\mathcal{C}} 0\right\} \tag{2.17}$$

and the cone

$$\mathcal{E}(A|\mathcal{C}) \triangleq \{pp^{\mathrm{T}} \in \mathbb{S}^n : p \in \odot\mathcal{C}^{\bullet},\ A^{\mathrm{T}}p \in \odot\mathcal{C}^{\circ}\}. \tag{2.18}$$

Then $\circledast\mathcal{E}(A|\mathcal{C})$ defines the set of extreme rays of $\circledast\mathcal{K}_{\text{lyap}}$.

Proof. Let $p \in \circledast\mathcal{E}(A|\mathcal{C})$ and be non-zero. Then $pp^{\mathrm{T}} \in \circledast\mathcal{K}_{\text{lyap}}$. To see this, we need to show that (i) $\langle x, pp^{\mathrm{T}}x\rangle \geq 0$ for all x $\in \mathcal{C}$ and (ii) $\langle x, (pp^{\mathrm{T}}A + A^{\mathrm{T}}pp^{\mathrm{T}})x\rangle \leq 0$ for all x $\in \mathcal{C}$. The first condition (i) is trivial. The right-hand side of the inequality in the second condition (ii) can be rewritten as $\langle x, (pp^{\mathrm{T}}A + A^{\mathrm{T}}pp^{\mathrm{T}})x\rangle = \langle x, p\rangle(\langle x, A^{\mathrm{T}}p\rangle + \langle p^{\mathrm{T}}A, x\rangle)$ which is nonpositive for all $x \in \mathcal{C}$. Now, suppose that $pp^{\mathrm{T}} = P_1 + P_2$ where $P_i \in \circledast\mathcal{K}_{\text{lyap}}$ for $i = 1, 2$. Let $w \in \mathcal{C} \subset \mathbb{R}^n$ be orthogonal to p such that $w^{\mathrm{T}}pp^{\mathrm{T}}w = w^{\mathrm{T}}(P_1 + P_2)w = 0$, which implies $w^{\mathrm{T}}P_iw = 0$ for all $i = 1, 2$. Since the space orthogonal to p has $n - 1$ dimension and $\mathcal{C}$ is a solid cone, the matrices P_i are at most of rank 1, i.e., $P_i = \alpha_i pp^{\mathrm{T}}$ for some $\alpha_i \geq 0$. In reverse, suppose that $E = pp^{\mathrm{T}} + P$ is an element of an extreme ray of $\circledast\mathcal{K}_{\text{lyap}}(A|\mathcal{C})$ and P is not aligned with pp^* such that rank$(E) \geq 2$. From $E \succeq_{\mathcal{C}} 0$, P necessarily has the form $P = \sum_{i=1}^{n_p} p_i p_i^{\mathrm{T}}$ for some $p_i \in \mathcal{C}^{\bullet}$ that are not on the same ray as p and $n_p \geq 1$. Then it directly follows that E does not generate an extreme ray of $\circledast\mathcal{K}_{\text{lyap}}$.QED

Lemma 2.7 (*Klee [147]*). Any closed convex set containing no lines can be expressed as the convex hull of its extreme points and extreme rays.

Theorem 2.9. For any $P \in \mathcal{K}_{\text{lyap}}(A|\mathcal{C})$ with a proper cone $\mathcal{C} \in \mathbb{R}^n$, we have a semi-spectral representation $P = \sum_{i=1}^{n_p} p_i p_i$ where $n_p \leq n$ and $p_i \in \mathcal{L}(A|\mathcal{C}) \triangleq \{p \in \mathbb{R}^n : p \in \odot\mathcal{C}^{\bullet}$ and $A^{\mathrm{T}}p \in \odot\mathcal{C}^{\circ}\}$ for all $i = 1, \ldots, n_p$.

Proof. We need to show

$$\mathcal{P}(A|\mathcal{C}) \triangleq \text{Conv}(\mathcal{E}(A|\mathcal{C})) \equiv \mathcal{K}_{\text{lyap}}(A|\mathcal{C}),$$

where $\mathcal{E}(A|\mathcal{C})$ is given in (2.18). Suppose that $P = \sum_{i=1}^{n_p} p_i p_i$ with $p_i \in \mathcal{L}(A|\mathcal{C})$ for all $i = 1, \ldots, n_p$ and $n_p \leq n$. Then it directly follows that $P \in \mathcal{K}_{\text{lyap}}(A|\mathcal{C})$, which implies $\mathcal{P}(A|\mathcal{C}) \subseteq \mathcal{K}_{\text{lyap}}(A|\mathcal{C})$. To show the reverse inclusion, note that $\mathcal{K}_{\text{lyap}}(A|\mathcal{C})$ is pointed such that it does not include a linear subspace. Lemma 2.7 implies that the convex hull of $\circledast\mathcal{E}(A|\mathcal{C})$ including the origin is $\circledast\mathcal{K}_{\text{lyap}}$. Thus, any $P \in \mathcal{K}_{\text{lyap}}$ can be written as $P = \sum_{i=1}^{\infty} z_i z_i^{\mathrm{T}}$ for some $z_i \in \mathcal{P}(A|\mathcal{C}) = \mathcal{E}(A|\mathcal{C})$. But, since $z_i \in \mathbb{R}^n$, there are at most n linearly independent vectors from the set $\mathcal{Z}_{\mathcal{E}}(P) \triangleq \{z_i \in \mathcal{E}(A|\mathcal{C}) : i = 1, \ldots\}$, which implies that P can be rewritten as $P = \sum_{i=1}^{n_p} p_i p_i^{\mathrm{T}}$ where $n_p \leq n$ and $p_i \in \{\alpha z \in \mathbb{R}^n : z \in \mathcal{Z}_{\mathcal{E}}(P), \alpha > 0\}$ are linearly independent for $i = 1, \ldots, n_p$. QED

2.5 Summary and Future Work

A characterization of the solutions for copositive Lyapunov inequalities was presented. We show that the extreme rays of solutions for copositive Lyapunov inequalities are indeed dyadic products of the co-state corresponding to Lagrangian dual variables that satisfy semi-algebraic conditions, which are polynomial-time verifiable under a mild assumption on the cone $\mathcal{C}$. As future research directions, we are interested in applying the presented conditions and characterizations of Lyapunov copositive solutions to verify stability of specific cone-preserving systems such as population models and quantum systems.

Chapter 3

Unified Analysis of Uncertain Linear Descriptor Systems

Abstract This chapter considers the impulsive behavior and robust stability and performance of continuous-time uncertain linear descriptor systems, which are described by a combination of differential and algebraic equations. We present necessary and sufficient conditions for robust stability and several dissipation performance indices of uncertain linear descriptor systems represented as generalized linear fractional transformations (gLFTs). The conditions are written as linear matrix inequalities (LMIs), which are computable in polynomial-time. Unified and generalizable convex conditions are provided for the analysis of robust stability and performance for linear descriptor systems with structured uncertainty. A necessary and sufficient condition for robust impulse-free dynamics for structured uncertain systems is derived from using structured singular value (μ) theory and incorporated into associated robust stability conditions.

3.1 Introduction

Descriptor systems are represented by differential and algebraic equations and are also known as *singular systems*, *implicit systems*, *generalized state-space systems*, *differential-algebraic equations* (DAEs), and *semi-state systems* in the literature. Research interest in analysis and control of descriptor systems has been largely motivated by applications in economic systems [162], large-scale systems [151], power systems, and other areas in engineering [181, 254]. Attempting to replace the algebraic equations by differential equations results in a loss of information, which is the main reason that methods to directly analyze the properties of the DAE system is of interest [187].

Considerable attention has been focused on the stability and performance analysis of descriptor systems in the absence and presence of model uncertainties. Robust stability and performance analysis and the control

of descriptor systems are more complicated than for standard state-space systems, i.e., systems in which the states are described only by differential equations. $\mathcal{H}_\infty$-performance of a linear descriptor system was studied in [172,285] without considering model uncertainty and in [218] with norm-bounded uncertainty that was not fully structured. $\mathcal{H}_2$-optimal and LQ-optimal control algorithms have been derived [23, 258] and passivity and the positive real lemma (and KYP lemma) have been extended to descriptor systems [47,86,95,171,303]. Results in the literature are limited in terms of allowed structure of the uncertainties.

This chapter exploits the structure of model uncertainties by employing the full-block S-procedure [233], which is an analogy of the quadratic separator [122] and integral quadratic constraints (IQCs) [178]. These approaches are well-developed for standard state-space systems, especially for linear parametric-varying (LPV) systems, but such a unified approach has not existed for descriptor systems, to our knowledge. Linear matrix equality (LME) and inequality (LMI) conditions are derived for several robust performance analyses of uncertain descriptor systems in which the plant-model mismatches are structured. System performances are represented as dissipation inequalities with respect to quadratic Lyapunov functions and various forms of quadratic supply rates, which are compatible with the use of the full block S-procedure for the analysis of robust performance. We also present convex optimization problems for which the system trajectory is required to satisfy some constraints such as a desired decay rate, confined into an ellipsoid or a polytope, in the presence of plant-model mismatch.

In addition to stability and performance analysis, many researchers have investigated the controllability and observability [57, 94, 278, 300], and impulse-free conditions [67, 159, 278] for descriptor systems. This chapter presents necessary and sufficient conditions for linear descriptor systems to be impulse-free in terms of the structured singular value (μ) in the presence of model uncertainties. Although the exact computation of μ is NP-hard [44], its upper and lower bounds can be computed in polynomial-time [8, 9, 308]. We also show the incorporation of this method for assessing impulse-freeness of structured uncertain linear systems into robustness analysis.

3.2 Preliminaries

3.2.1 Uncertain Descriptor Systems

This section describes a general representation for uncertain descriptor systems that can be interpreted as a generalized linear parameter-varying (gLPV) system. The representation covers most classes of uncertain descriptor systems including polynomial and rational uncertain descriptor systems.

Consider a continuous-time descriptor system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + B_1 w_u(t) + B_2 w_p(t), \\ z_u(t) &= C_1 x(t) + D_{11} w_u(t) + D_{12} w_p(t), \\ z_p(t) &= C_2 x(t) + D_{21} w_u(t) + D_{22} w_p(t), \end{aligned} \tag{3.1}$$

where $E, A \in \mathbb{R}^{n\times n}$ with $\text{rank}(E) \leq n$, $B_1 \in \mathbb{R}^{n\times m_1}$, $B_2 \in \mathbb{R}^{n\times m_2}$, $C_1 \in \mathbb{R}^{l_1\times n}$, $C_2 \in \mathbb{R}^{l_2\times n}$, and $D_{ij} \in \mathbb{R}^{l_i\times m_j}$ for $i = 1, 2$ and $j = 1, 2$, and the input-output pair $(w_u(t), z_u(t))$ satisfies the geometric implicit relation

$$\begin{bmatrix} w_u(t) \\ z_u(t) \end{bmatrix} \in \mathbf{Ker}(\Delta(t)), \quad \Delta(t) \in \mathbf{\Delta} \ \ \forall t. \tag{3.2}$$

If the uncertainty Δ is a function of the state variables then the system (3.1) is referred to as a *quasi-gLPV*. The relation (3.2) is a general representation of the uncertainty in the system that includes the well-known linear fractional transformation (LFT) [308]. For example, if the pair of signals $(w_u(t), z_u(t))$ satisfies the input-output relation $\Delta_u(t)z_u(t) = w_u(t)$ with a linear time-varying uncertain map $\Delta_u : \mathcal{T} \to \mathbb{R}^{m_1\times l_1}$ then set $\Delta(t) := \begin{bmatrix} \mathrm{I} & -\Delta_u(t) \end{bmatrix}$. For this case, it is not difficult to see that $\mathbf{Ker}(\Delta(t)) = \mathbf{Im}\left(\begin{bmatrix} \Delta_u(t) \\ \mathrm{I} \end{bmatrix}\right)$. To simplify the presentation, the time dependence of uncertainty is not shown explicitly in the remainder of this chapter.

3.2.2 Dissipation Inequalities

Here state-space interpretations of the small-gain and passivity approaches are briefly reviewed, with tests for robust versions of these properties derived later in this chapter. Consider a linear descriptor system given by

$$\Sigma(E, A, B, C, D) : \begin{cases} E\dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}, \tag{3.3}$$

where $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ with $m = p$.

On the space of input-output variables $(u, y) \in \mathcal{U} \times \mathcal{Y} \subseteq \mathbb{R}^m \times \mathbb{R}^p$, a real-valued scalar function $S_r : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$ called the *supply rate* is defined below (see [231, 289] for details).

Definition 3.1 (*Dissipation inequality*). A descriptor system $\Sigma(E, A, B, C, D)$ is said to be *dissipative* with respect to the supply rate S_r if there exists a real-valued nonnegative function $V : \mathcal{X} \to \mathbb{R}_+$, which is called the *storage function*, such that the *dissipation inequality*

$$\int_{t_1}^{t_2} S_r(u(t), y(t))dt \geq V(x(t_2)) - V(x(t_1)), \quad \forall t_2 \geq t_1 \geq 0 \tag{3.4}$$

holds, where $x(\cdot) \in \mathcal{X}$ is a solution and $(u(\cdot), y(\cdot)) \in \mathcal{U} \times \mathcal{Y}$ is the input-output pair of the system equa-

tion (3.3). If the equality holds in (3.4) then the descriptor system $\Sigma(E, A, B, C, D)$ is said to be *lossless* with respect to S_r. For a storage function V that is differentiable with respect to time, the dissipation inequality (3.4) can be rewritten as

$$S_r(u(t), y(t)) \geq \frac{d}{d\tau} V(x(\tau))\Big|_{\tau=t}, \quad \forall t \geq 0. \tag{3.5}$$

Remark 3.1. If the system representation $\Sigma(E, A, B, C, D)$ does not have fixed values for the system matrices (E, A, B, C, D), but has a set-valued uncertainty description $(E(\Delta), A(\Delta), B(\Delta), C(\Delta), D(\Delta))$ for $\Delta \in \mathbf{\Delta}$ where the support $\mathbf{\Delta}$ is compact, then an analogous definition of robust dissipative systems can be defined—conditions for robust dissipative descriptor systems are presented later.

Some popular dissipation properties are defined below (see [231, 289] for details).

Definition 3.2 (*Dissipation inequality*). A descriptor system $\Sigma(E, A, B, C, D)$ is said

- to be *passive* if the dissipation inequality (3.4) is satisfied with $S_r(u, y) = u^{\mathrm{T}} y$;
- to be *strictly input passive* if the dissipation inequality (3.4) is satisfied with $S_r(u, y) = u^{\mathrm{T}} y - \delta \|u\|^2$ for some $\delta > 0$;
- to be *strictly output passive* if the dissipation inequality (3.4) is satisfied with $S_r(u, y) = u^{\mathrm{T}} y - \delta \|y\|^2$ for some $\delta > 0$;
- to have $\mathcal{L}_2$*-gain* $\leq \gamma$ if the dissipation inequality (3.4) is satisfied with $S_r(u, y) = \gamma^2 \|u\|^2 - \|y\|^2$,

for some storage function $V : \mathcal{X} \to \mathbb{R}_+$.

The dissipation properties in Defn. 3.2 have quadratic supply rates that can be rewritten as

$$S_r(-\Pi) \triangleq - \begin{bmatrix} u \\ y \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^{\mathrm{T}} & \Pi_{22} \end{bmatrix} \begin{bmatrix} u \\ y \end{bmatrix} \tag{3.6}$$

with a properly chosen matrix Π, which is called a *supply rate matrix*, where the negative sign is used for notational convenience.

3.3 Generalized Full-Block S-Procedure

Let $\mathcal{S}$ be a subset of $\mathbb{R}^N$ and $V \in \mathbb{R}^{q \times N}$ be a full row-rank matrix with $q \leq N$. Suppose that $\mathbf{\Delta} \subset \mathbb{R}^{p \times q}$ is a compact set of matrices of full row rank such that $\mathbf{Ker}(\Delta) \subset \mathbb{R}^q$ depends continuously on the parameter Δ that varies in the compact path-connected set $\mathbf{\Delta}$. Define the family of subspaces

$$\mathcal{S}(\Delta) \triangleq \mathcal{S} \cap \mathbf{Ker}(\Delta V) \ \text{ for } \Delta \in \mathbf{\Delta}. \tag{3.7}$$

This set can be equivalently rewritten as $\mathcal{S}(\Delta) = \{\xi \in \mathcal{S} : V\xi \in \mathbf{Ker}(\Delta)\}$.

Theorem 3.1 (*Full block S-procedure [233]*). (A) The condition

$$P \prec 0 \quad \text{on } \mathcal{S}(\Delta),\ \forall \Delta \in \mathbf{\Delta} \tag{3.8}$$

holds if and only if there exists a symmetric multiplier Y such that

$$P + V^{\mathrm{T}} Y V \prec 0 \text{ on } \mathcal{S} \tag{3.9}$$

and

$$Y \succ 0 \text{ on } \mathbf{Ker}(\Delta),\ \forall \Delta \in \mathbf{\Delta}. \tag{3.10}$$

(B) Suppose that there exists a subspace $\mathcal{S}_0 \subset \mathcal{S}$ on which the matrix P is positive semidefinite, and whose dimension is large enough to satisfy $\dim(V\mathcal{S}_0) + \dim(\mathbf{Ker}(\Delta)) \geq q$. Then (3.9) and (3.10) imply that

$$V\mathcal{S}_o \oplus \mathbf{Ker}(\Delta) = \mathbb{R}^q,\ \forall \Delta \in \mathbf{\Delta}. \tag{3.11}$$

The proof of Thm. 3.1 is in [233]. The concatenation of the variables defining the system lie within the family of subspaces $\mathcal{S}(\Delta)$. Any system performance that is defined by a quadratic inequality with respect to the system variables can be represented as the matrix inequality (3.8), where the matrix P depends on a Lyapunov matrix X and a supply rate matrix Π. With $\mathcal{S}(\Delta)$ dependent of the uncertain mapping $\Delta \in \mathbf{\Delta}$, the feasibility of the matrix inequality (3.8) guarantees robust performance of the system. However, the matrix inequality (3.8) is not a testable condition, since $\mathcal{S}(\Delta)$ is infinite dimensional and depends on the uncertain mapping Δ. This motivates the introduction of the equivalent condition (3.9) with the matrix multiplier Y satisfying the inequality (3.10). The multiplier Y characterizes the uncertain mapping $\Delta \in \mathbf{\Delta}$ with which the structure and knowledge of the uncertainty can be exploited. The 'if'part of the proof may be trivial to see by some readers as it is just a Lagrangian relaxation [222]. The 'only if'part of the proof is not trivial and is lengthy. Note that the sets of feasible solutions for the constraints (3.8) and (3.9)–(3.10) are the same, i.e.,

$$\bigcup_{Y \in \mathcal{Y}_{\mathbf{\Delta}}} \left\{X \in \mathbb{R}^{n\times n} : P(X) + V^{\mathrm{T}} Y V \prec 0 \text{ on } \mathcal{S}\right\} = \left\{X \in \mathbb{R}^{n\times n} : P(X) \prec 0 \text{ on } \mathcal{S}(\Delta),\ \forall \Delta \in \mathbf{\Delta}\right\},$$

where

$$\mathcal{Y}_{\mathbf{\Delta}} \triangleq \left\{Y = Y^{\mathrm{T}} : Y \succ 0 \text{ on } \mathbf{Ker}(\Delta),\ \forall \Delta \in \mathbf{\Delta}\right\}. \tag{3.12}$$

This implies the following extension.

Corollary 3.1. The feasible solution set

$$\mathcal{X} \triangleq \left\{X \in \mathbb{R}^{n\times n} : U(X) \succeq 0\right\} \bigcap \left\{X \in \mathbb{R}^{n\times n} : P(X) \prec 0 \text{ on } \mathcal{S}(\Delta),\, \forall \Delta \in \mathbf{\Delta}\right\}$$

is non-empty if and only if the set

$$\left\{X \in \mathbb{R}^{n\times n} : U(X) \succeq 0\right\} \bigcap \left(\bigcup_{Y \in \mathcal{Y}_{\mathbf{\Delta}}} \left\{X \in \mathbb{R}^{n\times n} : P(X) + V^{\mathrm{T}} Y V \prec 0 \text{ on } \mathcal{S}\right\} \right)$$

is non-empty.

Corollary 3.1 is used later to derive robust stability and performance conditions for uncertain descriptor systems.

3.4 Robust Stability and Performance of Uncertain Descriptor Systems

3.4.1 Nominal Stability and Constrained State Properties of Linear Descriptor Systems

Consider the linear homogeneous descriptor system

$$E\dot{x}(t) = Ax(t). \tag{3.13}$$

For convenience, assume that $\mathcal{E}_{\mathrm{eq}}(E, A) = \{0\}$ and an equilibrium point at the origin could be a translation of a nonzero equilibrium point or a translation of a nonzero solution of the system [136]. The equilibrium point $x = 0$ of the system $\Sigma(E, A)$ is stable if and only if $\sigma(E, A) \in \mathbb{C}_-$, but the computation of the generalized eigenvalues of a matrix pencil (E, A) is known to be unreliable due to its ill-conditioning, i.e., a small perturbation in E or A can result in a large change in the generalized eigenvalues of the matrix pencil (E, A). An alternative way of checking stability is the use of Lyapunov theory [136] which gives computationally reliable tests for stability.

Theorem 3.2 (*[172, 218]*). Suppose that $\Sigma(E, A)$ is solvable. Then the descriptor system $\Sigma(E, A)$ is stable and impulse-free if and only if there exists a solution $X \in \mathbb{R}^{n\times n}$ to the generalized Lyapunov inequalities

$$E^{\mathrm{T}} X = X^{\mathrm{T}} E \succeq 0,\ A^{\mathrm{T}} X + X^{\mathrm{T}} A \prec 0. \tag{3.14}$$

The above theorem does not state the rank condition $\mathrm{rank}(E^{\mathrm{T}} X) = \mathrm{rank}(E)$, which is automatically satisfied by the feasibility of the condition $A^{\mathrm{T}} X + X^{\mathrm{T}} A \prec 0$. In particular, if X satisfies $A^{\mathrm{T}} X + X^{\mathrm{T}} A \prec 0$ then X must be nonsingular so that $\mathrm{rank}(E^{\mathrm{T}} X) = \mathrm{rank}(E)$ holds. The condition (3.14) can be interpreted in terms of the variables that define the system dynamics (3.13). Suppose that $\Sigma(E, A)$ be solvable. The

system in (3.13) is stable if and only if the solution x and its time-derivative $\dot{x}$ satisfies the inequality

$$\begin{bmatrix} x \\ E\dot{x} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} \\ X & 0 \end{bmatrix} \begin{bmatrix} x \\ E\dot{x} \end{bmatrix} < 0, \tag{3.15}$$

where X is a feasible solution to (3.14).

Lyapunov Exponent The Lyapunov exponent, also called the *decay rate*, of the system (3.13) is defined as the largest value α such that $\lim_{t\to\infty} e^{\alpha t}\|x(t)\| = 0$ holds for all solution trajectories, where $\|\cdot\|$ can be any vector norm. Exponential stability is equivalent to the condition $\alpha > 0$ and the condition $\dot{V}(x) \le -2d_r V(x)$ for all x on a differentiable Lyapunov function $V(x)$ can be used to establish a lower bound on the decay rate [35, 136]. The next lemma shows how to extend the concept of Lyapunov exponent to a homogeneous linear descriptor system (3.13).

Lemma 3.1. The solvable system (3.13) is globally exponentially stable (g.e.s.) with a Lyapunov exponent $d_r > 0$ if there exists $X \in \mathbb{R}^{n\times n}$ such that $X^{\mathrm{T}}E = E^{\mathrm{T}}X \succeq 0$ and

$$\begin{bmatrix} \mathrm{I} \\ A \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2d_r E^{\mathrm{T}}X & X^{\mathrm{T}} \\ X & 0 \end{bmatrix} \begin{bmatrix} \mathrm{I} \\ A \end{bmatrix} \prec 0. \tag{3.16}$$

Proof. Suppose that X is a solution of the matrix equality and inequalities in this lemma. Then the Lyapunov function $V(x) = x^{\mathrm{T}}E^{\mathrm{T}}Xx$ satisfies the relation $\dot{V}(x) < -2d_r V(x)$ for all $x \neq 0 \in \mathbb{R}^n$. QED

Remark 3.2. A positive decay rate is a necessary condition for some strict dissipation inequalities such as strict passivity and guarantees a nonzero robust stability margin for input-to-state stability (ISS) [136].

Remark 3.3. An equivalent definition for the Lyapunov exponent is

$$d_r \triangleq \max_i \lim_{t\to\infty} \frac{1}{t}\|X(t)e_i\| \tag{3.17}$$

where $X(t) \in C(\mathcal{T}, \mathbb{R}^{n\times n})$ is the solution matrix (or the state-transition matrix) and e_i is the ith unit vector.

Invariant Ellipsoids Quadratic stability and region of attraction can be geometrically characterized using the concept of *invariant ellipsoids*. For $Q \succ 0$, an ellipsoid $\mathcal{E}(Q) \triangleq \{\eta \in \mathbb{R}^n : \eta^{\mathrm{T}}Q\eta \le 1\}$ is centered at the origin and said to be *invariant* for a system equation, for example, (3.13), if every solution trajectory x starting from $x(t_0) \in \mathcal{E}(Q)$ remains $x(t) \in \mathcal{E}(Q)$ for all future time $t \ge t_0$. Consider a polytope described by its vertices, $\mathcal{P}(v) \triangleq \mathbf{Co}\{v_1, \cdots, v_p\}$. The ellipsoid $\mathcal{E}(Q) \triangleq \{\eta \in \mathbb{R}^n : \eta^{\mathrm{T}}Q\eta \le 1\}$ contains the polytope $\mathcal{P}(v)$ if and only if $v_i^{\mathrm{T}}Qv_i \le 1$ for all $i = 1, \ldots, p$ (see [35]).

Remark 3.4. An ellipsoid defined by $\mathcal{E}_c(E^{\mathrm{T}}X) \triangleq \{x \in \mathbb{R}^n : V(x) = x^{\mathrm{T}}E^{\mathrm{T}}Xx \leq c\}$ with $c > 0$ has nonzero co-dimension for a singular matrix E satisfying $\mathrm{rank}(E) = r < n$ and $E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0$ for some $X \in \mathbb{R}^{n\times n}$. Indeed, $\mathrm{codim}(\mathcal{E}_c(E^{\mathrm{T}}X)) = n - r$ for all $c > 0$ and nonsingular X.

Below is a characterization of the smallest invariant ellipsoid that contains the polytope $\mathcal{P}(v)$.

Proposition 3.1. Suppose that the matrix $X^\star$ solves the optimization

$$\begin{aligned} &\min_{X,\xi_i} \ -\log\det R^{\mathrm{T}}E^{\mathrm{T}}XR \\ &\text{s. t. } X^{\mathrm{T}}E = E^{\mathrm{T}}X \succeq 0, \quad X^TA + A^{\mathrm{T}}X \prec 0, \\ &\qquad v_i^{\mathrm{T}}E^{\mathrm{T}}Xv_i \leq 1, \quad v_i = R\xi_i, \quad i =, 1\ldots, p, \end{aligned} \tag{3.18}$$

where the columns of $R \in \mathbb{R}^{n\times r}$ are orthonormal bases of the subspace $[\mathbf{Ker}(E)]^c$ (the superscript c denotes the orthogonal complement of a subspace or subset). Then $\mathcal{E}(E^TX^\star) \triangleq \{\eta \in \mathbb{R}^n : \eta^{\mathrm{T}}E^{\mathrm{T}}X^\star\eta \leq 1\} \subset [\mathbf{Ker}(E)]^c$ is the invariant ellipsoid of smallest volume that contains the polytope $\mathcal{P}(v)$ defined with the vertices v_i, $i = 1, \ldots, p$.

Proof. The first two inequality constraints in (3.18) guarantee that an ellipsoid $\mathcal{E}_c(E^{\mathrm{T}}X)$ with $c > 0$ and a feasible solution X is invariant for the system (3.13). The feasibility of the third and fourth constraints is equivalent to the condition that the ellipsoid $\mathcal{E}(E^{\mathrm{T}}X)$ contains the polytope $\mathcal{P}(v)$. The equality constraints imply that the vertices defining the polytope $\mathcal{P}(v)$ must be in the subspace $[\mathbf{Ker}(E)]^c$ that is spanned by the columns of R. Furthermore, it is not difficult to see that $\mathcal{E}(E^{\mathrm{T}}X) = \{\xi \in \mathbb{R}^r : \xi^{\mathrm{T}}R^{\mathrm{T}}E^{\mathrm{T}}XR\xi \leq 1\}$. Minimizing the volume of the ellipsoid $\mathcal{E}(E^{\mathrm{T}}X)$ containing the polytope $\mathcal{P}(v)$, now, corresponds to the convex optimization problem in (3.18). QED

A similar characterization can be derived for the invariant ellipsoid of largest volume that is contained in the polytope $\mathcal{P}(h) \triangleq \{\eta \in \mathbb{R}^n : h_i^{\mathrm{T}}\eta \leq 1, \ i = 1, \ldots, p\}$, which is a bounded intersection of half spaces. The ellipsoid $\mathcal{E}(Q^{-1})$ is contained in the polytope $\mathcal{P}(h)$ if and only if $h_i^{\mathrm{T}}Qh_i \leq 1$ for all $i = 1, \ldots, p$ (see [35]), which is equivalent to the matrix inequality condition $\begin{bmatrix} 1 & h_i^{\mathrm{T}} \\ h_i & Q^{-1} \end{bmatrix} \succeq 0, \ i = 1, \ldots, p.$

Proposition 3.2. Suppose that the matrix $X^\star$ solves the optimization

$$\begin{aligned} &\min_{X} \ \log\det R^{\mathrm{T}}E^{\mathrm{T}}XR \\ &\text{s. t. } X^{\mathrm{T}}E = E^{\mathrm{T}}X \succeq 0, \quad X^TA + A^{\mathrm{T}}X \prec 0, \\ &\qquad \begin{bmatrix} 1 & h_i^{\mathrm{T}} \\ h_i & E^{\mathrm{T}}X \end{bmatrix} \succeq 0, \quad h_i = R\xi_i, \quad i =, 1\ldots, p. \end{aligned} \tag{3.19}$$

Then $\mathcal{E}(E^TX^\star) \triangleq \{\eta \in \mathbb{R}^n : \eta^{\mathrm{T}}E^{\mathrm{T}}X^\star\eta \leq 1\} \subset [\mathbf{Ker}(E)]^c$ is the largest invariant ellipsoid contained in the polytope $\mathcal{P}(h)$ defined with the vectors h_i, $i = 1, \ldots, p$.

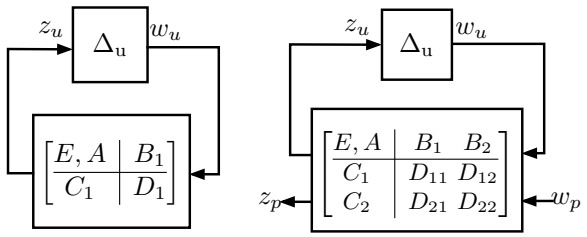


Figure 3.1: Generalized linear fractional transformations.

Proof. The proof is similar to Proposition 3.1. QED

3.4.2 Robust Stability of Uncertain Descriptor Systems

Consider the uncertain homogeneous descriptor system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + B_1 w_u(t) \\ z_u(t) &= C_1 x(t) + D_1 w_u(t) \\ w_u(t) &= \Delta_{\mathrm{u}}(t) z_u(t) \end{aligned} \tag{3.20}$$

or equivalently,

$$E\dot{x}(t) = A(\Delta_{\mathrm{u}})x(t) \tag{3.21}$$

where $A(\Delta_{\mathrm{u}}) = \mathcal{F}_u\left(\begin{bmatrix} A & B_1 \\ C_1 & D_1 \end{bmatrix}, \Delta_{\mathrm{u}}\right) = A + B_1\Delta_{\mathrm{u}}(\mathrm{I} - D_1\Delta_{\mathrm{u}})^{-1}C_1$ and $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$ and the singular matrix E is assumed to be independent of the uncertainty Δ_{u}.

Lemma 3.2. The uncertain homogeneous descriptor system in (3.20) is robustly stable if and only if there exists $X \in \mathbb{R}^{n\times n}$ such that

$$\begin{aligned} E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0 \text{ and} \\ \begin{bmatrix} \mathrm{I} \\ A(\Delta_{\mathrm{u}}) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} \\ X & 0 \end{bmatrix} \begin{bmatrix} \mathrm{I} \\ A(\Delta_{\mathrm{u}}) \end{bmatrix} \prec 0 \end{aligned} \tag{3.22}$$

hold for all $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$.

Theorem 3.3. The uncertain homogeneous descriptor system in (3.20) is robustly stable if and only if there

exists a matrix $X \in \mathbb{R}^{n\times n}$ and $Y \in \mathcal{Y}_{\mathbf{\Delta}} \triangleq \{Y = Y^{\mathrm{T}} : Y \succ 0 \text{ on } \mathbf{Ker}([\mathrm{I} \ -\Delta_{\mathrm{u}}])\}$ such that

$$E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0 \text{ and}$$
$$\begin{bmatrix} \mathrm{I} & 0 \\ A & B_1 \\ 0 & \mathrm{I} \\ C_1 & D_1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} & & \\ X & 0 & & \\ & & Y_{11} & Y_{12} \\ & & Y_{12}^{\mathrm{T}} & Y_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ A & B_1 \\ 0 & \mathrm{I} \\ C_1 & D_1 \end{bmatrix} \prec 0 \tag{3.23}$$

hold.

Proof. Consider the conditions (3.22) in Lemma 3.2. Then the second LMI is equivalent to the scalar inequalities $\xi^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} \\ X & 0 \end{bmatrix} \xi$ for all $\xi \in \mathbf{Im}\left(\begin{bmatrix} \mathrm{I} \\ A(\Delta_{\mathrm{u}}) \end{bmatrix}\right)$, $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$. Define an augmented vector $\bar{\xi} \triangleq \begin{bmatrix} x^{\mathrm{T}} & \dot{x}^{\mathrm{T}}E^{\mathrm{T}} & w_u^{\mathrm{T}} & z_u^{\mathrm{T}} \end{bmatrix}$. Then there exists $X \in \mathbb{R}^{n\times n}$ such that $\xi^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} \\ X & 0 \end{bmatrix} \xi$ holds for all $\xi \in \mathbf{Im}\left(\begin{bmatrix} \mathrm{I} \\ A(\Delta_{\mathrm{u}}) \end{bmatrix}\right)$, $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$, if and only if $\bar{\xi}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} & 0 & 0 \\ X & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \bar{\xi} < 0$ for all $\bar{\xi} \in \mathbf{Im}\left(\begin{bmatrix} \mathrm{I} & 0 \\ A & B_1 \\ 0 & \mathrm{I} \\ C_1 & D_1 \end{bmatrix}\right)$ such that $T\bar{\xi} \in \mathbf{Ker}\left(\begin{bmatrix} \mathrm{I} & -\Delta_{\mathrm{u}} \end{bmatrix}\right)$ for all $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$, where $T \triangleq \begin{bmatrix} 0 & 0 & \mathrm{I} & 0 \\ 0 & 0 & 0 & \mathrm{I} \end{bmatrix}$. Applying the full block S-procedure in Theorem 3.1 (or see [233]) with Corollary 3.1, we complete the proof. QED

3.4.3 Robust Performance of Uncertain Descriptor Systems

First we consider linear inhomogeneous descriptor systems without model uncertainty for which dissipation properties can be written in terms of LMIs. Similar to robust stability analysis, extensions to uncertain linear inhomogeneous descriptor systems can be performed using the full block S-procedure in Thm. 3.1.

Nominal Dissipativity of Linear Descriptor Systems Consider the linear inhomogeneous descriptor system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + B_2 w_p(t) \\ z_p(t) &= C_2 x(t) + D_2 w_p(t) \end{aligned} \tag{3.24}$$

Theorem 3.4. The linear descriptor system with input-output pair (w_p, z_p) in (3.24) is dissipative with respect to a quadratic supply rate $S_r(-\Pi)$ for a given symmetric matrix Π of compatible dimension if and

only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ such that

$$E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0 \text{ and}$$
$$\begin{bmatrix} \mathrm{I} & 0 \\ A & B_2 \\ 0 & \mathrm{I} \\ C_2 & D_2 \end{bmatrix}^{\mathrm{T}} \underbrace{\begin{bmatrix} 0 & X^{\mathrm{T}} & & \\ X & 0 & & \\ & & \Pi_{11} & \Pi_{12} \\ & & \Pi_{12}^{\mathrm{T}} & \Pi_{22} \end{bmatrix}}_{M(X,\Pi)} \begin{bmatrix} \mathrm{I} & 0 \\ A & B_2 \\ 0 & \mathrm{I} \\ C_2 & D_2 \end{bmatrix} \preceq 0 \tag{3.25}$$

hold.

Proof. The second LMI is equivalent to the inequality $\xi^{\mathrm{T}}M(X,\Pi)\xi \leq 0$ for all $\xi^{\mathrm{T}} \triangleq \begin{bmatrix} x^{\mathrm{T}} & \dot{x}^{\mathrm{T}}E^{\mathrm{T}} & w_p^{\mathrm{T}} & z_p^{\mathrm{T}} \end{bmatrix}$ satisfying the system equation (3.24), which is a concatenation of the system signals in (3.24). Consider a Lyapunov function $V(x) = x^{\mathrm{T}}E^{\mathrm{T}}Xx$. Then the conditions in (3.25) are equivalent to the dissipation inequality $s(-\Pi) \geq \dot{V}(x)$, and $V(x) \geq 0$ for all $x \in \mathbb{R}^n$ ($V(x) > 0$ for all $x \in \mathbb{R}^n - \mathcal{E}_{\mathrm{eq}}$, where $\mathcal{E}_{\mathrm{eq}}$ is the set of equilibrium points). QED

Below is a generalized positive real lemma for linear descriptor systems (3.24).

Corollary 3.2. The transfer function $G(s)$ is positive real if and only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ such that its minimal realization (E, A, B_2, C_2, D_2) satisfies the conditions in (3.25) with $\Pi = \begin{bmatrix} 0 & -\mathrm{I} \\ -\mathrm{I} & 0 \end{bmatrix}$.

Corollary 3.3. The transfer function $G(s)$ is strictly positive real if and only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ such that its minimal realization (E, A, B_2, C_2, D_2) satisfies the conditions in (3.25) with $\Pi = \begin{bmatrix} 0 & -\mathrm{I} \\ -\mathrm{I} & 0 \end{bmatrix}$ and the upper leftmost block $\begin{bmatrix} 0 & X^{\mathrm{T}} \\ X & 0 \end{bmatrix}$ replaced by $\begin{bmatrix} \epsilon E^{\mathrm{T}}X & X^{\mathrm{T}} \\ X & 0 \end{bmatrix}$ for some $\epsilon > 0$.

Remark 3.5. Several LMI-type conditions for passivity and PR of linear descriptor systems have been derived [86,171,303]. The well-known KYP lemma characterizes their intimate relation and has been studied in the behavioral framework [47].

Remark 3.6. The linear descriptor system with input-output pair (w_p, z_p) in (3.24) is strictly input- and output-passive (see Defn. 3.2) if and only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ such that the conditions (3.25) hold with $\Pi = \begin{bmatrix} \epsilon\mathrm{I} & -\mathrm{I} \\ -\mathrm{I} & 0 \end{bmatrix}$ and $\Pi = \begin{bmatrix} 0 & -\mathrm{I} \\ -\mathrm{I} & \epsilon\mathrm{I} \end{bmatrix}$ for some $\epsilon > 0$, respectively.

Robust Performance of Uncertain Descriptor Systems Consider the descriptor system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + B_1 w_u(t) + B_2 w_p(t), \\ z_u(t) &= C_1 x(t) + D_{11} w_u(t) + D_{12} w_p(t), \\ z_p(t) &= C_2 x(t) + D_{21} w_u(t) + D_{22} w_p(t), \end{aligned} \tag{3.26}$$

where $E, A \in \mathbb{R}^{n\times n}$ with $\mathrm{rank}(E) \leq n$, $B_1 \in \mathbb{R}^{n\times m_1}$, $B_1 \in \mathbb{R}^{n\times m_2}$, $C_1 \in \mathbb{R}^{l_1\times n}$, $C_2 \in \mathbb{R}^{l_2\times n}$, and $D_{ij} \in \mathbb{R}^{l_i\times m_j}$ for $i = 1, 2$ and $j = 1, 2$, and the pair of signals $(w_u(t), z_u(t))$ satisfies the geometric implicit relation given in (3.2). The system can be equivalently written as

$$\begin{aligned} E\dot{x}(t) &= A(\Delta)x(t) + B(\Delta)w_p(t), \\ z_p(t) &= C(\Delta)x(t) + D(\Delta)w_p(t), \end{aligned} \tag{3.27}$$

where

$$\begin{bmatrix} A(\Delta) & B(\Delta) \\ C(\Delta) & D(\Delta) \end{bmatrix} \triangleq \begin{bmatrix} A & B_2 \\ C_2 & D_{22} \end{bmatrix} + \begin{bmatrix} B_1 \\ D_{21} \end{bmatrix} S_1(\Delta)(S_2(\Delta) - D_{11}S_1(\Delta))^{-1} \begin{bmatrix} C_1 & D_{12} \end{bmatrix} \tag{3.28}$$

and $\begin{bmatrix} S_1(\Delta) \\ S_2(\Delta) \end{bmatrix} \triangleq \mathbf{Ker}(\Delta)$ for each $\Delta \in \mathbf{\Delta}$.

Suppose that the uncertainty in the system (3.26) is assumed to be described by an implicit equation $w_u(t) = \Delta_{\mathrm{u}}(t)z_u(t)$ with $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$, without loss of generality.

Lemma 3.3. The uncertain descriptor system with input-output pair (w_p, z_p) given in (3.26) is dissipative with respect to a quadratic supply rate $S_r(-\Pi)$ for a given symmetric matrix Π of compatible dimension if and only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ such that

$$E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0 \text{ and}$$
$$\begin{bmatrix} \mathrm{I} & 0 \\ A(\Delta) & B(\Delta) \\ 0 & \mathrm{I} \\ C(\Delta) & D(\Delta) \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} & & \\ X & 0 & & \\ & & \Pi_{11} & \Pi_{12} \\ & & \Pi_{12}^{\mathrm{T}} & \Pi_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 \\ A(\Delta) & B(\Delta) \\ 0 & \mathrm{I} \\ C(\Delta) & D(\Delta) \end{bmatrix} \prec 0 \tag{3.29}$$

hold for all $\Delta \in \mathbf{\Delta}$.

Theorem 3.5. The uncertain descriptor system with input-output pair (w_p, z_p) given in (3.26) is dissipative with respect to a quadratic supply rate $S_r(-\Pi)$ for a given symmetric matrix Π of compatible dimension if and only if there exists a matrix $X \in \mathbb{R}^{n\times n}$ and $Y \in \mathcal{Y}_{\mathbf{\Delta}} \triangleq \{Y = Y^{\mathrm{T}} : Y \succ 0 \text{ on } \mathbf{Ker}([\mathrm{I} \; -\Delta_{\mathrm{u}}]), \forall \Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}\}$

such that

$$E^{\mathrm{T}}X = X^{\mathrm{T}}E \succeq 0 \text{ and}$$

$$\begin{bmatrix} \mathrm{I} & 0 & 0 \\ A & B_1 & B_2 \\ 0 & \mathrm{I} & 0 \\ C_1 & D_{11} & D_{12} \\ 0 & 0 & \mathrm{I} \\ C_2 & D_{21} & D_{22} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 0 & X^{\mathrm{T}} & & & & \\ X & 0 & & & & \\ & & Y_{11} & Y_{12} & & \\ & & Y_{12}^{\mathrm{T}} & Y_{22} & & \\ & & & & \Pi_{11} & \Pi_{12} \\ & & & & \Pi_{12}^{\mathrm{T}} & \Pi_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0 & 0 \\ A & B_1 & B_2 \\ 0 & \mathrm{I} & 0 \\ C_1 & D_{11} & D_{12} \\ 0 & 0 & \mathrm{I} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \prec 0 \tag{3.30}$$

hold.

Proof. Directly follows from Thms. 3.3 and 3.4. QED

The LMI formalism presented in this section includes many existing results as special cases. For example, it is not hard to show that an LMI condition can be derived from (3.25) in Thm. 3.4 that is equivalent to a BMI condition in [218]. The bilinear terms in [218] results from an unnecessary use of the Schur complement lemma that produced a higher-dimensional BMI, whereas a simple congruence transformation and a permutation give an equivalent LMI

$$\begin{bmatrix} X^{\mathrm{T}}A + A^{\mathrm{T}}X & * & * & * & * \\ B_1^{\mathrm{T}}X & -\lambda\mathrm{I} & * & * & * \\ B_2^{\mathrm{T}}X & 0 & -\mathrm{I} & * & * \\ \lambda C_1 & 0 & \lambda D_{12} & -\lambda\mathrm{I} & * \\ C_2 & D_{21} & D_{22} & 0 & \frac{1}{\gamma^2}\mathrm{I} \end{bmatrix} \prec 0,$$

which can be derived by multiplication of the matrices in condition (3.30).

3.5 Robust Impulse-Free and Stable Uncertain Descriptor Systems: μ Approaches

Descriptor systems may undergo impulsive responses and show nonuniqueness of the solution trajectories that cause performance degradation, damage of system components, or even completely destroy the whole system. Such an undesirable impulsive behavior needs to be investigated and eliminated before implementation on a real system. This section provides necessary and sufficient conditions for uncertain linear descriptor systems for impulse-free dynamics using structured singular value theory.

Consider the uncertain linear descriptor system

$$E\dot{x}(t) = A(\Delta_{\rm u})x(t), \tag{3.31}$$

where $A(\Delta_{\rm u}) = \mathcal{F}_u\left(\begin{bmatrix} A & B \\ C & D \end{bmatrix}, \Delta_{\rm u}\right) = A + B\Delta_{\rm u}({\rm I} - D\Delta_{\rm u})^{-1}C$, $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$, and the set of matrices of block-diagonal perturbations given by

$$\mathbf{\Delta}_{\rm u} \triangleq \left\{ {\rm diag}\left(\delta_1 {\rm I}_{r_1}, \ldots, \delta_k {\rm I}_{r_k}, \Delta_{k+1}, \ldots, \Delta_{m_c}\right) : \delta_i \in \mathbb{C},\, \Delta_i \in \mathbb{C}^{r_i \times r_i},\, \sum_{i=1}^{m_c} r_i = m \right\} \tag{3.32}$$

and $\mathbf{B\Delta}_{\rm u}$ is the set of unity norm-bounded perturbations from $\mathbf{\Delta}_{\rm u}$, which is the time-invariant version of the uncertainty description in the previous section.

As a first property, we also need to characterize that the representation of uncertain descriptor systems in (3.31) is *well-posed*, that is, the interconnection of D and $\Delta_{\rm u}$ has nonsingular ${\rm I} - D\Delta_{\rm u}$ for all $\Delta_{\rm u} \in \mathbf{B\Delta}_{\rm u}$.

Next we derive robust impulse-free and stability conditions in terms of the structured singular value (μ) [308]: For a given matrix $M \in \mathbb{C}^{m\times m}$,

$$\mu_{\mathbf{\Delta}}(M) \triangleq \begin{cases} 0 \quad \text{if there exists no } \Delta \in \mathbf{\Delta} \text{ such that } \det({\rm I} - M\Delta) = 0 \\ \left(\min_{\Delta\in\mathbf{\Delta}} \{\bar{\sigma}(\Delta) \,:\, \det({\rm I} - M\Delta) = 0\}\right)^{-1} \text{ otherwise} \end{cases} \tag{3.33}$$

Our subsequent results will generalize the results of [159] for diagonal non-repeated real parametric uncertainty to general structured uncertainty.

3.5.1 Robust Impulse-Free Condition

Robust impulse-free descriptor systems are defined below, which is an extension of Defn. 3.3 to uncertain systems.

Definition 3.3 (*Robust impulse-free*). The uncertain system (3.31) is *robust impulse-free* if it is impulse-free for all $\Delta_{\rm u} \in \mathbf{B\Delta}_{\rm u}$.

The next result gives a necessary and sufficient condition for the uncertain system (3.31) to be robust impulse-free.

Theorem 3.6. Suppose that the matrix pencil (E, A) is regular and impulse-free. Then the uncertain system (3.31) is robust impulse-free if and only if

$$\mu_{\mathbf{\Delta}_{\rm u}}(CR_a(L_aAR_a)^{-1}L_aB - D) < 1, \tag{3.34}$$

where $L_a \in \mathcal{L}_\perp(E)$ and $R_a \in \mathcal{R}_\perp(E)$.

Proof. From Lemma **??** and Defn. 3.3, the uncertain system (3.31) is robust impulse-free if and only if $L_a(A + B\Delta_{\rm u}(\mathrm{I} - D\Delta_{\rm u})^{-1}C)R_a$ is invertible for any $\Delta_{\rm u} \in \mathbf{B}\mathbf{\Delta}_{\rm u}$, which is equivalent to the determinant condition, $\det(L_a(A + B\Delta_{\rm u}(\mathrm{I} - D\Delta_{\rm u})^{-1}C)R_a) \neq 0$, for any $\Delta_{\rm u} \in \mathbf{B}\mathbf{\Delta}_{\rm u}$. Using the matrix inversion lemma (also known as the Sherman-Morrison-Woodbury formula) [109] and that the nominal system $\Sigma(E, A)$ is assumed to be impulse-free such that $\det(L_aAR_a) \neq 0$, this relation is equivalent to $\det(L_a(A + B\Delta_{\rm u}(\mathrm{I} - D\Delta_{\rm u})^{-1}C)R_a) = \det(L_aAR_a)\det(\mathrm{I} + (CR_a(L_aAR_a)^{-1}L_aB - D)\Delta_{\rm u})$. This implies that the uncertain system (3.31) is robust impulse-free if and only if $\det(\mathrm{I} + (CR_a(L_aAR_a)^{-1}L_aB - D)\Delta_{\rm u})$ for any $\Delta_{\rm u} \in \mathbf{B}\mathbf{\Delta}_{\rm u}$, which is equivalent to the μ condition (3.34). QED

3.5.2 μ Condition for Robust Impulse-Free and Stable Systems

The next result is a necessary and sufficient condition for the uncertain system (3.31) to be robust impulse-free and stable.

Theorem 3.7. Suppose that the matrix pencil (E, A) is regular, impulse-free, and stable. Then the uncertain system (3.31) is robust impulse-free and stable if and only if

$$\mu_{\hat{\mathbf{\Delta}}_{\rm u}}\left(\hat{C}\begin{bmatrix} R_a(L_aAR_a)^{-1}L_a & 0 \\ 0 & (j\omega E - A)^{-1} \end{bmatrix}\hat{B} - \hat{D}\right) < 1 \tag{3.35}$$

for all $\omega \in \mathbb{R} \cup \{\infty\}$, where $\hat{\mathbf{\Delta}}_{\rm u} \triangleq \{\hat{\Delta} = \mathrm{diag}(\Delta, \Delta) : \Delta \in \mathbf{\Delta}_{\rm u}\}$, $\hat{B} = \mathrm{diag}(B, B)$, $\hat{C} = \mathrm{diag}(C, C)$, $\hat{D} = \mathrm{diag}(D, -D)$, and $L_a \in \mathcal{L}_\perp(E)$ and $R_a \in \mathcal{R}_\perp(E)$ are arbitrary.

Proof. A necessary condition for (3.35) is that the uncertain system (3.31) is robust impulse-free. Thus, it only needs to be shown that the robust stability of the system (3.31) is equivalent to the μ condition, $\mu_{\mathbf{\Delta}_{\rm u}}(G(j\omega)) \leq 1$ for all $\omega \in \mathbb{R} \cup \{\infty\}$, where $G(s) \triangleq C(sE - A)^{-1}B + D$. The uncertain system (3.31) is robustly stable if and only if $\{s \in \mathbb{C} : \det(sE - A - B\Delta_{\rm u}(s)(\mathrm{I} - D\Delta_{\rm u}(s))^{-1}C) = 0\} \subset \mathbb{C}_-$ for any $\Delta_{\rm u} \in \mathbf{B}\mathbf{\Delta}_{\rm u}$, where the argument $s \in \mathbb{C}$ is explicitly shown in $\Delta_{\rm u}$ to emphasize that the uncertainty $\Delta_{\rm u}$ has LTI dynamics in general. Using the matrix inversion lemma (aka Sherman-Morrison-Woodbury formula) [109] and that the nominal system $\Sigma(E, A)$ is assumed to be stable such that $\sigma(E, A) \subset \mathbb{C}_-$, this relation is equivalent to $\det(sE - A - B\Delta_{\rm u}(s)(\mathrm{I} - D\Delta_{\rm u}(s))^{-1}C) \neq 0$ if and only if $\det(sE - A)\det(\mathrm{I} - (C(sE - A)^{-1}B + D)\Delta_{\rm u}(s)) \neq 0$ for all $s \in \bar{\mathbb{C}}_+$. Since $\det(sE - A) \neq 0$ for all $s \in \bar{\mathbb{C}}_+$, $\det(sE - A - B\Delta_{\rm u}(s)(\mathrm{I} - D\Delta_{\rm u}(s))^{-1}C) \neq 0$ for all $s \in \bar{\mathbb{C}}_+$ and $\Delta_{\rm u} \in \mathbf{B}\mathbf{\Delta}_{\rm u}$ if and only if $\mu_{\mathbf{\Delta}_{\rm u}}(G(s)) < 1$ for all $s \in \bar{\mathbb{C}}_+$, which follows from the fact that $G(s)$ is proper (or impulse-free) and the definition of the structured singular value. QED

3.6 Summary and Future Work

The need for unified and generalizable conditions for robust stability and performance of uncertain linear descriptor systems motivates the systematic construction of convex optimization problems in this chapter. The presented tests are written in terms of LMIs that are computationally tractable via existing interior-point methods. The tests are obtained from an extension of the full block S-procedure with which a number of matrix inequalities for unknown variables can be rewritten as one matrix inequality. Applicability of the full block S-procedure to structured uncertain linear descriptor systems is supported. The conditions can be considered as a unification and generalization of several existing results that are distributed in many places, but in similar contexts. Apart from constructing the LMI conditions for robust stability and performance of linear uncertain descriptor systems, numerically reliable tests are proposed that are written in terms of the μ conditions for linear descriptor systems with structured uncertainty to be robust impulse-free.

Chapter 4

Robust Reliable Control of Uncertain Systems with Faults

Abstract This chapter provides necessary and sufficient conditions for several forms of controlled system reliability. For comparison purposes, past results on the reliability analysis of controlled systems are reviewed and it is shown that several of the past results are either conservative or have exponential complexity. For systems with real and complex uncertainties, conditions for *robust reliable stability and performance* are derived in terms of the structured singular values of certain transfer functions. The conditions are necessary and sufficient for the controller to stabilize the closed-loop system while retaining a desirable level of the closed-loop performance in the presence of actuator/sensor faults or failures, as well as plant-model mismatches. The resulting conditions based on the structured singular value are applied to the decentralized control for a high-purity distillation column and singular value decomposition-based optimal control for a parallel reactor with combined precooling. Tight polynomial-time bounds for the conditions can be evaluated by using available off-the-shelf software.

4.1 Introduction

An inevitable consequence of industrial practice is that actuators and sensors can become faulty or fail, which motivates the development of methods to evaluate the reliability of the closed-loop system to such imperfect operations. A feedback-controlled system is said to be *reliable* if it is guaranteed to retain desired closed-loop system properties while tolerating faults or failures of actuators and/or sensors. Maximizing the reliability of a system concerns minimizing its potential performance degradation while retaining closed-loop stability when a fault or failure occurs in a control/measurement channel. In addition to the possibility of actuator/sensor faults or failures, plant-model mismatches are also inevitable, which motivates their

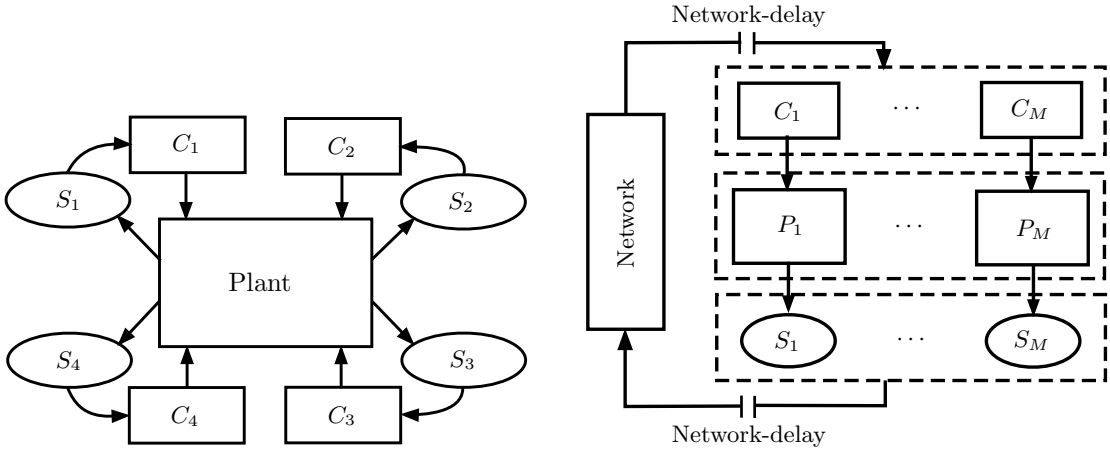
incorporation into reliability and integrity analysis. This chapter is motivated by the need for nonconservative testing conditions to ensure closed-loop stability and to retain a satisfactory closed-loop performance in the presence of both plant-model mismatches and actuator/sensor faults or failures.

This chapter primally considers decentralized controlled systems and studies their *robust reliable stability and performance* in the presence of possible actuator/sensor faults or failures with consideration of the overall plant-model mismatches that are described in terms of bounded set-valued linear operators. The main purpose of this chapter is to derive necessary and sufficient conditions for various types of *robust reliable stability and performance* of a set-valued plant model that is described by a linear fractional transformation (LFT) with structured uncertainties [308]. It is assumed that any failure of a local controller is detected and the controller is taken out of service whenever a failure occurs, so that any undesirable propagation of local failures to other parts of the system can be avoided. Even though we mostly concentrate on the analysis of decentralized control systems, the proposed approach does not depend on the structure of the selected control schemes and can be applied to any type of linear controller and actuator-sensor selection.

Decentralized control depicted in Figure 4.1(a) is ubiquitous in industrial applications, which is a special case of large-scale interconnected systems with interactions between subsystems and constraints on information flows. Extensive overviews on decentralized control are available [7, 243]. For decentralized controlled systems, actuator/sensor faults or failures can occur and the selection of a reliable actuator/sensor structure is an important consideration [40, 156]. One of the resurgent questions in systems and control theory related to *reliable decentralized control* is how to study the effect and propagation of communication link failures between several components of a *networked control system* (NCS) depicted in Figure 4.1(b) on the stability and performance of the overall system [119, 263, 284, 304]. Although studied for decades, NCSs have received a large surge of interest in recent years. As time delays and communication losses are inevitable in an NCS, reliability analysis in the presence of faults and failures in communication networks is also important.

In [241, 242], multi-controller systems were introduced for reliable control and since then reliable stabilization problems under various failure and fault scenarios have been studied using decentralized configurations [52, 101, 184, 185, 259]. In particular, the reliability of decentralized control with integral action was investigated in terms of steady-state gain matrices in [52, 184] and existence conditions for a reliably stabilizing decentralized integral controller have been derived in terms of the Niederlinski index (NI) and block relative gain (BRG) [131]. In [101] explicit conditions for reliable decentralized control of linear systems were derived for a two-channel decentralized feedback control configuration, and coprime factorization methods and a design method for such controllers were proposed [102].

In addition to the aforementioned frequency-domain approaches, some researchers have suggested design methods for reliable controllers in terms of state-space realizations of the plant and controller. Centralized reliable state feedback controllers were suggested in [124, 169] and design methods for decentralized reliable observer-based output-feedback controllers were presented in [69, 277]. A state feedback control design for dynamic systems in the presence of actuator failures has been proposed [306] in which robust pole placement



(a) Decentralized control systems (b) Networked control systems

Figure 4.1: Large-scale interconnected systems.

methods are adopted while requiring redundant actuators to recover the normal level of operation. The design method of [306] was only applicable to state feedback control problems without any plant-model mismatch, so that the proposed design methods may fail in real problems in which uncertainties are inevitable. In [236], a simple high-gain state feedback control based on a Riccati type equation was proposed with actuator redundancy. The model uncertainties were assumed to be time-varying, but not fully structured and no uncertainty was allowed in the input channel matrices.

The approaches proposed in this chapter are based on the structured singular value (μ) and a standard representation of uncertain systems known as the linear fractional transformation (LFT). Robust reliable control problems for large-scale systems with decentralized control are reformulated in terms of robustness analysis based on μ to model the effects of faults. Structures of interconnected sensors and actuators as well as structures of uncertainties can be fully exploited to perform nonconservative or less conservative analysis. Some of the results in this chapter were presented in [41] and subsequently there were many research efforts such as the aforementioned works to develop robust reliable controllers. The main objective of this chapter is to provide an efficient framework for the analysis and synthesis of robust reliability. Faults and failures in process components are treated as parametric uncertainties that are compatible with μ. Due to a resurgence of research interest in robust reliable control for systems with integral action, our past results [41] were extended to derive conditions for robust reliable stability of decentralized systems with integral action. Although the main focus of this chapter is on decentralized control problems, the methodology is not restricted to decentralized control and the results can be extended to general control structures in a straightforward manner.

4.1.1 Reliability of Decentralized Control

For notational convenience, the controller is assumed to be fully decentralized, i.e., the controller K is diagonal. Most of the results can be extended in an obvious manner to block-diagonal controllers and even to centralized controllers. Usually, the square plant P is assumed to be stable; the results do not carry over easily to systems that are open-loop unstable.

Several strong forms of reliability to failure of actuators or sensors are defined in the open literature for systems without plant-model mismatch. Below we review those forms of reliability and extend the definitions to uncertain systems. To simplify the presentation, we primarily focus on a discussion of reliability to actuator faults or failures, although very similar definitions and the results can be trivially extended to the other process equipment.

Integrity is defined as follows [41, 184, 185, 241, 242].

Definition 4.1. The closed-loop system demonstrates *integrity* if $K_{\mathrm{f}}(s) := EK(s)$ stabilizes $P(s)$ for all $E \in \mathcal{E}_{1/0} \triangleq \{\mathrm{diag}\{\epsilon_i\} : \epsilon_i \in \{0,1\},\, i = 1, \ldots, n\}$.

A closed-loop system that demonstrates integrity to actuator failures remains stable as actuators are arbitrarily brought in and out of service. For a system to demonstrate integrity, the nominal plant model $P(s)$ must be stable. To have actuator failure tolerance when the controller is unstable, the failures must be recognized and the corresponding columns of the controller taken off-line. It is clear that the integrity of a system can be tested through 2^n stability (eigenvalue) determinations.

The following definition extends integrity to uncertain systems.

Definition 4.2. The closed-loop system demonstrates *robust integrity* if $K_{\mathrm{f}}(s) := EK(s)$ stabilizes $P_\Delta(s)$ for all $E \in \mathcal{E}_{1/0}$ and all $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$ such that $\|\Delta_{\mathrm{u}}\|_\infty \leq 1$.

An uncertain system demonstrates robust integrity to actuator failures if it remains stabilized for any plant given by the uncertainty description, as actuators are arbitrarily brought in and out of service. For a system to demonstrate robust integrity, the plant must be stable for all allowed perturbations. To have actuator failure tolerance when the controller is unstable, the failures must be recognized and the corresponding columns of the controller taken off-line, just as in the nominal case. Note that robust integrity implies integrity. It is

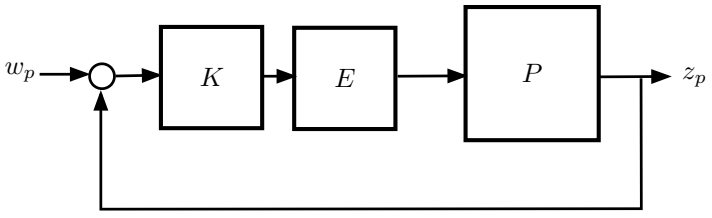


Figure 4.2: Integrity under actuator faults/failures. For robust integrity, replace P by the set of uncertain plants P_Δ.

clear that the robust integrity of a system can be tested through 2^n nominal stability (eigenvalue) and 2^n robust stability (μ) calculations.

A very strong notion of reliability was defined by Campo and Morari [52] for decentralized controllers. The requirement is that the nominal closed-loop system remains stable under arbitrary independent detuning of the controller gains. For decentralized control systems, this is equivalent to arbitrary detuning of the actuator/sensor gains to zero. Having stability with detuning allows the operators to safely change the closed-loop speed of response depending on process operating conditions.

Definition 4.3. The closed-loop system is *decentralized unconditionally stable (DUS)* if $K_{\mathrm{f}}(s) := EK(s)$ stabilizes $P(s)$ for all $E \in \mathcal{E}_D \triangleq \{\mathrm{diag}\{\epsilon_i\} : \epsilon_i \in (0,1)\}$.

The closed-loop system will not be DUS if either the plant $P(s)$ or controller $K(s)$ has poles in the open right-half plane (ORHP). To see this, let us consider the multivariable root locus [249] with equal detuning $\epsilon_i = \epsilon$ for all i. For small ϵ, the closed-loop poles approach the open-loop poles. Since the closed-loop poles are a continuous function of the controller gain, if any of the open-loop poles are in the ORHP then some of the closed-loop poles will be unstable for sufficiently small ϵ.

The following is the generalization to uncertain systems.

Definition 4.4. The closed-loop system is *robust decentralized unconditionally stable (RDUS)* if $K_{\mathrm{f}}(s) := EK(s)$ stabilizes $P_\Delta(s)$ for all $E \in \mathcal{E}_D$ and all $\Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}}$ such that $\|\Delta_{\mathrm{u}}\|_\infty \leq 1$.

By a similar argument as used for DUS, the closed-loop system will not be RDUS if any poles of the controller $K(s)$ or any plant given by the uncertainty description are in the ORHP. For open-loop unstable controllers or plants, some minimum amount of feedback is required for closed-loop stability.

Actually, the definition of DUS used by Campo and Morari [52] requires that the closed-loop system be stable for all $\epsilon \in [0,1]$. Here we refer to this notion as *closed decentralized unconditional stability (CDUS)*, with *closed robust decentralized unconditional stability (CRDUS)* defined similarly. These definitions of reliability require stability under total malfunctions of some actuators and allows perfect functioning of some actuators while other actuators are not working at all.

4.2 Analysis for Reliability of Decentralized Control using μ

This section primarily focuses on the nominal and robust fault tolerance of systems that are affected by real parametric uncertainties and complex dynamic uncertainties. The detuned control gains of decentralized controllers are assumed to be real constants, unknown but bounded by open or closed intervals.

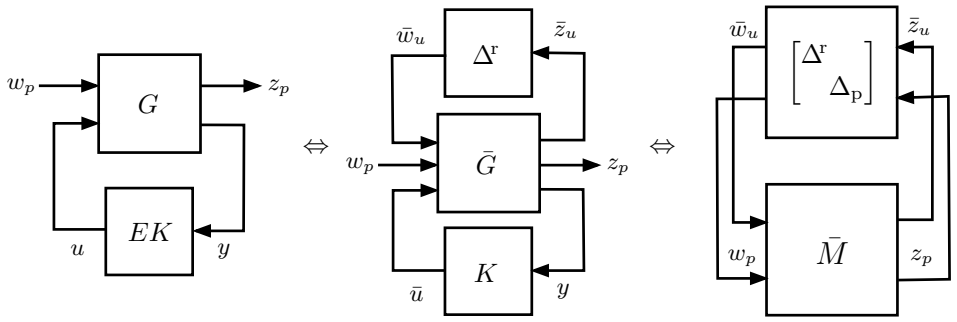


Figure 4.3: Equivalent LFTs of fault tolerance.

4.2.1 Modeling Faults using μ

Braatz [39] describes in some detail the modeling of faults with either uncertainty and/or performance descriptions. This modeling can be combined with requirements on the stability or performance during faulty operation to derive a μ condition that provides a test for system reliability. The following discussion illustrates how to model actuator gain variation for two cases: (i) without additional uncertainty (i.e., plant/model mismatch) and (ii) with additional uncertainty.

The nominal linear dynamic output feedback controller is defined to be $K(s) \in \mathbb{C}^{m\times\ell}$. Then the controller with gain variation can be described by $\tilde{K}(s) = EK(s)$, where $E \in \mathcal{E}[\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{upper}}] \triangleq \{\mathrm{diag}\{\epsilon_i\} : \epsilon_i \in [\epsilon_{i,\mathrm{low}}, \epsilon_{i,\mathrm{upper}}]\}$. Any $E \in \mathcal{E}[\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{upper}}]$ can be rewritten as

$$E \triangleq \bar{E} + W^{\mathrm{r}}\Delta^{\mathrm{r}} \tag{4.1}$$

where $\bar{E} = \mathrm{diag}(\bar{\epsilon}_i)$ with $\bar{\epsilon}_i \triangleq \frac{\epsilon_{i,\mathrm{low}}+\epsilon_{i,\mathrm{upper}}}{2}$, $W^{\mathrm{r}} = \mathrm{diag}\{\omega_i\}$ with $\omega_i \triangleq \frac{\epsilon_{i,\mathrm{upper}}-\epsilon_{i,\mathrm{low}}}{2}$, and Δ^{r} is a diagonal real independent uncertainty, i.e., $\Delta^{\mathrm{r}} = \mathrm{diag}\{\delta_i\}$ with $\delta_i \in [-1, 1]$, $i = 1, \ldots, m$.

Theorem 4.1. Suppose that the model of a system is represented by a transfer function matrix $P(s)$ without any additional uncertainty. The system remains stable under the gain variation defined with $E \in \mathcal{E}[\epsilon_{\mathrm{low}}, \epsilon_{\mathrm{upper}}]$ if and only if

$$\mu_{\mathbf{\Delta}^{\mathrm{r}}}(\bar{M}_{11}(j\omega)) < 1, \quad \forall\omega \in \mathbb{R} \cup \{\infty\}, \tag{4.2}$$

where $\bar{M}_{11}(s) = -K(s)(\mathrm{I} + P(s)\bar{E}K(s))^{-1}P(s)W^{\mathrm{r}}$ and $\Delta^{\mathrm{r}} \in \mathbf{\Delta}^{\mathrm{r}} \triangleq \{\mathrm{diag}\{\delta_i\} : \delta_i \in [-1, 1],\ i = 1, \ldots, m\}$.

Proof. The sequence of equivalent representations in Figure 4.3 is obtained with the system transfer function matrices

$$G := \begin{bmatrix} 0 & P \\ \mathrm{I} & -P \end{bmatrix},\ \bar{G} := \begin{bmatrix} 0 & 0 & \mathrm{I} \\ PW^{\mathrm{r}} & 0 & P\bar{E} \\ -PW^{\mathrm{r}} & \mathrm{I} & -P\bar{E} \end{bmatrix} \tag{4.3}$$

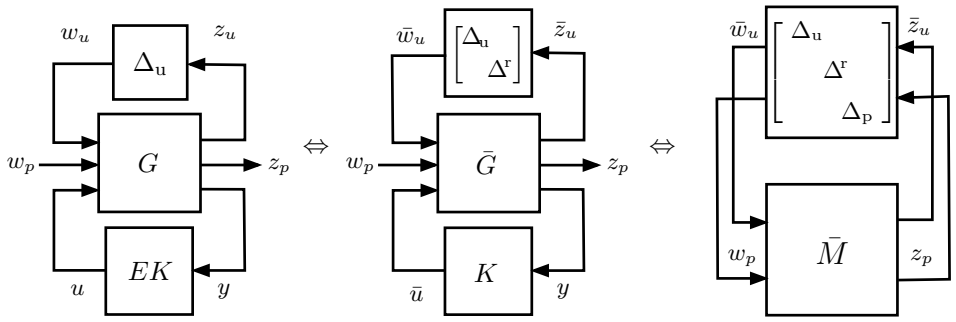


Figure 4.4: Equivalent LFTs of robust fault tolerance.

and

$$\begin{aligned}\bar{M} &:= \mathcal{F}_\ell(\bar{G}, K) \\ &= \begin{bmatrix} 0 & 0 \\ PW^{\rm r} & 0 \end{bmatrix} + \begin{bmatrix} \mathrm{I} \\ P\bar{E} \end{bmatrix} K(\mathrm{I} + P\bar{E}K)^{-1} \begin{bmatrix} -PW^{\rm r} & \mathrm{I} \end{bmatrix}.\end{aligned} \tag{4.4}$$

The definition of the structured singular value [72, 308] implies that the system is robustly stable under any gain variation $E \in \mathcal{E}[\epsilon_{\rm low}, \epsilon_{\rm upper}]$ if and only if $\mu_{\mathbf{\Delta}^{\rm r}}(\bar{M}_{11}(j\omega)) < 1$ for all $\omega \in \mathbb{R} \cup \{\infty\}$. QED

Theorem 4.2. Suppose that the model of a system is represented by a transfer function matrix $P(s)$ without any additional uncertainty. The system achieves the unity (reliable) $\mathcal{H}_\infty$ performance under the gain variation defined with $E \in \mathcal{E}[\epsilon_{\rm low}, \epsilon_{\rm upper}]$ if and only if

$$\mu_{\mathbf{\Delta}}(\bar{M}(j\omega)) < 1, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.5}$$

where $\Delta \in \mathbf{\Delta} \triangleq \{\mathrm{diag}(\Delta^{\rm r}, \Delta_{\rm p}) : \Delta^{\rm r} \in \mathbf{\Delta}^{\rm r} \text{ and } \Delta_{\rm p} \in \mathbb{C}^{m_2 \times \ell_2}\}$ and the matrix transfer function

$$\bar{M}(s) \triangleq \begin{bmatrix} -K(s)(\mathrm{I} + P(s)\bar{E}K(s))^{-1}P(s)W^{\rm r} & K(s)(\mathrm{I} + P(s)\bar{E}K(s))^{-1} \\ P(s)W^{\rm r} - P(s)\bar{E}K(s)(\mathrm{I} + P(s)\bar{E}K(s))^{-1}P(s)W^{\rm r} & P(s)\bar{E}K(s)(\mathrm{I} + P(s)\bar{E}K(s))^{-1} \end{bmatrix}. \tag{4.6}$$

Proof. Applying the main-loop theorem [308, Theorem 11.9] to the matrix transfer function $\bar{M}(s)$ given in (4.4) completes the proof. QED

Testing the maintainence of closed-loop stability and/or performance with respect to both actuator gain variation and additional perturbations like plant-model mismatch involves more complicated expressions for M and G.

Theorem 4.3. Suppose that the model of a system is represented by the standard LFT with uncertainty Δ_u, i.e., $P_\Delta = \mathcal{F}_{\rm u}(P, \Delta_u)$. The system remains stable under the gain variation defined with $E \in \mathcal{E}[\epsilon_{\rm low}, \epsilon_{\rm upper}]$ if and only if

$$\mu_{\mathbf{\Delta}_{\rm a}}(\bar{M}_{11}(j\omega)) < 1, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.7}$$

where $\bar{M}_{11}$ is the submatrix transfer function corresponding to the uncertainty block $\Delta_{\rm a} \triangleq \mathrm{diag}\{\Delta_{\rm u}, \Delta^{\rm r}\}$ of the total transfer function matrix

$$\bar{M} \triangleq \begin{bmatrix} P_{11} - P_{12}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{21} & P_{12}W^{\rm r} - P_{12}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{22}W^{\rm r} & P_{12}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1} \\ -K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{21} & -K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{22}W^{\rm r} & K(\mathrm{I} + P_{22}\bar{E}K)^{-1} \\ P_{21} - P_{22}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{21} & P_{22}W^{\rm r} - P_{22}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1}P_{22}W^{\rm r} & P_{22}\bar{E}K(\mathrm{I} + P_{22}\bar{E}K)^{-1} \end{bmatrix}. \tag{4.8}$$

Proof. The sequence of equivalent representations in Figure 4.4 is obtained with the system transfer function matrices

$$G := \begin{bmatrix} P_{11} & 0 & P_{12} \\ P_{21} & 0 & P_{22} \\ -P_{21} & \mathrm{I} & -P_{22} \end{bmatrix}, \quad \bar{G} := \begin{bmatrix} P_{11} & P_{12}W^{\rm r} & 0 & P_{12}\bar{E} \\ 0 & 0 & 0 & \mathrm{I} \\ P_{21} & P_{22}W^{\rm r} & 0 & P_{22}\bar{E} \\ -P_{21} & -P_{22}W^{\rm r} & \mathrm{I} & -P_{22}\bar{E} \end{bmatrix},$$

and $\bar{M}(s)$ is given in (4.8). which implies that the system is robustly stable for any uncertainty $\Delta_u \in \mathbf{\Delta_u}$ and under any gain variation $E \in \mathcal{E}[\epsilon_{\rm low}, \epsilon_{\rm upper}]$ if and only if $\mu_{\mathbf{\Delta}_{\rm a}}(\bar{M}_{11}(j\omega)) < 1$ for all $\omega \in \mathbb{R} \cup \{\infty\}$. QED

Theorem 4.4. Suppose that the model of a system is represented by a transfer function matrix $P(s)$ without any additional uncertainty. The system achieves an $\mathcal{H}_\infty$ performance, $\sup_{\|w_p\|_2 \le 1} \frac{\|z_p\|_2}{\|w_p\|_2} \le 1$, under the gain variation defined with $E \in \mathcal{E}[\epsilon_{\rm low}, \epsilon_{\rm upper}]$ if and only if

$$\mu_{\mathbf{\Delta}}(\bar{M}(j\omega)) < 1, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.9}$$

where $\Delta \in \mathbf{\Delta} \triangleq \{\mathrm{diag}\{\Delta_{\rm u}, \Delta^{\rm r}, \Delta_{\rm p}\} : \Delta_{\rm u} \in \mathbf{\Delta}_{\rm u},\ \Delta^{\rm r} \in \mathbf{\Delta}^{\rm r}, \text{ and } \Delta_{\rm p} \in \mathbb{C}^{m_2 \times \ell_2}, \|\Delta_{\rm p}\|_\infty \le 1\}$ and $\bar{M}(s)$ is given as (4.8).

Proof. Applying the main-loop theorem [308] to the matrix transfer function $\bar{M}(s)$ given in (4.8) completes the proof. QED

4.2.2 Conditions for Reliability using μ

DUS and RDUS The below necessary and sufficient conditions for DUS and RDUS can be tested approximately in polynomial time as a function of the plant dimension.

Corollary 4.1 (*DUS*). Suppose that $K(s)$ is decentralized. Define $\Delta^{\rm r}$ to be a diagonal Δ-block with independent real uncertainties. Then the closed-loop system is DUS if and only if $M(s)$ is internally stable and

$$\mu_{\mathbf{\Delta}^{\rm r}}(M(j\omega)) \le 1, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.10}$$

where $M(s) = -\frac{1}{2}K(s)(\mathrm{I} + \frac{1}{2}P(s)K(s))^{-1}P(s)$.

Proof. Set $\bar{E} = W^{\mathrm{r}} = \frac{1}{2}\mathrm{I}$ in (4.4). QED

Corollary 4.2 (*RDUS*). Suppose that $K(s)$ is decentralized and the uncertain system is described by $P(s)$ and Δ_{u}, i.e., $P_\Delta := \mathcal{F}_u(P, \Delta_{\mathrm{u}})$. Define Δ^{r} to be a diagonal Δ-block with independent real uncertainties. Then the closed-loop system is RDUS if and only if $M(s)$ is internally stable and

$$\mu_{\mathbf{\Delta}_{\mathrm{a}}}(M(j\omega)) \leq 1, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.11}$$

where $\Delta_{\mathrm{a}} \in \mathbf{\Delta}_{\mathrm{a}} = \{\mathrm{diag}\{\Delta_{\mathrm{u}}, \Delta^{\mathrm{r}}\} : \Delta_{\mathrm{u}} \in \mathbf{\Delta}_{\mathrm{u}},\ \Delta^{\mathrm{r}} \in \mathbf{\Delta}^{\mathrm{r}}\}$ and the transfer function matrix

$$M(s) = \begin{bmatrix} P_{11}(s) - \frac{1}{2}P_{12}(s)K(s)(\mathrm{I} + \frac{1}{2}P_{22}(s)K(s))^{-1}P_{21}(s) & \frac{1}{2}P_{12}(s) - \frac{1}{4}P_{12}(s)K(s)(\mathrm{I} + \frac{1}{2}P_{22}(s)K(s))^{-1}P_{22}(s) \\ -K(s)(\mathrm{I} + \frac{1}{2}P_{22}(s)K(s))^{-1}P_{21}(s) & -\frac{1}{2}K(s)(\mathrm{I} + \frac{1}{2}P_{22}(s)K(s))^{-1}P_{22}(s) \end{bmatrix}. \tag{4.12}$$

Proof. Set $\bar{E} = W^{\mathrm{r}} = \frac{1}{2}\mathrm{I}$ in (4.8). QED

CDUS and RCDUS When $K(s)$ is stable, a necessary and sufficient test for CDUS is given by Corollary 4.1, except with the condition $\mu < 1$ replacing $\mu \leq 1$ in (4.10). When $K(s)$ includes integral action in all channels, μ in (4.10) will be equal to 1 at $\omega = 0$, because setting the proportional gain to zero in a controller with integral action will remove the feedback around the integrators, which will then be a limit of instability. Thus, $\mu \leq 1$ in (4.10) is a tight necessary condition for CDUS. The following simple example shows that $\mu \leq 1$ is not sufficient for CDUS:

Example 4.1. Consider the following plant and controller:

$$P(s) = \frac{1}{s+1}\begin{bmatrix} s & -1 \\ 1 & 1 \end{bmatrix}, \quad K(s) = \frac{1}{s}\mathrm{I}.$$

It can be shown by using the Routh criterion that this system is DUS and $\mu \leq 1$. Loop #1 is not stable (for any ϵ_1) when Loop #2 is open (due to a pole-zero cancellation at $s = 0$), and so the system does not possess integrity and is not CDUS.

The following more involved example illustrates that a system can possess integrity and be DUS without being CDUS.

Example 4.2. Consider the plant and controller:

$$P(s) = \frac{1}{s+4}\begin{bmatrix} \frac{\gamma(s^2+s+10)}{s+\alpha} & 1 \\ 1 & 1 \end{bmatrix}, \quad K(s) = \frac{1}{s}\mathrm{I},$$

where $\gamma = (4 - \sqrt{55} - 256)/9$ and $\alpha = (62 - 8\sqrt{55})/9$. It can be shown by using the Routh criterion that this system is DUS and $\mu \leq 1$. It can also be shown that the first loop is not stable for $\epsilon_1 = 1/2$ and $\epsilon_2 = 0$ though it is stable for all other $\epsilon_i \in [0, 1]$.

CDUS can be checked through a finite number of stability and μ tests, by using Corollary 4.1 to check the interior of the ϵ-hypercube, and testing the boundary (the points, edges, faces, etc.) through additional μ tests. The number of μ tests required grows rapidly with the number of actuators/sensors in the system. Though the above examples show that CDUS is not equivalent to DUS, the set of plants that are DUS but not CDUS is non-generic, i.e., any perturbation in such a plant will likely cause the plant to either become DUS or not be DUS. Since Corollary 4.1 provides an exact condition for DUS, finding computable exact conditions for CDUS is of diminished importance. A similar discussion applies for RDUS vs. CRDUS.

4.2.3 Sufficient Conditions for Robust Reliability of Decentralized Integral Control using μ

Now consider a special case of decentralized control in which there exists integral control action in each control loop. Its integrity is defined as follows.

Definition 4.5 (*Definition 14.2-2 in [185]*). The system $L(s) = P(s)C(s)$ is *integral controllable* (IC) if there exists a $k > 0$ such that (a) the closed-loop system in Figure 4.5 is stable for $K = k\mathrm{I}$ and (b) the gains of the loops can be reduced to $K_\epsilon = \epsilon k\mathrm{I}$, $\epsilon \in (0, 1]$ without affecting the closed-loop stability.

In decentralized integral control, the integral controllability of the closed-loop system can be related to the eigenvalues of the open-loop steady-state gain matrix.

Theorem 4.5 (*Theorems 14.3-2 in [185] or [184]*). Suppose that $K \in \mathbb{R}^{m \times m}$ is a diagonal constant gain matrix with positive entries, i.e., $K = \mathrm{diag}\{k_i\}$, $k_i > 0$, $i = 1, \ldots, m$ and $\Delta_\mathrm{u} = 0$ such that $P_\Delta(s) = P_{22}(s)$ that is the lower right block transfer function of $P(s)$. The closed-loop system is IC if the steady-state gain matrix $L(0) = P_{22}(0)C(0)$ is anti-Hurwitz, i.e., $\sigma(L(0)) \subset \mathbb{C}_+$.

A natural extension of integral controllability to uncertain systems can be defined as follows.

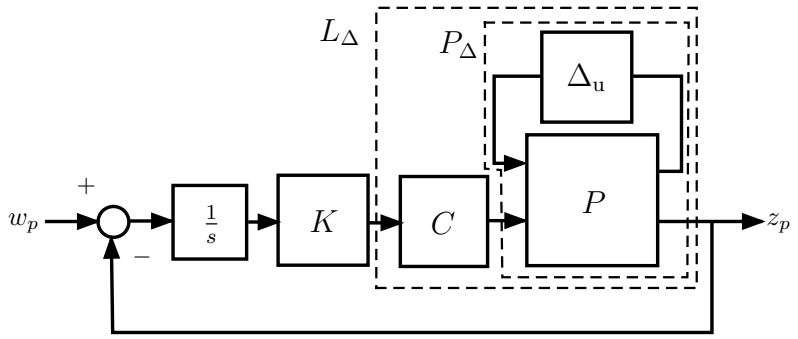


Figure 4.5: Closed-loop uncertain system with integrator and diagonal compensator.

Definition 4.6. The system $L_\Delta(s) = P_\Delta(s)C(s)$ is *robust integral controllable* (RIC) if there exists a $k > 0$ such that, for any $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$, (a) the closed-loop system shown in Figure 4.5 is stable for $K = k{\rm I}$ and (b) the gains of the loops can be reduced to $K_\epsilon = \epsilon k{\rm I}$, $\epsilon \in (0, 1]$ without affecting the closed-loop stability.

Similar to the integral controllability, the robust integral controllability of the closed-loop system can be related to the eigenvalues of the open-loop steady-state gain matrix of which robustness is required.

Corollary 4.3. Suppose that $K \in \mathbb{R}^{m\times m}$ is a diagonal constant gain matrix with positive entries, i.e., $K = \text{diag}\{k_i\}$, $k_i > 0$, $i = 1, \ldots, m$, and the uncertainty $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$. The closed-loop system in Figure 4.5 is RIC if the steady-state gain matrix $L_\Delta(0)$ is anti-Hurwitz for all $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$, i.e., $\sigma(L_\Delta(0)) \subset \mathbb{C}_+$ for all $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$.

The proof of Corollary 4.3 follows from the application of Theorem 4.5 to each plant in the set of uncertain plants. The next result is a sufficient condition for RIC in terms of μ.

Theorem 4.6. Suppose that $K \in \mathbb{R}^{m\times \ell}$ is a diagonal constant gain matrix with positive entries, i.e., $K = \text{diag}\{k_i\}$, $k_i > 0$, $i = 1, \ldots, m$, and the uncertainty $\Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}$. The closed-loop system in Figure 4.5 is RIC if

$$\mu_{\mathbf{\Delta}_{\rm u}^0}(M(j\omega)) < \left(\sup_{\Delta\in\mathbf{\Delta}_{\rm u}^0} \bar\sigma(\Delta) \right)^{-1}, \quad \forall \omega \in \mathbb{R} \cup \{\infty\}, \tag{4.13}$$

where $\mathbf{\Delta}_{\rm u}^0 \triangleq \{\Delta_{\rm u}(0) : \Delta_{\rm u} \in \mathbf{\Delta}_{\rm u}\}$ and

$$M(s) \triangleq \mathcal{F}_u \left(\begin{bmatrix} -P_{11}(0)C(0) & -P_{12}(0) \\ P_{21}(0)C(0) & P_{22}(0) \end{bmatrix}, \frac{1}{s}{\rm I} \right).$$

Proof. The steady-state gain matrix $L_\Delta(0)$ is anti-Hurwitz if and only if the linear system $\dot x = -L_\Delta(0)x$ is globally asymptotically stable (g.a.s.) (or equivalently, globally exponentially stable (g.e.s.)). Furthermore, $L_\Delta(0)$ can be rewritten as

$$\mathcal{F}_\ell \left(\begin{bmatrix} P_{11}(0)C(0) & P_{12}(0) \\ P_{21}(0)C(0) & P_{22}(0) \end{bmatrix}, \Delta_{\rm u}(0) \right).$$

Now, for each $\Delta_u(0) \in \mathbf{\Delta}_{\rm u}^0$, $\dot x = -L_\Delta(0)x$ is g.a.s. if and only if $\det({\rm I} + M(s)\Delta_{\rm u}(0)) \neq 0$ for all $s \in \mathbb{C}_+$. From the subharmonic property of μ and the homotopy condition on the uncertainty set $\mathbf{\Delta}_{\rm u}^0$ (i.e., $\Delta_u(0) \in \mathbf{\Delta}_{\rm u}^0$ implies that $\tau\Delta_u(0) \in \mathbf{\Delta}_{\rm u}^0$ for any $\tau \in [0, 1]$), the determinant condition can be reduced to the frequency-domain condition on μ in (4.13). QED

4.2.4 Remarks on Decentralized Detunability

Detuning a controller refers to changing some parameter in the controller or in the control synthesis procedure so that the control action becomes less aggressive. For example, in linear quadratic (LQ) optimal control,

detuning refers to increasing the magnitude of the weight of control action in the quadratic cost function—exactly opposite of *cheap control* in which control weights are very small [237]. In decentralized internal model control (IMC), detuning refers to increasing the IMC filter time constants (or equivalently, decreasing the bandwidth of the IMC filter) in each single-loop controller [112, 113]. The special case of detuning the single-loop controller gains in a decentralized controller was discussed earlier in the sections on DUS and RDUS.

Hovd [112] introduced the following very general definition for robust decentralized detunability.

Definition 4.7. For a given design method, a closed-loop system is robust decentralized detunable (RDD) if each single-loop controller can be detunable independently by an arbitrary amount without losing robust stability in the closed-loop system.

Whenever a controller is detuned by varying a parameter in the controller, RDD can be tested via a μ test where the variation in parameters is covered by real uncertainty (the real uncertainty must be independent for arbitrary detuning). Both the robust performance and the RDD loopshaping bounds are plotted and the most restrictive of the bounds are used in the design. The resulting controller meets robust performance and gives a system that is RDD. This loopshaping design procedure is illustrated in Braatz [39], where interested readers can go for details and examples.

4.3 Further Remarks

4.3.1 Review of Previous Research with Illustrative Examples

Integrity Most research on reliability analysis considers only system integrity without considering plant-model mismatch [70, 91, 184, 185]. Controller-independent conditions that can establish necessary and sufficient conditions for the existence or non-existence of a controller such that the system possesses integrity have been derived [102, 131], but these conditions are also only applicable to perfectly known LTI systems.

Fujita and Shimemura [91] state that a necessary and sufficient condition for integrity with stable controllers is that all the principal minors of $\mathrm{I} + PK$ are minimum phase. This condition is theoretically interesting, because this test does not require the calculation of matrix inverses. However, since the number of principal minors of matrix grows exponentially with its dimension, the calculation required by this test grows exponentially as a function of the plant dimension. Fujita and Shimemura [91] also provide a sufficient condition for integrity when the controller is stable, in terms of the generalized diagonal dominance of $\mathrm{I} + P(j\omega)K(j\omega)$. Applying the Perron-Frobenius Theorem [109] gives the following lemma (for details, see Delich [70]).

Lemma 4.1. Assume $P(s)$ and $K(s)$ are stable, the diagonal elements of $\mathrm{I} + P(s)K(s)$ are minimum phase,

and $P(s)$ is irreducible. Then the closed-loop system demonstrates integrity if

$$\rho\left(\left|H(j\omega)\left(\bar{H}(j\omega)\right)^{-1}\right|\right) < 2, \quad \forall\omega \in \mathbb{R} \cup \{\infty\}, \tag{4.14}$$

where $H = \mathrm{I} + PK$, $\bar{H}$ refers to the matrix with all off-diagonal elements of H replaced by zeros, and $|A|$ denotes the matrix with each element of A replaced by its magnitude.

The above assumption that P is irreducible can be removed with some added complexity in the theorem statement [70]. The spectral radius is readily computable with polynomial growth ($\sim n^3$) as a function of the plant dimension. However, the lemma might be conservative as shown in the following example.

Example 4.3. Consider the closed-loop system with the plant and controller:

$$P(s) = \frac{1}{75s+1}\begin{bmatrix} -0.878 & 0.014 \\ -1.082 & -0.014 \end{bmatrix};\ K(s) = \frac{75s+1}{\lambda s+1}\begin{bmatrix} -\frac{1}{0.878} & 0 \\ 0 & -\frac{1}{0.014} \end{bmatrix};\ \lambda = 4.$$

The system demonstrates integrity but the condition in (4.14) is not satisfied for this system ($\rho \approx 2.1 < 2$), which indicates that the test (4.14) can be conservative, even for 2×2 systems.

Robust Integrity Laughlin et al. [155] provide computationally simple tests for robust integrity that are useful for cross-directional processes (see [273] for a review of cross-directional process control problems). Their results do not extend to general plants and so are not further discussed here.

Decentralized Unconditional Stability Morari [184] considers stability with simultaneous detuning of all loops, which leads to a number of computationally simple necessary conditions for DUS that are surveyed in the monograph by Morari and Zafiriou [185]. However, all these conditions can be conservative for testing DUS, as illustrated by numerous examples in that monograph.

CDUS Nwokah and Perez [204] considered conditions for which a system with controller $K(s) = \frac{1}{s}\mathrm{I}$ is CDUS, including the claim that a necessary condition for $K(s) = \frac{1}{s}\mathrm{I}$ to provide CDUS is that the steady-state matrix $P(0)$ is *all gain positive stable*. A matrix P is all gain positive stable if P, P^{-1}, and all their corresponding principal submatrices are D-stable. A matrix P is D-stable if $\sigma(PD) \subset \mathbb{C}_+$ for all positive diagonal matrices D. Example 4.4 shows that the condition in [204] is not necessary.

Example 4.4. Consider the plant [52]:

$$P(s) = \begin{bmatrix} 1 & 0 & 2 \\ \frac{1}{s+1} & 1 & \frac{-4s}{s+1} \\ 0 & 4 & 1 \end{bmatrix}.$$

It can be shown that the Routh-Hurwitz stability criteria that the closed-loop system for the above plant is stable for $K(s) = \frac{1}{s}\mathrm{I}$ and remains stable with arbitrary detuning of the SISO loop gains. But, $\sigma(P(0)) = \{\pm i\sqrt{3}, 3\}$, so $P(0)$ is not D-stable, and $P(0)$ is not all gain positive stable. We note here without details that this plant also shows that all of the theorems in [204] regarding decentralized integral controllability are also not necessary.

RDUS and RCDUS To our knowledge, it seems that RDUS and RCDUS have not been considered in the open literature, except for a thesis [39] and the proceedings paper [41] that contains some of the results of this manuscript. Note that it is previously shown that conditions for RDUS and RCDUS can be represented as evaluating the SSV of the associated transfer function. Upper and lower bounds on μ are computable in polynomial-time [8], even though its exact computation is NP-hard [44].

4.3.2 Related Topics

Fault Detection and Diagnosis For systems affected by time-varying parametric uncertainties and time-varying detuned gain of decentralized controllers, it might be natural to discuss the design of linear parametrically varying (LPV) controllers or gain-scheduled controllers when the time-varying parameters are not known *a priori*, but are online measurable. In that control framework, faults in the actuators and/or sensors can be detected and LPV control laws give a natural way to remedy those faults.

Reliable Networked Control Systems In a networked control system, most communication links introduce variable and unpredictable time delays in the information flow, which are called *network-induced delays* [304]. This application problem has motivated the analysis of the effects of time delays among interconnecting elements of a decentralized or distributed network control system on the closed-loop system stability and performance. The problem formulation and conditions for robust reliability analysis of decentralized control systems can be extended to the robust stability and performance analysis of networked control systems under intermittent communication losses between distributed sensors and actuators.

4.4 Illustrative Examples: Fault-tolerant Decentralized Control

High-purity Distillation Column We now illustrate the investigation of robust stability and performance of a decentralized controller for the high-purity distillation column under fault/failure scenarios. A high-purity distillation column is given in [247] and discussed in more detail in [248]. The nominal model is

$$P_{\mathrm{n}}(s) = \frac{1}{75s+1}\begin{bmatrix} -0.878 & 0.014 \\ -1.082 & -0.014 \end{bmatrix},$$

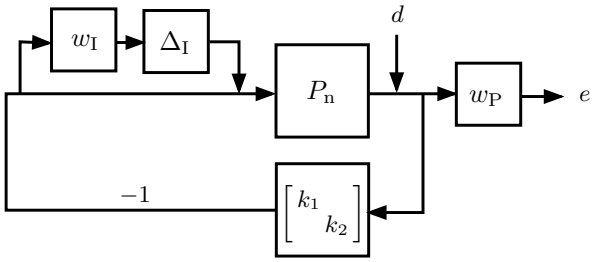


Figure 4.6: The plant with input uncertainty Δ_{I} of magnitude $w_{\mathrm{I}}(s)$ and the performance specification $w_{\mathrm{P}}(s)$.

which uses distillate and boilup as manipulated inputs to control top and bottom composition using measurements of the top and bottom compositions. The plant has a large condition number, so input uncertainty strongly affects robust performance [248]. The uncertainty and performance weights are

$$w_{\mathrm{I}}(s) = 0.1\frac{5s+1}{0.25s+1}, \quad w_{\mathrm{P}}(s) = 0.025\frac{7s+1}{7s}.$$

The input uncertainty includes actuator uncertainty and neglected right-half plane zeros of the plant. The performance bound implies zero steady-state error and a closed-loop time constant of 7 minutes. The uncertainty block Δ_{I} is a diagonal 2×2 matrix (independent actuators) and the performance block Δ_{P} is a full 2×2 matrix.

In [39], loopshaping bounds are used to design the decentralized controller

$$K(s) = \frac{75s+1}{4s}\begin{bmatrix} -\frac{1}{0.878} & 0 \\ 0 & -\frac{1}{0.014} \end{bmatrix}.$$

We will now analyze the closed-loop system with the designed controller to show that it satisfies integrity, robust integrity, DUS, and RDUS.

Integrity The following four transfer functions are stable:

$$\begin{aligned}
(\epsilon_1, \epsilon_2) = (0,0) &\Rightarrow P_{\mathrm{n}}, \\
(\epsilon_1, \epsilon_2) = (1,1) &\Rightarrow -w_{\mathrm{I}}K(\mathrm{I} + P_{\mathrm{n}}K)^{-1}P_{\mathrm{n}}, \\
(\epsilon_1, \epsilon_2) = (1,0) &\Rightarrow -w_{\mathrm{I}}K_1(\mathrm{I} + P_{\mathrm{n},11}K_1)^{-1}P_{\mathrm{n},11}, \\
(\epsilon_1, \epsilon_2) = (0,1) &\Rightarrow -w_{\mathrm{I}}K_2(\mathrm{I} + P_{\mathrm{n},22}K_2)^{-1}P_{\mathrm{n},22},
\end{aligned}$$

so the closed-loop system has integrity.

Robust integrity Robust integrity for a 2×2 system can be evaluated by checking the robust stability for four conditions. Nominal stability was tested above (for testing integrity), so only the μ conditions are tested here. The system has robust stability when all loops are turned off provided that $P_{\mathrm{n}}(\mathrm{I} + w_{\mathrm{I}}\Delta_{\mathrm{I}})$ is

stable. Since $P_{\rm n}$, $w_{\rm I}$, and $\Delta_{\rm I}$ are stable, $P_{\rm n}(\mathrm{I}+w_{\rm I}\Delta_{\rm I})$ is stable. Robust stability for the overall system is satisfied since $\mu_{\mathbf{\Delta}_{\rm I}}(-w_{\rm I}K(\mathrm{I}+P_{\rm n}K)^{-1}P_{\rm n}) = 0.3 < 1$. Robust stability for the cases when exactly one loop has failed is satisfied since

$$(\epsilon_1,\epsilon_2)=(1,0)\Rightarrow \mu_{\mathbf{\Delta}_{\mathrm{I},\mathbf{11}}}(-w_{\rm I}K_1(\mathrm{I}+P_{\mathrm{n},11}K_1)^{-1}P_{\mathrm{n},11}) = 0.12 < 1;$$
$$(\epsilon_1,\epsilon_2)=(0,1)\Rightarrow \mu_{\mathbf{\Delta}_{\mathrm{I},\mathbf{22}}}(-w_{\rm I}K_2(\mathrm{I}+P_{\mathrm{n},22}K_1)^{-1}P_{\mathrm{n},22}) = 0.12 < 1.$$

Since all four μ conditions are satisfied, the system demonstrates robust integrity.

DUS and RDUS First let's test RDUS. The transfer function matrices P, G, $\bar{G}$, and $\Delta_{\rm a}$ needed to apply Theorems. 4.3 and 4.4 are derived directly from the block diagram in Figure 4.6:

$$P=\begin{bmatrix}0 & -w_{\rm I}\mathrm{I}\\ P_{\rm n} & -P_{\rm n}\end{bmatrix},\quad G=\begin{bmatrix}0 & 0 & -w_{\rm I}\mathrm{I}\\ w_{\rm P}P_{\rm n} & 0 & -w_{\rm P}P_{\rm n}\\ -P_{\rm n} & \mathrm{I} & P_{\rm n}\end{bmatrix}, \bar{G}=\begin{bmatrix}0 & -w_{\rm I}W^{\rm r} & 0 & -w_{\rm I}\bar{E}\\ 0 & 0 & 0 & \mathrm{I}\\ w_{\rm P}P_{\rm n} & -w_{\rm P}P_{\rm n}W^{\rm r} & 0 & -w_{\rm P}P_{\rm n}\bar{E}\\ -P_{\rm n} & P_{\rm n}W^{\rm r} & \mathrm{I} & P_{\rm n}\bar{E}\end{bmatrix}, \tag{4.15}$$

$\Delta_{\rm a}=\mathrm{diag}\{\Delta_{\rm I},\Delta^{\rm r}\}$.

Figure 4.7(a) is the μ plot to test condition (4.7) in Theorem 4.3 for evaluating RDUS. As expected, the value of μ approaches 1 at zero frequency due to the integrators as either of the ϵ_i approach zero. We see that $\mu \ll 1$ for all frequencies away from $\omega = 0$. Since $\mu \leq 1$, the system demonstrates RDUS. Since DUS is implied by RDUS, DUS does not need to be numerically tested for this example.

Robust performance under arbitrary detuned control gains ($0 \sim 100\%$ of the nominal value) can also be studied using the condition (4.9) in Theorem 4.4. Consider the analysis of robust performance under a reduced performance defined by $w_{\rm P}(s) = 0.025\frac{7s+1}{7s}$, which is 1/10 of the nominal performance specified by [248] when there is no faults in the system. Figure 4.7(b) shows that $\mu_{\mathbf{\Delta}}(j\omega) \leq 1$ for all frequency, so the reduced level of robust performance is achieved for the specified range of detuned control gains.

Parallel Reactors with Combined Precooling Now consider the robust stability and performance of an SVD optimal controller for a parallel reactor with combined precooling. In [114], a simplified model of four parallel reactors with combined precooling is

$$G(s)=\frac{1}{100s+1}\begin{bmatrix}1 & 0.7 & 0.7 & 0.7\\ 0.7 & 1 & 0.7 & 0.7\\ 0.7 & 0.7 & 1 & 0.7\\ 0.7 & 0.7 & 0.7 & 1\end{bmatrix}.$$

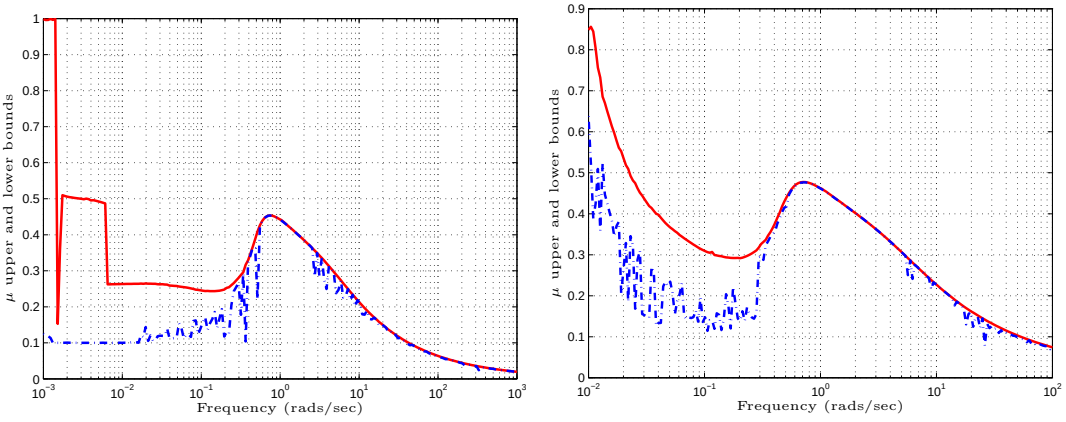
Consider the input and output uncertainty in the system shown in Figure 4.8. The input uncertainty $\Delta_{\rm I}$ and output uncertainty $\Delta_{\rm O}$ are assumed to have independent diagonal and uncertainty weights given by

$w_{\rm I} := 0.125\frac{5s+1}{0.5s+1}{\rm I}$ and $w_{\rm O} := 0.125\frac{2.5s+1}{0.25s+1}{\rm I}$, respectively. To reject disturbances at the system output, the weighted performance specification is $\|w_{\rm P}S_{\rm p}\|_\infty < 1$, where $S_{\rm p}$ is the transfer function mapping d to y, with the performance weight $w_{\rm P}(s) := 0.125\frac{125s+1}{125s}{\rm I}$. An SVD optimal controller was designed using DK-iteration and reported in [114]. This example considers the reliability of this controller design to 80% independent detuning of the controller gains. The performance weight models partially degraded performance, compared to the case when there is no fault or failure of controllers in [114].

Robust reliability The transfer function matrices P, G, $\bar{G}$, and $\Delta_{\rm a}$ needed to apply Theorems. 4.3 and 4.4 are derived directly from the block diagram in Figure 4.8:

$$P = \begin{bmatrix} 0 & 0 & -w_{\rm I}{\rm I} \\ w_{\rm O}P_{\rm n} & 0 & -w_{\rm O}P_{\rm n} \\ P_{\rm n} & {\rm I} & -P_{\rm n} \end{bmatrix},\ G = \begin{bmatrix} 0 & 0 & 0 & -w_{\rm I}{\rm I} \\ w_{\rm O}P_{\rm n} & 0 & 0 & -w_{\rm O}P_{\rm n} \\ w_{\rm P}P_{\rm n} & w_{\rm P}{\rm I} & 0 & -w_{\rm P}P_{\rm n} \\ -P_{\rm n} & -{\rm I} & {\rm I} & P_{\rm n} \end{bmatrix},$$
$$\bar{G} = \begin{bmatrix} 0 & 0 & -w_{\rm I}W^{\rm r} & 0 & -w_{\rm I}\bar{E} \\ w_{\rm O}P_{\rm n} & 0 & w_{\rm O}P_{\rm n}W^{\rm r} & 0 & -w_{\rm O}P_{\rm n}\bar{E} \\ 0 & 0 & 0 & 0 & {\rm I} \\ w_{\rm P}P_{\rm n} & w_{\rm P}{\rm I} & w_{\rm P}P_{\rm n}W^{\rm r} & 0 & -w_{\rm P}P_{\rm n}\bar{E} \\ -P_{\rm n} & -{\rm I} & -P_{\rm n}W^{\rm r} & {\rm I} & P_{\rm n}\bar{E} \end{bmatrix},\ \Delta_{\rm a} = {\rm diag}\{\Delta_{\rm I}, \Delta_{\rm O}, \Delta^{\rm r}\}. \tag{4.16}$$

To assess whether the closed-loop uncertain system remains stable with up to 80% independent detuning of the actuator/sensor/controller gains, set $\epsilon_i \in [0.2, 1]$ for all i and $W^{\rm r} = 0.4{\rm I}$ and $\bar{E} = 0.6{\rm I}$. The μ plot in Figure 4.9(a) to test condition (4.7) in Theorem 4.3 shows that $\mu(j\omega) < 1$ for all frequencies, which implies



(a) μ plot for evaluation of RDUS (b) μ plot for evaluation of robust performance

Figure 4.7: μ plots for evaluating reliability to uncertainties and reduction of actuator/sensor/controller gains for the high-purity distillation column. The smooth red curve is the upper bound for μ and the rough blue curve is its lower bound.

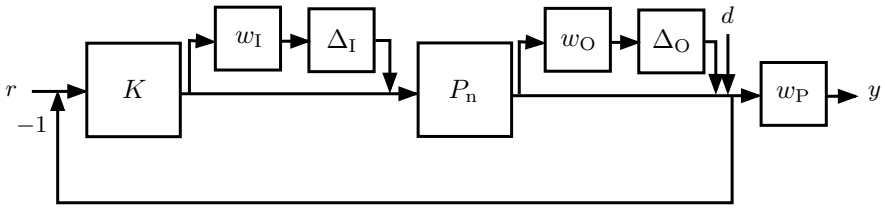
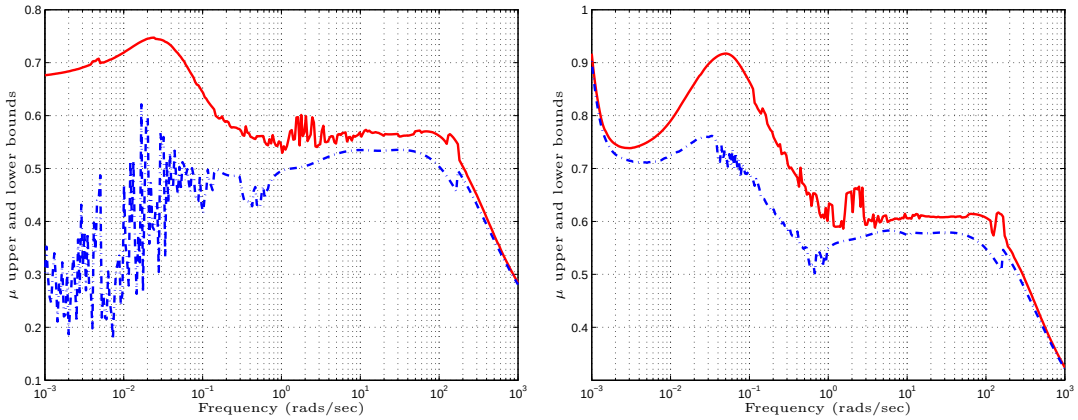


Figure 4.8: The plant with input and output uncertainties Δ_{I} and Δ_{O} of magnitude $w_{\mathrm{I}}(s)$ and $w_{\mathrm{O}}(s)$, and the performance specification $w_{\mathrm{P}}(s)$.



(a) μ plot for robust stability

(b) μ plot for robust performance

Figure 4.9: μ plots for evaluating reliability to uncertainties and up to 80% independent detuning of controller gains for the parallel reactors with combined precooling. The smooth red curve is the upper bound for μ and the rough blue curve is its lower bound.

that the system is robust to this degree of control detuning and to model uncertainties. This μ plot also implies that the nominal system is reliable to 80% independent detuning of the actuator/sensor/controller gains.

Robust performance under detuned control gains can be studied using the condition (4.9) in Theorem 4.4. Figure 4.9(b) shows that $\mu_{\mathbf{\Delta}}(j\omega) \leq 1$ for all frequencies, so robust performance is achieved with up to 80% independent detuning of the controller gains. Various degrees of degraded closed-loop performance could be defined for different degrees of detuning, by plotting a different μ plot for each performance weight and range of detuning.

4.5 Summary and Future Work

Robust reliability of closed-loop systems is an important issue in control systems engineering and for large-scale interconnected systems. This chapter considers the analysis of the reliability of controlled systems with and without model uncertainties. Necessary and sufficient conditions for robust fault tolerant stability and performance under constant but unknown gain variation are derived for uncertain systems that are

affected by real parametric and complex dynamic uncertainties. The proposed conditions are represented in terms of the structured singular value and are nonconservative in the sense that locations and structures of potential faults and failures can be fully exploited, and structured plant-model mismatches are considered to derive necessary and sufficient conditions for system reliability. Upper- and lower-bounds on μ can be computed in polynomial-time by using off-the-shelf software [8] and provide computationally tractable tools for verifying reliability of the controllers. Numerical case studies for high-purity distillation column and parallel reactors with combined precooling are presented for illustration of the application of the proposed reliability conditions.

Chapter 5

Robust Nonlinear Internal Model Control of Wiener Systems

Abstract Many process systems can be modeled as a stable Wiener system, which is a stable linear system followed by a static nonlinearity. A nonlinear control design procedure is presented that provides robustness to uncertainties while being applicable to systems with unstable zero dynamics, unmeasured states, disturbances, and measurement noise. The design procedure combines nonlinear internal model control with linear matrix inequality feasibility or optimization problems, such that all robust stability and performance criteria are computable in polynomial-time using readily available software. Application to a pH neutralization case study demonstrates the importance of taking uncertainty into account during the design of controllers for Wiener systems. The approach is generalizable to Hammerstein and sandwich systems, whether well- or poorly-conditioned, and to systems with actuator constraints.

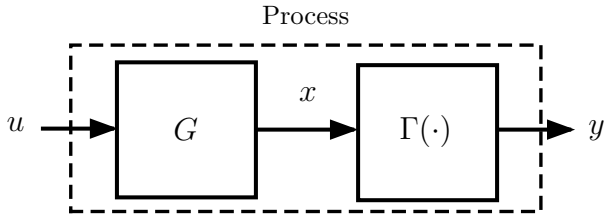


Figure 5.1: Wiener model structure where G is linear time-invariant and $\Gamma(\cdot)$ is a static (nonlinear) operator that is included in a certain family of nonlinear functions.

5.1 Introduction

A Wiener model consists of a linear dynamic system G followed by a static nonlinear operator Γ (see Figure 5.1). Many process systems have been described by Wiener models including distillation columns [76], heat exchangers [76], pH neutralization [98], packed bed reactors [77], and plasma reactors [271, 272]. Various control strategies have been developed for Wiener systems, including adaptive control [210], linearizing feedforward-feedback control [128], and model predictive control [203, 212] strategies, many of which have been evaluated by application to pH neutralization processes.

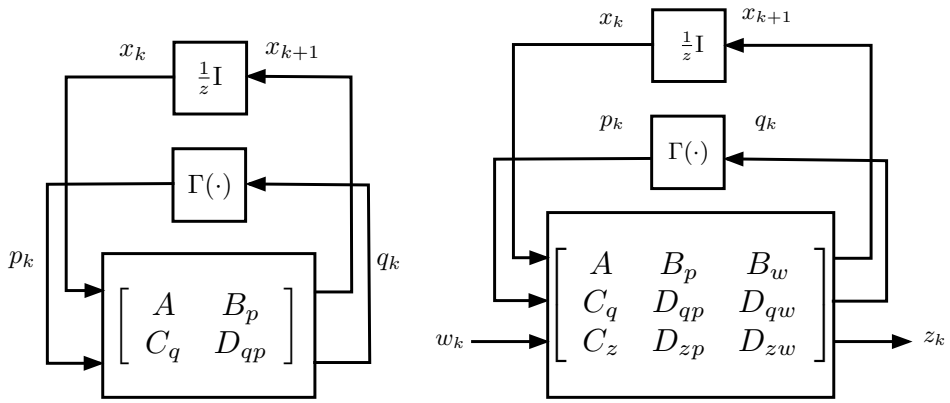
The static nonlinearity in a Wiener model for a practical application is always uncertain, and most existing methods for the control of Wiener systems ignore this uncertainty. As will be demonstrated in a case study later in this chapter, the uncertainty in the nonlinearity can have major effects on the closed-loop stability and performance. While robust optimal controllers for Wiener systems have been designed by formulating the optimal control problem in terms of bilinear matrix inequalities (BMIs) [271,272], a drawback of such an approach is that optimization over BMIs is an NP-hard problem [165,266]. This section proposes an approach that only involves convex programs that can be solved using off-the-shelf software in polynomial-time. The proposed approach is applicable to stable Wiener systems with unstable zero dynamics, unmeasured states, disturbances, and measurement noise. A novel aspect of the approach is that various characteristics of the nonlinearities can be taken explicitly into account in the closed-loop stability and performance analyses. The approach is applied to a case study involving the control of pH in which the Wiener model is identified from experimental data. The control of pH is an important industrial problem that has been extensively studied [12, 78, 123, 239, 292].

5.2 Theory and Methods: Stability and Performance Criteria

5.2.1 Problem Statement

Standard Nonlinear Operator Form The proposed approach employs the standard nonlinear operator form (SNOF) in Figure 5.2, which has its roots in the 1940s Russian control literature [85, 164, 179]. The SNOF consists of a linear system with a static nonlinear operator in feedback, where the static nonlinearities of the operator can be further restricted to be diagonal, monotonic, and locally slope-restricted. Nearly any arbitrary nonlinear system (including unstable zero dynamics, chaotic, and quasi-periodic behavior) can be approximated with arbitrary accuracy by a model in standard nonlinear operator form [141]. Furthermore, all dynamic artificial neural networks can be transformed into the SNOF, so that any of the software packages available for fitting DANNs to experimental data[1] produces models that can be written in SNOF. The static, monotonic, and locally slope-restricted nature of the nonlinearities can be exploited to produce polynomial-

[1] e.g., the Matlab Neural Network toolbox



(a) Stability Analysis (b) Performance Analysis

Figure 5.2: Standard nonlinear operator form for discrete-time systems. The structure for continuous-time systems is obtained by replacing z with s, replacing x_{k+1} with dx/dt, and redefining the other variables to be continuous-time.

time tools to analyze the stability and performance of these systems (e.g., see [139,140] and references cited therein). The analysis tools can be written in terms of linear matrix inequalities (LMIs) [35], which are computable using available software (e.g., [160, 255]), much of which can be run in Matlab. The proposed nonlinear control design stategy, described below, weds the above analysis tools with internal model control.

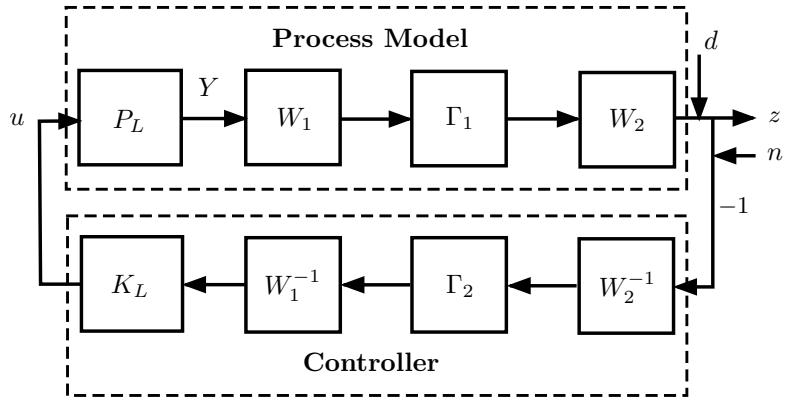


Figure 5.3: Block diagram for the nonlinear closed-loop system, with output z, disturbance d, and noise n. The linear time-invariant and nonlinear static monotonic operators of the process are P_L (assumed to be stable) and Γ_1, respectively. The linear and nonlinear internal model controllers have $\Gamma_2 = 1$ and $\Gamma_2 \approx \Gamma_1^{-1}$, respectively.

Nonlinear Internal Model Control Strategy As is standard in inversion-based control strategies for the control of Wiener systems (e.g., see [257] and citations therein), the control structure has the form in Figure 5.3, where Γ_2 is selected to be either the identity or the inverse of the process nonlinearity Γ_1, depending on whether the overall controller is desired to be linear or nonlinear. For nonlinear Γ_2, the

controller in Figure 5.3 has the Hammerstein structure, in which a dynamic linear controller K_L is augmented with the inverse of the nonlinear operator combined with the identified parameters of the process W_1 and W_2.

The remaining task is to determine the linear time-invariant controller K_L based on some closed-loop criteria. For example, an example of an $\mathcal{L}_2$-optimal control problem would be to determine the K_L that minimizes a weighted combination of the worst-case effects of the disturbance d and noise n on the output z:

$$\begin{aligned} \textbf{Problem}: \quad \inf_{K_L} \; & \alpha \left(\sup_{\|d\|_2 \leq 1} \|z\|_2 \right) + (1-\alpha) \left(\sup_{\|n\|_2 \leq 1} \|z\|_2 \right) \\ \text{s.t.} \; & \text{the closed-loop system in Figure 5.3 is stable,} \end{aligned} \tag{5.1}$$

where the weight $\alpha \in [0,1]$. For a linear time-invariant process ($\Gamma_1 = \mathrm{I}$), it is well known that the above optimal control problem is equivalent to a weighted $\mathcal{H}_\infty$-control objective, $\|w_u S + w_p T\|_\infty$, where S is the sensitivity function that maps the output disturbance d to the controlled output z, T is the complementary sensitivity function that maps the measurement noise n to z, and w_u and w_p are weights that define the tradeoff between disturbance suppression and insensitivity to measurement noise (e.g., [185, 309]).

While there are several approaches available for solving $\mathcal{L}_2$-optimal control problems for linear time-invariant systems, solving such problems for nonlinear systems is much more challenging [13] and would require extensive software generation, even in the case where there is no uncertainty in the system. The nonlinearity Γ has associated uncertainty, so that the nonlinear inversion introduces nonlinear uncertainty into the closed-loop system. Rigorously taking that uncertainty into account while solving (5.1) results in a nonconvex optimization over bilinear matrix inequality constraints [271, 272]. It is straightforward to show that optimizations over bilinear matrix inequality constraints are NP-hard,[2] either by reduction to the knapsack problem [165, 266] or to an indefinite quadratic program [44]. An alternative approach is to parameterize K_L in Figure 5.3 in terms of the well-known Youla parameterization (e.g., [185, 309]), $K_L = \frac{Q}{1-P_L Q}$, where the stable transfer function Q provides degrees of freedom for controller design. The engineer can parameterize Q in any way that maintains stability, and then tune the control parameters to optimize the objective (5.1).

Internal model control (IMC) restricts the degrees of freedom in Q so that the control tuning parameters are few and have a direct relationship to setpoint tracking response, disturbance suppression, insensitivity to measurement noise, and robustness to model uncertainties (e.g., [146, 185]). The form for Q is selected as a low-pass filter F in series with the inverse of a minimum-phase approximation of the linear model of the stable process being controlled. For the notation used here, $Q = P_{L,\mathrm{m}}^{-1}$, where $P_{L,\mathrm{m}}$ is the minimum-phase approximation of the stable transfer function P_L.

Instead of directly solving an optimization such as (5.1) for the control tuning parameters, a common

[2] except for very specialized matrix structures

alternative is to tune the controller as fast as possible while satisfying all of the control objectives, such as guaranteed closed-loop stability with respect to all uncertainties with a prescribed set, effectiveness at disturbance suppression, or insensitivity to measurement noise (e.g., [249]). This approach avoids having to explicitly define a performance weight, and avoids having to balance the weights with respect to each other.

The above approaches apply to both continuous- or discrete-time systems; for brevity only the discrete-time equations will be presented below. To parametrize the controller, consider the low-pass filter

$$F(z) = \frac{1}{\left(\lambda \frac{z-1}{z+1} + 1\right)^m} \mathrm{I} \tag{5.2}$$

where λ specifies the response speed of the low-pass filter and m is an integer that defines the order of the transfer function. This form for F is obtained from Tustin's discretization [54] of the low-pass filter $F(s) = \frac{1}{(\lambda s+1)^m}$ in the Laplace domain. The order of the low-pass filter is fixed and the control tuning parameter is the IMC filter time constant $\lambda > 0$. If needed, this filter form can be generalized to include numerator dynamics or different time constants in each diagonal element [24, 108, 185].

The closed-loop system in Figure 5.3 can be rearranged into the SNOF in Figure 5.2 by using block-diagram algebra as described in standard textbooks [185, 249] or by using the *sysic* program in the Matlab Robust Control Toolbox. The next section presents methods to quantify robust stability and performance criteria for the closed-loop system in terms of linear matrix inequalities, which can be computed using off-the-shelf software that have Matlab interfaces (e.g., [255]). These quantifications can be inserted into an optimization formulation for λ using a weighted control objective such as (5.1) or can be used to determine the minimum value for λ that satisfies all of the robust stability and performance criteria.

Remark 5.1. The authors of [75] proposed an augmentation of the nonlinear controller with a linear filter, in which the model inverse was constructed using numerical procedures based on the contraction mapping principle and Newton's method. The same nonlinear IMC structure was later used [103], in which the model inverse was determined using differential geometry. The aforementioned nonlinear control structure is very similar to those used in these and other past publications. As described in the next section, the proposed design method will differ from past works by rigorously taking uncertainties associated with nonlinear inversion into account.

5.2.2 Stability Analysis

This subsection describes a necessary condition and sufficient conditions for the analysis of stability of a system in SNOF. To simplify the notation, (B, C, D) is used as a shorthand notation for (B_p, C_q, D_{qp}).

The following necessary condition for stability of an SNOF is obtained from linearization of a nonlinear process model [136].

Theorem 5.1 (*Necessary Stability Condition*). Consider a nonlinear system in SNOF as shown in Figure 5.2(a):

$$\begin{aligned} x_{k+1} &= Ax_k + Bp_k \\ q_k &= Cx_k + Dp_k \\ p_k &= \Gamma(q_k) \end{aligned} \tag{5.3}$$

where $x \in \mathbb{R}^n$, $p \in \mathbb{R}^h$, $q \in \mathbb{R}^h$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times h}$, $C \in \mathbb{R}^{h\times n}$, $D \in \mathbb{R}^{h\times h}$, Γ is a diagonal nonlinear operator, n is the number of states, and h is the input-output dimension of Γ. A necessary condition for asymptotic stability of the steady-state x_{ss} is that the eigenvalues of the matrix

$$A_L(x_{\text{ss}}) \triangleq A + B\left(I - \left.\frac{\partial\Gamma}{\partial q}\right|_{\text{ss}} D\right)^{-1} \left.\frac{\partial\Gamma}{\partial q}\right|_{\text{ss}} C, \tag{5.4}$$

all have magnitude less than or equal to one, i.e., $\rho(A_L(x_{\text{ss}})) \leq 1$ where $\rho(\cdot)$ denotes the spectral radius of a matrix.

The term $\left.\frac{\partial\Gamma}{\partial q}\right|_{\text{ss}}$ is the Jacobian of Γ evaluated at the steady-state value for the state. For a diagonal $\Gamma =: \text{diag}\{\gamma_i\}$, this Jacobian has a rather simple form, $\left.\frac{\partial\Gamma}{\partial q}\right|_{\text{ss}} = \text{diag}\left\{\left.\frac{\partial\gamma_i}{\partial q_i}\right|_{x=x_{\text{ss}}}\right\}$.

Any of the sufficient conditions for analyzing the stability of systems in SNOF using linear matrix inequalities (e.g., see [35, 180, 270] and citations therein) can be applied to this approach. Which stability condition to apply depends on the assumptions made on the nonlinearities concerning the matrix structure (e.g., full-block, block-diagonal, or diagonal) and the extent of time variation (e.g., arbitrarily fast time-varying, arbitrarily slow time-varying, static). Which condition to use depends on the nature of the specific problem. For example, if the assumption that the process nonlinearity is static was only an approximation during the identification of the Wiener model, then a stability condition can be selected that allows the uncertainty in the nonlinear to be dynamic. If the nonlinearity in the Wiener model is multivariable, then a full-block uncertainty structure should be used. If the nonlinearities in the Wiener model are distinct and isolated, then a diagonal uncertainty structure should be used to reduce conservatism. Few of the published conditions, however, take into account the static, monotonic, and slope-restricted nature of most nonlinearities and such conditions that have been derived either require restrictive assumptions (see [139, 140] for details). The following sufficient condition, which is computable for large-scale systems while taking into account these characteristics of nonlinearities, will be applied in the pH control case study.

Theorem 5.2 (*A Sufficient Stability Condition*[3]). Consider a system described in Figure 5.2(a) with

$$\begin{aligned} x_{k+1} &= Ax_k + Bp_k \\ q_k &= Cx_k + Dp_k, \end{aligned} \tag{5.5}$$

and $p_k = \Gamma(q_k)$ subject to the sector-bounded and slope-restricted conditions

$$\gamma_i(q_{k,i})\left[\gamma_i(q_{k,i}) - \xi_i q_{k,i}\right] \leq 0, \quad \forall q_{k,i} \in \mathbb{R},\, i = 1, \ldots, h \tag{5.6}$$

and

$$0 \leq \frac{\gamma_i(q_{k+1,i}) - \gamma_i(q_{k,i})}{q_{k+1,i} - q_{k,i}} \leq \mu_i, \quad \forall q_{k,i} \in \mathbb{R},\, i = 1, \ldots, h \tag{5.7}$$

where ξ_i and μ_i is the maximum sector bound and slope of the i^{th} nonlinearity, respectively. A sufficient condition for global asymptotic stability is the existence of a positive-semidefinite matrix $P = P^{\mathrm{T}}$ with a positive-definite submatrix $P_{11} = P_{11}^{\mathrm{T}}$ and diagonal positive-semidefinite matrices Q, $\tilde{Q}$, T, $\tilde{T}$, $N \in \mathbb{R}^{n_q \times n_q}$ such that the LMI

$$G \triangleq A_a^T P A_a - E_a^T P E_a + U_1 + U_2 - S_1 - S_2 - S_3 < 0 \tag{5.8}$$

holds, where the matrices A_a, E_a, U_1, U_2, S_1, S_2, and S_3 are defined in (5.23).

Although stated as a stability condition, Theorem 5.2 is also a robust stability condition, in that the existence of a feasible solution to the LMI (5.8) implies that the system is stable for all nonlinearities that satisfy the sector and slope bounds (5.6) and (5.7), respectively. Uncertainties in the parameters W_1 and W_2 in Figure 5.3 can be combined with the uncertainty in the nonlinearity when applying the robust stability condition.

5.2.3 Performance Analysis

The input w of the system described in Figure 5.2(b) is assumed to belong to a set of $\mathcal{L}_2$-norm-bounded functions. Sufficient conditions for the $\mathcal{L}_2$-gain of a system in SNOF to be finite and less than some bound have been derived in terms of convex optimizations with LMI constraints (e.g., see [35,180,270] and citations therein). As for stability conditions, which performance condition to use depends on the assumptions made on the uncertainty in the nonlinearity inversion. The following sufficient condition, which is applied in the pH control case study, quantifies the performance for nonlinearities that are diagonal, static, sector-bounded, and slope-restricted.

Theorem 5.3 (*$\mathcal{L}_2$-gain Performance Condition*). Consider the system described in Figure 5.2(b) with

$$\begin{aligned} x_{k+1} &= A x_k + B_p p_k + B_w w_k \\ q_k &= C_q x_k + D_{qp} p_k + D_{qw} w_k \\ z_k &= C_z x_k + D_{zp} p_k + D_{zw} w_k \end{aligned} \tag{5.9}$$

where $p_k = \Gamma(q_k)$ satisfies the sector-bounded and slope-restricted conditions (5.6) and (5.7), $A \in \mathbb{R}^{n \times n}$, $B_p \in \mathbb{R}^{n \times h}$, $B_w \in \mathbb{R}^{n \times m}$, $C_q \in \mathbb{R}^{h \times n}$, $C_z \in \mathbb{R}^{r \times n}$, $D_{qp} \in \mathbb{R}^{h \times h}$, $D_{zw} \in \mathbb{R}^{r \times m}$, $D_{qw} \in \mathbb{R}^{h \times m}$, $D_{zp} \in \mathbb{R}^{r \times h}$, n

is the number of states, h is the number of nonlinearities, m is the dimension of the input vector, and r is the dimension of the output vector.

The system of the form described in Figure 5.2(b) and (5.9) is stable and has an upper bound on the induced $\mathcal{L}_2$-norm (or $\mathcal{L}_2$-gain) $\eta^\star$ if the optimization problem

$$\begin{aligned} \eta^\star = & \min_{P,Q,\Lambda,\bar{\Lambda}} \eta \\ & \text{s.t. } P,\, P_{11} = P_{11}^{\mathrm{T}} > 0,\, Q \geq 0,\, \tilde{Q} \geq 0,\, T \geq 0,\, \tilde{T} \geq 0,\, N \geq 0,\, \bar{G} \leq 0,\, \eta > 0 \end{aligned} \tag{5.10}$$

has feasible solutions, where Q, $\tilde{Q}$, T, $\tilde{T}$, and N are diagonal, $\xi = \mathrm{diag}(\xi_i)$, $\mu = \mathrm{diag}(\mu_i)$, the matrix $\bar{G} = \bar{G}^{\mathrm{T}}$ is defined by

$$\begin{aligned} \bar{G} \triangleq & \bar{A}_a^{\mathrm{T}} P \bar{A}_a - \bar{E}_a^T P \bar{E}_a + \bar{U}_1 + \bar{U}_2 - \bar{S}_1 - \bar{S}_2 - \bar{S}_3 \\ & - \begin{bmatrix} 0 \\ 0 \\ 0 \\ \mathrm{I} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & \mathrm{I} \end{bmatrix} + \frac{1}{\eta^2} \begin{bmatrix} C_z^{\mathrm{T}} \\ D_{zp}^{\mathrm{T}} \\ 0 \\ D_{zw}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} C_z & D_{zp} & 0 & D_{zw} \end{bmatrix}, \end{aligned} \tag{5.11}$$

and the matrices $\bar{A}_a$, $\bar{E}_a$, $\bar{U}_1$, $\bar{U}_2$, $\bar{S}_1$, $\bar{S}_2$, and $\bar{S}_3$ are defined in (5.24).

5.3 Application to pH Neutralization

Consider the continuous pH neutralization of an acid stream by a highly concentrated basic stream (see Figure 5.4). The only measured signal is the controlled variable, which is the pH, and the manipulated variable is the flow rate of basic solution. The tank has a volume of 5 liters, and the 0.01 M hydrochloric (HCl) and 0.1 M caustic soda (NaOH) solutions are pumped from 200-liter tanks into the mixing tank. Solutions are prepared with tap water, which contains a significant amount of dissolved carbon dioxide (in the form of aqueous HCO_3^- and CO_3^{2-}). Unmeasured disturbances include the buffering species (carbonates)

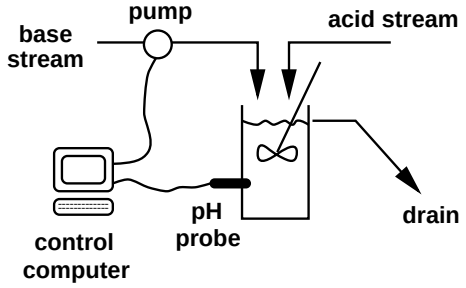


Figure 5.4: pH neutralization apparatus.

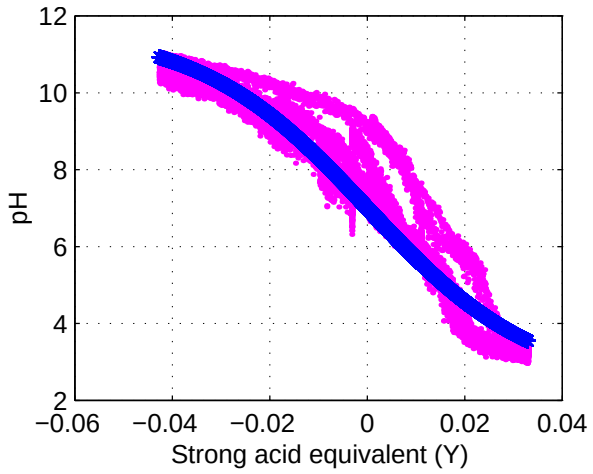


Figure 5.5: Titration curve nonlinearity (W_1 and W_2 were determined by nonlinear least-squares fitting). The model is the thick blue line; experimental data points are purple dots.

in the base and acid flows, nonideal mixing in the main tank, nonideal mixing in the acid and base storage tanks, and air bubbles in the tubes through which the acid and base streams flow.

In a pH process represented as a Wiener model in Figure 5.1, the dynamic linear system describes the mixing dynamics and the static nonlinearity describes the titration curve [127]. Other than having a different analytical expression for the nonlinearity, the same process model was used as in [205]:

$$\begin{aligned} V\frac{dY}{dt} &= -FY - u \\ \text{pH} &= W_2 \tanh(W_1 Y) \end{aligned} \tag{5.12}$$

where V is the volume of the mixing tank, u is the base flow rate, F is the acid flow rate, W_1 and W_2 are weights, and Y is the dimensionless strong acid equivalent [205].

For the pH process, each term in the SNOF has clear physical meaning. The nonlinearity directly corresponds to the titration curve, and the linear term directly corresponds to the mixing dynamics. Figure 5.5 shows the form of the nonlinear relation between Y and pH, with some experimental data collected for a pH experimental apparatus at the University of Illinois. The process disturbances result in significant uncertainty in the nonlinearity, as shown in Figure 5.5.

An exact discretization [54] of (5.12) leads to:

$$\begin{aligned} x_{k+1} &= e^{-F\Delta t/V} x_k + \frac{1}{F}\left(e^{-F\Delta t/V} - 1\right) u_k \\ Y_k &= x_k \\ \text{pH}_k &= W_2 \tanh(W_1 x_k) \end{aligned} \tag{5.13}$$

where Δt is the sampling-time. In addition, there is a time delay θ due to the sensor location. The process

θ	Necessary (Thm. 5.1)	Sufficient (Thm. 5.2)
0	0.91	0.91
2	2.08	2.08
6	4.63	4.63
10	7.17	7.17
12	8.46	8.46
16	11.00	11.00

Table 5.1: Lowest IMC filter parameter λ (in seconds) that indicates stability for the closed-loop system with the linear IMC controller. Perfect model information is assumed. Theorem 5.2 was applied with $\xi = \mu = 1$. All times are in seconds.

time delay $\theta = 0.27$ minutes (= 16 seconds) and effective time constant $\tau = 3$ minutes were determined from experimental data by nonlinear least-squares estimation.

The linear and nonlinear IMC structures are shown in Figure 5.3. The block diagram in Figure 5.3 can be written directly as an SNOF.

Linear IMC design with no nonlinearity cancellation First consider the case where the IMC is linear and the process nonlinearity Γ_1 is perfectly known and equal to tanh. Closed-loop stability results for a range of time delays are included to provide an indication as to the potential conservatism of the analysis tools for the pH neutralization problem (see Table 5.1). The necessary and the sufficient stability analysis results gave identical stability limits, indicating no conservatism for this closed-loop system. For this problem, the smallest stabilizing IMC filter parameter increased linearly with the time delay.

Now the linear internal model controller was designed that minimizes the desired closed-loop response time (λ) while requiring that the effect of worst-case disturbances d on the output y cannot be magnified by more than a factor of 2.5. This is a direct nonlinear generalization of the linear IMC design procedure for the specification that the peak sensitivity is less than 2.5 [185, 249]. The performance measure for a range of controller tuning parameter λ for the linear IMC controller is shown in Figure 5.6. The optimal λ is equal to 8.5 minutes. This performance condition places a much stronger restriction on the closed-loop speed of response than the requirement of nominal stability.

Nonlinear IMC design with perfect nonlinearity cancellation The geometric control literature commonly assumes that the nonlinear process is perfectly known. This assumption would imply that the controller nonlinearity Γ_2 (in Figure 5.3) perfectly cancels the process nonlinearity Γ_1, and the closed-loop stability could be determined from linear stability analysis. The resulting stability conditions are exactly the same as those used to compute the necessary condition in Table 5.1. Hence for the pH neutralization process under the assumption of a perfect model, the stability limit for the linear IMC is equal to the stability limit for the nonlinear IMC, which is $\lambda = 11$ seconds for the time delay $\theta = 16$ seconds.

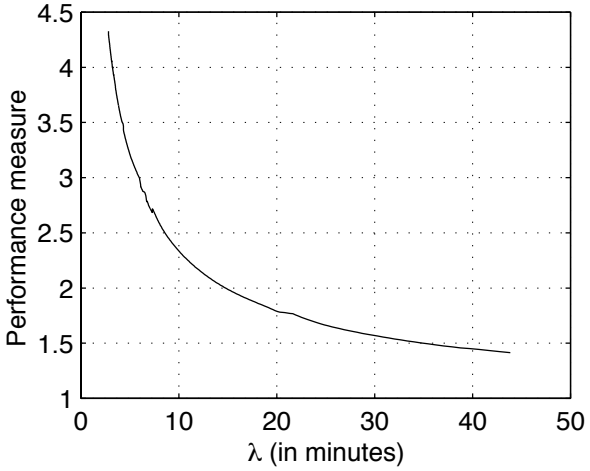


Figure 5.6: Performance measure $\eta^\star$ for the closed-loop system with the linear internal model controller with tuning parameter λ (for a process time delay $\theta = 0.27$ minutes).

θ	Necessary (Thm. 5.1)	Sufficient (Thm. 5.2)
0	1.44	1.44
2	3.71	4.28
6	8.79	9.38
10	13.88	14.49
12	16.44	17.04
16	21.52	22.17

Table 5.2: Lowest IMC filter parameter λ (in seconds) that indicates robust stability for the closed loop system with the nonlinear IMC controller and significant uncertainty in the process nonlinearity.

Nonlinear IMC design with robustness to uncertainty in nonlinearity inversion Nonlinear models are rarely of very high accuracy, and this is certainly true for the pH neutralization process as demonstrated in Figure 5.5. Now a nonlinear IMC controller is designed that minimizes the desired closed-loop response time (λ) while requiring stability to uncertainties in the cancellation of the process nonlinearity, which is mathematically represented as deviations in $\Gamma_2\Gamma_1$ from one. Based on close inspection of Figure 5.5, it was assumed that the maximum overall slope of $\Gamma_2\Gamma_1$ could be as high as two, while its instantaneous slope could be off as much as a factor of four. Ensuring stability for this range of uncertainties is a nonlinear generalization of the common objective used in the design of controllers for linear systems of providing a gain margin of 2. Table 5.2 gives stability limits for a range of time delays to provide some indication of potential conservatism of the stability analysis results. The minimum IMC filter parameter λ that provides robust stability increases linearly with the time delay. The minimize filter parameter is $\lambda \approx 22$ seconds for the time delay $\theta = 16$ seconds, which is twice the value computed for the nonlinear IMC design that ignored uncertainty in the nonlinearity inversion.

5.4 Discussion

Whether a performance or stability constraint was used in the IMC design had a significant influence on the filter parameter λ in the nonlinear IMC-based controller. For the pH neutralization process, controllers tuned based on a nominal stability or robust stability constraint provided much faster closed-loop speed of response than the controller based on the worst-case performance constraint. This observation indicates the importance of carefully choosing the controller design criteria.

The sufficient stability condition in Theorem 2 was nonconservative to three significant figures for the closed-loop system controlled by the linear internal model controller, for a range of time delays (see Table 5.1). For the stability analysis that took the uncertainty in the nonlinearity inversion into account, the sufficient condition could potentially be somewhat conservative, as there is a gap in the values for the minimize allowable filter parameter λ computed from the necessary condition and sufficient conditions (see Table 5.2). The gap is less than 0.1% for a system with no time delay and about 3% for the time delay of 16 seconds identified in the experiments. The gap in the values of the minimum λ for the other time delays are all about 0.6 seconds (see Table 5.2). Although the limits computed from the necessary and the sufficient conditions are not exactly equal, they are certainly close enough for practical application.

The stability analysis that took uncertainty in the nonlinearity inversion into account indicated that the minimum stabilizing values for the filter parameter λ were twice as large as the values that were computed that ignored the uncertainty in the nonlinearity inversion. This observation indicates that importance of taking uncertainty during nonlinearity inversion into account when designed nonlinear inversion-based controllers.

The approach in this chapter can be extended to nonlinear operators Γ in which each of its outputs is

related to each of its inputs from conditions such as shown in Theorem 2. The simplest way to implement this generalization is to rearrange the scalar nonlinearities to form a larger diagonal nonlinear operator in Figure 2. The expressions for the state-space matrices in Figure 2 are messier. The generalization to full-block or block-diagonal nonlinear operators Γ follows the same derivations, but with much messier nomenclature.

The approach in this chapter applies to systems with larger time delays. Proposition 5.1 shows the transformation of a system with potentially large time delay into the standard state-space system description; this transformation is standard in the control literature. The computational cost of analyzing systems with larger time delays is much smaller than suggested by the increase in dimensions, because the state-space matrices for the extended system are highly sparse, and many existing LMI solvers are effective at exploiting sparsity (e.g., [160, 255]).

5.5 Summary and Future Work

A nonlinear internal model control procedure was presented for stable Wiener systems that ensures robustness of closed-loop stability and performance to uncertainties in the inversion of the static nonlinearity, while having polynomial-time computational cost. Several more general observations can be made based on a pH control case study. Assuming perfect nonlinearity inversion when controlling pH processes led to overly optimistic predictions on the achievable closed-loop performance, which indicates that the commonly made assumption of perfect nonlinearity inversion can produce poor results in practical applications. A comparison of pH controllers designed to satisfy robust stability or disturbance suppression constraints showed that the closed-loop response speed could significantly change depending on the design criteria. A comparison of the sufficient robust stability condition with a necessary condition showed that the sufficient robust stability condition was nonconservative for this particular application.

The nonlinear IMC procedure is applicable to stable Wiener systems with unstable zero dynamics, unmeasured states, disturbances, and measurement noise. This is in contrast to many nonlinear control methods that require stable zero dynamics and/or ignore disturbances and measurement noise. The generalization of the approach to Hammerstein and Sandwich models is straightforward, and can be used to explicitly incorporate actuator constraints into the nonlinear controller design, by combining these static nonlinearities with any other static nonlinearity associated with the input to the process. The approach can also be combined with directionality compensation, which can improve the closed-loop dynamics for ill-conditioned processes [51, 250].

Appendix

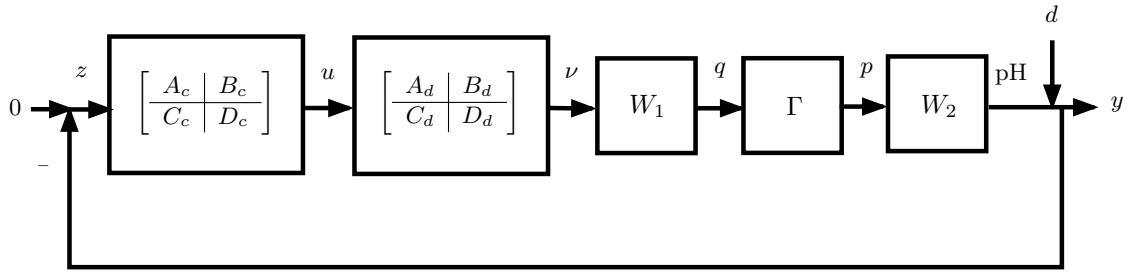


Figure 5.7: Block diagram for pH system.

Computation of SNOF for pH system Consider the block diagram in Figure 5.7 where $\left[\begin{array}{c|c} A_d & B_d \\ \hline C_d & D_d \end{array}\right] \triangleq C_d(z\mathrm{I} - A_d)^{-1}B_d + D_d$ and $w = d$. The system equation is given by

$$\begin{aligned}
z_k &= -(\mathrm{pH}_k + w_k) = -W_2 p_k - w_k \\
q_k &= W_1 \nu_k \\
&= W_1 \left(C_d Y_k + D_d C_c x_k - D_d D_c W_2 p_k - D_d D_c w_k\right) \\
Y_{k+1} &= A_d Y_k + B_d C_c x_k - B_d D_c W_2 p_k - B_d D_c w_k \\
x_{k+1} &= A_c x_k - B_c W_2 p_k - B_c w_k \\
p_k &= \Gamma(q_k).
\end{aligned}$$

Define the concatenated vector $\hat{x}_k \triangleq (Y_k, x_k)^{\mathrm{T}}$, then the SNOF in Figure 5.2(b) for the pH system is given by

$$\begin{aligned}
\hat{x}_{k+1} &= \underbrace{\begin{bmatrix} A_d & B_d C_c \\ 0 & A_c \end{bmatrix}}_{A} \hat{x}_k + \underbrace{\begin{bmatrix} -B_d D_c W_2 \\ -B_c W_2 \end{bmatrix}}_{B_p} p_k + \underbrace{\begin{bmatrix} -B_d D_c \\ -B_c \end{bmatrix}}_{B_w} w_k \\
q_k &= \underbrace{\begin{bmatrix} W_1 C_d & C_1 D_d C_c \end{bmatrix}}_{C_q} \hat{x}_k + \underbrace{\begin{bmatrix} -W_1 D_d D_c W_2 \end{bmatrix}}_{D_{qp}} p_k \\
&\quad + \underbrace{\begin{bmatrix} -W_1 D_d D_c \end{bmatrix}}_{D_{qw}} w_k \\
z_k &= \underbrace{\begin{bmatrix} 0 \end{bmatrix}}_{C_z} \hat{x}_k + \underbrace{\begin{bmatrix} -W_2 \end{bmatrix}}_{D_{zp}} p_k + \underbrace{\begin{bmatrix} -\mathrm{I} \end{bmatrix}}_{D_{zw}} w_k.
\end{aligned}$$

Extended State-space Representation for Systems with Time-delay in Input Channels Consider the discrete-time system model

$$x_{k+1} = A_d(\Delta t)x_k + B_d(\Delta t)\tilde{u}_k(\theta) \tag{5.14}$$

where Δt is the sampling interval for discretization such that $x_k = x(k\Delta t)$ and $\tilde{u}_k(\theta) \triangleq u(k\Delta t - \theta)$ with time delay θ. Suppose that $\theta = m\Delta t$ where m is an integer. Then the system equation can be written as

$$x_{k+1} = A_d(\Delta t)x_k + B_d(\Delta t)u_{k-m} \tag{5.15}$$

Define the auxiliary states

$$\begin{aligned} \zeta_{k+1}^0 &\triangleq A_d(\Delta t)\zeta_k^0 + B_d(\Delta t)u_k, \\ \zeta_{k+1}^{\ell+1} &\triangleq \zeta_k^\ell, \quad \ell = 0, \ldots, m-1. \end{aligned} \tag{5.16}$$

Then the state of the system (5.14) is the output of the extended state-space model:

$$\begin{aligned} \zeta_{k+1} &= \mathcal{A}_d(\Delta t)\zeta_k + \mathcal{B}_d(\Delta t)u_k \\ x_k &= \mathcal{C}_d\zeta_k \end{aligned} \tag{5.17}$$

where $\zeta \triangleq \mathrm{vec}(\zeta^0, \zeta^1, \cdots, \zeta^m) \in \mathbb{R}^{nm}$ and[4]

$$\mathcal{A}_d(\Delta t) \triangleq \begin{bmatrix} A_d(\Delta t) & 0 & \cdots & \cdots & \cdots & 0 \\ \mathrm{I} & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \mathrm{I} & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \mathrm{I} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \mathrm{I} & 0 \end{bmatrix}, \mathcal{B}_d(\Delta t) \triangleq \begin{bmatrix} B_d(\Delta t) \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

$$\mathcal{C}_d \triangleq \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 & \mathrm{I} \end{bmatrix}.$$

Proposition 5.1. The system (5.14) is stable if and only if the system (5.17) is stable.

Proof. ($\Leftarrow$): This direction is shown by observing that the state vector in (5.14) is a subset of the state vector in (5.17), that is, $\zeta^m = x$. ($\Rightarrow$): Suppose that the system (5.14) has a unique stable steady-state x^{eq}. Then, there exists a sufficiently large $K \in \mathbb{Z}_+$ such that $\zeta_k^m = x^{\mathrm{eq}}$ for all $k \geq K$, where it can be assumed that $K \gg m$ without loss of generality. This implies that $\zeta_\kappa^\ell = x^{\mathrm{eq}}$, $\ell = 0, \ldots, m-1$ for all $\kappa \geq K - m$, which is equivalent to stability of the system (5.17). QED

[4] $\mathrm{vec}(a, b) \in \mathbb{R}^{n_1+n_2}$ refers to the concatenation of vectors $a \in \mathbb{R}^{n_1}$ and $b \in \mathbb{R}^{n_2}$, i.e., its first n_1 entries are equal to a and the remaining entries are equal to b

Proof of Theorem 5.2

Proof. Consider the Lyapunov function

$$V(x_k) = \bar{x}_k^T P \bar{x}_k + 2\sum_{i=1}^{n_q} Q_{ii} \int_0^{q_{k,i}} \phi_i(\sigma) d\sigma + 2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \int_0^{q_{k,i}} [\xi_i \sigma - \phi_i(\sigma)]\, d\sigma,$$

where

$$\bar{x}_k \triangleq \begin{bmatrix} x_k \\ p_k \\ q_k \end{bmatrix},\ P^T = P \triangleq \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^T & P_{22} & P_{23} \\ P_{13}^T & P_{23}^T & P_{33} \end{bmatrix} \geq 0,\ P_{11} > 0,$$
$$Q_{ii} \geq 0,\ \tilde{Q}_{ii} \geq 0,\ \forall i = 1, \ldots, n_q,$$

and the subscript k indicates a sampling instance. Both p_k and q_k are functions of the state variable vector x_k, and the above Lyapunov function is radially unbounded and positive for all nonzero $x_k \in \mathbb{R}^n$. The difference in the Lyapunov function between the $k+1$ and k sampling instances is

$$\begin{aligned} \Delta V(x_k) = {} & \zeta_k^T (A_a^T P A_a - E_a^T P E_a) \zeta_k \\ & + \ 2\sum_{i=1}^{n_q} Q_{ii} \int_{q_{k,i}}^{q_{k+1,i}} \phi_i(\sigma) d\sigma + \ 2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \int_{q_{k,i}}^{q_{k+1,i}} [\xi_i \sigma - \phi_i(\sigma)]\, d\sigma, \end{aligned} \tag{5.18}$$

where

$$\zeta_k \triangleq \begin{bmatrix} x_k \\ p_k \\ p_{k+1} \end{bmatrix}, A_a \triangleq \begin{bmatrix} A & B & 0 \\ 0 & 0 & I \\ CA & CB & D \end{bmatrix}, E_a \triangleq \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ C & D & 0 \end{bmatrix}$$

Slope restrictions on the nonlinearities place an upper bound on the first integral:

$$\begin{aligned} 2\sum_{i=1}^{n_q} Q_{ii} \int_{q_{k,i}}^{q_{k+1,i}} \phi(\sigma) d\sigma &\leq 2\sum_{i=1}^{n_q} Q_{ii} \left\{ \begin{array}{c} (\phi_{k+1,i} - \phi_{k,i})(q_{k+1,i} - q_{k,i}) \\ -\frac{1}{2\mu_i}(\phi_{k+1,i} - \phi_{k,i})^2 \end{array} \right\} \\ &= \zeta_k^T U_1 \zeta_k, \end{aligned}$$

where U_1 is given in (5.23). Similarly, an upper bound can be derived on the second integral:

$$\begin{aligned} 2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \int_{q_{k,i}}^{q_{k+1,i}} [\xi_i \sigma - \phi(\sigma)]\, d\sigma &= -2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \int_{q_{k,i}}^{q_{k+1,i}} \phi(\sigma) d\sigma + 2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \int_{q_{k,i}}^{q_{k+1,i}} \xi_i \sigma d\sigma \\ &\leq -2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \left\{ \frac{1}{2\mu_i}(\phi_{k+1,i} - \phi_{k,i})^2 + \phi_{k,i}(q_{k+1,i} - q_{k,i}) \right\} \\ &\quad + 2\sum_{i=1}^{n_q} \tilde{Q}_{ii} \xi_i \left[q_{k+1,i}^2 - q_{k,i}^2 \right] \end{aligned}$$

$$= \zeta_k^T U_2 \zeta_k,$$

where U_2 is given in (5.23).

Since the (negative) feedback-connected nonlinearity is monotonic with slope restriction in addition to being $[0, \xi]$ sector-bounded, i.e., $\phi \in \Phi_{sb}^{[0,\xi]} \cap \Phi_{sr}^{[0,\mu]}$, it can be shown that the following inequalities are satisfied at each sampling instance k and all indices $i = 1, \ldots, n_q$:

$$\phi_{k,i}[\xi_i^{-1}\phi_{k,i} - q_{k,i}] \leq 0, \tag{5.19}$$

$$(\phi_{k+1,i} - \phi_{k,i})[\mu_i^{-1}(\phi_{k+1,i} - \phi_{k,i}) - (q_{k+1,i} - q_{k,i})] \leq 0. \tag{5.20}$$

The following notations based on (5.19) are useful when applying the S-procedure:

$$\sum_{i=1}^{n_q} 2\tau_i \phi_{k,i}[\xi_i^{-1}\phi_{k,i} - q_{k,i}] = \zeta_k^T S_1 \zeta_k, \tag{5.21}$$

$$\sum_{i=1}^{n_q} 2\tilde{\tau}_i \phi_{k+1,i}[\xi_i^{-1}\phi_{k+1,i} - q_{k+1,i}] = \zeta_k^T S_2 \zeta_k, \tag{5.22}$$

where S_1 and S_2 are given in (5.23). A similar notation based on the inequality (5.20) is:

$$\sum_{i=1}^{n_q} 2N_{ii}(\phi_{k+1,i} - \phi_{k,i}) \left[\mu_i^{-1}(\phi_{k+1,i} - \phi_{k,i}) - (q_{k+1,i} - q_{k,i})\right] = \zeta_k^T S_3 \zeta_k,$$

where S_3 is given in (5.23).

Applying the S-procedure, if the LMI $G \triangleq A_a^T P A_a - E_a^T P E_a + U_1 + U_2 - S_1 - S_2 - S_3 < 0$ is feasible then $\Delta V(x_k) < 0$ is satisfied for the specific class of feedback-connected nonlinearities $\phi \in \Phi_{sb}^{[0,\xi]} \cap \Phi_{sr}^{[0,\mu]}$. All of introduced matrix (decision) variables are of compatible dimensions. QED

Proof of Theorem 5.3

Proof. It is not difficult to see that the system given in (5.9) is g.a.s. and is dissipative with respect to the supply rate

$$s(w_k, z_k) \triangleq \begin{bmatrix} w_k \\ z_k \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \mathrm{I} & 0 \\ 0 & -\frac{1}{\eta^2}\mathrm{I} \end{bmatrix} \begin{bmatrix} w_k \\ z_k \end{bmatrix}$$

and the Lyapunov function $V(x_k)$ if and only if $\Delta V(x_k) \leq s(w_k, x_k)$ holds for all $k \in \mathbb{Z}$, which is equivalent to the $\mathcal{L}_2$-gain performance bound, $\sup_{\|w\|_2 \leq 1} \|z\|_2 \leq \eta$. Consider the Lyapunov function (5.18) and the system equation (5.9). Then the condition $\Delta V(x_k) - s(w_k, z_k) \leq 0$ can be rewritten as $\bar{\zeta_k}^{\mathrm{T}} \bar{G} \bar{\zeta_k}$ where $\bar{G}$ is given in (5.11) and $\bar{\zeta}^{\mathrm{T}} \triangleq \begin{bmatrix} x_k^{\mathrm{T}} & p_k^{\mathrm{T}} & p_{k+1}^{\mathrm{T}} & w_k^{\mathrm{T}} \end{bmatrix}$. The computation of the matrix $\bar{G}$ can be performed by using the S-procedure and the derivation is similar to the proof of Theorem 5.2. QED

Matrices for the Application of the S-Procedure in Theorem 5.2

$$
\begin{aligned}
U_1 &\triangleq \begin{bmatrix} 0 & -(CA-C)^{\mathrm{T}}Q & (CA-C)^{\mathrm{T}}Q \\ * & -Q(CB-D)-(CB-D)^{\mathrm{T}}Q-Q\mu^{-1} & -QD+Q\mu^{-1} \\ * & * & -QD-D^{\mathrm{T}}Q-Q\mu^{-1} \end{bmatrix} \\
U_2 &\triangleq \begin{bmatrix} A^{\mathrm{T}}C^{\mathrm{T}}\tilde{Q}\xi CA-C^{\mathrm{T}}\tilde{Q}\xi C & A^{\mathrm{T}}C^{\mathrm{T}}\tilde{Q}\xi CB-(CA-C)^{\mathrm{T}}\tilde{Q}-C^{\mathrm{T}}\tilde{Q}\xi D & A^{\mathrm{T}}C^{\mathrm{T}}\tilde{Q}\xi D \\ * & \begin{pmatrix} B^{\mathrm{T}}C^{\mathrm{T}}\tilde{Q}\xi CB+D^{\mathrm{T}}\tilde{Q}\xi D \\ -(CB-D)^{\mathrm{T}}\tilde{Q}-\tilde{Q}(CB-D)-\mu^{-1}\tilde{Q} \end{pmatrix} & \mu^{-1}\tilde{Q}-\tilde{Q}D+B^{\mathrm{T}}C^{\mathrm{T}}\tilde{Q}\xi D \\ * & * & -\mu^{-1}\tilde{Q}+2D^{\mathrm{T}}\tilde{Q}\xi D \end{bmatrix} \\
S_1 &\triangleq \begin{bmatrix} 0 & -C^{\mathrm{T}}T & 0 \\ * & 2\xi^{-1}T-TD-D^{\mathrm{T}}T & 0 \\ * & * & 0 \end{bmatrix}, \quad S_2 \triangleq \begin{bmatrix} 0 & 0 & -A^{\mathrm{T}}C^{\mathrm{T}}\tilde{T} \\ * & 0 & -B^{\mathrm{T}}C^{\mathrm{T}}\tilde{T} \\ * & * & 2\xi^{-1}\tilde{T}-\tilde{T}D-D^{\mathrm{T}}\tilde{T} \end{bmatrix} \\
S_3 &\triangleq \begin{bmatrix} 0 & -(CA-C)^{\mathrm{T}}N & (CA-C)^{\mathrm{T}}N \\ * & 2N\mu^{-1}+(CB-D)^{\mathrm{T}}N+N(CB-D) & -2N\mu^{-1}+(CB-D)^{\mathrm{T}}N+ND \\ * & * & 2N\mu^{-1}-D^{\mathrm{T}}N-ND \end{bmatrix}.
\end{aligned} \tag{5.23}
$$

Matrices in Theorem 5.3

$$
\begin{aligned}
\bar{A}_a &\triangleq \begin{bmatrix} A & B_p & 0 & B_w \\ 0 & 0 & \mathrm{I} & 0 \\ C_qA & C_qB_p & D_{qp} & CB_w \end{bmatrix}, \quad \bar{E}_a \triangleq \begin{bmatrix} \mathrm{I} & 0 & 0 & 0 \\ 0 & \mathrm{I} & 0 & 0 \\ C_q & D_{qp} & 0 & 0 \end{bmatrix}, \\
\bar{U}_1 &\triangleq \begin{bmatrix} 0 & -(C_qA-C_q)^{\mathrm{T}}Q & (C_qA-C_q)^{\mathrm{T}}Q & 0 \\ * & \begin{pmatrix} -Q(C_qB_p-D_{qp}) \\ -(C_qB_p-D_{qp})^{\mathrm{T}}Q \\ -Q\mu^{-1} \end{pmatrix} & \begin{pmatrix} (C_qB_p-D_{qp})^{\mathrm{T}}Q \\ -QD_{qp}+Q\mu^{-1} \end{pmatrix} & -QC_qB_w \\ * & * & \begin{pmatrix} Q(C_qB_p-D_{qp}) \\ +(C_qB_p-D_{qp})^{\mathrm{T}}Q \\ -Q\mu^{-1} \end{pmatrix} & QC_qB_w \\ * & * & * & 0 \end{bmatrix}, \\
\bar{U}_2 &\triangleq \begin{bmatrix} \begin{pmatrix} A^{\mathrm{T}}C_q^{\mathrm{T}}\xi\tilde{Q}C_qA \\ -C_q^{\mathrm{T}}\xi\tilde{Q}C_q \end{pmatrix} & \begin{pmatrix} (C_qA-C_q)^{\mathrm{T}}\xi \\ \cdot\tilde{Q}(C_qB_p+D_{qp}) \\ -(C_qA-C_q)^{\mathrm{T}}\tilde{Q} \end{pmatrix} & \begin{pmatrix} (C_qA-C_q)^{\mathrm{T}} \\ \cdot\xi\tilde{Q}D_{qp} \end{pmatrix} & \begin{pmatrix} (C_qA-C_q)^{\mathrm{T}} \\ \cdot\xi\tilde{Q}C_qB_w \end{pmatrix} \\ * & \begin{pmatrix} 2(C_qB_p+D_{qp})^{\mathrm{T}}\xi \\ \cdot\tilde{Q}(C_qB_p+D_{qp}) \\ -\tilde{Q}(C_qB_p-D_{qp}) \\ -(C_qB_p-D_{qp})^{\mathrm{T}}\tilde{Q} \\ -\tilde{Q}\mu^{-1} \end{pmatrix} & \begin{pmatrix} (C_qB_p+D_{qp})^{\mathrm{T}} \\ \cdot\xi\tilde{Q}D_{qp} \\ -\tilde{Q}D_{qp}+\tilde{Q}\mu^{-1} \end{pmatrix} & \begin{pmatrix} (C_qB_p+D_{qp})^{\mathrm{T}} \\ \cdot\xi\tilde{Q}C_qB_w \\ -\tilde{Q}C_qB_w \end{pmatrix} \\ * & * & 2D_{qp}\xi\tilde{Q}D_{qp}-\tilde{Q}\mu^{-1} & D_{qp}^{\mathrm{T}}\xi\tilde{Q}C_qB_w \\ * & * & * & 2B_w^{\mathrm{T}}C_q^{\mathrm{T}}\xi\tilde{Q}C_qB_w \end{bmatrix},
\end{aligned} \tag{5.24}
$$

$$\bar{S}_1 \triangleq \begin{bmatrix} 0 & -C_q^{\mathrm{T}}T & 0 & 0 \\ * & 2\xi^{-1}T - TD_{qp} - D_{qp}^{\mathrm{T}}T & 0 & 0 \\ * & * & 0 & 0 \\ * & * & * & 0 \end{bmatrix}, \; \bar{S}_2 \triangleq \begin{bmatrix} 0 & -A^{\mathrm{T}}C_q^{\mathrm{T}}\tilde{T} & 0 & 0 \\ * & -B_q^{\mathrm{T}}C_q^{\mathrm{T}}\tilde{T} - \tilde{T}C_qB_p & -\tilde{T}D_{qp} & \tilde{T}C_qB_w \\ * & * & 2\xi^{-1}\tilde{T} & 0 \\ * & * & * & -B_q^{\mathrm{T}}C_q^{\mathrm{T}}\tilde{T} \end{bmatrix},$$

$$\bar{S}_3 \triangleq \begin{bmatrix} 0 & -(C_qA - C_q)^{\mathrm{T}}N & 0 & 0 \\ * & \begin{pmatrix} (C_qB_p - D_{qp})^{\mathrm{T}}N \\ +N(C_qB_p - D_{qp}) \\ +2N\mu^{-1} \end{pmatrix} & \begin{pmatrix} (C_qB_p - D_{qp})^{\mathrm{T}}N \\ -2N\mu^{-1} + ND_{qp} \end{pmatrix} & NC_qB_w \\ * & * & 2N\mu^{-1} - D_{qp}^{\mathrm{T}}N - ND_{qp} & -NC_qB_w \\ * & * & * & 0 \end{bmatrix}. \tag{5.25}$$

Chapter 6

Computational Complexity of Robust Control: An Overview

Abstract The main role of feedback control is to address the effects of uncertainties, and much of the process control literature since the 1980s has involved the analysis and design of uncertain systems. As the complexity of the systems that are being controlled continues to increase, a practical consideration is the computational cost of control analyses and design methods as the system size increase. This chapter reviews results on the computational complexity of robust control problems, starting with well-known results and then moving to lesser known results that have broad implications. This chapter ends with a look to the future, towards stochastic robustness analysis.

6.1 Introduction

The structured singular value μ has been widely used to analyze the effects of uncertainties on the stability and performance of multivariable systems. John Doyle [72] and Michael Safonov [228] were largely responsible for the introduction of μ for the analysis of robust stability and robust performance margins in the early 1980s. The main advantage of μ compared to earlier approaches was its ability to explicitly take into account the *structure* of the uncertainties, with the value of μ quantifying how much the magnitude of the uncertainties can be increased until the system first becomes unstable. The μ analysis also produces deterministic worst-case values for the uncertainties, which focuses attention on combinations of uncertain parameters that result in the worst-case behavior of the system.

Historically, the graphical approach for the robust stability analysis of single-input single-output (SISO) systems was already applied by the early 1960s, which mapped uncertain gains, phases, and parameters into the Nyquist plot [111]. By mid 1960 George Zames [301, 302] had applied functional analysis to analyze

stability for input-output problems where two elements were interconnected in a feedback loop. His conic relation stability theorem provided conditions to ensure that the overall closed-loop system was stable. Safonov [227,229] extended this theorem to multi-input multi-output (MIMO) systems. Zames' and Safonov's proposed approaches were based on the concept of *topological separation*. The feedback interconnection of two elements H_1 and H_2 is robustly stable if and only if the graph of H_1 and the inverse graph of H_2 are topologically separated for any allowed variations in H_1 or H_2. The main goal in this approach to robust stability analysis was to find a separator or a set of separators that characterize the input-output relation (i.e., graph) of either of the system elements H_1 or H_2. Similar approaches were investigated and further developed by many researchers including Goh & Safonov [97, IQC separator], Megretski & Rantzer [178,217, IQC theory], Scherer [232,233, full-block S-procedure], and Iwasaki & Hara [121, quadratic separator], to name a few. Robust stability analysis based on topological separation can be used to compute upper bounds on μ, and these upper bounds tend to be readily computable for systems with a fairly large number of uncertainties.

An alternative to topological approaches for robustness analysis was the algebraic approach by Kharitonov and others [28, 138]. A simple stability criterion was derived for uncertain polynomials in which the coefficients are allowed to independently vary within bounded intervals. Many extensions to Kharitonov's results were published (see [1] and references therein) whose goals were to extend the results to successively more complicated uncertainty descriptions while retaining very high computational efficiency. The result by Manfred Morari and coworkers [43, 44] showed that the computationally efficient algebraic approaches could not possibly be applicable to general uncertainty descriptions, which resulted in a severe drop in interest in algebraic approaches.

Lyapunov methods for stability analysis have a very long history and many very good introductory descriptions are available [136, 279]. Lyapunov methods for robustness analysis are arguably the most flexible, and can be used to derive many of the robustness analysis tools derived using other methods. Many of the papers in the last 20 years that employ Lyapunov methods derive robustness analysis conditions in terms of optimizations over linear or bilinear matrix inequality constraints, the computational consequences of which are discussed later in this chapter.

The main purpose of this chapter is to provide an overview of results related to the computational complexity of robustness analysis. As the structured singular value μ is an exact measure of the robustness of systems with structured uncertainties, numerous efforts have been made to develop efficient algorithms for its computation. Several researchers showed that the exact calculation of μ for systems with purely real [43, 44, 64, 195, 215], mixed real and complex [43, 44], and purely complex uncertainties is *NP-hard* [267], implying that its exact computation is very expensive for large-scale systems. It was also shown that the design of μ-optimal controller for systems with purely complex uncertainties is NP-hard [267] by selection of performance weights so that the optimal closed-loop performance objective is equal to the value for μ for an equivalent robustness analysis problem that is NP-hard. This same proof technique can be used to show

that the robust control design problem for any particular uncertainty class is NP-hard when its robustness analysis problem is NP-hard for the same uncertainty class. In particular, the design of robust optimal controllers is NP-hard for uncertainties that are purely real, purely complex, or mixed.

After the exact μ-calculation was proven to be NP-hard, it was shown that the approximation of μ within a specified accuracy is also NP-hard for systems with purely real [88][1] and mixed real and complex [42] uncertainties. The greater ease in the computation of upper bounds on μ motivated the analysis of their conservatism, which was investigated by Alexandre Megretski and others [174, 175, 179, 224, 268].

The high cost of μ-calculation motivated the development of polynomial-time dimensional reduction methods [16,225,226]. The main idea behind these algorithms is to reduce the dimension of the set of uncertainties as much as possible using polynomial-time methods, so that the subsequent robustness analysis is less costly.

The NP-hardness of μ-calculation increased interest in probabilistic randomized algorithms [137, 253, 262, 280–282] as computationally more efficient alternative approaches to define and compute robustness margins. In these approaches, robustness is evaluated in a probabilistic sense, instead of trying to compute hard bounds based on worst-case uncertainties as done in μ. The probabilistic randomized algorithms provide certain levels of accuracy and confidence in estimates for robust stability and performance margins that depend on a number of uncertainties that have been samplied within the set of allowable uncertainties.

The computational complexity results of [43, 44] motivated many subsequent research efforts to study the computational complexity of systems and control problems. Blondel and Tsitsiklis [33] surveyed computational complexity results in systems and control up to 2000. Of especial interest is that many important control analysis and synthesis problems can be formulated as optimizations over bilinear matrix inequalities (BMIs) (see [270] for a tutorial on such problems) and the problem of checking the solvability of a BMI was also shown to be *NP-hard* [266]. The latter result can be proved in many ways, with one approach being to first show that checking the existence of a Hurwitz-stable matrix in a given affine space is NP-hard, and then showing that this problem polynomially reduces to a BMI feasibility problem [266]. A more straightforward approach to proving that BMI optimization is NP-hard is to just write an indefinite quadratic program in the form of a BMI optimization.

A commonly used method for the design of robust controllers is *DK-iteration*, in which an upper bound of μ is computed while alternating with the solution of polynomial-time optimal control problems. The optimization that DK-iteration attempts to solve is equivalent to a BMI problem, which suggested[2] that replacing μ by a polynomial-time upper bound in the optimization formulation of a robust control design procedure did not necessarily produce an optimization that can be solved in polynomial time.

This section is organized as follows. A brief history of robustness analysis is followed by an introduction to computational complexity theory. Next is a review on the computational complexity of the exact calculation of μ, and on the approximate calculation of μ, which discusses and shows connections between the

[1] We will later discuss how the formulation of [88] was intrinsically flawed.
[2] But did not prove.

various results. These results motivate the subsequent discussion on analyses of the gap between μ and its polynomial-time computable upper bound, and alternative approaches to robustness margin computation such as probabilistic randomized algorithms. Then the section concludes.

Notation Define the set of matrices of block-diagonal perturbations by

$$\begin{aligned}\mathbf{\Delta} \triangleq \Big\{ \mathrm{diag}\Big\{ \delta_1^r \mathrm{I}_{r_1}, \cdots, \delta_k^r \mathrm{I}_{r_k}, \delta_{k+1}^c \mathrm{I}_{r_{k+1}}, \cdots, \delta_m^c \mathrm{I}_{r_m}, \Delta_{r_{m+1}}, \cdots, \Delta_{r_{m_C}} \Big\} \\ : \delta_i^r \in \mathbb{R}, \delta_i^c \in \mathbb{C}, \, \Delta_i \in \mathbb{C}^{r_i \times r_i}, \, \sum r_i = \ell \Big\},\end{aligned} \tag{6.1}$$

which includes real scalar perturbations, complex scalar perturbations, and complex matrix perturbations. The real scalar perturbations can model variations in such parameters as spring constants and time constants, and the complex perturbations can be used to represent *unmodeled dynamics*, in which the set of plants described by the uncertainty description can have much higher order than the nominal process model [208, 209]. Define $\mathbf{B\Delta}$ to be the set of unity norm-bounded perturbations with structure given by $\mathbf{\Delta}$.

For a given matrix $M \in \mathbb{C}^{m \times m}$, the structured singular value [72, 79] is defined as

$$\mu_{\mathbf{\Delta}}(M) \triangleq \begin{cases} 0 \quad \text{if there exists no } \Delta \in \mathbf{\Delta} \text{ such that } \det(\mathrm{I} - M\Delta) = 0 \\ (\min_{\Delta \in \mathbf{\Delta}} \{ \bar{\sigma}(\Delta) \, : \, \det(\mathrm{I} - M\Delta) = 0 \})^{-1} \quad \text{otherwise} \end{cases} \tag{6.2}$$

in which more general classes of structured uncertainties can be handled without introducing any conservatism. Without loss of generality, the matrix M and each subblock of Δ have been assumed to be square. In this context, $\mu_{\mathbf{\Delta}}(M)$ defines a measure of the smallest structured $\Delta^\star \in \mathbf{\Delta}$ that destabilizes of the feedback interconnected system depicted in Figure 6.1, and the norm of this destabilizing uncertainty $\Delta^\star$ is quantified as $(\mu_{\mathbf{\Delta}}(M))^{-1}$.

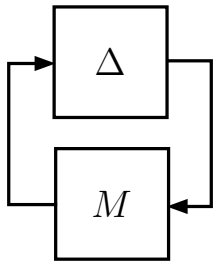


Figure 6.1: Feedback interconnected system.

6.2 Computational Complexity of Exact μ-calculation

6.2.1 NP-hardness of Purely Real μ-calculation [43, 44]

Braatz et al. [43,44] showed that, for a fixed k, the recognition problem "$q^\star_{\rm qp} \geq k$" for the quadratic program (QP)

$$q^\star_{\rm qp} := \max_{x \in \mathcal{X}_1} \left| x^{\rm T} A x + p^{\rm T} x + c \right| \geq k, \tag{6.3}$$

where $\mathcal{X}_1 \triangleq \{x \in \mathbb{R}^n : b^{\rm l} \leq x \leq b^{\rm u}\}$ polynomially reduces to a purely real μ recognition problem.

Theorem 6.1. [44, Thm. 2.1] Define the matrices

$$M^{\rm r}(k) \triangleq \begin{bmatrix} 0 & 0 & kw \\ kA & 0 & kA\bar{x} \\ \bar{x}^{\rm T}A + p^{\rm T} & w^{\rm T} & \bar{x}^{\rm T}A\bar{x} + p^{\rm T}\bar{x} + c \end{bmatrix}, \tag{6.4}$$

$$\hat{\mathbf{\Delta}}^{\rm r}_{\rm p} \triangleq \{{\rm diag}(\delta^{\rm r}_1, \cdots, \delta^{\rm r}_n, \delta^{\rm r}_1, \cdots, \delta^{\rm r}_n, \delta^{\rm c}) : \delta^{\rm r}_i \in \mathbb{R}, \delta^{\rm c} \in \mathbb{C}\}, \tag{6.5}$$

$$\tilde{\mathbf{\Delta}}^{\rm r}_{\rm p} \triangleq \{{\rm diag}(\delta^{\rm r}_1, \cdots, \delta^{\rm r}_n, \delta^{\rm r}_1, \cdots, \delta^{\rm r}_n, \delta^{\rm r}_{n+1}) : \delta^{\rm r}_i \in \mathbb{R}\}, \tag{6.6}$$

$$\bar{x} \triangleq \frac{1}{2}(b^{\rm u} + b^{\rm l}), \quad w \triangleq \frac{1}{2}(b^{\rm u} - b^{\rm l}). \tag{6.7}$$

Then, $\mu_{\hat{\mathbf{\Delta}}^{\rm r}_{\rm p}}(M^{\rm r}(k)) = \mu_{\tilde{\mathbf{\Delta}}^{\rm r}_{\rm p}}(M^{\rm r}(k))$ and, for a fixed constant $k > 0$,

$$\mu_{\hat{\mathbf{\Delta}}^{\rm r}_{\rm p}}(M^{\rm r}(k)) \geq k \Longleftrightarrow q^\star_{\rm qp} \geq k \tag{6.8}$$

In particular, the recognition version "$q \geq k$" of the QP given in (A.2) can polynomially reduce to the real μ recognition problem $\mu_{\hat{\mathbf{\Delta}}^{\rm r}}(M^{\rm r}(k)) \geq k$ with

$$M^{\rm r}(k) = \begin{bmatrix} 0 & 0 & \frac{1}{2}ke \\ k(rr^{\rm T} - {\rm I}) & 0 & \frac{1}{2}k(rr^{\rm T} - {\rm I})e \\ \frac{1}{2}e^{\rm T}(rr^{\rm T} + {\rm I}) - 2r_0 r^{\rm T} & \frac{1}{2}e^{\rm T} & \frac{1}{4}e^{\rm T}(rr^{\rm T} + {\rm I})e - r_0 r^{\rm T} e \end{bmatrix}$$

where $e = [1, \cdots, 1]^{\rm T} \in \mathbb{R}^n$. Since the QP (A.2) is an NP-hard problem, this implies that μ-calculation is NP-hard for both purely real and mixed real and complex uncertainties (see [44, Thms. 2.6 and 2.7]).

Poljak and Rohn's Approach [215]: (Independent Real Perturbations) Poljak and Rohn [215] considered the computational complexity of a problem concerning robust nonsingularity of matrices. For a square matrix $A \in \mathbb{R}^{n\times n}$ and a nonnegative matrix $0 \preceq D \in \mathbb{R}^{n\times n}$, the radius of nonsingularity is defined as

$$d(A, D) \triangleq \min\{k \geq 0 : \exists \text{a singular matrix } A_0 \text{ s.t.} A - kD \preceq A_0 \preceq A + kD\}.$$

Theorem 6.2. [215, Thm. 2.1] Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and $0 \preceq D \in \mathbb{R}^{n\times n}$. Then $d(A, D)$ can be represented as the maximal value

$$d(A, D) = \left(\max\{\rho_{\mathbb{R}}(A^{-1}\Delta_1 D\Delta_2) : \Delta_1, \Delta_2 \in \mathbf{B}\mathbf{\Delta}^{\mathrm{r}}\}\right)^{-1}.$$

The next corollary shows an explicit relation between computation of the radius of nonsingularity $d(\cdot, \cdot)$ and the real structured singular value μ. Its proof directly follows from the definition of μ.

Corollary 6.1. Let $A \in \mathbb{R}^{n\times n}$ be nonsingular and $0 \preceq D \in \mathbb{R}^{n\times n}$. Then the computation of $d(A, D)$ is equilvalent to a real μ-calculation:

$$d(A, D) = \left(\mu_{\hat{\mathbf{\Delta}}^{\mathrm{r}}}(N)\right)^{-2}$$

with $\hat{\mathbf{\Delta}}^{\mathrm{r}} \triangleq \{\mathrm{blkdiag}\{\Delta_1, \Delta_2\} \in \mathbb{R}^{2n\times 2n} : \Delta_1, \Delta_2 \in \mathbf{\Delta}^{\mathrm{r}}\}$ and $N \triangleq \begin{bmatrix} 0 & D \\ A^{-1} & 0 \end{bmatrix}$.

Proof. The definitions of real μ and the real spectral radius imply that

$$\begin{aligned}
\mu_{\hat{\mathbf{\Delta}}^{\mathrm{r}}}(N) &= \max\{\rho_{\mathbb{R}}(N\hat{\Delta}^{\mathrm{r}}) : \hat{\Delta}^{\mathrm{r}} \in \mathbf{B}\hat{\mathbf{\Delta}}^{\mathrm{r}}\} \\
&= \max_{\Delta_1,\Delta_2\in\mathbf{B}\mathbf{\Delta}^{\mathrm{r}}} \max\left\{|\lambda| : \lambda \in \mathbb{R},\ \det\left\{\begin{bmatrix} \lambda\mathrm{I} & -D\Delta_2 \\ -A^{-1}\Delta_1 & \lambda\mathrm{I} \end{bmatrix}\right\} = 0\right\} \\
&= \max_{\Delta_1,\Delta_2\in\mathbf{B}\mathbf{\Delta}^{\mathrm{r}}} \max\left\{|\lambda| : \lambda \in \mathbb{R},\ \det(\lambda\mathrm{I})\det(\lambda\mathrm{I} - \frac{1}{\lambda}A^{-1}\Delta_1 D\Delta_2) = 0\right\} \\
&= \max_{\Delta_1,\Delta_2\in\mathbf{B}\mathbf{\Delta}^{\mathrm{r}}} \max\left\{|\lambda| : \lambda \in \mathbb{R},\ \det(\lambda^2\mathrm{I} - A^{-1}\Delta_1 D\Delta_2) = 0\right\} \\
&= \sqrt{\max_{\Delta_1,\Delta_2\in\mathbf{B}\mathbf{\Delta}^{\mathrm{r}}} \rho_{\mathbb{R}}(A^{-1}\Delta_1 D\Delta_2)}
\end{aligned}$$

where the third equality follows from the Sylvester's theorem for determinants (see [109], for example), λ is assumed to be nonzero (for $d(A, D) = 0$, the proof is trivial.), and the last equality follows from the fact that $\Delta_i \in \mathbf{B}\mathbf{\Delta}^{\mathrm{r}}$ implies $-\Delta_i \in \mathbf{B}\mathbf{\Delta}^{\mathrm{r}}$ so that the real spectral radius of $A^{-1}\Delta_1 D\Delta_2$ is achieved by a positive eigenvalue. QED

A similar result of Corollary 6.1 was presented in [42] and is summarized here in Theorem 6.5. In order to show that the computation of $d(A, D)$ is NP-hard, Poljak and Rohn [215] considered the special case when $D = ee^{\mathrm{T}}$ where $e \in \mathbb{R}^n$ is the vector whose entries are all ones. For this case, it is not hard to see that the polytope $\mathbf{B}\mathbf{\Delta}^{\mathrm{r}}$ can be replaced by its vertices, without changing the optimal value, i.e., $d(A, ee^{\mathrm{T}}) = \left(\max\{z^{\mathrm{T}}A^{-1}y : z, y \in \{-1, 1\}^n\}\right)^{-1}$.

Theorem 6.3. [215, Thm. 2.5] Let $A \in \mathbb{R}^{n\times n}$ be a rational matrix. Then the computation of $d(A, D)$ is equivalent to the computation of the max-cut problem

$$d(A, ee^{\mathrm{T}}) = \left(2\mathrm{MC}(B_{A^{-1}}) - e^{\mathrm{T}}A^{-1}e\right)^{-1}$$

where B_M is the bipartite graph of a matrix M defined as the weighted bipartite $B_M = (Y \bigcup Z)$ where Y and Z are two copies of the index set $\{1, \ldots, n\}$.

Since the max-cut problem is NP-hard, computing $d(A, ee^{\mathrm{T}})$ is also NP-hard for any rational nonsingular matrix A.

Nemirovskii's Approach [195]: (Independent Real Perturbations) To show that it is NP-hard to determine whether all representatives of a square interval matrix set are all nonsingular, Nemirovskii [195] showed that the NP-complete knapsack problem A.4 can be reduced to this nonsingularity test.

Theorem 6.4. [195, Lemma 2.1] Consider the set of matrices

$$\mathcal{M}(k) \triangleq \left\{ \begin{bmatrix} C & z \\ y^{\mathrm{T}} & 1 \end{bmatrix} : y, z \in \mathbb{R}^n,\ \|y\|_\infty \leq 1/k,\ \|z\|_\infty \leq 1/k \right\} \tag{6.9}$$

where C is a rational matrix. For a given constant $k > 0$, the task of determining whether $\mathcal{M}(k)$ includes a singular matrix is NP-hard.

This approach was also applied by Braatz and Russell [42] to prove NP-hardness of the computation of μ with independent real perturbations.

Theorem 6.5. [42, Thm. 1] For a given constant $k > 0$, all matrices in $\mathcal{M}(k)$ are nonsingular if and only if

$$\mu_{\hat{\mathbf{\Delta}}^{\mathrm{r}}}(N) < k$$

where $\hat{\mathbf{\Delta}}^{\mathrm{r}} \triangleq \{\mathrm{blkdiag}\{\Delta_1, \Delta_2\} \in \mathbb{R}^{2n \times 2n} : \Delta_1, \Delta_2 \in \mathbf{\Delta}^{\mathrm{r}}\}$ and $N \triangleq \begin{bmatrix} 0 & C^{-1} \\ ee^{\mathrm{T}} & 0 \end{bmatrix}$.

Coxson and DeMarco's Approach [64] (Independent Real Perturbations) Coxson and DeMarco [64] generalized Nemirovskii's results [195].

Theorem 6.6. [64, Prop. 1 and Thm. 1] Consider the set of matrices $\mathcal{M}(1)$ defined in (6.9). All matrices in $\mathcal{M}(1)$ are nonsingular if and only if

$$\max\left\{ y^{\mathrm{T}} C^{-1} z : y, z \in \mathbb{R}^n,\ \|y\|_\infty \leq 1,\ \|z\|_\infty \leq 1 \right\} \geq 1.$$

Furthermore, the above recognition problem above can be transformed to a max-cut problem in polynomial-time, which implies that this problem is also NP-hard.

6.2.2 NP-hardness of Mixed μ-calculation

Since any μ-calculation containing real uncertainty is NP-hard, NP-hardness of the real μ recognition roblem directly implies the NP-hardness of mixed μ recognition. In other words, NP-hardness of mixed μ recognition can be considered as a corollary of NP-hardness of real μ recognition [42, Cor. 1]. Below is a separate proof for the NP-hardness of mixed μ recognition that can be extended to show the NP-hardness of μ for purely complex perturbations.

Consider a QCQP of the form

$$q^{\star}_{\mathrm{qcqp}} := \max_{x \in \mathcal{X}_2} \left| x^* A x + p^{\mathrm{T}} x + c \right| \tag{6.10}$$

where $A \in \mathbb{R}^{n \times n}$, $p \in \mathbb{R}^n$, $c \in \mathbb{R}$, and $\mathcal{X}_2 \triangleq \{x \in \mathbb{C}^n : 0 < b_i^{\mathrm{l}} \leq |x^* E_i x| \leq b_i^{\mathrm{u}},\ E_i = e_i e_i^{\mathrm{T}},\ i = 1, \ldots, m\}$. Now, we show that for a fixed constant $k > 0$, the recognition problem "$q^{\star}_{\mathrm{qcqp}} \geq k$" polynomially reduces to a mixed μ recognition problem.

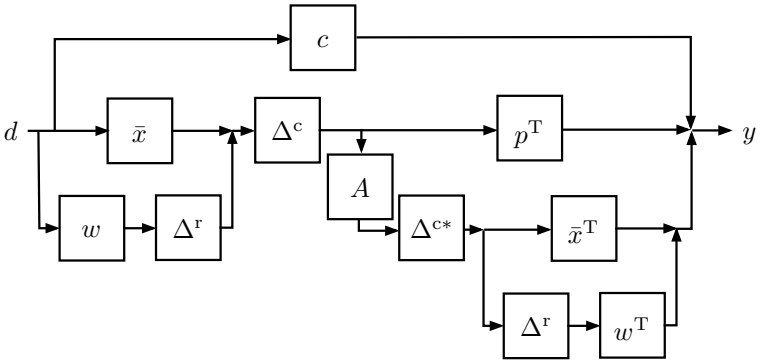


Figure 6.2: Equivalent block diagram for a QCQP.

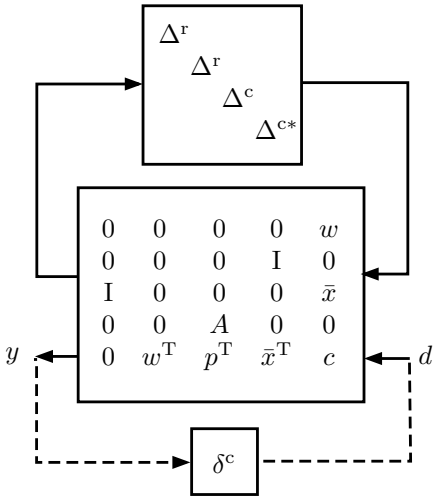
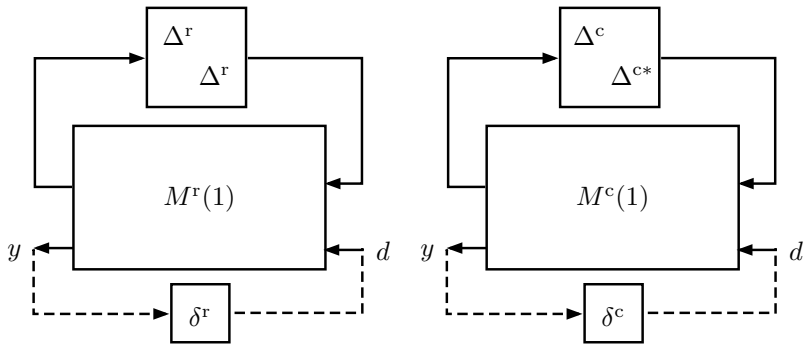


Figure 6.3: QCQP as a robustness problem.



(a) Repeated real uncertainties. (b) Repeated complex uncertainties.

Figure 6.4: Pure real and complex robustness problems.

Theorem 6.7. Define the matrices

$$M^{\mathrm{m}}(k) \triangleq \left[\begin{array}{cc|cc|c} 0 & 0 & 0 & 0 & kw \\ 0 & 0 & 0 & k\mathrm{I} & 0 \\ \hline k\mathrm{I} & 0 & 0 & 0 & k\bar{x} \\ 0 & 0 & kA & 0 & 0 \\ \hline 0 & w^{\mathrm{T}} & p^{\mathrm{T}} & \bar{x}^{\mathrm{T}} & c \end{array}\right], \tag{6.11}$$

$$\hat{\mathbf{\Delta}}^{\mathrm{m}}_{\mathrm{p}} \triangleq \{\mathrm{diag}(\delta^{\mathrm{r}}_{1:n}, \delta^{\mathrm{r}}_{1:n}, \delta^{\mathrm{c}}_{1:n}, \delta^{\mathrm{c}}_{1:n}, \delta^{\mathrm{c}}) : \delta^{\mathrm{r}}_i \in \mathbb{R}, \delta^{\mathrm{c}} \in \mathbb{C}\}, \tag{6.12}$$

$$\bar{x} \triangleq \frac{1}{2}(b^{\mathrm{u}} + b^{\mathrm{l}}), \quad w \triangleq \frac{1}{2}(b^{\mathrm{u}} - b^{\mathrm{l}}). \tag{6.13}$$

For any fixed constant $k > 0$,

$$\mu_{\hat{\mathbf{\Delta}}^{\mathrm{m}}_{\mathrm{p}}}(M^{\mathrm{m}}(k)) \geq k \Longleftrightarrow q^{\star}_{\mathrm{qcqp}} \geq k. \tag{6.14}$$

Proof. The proof is similar to the proof of the main theorem in [44, Thm. 2.1]. The constraint set can be rewritten as

$$\mathcal{X}_2 = \{x \in \mathbb{C}^n : x = \Delta^{\mathrm{c}}(\bar{x} + \Delta^{\mathrm{r}} w), \Delta^{\mathrm{r}} \in \mathcal{B}\mathbf{\Delta}^{\mathrm{r}}, \Delta^{\mathrm{c}} \in \mathcal{B}\mathbf{\Delta}^{\mathrm{c}}\}. \tag{6.15}$$

By defining dummy input and output $d, y \in \mathbb{R}$, the QCQP (6.10) can be represented as the block diagram in Figure 6.2 in which the optimization objective is the input-output relation between d and y. A simple block diagram manipulation produces its equivalent linear fractional transformation (LFT) in Figure 6.3, augmented with a performance block δ^{c} (following from the so-called main-loop theorem [208, Thm. 4.3]). Then

$$q^{\star}_{\mathrm{qcqp}} = \max_{\Delta \in \hat{\mathbf{\Delta}}^{\mathrm{m}}} |\mathcal{F}_{\mathrm{u}}(M^{\mathrm{m}}(1), \Delta)| \tag{6.16}$$

where $\hat{\mathbf{\Delta}}^{\mathrm{m}} \triangleq \{\mathrm{blkdiag}\{\Delta_1, \Delta_1, \Delta_2, \Delta_2\} : \Delta_1 \in \mathbf{\Delta}^{\mathrm{r}},\ \Delta_2 \in \mathbf{\Delta}^{\mathrm{c}}\}$. The remaining part of the proof is straightforward from the main loop theorem [208, Thm. 4.3]. QED

Remark 6.1. NP-hardness of mixed μ-calculation is a direct consequence of the fact that the QCQP prob-

lem (6.10) is NP-hard.

Remark 6.2. The mixed μ-calculation problem in Theorem 6.7 reduces to the real μ-calculation problem in Theorem 6.1 as a special case. Setting the complex uncertainty $\Delta^{\mathrm{c}} = \mathrm{I}$ in Figure 6.2 gives

$$M^{\mathrm{r}}(k) = \begin{bmatrix} M_{11}^{\mathrm{m}}(k) & M_{13}^{\mathrm{m}}(k) \\ M_{31}^{\mathrm{m}}(k) & M_{33}^{\mathrm{m}}(k) \end{bmatrix} + \begin{bmatrix} M_{12}^{\mathrm{m}}(k) \\ M_{32}^{\mathrm{m}}(k) \end{bmatrix} (\mathrm{I} - M_{22}^{\mathrm{m}}(k))^{-1} \begin{bmatrix} M_{21}^{\mathrm{m}}(k) & M_{23}^{\mathrm{m}}(k) \end{bmatrix}, \tag{6.17}$$

with the corresponding LFT in Figure 6.4(a).

6.2.3 NP-hardness of Complex μ-calculation

Toker and Özbay [267] used a similar approach as in [43,44] to show that μ computation with purely complex uncertainties is NP-hard. More specifically, they showed that a certain recognition problem to determine whether the inequality $\mu_{\boldsymbol{\Delta}}(M) < 1$ polynomially reduces to an NP-hard complex program.

Theorem 6.8. [267, Thm. 1] Consider the matrix

$$M = \left[\begin{array}{cc} P & 0 \\ 0 & e^{\mathrm{T}} \end{array}\right] \left[\begin{array}{cc|c} 0 & A & 0 \\ 0 & 0 & \mathrm{I} \\ \hline \mathrm{I} & 0 & 0 \end{array}\right] \begin{bmatrix} P^{-1} & 0 \\ 0 & e \end{bmatrix} \tag{6.18}$$

where $P = [e_1, e_3, \cdots, e_{2n-3}, e_{2n-1}, e_2, e_4, \cdots, e_{2n-2}, e_{2n}] \in \mathbb{R}^{2n \times 2n}$ and $e_k \in \mathbb{R}^{2n}$ are the standard basis vectors. Then the following recognition problems are equivalent:

$$\mu_{\boldsymbol{\Delta}}(M) < 1 \Longleftrightarrow \sup_{x_i \in \mathbb{C}, |x_i| \leq 1} |x^{\mathrm{T}} A x| < 1. \tag{6.19}$$

In [267], it was shown that a version of the NP-hard knapsack problem (see Problem A.5) can be written in the form of the complex program in the right-hand side of (6.19), which implies that pure complex μ-calculation is also NP-hard.

Remark 6.3. A special case of the mixed μ-calculation problem in Theorem 6.7 can be used to derive a purely complex μ-calculation problem. Setting the real uncertainty $\Delta^{\mathrm{r}} = \mathrm{I}$ gives

$$\begin{aligned} M^{\mathrm{c}}(k) &= \begin{bmatrix} M_{22}^{\mathrm{m}}(k) & M_{23}^{\mathrm{m}}(k) \\ M_{32}^{\mathrm{m}}(k) & M_{33}^{\mathrm{m}}(k) \end{bmatrix} + \begin{bmatrix} M_{21}^{\mathrm{m}}(k) \\ M_{31}^{\mathrm{m}}(k) \end{bmatrix} (\mathrm{I} - M_{11}^{\mathrm{m}}(k))^{-1} \begin{bmatrix} M_{12}^{\mathrm{m}}(k) & M_{13}^{\mathrm{m}}(k) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & k(\bar{x} + w) \\ kA & 0 & kp \\ p^{\mathrm{T}} & \bar{x}^{\mathrm{T}} + w^{\mathrm{T}} & c \end{bmatrix}, \end{aligned} \tag{6.20}$$

with the corresponding block diagram in Figure 6.4(b). Setting $k = 1$, $p = 0$, $c = 0$, and $\bar{x} = w = \frac{1}{2}e$ (corresponding to $b^{\mathrm{l}} = 0$ and $b^{\mathrm{u}} = e$) produces the same expression as the left-hand side of (6.19) but with

the objective function replaced by $|x^*Ax|$ in the right-hand side of (6.19). This relationship provides an alternative route to proving that the purely complex μ-calculation problem is NP-hard.

6.3 Computational Complexity of Approximate μ-calculation

This section considers the computational complexity of the problem of approximating the value of μ. A mathematical description of approximation problems is provided and a class of polynomial programs with box constraints is introduced whose exact computation is NP-hard and whose arbitrarily close approximation is also NP-hard. NP-hardness of the approximation problem for this class of polynomial programs is used to show that approximation problems for real and mixed μ-calculations are NP-hard [42].

6.3.1 Computational Complexity of ϵ-approximation Problems

Suppose that the set of feasible solutions $\mathcal{X}$ is non-empty and compact. Following the Weierstrass theorem [161], there exists $x_\star \in \mathcal{X}$ such that $c_\star \triangleq c(x_\star) \leq c(x)$ for all $x \in \mathcal{X}$. The *ϵ-approximation problem* for (A.1) is to compute a value $\hat{c} \in \mathbb{R}$ such that for a given $\epsilon > 0$,

$$|c^\star - \hat{c}| \leq \epsilon |c^\star - c_\star|. \tag{6.21}$$

This definition of ϵ-approximation is strongly preferred because it is invariant under translation and dilation of the objective function, and quantifying the optimization objective in different units does not affect the quality of the approximation as computed from (6.21) (see [18, 211, 276], for example).

A 0-approximation algorithm provides the exact optimal solution, while a 1-approximation algorithm need only find any feasible point and compute its objective. Now consider the existence of polynomial-time algorithms for computing an ϵ-approximation. Let n be a measure of the quantity of data needed to describe an instance of an optimization problem. To provide the strongest results, ϵ will be selected to be a function of n, which allows the consideration of algorithms for which the accuracy of the approximation degrades as the size of the problem (measured by n) increases.

For an optimization whose exact computation is NP-hard, the ϵ-approximation problem may or may not have an algorithm that computes an ϵ-approximation in polynomial-time as a function of problem size. An example is the knapsack problem (Problems A.4 and A.5). There exists an algorithm such that its solution can be approximated in polynomial-time within any factor of the optimum (see [214, Thm. 5.1] and [10, Thm. 7.4], for examples of randomized algorithms for solving the approximate knapsack problems), although its exact computation is known to be NP-hard.

The traveling salesmans problem is NP-complete and its ϵ-approximation problem is NP-hard [211, Thm.

17.10]. Another example is the class of polynomial programs with polytope (box, in particular) constraints

$$q^{\star}_{\text{poly}} := \max_{0 \le x_i \le 1} \left\{ \sum_{t=1}^{T} \left(\prod_{i \in \mathcal{I}_t^1} x_i \right) \left(\prod_{j \in \mathcal{I}_t^2} (1 - x_j) \right) \right\} \tag{6.22}$$

where T is the number of terms in the polynomial objective, n is the number of optimization variables, and for each t, $\bigcap \mathcal{I}_t^i = \emptyset$ and $\bigcup \mathcal{I}_t^i = \{1, \ldots, n\}$. The exact optimization problem is NP-hard and the ϵ-approximation problem is also hard. More specifically, finding a polynomial-time algorithm that approximates the optimum within a given constant is equivalent to establishing that P$\neq$NP as stated in the following lemma.

Lemma 6.1. [18, Thm. 3.1] There is a constant $\delta > 0$ such that if there exists a polynomial-time algorithm that can compute an ϵ-approximation for the optimization problem in (6.22) with $\epsilon = \epsilon(n) = 1 - n^{-\delta}$, then P=NP.

The general consensus in the computational community is that P$\neq$NP, which implies that an ϵ-approximation problem for polynomial programming with box constraints is also hard. Note that the particular form of $\epsilon(n)$ in Lemma 6.1 considers very weak forms of approximation, as it allows the quality of the approximation to degrade as a function of problem size. In particular, for fixed δ and large n, $\epsilon(n)$ approaches one, which only requires that the approximation algorithm be able to find a feasible point and evaluate the corresponding objective function. Thus Lemma 6.1 indicates that the existence of even a weak polynomial-time approximation algorithm for polynomial programs with box constraints is highly unlikely.

6.3.2 Approximate Real and Mixed μ-calculation

Braatz and Russell [42] showed that computing robust stability or performance margin within a given ϵ accuracy is a computationally expensive problem. To do this, they first showed that the polynomial program with polytope constraints in (6.22) can be represented as a skewed-μ problem.

Lemma 6.2. [42, Lemma 4] Consider the optimization problem in (6.22). For any fixed constant $k > 0$, the recognition problem "$q^{\star}_{\text{poly}} \ge k$" is polynomial-time reducible to the μ recognition problem "$\mu_{\mathbf{\Delta}^{\text{r}}}(M) \ge k$" for some rational matrix M and a set of real diagonal perturbations $\mathbf{\Delta}^{\text{r}}$.

The direct consequence is that the ϵ-approximation of robust stability or performance margins is computationally expensive.

Theorem 6.9. [42, Thm. 2] Consider the skewed-μ problem with any super set of uncertainties that includes either (i) all real scalar uncertainties or (ii) one complex and the rest real scalar uncertainties. There is a constant $\delta > 0$ such that if there exists a polynomial-time algorithm that can compute an ϵ-approximation for the skewed-mu with $\epsilon = \epsilon(n) = 1 - n^{-\delta}$, then P=NP.

Some parallel results were presented in the literature. In [64, Thms. 1 and 2], it was shown that there exists some arbitrarily small $\epsilon > 0$ such that the ϵ-approximation for real μ-calculation is NP-hard, unless P=NP. Toker [265, Thm. 3.1] showed that computing an ϵ-approximation for real μ-calculation with $\epsilon(n) = 1 - n^{-\delta}$ for some $\delta > 0$ is an NP-hard problem. In [88, Thm. 1.2], it was shown that for any $\epsilon > 0$, the ϵ-approximation for real μ-calculation is NP-hard, and its results were extended to any p-norm cases of real uncertainty [90, Thm. 2.2].

Unfortunately, all of the above parallel results employed definitions of ϵ-approximation that are not practically or numerically meaningful when applied to robustness margin problems. First consider the results of [88,90], which were based on the definition for the ϵ-approximation problem: $|c^\star - \hat{c}| \leq \epsilon |c^\star|$. The problem with this definition is that it is not invariant to scaling and dilation [276]. It has been well understood by the computational complexity community at least since the early 1990s that the definition of the ϵ-approximation problem for continuous optimization problems must be scaling invariant to be meaningful [276]. To illustrate this problem, consider any particular μ problem in which a very accurate approximation is available according to the definition $|c^\star - \hat{c}| \leq \epsilon |c^\star|$, let's say $|\mu_{\mathbf{\Delta}}(M) - \hat{\mu}| \approx 10^{-10} |\mu_{\mathbf{\Delta}}(M)|$ where $\hat{\mu}$ is our approximate value for $\mu_{\mathbf{\Delta}}(M)$ and $\mu_{\mathbf{\Delta}}(M) \approx 1$. A mathematically equivalent μ problem can be constructed that is exactly the same as the original μ problem, but with the matrix M modified to add a known constant of 10^{100} to the value of $\mu_{\mathbf{\Delta}}(M)$. The same procedure can be employed for the approximation, which adds the same constant to the value of $\hat{\mu}$. The computed accuracy for our mathematically equivalent approximation to the μ problem is $|\mu_{\mathbf{\Delta}}(M) - \hat{\mu}| \approx 10^{-10} |\mu_{\mathbf{\Delta}}(M) + 10^{100}| \approx 10^{10} |\mu_{\mathbf{\Delta}}(M)|$. According to this definition of the ϵ-approximation problem, one μ problem has a very highly accurate approximate solution with $\epsilon = 10^{-10}$ whereas a μ problem and its approximation that *are completely mathematically equivalent* to the original μ problem is indicated to have a very highly inaccurate approximate solution, with $\epsilon = 10^{10}$, using the same error definition.

While [64, 265] use a different ϵ-approximation definition than [88, 90], their definition also has a scaling problem, in that an equivalent μ problem with an additive large constant causes their definition is have vanishingly small errors regardless of the original value of ϵ. That is, a particular robustness margin problem will have a vanishingly small or absurdly large value for the error ϵ depending on the details of how the problem was mathematically represented rather than on the underlying problem.

6.3.3 Approximate Complex μ-calculation

Unlike real and mixed μ-calculation, it seems that there is no available affirmative results on the computational complexity of approximate complex μ-calculation. Readers are referred to [89, Prob. 22] for some discussions about this problem.

6.4 Approximate BMI optimization and μ-synthesis calculations

It is straightforward to show that ϵ-approximation for BMI optimization problems and μ-synthesis problems are computationally complex using the definition (6.21), following a very similar proof technique as used by [42] for μ-calculation.

6.5 Results on the Gap between Exact and Upper-bounds of μ-calculations

This section reviews results on the gap between the exact μ and its convex upper bounds. Before studying the conservatism of robust stability margin computation using convex upper bounds, some absolute stability criteria are reviewed that provide sufficient conditions for robust stability and whose conservatisms have been studied. A reason for studying absolute stability criteria and their conservatism is because they can be also used to compute or approximate upper bounds on exact μ and give lower bounds on robust stability margin computation.

6.5.1 Absolute Stability as Sufficient Conditions of Robust Stability

Suppose that a closed-loop system can be represented as a feedback interconnection of a linear time-invariant (LTI) system and a structured (block-diagonal) memoryless sector-bounded nonlinear operator, denoted by $M - \Gamma$. Its description will be the same as the $M - \Delta$ representation in Figure 6.1 where the only difference is that the uncertain operator Δ is now a sector-bounded memoryless (or static) nonlinear operator Γ. The closed-loop system is said to be *absolutely stable* if it is globally uniformly asymptotically stable at the origin for all memoryless nonlinearities in a given sector (see [136,279] for additional background on absolute stability and related works). The circle and Popov criteria provide sufficient conditions for absolute stability in terms of the strict positive realness of certain transfer functions, with these conditions representable in terms of linear matrix inequalities (LMIs). The conditions also provide graphical tests in the Nyquist plot for single-input single-output (SISO) systems. Note that a sufficient condition of robust stability provides a lower bound on the robust stability margin of uncertain systems.

Circle Criterion and Its Conservatism

Lemma 6.3. [136, Thm. 7.1] Consider the nonlinear operator $\Gamma \in \text{Sector}([0, k_i])$, $k_i > 0$, $i = 1, \ldots, n$. Then the feedback interconnected system $M(s) - \Gamma$ is absolutely stable if the transfer function $K^{-1} + M(s)$ is strictly positive real (SPR), where $K = \text{diag}\{k_1, \cdots, k_n\}$.

To simplify the presentation, suppose that $k_i = k > 0$ for all $i = 1, \ldots, n$. Define an optimal value by

$$\hat{\mu}_{\text{circ}}(M) \triangleq \left(\max\{k > 0 : \mathrm{I} + kM(s) \text{ is SPR}\}\right)^{-1} \tag{6.23}$$

whose inverse provides a lower bound on the robust stability margin for the $M-\Delta$ feedback interconnected system with the nonlinear operator Γ in the place of Δ, which will be denoted by $1/\mu^\star(M)$. Megretski [176, Problem 3] posed the question of conservatism of the circle criterion and [177, 224] provided an answer. In [224, Thm. 1], it was shown that there exists a sequence of matrices M_n whose dimension goes to infinity as $n \to \infty$ such that

$$\lim_{n\to\infty} \frac{\hat{\mu}_{\rm circ}(M_n)}{\mu^\star(M_n)} = \infty, \tag{6.24}$$

which implies that the circle criterion cannot be sharp up to a constant that does not depend on the order of a system. Note that the argument of [224] also holds for the cases when the operator Γ is replaced by $\Delta \in \mathbf{\Delta}^{\rm r}$ or $\mathbf{\Delta}^{\rm c}$. For such cases, $\mu^\star(M) = \max\left\{k>0 : \mu_{\mathbf{\Delta}}\left((2{\rm I}-kM)^{-1}M\right) \le 1/k\right\}$ where $\mathbf{\Delta} \in \{\mathbf{\Delta}^{\rm r}, \mathbf{\Delta}^{\rm c}\}$, which implies that there exists a sequence of matrices $M_n \in \mathbb{C}^{n\times n}$ such that

$$\lim_{n\to\infty} \frac{\hat{\mu}_{\rm circ}(M_n)}{\mu_{\mathbf{\Delta}}(M_n)} = \infty, \quad \mathbf{\Delta} \in \{\mathbf{\Delta}^{\rm r}, \mathbf{\Delta}^{\rm c}\}. \tag{6.25}$$

Popov Criterion and Its Conservatism The Popov criterion provides a less conservative stability test than the circle criterion by exploiting the memoryless (static) property of the nonlinear operator Γ and an LTI (dynamic) multiplier $({\rm I}+s\Lambda)^{-1}$ that does not change the sector boundedness.

Lemma 6.4. [136, Thm. 7.3] Consider the nonlinear operator $\Gamma \in {\rm Sector}([0, k_i])$, $k_i > 0$, $i = 1, \ldots, n$. Then the feedback interconnected system $M(s) - \Gamma$ is absolutely stable if a positive-semidefinite diagonal matrix $\Lambda = {\rm diag}\{\lambda_1, \cdots, \lambda_n\}$ such that the transfer function $K^{-1} + ({\rm I}+s\Lambda)M(s)$ is SPR, where $K = {\rm diag}\{k_1, \cdots, k_n\}$.

To simplify the presentation, suppose that $k_i = k > 0$ for all $i = 1, \ldots, n$. Similar to the circle criterion, define an optimal value by

$$\hat{\mu}_{\rm ppv}(M) \triangleq \left(\max_{\lambda_i \ge 0,\ \Lambda = {\rm diag}\{\lambda_i\}} \{k>0 : {\rm I} + k({\rm I}+s\Lambda)M(s) \text{ is SPR}\}\right)^{-1}. \tag{6.26}$$

whose inverse also provides a lower bound on the robust stability margin for the $M-\Delta$ feedback interconnected system with the nonlinear operator Γ in the place of Δ, which is denoted by $1/\mu^\star(M)$. It is not known whether there exists a sequence of matrices $M_n \in \mathbb{C}^{n\times n}$ such that

$$\lim_{n\to\infty} \frac{\hat{\mu}_{\rm ppv}(M_n)}{\mu^\star(M_n)} = \infty. \tag{6.27}$$

This Popov criterion can be extended to norm-bounded real parametric uncertain matrices and provides an upper bound on the structured singular value with real parametric uncertainties [26, 251]. From an equivalence between the Popov criterion and the D-scaling method [117], it is provable that the limit in (6.27) holds. Conservatism of the D-scaling upper bound will be discussed shortly in the next section.

6.5.2 Conservatism of Upper Bounds for Complex μ-calculation

To evaluate the conservatism of convex upper bounds of μ, consider the diagonal non-repeated complex uncertainty $\mathbf{\Delta}^{\mathrm{c}}$ and the D-scaling upper bound defined as

$$\hat{\mu}_{\mathbf{\Delta}^{\mathrm{c}}}(M) \triangleq \inf\left\{\bar{\sigma}(DMD^{-1}) : D \in \mathcal{D}_{\mathbf{\Delta}^{\mathrm{c}}}, D = D^* > 0\right\} \tag{6.28}$$

where $\mathcal{D}_{\mathbf{\Delta}^{\mathrm{c}}} \triangleq \{D \in \mathbb{C}^{n\times n} : D\Delta = \Delta D,\ \forall \Delta \in \mathbf{\Delta}^{\mathrm{c}}\}$.

In [208], many numerical simulations were performed to estimate the ratio of $\hat{\mu}_{\mathbf{\Delta}^{\mathrm{c}}}(M)/\mu_{\mathbf{\Delta}^{\mathrm{c}}}(M)$ over wide range of the dimension of M and [208] conjectured that this ratio is bounded by some constant that is independent of the system dimension. This conjection was later disproved. Define the worst-case ratio of μ and its upper bound by

$$r_n \triangleq \sup_{M\in\mathbb{C}^{n\times n}} \frac{\hat{\mu}_{\mathbf{\Delta}^{\mathrm{c}}}(M)}{\mu_{\mathbf{\Delta}^{\mathrm{c}}}(M)} \tag{6.29}$$

which is defined to be zero if both the numerator and denominator are zero. Megretski [175, Thm. 1] showed that $r_n \leq cn$ for some constant $c > 0$, which did not disprove or prove the conjecture. However, in [268, Thm. 1.0], it was proven that $\lim_{n\to\infty} r_n \to \infty$ and conjectured that the growth of r_n is sublinear with problem size n.

6.6 Alternative Approaches to Robustness Margin Computation

The above results indicate that both the exact computation of robustness margins is computationally expensive in worst-case, and that the approximate computation of robustness margins is either computationally expensive or can be conservative in worst-case. These theoretical results motivate the development of alternative approaches to compute or approximate robust stability margins. This section reviews two alternative approaches to those problems: (a) deterministic polynomial-time model reduction algorithms and (b) randomized algorithms using statistical learning theory.

6.6.1 Polynomial-time Model Reduction

As previously steted, the exact and approximate calculation of μ-calculation are hard problems. Furthermore, the μ-problems tend to involve poorly conditioned matrices as the size of the problem increases [225]. For those reasons, there have been several efforts to develop reduced models that have exactly or approximately the same robustness margins as the original model. In [100], four dimensionality reduction methods for uncertain systems were reviewed and their theoretical and computational characteristics were compared: (a) singular value decomposition (SVD)-based algorithms [225]; (b) successive realization algorithms [226]; (c) balanced truncation [15–17]; and (d) Kalman decomposition [68].

Among the model reduction methods for uncertain systems, this section focuses on the SVD-based algorithms as they behave in a numerically well-conditioned manner and are relatively easy to understand and implement. The key idea is to apply SVDs (aka principal component analysis (PCA) [183]) to remove the subspace components that do not affect the value of the robustness margin, or have negligible effects on it. The only reason for the complexity in writing the algorithm is to do remove these components while respecting the structure of the uncertainty.

6.6.2 Polynomial-time Probabilistic Randomized Algorithms

Proving that a problem is NP-complete or NP-hard avoids the investment of further futile efforts to solve this problem algorithmically to find an exact solution in polynomial time. One alternative is to search for an approximation to the exact solution that is *plausible* in polynomial time, where statistics are used to more preicsely define what is meant by *plausible*. Randomized algorithms for robustness analysis have been considered by many researchers in recent years, with a large increase in interest occurring once it was shown that the calculation of μ is NP-hard (see [11, 137, 253, 262, 281, 282], for example).

Polynomial-time randomized algorithms exist for computing approximate values for μ. Consider the feedback interconnected uncertain system in Figure 6.1 and a set of randomly generated uncertain matrices $\mathbf{\Delta}_{s(n)} \triangleq \left\{\Delta^{(1)}, \cdots, \Delta^{(s(n))}\right\} \subset \mathcal{B}\mathbf{\Delta}$ where n denotes the dimension of the M and $s(n)$ is the corresponding number of samples. Define a sequence of values $\check{\mu}_{s(n)}(M) \triangleq \max_{i=1}^{s(n)} \varrho(M\Delta^{(i)})$ where $\varrho(\cdot) = \rho(\cdot)$ for complex uncertainty set $\mathcal{B}\mathbf{\Delta}$, $\varrho(\cdot) = \rho_{\mathbb{R}}(\cdot)$ for real/mixed uncertainty set $\mathcal{B}\mathbf{\Delta}$. The $\check{\mu}_{s(n)}(M)$ are random variables and provide lower bounds on μ, $\check{\mu}_{s(n)}(M) \leq \mu_{\mathbf{\Delta}}(M)$ with probability one, for all $s(n) \in \mathbb{N}$ and $n \in \mathbb{N}$. Define a random sequence $q_{s(n)} \triangleq \frac{\check{\mu}_{s(n)}(M)}{\mu_{\mathbf{\Delta}}(M)}$ that is nondecreasing and less than or equal to 1 with probability one. Furthermore, $\lim_{s(n)\to\infty} q_{s(n)} = 1$ in probability, which implies that for all $\epsilon > 0$, $\text{Prob}_{\mathbf{\Delta}_{s(n)}}[q_{s(n)} \geq 1-\epsilon] \to 1$ as $s(n) \to \infty$. To estimate $\mu_{\mathbf{\Delta}}(M)$ from $\check{\mu}_{s(n)}(M)$, it is required to find the minimal number of samples $S(n)$ such that the following inequality is satisfied:

$$\text{Prob}_{\mathbf{\Delta}_{S(n)}}[q_{S(n)} \geq 1 - \epsilon] \geq 1 - \delta \tag{6.30}$$

where $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$ correspond to the accuracy and confidence, respectively, in the conclusion being made, based on the computations of $\check{\mu}_{S(n)}(M)$. A lower bound on such $S(n)$ is the Chernoff bound $S(n) \geq \frac{1}{2\epsilon^2} \log \frac{1}{\delta}$ [281]. This lower bound is uniform in the sense that it does not depend on the dimension of M, i.e., $S(n) = S$ for any $n \in \mathbb{N}$. Because the spectral radius and real spectral radius can be computed in polynomial time, this implies that this lower bound grows in polynomial time with the size of the robustness problem (that is, the dimension of M). The detailed theoretical backgrounds on and applications of statistical learning theory and randomized algorithms are beyond the scope of this chapter and the readers are referred to some research monographs [262, 274, 280] for more information on these topics.

Another interesting approach to stochastic robust analysis and control is a numerical approximation method called polynomial chaos (PC) expansion that was first introduced by Wiener [288] for turbulence modeling for uncertainties that are Gaussian random variables, which was later extended to other random variables [294]. In the polynomial chaos expansion framework and its extensions, the main idea is to use orthogonal basis functions to approximate the uncertainty propagation in terms of probability distributions by projecting the functional onto the space spanned by the set of bases [96, 167, 293]. Recently, many researchers have shown that spectral methods based on PC expansions can be computationally efficient approaches for control systems analysis and design, in which the parametric uncertainties are treated as random variables and the probability distributions of the system properties of consideration are approximated by determining the coefficients of associated basis functions [82, 84, 115, 116, 142, 143, 191, 192].

6.7 Summary and Future Work

A comprehensive overview is provided of research related to the computational complexity of robustness margin calculations. Followed by the pioneering papers on the structured singular value (μ), there have been numerous efforts to develop efficient algorithms for computing μ for purely real, mixed real and complex, and purely complex uncertainties. Results on the NP-hardness of the exact computation of μ motivated interest on the computational complexity of and potential conservatism in the approximation of μ. This chapter collects together many results that are not well known in the literature, including that the cost of μ calculation scales by the rank of the M matrix, and that in worst case the widely used upper bound for μ can be arbitrarily far off. The chapter also describes approaches for the extension of past results. The role of probabilistic randomized algorithms is also discussed, including their favorable scaling with problem size. Polynomial chaos expansion-based methods are described as a computationally efficient alternative for sampling-based stochastic robustness analysis and controller synthesis (aka Monte Carlo simulation).

Some open problems remain, such as the computational complexity of the ϵ-approximation of μ with purely complex uncertainties, and the degree of conservatism of other convex upper bounds for μ such as integral quadratic separators [121, 178]. Much of the literature in post 1990s moved into the development of numerical algorithms for robust control analysis and design based on linear and bilinear matrix inequalities, but a review of those developments would require another paper.

Part II

SPECTRAL METHODS FOR STOCHASTIC CONTROL

Chapter 7

Probabilistic Analysis and Control of Uncertain Systems

Abstract Uncertainties are ubiquitous in mathematical models of complex systems and this chapter considers the incorporation of generalized polynomial chaos expansions for uncertainty propagation and quantification into robust control design. Generalized polynomial chaos expansions are more computationally efficient than Monte Carlo simulation for quantifying the influence of stochastic parametric uncertainties on the states and outputs. Approximate surrogate models based on generalized polynomial chaos expansions are applied to design optimal controllers by solving stochastic optimizations in which the control laws are suitably parameterized, and the cost functions and probabilistic (chance) constraints are approximated by spectral representations. The approximation error is shown to converge to zero as the number of terms in the generalized polynomial chaos expansions increases. Several proposed approximate stochastic optimization problem formulations are demonstrated for a probabilistic robust optimal IMC control problem.

7.1 Introduction

Robust control theories [72, 308] analyze the stability and performance of uncertain systems against the worst case, which may have a vanishingly small probability of occurrence [281]. Analysis or design based on worst-case uncertainies can be too conservative to be applied in practice or may result in an overdesign of process equipment. In a practical point of view, it is rare that an engineer can exactly quantify hard bounds on the uncertainty, and a probabilistic description of uncertain parameters may be available instead. For probabilistic uncertain models, the conclusions on robustness are intrinsically stochastic and can be obtained in terms of a probability distribution or a level of confidence in estimates with probabilistic risk of incorrectness. Research monographs are available that describe probabilistic approaches to tackle robust

analysis and design problems for uncertain models [96, 167, 262, 293].

Most commonly used probabilistic analysis approaches are Monte Carlo (MC) methods in which many simulations are run with sampled uncertain parameters. The computational effort needed to propagate the uncertainty through the system model essentially amounts to simulating a large number of individual deterministic model realizations. While such an MC approach is applicable to many systems, the computational cost can be prohibitively expensive more complex systems, especially in real-time control implementations. The high computational cost of MC methods has motivated the development of computationally efficient methods for uncertainty propagation and quantification that replace or accelerate MC methods [96,167,293]. This chapter considers generalized polynomial chaos (gPC) expansions as a functional approximation and surrogate model of a mathematical model in the presence of uncertainty. Polynomial chaos expansions were first introduced by Wiener [288] for turbulence modeling for uncertainties that are Gaussian random variables, which was later extended to other random variables [294]. Recently, many researchers have shown that spectral methods based on PC expansions can be a computationally efficient alternative to MC approaches for control systems design [82, 84, 115, 116, 190–192].

Although uncertainty propagation and quantification using PC expansions has been extensively studied, the application of PC expansions to probabilistically robust and optimal controller design is a relatively new research topic. When the system parameters are assumed to be random variables and the exogenous disturbance are random processes, the solution and output trajectories are stochastic and controller design reduces to solving stochastic optimization problems with probabilistic constraints, also known as chance constraints. PC expansion approaches can be used to approximate chance constraints and we show the convergence of the approximation. Several examples and case studies are presented to illustrate PC expansion analysis and controller design of uncertain dynamical systems.

7.2 Probabilistic Uncertainty Quantification in Dynamic Systems

7.2.1 Uncertainty Quantification using gPC

The analysis of uncertainty propagation and quantification in models has several applications in systems engineering. The validation (or invalidation) of models against experiment data must take into account the effects of model uncertainties and measurement noise. In the presence of stochastic uncertainties, it is important to determine the probabilities of the system properties exceeding specified critical values or operation limits, and such evaluation can be used to conduct reliability and risk analyses. In a gPC approach for addressing these problems, the system model is replaced by a surrogate model whose solutions are represented by a gPC expansion and the surrogate model analyzed or simulated to quantify the propagation of uncertainty through the system.

7.2.2 Probabilistic Quantification of Uncertainty Propagation in System Performance

We study the probabilistic $\mathcal{H}_\infty$ and $\mathcal{H}_2$ performances of a transfer function between the external disturbance w_p and the induced output z_p.Let Δ be the support of the uncertainty δ and define a performance index $J : \Delta \to \mathbb{R}$ that is assumed to be measurable. For probabilistic robust performance analysis, the goal is to compute or approximate the probability distribution of $J(\delta)$.

Example 7.1 (*Probabilistic robust performance*). Consider the continuous-time linear parametric uncertain system introduced in [262, Ch. 6]

$$\begin{aligned} \frac{d}{dt}x(t) &= A(\delta)x(t) + Bw_p \\ z_p(t) &= Cx(t) \end{aligned} \tag{7.1}$$

with system matrices

$$A(\delta) = \begin{bmatrix} -2+\delta_1 & \delta_1\delta_2 \\ 0 & -4+\delta_2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

with the independent random variables δ_1, $\delta_2 \sim \mathcal{U}([-1, 1])$. Consider the quantification of uncertainty in system performances that are specified in terms of $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norms of the uncertain system transfer function $G(s;\delta) \triangleq C(s\mathrm{I} - A(\delta))^{-1}B$. Define the performance indices (a) $J_2(\delta) \triangleq \|G(s;\delta)\|_2^2$ and (b) $J_\infty(\delta) \triangleq \|G(s;\delta)\|_\infty^2$. Figs. 7.2(a) and 7.2(b) compare the Legendre PC expansion (L-PCE) approach with a heuristic MC method, where 10,000 samples of the random vector δ was used to compute the first and second moments, and to generate the histograms. The order of the L-PCE was set to 5 and the coefficients were determined from non-intrusive least-squares minimization with 100 importance samples [167, Sec. 3.2]. The L-PCE is indistinguishable from the Monte Carlo method in quantifying the PDF of the $\mathcal{H}_2$- and $\mathcal{H}_\infty$-norms.

Remark 7.1. Similar approaches can be applied to the analysis of frequency-dependent specifications for a system transfer function.

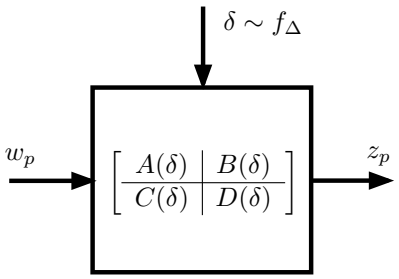
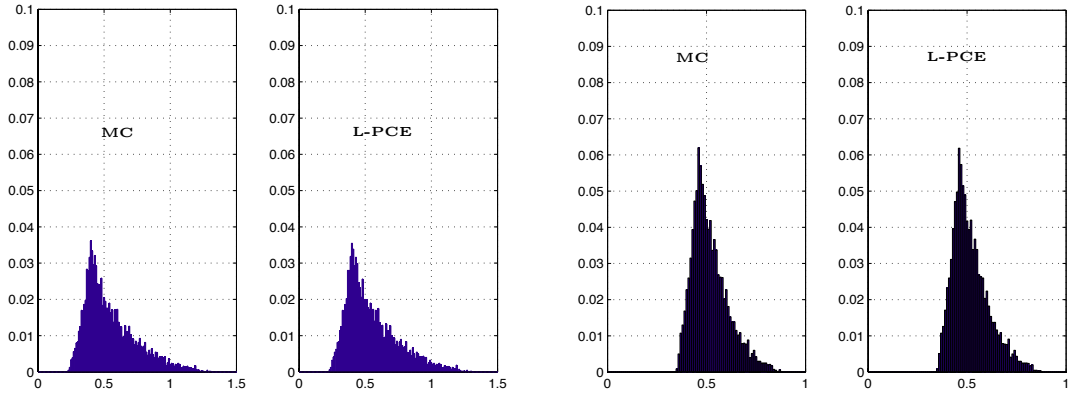


Figure 7.1: Probabilistic analysis of robust performance.



(a) $\mathcal{H}_2$-norm (b) $\mathcal{H}_\infty$-norm

Figure 7.2: PDF of system performances.

7.2.3 Spectral Analysis of Uncertainty Propagation under Differential Equations

Consider a general stochastic differential equation (SDE) of the form

$$L(z,t,\theta;y) = g(z,t,\theta) \tag{7.2}$$

where $z \in \mathcal{Z}$ and $t \in \mathcal{T}$ are the spatial and temporal variables, $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ is the concatenation of the random variables, the function $g : \mathcal{Z} \times \mathcal{T} \times \Theta \to \mathbb{R}$ is a forcing term, and $y : \mathcal{Z} \times \mathcal{T} \to \mathbb{R}$ is a solution of the equation, which also defines a random field over the spatial and temporal spaces $\mathcal{Z} \times \mathcal{T}$ due to the random variable vector θ. Suppose that there exists a bijective transformation (not necessarily diffeomorphism) $T : \Theta \to \Xi$ such that $\zeta(\theta;\omega) = T(\theta(\omega))$ for all $\theta \in \Theta$ and $\omega \in \Omega$ and the transformed random variable ζ is a standard (optimal in the sense of convergence rate) random variable for the set of polynomial basis functions $\{\phi_i\}$. For application of the spectral method based on polynomial chaos expansions, assume that the solution of the SDE (A.62) has the form

$$y \approx y^{N_p} \triangleq \sum_{i=1}^{N_p} y_i(z,t)\phi_i(\zeta(\theta;\omega)) \tag{7.3}$$

which is an approximation of the true solution y with N_p+1 basis functions from the set $\{\phi_i\}$. Obtaining the approximated solution y^{N_p} requires determining the spatial- and temporal-varying deterministic coefficients $y_i(z,t)$. To do this, substitute the approximation y^{N_p} to y of the SDE (A.62)

$$L\left(z,t,\theta;\sum_{i=1}^{N_p} y_i(z,t)\phi_i(\zeta(\theta;\omega))\right) = g(z,t,\theta) \tag{7.4}$$

and solve for the spatial- and temporal-dependent coefficients $y_i(z,t)$ by intrusive or non-intrusive projections onto the probability space of the random variable θ or ζ.

7.2.4 Case Study: Risk Analysis of Optimal Dynamic Portfolio Selection

Consider a capital market with $n+1$ risky securities with random rates of returns $\{r^i\}$, where the superscript i denotes the ith random return rate. The rate of returns of the risky securities at time instance t are denoted by the concatenated vector $r_t \triangleq [r_t^0, \cdots, r_t^n]^{\mathrm{T}}$. Assume that the probabilistic distribution of the random process r_t is known. The wealth dynamics [310] is given as

$$x_{t+1} = \sum_{i=1}^{n} r_t^i u_t^i + \left(x_t - \sum_{i=1}^{n} u_t^i \right) r_t^0 \tag{7.5}$$

where x_t is the wealth of the investor at time instance t and u_t^i is the amount invested in the ith risky asset at time instance t. The wealth dynamics (7.5) can be rewritten as

$$x_{t+1} = r_t^0 x_t + R_t^{\mathrm{T}} u_t \tag{7.6}$$

where $R_t \triangleq [r_t^1 - r_t^0, \cdots, r_t^n - r_t^0]^{\mathrm{T}}$ and $u_t \triangleq [u_t^1, \cdots, u_t^n]^{\mathrm{T}}$.

Further assume a stationary multiperiod process with the period $t_{\mathrm{f}} = 4$ and that the rate of return r_t is time-invariant in that period, i.e., $r_t = r$ for all $t = 0, \ldots, t_{\mathrm{f}} - 1$. The probabilistic rate of return r is assumed to be a Gaussian random variable with mean $\bar{r}$ and covariance matrix Σ_r, i.e., $r \sim \mathcal{N}(\bar{r}, \Sigma_r)$. Consider an investment policy given as $u_t = -K_t x_t + v_t$ where $K_t = [1.6238\ 4.2907]^{\mathrm{T}}$ for all $t = 0, \cdots, 3$ and $v_0 = [4.3548\ 11.9327]^{\mathrm{T}}$, $v_1 = [5.1094\ 14.0004]^{\mathrm{T}}$, $v_2 = [5.9948\ 16.4263]^{\mathrm{T}}$, and $v_3 = [7.0035\ 19.2726]^{\mathrm{T}}$. Figure 7.3 shows the probabilistic uncertainty propagation of the wealth dynamics using a Monte Carlo simulation and the PC expansion method based on the Hermite polynomial and the Galerkin projection. The distributions are indistinguishable.

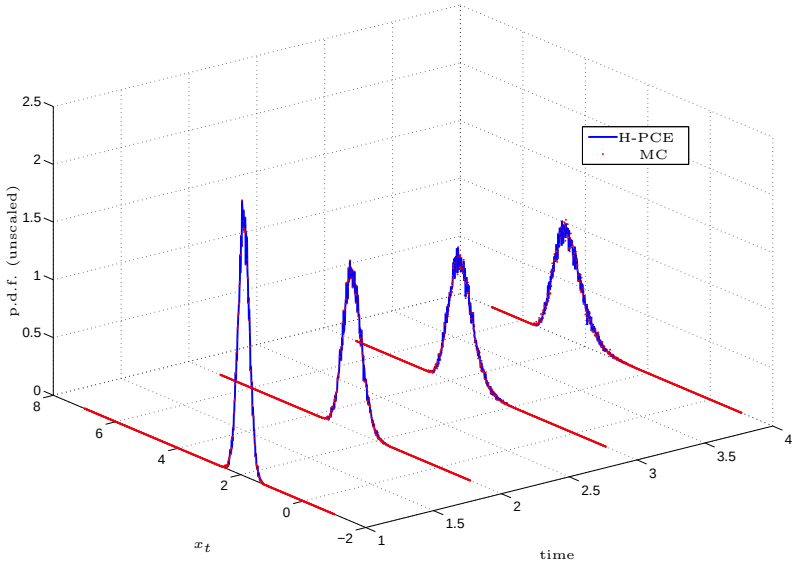


Figure 7.3: Probabilistic distribution and time propagation of the wealth of the investor.

7.3 Approximation of Probabilistic Bounds

7.3.1 Probabilistic (Chance) Constraints

In systems and control theory, most decision processes can be written in terms of finding feasible solutions for constraints in the presence of uncertainty. For instance, a requirement of the system properties might be represented as a nonlinear inequality $h(x, \delta) \leq \gamma$ where x is the design variable and $\delta \in \Delta$ corresponds to the concatenation of internal and environmental uncertainty. In a deterministic worst-case approach, a feasible solution x must satisfy the constraint $h(x, \delta) \leq \gamma$ for all $\delta \in \Delta$. This approach can be too conservative in that it does not allow for a single realization of uncertainty δ to violate the constraint and may result in there being no feasible solution x for the problem. Alternatively, probabilistic analysis considers the probabilistic risk of violation of the constraints in terms of the probabilistic constraint $\mathbf{Pr}_{\Delta}[h(x, \delta) \leq \gamma] \geq \beta$ where δ is a random variable with known probability distribution and $\beta \in [0, 1]$ corresponds to the level of confidence in feasibility of the constraint.

It is not, in general, a trivial problem to evaluate the quantity $\mathbf{Pr}_{\Delta}[h(x, \delta) \leq \gamma]$ for a fixed value $x \in \mathcal{X}$ unless the nonlinear function $h : \mathcal{X} \times \Delta \to \mathbb{R}$ has a simple form such as a linear function in both arguments, even though the probability distribution of δ is given. gPCE approaches can be used to approximate the probabilistic feasibility constraint $\mathbf{Pr}_{\Delta}[h(x, \delta) \leq \gamma]$ for each fixed $x \in \mathcal{X}$. Its convergence is demonstrated and illustrated below.

7.3.2 Approximation of Chance Constraints

The simplest approach to evaluate or approximate probabilistic constraints is Monte Carlo simulation in which the probability of feasibility is computed from

$$p_h(x, \gamma) \triangleq \mathbf{Pr}_{\Delta}[h(x, \delta) \leq \gamma] \approx \frac{1}{N_s} \sum_j I_{\{h(x, \delta^j) \leq \gamma\}} \tag{7.7}$$

where N_s is the number of samplings and $I_{\{.\}}$ is the indicator function that is 1 if a sampled uncertain parameter δ^j satisfies the constrained $h(x, \delta^j) \leq \gamma$ and 0 otherwise.

An alternative to the Monte Carlo approach for the approximation of chance constraints is gPCE-based spectral approximation in which the nonlinear uncertainty propagation is approximated by a gPCE and the probabilistic feasibility of constraints is evaluated with a relatively cheap computational cost compared to Monte Carlo simulation. The next theorem shows that such approximation of probabilistic bounds using gPCEs converges to the true probability of feasibility.

Theorem 7.1. Define the function $F(x, \gamma) \triangleq \mathbf{Pr}_{\Delta}[h(x; \delta) \leq \gamma]$. For any $\gamma \in \mathbb{R}$ and $h(x; \cdot) : \Delta \to \mathcal{Y} \subset \mathbb{R}$

such that $h(x;\cdot) \in \mathcal{L}_2(\Delta)$ for any $x \in \mathcal{X}$, there exist sets of coefficients $\{a_i\}$ and polynomials $\{\phi_i\}$ such that

$$\lim_{n\to\infty} \left| F(x,\gamma) - \hat{F}_n(x,\gamma) \right| = 0 \tag{7.8}$$

holds pointwisely in $x \in \mathcal{X}$, where $\hat{F}_n(x,\gamma) \triangleq \mathbf{Pr}_\Delta[h_n(x;\delta) \triangleq \sum_{i=0}^n a_i(x)\phi_i(\delta) \leq \gamma]$.

Proof. The proof is straightforward. From Theorem A.1, for any $h \in \mathcal{L}_2(\Delta)$, there exists a gPCE of the form $h_n(x;\delta)$ such that $h_n \to_{m.s.} h$. The fact that m.s.-convergence implies the d-convergence (convergence in distribution) (see [252, Ch. 6.7] for the definitions and relations of probabilistic convergence), implies that $F(x,\gamma) \to \hat{F}_n(x,\gamma)$ as $n \to \infty$ for any fixed $x \in \mathcal{X}$. QED

The convergence rate depends on the smoothness (differentiability) of the function $h(x;\cdot)$.

In general the computation or approximation of the probability $\mathbf{Pr}_\Delta[h(x;\delta) \leq \gamma]$ still requires generating random samples, since no analytical solution exists in general. A more computationally efficient approach is to exploit a useful property of gPCE: All of the moments for a gPCE have analytical forms in terms of the coefficients. In particular, we use the well-known Chebyshev inequality $\mathbf{Pr}[|X - \mathbf{E}[X]| \geq \lambda] \leq \frac{\mathbf{Var}[x]}{\lambda^2}$ or $\mathbf{Pr}[X \geq \lambda] \leq \frac{\mathbf{Var}[x]}{(\mathbf{E}[X]-\lambda)^2}$.

Example 7.2. To illustrate the proposed approach, consider the uncertain system (7.1) given in Example 7.1. We compute the probabilities $\mathbf{Pr}_\Delta[\|G(s;\delta)\|_2^2 \leq \gamma_2]$ and $\mathbf{Pr}_\Delta[\|G(s;\delta)\|_\infty^2 \leq \gamma_\infty]$ for $\gamma_2 = 1.00$ and $\gamma_\infty = 0.75$. Table 7.1 shows the approximation of these values using different approaches for an order of 5 for the L-PCE. For the third approach using the computed coefficients of the L-PCE, the Chebyshev inequality gives a lower bound for the probability that is very tight for the H_2-norm and somewhat less tight for the H_∞-norm.

7.4 Optimal Controller Design with Chance Constraints

7.4.1 Parameterization of Control Laws

This section introduces three different approaches to controller design using gPCE methods for parametric uncertain systems that are illustrated for a specific application to provide a clear comparison between approaches.

Approaches	$\mathbf{Pr}_\Delta[\|G(s;\delta)\|_2^2 \leq \gamma_2]$	$\mathbf{Pr}_\Delta[\|G(s;\delta)\|_\infty^2 \leq \gamma_\infty]$
MC	0.9691	0.9800
PCE+MC	0.9691	0.9801
PCE+Coeff	≥ 0.9641	≥ 0.9574

Table 7.1: Approximation of probabilistic bounds.

Quantization of control input Parameterize an open-loop control law by

$$u(z,t) = \sum_{i=1}^{M} u_i \mathcal{I}_{\chi_i}(z,t) \tag{7.9}$$

where

$$\mathcal{I}_{\chi_i}(z,t) \triangleq \begin{cases} 1 & (z,t) \in \chi_i \\ 0 & (z,t) \notin \chi_i \end{cases} \tag{7.10}$$

such that $\bigcup_i \chi_i = \chi \subset \mathcal{Z} \times \mathcal{T}$ is a compact set.

Example 7.3 (*Control vector parameterization (CVP)*). CVP is a simple and popular approach to solving infinite-dimensional optimization problems that simultaneously optimizes multiple parameters describing a spatial profile of control inputs (variables). Consider a model-based optimal porosity distribution control problem that minimizes ohmic drop across a porous electrode in a lithium-ion battery. A mathematical model of a porous electrode is [92, 216]

$$\begin{aligned} \frac{di_1}{dz} &= i_0 \frac{F}{RT} \frac{3(1-\nu(z))}{R_p} (h_1(z) - h_2(z)), \\ \frac{dh_1}{dz} &= -\frac{1}{\sigma_0 (1-\nu(z))^b} i_1(z), \\ \frac{dh_2}{dz} &= \frac{1}{\kappa_0 \nu(z)^b} \left(\sigma_0 (1-\nu(z))^b \left. \frac{dh_1}{dz} \right|_{z=0} - i_1(z) \right). \end{aligned} \tag{7.11}$$

The boundary conditions for solution of these equations are given by

$$h_1|_{z=0} = 1, \;\; h_2|_{z=\ell_p} = 0, \;\; i_1|_{z=\ell_p} = 0. \tag{7.12}$$

In CVP, the control variable $\nu(z)$ is parameterized by a finite number of parameters by partitioning its values over the spatial domain, i.e.,

$$\nu(z) = \sum_{i=1}^{M} \nu_i \mathcal{I}_{\mathcal{Z}_i}(z) \tag{7.13}$$

where $\mathcal{Z}_i$ is the partition of the interval $[0, \ell_p]$ satisfying $\bigcup_i \mathcal{Z}_i = [0, \ell_p]$ and $\mathcal{Z}_i \bigcap \mathcal{Z}_j = \emptyset$, and $\mathcal{I}_{\mathcal{Z}_i}(z)$ is the indicator function which is 1 for $z \in \mathcal{Z}_i$ and 0 otherwise. This parameterization reduces the optimization to being finite-dimensional:

$$\begin{aligned} &\min_{\{\nu_i\}_{i=1}^M} \; J \\ &\text{subject to (7.11) with (7.12)}, \\ &\qquad 0 < \nu_i < 1, \; i = 1, \ldots, M \end{aligned} \tag{7.14}$$

where the cost function J is set are the battery design objective to be minimized and the spatially quantized

control variable is (7.13).

For robust optimization, consider the model parameters $\theta \triangleq [F, R, T, i_0, R_p, \sigma_0, \kappa_0, b]$ that are assumed to be random variables. A common robust optimization objective is the weighted sum of the mean and variance, $\mathbf{E}[J(\boldsymbol{\nu};\theta)]+w\mathbf{Var}[J(\boldsymbol{\nu};\theta)]$, where $\boldsymbol{\nu}$ is the concatenation of the decision variables $\{\nu_i\}_{i=1}^{M}$ and $w > 0$ is a user-defined weight specifying the tradeoff between nominal and robust performance. Alternatively, J in (7.14) can be replaced by $\mathbf{E}[J(\boldsymbol{\nu};\theta)]$ and an additional constraint $\mathbf{Var}[J(\boldsymbol{\nu};\theta)] \leq \beta_J$ can be introduced where β_J is a user-defined bound on the variance of the cost that plays the same role as w.

Spectral representation of control input Consider the open-loop control law

$$u(z,t) = \sum_{i=1}^{M} u_i \psi_i(z,t) \tag{7.15}$$

where $\{\psi_i\}$ is a set of appropriately chosen basis functions. Note that elements of the set $\{\psi_i\}$ could be neither orthogonal nor orthonormal.

Example 7.4 (*CVP based on sinusoidal basis functions*). Consider the same mathematical model given in (7.11) and a spectral representation of the control variable $\nu(z) = \sum_{k=1}^{M} \nu_k \sin\left(\frac{2\pi}{k} z + \nu_{M+k}\right)$. With this finite parameterization of the control variable as $\{\nu_i\}_{i=1}^{2M}$, formulation of the robust optimization based on PCE is similar to Ex. 4.

Fixed structure of controllers Consider a closed-loop control law

$$u(z,t) = k(y(z,t); u_i) \tag{7.16}$$

where y denotes the measured output of the system, the mapping $k : \mathcal{Y} \to \mathcal{U}$ is a feedback control law, and the u_i are the parameters of the controller. With this finite parameterization of the control variable, formulation of robust optimizations based on PCE is similar to Ex. 4.

7.4.2 Case Study: IMC-based Robust Optimal Controller Design

This section illustrates probabilistic controller design approaches based on gPC expansions to achieve robust performance and stability in the presence of parametric uncertainties. Three problem formulations for stochastic programming are considered: (a) *probabilistically guaranteed cost*; (b) *probabilistic cost minimization*; and (c) *mean-variance optimization*. We also present approximate stochastic programmings based on gPC to handle chance constraints and probabilistic bounds on cost functions for the three different stochastic optimizations.

Some industrial processes are described by a first-order biproper nonminimum-phase (FBNP) model

$$G(s) = \frac{k_1(1-\tau_d s)}{1+\tau_c s} \tag{7.17}$$

where each parameter is assumed to be uncertain, i.e., $k_1 = k_1^o(1+\rho\delta_1)$, $\tau_d = \tau_d^o(1+\rho\delta_2)$, and $\tau_c = \tau_c^o(1+\rho\delta_3)$ with the percentile of uncertainty $\rho \in [0,1]$ and the uncertainty $\delta_i \in [-1,1]$, $i = 1,2,3$.

Robustness considerations can be taken into account by posing the design problem as the min-max optimization:

$$\begin{aligned} &\min_{C(s)} \max_{\delta\in\Delta} \ \left\|W_1(s)(M(s)-G_c(s))\right\|_\infty \\ &\text{subject to } \ \left\|W_2(s)T(s)\right\|_\infty < 1, \quad \forall \delta \in \Delta \end{aligned} \tag{7.18}$$

where $M(s)$ is a reference model for the closed-loop system response, $G_c(s)$ is the closed-loop system response with the uncertain plant $G(s)$ and controller $C(s)$, $W_2(s)$ is a weighting transfer function for robust stability, $T(s)$ is the complementary sensitivity function, and $\Delta \triangleq [-1,1]^3$ is the set of parametric uncertainties. The $\mathcal{H}_\infty$-norm of the irrational transfer functions in (7.18) was computed using the third-order Padé approximation.

For design of a controller solving the optimization (7.18), consider an internal model controller [185]

$$C(s) \triangleq \frac{J(s)}{G_{om}(s)(1-J(s))} \tag{7.19}$$

where $G_{om}(s) \triangleq \frac{k_1^o}{1+\tau_c^o s}$ is the minimum-phase part of the nominal system transfer function $G_o(s) \triangleq \frac{k_1^o(1-\tau_d^o s)}{1+\tau_c^o s}$ and $J(s) \triangleq \frac{1}{(1+\tau s)^k}$ is a low-pass filter with two tuning parameters $\tau > 0$ and $k \in \mathbb{N}$. With this parameterization of controllers, the infinite-dimensional optimization (7.18) reduces to a finite-dimensional optimization in which the decision variables are τ and k. Now define two nonlinear functions $f_1(\tau,k;\delta) \triangleq \left\|W_1(s)(M(s)-G_c(s))\right\|_\infty$ and $f_2(\tau,k;\delta) \triangleq \left\|W_2(s)T(s)\right\|_\infty$ and assume that the uncertain parameters δ_i are uniformly distributed independent random variables over the support Δ.

Probabilistically guaranteed cost For this problem, the goal is to ensure a certain level of achievement for the cost function (denoted by γ) with minimized probability of violation (denoted by $1-\beta_1$) while satisfying the constraint of robust stability with a given probability (denoted by β_2). Its mathematical description is given by

$$\begin{aligned} &\max_{\tau>0,k\in\mathbb{N}} \ \beta_1 \\ &\text{subject to } \mathbf{Pr}_\Delta[f_1(\tau,k;\delta) \geq \gamma] \leq 1-\beta_1, \\ &\qquad\qquad \mathbf{Pr}_\Delta[f_2(\tau,k;\delta) \geq 1] \leq 1-\beta_2. \end{aligned} \tag{P1}$$

The integer variable can be removed by fixing the order of the low-pass filter k and using $\tau > 0$ as the tuning parameter. Now replace the nonlinear functions f_j by their approximations using the multivariate Legendre PCEs:

$$f_j(\tau;\delta) \approx \hat{f}_j(\tau;\delta) \triangleq \sum_{i=0}^{N_p^j} a_i^j(\tau) L_i(\delta) \tag{7.20}$$

where L_i are the Legendre polynomials.

With this approximation of uncertainty propagation under nonlinear functions, solve an approximate optimization

$$\text{(P1}') \qquad \begin{aligned} &\max_{\tau>0} \ \beta_1 \\ &\text{subject to } \hat{h}_1(\tau;\gamma) \le 1-\beta_1,\ \hat{h}_2(\tau) \le 1-\beta_2, \end{aligned}$$

where $\hat{h}_1(\tau;\gamma) \triangleq \mathbf{Pr}_\Delta[\hat{f}_1(\tau;\delta) \ge \gamma]$ and $\hat{h}_2(\tau) \triangleq \mathbf{Pr}_\Delta[\hat{f}_2(\tau;\delta) \ge 1]$.

Probabilistic cost minimization problem For this problem, the bound γ on the cost function is minimized with a user-defined probabilistic violation $1-\beta_1$ while satisfying the constraint of robust stability with probability β_2:

$$\text{(P2)} \qquad \begin{aligned} &\min_{\tau>0,k\in\mathbb{N}} \ \gamma \\ &\text{subject to } \mathbf{Pr}_\Delta[f_1(\tau,k;\delta) \ge \gamma] \le 1-\beta_1, \\ &\qquad\qquad \mathbf{Pr}_\Delta[f_2(\tau,k;\delta) \ge 1] \le 1-\beta_2. \end{aligned}$$

This optimization can be considered as a probabilistic relaxation of the worst-case optimization and feasible solutions with confidence levels $\beta_1 = \beta_2 = 1$ correspond to the deterministic solutions in (7.18).

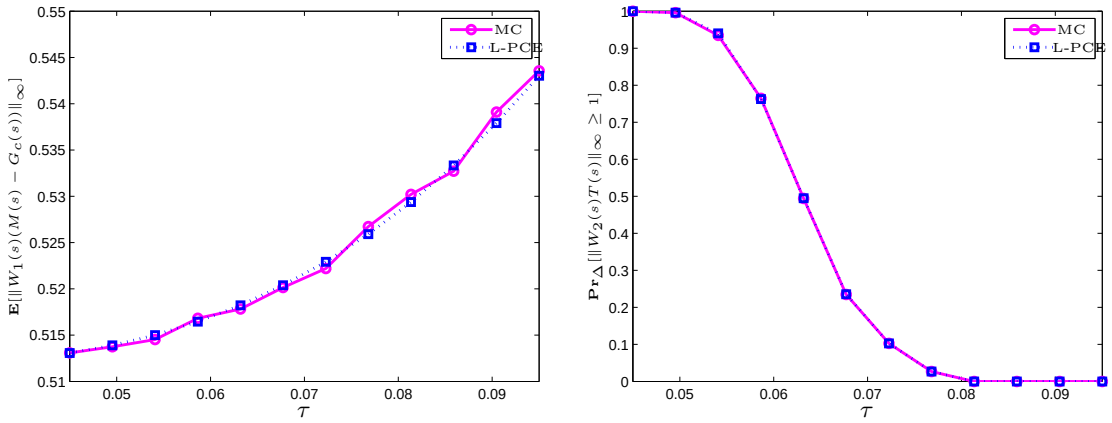
For a fixed order of the low-pass filter k, $\tau > 0$ is the only decision variable and the multivariate Legendre PCE (7.20) can be used to approximate the chance constraints in (P2):

$$\text{(P2}') \qquad \begin{aligned} &\min_{\tau>0} \ \gamma \\ &\text{subject to } \hat{h}_1(\tau,\gamma) \le 1-\beta_1,\ \hat{h}_2(\tau) \le 1-\beta_2, \end{aligned}$$

where $\hat{h}_1(\tau,\gamma)$ and $\hat{h}_2(\tau)$ are previously defined.

Mean-variance optimization In this problem, the weighted mean-variance cost is minimized while satisfying the constraint of robust stability with probability β. Its mathematical description is

$$\text{(P3)} \qquad \begin{aligned} &\min_{\tau>0,k\in\mathbb{N}} \ \mathbf{E}[f_1(\tau,k;\delta)] + w\mathbf{Var}[f_1(\tau,k;\delta)] \\ &\text{subject to } \mathbf{Pr}_\Delta[f_2(\tau,k;\delta) \ge 1] \le 1-\beta. \end{aligned}$$



(a) Expectation of performance

(b) Probability of infeasibility

Figure 7.4: Comparison of L-PCE and MC approaches that can be used to trade off two objectives in stochastic optimization in Ex. 7.5.

Same as before, fix the order of the low-pass filter k and use $\tau > 0$ as the only decision variable. The multivariate Legendre PCE (7.20) can be used to approximate the chance constraints in (P3):

$$\text{(P3}') \qquad \begin{aligned} \min_{\tau>0} \quad & \mathbf{E}[\hat{f}_1(\tau;\delta)] + w\mathbf{Var}[\hat{f}_1(\tau;\delta)] \\ \text{subject to} \quad & \hat{h}(\tau) \leq 1-\beta, \end{aligned}$$

where $\hat{h}(\tau) \triangleq \mathbf{Pr}_\Delta[\hat{f}_2(\tau;\delta) \geq 1]$. From orthonormality of the Legendre PCE, the closed form for the cost function in (P3′) is $\mathbf{E}[\hat{f}_1(\tau;\delta)] + w\mathbf{Var}[\hat{f}_1(\tau;\delta)] = a_0^1(\tau) + w\sum_{i=1}^{N_p^1}(a_i^1(\tau))^2$.

Example 7.5. Consider an FBNP model (7.17) with the nominal parameter values $k_1^o = \tau_d^o = \tau_c^o = 1$ and 30% uncertainty $\rho = 0.3$. The reference model is assumed to be $M(s) = \frac{1}{1+s}$ and the weighting transfer functions are $W_1(s) = 1.25\frac{1}{1+\sqrt{2}s}$ and $W_2(s) = 0.125\frac{1+0.25s}{1+0.0025s}$. The IMC controller (7.19) is considered with the low-pass filter $J(s) = \frac{1}{(1+\tau s)^3}$. The probabilistic violation of the chance constraint $\mathbf{Pr}_\Delta[\|W_2(s)T(s)\|_\infty \geq 1]$ using the Legendre-PCE (L-PCE) of order 5 is nearly identical to MC simulations (see Fig. 7.4).

7.5 Summary and Future Work

This chapter studies generalized polynomial chaos expansion approaches to approximate the functional dependence of dynamical system properties on uncertainties that are random variables for stochastic control problems as a means of replacing or facilitating MC simulation methods. Stochastic optimal control problems were formulated using gPC in which the cost function and probabilistic constraints can be reformulated as the constraints over the coefficients of gPC expansions. Several numerical examples and case studies were presented to illustrate the proposed approaches. Since closed form expressions for the dependence of the coefficients on the decision variables are unavailable in general, the optimizations may be solved by derivative-

free or approximate gradient methods for nonlinear programming.

Chapter 8

Approximate Stochastic Model Predictive Control

Abstract This chapter considers the model predictive control of dynamic systems subject to stochastic uncertainties due to parametric uncertainties and exogenous disturbance. The effects of uncertainties are quantified using generalized polynomial chaos expansions with an additive Gaussian random process as the exogenous disturbance. With Gaussian approximation of the resulting solution trajectory of a stochastic differential equation using generalized polynomial chaos expansion, convex finite-horizon model predictive control problems are solved that are amenable to online computation of a stochastically robust control policy over the time horizon. Using generalized polynomial chaos expansions combined with convex relaxation methods, the probabilistic constraints are replaced by convex deterministic constraints that approximate the probabilistic violations. This approach to chance-constrained model predictive control provides an explicit way to handle a stochastic system model in the presence of both model uncertainty and exogenous disturbances.

8.1 Introduction

In recent years, stochastic programming formulations for model predictive control (MPC, aka receding horizon control) have been intensively studied in the context of many different areas of application including robot and vehicle path planning [29–31], network traffic control [296], chemical processes [158, 235, 269], and economics [60, 105, 310]. In such control problems, stochastic models are represented in terms of stochastic differential equations (SDEs) with the stochasticity resulting from exogenous disturbances, plant/model mismatches, and sensor noise.

Robust MPC formulations can be categorized as being either deterministic or stochastic, based on the

representation of the uncertainties. Deterministic robust MPC (e.g., see [20, 50, 286] and references therein) analyzes the stability and performance of systems against worst-case perturbations with the resulting optimizations being min-max problems that are computationally demanding to solve directly and so are typically replaced by approximate solutions that are more amenable to implementation. The worst-case perturbations may have a vanishingly small probability of occurring in practice, but any such information on probabilities is not taken into account in a deterministic formulation. Analysis or design based on worst-case uncertainties can be too conservative to be applied in practice, may result in an overdesign of process equipment, or can result in infeasibility during real-time optimization. From a practical point of view, it is rare that an engineer knows exactly what value for hard bounds to specify on the uncertainty (e.g., knows that the hard bound on uncertainty in a parameter should be exactly 10.6% instead of 11.3%), and a small perturbation in these bounds can mean the difference because a closed-loop system being robust to the uncertainties or being unstable.

Most parameter estimation algorithms generate models with probabilistic descriptions of the uncertainties. For such models, robustness characterizations are intrinsically stochastic and can be written in terms of a probability distribution or a level of confidence in estimates with probabilistic risk of incorrectness. Contrary to deterministic robust MPC, stochastic robust MPC incorporates such probabilistic uncertainties and probabilistic violations of constraints, and allows for specified levels of risk during operation. Commonly used probabilistic analysis approaches are Monte Carlo (MC) methods, in which many simulations are run with sampled random variables or random sequences. The effects of uncertainty on the closed-loop system are quantified by simulating a large number of individual deterministic model realizations. While such MC approaches are applicable to most systems, the computational cost can be prohibitively expensive, especially in real-time optimal control algorithms such as MPC. Apart from simulation-based methods, convex approximations for a receding horizon method of the constrained discrete-time stochastic control are considered in [55], in which convexity of the resultant optimization is carried out in the basis of robust optimization [22, 27] that includes robust linear programs and more generally robust convex programs (see [21] for details of robust convex optimization). However, such robust optimization formulations of chance constraints are not applicable to the cases when the stochastic dynamical system has nonlinear parametric uncertainties, whereas this paper can manage the system model that is linear parameter-varying Gaussian, for which the system matrices have nonlinear dependence of random variables and there are additive Gaussian random processes corresponding to external disturbance and measurement noise.

The high computational cost of simulation-based methods has motivated the development of computationally efficient methods for uncertainty analysis that replace or accelerate MC methods [96, 167, 293]. The MPC formulation in this chapter uses generalized polynomial chaos (gPC) expansions, which is a spectral method to approximate the solution of an SDE that has stochastic parametric uncertainties and exogenous disturbances. Polynomial chaos expansions were first introduced for turbulence modeling with the uncertainties being Gaussian random variables [288], with later extensions considering other types of common

probability distributions [294]. Recently, many researchers have demonstrated the use of (generalized) polynomial chaos expansions as a computationally efficient alternative to MC approaches for the analysis and control of uncertain systems [82,115,116,143,191,192]. In [83,84], the gPC expansion is applied to formulate optimal trajectory generation problems in the presence of random uncertain parameters.

This chapter also presents several probabilistic collision conditions that are functions of the mean and covariance of the trajectory. We show that a gPC expansion that is an approximation of the solution of an SDE, in which both system parameters and exogenous disturbances are stochastic, converges in the mean-square sense as the number of terms in the expansion increases. The proposed approximation results in a convex optimization for the control policy that does not use any sampling and is amenable to online computation.

8.1.1 Problem Statement

Consider a stochastic discrete-time linear parameter-varying system:

$$x_{t+1} = A(\delta)x_t + B_u(\delta)u_t + B_w(\delta)w_t, \tag{8.1}$$

where $\delta \in \Delta$ denotes the concatenation of the parametric uncertainties and $A : \Delta \to \mathbb{R}^{n\times n}$, $B_u : \Delta \to \mathbb{R}^{n\times m}$, and $B_w : \Delta \to \mathbb{R}^{n\times n_w}$ are uncertain system matrices. Assume that $w \in \mathbb{R}$ is a Gaussian white noise process with known distribution, and the initial state x_0 and uncertainty δ are random variables with known probability density functions (pdfs). Under this stochasticity of parameters and disturbance, the solution trajectory of the system (8.1) is a random process for which the main goal of analysis is to compute or approximate the statistical properties and the main goal of synthesis is to drive the random process x_t to have a desirable statistic.

In finite-horizon stochastic MPC, the goal is to determine a control policy $\mu_T \triangleq (u_0, \cdots, u_T)$ that solves the optimization

$$\begin{aligned}
\min_{\mu_T}\ & J(\bar{x}_0, \Sigma_{x_0}, \mu_T) \\
\text{s.t.}\ \ & x_{t+1} = A(\delta)x_t + B_u(\delta)u_t + B_w(\delta)w_t, \\
& y_t = Cx_t,\ \mathbf{Pr}[y_t \notin \mathcal{F}_y] \geq \beta, \\
& u_t \in \mathcal{U},\ \text{for } t = 0, \cdots, T, \\
& w_t \sim f_w,\ \delta \sim f_\delta,\ x_0 \sim f_{x_0},
\end{aligned} \tag{8.2}$$

where $\mathcal{F}_y$ denotes the forbidden region for the output y_t and β is a lower bound of probabilistic collision avoidance.[1]

[1] $\mathcal{F}_y$ and β can be time-varying, where the forbidden region might correspond to moving objectives and time-varying β can be used to assign different risk of collision in different time sequences in the predicted motions.

8.2 Feasibility of Chance Constraints: Probabilistic Collision Checking

This section presents four different ways of formulation of chance constraints corresponding to probabilistic collision avoidance. In particular, for a motion-planning problem for a mobile system it is necessary to impose constraints on the (controlled) state or output variables and such constraints have the form $\eta(x) \leq 0$ where $x \in \mathcal{X} \subset \mathbb{R}^n$ refers to the state variables and the function $\eta : \mathbb{R}^n \to \mathbb{R}^m$ is of vector-value. Due to stochastic nature of the state variables, it is natural to introduce the so-called chance constraints that are of the form $\Pr[\eta(x) \leq 0] \geq \alpha$ where $\alpha \in (0, 1)$ denotes a level of confidence. For a probabilistic collision avoidance problem, the formulation of chance constraints depends on the representation of obstacles and mobile agents that have stochastic uncertainty.

8.2.1 Obstacles as Point Masses in a Large Work Space

The probability of collision to obstacles at time t and in the work space $W \subset \mathbb{R}^{n_s}$, $n_s \leq 3$, can be defined as [154, 264]

$$P_t^c \triangleq \int_{x^v} \int_{x^a} I_c(x_t^v, x_t^a) dF_{va}(x_t^v, x_t^a), \tag{8.3}$$

where $F_{va}(\cdot, \cdot)$ is the joint cumulative distribution function (cdf), the indicator function for collision is defined by

$$I_c(x^v, x^a) \triangleq \begin{cases} 1, & \text{for } \mathcal{X}_v(x^v) \bigcap \mathcal{X}_a(x^a) \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases}$$

$\mathcal{X}_v(x^v)$ and $\mathcal{X}_a(x^a)$ are the regions occupied by the vehicle and the obstacle whose global reference coordinates are x^v and x^a, respectively. Equipped with this definition of probabilistic collision, the chance constraint $\mathbf{Pr}[y_t \notin \mathcal{F}_y] \geq \beta$ in (8.2) can be rewritten as $P_t^c \leq 1 - \beta$. Consider the obstacles as point masses, which occurs when the volume of $\mathcal{X}_v(x^v)$ is much smaller than the work space W for all $x^v \in W$ and the volume of $\mathcal{X}_a(x^a)$ is 0 for all $x^a \in W$.

Lemma 8.1 (*Lemma 1 in [264]*). For obstacles as point masses, suppose that $x^v \sim \mathcal{N}(\bar{x}^v, \Sigma_{x^v})$ and $x^a \sim \mathcal{N}(\bar{x}^a, \Sigma_{x^a})$ are independent Gaussian random variables. Then $P^c \leq 1 - \beta$ can be rewritten as the constraint on $(\bar{x}^v, \Sigma_{x^v}, \bar{x}^a, \Sigma_{x^a})$:

$$(\bar{x}^v - \bar{x}^a)^{\mathrm{T}} \Sigma_x^{-1} (\bar{x}^v - \bar{x}^a) \geq -2 \ln \left(\frac{1 - \beta}{V_v} \sqrt{\det(2\pi\Sigma_x)} \right), \tag{8.4}$$

where $\Sigma_x = \Sigma_{x^v} + \Sigma_{x^a}$ and V_v is the volume of the vehicle.

The constraint (8.4) is not convex in $(\bar{x}^v, \bar{x}^a)$ even for a fixed β, but is concave in $(\bar{x}^v, \bar{x}^a)$ due to positive definiteness of the inverse covariance matrix Σ_x^{-1}. However, a method of semidefinite programming (SDP) relaxation can be used to check its feasibility and solve related optimizations approximately.

Convex relaxation: Suppose that $\bar{x}^v$ is affine in the control input u, i.e., $\bar{x}^v = Mu+b$ with an appropriate matrix M and a vector b of compatible dimensions. Then the inequality (8.4) can be rewritten as

$$\begin{aligned}\begin{bmatrix} 1 \\ u \end{bmatrix}^{\mathrm{T}} \mathcal{Q} \begin{bmatrix} 1 \\ u \end{bmatrix} \geq \gamma \iff \mathrm{Tr}\left(\mathcal{Q} \begin{bmatrix} 1 \\ u \end{bmatrix} \begin{bmatrix} 1 \\ u \end{bmatrix}^{\mathrm{T}} \right) \geq \gamma \\ \iff \mathrm{Tr}(\mathcal{Q}U) \geq \gamma,\ U \succeq 0,\ U_{11} = 1,\ \mathrm{Rank}(U) = 1\end{aligned} \tag{8.5}$$

where $Q \succeq 0$ and γ can be appropriately computed from (8.4). Suppose that a convex quadratic constraint of the form $u^{\mathrm{T}}Q_1u + q_1^{\mathrm{T}}u + q_{10} \leq 0$ with $Q_1 \succeq 0$ is imposed on the control input. Minimizing the probability of collision under that quadratic constraint can be represented as the optimization

$$\begin{aligned}&\max_U \mathrm{Tr}(\mathcal{Q}U) \\ &\text{s.t. } \mathrm{Tr}(\mathcal{Q}_1U) \leq 0,\ U \succeq 0,\ U_{11} = 1,\ \mathrm{Rank}(U) = 1\end{aligned} \tag{8.6}$$

where the symmetric matrix $\mathcal{Q}_1$ satisfies the relation $\begin{bmatrix} 1 \\ u \end{bmatrix}^{\mathrm{T}} \mathcal{Q}_1 \begin{bmatrix} 1 \\ u \end{bmatrix} = u^{\mathrm{T}}Q_1u + q_1^{\mathrm{T}}u + q_{10}$. It is well known that removing the rank constraint $\mathrm{Rank}(U) = 1$ in this particular optimization does not change the optimum value, i.e, the corresponding SDP relaxation is exact [198]. Next, consider a similar problem with the box-type constraints $|u_i| \leq 1$ for $i = 1, \dots n_u$ in the place of the quadratic constraint on u. Then minimization of the probability of collision can be represented as the optimization

$$\begin{aligned}&\max_U \mathrm{Tr}(\mathcal{Q}U) \\ &\text{s.t. } U_{ii} \leq 1,\ i = 2, \dots, n_u + 1, \\ &\qquad U \succeq 0,\ U_{11} = 1,\ \mathrm{Rank}(U) = 1.\end{aligned} \tag{8.7}$$

The associated primal SDP relaxation is the same as the optimization (8.7) without the rank constraint, and its suboptimality is bounded by

$$\gamma^\star \leq \gamma^\star_{\mathrm{sdp}} \leq \frac{\pi}{2}\gamma^\star \tag{8.8}$$

where $\gamma^\star$ refers to the optimal value of (8.7) and $\gamma^\star_{\mathrm{sdp}}$ refers to the optimal value of the primal SDP relaxation [196, 198].

8.2.2 Probabilistic Safety Regions

Instead of quantifying the probability of safety by $1 - P_c$, consider the dual definition of probability of safety:

$$P_t^s \triangleq \int_{x^v} \int_{x^s} I_s(x_t^v, x_t^s) dF_{vs}(x_t^v, x_t^s), \tag{8.9}$$

where x^s is a global reference coordinate that characterizes a virtual safety region $\mathcal{X}_s(x^s)$ and the joint cdf F_{vs} and the indicator function I_s follow similar definitions as in the previous section. With this definition of probabilistic safety regions, the chance constraint $\mathbf{Pr}[y_t \notin \mathcal{F}_y] \geq \beta$ in (8.2) can be rewritten as $P_t^s \geq \beta$. Consider the obstacles as point masses, which is a case when the volume of $\mathcal{X}_v(x^v)$ is much smaller than the work space W for all $x^v \in W$ and the volume of $\mathcal{X}_s(x^s)$ defines a point in W.

Lemma 8.2. For point-mass obstacles, suppose that $x^v \sim \mathcal{N}(\bar{x}^v, \Sigma_{x^v})$ and $x^s \sim \mathcal{N}(\bar{x}^s, \Sigma_{x^s})$ are independent Gaussian random variables. Then $P_s \geq \beta$ can be rewritten as the constraint on $(\bar{x}^v, \Sigma_{x^v}, \bar{x}^s, \Sigma_{x^s})$:

$$(\bar{x}^v - \bar{x}^s)^{\mathrm{T}} \Sigma_x^{-1} (\bar{x}^v - \bar{x}^s) \leq -2 \ln \left(\frac{\beta}{V_v} \sqrt{\det(2\pi\Sigma_x)} \right), \tag{8.10}$$

where $\Sigma_x = \Sigma_{x^v} + \Sigma_{x^s}$ and V_v is the volume of the vehicle.

The constraint (8.10) is convex in $(\bar{x}^v, \bar{x}^s)$ for a fixed $\beta \in [0, 1]$.

8.2.3 Obstacles as Linear Constraints in a Work Space

Consider the lifted system output y_t. The forbidden region for the system output can be defined as a union of N linear inequality constraints that is a nonconvex polyhedral set

$$\mathcal{F}_y \triangleq \bigcup_{i=1}^{N} \left\{ y : h_i^{\mathrm{T}} y \geq b_i \right\}. \tag{8.11}$$

Assume that $y \sim \mathcal{N}(\bar{y}, \Sigma_y)$ and define $\eta_i \triangleq h_i^{\mathrm{T}} y$, which is a univariate Gausian random variable with mean $\bar{\eta}_i = h_i^{\mathrm{T}} \bar{y}$ and variance $\Sigma_{\eta_i} = h_i^{\mathrm{T}} \Sigma_y h_i$. The idea of risk allocation proposed by [29,207] can be used to derive a conservative convex condition for the constraint (8.11).

Lemma 8.3 (*Lemmas 1, 2, and 3 in [29]*). Consider a chance constraint $\mathbf{Pr}[y \notin \mathcal{F}_y] \geq \beta$ or the equivalent condition $\mathbf{Pr}[y \in \mathcal{F}_y] \leq 1 - \beta$ where $\mathcal{F}_y$ is defined in (8.11). Then the feasibility of the constraint

$$\mathbf{Pr}[\eta_i \geq b_i] \leq \epsilon_i, \ \epsilon_i \in (0, 1), \text{ and } \sum_i \epsilon_i = 1 - \beta \tag{8.12}$$

implies the feasibility of the constraint $\mathbf{Pr}[y \notin \mathcal{F}_y] \geq \beta$. Furthermore, $\mathbf{Pr}[\eta_i \geq b_i] \leq \epsilon_i \Leftrightarrow \frac{1}{2} \left(1 - \mathrm{erf} \left(\frac{b_i - \bar{\eta}_i}{\sqrt{2\Sigma_{\eta_i}}} \right) \right) \leq \epsilon_i$, and the constraint (8.12) is convex in $(\bar{\eta}_i, \epsilon_i)$ for $\beta \geq 0.5$.

Alternatively, consider the forbidden region for the system output defined as an intersection of N linear inequality constraints that is a convex polyhedral set

$$\mathcal{F}'_y \triangleq \bigcap_{i=1}^{N} \left\{ y : h_i^{\mathrm{T}} y \leq b_i \right\}. \tag{8.13}$$

In this case, the nonconvex chance constraint $\mathbf{Pr}[y \notin \mathcal{F}'_y] \geq \beta$ can be replaced by the relaxation

$$\mathbf{Pr}[\eta_i > b_i] \geq \epsilon_i,\ \epsilon_i \in (0,1),\ \text{for } i = 1, \ldots, N \tag{8.14}$$

where ϵ_i are appropriately defined functions of β.

Lemma 8.4. With the condition $\epsilon_i \geq 0.5$ incorporated into the constraint (8.14), the combined constraint is convex in $(\bar{\eta}_i, \epsilon_i)$.

Proof. $\mathbf{Pr}[\eta_i > b_i]$ is a concave function in $\bar{\eta}_i \geq b_i$, and for $\epsilon_i \geq 0.5$, the feasibility of the constraint (8.14) necessarily requires $\bar{\eta}_i \geq b_i$. Thus, if $(\bar{\eta}_i^1, \epsilon_i^1)$ and $(\bar{\eta}_i^2, \epsilon_i^2)$ are feasible solutions of the constraint (8.14) and $\epsilon_i^j \geq 0.5$, $j = 1, 2$, then $(\bar{\eta}_i^\lambda, \epsilon_i^\lambda)$ is also a feasible solution for all $\lambda \in [0, 1]$ where the superscript λ refers to the λ-convex combination of the feasible solutions with the superscripts 1 and 2. QED

Imposing additional linear constraints on ϵ_i does not change the convexity of the combined constraint. For example, additional constraints $\epsilon_i \geq \ell(\beta)$ could be introduced in which $\ell : \beta \mapsto [0.5, 1)$ is a nondecreasing function. However, the most practically useful functional form for the ℓ is not obvious. One functional form that may be useful is $\ell(\beta) = \sqrt[N]{\beta}$, in which case $\beta \geq 0.5^N$ would satisfy the constraint $\epsilon_i \geq 0.5$.

8.3 Efficient Approximation of Feasibility of Probabilistic Constraints

In the previous section, we presented the methods of formulating chance constraints for probabilistic collision avoidance under stochastic uncertain circumstance and model uncertainty. Under the assumption that the state (or output) variables are jointly Gaussian random variables, the resultant chance constraints are to impose the constraints on the mean and covariance of the state. However, the state of the system model (8.1) is not Gaussian and even computations of its mean and covariance can necessitate sampling-based evaluation such as Monte Carlo simulation. This section presents and analyzes methods of approximating uncertainty propagation in a stochastic dynamical system (8.1) that are in the basis of generalized polynomial chaos expansions. The methods provides numerically tractable computations of the mean and covariance of controlled state variables for which closed-forms of approximate mean and covariance can be obtained and the associated chance constraints can be efficiently evaluated.

8.3.1 Approximation of Uncertainty Propagation

Consider the concatenated parametric uncertainty $\theta := [x_0^{\mathrm{T}}, \delta^{\mathrm{T}}]^{\mathrm{T}}$. Suppose that there exists a diffeomorphism $T : \Xi \to \Theta = \Delta \times \mathcal{X}$ such that $\theta = T(\zeta)$ and $\zeta \in \Xi$ is a standard random variable. For application of the

spectral method based on gPC expansions, assume that the solution of the SDE (8.1) has the form

$$x_t \approx \hat{x}_t \triangleq \sum_{i=0}^{p-1} \phi_i(\zeta) X_t^i \tag{8.15}$$

which is an approximation of the true solution x with p basis functions from the set $\{\phi_i\}$. Obtaining the approximate solution $\hat{x}$ involves determining the time-varying deterministic coefficients X_t^i. To do this, substitute the approximation $\hat{x}$ into x of the SDE (8.1) and solve for the X_t^i by intrusive or non-intrusive projections onto the probability space of the random variable ζ whose cdf is given by F_ζ. In particular, applying Galerkin projection [167] results in another SDE

$$X_{t+1} = G_X X_t + G_u u_t + G_w w_t, \tag{8.16}$$

where $X_t := \mathrm{col}(X_t^i) \in \mathbb{R}^{np}$, the initial condition $X_0^i = \langle \phi_i(\zeta), x_0(\zeta) \rangle$, and the matrices $G_{(\cdot)}$ are computed from the inner product (A.41) defined on a measure space $(\Xi, \mathcal{M}, F_\zeta)$ for the Galerkin projection. The lifted variables of interest over the time-horizon satisfy the equation

$$X_{0:T} = H_X X_0 + H_u u_{0:T} + H_w w_{0:T}, \tag{8.17}$$

where $X_{0:T} := \mathrm{col}(X_0, \cdots, X_T)$ and the matrices $H_{(\cdot)}$ have closed forms in terms of the matrices $G_{(\cdot)}$ in (8.16). From the assumption of Gaussian white noise w_t, the lifted coefficients X_t is a Gaussian random process resulting in a Gaussian random variable $X_{0:T}$ with mean and covariance

$$\begin{aligned} \bar{X}_{0:T} &= H_X X_0 + H_u u_{0:T} + H_w \bar{w}_{0:T}, \\ \Sigma_{X_{0:T}} &= H_X \Sigma_{X_0} H_X^{\mathrm{T}} + H_w \Sigma_{w_{0:T}} H_w^{\mathrm{T}}. \end{aligned} \tag{8.18}$$

The following proposition shows that the mean and covariance of the approximate solution $\hat{x}_t$ have closed-forms with respect to the mean and covariance of the coefficients of a generalized polynomial chaos expansion given in (8.18).

Proposition 8.1. The lifted approximation of the solution $\hat{x}_{0:T}$ satisfies

$$\mathbf{E}[\hat{x}_{0:T}] = K_X X_0 + K_u u_{0:T} + K_w \bar{w}_{0:T}, \tag{8.19}$$

where the matrices $K_{(\cdot)}$ are functions of $G_{(\cdot)}$ and $H_{(\cdot)}$, and there exists an affine surjective map $\Omega : \mathbb{S}^{np(T+1)} \to \mathbb{S}^{n(T+1)}$ such that

$$\Sigma_{\hat{x}_{0:T}} = \Omega(\Sigma_{X_{0:T}}). \tag{8.20}$$

Proof. The proof is straightforward. Consider an approximate solution using a polynomial expansion

(8.15). Due to independence of the random parameter θ and the random process w_t, its expectation is $\mathbf{E}[\hat{x}_t] = \mathbf{E}[\phi(\zeta)^{\mathrm{T}} \otimes \mathrm{I}_n]\mathbf{E}[X_t]$ where the first expectation is computed w.r.t. the random vector ζ and the second expectation is computed w.r.t. the random process w_t. Since the coefficient X_t is linear and has affine dependence on the control input u_t and the external disturbance w_t, the lifted approximate state $\mathbf{E}[\hat{x}_{0:T}]$ is of the form given in (8.19). Similarly, the variance of the approximate state $\mathbf{E}[\hat{x}_t\hat{x}_t^{\mathrm{T}}]$ can be rewritten as $\mathbf{E}[(\phi(\zeta)^{\mathrm{T}}\otimes\mathrm{I}_n)X_tX_t^{\mathrm{T}}(\phi(\zeta)^{\mathrm{T}}\otimes\mathrm{I}_n)^{\mathrm{T}}]$ or equivalently, $\mathbf{E}[\underline{\mathrm{X}}_t\phi(\zeta)\phi(\zeta)^{\mathrm{T}}\underline{\mathrm{X}}_t^{\mathrm{T}}] = \mathbf{E}[\underline{\mathrm{X}}_t\Phi\underline{\mathrm{X}}_t^{\mathrm{T}}]$ where $\Phi \triangleq \mathbf{E}[\phi(\zeta)\phi(\zeta)^{\mathrm{T}}]$ and $\underline{\mathrm{X}}_t \triangleq \left[X_t^0, \cdots, X_t^{p-1}\right] \in \mathbb{R}^{n\times p}$. From orthonormality of the basis functions, let $\Phi = \mathrm{I}_p$ without loss of generality. The (k,ℓ) element of the matrix $\mathbf{E}[\hat{x}_t\hat{x}_t^{\mathrm{T}}] = \mathbf{E}[\underline{\mathrm{X}}_t\underline{\mathrm{X}}_t^{\mathrm{T}}]$ is $\sum_{j=0}^{p-1} X_{t,k}^j X_{t,\ell}^j$ and $X_{t,k}^jX_{t,\ell}^j$ is an element of the matrix $\mathbf{E}[X_tX_t^{\mathrm{T}}]$. Therefore, $\mathbf{E}[\hat{x}_t\hat{x}_t^{\mathrm{T}}]$ is an affine function of $\mathbf{E}[X_tX_t^{\mathrm{T}}]$, which is equivalent to the lifted covariance matrix $\Sigma_{\hat{x}_{0:T}}$ being an affine function of $\Sigma_{X_{0:T}}$. The corresponding mapping is a projection that is surjective. QED

The random process X_t is Gaussian such that the mean and covariance given by (8.18) exactly characterize the probability distribution of X_t for all t, whereas the approximation $\hat{x}_t$ to the solution x_t is not necessarily Gaussian, due to the additional randomness of the parameters (x_0, δ). However, the mean and covariance of x_t can be approximated by the mean and covariance of $\hat{x}_t$ given by (8.19) and (8.20). More precisely, the next proposition shows the convergence of the approximation error in the mean-square sense.

Proposition 8.2. Consider the solution trajectory x_t of the system (8.1) and its approximation $\hat{x}_t$ using a gPC expansion given by (8.15) whose coefficients X_t solve (8.16). Assume that the random variables (x_0, δ) are independent of the random process w_t and X_t is a second-moment process.[2] Then $\|x_t - \hat{x}_t\| \to_{\mathrm{m.s.}} 0$ pointwisely in t as $p \to \infty$, where $\|\cdot\|$ can be any vector p-norm.

Proof. An approximation $\hat{x}_t$ can be explicitly rewritten as $\sum_{i=0}^{p-1}\phi_i(\zeta(x_0,\delta))X_t^i(w_{0:t-1})$. From Thm. A.1, for any realization of the random variable $w_{0:t-1} \in \mathcal{W}^t$, where $\mathcal{W}$ is the support of w_t and $\bar{\epsilon}$ is greater than zero, there exists $\bar{p} \in \mathbb{N}$ such that $\int_\Xi \|x_t - \sum_{i=0}^{p-1}\phi_i(\zeta)X_t^i(w_{0:t-1})\|^2 d\mu_\zeta(\zeta) \leq \bar{\epsilon}$ for all $p \geq \bar{p}$, where μ_ζ is the probability measure of ζ. The $\bar{\epsilon}$ is a function of $w_{0:t-1}$. Due to the linear dependence of x_t and X_t^i on $w_{0:t-1}$, which follows from (8.1) and (8.16), $\bar{\epsilon} = \epsilon w_{0:t-1}^{\mathrm{T}}w_{0:t-1}$ where $\epsilon > 0$ is an arbitrary constant that is independent of $w_{0:t-1}$. This implies that the mean-square approximation error is bounded above:

$$\int_{\mathcal{W}^t}\int_\Xi \left\| x_t - \sum_{i=0}^{p-1}\phi_i(\zeta)X_t^i(w_{0:t-1})\right\|^2 d\mu_\zeta(\zeta)d\mu_w(w_{0:t-1}) \leq \epsilon\int_{\mathcal{W}^t} w_{0:t-1}^{\mathrm{T}}w_{0:t-1}d\mu_w(w_{0:t-1}) \leq \epsilon M,$$

where μ_w is the corresponding probability measure of the random variable $w_{0:t-1}$ and $M < \infty$ whose boundedness follows from the second-moment assumption of the random process w_t. Since $\epsilon > 0$ is arbitrary, the convergence is ensured. QED

Furthermore, if the system matrices are analytic functions of the random variables (x_0, δ) then the convergence rate of the approximation error $\|x_t - \hat{x}_t\|$ to 0 in mean-square is exponential, which follows from the

[2]Consider the time interval $[0, T]$ in which X_t is a second-moment process.

solution trajectory x_t being an analytic function of (x_0, δ) under those assumptions.

8.3.2 Gaussian Approximation and Convexifications of Chance-constrained MPC: Information Theoretic Justification

For a Gaussian random process y_t (or x_t), we previously showed that the chance constraint corresponding to probabilistic collision avoidance $\mathbf{Pr}[y_t \notin \mathcal{F}_y] \geq \beta$ can be rewritten as conditions in terms of its mean $\bar{y}_t$ and covariance Σ_{y_t}. In particular, the conditions (8.10), (8.12), and (8.14) are jointly convex in y_t (or x_t) and the other decision variables (ϵ_i), under some mild assumptions.

However, the solution x_t of the system dynamics (8.1) and its spectral approximation $\hat{x}_t$ given in (8.15) are not generally Gaussian random processes, which make the optimization (8.2) difficult to solve in the sense that the chance constraint does not have a closed-form expression and its feasibility is hard to check. To avoid the use of any sampling or simulation-based methods to evaluate the feasibility of the chance constraint $\mathbf{Pr}[y_t \notin \mathcal{F}_y] \geq \beta$, the approximate solution $\hat{x}_t$ is substituted in the place of x_t and Gaussian fitting of the random variables under consideration is applied. More specifically, assume that $\hat{x}_t \sim \mathcal{N}(\bar{\hat{x}}_t, \Sigma_{\hat{x}_t})$, for which there are closed-form expressions given by (8.19) and (8.20).[3] A theoretical justification of this assumption $\hat{x}_t \sim \mathcal{N}(\bar{\hat{x}}_t, \Sigma_{\hat{x}_t})$ can be made from the principle of maximum entropy [61, Chap. 12]. Maximum entropy can be used to determine or approximate a probability distribution that incorporates only known information. If only the first and second moments of $\hat{x}_t$ are used to approximate its probability distribution then the maximum entropy distribution has the form $\mathcal{N}(\bar{\hat{x}}_t, \Sigma_{\hat{x}_t})$, i.e., a Gaussian distribution. Furthermore, since $\hat{x}_t$ converges to x_t in the mean-square sense as the number of basis functions increases, the approximate probability distribution $\mathcal{N}(\bar{\hat{x}}_t, \Sigma_{\hat{x}_t})$ can be made arbitrarily close to the probability distribution of x_t that maximizes entropy subject to the constraints corresponding to the first and second moments.

Proposition 8.3. Consider the solution trajectory x_t of the system (8.1) and its approximation $\hat{x}_t$ using a gPC expansion given by (8.15) whose coefficients X_t solve (8.16). Assume that the random variables (x_0, δ) are independent of the random process w_t and X_t is a second-moment process (for notation convenience, the subscript t is dropped from here on). Suppose that a probability density $f^\star$ solves the optimization

$$\begin{aligned} \max_f \ & -\int_S f(x) \log f(x) dx \\ \text{s.t. } & f(x) \geq 0, \ \int_S f(x) dx = 1, \ \int_S f(x) x^i dx = M_i, \ i = 1, 2, \end{aligned} \tag{8.21}$$

where S denotes the support for the random variable x, and M_1 and M_2 correspond to the given first and second moments, respectively. Then an approximate Gaussian distribution $\hat{f}_2 \triangleq \mathcal{N}(\bar{\hat{x}}, \Sigma_{\hat{x}})$ obtained from

[3] The computation of deterministic constant matrices $K_{(\cdot)}$ and $\Sigma_{X_{0:T}}$ (or $\Sigma_{\hat{x}_{0:T}}$) can be performed off-line.

the solution of (8.16) converges to $f^\star$ as $p \to \infty$ in the $\mathcal{L}_1$-norm sense, i.e.,

$$\lim_{p\to\infty} \int_S \left| f^\star(x) - \hat{f}_2(x) \right| dx = 0.$$

Proof. Due to limited space, consider the scalar case (the extension to the multivariable case is straightforward). From the principle of maximum entropy, a unique $f^\star$ has the form of $e^{\lambda_0+\lambda_1 x+\lambda_2 x^2}$ that corresponds to a Gaussian distribution. Similarly, $\hat{f}_2$ is a unique maximum entropy distribution that solves the optimization (8.21) with given approximate moments $\hat{M}_i$, $i = 1, 2$, in the place of M_i and can be rewritten as $e^{\hat{\lambda}_0+\hat{\lambda}_1 x+\hat{\lambda}_2 x^2}$ for some constants $\hat{\lambda}_j$, $j = 0, 1, 2$. Since convergence in mean-square implies convergence in distribution and $\hat{M}_i$ can be arbitrarily close to M_i as $p \to \infty$ from Prop. 8.2, this implies that $\lim_{p\to\infty} \max_j |\lambda_j - \hat{\lambda}_j| = 0$. Therefore, for any arbitrary constant $\epsilon > 0$, there exists $\bar{p} \in \mathbb{N}$ such that $\min\{e^\epsilon, e^{-\epsilon}\} \le \hat{f}_2(x)/f^\star(x) \le \max\{e^\epsilon, e^{-\epsilon}\}$ uniformly in $x \in S$ for all $p \ge \bar{p}$. This implies that $\hat{f}_2$ converges to $f^\star$ in the $\mathcal{L}_1$-norm sense as $p \to \infty$. QED

Remark 8.1. The above Gaussian approximation is a suboptimal way to estimate the probability distribution of x_t, which produces convex chance constraints that are more computationally tractable by ignoring the extra information in the higher-order moments of $\hat{x}_t$. This method of approximation has the same characteristics as the extended Kalman filter (EKF) and unscented Kalman filter (UKF) that are widely used in practical applications although there are no theoretical guarantees that those estimation methods will always work well or even converge.

Using the Gaussian approximation, the design problem reduces to finding a control policy μ_T (or $u_{0:T}$) that solves the optimization

$$\begin{aligned}
&\min_{\mu_T} J(\bar{x}_0, \Sigma_{x_0}, \mu_T) \\
&\text{s.t. } \bar{\hat{x}}_{0:T} = K_X X_0 + K_u u_{0:T} + K_w \bar{w}_{0:T}, \\
&\quad \Sigma_{\hat{x}_{0:T}} = \Omega(\Sigma_{X_{0:T}}), \\
&\quad \hat{x}_{0:T} \sim \mathcal{N}(\bar{\hat{x}}_{0:T}, \Sigma_{\hat{x}_{0:T}}),\ y_{0:T} = (\oplus_{i=0}^{T} C)\hat{x}_{0:T}, \\
&\quad (\bar{y}_{0:T}, \Sigma_{y_{0:T}}) \in \mathcal{F}(\beta) \text{ or } (\bar{y}_{0:T}, \Sigma_{y_{0:T}}, \epsilon) \in \mathcal{F}(\beta), \\
&\quad u_{0:T} \in \mathcal{U}^{T+1},
\end{aligned} \tag{8.22}$$

where the matrices $K_{(\cdot)}$ and $\Sigma_{X_{0:T}}$, and the injection map Ω are precomputed, $\oplus_{i=0}^{T} C \triangleq \text{diag}(C, \cdots, C)$, and the constraints $\mathcal{F}(\beta)$ can be one of the sets:

$$\{(y, \Sigma_y) : \text{Eq. (8.10)}, \bar{x}^v = y,\ \Sigma_{x^v} = \Sigma_y\}; \tag{8.23}$$

$$\{(y, \Sigma_y, \epsilon) : \text{Eq. (8.12)}, \bar{\eta} = y,\ \Sigma_\eta = \Sigma_y\}; \tag{8.24}$$

$$\{(y, \Sigma_y, \epsilon) : \text{Eq. (8.14)}, \bar{\eta} = y, \Sigma_\eta = \Sigma_y, \epsilon \geq 0.5\}, \tag{8.25}$$

where $\epsilon = \text{col}(\epsilon_i)$, $\beta \geq 0.5$ is required for the second feasible solution set to be convex in (y, ϵ), and the first and the third sets are convex in y and (y, ϵ), respectively. With the standard performance specification that the objective function J is convex quadratic in μ_T and the set $\mathcal{U}$ is a convex polytope, the optimization (8.22) is a convex quadratically constrained quadratic program (QCQP) when $\mathcal{F}(\beta)$ is given by (8.23) and a convex nonlinear program when $\mathcal{F}(\beta)$ is given by (8.24) or (8.25).

8.3.3 A Demonstration Example

This section compares the accuracy of the proposed gPC-based MPC formulations for a numerical example. Consider the parametric uncertain linear time-invariant system

$$x_{t+1} = \begin{bmatrix} 0.9 + \rho_1\delta_1 & 0.1 \\ 0.1 & 0.85 \end{bmatrix} x_t + \begin{bmatrix} 0.25 - \rho_2\delta_2 \\ 0.75 + \rho_2\delta_2 \end{bmatrix} u_t + \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} w_t$$

with initial condition $x_0 = [20,\ 10]^{\mathrm{T}}$, $\rho_1 = 0.001$ and $\rho_2 = 0.05$ are weights on normalized standard random variables $\delta_1 \sim \mathcal{N}(0, 1)$ and $\delta_2 \sim \mathcal{N}(0, 1)$, respectively, and the exogenous process noise $w_t \sim \mathcal{N}(0, 0.001)$ is assumed to have autocorrelation $\mathbf{E}[w_t w_s] = 0$ for all $t \neq s$. This example computes the control inputs by solving the optimization (8.28), for which the covariance constraints are imposed based on the 99% level of confidence for collision avoidance, and comparing the probability of collisions obtained from different methods presented in the paper. Consider $Q_t = \text{diag}\{100, 100\}$ and $R_t = 1$ for all t, a prediction horizon of $T = 4$, and input constraint $u_t \in [-0.5, 0.5]$. The constraints are constructed from the obstacle shown in Fig. 8.1. The resultant controlled system trajectory generated by a system model with fixed parameters $\delta = [0.01,\ 0.05]$ and randomly chosen exogenous disturbances w_t is shown in Fig. 8.1, which avoids the obstacle as desired. Fig. 8.2 shows Monte Carlo simulations with 5000 samples of (δ, w), which indicates that the stochastic MPC algorithm was effective in avoiding the obstacle while allowing the closed-loop trajectory to become rather close to the obstacle so as to optimize the closed-loop performance objective. Fig. 8.3 compares the computed probabilities of collision using the methods presented in this chapter with the probabilities quantified by the Monte Carlo simulations. At each time the probability of collision obtained by the gPC expansion is very close to the value computed using either Monte Carlo applied to the original system or Monte Carlo applied to the convex relaxation. The approximate probabilities of collision follow nearly identical trends to the true probabilities while enabling the optimal control problem at each time instance of MPC to be computed from a convex program that can be solved in polynomial-time.[4]

[4]In particular, the computational complexity using a standard interior-point method [36] is at most $O(\ell M^4 \log M)$ where $M = npT$ (n is the dimension of the state variables, p is the number of basis functions for a gPC expansion, T is the length of prediction horizon), and ℓ denotes the number of probabilistic polyhedral constraints. The average computation time at each sampling instance was ≈ 0.36 CPU seconds for $n = 2$, $p = 3$, and $T = 5$. This computation time includes the computation of the optimization data, i.e., the time for computing matrices associated with the objective function and constraints, as well as

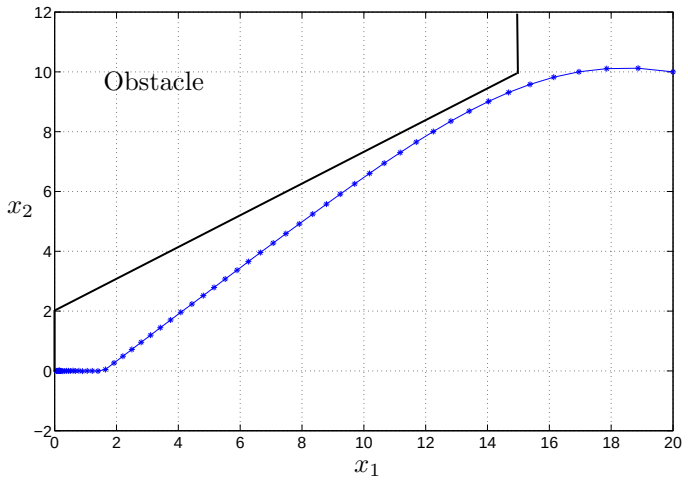


Figure 8.1: A controlled trajectory produced by the proposed stochastic MPC formulation.

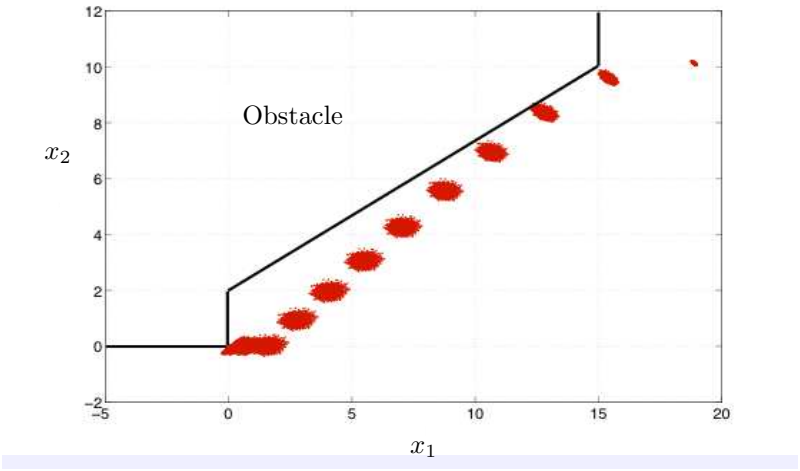


Figure 8.2: Monte Carlo simulations: the red dots correspond to simulated states at each 4^{th} sampling instance for 5000 samples.

To further assess the accuracy of the gPC expansion, let $\Sigma_x^{\mathrm{mc}}(t)$ and $\Sigma_x^{\mathrm{pce}}(t)$ be the computed covariance of the controlled system trajectory obtained from Monte Carlo simulations with 5000 samples of (δ, w) and the polynomial chaos expansion with a specified order of Hermite polynomials, respectively. Table 8.1 compares the worst-case deviation of $\max_{0\leq t\leq 60} \|\Sigma_x^{\mathrm{mc}}(t) - \Sigma_x^{\mathrm{pce}}(t)\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm, for different degrees of Hermite polynomials. The approximation error of the covariance matrix is small and, as expected from the theoretical analysis, the error in the state covariance matrix decreases as the number of terms in the polynomial expansion increases.

the time for solving the resulting constrained optimization. Optimization is performed by the CVX toolbox [99] on a MacBook Pro laptop (2.53 GHz Intel Core 2 Duo, 4GB DDR3).

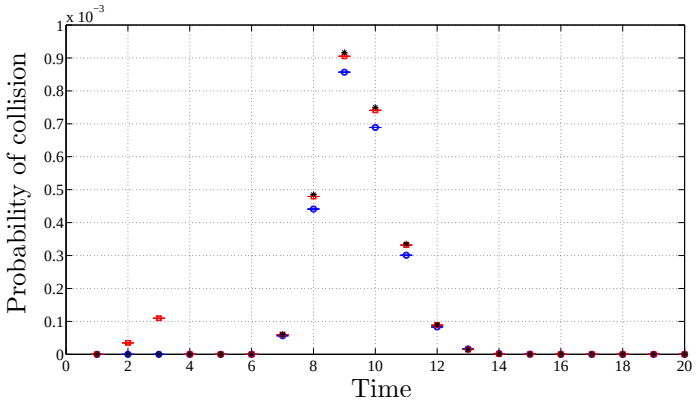


Figure 8.3: A comparison of the computed probability of collision for the true system and the approximation using a gPC expansion, estimated using Monte Carlo (MC) simulations where the error bars were obtained from 1000 Monte Carlo simulations with different sets of 5000 samples. The blue circle refers to the MC simulation result and the red box refers to the collision probability that is obtained from the MC simulations with the convex relaxation (8.12). For both computations, the corresponding error bars were generated at the 95% confidence level. The widths of the computed confidence intervals were smaller than 10^{-7}, which is negligible compared to collision probability. The black star refers to the collision probability obtained from the presented gPC method that incorporates with the convex relaxation (8.12).

Table 8.1: Covariance approximation errors for different degrees of Hermite polynomial expansions.

Degree of Hermite polys.	$\max_{0\leq t\leq 60} \|\Sigma_x^{\mathrm{mc}}(t) - \Sigma_x^{\mathrm{pce}}(t)\|_F$
1^{st}	$\approx 1.0001 \times 10^{-3}$
$2^{\mathrm{n}d}$	$\approx 6.0878 \times 10^{-4}$
$3^{\mathrm{r}d}$	$\approx 5.3627 \times 10^{-4}$
4^{th}	$\approx 2.8365 \times 10^{-4}$

8.3.4 Discussions and Further Remarks

Approximate Solution using Spectral Methods with the KL Expansions This section consider two sources of uncertainties: (a) parametric uncertainty and (b) exogenous disturbance. Uncertainty propagations induced by parametric uncertainty are approximated by using a gPC expansion and additive exogenous disturbances affect the coefficients of the resultant gPC expansion. Another possible approach to the same problem is to use a KL expansion to approximate the random process w_t, that is, replace the random process w_t by its principal component approximation with random variables and solve a larger dimension deterministic ordinary differential equation (ODE) to approximate the true solution x_t. This approach requires higher online computational expense as the dimension of a deterministic ODE increases, even though the system data for such an ODE can be precomputed.

Heuristic Convexification Methods for Chance Constraints with Stochastic Parametric Uncertainty Here the convexification methods are illustrated for the prototypical stochastic MPC problem

$$\begin{aligned}\min_{u_{0:T-1}} \ & \mathbf{E}\left[\sum_{t=1}^{T} x_t^{\mathrm{T}} Q_t x_t + u_{t-1}^{\mathrm{T}} R_{t-1} u_{t-1}\right] \\ \text{s.t.} \ & x_{t+1} = A(\delta)x_t + B(\delta)u_t, \\ & \Pr_{\Delta}[H_t x_t \geq b_t] \leq \epsilon_t, \\ & u_{\min,t} \leq u_t \leq u_{\max,t},\end{aligned} \tag{8.26}$$

with the stochastic uncertainties δ (the incorporation of the external noise perturbation is straightforward as described in the theoretical parts of this chapter but not included in this example to shorten the presentation). The constraints are defined over the time interval $[1, T]$ for x_t and $[0, T-1]$ for u_t, and this time interval consideration is omitted here for notational convenience and will always be clear from the context. The optimization (8.26) is further simplified by replacing the chance constraint $\Pr_{\Delta}[H_t x_t \geq b_t] \leq \epsilon_t$ by $H_t \mathbf{E}[x_t] \leq b_t - \beta_t$ where $\beta_t > 0$ is an additional decision variable. By doing this, the optimization (8.26), in which the dynamic constraint is an SDE, reduces to the deterministic optimization

$$\begin{aligned}\min_{u_{0:T-1},\, \beta_{1:T}} \ & \sum_{t=1}^{T} \left(X_t^{\mathrm{T}} \bar{Q}_t X_t + u_{t-1}^{\mathrm{T}} R_{t-1} u_{t-1}\right) - \gamma \sum_{t=1}^{T} \ell(\beta_t) \\ \text{s.t.} \ & X_{t+1} = F X_t + G u_t, \\ & c_0 H_t X_t^0 + \beta_t \leq b_t, \ \beta_t > 0, \\ & u_{\min,t} \leq u_t \leq u_{\max,t},\end{aligned} \tag{8.27}$$

where x_t in the constraint of the optimization (8.26) is approximated by $\hat{x}_t$ in (8.15), $\bar{Q}_t \triangleq \mathbf{E}\left[(\phi(\zeta) \otimes \mathrm{I}_n)^{\mathrm{T}} Q_t (\phi(\zeta) \otimes \mathrm{I}_n)\right]$, $c_0 \triangleq \langle 1, \phi_0(\zeta) \rangle$, $\phi(\zeta) \triangleq \mathrm{col}(\phi_i)$, $\gamma > 0$ is a user-defined weight in the optimization that corresponds to the maximization of the feasibility of the chance constraint $\Pr_{\Delta}[H_t x_t \geq b_t] \leq \epsilon_t$, and $\ell(\beta_t)$ is an incentive for decision variables to maximize the feasibility of the chance constraint $\Pr_{\Delta}[H_t x_t \geq b_t] \leq \epsilon_t$; a typical choice can be $\sum_{i=1}^{m} \beta_{t,i}$ or $\max_i \beta_{t,i}$ that is linear in β_t, where $\beta_{t,i}$ is the ith entry of β_t.[5] The constrained optimization (8.27) is a convex QP that can be solved efficiently.[6]

For a different formulation of constraints, consider constraints on the deviation of the solution trajectory

[5] To be a convex program, $\|\beta_t\|_2$ cannot be used, since it results in a concave function term in the objective function in a minimization problem.

[6] By *efficiency*, it is meant that there is a numerical algorithm whose convergence is guaranteed and that provides an infeasibility certificate. Convex programs are such cases.

from the expectation:

$$\begin{aligned}\min_{u_{0:T-1}} \quad & \sum_{t=1}^{T} X_t^{\mathrm{T}} \bar{Q}_t X_t + u_{t-1}^{\mathrm{T}} R_{t-1} u_{t-1} \\ \text{s.t.} \quad & X_{t+1} = F X_t + G u_t, \\ & c_0 H_t X_t^0 \leq b_t, \\ & (\bar{X}_t^i)^{\mathrm{T}} W \bar{X}_t^i - (c_0 X_{t,i}^0)^2 \leq \sigma_{t,i}^2, \\ & u_{\min,t} \leq u_t \leq u_{\max,t},\end{aligned} \tag{8.28}$$

where $\bar{Q}_t$ and c_0 are the same as defined before, $W \triangleq \mathrm{diag}(\|\phi_i\|^2)$, and $\bar{X}_t^i$ denote the concatenation of coefficients of the polynomial expansion for the ith state. The constant vector c_0 and matrix W can be assumed to be normalized to be $\mathbf{1}_n$ and I_n without loss of generality. The constrained optimization (8.28) is a convex QCQP that can be solved efficiently. More precisely, it is not hard to see that the optimization (8.28) can be rewritten as

$$\min_u \mathcal{Q}_0(u) \;\; \text{s.t.} \;\; \mathcal{Q}_i(u) \leq 0, \;\; i = 1, \ldots, m_q \tag{8.29}$$

where $u \triangleq u_{0:T-1}$ and $\mathcal{Q}_i$ are convex quadratic forms for all $i = 0, 1, \ldots, m_q$. From [179],[7] if the constraints $\mathcal{Q}_i \leq 0$ are regular, i.e., satisfy a constraint qualification such as Slater's condition [36, Sec. 5.2.3], then the static optimization (8.29) has the same optimum as the optimization

$$\max_{\lambda \geq 0} \min_u \mathcal{Q}_0(u) + \sum_i^{m_q} \lambda_i \mathcal{Q}_i(u) \tag{8.30}$$

for which fixing arbitrarily large $\lambda_i > 0$ results in the same optimal solution u^* as obtained from solving the constrained optimization (8.29).

The conceptual picture of a constrained trajectory in Fig. 8.4 shows how the constraints in (8.28) can be used to impose desired bounds on the controlled trajectory.

8.3.5 The Use of Concentration-of-Measure Inequalities for Probabilistic Validation Certificates of Joint Chance Constraints

This subsection shows that the Boole inequality can be incorporated into some well-known concentration-of-measure inequalities to provide probabilistic validation certificates for joint chance constraints. Consider the constraint $H^{\mathrm{T}} X \leq b$, or equivalently $h_i^{\mathrm{T}} X \leq b_i$ for $i = 1, \ldots, m$ where X is a random vector and h_i denotes the ith column of the matrix H. The associated probabilistic constraint can be written as $\Pr[H^{\mathrm{T}} X > b] = \Pr\left[\cup_{i=1}^m \{h_i^{\mathrm{T}} X > b_i\}\right]$. The Boole inequality gives an upper bound on this probabilistic

[7] They applied a method of relaxation called the *S-procedure* [295]. Our case is a special case in which all the quadratic forms are convex, whereas [179] considered more general cases where some of quadratic forms might be nonconvex. They proposed a sufficient condition for the set of quadratic form constraints to be lossless, i.e., the resultant relaxation obtained from the S-procedure gives an exact optimum.

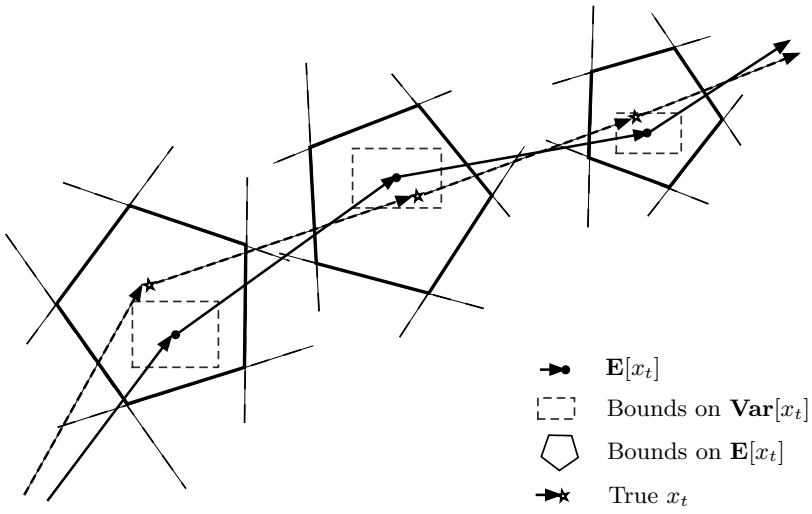


Figure 8.4: A schematic cartoon of constrained state trajectories with semi-chance constraints.

violation of constraints:

$$\Pr\left[\bigcup_{i=1}^{m}\{h_i^{\mathrm{T}}X > b_i\}\right] \leq \sum_{i=1}^{m}\Pr\left[h_i^{\mathrm{T}}X > b_i\right]. \tag{8.31}$$

Suppose that $b_i > 0$ for all $i = 1, \ldots, m$ without loss of generality. Some concentration-of-measure inequalities can be used for upper bounds on the right-hand side of (8.31) [34]:

- The Chernoff's bound: $\Pr\left[h_i^{\mathrm{T}}X > b_i\right] \leq \frac{\mathbf{E}[e^{s_i h_i^{\mathrm{T}}X}]}{e^{s_i b_i}}$ where $s_i > 0$ for all $i = 1, \ldots, m$.
- The generalized Markov inequality: $\Pr\left[h_i^{\mathrm{T}}X > b_i\right] \leq \frac{\mathbf{E}[\phi_i(h_i^{\mathrm{T}}X)]}{\phi_i(b_i)}$ where $\phi_i : \mathbb{R} \to \mathbb{R}_+$ for all $i = 1, \ldots, m$.
- The Chebyshev inequality: $\Pr\left[|h_i^{\mathrm{T}}X - \mathbf{E}[h_i^{\mathrm{T}}X]| > t_i\right] \leq \frac{\mathrm{Var}(h_i^{\mathrm{T}}X)}{t_i^2} = \frac{h_i^{\mathrm{T}}\mathrm{Var}(X)h_i}{t_i^2}$.

Here we use the Chebyshev inequality.

Proposition 8.4. If the random vector X satisfies the constraints on its expectation and variance

$$h_i^{\mathrm{T}}X \leq b_i,\ h_i^{\mathrm{T}}\mathrm{Var}(X)h_i \leq t_i^2\epsilon_i, \quad i = 1, \ldots, m \tag{8.32}$$

then the inequality $H^{\mathrm{T}}X \leq b + t$ is satisfied with at least probability $1 - \sum_{i=1}^{m}\epsilon_i$, i.e., $\Pr\left[H^{\mathrm{T}}X \leq b + t\right] \geq 1 - \sum_{i=1}^{m}\epsilon_i$.

Fig. 8.5 illustrates the outer polytopic certificate (colored in red) for the associated chance constraint $\Pr[H^{\mathrm{T}}X > b] \leq \epsilon$ with $\sum_{i=1}^{m}\epsilon_i \leq \epsilon$. The polytope colored in blue corresponds to the constraints on the expectation of the trajectories. Such certificates can be computed from the results in Prop. 8.4.

From (8.18) and the results in Prop. 8.1, gPC expansions can provide closed forms for the expectation and variance of the controlled predicted state and output trajectories. This implies that any chance constraints of polyhedral inequalities can be certificated by deterministic polyhedral inequalities that are obtained from

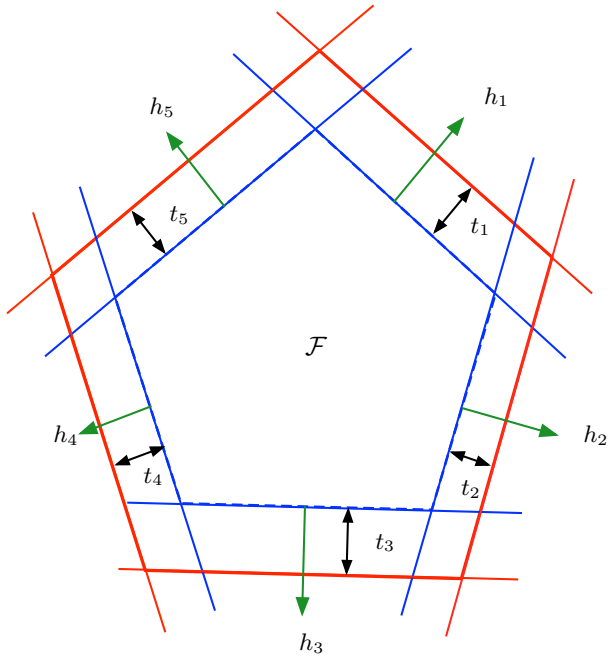


Figure 8.5: A cartoon of the outer bound that can be obtained from the Boole inequality and concentration-of-measure inequalities.

incorporating gPC expansions into the Boole inequality and the Chebyshev inequality of the form presented in Prop. 8.4.

8.3.6 Affine Feedback Control Policy

The aforementioned methods for computation of a suboptimal control policy use the new measurements to compute a control action as well as to initialize the state-transition constraint in the optimizations at each step of prediction. In the presence of model/plant mismatch and external disturbances, the predicted control trajectory at time k can significantly deviate from the true controlled trajectory and the variance of the trajectory can increase such that the optimization is feasible only for a short time horizon, which is undesirable in terms of closed-loop stability. This situation can be avoided by incorporating a feedback control in each step of solving the optimization, as has been done in many deterministic robust MPC formulations (e.g., see [149]). For example, the affine control law $u_t = K_t z_t + \nu_t$ can be inserted into the optimization, where z_t is an estimated state or measured output. For a precomputed K_t (or a stationary control gain K), the resultant problem is exactly same as the open-loop feedback control in which ν_t is the only decision variable in each step of optimization. If K_t is considered as an additional decision variable in each step of optimization, then the resulting optimization will retain the same degree of convexification as for u_t considered before.

8.3.7 Time-varying Uncertain Parameters

Consider the system dynamics given in (8.1) where the uncertain parameter vector $\delta \in \Delta$ (or $\theta = (\delta, x_0) \in \Theta$, if the initial condition is considered to be uncertain) is assumed to be an unknown constant vector. In the aforementioned MPC formulations, the uncertain parameters were assumed to be fixed only in the prediction step. The approaches can also be applied to slowly time-varying uncertain parameters, that is, when the prediction horizon multiplied by the sampling interval is less than the time interval of significant parameter variation. A more accurate study of time-varying uncertain parameters can be performed by considering a large dimensional space of uncertain parameters. In particular, for the time-varying uncertain parameter vector $\delta_t \in \Delta$, consider the stacked vector $\delta_{0:T-1} \triangleq [\delta_0', \cdots, \delta_{T-1}']' \in \Delta^T$ where the superscript $'$ denotes the transpose and T denotes the prediction horizon. Then the approximation based on a polynomial chaos expansion is represented in terms of the stacked uncertain parameter vector $\delta_{0:T-1}$. This approach requires more basis functions for the corresponding spectral representation, but the time-dependent coefficients corresponding to the uncertain parameters in future can be set to zeros, which reduces the computation of projections to determine the coefficients of the gPC expansion.

8.4 Summary and Future Work

This chapter considers a new approach for stochastic MPC problems in the presence of both parametric model uncertainty and exogenous stochastic disturbances. To approximate the solution of a stochastic differential equation and solve the corresponding stochastic MPC problem, a spectral method known as generalized polynomial chaos expansion is applied and constraints corresponding to the probability of safety/collision are imposed on the approximately predicted controlled trajectories, based on the model of a stochastic differential equation. The first and second moments of the approximate solution were exploited to estimate the probability distribution of the true solution. Under these technical assumptions, the chance constraints were replaced by convex constraints for the mean and covariance of the trajectory that are analytically computed from the gPC expansion. It was also shown that concentration-of-measure inequalities combined with the Boole inequality can provide conservative probabilistic certificates for chance constraints of polyhedral inequalities, for which applications of the gPC expansions are straightforward. Further studies to follow are to apply the presented methods to more complicated case studies and compare the heuristic convexification methods to convex nonlinear programs in terms of the tradeoff between complexity and accuracy.

Part III

STATISTICAL INFERENCE FOR FAULT DIAGNOSIS

Chapter 9

Bayesian Hypothesis Tests: An Overview

Abstract This chapter provides a concise introduction to methods of Bayesian hypothesis testing and generalizes some of the existing work in the literature. For details on Bayesian theory and its applications to statistical hypothesis testing, readers are referred to [166, 220], for example. The main applications of consideration are change detection and fault detection and diagnosis in stochastic dynamical systems, for which statistical inference problems based on the Bayesian theory of statistics are formulated as mathematical programs. Computational tractability, scalability, reliability, and robustness of general Bayesian hypothesis tests for large-scale inference problems are active research topics. Simulation results are presented to illustrate a certain type of Bayesian hypothesis testing known as likelihood ratio tests, for which a non-interacting two-tanks model is considered with several fault scenarios. In addition, some open research directions for robust Bayesian hypothesis tests and integrated model-based real-time optimal control in the presence of system component faults are discussed.

9.1 Bayesian Hypothesis Testing

Hypothesis testing has been extensively studied in signal processing, communications, biological statistics, observational astronomy, and so on. This area can be considered as a devision of decision science in which a choice needs to be made among multiple hypotheses on the basis of limited and noisy data. Developing a method of hypothesis tests is to provide a decision rule for the selection among multiple probable hypotheses in some principled or optimal criterion (e.g., minimization of decision error, minimization of expected cost or risk in the mismatched estimation, etc.).

9.1.1 Optimum Decision Rules

Consider the set of hypotheses $\mathcal{H} \triangleq \{H_0, H_1, \cdots, H_m\}$ in which the i^{th} hypothesis H_i corresponds to a model to explain the observed data $z \in \mathcal{Z}$. A Bayesian hypothesis testing problem is to find an optimal hypothesis $H^\star$ that is most consistent with the observed data z in the sense that it maximizes the associated posterior distribution $p_{\mathbf{H}|\mathbf{z}}(h(z)|z)$ where $h : \mathcal{Z} \to \mathcal{H}$ is a deterministic decision rule. This problem can be formulated as the optimization

$$H^\star(z) := \arg \max_{H_i \in \mathcal{H}} p_{\mathbf{H}|\mathbf{z}}(H_i|z), \tag{9.1}$$

which is known as the *maximum a posteriori* (MAP) decision rule.

For a more general problem formulation, a similar optimization can be written as

$$\begin{aligned} &\min_{h(\cdot)\in\mathcal{H}} \underbrace{\mathbf{E}_{\mathbf{H}|\mathbf{z}}[C(H, f(z))|\mathbf{z} = z]}_{\tilde{J}(H:z)} \\ &\quad = \min_{H_i \in \mathcal{H}} \left\{ \sum_{j=0}^{m} C(H_j, H_i) p_{\mathbf{H}|\mathbf{z}}(H_j|z) \right\} \\ &\quad = c \min_{H_i \in \mathcal{H}} \left\{ \sum_{j=0}^{m} C(H_j, H_i) p_{\mathbf{z}|\mathbf{H}}(z|H_j) p_{\mathbf{H}}(H_j) \right\} \end{aligned} \tag{9.2}$$

and

$$H^\star(z) = \arg \min_{H_i \in \mathcal{H}} \left\{ \sum_{j=0}^{m} C(H_j, H_i) p_{\mathbf{z}|\mathbf{H}}(z|H_j) p_{\mathbf{H}}(H_j) \right\} \tag{9.3}$$

where $C : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$, $C(H, h(\cdot))$ refers to the cost of deciding that the hypothesis is $h(\cdot)$ when the correct hypothesis is H, $c = \sum_{k=0}^{m} p_{\mathbf{z}|\mathbf{H}}(z|H_k) p_{\mathbf{H}}(H_k)$ denotes the normalization factor that is independent of the hypothesis, and $p_{\mathbf{z}|\mathbf{H}}(z|H_j)$ and $p_{\mathbf{H}}(H_j)$ (shortly, p_j) are the likelihood function and prior distribution associated with the hypothesis H_j, respectively.

For measurement-independent optimum decision rules, consider the average cost of deciding that the hypothesis is f that is defined by

$$\begin{aligned} J(h) &\triangleq \mathbf{E}_{\mathbf{H},\mathbf{z}}[C(H, f(z))] \\ &= \mathbf{E}_{\mathbf{z}}\big[\mathbf{E}_{\mathbf{H}|\mathbf{z}}[C(H, f(z))|\mathbf{z} = z]\big] \\ &= \int_{\mathcal{Z}} \tilde{J}(h(z):z) p_{\mathbf{z}}(z) dz \\ &= \sum_{i,j} C(H_j, H_i) \underbrace{\Pr\left[h(\mathbf{z}) = H_i | \mathbf{H} = H_j\right]}_{\delta_{ji}(h)} p_j \end{aligned} \tag{9.4}$$

which is called the *Bayes risk*, where $h : \mathcal{Z} \to \mathcal{H}$ can be any type of deterministic decision rules, in general.

The quantities $\delta_{ji}(h)$ can be called the *discrimination probabilities* of the decision rule $h(\cdot)$ and can be rewritten as

$$\delta_{ji}(h) = \int_{\mathcal{Z}_i(h)} p_{\mathbf{z}|\mathbf{H}}(z|H_j)dz \tag{9.5}$$

where the sets $\mathcal{Z}_i(h)$ are the *decision regions* corresponding to the values of z for which the decision for an optimal hypothesis is H_i, i.e., $h(z) = H_i$, and can be written as

$$\mathcal{Z}_i(h) \triangleq \{z \in \mathcal{Z} : h(z) = H_i\} \tag{9.6}$$

These sets are disjoint, i.e., $\mathcal{Z}_i(h) \bigcap \mathcal{Z}_j(h) = \emptyset$ for all $i \neq j$ and $\bigcup_i \mathcal{Z}_i(h) = \mathcal{Z}$, for any deterministic decision rule $h(\cdot)$.

9.1.2 Related Work

Generalized Likelihood Ratio Test Kalman filtering and generalized likelihood ratio (GLR) tests have been used for the detection and estimation of jumps (or switches) in dynamical systems [230, 240, 290, 291]. The GLR tests for online detection of faults and parameter changes in control systems have been further developed for the optimal/suboptimal choice of threshold and the monitoring window size [153], multiple hypothesis testing problems [14] have been considered, and information-theoretic bounds for detection performance have been obtained [152]. In addition, particle filtering techniques based on simulations and samplings have been incorporated into real-time fault and parameter change detections [3, 6, 157].

Two-ellipsoid Overlap Test for Real-time Fault Detection As an alternative to GLR tests, a simple geometric test for two-ellipsoid overlap was considered [133–135, 311, 312] which is also called a *chi-squared test* for fault detection. In particular, [133–135, 311] consider two confidence regions of ellipsoidal cross section that are associated with the (monitoring) state trajectory of the normal (no-failure) operation and the Kalman estimate computed from the online observables, while [312] consider two confidence regions of ellipsoidal cross section that are associated with the (monitoring) system parameters of the normal (no-failure) operation and the recursive least squares (RLS) estimate computed from the online observables.

Hybrid Automata and Mode Estimation In [106, 107, 193], the so-called hybrid estimation problem was introduced to detect the onset of subtle faults or failures. A *hybrid estimation problem* is that, for given a probabilistic hybrid automaton (PHA) (see [104, 107] for details of PHA), compute the most likely hybrid state together with the associated system mode at the time instance t which is inferred from a sequence of control inputs $\{u_0, \cdots, u_{t-1}\}$ and measurement outputs $\{y_0, \cdots, y_t\}$.

Adaptive Filtering for State Estimation Performance of a model-based state estimator can highly rely on the accuracy of the model. In other words, the closer model to the true system, the better the performance

of state estimation. For this purpose, system mode estimation can be embedded in a state estimator such as Kalman filtering or moving horizon estimation (MHE) that uses a system model to estimate the state variables, and using a more accurate model for the true plant dynamics can achieve a better state estimation performance [246].

9.2 Performance Analysis of Bayesian Hypothesis Tests

9.2.1 Operating Characteristic

Operating Characteristic of the LRT Consider a binary hypothesis test, i.e., $\mathcal{H} = \{H_0, H_1\}$. To quantify the performance of a decision rule $h(\cdot)$, the associated discrimination probabilities (9.5) are

$$P_{\mathrm{d}}(h) := \delta_{11}(h) \quad \text{and} \quad P_{\mathrm{f}}(h) := \delta_{01}(h), \tag{9.7}$$

which are also referred to as the *detection* and *false-alarm* probabilities, respectively. The so-called *probabilities of error of the first and second kind* are defined respectively as

$$P_{\mathrm{e}}^1(h) := P_{\mathrm{f}}(h) = \delta_{01}(h) \quad \text{and} \quad P_{\mathrm{e}}^2(h) := 1 - P_{\mathrm{d}}(h) = \delta_{10}(h). \tag{9.8}$$

Neyman-Pearson Criterion In many applications, it might not be obvious to assign the costs $C(H_j, H_i)$ for the pairs of (i, j) and it is often unrealistic to assume that prior probabilities are known. The classical Neyman-Pearson criterion [199] is to choose the decision rule $h(\cdot)$ to maximize $P_{\mathrm{d}}(h)$ while $P_{\mathrm{f}}(h)$ is constrained to be smaller than a certain value:

$$\begin{aligned} &\max_{h(\cdot)\in\mathcal{H}} P_{\mathrm{d}}(h) \\ &\text{s.t. } P_{\mathrm{f}}(h) \le \alpha \end{aligned} \tag{9.9}$$

where $\alpha > 0$ is a user-defined threshold. It is not difficult to see that the optimization (9.9) can be equivalently represented by

$$\max_{h(\cdot)\in\mathcal{H}} \alpha_1 P_{\mathrm{d}}(h) - \alpha_2 P_{\mathrm{f}}(h) \tag{9.10}$$

for some appropriate positive constants α_1 and α_2 such that the decision rules based on (9.9) and (9.10) have the same optimal decision. From the definitions of P_{e}^i, another equivalent decision rule is

$$\min_{h(\cdot)\in\mathcal{H}} \sum_{i=1}^{2} w_i P_{\mathrm{e}}^i(h) \tag{9.11}$$

where $w_i > 0$ are user-defined weights for each error kind that can chosen to be functions of α such that the decision rules based on (9.9) and (9.11) have the same optimal decision.

Generalized Neyman-Pearson Criterion The aforementioned Neyman-Pearson criterion can be extended to multiple hypotheses. To maximize the ability of detecting the true hypothesis H_i while requiring the false-alarm probabilities is smaller than certain values ($\alpha_j > 0$), solve the optimization

$$\max_{h(\cdot)\in\mathcal{H}} \delta_{ii}(h) \text{ subject to } \delta_{ji}(h) \leq \alpha_j,\ j \neq i. \tag{9.12}$$

An alternative decision rule can be obtained from minimizing the weighted-sum of probabilities of errors:

$$\min_{h(\cdot)\in\mathcal{H}} \sum_{j=0}^{m} \sum_{i\neq j} w_{ji}\delta_{ji}(h) \tag{9.13}$$

where $w_{ji} > 0$ for $j \neq i$ are the user-defined weights.

9.2.2 The Bayes Risk Error

Bayesian hypothesis tests that find a decision rule minimizing the Bayes risk (9.4) rely on precise knowledge of the prior probabilities of the hypotheses, which implies that the optimal decision is sensitive to the choice of the prior probabilities. To analyze the sensitivity of the decision rule to the prior probabilities, consider the *mismatched* Bayes risk defined as

$$\tilde{J}(p,\bar{p}) \triangleq \sum_{i,j} C(H_j,H_i) \underbrace{\Pr\left[\hat{h}(\mathbf{z},p) = H_i | \mathbf{H} = H_j\right]}_{\delta_{ji}(h)} \bar{p}_j \tag{9.14}$$

where the decision rule $\hat{h}(\cdot,p)$ refers to the hypothesis minimizing the *virtual* Bayes risk $\tilde{J}(p,p)$ with the virtual prior probability p while the true prior probability is $\bar{p}$. Define the Bayes risk error as

$$\varphi(p,\bar{p}) \triangleq \tilde{J}(p,\bar{p}) - \tilde{J}(\bar{p},\bar{p}). \tag{9.15}$$

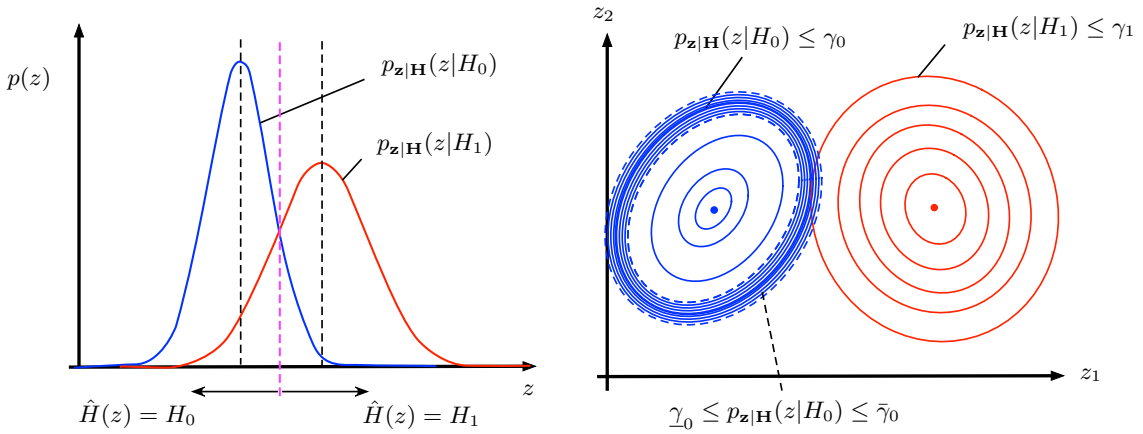
The Bayes risk error (9.15) has the following properties [275]:

i. $\varphi(p,q) \geq 0$ for all $p,q \in \mathcal{P}$ where $\mathcal{P} \triangleq \left\{p \in \mathbb{R}_+^{m+1} : \sum_i p_i = 1,\ p_i \in [0,1],\ \forall i\right\}$;

ii. $\varphi(p,q)$ is strictly convex in $q \in \mathcal{P}$ and quasi-convex in $p \in \mathcal{P}$ for deterministic LRT;

iii. $\varphi(p,q) = \tilde{J}(p,p) - \tilde{J}(q,q) - \langle p-q, \nabla_1 \tilde{J}(q,q)\rangle$ where ∇_1 denotes the partial derivative with respect to the first argument of a function.

Note that this function is not symmetric, i.e., $\varphi(p,q) \neq \varphi(q,p)$ for $p,q \in \mathcal{P}$, in general.



(a) Decision rule for a likelihood ratio test

(b) Bayes risk error with respect to level set descrptions

Figure 9.1: A likelihood ratio test, the Bayes risk, and sensitivity of the associated Bayesian hypothesis test: Statistical distance measure can be used for quantification of the Bayes risk and the sensitivity of the associated test with respect to the choice of prior probabilities and the costs of decision errors.

9.3 A Case Study of Two Tanks in Series

This section presents simulation results for the two-tank configuration in Figure 9.2, in which several fault scenarios are presumed and generalized likelihood ratio tests[1] are applied for model selection and checking. The system in the disturbance-free and fault-free case is governed by the material balance equations

$$\begin{aligned} A_{c1}\frac{dh_1}{dt} &= F_i - c_1 h_1 \\ A_{c2}\frac{dh_2}{dt} &= c_1 h_1 - c_2 h_2 \end{aligned} \tag{9.16}$$

where A_{c1} and A_{c2} are the cross-sectional areas of Tanks 1 and 2, h_1 and h_2 are the liquid levels for Tanks 1 and 2, c_1 and c_2 are constants which depend on the valves, and F_i is the measured inlet flow rate. The outlet flow rates are F_{o1} and F_{o2} are measured, and are nominally equal to $c_1 h_1$ and $c_2 h_2$, respectively. Eq. (9.16) can be rewritten in state-space form

$$\begin{aligned} \frac{dx}{dt}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \end{aligned} \tag{9.17}$$

where

$$A = \begin{bmatrix} -\frac{c_1}{A_{c1}} & 0 \\ \frac{c_1}{A_{c2}} & -\frac{c_2}{A_{c2}} \end{bmatrix}, \; B = \begin{bmatrix} \frac{1}{A_{c1}} \\ 0 \end{bmatrix}, \; C = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix},$$

$u = F_i$, $y = [F_{o1}, F_{o2}]^{\mathrm{T}}$, and $x = [h_1, h_2]^{\mathrm{T}}$.

[1] It is known that every aforementioned Bayesian hypothesis test can be equivalently transformed into a likelihood ratio test with a properly chosen value of threshold [220]. Due to this equivalence, the focus of this simulation study is on the use of likelihood ratio tests for detection of changes and faults.

Consider the case where process and measurement noise are explicitly introduced into the state-space equation (9.17) and the corresponding state-space form is

$$\begin{aligned} \frac{dx}{dt}(t) &= Ax(t) + Bu(t) + Ew(t) \\ y(t) &= Cx(t) + Fv(t) \end{aligned} \tag{9.18}$$

where $E = F = \mathrm{I}$, and $w(t)$ and $v(t)$ are independent zero-mean Gaussian white noise having correlation matrices Σ_w and Σ_v, respectively, i.e., $w_t \sim \mathcal{N}(0, \Sigma_w)$ and $v_t \sim \mathcal{N}(0, \Sigma_v)$ for all t, and $\mathbf{E}[w_t w_s^{\mathrm{T}}] = 0 = \mathbf{E}[v_t v_s^{\mathrm{T}}]$ for $t \neq s$. Consider discrete-time system dynamics for which the first-order hold discretization method is applied to 9.18 and the control inputs are assumed piecewise linear over the sampling period $t_s = 0.1$ seconds.

The likelihood ratio tests are applied for the following fault scenarios:

S1) *A leak in the injet flow:* In this fault scenario, suppose that there is a leak in Stream 1 entering Tank 1 (see Figure 9.2). Due to this leak, the inlet flow to the upper tank decreases by 30%, compared to the normal operation case. Figures 9.3(a) and 9.3(b) show the corresponding output trajectories and log likelihood ratio tests, respectively.

S2) *A biased output measurement:* Another probable fault is a sensor fault in which the sensor measuring the output flow rate F_{o2} is malfunctioning and shows a bias. Figures 9.4(a) and 9.4(b) show the corresponding output trajectories and log likelihood ratio tests, respectively.

S3) *Two competing fault scenarios (S1 vs. S2):* In this scenario, suppose that two aforementioned fault scenarios S1 and S2 are presumed to be a current fault and compared with respect to likelihood ratio tests. Figures 9.5(a) and 9.5(b) show the corresponding output trajectories and log likelihood ratio tests, respectively.

S4) *Three competing fault scenarios (normal operation vs. S1 vs. S2):* For the final test, compare and assess three fault scenarios, viz., the normal operation, and the faults scenarios S1 and S2. Figures 9.6(a) and 9.6(b) show the corresponding output trajectories and log likelihood ratio tests, respectively.

For the (log) likelihood ratio tests in the simulations presented in this chapter, the moving monitoring window with a length of 10 sampling intervals is used. The longer window length, the longer delay in fault detection using likelihood ratio tests. For given hypothesis model, the corresponding likelihood function depends on the measurements that are realizations of random processes from unknown probability distributions and is indeed a random process, same as the associated likelihood ratio test, which implies that the fault detection instance—the time instance when the likelihood ratio curve crosses the threshold bar—also has a probabilistic nature.

The likelihood ratio tests correctly identify the change in the dynamic behavior in all of the scenarios, with a short detection delay.

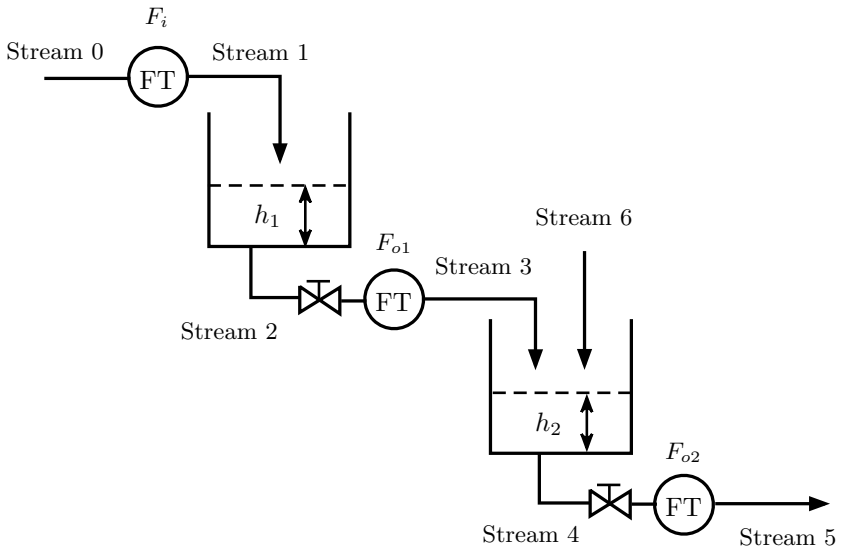


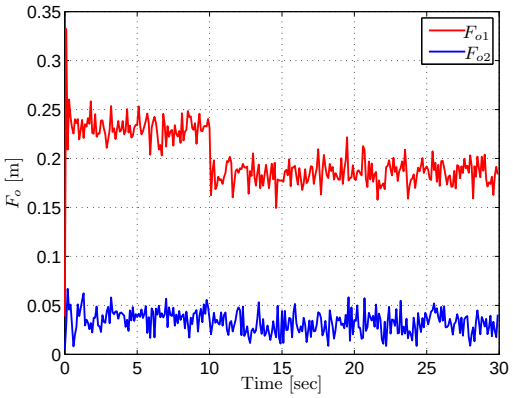
Figure 9.2: Two non-interacting flow tanks in series.

9.4 Discussions

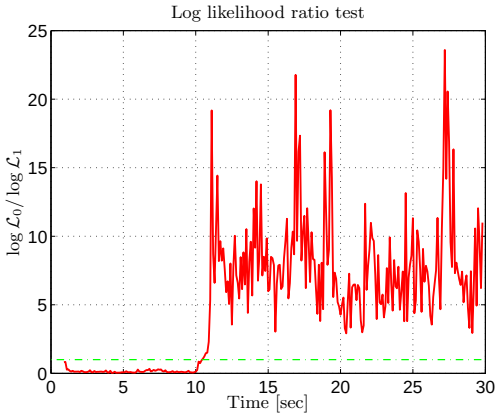
This section discusses (i) a generalization of Bayesian hypothesis tests in which discrete hypotheses are parameterized by continuous uncertain random variables, and (ii) incorporation of fault detection and diagnosis based on Bayesian hypothesis tests into real-time model-based (sub-)optimal control known as model predictive control and approximate dynamic programming. The discussions are somewhat intuitive and abstract. The objective of presenting these discussions is not to present a concrete support for the ideas, but to provide some open discussions for future research directions.

9.4.1 Parameterized Bayesian Hypothesis Testing

It is straightforward that establishing an estimate for a continuous random variable involves carrying out a hypothesis test for a continuum of hypotheses, rather than discrete hypotheses, with principled or optimal decision rules. For fault scenarios in which there are a finite number of discrete hypothesized models of candidates, an optimal decision rule is constructed such that the associated optimality criterion is evaluated for such discrete hypotheses. However, it is often difficult to know precise models corresponding to certain fault scenarios and it is necessary to parameterize models of hypotheses to capture variations of faults. Such further parameterization can be referred to as *fault-parameterization* or *fault-parameters* and this parameterization has essential importance in robust fault detection and diagnosis in the presence of additional uncertainty in hypotheses, as well as disturbances and measurement noise.

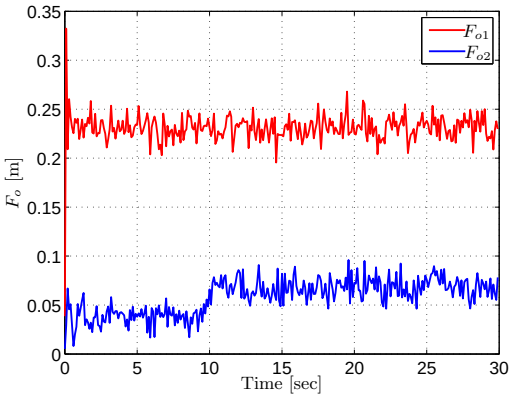


(a) The measured output of two tanks in series.

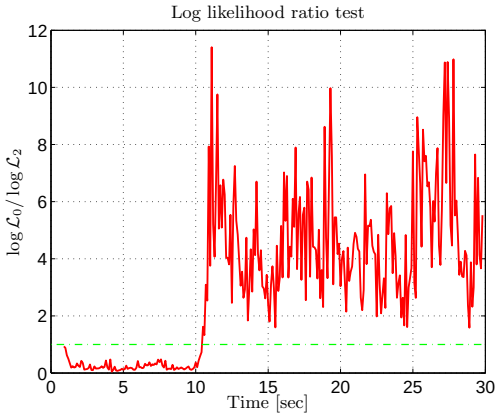


(b) Log likelihood ratio of the models for the normal operation and Fault 1.

Figure 9.3: Fault scenario 1: A leak in Stream 1 occurs at $t = 10.0$ sec.



(a) The measured output of two tanks in series.



(b) Log likelihood ratio of the models for the normal operation and Fault 2.

Figure 9.4: Fault scenario 2: The output F_{o2} gives a biased reading that occurs at $t = 10.0$ sec.

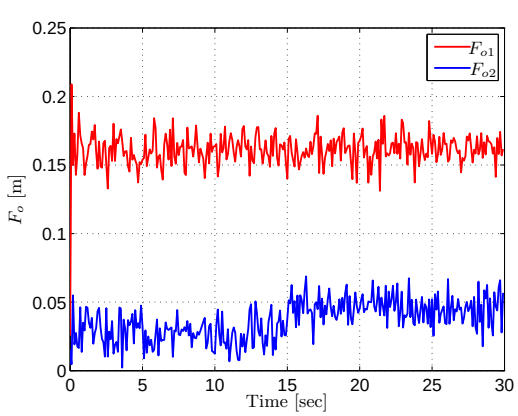
A set-valued model representation is

$$\mathcal{M}^k_{\Theta_k} \triangleq \{M_k(\theta_k) : \theta_k \in \Theta_k\}$$

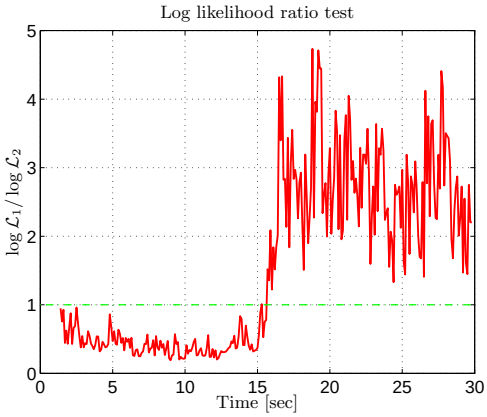
where Θ_k denotes the set of fault-parameters for the kth hypothesis. For given measures $\mathbf{y}$ and fixed hypothesis of model k, the maximum likelihood estimation is to find an optimal solution for

$$\max_{\theta_k \in \Theta_k;\, k=0,\cdots,m} \mathcal{L}_k(\theta_k; \mathbf{y}).$$

For designating the robustly most probable model explaining the measurements $\mathbf{y}$, determine an index $k^\star$



(a) The measured output of two tanks in series.



(b) Log likelihood ratio of the models for Fault 1 and Fault 2.

Figure 9.5: Fault scenario 3: A fault switching from Fault 1 to Fault 2 at $t = 15.0$ sec.

satisfying the robust optimality relation

$$\sup_{\theta_{k^\star} \in \Theta_{k^\star}} \mathcal{L}_{k^\star}(\theta; \mathbf{y}) \geq \sup_{\theta_k \in \Theta_k} \mathcal{L}_k(\theta_k; \mathbf{y}) \quad \text{for all } k \neq k^\star.$$

A resultant optimal solution among families of fault-parameter sets

$$\theta^\star_{k^\star} \triangleq \arg \max_{\theta_{k^\star} \in \Theta_{k^\star}} \mathcal{L}_{k^\star}(\theta_{k^\star}; \mathbf{y})$$

can be used to determine the associated most likely system mode

$$M_{k^\star}(\theta^\star_{k^\star}) \in \mathcal{M}^{k^\star}_{\Theta_{k^\star}}.$$

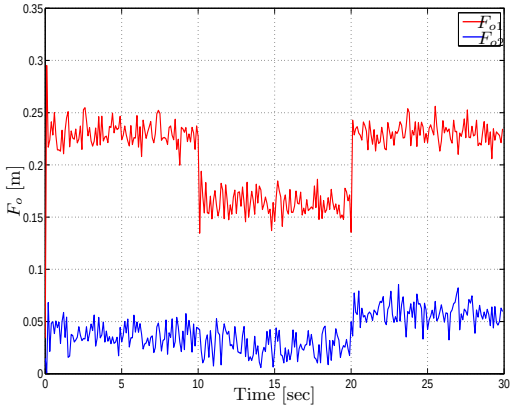
9.4.2 Integrated Real-time Model-based Optimal Control

Consider the discrete-time linear time-varying stochastic system

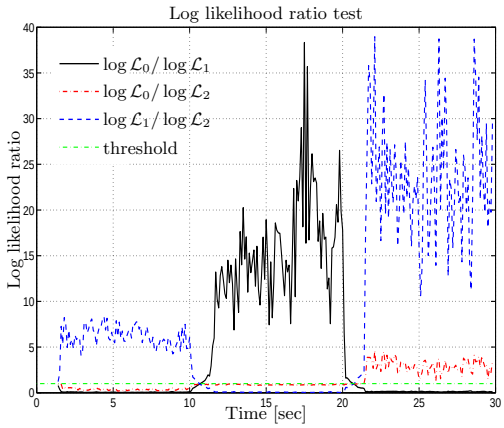
$$\begin{aligned} x_{t+1} &= A_t x_t + B_t u_t + E_t w_t \\ y_t &= C_t x_t + D_t u_t + F_t v_t \end{aligned} \tag{9.19}$$

where w and v are independent Wiener processes, i.e., $w_t \sim \mathcal{N}(\mu_w, \Sigma_w)$, $v_t \sim \mathcal{N}(\mu_v, \Sigma_v)$ for all t, $\mathbf{E}[w_t w_s^{\mathrm{T}}] = 0$ and $\mathbf{E}[v_t v_s^{\mathrm{T}}] = 0$ for all $t \neq s$, and $\mathbf{E}[w_t v_s^{\mathrm{T}}] = 0$ for all t and s.

Suppose that there are m scenarios of faults and the model associated with the i^{th} fault scenario is given



(a) The measured output of two tanks in series during model switchings.



(b) Log likelihood ratio of the models for non-fault, Fault 1, and Fault 2.

Figure 9.6: Fault scenario 4: Model switchings: (a) Normal operation (0.0–10.0 sec), (b) Fault 1 (10.0–20.0 sec), and (c) Fault 2 (20.0–30.0 sec).

by

$$M_i : \begin{cases} x_{t+1}^i = A_t^i x_t^i + B_t^i u_t^i + E_t^i w_t^i \\ y_t^i = C_t^i x_t^i + D_t^i u_t^i + F_t^i v_t^i \end{cases} \tag{9.20}$$

where the superscript i refers to the occurrence of the i^{th} hypothesized fault. Roughly speaking, a procedure of fault detection and diagnosis is to find the most probable model from the set of hypothesized models $\mathcal{M} \triangleq \{M_0, \cdots, M_m\}$ and Bayes' rule is applied to quantified probabilistic confidence levels as functions of the measurements for all the hypothesized models.

For a control strategy, consider the standard stochastic model predictive control (MPC) problem for a discrete-time linear time-varying system (10.1):

$$\begin{aligned} \min_{u_{0:h-1}} \ & \mathbf{E}\left[\sum_{t=1}^{h} \left(x_t^{\mathrm{T}} Q_t x_t + u_{t-1}^{\mathrm{T}} R_{t-1} u_{t-1}\right)\right] \\ \text{s.t.} \ \ & x_{t+1} = A_t x_t + B_t u_t + E_t w_t, \\ & y_t = C_t x_t + D_t u_t + F_t v_t, \\ & \Pr[H_{y,t} y_t + H_{u,t} u_t \le b_t] \ge \beta_t, \\ & w_t \sim p_w, \ v_t \sim p_v, \ x_0 \sim p_{x_0}, \end{aligned} \tag{9.21}$$

where the constraints are imposed for all t from 0 to $h-1$ or h whose commitment is clear from the context and its explicit description is omitted for notational convenience. Note that $t = 0$ is the time at which a new (current) measurement is obtained and a new prediction step for optimization starts. For the sake of simplicity and without loss of generality, always consider the interval $[0, h]$ in the remainder of this thesis. As a standard receding-horizon control scheme, only the first step of the computed control strategy

is implemented. The calculations are repeated starting from the current measured output and estimated state, yielding a new control and new predicted output and state paths. The prediction horizon keeps being shifted forward.

Figure 9.7 represents an integrated real-time control scheme in which model-based real-time predictive control incorporates or is combined with fault detection and diagnosis (FDD), and a state estimator. Figure 9.8 is an extension of an integrated control scheme in which an additional active probing input design mechanism is included to increase the performance of statistical inference associated with FDD. Each component in such an integrated real-time control scheme shares continuous and/or discrete signals that include information required for the components to achieve their own functionality to maintain the integrity and operational objectives of the overall system. Table 9.1 shows signal flows through the components of overall closed-loop system.

The associated MPC problem can be written as an optimization

$$\begin{aligned}
&\min_{u_{0:h-1}} \ \mathbf{E}[J_{i^\star}(x_0, u_{0:h-1})] \\
&\text{s.t. } x_{t+1}^j = A_t^j x_t^j + B_t^j u_t + E_t^j w_t^j;\ j = 0, \ldots, m, \\
&\quad y_t^j = C_t^j x_t^j + D_t^j u_t + F_t^j v_{tj};\ j = 0, \ldots, m, \\
&\quad \Pr[H_{y,t}^{i^\star} y_t^{i^\star} + H_{u,t}^{i^\star} u_t \le b_t^{i^\star}] \ge \beta_t^{i^\star}, \\
&\quad \mathbf{E}[J_j(x_0, u_{0:h-1})] \le \gamma_t^j;\ j \ne i^\star, \\
&\quad w_t^j \sim p_w^j,\ v_t^j \sim p_v^j,\ x_0 \sim p_{x_0};\ j = 0, \ldots, m,
\end{aligned} \tag{9.22}$$

where the index $i^\star$ refers to the system mode corresponding the most probable model determined from a decision process in the FDD procedure and

$$J_j(x_0, u_{0:h-1}) \triangleq \sum_{t=1}^{h} \left(x_t^{j\mathrm{T}} Q_t^j x_t^j + u_{t-1}^{\mathrm{T}} R_{t-1}^j u_{t-1} \right)$$

denotes the performance criterion related to the j^{th} system mode. An estimated initial condition x_0 of each receding horizon problem is assumed to be available from a state estimator and is considered as a random variable without loss of generality.

Types (Outputs\Inputs)	State Estimator	FDD	Active FDD	MPC	Command	Output
State Estimator		s	u_a	u		y
FDD	p		u_a	u		y
Active FDD	$\hat{x}$	q				
MPC	$\hat{x}$	q			r	
Plant			u_a	u		

Table 9.1: Signal flows.

Remark 9.1. The constraint $\mathbf{E}[J_j(x_0, u_{0:h-1})] \leq \gamma_t^j$ for $j \neq i^\star$ is used to impose the expected outputs of the j^{th} system mode on a compact level set corresponding to its performance criterion, and a feasible solution enables the state trajectory of the j^{th} system mode to be kept inside a feasibility set. Similar to the chance constraint on the output trajectory of the (primary) $i^\star$ system mode, this constraint could be replaced by a chance constraint $\Pr[H_{y,t}^j y_t^j + H_{u,t}^j u_t \leq b_t^j] \geq \beta_t^j$ for all $t \in [1, h]$ and for $j \neq i^\star$.

9.5 Summary and Remarks

This chapter provides a concise overview of Bayesian hypothesis testing, and discusses some open research directions for robust hypothesis tests and model-based real-time reliable optimal control methods integrating estimation and control tools. A Bayesian hypothesis test is a statistical method for making an optimal decision that reduces to finding a decision rule that minimizes the Bayes risk. The Bayes risk is sensitive to a specific choice of the prior probabilities and the performance of a decision rule based on Bayesian hypothesis testing depends on the statistical properties of the hypotheses. The next chapter considers design problems of optimal probing inputs for maximizing the performance of Bayesian hypothesis testing, in particular for fault detection and diagnosis of stochastic LTI dynamical systems.

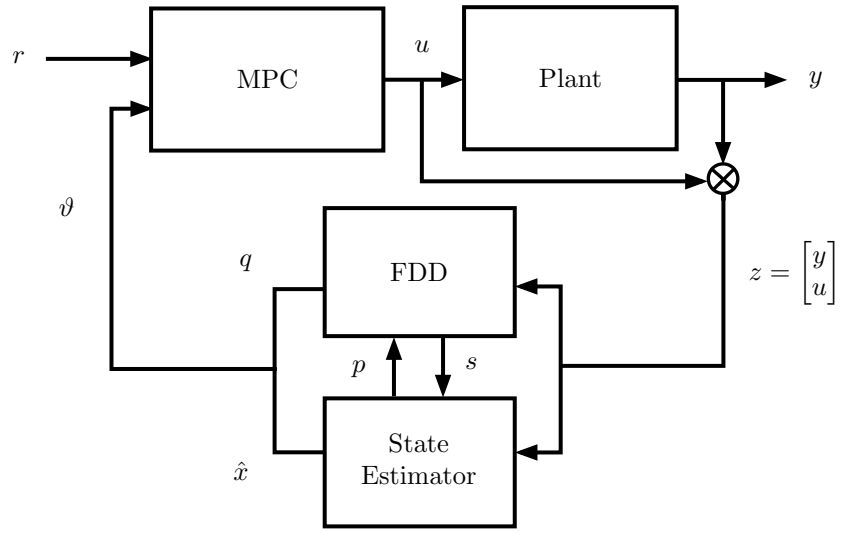


Figure 9.7: An integration of state estimator, fault detection and diagnosis algorithm, and model predictive control.

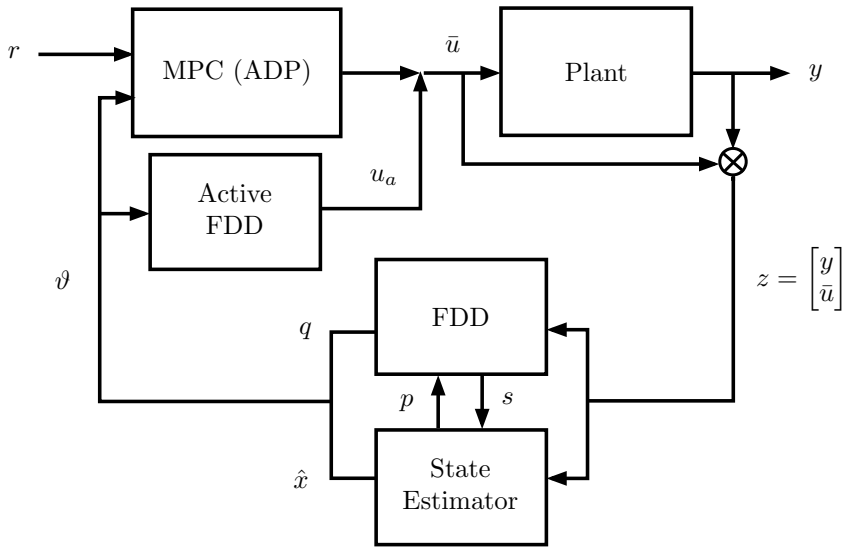


Figure 9.8: A general integration of state estimator, fault detection and diagnosis algorithm, and model predictive control, equipped with active probing input design.

Chapter 10

Optimal Probing Inputs for Statistical FDD

Abstract This chapter considers optimal/suboptimal *active* input design problems for fault detection and diagnosis (FDD). The design problems are formulated as optimizations in which an optimal sequence of inputs within a prediction horizon is computed for maximizing the statistical discrimination of different models of fault scenarios. The optimality criteria are information theoretic measures of the statistical distance between probability distributions and constraints on the predicted controlled output trajectory are imposed for ensuring operational safety as well as the input constraints that correspond to hardware limitations. Two different approaches to such constrained optimal input design problems are presented. The first scheme is to compute an optima input sequence for maximizing discrimination between system models of fault scenarios in a statistical sense. Two different measures quantifying the degree of distinguishability between two stochastic LTI system models are considered, and their geometric properties are investigated. Their connection to the generalized likelihood ratio tests are also presented. The resultant constrained open- and closed-loop feedback input design problems are shown to be concave programs and an iteration algorithm to solve these special families of nonlinear programs is presented, in which semidefinite programs are sequentially solved and a local optimum can be achieved. The second scheme is semidefinite programming (SDP) relaxation in which three different measures for the degree of statistical discrimination between two hypothesized stochastic dynamical system models are considered and their mathematical properties that are related to Bayesian hypothesis tests are studied. The resulting input design problems are non-convex and we propose associated convex relaxation methods that can be solved in polynomial time using interior point methods. In addition, an upper bound on the sub-optimality of the proposed convex relaxation is presented for the case when there is only the input amplitude constraint, and randomized algorithms are presented to compute a suboptimal solution from an optimal solution of the convex relaxation problem. Receding horizon method is used to implement the computed inputs for both approaches for constrained optimal probing input design. Numerical simulations with an aircraft model are provided to illustrate and demonstrate the presented methods of optimal input design for FDD.

10.1 Introduction

The complexity of devices and processes implies that faults are inevitable, and the tight interactions between instrumentation and other components of the overall system can result in cascading effects with significant economic, environmental, and human damages. To properly and safely operate the facilities and devices in real-time while preventing any unallowable behaviors of the system, reliable FDD algorithms are needed that monitor the inputs and outputs of the system and determines whether a fault occurs and to point to the location of the fault (aka *fault diagnosis*). In addition, without an optimal integration between the monitoring and control systems, the response to faults can reduce reliability and profit or can be overly conservative, for example, by initiating an unnecessary automated shutdown of the facility due to false alarm.

The design of FDD procedures are challenged by the presence of disturbances, noise, and model uncertainties that can make the symptoms of faults/failures indiscernible. Classical *passive* FDD methods monitor the observables of the system and make a decision on whether and where a fault has occurred, whereas *active* FDD approaches intentionally intervene in the operations and attempt to excite or perturb the observables such that abnormal behaviors are exhibited [49, 132, 200, 244, 305]. There have been many research efforts to suggest systematic methods for such auxiliary input design, both in the stochastic system setting [32] and in the deterministic uncertain system setting [201, 202, 245]. The effects of feedback in terms of the performance of FDD and quadratic cost optimality criteria have been investigated [4,5], as have been finite- or infinite-horizon control methods for the design of active input signals for FDD [32, 244, 245].

This chapter considers three different measures for the difference between two hypothesized dynamical system models. A measure for difference between two hypothesized dynamical system models is to quantify the difference between two probability distribution functions associated with random processes that are the solutions of stochastic differential (or difference) equations. The optimality criteria for design problems are information theoretic measures of the statistical distance between probability distributions and controlled state constraints are imposed for ensuring operational safety as well as the input constraints that correspond to hardware limitations. The active input design problems are formulated as optimizations for which an optimal input or a sequence of inputs is computed to maximize distinguishability (or discrimination) of different models of fault scenarios. To quantify discrimination of two stochastic dynamical system models, we use information theoretic measures that compute statistical distances between the solutions of the associated stochastic differential equations. The resultant optimizations are non-convex (more precisely, concave programs) that are NP-hard. Two different approaches to such constrained optimal input design problems are presented. The first approach is to compute open- and closed-loop feedback input sequences from solving concave nonlinear programs, for which a sequential method of semidefinite programs (SDPs) is applied. It is observed that the approximate distance measure using geometric properties of Gaussian distributions can be compatible with a closed-loop static state feedback controller that might outperform open-loop input design

methods, whereas the direct use of the KL-divergence might not be trivial. An a byproduct, we investigate the properties of the two approximate measures of statistical distance between probability distributions associated with two different fault scenario models. The second approach is to use semidefinite programming relaxations for those non-convex optimizations, which provide suboptimal solutions for active input design problems with guaranteed bounds of performance degradation for certain special cases (e.g., when there are only box constraints on the inputs).[1] In both approaches to constrained optimal *probing* input design problems, an underlying assumption considered in this chapter is that the procedures of FDD are based on a statistical decision that solves Bayesian inference problems for which the measurements of observables are used to infer a hidden process. In this chapter, robustness is considered with respect to stochastic uncertainties, disturbances, and noises. Robust FDD can be considered as maximizing the confidence of a binary decision (for fault detection) and locating a correct hypothesis (for fault diagnosis) among many candidate fault scenarios in the presence of uncertainty in a given data set. The objective of active input design for FDD is to facilitate the associated statistical decision and maximize its robustness. In addition, as theoretical contributions for the problem formulation of optimum input design for FDD, we investigate the relations of the proposed measures of statistical distance to the generalized likelihood ratio tests and justify the use of such measures of statistical distance for quantifying the degree of discrimination between two hypothesized models.

Bayesian Inference for FDD using Multiple Models of Fault Scenarios In this chapter, we consider the discrete-time linear time-varying stochastic system

$$\begin{aligned} x_{t+1} &= A_t x_t + B_t u_t + E_t w_t \\ y_t &= C_t x_t + D_t u_t + F_t v_t \end{aligned} \tag{10.1}$$

where w and v are independent Wiener processes, i.e., $w_t \sim \mathcal{N}(\mu_w, \Sigma_w)$, $v_t \sim \mathcal{N}(\mu_v, \Sigma_v)$ for all t, $\mathbf{E}[w_t w_s^{\mathrm{T}}] = 0$ and $\mathbf{E}[v_t v_s^{\mathrm{T}}] = 0$ for all $t \neq s$, and $\mathbf{E}[w_t v_s^{\mathrm{T}}] = 0$ for all t and s.

Suppose that there are m fault scenarios and the model associated with the i^{th} fault scenario is given by

$$M_i : \begin{cases} x_{t+1}^i = A_t^i x_t^i + B_t^i u_t^i + E_t^i w_t^i \\ y_t^i = C_t^i x_t^i + D_t^i u_t^i + F_t^i v_t^i \end{cases} \tag{10.2}$$

where the superscript i refers to the occurrence of the i^{th} hypothesized fault. Roughly speaking, a procedure of fault detection and diagnosis is to find the most probable model from the set of hypothesized models $\mathcal{M} \triangleq \{M_0, \cdots, M_m\}$ and Bayes' rule is applied to quantify probabilistic confidence levels as functions of the measurements for all the hypothesized models.

[1] Similar convex relaxation methods have been applied to the optimal experiment design for system identification [144, 145, 168].

10.2 Statistical Distance Measures for Hypothesis Testing

10.2.1 Distance Measure Between Gaussian Hypotheses

A measure of distance between two Gaussian hypotheses can be used to characterize the performance limitation of the decision process based on the Bayesian approach.

Symmetrized Relative Entropy as a Distance Measure A common measure for the statistical distance between two probability distributions is the relative entropy, which is also called the Kullback-Leibler distance (or divergence).

Definition 10.1 (*See also [61] for details*). For two probability density functions f and g, relative entropy is defined by

$$d_{\mathrm{KL}}(f||g) \triangleq \int_{\mathbb{R}^n} f(x) \ln \frac{f(x)}{g(x)} dx. \tag{10.3}$$

For two probability mass functions f and g, relative entropy is defined by

$$d_{\mathrm{KL}}(f||g) \triangleq \sum_{x \in \mathcal{X}} f(x) \ln \frac{f(x)}{g(x)}, \tag{10.4}$$

where the support set $\mathcal{X}$ is assumed to be countable. For a probability mass function f and a probability density function g, relative entropy is defined by (10.4). For a probability density function f and a probability mass function g, relative entropy is defined by (10.3) and its value is indeed infinity, since the integrand is finite only if the support of f is contained in the support of g.[2]

This measure of distance between two probability distribution is not symmetric, i.e., $d_{\mathrm{KL}}(f||g) \neq d_{\mathrm{KL}}(g||f)$, in general. For example, consider two different Gaussian distribution functions $f : \mathbb{R}^n \to [0, 1]$ and $g : \mathbb{R}^n \to [0, 1]$. In particular, $f = \mathcal{N}(\mu_1, \Sigma_1)$ and $g = \mathcal{N}(\mu_2, \Sigma_2)$. Then the KL distances are

$$d_{\mathrm{KL}}(f||g) = \tfrac{1}{2} \left(\ln \det \Sigma_2 - \ln \det \Sigma_1 + \mathrm{Tr}\, \Sigma_2^{-1} \Sigma_1 \ + (\mu_1 - \mu_2)^{\mathrm{T}} \Sigma_2^{-1} (\mu_1 - \mu_2) - n \right) \tag{10.5}$$

and

$$d_{\mathrm{KL}}(g||f) = \tfrac{1}{2} \left(\ln \det \Sigma_1 - \ln \det \Sigma_2 + \mathrm{Tr}\, \Sigma_1^{-1} \Sigma_2 \ + (\mu_1 - \mu_2)^{\mathrm{T}} \Sigma_1^{-1} (\mu_1 - \mu_2) - n \right). \tag{10.6}$$

They are the same if $\Sigma_1 = \Sigma_2$, but not the same in general.

[2] Definitions for relative entropy of two probability distributions of different types of supports, i.e., continuous and discrete support sets, were not studied in classical information theory. However, relative entropy is defined with a measure function $f(\cdot)$ for both of the continuous support case (10.3) and the discrete support case (10.4) so that it is natural to follow the support of f for definition, provided that the other probability distribution $g(\cdot)$ is well-defined over that support.

For a symmetric distance measure, consider one of the followings:

$$\begin{aligned}\rho_{\mathrm{KL}}^{\min}(f,g) &\triangleq \min\{d_{\mathrm{KL}}(f||g), d_{\mathrm{KL}}(g||f)\},\\ \rho_{\mathrm{KL}}^{\max}(f,g) &\triangleq \max\{d_{\mathrm{KL}}(f||g), d_{\mathrm{KL}}(g||f)\},\\ \rho_{\mathrm{KL}}^{\mathrm{ave}}(f,g) &\triangleq \frac{1}{2}\left(d_{\mathrm{KL}}(f||g) + d_{\mathrm{KL}}(g||f)\right).\end{aligned} \tag{10.7}$$

We now show the relations between the KL-divergence and the likelihood ratio test. For notational convenience, rewrite the likelihood function corresponding to the k^{th} model of the hypothesis H_k as

$$\mathcal{L}_k(z) = p_{\mathbf{z}|\mathbf{H}}(z|H_k)$$

where z refers to the concatenation of all observables. Define the ratio of two likelihood functions by

$$R_{i,j}(z) \triangleq \frac{\mathcal{L}_i(z)}{\mathcal{L}_j(z)}$$

and define the (probability) measure-dependent quantity

$$T_{i,j}(\mu') \triangleq \int_{\mathcal{Z}} \ln(\max\{R_{i,j}(z), R_{j,i}(z)\})\mathrm{d}\mu(z) \tag{10.8}$$

where μ' denotes the first-order derivative of the measure μ, provided it is differentiable. Note that since $\max\{R_{i,j}(z), R_{j,i}(z)\} \geq 1$ for all $z \in \mathcal{Z}$, $T_{i,j}(\mu') \geq 0$ for any measure satisfying $\mu'(z) \geq 0$ for all $z \in \mathcal{Z}$. It is straightforward to show that

$$\begin{aligned}T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_i)) &= d(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)),\\ T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_j)) &= d(p_{\mathbf{z}|\mathbf{H}}(z|H_j), p_{\mathbf{z}|\mathbf{H}}(z|H_i)),\end{aligned} \tag{10.9}$$

which also implies

$$\rho_{\mathrm{KL}}^{\min}(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)) = \min\left\{T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_i)), T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_j))\right\}. \tag{10.10}$$

Therefore, the symmetric measure using the KL divergence $\rho_{\mathrm{KL}}^{\min}$ can be interpreted as the expectation of the logarithm of the likelihood ratio with respect to the likelihood function corresponding to the minimum value. Similarly, we have the following relations:

$$\rho_{\mathrm{KL}}^{\max}(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)) = \max\left\{T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_i)), T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_j))\right\} \tag{10.11}$$

and

$$\rho_{\mathrm{KL}}^{\mathrm{ave}}(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)) = \frac{1}{2}\left(T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_i)) + T_{i,j}(p_{\mathbf{z}|\mathbf{H}}(z|H_j))\right). \tag{10.12}$$

Bhattacharyya Bound as an Upper Bound on the Bayes Risk A Bayesian hypothesis test has a finite probability of selecting the incorrect model, which is called the *Bayes risk*, given by

$$p_{\mathrm{err}} \triangleq \sum_i \sum_{j \neq i} \int_{\mathcal{R}_j} p_{\mathbf{z}|\mathbf{H}}(z|H_i) p(H_i) dz \tag{10.13}$$

where $\mathcal{R}_j \triangleq \{z \in \mathcal{Z} : p_{\mathbf{H}|\mathbf{z}}(H_j|z) > p_{\mathbf{H}|\mathbf{z}}(H_k|z),\, \forall k \neq j\}$ is the region in which the hypothesis H_j is the most probable. Computing the exact p_{err} requires high computational demand that corresponds to sums of many multi-dimensional integrals. Computation can be relaxed by using the Bhattacharyya bound that provides an upper bound on the Bayes risk and is defined as

$$\bar{e}_{\mathrm{Bhat}}(H_i, H_j) \triangleq p_{ij} \int \sqrt{p_{\mathbf{z}|\mathbf{H}}(z|H_i) p_{\mathbf{z}|\mathbf{H}}(z|H_j)} dz \tag{10.14}$$

where $p_{ij} \triangleq \sqrt{p(H_i)p(H_j)}$ are constants related to the prior distributions of the hypotheses H_i and H_j.

Define a distance measure between two probability distributions $f = \mathcal{N}(\mu_1, \Sigma_1)$ and $g = \mathcal{N}(\mu_2, \Sigma_2)$ by

$$\rho_{\mathrm{Bhat}}(f, g) \triangleq \frac{1}{2}\left(\ln \frac{\det(\Sigma_1 + \Sigma_2)}{\det \Sigma_1 \det \Sigma_2} + n \ln \frac{1}{2} \; + \frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}} (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)\right). \tag{10.15}$$

Consider the set of linear hypothesized models $\mathcal{M}$ given in (10.2) that are associated with the set of hypotheses $\mathcal{H}$. Then, the Bhattacharyya bound (10.14) has the following relation with the distance measure given in (10.15):

$$\ln \bar{e}_{\mathrm{Bhat}}(H_i, H_j) \leq \ln p_{ij} - \rho_{\mathrm{Bhat}}(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j))).$$

Geometric Interpretations of Relative Entropy for Gaussian Distributions From a geometric point of view, $d_{\mathrm{KL}}(f||g)$ and $d_{\mathrm{KL}}(f||g)$ for Gaussian distributions f and g can be interpreted as measures of geometric disagreement between two ellipsoids associated with each distribution. To see this, the first two terms in (10.5) and (10.6)

$$\begin{aligned} \ln \det \Sigma_1^{-1} - \ln \det \Sigma_2^{-1} &= \ln \frac{\det \Sigma_1^{-1}}{\det \Sigma_2^{-1}} \\ \ln \det \Sigma_2^{-1} - \ln \det \Sigma_1^{-1} &= \ln \frac{\det \Sigma_2^{-1}}{\det \Sigma_1^{-1}} \end{aligned} \tag{10.16}$$

are the volume ratio of the two ellipsoids $\mathcal{E}_1(\Sigma_1, 1) \triangleq \{x \in \mathbb{R}^n : x^{\mathrm{T}}\Sigma_1^{-1}x \leq 1\}$ and $\mathcal{E}_2(\Sigma_2, 1) \triangleq \{x \in \mathbb{R}^n : x^{\mathrm{T}}\Sigma_2^{-1}x \leq 1\}$. In addition, the quantities $\operatorname{Tr}\Sigma_2^{-1}\Sigma_1 - n$ and $\operatorname{Tr}\Sigma_1^{-1}\Sigma_2 - n$ are the projections of differences $(\Sigma_1 - \Sigma_2)$ and $(\Sigma_2 - \Sigma_1)$ onto Σ_2^{-1} and Σ_1^{-1}, respectively:

$$\begin{aligned} \operatorname{Tr}\Sigma_2^{-1}\Sigma_1 - n &= \operatorname{Tr}\Sigma_2^{-1}(\Sigma_1 - \Sigma_2), \\ \operatorname{Tr}\Sigma_1^{-1}\Sigma_2 - n &= \operatorname{Tr}\Sigma_1^{-1}(\Sigma_2 - \Sigma_1). \end{aligned} \tag{10.17}$$

Furthermore, an optimum minimizing the sum of those values

$$(\Sigma_1^*, \Sigma_2^*) \triangleq \arg \min_{\Sigma_1 \succ 0, \Sigma_2 \succ 0} \operatorname{Tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_1\Sigma_2^{-1}\right) \tag{10.18}$$

satisfies the relation $\Sigma_1^* = \Sigma_2^*$, which follows from Lem. 10.1. Therefore, $\operatorname{Tr}(\Sigma_1^{-1}\Sigma_2 + \Sigma_1\Sigma_2^{-1})$ can be interpreted as a degree of disagreement between Σ_1 and Σ_2, which is measured in the Hilbert subspace of $\mathbb{S}_{++}^n$ equipped with the inner product $\langle A, B\rangle \triangleq \operatorname{Tr}(AB)$.

Lemma 10.1. For $X \in \mathbb{S}_{++}^n$, $\operatorname{Tr}(X + X^{-1}) \geq 2n$. Furthermore, equality holds if and only if $X = \mathrm{I}$.

The last terms

$$(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma_2^{-1}(\mu_1 - \mu_2) \text{ and } (\mu_1 - \mu_2)^{\mathrm{T}}\Sigma_1^{-1}(\mu_1 - \mu_2) \tag{10.19}$$

are the Euclidean distances between the points μ_1 and μ_2 with respect to Σ_2^{-1} and Σ_1^{-1}, respectively, i.e., they are weighted matrix 2-norms denoted by $\|\mu_1 - \mu_2\|_{\Sigma_2^{-1}}^2$ and $\|\mu_1 - \mu_2\|_{\Sigma_1^{-1}}^2$, respectively.

For a symmetric distance measure, define

$$\rho_{\mathrm{KL}}(f, g) \triangleq \min\{d_{\mathrm{KL}}(f||g), d_{\mathrm{KL}}(g||f)\} \tag{10.20}$$

such that $\rho_{\mathrm{KL}}(f, g) = \rho_{\mathrm{KL}}(g, f)$. For pedagogical purposes to see how this measure of distance between two probability distribution can be used for estimation or hypothesis testing problems, consider two Gaussian distributions f and g. For a given Gaussian distribution $f = \mathcal{N}(\mu_1, \Sigma_1)$, solve the minimization problem

$$\min_{g \in \mathcal{G}} d_{\mathrm{KL}}(f||g) \tag{10.21}$$

where $\mathcal{G} \triangleq \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathcal{M},\, \sigma \in \mathcal{S}\}$ with the convex compact sets $\mathcal{M} \subset \mathbb{R}^n$ and $\mathcal{S} \in \mathbb{S}_{++}^n$. More explicitly, the minimization problem is written as

$$\min_{(\mu, \Sigma) \in \mathcal{M} \times \mathcal{S}} -\ln\det\Sigma^{-1} + \operatorname{Tr}\Sigma^{-1}\Sigma_1 + (\mu_1 - \mu)^{\mathrm{T}}\Sigma^{-1}(\mu_1 - \mu). \tag{10.22}$$

Assume further that $\mathcal{S}^{-1} \triangleq \{S : S = \Sigma^{-1}, \Sigma \in \mathcal{S}\} \subset \mathbb{S}^n_{++}$ is also a convex compact set. Then an equivalent optimization is

$$\begin{aligned} \min_{t\geq 0,(\mu,X)\in\mathcal{M}\times\mathcal{S}^{-1}} & -\ln\det X + \operatorname{Tr} X\Sigma_1 + t \\ \text{s.t.} \quad & (\mu_1-\mu)^{\mathrm{T}} X(\mu_1-\mu) \leq t \end{aligned} \tag{10.23}$$

which is also equivalent to

$$\begin{aligned} \min_{t\geq 0,(\mu,X)\in\mathcal{M}\times\mathcal{S}^{-1}} & -\ln\det X + \operatorname{Tr} X\Sigma_1 + t \\ \text{s.t.} \quad & \begin{bmatrix} t & (\mu-\mu_1)^{\mathrm{T}} X \\ X(\mu-\mu_1) & X \end{bmatrix} \succeq 0 \end{aligned} \tag{10.24}$$

If there is no constraint on the choice of mean μ, i.e., $\mathcal{M} \equiv \mathbb{R}^n$, then the minimization can be rewritten as the SDP

$$\begin{aligned} \min_{t\geq 0,\zeta,X\in\mathcal{S}^{-1}} & -\ln\det X + \operatorname{Tr} X\Sigma_1 + t \\ \text{s.t.} \quad & \begin{bmatrix} t & \zeta^{\mathrm{T}} - \mu_1^{\mathrm{T}} X \\ \zeta - X\mu_1 & X \end{bmatrix} \succeq 0, \end{aligned} \tag{10.25}$$

provided $\mathcal{S}^{-1}$ can be represented by a positive-semidefinite cone.

Geometric Distance Measure Between Two Gaussian Distributions Consider two ellipsoids $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma)$ and $\mathcal{E}_2(\mu_2, \Sigma_2, \gamma)$ where the positive constant γ corresponds to the scaling factor of the volume of the corresponding ellipsoid. For these confidence ellipsoids for two Gaussian random variables $x \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $y \sim \mathcal{N}(\mu_2, \Sigma_2)$, the following conditions are equivalent:

i. The intersection of two ellipsoids $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma) \triangleq \{x \in \mathbb{R}^n : (x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1) \leq \gamma\}$ and $\mathcal{E}_2(\mu_2, \Sigma_2, \gamma) \triangleq \{x \in \mathbb{R}^n : (x-\mu_2)^{\mathrm{T}}\Sigma_2^{-1}(x-\mu_2) \leq \gamma\}$ is empty;

ii. There exists a separating hyperplane between two ellipsoids $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma)$ and $\mathcal{E}_2(\mu_2, \Sigma_2, \gamma)$;

iii. The optimal value of the semidefinite program (SDP)

$$\begin{aligned} \min_{t,x,y} \; & t \\ \text{s.t.} \; & \begin{bmatrix} \gamma & (x-\mu_i)^{\mathrm{T}} \\ (x-\mu_i) & \Sigma_i \end{bmatrix} \succeq 0, \quad i = 1,2, \\ & \begin{bmatrix} t & (x-y)^{\mathrm{T}} \\ (x-y) & \mathrm{I} \end{bmatrix} \succeq 0 \end{aligned} \tag{10.26}$$

is strictly positive.

To quantify the statistical distance between two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$, compute the largest value of $\gamma > 0$ such that $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma) \cap \mathcal{E}_2(\mu_2, \Sigma_2, \gamma) = \emptyset$, which can be formulated as the two-stage SDP

$$\begin{aligned} &\max_{\gamma} \ \gamma \\ &\text{s.t.} \quad 0 < \min_{t,x,y} \quad t \\ &\qquad\qquad \text{s.t.} \quad \begin{bmatrix} \gamma & (x-\mu_i)^{\mathrm{T}} \\ (x-\mu_i) & \Sigma_i \end{bmatrix} \succeq 0, \quad i = 1, 2, \\ &\qquad\qquad\qquad\quad \begin{bmatrix} t & (x-y)^{\mathrm{T}} \\ (x-y) & \mathrm{I} \end{bmatrix} \succeq 0. \end{aligned} \tag{10.27}$$

The optima in this optimization are not actually achieved since the sets $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma)$ and $\mathcal{E}_2(\mu_2, \Sigma_2, \gamma)$ are compact for all $0 \leq \gamma < \infty$, provided $\Sigma_i \succ 0$ for $i = 1, 2$. This problem can be arbitrarily accurately solved by using a bisection method for which each step reduces to a recognition problem to check if $t^* > 0$ and a stopping criterion can be used to impose an allowed error of approximation. For an alternative way to compute the largest value of $\gamma > 0$ such that $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma) \cap \mathcal{E}_2(\mu_2, \Sigma_2, \gamma) = \emptyset$, consider the SDP

$$\begin{aligned} &\min_{\gamma, x} \ \gamma \\ &\text{s.t.} \quad \begin{bmatrix} \gamma & (x-\mu_i)^{\mathrm{T}} \\ (x-\mu_i) & \Sigma_i \end{bmatrix} \succeq 0, \quad i = 1, 2. \end{aligned} \tag{10.28}$$

Then for any $\gamma < \gamma^*$ where γ^* is the optimal solution for (10.28), $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma) \cap \mathcal{E}_2(\mu_2, \Sigma_2, \gamma) = \emptyset$. Furthermore, for $\Sigma_1, \Sigma_2 \in \mathbb{S}^n_{++}$, $\mathcal{E}_1(\mu_1, \Sigma_1, \gamma) \cap \mathcal{E}_2(\mu_2, \Sigma_2, \gamma)$ is a unique singleton. Define the distance measure for two Gaussian distributions f and g

$$\rho_{\mathrm{geo}}(f, g) \triangleq \gamma^* \tag{10.29}$$

where $f = \mathcal{N}(\mu_1, \Sigma_1)$, $g = \mathcal{N}(\mu_2, \Sigma_2)$, and γ^* is the optimal value of the SDP (10.28).

This geometric measure has several interesting relations with some existing distance metrics.

Remark 10.1. The measure of distance $\rho_{\mathrm{geo}}(f, g)$ for two Gaussian distributions f and g is related to the Minkowski functional of a convex set $\mathcal{K} \in \mathbb{X}$ ($0 \in \operatorname{int} \mathcal{K}$) defined as

$$\rho_{\mathrm{Mk}}(x, \mathcal{K}) \triangleq \inf\{r : \frac{x}{r} \in \mathcal{K},\, r > 0\}$$

which is a kind of measure of distance from the origin to the point x in a normed linear vector space $\mathbb{X}$ measured with respect to $\mathcal{K}$—it is the scaling factor by which $\mathcal{K}$ needs to be expanded so as to include x

(see [161] for details). The explicit relation is that, for $f = \mathcal{N}(\mu_1, \Sigma_1)$ and $g = \mathcal{N}(0, \Sigma_2)$,

$$\lim_{\|\Sigma_1\| \to 0} \rho_{\text{geo}}(f, g) = \rho_{\text{Mk}}(\mu_1, \mathcal{E}_2(0, \Sigma_2, 1)),$$

where $\|\cdot\|$ can be any arbitrary matrix norm.

Remark 10.2. Another related distance measure is $n/2$ times the so-called Mahalanobis distance between two points μ_1 and μ_2 that is defined by

$$\rho_{\text{Mh}}(\mu_1, \mu_2 | \Sigma) \triangleq \frac{n}{2}(\mu_1 - \mu_2)^{\text{T}} \Sigma^{-1} (\mu_1 - \mu_2).$$

This measure of distance satisfies the relation $\rho_{\text{Mh}}(\mu_1, \mu_2|\Sigma) = \rho_{\text{Mk}}(\mu_1 - \mu_2, \mathcal{E}(0, \Sigma, 2/n))$.

Remark 10.3. The measure of distance defined by (10.28) and (10.29) can be unbounded for covariances with infinite condition number. For example, consider $\mu_1 \neq \mu_2$, $\Sigma_1 = [1, 0; 0, \epsilon_1]$, and $\Sigma_2 = [1, 0; 0, \epsilon_2]$ where $0 < \epsilon_1, \epsilon_2 \ll 1$. Then $\lim_{\max\{\epsilon_i\} \to \infty} \rho_{\text{geo}}(f, g) = \infty$, where $f = \mathcal{N}(\mu_1, \Sigma_1)$, $g = \mathcal{N}(\mu_2, \Sigma_2)$.

Remark 10.4. $\rho_{\text{geo}}(f, g_1) < \infty$ and $\rho_{\text{geo}}(f, g_2) < \infty$ does not imply $\rho_{\text{geo}}(g_1, g_2) < \infty$. For example, consider $f = \mathcal{N}([0; 0], \text{I})$, $g_1 = \mathcal{N}([0; 1], [1, 0; 0, \epsilon_1])$, and $g_2 = \mathcal{N}([0; -1], [1, 0; 0, \epsilon_2])$. Then $\lim_{\epsilon_i \to \infty} \rho_{\text{geo}}(f, g_i) = 1$ for $i = 1, 2$, but $\lim_{\max\{\epsilon_i\} \to \infty} \rho_{\text{geo}}(g_1, g_2) = \infty$.

In spite of such improper cases, we have experienced some success in well-posed problems in which the shape of the probability distributions to be compared are neither drastically ill-conditioned nor have very large distance.

Statistical Robust Fault Detectability The statistical distance of a trajectory induced by a fault from the no-fault trajectory can be measured by ρ_{KL} or ρ_{geo}.

Lemma 10.2. Consider the probability mass function $f(x) = 1$ for $x = \mu$ and $f(x) = 0$ otherwise. Then for any probability distribution $g(x)$,

$$\rho_{\text{KL}}(f, g) = d_{\text{KL}}(f||g) = -\ln g(\mu).$$

Proof. From the definition of the KL divergence in Def. 10.1, $d_{\text{KL}}(f||g) = -\ln g(\mu)$ and $d_{\text{KL}}(g||f) = \infty$. From the definition of $\rho_{\text{KL}}(f, g)$ in (10.20), $\rho_{\text{KL}}(f, g) = d_{\text{KL}}(f||g)$. QED

Lemma 10.3. Consider the probability mass function $f(x) = 1$ for $x = \mu$ and $f(x) = 0$ otherwise and the Gaussian distribution $g = \mathcal{N}(\nu, \Sigma)$. Then ρ_{KL} and ρ_{geo} satisfy the relation

$$\rho_{\text{KL}}(f, g) = \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln \det \Sigma^{-1} + \frac{1}{2} \rho_{\text{geo}}(f, g)$$

where $\rho_{\text{geo}}(f,g) = (\nu-\mu)^{\mathrm{T}}\Sigma^{-1}(\nu-\mu)$.

Proof. From Lem. 10.2, $\rho_{\text{KL}}(f,g) = -\ln g(\mu) = \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma^{-1} + \frac{1}{2}(\nu-\mu)^{\mathrm{T}}\Sigma^{-1}(\nu-\mu)$ for $g = \mathcal{N}(\nu,\Sigma)$. QED

Remark 10.5. Note that the difference between ρ_{KL} and $\rho_{\text{geo}}/2$, i.e., $\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det\Sigma^{-1} = \frac{1}{2}\ln\frac{(2\pi)^n}{\det\Sigma^{-1}}$, is the logarithm of the volume ratio of the n-dimensional sphere to the ellipsoid $\mathcal{E}(0,\Sigma,1) \triangleq \{x \in \mathbb{R}^n : x^{\mathrm{T}}\Sigma^{-1}x \leq 1\}$.

10.2.2 Gaussian m-array Hypothesis Testing for FDD

To formulate an optimization problem for Bayesian hypothesis testing, there are three elementary components that are required to be predetermined for multiple hypothesis tests (or m-array hypothesis tests) [166, 220]:

- m-array hypotheses with associated priors: Define m hypotheses. Without loss of generality, include the null hypothesis H_0 that corresponds to the normal operation, i.e., non-faulty model, for FDD. The total number of hypotheses to be tested are $m+1$. The corresponding a priori probability distributions are given by $P_i \triangleq \Pr[H = H_i]$.[3]
- Penalties for wrong decisions: Assign the penalty $C_{ij} \geq 0$ that corresponds to the cost to pay when the decision is $\hat{H} = H_i$, but the truth is $H = H_j$.
- Likelihood functions: Specify the closed-form of the propagation of hypothesis to the observables $p_{\mathbf{z}|\mathbf{H}}(z|H_i)$[4] for each hypothesis H_i.

For FDD using a set of multiple models $\{M_0, \cdots, M_m\}$, each hypothesis is assigned to each associated model, i.e., $H = H_i \Leftrightarrow M = M_i$. For abuse of notation, $M = M_i$ is also used to refer to the associated hypothesis $H = H_i$.

The resulting optimization has the cost

$$\begin{aligned} J(H_i, z) &= \sum_{j=0}^{m} C_{ij} \Pr[H = H_j | \mathbf{z} = z] \\ &= c \sum_{j=0}^{m} C_{ij} p_{\mathbf{z}|\mathbf{H}}(z|H_j) P_j \end{aligned} \tag{10.30}$$

where $c \triangleq \sum_{j=0}^{m} p_{\mathbf{z}|\mathbf{H}}(z|H_j)P_j$ is the marginal probability that is independent of H_i and depends only on the (observable) data z. For a fixed realization z of the random variable $\mathbf{z}$, an optimal decision for hypothesis

[3] The choice of prior probability distributions might be essentially subjective and the *best* P_i is, in general, hard to determine in an objective way.

[4] The observable vector z is assumed to consist of the variables that can be used for hypothesis testing, e.g., the inputs, measurements, and controlled outputs.

selection is

$$\hat{H}(z) = \arg \min_{H_i : i=0,\cdots,m} J(H_i, z), \tag{10.31}$$

which is a finite-state optimization for which a set of hypotheses are assessed and compared with each other. This hypothesis testing is called *Series of games for Bayesian hypotheses* (SGBH). The required number of comparisons or tests increases as $O(m(m+1)/2)$ where m is the number of hypotheses to be assessed.

In dynamical systems, such Bayesian hypothesis testing is more likely comparing the predictions of the observables for multiple competing fault models (e.g., models given in (10.2)). For simplicity, assume that $z = y$, i.e., the measurement outputs are the only observables for hypothesis testing. The posterior distribution of the predicted output at time t with the system model M_k is

$$\begin{aligned}\Pr[M = M_k, \eta_{0:t-1}|y_t] &= \Pr[y_t|\eta_{0:t-1}, M = M_k] \Pr[M_k] \\ &= \Pr[y_t|\eta_{t-1}, M = M_k] \Pr[M_k]\end{aligned}$$

where $\eta \triangleq (x, u)$ and the last equality is due to the Markovian property induced by whiteness of process noise w_t and measurement noise v_t. Suppose that each hypothesized fault scenario has the system dynamics given by (10.2). Then the closed-forms of the likelihood functions $\Pr[y_t|\eta_{t-1}, M = M_k]$ can be computed, provided the previous system information compressed into η_{t-1} (or more precisely, its distribution $p(\eta_{t-1})$) can be accessed to every hypothesized model of a fault. In addition, we assume reasonable accuracy of a state estimator such as Kalman filter or moving-horizon estimator (MHE).

To improve reliability of a Bayesian hypothesis testing, we can use a finite-time monitoring in which a finite sequence of the measurements is used to compute the posteriori distribution or more precisely to compute the likelihood function. Consider the monitoring window of length ℓ_m for which the sequence of the measurements $\{y_t, y_{t-1}, \cdots, y_{t-\ell_m+1}\}$ is monitored. The posterior distribution of the predicted measurements in this monitoring window with the system model M_k is

$$\begin{aligned}&\Pr[M = M_k, x_{0:t-\ell_m}, u_{0:t-1}|y_{t-\ell_m+1:t}] \\ &\quad = \Pr[y_{t-\ell_m+1:t}|x_{0:t-\ell_m}, u_{0:t-1} M = M_k] \Pr[M_k] \\ &\quad = \Pr[y_{t-\ell_m+1:t}|x_{t-\ell_m}, u_{t-\ell_m:t-1}, M = M_k] \Pr[M_k].\end{aligned}$$

Assume that $\Pr[M_k] = 1/(m+1)$ for all $k = 0, \ldots, m$. Formally, the likelihood function for the k^{th} model M_k is defined as

$$\mathcal{L}_k(y) = \Pr[y_{t-\ell_m+1:t} = y|x_{t-\ell_m}, u_{t-\ell_m:t-1}, M = M_k] \tag{10.32}$$

where $y \in \mathbb{R}^{n_y \times \ell_m}$ refers to the measurements during the time interval corresponding to the monitoring window. Therefore,

$$k^\star(y) := \arg \max_{k=0,\cdots,m} \mathcal{L}_k(y)$$

where the argument y is explicitly represented to emphasize its dependence on the measurements that are indeed realizations of a random process in a finite interval.

Remark 10.6. When the monitoring window is moved forward, the previous (normalized) likelihood functions may be used as the prior distributions. In this case, there can be a longer delay in fault detection and diagnosis, but the probability of false alarm can decrease.

Remark 10.7. Consider the assumption that the initial condition at the starting point of the monitoring window and the applied control inputs are the same for all the system modes. Then the likelihood function can be rewritten as $\mathcal{L}_k(y; x_{t-\ell_m}, u_{t-\ell_m:t-1})$ and the corresponding optimal system mode rewritten as $k^\star(y; x_{t-\ell_m}, u_{t-\ell_m:t-1})$.

10.3 Optimal Input Design for FDD: Approximation Methods

It can occur that two or more hypotheses are almost equally probable so are not distinguishable from the current observable data because their predicted (hypothesized) distributions are quantitatively very close. To resolve such difficult decision-making situations, consider optimal input design problems for which the control input maximizing detectability of faults is constructed while retaining desirable system behaviors or minimizing degradation of system performance incurred by *active* FDD. Many of the existing fault diagnosis methods are *passive* in the sense that those diagnostic procedures are based on the observed data for given inputs. Input design for fault diagnosis presented in this chapter is an active approach to determine the true fault. For two different models corresponding to two different fault scenarios, the sensitivity of the observables' statistics to input changes can be substantially different from each other. Consider varying the inputs within an allowable range of operation so that the resultant statistics of observables predicted by different fault scenarios are notably different and the more probable fault scenario in a likelihood ratio hypothesis test is diagnosed as an estimated fault. A quantified measure of distinguishability between two models of fault scenarios is

$$\delta_{ij}(z) \triangleq \rho(p_{\mathbf{H}|\mathbf{z}}(H_i|z), p_{\mathbf{H}|\mathbf{z}}(H_j|z)) = \delta_{ji}(z)$$

where $\rho(\cdot,\cdot)$ denotes a certain (symmetric) measure of distance between two probability distributions. We consider the previously defined two measures of statistical distance ρ_{KL} and ρ_{geo} to quantify distinguishability of faults. Suppose that the input constraint $u \in \mathcal{U}$ is defined over a convex compact set $\mathcal{U}$ such as a polytope or ellipsoid.

10.3.1 Method Based on the Measure ρ_{KL}

Consider the measure of statistical distance ρ_{KL} between two Gaussian distributions. Maximizing ρ_{KL} can be formulated as the optimization

$$\begin{aligned}
&\max_{u\in\mathcal{U}} \ \min\{\gamma_1,\gamma_2\} \\
&\text{s.t.} \ \frac{1}{2}\left(\ln\det\Sigma_2-\ln\det\Sigma_1+\operatorname{Tr}\Sigma_2^{-1}\Sigma_1 \ +(\mu_1-\mu_2)^{\mathrm{T}}\Sigma_2^{-1}(\mu_1-\mu_2)-n\right)\geq\gamma_1, \\
&\quad\ \frac{1}{2}\left(\ln\det\Sigma_1-\ln\det\Sigma_2+\operatorname{Tr}\Sigma_1^{-1}\Sigma_2 \ +(\mu_1-\mu_2)^{\mathrm{T}}\Sigma_1^{-1}(\mu_1-\mu_2)-n\right)\geq\gamma_2,
\end{aligned} \tag{10.33}$$

where the mean and covariance are convex functions of the control input u. However, neither optimization (10.33) is not convex, even for the case when the expectations μ_1 and μ_2 are linearly dependent on $u\in\mathcal{U}$, and the covariances Σ_1 and Σ_2 are independent of $u\in\mathcal{U}$.

10.3.2 Method Based on the Measure ρ_{geo}

Consider the measure of statistical distance ρ_{geo} between two Gaussian distributions. Maximizing ρ_{geo} can be formulated as the max-min problem

$$\begin{aligned}
&\max_{u\in\mathcal{U}} \min_{\mu,\gamma} \ \gamma \\
&\qquad \text{s.t.} \ \begin{bmatrix} \gamma & (\mu-\mu_1(u))^{\mathrm{T}} \\ (\mu-\mu_1(u)) & \Sigma_1(u) \end{bmatrix} \succeq 0, \\
&\qquad\qquad\ \ \begin{bmatrix} \gamma & (\mu-\mu_2(u))^{\mathrm{T}} \\ (\mu-\mu_2(u)) & \Sigma_2(u) \end{bmatrix} \succeq 0,
\end{aligned} \tag{10.34}$$

where μ_i and Σ_i are the mean and covariance corresponding to the i^{th} Gaussian distribution for $i=1,2$, which are functions of the control input u. This max-min problem can be rewritten as

$$\begin{aligned}
&\min_{u\in\mathcal{U}} \max_{\mu,\gamma} \ -\gamma \\
&\qquad \text{s.t.} \ \begin{bmatrix} \gamma & (\mu-\mu_i(u))^{\mathrm{T}} \\ (\mu-\mu_i(u)) & \Sigma_i(u) \end{bmatrix} \succeq 0, \ \imath=1,2.
\end{aligned} \tag{10.35}$$

Lemma 10.4. Suppose that $\mu_i:\mathcal{U}\to\mathbb{R}^n$ is an affine function and $\Sigma_i:\mathcal{U}\to\mathbb{S}^n_+$ is an affine or concave quadratic function for each $i=1,2$. Then, the optimization (10.35) is a convex-constrained concave program, i.e., the objective function is concave and the constraint set is convex.

Since the optimization (10.35) has a concave objective function and the constraint set is convex and compact, a global optimum is achieved on the boundary of the closed convex constraint set, i.e., $u^\star\in\partial\mathcal{U}$ where $\partial\mathcal{U}$ refers to the boundary of $\mathcal{U}$, or must be constant over $\mathcal{U}$. An iteration algorithm is described in

Algorithm 1. Furthermore, if the input constraint set $\mathcal{U}$ is a polytope then an global optimal solution $u^\star$ is in the set of vertices, denoted by $\mathcal{U}_{\rm v}$. This implies that for a polytope $\mathcal{U}$, we only need to compare a finite number of candidates for optimal inputs:

$$u^\star = \arg\max_{u\in\mathcal{U}_{\rm v}} \rho_{\rm geo}(\mathcal{N}(\mu_1(u),\Sigma_1(u)),\mathcal{N}(\mu_2(u),\Sigma_2(u))).$$

10.3.3 One-step Maximization Detectability or Distinguishability of Two Competing Faults Using $\rho_{\rm geo}$

Consider two competing fault models M_i $(i=1,2)$ in (10.2) and assume perfect information of the state variables[5] and that the output transition maps are dropped for simplicity (the extension to the general case is straightforward). Suppose that the current state of the true process has a known Gaussian distribution $\mathcal{N}(\bar{x}_t,\Sigma_{x_t})$, the dimensions of x^i are the same as the true state's, and $x_t = x_t^i$ for $i=1,2$. Due to linearity of the state transition map of models M_i, the one-step lookahead trajectories of the models M_i $(i=1,2)$ corresponding to two hypothesized faults are also Gaussian, provided an affine state feedback or open-loop control u_t. An optimal input design can be formulated as an optimization of finding $u\in\mathcal{U}$ maximizing $\rho_{\rm geo}(\mathcal{N}(\bar{x}_{t+1}^{f_1},\Sigma_{x_{t+1}^{f_1}}),\mathcal{N}(\bar{x}_{t+1}^{f_2},\Sigma_{x_{t+1}^{f_2}}))$. Consider an affine state feedback control

$$u_t = K_t x_t + \nu_t. \tag{10.37}$$

[5]By *perfect information*, we mean that the state vector is measurable and its distribution is known or an unbiased estimation can be obtained with the associated computable error covariance.

Algorithm 1 Iteration algorithm for solving (10.34) and (10.35).

Input: $\mu_i(\cdot),\Sigma_i(\cdot)$; $i=1,2$, $u^{(0)}$, and $\{\delta^{(j)}\}\subset\mathbb{R}_{++}$.
Onput: $\hat{u}^*,\hat{\gamma}^*$, and δ.

Step 0: Set $j=0$.
Step 1: For $u:=u^{(j)}$, solve the minimization part in (10.35) and assign its optimal value by $\gamma^{(j)}:=\gamma^*$.
Step 2: Set $u:=u^{(j)}+du$.
Step 3: Solve the minimization

$$\begin{aligned}\min_{u\in\mathcal{U},\mu,\gamma}\ & \gamma\\ \text{s.t. }& \begin{bmatrix}\gamma & (\mu-\mu_i(u))^{\rm T}\\ (\mu-\mu_i(u)) & \Sigma_i(u)\end{bmatrix}\succeq 0,\quad i=1,2.\\ & \gamma\geq(1+\delta^{(j)})\gamma^{(j)}.\end{aligned} \tag{10.36}$$

if (10.36) has a feasible optimal solution (γ^*,du^*,μ^*) **then**
Assign the optimum and optimal value by $u^{(j+1)}:=u^{(j)}+du^*$ and $\gamma^{(j+1)}:=\gamma^*$. Set $j:=j+1$ and go to Step 2.
else
Set $\hat{u}^*:=u^{(j)}$, $\hat{\gamma}^*:=\gamma^{(j)}$, and $\delta:=\delta^{(j)}$.
end if

$$\begin{aligned}
&\max_{K_t,\nu_t}\ \min_{\mu,\gamma}\ \gamma \\
&\quad \text{s.t.}\ \begin{bmatrix} \gamma & (\mu-(A_t^i+B_t^iK_t)\bar{x}_t+B_t^i\nu_t)^{\mathrm{T}} \\ (\mu-(A_t^i+B_t^iK_t)\bar{x}_t+B_t^i\nu_t) & (A_t^i+B_t^iK_t)\Sigma_{x_t}(A_t^i+B_t^iK_t)^{\mathrm{T}}+E_t^i\Sigma_w(E_t^i)^{\mathrm{T}} \end{bmatrix} \succeq 0, \\
&\qquad \begin{bmatrix} \Sigma_{u,t}^{\max} & K_t \\ K_t^{\mathrm{T}} & \Sigma_{x_t}^{-1} \end{bmatrix} \succeq 0,\ b_{u,t}-H_{u,t}K_t\bar{x}_t-H_{u,t}\nu_t \geq 0.
\end{aligned} \tag{10.40}$$

$$\begin{aligned}
&\max_{K_t,\nu_t}\ \min_{\mu,\gamma}\ \gamma \\
&\quad \text{s.t.}\ \begin{bmatrix} \gamma & (\mu-(A_t^i+B_t^iK_t)\bar{x}_t+B_t^i\nu_t)^{\mathrm{T}} \\ (\mu-(A_t^i+B_t^iK_t)\bar{x}_t+B_t^i\nu_t) & (A_t^i+B_t^iK_t)\Sigma_{x_t}(A_t^i+B_t^iK_t)^{\mathrm{T}}+E_t^i\Sigma_w(E_t^i)^{\mathrm{T}} \end{bmatrix} \succeq 0, \\
&\qquad \begin{bmatrix} \Sigma_{x,t}^{\max}-E_t^i\Sigma_w(E_t^i)^{\mathrm{T}} & (A_t^i+B_t^iK_t) \\ (A_t^i+B_t^iK_t)^{\mathrm{T}} & \Sigma_{x_t}^{-1} \end{bmatrix} \succeq 0,\ b_{x,t}-H_{x,t}(A_t^i+B_t^iK_t)\bar{x}_t-H_{x,t}B_t^i\nu_t \geq 0, \\
&\qquad \begin{bmatrix} \Sigma_{u,t}^{\max} & K_t \\ K_t^{\mathrm{T}} & \Sigma_{x_t}^{-1} \end{bmatrix} \succeq 0,\ b_{u,t}-H_{u,t}K_t\bar{x}_t-H_{u,t}\nu_t \geq 0.
\end{aligned} \tag{10.41}$$

Then, the mean and covariance of the one-step lookahead state of the model M_i are

$$\begin{aligned}
\bar{x}_{t+1}^i &= (A_t^i+B_t^iK_t)\bar{x}_t+B_t^i\nu_t \\
\Sigma_{x_{t+1}^i} &= (A_t^i+B_t^iK_t)\Sigma_{x_t}(A_t^i+B_t^iK_t)^{\mathrm{T}}+E_t^i\Sigma_w(E_t^i)^{\mathrm{T}}
\end{aligned} \tag{10.38}$$

for each $i=1,2$. For input constraints, consider $\mathcal{U}_t \triangleq \{u\in\mathbb{R}^{n_u} : H_{u,t}\mathbf{E}[u]\leq b_{u,t} \text{ and } \mathbf{Var}[u]\preceq\Sigma_{u,t}^{\max}\}$ where the subscript t denotes time dependence and $\Sigma_{u,t}^{\max}$ refers to an upper bound on the covariance of interest. For an affine state feedback control (10.37), $u_t\in\mathcal{U}_t$ if and only if K_t and ν_t satisfy the inequalities

$$\begin{aligned}
H_{u,t}K_t\bar{x}_t+H_{u,t}\nu_t &\leq b_{u,t} \\
\begin{bmatrix} \Sigma_{u,t}^{\max} & K_t \\ K_t^{\mathrm{T}} & \Sigma_{x_t}^{-1} \end{bmatrix} &\succeq 0
\end{aligned} \tag{10.39}$$

that are convex in (K_t,ν_t). The resulting optimization for an optimal input design with the input constraint (10.39) is given by (10.40).

In addition to input constraints, to avoid instability and performance degradation of the closed-loop system, the state constraints for x_{t+1}^i $(i=1,2)$ should also be considered. Since the predicted state trajectories are essentially stochastic, the corresponding state constraints are written in terms of chance constraints. Similar to input constraints, consider $\mathcal{X}_t \triangleq \{x\in\mathbb{R}^n : H_{x,t}\mathbf{E}[x]\leq b_{x,t} \text{ and } \mathbf{Var}[x]\preceq\Sigma_{x,t}^{\max}\}$. From (10.38),

$x_{t+1}^i \in \mathcal{X}_t$ if and only if

$$\begin{aligned} H_{x,t}(A_t^i + B_t^i K_t)\bar{x}_t + H_{x,t}B_t^i \nu_t &\le b_{x,t} \\ \begin{bmatrix} \Sigma_{x,t}^{\max} - E_t^i \Sigma_w (E_t^i)^{\mathrm{T}} & (A_t^i + B_t^i K_t) \\ (A_t^i + B_t^i K_t)^{\mathrm{T}} & \Sigma_{x_t}^{-1} \end{bmatrix} &\succeq 0. \end{aligned} \tag{10.42}$$

that are convex in (K_t, ν_t). The resulting optimization for an optimal input design with the input and state constraints (10.39) and (10.42) is given by (10.41).

Remark 10.8. The convex constraints (10.39) and (10.42) are the intersections of a polytope and a positive-semidefinite cone (basically, a mixed linear-conic constraint).

Algorithm 1 cannot be directly applied to solve the optimization problems (10.40) and (10.41) for the decision variable (K_t, ν_t). However, if a state feedback gain K_t is fixed then Algorithm 1 can be used to solve those optimizations for ν_t. Computing a state-feedback gain K_t in the optimizations (10.40) and (10.41) can be performed separately by solving the LMIs in (10.39) and (10.42), respectively. Note that solutions of such convex constraints are not generally unique. To resolve this non-uniqueness of feasible solutions K_t, the associated symmetric matrices defining the LMIs could be forced to be close to the extreme rays of positive-semidefinite cone, which is the set of rank-one symmetric matrices. This could be achieved by minimizing the rank of the resulting symmetric matrices and there are several ways to approximately perform rank-minimization using smooth approximation to the rank operator for positive-semidefinite matrices. For example, $-\log\det(X)$, which is convex in $X \in \mathbb{S}_{++}^n$ (or $X \in \mathbb{S}_{+}^n$),[6] or $\mathrm{Tr}(X)$ which is linear can be used.

10.3.4 A Separate Design Method of State Feedback using $\mathcal{H}_2$ Optimal Control

Consider the fault scenario models (10.2) for $i = 1, 2$, where the measurement noise v_t^i is ignored without loss of generality.[7] Since the $\mathcal{H}_2$ norm can be interpreted as the maximum output variance excited by white noise of the unit $\mathcal{L}_2$ norm, a natural way to compute a state feedback control gain K_t for a fixed $t \ge 0$ satisfying the constraints on the variance is $\mathcal{H}_2$-optimal control [73]. Utilizing the LMI conditions for state feedback $\mathcal{H}_2$ synthesis,[8] for each time instance $t \ge 0$, design of a state feedback gain K_t can be performed by the following procedure.

[6] If we consider $\mathbb{S}_{+}^n$ as the domain of the function $-\log\det(\cdot)$ then an extended real line is considered as the range of $-\log\det(\cdot)$, i.e., $-\log\det : \mathbb{S}_{+}^n \to (-\infty, \infty]$.

[7] The perturbation or variation due to v_t^i is not controllable by using a state feedback controller.

[8] See [73] for details on $\mathcal{H}_2$-synthesis problems.

S1. Solve the following SDP for Z, X, and W: Minimize η subject to

$$\begin{aligned}
\begin{bmatrix} A_t^i & B_t^i \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix} + \begin{bmatrix} X & Z^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A_t^{i\,\mathrm{T}} \\ B_t^{i\,\mathrm{T}} \end{bmatrix} + E_t^i E_t^{i\,\mathrm{T}} &\prec 0, \\
\begin{bmatrix} X & (C_t^i + D_t^i Z)^{\mathrm{T}} \\ (C_t^i + D_t^i Z) & W \end{bmatrix} &\succ 0, \\
\mathrm{Tr}(W) &< \eta, \\
X &\succ 0.
\end{aligned} \tag{10.43}$$

Define the optimal solutions as $Z_\star$, $X_\star$, and $W_\star$.

S2. Solve the following SDP for Z and W: Minimize η subject to

$$\begin{aligned}
\begin{bmatrix} A_t^i & B_t^i \end{bmatrix} \begin{bmatrix} X_\star \\ Z \end{bmatrix} + \begin{bmatrix} X_\star & Z^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} A_t^{i\,\mathrm{T}} \\ B_t^{i\,\mathrm{T}} \end{bmatrix} + E_t^i E_t^{i\,\mathrm{T}} &\prec 0, \\
\begin{bmatrix} X_\star & (C_t^i + D_t^i Z)^{\mathrm{T}} \\ (C_t^i + D_t^i Z) & W \end{bmatrix} &\succ 0, \\
\mathrm{Tr}(W) &< \eta, \\
\begin{bmatrix} \Sigma_{u,t}^{\max} & Z \\ Z^{\mathrm{T}} & X_\star \Sigma_{x_t}^{-1} X_\star \end{bmatrix} &\succeq 0, \\
\begin{bmatrix} \Sigma_{x,t}^{\max} - E_t^i \Sigma_w (E_t^i)^{\mathrm{T}} & (A_t^i X_\star + B_t^i Z) \\ (A_t^i X_\star + B_t^i Z)^{\mathrm{T}} & X_\star \Sigma_{x_t}^{-1} X_\star \end{bmatrix} &\succeq 0,
\end{aligned} \tag{10.44}$$

for $i = 1, 2$. Define the optimal solutions as $Z^\star$ and $W^\star$.

S3. Compute the state feedback control gain $K_t = Z^\star X_\star^{-1}$.

Once the state feedback gain K_t is fixed, the remaining problem is to determine the affine term ν_t solving the constrained optimization (10.40) or (10.41), which can be performed, again, by using Algorithm 1.

10.3.5 Multi-step Maximization of Discrimination between Multiple Competing Faults Using ρ_{geo}

Extension to when there are more than two probable fault scenario models is more complicated. One strategy is to compute an optimal input u^j for the j^{th} pair of faults and consider a convex combination $u_\lambda = \sum_{j=1}^{n_p} \lambda_j u^j$ where $n_p = p(p-1)/2$ refers to the number of pairs from $p(\geq 2)$ models of faults under consideration, and λ_js satisfy $\lambda_j \geq 0$ and $\sum_{j=1}^{n_p} \lambda_j = 1$. Due to convexity of the input constraint set $\mathcal{U}$, $u_\lambda \in \mathcal{U}$ for every λ_j satisfying $\lambda_j \geq 0$ and $\sum_{j=1}^{n_p} \lambda_j = 1$. If the objective function is replaced by a weighted

sum of distances γ_j^* of the j^{th} pair, then a new objective is to find an optimal coefficients λ_j's satisfying the convexity condition for u_λ. For notational convenience, consider a simpler form of optimization (10.35). The resulting optimization of an optimal input design for maximizing fault isolation (discrimination) is

$$\begin{aligned}
&\max_{\lambda_j;\, j=1,\ldots n_p} \sum_{j=1}^{n_p} \omega_j \gamma_j^* \\
&\gamma_j^* := \min_{\mu^j, \gamma_j} \gamma_j \\
&\qquad \text{s.t.} \begin{bmatrix} \gamma_j & (\mu^j - \mu_{j_1}(u_\lambda))^{\mathrm{T}} \\ (\mu^j - \mu_{j_1}(u_\lambda)) & \Sigma_{j_1}(u_\lambda) \end{bmatrix} \succeq 0, \\
&\qquad\qquad \begin{bmatrix} \gamma_j & (\mu^j - \mu_{j_2}(u_\lambda))^{\mathrm{T}} \\ (\mu^j - \mu_{j_2}(u_\lambda)) & \Sigma_{j_2}(u_\lambda) \end{bmatrix} \succeq 0, \\
&\qquad\qquad u_\lambda = \sum_{j=1}^{n_p} \lambda_j u^j; \sum_{j=1}^{n_p} \lambda_j = 1,\ \lambda_j \geq 0,
\end{aligned} \tag{10.45}$$

where ω_j is the user-defined weight on the j^{th} distance γ_j^* for each $j = 1, \ldots, n_p$, and the subscripts j_1 and j_2 refer to the quantities associated with the j^{th} pair of two Gaussian distributions. The optimization (10.45) can be solved by a similar iteration algorithm as Algorithm 1.

Remark 10.9. Extension to multi-step lookahead input design is not difficult and the receding horizon method can be applied. For the state feedback gains K_t in the multi-step prediction, the same aforementioned design method can be used with which K_t is independent of K_s for all $s > t$ such that the sequence of feedback gains K_t can be successively designed, since the open-loop input ν_t does not change the variance of the state and the output. However, note that time-varying state feedback can have high computational demand as the number of decision variables increases linearly in the length of the prediction horizon.[9]

10.4 Optimal Input Design for FDD: Convex Relaxation

There can be situations when two or more hypotheses are nearly equally probable and so are not distinguishable from the current observable data because their predicted (hypothesized) distributions are quantitatively very close. To resolve such difficult decision-making situations, we consider optimal input design problems for which the control input maximizing detectability of faults is constructed while retaining desirable system behaviors or minimizing degradation of system performance incurred by FDD. Most of the existing fault diagnosis methods are *passive* in the sense that those diagnostic procedures are based on the observed data for given inputs. The input design for fault diagnosis considered here is an active approach to facilitate statistical decision location of the true fault. For two different models corresponding to different faults,

[9] Note that computing K_t using the method presented in Section 10.3.4 requires to solve LMIs and can be done in polynomial-time. This implies that time-varying state feedback can be computed in real-time with a moderate size of the state variables and the prediction horizon.

sensitivity of the observables' statistics to input changes can be substantially different from each other. The inputs are changed within an allowable range of operation so that the resultant statistics of observables predicted from different fault scenarios are notably different and the more probable fault scenario obtained from hypothesis testing is diagnosed as a most likely estimated fault. The distinguishability between two models of fault scenarios can be quantified by

$$\begin{aligned} \delta_{ij}(z) &\triangleq \rho(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)) = \delta_{ji}(z) \text{ or} \\ \delta_{ij}(z) &\triangleq \rho(p_{\mathbf{H}|\mathbf{z}}(H_i|z), p_{\mathbf{H}|\mathbf{z}}(H_j|z)) = \delta_{ji}(z) \end{aligned} \tag{10.46}$$

where $\rho(\cdot,\cdot)$ refers to a certain (symmetric) measure (ρ_{KL} or ρ_{geo}) of distance between two probability distributions.

10.4.1 Constraints on Predicted Controlled Trajectories

The constraints $\mathcal{U}$ and $\mathcal{Y}$ are assumed to be convex and several classes of those constraints considered here are

$$\begin{aligned} \mathcal{U}_1(\gamma_1^u) &\triangleq \{u : \|u\|_1 \le \gamma_1^u\}, \\ \mathcal{U}_\infty(u_{\min}, u_{\max}) &\triangleq \{u : u_{\min} \le u \le u_{\max}\}, \\ \mathcal{U}_2(\gamma_{\mathrm{p}}^u) &\triangleq \{u : \|u\|_2 \le \gamma_{\mathrm{p}}^u\}, \\ \mathcal{U}_\delta(\delta u_{\min}, \delta u_{\max}) &\triangleq \{u : \delta u_{\min} \le u_{i+1} - u_i \le \delta u_{\max}\}, \end{aligned} \tag{10.47}$$

and

$$\begin{aligned} \mathcal{Y}_\infty(y_{\min}, y_{\max}) &\triangleq \{y : y_{\min} \le y \le y_{\max}\}, \\ \mathcal{Y}_2(\gamma_{\mathrm{p}}^y) &\triangleq \{y : \|y\|_2 \le \gamma_{\mathrm{p}}^y\}, \\ \mathcal{Y}_\delta(\delta y_{\min}, \delta y_{\max}) &\triangleq \{y : \delta y_{\min} \le y_{i+1} - y_i \le \delta y_{\max}\}. \end{aligned} \tag{10.48}$$

The constraint sets $\mathcal{U}$ and $\mathcal{Y}$ can also be intersections of the constraints in (10.47) and (10.48), respectively.

Remark 10.10. The constraint $\mathcal{U}_\delta$ is also known as the plant-friendly constraint in literature [58, 81, 219, 307], which is mostly studied in the literature of optimal input design for system identification. The methods in this chapter can handle the plant-friendly constraint in the time domain in its original form, which does not introduce any further conservatism.

Remark 10.11. The ℓ_1-norm constraint on the control input, defined by $\|u\|_1 = \sum_{i=1}^{N_h} |u_i|$, can be imposed to restrict or minimize the number of applied control actions. This can be considered as a convex relaxation of the nonconvex constraint $\|u\|_0 = \sum_{i=1}^{N_h} \iota_{\{0\}}(u_i)$ where ι refers to the indicator function defined by $\iota_{\mathcal{C}}(x) = 1$ if $x \in \mathcal{C}$ and $\iota_{\mathcal{C}}(x) = 0$ otherwise. This is called a *parsimonious* input constraint that corresponds to

minimizing unnecessary interruption to process operation.

Constrained Optimization for Maximizing Model Discrimination Consider two models of faults M_i and M_j corresponding to hypotheses H_i and H_j. The goal is to find optimum data z that solve the constrained optimization

$$\max_{z\in\mathcal{U}\times\mathcal{Y}} \rho(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j)) \tag{10.49}$$

where $\mathcal{U}$ and $\mathcal{Y}$ refer to the input and output constraints, respectively, given in Section 10.4.1.

Lemma 10.5. Suppose that the control input u is independent of the output y, i.e., u is an open-loop control. Then $\rho(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j))$ is convex in z, or equivalently in u, for any distance measure $\rho \in \{\rho_{\mathrm{KL}}^{\min}, \rho_{\mathrm{KL}}^{\mathrm{ave}}, \rho_{\mathrm{Bhat}}\}$.

10.4.2 Multi-step Lookahead Maximization of Distinguishability between Two Competing Hypothesized Models

Consider the LTI system models (10.2). Then the concatenated output trajectory within the time interval of prediction horizon $[\kappa+1, \kappa+m_h]$ is

$$\mathbf{y}^i_{\kappa,m_h} = \mathcal{F}^i_u \mathbf{u}_{\kappa,m_h} + \mathcal{F}^i_x x_\kappa + \mathcal{F}^i_w \mathbf{w}^i_{\kappa,m_h} + \mathcal{F}^i_v \mathbf{v}^i_{\kappa,m_h} \tag{10.50}$$

where $\mathbf{y}^i_{\kappa,m_h} = y^i_{\kappa+1:\kappa+m_h}$, $\mathbf{u}_{\kappa,m_h} = u_{\kappa:\kappa+m_h-1}$, $\mathbf{w}^i_{\kappa,m_h} = w^i_{\kappa:\kappa+m_h-1}$, $\mathbf{v}^i_{\kappa,m_h} = v^i_{\kappa+1:\kappa+m_h}$.

The corresponding mean and covariance of the predicted controlled output trajectory are given by

$$\mu^i_{\kappa,m_h} \triangleq \mathbf{E}[\mathbf{y}^i_{\kappa,m_h}] \tag{10.51}$$

and

$$\Sigma^i_{\kappa,m_h} \triangleq \mathbf{E}\big[(\mathbf{y}^i_{\kappa,m_h} - \mu^i_{\kappa,m_h})(\mathbf{y}^i_{\kappa,m_h} - \mu^i_{\kappa,m_h})^{\mathrm{T}}\big]\,. \tag{10.52}$$

The covariance is independent of the control input sequence $\mathbf{u}_{\kappa,m_h}$ provided that it is open-loop control. For notational convenience, rewrite

$$\mu^i_{\kappa,m_h} = \mathcal{G}^i_u \mathbf{u}_{\kappa,m_h} + g^i_\kappa. \tag{10.53}$$

For optimality criteria for the control input maximizing the statistical distance between two hypothesized system models, consider

$$J_{ij}(\mathbf{u}_{\kappa,m_h}; \rho) = (\tilde{\mu}^{ij}_{\kappa,m_h})^{\mathrm{T}} \mathcal{P}^{ij}_\rho (\tilde{\mu}^{ij}_{\kappa,m_h}) \tag{10.54}$$

where $\tilde{\mu}^{ij}_{\kappa,m_h} \triangleq \mu^{i}_{\kappa,m_h} - \mu^{j}_{\kappa,m_h}$ and each $\mathcal{P}^{ij}_{\rho}$ is defined by one of the positive-definite matrices:

$$\begin{cases} (\Sigma^{i}_{\kappa,m_h})^{-1} + (\Sigma^{j}_{\kappa,m_h})^{-1} & \text{for } \rho = \rho^{\text{ave}}_{\text{KL}}, \\ (\Sigma^{i}_{\kappa,m_h} + \Sigma^{j}_{\kappa,m_h})^{-1} & \text{for } \rho = \rho_{\text{Bhat}}. \end{cases} \tag{10.55}$$

Similarly, consider

$$\begin{aligned} J_{ij}(\mathbf{u}_{\kappa,m_h}; \rho^{\max}_{\text{KL}}) &= \max_{k\in\{i,j\}} (\tilde{\mu}^{ij}_{\kappa,m_h})^{\mathrm{T}} \Sigma^{k}_{\kappa,m_h} (\tilde{\mu}^{ij}_{\kappa,m_h}), \\ J_{ij}(\mathbf{u}_{\kappa,m_h}; \rho^{\min}_{\text{KL}}) &= \min_{k\in\{i,j\}} (\tilde{\mu}^{ij}_{\kappa,m_h})^{\mathrm{T}} \Sigma^{k}_{\kappa,m_h} (\tilde{\mu}^{ij}_{\kappa,m_h}), \end{aligned} \tag{10.56}$$

which can be rewritten as the quadratic form

$$J(\mathbf{u}_{\kappa,m_h}; \rho) \triangleq \mathbf{u}^{\mathrm{T}}_{\kappa,m_h} \mathcal{Q}_{\rho} \mathbf{u}_{\kappa,m_h} + q^{\mathrm{T}}_{\rho} \mathbf{u}_{\kappa,m_h} + q_{\rho,0} \tag{10.57}$$

where the super- and subscripts ij are dropped due to simplify notation. With this general form of optimality measure, the resultant constrained optimization can be represented as

$$\begin{aligned} \min_{\mathbf{u}_{\kappa,m_h}} \quad & -J(\mathbf{u}_{\kappa,m_h}; \rho) \\ \text{s.t.} \quad & \mathbf{u}_{\kappa,m_h} \in \mathcal{U}, \\ & \mathcal{G}^{i}_{u} \mathbf{u}_{\kappa,m_h} + g^{i}_{\kappa} \in \mathcal{Y}, \end{aligned} \tag{10.58}$$

where the output constraints are imposed on the expected output trajectory within the prediction horizon interval $[\kappa+1, \kappa+m_h]$. The symmetric matrix $\mathcal{Q}_{\rho}$ in $J(\mathbf{u}_{\kappa,m_h}; \rho)$ is positive definite for all ρ under consideration, which implies that the optimization problem (10.58) is nonconvex since the objective function is concave in $\mathbf{u}_{\kappa,m_h}$.

10.4.3 Semidefinite Relaxation

The optimization problem (10.58) can be rewritten as

$$\begin{aligned} \min_{\mathbf{u}_{\kappa,m_h}} \quad & -\operatorname{Tr}(\mathcal{Q}_{\rho} \mathbf{U}_{\kappa,m_h}) - q^{\mathrm{T}}_{\rho} \mathbf{u}_{\kappa,m_h} - q_{\rho,0} \\ \text{s.t.} \quad & \mathbf{U}_{\kappa,m_h} = \mathbf{u}_{\kappa,m_h} \mathbf{u}^{\mathrm{T}}_{\kappa,m_h}, \\ & \mathbf{u}_{\kappa,m_h} \in \mathcal{U}, \\ & \mathcal{G}^{k}_{u} \mathbf{u}_{\kappa,m_h} + g^{k}_{\kappa} \in \mathcal{Y}, \ k \in \{i,j\}, \end{aligned} \tag{10.59}$$

where a dummy matrix variable $\mathbf{U}_{\kappa,m_h}$ is introduced and the first equality corresponds to the only nonconvex relation. More general and explicit form of the optimization can be written as a (nonconvex) QCQP

$$\begin{aligned} \min_{\mathbf{x}} \ & \mathbf{x}^{\mathrm{T}}\mathcal{Q}\mathbf{x} \\ \text{s.t. } & \mathbf{x}^{\mathrm{T}}\mathcal{A}_\ell\mathbf{x} \geq 0, \ \ell = 1,\ldots,m_q, \\ & \mathcal{B}\mathbf{x} \geq 0, \\ & \mathbf{x} = [1 \ \mathbf{u}_{\kappa,m_h}^{\mathrm{T}}]^{\mathrm{T}}, \end{aligned} \tag{10.60}$$

where $\mathcal{Q} \triangleq \begin{bmatrix} -q_{\rho,0} & -1/2q_\rho^{\mathrm{T}} \\ -1/2q_\rho & -\mathcal{Q}_\rho \end{bmatrix}$, and the matrices $\mathcal{A}_\ell$ and $\mathcal{B}$ can be explicitly obtained from the constraint sets $\mathcal{U}$ and $\mathcal{Y}$ and the system matrices associated with the hypothesized models indexed by $k \in \{i,j\}$. This (nonconvex) QCQP can be rewritten as

$$\begin{aligned} \min_{\mathbf{X}} \ & \mathrm{Tr}(\mathcal{Q}\mathbf{X}) \\ \text{s.t. } & \mathrm{Tr}(\mathcal{A}_\ell\mathbf{X}) \geq 0, \ \ell = 1,\ldots,m_q, \\ & \mathcal{B}\mathbf{X}e_1 \geq 0, \\ & \mathcal{B}\mathbf{X}\mathcal{B}^{\mathrm{T}} \geq 0, \\ & e_1^{\mathrm{T}}\mathbf{X}e_1 = 1, \\ & \mathbf{X} \succeq 0, \\ & \mathrm{rank}(\mathbf{X}) = 1, \end{aligned} \tag{10.61}$$

where e_1 denotes the first standard basis vector in $\mathbb{R}^{m_h+1}$.

By removing the (nonconvex) rank constraint, $\mathrm{rank}(\mathbf{X}) = 1$, the corresponding primal SDP relaxation is

$$\begin{aligned} \min_{\mathbf{X}} \ & \mathrm{Tr}(\mathcal{Q}\mathbf{X}) \\ \text{s.t. } & \mathrm{Tr}(\mathcal{A}_\ell\mathbf{X}) \geq 0, \ \ell = 1,\ldots,m_q, \\ & \mathcal{B}\mathbf{X}e_1 \geq 0, \\ & \mathcal{B}\mathbf{X}\mathcal{B}^{\mathrm{T}} \succeq 0, \\ & e_1^{\mathrm{T}}\mathbf{X}e_1 = 1, \\ & \mathbf{X} \succeq 0, \end{aligned} \tag{10.62}$$

where e_1 denotes the first standard basis vector in $\mathbb{R}^{m_h+1}$. Its dual problem can be written as

$$\begin{aligned}
&\max_{\gamma,\lambda,\mu,\mathbf{Y}} \gamma \\
&\quad \text{s.t. } \mathcal{Q} \succeq \gamma e_1 e_1^{\mathrm{T}} + \sum_{\ell=1}^{m_q} \lambda_\ell \mathcal{A}_\ell + \mathcal{B}^{\mathrm{T}} \mu e_1^{\mathrm{T}} + e_1 \mu^{\mathrm{T}} \mathcal{B} + \mathcal{B}^{\mathrm{T}} \mathbf{Y} \mathcal{B}, \\
&\qquad \lambda_\ell \geq 0,\ \ell = 1, \ldots, m_q, \\
&\qquad \mu \geq 0, \\
&\qquad \mathbf{Y} \succeq 0, \\
&\qquad \mathbf{Y}_{ii} = 0,\ i = 1, \ldots, m_l.
\end{aligned} \tag{10.63}$$

Remark 10.12. The semidefinite program (10.63) is the Lagrangian dual of the QCQP (10.60).

Using different measures defined in (10.56), the objective function in the SDP relaxation (10.62) can be replaced by $\min_{\mathbf{X}} \max_{k\in\{i,j\}} \mathrm{Tr}(\mathcal{Q}^k \mathbf{X})$ or $\min_{\mathbf{X}} \min_{k\in\{i,j\}} \mathrm{Tr}(\mathcal{Q}^k \mathbf{X})$ where the symmetric matrices $\mathcal{Q}^k$ can be computed from (10.56) and the associated system matrices for the hypothesized models indexed by $k \in \{i, j\}$. Notice that $\max_{k\in\{i,j\}} \mathrm{Tr}(\mathcal{Q}^k \mathbf{X})$ is convex whereas $\min_{k\in\{i,j\}} \mathrm{Tr}(\mathcal{Q}^k \mathbf{X})$ is concave, which indicates a preference for $\rho_{\mathrm{KL}}^{\min}$ instead of $\rho_{\mathrm{KL}}^{\max}$.

Lemma 10.6 (*[196]*). Consider a hypercube constraint $\mathcal{U} = \mathcal{U}_\infty(\cdot,\cdot)$ and $\mathcal{Y} = \mathbb{R}^{N_h}$. The performance bounds achieved by the SDP relaxation is

$$J_{\mathrm{sdp}}^\star \leq J_{\mathrm{qcqp}}^\star \leq \frac{\pi}{2} J_{\mathrm{sdp}}^\star.$$

Claim 10.1. For any constraints given in (10.47) and (10.48), there exists a constant $c > 0$[10] such that

$$J_{\mathrm{sdp}}^\star \leq J_{\mathrm{qcqp}}^\star \leq c J_{\mathrm{sdp}}^\star.$$

10.4.4 Optimality Criteria for Model Discrimination with Multiple Hypotheses

Suppose that there are more than two candidate fault scenarios including the nominal operation, $\mathcal{H}_N \triangleq \{H_1, \cdots, H_N\}$. For an optimality criterion that quantifies information included in the predicted input-output data $z \in \mathcal{U} \times \mathcal{Y}$, we propose to consider one of the following:

i. Maximizing the minimum statistical distance among the hypothesized models

$$\max_{z\in\mathcal{U}\times\mathcal{Y}} \min_{\{j>i, i=1,\ldots,N\}} \rho_{ij}(z); \tag{10.64}$$

[10] We observe that the constant c could depend on the number of ellipsoidal constraints.

ii. Maximizing the average statistical distance of the hypothesized models

$$\max_{z\in\mathcal{U}\times\mathcal{Y}} \sum_{i=1}^{N-1}\sum_{j>i}^{N} \rho_{ij}(z); \tag{10.65}$$

iii. Maximizing the weighted average statistical distance of the hypothesized models

$$\max_{z\in\mathcal{U}\times\mathcal{Y}} \sum_{i=1}^{N-1}\sum_{j>i}^{N} \gamma_{ij}\rho_{ij}(z),\ \gamma_{ij}\in[0,1],\ \sum_{i=1}^{N-1}\sum_{j>i}^{N}\gamma_{ij}=1, \tag{10.66}$$

where $\rho_{ij}(z) \triangleq \rho(p_{\mathbf{z}|\mathbf{H}}(z|H_i), p_{\mathbf{z}|\mathbf{H}}(z|H_j))$ with $\rho \in \{\rho_{\mathrm{KL}}^{\min}, \rho_{\mathrm{KL}}^{\mathrm{ave}}, \rho_{\mathrm{Bhat}}\}$.

Remark 10.13. Note that if $\rho_{ij}(\cdot)$ are concave for all indices then the aforementioned optimizations (10.64), (10.65), and (10.66) are all convex, provided that the constraint $\mathcal{U}\times\mathcal{Y}$ is convex.

10.4.5 Randomized Algorithms: Suboptimal Solutions

A randomized algorithm for computing a rank 1 solution from the optimal solutions $\mathbf{X}_\ell^\star$, $\ell = 1,\ldots,N$, of the SDP relaxation (10.62) is presented in [196, 198]. Using a Cholesky factorization $P\mathbf{X}_\ell^\star P^{\mathrm{T}} = S_\ell^{\mathrm{T}} S_\ell$,[11]

$$\tilde{\mathbf{x}}_\ell := D\mathrm{sgn}(\mathrm{S}_\ell^{\mathrm{T}}\xi),$$

where ξ is a Gaussian random vector whose distribution is $\mathcal{N}(0,\mathrm{I})$ and $D>0$ is the diagonal scaling matrix such that $\tilde{\mathbf{x}}$ satisfies the constraints in (10.60). An optimal scaling matrix may be defined by the convex optimization

$$\begin{aligned} &\max\ \|\mathrm{diag}(D)\|_p \\ &\text{s.t. } \tilde{\mathbf{x}}_\ell = D\mathrm{sgn}(\mathrm{S}_\ell^{\mathrm{T}}\xi),\ \tilde{\mathbf{x}}_\ell\in\mathcal{C} \end{aligned} \tag{10.67}$$

where $p\in[1,\infty]$ refers in the vector p-norm, $\mathcal{C}$ is the intersection of the quadratic and linear constraints given in (10.60), and ξ is a realization from the distribution $\mathcal{N}(0,\mathrm{I})$. Another method for computing a rank 1 solution from the optimal solutions $\mathbf{X}_\ell^\star$ of the SDP relaxation (10.62) is a biased randomized algorithm:

$$\tilde{\mathbf{x}}_\ell := \bar{\mathbf{x}}_\ell + D\mathrm{sgn}(\mathrm{S}_\ell^{\mathrm{T}}\xi)$$

where $\bar{\mathbf{x}}_\ell$ is the singular vector corresponding to the largest singular value of $P\mathbf{X}_\ell^\star P^{\mathrm{T}}$. Another method for computing a rank 1 solution from the optimal solutions $\mathbf{X}_\ell^\star$ of the SDP relaxation (10.62) is

$$\tilde{\mathbf{x}}_\ell := \bar{\mathbf{x}}'_\ell + D\mathrm{sgn}(\mathrm{S}_\ell^{\mathrm{T}}\xi)$$

[11] $P\mathbf{X}_\ell^\star P^{\mathrm{T}}$ is the matrix obtained by removing the first row and column of $\mathbf{X}_\ell^\star$.

where $\bar{\mathbf{x}}'_\ell = (\mathbf{X}^\star_\ell)_{2:N_h,1}$ is the vector obtained from the first column vector of $\mathbf{X}^\star_\ell$ from which the first element is excluded.

Remark 10.14. Since D depends on the random vector $\xi \sim \mathcal{N}(0, \mathrm{I})$, it is also a stochastic matrix.

By generating N_s samples $\{\xi_n\}_{n=1}^{N_\mathrm{s}}$ of ξ from the distribution $\mathcal{N}(0, \mathrm{I})$, compute an approximate suboptimal solution

$$\hat{\mathbf{x}} := \arg \min_{n=1,\cdots,N_\mathrm{s}} \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_n \end{bmatrix}^\mathrm{T} \mathcal{Q} \begin{bmatrix} 1 \\ \tilde{\mathbf{x}}_n \end{bmatrix} \tag{10.68}$$

where $\tilde{\mathbf{x}}_n$ is a feasible solution for the constraints in (10.60) associated with the nth sample ξ_n. The notation $\hat{\mathbf{x}}(N_\mathrm{s})$ can be used to denote its dependence on the number of samples.

10.5 Discussion

10.5.1 Simulation Results

To illustrate and compare the input design methods, consider fault scenarios for an aircraft system. The numerical data are adopted from [32]. Consider a discretized dynamical system model for the longitudinal aircraft dynamics, in which the sampling time is 0.5 sec. The state variables are $x = [V_y,\, V_x,\, \omega,\, \theta,\, L]^\mathrm{T}$ where V_y, V_x, $\omega = \dot{\theta}$, θ, and L refer to the vertical velocity, horizontal velocity, pitch rate, pitch angle, and altitude, respectively. The measurable outputs are $y = [V_y,\, \omega]^\mathrm{T}$. For nominal operation without any faults corresponding to the hypothesis H_0, the resulting system matrices are

$$A^0 = \begin{bmatrix} 0.9985 & 0.1950 & 0 & -0.161 & 0 \\ -0.0325 & 0.8405 & 3.87 & 0 & 0 \\ 0.01 & -0.0505 & 0.7855 & 0 & 0 \\ 0 & 0 & 0.5 & 1.0 & 0 \\ 0.5 & 0 & 0 & 0 & 1.0 \end{bmatrix},\ B^0 = \begin{bmatrix} 0.005 \\ -0.09 \\ -0.58 \\ 0 \\ 0 \end{bmatrix},\ E^0 = \mathrm{I}_5,$$

$$C^0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},\ D^0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ F^0 = \mathrm{I}_2.$$

Consider three different types of fault scenarios of the hypotheses and the associated system matrices given by the followings:

◦ H_1: Failure of the vertical velocity sensor

- $M_1 = (A^1, B^1, C^1, D^1, E^1, F^1)$ where $A^1 = A^0$, $B^1 = B^0$, $D^1 = D^0$, $E^1 = E^0$, $F^1 = F^0$, and $C^1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$;

- H_2: Failure of the pitch rate sensor

  - $M_2 = (A^2, B^2, C^2, D^2, E^2, F^2)$ where $A^2 = A^0$, $B^2 = B^0$, $D^2 = D^0$, $E^2 = E^0$, $F^2 = F^0$, and $C^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$;

- H_3: Failure of the elevator actuator

  - $M_3 = (A^3, B^3, C^3, D^3, E^3, F^3)$ where $A^3 = A^0$, $C^3 = C^0$, $D^3 = D^0$, $E^3 = E^0$, $F^3 = F^0$, and $B^3 = 0_{5\times 1}$.

The inputs are assumed to be bounded in amplitude, $\mathcal{U}_\infty(-1/2, 1/2)$.

Example 10.1. Consider the two hypotheses $\mathcal{H} = \{H_0, H_1\}$. Fig. 10.1 shows the trajectories of inputs that are computed from solving the associated convex relaxations of optimal input design problems with different objective functions of statistical distance measures. Fig. 10.2 presents the resultant trajectories of V_y. Two input designs induce an oscillation about the nominal value whereas one input design induces a bias.

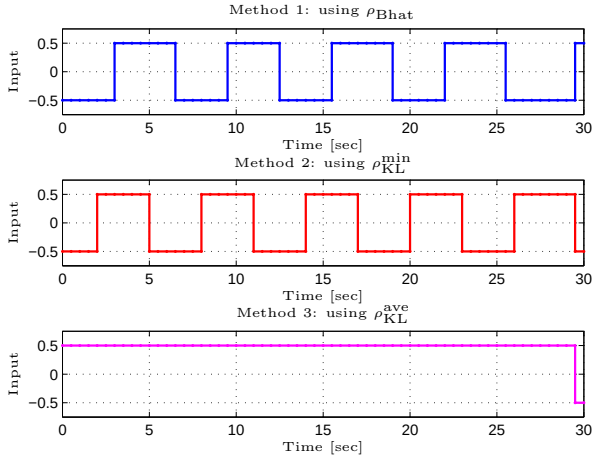


Figure 10.1: Input sequences obtained from the three design methods for $\mathcal{H} = \{H_0, H_1\}$.

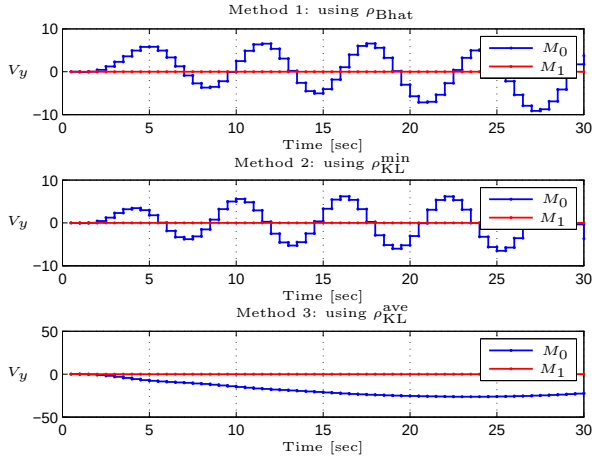


Figure 10.2: The expected trajectory of V_y generated by the input design methods for $\mathcal{H} = \{H_0, H_1\}$.

Example 10.2. Consider the two hypotheses $\mathcal{H} = \{H_0, H_2\}$. Fig. 10.3 shows the trajectories of inputs that are computed from solving the convex relaxations of the optimal input design problems with different objective functions of statistical distance measures. Fig. 10.4 presents the resulting trajectories of ω.

Example 10.3. Consider the two hypotheses $\mathcal{H} = \{H_0, H_3\}$. Fig. 10.5 shows the input trajectories computed from solving the associated convex relaxations of optimal input design problems with different statistical distance measures. Figs. 10.6 and 10.7 present the resulting trajectories of V_y and ω, respectively. The optimal solutions obtained from the three different methods are identical for this example.

Example 10.4. Consider $\mathcal{H} = \{H_0, H_1, H_2, H_3\}$. Fig. 10.8 shows the input trajectories computed from solving the associated convex relaxations of optimal input design problems with different statistical distance measures. Since there are more than two models of hypotheses to compare, we solve a mini-max problem for which an optimal solution minimizes the maximum among the statistical distance measures of each pair of hypotheses. The number of pairs is $N(N+1)/2$ where N is the number of system models (or modes) for hypotheses, which implies that the associated convex relaxation for a multiple hypotheses test can be still

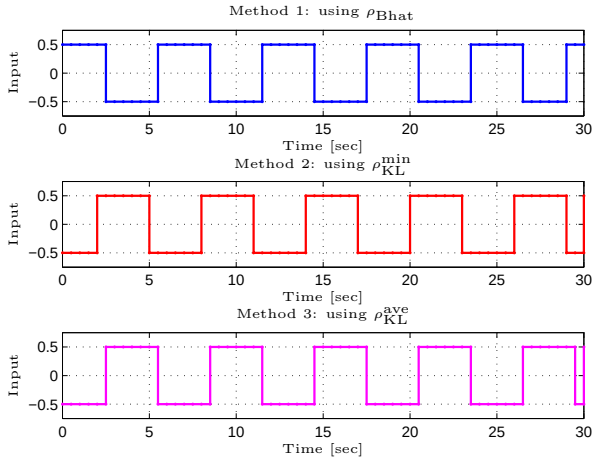


Figure 10.3: Input sequences obtained from the design methods for $\mathcal{H} = \{H_0, H_2\}$.

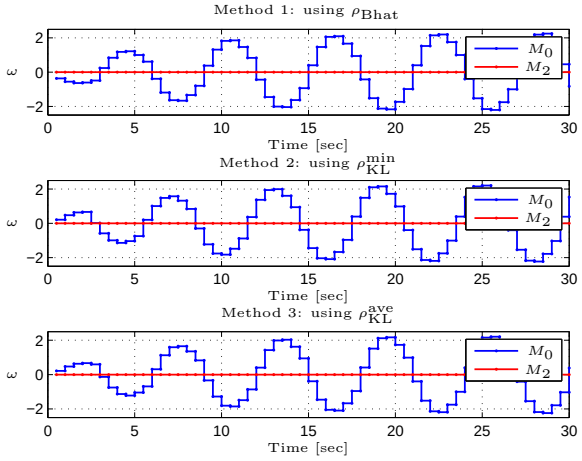


Figure 10.4: The expected trajectory of ω generated by the input design methods for $\mathcal{H} = \{H_0, H_2\}$.

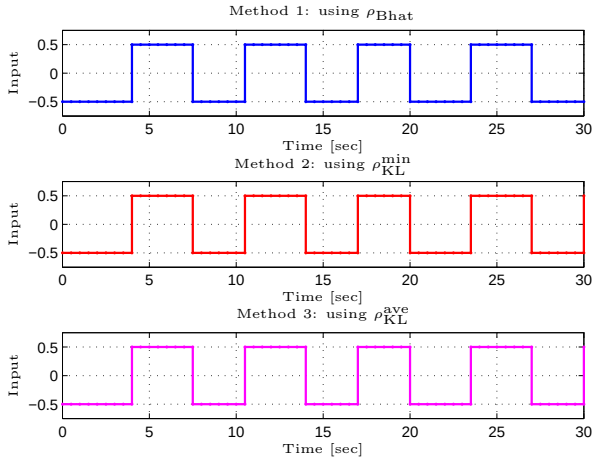


Figure 10.5: Input sequences obtained from the design methods for $\mathcal{H} = \{H_0, H_3\}$.

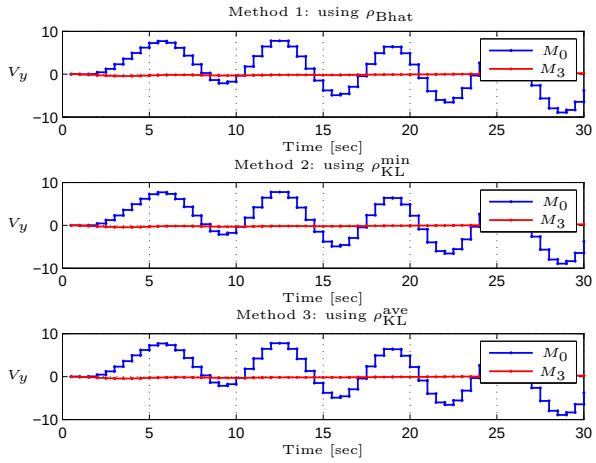


Figure 10.6: The expected trajectory of V_y generated by the input design methods for $\mathcal{H} = \{H_0, H_3\}$.

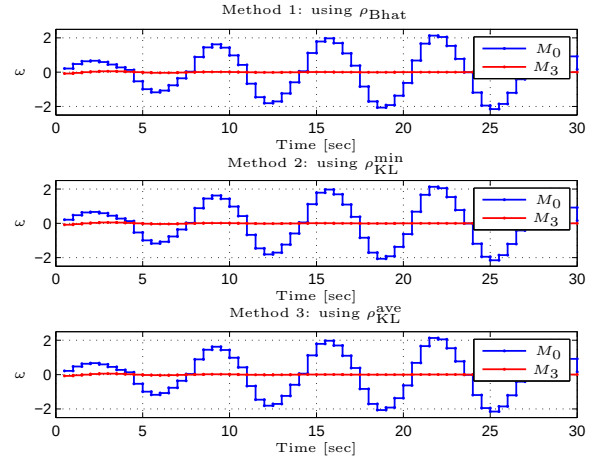


Figure 10.7: The expected trajectory of ω generated by the input design methods for $\mathcal{H} = \{H_0, H_3\}$.

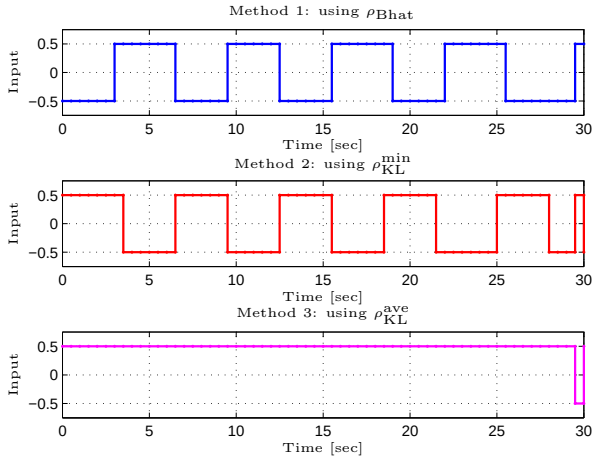


Figure 10.8: Input sequences obtained from the design methods for multiple hypotheses $\mathcal{H} = \{H_0, H_1, H_2, H_3\}$.

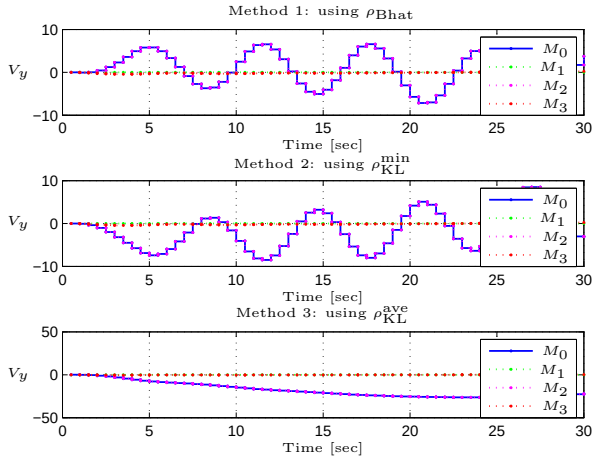


Figure 10.9: The expected trajectory of V_y generated by the input design methods for multiple hypotheses $\mathcal{H} = \{H_0, H_1, H_2, H_3\}$.

solved in polynomial-time. Figs. 10.9 and 10.10 present the resulting trajectories of V_y and ω, respectively, in which sharp distinctions between the four models can be observed for all three input design methods. The output V_y shows the distinct behaviors for the models M_0 and M_2 whereas the models M_1 and M_3 have the same trajectory. The output ω shows the different behaviors for the models M_1 and M_3 whereas the models M_0 and M_2 have the same trajectory.

10.6 Summary and Future Work

In this chapter, we considered optimal *active* input design problems for fault detection and diagnosis based on Bayesian inference. The resulting optimization for input design is to maximize statistical discrimination between models of hypotheses corresponding to fault scenarios, while requiring the controlled state/output

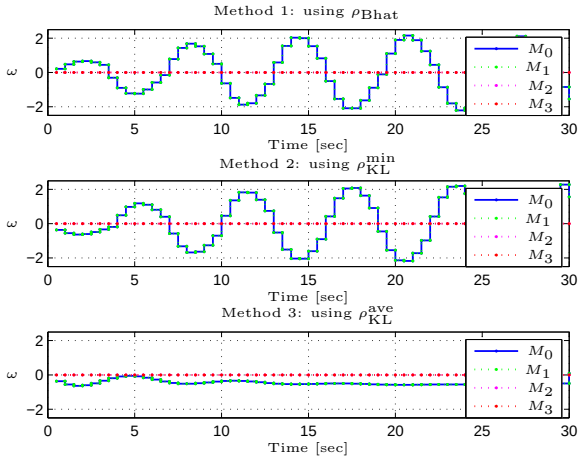


Figure 10.10: The expected trajectory of ω generated by the presented design methods for multiple hypotheses $\mathcal{H} = \{H_0, H_1, H_2, H_3\}$.

trajectories as well as the inputs remain certain bounds. Each model of a fault scenario characterizes a random process of the measurable outputs and, to quantify the quality of the measurable data for FDD, three different measures for the statistical distance between two random processes and two approximate measures of them are considered. With such statistical distance measures, the original optimization is non-convex even without any constraints, which would be computationally expensive, especially for multiple hypothesis tests. First, we proposed a sequential SDP method to find a local optimum that can be further improved by using multiple shooting or warm starts. In addition, closed-loop state feedback input design problems are proposed, for which semi-chance constraints are introduced to impose bounds on the expected controlled trajectories and their variances. Second, Convex relaxation methods were proposed that compute approximate solutions for which the potential degree of sub-optimality is known in some special cases. Simulation results are included to demonstrate the proposed active input design methods for FDD.

Chapter 11

Belief Propagation and Optimization for Distributed Fault Detection and Diagnosis

Abstract This chapter develops distributed Bayesian hypothesis tests for fault detection and diagnosis that are based on belief propagation and optimization in graphical models. The main challenges in developing distributed statistical estimation algorithms are (i) difficulties in ensuring convergence and consensus for solutions of distributed inference problems, (ii) increasing computational costs due to lack of scalability, and (iii) communication constraints for networked multi-agent systems. To cope with those challenges, this chapter considers (i) belief propagation and optimization in graphical models of complex distributed systems, (ii) decomposition methods of optimization for parallel and iterative computations, and (iii) distributed decision-making protocols. This chapter discusses further research directions for efficient and proper use of the proposed methods in distributed statistical inference.

11.1 Introduction

Stochastic inference using graphical models [63, 283] have been important research topics in a variety of disciplines that include signal processing [256], machine learning [87], and artificial intelligence [213]. For the use of graphical models in statistical inference problems, optimal fusion of information and/or data over networked agents that are individual decision makers or processors and the design of compromised inference methods for distributed decision makers have far significant importance.

Pearl [213] referred to *belief propagation* (BP) as a message-passing algorithm for which local evidences are exchanged as messages that are used to update local beliefs and to find fixed-points of iterations, corresponding to marginal probability distributions of the node states. In a standard BP method for statistical inference in a graphical model, agents on the nodes exchange messages with neighboring agents connected

over the edges. The BP algorithm is known to provide exact marginal distributions when the graphical model are tree-structured, i.e., of no cyclic loops [213]. In the presence of cyclic loops in a graphical model, neither convergence nor optimally of BP methods can be, in general, guaranteed, although some empirical studies on performance of loopy BP [188] and conversion to equivalent cycle-free graphical models [62] are available.

The main challenges in the development of BP algorithms with general Markov and Bayesian graphical models are

(i) **Convergence Analysis**: As previously mentioned, message-passing algorithms of BP do not generally converge to a fixed point in the presence of cyclic loops.

(ii) **Scalability**: In a tree-structured graphical model, BP algorithms can find a fixed point in $O(n)$ iterations, where n is the diameter of the graph. However, calculation of posterior marginal probabilities on nodes in an arbitrary Bayesian network is known to be NP-complete [59, 238] and even an approximate computation of posterior marginal probabilities is NP-hard [66].

(iii) **Communication Constraints**: Message-passing or information exchange over communication networks are not necessarily reliable, and communication bandwidth and energy constraints are typical sources of degrading performance of networked inference algorithms [53].

To cope with the aforementioned difficulties confronted to BP methods for statistical inference in graphical models, consider

→(i) **Belief Optimization**: In [299], it was shown that BP fixed points correspond to the stationary points of the Bethe free energy approximation for a factor graph. The associated constrained minimization is called *belief optimization* (BO). This chapter presents statistical inference methods based on the same principle that the joint probability distribution of the node states in a graphical model is a minimizer of the free energy, and the beliefs, corresponding to marginal probabilities of the node states, can be computed from minimizing approximate free energy such the mean field and Bethe free energies. The resultant statistical inference problems are formulated as constrained minimizations.

→(ii) **Decomposition Methods of Optimization**: Belief optimization is large-scale constrained minimization that becomes intractable and non-scalable as the number of nodes and cardinality of the node states increase. Since the coupling between marginal probabilities to be determined are constrained on the edges in graphical models, natural ways of reducing computational demand are to use decomposition methods for optimization.

→(iii) **Distributed Decision Processes**: In the presence of communication constraints, decision processes and information exchange need to be localized and distributed for reliable statistical inference over graphical models.

The main applications of BP/BO methods of interest in this chapter are in the derivation of distributed hypothesis tests for fault detection and diagnosis (FDD) in large-scale distributed dynamical systems. Developing automatic monitoring, detection, and diagnosis of system faults has rapidly growing importance as the size and complexity of systems increase. Most of existing methods for model-based FDD are centralized schemes in the sense that the central decision maker can access all measurements and the decision goal is to decide whether faults occur and determine the types and locations of faults. Distributed FDD is suitable for large-scale interconnected and networked dynamical systems such as multi-agent systems and power grids. Furthermore, since not all measurements are accessible to local processors and computation nodes, centralized FDD schemes may not be applicable to distributed systems. Belief propagation and optimization provide naturally suitable ways of distributed statistical inference and decision making, for which graphical models are used for representation of interconnections and networks of local sensors (measurements) and processors (data/information processing), and belief consensus constraints are required to be satisfied by exchanging messages for BP and by imposing public variable constraints for BO.

11.2 Belief Propagation in Graphical Models

BP algorithms are developed for graphical models. This section provides a concise discussion of graphical representations and the corresponding BP methods for distributed inference problems. There are two types of graphical models that are used to represent probabilistic and informational dependencies of random variables—Markov networks and Bayesian networks. A *Markov network* is defined with an undirected graph whose nodes correspond to random variables and the edges correspond to their probabilistic and information dependencies. A *Bayesian network* is defined with a directed graph whose nodes correspond to random variables and the arrows are used to denote causality constraints or class-property relations. Since a focus on developing distributed Bayesian hypothesis tests for FDD using BP/BO, this chapter only considers Markov network models. Many research monographs are available that provide a tutorial on graphical models (see [56, 63, 283], for example).

11.2.1 Pairwise MRF

Markov networks (aka Markov random field (MRF) models) are suitable for representing conditional dependencies of the node states.

Definition 11.1 (*MRF*). The random vector X is Markov with respect to the graph $G = (V, E)$ if, for any partition of the node set V into disjoint sets A, B, C in which B separates A and C, the degenerate random vectors X_A, X_B, X_C corresponding to each node set are conditionally independent in the sense that $P_{AB|C}(x_a, x_b|x_c) = P_{A|B}(x_a|x_b)P_{C|B}(x_c|x_b)$, or equivalently $P_{A|BC}(x_a|x_b, x_c) = P_{A|B}(x_a|x_b)$ (or symmetrically, $P_{C|AB}(x_c|x_a, x_b) = P_{C|B}(x_c|x_b)$).

The Hammersley-Clifford theorem provides a sufficient (and necessary) condition for which the joint probability distribution of the node states can be represented as an MRF.

Theorem 11.1 (*Hammersley-Clifford Theorem (see [56, 213])*). The random vector X is Markov w.r.t. the graph G if (and only if for strictly positive probability distributions) its distribution can be factorized by a product of variables restricted to cliques, i.e., the joint probability can be factorized as

$$P(\mathbf{x}) = \gamma \prod_{C \in \mathcal{C}} \psi_C(x_c) \tag{11.1}$$

where $\gamma = (\sum_x \prod_{C \in \mathcal{C}} \psi_C(x_c))^{-1}$ and $\mathcal{C}$ refers to the set of cliques in G.

The $\psi_C(x_c)$ are called the *compatibility functions* that correspond to the marginal probabilities, and their negative logarithms are referred to as *potentials* or *potential functions*, $V_C(x_c) := -\ln \psi_C(x_c) \geq 0$. The factorization (11.1) can be rewritten as

$$P(\mathbf{x}) = \gamma \left(\prod_{k \in V} \psi_k(x_k) \right) \left(\prod_{(i,j) \in E} \psi_{ij}(x_i, x_j) \right) \left(\prod_{C \in \mathcal{C} \setminus V, E} \psi_C(x_c) \right). \tag{11.2}$$

Assumption 11.1 (*Pairwise Potentials*). Assume that either

(i) there is no clique with more than two nodes in the graph G, or

(ii) the potentials are only defined by the variable as a single node in V or by the two variables as a pair of nodes on an edge in E.

Under Assumption 11.1, there is no contribution of the last term in (11.2), i.e.,

$$P(\mathbf{x}) \equiv \hat{P}(\mathbf{x}) \triangleq \gamma \left(\prod_{k \in V} \psi_k(x_k) \right) \left(\prod_{(i,j) \in E} \psi_{ij}(x_i, x_j) \right) \tag{11.3}$$

where $\hat{P}(\mathbf{x})$ can be interpreted as an approximation of the joint probability distribution $P(x)$ of the random variable X that is Markov w.r.t. $G = (E, V)$, up to the 2-cliques.

11.2.1.1 Graphical Models for Distributed Inference

The rest of this chapter assumes that there are local measurements (or evidences) $y_k \in \mathcal{Y}_k$ that are associated with the node $k \in V$. For any non-loopy graph, i.e, graphical models on trees, the compatibility functions can be represented in terms of the marginal probabilities up to the 2-cliques: $\psi_k(x_k) = p_k(x_k)p(y_k|x_k)$ for $k \in V$ and $\psi_{ij}(x_i, x_j) = p_{ij}(x_i, x_j)p(y_i, y_j|x_i, x_j)/p_i(x_i)p(y_i|x_i)p_j(x_j)p(y_j|x_j)$ for $(i, j) \in E$. With this

representation of the compatibility functions, $\hat{P}(X)$ can be rewritten as

$$\hat{P}(\mathbf{x}) = \gamma \left(\prod_{k \in V} p_k(x_k) p(y_k|x_k) \right) \left(\prod_{(i,j) \in E} \frac{p_{ij}(x_i, x_j) p(y_i, y_j|x_i, x_j)}{p_i(x_i) p(y_i|x_i) p_j(x_j) p(y_j|x_j)} \right) \tag{11.4}$$

or

$$\hat{P}(\mathbf{x}) = \gamma \left(\prod_{k \in V} p(x_k|y_k) \right) \left(\prod_{(i,j) \in E} \frac{p(x_i, x_j|y_i, y_j)}{p(x_i|y_i) p(x_j|y_j)} \right) \tag{11.5}$$

where, with a minor abuse of notation, γ might not be the same as the γ in (11.3), but can be considered as an equivalent partition function (value).

For the purpose of distributed statistical inference in a graphical model, a goal is to estimate the posterior marginal probabilities, for which messages from the neighboring nodes are required to have sufficient statistics of local measurements that can be considered as realizations from unknown probability distributions.

Problem 11.1. Consider an undirected graph $G = (V, E)$. Compute (or approximate) the posterior marginal probabilities

$$p_k(x_k|y_1, \cdots, y_N), \quad k \in V \tag{11.6}$$

where $N = |V|$.

To exactly solve Problem 11.1, the required property of a BP method is the relation of sufficient statistics

$$p_k(x_k|y_1, \mu_k) \equiv p_k(x_k|y_1, \cdots, y_N), \quad k \in V \tag{11.7}$$

where μ_k refers to the total messages delivered to the agent at the node k.[1]

11.2.1.2 Distributed Belief Propagation

In the aforementioned BP algorithms, there are slightly different methods of computing messages to be transmitted, which have different interests [299]: (a) the *max-product BP message* is to obtain a global state that is most probable in the Bayesian sense and consists of a local state maximizing the local belief, and (b) the *sum-product BP message* is to compute marginal posterior probabilities, given the total evidence or measurements that are available in the system. Their properties are clarified below.

The Max-Product BP A goal of a belief propagation algorithm for Bayesian estimation, particularly for maximum a posteriori estimation, can be to achieve the relation

$$\beta_k(x_k) = \alpha_k \max_{x_{-k}} p_k\left(x_k, x_{-k}|y_1, \cdots, y_N\right), \quad \forall x_k, \ \forall k \in V, \tag{11.11}$$

[1] Agent k refers to a processor or decision maker at the node k.

In a belief propagation algorithm, the belief at the node k in its state x_k is

$$\beta_k(x_k) \propto \psi_k(x_k) \prod_{\ell \in \mathcal{N}(k)} \mu_{\ell \to k}(x_k) \tag{11.8}$$

and the message from the node ℓ to the node k about the state x_k can be either the sum-product BP message

$$\mu_{\ell \to k}(x_k) \propto \sum_{x_\ell} \psi_{\ell k}(x_\ell, x_k) \psi_\ell(x_\ell) \prod_{u \in \mathcal{N}(\ell) \setminus \{k\}} \mu_{u \to \ell}(x_\ell) \tag{11.9}$$

or the max-product BP message

$$\mu_{\ell \to k}(x_k) \propto \max_{x_\ell} \psi_{\ell k}(x_\ell, x_k) \psi_\ell(x_\ell) \prod_{u \in \mathcal{N}(\ell) \setminus \{k\}} \mu_{u \to \ell}(x_\ell), \tag{11.10}$$

where conditional dependence of the beliefs and messages on measurements $Y = \{y_i\}_{i=1}^n$ is dropped to simplify the notation.

for given total measurement data $\{y_k\} \in Y$, where each α_k is a positive constant that is independent of the value of x_k and results in $\beta_k(\cdot) \in [0, 1]$. Alternatively, a slightly weaker relation is that, for given measurement data $\{y_k\}$,

$$\beta_k(x) \leq \beta_k(z) \;\Rightarrow\; \max_{x_{-k}} p_k\left(x, x_{-k} | y_1, \cdots, y_N\right) \leq \max_{x_{-k}} p_k\left(z, x_{-k} | y_1, \cdots, y_N\right), \tag{11.12}$$

for all nodes $k \in V$. Note that this relation ensures marginal maximum a posteriori (m-MAP) estimation, i.e.,

$$\begin{aligned} x_k^\star &= \arg\max_x \beta_k(x) \\ &= \arg\max_x p_k\left(x, x_{-k}^\star | y_1, \cdots, y_N\right) \end{aligned} \tag{11.13}$$

and results in the joint MAP (j-MAP) estimator satisfying the relation

$$\{x_k^\star\} = \arg\max_{\{x_i\}} p\left(x_1, \cdots, x_N | y_1, \cdots, y_N\right). \tag{11.14}$$

The Sum-Product BP Similar to the max-product BP algorithm, the goal of the sum-product BP is to achieve the relation

$$\beta_k(x_k) = \alpha_k \sum_{x_{-k}} p_k(x_k, x_{-k} | y_1, \cdots, y_N), \;\; \forall k \in V, \tag{11.15}$$

where the summation is computed for all realizations of the compound random vector x_{-k} and each α_k is a positive constant that is independent of the value of x_k and results in $\beta_k(\cdot) \in [0, 1]$. This algorithm estimates the marginal posterior probabilities, for given total measurements.

Remark 11.1. A notable discrimination of the sum-product BP against the max-product BP is that the combination of optimal m-MAP estimators $x_k^\star = \arg\max_x \beta_k(x)$, where the beliefs are obtained from the

sum-product BP, is not necessarily an optimal j-MAP estimation.

Iterative Message-Passing and Fixed Points The following algorithm is a standard asynchronous iterative message-passing algorithm for belief propagation.

The belief at the node k in its state x_k at time t is

$$\beta_k^{(t)}(x_k) \propto \psi_k(x_k) \prod_{\ell \in \mathcal{N}(k)} \mu_{\ell \to k}^{(t)}(x_k) \tag{11.16}$$

and the message from the node ℓ to the node k about the state x_k at time t can be either the sum-product BP message update

$$\mu_{\ell \to k}^{(t)}(x_k) \propto \sum_{x_\ell} \psi_{\ell k}(x_\ell, x_k) \psi_\ell(x_\ell) \prod_{u \in \mathcal{N}(\ell) \setminus \{k\}} \mu_{u \to \ell}^{(t-1)}(x_\ell) \tag{11.17}$$

or the max-product BP message update

$$\mu_{\ell \to k}^{(t)}(x_k) \propto \max_{x_\ell} \psi_{\ell k}(x_\ell, x_k) \psi_\ell(x_\ell) \prod_{u \in \mathcal{N}(\ell) \setminus \{k\}} \mu_{u \to \ell}^{(t-1)}(x_\ell). \tag{11.18}$$

11.3 Belief Optimization in Graphical Models

11.3.1 Bethe Peirerls Approximation to the Free Energy

In [297–299], it was shown that the fixed points of BP and its generalization are associated with extrema of the Bethe and Kikuchi free energies, respectively. Below is a concise overview of some useful results from statistical physics. In particular, the observation that statistical inference problems can be represented as minimization of (approximate) free energy (see also [297, 299]) motivates the study of various approximate free energies.

11.3.1.1 Gibbs Free Energy in Statistical Physics

In statistical physics, the Boltzmann distribution law indicates that, for the energy $E(\mathbf{x})$ associated with some state or condition $\mathbf{x}$ of a system, the probability distribution of its occurrence is given by

$$p(\mathbf{x}) = \frac{1}{Z} \exp(-E(\mathbf{x})/T) \tag{11.19}$$

where Z denotes the partition function (constant) and T is the temperature that can be set to be 1 without loss of generality. Comparing this expression to the factorization (11.1) gives $\gamma = 1/Z$ and $E(x) = -\sum_{C \in \mathcal{C}} \ln \psi_C(x_c) = \sum_{C \in \mathcal{C}} V_C(x_c)$, i.e., the total energy is the sum of the potentials over the system. To compute the distance between the belief $\beta(\mathbf{x})$ and the true joint probability distribution, use the

Kullback-Leibler (KL) distance defined by

$$\begin{aligned} D(\beta||p) &= \sum_{\mathbf{x}} \beta(\mathbf{x}) \ln \frac{\beta(\mathbf{x})}{p(\mathbf{x})} \\ &= \sum_{\mathbf{x}} \beta(\mathbf{x}) E(\mathbf{x}) + \sum_{\mathbf{x}} \beta(\mathbf{x}) \ln \beta(\mathbf{x}) + \ln Z \end{aligned} \tag{11.20}$$

such that $D(\beta||p) = 0$ if and only if $\beta \equiv p$ and $D(\beta||p) \geq 0$ for all $\beta \in \Delta$ where Δ refers to the set of probabilities. Define the Gibbs free energy by

$$G(\beta) \triangleq \sum_{\mathbf{x}} \beta(\mathbf{x}) E(\mathbf{x}) + \sum_{\mathbf{x}} \beta(\mathbf{x}) \ln \beta(\mathbf{x}) = U(\beta) - H(\beta) \tag{11.21}$$

such that $D(\beta||p) = G(\beta) - F$ where $F \triangleq -\ln Z$ is called the *Helmholtz free energy*, and $U(\beta)$ and $H(\beta)$ refer to the average energy and the entropy, respectively.

11.3.1.2 Approximate Free Energy

Previously, it was assumed that the joint probability $p(\mathbf{x})$ is a function of the total energy function $E(\mathbf{x})$. Suppose that the system is of a pairwise MRF with the graph $G(V, E)$ in which there is no potential related to cliques with more than two nodes. Then the corresponding energy of such a configuration is

$$E(\mathbf{x}) = -\sum_{k \in V} \ln \psi_k(x_k) - \sum_{(i,j) \in E} \ln \psi_{ij}(x_i, x_j). \tag{11.22}$$

A. The Mean Field Free Energy In mean-field theory, the joint distribution $\beta(\mathbf{x})$ is approximated by complete factorization, i.e,

$$\beta(\mathbf{x}) \approx \prod_{k \in V} \beta_k(x_k). \tag{11.23}$$

With this approximate joint distribution under a pairwise MRF configuration, the mean-field average energy is

$$\tilde{U}(\{\beta_\ell\}_{\ell \in V}) = -\sum_{k \in V} \sum_{x_k} \beta_k(x_k) \ln \psi_k(x_k) - \sum_{(i,j) \in E} \sum_{x_i, x_j} \beta_i(x_i) \beta_j(x_j) \ln \psi_{ij}(x_i, x_j) \tag{11.24}$$

and similarly the mean-field entropy is

$$\tilde{H}(\{\beta_\ell\}_{\ell \in V}) = -\sum_{k \in V} \sum_{x_k} \beta_k(x_k) \ln \beta_k(x_k). \tag{11.25}$$

Note that the mean field free energy $\tilde{G} = \tilde{U} - \tilde{H}$ is a function of the separate one-node beliefs $\beta_k(\cdot)$.

B. The Bethe Free Energy For more general approximation, the joint distribution $\beta(\mathbf{x})$ can be approximated by the factorization with one- and two-node beliefs, viz.,

$$\beta(\mathbf{x}) \approx \frac{\prod_{(i,j)\in E} \beta_{ij}(x_i, x_j)}{\prod_{k\in V} \beta_k(x_k)^{q_k - 1}} \tag{11.26}$$

where $q_k = |\mathcal{N}(k)|$. With this approximate joint distribution under a pairwise MRF configuration, the Bethe average energy is

$$\begin{aligned} \tilde{U}(\{\beta_k\}_{k\in V}, \{\beta_{ij}\}_{(i,j)\in E}) = &- \sum_{k\in V} \sum_{x_k} \beta_k(x_k) \ln \psi_k(x_k) \\ &- \sum_{(i,j)\in E} \sum_{x_i, x_j} \beta_{ij}(x_i, x_j) \ln \psi_{ij}(x_i, x_j) \end{aligned} \tag{11.27}$$

and similarly the Bethe entropy is

$$\begin{aligned} \tilde{H}(\{\beta_k\}_{k\in V}, \{\beta_{ij}\}_{(i,j)\in E}) = &\sum_{k\in V} (q_k - 1) \sum_{x_k} \beta_k(x_k) \ln \beta_k(x_k) \\ &- \sum_{(i,j)\in E} \sum_{x_i, x_j} \beta_{ij}(x_i, x_j) \ln \beta_{ij}(x_i, x_j). \end{aligned} \tag{11.28}$$

Remark 11.2. In contrast to mean-field energy, the Bethe free energy is not generally an upper bound on the true Gibbs free energy [299].

11.3.2 Belief Optimization

Consider the discrete random variables $X_k \in \mathcal{X}_k \triangleq \{x_{k1}, x_{k2}, \cdots, x_{kn_k}\}$ with probability one and $|\mathcal{X}_k| = n_k$ for each $k \in V$. For the sake of notation, assume that all the nodes have the same cardinality of their supports, i.e., $n_k = n$ for all $k \in V$. Define the probability vector and matrix by

$$\underline{\beta_k} \triangleq \begin{bmatrix} \beta_k(x_{ki}) \\ \vdots \\ \beta_k(x_{kn}) \end{bmatrix}, \text{ for } k \in V \tag{11.29}$$

and

$$\underline{\underline{\beta_{ij}}} \triangleq \begin{bmatrix} \beta_{ij}(x_{i1}, x_{j1}) & \cdots & \beta_{ij}(x_{i1}, x_{jn}) \\ \vdots & \ddots & \vdots \\ \beta_{ij}(x_{in}, x_{j1}) & \cdots & \beta_{ij}(x_{in}, x_{jn}) \end{bmatrix}, \text{ for } (i,j) \in E, \tag{11.30}$$

respectively. The Belief Optimization (BO) is to find $\{\underline{\beta_k}\}_{k\in V}$ minimizing $\tilde{G}$ for the mean-field free energy approximation or $(\{\underline{\beta_k}\}_{k\in V}, \{\underline{\underline{\beta_{ij}}}\}_{(i,j)\in E})$ minimizing $\tilde{G}$ for the Bethe free energy approximation.

11.3.2.1 Minimization of The Mean Field Free Energy

A popular method of approximating a free energy is the aforementioned mean-field approach for which an optimal configuration of beliefs, that is, an approximation of joint probability distribution, can be obtained as a factorization (11.23) and the associated factors $\{\underline{\beta_k}\}_{k\in V}$ are optimal solutions of the constrained minimization

$$\begin{aligned} \min \;\; & \tilde{G} = \tilde{U} - \tilde{H} \\ \text{s.t.} \;\; & e^{\mathrm{T}}\underline{\beta_k} = 1, \; k \in V \\ & 0 \leq \underline{\beta_k} \leq 1, \; k \in V \end{aligned} \tag{11.31}$$

where $\tilde{U}$ and $\tilde{H}$ are given by (11.24) and (11.25), respectively. This optimization can be explicitly rewritten as

$$\begin{aligned} \min \quad & -\sum_{k\in V} \underline{\beta_k}^{\mathrm{T}} \ln \underline{\psi_k} - \sum_{(i,j)\in E} \underline{\beta_i}^{\mathrm{T}} \ln \underline{\underline{\psi_{ij}}}\, \underline{\beta_j} + \sum_{k\in V} \underline{\beta_k}^{\mathrm{T}} \ln \underline{\beta_k} \\ \text{s.t.} \quad & \underline{\beta_k} \in \Delta, \, k \in V \end{aligned} \tag{11.32}$$

where $\Delta \triangleq \{p \in \mathbb{R}^n : e^{\mathrm{T}}p = 1, \, p_i \in [0,1], \, \forall i\}$.

11.3.2.2 Minimization of The Bethe Free Energy

Similar to minimization of the mean-field free energy, an optimal configuration of beliefs, that is, an approximation of joint probability distribution, can be obtained as a factorization (11.26) and the associated factors $(\{\underline{\beta_k}\}_{k\in V}, \{\underline{\underline{\beta_{ij}}}\}_{(i,j)\in E})$ are optimal solutions of the constrained minimization

$$\begin{aligned} \min \;\; & \tilde{G} = \tilde{U} - \tilde{H} \\ \text{s.t.} \;\; & e^{\mathrm{T}}\underline{\beta_k} = 1, \; k \in V \\ & 0 \leq \underline{\beta_k} \leq 1, \; k \in V \\ & e^{\mathrm{T}}\underline{\underline{\beta_{ij}}} = \underline{\beta_j}^{\mathrm{T}}, \, \underline{\underline{\beta_{ij}}} e = \underline{\beta_i}, \, (i,j) \in E \end{aligned} \tag{11.33}$$

where $\tilde{U}$ and $\tilde{H}$ are given by (11.27) and (11.28), respectively. The optimization can be explicitly rewritten as

$$\begin{aligned} \min \quad & -\sum_{k\in V} \underline{\beta_k}^{\mathrm{T}} \ln \underline{\psi_k} - \sum_{(i,j)\in E} [\underline{\underline{\beta_{ij}}} \circ \ln \underline{\underline{\psi_{ij}}}] \\ & + \sum_{k\in V} (1-q_k) \underline{\beta_k}^{\mathrm{T}} \ln \underline{\beta_k} + \sum_{(i,j)\in E} [\underline{\underline{\beta_{ij}}} \circ \ln \underline{\underline{\beta_{ij}}}] \\ \text{s.t.} \quad & \underline{\beta_k} \in \Delta,\ k \in V \\ & e^{\mathrm{T}} \underline{\underline{\beta_{ij}}} = \underline{\beta_j}^{\mathrm{T}},\ \underline{\underline{\beta_{ij}}} e = \underline{\beta_i},\ (i,j) \in E \end{aligned} \tag{11.34}$$

where $[A \circ B] = \mathrm{Tr}(A^{\mathrm{T}} B)$ refers to entry-wise sum of the Hadamard (aka Schur) product $A \circ B$.

11.3.2.3 Minimization of The TAP Free Energy

The TAP (Thouless-Anderson-Palmer) approach that is used to approximate free energy in statistical mechanics has been adopted to robust decoding and statistical inference based on belief propagation (see [65, 125, 126], for example). An optimal configuration of beliefs, that is, an approximation of joint probability distribution, can be obtained as a factorization (11.23) and the associated factors $\{\underline{\beta_k}\}_{k\in V}$ are optimal solutions of the constrained minimization

$$\begin{aligned} \min \quad & \tilde{G} = \tilde{U} - \tilde{H} - \tilde{T} \\ \text{s.t.} \quad & \underline{\beta_k} \in \Delta,\ k \in V \end{aligned} \tag{11.35}$$

where $\tilde{T}$ refers to the TAP-correction to the mean field free energy. This belief optimization based on the TAP free energy approximation is similar to the mean-field free energy approach for which the marginal probability distributions are assumed to be independent. In addition, the TAP free energy approach can be considered as an approximation of the Bethe free energy approach up to the second order moment [287]. Due to its similarity to the mean-field energy approach and lack of accuracy compared to the Bethe free energy approach, this chapter focuses only on using the mean field and the Bethe free energy approaches and solving the corresponding constrained minimization problems.

11.4 BP/BO Approaches to Decentralized/Distributed FDD

This section develops decomposed methods to solve the optimizations in Section 11.3.2. In particular, methods of dual decomposition (see Appendix B.1) that solve the associated large-scale optimization are used for decentralized/distributed computations.

11.4.1 Decentralized FDD: Dual Decomposition

11.4.1.1 Minimization of The Mean Field Free Energy

Consider the constrained minimization (11.32). This large-scale optimization over a graphical model can be decomposed into separated constrained minimizations for which Agent i solves the optimization

$$\begin{aligned} \min_{\underline{\beta}_i,\, \{\underline{\beta}_j\}_{j\in\mathcal{N}(i)}} \quad & -\underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\psi}_i - \sum_{j\in\mathcal{N}(i)} \underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\underline{\psi}}_{ij}\, \underline{\beta}_j + \underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\beta}_i \\ \text{s.t.} \quad & \underline{\beta}_i \in \Delta, \\ & \underline{\beta}_j = \underline{\beta}_i,\ \forall j \in \mathcal{N}(i), \end{aligned} \tag{11.36}$$

where the second constraint corresponds to the consensus between the agents on edges connecting the node of Agent i and $\mathcal{N}(i)$ refers to the set of agents neighboring Agent i.

For fault detection and diagnosis with multiple hypotheses, assume that each agent has the same bank of hypothesized models and the objective of resultant distributed decision-making is to obtain optimal marginal beliefs $\{\underline{\beta}_i\}_{i\in V}$ that achieve consistency in localized estimations, i.e.,

$$\textbf{Marginal Belief Consensus I: } \beta_i(x) = \beta(x),\ \forall x \in \mathcal{X}_i,\ \forall i \in V, \tag{11.37}$$

which can be rewritten as

$$\underline{\beta}_i = \underline{\beta},\ \forall i \in V,\ \text{ for some } \underline{\beta} \in \Delta. \tag{11.38}$$

Incorporating the consensus requirement (11.38) into (11.36) results in a decomposed optimization for which Agent i solves

$$\begin{aligned} \min_{\underline{\beta}_i,\, \underline{\beta}} \quad & -\underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\psi}_i + \underline{\beta}_i{}^{\mathrm{T}} \underline{\underline{M}}_i \underline{\beta}_i + \underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\beta}_i \\ \text{s.t.} \quad & \underline{\beta}_i = \underline{\beta} \in \Delta, \end{aligned} \tag{11.39}$$

where $\underline{\underline{M}}_i \triangleq -\sum_{j\in\mathcal{N}(i)} \ln \underline{\underline{\psi}}_{ij}$ are nonnegative matrices since their entries correspond to compatibility functions or constraints and can be normalized to be in the interval $[0, 1]$ without deforming configuration of the free energy with respect to the beliefs. Notice that the $\underline{\beta}$ is a global variable that is required to be the same in all of the decomposed optimizations.

Case 1: [For $\underline{\underline{M}}_i \succeq 0$] If the pairwise compatibility matrix $\underline{\underline{M}}_i$ is positive semidefinite then the optimization (11.46) is convex and can be solved by using iterative dual decomposition methods, for which computations are decentralized for each Agent i and belief consensus is achieved by iterations to find an optimal Lagrange multipliers. For details of the use of dual decomposition methods and underlying theories, see Appendix B.1.

Case 2: [**For** $\underline{\underline{M_i}} \succeq_\Delta 0$] If the pairwise compatibility matrix $\underline{\underline{M_i}}$ is conditionally positive semidefinite over the standard simplex Δ then the optimization (11.46) is convex, but checking if $\underline{\underline{M_i}} \succeq_\Delta 0$ is NP-hard [189]. If a prior knowledge of $\underline{\underline{M_i}} \succeq_\Delta 0$ is available, then the same dual decomposition methods can be used as in Case 1. If there is no condition $\underline{\underline{M_i}} \succeq_\Delta 0$ a priori, then semidefinite programming relaxation can be used, in the same manner as described in Case 3 below.

Case 3: [**Indefinite** $\underline{\underline{M_i}}$] The optimization (11.46) can be rewritten as

$$\begin{aligned} \min_{\underline{\beta_i},\, \underline{\beta},\, B_i} \quad & -\underline{\beta_i}^{\mathrm{T}} \ln \underline{\psi_i} + \langle \underline{\underline{M_i}}, B_i \rangle + \underline{\beta_i}^{\mathrm{T}} \ln \underline{\beta_i} \\ \text{s.t.} \quad & \underline{\beta_i} = \underline{\beta} \in \Delta, \\ & \underline{\beta_i}\, \underline{\beta_i}^{\mathrm{T}} = B_i, \end{aligned} \tag{11.40}$$

where $\langle X, Y \rangle = \mathrm{Tr}(X^{\mathrm{T}} Y)$. Since for any $\beta_i \in \Delta$,

$$\underline{\beta_i}\, \underline{\beta_i}^{\mathrm{T}} = B_i \iff B_i e = \beta_i,\, B_i \succeq 0,\, \mathrm{rank}(B_i) = 1,\, e^{\mathrm{T}} B_i e = 1, \tag{11.41}$$

a convex relaxation of (11.40) can be

$$\begin{aligned} \min_{\underline{\beta_i},\, \underline{\beta},\, B_i,\, B} \quad & -\underline{\beta_i}^{\mathrm{T}} \ln \underline{\psi_i} + \langle \underline{\underline{M_i}}, B_i \rangle + \underline{\beta_i}^{\mathrm{T}} \ln \underline{\beta_i} \\ \text{s.t.} \quad & \underline{\beta_i} = \underline{\beta} \in \Delta, \\ & B_i e = \beta_i,\, e^{\mathrm{T}} B_i e = 1,\, B_i = B \succeq 0, \end{aligned} \tag{11.42}$$

where the rank constraint is not imposed and B is a variable that all agents share, i.e., it is a global variable that is required to be the same in the all of the decomposed optimizations. The optimization (11.42) provides a suboptimal solution for (11.40) and the corresponding suboptimal value is a lower bound on the optimal value of (11.40). The resultant optimization (11.42) is convex and can be efficiently solved to find suboptimal solutions $\underline{\beta_i}^\star = \underline{\beta}$ for all Agents $i \in V$, such as by using dual decomposition methods (see Appendix B.1).

11.4.1.2 Minimization of The Bethe Free Energy

Consider the constrained minimization (11.34). This large-scale optimization over a graphical model can be decomposed into separated constrained minimizations for which Agent i solves the optimization

$$\begin{aligned} \min_{\underline{\beta}_i,\, \underline{\underline{\beta}}_{ij}} \quad & -\underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\psi}_i - \sum_{j\in\mathcal{N}(i)} [\underline{\underline{\beta}}_{ij} \circ \ln \underline{\underline{\psi}}_{ij}] \\ & + (1-q_i)\underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\beta}_i + \sum_{j\in\mathcal{N}(i)} [\underline{\underline{\beta}}_{ij} \circ \ln \underline{\underline{\beta}}_{ij}] \\ \text{s.t.} \quad & \underline{\beta}_i \in \Delta, \\ & \underline{\underline{\beta}}_{ij} e = \underline{\beta}_i,\ j \in \mathcal{N}(i) \end{aligned} \tag{11.43}$$

where the second constraint corresponds to the marginal probability constraint for the agents on edges connecting the node of Agent i.

Similar to the mean-field energy approach, for fault detection and diagnosis with multiple hypotheses, assume that each agent has the same bank of hypothesized models and the objective of the resultant distributed decision-making is to obtain optimal marginal and pairwise marginal beliefs $(\{\underline{\beta}_k\}_{k\in V}, \{\underline{\underline{\beta}}_{ij}\}_{(i,j)\in E})$ that achieve the consistency in localized estimations, i.e.,

$$\begin{aligned} \textbf{Marginal Belief Consensus II: } & \beta_i(x) = \beta(x),\ \forall x \in \mathcal{X}_i,\ \forall i \in V \\ & \beta_{ij}(x,y) = b(x,y),\ \forall x \in \mathcal{X}_i,\ \forall y \in \mathcal{X}_j,\ \forall (i,j) \in E, \end{aligned} \tag{11.44}$$

which can be rewritten as

$$\begin{aligned} & \underline{\beta}_i = \underline{\beta},\ \forall i \in V, \ \text{ for some } \underline{\beta} \in \Delta \\ & \underline{\underline{\beta}}_{ij} = B,\ \forall (i,j) \in E \ \text{ for some } B \in \Omega \end{aligned} \tag{11.45}$$

where $\Omega \triangleq \{A \in \mathbb{R}_+^{n\times n} : Ae = p \text{ and } e^{\mathrm{T}}A = q \text{ for some } p, q \in \Delta\}$.

The use of Bayesian hypothesis tests for FDD needs special attention, for which the hypotheses at the nodes are homogeneous. The pairwise marginal distributions $\{\underline{\underline{\beta}}_{ij}\}_{(i,j)\in E}$ are required to satisfy the conditions $\beta_{ij}(x,y) = 0$ for all $x \neq y$ for all $(i,j) \in E$, which implies that the off-diagonal entries of $\underline{\underline{\beta}}_{ij}$ are zeros for all $(i,j) \in E$, or equivalently, the matrix B in (11.45) is a diagonal matrix.

Incorporating the consensus requirement (11.45) into (11.43) results in a decomposed optimization for which Agent i solves

$$\begin{aligned} \min_{\underline{\beta}_i,\, \underline{\beta}} \quad & -\underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\psi}_i - \underline{\beta}_i{}^{\mathrm{T}} \underline{a}_i + \underline{\beta}_i{}^{\mathrm{T}} \ln \underline{\beta}_i \\ \text{s.t.} \quad & \underline{\beta}_i = \underline{\beta} \in \Delta, \end{aligned} \tag{11.46}$$

where $\underline{a_i} \triangleq \sum_{j \in \mathcal{N}(i)} \ln(\mathrm{diag}[\underline{\underline{\psi_{ij}}}])$ and $\mathrm{diag}[A]$ denotes the vector whose elements are the diagonal entries of A in order. The resultant optimizations (11.46) are convex and can be efficiently solved to find global consensus optima $\underline{\beta_i}^\star = \underline{\beta}$ for all Agents $i \in V$, such as by using dual decomposition methods (see Appendix B.1).

11.5 Discussion

This section discusses several issues on the use of belief propagation for distributed statistical inference, and presents some open questions that are not fully answered in this chapter. The purpose of these discussions is to suggest future research directions for extensions and applications of BP/BO methods.

11.5.1 Unresolved Problems

For proper usage of belief propagation and optimization to tackle distributed statistical inference problems, some underlying assumptions of BP/BO methods need to be further investigated.

11.5.1.1 Correlated Measurements

Most of research works in the literature of belief propagation assume that each local measurement is conditionally independent given the other states at V (even given the states at its neighborhood). In other words, the likelihood functions have the relations

$$p(y_k|x_k, x_{-k}) = p(y_k|x_k), \quad \forall k \in V. \tag{11.47}$$

This assumption would be valid only for some special cases such as when the sensors are static (memoryless) and each source of uncertainty is localized. To see the role of this assumption of conditional independence in belief propagation, consider the next example of sensor fusion.

Example 11.1. Consider the Markov network model of sensor fusion depicted in Figure 11.1. Messages from Agents 2 and 3 to Agent 1 are computed by

$$\mu_{j\to 1}(\mathbf{x}_1) \propto \sum_{\mathbf{x}_j \in \mathcal{X}_j} p(\mathbf{x}_1|\mathbf{x}_j) p(\mathbf{x}_j|y_j), \text{ for } j = 1, 2, \tag{11.48}$$

where $\{y_j\}$ are the local measurements that are available to Agents j. Note that this is a marginalization and results in

$$\mu_{j\to 1}(\mathbf{x}_1) \propto p(\mathbf{x}_1|y_j), \text{ for } j = 1, 2, \tag{11.49}$$

with the resultant belief

$$\begin{aligned}\beta_1(\mathbf{x}_1) &\propto p(\mathbf{x}_1|y_1)\mu_{2\to 1}(\mathbf{x}_1)\mu_{3\to 1}(\mathbf{x}_1)\\ &\propto p(\mathbf{x}_1|y_1)p(\mathbf{x}_1|y_2)p(\mathbf{x}_1|y_3).\end{aligned} \tag{11.50}$$

Under the assumption of conditional independence (11.47), the belief can be rewritten as

$$\beta_1(\mathbf{x}_1) \propto p(\mathbf{x}_1|y_1, y_2, y_3) \tag{11.51}$$

that is, the marginal probability of the state of Agent 1 for given total measurements. The marginal probabilities of Agents 2 and 3 can be computed in similar ways, viz., $\beta_2(\mathbf{x}_2) \propto p(\mathbf{x}_2|y_1, y_2, y_3)$ and $\beta_3(\mathbf{x}_3) \propto p(\mathbf{x}_3|y_1, y_2, y_3)$.

In the previous example, notice that without assuming or guaranteeing conditional independence, the messages $\mu_{j\to i}(\mathbf{x}_i)$ for $i \neq j = 1, 2, 3$ result in the beliefs $\beta_i(\mathbf{x}_i) \propto \prod_{j=1}^{3} p(\mathbf{x}_i|y_j)$ that are not the same as the desired relations $\beta_i(\mathbf{x}_i) \propto p(\mathbf{x}_i|y_1, y_2, y_3)$.

Fortunately, for the case of homogeneous hypotheses in graphical models, the likelihood functions (11.47) have the relations

$$p(y_k|x_k, x_{-k}) = p(y_k|x_k)\prod_{j=1}^{n}\delta(x_k, x_j), \quad \forall k \in V, \tag{11.52}$$

where $\delta(x, y)$ refers to the standard scalar Dirac delta function. This fact implies that, for Example 11.1, the message-passing algorithms (11.48) achieve the correct beliefs (11.51) only if they satisfy the additional conditions of marginal belief consensus, viz., $\beta_1(\mathbf{x}) = \beta_2(\mathbf{x}) = \beta_3(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

11.5.1.2 Pre vs. Post Data Processing and Information Fusion

The primary goal of message-passing algorithms is to provide sufficient statistics for computations of marginal probabilities. In the context of belief propagation, sufficient statistics of messages are properties that ensure the relations

$$\beta_i\big(\mathbf{x}_i|y_i, \{\mu_{j\to i}\}_{j\in\mathcal{N}(i)}\big) = p\big(\mathbf{x}_i|Y = \{y_j\}_{j=1}^{n}\big), \ \forall \mathbf{x}_i \in \mathcal{X}_i, \ \forall i = 1, \cdots, n. \tag{11.53}$$

Message-passing algorithms can be considered as post data processing for information fusion, whereas transmitting raw data, not subject to any data processing, is a naive method for computations of marginal probabilities. Due to communication bandwidth limitations and cost of data storage, transmission of raw data is not practical nor efficient.

In belief propagation algorithms based on graphical networked models, reducing communication costs is of primary interest. Reducing the size of transmitting messages with guaranteed exactness of resultant

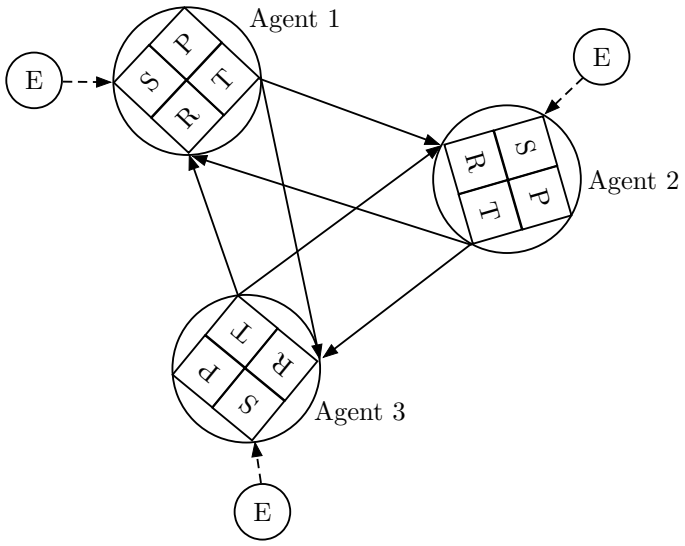


Figure 11.1: A schematic of a Markov network for sensor fusion. The solid arrows correspond to communication links and the dotted arrows correspond to measurement mechanism. S: Sensor, P: Processor, S: Receiver, T: Transmitter, and E: Evidence (or Observational Event).

statistical inference is to compute the smallest sufficient statistics.

11.5.1.3 Suboptimality of Consensus Algorithms

There was much research effort that studies convergence of message-passing algorithms in terms of properties of the graph $G = (V, E)$ (see [2, 182, 206], for example). Notice, however, that convergence does not imply optimality in general. Furthermore, such suboptimality can result in an arbitrarily bad decision whenever the estimation problem is connected to an optimal control problem, in which inaccurate belief can deviate the resultant decision from an optimal decision such that the achieved performance can be significantly worse off. In [206], the average-consensus algorithm and a belief propagation method are combined—such an algorithm was refereed to as *belief consensus*. This belief consensus has many benefits such as scalability and convergence under varying network topology. However, [206] did not provide any analysis of optimality and sub-optimality of their methods for distributed hypothesis tests. Notice that convergence or consensus of beliefs or messages does not necessarily imply optimality of the resultant hypothesis testing.

11.5.2 MAP Consensus

In Section 11.4, belief consensus constraints—conditions of (11.37) for the mean-field energy minimization and conditions of (11.44) for the Bethe free energy minimization—are incorporated into belief optimization to reach agreement in marginal and pairwise marginal probability distributions of multiple hypotheses for given total measurements.

A popular statistical inference problem is to find a state that is the most probable from a probability

distribution for given measurements. For graphical models of distributed hypothesis testing, such a state can be obtained from m-MAP or j-MAP estimation. Recall that an m-MAP estimator is a process to find state variables associated with the nodes in a graphical model such that the corresponding marginal posterior probabilities have maximum values for given total measurements. Similarly, but slightly differently, a j-MAP estimator is a process to find a configuration of state variables in a graphical model such that the corresponding joint posterior probability is a maximum for given total measurements. For this purpose of inference, the aforementioned max-product BP algorithms can be beneficial—using the max-product BP can reduce the communication costs, while the computational burdens of local processors would increase.

11.6 Summary and Future Work

This chapter has developed methods for distributed Bayesian hypothesis testing, particularly, for applications to distributed fault detection and diagnosis in large-scale networked systems. The presented methods are based on belief propagation and optimization and use graphical models to represent the systems of consideration. The resultant estimation problems reduce to the solution of distributed optimization for which the idea of belief optimization is adopted to use the concept of minimization of free energy to find an optimal probabilistic configuration of the state variables in Markov random fields. For distributed computations of the associated constrained minimization problems, dual decomposition methods are used, which provide benefits of scalability and convergence.

Several issues in the efficient and proper use of belief propagation and optimization for distributed statistical inference problems are discussed. Future research directions would be (a) to develop further generalization of belief optimization using the concepts of region-based free energy representations, which are extensions of pairwise potential energy descriptions, and (b) to evaluate exactness and compute approximation errors of an estimator that is obtained from minimizing an approximate free energy.

Chapter 12

CONCLUSIONS

This thesis has developed several model-based analysis, control, and estimation methods for both deterministic and stochastic uncertain dynamical systems. Part I extended and customized existing theoretical developments to delve further into the analysis and control of certain structured and characterized deterministic uncertain systems having additional properties that can be exploited. Part II employed spectral methods known as generalized polynomial chaos expansions for approximate quantification of stochastic uncertainties that are propagated through stochastic dynamical systems. In Part III, statistical inference problems in the basis of Bayesian theory were considered for the purpose of fault detection and diagnosis, for which multiple hypotheses of stochastic dynamical systems are assessed to find an optimally probable model of an event of faults, based on available observations and monitoring processes.

Chapter 2 presented a characterization of the solutions for copositive Lyapunov inequalities. It was shown that the extreme rays of solutions for copositive Lyapunov inequalities are indeed dyadic products of the co-state corresponding to Lagrangian dual variables that satisfy semi-algebraic conditions, which are polynomial-time verifiable under a mild assumption on the cone.

Chapter 3 considered uncertain linear descriptor systems of which unified and generalizable conditions for robust stability and performance were developed. The presented tests were coupled linear matrix inequalities and equalities that are computationally tractable via existing interior-point methods. Applicability of the full block S-procedure and its extensions to structured uncertain linear descriptor systems was supported.

Chapter 4 considered the reliability analysis of controlled systems with and without model uncertainties and provided necessary and sufficient conditions for robust fault-tolerant stability and performance under constant but unknown gain variation for uncertain systems that are affected by real parametric and complex dynamic uncertainties. The proposed conditions were represented in terms of the structured singular value whose upper- and lower-bounds can be computed in polynomial-time by using existing numerical methods. For illustration of the application of the proposed reliability conditions, numerical case studies for high-purity distillation column and parallel reactors with combined precooling were provided.

Chapter 5 presented a nonlinear internal model control scheme for stable Wiener systems that ensures robustness of closed-loop stability and performance to uncertainties in the inversion of the static nonlinearity. The proposed nonlinear IMC procedure was shown to be computationally tractable and applicable to stable Wiener systems with unstable zero dynamics, unmeasured states, disturbances, and measurement noise. The generalization of our approach to Hammerstein and Sandwich models is straightforward, and can be used to explicitly incorporate actuator constraints into the nonlinear controller design.

Chapter 6 provided a comprehensive overview of research related to the computational complexity of robustness margin calculations. This chapter collected together many results that are not well known in the literature, including that the cost of the structured singular value calculation scales by the rank of the nominal matrix, and that in worst case the widely used upper bound for the structured singular value can be arbitrarily far off. The chapter also represented approaches for the extension of past results, including randomized algorithms and polynomial chaos.

Chapter 7 considered generalized polynomial chaos (gPC) expansion approaches that can be used to approximate the functional dependence of dynamical system properties on uncertainties that are random variables for stochastic control problems as a means of replacing or facilitating MC simulation methods. This chapter presented stochastic optimal control problems by using gPC in which the cost function and probabilistic constraints can be reformulated as the constraints over the coefficients of gPC expansions.

Chapter 8 developed a new approach for stochastic model predictive control (MPC) problems in the presence of both parametric model uncertainty and exogenous stochastic disturbances. To approximate the solution of a stochastic differential equation and solve the corresponding stochastic MPC problem, gPC expansions were applied and constraints corresponding to the probability of safety/collision were imposed on the approximately predicted controlled trajectories, based on the model of a stochastic differential equation. It was also shown that concentration-of-measure inequalities combined with the Boole inequality can provide conservative probabilistic certificates for chance constraints of polyhedral inequalities, for which applications of the gPC expansions can be straightforward.

Chapter 9 presented a concise overview of Bayesian hypothesis testing, and discussed some open research directions for robust hypothesis tests and model-based real-time reliable optimal control methods that integrate estimation and control tools. Due to sensitivity of Bayesian hypothesis tests against lack of precise knowledge of the priors and proper choice of misdecision penalties, careful assessment of performance of the resultant statistical inference and decision is required.

Chapter 10 developed optimal active input design methods for fault detection and diagnosis (FDD) of Bayesian inference. The proposed design of optimal probing inputs was to maximize statistical discrimination between models of hypotheses corresponding to fault scenarios, while requiring the controlled state/output trajectories as well as the inputs remain within certain bounds. Different measures of the statistical discrimination between random processes or stochastic dynamical systems were considered. The first approach to design optimal probing inputs for FDD was to use a sequential SDP method to find a local optimum that

can be further improved by using multiple shooting or warm starts. The second approach was to use convex relaxation methods that compute approximate solutions for which the potential degree of sub-optimality is known in some special cases.

Chapter 11 developed methods for distributed Bayesian hypothesis testing in consideration of applications to distributed fault detection and diagnosis in large-scale networked systems. The methods were developed in the basis of belief propagation and optimization in which graphical models were used to represent the systems of consideration. For distributed computations of the proposed constrained minimization problems to find an optimum hypothesis or achieve belief consensus in probabilities of hypotheses, dual decomposition methods are used, which provide benefits of scalability and convergence.

As a concluding remark, the main research themes of this thesis were to exploit characteristics and properties of uncertainties in dynamical system models to construct computationally efficient analysis and control methods for the actual systems that are presumed to be in the set of uncertain models, without changing their fundamental physical nature.

Appendix A

Mathematical Backgrounds

A.1 Background on Computational Complexity Theory

This Appendix provides a compact introduction to computational complexity theory. Computational complexity theory allows a characterization of the inherent difficulty of calculating the solution for a problem under study. Problems (or equivalent versions of the same problem) are generally characterized as being in one of two classes: *P* or *NP-hard*. The class P refers to problems in which the exact time needed to solve the problem can always be bounded by a single function that is polynomial in the amount of data needed to define the problem. Such problems are said to be solvable in *polynomial time*. Although the exact consequences of a problem being *NP-hard* is still a fundamental open question in the theory of computational complexity, it is generally accepted that a problem being NP-hard means that its solution cannot be computed in polynomial time in the worst case. It is important to understand that being NP-hard is a property of the problem itself, not of any particular algorithm. It is also important to understand that having a problem be NP-hard does not imply that practical algorithms are not possible. Practical algorithms for NP-hard problems exist and typically involve approximation, heuristics, branch-and-bound, or local search.

Determining whether a problem is polynomial-time or NP-hard informs an engineer working on large-scale systems of the computational efficiency that can be expected by the best algorithms, and what kinds of algorithms to investigate for providing practical solutions to the problem.

A.1.1 Optimization Problems

An *instance* is defined to be all of the information needed to define a computational problem, whereas the size of the problem can be defined in a number of ways, such as the number of elements in a vector that contains all of the input data for the problem, or the number of rows in a matrix which contains most of the

data. Consider the optimization problem defined by

$$\sup_{x \in \mathcal{X}} c(x) \tag{A.1}$$

where each *instance* is a pair $(\mathcal{X}, c)$ and $\mathcal{X}$ is the set of feasible solutions and c is a cost function $c : \mathcal{X} \to \mathbb{R}$. Assume that $\mathcal{X}$ is non-empty and compact.

From this assumption and the Weierstrass theorem [161], the supremum is achieved by at least one $x \in \mathcal{X}$ so that the supremum can be replaced by the maximum.

A maximization problem can be written in one of the following versions [211, Chapter 15]:

P1. The optimization version: Find an optimal solution $x^\star \in \mathcal{X}$ such that $c(x) \leq c(x^\star)$ for all $x \in \mathcal{X}$.

P2. The evaluation version: Compute the optimal value $c^\star = c(x^\star)$ of an optimal solution.

P3. The recognition version: Given k, determine whether there is a feasible solution $x \in \mathcal{X}$ such that $c(x) \geq k$.

The recognition version is important for studying the complexity of an optimization problem, since it is the type of problem traditionally studied by the theory of computation [211, Ch. 15]. Unlike the first two versions, the recognition version is a question, which is answered by true (1) or false (0). Namely, difficulty of solving problems are ordered as P3 $\leq$ P2 $\leq$ P1.

A.1.1.1 The Classes P and NP

The definition of the recognition version of optimization problems allows the classification of the kinds of problems according to their computational complexity. The class P denotes the class of recognition problems that can be solved by a polynomial-time algorithm, i.e., given an instance, there is an efficient way for telling whether the answer is true or false. NP is a seemingly richer class of recognition problems and, for a problem to be in NP, it is not required that it can be answered in polynomial-time by an algorithm. It is only required that, for a yes instance of x, there exists a concise certificate for this x such that it can be checked in polynomial-time for validity [211, Chapter 15.3]. This condition is referred to as *polynomial-time verifiability* and this definition naturally implies that P is a subset of NP, i.e., P$\subset$NP, and it is believed that this relation is strict, i.e., P$\neq$NP. The complement of NP is denoted by *co-NP* and a co-NP problem would be related to the optimization problem to determine $c(x^*) < k$ for a given constant k. It can be only answered when the optimal value $c(x^*)$ is evaluated. This problem appears to be more difficult than verifying if the recognition version is in NP. For this reason, researchers believe that NP$\neq$co-NP. Figure A.1 shows the relations between complexity classes of recognition problems.

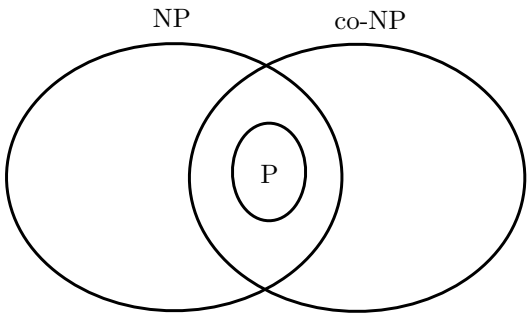


Figure A.1: Relations between complexity classes [211, Fig. 16.1].

A.1.1.2 Polynomial-time Reductions

It might be the case that solving an optimization problem becomes easy once an efficient algorithm is available for solving another problem.

Definition A.1. [211, Defn. 15.2] Let R_1 and R_2 be recognition problems. Then R_1 is said to *reduce in polynomial time* to R_2 if there exists a polynomial-time algorithm for R_1 that uses an algorithm for R_2 in a subroutine.

If R_1 reduces to R_2 and there is a polynomial-time algorithm for R_2 then there is a polynomial-time algorithm for R_1 too. Next consider the concept of polynomial-time transformation.

Definition A.2. [211, Defn. 15.3] A recognition problem R_1 *polynomially transforms* to another recognition problem R_2, if, for a given instance x, a new instance y for R_2 can be constructed within polynomial time (in the size of x) such that x yields a true instance of R_1 if and only if y gives a true instance of R_2.

A.1.1.3 NP-complete and NP-hard Problems

NP-complete problems are the hardest problems in NP. Examples of problems that are NP-complete are the traveling salesman problem, the max-cut problem, and the indefinite quadratic programming problem. Essentially everyone familiar with computational complexity believes that NP-complete problems are harder than P problems.[1] Any problem that is at least as hard as an NP-complete problem is said to be *NP-hard*. An NP-hard problem can refer to much broader classes of problems than recognition problems. In particular, if the recognition version of an optimization is NP-complete, then the corresponding evaluation and optimization versions are NP-hard.

A.1.2 Three Well-known Example Problems that are NP-complete or NP-hard

In order to establish NP-hardness of a problem $\mathcal{P}$, it suffices to demonstrate that a certain problem $\mathcal{P}_c$ that has been proven to be NP-complete can be reduced to the problem $\mathcal{P}$ in polynomial time. Some well-

[1] Although there is no formal proof.

known NP-hard and NP-complete problems are described that are useful in showing that robustness margin problems are NP-hard.

A.1.2.1 Indefinite QPs and QCQPs

This section defines the real quadratic program (QP), real quadratically constrained quadratic program (QCQP), and complex QCQP.

Problem A.1 (*QP*). Given a real symmetric matrix $A \in \mathbb{R}^{n\times n}$, a real vector $p \in \mathbb{R}^n$, a real scalar $c \in \mathbb{R}$, and real vectors b^{l} and b^{u} with $b^{\mathrm{l}} \leq b^{\mathrm{u}}$, determine whether there exists $x \in \mathbb{R}^n$ such that $b^{\mathrm{l}} \leq x \leq b^{\mathrm{u}}$ and $\left|x^{\mathrm{T}}Ax + 2p^{\mathrm{T}}x + c\right| \geq k$ for a fixed constant $k > 0$.

Problem A.2 (*QCQP*). Given real symmetric matrices $Q_i \in \mathbb{R}^{n\times n}$, complex vectors $q_i \in \mathbb{R}^n$, and complex scalars $d_i \in \mathbb{R}$ for $i = 0, \ldots, m$, determine whether there exists $x \in \mathbb{R}^n$ such that $b^{\mathrm{l}} \leq x^{\mathrm{T}}Q_i x + 2q_i^{\mathrm{T}}x + d_i \leq b^{\mathrm{u}}$ for all $i = 1, \ldots, n$, and $\left|x^{\mathrm{T}}Q_0 x + 2q_0^{\mathrm{T}}x + d_0\right| \geq k$ for a fixed constant $k > 0$.

Problem A.3 (*Complex QCQP*). Given complex hermitian matrices $Q_i \in \mathbb{C}^{n\times n}$, complex vectors $q_i \in \mathbb{C}^n$, and complex scalars $d_i \in \mathbb{C}$ for $i = 0, \ldots, m$, determine whether there exists $x \in \mathbb{C}^n$ such that $b^{\mathrm{l}} \leq |x^*Q_i x + q_i^*x + d_i| \leq b^{\mathrm{u}}$ for all $i = 1, \ldots, n$, and $|x^*Q_0 x + q_0^*x + d_0| \geq k$ for a fixed constant $k > 0$.

Remark A.1. No "complex QP" problem is defined since any constraint on the magnitude of a complex vector $x \in \mathbb{C}^n$ or a linear combination of its entries reduces to a quadratic constraint on x, resulting in a complex QCQP.

Remark A.2. Notice that Problem A.1 is not a special case of Problem A.3, since they have different fields, $\mathbb{R}$ and $\mathbb{C}$, respectively. The complexity of decision problems can be switched from NP to P or from P to NP, when the underlying field is changed.

Lemma A.1. The QP problem A.1 is NP-hard.

Proof. Consider an indefinite real quadratic program

$$q := \max_{0\leq x_i\leq 1} \left(\sum_{i=1}^{n} r_i x_i - r_0\right)^2 + \sum_{i=1}^{n} x_i(1 - x_i) \tag{A.2}$$

where $r_i \in \mathbb{Q}$ for $i = 0, \ldots, n$. Murty and Kabadi [189, Lemma 1] show that the recognition problem "$q \geq k$" for a fixed constant k is NP-hard. Since this problem is a special case of Problem A.1, Problem A.1 is NP-hard. QED

Lemma A.2. The QCQP problem A.2 is NP-hard.

Proof. Consider an indefinite real quadratic program

$$q := \max_{0 \le x_i \le 1} \left(\sum_{i=1}^{n} r_i x_i - r_0 \right)^2 \quad \text{subject to } x_i(1 - x_i) = 0,\ i = 1, \ldots, n \tag{A.3}$$

where $r_i \in \mathbb{Q}$ for $i = 0, \ldots, n$. This problem is indeed NP-hard, since it is polynomially equivalent to (A.2) [211]. The quadratic equality constraint $x_i(1 - x_i) = 0$ is equivalent to two quadratic inequalities $x_i(1 - x_i) \le 0$ and $x_i(1 - x_i) \ge 0$. Therefore, this NP-hard problem can be reformulated as a QCQP in polynomial time, which implies that the real QCQP problem A.2 is also NP-hard [36, 211]. QED

Lemma A.3. The complex QCQP problem A.3 is NP-hard.

Proof. Consider an indefinite complex quadratic program

$$q := \max_{|x_i| \le 1, x_i \in \mathbb{C}} |x^* A x| \tag{A.4}$$

where $A \in \mathbb{C}^{n \times n}$. Toker and Özbay [267, Lemma A.3] show that the knapsack problem, which is NP-hard, can be written as problem (A.4). This observation implies that the recognition problem "$q \ge k$" for a fixed constant k is NP-hard, which implies that the more general complex QCQP in Problem A.3 is also NP-hard. QED

A.1.2.2 Knapsack Problems

Define the following two knapsack problems.[2]

Problem A.4 (*Knapsack 1*). Given a positive integer n and a rational positive vector $a \in \mathbb{R}^n$ with $\|a\|_2 \le 0.1$, determine whether the equation $\sum_{i=1}^{n} a_i x_i = 0$ has a solution with $x \in \{-1, 1\}^n$.

Problem A.5 (*Knapsack 2*). Given a positive integer n and a positive integer vector $a \in \mathbb{Z}^n$, determine whether the equation $\sum_{i=1}^{n} a_i x_i = 0$ has a solution $x \in \{-1, 1\}^n$.

Lemma A.4. Problems A.4 and A.5 are NP-complete.

Proof. See [93, Sec. 3.2] for proofs. QED

A.1.2.3 Max-cut Problem

Another well-known NP-complete problem is the *max-cut problem*. Let $G = (V, E)$ be a graph with vertices V and edges E, and $w : E \to \mathbb{R}$ be a weight function defined on the edges of the graph. For a given subset

[2] These problems are also known as *subset sum problems*.

of vertices, $N \subset V$, the *maximum cut* in the graph with respect to the weight function w is defined as $\mathrm{MC}(G) = \max_{S \subset N} w(\delta S)$, where δS is the set of edges with one vertex in S and the other vertex in $N \backslash S$, and $w(D) := \sum_{d \in D} w(d)$ for a subset $D \subset E$.

Problem A.6 (*Max-cut*). Given a graph $G = (V, E)$ and a weight function $w : E \to \mathbb{R}$, compute $\mathrm{MC}(G)$.

The max-cut problem is known to be NP-hard [93].

A.2 Generalized Copositive Programs in Control and Systems Theory

This appendix collects mathematical results associated with a promising research direction on the use of copositive programming in systems and control theory. In particular, we focus on mathematical programs related to tests of conditional definiteness of given matrices. Computational challenges of checking cone-positive definiteness motivates development of convex relaxation in the basis of copositive programming. However, the associated copositive programs still remain to be NP-hard and polynomial-time computable semidefinite programming (SDP) relaxation can be considered as an efficient certificate, while the resultant conservatism needs to be further investigated.

A.2.1 Generalized Copositive Programming

Copositive programming is a class of conic programming that generalizes linear programming and semidefinite programming, A copositive program is optimization over the cone of the so-called copositive matrices. Similar to semidefinite programming, copositive programming has particular importance in application to combinatorial and quadratic optimization (see [38, 74, 194], for example).

The set of copositive matrices defines a cone.

Definition A.3 (*Cone of copositive matrices*). The cone of copositive matrices is defined as the set of matrices whose quadratic form takes nonnegative values on the positive orthant $\mathbb{R}^n_+$, viz.,

$$\mathcal{K}(\mathbb{R}^n_+) \triangleq \left\{ A \in \mathbb{S}^{n \times n} : x^{\mathrm{T}} A x \geq 0, \forall x \in \mathbb{R}^n_+ \right\}. \tag{A.5}$$

The dual cone of $\mathcal{K}(\mathbb{R}^n_+)$ is obtained by

$$\mathcal{K}(\mathbb{R}^n_+)^\bullet = \mathrm{Conv}\left\{ x x^{\mathrm{T}} : x \in \mathbb{R}^n_+ \right\}. \tag{A.6}$$

Let define the cone of positive semidefinite matrices and the cone of nonnegative matrices in $\mathbb{S}^{n \times n}$ as $\mathcal{S}_+$ and $\mathcal{P}_+$, respectively. It is straightforward to see that these two cones are self-dual, i.e., $\mathcal{S}^\bullet_+ = \mathcal{S}_+$ and $\mathcal{P}^\bullet_+ = \mathcal{P}_+$, whereas the cone $\mathcal{K}(\mathbb{R}^n_+)$ is not self-dual, but indeed satisfies the relation $\mathcal{K}(\mathbb{R}^n_+)^\bullet \subset \mathcal{K}(\mathbb{R}^n_+)$. In

particular, we have the relations

$$\mathcal{K}(\mathbb{R}^n_+)^\bullet \subseteq (\mathcal{S}_+ \cap \mathcal{P}_+) \quad \text{and} \quad \mathcal{S}_+ + \mathcal{P}_+ \subseteq \mathcal{K}(\mathbb{R}^n_+) \tag{A.7}$$

where the subset relations are known to be strict for $n \geq 5$, while we have the equality in the above relations for $n \leq 4$ [173].

Definition of the aforementioned cone of copositive matrices can be generalized to define a cone of matrices whose quadratic form takes nonnegative values on the cone $\mathcal{C} \subset \mathbb{R}^n$, namely,

$$\mathcal{K}(\mathcal{C}) \triangleq \{A \in \mathbb{S}^{n\times n} : x^{\mathrm{T}} A x \geq 0, \forall x \in \mathcal{C}\} \tag{A.8}$$

and its dual is similarly obtained as

$$\mathcal{K}(\mathcal{C})^\bullet = \mathrm{Conv}\{xx^{\mathrm{T}} : x \in \mathcal{C}\}. \tag{A.9}$$

A cone program is a linear optimization problem in matrix variables of the standard form

$$\begin{aligned} &\min \ \langle C, X\rangle \\ &\text{s.t. } \langle A_i, X\rangle = b_i, \ i = 1, \ldots, m, \\ &\quad X \in \mathcal{K} \end{aligned} \tag{A.10}$$

where $\mathcal{K}$ refers to a cone. With regard to the types of the cone $\mathcal{K}$, we have the hierarchy relations

$$\text{LP} \subset \text{SOCP} \subset \text{SDP} \subset \text{COP} \subset \text{g-COP} \subset \text{CP} \tag{A.11}$$

where classifications of cone programming follow

$$\begin{aligned} \text{LP: (A.10) with } \mathcal{K} &= \mathcal{P}_+, \\ \text{SOCP: (A.10) with } \mathcal{K} &= \mathcal{S}_2, \\ \text{SDP: (A.10) with } \mathcal{K} &= \mathcal{S}_+, \\ \text{COP: (A.10) with } \mathcal{K} &= \mathcal{K}(\mathbb{R}^n_+), \\ \text{g-COP: (A.10) with } \mathcal{K} &= \mathcal{K}(\mathcal{C}), \\ \text{CP: (A.10) with } \mathcal{K} & \end{aligned} \tag{A.12}$$

with $\mathcal{S}_p \triangleq \{(x, r) \in \mathbb{R}^{n-1} \times \mathbb{R} : \|x\|_p \leq r\}$ known as the Lorentz cone or ice cream cone (for $p = 2$) and the cone $\mathcal{K}$ considered in CP can be any one of types of aforementioned cones or, more generally, a Cartesian product of aforementioned cones.

The dual program of a cone program that can be straightforwardly obtained from the Lagrangian approach is a cone program having the form

$$\begin{aligned} &\min \sum_{i=1}^{m} b_i y_i \\ &\text{s.t. } C - \sum_{i=1}^{m} y_i A_i \in \mathcal{K}^\bullet \end{aligned} \tag{A.13}$$

where $\mathcal{K}^\bullet$ refers to the dual cone of $\mathcal{K}$.

With regard to the use of interior-point methods to solve the primal (A.10) and dual (A.13) optimization of cone programming, the associated optimization is polynomial-time solvable only if there exists a self-concordant barrier function, that can be evaluated in polynomial-time, for the cones $\mathcal{K}$ and $\mathcal{K}^\bullet$ (see [197] for further details of polynomial-time interior-point methods for convex optimization).

A.2.1.1 Tests for Conditionally Definite Matrices

This subsection presents several results related with conditional definite matrices. The readers are further referred to [118] for an extensive survey on the results related with conditionally definite matrices and [129, 130] for tests of copositiveness of symmetric matrices.

Definiteness on a Subspace Consider a subspace $\mathcal{S}(B) \subset \mathbb{R}^n$ defined by

$$\mathcal{S}(B) \triangleq \{x \in \mathbb{R}^n : Bx = 0\} \tag{A.14}$$

where $B \in \mathbb{R}^{m \times n}$ with $m < n$ and $\text{Rank}(B) = m$.

Definition A.4. A matrix $P \in \mathbb{S}^{n \times n}$ is said to be $\mathcal{S}(B)$-positive definite and denoted by $P \succeq_{\mathcal{S}(B)} 0$ if

$$\langle x, Px \rangle > 0, \ \forall x \in \mathcal{S}(B) \setminus \{0\}. \tag{A.15}$$

Similarly, if

$$\langle x, Px \rangle \geq 0, \ \forall x \in \mathcal{S}(B) \tag{A.16}$$

then $P \in \mathbb{S}^{n \times n}$ is said to be $\mathcal{S}(B)$-positive semidefinite.

Lemma A.5 (*Finsler's Lemma [19]*). Consider a symmetric matrix $P \in \mathbb{S}^{n \times n}$ and a subspace $\mathcal{S}(B) \in \mathbb{R}^n$ defined above. Then the following statements are equivalent:

i) The matrix P is $\mathcal{S}(B)$-positive definite, i.e., $\langle x, Px \rangle > 0$ for all $x \neq 0$ such that $Bx = 0$.

ii) $(B^\perp)' P B^\perp \succ 0$ where $BB^\perp = 0$.

iii) There exists a constant $\rho_0 > 0$ such that $P + \rho B^{\mathrm{T}} B \succ 0$ for all $\rho \geq \rho_0$.

iv) There exists a matrix $X \in \mathbb{R}^{n \times m}$ such that $P + XB + B^{\mathrm{T}} X^{\mathrm{T}} \succ 0$.

Definiteness on an Affine Set Consider an affine set $\mathcal{A}(B, b) \subset \mathbb{R}^n$ defined by

$$\mathcal{A}(B, b) \triangleq \{x \in \mathbb{R}^n : Bx = b\} \tag{A.17}$$

where $B \in \mathbb{R}^{m \times n}$ with $m < n$ and $\mathrm{Rank}(B) = m$, and $b \in \mathbb{R}^m$.[3] Suppose that the affine set $\mathcal{A}(B, b)$ is nonempty.[4] Then there exists $x_0 \in \mathbb{R}^n$ such that $Bx_0 = b$.

Lemma A.6. Consider a symmetric matrix $P \in \mathbb{S}^{n \times n}$ and an affine set $\mathcal{A}(B, b) \in \mathbb{R}^n$ defined above. The matrix P is $\mathcal{A}(B, b)$-positive definite, i.e., $\langle x, Px \rangle > 0$ for all x such that $Bx = b$, if and only if

$$\begin{bmatrix} \langle x_0 | Px_0 \rangle & \langle Px_0 | \\ | Px_0 \rangle & P \end{bmatrix} \succ_{\mathcal{S}(\bar{B})} 0, \quad \bar{B} \triangleq \begin{bmatrix} 0 & B \end{bmatrix}, \tag{A.18}$$

where $x_0 \in \mathcal{A}(B, b)$ can be an arbitrary finite vector and the bra-ket notation is used to refer to the inner product of two vectors for convenience.

Remark A.3. For a given symmetric matrix $P \in \mathbb{S}^{n \times n}$, the tests for $\mathcal{S}(B)$ and $\mathcal{A}(B, b)$ definiteness are polynomial-time solvable, which follows from Finsler's lemma and the existence of polynomial-time evaluable self-concordant barrier function for the positive semidefinite and definite cones of symmetric matrices.

Definiteness on a Positive Orthant Motzkin [186] introduced the concept of a copositive matrix (see Definition A.3). For a copositive matrix $P \in \mathbb{S}^{n \times n}$ denoted as $P \in \mathcal{K}(\mathbb{R}^n_+)$, use the notation $P \succeq_{\mathbb{R}^n_+} 0$. For abuse of notation, a copositive matrix P is said to be $\mathbb{R}^n_+$-positive semidefinite.

Testing whether a given matrix P is not a copositive matrix involves the decision problem with

$$\begin{aligned} &\textbf{Instance} : P \in \mathbb{S}^{n \times n}, \\ &\textbf{Question} : \text{Is there a } x \geq 0 \text{ such that } \langle x, Px \rangle < 0? \end{aligned} \tag{A.19}$$

This decision problem can be naturally answered from the optimal value of the quadratic program

$$\begin{aligned} &\min \ \langle x, Px \rangle \\ &\text{s.t. } x \geq 0. \end{aligned} \tag{A.20}$$

This QP is known to be NP-hard [189, Thm. 1] and a test for copositive matrices is indeed NP-complete [189, Thm. 3].

[3] Every affine set can be represented in this way for properly chosen–may not be unique–parameters B and b [221, Thm. 1.4].

[4] Note that the affine set $\mathcal{A}(B, b)$ is nonempty if and only if $b \in \mathrm{Image}(B)$.

Definiteness on a Polyhedral Cone Consider a nonempty polyhedral cone $\mathcal{C}(B) \subset \mathbb{R}^n$ defined by

$$\mathcal{C}(B) \triangleq \{x \in \mathbb{R}^n : Bx \geq 0\} \tag{A.21}$$

where $B \in \mathbb{R}^{m\times n}$. A matrix $P \in \mathbb{S}^{n\times n}$ is said to be $\mathcal{C}(B)$-positive semidefinite if $P \in \mathcal{K}(\mathcal{C}(B))$ (see (A.8) for the definition of $\mathcal{K}(\mathcal{C}(B))$) and denoted by $P \succeq_{\mathcal{K}(\mathcal{C}(B))} 0$.

Testing whether a given matrix P is not a copositive matrix involves the decision problem with

$$\begin{aligned} &\textbf{Instance} : P \in \mathbb{S}^{n\times n},\, B \in \mathbb{R}^{m\times n}, \\ &\textbf{Question} : \text{Is there a } x \in \mathbb{R}^n \text{ satisfying } Bx \geq 0 \text{ such that } \langle x, Px\rangle < 0? \end{aligned} \tag{A.22}$$

This decision problem can be naturally answered from the optimal value of the quadratic program

$$\begin{aligned} &\min\ \langle x, Px\rangle \\ &\text{s.t. } Bx \geq 0. \end{aligned} \tag{A.23}$$

This QP is known to be NP-hard [189, Thm. 2] and a test for $\mathcal{C}(B)$-positive semidefinite matrices is indeed co-NP-complete [189, Thm. 4].

Definiteness on a Cone Consider a nonempty cone $\mathcal{C} \subset \mathbb{R}^n$. A matrix $P \in \mathbb{S}^{n\times n}$ is said to be $\mathcal{C}$-positive semidefinite if $P \in \mathcal{K}(\mathcal{C})$ (see (A.8) for the definition of $\mathcal{K}(\mathcal{C})$) and denoted by $P \succeq_{\mathcal{K}(\mathcal{C})} 0$.

Testing whether a given matrix P is not a copositive matrix involves the decision problem with

$$\begin{aligned} &\textbf{Instance} : P \in \mathbb{S}^{n\times n},\, \mathcal{C} \subset \mathbb{R}^n, \\ &\textbf{Question} : \text{Is there a } x \in \mathcal{C} \text{ such that } \langle x, Px\rangle < 0? \end{aligned} \tag{A.24}$$

This decision problem can be naturally answered from the optimal value of the quadratic program

$$\begin{aligned} &\min\ \langle x, Px\rangle \\ &\text{s.t. } x \in \mathcal{C}. \end{aligned} \tag{A.25}$$

This QP is NP-hard, since it includes the QPs (A.20) and (A.23) as special cases. Furthermore, it is straightforward that the QP (A.25) is equivalent[5] to the completely cone positive program (CCPP)

$$\begin{aligned} &\min\ \langle P, X\rangle \\ &\text{s.t. } X \in \mathcal{K}(\mathcal{C})^\bullet, \end{aligned} \tag{A.26}$$

for which the optimal value is zero if and only if $P \in \mathcal{K}(\mathcal{C})$ that directly follows from the definition of the

[5]Here, equivalence between two programs implies that their optimal solutions or optimal values have a one-to-one relation.

dual cone. Note that a minimum of the above optimization (A.26) is achieved at an extreme point of the convex cone $\mathcal{K}(\mathcal{C})^\bullet$, i.e., an optimal solution is a rank-one matrix $X^\star = xx^{\mathrm{T}}$ with some $x \in \mathcal{C}$. Therefore, the optimization (A.26) is an *exact* convex relaxation of (A.25).

A.2.1.2 Copositive Programs as Convex Relaxations

Copositive Programming Relaxation for Standard QP Consider a quadratic program whose supporting cone is the positive orthant:

$$\begin{aligned} &\min\ \langle x, Cx\rangle \\ &\text{s.t. } e'x = 1,\ x \geq 0, \end{aligned} \tag{A.27}$$

where e denotes the all-ones vector in $\mathbb{R}^n$. We refer to the optimization of the form (A.27) as *pos-QP*. This QP can be equivalently rewritten as

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } \langle E, X\rangle = 1,\ X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ \mathrm{Rank}(X) = 1, \end{aligned} \tag{A.28}$$

where $E = ee^{\mathrm{T}}$ is the all-ones rank-one matrix in $\mathbb{R}^{n\times n}$. It is not hard to see that the rank constraint does not change the optimal value nor the optimal solution of (A.28). In other words, the convex relaxation (A.28) without the rank constraint is exact. The resultant convex relaxation is still a hard problem, but can be further relaxed from replacing the completely positive cone by its superset $\mathcal{S}_+ \cap \mathcal{P}_+$, namely,

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } \langle E, X\rangle = 1,\ X \in \mathcal{S}_+ \cap \mathcal{P}_+. \end{aligned} \tag{A.29}$$

Since the constraint $X \in \mathcal{P}_+$ can be rewritten as the constraints $\mathcal{L}_k(X) \in \mathcal{S}_+$ with linear operators $\mathcal{L}_k : \mathbb{S}^{n\times n} \to \mathbb{S}^{n\times n}$ for $k = 1, \ldots, \frac{n(n+1)}{2}$, the optimization (A.29) is a semidefinite program and can be solved in polynomial-time by using interior-point methods. Alternatively, but equivalently, one can rewrite (A.29) by

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } \langle E, X\rangle = 1, \\ &\quad \langle E_k, X\rangle = z_k,\ k = 1, \ldots, m, \\ &\quad X \succeq 0,\ z \geq 0, \end{aligned} \tag{A.30}$$

where $E_k = e_i e_j^{\mathrm{T}}$ with the standard basis vectors e_i in $\mathbb{R}^n$ and $i \leq j$ such that there exists an injective mapping from a pair (i, j) to an index k, and $m = \frac{n(n+1)}{2}$.

For more general QP, consider the optimization

$$\begin{aligned} &\min\ \langle x, Cx\rangle \\ &\text{s.t. } b - Ax \geq 0,\ x \geq 0, \end{aligned} \tag{A.31}$$

which can be equivalently rewritten as

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } bb^{\mathrm{T}} - Axb^{\mathrm{T}} - bx^{\mathrm{T}}A^{\mathrm{T}} + AXA^{\mathrm{T}} \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ X = xx^{\mathrm{T}}. \end{aligned} \tag{A.32}$$

A natural convex relaxation can be

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } bb^{\mathrm{T}} - Axb^{\mathrm{T}} - bx^{\mathrm{T}}A^{\mathrm{T}} + AXA^{\mathrm{T}} \in \mathcal{K}(\mathbb{R}^n_+)^\bullet, \\ &\quad X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ X - xx^{\mathrm{T}} \in \mathcal{K}(\mathbb{R}^n_+), \end{aligned} \tag{A.33}$$

and an associated SDP relaxation can be

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } bb^{\mathrm{T}} - Axb^{\mathrm{T}} - bx^{\mathrm{T}}A^{\mathrm{T}} + AXA^{\mathrm{T}} \geq 0, \\ &\quad bb^{\mathrm{T}} - Axb^{\mathrm{T}} - bx^{\mathrm{T}}A^{\mathrm{T}} + AXA^{\mathrm{T}} \succeq 0, \\ &\quad \begin{bmatrix} X & x \\ x^{\mathrm{T}} & 1 \end{bmatrix} \succeq 0,\ X \geq 0,\ x \geq 0. \end{aligned} \tag{A.34}$$

Copositive Programming Relaxation for Standard QCQP Consider a quadratically constrained quadratic program (QCQP) whose supporting cone is the positive orthant:

$$\begin{aligned} &\min\ \langle x, Cx\rangle \\ &\text{s.t. } \langle x, Ax\rangle = 1,\ x \geq 0, \end{aligned} \tag{A.35}$$

where $A \in \mathbb{R}^{n\times n}$. This QCQP can be equivalently rewritten as

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } \langle A, X\rangle = 1,\ X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ \mathrm{Rank}(X) = 1. \end{aligned} \tag{A.36}$$

Lemma A.7. For the QCQP of the form (A.35), the following convex relaxation

$$\begin{aligned} &\min\ \langle C, X\rangle \\ &\text{s.t. } \langle A, X\rangle = 1,\ X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet \end{aligned} \tag{A.37}$$

is exact.

Proof. It is not hard to see that there must be an optimal solution $X^\star$ that is an extreme ray of the closed convex set $\{X \in \mathbb{S}^{n\times n} : \langle A, X\rangle = 1,\, X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet\}$–it can be proved by contradiction–such that $\mathrm{Rank}(X^\star) = 1$. To show this rigorously, suppose that $X^\star$ can be decomposed as $X^\star = \lambda\xi\xi^{\mathrm{T}} + (1-\lambda)Y$ for some $\lambda \in [0,1]$ where $\xi \in \mathbb{R}^n_+$ satisfies $\langle \xi, A\xi\rangle = 1$, $Y \in \mathcal{K}(\mathbb{R}^n_+)^\bullet$ satisfies $\langle A, Y\rangle = 1$, and $Y - \rho\xi\xi^{\mathrm{T}} \neq 0$ for any $\rho \geq 0$. Then $\langle C, X^\star\rangle = \langle C, \lambda\xi\xi^{\mathrm{T}} + (1-\lambda)Y\rangle = \lambda\langle \xi, C\xi\rangle + (1-\lambda)\langle C, Y\rangle \geq \min\{\langle \xi, C\xi\rangle, \langle C, Y\rangle\} \geq \mathrm{OPT(A.35)}$ where $\mathrm{OPT}(\cdot)$ refers to the optimal value of the associated optimization. Due to the relaxation nature, $\mathrm{OPT(A.35)} \equiv \mathrm{OPT(A.36)} \geq \mathrm{OPT(A.37)}$. Thus, we have $\mathrm{OPT(A.35)} \equiv \mathrm{OPT(A.37)}$. QED

For more general QCQP, consider the optimization

$$\begin{aligned} \min\ & \langle x, Cx\rangle \\ \text{s.t. } & \langle x, A_i x\rangle \leq b_i, \ , i = 1, \ldots, m, \\ & x \geq 0, \end{aligned} \tag{A.38}$$

which can be equivalently rewritten as

$$\begin{aligned} \min\ & \langle C, X\rangle \\ \text{s.t. } & \langle A_i, X\rangle \leq b_i,\ i = 1, \ldots, m, \\ & X \in \mathcal{K}(\mathbb{R}^n_+)^\bullet,\ \mathrm{Rank}(X) = 1. \end{aligned} \tag{A.39}$$

A natural SDP relaxation can be

$$\begin{aligned} \min\ & \langle C, X\rangle \\ \text{s.t. } & \langle A_i, X\rangle \leq b_i,\ i = 1, \ldots, m, \\ & X \geq 0,\ X \succeq 0. \end{aligned} \tag{A.40}$$

A.3 Background on Spectral Methods for Uncertainty Quantification

This Appendix provides a review for fundamentals of stochastic spectral methods using polynomial expansions for the system states or outputs with random system parameters and inputs. To do this, we start from representing important characteristics of certain polynomials that are used for spectral methods.

A.3.1 Orthogonal Polynomials

Polynomial approximations are almost always used when implementing functions on a computing system and the basic assumption on this discipline is that a finite sum of polynomials can accurately approximate

a function of interest. For polynomial approximations, orthogonal polynomials play a crucial role and we briefly review their properties.

A.3.1.1 Orthogonality

Consider a measure space $(\mathcal{X}, \mathcal{M}, \mu)$ where $\mathcal{X}$ is a nonempty set equipped with a σ-algebra $\mathcal{M}$ and a measure μ. A set or orthogonal polynomials $\{\phi_n(s)\}$ for $x \in \mathcal{M}$ is defined by their orthonormality relation

$$\langle \phi_n, \phi_m \rangle \triangleq \int_{\mathcal{X}} \phi_n(x)\phi_m(x)d\mu(x) = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{otherwise.} \end{cases} \tag{A.41}$$

We might use a short notation for this relation: $\langle \phi_n, \phi_m \rangle = \delta_{nm}$ where δ_{nm} is the Kronecker delta function. For each family of orthogonal polynomials there is a corresponding integration rules with different measures. Table A.1 shows several common orthogonal polynomials and their measures.

A.3.1.2 Recurrence Relation

It is know that any set of orthogonal polynomials $\{\phi_n(s)\}$ on the real iine satisfies a three-term recurrence formula

$$x\phi_n(x) = a_{n+1}\phi_{n+1}(x) + b_n\phi_n(x) + a_n\phi_{n-1}(x) \tag{A.42}$$

for $n = 0, 1, \ldots,$. Along with $\phi_{-1}(x) = 0$, this formula holds consistently and ϕ_0 is always a constant. This recurrence formula can be also represented by a matrix equation

$$x \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{p-2}(x) \\ \phi_{p-1}(x) \end{pmatrix} = \begin{pmatrix} b_0 & a_1 & & & \\ a_1 & b_1 & a_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{p-2} & b_{p-2} & a_{p-1} \\ & & & a_{p-1} & b_{p-1} \end{pmatrix} \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{p-2}(x) \\ \phi_{p-1}(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ a_p\phi_p(x) \end{pmatrix}. \tag{A.43}$$

Cooperating the recurrence formula (A.42) or (A.43) with numerically stable algorithms we can produce a set of orthogonal polynomials.

Polynomial	Support	Measure ($\mu(x)$)
Legendre	$(-1, 1)$	1
Laguerre	$(0, \infty)$	$x^\alpha e^{-x}$
Hermite	$(-\infty, \infty)$	e^{-x^2}
Chebyshev	$(-1, 1)$	$(1 - x^2)^{-1/2}$
Jacobi	$(-1, 1)$	$(1 - x)^\alpha(1 + x)^\beta$

Table A.1: Supports and measures of common orthogonal polynomials.

A.3.2 Parameterization of Random Inputs

In this section, we introduce how to properly characterize a stochastic model to study uncertainty propagation in an input-output static and dynamic system models, where inputs include system parameters and external inputs to the system. For any analysis of a stochastic system model based on simulations and/or experiments, a critical step is to specify and characterize random inputs appropriately. To reduce an infinite-dimensional probability space to a finite-dimensional space of random inputs, we parameterize the probability space by a set of random variables that might be required to be mutually independent for accuracy and convenience of analysis.

A.3.2.1 Random Variables

For a state-space parameterized system model, the parameterization of the probability space $(\Omega, \mathcal{F}, P)$ is straightforward. Consider the concatenated system parameter vector $\theta : \Omega \to \Theta \subseteq \mathbb{R}^{n_\theta}$ that is a random variable defined on the events Ω, where the set Θ is assumed to be also known and the true system parameter θ^* that is a realization of a random variable θ is supposed to be in the set. We also suppose that the statistics of the random variable θ is known, i.e., the joint probability distribution of θ is given. For a given probability distribution of random parameter of a system, the first step of analysis using polynomial chaos is to transform the parameters to a set of independent random variables and normalize them. We might call such transformed random variables *standard random variables* [120]. Our objective is to find a diffeomorphism $T : \Xi \to \Theta$ such that $\theta = T(\zeta)$ for $\zeta \in \Xi$ and the resulting state/output variables of a stochastic model x have equivalent representations $x(z, t; \theta(\omega)) = x(z, t; \zeta(\omega))$.

Gaussian parameters Suppose that the system parameter vector θ is Gaussian: $\theta \sim \mathcal{N}(b, C)$. Then an affine transformation $T(\zeta) \triangleq A\zeta + b$ of the standard random variable $\zeta \sim \mathcal{N}(0, \mathrm{I})$, where $A \in \mathbb{R}^{n_\theta \times n_\theta}$ satisfies $AA^{\mathrm{T}} = C$, yields a reparameterization of θ in terms of the standard random variable ζ.

Non-Gaussian parameters Rosenblatt [223] suggested a simple transformation of an absolute continuous random variable $\theta \in \Omega \to \mathbb{R}^{n_\theta}$ into the uniform distribution on the n_θ-dimensional hypercube $[0, 1]^{n_\theta}$. Furthermore, in some cases the uncertainty in a system input θ is specified by an empirical cumulative distribution function (cdf). Let $\theta = (\theta_1, \cdots, \theta_{n_\theta})$ be a random vector with cdf $F_\theta(\theta^1, \cdots, \theta^{n_\theta})$. Define a transformation $T : \Theta \to [0, 1]^{n_\theta}$ by

$$\begin{aligned} \zeta_1 &= P[\theta_1 \le \theta^1] = F_1(\theta^1), \\ &\vdots \\ \zeta_{n_\theta} &= P[\theta_{n_\theta} \le \theta^{n_\theta} | \theta_1 = \theta^1, \cdots, \theta_{n_\theta - 1} = \theta^{n_\theta - 1}] = F_{n_\theta}(\theta^{n_\theta} | \theta^1, \cdots, \theta^{n_\theta - 1}). \end{aligned} \tag{A.44}$$

Then, $\zeta_i \sim \mathcal{U}(0, 1)$ for all i and $\{\zeta_i\}$ are independent.

Distribution Type of $\theta \in \Theta$	Transformation: $T : \Xi \to \Theta$	Transformation: $T^{-1} : \Theta \to \Xi$
Uniform (a, b)	$a + (b-a)\left(\frac{1}{2} + \frac{1}{2}\text{erf}(\zeta/\sqrt{2})\right)$	$\sqrt{2}\text{erf}^{-1}\left(\frac{2\theta-(\text{b+a})}{\text{b-a}}\right)$
Normal (μ, σ)	$\mu + \sigma\zeta$	$\frac{\theta-\mu}{\sigma}$
Lognormal (μ, σ)	$\exp(\mu + \sigma\zeta)$	$\frac{\ln\theta-\mu}{\sigma}$
Gamma (a, b)	$ab\left(\zeta\sqrt{\frac{1}{9a}} + 1 - \frac{1}{9a}\right)^3$	$\sqrt{9a}\left(\sqrt[3]{\frac{\theta}{ab}} + \frac{1}{9a} - 1\right)$
Exponential (λ)	$-\frac{1}{\lambda}\log\left(\frac{1}{2} + \frac{1}{2}\text{erf}(\zeta/\sqrt{2})\right)$	$\sqrt{2}\text{erf}^{-1}\left(2\exp(-\lambda\theta) - 1\right)$
Weibull (a)	$y^{1/a}$, where y is Exponential(1)	θ^a
Extreme Value	$-\log(y)$, where y is Exponential(1)	$\exp(-\theta)$

Table A.2: Transformation between the standard normal random variable ζ and several common univariate distributions θ.

Transformation of random variables Some transformations from common univariate distributions to the standard normal random variables were presented by Devroye [71] and extended by [120]. Table A.2 shows a list of transformations for some probability distributions commonly used in system analysis.

Remark A.4. A transformation of random input variables might result in a high-order or non-polynomial representation of spectral representation (e.g. polynomial chaos expansions). For example, consider a standard uniform random variables $\theta \sim \mathcal{U}(0, 1)$ and transform θ to the standard normal random variable $\zeta \sim \mathcal{N}(0, 1)$ using Table A.2. Then the first order polynomial $p_1(\theta) \triangleq a_0 + a_1\theta$ becomes a non-polynomial function of ζ: $p_1(\zeta) = a_0 + a_1\sqrt{2}\text{erf}^{-1}(2\zeta - 1)$.

A.3.2.2 Random Sequences

It is common that the stochastic inputs of a system are random processes. The Karhunen-Loeve (KL) expansion has proved to be useful to represent the stochastic input quantities in the stochastic system models and be compatible to spectral methods of system identification and analysis using polynomial chaos– it is because that the KL expansion gives a natural way to parameterize the random process inputs so that such a parameterization can be exploited in the spectral analysis to construct basis functions.

An essential idea behind the KL decomposition is to represent a stochastic process by a spectral decomposition of its correlated function. Consider a spatially or temporally varying random filed $\alpha(z, t, \omega)$ over a the spatial domain $\mathcal{Z}$ and time domain $\mathcal{T}$ with the mean $\bar{\alpha}(z, t)$ and covariance function $C_\alpha(\eta_1, \eta_2)$ where $\eta \triangleq (z, t)$. Then the KL expansion of the random process $\alpha(\eta, \omega)$is given by

$$\alpha(\eta, \omega) = \bar{\alpha}(\eta) + \sum_{i=0}^{\infty} \sqrt{\lambda_i}\phi_i(\eta)\alpha_i(\omega) \tag{A.45}$$

where ϕ_i are the orthogonal eigenfunctions corresponding to the eigenvalues λ_i of the integral operator $\mathbf{T}_{C_\alpha}$ defined as

$$\mathbf{T}_{C_\alpha}\phi(\eta) \triangleq \int_{\mathcal{Z}\times\mathcal{T}} C_\alpha(\eta, s)\phi(s)d\mu(s) \tag{A.46}$$

with a properly chosen measure $\mu : \mathcal{Z} \times \mathcal{T} \to \mathbb{R}$. The set of random variables $\{\alpha_i(\omega)\}$ is defined over the event $\omega \in \Omega$ are jointly uncorrelated and have zero means, i.e., $\mathbf{E}[\alpha_i] = 0$ and $\mathbf{E}[\alpha_i \alpha_j] = \delta_{ij}$. It can be computed by the following equation using the orthogonality of $\{\phi_i\}$:

$$\alpha_i(\omega) = \frac{1}{\sqrt{\lambda_i}} \int_{\mathcal{Z}\times\mathcal{T}} (\alpha(\eta, \omega) - \bar{\alpha}(\eta))\, \phi_i(\eta) d\mu(\eta) \tag{A.47}$$

for each $i \in \mathbb{N}$. Relying on principal component analysis (PCA), the truncated series after a finite number of terms, we use an approximate representation of the KL expansion

$$\alpha^N(\eta, \omega) \triangleq \bar{\alpha}(\eta) + \sum_{i=0}^{N} \sqrt{\lambda_i} \phi_i(\eta) \alpha_i(\omega) \tag{A.48}$$

where $N+1$ is the number of basis functions and the sequence of eigenvalues $\{\lambda_i\}$ is assumed to be chosen as being non-increasing. It is known that the truncated KL expansion is the finite spectral representation with the minimal mean-square error over any finite number of basis functions. We also note that the probability space is decoupled from the *deterministic* spatial and temporal spaces in the KL expansion for the random field $\alpha(\eta, \omega)$, which is important for our approaches to analysis and synthesis problems of stochastic system models and experiments using polynomial chaos.

Remark A.5. It is important to note that in the KL expansion (A.45) of a random field $\alpha(\eta, \omega)$, the set of parameterized random variables $\{\alpha_i(\omega)\}$ is not jointly independent, but uncorrelated. For Gaussian random processes, they are equivalent. However, independence and uncorrelatedness are not equivalent in general. it should be mentioned that uncorrelated random variables are used in generalized polynomial chaos, even though it is not a theoretically rigorous approach. In many practical situation, this assumption (or heuristic justification) might be fine.

A.3.3 Generalized Polynomial Chaos Expansions

Here, we discuss the PC expansion and its generalization to other types of orthogonal polynomials called generalized PC (gPC) expansions.

A.3.3.1 Universal Approximation Property of PC expansion

The homogeneous chaos expansion was first proposed by Wiener [288]. It employs the Hermite polynomials in terms of Gaussian random variables. It was shown that it can approximate any functionals in $\mathcal{L}_2$ and converges in the $\mathcal{L}_2$ sense [46] (see Theorem A.1). Therefore, Hermite polynomial chaos (H-PC) expansion has a universal approximation property for expanding second-order random processes in terms of orthogonal polynomials. Second-order random processes are processes with finite variance, and this applies to most

physical process [294].

$$X(\omega) = \sum_{i=0}^{N_p} \hat{a}_i \psi_i(\zeta(\omega)) \tag{A.49}$$

$$\begin{aligned} X(\omega) = a_0 H_0 + \sum_{i_1}^{N_p} a_{i_1} H_1(\zeta_{i_1}(\omega)) + \sum_{i_1}^{N_p} \sum_{i_2}^{i_1} a_{i_1 i_2} H_2(\zeta_{i_1}(\omega), \zeta_{i_2}(\omega)) \\ + \sum_{i_1}^{N_p} \sum_{i_2}^{i_1} \sum_{i_3}^{i_2} a_{i_1 i_2 i_3} H_3(\zeta_{i_1}(\omega), \zeta_{i_2}(\omega), \zeta_{i_3}(\omega)) + \cdots \end{aligned} \tag{A.50}$$

where

$$H_n(\zeta_{i_1}, \cdots, \zeta_{i_n}) = e^{1/2\zeta^{\mathrm{T}}\zeta}(-1)^n \frac{\partial^n}{\partial\zeta_{i_1} \cdots \partial\zeta_{i_n}} e^{-1/2\zeta^{\mathrm{T}}\zeta}$$

denotes the H-PCE of order n in the variable $(\zeta_{i_1}, \cdots, \zeta_{i_n})$, i.e., the H_n are Hermite polynomials in terms of the standard Gaussian random vector $\zeta \sim \mathcal{N}(0, \mathrm{I})$.

Definition A.5. A function $f : \Theta_x \to \Theta_y$ is said to be in a Hilbert space $G\mathcal{L}_2(\Theta_x)$ with the inner product defined as $\langle f, f \rangle_\mu \triangleq \int_{\Theta_x} f^*(x) f(x) d\mu(x)$, if $\langle f, f \rangle_\mu < \infty$.

Theorem A.1 (*Cameron-Martin Theorem [46]*). For any functional $F(x) \in G\mathcal{L}_2(\Theta_x)$,

$$\lim_{N \to \infty} \int_{\Theta_x} \left(F(x) - \sum_{j=0}^{N} \hat{a}_j \psi_j(x) \right)^2 \rho(x) dx = 0 \tag{A.51}$$

where $\hat{a}_j$ is obtained from the Galerkin projection, i.e., $\hat{a}_j := \frac{\langle F, \psi_j \rangle}{\|\psi_j\|^2_{G\mathcal{L}_2}}$.

A.3.3.2 Generalized PC expansion: The Wiener-Askey Polynomial Chaos

In order to deal with more general random variables, the Wiener-Askey polynomial chaos expansion has been introduced [294] as a generalization of the original Wiener chaos expansion.

$$X(\theta) = \sum_{i=0}^{N_p} \hat{c}_i \phi_i(\zeta(\theta)) \tag{A.52}$$

$$\begin{aligned} X(\theta) = c_0 I_0 + \sum_{i_1}^{N_p} c_{i_1} I_1(\zeta_{i_1}(\theta)) + \sum_{i_1}^{N_p} \sum_{i_2}^{i_1} c_{i_1 i_2} I_2(\xi_{i_1}(\theta), \zeta_{i_2}(\theta)) \\ + \sum_{i_1}^{N_p} \sum_{i_2}^{i_1} \sum_{i_3}^{i_2} c_{i_1 i_2 i_3} I_3(\zeta_{i_1}(\theta), \zeta_{i_2}(\theta), \zeta_{i_3}(\theta)) + \cdots \end{aligned} \tag{A.53}$$

	Random Variables ζ	Wiener-Askey Polynomial Chaos ζ	Support
Continuous	Gaussian	Hermite PCE	$(-\infty, \infty)$
	Gamma	Laguerre PCE	$[0, \infty)$
	Beta	Jacobi PCE	$[a, b]$
	Uniform	Legendre PCE	$[a, b]$
Discrete	Poisson	Charlier PCE	$\{0, 1, 2, \cdots\}$
	Binomial	Krawtchouk PCE	$\{0, 1, \cdots, N\}$
	Negative Binomial	Meixner PCE	$\{0, 1, 2, \cdots\}$
	Hypergeometric	Hahn PCE	$\{0, 1, \cdots, N\}$

Table A.3: Types of gPC expansions and the corresponding standard random variables.

where $I_n(\zeta_{i_1}, \cdots, \zeta_{i_n})$ denotes the Wiener-Askey polynomial chaos of order n in terms of the random vector $\zeta = (\zeta_{i_1}, \cdots, \zeta_{i_n})$ and it is not restricted to Hermite polynomials but rather can be all types of the orthogonal polynomials from the Askey scheme (see [294] for details).

A.3.3.3 Extensions to Heterogeneous Random Variables

We consider a function of random variables $X : \Theta_1 \times \Theta_2 \to \Theta_x$ where the two random variables $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ are independent and have different types of probabilistic distributions. For example, θ_1 is uniformly distributed in a bounded hypercube and θ_2 is a Gaussian random vector, $\theta_2 \sim \mathcal{N}(\bar{\theta}_2, \Sigma_{\theta_2})$.

$$\begin{aligned} X(\theta_1, \theta_2) &= X_1(\theta_1) X_2(\theta_2) \\ &= \left(\sum_{i=0}^{N_1} \hat{a}_i \psi_i(\xi(\theta_1)) \right) \left(\sum_{j=0}^{N_2} \hat{c}_j \phi_i(\zeta(\theta_2)) \right) \\ &= \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \hat{a}_i \hat{c}_j \psi_i(\xi(\theta_1) \phi_i(\zeta(\theta_2)) \\ &= \sum_{l=0}^{N_p} \hat{b}_l \Gamma_l(\xi(\theta_1), \zeta(\theta_2)) \end{aligned} \tag{A.54}$$

where $N_p = N_1 N_2$ and the $\hat{b}_l$ and Γ_l are defined as the followings with an one-to-one index mapping $\iota : \{0, \cdots, N_1\} \times \{0, \cdots, N_2\} \to \{0, \cdots, N_p\}$:

$$\begin{aligned} \hat{b}_l &:= \hat{a}_i \hat{c}_j \\ \Gamma_l(\xi(\theta_1), \zeta(\theta_2)) &:= \psi_i(\xi(\theta_1)) \phi_j(\zeta(\theta_2)) \end{aligned} \tag{A.55}$$

where the indices (i, j, l) satisfies $\iota(i, j) = l$.

Proposition A.1. The new polynomial basis $\{\Theta_l\}$ forms a complete orthogonal basis on the Hilbert space

$\mathcal{L}_2(\Theta)$, where $\Theta \triangleq \Theta_1 \times \Theta_2$, with the inner product defied by

$$\langle f, g \rangle \triangleq \int_{\Theta_1} \int_{\Theta_2} f(x_1, x_2) g(x_1, x_2) \rho_1(x_1) \rho_2(x_2) dx_1 dx_2 \quad \text{for } f, g \in \mathcal{L}_2(\Theta)$$

In other words, these polynomials also have the orthogonal property:

$$\langle \Gamma_m, \Gamma_n \rangle = \|\Gamma_m\|^2_{\mathcal{L}_2} \delta_{mn}. \tag{A.56}$$

Proof. The proof is straightforward:

$$\begin{aligned}
\langle \Gamma_m, \Gamma_n \rangle &= \langle \psi_{i_m} \phi_{j_m}, \psi_{i_n} \phi_{j_n} \rangle \quad \text{where } \iota(i_m, j_m) = m \text{ and } \iota(i_n, j_n) = n, \\
&= \langle \psi_{i_m}, \psi_{i_n} \rangle \langle \phi_{j_m}, \phi_{j_n} \rangle, \\
&= (\|\psi_{i_m}\|^2_{\mathcal{L}_2^1} \delta_{i_m i_n})(\|\phi_{j_m}\|^2_{\mathcal{L}_2^2} \delta_{j_m j_n}), \\
&= \|\psi_{i_m}\|^2_{\mathcal{L}_2^1} \|\phi_{j_m}\|^2_{\mathcal{L}_2^2} \delta_{mn} \quad (\because \iota \text{ is one-to-one}), \\
&= \|\Gamma_m\|^2_{\mathcal{L}_2} \delta_{mn}.
\end{aligned} \tag{A.57}$$

The next proposition is only for a multilinear map. For general nonlinear map $F(x_1, x_2)$, the proof for being uniform approximation might be not simple.

Proposition A.2. Same as Cameron-Martin Theorem, for any functional $F(x_1, x_2) = (F_1 \otimes F_2)(x_1, x_2) \in \mathcal{L}_2(\Theta_1 \times \Theta_2)$, where $\otimes$ denotes the tensor product of two functionals $F_1 : \Theta_1 \to \mathcal{X}_1$ and $F_3 : \Theta_2 \to \mathcal{X}_2$,

$$\lim_{N \to \infty} \int_{\Theta_1} \int_{\Theta_2} |F(x_1, x_2) - \sum_{j=0}^{N} \hat{b}_j \Gamma_j(x_1, x_2)|^2 \rho_1(x_1) \rho_2(x_2) dx_1 dx_2 = 0$$

where $\hat{b}_j$ is obtained from the Galerkin projection, i.e., $\hat{b}_j := \frac{\langle F, \Gamma_j \rangle}{\|\Gamma_j\|^2_{\mathcal{L}_2}}$.

Proof. Proof is easy. Noting the relation $X_1 X_2 - \hat{X}_1 \hat{X}_2 = (X_1 - \hat{X}_1) X_2 + (X_2 - \hat{X}_2) \hat{X}_1$,

$$\begin{aligned}
&\int_{\Theta_1} \int_{\Theta_2} |F(x_1, x_2) - \sum_{j=0}^{N} \hat{b}_j \Gamma_j(x_1, x_2)| \rho_1(x_1) \rho_2(x_2) dx_1 dx_2, \\
&= \int_{\Theta_1} \int_{\Theta_2} |F_1(x_1) F_2(x_2) - \hat{F}_1(x_1) \hat{F}_2(x_2)| \rho_1(x_1) \rho_2(x_2) dx_1 dx_2, \\
&= \int_{\Theta_1} \int_{\Theta_2} |(F_1(x_1) - \hat{F}_1(x_1)) F_2(x_2) + (F_2(x_2) - \hat{F}_2(x_2)) - \hat{F}_1(x_1)| \rho_1(x_1) \rho_2(x_2) dx_1 dx_2, \\
&\leq \int_{\Theta_2} \left(\int_{\Theta_1} |F_1(x_1) - \hat{F}_1(x_1)| \rho_1(x_1) dx_1 \right) |F_2(x_2)| \rho_2(x_2) dx_2, \\
&\quad + \int_{\Theta_1} \left(\int_{\Theta_2} |F_2(x_2) - \hat{F}_2(x_2)| \rho_2(x_2) dx_2 \right) |\hat{F}_1(x_1)| \rho_1(x_1) dx_1 \\
&\to 0 \ \text{ as } N \to \infty (\text{ in the sense previously explained}).
\end{aligned} \tag{A.58}$$

We implicitly suppose that $N \to \infty$ means $N_1 \to \infty$ and $N_2 \to \infty$, i.e., both of degrees of basis polynomials for $\mathcal{L}_2^1$ and $\mathcal{L}_2^2$ are increased at the same rate, at least in the asymptotes.

A.3.4 Determination of Coefficients: Eigendecomposition

Once we select an appropriate (optimal in the convergence rate) set of basis functions, then the next step in stochastic spectral methods is to find a set of coefficients $\mathbf{a} \triangleq \{a_i\}$ minimizing the distance between the true function $y(x)$ and its approximator $\hat{y}(x) = \sum_{i=0}^{N_P} a_i \phi_i(x)$. For sake of simplicity, we only consider a scalar output $y \in \mathbb{R}$, but extensions to higher dimensional cases are not difficult. It is also important to select a proper definition for the distance between two functions and to project the true solution onto the space of spectral expansions.

A.3.4.1 Non-intrusive Projection: The Least-Squares Fitting

Standard Least-Squares Least-squares fitting is a simple and popular method in parameter estimation where one seeks system (model) parameters minimizing the sum of the squares of the residuals from a set of measurements or observations. To do this, we first choose a set of fitting data points, $\{x_i\}_{i=1}^{N_s}$ and $\{\bar{y}_i\}_{i=1}^{N_s}$, where $\bar{y}_i = y(x_i)$ and construct the input matrix $\Phi \in \mathbb{R}^{N_s \times (N_p+1)}$ whose elements are defined as $\Phi_{ij} = \Phi_j(x_i)$. Now, we need to solve the least-square problem

$$\min_{a \in \mathbb{R}^{N_p+1}} \|\Phi \mathbf{a} - \bar{y}\|_2 \tag{A.59}$$

where $\mathbf{a} \triangleq [a_0, \cdots, a_{N_P}]^{\mathrm{T}}$ is the concatenation of the coefficients and $\bar{y} \triangleq [y_1, \cdots, y_{N_s}]^{\mathrm{T}}$ is the concatenation of the measurements corresponding to the set of input data $\{x_i\}_{i=1}^{N_s}$. The unique optimal solution is obtained as $\mathbf{a}^\star = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\bar{y}$, provided $\Phi^{\mathrm{T}}\Phi$ is invertible. More generally, $\mathbf{a}^\star = (\Phi^{\mathrm{T}}\Phi)^{\dagger}\Phi^{\mathrm{T}}\bar{y}$ where $\dagger$ denotes the Moore-Penrose pseudoinverse.

Remark A.6. There might be no assumption on the parameter x, possibly except for its support. For the Bayesian approach, one might have (a priori) information (guessed or by intuition) on the probabilistic distribution of x. However, if this a priori information is not really close to the true probability distribution of x, then a least-squares fitting might not be successful. For example, suppose that $x \sim \mathcal{N}(0,1)$, but we use the uniform distribution $x \in [x^{\mathrm{l}}, x^{\mathrm{u}}]$ as a priori distribution. Then the set of test data $\{(x_i, y_i), \cdots, (x_{N_s}, y_{N_s})\}$ might end up with a poor estimation, in the sense that it minimizes the error in the finite dimensional space constructed from the selected and observed data points. This can happen because the importance weights of the training data are misplaced and it is also related with the well-known Runge phenomenon.

Regularized Least-Squares When the measurements or observations are corrupted by noise or disturbance, regularization methods must be applied to prevent over-parameterization. To do this, we solve the

optimization problems in Table A.4 where the matrix of reproducing kernels $K \in \mathbb{R}^{N_s \times N_s}$ has the elements $K_{ij} \triangleq \sum_{s=0}^{N_p} \phi_s(x_i)\phi_s(x_j) = K_{ji}$ and for the indirect regularization approach, an optimal coefficient vector $\mathbf{a}^\star$ can be reconstructed as $\Phi^{\mathrm{T}}\mathbf{b}^\star$. When $p = 2$, it is also called the Tikhonov regularization and analytical solutions for each optimization problems can be obtained. For $p = 1, \infty$, one needs to rely on numerical computations to compute optimal solutions.

A.3.5 Non-Intrusive Interpolation: The Principle of Collocation

Unlike non-intrusive regression methods based on least-squares, the collocation methods are not minimizing the residual between the true system and the surrogate model based on PCEs, but rely on interpolation of the properly chosen input-output data points. The approximation is exact at the chosen N_s collocation points $\{x_i\}_{i=1}^{N_s}$, i.e., an approximation $\hat{y}$ is required to satisfy $\hat{y}(x_i) = y(x_i)$ for each $i = 1, \ldots, N_s$, and the resulting approximate solution is a linear combination of the interpolation polynomials with the coefficients $\{y(x_i)\}_{i=1}^{N_s}$, i.e., $\hat{y}(x) = \sum_{i=1}^{N_s} y(x_i)I_i(x)$ where $I_i(x_j) = \delta_{ij}$ for $i, j = 1, \ldots, N_s$. Since a set of nodes (or collocation points) are defined in a random space, we might label this collocation method as probabilistic polynomial collocation method (PPCM). There are several techniques to determine an interpolation polynomial $\{I_i(s)\}_{i=1}^{N_s}$ and we refer the readers to research monographs [167, 293], for detailed discussion of numerical computations.

A.3.6 Non-intrusive Spectral Decomposition

Unlike the previously introduced methods, the non-intrusive spectral decomposition does not exploit any input-output data, but computes the projected coefficients of random model output $y(x)$ on a finite dimensional subspace that is spanned by the set of basis functions, i.e., $\mathbf{Span}\{\phi_i\}_{i=0}^{N_p} \in \mathcal{L}_2(\mathcal{X})$. Suppose that the output is a second-order random variable, i.e., $\|y\|_{\mathcal{L}_2} < \infty$, then the approximation of y in the subspace $\mathbf{Span}\{\phi_i\}_{i=0}^{N_p}$ is represented as

$$y^{N_P}(x) \triangleq \sum_{i=0}^{N_p} a_i\phi_i(x) \tag{A.60}$$

Classification	Optimization Problem	Analytic Solution for $p = 2$
Direct	$\min_{\mathbf{a}\in\mathbb{R}^{N_p+1}} \frac{1}{N_s}\sum_{i=1}^{N_s} \|\Phi\mathbf{a} - \bar{y}\|_2^2 + \lambda_{\mathrm{r}}\|\mathbf{a}\|_p$	$\mathbf{a}^\star = (\Phi^{\mathrm{T}}\Phi + N_s\lambda_{\mathrm{r}}\Phi)^{-1}\Phi^{\mathrm{T}}\bar{y}$
Indirect	$\min_{\mathbf{b}\in\mathbb{R}^{N_s}} \frac{1}{N_s}\sum_{i=1}^{N_s} \|K\mathbf{b} - \bar{y}\|_2^2 + \lambda_{\mathrm{r}}\|\mathbf{b}\|_p$	$\mathbf{b}^\star = (K^{\mathrm{T}}K + N_s\lambda_{\mathrm{r}}K)^{-1}K^{\mathrm{T}}\bar{y}$

Table A.4: Direct and indirect regularization methods with the p-norm regularization term.

Integration methods	Examples
Simulation based approaches	Monte Carlo method
	Improved sampling strategies
Deterministic Approaches	Quadratic formulas
	Tensor product
	Sparse grid cubatures
	Adaptive sparse grids

Table A.5: Numerical integration methods.

where the projection coefficients a_i are given by

$$a_i \triangleq \frac{\langle y, \phi_i \rangle}{\langle \phi_i, \phi_i \rangle}, \quad i = 0, \ldots, N_p \tag{A.61}$$

and the inner product is defined as the integration $\int_{\mathcal{X}} y(x)\phi_i(x)d\mu(x) = \int_{\mathcal{X}} y(x)\phi_i(x)\rho(x)dx$. Table A.5 shows different computation schemes of multivariable integration that have been proposed to numerically estimate the inner products.

A.3.7 Intrusive Galerkin Projection

Consider a general stochastic differential equation (SDE) of the form

$$L(z, t, \theta; y) = g(z, t, \theta) \tag{A.62}$$

where $z \in \mathcal{Z}$ and $t \in \mathcal{T}$ are the spatial and temporal variables, $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ is the concatenation of the random variables, the function $g : \mathcal{Z} \times \mathcal{T} \times \Theta \to \mathbb{R}$ is a forcing term, and $y : \mathcal{Z} \times \mathcal{T} \to \mathbb{R}$ is a solution of the equation, which also defines a random field over the spatial and temporal spaces $\mathcal{Z} \times \mathcal{T}$ due to the random variable vector θ.

Suppose that there exists a bijective transformation (not necessarily diffeomorphism) $T : \Theta \to \Xi$ such that $\zeta(\theta; \omega) = T(\theta(\omega))$ for all $\theta \in \Theta$ and $\omega \in \Omega$ and the transformed random variable ζ is a standard (optimal in the sense of convergence rate) random variable for the set of polynomial basis functions $\{\phi_i\}$.

For application of the spectral method based on polynomial chaos expansions, we assume that the solution of the SDE given in (A.62) has the form

$$y \approx y^{N_p} \triangleq \sum_{i=1}^{N_p} y_i(z, t)\phi_i(\zeta(\theta; \omega)) \tag{A.63}$$

which is an approximation of the true solution y with $N_p + 1$ basis functions from the set $\{\phi_i\}$. To obtain the approximated solution y^{N_p}, we need to determine the spatial- and temporal-varying deterministic coefficients

$y_i(z,t)$. To do this, substitute the approximation y^{N_p} to y of the SDE (A.62)

$$L\left(z,t,\theta;\sum_{i=1}^{N_p} y_i(z,t)\phi_i(\zeta(\theta;\omega))\right) = g(z,t,\theta) \tag{A.64}$$

and solve it for the spatial- and temporal-dependent coefficients $y_i(z,t)$ by intrusive or non-intrusive projections onto the probability space of the random variable θ or ζ. In particular, a Galerkin projection for the above equation can be conducted such that the approximation error is orthogonal to the functional space that is spanned by the finite dimensional bases $\{\phi_i\}$:

$$\left\langle L\left(z,t,\theta;\sum_{i=1}^{N_p} y_i(z,t)\phi_i(\zeta(\theta;\omega))\right), \phi_k(\zeta(\theta;\omega))\right\rangle = \langle g(z,t,\theta), \phi_k(\zeta(\theta;\omega))\rangle \tag{A.65}$$

for each $k = 1,\ldots,N_p$. The resulting equations are the governing equations for the coefficients $y_i(z,t)$, which are deterministic.

We provide a detailed treatment of an elementary example using intrusive Galerkin projection method for a gPCE.

Example A.1. consider the system equation given by

$$\frac{d}{dt}x(t) + kx(t) = u_m h(t), \quad x(0) = x_0 \tag{A.66}$$

where $1/k$ corresponds to the time-constant of the equation, u_m is the magnitude of the Heaviside step input $h(t)$, and x_0 is the initial condition for the solution x. It is not difficult to see that the unique analytical solution has the form

$$x(t) = \left(x_0 - \frac{1}{k}u_m\right)e^{-kt} + \frac{1}{k}u_m. \tag{A.67}$$

Suppose that the parameters x_0. k, and u_m are random variables and there exist bijective transformation s_i such that $\theta_i = s_i(\zeta_i)$ such that $\zeta_i \sim$ Uniform$(-1,1)$ for each $i = 1,2,3$ and $\theta_1 := x_0$, $\theta_2 := k$, and $\theta_3 := u_m$. In particular, we consider $x_0 = \bar{x}_0(1+\rho\zeta_1)$, $k = \bar{k}_0(1+\rho\zeta_2)$, and $u_m = \bar{u}_{m0}(1+\rho\zeta_3)$ where $\bar{x}_0$, $\bar{k}_0$, and $\bar{u}_{m0}$ are the nominal values of system parameters, and $\rho = 0.5$ denotes 50% uncertainty in system parameters. Consider an approximation $x^{N_p}(t) \triangleq \sum_{i=0}^{N_p} x_i(t)\phi_i(\zeta)$ to the solution $x(t)$. Substituting $x^{N_p}(t)$ into $x(t)$ of (A.66) yields

$$\frac{d}{dt}\sum_{i=0}^{N_p} x_i(t)\phi_i(\zeta) + s_2(\zeta_2)\sum_{i=0}^{N_p} x_i(t)\phi_i(\zeta) = s_3(\zeta_3)h(t), \tag{A.68}$$

where the initial condition is rewritten as $x(0) = s_1(\zeta_1)$. Then we apply the Galerkin projection to determine

the coefficients $\{x_i(t)\}$:

$$\langle \phi_j(\zeta), \text{Eq. (A.68)} \rangle$$
$$\Rightarrow \quad \langle \phi_j(\zeta), \phi_j(\zeta) \rangle \, \tfrac{d}{dt} x_j(t) + \sum_{i=0}^{N_p} \langle \phi_j(\zeta), s_3(\zeta_3)\phi_i \rangle \, x_i(t) = \langle \phi_j(\zeta), s_2(\zeta_2) \rangle \, h(t); \;\; x(0) = \langle \phi_j(\zeta), s_1(\zeta_1) \rangle \, .$$

The resulting deterministic ordinary differential equation (ODE) for the coefficients $\{x_i(t)\}$ becomes

$$E\dot{X}(t) + AX(t) = Bh(t), \quad X(0) = C \tag{A.69}$$

where $X(t) \triangleq (x_0(t), \cdots, x_{N_p}(t)) \in \mathbb{R}^{N_p+1}$ and

$$\begin{aligned}
E &= [E_{ij}], \quad E_{ij} \triangleq \langle \phi_i(\zeta), \phi_j(\zeta) \rangle; \\
A &= [A_{ij}], \quad A_{ij} \triangleq \langle \phi_i(\zeta), s_3(\zeta_3)\phi_j(\zeta) \rangle; \\
B &= [B_i], \quad B_i \triangleq \langle \phi_i(\zeta), s_2(\zeta_2) \rangle; \\
C &= [C_i], \quad C_i \triangleq \langle \phi_i(\zeta), s_1(\zeta_1) \rangle.
\end{aligned}$$

Indeed, $E = \text{diag}\left(\langle \phi_i(\zeta), \phi_i(\zeta) \rangle\right)$ is diagonal and invertible and each element of the matrices above are polynomial such that the inner products that correspond to multivariate integrals can be computed exactly by using Gaussian quadrature-rules.

Appendix B

Decomposition Methods for Optimization

B.1 Iterative Dual Decomposition

This section primarily focuses on two problems. The first problem is a standard form of decomposable optimization with linear consistency (or complicating) constraints and the second problem has separable costs and constraints with coupled linear inequalities. The mathematical formulation of the first class of decomposable optimization is the following.

Problem B.1. Consider an optimization with the separable payoff function, separable constraints, and equality consistency constraints of the form

$$\begin{aligned} &\text{maximize } J(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{N} \ell_k(x_k, y_k) \\ &\text{subject to } (x_k, y_k) \in \mathcal{F}_k,\ k = 1, \ldots, N, \\ &\qquad y_k = C_k \mathbf{z},\ k = 1, \ldots, N, \end{aligned} \tag{B.1}$$

where (x_k, y_k) is the kth pair of separable decision variables that correspond to the separated convex cost functions $\ell_k(x_k, y_k)$, $\mathcal{F}_k$ denotes the kth constraint for the separated decision variable pair (x_k, y_k), and $y_k = C_k\mathbf{z}$ for $k = 1, \ldots, N$ are the consistency constraints, which are the only coupled constraints over the separated decision variables.

B.1.1 Lagrangian Method and Decomposition

Consider the optimization (B.1). An associated augmented Lagrangian to relax the consistency constraint is given by

$$\begin{aligned} L(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) &= \sum_{k=1}^{N} \ell_k(x_k,y_k) - \sum_{k=1}^{N} \langle v_k, y_k - C_k\mathbf{z}\rangle \\ &= \sum_{k=1}^{N} \left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right) + \sum_{k=1}^{N} \langle C_k^* v_k, \mathbf{z}\rangle \end{aligned} \tag{B.2}$$

where $\mathbf{x} = [x_1,\cdots,x_N]^{\mathrm{T}}$, $\mathbf{y} = [y_1,\cdots,y_N]^{\mathrm{T}}$, $\mathbf{v} = [v_1,\cdots,v_N]^{\mathrm{T}}$,[1] and the superscript refers to the associated adjoint operator.

Finding a saddle point that is a global optimal solution requires solving the two-stage optimization

$$\inf_{\mathbf{v}} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{F},\,\mathbf{z}} L(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{v}) \tag{B.3}$$

where $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_N$ refers to the product set of local (i.e., subsystem) constraints. From (B.2), the optimization (B.3) can be rewritten as

$$\begin{aligned} &\inf_{\mathbf{v}} \underbrace{\sup_{(\mathbf{x},\mathbf{y})\in\mathcal{F},\,\mathbf{z}} \left(\sum_{k=1}^{N} \left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right) + \langle C_k^* v_k, \mathbf{z}\rangle\right)}_{\begin{cases} \inf\limits_{\mathbf{v}} \sup\limits_{(\mathbf{x},\mathbf{y})\in\mathcal{F}} \left(\sum\limits_{k=1}^{N} \left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right)\right) \\ \qquad\qquad \text{if } \mathbf{C}^*\mathbf{v} = 0 \\ +\infty \qquad \text{otherwise} \end{cases}} \\ &= \inf_{\mathbf{C}^{\mathrm{T}}\mathbf{v}=0} \sup_{(\mathbf{x},\mathbf{y})\in\mathcal{F}} \left(\sum_{k=1}^{N} \left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right)\right) \\ &= \inf_{\mathbf{C}^{\mathrm{T}}\mathbf{v}=0} \sum_{k=1}^{N} \left(\sup_{(x_k,y_k)\in\mathcal{F}_k} \left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right)\right) \end{aligned} \tag{B.4}$$

where $\mathbf{C}^{\mathrm{T}} = [C_1^{\mathrm{T}},\cdots,C_N^{\mathrm{T}}]$.

The optimization (B.3) can be decomposed into two convex programs:

$$\textbf{Slave Problem:} \quad \sup_{(x_k,y_k)\in\mathcal{F}_k} \underbrace{\left(\ell_k(x_k,y_k) - \langle v_k, y_k\rangle\right)}_{S_k(x_k,y_k|v_k)} \tag{B.5}$$

for $k = 1,\ldots,N$, and

$$\textbf{Master Problem:} \quad \inf_{\mathbf{C}^{\mathrm{T}}\mathbf{v}=0} \sum_{k=1}^{N} \underbrace{S_k(x_k^\star, y_k^\star|v_k)}_{Q_k(v_k)} \tag{B.6}$$

[1] The bold refers to global variables while the non-bold refers to local variables.

where $(x_k^\star, y_k^\star)$ refers to the optimal solution pair for the Slave Problem (B.5) for given v_k.

B.1.2 Projection-(Sub)gradient Method

The Master Problem (B.6) can be solved using a first-order (sub-)gradient projection method, whereas the Slave Problem (B.5) is of much smaller size and can be accurately and efficiently solved by a second-order method such as an interior-point algorithm [37, 197]. Consider that the Master Problem (B.6) can be rewritten as

$$\inf_{\mathbf{C}^{\mathrm{T}}\mathbf{v}=0} \sum_{k=1}^{N} Q_k(v_k) \tag{B.7}$$

where $Q_k(v_k)$ is convex in v_k for all $k = 1, \ldots, N$. Define a linear subspace $\mathcal{M} \triangleq \{\mathbf{v} \in \mathbb{R}^\bullet : \mathbf{C}^{\mathrm{T}}\mathbf{v} = 0\}$, which is the null-space of the matrix $\mathbf{C}^{\mathrm{T}}$. The optimization (B.7) can be solved using a subgradient-projection method:

$$\mathbf{v}^{(n+1)} := \mathcal{P}_{\mathcal{M}}(\mathbf{v}^{(n)} - \alpha_n g^{(n)}(\mathbf{v}^{(n)})) \tag{B.8}$$

where $\mathcal{P}_{\mathcal{M}} : \mathbb{R}^\bullet \to \mathcal{M}$ refers to the projection on the subspace $\mathcal{M}$, $g^{(n)} : \mathbb{R}^\bullet \to \mathbb{R}^\bullet$ denoted the subgradient, i.e., $g \in \partial_{\mathbf{v}} \sum Q_k(v_k)$, and α_n is a step size that can be selected in any of standard ways (e.g., constant, diminishing, etc.). The sub-differential can be represented as

$$\begin{aligned} \partial_{\mathbf{v}} \left(\sum_{k=1}^{N} Q_k(v_k) \right) &= \partial_{v_1} Q_1(v_1) \times \cdots \times \partial_{v_N} Q_N(v_N) \\ &= \{-\mathbf{y}^\star\} \end{aligned} \tag{B.9}$$

where $\mathbf{y}^\star$ denotes the concatenation of optimal solutions of the Slave Problem (B.5) for a given sequence $\{v_k\}$. In other words, local subsystems are required to sequentially report the computed public variables to the supervisor (or price-planner). Therefore, the update rule for the subgradient-projection method (B.8)

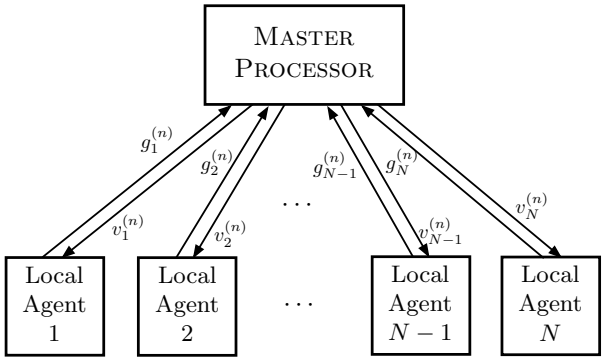


Figure B.1: Iterative dual decomposition of sequentially reporting public variables $\{g_k^{(n)}\}$ and assigning prices $\{v_k^{(n)}\}$. The superscript (n) refers to the iteration sequence.

can be rewritten as

$$\mathbf{v}^{(n+1)} := \mathcal{P}_{\mathcal{M}}\Big(\mathbf{v}^{(n)} + \alpha_n \mathbf{y}^{(n)}\Big) \tag{B.10}$$

where the superscript $\star$ of $\mathbf{y}$ is removed for notational convenience. Furthermore, it is not hard to see that

$$\mathcal{P}_{\mathcal{M}}(z) = \big(\mathrm{I} - C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\big)\, z, \tag{B.11}$$

so that

$$\begin{aligned} \mathbf{v}^{(n+1)} &:= \big(\mathrm{I} - C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\big)\, \Big(\mathbf{v}^{(n)} + \alpha_n \mathbf{y}^{(n)}\Big) \\ &:= \mathbf{v}^{(n)} + \alpha_n \underbrace{\big(\mathrm{I} - C(C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}\big)}_{U} \mathbf{y}^{(n)} \end{aligned} \tag{B.12}$$

where the computation of the matrix U needs to be performed only once and can be done offline (before performing optimization).

B.1.3 Separable Cost with Coupled Inequalities

Problem B.2. Consider an optimization with the separable payoff function, separable constraints, and coupled inequality constraints of the form

$$\begin{aligned} &\text{maximize } J(\mathbf{x}) = \sum_{k=1}^{N} \ell_k(x_k) \\ &\text{subject to } x_k \in \mathcal{F}_k,\ k = 1, \ldots, N, \\ &\qquad\qquad \mathbf{C}\mathbf{x} \geq 0 \end{aligned} \tag{B.13}$$

where x_k is the kth separable decision variable that corresponds to the separated *convex* cost functions $\ell_k(x_k)$, $\mathcal{F}_k$ denotes the kth constraint for x_k, and $\mathbf{C}\mathbf{x} = \sum_{k=1}^{N} C_k x_k$ for $k = 1, \ldots, N$ are coupled inequality constraints.

An associated augmented Lagrangian is

$$L(\mathbf{x}, \mathbf{v}) = \sum_{k=1}^{N} \ell_k(x_k) - \langle \mathbf{v}, \mathbf{C}\mathbf{x} \rangle = \sum_{k=1}^{N} \ell_k(x_k) - \sum_{k=1}^{N} \langle \mathbf{v}, C_k x_k \rangle \tag{B.14}$$

where $\mathbf{v} \geq 0$. The constrained optimization (B.13) can be decomposed into the two-stage optimization

$$\begin{aligned}\inf_{\mathbf{v}\geq 0}\sup_{\mathbf{x}\in\mathcal{F}} L(\mathbf{x},\mathbf{v}) &= \inf_{\mathbf{v}\geq 0}\left(\sup_{\mathbf{x}\in\mathcal{F}}\left(\sum_{k=1}^{N}\ell_k(x_k) - \sum_{k=1}^{N}\langle \mathbf{v}, C_k x_k\rangle\right)\right)\\ &= \inf_{\mathbf{v}\geq 0}\left(\sum_{k=1}^{N}\underbrace{\sup_{x_k\in\mathcal{F}_k}\left(\ell_k(x_k) - \langle \mathbf{v}, C_k x_k\rangle\right)}_{Q_k(\mathbf{v})}\right)\end{aligned} \tag{B.15}$$

where $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_N$ refers to the product set of local (i.e., subsystem) constraints. The optimization (B.15) can be decomposed into two convex programs:

$$\textbf{Slave Problem:}\quad \sup_{x_k\in\mathcal{F}_k}\underbrace{\left(\ell_k(x_k) - \langle \mathbf{v}, C_k x_k\rangle\right)}_{S_k(x_k|\mathbf{v})} \tag{B.16}$$

for $k = 1, \ldots, N$, and

$$\textbf{Master Problem:}\quad \inf_{\mathbf{v}\geq 0}\sum_{k=1}^{N}\underbrace{S_k(x_k^\star|\mathbf{v})}_{Q_k(\mathbf{v})} \tag{B.17}$$

where $x_k^\star$ refers to the optimal solution pair for the Slave Problem (B.16) for given $\mathbf{v}$. A similar projection-subgradient method as aforementioned can be used to solve this problem.

Projection-(Sub)gradient Method: Starting from a feasible dual variable $\mathbf{v}^{(0)} \geq 0$, the sequences of primal-dual solutions can be computed as follows:

$$x_k^{(n)} := \arg\max_{x_k\in\mathcal{F}_k}\left(\ell_k(x_k) - \langle \mathbf{v}^{(n)}, C_k x_k\rangle\right) \tag{B.18}$$

and

$$\mathbf{v}^{(n+1)} := \left(\mathbf{v}^{(n)} + \alpha_n\sum_{k=1}^{N} C_k x_k^{(n)}\right)_+ \tag{B.19}$$

where $(a)_+$ has the ith element defined as a_i if $a_i \geq 0$ and 0 otherwise.

B.1.4 An Application: Overlapped Data Processing with Redundant Sensors

Consider the constrained least-squares estimation problem

$$\begin{aligned}&\min_{\theta}\ \|b - A\theta\|_2^2\\ &\text{s.t. } \ell \leq \theta \leq u\end{aligned} \tag{B.20}$$

where the matrix A has a coupled-structure of the form

$$A = \left[\begin{array}{c|c|c|c|c|c|c} A_{11} & A_{12} & 0 & 0 & \cdots & 0 & A_{1N} \\ \hline A_{21} & A_{22} & A_{23} & 0 & \ddots & \cdots & 0 \\ \hline 0 & A_{21} & A_{22} & A_{23} & 0 & \ddots & \vdots \\ \hline \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \hline \vdots & 0 & 0 & \bullet & \bullet & \bullet & 0 \\ \hline 0 & \vdots & \ddots & \ddots & \bullet & \bullet & \bullet \\ \hline A_{N1} & 0 & \cdots & \cdots & 0 & A_{NN-1} & A_{NN} \end{array}\right]. \tag{B.21}$$

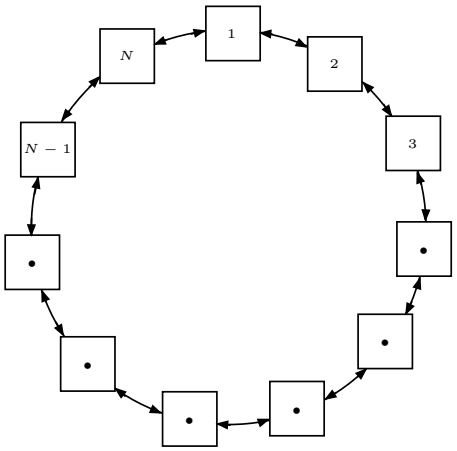


Figure B.2: Circularly interconnected sensor network

Rewrite $A\theta$ by

$$A\theta = \begin{bmatrix} A_{11}\theta_1 + A_{12}\theta_2 + A_{1N}\theta_N \\ A_{22}\theta_2 + A_{21}\theta_1 + A_{23}\theta_3 \\ \vdots \\ A_{N-1N-1}\theta_{N-1} + A_{N-1N-2}\theta_{N-2} + A_{N-1N}\theta_N \\ A_{NN}\theta_N + A_{NN-1}\theta_{N-1} + A_{N1}\theta_1 \end{bmatrix} = \begin{bmatrix} A_{11}\underbrace{\theta_1}_{x_1} + [A_{12}\ A_{1N}]\underbrace{\begin{bmatrix}\theta_2 \\ \theta_N\end{bmatrix}}_{y_1} \\ \vdots \\ A_{N-1N-1}\underbrace{\theta_{N-1}}_{x_{N-1}} + [A_{N-1N-2}\ A_{N-1N}]\underbrace{\begin{bmatrix}\theta_{N-2} \\ \theta_N\end{bmatrix}}_{y_{N-1}} \\ A_{NN}\underbrace{\theta_N}_{x_N} + [A_{N1}\ A_{NN-1}]\underbrace{\begin{bmatrix}\theta_1 \\ \theta_{N-1}\end{bmatrix}}_{y_N} \end{bmatrix}. \tag{B.22}$$

Using the reparameterization (lifting) of the decision variables (B.22), the optimization (B.20) can be rewritten as

$$\begin{aligned} \min_{\hat{\theta},\, \mathbf{z}} \ & \sum_{i=1}^{N} \left\| b_i - \hat{A}_i \hat{\theta}_i \right\|_2^2 \\ \text{s.t. } & \hat{\ell}_i \le \hat{\theta}_i \le \hat{u}_i,\ i = 1, \ldots, N, \\ & y_i = C_i \mathbf{z},\ i = 1, \ldots, N, \end{aligned} \tag{B.23}$$

where $\hat{\theta}_i^{\mathrm{T}} = [x_i^{\mathrm{T}}, y_i^{\mathrm{T}}]$, $\mathbf{z}$ is a dummy variable that will eventually vanish in dual decomposition,

$$\hat{A}_i \triangleq \begin{bmatrix} A_{ii} & A_{ii-1} & A_{ii+1} \end{bmatrix} \quad \text{for } i = 1, \ldots, N \tag{B.24}$$

with setting the indices $0 := N$ and $N + 1 := 1$, and C_i is the ith row-block of the matrix A with replacing its ith diagonal term by the zero matrix of compatible dimension.

References

[1] J. Ackermann, P. Blue, T. Buente, L. Guevenc, D. Kaesbauer, M. Kordt, M. Muhler, and D. Odenthal. *Robust Control–The Parametric Approach*. Springer–Verlag, London, U.K., 2002.

[2] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron. Distributed Bayesian hypothesis testing in sensor networks. In *American Control Conference, 2004. Proceedings of the 2004*, volume 6, pages 5369–5374. IEEE, 2004.

[3] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92:423–438, 2004.

[4] A. E. Ashari, R. Nikoukhah, and S. Campbell. Active robust fault detection in closed-loop systems: Qquadratic optimization approach. *IEEE Transactions on Automatic Control*, 54:2532–2544, 2012.

[5] A. E. Ashari, R. Nikoukhah, and S. Campbell. Effects of feedback on active fault detection. *Automatica*, 48:866–872, 2012.

[6] B. Azimi-Sadjadi and P. S. Krishnaprasad. Change detection for nonlinear systems; a particle filtering approach. In *Proc. of American Control Conference*, volume 5, pages 4074–4079, 2002.

[7] L. Bakule. Decentralized control: An overview. *Annual Reviews in Control*, 32:87–98, 2008.

[8] G. Balas, R. Chiang, A. Packard, and M. Safonov. *MATLAB — Robust Control Toolbox, Ver. 3*. The MathWorks Inc., Natick, MA, 1998.

[9] G. J. Balas, J. C. Doyle, K. Glover, A. Packard, and R. Smith. *μ–Analysis and Synthesis Toolbox*. The MathWorks Inc., Natick, MA, 1998.

[10] I. Barland, P. G. Kolaitis, and M. N. Thakur. Integer programming as a framework for optimization and approximability. *Journal of Computer and System Sciences*, 57:144–161, 1998.

[11] B. R. Barmish and C. M. Lagoa. The uniform distribution: A rigorous justification for its use in robustness analysis. *Mathematics of Control, Signals, and Systems*, 10:203–222, 1997.

[12] J. Barraud, Y. Creff, and N. Petit. pH control of a fed batch reactor with precipitation. *Journal of Process Control*, 19:888–895, 2009.

[13] T. Basar and P. Bernhard. *H_∞-optimal control and related minimax design problems: A dynamic game approach*. Birkhauser, Boston, 2nd edition, 1995.

[14] C. W. Baum and V. V. Veeravalli. A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory*, 40:1994–2007, 1994.

[15] C. Beck and J. Doyle. A necessary and sufficient minimality condition for uncertain systems. *IEEE Transactions on Automatic Control*, 44:1802–1813, 1999.

[16] C. Beck, J. C. Doyle, and K. Glover. Model reduction of multidimensional and uncertain systems. *IEEE Transactions on Automatic Control*, 41:1466–1477, 1996.

[17] C. L. Beck. *Model Reduction and Minimality for Uncertain Systems*. PhD thesis, California Institute of Technology, Pasadena, California, 1996.

[18] M. Bellare and P. Rogaway. The complexity of approximating a nonlinear program. *Mathematical Programming*, 69:429–441, 1995.

[19] R. E. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1970.

[20] A. Bemporad and M. Morari. Robust model predictive control: A survey. In F. W. Vaandrager and J. H. van Schuppen, editors, *Hybrid Systems: Computation and Control*, pages 31–45. Lecture Notes in Computer Science, 1999.

[21] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, New Jersey, 2009.

[22] A. Ben-Tal and A. Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

[23] D. J. Bender and A. J. Laub. The linear-quadratic optimal regulator for descriptor systems. *IEEE Transactions on Automatic Control*, 32:672–688, 1987.

[24] B. W. Bequette. *Process Control: Modeling, Design and Simulation*. Prentice-Hall, Upper Saddle River, NJ, 2003.

[25] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[26] D. S. Bernstein, W. Haddad, and A. G. Sparks. A simplified proof of the multivariable Popov criterion and an upper bound for the structured singular value with real parametric uncertainty. In *IEEE Conference on Decision and Control*, pages 2139–2140, Lake Buena Vista, FL, 1994.

[27] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[28] S. Bialas and J. Garloff. Stability of polynomials under coefficient perturbation. *IEEE Transactions on Automatic Control*, 30:310–312, 1985.

[29] L. Blackmore and M. Ono. Convex chance constrained predictive control without sampling. In *Proc. of the AIAA GNC*, Chicago, 2009.

[30] L. Blackmore, M. Ono, and B. C. Williams. A probabilistic particle-control approximation of chance-constrained stochastic predictive control. *IEEE Transactions on Robotics*, 26(3):502–517, 2010.

[31] L. Blackmore, M. Ono, and B. C. Williams. Chance-constrained optimal path planning with obstacles. *IEEE Transactions on Robotics*, 27:1080–1094, 2011.

[32] L. Blackmore, S. Rajamanoharan, and B. Williams. Active estimation for jump Markov linear systems. *IEEE Transactions on Automatic Control*, 53:2223–2236, 2008.

[33] V. D. Blondel and J. N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36:1249–1274, 2000.

[34] S. Boucheron, O. Bousquet, and G. Lugosi. Concentration inequalities. In O. Bousquet, U. V. Luxburg, and G. Rtsch, editors, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

[35] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*. SIAM, Philadelphia, 1994.

[36] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, Cambridge, UK, 2004.

[37] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, NY, USA, 2004.

[38] S. Boyd, L. Vandenberghe, et al. Semidefinite programming relaxations of non-convex problems in control and combinatorial optimization. *communications, computation, control and signal processing: a tribute to Thomas Kailath*, pages 279–288, 1997.

[39] R. D. Braatz. *Robust Loopshaping for Process Control*. PhD thesis, California Institute of Technology, Pasadena, CA, 1993.

[40] R. D. Braatz, J. H. Lee, and M. Morari. Screening plant designs and control structures for uncertain systems. *Computers & Chemical Engineering*, 20:463–468, 1996.

[41] R. D. Braatz, M. Morari, and S. Skogestad. Robust reliable decentralized control. In *Proc. of American Control Conference*, pages 3384–3388, 1994.

[42] R. D. Braatz and E. L. Russell. Robustness margin computation for large scale systems. *Computers & Chemical Engineering*, 23:1021–1030, 1999.

[43] R. D. Braatz, P. M. Young, J. C. Doyle, and M. Morari. Computational complexity of μ calculation. In *Proc. of American Control Conference*, pages 1682–1683, San Francisco, CA, 1993.

[44] R. D. Braatz, P. M. Young, J. C. Doyle, and M. Morari. Computational complexity of μ calculation. *IEEE Transactions on Automatic Control*, 39:1000–1002, 1994.

[45] S. Bundfuss and M. Dur. Copositive Lyapunov functions for switched systems over cones. *Systems & Control Letters*, 58:342–345, 2009.

[46] R. Cameron and W. Martin. The orthogonal development of nonlinear functionals in series of Fourier-Hermite functionals. *The Annals of Mathematics*, 48:385, 1947.

[47] M. K. Camlibel and R. Frasca. Extension of Kalman-Yakubovich-Popov lemma to descriptor systems. *Systems & Control Letters*, 58:795–803, 2009.

[48] M. K. Camlibel and J. M. Schumacher. Copositive Lapunov functions. In V. D. Blondel and A. .Megretski, editors, *Unsolved Problems in Mathematical Systems and Control Theory*, pages 189–193. Princeton University Press, 2004.

[49] S. L. Campbell and R. Nikoukhah. *Auxiliary Signal Design for Failure Detection*. Princeton University Press, Princeton, NJ, 2004.

[50] P. J. Campo and M. Morari. Robust model predictive control. In *Proc. of American Control Conference*, pages 1021–1026, Minneapolis, MN, USA, 1987.

[51] P. J. Campo and M. Morari. Robust control of processes subject to saturation nonlinearities. *Computers & Chemical Engineering*, 14:343–358, 1990.

[52] P. J. Campo and M. Morari. Achievable closed loop properties of systems under decentralized control: Conditions involving the steady state gain. *IEEE Transactions on Automatic Control*, 39:932–943, 1994.

[53] M. Cetin, L. Chen, J. W. Fisher III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky. Distributed fusion in sensor networks. *Signal Processing Magazine, IEEE*, 23(4):42–55, 2006.

[54] T. Chen and B. A. Francis. *Optimal Sampled-data Control Systems*. Springer, New York, 1995.

[55] E. Cinquemani, M. Agarwal, D. Chatterjee, and J. Lygeros. Convexity and convex approximations of discrete-time stochastic control problems with constraints. *Automatica*, 47(9):2082–2087, 2011.

[56] P. Clifford. Markov random fields in statistics. *Disorder in physical systems*, pages 19–32, 1990.

[57] D. Cobb. Controllability, observability, and duality in singular systems. *IEEE Transactions on Automatic Control*, 29:1076–1082, 1984.

[58] B. Cooley, J. Lee, and S. Boyd. Control-relevant experiment design: A plant-friendly, lmi-based approach. In *Proc. of American Control Conference*, pages 1240–1244, 1998.

[59] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2):393–405, 1990.

[60] P. D. Couchman, M. Cannon, and B. Kouvaritakis. MPC as a tool for sustainable development integrated policy assessment. *IEEE Transactions on Automatic Control*, 51:145–149, 2006.

[61] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiely & Sons, Inc., Hoboken, NJ, USA, 2006.

[62] R. Cowell. Advanced inference in bayesian networks. *NATO ASI Series D Behavioural And Social Sciences*, 89:27–50, 1998.

[63] R. Cowell. Introduction to inference for bayesian networks. *NATO ASI Series D Behavioural And Social Sciences*, 89:9–26, 1998.

[64] G. E. Coxson and C. L. DeMarco. The computational complexity of approximating the minimal perturbation scaling to achieve instability in an interval matrix. *Mathematics of Control, Signals, and Systems*, 7:279–292, 1994.

[65] L. Csató, M. Opper, and O. Winther. TAP Gibbs free energy, belief propagation and sparsity. *Advances in Neural Information Processing Systems*, 14:657–663, 2001.

[66] P. Dagum and M. Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial intelligence*, 60(1):141–153, 1993.

[67] L. Dai. Impulsive modes and causality in singular systems. *International J. of Control*, 50:1267–1281, 1989.

[68] R. D'Andrea and S. Khatri. Kalman decomposition of linear fractional transformation representations and minimality. In *Proc. of American Control Conference*, pages 3557–3561, Albuquerque, New Mexico, 1997.

[69] R. A. Date and J. H. Chow. A reliable coordinated decentralized control system design. In *IEEE Conference on Decision and Control*, pages 1295–1300, Tampa, FL, 1989.

[70] J. Delich. *The Role of Excess Manipulated Variables within Control System Development*. PhD thesis, University of Sydney, Australia, 1992.

[71] L. Devroye. *Non-uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[72] J. C. Doyle. Analysis of feedback systems with structured uncertainties. *IEE Proceedings Part D*, 129:242–250, 1982.

[73] G. E. Dullerud and F. Paganini. *A Course in Robust Control Theory: A Convex Approach*. Springer, New York, 2000.

[74] M. Dür. Copositive programming–a survey. *Recent advances in optimization and its applications in engineering*, pages 3–20, 2010.

[75] C. G. Economou, M. Morari, and B. O. Palsson. Internal model control: Extension to nonlinear system. *Industrial & Engineering Chemistry Process Design and Development*, 25:403–411, 1986.

[76] E. Eskinat, S. H. Johnson, and W. L. Luyben. Use of Hammerstein models in identification of nonlinear-systems. *AIChE J.*, 37:255–268, 1991.

[77] F. J. Doyle III. *Robustness Properties of Nonlinear Process Control and Implications for the Design and Control of a Packed Reactor*. PhD thesis, California Institute of Technology, Pasadena, CA, 1991.

[78] A. Faanes and S. Skogestad. pH-neutralization: Integrated process and control design. *Computers & Chemical Engineering*, 28:1475–1487, 2004.

[79] M. K. H. Fan, A. L. Tits, and J. C. Doyle. Robustness in the presence of mixed parametric uncertainty and unmodeled dynamics. *IEEE Transactions on Automatic Control*, 36:25–38, 1991.

[80] L. Farina and S. Rinaldi. *Positive Linear Systems: Theory and Applications*. Wiley-Interscience, New York, 2000.

[81] A. P. Featherstone and R. D. Braatz. Input design for large scale sheet and film processes. *Industrial & Engineering Chemistry Research*, 37:449–454, 1998.

[82] J. Fisher and R. Bhattacharya. Linear quadratic regulation of systems with stochastic parameter uncertainties. *Automatica*, 45:2831–2841, 2009.

[83] J. Fisher and R. Bhattacharya. Optimal trajectory generation with probabilistic system uncertainty using polynomial chaos. *Journal of Dynamic Systems, Measurement, and Control*, 133, 2011.

[84] J. R. Fisher. *Stability Analysis and Control of Stochastic Dynamic Systems Using Polynomial Chaos*. PhD dissertation, Texas A&M University, College Station, TX, 2008.

[85] A. L. Fradkov and V. A. Yakubovich. The S-procedure and a duality relations in nonconvex problems of quadratic programming. *Vestnik Leningrad Univ. Math.*, 5:101–109, 1973.

[86] R. W. Freund and F. Jarre. An extension of the positive real lemma to descriptor systems. *Optimization Methods and Software*, 19:69–87, 2004.

[87] B. J. Frey. *Graphical models for machine learning and digital communication*. MIT press, 1998.

[88] M. Fu. The real structured singular value is hardly approximable. *IEEE Transactions on Automatic Control*, 42:1286–1288, 1997.

[89] M. Fu. Approximation of complex μ. In V. D. Blondel, E. D. Sontag, M. Vidyasagar, and J. C. Willems, editors, *Open Problems in Mathematical Systems and Control Theory*. Springer Verlag, London, UK, 1999.

[90] M. Fu and S. Dasgupta. Computational complexity of real structured singular value in ℓ_p setting. *IEEE Transactions on Automatic Control*, 45:2173–2176, 2000.

[91] M. Fujita and E. Shimemurab. Integrity against arbitrary feedback-loop failure in linear multivariable control systems. *Automatica*, 24:765–772, 1988.

[92] R. E. Garcia, Y. M. Chiang, W. C. Carter, P. Limthongkul, and C. M. Bishop. Microstructural modeling and design of rechargeable lithium-ion batteries. *J. Electrochem. Soc.*, 152:A255–A263, 2005.

[93] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, 1979.

[94] T. Geerts. Invariant subspaces and invertibility properties for singular systems: The general case. *Linear Algebra and its Applications*, 183:61–88, 1993.

[95] R. Geest and H. Trentelman. The Kalman-Yakubovich-Popov lemma in a behavioral framework. *Systems & Control Letters*, 32:283–290, 1997.

[96] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, New York, 1991.

[97] K.-C. Goh and M. G. Safonov. Robust analysis, sectors, and quadratic functionals. In *IEEE Conference on Decision and Control*, pages 1988–1993, New Orleans, LA, 1995.

[98] J. C. Gomez and E. Baeyens. Identification of block-oriented nonlinear systems using orthonormal bases. *Journal of Process Control*, 14:685–697, 2004.

[99] M. Grant, S. Boyd, and Y. Ye. cvx users guide. Technical report, Technical Report Build 711, Citeseer. Available at: http://citeseerx. ist. psu. edu/viewdoc/download, 2009.

[100] R. Gunawan, E. L. Russell, and R. D. Braatz. Comparison of theoretical and computational characteristics of dimensionality reduction methods for large-scale uncertain systems. *Journal of Process Control*, 11:543–552, 2001.

[101] A. N. Gündes. Reliable decentralized stabilization of linear systems. *IEEE Transactions on Automatic Control*, 43:1733–1739, 1998.

[102] A. N. Gündes and M. G. Kabuli. Reliable decentralized integral-action controller design. *IEEE Transactions on Automatic Control*, 46:296–301, 2001.

[103] M. A. Henson and D. E. Seborg. An internal model control strategy for nonlinear systems. *AIChE J.*, 37:1065–1081, 1991.

[104] T. A. Henzinger. The theory of hybrid automata. In *Logic in Computer Science, 1996. LICS'96. Proceedings., Eleventh Annual IEEE Symposium on*, pages 278–292. IEEE, 1996.

[105] F. Herzog, G. Dondi, and H. Geering. Stochastic model predictive control and portfolio optimization. *International J. of Theoretical and Applied Finance*, 10:203–233, 2007.

[106] M. Hofbaur and B. Williams. Mode estimation of probabilistic hybrid systems. *Hybrid Systems: Computation and Control*, pages 81–91, 2002.

[107] M. Hofbaur and B. C. Williams. Hybrid estimation of complex systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(5):2178–2191, 2004.

[108] I. G. Horn, J. R. Arulandu, C. J. Gombas, J. G. VanAntwerp, and R. D. Braatz. Improved filter design in internal model control. *Ind. Eng. Chem. Res.*, 35:3437–3441, 1996.

[109] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, 1985.

[110] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, 1991.

[111] I. M. Horowitz. *Synthesis of Feedback Systems*. Academic, New York, NY, 1963.

[112] M. Hovd. *Studies on Control Structure Selection and Design of Robust Decentralized and SVD Controllers*. PhD thesis, Norwegian Institute of Technology, Trondheim, Norway, 1992.

[113] M. Hovd, R. D. Braatz, and S. Skogestad. On the structure of the robust optimal controller for a class of problems. In *Proc. of the IFAC World Congress*, Sydney, 1993.

[114] M. Hovd, R. D. Braatz, and S. Skogestad. SVD controllers for H_2-, H_∞- and μ-optimal control. *Automatica*, 33:433–439, 1997.

[115] F. S. Hover. Application of polynomial chaos in stability and control. *Automatica*, 42:789–795, 2006.

[116] F. S. Hover. Gradient dynamic optimization with Legendre chaos. *Automatica*, 44:135–140, 2008.

[117] J. P. How and S. R. Hall. Connections between the Popov stability criterion and bounds for real parameter uncertainty. In *Proc. of American Control Conference*, pages 1084–1089, San Francisco, CA, 1993.

[118] K. D. Ikramov and N. Savel'eva. Conditionally definite matrices. *Journal of Mathematical Sciences*, 98(1):1–50, 2000.

[119] O. C. Imer, S. Yüskel, and T. Basar. Optimal control of LTI systems over unreliable communication links. *Automatica*, 42:1429–1439, 2006.

[120] S. S. Isukapalli. *Uncertainty Analysis of Transport-Transformation Models*. PhD thesis, The State University of New Jersey, New Brunswick, NJ, 1999.

[121] T. Iwasaki and S. Hara. Well–posedness of feedback systems: Insights into exact robustness analysis and approximate computations. *IEEE Transactions on Automatic Control*, 43:619–630, 1998.

[122] T. Iwasaki and G. Shibata. LPV system analysis via quadratic separator for uncertain implicit systems. *IEEE Transactions on Automatic Control*, 46:1195–1208, 2001.

[123] A. Johnson. The control of fed-batch fermentation processes–A survey. *Automatica*, 23:691–705, 1987.

[124] S. M. Joshi. Failure-accommodating control of large flexible spacecraft. In *Proc. of American Control Conference*, pages 156–161, Seattle, WA, 1986.

[125] Y. Kabashima and D. Saad. The belief in TAP. *Advances in Neural Information Processing Systems 11*, 11:246, 1999.

[126] Y. Kabashima and D. Saad. Belief propagation vs. TAP for decoding corrupted messages. *EPL (Europhysics Letters)*, 44(5):668, 2007.

[127] A. Kalafatis, N. Arifin, L. Wang, and W. R. Cluett. A new approach to identification of pH processes based on the Wiener model. *Chemical Engineering Science*, 50:3693–3701, 1995.

[128] A. Kalafatis, L. Wang, and W. R. Cluett. Linearizing feedforward-feedback control of pH processes based on the Wiener model. *Journal of Process Control*, 15:103–112, 2005.

[129] W. Kaplan. A test for copositive matrices. *Linear Algebra and its Applications*, 313(1):203–206, 2000.

[130] W. Kaplan. A copositivity probe. *Linear Algebra and Its Applications*, 337(1):237–251, 2001.

[131] V. Kariwala, J. F. Forbes, and E. S. Meadows. Integrity of systems under decentralized integral control. *Automatica*, 41:1575–1581, 2005.

[132] F. Keresteciogl u and M. B. Zarrop. Input design for detection of abrupt changes in dynamical systems. *International J. of Control*, 59:1063–1084, 1994.

[133] T. H. Kerr. Real-time failure detection: A static nonlinear optimization problem that yields a two ellipsoid overlap test. *J. Optimization Theory and Applications*, 22:509–535, 1977.

[134] T. H. Kerr. Statistical analysis of a two ellipsoid overlap test for real-time failure detection. *IEEE Transactions on Automatic Control*, 25:762–773, 1980.

[135] T. H. Kerr. False alarm and correct detection probabilities over a time interval for restricted classes of failure detection algorithms. *IEEE Transactions on Information Theory*, 28:619–631, 1982.

[136] H. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 2002.

[137] P. P. Khargonekar and A. Tikku. Randomized algorithms for robust control analysis and synthesis have polynomial complexity. In *IEEE Conference on Decision and Control*, pages 3470–3475, Kobe, Japan, 1996.

[138] V. L. Kharitonov. Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differential'nye Uraveniya*, 14:1483–1485, 1978.

[139] K. K. Kim. Robust control for systems with sector-bounded, slope-restricted, and odd monotonic nonlinearities using linear matrix inequalities. Master's thesis, University of Illinois at Urbana-Champaign, Illinois, USA, 2009.

[140] K.-K. K. Kim and R. D. Braatz. Standard representation and stability analysis of dynamic artificial neural networks: A unified approach. In *Proc. of the IEEE Multi-Conference on Systems and Control*, pages 840–845, Piscataway, NJ: IEEE Press, 2011.

[141] K.-K. K. Kim and R. D. Braatz. Universal approximation with error bounds for dynamic artificial neural network models: A tutorial and some new results. In *Proc. of the IEEE Multi-Conference on Systems and Control*, pages 834–839, Piscataway, NJ: IEEE Press, 2011.

[142] K.-K. K. Kim and R. D. Braatz. Generalized polynomial chaos expansion approaches to approximate stochastic receding horizon control with applications to probabilistic collision checking and avoidance. In *IEEE Multi-Conference on Systems and Control*, pages 350–355, Dubrovnik, Croatia, 2012.

[143] K.-K. K. Kim and R. D. Braatz. Probabilistic analysis and control of uncertain dynamic systems: Generalized polynomial chaos expansion approaches. In *Proc. of American Control Conference*, pages 44–49, Montreal, Canada, 2012.

[144] K.-K. K. Kim and R. D. Braatz. Convex relaxation of sequential optimal input/experiment design for a class of structured large-scale systems: Process gain estimation. In *Proc. of American Control Conference*, 2013. To appear.

[145] K.-K. K. Kim and R. D. Braatz. Optimal input design for system identification via adaptation and receding horizon methods: Semidefinite programming relaxation. under preparation, March 2013.

[146] M. Kishida and R. D. Braatz. Internal model control. In W. S. Levine, editor, *The Control Handbook, 2nd Edition*, pages 9:100–9:123. CRC Press, Boca Raton, Florida, 2011.

[147] V. Klee. Convex sets in linear spaces. *Duke Math. J*, 18(2):443–466, 1951.

[148] F. Knorn, O. Mason, and R. N. Shorten. On linear co-positive Lyapunov functions for sets of linear positive systems. *Automatica*, 45(8):1943–1947, 2009.

[149] M. V. Kothera, V. Balakrishnan, and M. Morari. Robust constrained model predictive control using linear matrix inequalities. *Automatica*, 32:1361–1379, 1996.

[150] M. G. Krein and M. A. Rutman. Linear operators leaving invariant a cone in a Banach space. *Usp. Mat. Nauk.*, 3:3–95, 1948. (English transl.: Amer. Math. Soc. Transl. Ser. I, vol. 10, pp. 199–325, 1962.).

[151] A. Kumar and P. Daoutidis. Feedback control of nonlinear differential-algebraic equation systems. *AlChE J.*, 41:619–636, 1995.

[152] T. L. Lai. Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE Transactions on Information Theory*, 46:595–608, 2000.

[153] T. L. Lai and J. Z. Shan. Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, 44:952–966, 1999.

[154] A. Lambert, D. Gruyer, and G. S. Pierre. A fast Monte Carlo algorithm for collision probability estimation. In *Proc. Int. Conf. Control, Autom., Robot. Vis.*, pages 406–411, 2008.

[155] D. L. Laughlin, M. Morari, and R. D. Braatz. Robust performance of cross directional basis-weight control in paper machines. *Automatica*, 29:1395–1410, 1993.

[156] J. H. Lee, R. D. Braatz, M. Morari, and A. Packard. Screening tools for robust control structure selection. *Automatica*, 31:229–235, 1995.

[157] P. Li and V. Kadirkamanathan. Particle filtering based likelihood ratio approach to fault diagnosis in nonlinear stochastic systems. *IEEE Transactions on Systems Man and Cybernetics Part c–Applications and Reviews*, 31:337–343, 2001.

[158] P. Li, M. Wendt, and G. Wozny. Robust model predictive control under chance constraints. *Computers & Chemical Engineering*, 24:829–834, 2000.

[159] C. Lin, J. Wang, D. Wang, and C. B. Soh. Robustness of uncertain descriptor systems. *Systems & Control Letters*, 31:129–138, 1997.

[160] J. Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proc. Computer-Aided Control System Design Conf. (2004)*, pages 284–289, Piscataway, NJ: IEEE Press, 2004.

[161] D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley & Sons, Inc., New York, NY, 1969.

[162] D. G. Luenberger. Nonlinear descriptor systems. *J. of Economic Dynamics & Control*, 1:219–242, 1979.

[163] D. G. Luenburger. *Introduction to Dynamic Systems: Theory, Models and Applications*. Wiley, New York, 1979.

[164] I. A. Lur'e and V. N. Postnikov. On the theory of stability of control systems. *Applied Mathematics and Mechanics*, 8:246–248, 1944.

[165] B. M., C. M., and K. B. Feedback linearization mpc for discrete-time bilinear systems. In *Proc. of 15th IFAC World Congress*, Barcelona, Spain, 2002.

[166] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, UK, 2003.

[167] O. P. L. Maître and O. M. Kino. *Spectral Methods for Uncertainty Quantification*. Springer, Dordrecht, NY, 2010.

[168] I. R. Manchester. Input design for system identification via convex relaxation. In *IEEE Conference on Decision and Control*, pages 2041–2046, Atlanta, Georgia, 2010.

[169] M. Mariton and P. Bertrand. Improved multiplex control systems: Dynamic reliability and stochastic optimality. *International J. of Control*, 44:219–234, 1986.

[170] O. Mason and R. N. Shorten. On linear copositive Lyapunov functions and the stability of switched positive linear systems. *IEEE Transactions on Automatic Control*, 52:1346–1349, 2007.

[171] I. Masubuchi. Dissipativity inequalities for continuous-time descriptor systems with applications to synthesis of control gains. *Systems & Control Letters*, 55:158–164, 2006.

[172] I. Masubuchi, Y. Katamine, A. Ohara, and N. Suda. $\mathcal{H}_\infty$ control for descriptor systems: A matrix inequalities approach. *Automatica*, 33:669–673, 1997.

[173] J. E. Maxfield and H. Minc. On the matrix equation xx= a. In *Proc. Edinburgh Math. Soc*, volume 13, pages 125–129. Cambridge Univ Press, 1962.

[174] A. Megretski. Necessary and sufficient conditions of stability: A multiloop generalization of the circle criterion. *IEEE Transactions on Automatic Control*, 38:753–756, 1993.

[175] A. Megretski. On the gap between structured singular values and their upper bounds. In *IEEE Conference on Decision and Control*, pages 3461–3462, San Antonio, TX, 1993.

[176] A. Megretski. How conservative is the circle criterion? *Open Problems in Mathematical Systems Theory and Control, V. Blondel, E. Sontag, M. Vidyasagar and J. Willems Editors, Springer-Verlag*, pages 149–151, 1999.

[177] A. Megretski. How conservative is the circle criterion? `http://perso.uclouvain.be/vincent.blondel/books/openprobs/list.html`, 2001.

[178] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42:819–830, 1997.

[179] A. Megretski and S. Treil. Power distribution inequalities in optimization and robustness of uncertain systems. *Journal of Math. Systems, Estimation, Control*, 3:301–319, 1993.

[180] A. N. Michel and D. Liu. *Qualitative Analysis and Synthesis of Recurrent Neural Networks*. Marcel Dekker, New York, 2002.

[181] J. K. Mills and A. A. Goldenberg. Force and position control of manipulators during constrained motion tasks. *IEEE Transactions on Robotics and Automation*, 38:30–46, 1989.

[182] C. C. Moallemi and B. Van Roy. Consensus propagation. *Information Theory, IEEE Transactions on*, 52(11):4753–4766, 2006.

[183] B. C. Moore. Principle component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26:17–32, 1981.

[184] M. Morari. Robust stability of systems with integral control. *IEEE Transactions on Automatic Control*, 30:574–577, 1985.

[185] M. Morari and E. Zafiriou. *Robust Process Control*. Prentice-Hall, Inc., Englewood, Cliffs, NJ, 1989.

[186] T. Motzkin. Copositive quadratic forms. *National Bureau of Standards Report*, 1818:11–12, 1952.

[187] V. V. Murata and E. C. J. Biscaia. Structural and symbolic techniques for automatic characterization of differential-algebraic equations. *Comput. Chem. Eng.*, 21:829–834, 1997.

[188] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[189] K. Murty and S. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[190] Z. K. Nagy and R. D. Braatz. Distributional uncertainty analysis of a batch crystallization process using power series and polynomial chaos expansions. In *Proc. of the 8th IFAC Symposium on Advanced Control of Chemical Processes*, pages 655–660, Gramado, Brazil, 2006.

[191] Z. K. Nagy and R. D. Braatz. Distribution uncertainty analysis using power series and polynomial chaos expansions. *Journal of Process Control*, 17:229–240, 2007.

[192] Z. K. Nagy and R. D. Braatz. Distributed uncertainty analysis using polynomial chaos expansions. In *IEEE Multi-Conference on Systems and Control*, pages 1103–1108, Yokohama, Japan, 2010.

[193] S. Narasimhan and G. Biswas. Model-based diagnosis of hybrid systems. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(3):348–361, 2007.

[194] A. Nemirovski. Advances in convex optimization: conic programming. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 413–444, 2006.

[195] A. Nemirovskii. Several NP-hard problems arising in robust stability analysis. *Mathematics of Control, Signals, and Systems*, 6:99–105, 1993.

[196] Y. Nesterov. Semidefinite relaxation and nonconvex quadratic optimization. *Optimization Methods and Software*, 9:141–160, 1998.

[197] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, USA, 1994.

[198] Y. Nesterov, H. Wolkowicz, and Y. Ye. Nonconvex quadratic optimization. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 361–395. Kluwer Academic Publishers, Boston, MA, 2000.

[199] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[200] H. H. Niemann. A setup for active fault diagnosis. *IEEE Transactions on Automatic Control*, 51:1572–1578, 2006.

[201] H. H. Niemann and J. Stoustrup. Robust fault detection in open loop vs. closed loop. In *IEEE Conference on Decision and Control*, pages 4496–4497, 1997.

[202] R. Nikoukhah, S. L. Campbell, K. G. Horton, and F. Delebecque. Auxiliary signal design for robust multimodel identification. *IEEE Transactions on Automatic Control*, 47:158–164, 2002.

[203] S. J. Norquay, A. Palazoglu, and J. A. Romagnoli. Application of Wiener model predictive control (wmpc) to a pH neutralization experiment. *IEEE Transactions on Control System Technology*, 7:437–445, 1999.

[204] O. D. I. Nwokah and R. Perez. On multivariable stability in the gain space. *Automatica*, 27:975–983, 1991.

[205] B. Ogunnaike and R. A. Wright. Industrial applications of nonlinear control. In *Fifth Int. Conf. on Chemical Process Control (1997)*, pages 46–59, 1997.

[206] R. Olfati-Saber, E. Franco, E. Frazzoli, and J. Shamma. Belief consensus and distributed hypothesis testing in sensor networks. *Networked Embedded Sensing and Control*, pages 169–182, 2006.

[207] M. Ono and B. C. Williams. Iterative risk allocation: A new approach to robust model predictive control with a joint chance constraint. In *IEEE Conference on Decision and Control*, pages 3427–3432, Cancun, Mexico, 2008.

[208] A. Packard and J. Doyle. The complex structured singular value. *Automatica*, 29:71–109, 1993.

[209] A. K. Packard. *Whats new with μ structured uncertainty in multivariable control*. PhD thesis, University of California, Berkeley, California, 1988.

[210] G. Pajunen. Adaptive control of Wiener type nonlinear systems. *Automatica*, 28:781–785, 1992.

[211] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Dover Pub., Inc., Mineola, NY, 1998.

[212] R. S. Patwardhan, S. Lakshminarayanan, and S. L. Shah. Constrained nonlinear MPC using Hammerstein and Wiener models: PLS framework. *AIChE J.*, 44:1611–1622, 1998.

[213] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

[214] D. Peleg, G. Schechtman, and A. Wool. Approximating bounded $0-1$ integer linear programs. In *Proc. 2nd Israel Symp. on the Theory of Computing Systems*, pages 69–77, Netanya, Israel, 1993.

[215] S. Poljak and J. Rohn. Checking robust non-singularity is NP-hard. *Mathematics of Control, Signals, and Systems*, 6:1–9, 1993.

[216] V. Ramadesigan, R. N. Methekar, V. R. Subramanian, F. Latinwo, and R. D. Braatz. Optimal porosity distribution for minimized Ohmic drop across a porous electrode. *J. Electrochem. Soc.*, 157:A1328–A1334, 2010.

[217] A. Rantzer and A. Megretski. System analysis via integral quadratic constraints. In *IEEE Conference on Decision and Control*, pages 3062–3067, Lake Buena Vista, FL, 1994.

[218] A. Rehm and F. Allgöwer. $\mathcal{H}_\infty$ control of descriptor systems with norm-bounded uncertainties in the system matrices. In *Proc. of American Control Conference*, pages 3244–3248, 2000.

[219] D. Rivera, H. Lee, M. Braun, and H. Mittelmann. Plant friendly system identification: A challenge for the process industries. In *IFAC SYSID*, pages 917–922, 2003.

[220] C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Motivations to Computational Implementation*. Springer-Verlag, NY, USA, 2001.

[221] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[222] R. T. Rockafellar. Lagrange multipliers and optimality. *SIAM Rev.*, 35:183–238, 1993.

[223] M. Rosenblatt. Remarks on multivariate transformation. *The Annals of Mathematical Statistics*, 23:470–472, 1952.

[224] S. M. Rump. Conservatism of the circle criterion–Solution of a problem posed by A. Megretski. *IEEE Transactions on Automatic Control*, 46:1605–1608, 2001.

[225] E. L. Russell and R. D. Braatz. Model reduction for the robustness margin computation of large scale uncertain systems. *Computers & Chemical Engineering*, 22:913–926, 1998.

[226] E. L. Russell, C. P. H. Power, and R. D. Braatz. Multidimensional realizations of large scale uncertain systems for multivariable stability margin computation. *International J. of Robust and Nonlinear Control*, 7:113–125, 1997.

[227] M. G. Safonov. *Stability and Robustness of Multivariable Feedback Systems*. MIT Press, Cambridge, MA, 1980.

[228] M. G. Safonov. Stability margins of diagonally perturbed multivariable feedback systems. *IEE Proceedings Part D*, 129:252–255, 1982.

[229] M. G. Safonov and M. Athans. Gain and phase margin for multiloop LQG regulators. *IEEE Transactions on Automatic Control*, 22:173–179, 1977.

[230] P. Sanyal and C. N. Shen. Bayes' decision rule for rapid detection and adaptive estimation scheme with space applications. *IEEE Transactions on Automatic Control*, 19:228–231, 1974.

[231] A. Schaft. *$\mathcal{L}_2$-Gain and Passivity Techniques in Nonlinear Control*. Springer-Verlag, London, 2000.

[232] C. Scherer. A full block S-procedure with applications. In *IEEE Conference on Decision and Control*, pages 2602–2607, San Diego, CA, 1997.

[233] C. W. Scherer. LPV control and full block multipliers. *Automatica*, 37:361–375, 2001.

[234] H. Schneider and M. Vidyasagar. Cross-positive matrices. *SIAM J. Numer. Anal.*, 7:508–519, 1970.

[235] A. T. Schwarm and M. Nikolaou. Chance-constrained model predictive control. *AIChE J.*, 45:1743–1752, 1999.

[236] C. J. Seo and B. K. Kim. Robust and reliable h_∞ control for linear systems with parameter uncertainty and actuator failure. *Automatica*, 32:465–467, 1996.

[237] M. M. Seron, J. H. Braslavsky, P. V. Kokotovic, and D. Q. Mayne. Feedback limitations in nonlinear systems: From Bode integrals to cheap control. *IEEE Transactions on Automatic Control*, 44:829–833, 1999.

[238] S. E. Shimony. Finding maps for belief networks is np-hard. *Artificial Intelligence*, 68(2):399–410, 1994.

[239] F. G. Shinskey. *pH and pION Control in Process and Waste Streams*. Wiley, New York, 1973.

[240] D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistics for sequential detection of a change-point. *Annals of Statistics*, 23:255–271, 1995.

[241] D. D. Siljak. On reliability of control. In *IEEE Conference on Decision and Control*, pages 687–694, 1978.

[242] D. D. Siljak. Reliable control using multiple control systems. *International J. of Control*, 31:303–329, 1980.

[243] D. D. Siljak. Decentralized control and comptations: Status and prospects. *Annual Reviews in Control*, 20:131–141, 1996.

[244] M. Simandl and I. Puncochar. Active fault detection and control: unified formulation and optimal design. *Automatica*, 45:2052–2059, 2009.

[245] M. Simandl, I. Puncochar, and J. Kralovec. Rolling horizon for active fault detection. In *IEEE Conference on Decision and Control*, pages 3789–3794, 2005.

[246] K. S. Sin and G. C. Goodwin. *Adaptive Filtering Prediction and Control*. Prentice Hall Publishers, Englewood, NJ, 1984.

[247] S. Skogestad and M. Morari. Robust performance of decentralized control systems by independent designs. *Automatica*, 29:119–125, 1989.

[248] S. Skogestad, M. Morari, and J. C. Doyle. Robust control of ill-conditioned plants: High purity distillation. *IEEE Transactions on Automatic Control*, 33:1092–1105, 1988.

[249] S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control*. John Wiley & Sons Ltd, Southern Gate, Chichester, UK, 2005.

[250] M. Soroush and K. R. Muske. Analytical model predictive control. In F. Allgower and A. Zheng, editors, *Nonlinear Model Predictive Control*, pages 163–179. Springer, Upper Saddle River, NJ, 2000.

[251] A. G. Sparks and D. S. Bernstein. The scaled Popov criterion and bounds for the real structured singular value. In *IEEE Conference on Decision and Control*, pages 2998–3002, Lake Buena Vista, FL, 1994.

[252] H. Stark and J. W. Woods. *Probability and Random Processes with Applications to Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2002.

[253] R. F. Stengel and L. R. Ray. Stochastic robustness of linear time-invariant control systems. *IEEE Transactions on Automatic Control*, 36:82–87, 1991.

[254] B. L. Stevens and F. L. Lewis. *Aircraft Modeling, Dynamics and Control*. Wiley, New York, 1991.

[255] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones, 1999.

[256] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky. Embedded trees: Estimation of gaussian processes on graphs with cycles. *Signal Processing, IEEE Transactions on*, 52(11):3136–3150, 2004.

[257] Z. Szabo, P. Gaspar, and J. Bokor. Reference tracking for wiener systems using dynamic inversion. In *Proc. of the Joint Conf. of the 20th IEEE International Symposium on Intelligent Contol and 13th Mediterranean Conf. on Control and Automation*, pages 1196–1200, 2005.

[258] K. Takaba. Robust $\mathcal{H}_2$ control of descriptor system with time-varying uncertainty. *International J. of Control*, 71:559–579, 1998.

[259] X. L. Tan, D. D. Siljak, and M. Ikeda. Reliable stabilization via factorization methods. *IEEE Transactions on Automatic Control*, 37:1786–1791, 1992.

[260] S. Tarbouriech and C. Burgat. Positively invariant sets for continuous-time systems with cone-preserving property. *International J. of Systems & Science*, 24:1037–1047, 1993.

[261] S. Tarbouriech and C. Burgat. Positively invariant sets for constrained continuous time systems with cone properties. *IEEE Transactions on Automatic Control*, 39:401–405, 1994.

[262] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Springer-Verlag, London, UK, 2005.

[263] Y. Tipsuwan and M.-Y. Chow. Control methodologies in networked control systems. *Control Engineering Pracice*, 11:1099–1111, 2003.

[264] N. E. D. Toit and J. W. Burdick. Probabilistic collision checking with chance constraints. *IEEE Transactions on Robotics*, 27:809–815, 2011.

[265] O. Toker. On the conservatism of upper bound tests for structured singular value analysis. In *IEEE Conference on Decision and Control*, pages 1295–1300, Kobe, Japan, 1996.

[266] O. Toker and H. Özbay. On the NP-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback. In *Proc. of American Control Conference*, pages 2525–2526, Seattle, WA, 1995.

[267] O. Toker and H. Özbay. On the complexity of purely complex μ computation and related problems in multidimensional systems. *IEEE Transactions on Automatic Control*, 43:409–414, 1998.

[268] S. Treil. The gap between complex structured singular value μ and its upper bound is infinite. Technical report, MIT LIDS, Cambridge, MA, 2000.

[269] D. H. van Hessem and O. H. Bosgra. Closed-loop stochastic model predictive control in a receding horizon implementation on a continuous polymerization reactor example. In *Proc. of American Control Conference*, pages 914–919, Boston, 2004.

[270] J. G. VanAntwerp and R. D. Braatz. A tutorial on linear and bilinear matrix inequalities. *Journal of Process Control*, 10:363–385, 2000.

[271] J. G. VanAntwerp, R. D. Braatz, and N. V. Sahinidis. Robust nonlinear control of plasma etching. In *Proc. of Electrochem. Soc. (1997)*, pages 454–462, 1997.

[272] J. G. VanAntwerp, R. D. Braatz, and N. V. Sahinidis. Globally optimal robust process control. *Journal of Process Control*, 9:374–383, 1999.

[273] J. G. VanAntwerp, A. P. Featherstone, B. A. Ogunnaike, and R. D. Braatz. Cross-directional control of sheet and film processes. *Automatica*, 43:191–211, 2007.

[274] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, 1995.

[275] K. R. Varshney and L. R. Varshney. Quantization of prior probabilities for hypothesis testing. *Signal Processing, IEEE Transactions on*, 56(10):4553–4562, 2008.

[276] S. A. Vavasis. On approximation algorithms for concave programming. In *Recent advances in global optimization*, pages 3–18. Princeton University Press, Princeton, NJ, 1992.

[277] R. J. Veillette, J. V. Medanic, and W. R. Perkins. Design of reliable control systems. *IEEE Transactions on Automatic Control*, 37:290–304, 1992.

[278] G. C. Verghese, B. C. Lévy, and T. Kailath. A generalized state-space for singular systems. *IEEE Transactions on Automatic Control*, 26:811–831, 1981.

[279] M. Vidyasagar. *Nonlinear Systems Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[280] M. Vidyasagar. *A Theory of Learning and Generalization: with Application to Neural Networks and Control Systems*. Springer-Verlag, London, UK, 1997.

[281] M. Vidyasagar. Statistical learning theory and randomized algorithms for control. *IEEE Control Systems Magazine*, 18:69–85, 1998.

[282] M. Vidyasagar and V. D. Blondel. Probabilistic solutions to some NP-hard matrix problems. *Automatica*, 37:1397–1405, 2001.

[283] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[284] G. C. Walsh, Y. Hong, and L. G. Bushnell. Stability analysis of networked control systems. *IEEE Transactions on Control System Technology*, 10:438–446, 2002.

[285] H.-S. Wang, C.-F. Yung, and F.-R. Chang. Bounded real lemma and $\mathcal{H}_\infty$ control for descriptor systems. *IEE Proceedings Part D*, 145:316–322, 1998.

[286] Y. Wang. *Robust Model Predictive Control*. PhD dissertation, University of Wisconsin-Madison, 2002.

[287] M. Welling and Y. W. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 554–561. Morgan Kaufmann Publishers Inc., 2001.

[288] N. Wiener. The homogeneous chaos. *American J. of Mathematics*, 60:897–936, 1938.

[289] J. C. Willems. Dissipative dynamical systems–Part I: General theory. *Arch. Ration. Mech. An.*, 45:321–351, 1972.

[290] A. S. Willsky. A survey of design methods for failure detection in dynamic systems. *Automatica*, 12:601–611, 1976.

[291] A. S. Willsky and H. L. Jones. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21:108–112, 1976.

[292] R. A. Wright and C. Kravaris. On-line identification and nonlinear control of an industrial pH process. *Journal of Process Control*, 11:361–374, 2001.

[293] D. Xiu. *Numerical Methods for Stochastic Computations : A Spectral Method Approach*. Princeton University Press, Princeton, NJ, 2010.

[294] D. Xiu and G. E. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24:619–644, 2002.

[295] V. A. Yakubovich. S-procedure in nonlinear control theory. *Vestnik Leningrad University*, 2(7):62–77, 1971. (English translation in Vestnik Leningrad Univ. 4:73–93, 1977).

[296] Y. Yan and R. R. Bitmead. Incorporating state estimation into model predictive control and its application to network traffic control. *Automatica*, 41:595–604, 2005.

[297] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13, 2001.

[298] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[299] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[300] E. L. Yip and R. F. Sincovec. Solvability, controllability, and observability of continuous descriptor systems. *IEEE Transactions on Automatic Control*, 26:702–707, 1981.

[301] G. Zames. On the input-output stability of time-varying nonlinear feedback systems, Part I: Conditions using concepts of loop gain, conicity, and positivity. *IEEE Transactions on Automatic Control*, 11:228–238, 1966.

[302] G. Zames. On the input-output stability of time-varying nonlinear feedback systems, Part II: Conditions involving circles in the frequency plane and sector nonlinearities. *IEEE Transactions on Automatic Control*, 11:465–476, 1966.

[303] L. Q. Zhang, J. Lam, and S. Y. Xu. On positive realness of descriptor systems. *IEEE Transactions Circuits Systems–I: Fundamental Theory and Applications*, 49:401–407, 2002.

[304] W. Zhang, M. S. Branicky, and S. Phillips. Stability of networked control systems. *IEEE Control Systems Magazine*, 21:84–99, 2001.

[305] X. J. Zhang. Auxiliary signal design in fault detection and diagnosis. In *Lecture Notes in Control and Information Sciences, Vol. 134*. Springer-Verlag, Berlin, 1989.

[306] Q. Zhao and J. Jiang. Reliable state feedback control system design against actuator failures. *Automatica*, 34:1267–1272, 1998.

[307] Q. Zheng. *A Volterra Series Approach to Nonlinear Process Control and Control-Relevant Identification*. Ph.d. dissertation, Univ. of Maryland, College Park, 1995.

[308] K. Zhou, J. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice–Hall, Upper Saddle River, NJ, 1996.

[309] K. Zhou and J. C. Doyle. *Essentials of Robust Control*. Prentice-Hall, Upper Saddle River, NJ, 1998.

[310] S.-S. Zhu, D. Li, and S.-Y. Wang. Risk control over bankruptcy in dynamic portfolio selection: A generalized mean-variance formulation. *IEEE Transactions on Automatic Control*, 49:447–457, 2004.

[311] A. Zolghadri. An algorithm for real-time failure detection in Kalman filters. *IEEE Transactions on Automatic Control*, 41:1537–1539, 1996.

[312] A. Zolghadri, B. Bergeon, and M. Monison. A two ellipsoid overlap test for on-line failure detection. *Automatica*, 29:1517–1522, 1993.