A Robust Control Perspective on Optimization of Strongly-Convex Functions

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Bin Hu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Peter J. Seiler, advisor

July, 2016

Acknowledgements

My greatest gratitude goes to my advisor, Peter Seiler. I really want to thank him for being the best advisor I have ever met in my life. It is a huge honor for me to work with such exceptional mind. Peter is a great person and also a closed friend. I am sincerely thankful for his advice and help.

I also want to thank Gary Balas. Gary is such a great leader. He influenced me in a lot of positive ways. I completely changed my presentation style due to his comments. I am always grateful to him.

I would like to thank Anders Rantzer for supportively collaborating with me and showing me the beauty of positive systems. I want to thank Tom Luo for teaching me so much on optimization. His insights are deep and inspiring. Several of my current research topics are motivated by Tom, and I am really grateful for that. Special thanks to Laurent Lessard. His work has huge impacts on my research and inspires the whole story of my Ph.D dissertation. I also want to thank Laurent for collaborating with me and offering me the opportunity to visit the optimization group in Madison.

Thanks to Demoz Gebre-Egziabher, Maziar S. Hemati, and Gongjun Xu for serving on my doctoral committee. I would also like to acknowledge Tryphon T. Georgiou, Mihailo Jovanovic, and Sairaj Dhople for their help during my Ph.D study. I want to thank Sei Zhen Khong, Harald Pfifer, Raghu Venkataraman, and Shu Wang for many useful discussions on control theory. I should also acknowledge Yongxin Chen for a lot of fun discussions on mathematics.

Finally I want to thank all my family and friends for their love and support. My sincere gratitude goes to my aunt, Fuqin Zhou. She took good care of me when I was a child. I am always grateful for that. I owe thanks to my dearest Yingying Luo for the color and warmth she brings to my life.

Dedication

To My Mom Yanqin Zhou

Abstract

Large-scale optimization is a central topic in big data science. First-order black-box optimization methods have been widely applied in machine learning problems, since the oracle complexity of these methods can be independent of the parameter dimension. In this dissertation, we formulate linear matrix inequality (LMI) conditions to analyze the convergence rates of various deterministic and stochastic optimization methods. We derive these LMIs using integral quadratic constraints (IQCs) and dissipation inequalities.

The first part of this dissertation analyzes deterministic first-order methods (gradient descent, Nesterov's method, etc) as generalized eigenvalue problems (GEVPs). A standard dissipation inequality requires a non-negative definite storage function and "hard" IQCs which must hold over all finite time horizons. We develop a modified dissipation inequality that requires neither non-negative definite storage functions nor hard IQCs. Then we show that linear rate analysis of a given deterministic first-order method is equivalent to uniform stability analysis of a related scaled system. This enables derivation of linear rate analysis conditions using standard IQCs for a scaled operator. A soft Zames-Falb IQC is derived and used in the modified dissipation inequality, leading to a GEVP formulation for linear rate analysis of first-order optimization methods.

In the second part of this dissertation, we extend the IQC framework to analyze stochastic optimization methods which have been widely applied in empirical risk minimization and machine learning problems. We first combine jump system theory with IQCs to derive LMI conditions for rate analysis of the stochastic average gradient (SAG) method and its variants (SAGA, etc). The resultant LMI conditions can be used to analyze the convergence rates of SAG, SAGA, and other related variants with uniform or non-uniform sampling strategies. Then we develop LMI conditions to analyze the stochastic gradient (SG) method and its variants. The SG method with a constant stepsize typically achieves a linear convergence rate only up to some fixed tolerance. We develop stochastically averaged quadratic constraints with disturbance terms quantifying the inaccuracy of the SG method. Several known results about the SG method have been recovered using our proposed LMI conditions. We also obtain new results regarding the convergence of the SG method under different conditions.

Contents

Acknowledgements **i**

Dedication **ii**

Abstract **iii**

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**

1.1 Outline and Contributions 3

2 Background **6**

2.1 Notation 6

2.2 Basic Facts about Gradients of Convex Functions 8

2.3 Review of First-Order Optimization Methods 9

2.4 Operator Theory 10

2.5 Linear State-Space Models 12

2.6 Feedback Interconnection and Stability Concepts 15

2.7 Integral Quadratic Constraints 17

2.8 Uniform Stability Analysis Using IQCs 21

2.9 ρ-Hard IQCs and ρ-Exponential Stability Analysis 26

2.10 Linear Rate Analysis of Optimization Methods 29

2.11 Related Operators and Existing IQCs 31

2.11.1 Memoryless Nonlinearity in a Sector 31
2.11.2 Static Nonlinearity 32
2.11.3 Gradients of Smooth Strongly-Convex Functions 33
2.11.4 Multiplication with an Uncertain Parameter 35
2.11.5 Time Delay 36

3 Hard Factorizations of Frequency Domain IQCs 38
3.1 J-Spectral Factorizations and Related Games 40
3.1.1 Open-loop Dynamic Games and IQC Factorizations 40
3.1.2 Nash Equilibrium for the Two-Player Game 44
3.1.3 J-Spectral Factorization for Strict-PN Multipliers 47
3.2 Modified Dissipation Inequality 50
3.3 Related Work 52

4 Linear Rate Analysis Using Internal Uniform Stability Tests 53
4.1 Equivalence between ρ-Exponential Stability Analysis and Uniform Stability Analysis 54
4.2 Boundedness and IQCs for Scaled Perturbation 56
4.2.1 Scaled Operator for Memoryless Nonlinearity in a Sector 57
4.2.2 Scaled Operator for Static Nonlinearity 57
4.2.3 Scaled Operator for Gradients of Smooth Strongly-Convex Functions 61
4.2.4 Scaled Operator for Multiplicative Perturbation 62
4.2.5 Scaled Operator for Delay 63
4.3 Equivalence between ρ-Hard IQCs on Δ and Hard IQCs on Δ_ρ 63
4.4 A GEVP Formulation for Linear Rate Analysis of Deterministic First-Order Optimization Methods 65
4.5 Numerical Example: Analysis of Nesterov's Accelerated Method 68

5 Analysis for SAG and Related Variants Using IQCs and Jump System Formulations 72
5.1 Background of Dynamic Jump Systems 74
5.2 IQC-Based Analysis for SAG 75

5.2.1 A Jump System Formulation for SAG 75
5.2.2 ρ-Hard IQCs on Δ_f 78
5.2.3 Convergence Rate Analysis of SAG Using Semidefinite Programs 80
5.2.4 Numerical Results on SAG 82
5.3 Generalizations for Variants of SAG 85
5.3.1 Variants of SAG 85
5.3.2 Jump System Formulations for Variants of SAG 86
5.3.3 Numerical Results on SAGA 87
5.3.4 Further Discussions 88
5.3.5 Remarks on Increment Aggregated Gradient 89

6 IQC Analysis of Stochastic Gradient with a Constant Stepsize 91
6.1 An IQC-Based Proof for Proposition 1 93
6.1.1 Stochastic Quadratic Constraints 93
6.1.2 Recovery of Proposition 1 95
6.2 IQC Analysis for SG with General Cost Functions 97
6.2.1 A General Construction of Stochastic Quadratic Constraints . . 97
6.2.2 Analysis Results 100
6.3 Robustness of SG with respect to Deterministic Noise 105
6.4 A General Analysis Framework for Variants of SG 107

7 Conclusions and Future Work 110

References 114

Appendix A. IQC Multipliers and DARE Stabilizing Solutions 124

List of Tables

4.1 Various Numerical Rate Results for Nesterov's Accelerated Method . . . 71

5.1 Values of $\left(1 - \frac{1}{cn}\right)^n$ for $c \in \{8, 6, 2, 1.5, 1\}$ 83

5.2 Numerical Rate Results for SAG with $m = 1$, $m_i = 0$, $L_i = L = 10$, and $\alpha = \frac{1}{16L}$ 84

5.3 Numerical Rate Results for SAG with $m = 1$, $m_i = 0$, $L_i = L = 100$, and $\alpha = \frac{1}{16L}$ 85

5.4 Numerical Rate Results for SAGA with $m = 1$, $m_i = 0$, $L_i = L = 10$, and $\alpha = \frac{1}{3L}$ 88

List of Figures

2.1 The Block-Diagram for an Operator Δ 10
2.2 The Block-Diagram for the Composition Operator $\Delta_1 \circ \Delta_2$ 11
2.3 The Block-Diagram for a Dynamic System G 13
2.4 The Block-Diagram for Series Connection $G_2 \circ G_1$ 14
2.5 The Block-Diagram for Feedback Interconnection $F_u(G, \Delta)$ 16
2.6 Graphical Interpretation for Time Domain IQCs 19
2.7 Uncertain LTI System Extended to Include Filter Ψ 22
4.1 Transformed Interconnection 54
5.1 Removing Δ by Enforcing a Constraint on the Output of Ψ 80
5.2 The Block-Diagram for Variants of SAG 86
5.3 The Block-Diagram for Reducing-Order Modeling 89

Chapter 1

Introduction

Large-scale optimization has become an important topic for machine learning research and big data science [1]. Many machine learning problems, e.g. ridge regression [2], support vector machine [3], logistic regression [4], LASSO [5], matrix completion [6], etc, can essentially be formulated as convex optimization problems. A detailed review can be found in [7, Chapter 1.2]. In principle, convexity can be exploited to obtain global optimum solutions for such problems [8]. The real difficulty is posed by the scale of these optimization problems associated with the big data applications. The so-called black-box methods have become popular in such large-scale optimization since the oracle complexity of these methods can be free of the problem dimension [7, Section 1.4]. Various first-order black-box methods have been proposed and applied in practice.

In this dissertation, we focus on the performance of first-order black-box methods when applied to optimization of strongly-convex functions. Several popular optimization problems with strongly-convex objective functions include ridge regression [7], ℓ_2-regularized logistic regression [9], and smooth support vector machine [10]. First-order methods for such problems can be either deterministic or stochastic. Several commonly-used deterministic first-order methods are the gradient descent method, Nesterov's accelerated method and the Heavy-ball method [11, 12]. Several examples for stochastic first-order methods include the stochastic gradient (SG) method [13], the stochastic average gradient (SAG) method [14, 15], and the SAGA method [16]. Although there exist convergence rate analysis for all these methods, the standard proof techniques are developed in a case-by-case manner. There lacks a unified framework which can

automate the analysis for more complicated first-order methods, when subtle design trade-offs between accuracy, speed, memory size, and robustness need to be carefully addressed.

Recently, semidefinite programs have been used to certify the convergence rates of deterministic first-order methods [12,17,18]. In [12], a general analysis for deterministic first-order optimization methods (the gradient descent method, Nesterov's accelerated method, the Heavy-ball method, etc) is provided by adapting the integral quadratic constraint (IQC) framework [19] from robust control theory. Linear matrix inequality (LMI) conditions are formulated using the IQC framework, and numerically tested to obtain the convergence rate bounds of various optimization methods. For the first-order methods considered in [12], the convergence rates obtained by the IQC analysis outperform or at least recover the existing rates in the literature.

The key insight in [12] is that many deterministic first-order methods can be viewed as interconnections of a linear time-invariant dynamic system and a static nonlinearity. This type of interconnection has been studied extensively in the control community [20]. A large body of the controls literature even studies the more general interconnection structure which consists of a linear system and a troublesome perturbation [21,22]. IQCs provide a general framework for analysis of such interconnections [19]. The IQC framework builds on a long history of classical multiplier results, e.g. the use of Zames-Falb multipliers [20, 23]. The original IQC theory developed in [19] addresses input-output stability based on homotopy arguments and frequency domain inequalities. The original IQC theory can be used to prove linear convergence but does not provide an accurate estimate/bound for the convergence rate [24]. Similar results have been obtained in [25,26] based on connections between IQCs and dissipativity theory [27,28]. Significant progress has been made in [12, 29] by developing new notions of IQCs which are specifically tailored for linear rate analysis of optimization methods. The work in [12,29] is built upon the classical results in [19, 27, 28, 30].

The main advantage of the IQC framework in [12] is that the analysis can be automated for different deterministic first-order methods. More specifically, the LMI conditions for different first-order methods are derived in the same way. Notice the "answer" given by the IQC analysis is an LMI condition, whose feasible set is convex and can be efficiently searched using standard solvers. Hence in general, the IQC approach is

subject to numerical errors. However, due to the increasing power of modern computers, the LMI methods could potentially lead to numerical solutions for many problems which cannot be handled by analytical derivations.

In this dissertation, we follow the direction in [12]. We develop LMI conditions for linear rate analysis of various deterministic and stochastic first-order methods. All the LMI conditions in this dissertation are derived using IQCs and dissipation inequalities. Although the dissipation inequalities for deterministic first-order methods and stochastic optimization algorithms are not completely the same, the underlying machinery is always the Lyapunov theory. The derivations of the LMI conditions in this dissertation adopt repeatable patterns. Consequently, our proposed IQC analysis can be automated for more complicated first-order methods.

1.1 Outline and Contributions

Chapter 2 presents the notation and required background materials from both the optimization and controls literature. Several commonly-used first-order methods are introduced. The standard dissipativity-based IQC approach for uniform stability analysis is summarized, and the ρ-hard IQC approach for linear rate analysis of deterministic first-order methods is reviewed. Existing IQCs for various operators are briefly discussed.

Chapter 3 focuses on the IQC theory for uniform stability analysis of feedback interconnections. The standard dissipation inequality approach for uniform stability analysis requires a non-negative definite storage function and "hard" IQCs. The term "hard" means that the IQCs must hold over all finite time horizons. Chapter 3 presents a J-spectral factorization result for hard IQC constructions. Several lemmas regarding discrete-time IQC factorizations and a related open-loop linear quadratic difference game are obtained to support the proof of this main J-spectral factorization result. Then, the J-spectral factorization is applied to prove a discrete-time modified dissipation inequality result. The modified dissipation inequality removes the constraints of non-negative storage functions and hard IQCs that appear in standard dissipation inequalities. This allows more general IQC parameterizations [31, 32], leading to more flexible LMI formulations for uniform stability analysis.

In Chapter 4, we first show that linear rate analysis of a deterministic first-order

optimization method is equivalent to uniform stability analysis of a related scaled system. This enables derivation of linear rate analysis conditions from uniform stability tests using standard IQCs on a scaled operator. A list of IQCs is derived for various scaled operators using the detailed information of the original operators. Connections between the proposed approach and the existing ρ-hard IQC approach are discussed. Then a new soft Zames-Falb IQC is derived and embedded into a modified dissipation inequality, yielding a generalized eigenvalue problem (GEVP) formulation [33] for linear rate analysis of deterministic first-order methods.

In Chapter 5, we extend the IQC framework to analyze the convergence rates of SAG and its variants with arbitrary constant stepsizes and possibly non-uniform sampling strategies. We combine jump system theory with IQCs to derive sufficient conditions, which can be used to certify the convergence rates of SAG and its variants. The derived conditions can be checked by numerically solving semidefinite programs. Based on these conditions, we obtain new numerical upper bounds on the convergence rates of SAG and SAGA.

Chapter 6 develops the concept of stochastically averaged quadratic constraints to formulate LMI conditions for analysis of the SG method and its variants. The SG method with a constant stepsize converges linearly only up to a tolerance. The stepsize selection involves a trade-off between the convergence rate and the computation accuracy. To capture this trade-off, we develop stochastically averaged quadratic constraints with disturbance terms for the stochastic gradient operator. The disturbance terms in the resultant constraints are included as hidden energy in the dissipation inequality to quantify the inaccuracy of the SG method. Several known and new results about the SG method have been derived using the proposed LMI conditions.

The contributions of this dissertation are summarized as follows.

- A J-spectral factorization result is proved to construct hard, discrete-time IQCs. The J-spectral factorization result is also applied to prove a discrete-time modified dissipation inequality that requires neither non-negative storage function nor hard IQCs.

- Linear rate analysis of a deterministic first-order method is shown to be equivalent to uniform stability analysis of a related scaled system. Consequently, a GEVP

formulation for linear rate analysis of deterministic optimization methods has been obtained using the modified dissipation inequality and a new soft Zames-Falb IQC for a scaled nonlinear operator.

- LMI conditions for linear rate analysis of SAG and related variants with uniform or non-uniform sampling strategies have been formulated by combining jump system theory and IQCs. The derived conditions are numerically solved via semidefinite programing, and new numerical rate bounds for SAG and SAGA are obtained.

- LMI conditions for analysis of the SG method and its variants have been proposed using the concept of averaged quadratic constraints. Several known and new results about the SG method have been derived using the proposed LMI conditions.

Chapter 2

Background

We summarize the required background in this section.

2.1 Notation

The set of non-negative integers is denoted as $\mathbb{Z}^+$. The set of p-dimensional real vectors is denoted as $\mathbb{R}^p$. The set of complex numbers is denoted as $\mathbb{C}$.

The $p \times p$ identity matrix and the $p \times p$ zero matrix are denoted as I_p and 0_p, respectively. The $n \times n$ identity matrix is denoted as I_n. Let e_i denote the n-dimensional vector whose entries are all 0 except the i-th entry which is 1. Let e denote the n-dimensional vector whose entries are all 1. Let $\tilde{0}$ denote the n-dimensional vector whose entries are all 0. For simplicity, 0 is occasionally used to denote a zero vector or a zero matrix when there is no confusion on the dimension. The Kronecker product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$ is denoted by $A \otimes B \in \mathbb{R}^{(mp) \times (nq)}$. Notice $(A \otimes B)^T = A^T \otimes B^T$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ when the matrices have the compatible dimensions. A square matrix is said to be Schur stable if all of its eigenvalues are strictly inside the unit circle. When a matrix P is negative semidefinite (definite), we will use the notation $P \leq (<) 0$. Similarly, when P is positive definite, we will use the notation $P > 0$.

Consider a (real) sequence $u := (u^0, u^1, \ldots)$. This sequence is said to be in ℓ_{2e}^p if $u^k \in \mathbb{R}^p$ for all k. In addition, this sequence is said to be in ℓ_2^p if $\sum_{k=0}^{\infty} \|u^k\|^2 < \infty$ where $\|u^k\| := (u^k)^T u^k$ denotes the standard (vector) 2-norm of u^k. For simplicity, the

superscripts of ℓ_{2e}^p and ℓ_2^p may be omitted when there is no confusion. The ℓ_2-norm for $u \in \ell_2^p$ is defined as $\|u\|^2 := \sum_{k=0}^{\infty} \|u^k\|^2$. An inner product on ℓ_2^p is defined as $\langle u, v\rangle = \sum_{k=0}^{\infty} (u^k)^T v^k$ for any $u, v \in \ell_2^p$.

A continuously differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ has Lipschitz-continuous gradients with parameter L if the following inequality holds for all $x, y \in \mathbb{R}^p$

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|. \tag{2.1}$$

The continuously differentiable function g is said to be strongly-convex with parameter $m > 0$ if the following inequality holds for all $x, y \in \mathbb{R}^p$

$$g(x) \geq g(y) + \nabla g(y)^T (x - y) + \frac{m}{2}\|x - y\|^2. \tag{2.2}$$

Notice g is said to be convex if (2.2) holds with $m = 0$. We define $\mathcal{F}(m, L)$ to be the set of continuously differentiable functions $g : \mathbb{R}^p \to \mathbb{R}$ that are strongly convex with parameter m and have Lipschitz gradients with parameter L. We will use $\mathcal{F}(0, L)$ to denote the set of continuously differentiable functions that are convex and have Lipschitz gradients with parameter L.

Let $\mathbb{RL}_\infty$ denote the set of rational functions with real coefficients that have no poles on the unit circle. $\mathbb{RH}_\infty$ is the subset of functions in $\mathbb{RL}_\infty$ that are proper and analytic outside the unit disk of the complex plane. The para-Hermitian conjugate of $\Pi \in \mathbb{RL}_\infty^{m \times n}$, denoted as $\Pi^\sim$, is defined by $\Pi^\sim(z) := \Pi^T(z^{-1})$. Hence $\Pi^\sim(e^{j\omega}) = \Pi^*(e^{j\omega})$ holds on the unit circle. Notice $\mathbb{RL}_\infty$ contains improper functions, e.g. polynomials in z, while $\mathbb{RH}_\infty$ contains only proper functions. Thus functions in $\mathbb{RH}_\infty$ have standard, discrete-time state space representations but descriptor systems are required, in general, to represent functions in $\mathbb{RL}_\infty$ [34]. The use of descriptor systems is limited to one technical result (Lemma 26 in the appendix).

Finally, $DARE(A, B, Q, R, S)$ denotes the following discrete-time Algebraic Riccati Equation (DARE)

$$A^T X A - X - (A^T X B + S)(R + B^T X B)^{-1}(A^T X B + S)^T + Q = 0 \tag{2.3}$$

The stabilizing solution $X = X^T$, if it exists, is such that $(R + B^T X B)$ is nonsingular and $(A - BK)$ is a Schur stable matrix where $K := (R + B^T X B)^{-1}(A^T X B + S)^T$ is the stabilizing DARE gain.

2.2 Basic Facts about Gradients of Convex Functions

We first review several basic facts about gradients of convex functions.

Lemma 1. *Suppose $g \in \mathcal{F}(m, \infty)$ with $m \geq 0$.*

1. *Given arbitrary $x, y \in \mathbb{R}^p$, the following inequality holds:*

$$[\nabla g(x) - \nabla g(y)]^T (x - y) \geq m\|x - y\|^2 \geq 0 \tag{2.4}$$

In addition, if $g \in \mathcal{F}(m, L)$, then

$$[\nabla g(x) - \nabla g(y) - m(x - y)]^T [\nabla g(x) - \nabla g(y) - L(x - y)] \leq 0. \tag{2.5}$$

2. *If $m > 0$, then there exists a unique $x^* \in \mathbb{R}^p$ such that $g(x^*) \leq g(x)$ for all $x \in \mathbb{R}^p$. In addition, x^* is also the only point satisfying $\nabla g(x^*) = 0$. Given arbitrary $x \in \mathbb{R}^p$, the following inequality holds:*

$$\begin{bmatrix} x - x^* \\ \nabla g(x) \end{bmatrix}^T \begin{bmatrix} -2mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} x - x^* \\ \nabla g(x) \end{bmatrix} \geq 0 \tag{2.6}$$

In addition, if $g \in \mathcal{F}(m, L)$, then

$$\begin{bmatrix} x - x^* \\ \nabla g(x) \end{bmatrix}^T \begin{bmatrix} -2mLI_p & (m + L)I_p \\ (m + L)I_p & -2I_p \end{bmatrix} \begin{bmatrix} x - x^* \\ \nabla g(x) \end{bmatrix} \geq 0 \tag{2.7}$$

Proof. Statement 1 has been proved in [11]. Notice (2.4) is a restatement of (2.1.17) in [11], while (2.5) is equivalent to (2.1.24) in [11].

The proof of Statement 2 is briefly sketched as follows. The strong convexity condition (2.2) implies that g is lower bounded by a quadratic function. Hence any sublevel set of g is bounded. The continuity of g implies that g is a proper closed function such that its sublevel sets are closed. Hence the sublevel sets of g are always compact. Based on the well-known Bolzano-Weierstrass Theorem, there exists some x^* satisfying $g(x^*) \leq g(x)$ for all $x \in \mathbb{R}^p$. The convexity of g implies that $\nabla g(x^*) = 0$. The uniqueness of x^* is a direct consequence of Inequality (9.11) in [8]. Finally, we can set $y = x^*$ in (2.4) and (2.5) to obtain (2.6) and (2.7). □

2.3 Review of First-Order Optimization Methods

Consider the unconstrained minimization problem

$$\min_{x\in\mathbb{R}^p} \; g(x) \tag{2.8}$$

where $g \in \mathcal{F}(m, L)$ with $m > 0$. The strong convexity of g guarantees that there exists a unique $x^* \in \mathbb{R}^p$ satisfying $\nabla g(x^*) = 0$. One can solve (2.8) by finding x^*.

A classical way to solve (2.8) is the gradient descent method, which uses the following iteration to find x^*:

$$x^{k+1} = x^k - \alpha \nabla g(x^k) \tag{2.9}$$

Since g is strongly convex, the gradient method with a well-chosen constant stepsize α achieves a linear convergence rate [11]. Specifically, if α is chosen well, then there exists a constant $\rho \in [0, 1]$ and a constant $c \in \mathbb{R}$ such that

$$\|x^k - x^*\|^2 \leq c\rho^{2k}\|x^0 - x^*\|^2 \tag{2.10}$$

For example, we can choose $\alpha = \frac{2}{L+m}$ and obtain $\rho = \frac{L-m}{L+m}$. Another popular choice for α is that $\alpha = \frac{1}{L}$, which leads to the convergence rate $\rho = \sqrt{\frac{L-m}{L+m}}$. These results were formally documented in [11, Theorem 2.1.15].

The gradient descent method can be further accelerated by incorporating memory into the algorithm. Nesterov's accelerated method uses the following iteration rule:

$$\begin{aligned} x^{k+1} &= \zeta^k - \alpha \nabla g(\zeta^k) \\ \zeta^k &= (1+\beta)x^k - \beta x^{k-1} \end{aligned} \tag{2.11}$$

Nesterov's accelerated method with a well-chosen constant stepsize achieves a faster linear convergence rate than the gradient descent method [11]. For example, if we choose $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$, then we can obtain a linear rate $\rho = \sqrt{1 - \sqrt{\frac{m}{L}}}$, which is faster than the rates obtained by the gradient descent method. This fact was stated in [11, Theorem 2.2.3].

Another popular method is the Heavy-ball method, which incorporates a momentum term into the iteration:

$$x^{k+1} = x^k - \alpha \nabla g(x^k) + \beta(x^k - x^{k-1}) \tag{2.12}$$

Although the local convergence rate of the Heavy-ball method is very fast, there is no global linear convergence guarantee for this method.

2.4 Operator Theory

Several basic facts from operator theory are required in this dissertation. Most facts here can be found in [35,36]. An operator is a mapping from one vector space to another. We mainly consider the case where both spaces are ℓ_{2e}, i.e. $\Delta : \ell_{2e} \to \ell_{2e}$.

Now we define the truncation operator $P_N : \ell_{2e} \to \ell_{2e}$, which maps a sequence $u \in \ell_{2e}$ to $v = P_N(u)$ as follows:

$$v^k := \begin{cases} u^k & \text{for } k \leq N \\ 0 & \text{for } k > N \end{cases} \tag{2.13}$$

For simplicity, $P_N(u)$ is occasionally abbreviated as $(u)_N$. It is clear that the extended space ℓ_{2e} is the set of sequences u such that $P_N(u) \in \ell_2$ for all $N \geq 0$. [1]

An operator represents any input-output relationship. In the controls field, a block diagram is usually used to graphically represent this input-output characteristic. For example, Figure 2.1 presents a block-diagram for an operator $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ which maps v to w.

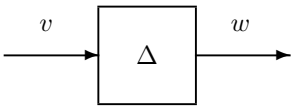

Figure 2.1: The Block-Diagram for an Operator Δ

Let $\Delta_1 : \ell_{2e} \to \ell_{2e}$ and $\Delta_2 : \ell_{2e} \to \ell_{2e}$ be two operators. Then the composition of Δ_1 and Δ_2 is also an operator which maps ℓ_{2e} to ℓ_{2e}. We denote the composition of Δ_1 and Δ_2 as $\Delta_1 \circ \Delta_2$, which satisfies $(\Delta_1 \circ \Delta_2)(v) = \Delta_1(\Delta_2(v))$ for any $v \in \ell_{2e}$. A block diagram for the composition $u = (\Delta_1 \circ \Delta_2)(v)$ is shown in Figure 2.2.

The next fact is also important. Given any $c_1, c_2 \in \mathbb{R}$, then $c_1\Delta_1 + c_2\Delta_2$ is also an operator which maps ℓ_{2e} to ℓ_{2e} via $(c_1\Delta_1 + c_2\Delta_2)(v) = c_1\Delta_1(v) + c_2\Delta_2(v)$.

An operator $\Delta : \ell_{2e} \to \ell_{2e}$ is said to be *causal* if $P_N \circ \Delta \circ P_N = P_N \circ \Delta$ for all $N \geq 0$. A key point stated in [35, Section 2.4] is that the concept of boundedness for

[1] Note that a sequence having a finite escape time in the ℓ_2-norm will have a finite escape time in any other ℓ_p-norm. Hence any ℓ_p space can be extended to generate the same extended space. Here the notation ℓ_{2e} is adopted to emphasize that the norms of the operators on ℓ_{2e} are induced by ℓ_2-norms.

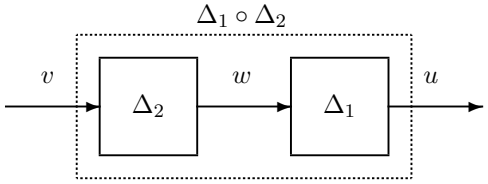

Figure 2.2: The Block-Diagram for the Composition Operator $\Delta_1 \circ \Delta_2$

causal operators on ℓ_{2e} can be naturally defined even though ℓ_{2e} itself is not a normed space. Specifically, a causal Δ is called *bounded* if $\Delta(0) = 0$ and the following holds

$$\|\Delta\| := \sup_{N>0;\|P_N(v)\|_2 \neq 0} \frac{\|(P_N \circ \Delta)(v)\|_2}{\|P_N(v)\|_2} < \infty. \tag{2.14}$$

Notice $c_1 v_1 + c_2 v_2 \in \ell_{2e}$ if $c_1, c_2 \in \mathbb{R}$ and $v_1, v_2 \in \ell_{2e}$. An operator Δ is linear if $\Delta(c_1 v_1 + c_2 v_2) = c_1 \Delta(v_1) + c_2 \Delta(v_2)$ for all $c_1, c_2 \in \mathbb{R}$ and $v_1, v_2 \in \ell_{2e}$. One example for linear bounded operators is the truncation operator P_N.

The triangle inequality for linear bounded operators is well-known. Similarly, the triangle inequality for causal bounded operators also holds. The following two inequalities are useful.

Lemma 2. *Given any two (possibly nonlinear) causal bounded operators $\Delta_0 : \ell_{2e} \to \ell_{2e}$ and $\Delta_1 : \ell_{2e} \to \ell_{2e}$, the following inequality holds:*

$$\begin{aligned} \|\Delta_0 + \Delta_1\| &\leq \|\Delta_0\| + \|\Delta_1\| \\ \|\Delta_1 \circ \Delta_2\| &\leq \|\Delta_1\| \|\Delta_2\| \end{aligned} \tag{2.15}$$

Proof. For any $v \in \ell_{2e}$, we have $P_N(v) \in \ell_2$ and $(P_N \circ \Delta \circ P_N)(v) \in \ell_2$. Then the above lemma can be easily proved using properties of operators on ℓ_2. □

In functional analysis, bounded linear operators are well studied [37, 38]. However, the bounded operators defined here need not be linear. Most bounded operators used in this dissertation have been studied in the robust control literature [19]. Two particularly important operators for analysis of optimization methods are now introduced. Given a function $g \in \mathcal{F}(m, L)$ with $m \geq 0$, we define the operator $\bar{\Delta}_g : \ell_{2e}^p \to \ell_{2e}^p$ that maps $v \in \ell_{2e}^p$ to $w = \bar{\Delta}_g(v)$ as

$$w^k = \nabla g(v^k). \tag{2.16}$$

Since $m \geq 0$, there exists x^* such that $\nabla g(x^*) = 0$. Consequently, $\|w^k\| = \|\nabla g(v^k)\| \leq L\|v^k - x^*\| < \infty$, and $\bar{\Delta}_g$ is a well-defined operator mapping ℓ_{2e}^p to ℓ_{2e}^p. It is straightforward to check that $\bar{\Delta}_g$ is causal. Since w^k is completely determined by v^k and independent of v^t $(t \neq k)$, this operator is said to be static (or memoryless). Notice that $\bar{\Delta}_g$ is not bounded in general since $\nabla g(0)$ may not be equal to 0. On the other hand, a shifted version of $\bar{\Delta}_g$ (denoted as Δ_g) is bounded. The operator $\Delta_g : \ell_{2e}^p \to \ell_{2e}^p$ is defined to map $v \in \ell_{2e}^p$ to $w = \Delta_g(v)$ as

$$w^k = \nabla g(v^k + v^*) - \nabla g(v^*) \tag{2.17}$$

where v^* is an arbitrary referencing point. We can verify that $\Delta_g(0) = 0$ and $\|w^k\| \leq L\|v^k\|$ such that Δ_g is a bounded operator satisfying $\|\Delta_g\| \leq L$. When v^* satisfies $\nabla g(v^*) = 0$, Equation (2.17) becomes $w^k = \nabla g(v^k + v^*)$, which can also be constructed from (2.16) by shifting v^k to $(v^k + v^*)$. Hence Δ_g can be viewed as a shifted version of $\bar{\Delta}_g$ in this case. The connection between $\bar{\Delta}_g$ and Δ_g is important for the analysis in this dissertation.

A discrete-time state-space model with a known initial condition can be viewed as an operator. Section 2.5 reviews the required background on linear state-space models. Several other operators which are useful for robustness analysis of optimization methods are summarized in Section 2.11.

2.5 Linear State-Space Models

Now we briefly review some basic concepts regarding dynamic system theory. Let a dynamic system G be governed by a linear state-space model, which is described by the following recursive iteration:

$$\begin{aligned} \xi^{k+1} &= A\xi^k + Bw^k \\ v^k &= C\xi^k + Dw^k \end{aligned} \tag{2.18}$$

where $\xi^k \in \mathbb{R}^{n_\xi}$, $w^k \in \mathbb{R}^{n_w}$, $v^k \in \mathbb{R}^{n_v}$, $A \in \mathbb{R}^{n_\xi \times n_\xi}$, $B \in \mathbb{R}^{n_\xi \times n_w}$, $C \in \mathbb{R}^{n_v \times n_\xi}$, and $D \in \mathbb{R}^{n_v \times n_w}$. In the controls literature, the step k is typically denoted as a subscript. Here we adopt the convention in the optimization literature and write k as a superscript. At each step k, the variables ξ^k, w^k, and v^k are referred to as the state, input, and

output of the system G. When the initial condition ξ^0 is given, one can use the state space model (2.18) to determine the state ξ and the output v for any given input sequence w. The state-space model (2.18) for G is linear in the sense that G becomes a linear operator given zero initial conditions ($\xi^0 = 0$). The dynamic system (2.18) is completely determined by the matrices A, B, C, and D. We say that G is determined by (A, B, C, D) or equivalently

$$G := \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]. \tag{2.19}$$

In the simplest case, A, B, C, and D are constant matrices which do not change with k. Then the dynamic system G is said to be linear time-invariant (LTI), and has a frequency domain representation $G = C(zI - A)^{-1}B + D$. A thorough treatment of the frequency domain characterizations of LTI systems can be found in [39]. It is worth mentioning that the poles of G are the eigenvalues of A. Hence $G \in \mathbb{RL}_\infty^{n_v \times n_w}$ if all the eigenvalues of A are not on the unit circle. In addition, $G \in \mathbb{RH}_\infty^{n_v \times n_w}$ if A is a Schur stable matrix.

From an input-output viewpoint, the model (2.18) defines a system G that maps input w to output v. We denote this input-output relationship as $v = G(w)$. One can easily use induction to show that a state-space model (2.18) with an initial condition ξ^0 is an operator mapping from $\ell_{2e}^{n_v}$ to $\ell_{2e}^{n_w}$. A block diagram can be used to graphically represent the input-output characteristic of a dynamic system. For example, the dynamic system G described by the state-space model (2.18) can be represented by the block-diagram, as shown in Figure 2.3. Although the states of G are not explicitly shown in the block-diagram, the block-diagram captures the relationship $v = G(w)$.

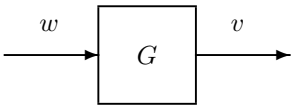

Figure 2.3: The Block-Diagram for a Dynamic System G

A complex system can be modeled by augmenting the state-space models of its subsystems via several basic rules. The augmenting rule for series connection is now

reviewed. The series connection of G_1 and G_2 in Figure 2.4 can be viewed as a composition of G_1 and G_2, and is denoted as $G_2 \circ G_1$. Specifically, we have $u = G_1(w)$ and $v = G_2(u)$. Hence $v = G_2\left(G_1(w)\right) = (G_2 \circ G_1)(w)$. Suppose G_1 has a state-space model (A_1, B_1, C_1, D_1) and G_2 has a state-space model (A_2, B_2, C_2, D_2). Then $G_2 \circ G_1$ also has a state-space model. The input of $G_2 \circ G_1$ is the input to G_1, and the output of $G_2 \circ G_1$ is the output of G_2. The state of $G_2 \circ G_1$ is the augmentation of the state of G_1 and the state of G_2. The state-space model for $G_2 \circ G_1$ is determined by

$$\left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ B_2C_1 & A_2 & B_2D_1 \\ \hline D_2C_1 & C_2 & D_2D_1 \end{array}\right]. \tag{2.20}$$

This is a standard result in the controls literature. More details can be found in [39].

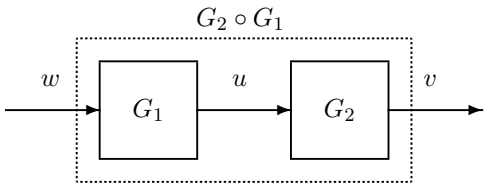

Figure 2.4: The Block-Diagram for Series Connection $G_2 \circ G_1$

Given two state-space models G and Ψ, we define a specific augmentation $\mathcal{G}^{(G,\Psi)}$ as

$$\mathcal{G}^{(G,\Psi)} = \Psi \circ \begin{bmatrix} G \\ I \end{bmatrix}. \tag{2.21}$$

Here the input dimension of Ψ is assumed to be equal to the sum of the input dimension and output dimension of G. The state-space model for $\mathcal{G}^{(G,\Psi)}$ is now presented. Suppose G is described by (2.18), and the output of $\left[\begin{smallmatrix} G \\ I \end{smallmatrix}\right]$ (hence the input for Ψ) is $\left[\begin{smallmatrix} v \\ w \end{smallmatrix}\right]$. Suppose the state-space model for Ψ is described by

$$\begin{aligned} \psi^{k+1} &= A_\psi \psi^k + \begin{bmatrix} B_{\psi 1} & B_{\psi 2} \end{bmatrix} \begin{bmatrix} v^k \\ w^k \end{bmatrix} \\ r^k &= C_\psi \psi^k + \begin{bmatrix} D_{\psi 1} & D_{\psi 2} \end{bmatrix} \begin{bmatrix} v^k \\ w^k \end{bmatrix} \end{aligned} \tag{2.22}$$

Since $v^k = C\xi^k + Dw^k$, we can rewrite (2.22) as

$$\begin{aligned} \psi^{k+1} &= A_\psi \psi^k + \begin{bmatrix} B_{\psi 1}C & B_{\psi 1}D + B_{\psi 2} \end{bmatrix} \begin{bmatrix} \xi^k \\ w^k \end{bmatrix} \\ r^k &= C_\psi \psi^k + \begin{bmatrix} D_{\psi 1}C & D_{\psi 1}D + D_{\psi 2} \end{bmatrix} \begin{bmatrix} \xi^k \\ w^k \end{bmatrix}. \end{aligned} \tag{2.23}$$

Notice the state for $\mathcal{G}^{(G,\Psi)}$ is $\begin{bmatrix} \xi \\ \psi \end{bmatrix}$. We can augment (2.23) with (2.18) to obtain the following state-space model for $\mathcal{G}^{(G,\Psi)}$:

$$\begin{aligned} \begin{bmatrix} \xi^{k+1} \\ \psi^{k+1} \end{bmatrix} &= \begin{bmatrix} A & 0 \\ B_{\psi 1}C & A_\psi \end{bmatrix} \begin{bmatrix} \xi^k \\ \psi^k \end{bmatrix} + \begin{bmatrix} B \\ B_{\psi 1}D + B_{\psi 2} \end{bmatrix} w^k \\ r^k &= \begin{bmatrix} D_{\psi 1}C & C_\psi \end{bmatrix} \begin{bmatrix} \xi^k \\ \psi^k \end{bmatrix} + (D_{\psi 1}D + D_{\psi 2})w^k \end{aligned} \tag{2.24}$$

In other words, the state matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ for the augmented system $\mathcal{G}^{(G,\Psi)}$ are determined by the state matrices of G and Ψ as follows:

$$\mathcal{A} := \begin{bmatrix} A & 0 \\ B_{\psi 1}C & A_\psi \end{bmatrix}, \qquad \mathcal{B} := \begin{bmatrix} B \\ B_{\psi 1}D + B_{\psi 2} \end{bmatrix} \tag{2.25}$$

$$\mathcal{C} := \begin{bmatrix} D_{\psi 1}C & C_\psi \end{bmatrix}, \qquad \mathcal{D} := D_{\psi 1}D + D_{\psi 2} \tag{2.26}$$

An alternative way to derive the above result is first writing out the state-space model for $\begin{bmatrix} G \\ I \end{bmatrix}$ and then augmenting with Ψ using the series connection rule (2.20).

2.6 Feedback Interconnection and Stability Concepts

In this section, we introduce the concept of an feedback interconnection. For a dynamic system G and a causal operator Δ, a feedback interconnection of G and Δ is shown in Figure 2.5 and denoted as $F_u(G, \Delta)$. The feedback connection states that v and w musty satisfy $v = G(w)$ and $w = \Delta(v)$ simultaneously.

To clarify what the feedback interconnection really stands for, consider the following example. Recall Δ_g maps v to w via $w^k = \nabla g(v^k + v^*) - \nabla g(v^*)$, and G is described by

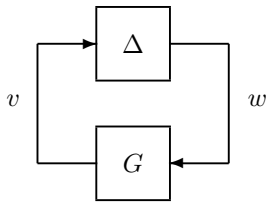

Figure 2.5: The Block-Diagram for Feedback Interconnection $F_u(G, \Delta)$

the state-space model (2.18). The feedback interconnection $F_u(G, \Delta_g)$ represents the following iteration:

$$\begin{aligned} \xi^{k+1} &= A\xi^k + Bw^k \\ v^k &= C\xi^k + Dw^k \\ w^k &= \nabla g(v^k + v^*) - \nabla g(v^*) \end{aligned} \tag{2.27}$$

In the above iteration, the linear state-space model part states $v = G(w)$, and the nonlinear mapping part states $w = \Delta_g(v)$. Clearly there is an issue on whether there exist $(\xi, w, v) \in \ell_{2e}$ satisfying (2.27) such that the feedback interconnection $F_u(G, \Delta_g)$ even makes sense. This is the so-called well-posedness issue in the controls literature.

Definition 1. *The interconnection $F_u(G, \Delta)$ is well-posed if for each $\xi^0 \in \mathbb{R}^{n_\xi}$ there exists a unique solution $\xi \in \ell_{2e}^{n_\xi}$, $v \in \ell_{2e}^{n_v}$ and $w \in \ell_{2e}^{n_w}$ satisfying Equation (2.18) and $w = \Delta(v)$.*

One commonly-used proof technique for well-posedness of discrete-time interconnections is induction. In general, the detailed proof of well-posedness is case-dependent. In this dissertation, the feedback interconnection is mainly used to model optimization methods. As we will later see, we mainly consider feedback interconnection $F_u(G, \Delta)$ with $D = 0$. In this case, the proof of well-posedness becomes straightforward. For example, to prove that $F_u(G, \Delta_g)$ is well-posed, we can rewrite (2.27) as $\xi^{k+1} = A\xi^k + B\nabla g(C\xi^k + v^*) - B\nabla g(v^*)$. Suppose ∇g does not map any finite number to infinity. By induction, there always exists a unique ℓ_{2e} solution (ξ, v, w) satisfying (2.27) for any initial condition $\xi^0 \in \mathbb{R}^{n_\xi}$. Therefore, $F_u(G, \Delta_g)$ is well-posed. Similarly, we can use induction to show that $F_u(G, \Delta)$ is well-posed given $D = 0$ and Δ being other causal operators mapping ℓ_{2e} to ℓ_{2e}.

The feedback interconnection provides a compact expression for systems which are pieced together by linear state-space models and other more troublesome operators. This structure has been extensively used in the controls literature. In this dissertation, we will use it to represent optimization methods. Two notions of stability are studied in this dissertation.

Definition 2. *The interconnection $F_u(G, \Delta)$ is uniformly stable*[2] *if it is well-posed and if $\exists\ c \geq 0$ such that $\|\xi^k\| \leq c\|\xi^0\|$ for all $\xi^0 \in \mathbb{R}^{n_\xi}$ and $k \geq 0$.*

Definition 3. *$F_u(G, \Delta)$ is exponentially stable with rate ρ (≥ 0) if it is well-posed and if $\exists\ c \geq 0$ such that $\|\xi^k\| \leq c\rho^k\|\xi^0\|$ for all $\xi^0 \in \mathbb{R}^{n_\xi}$ and $k \geq 0$.*

In Definition 2, the stability is "uniform" in the sense that the constant c does not depend on ξ^0. In the traditional IQC setup [19], the feedback interconnection involves two exogenous inputs for the purpose of input-output stability analysis. This dissertation focuses on the analysis of optimization methods, and stability regarding internal states is of interest. Hence the exogenous inputs are dropped in our current setup. The stability analysis of $F_u(G, \Delta)$ is typically non-trivial due to the troublesome element Δ. For example, if g is not a quadratic function, then the operator Δ_g becomes nonlinear. As a result, we cannot directly apply the linear system theory in [39] to study the stability of $F_u(G, \Delta_g)$. We will show how to use semidefinite programs to check uniform stability and ρ-exponential stability of $F_u(G, \Delta)$ for various G and Δ.

2.7 Integral Quadratic Constraints

One general framework for analysis of $F_u(G, \Delta)$ is provided by integral quadratic constraints (IQCs) [19, 26]. The key idea is to replace the troublesome element Δ with quadratic constraints on its inputs and outputs. IQCs can be specified either in the frequency or time domain. The definitions of IQCs are given as follows.

[2] *The notion of stability we use here is a special case of the so-called global uniform stability [40, Lemma 4.5] when the required class $\mathcal{K}$ function is a linear function.*

Definition 4. *Let $\Pi = \Pi^{\sim} \in \mathbb{RL}_{\infty}^{(n_v+n_w)\times(n_v+n_w)}$ be given. A bounded, causal operator $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ satisfies the frequency domain IQC defined by the multiplier Π, if the following inequality holds for all $v \in \ell_2^{n_v}$ and $w = \Delta(v)$*

$$\int_0^{2\pi} \begin{bmatrix} V(e^{j\omega}) \\ W(e^{j\omega}) \end{bmatrix}^* \Pi(e^{j\omega}) \begin{bmatrix} V(e^{j\omega}) \\ W(e^{j\omega}) \end{bmatrix} d\omega \geq 0 \tag{2.28}$$

where V and W are discrete-time Fourier transforms of v and w.

Definition 5. *Let Ψ be an $n_r \times (n_v + n_w)$ LTI system governed by the state-space model (2.22), and $M = M^T \in \mathbb{R}^{n_r \times n_r}$.*

(a) Assume A_ψ is a Schur stable matrix. A bounded, causal operator $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ satisfies the time domain soft IQC defined by (Ψ, M) if the following inequality holds for all $v \in \ell_2^{n_v}$ and $w = \Delta(v)$

$$\sum_{k=0}^{\infty} (r^k)^T M r^k \geq 0 \tag{2.29}$$

where r is the output of the state-space model (2.22) with inputs (v, w) and zero initial conditions.

(b) A causal operator $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ satisfies the time domain hard IQC defined by (Ψ, M) if the following inequality holds for all $v \in \ell_{2e}^{n_v}$, $w = \Delta(v)$ and $N \geq 0$

$$\sum_{k=0}^{N} (r^k)^T M r^k \geq 0 \tag{2.30}$$

where r is the output of the state-space model (2.22) with inputs (v, w) and zero initial conditions.

The notation $\Delta \in \text{IQC}(\Pi)$, $\Delta \in \text{SoftIQC}(\Psi, M)$ and $\Delta \in \text{HardIQC}(\Psi, M)$ will be used when Δ satisfies the corresponding frequency domain, time domain soft, or time domain hard IQC, respectively. The definition of time domain hard IQCs does not require Δ to be bounded, while frequency domain IQCs and time domain soft IQCs can only be defined for bounded operators. Time domain IQCs yield a graphical interpretation as shown in Figure 2.6. Let the input and output signals of Δ be filtered

through Ψ with zero initial conditions. If $\Delta \in \text{SoftIQC}(\Psi, M)$, then sequence r must satisfy the infinite time horizon constraint in (2.29) for any $v \in \ell_2^{n_v}$ and $w = \Delta(v)$. A similar interpretation holds for time domain hard IQCs.

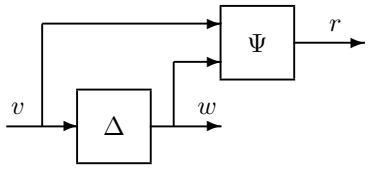

Figure 2.6: Graphical Interpretation for Time Domain IQCs

A library of frequency domain IQCs for different continuous-time bounded operators was summarized in [19]. Additional frequency domain IQCs have been developed for time-varying delays [41–43] and nonlinearities [44]. Many of these continuous-time IQCs have discrete-time counterparts [44–46]. Some IQCs were originally derived in the time domain while many other IQCs were developed in the frequency domain. In Section 2.11, we document existing IQCs for several discrete-time operators which are useful for analysis of first-order optimization methods. The stability analysis in this dissertation requires time domain hard IQCs. It is useful to connect frequency and time domain IQCs so that the full library of known IQCs can be used within the proposed analysis framework. This connection relies on factorizing a frequency domain multiplier as $\Pi = \Psi^{\sim} M \Psi$. Such a factorization is always possible as stated in the next lemma although it is not unique.

Lemma 3. *Suppose* $\Pi = \Pi^{\sim} \in \mathbb{RL}_{\infty}^{(n_v+n_w)\times(n_v+n_w)}$. *Then there exists real matrices* A_ψ, B_ψ, Q, S, *and* R *of compatible dimensions with* A_ψ *being Schur stable,* $Q = Q^T$, *and* $R = R^T$ *such that*

$$\Pi(z) := \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix}^{\sim} \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix} \tag{2.31}$$

Proof. The proof given here is a modification of the continuous-time result presented in [47, Section 7.3]. Separate $\Pi = G_S + G_U$ where G_S and G_U are uniformly bounded outside and inside the closed unit disk, respectively. In addition, without loss of generality, one can choose a specific G_S satisfying $G_S(\infty) = 0$. Let $(A_S, B_S, C_S, 0)$ denote a

realization for this G_S, i.e. $G_S(z) = C_S(zI - A_S)^{-1}B_S$. Here A_S is a Schur stable matrix since G_S is bounded outside the closed unit disk. The assumption $\Pi = \Pi^\sim$ implies that $G_S + G_U = G_S^\sim + G_U^\sim$. This can be rewritten as $G_S - G_U^\sim = G_S^\sim - G_U$ where the left and right sides are analytic outside and inside the (closed) unit disk, respectively. Hence both sides must be analytic in the entire complex plane. By Liouville's theorem, one can conclude that $G_S^\sim - G_U$ is a constant, i.e. there exists matrix R such that $G_U(z) = G_S^\sim(z) + R$ for all $z \in \mathbb{C}$. This implies $\Pi = G_S + G_S^\sim + R$, and $R = R^T$ follows immediately from $\Pi = \Pi^\sim$. Thus Π can be written as in Equation (2.31) with $A_\psi = A_S$, $B_\psi = B_S$, $Q = 0$, $S = C_S^T$ and the constant matrix R. □

Existing numerical algorithms can be used to construct the factorization presented in Lemma 3. If Π is proper then the Matlab function `stabsep` can be used to separate $\Pi = G_S + G_U$ where G_S is stable and causal. However, $\Pi \in \mathbb{RL}_\infty^{m\times m}$ may be a non-proper (polynomial) function of z, e.g. the multipliers used in [12, 48]. A descriptor system representation of Π is required in such cases. If Π is a non-proper (descriptor) system then the algorithm in [49] can be used to separate out the stable part. This algorithm is easily implemented in Matlab and is based on LAPACK linear algebra routines described in [50, 51]. The stable part G_S in this construction is strictly proper and hence it has a standard state-space description. Finally, the matrix R can be explicitly computed by evaluating $R = \Pi(z_0) - G_S(z_0) - G_S(z_0^{-1})^T$ for some $z_0 \in \mathbb{C}$. For example, evaluating at $z_0 = 1$ is useful as both Π and G_S are bounded on the unit circle.

As mentioned in Section 2.5, LTI systems can be represented by rational functions. Then frequency and time domain IQCs can be connected by the (non-unique) factorizations $\Pi = \Psi^\sim M\Psi$. This is formalized in the next lemma.

Lemma 4. *Let $\Pi = \Psi^\sim M\Psi$ with $\Psi \in \mathbb{RH}_\infty^{n_r\times(n_v+n_w)}$ and $M = M^T \in \mathbb{R}^{n_r\times n_r}$. Let $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a bounded, causal operator. Then*

1. *$\Delta \in IQC(\Pi)$ if and only if $\Delta \in SoftIQC(\Psi, M)$.*
2. *$\Delta \in IQC(\Pi)$ if $\Delta \in HardIQC(\Psi, M)$.*

Proof. To prove Statement 1, first assume $\Delta \in \mathrm{IQC}(\Pi)$. For any $v \in \ell_2^{n_v}$ and $w = \Delta(v)$,

the frequency domain IQC inequality (2.28) can be written as:

$$\int_0^{2\pi} R(e^{j\omega})^* M R(e^{j\omega}) d\omega \geq 0 \tag{2.32}$$

where $R(e^{j\omega}) = \Psi(e^{j\omega}) \begin{bmatrix} V(e^{j\omega}) \\ W(e^{j\omega}) \end{bmatrix}$ is the Fourier transform of an ℓ_2 signal since Ψ is stable.[3] By Parseval's theorem [21], Δ satisfies the time domain soft IQC defined by (Ψ, M). The reverse implication of Statement 1 also follows via Parseval's theorem.

To prove Statement 2, assume $\Delta \in \text{HardIQC}(\Psi, M)$. Let r be the output of Ψ driven by inputs $v \in \ell_2^{n_v}$ and $w \in \ell_2^{n_w}$ with zero initial conditions. Stability of Ψ implies $r \in \ell_2^{n_r}$. Hence the hard IQC inequality (2.30) holds as $N \to \infty$ and $\Delta \in \text{SoftIQC}(\Psi, M)$. This implies $\Delta \in \text{IQC}(\Pi)$ by Statement 1. □

Statement 2 of Lemma 4 states that a time domain hard IQC with a stable filter Ψ always leads to a frequency domain IQC. The reverse implication does not hold in general. It is important to emphasize that factorizations of Π are not unique. Some factorizations of Π may yield time domain hard IQCs while others do not. Thus the hard/soft property is not inherent to the multiplier Π but depends on the factorization (Ψ, M). The factorization introduced by Lemma 3 does not, in general, yield a valid time domain hard IQC. Later in Section 3.1, we will develop a J-spectral factorization approach, which can be used to systematically factorize frequency domain IQCs into time domain hard IQCs under mild technical conditions.

2.8 Uniform Stability Analysis Using IQCs

The original work in [19] addresses input-output stability using IQCs. For the purpose of this dissertation, uniform stability and ρ-exponential stability of $F_u(G, \Delta)$ are considered. This section presents a standard dissipation inequality approach for uncertainty analysis [27, 28, 40, 52]. Linear matrix inequality (LMI) conditions for uniform stability of $F_u(G, \Delta)$ are formulated using time domain hard IQCs.

First, the core idea of the IQC analysis is briefly explained. To better explain the

[3] The transform $R(e^{j\omega})$ is unrelated to the constant matrix R appearing in the basic IQC factorization.

IQC analysis, the notation $(v, w) \in \text{HardIQC}(\Psi, M)$ is used when the sequence r generated by (2.22) with input pair (v, w) and zero initial conditions always satisfies the constraint (2.30). Therefore, $\Delta \in \text{HardIQC}(\Psi, M)$ if and only if $(v, w) \in \text{HardIQC}(\Psi, M)$ for all $v \in \ell_{2e}^{n_v}$ and $w = \Delta(v)$. When analyzing $F_u(G, \Delta)$, one aims to draw conclusions on the pair (v, w) in the set $\{(v, w) \in \ell_{2e} : v = G(w),\ w = \Delta(v)\}$. Suppose $\Delta \in \text{HardIQC}(\Psi, M)$, then the set $\{(v, w) \in \ell_{2e} : v = G(w),\ w = \Delta(v)\}$ is a subset of $\{(v, w) \in \ell_{2e} : v = G(w),\ (v, w) \in \text{HardIQC}(\Psi, M)\}$. If one can prove that the state of G is uniformly bounded for any pair (v, w) in the set $\{(v, w) \in \ell_{2e} : v = G(w),\ (v, w) \in \text{HardIQC}(\Psi, M)\}$, then it is guaranteed that the state of G is uniformly bounded for any pair (v, w) satisfying $v = G(w)$ and $w = \Delta(v)$ simultaneously. Equivalently, the uniform stability of $F_u(G, \Delta)$ is guaranteed. Hence one can completely remove Δ from the analysis by enforcing the condition $(v, w) \in \text{HardIQC}(\Psi, M)$. A graphical interpretation is shown in Figure 2.7. After replacing Δ with the IQC condition, the pair (v, w) still satisfies $v = G(w)$. In addition, let $r = \Psi(v, w) = \Psi(G(w), w)$. Then r must satisfy the constraint (2.30). From (2.21), we have $r = \mathcal{G}^{(G,\Psi)}(w)$. Eventually we only need to analyze $\mathcal{G}^{(G,\Psi)}$ with input $w \in \ell_{2e}^{n_w}$ and the output r. By induction, w has to be in $\ell_{2e}^{n_w}$, and r always satisfies the constraint (2.30) given the condition $\Delta \in \text{HardIQC}(\Psi, M)$.

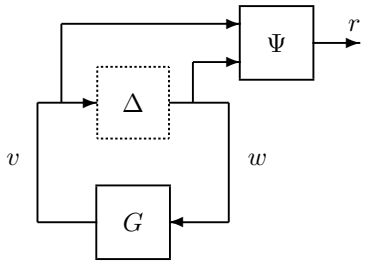

Figure 2.7: Uncertain LTI System Extended to Include Filter Ψ

The detailed analysis is now presented. Notice conservatism may be introduced in the IQC analysis since the full information of Δ is completely replaced by the IQC condition. Multiple IQCs are typically used to reduce the potential conservatism. Now the operator Δ is assumed to satisfy multiple time domain hard IQCs defined by $\{(\Psi_j, M_j)\}_{j=1}^{N_J}$. All $\{\Psi_j\}_{j=1}^{N_J}$ are first aggregated into a single filter denoted Ψ with

the following state-space realization:

$$\begin{bmatrix} \psi^{k+1} \\ r^k \end{bmatrix} = \begin{bmatrix} A_\psi & B_{\psi 1} & B_{\psi 2} \\ C_\psi & D_{\psi 1} & D_{\psi 2} \end{bmatrix} \begin{bmatrix} \psi^k \\ v^k \\ w^k \end{bmatrix} \tag{2.33}$$

where $r := [r_1^T \dots r_{N_J}^T]^T$ and r_j is the output of the filter Ψ_j. As mentioned previously, the uniform stability analysis of $F_u(G, \Delta)$ is based on the extended system shown in Figure 2.7. Consider the extended system $\mathcal{G}^{(G,\Psi)}$ defined in Equation (2.24). Notice $\mathcal{G}^{(G,\Psi)}$ has the following state-space model:

$$\begin{bmatrix} \eta^{k+1} \\ r^k \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} \eta^k \\ w^k \end{bmatrix} \tag{2.34}$$

The extended state vector is $\eta^k := \begin{bmatrix} \xi^k \\ \psi^k \end{bmatrix} \in \mathbb{R}^{n_\xi + n_\psi}$. Based on (2.25) and (2.26), the state matrices for the extended system $\mathcal{G}^{(G,\Psi)}$ can be computed from the state matrices of G and Ψ.

Now define $M_\lambda := diag(\lambda_1 M_1, \dots, \lambda_{N_J} M_{N_J})$, where the "*diag*" notation means block diagonal concatenation. The next theorem presents LMI conditions for uniform stability of $F_u(G, \Delta)$ using time domain hard IQCs and a standard dissipation inequality. This theorem uses an LMI defined by G and $\{(\Psi_j, M_j)\}_{j=1}^{N_J}$:

$$LMI_{(G,\Psi)}(P, M_\lambda) := \begin{bmatrix} \mathcal{A}^T P \mathcal{A} - P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \tag{2.35}$$

Theorem 1. *Let G be an LTI system defined by (2.18) and $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a causal operator. Assume $F_u(G, \Delta)$ is well-posed and $\Delta \in HardIQC(\Psi_j, M_j)$ for $j = 1, \dots, N_J$. If one of the following conditions holds*

(a) $\exists$ a matrix $P = P^T > 0$ and scalars $\lambda_j \geq 0$ such that $LMI_{(G,\Psi)}(P, M_\lambda) \leq 0$.

(b) $\exists$ a matrix $P = P^T \geq 0$ and scalars $\lambda_j \geq 0$ such that $LMI_{(G,\Psi)}(P, M_\lambda) < 0$.

Then $F_u(G, \Delta)$ is uniformly stable.

Proof. Given any fixed ξ^0, let (ξ, v, w) be the unique ℓ_{2e} solution satisfying the feedback iteration $F_u(G, \Delta)$ with this initial condition ξ^0, and let (ψ, r) be generated by Ψ with inputs (v, w) and zero initial conditions. Since $\eta^k = \begin{bmatrix} \xi^k \\ \psi^k \end{bmatrix}$, we can verify that (2.34) holds with the current choice of (η, w, r) and the initial condition $\eta^0 = \begin{bmatrix} \xi^0 \\ \psi^0 \end{bmatrix}$. Here ψ^0 is the zero vector, since Ψ has zero initial conditions. Define a storage function by $V(\eta^k) = (\eta^k)^T P \eta^k$. Since $\eta^{k+1} = \mathcal{A}\eta^k + \mathcal{B}w^k$, one can show the following holds

$$\begin{aligned} V(\eta^{k+1}) - V(\eta^k) &= (\mathcal{A}\eta^k + \mathcal{B}w^k)^T P(\mathcal{A}\eta^k + \mathcal{B}w^k) - (\eta^k)^T P \eta^k \\ &= \begin{bmatrix} \eta^k \\ w^k \end{bmatrix}^T \begin{bmatrix} \mathcal{A}^T P \mathcal{A} - P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} \begin{bmatrix} \eta^k \\ w^k \end{bmatrix}. \end{aligned} \tag{2.36}$$

Similarly, since $r^k = \mathcal{C}\eta^k + \mathcal{D}w^k$, one can show

$$\sum_{j=1}^{N_J} \lambda_j (r_j^k)^T M_j r_j^k = (r^k)^T M_\lambda r^k = \begin{bmatrix} (\eta^k)^T & (w^k)^T \end{bmatrix} \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} \eta^k \\ w^k \end{bmatrix}. \tag{2.37}$$

Now assume Condition (a) holds. Left and right multiply $LMI_{(G,\Psi)}(P, M_\lambda) \leq 0$ by $[\eta^T, w^T]$ and $[\eta^T, w^T]^T$ to show that V satisfies:

$$V\left(\eta^{k+1}\right) - V\left(\eta^k\right) + \sum_{j=1}^{N_J} \lambda_j (r_j^k)^T M_j r_j^k \leq 0 \tag{2.38}$$

The above inequality can be summed from $k = 0$ to $k = N$ to yield:

$$V\left(\eta^{N+1}\right) - V\left(\eta^0\right) + \sum_{j=1}^{N_J} \lambda_j \left(\sum_{k=0}^{N} (r_j^k)^T M_j r_j^k \right) \leq 0 \tag{2.39}$$

Applying the time domain hard IQC conditions with the fact $\lambda_j \geq 0$, we directly get $V\left(\eta^{N+1}\right) \leq V(\eta^0)$, which is equivalent to $V(\eta^k) \leq V(\eta^0)$. The zero initial condition for Ψ implies $\|\xi^k\|^2 \leq \|\eta^k\|^2 \leq \text{cond}(P)\|\eta^0\|^2 = \text{cond}(P)\|\xi^0\|^2$. Thus $\|\xi^k\| \leq \sqrt{\text{cond}(P)}\|\xi^0\|$, and $F_u(G, \Delta)$ is uniformly stable.

Now assume Condition (b) holds. Since $LMI_{(G,\Psi)}(P, M_\lambda) < 0$, $\exists\ \epsilon > 0$ such that $LMI_{(G,\Psi)}(P + \epsilon I, M_\lambda) \leq 0$. Uniform stability follows from Condition (a) due to the fact $P + \epsilon I > 0$. □

Given $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ and M_j, the matrix $LMI_{(G,\Psi)}(P, M_\lambda)$ is linear in P and λ_j. Therefore, the stability tests in the above theorem are presented as LMI conditions,

whose feasible sets are convex and can be effectively searched using the state-of-the-art convex optimization techniques, e.g. interior point method. Many optimization solvers are available such that coding the LMI conditions in Theorem 1 is a straightforward task. In this dissertation, the numerical calculations of LMI conditions were performed using CVX [53, 54] with the solver SDPT3 [55, 56].

The dissipation inequality approach in Theorem 1 relies on the fact that the constraint (2.30) holds for any finite-horizon $N \geq 0$. It does not require either G or Ψ to be stable. It only requires that the states of $\mathcal{G}^{(G,\Psi)}$ have no finite escape time. Hence, the definition of time domain hard IQCs does not enforce the stability of Ψ. In principle, one can use time domain hard IQCs with unstable Ψ, although the J-spectral factorization of any frequency domain IQC always leads to stable Ψ.

In Theorem 1, Statement (b) is stronger than Statement (a) in the sense that Statement (b) is a sufficient condition for Statement (a). However, Statement (b) can be generalized to formulate other uniform stability theorems under extra assumptions. In Section 3.2, we will present one such generalization using a modified dissipation inequality. More specifically, when $G \in \mathbb{RH}_\infty^{n_v \times n_w}$, a modified dissipation inequality can be used to drop the constraint $P \geq 0$ in Condition (b) of the above theorem. In addition, the conic combination of time domain hard IQCs can be extended to more general IQC parameterizations where soft IQCs are involved and M_λ is an affine function of λ [57]. This leads to potentially more flexible formulations of uniform stability tests.

It is possible to perform uniform stability analysis of $F_u(G, \Delta)$ using some alternative procedures (ν-gap metric theory in [58], dissipation inequality in [25], etc). These procedures were originally developed for input-output stability analysis. A detailed discussion on the adaption of these procedures for uniform stability analysis is beyond the scope of this dissertation.

Remark 1. *It is worth mentioning that the composition of two operators can also be handled by the IQC framework. Specifically, the analysis of $F_u(G, \Delta_1 \circ \Delta_2)$ can be performed when IQCs are posed on Δ_1 and Δ_2 separately. In the controls literature, the interconnection $F_u(G, \Delta_1 \circ \Delta_2)$ is typically transformed into a new interconnection with block diagonal perturbation $diag(\Delta_1, \Delta_2)$. Then IQCs on $diag(\Delta_1, \Delta_2)$ can be directly constructed from the IQCs on Δ_1 and Δ_2, and used to formulate stability theorems. An alternative approach was presented in [12, Section 5.2]. The idea is as follows. One can*

define an augmented operator $\Delta = \begin{bmatrix} \Delta_2 \\ \Delta_1 \circ \Delta_2 \end{bmatrix}$. *Notice* $\Delta_1 \circ \Delta_2 = [\, 0 \;\; I \,] \circ \Delta$. *One can show that* $F_u(G, \Delta_1 \circ \Delta_2)$ *is uniformly stable if and only if* $F_u(G \circ [\, 0 \;\; I \,], \Delta)$ *is uniformly stable. Notice* $G \circ [\, 0 \;\; I \,]$ *is still governed by a linear state-space model. In addition, the IQCs on* Δ *can be efficiently constructed from the IQCs on* Δ_1 *and* Δ_2. *Hence the uniform stability analysis of* $F_u(G \circ [\, 0 \;\; I \,], \Delta)$ *can be directly handled by Theorem 1.*

2.9 ρ-Hard IQCs and ρ-Exponential Stability Analysis

Recently, it is recognized that many deterministic first-order optimization methods can be cast as $F_u(G, \Delta)$, and the concept of ρ-hard IQCs is further developed to analyze the convergence rates of such interconnections [12]. Later we will use a state shifting argument to show that the linear convergence rate analysis of an optimization method is equivalent to the ρ-exponential stability analysis of a related interconnection. This section reviews the ρ-hard IQC approach, which is inspired by early results on sector-bounded nonlinearities [30] and formalized in [12]. The concept of time domain ρ-hard IQCs is introduced, and a related dissipation inequality for ρ-exponential stability of $F_u(G, \Delta)$ is briefly reviewed.

Definition 6. *Let* Ψ *be an* $n_r \times (n_v + n_w)$ *LTI system governed by the state-space model* (2.22), *and* $M = M^T \in \mathbb{R}^{n_r \times n_r}$. *Suppose* $0 < \rho \leq 1$. *A causal operator* $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ *satisfies the time domain ρ-hard IQC defined by* (Ψ, M, ρ), *if the following inequality holds for all* $v \in \ell_{2e}^{n_v}$, $w = \Delta(v)$ *and* $N \geq 0$

$$\sum_{k=0}^{N} \rho^{-2k} (r^k)^T M r^k \geq 0 \tag{2.40}$$

where r *is the output of* Ψ *driven by inputs* (v, w) *with zero initial conditions.*

The notation $\Delta \in \rho\text{-HardIQC}(\Psi, M, \rho)$ will be used when Δ satisfies the corresponding time domain ρ-hard IQC. Suppose $\Delta \in \rho\text{-HardIQC}(\Psi_j, M_j, \rho)$ for $j = 1, \ldots, N_J$. All $\{\Psi_j\}_{j=1}^{N_J}$ are aggregated into a filter Ψ governed by Equation (2.33). Let $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ denote the state space realization of $\mathcal{G}^{(G,\Psi)}$. Theorem 4 in [12] essentially states the following result:

Theorem 2. *Let G be an LTI system defined by (2.18) and $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a causal operator. Assume $F_u(G,\Delta)$ is well-posed, and $\Delta \in \rho$-HardIQC(Ψ_j, M_j, ρ) for $j = 1, \ldots, N_J$. If one of the following conditions holds*

(a) $\exists$ a matrix $P = P^T > 0$ and scalars $\lambda_j \geq 0$ such that

$$\begin{bmatrix} \mathcal{A}^T P \mathcal{A} - \rho^2 P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \leq 0 \tag{2.41}$$

(b) $\exists$ a matrix $P = P^T \geq 0$ and scalars $\lambda_j \geq 0$ such that the left side of (2.41) is negative definite.

Then $F_u(G,\Delta)$ is exponentially stable with rate ρ.

Proof. Statement (a) has been proved in [12]. The proof is similar to the proof of Theorem 1. The basic idea is sketched as follows. Set a storage function as $V(\eta^k) = (\eta^k)^T P \eta^k$. Left and right multiply (2.41) by $[\eta^T, w^T]$ and $[\eta^T, w^T]^T$ to show that V satisfies: $\rho^{-2k} V(\eta^{k+1}) - \rho^{2-2k} V(\eta^k) + \sum_{j=1}^{N_J} \lambda_j \rho^{-2k} (r_j^k)^T M_j r_j^k \leq 0$. Summing this inequality from $k = 0$ to $k = N$ with initial condition $\eta^0 = \begin{bmatrix} \xi^0 \\ 0 \end{bmatrix}$ and applying the time domain ρ-hard IQC conditions yields $\rho^{-2N} V\left(\eta^{N+1}\right) \leq \rho^2 V\left(\eta^0\right)$, which is equivalent to $\rho^{2-2k} V\left(\eta^k\right) \leq \rho^2 V\left(\eta^0\right)$. Therefore, $\|\xi^k\| \leq \|\eta^k\| \leq \sqrt{\text{cond}(P)} \|\eta^0\| \rho^k = \sqrt{\text{cond}(P)} \|\xi^0\| \rho^k$, and $F_u(G,\Delta)$ is exponentially stable with rate ρ. When Condition (b) holds, the perturbation argument in the proof of Theorem 1 can be used again to conclude the desired conclusion. □

When formulating (2.41), M_j and the state matrices of G do not depend on ρ. However, the state-space realization of Ψ may depend on ρ, e.g. see [12, Lemma 10]. Hence, $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ may depend on ρ. In addition, the term $\rho^2 P$ is bilinear in ρ^2 and P. Hence, one cannot treat (2.41) as a single LMI when trying to find the smallest ρ such that (2.41) is feasible. When ρ is fixed, (2.41) becomes an LMI with decision variables P and λ_j. Hence Theorem 2 can be used to check whether an interconnection $F_u(G,\Delta)$ is exponentially stable with a given rate ρ. A bisection on ρ^2 can then be used to find the best (i.e. smallest) exponential rate bound for $F_u(G,\Delta)$. If the state-space realization of Ψ does not depend on ρ, the only bilinear term in (2.41) is $\rho^2 P$. One can treat ρ^2 as one variable. Solving smallest ρ with (2.41) becomes a generalized eigenvalue

problem (GEVP) [33]. In this case, a bisection search on ρ is no longer required, and more efficient algorithms are available [59].

The size of (2.41) depends on the state dimensions of G and Ψ. When applying Theorem 2 to analyze deterministic first-order optimization methods, a dimension reduction step is involved and the resultant semidefinite programs typically have small sizes. The dimension reduction relies on the following lemma.

Lemma 5. *Suppose the matrices (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ satisfy*

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{c|c} \tilde{A} \otimes I_p & \tilde{B} \otimes I_p \\ \hline \tilde{C} \otimes I_p & \tilde{D} \otimes I_p \end{array}\right]. \tag{2.42}$$

Similarly, suppose $A_\psi = \tilde{A}_\psi \otimes I_p$, $B_{\psi 1} = \tilde{B}_{\psi 1} \otimes I_p$, $B_{\psi 2} = \tilde{B}_{\psi 2} \otimes I_p$, $C_\psi = \tilde{C}_\psi \otimes I_p$, $D_{\psi 1} = \tilde{D}_{\psi 1} \otimes I_p$, and $D_{\psi 2} = \tilde{D}_{\psi 2} \otimes I_p$. Let $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be determined by (2.25) and (2.26), and let $(\tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}, \tilde{\mathcal{D}})$ be calculated from

$$\tilde{\mathcal{A}} := \begin{bmatrix} \tilde{A} & 0 \\ \tilde{B}_{\psi 1}\tilde{C} & \tilde{A}_\psi \end{bmatrix}, \qquad \tilde{\mathcal{B}} := \begin{bmatrix} \tilde{B} \\ \tilde{B}_{\psi 1}\tilde{D} + \tilde{B}_{\psi 2} \end{bmatrix} \tag{2.43}$$

$$\tilde{\mathcal{C}} := \begin{bmatrix} \tilde{D}_{\psi 1}\tilde{C} & \tilde{C}_\psi \end{bmatrix}, \qquad \tilde{\mathcal{D}} := \tilde{D}_{\psi 1}\tilde{D} + \tilde{D}_{\psi 2} \tag{2.44}$$

Suppose $P = \tilde{P} \otimes I_p$ and $M_\lambda = \tilde{M}_\lambda \otimes I_p$. Then

$$\begin{aligned} &\begin{bmatrix} \mathcal{A}^T P \mathcal{A} - \rho^2 P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} = \\ &\qquad \left(\begin{bmatrix} \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{A}} - \rho^2 \tilde{P} & \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{B}} \\ \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{A}} & \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{B}} \end{bmatrix} + \begin{bmatrix} \tilde{\mathcal{C}}^T \\ \tilde{\mathcal{D}}^T \end{bmatrix} \tilde{M}_\lambda \begin{bmatrix} \tilde{\mathcal{C}} & \tilde{\mathcal{D}} \end{bmatrix} \right) \otimes I_p \end{aligned} \tag{2.45}$$

where the matrix dimensions are assumed to be compatible.

Proof. Based on the basic property of the Kronecker product, one can first verify

$$\left[\begin{array}{c|c} \mathcal{A} & \mathcal{B} \\ \hline \mathcal{C} & \mathcal{D} \end{array}\right] = \left[\begin{array}{c|c} \tilde{\mathcal{A}} \otimes I_p & \tilde{\mathcal{B}} \otimes I_p \\ \hline \tilde{\mathcal{C}} \otimes I_p & \tilde{\mathcal{D}} \otimes I_p \end{array}\right]. \tag{2.46}$$

Then one can combine the above relation with the facts $P = \tilde{P} \otimes I_p$ and $M_\lambda = \tilde{M}_\lambda \otimes I_p$. The rest of the proof follows directly from the basic property of the Kronecker product. □

Based on the above lemma, the dimension of the LMI condition (2.41) can be significantly reduced in many optimization analysis problems. See Remark 2 in Section 2.11.3 and [12, Section 4.2] for further explanations.

2.10 Linear Rate Analysis of Optimization Methods

As presented in [12, Section 2], we can represent various first-order optimization methods as feedback interconnections. For example, to rewrite the standard gradient descent method $x^{k+1} = x^k - \alpha \nabla g(x^k)$, we can set $\bar{\xi}^k = x^k$ and use the following feedback interconnection:

$$\begin{aligned} \bar{\xi}^{k+1} &= \bar{\xi}^k - \alpha \bar{w}^k \\ \bar{v}^k &= \bar{\xi}^k \\ \bar{w}^k &= \nabla g(\bar{v}^k) \end{aligned} \tag{2.47}$$

which is equivalent to $F_u(\bar{G}, \bar{\Delta}_g)$ with

$$\bar{G} = \left[\begin{array}{c|c} I_p & -\alpha I_p \\ \hline I_p & 0_p \end{array}\right]. \tag{2.48}$$

Recall that the operator $\bar{\Delta}_g$ is defined by (2.16).

Similarly, to represent Nesterov's accelerated method (2.11), we can use $F_u(\bar{G}, \bar{\Delta}_g)$ with

$$\bar{G} = \left[\begin{array}{cc|c} (1+\beta) I_p & -\beta I_p & -\alpha I_p \\ I_p & 0_p & 0_p \\ \hline (1+\beta) I_p & -\beta I_p & 0_p \end{array}\right]. \tag{2.49}$$

To represent the Heavy-ball method (2.12), we can use $F_u(\bar{G}, \bar{\Delta}_g)$ with

$$\bar{G} = \left[\begin{array}{cc|c} (1+\beta) I_p & -\beta I_p & -\alpha I_p \\ I_p & 0_p & 0_p \\ \hline I_p & 0_p & 0_p \end{array}\right]. \tag{2.50}$$

The optimization method $F_u(\bar{G}, \bar{\Delta}_g)$ is used to find the point x^* satisfying $\nabla g(x^*) = 0$. Hence ideally the gradient descent method should converge to $\xi^* = x^*$. Nesterov's accelerated method and the Heavy-ball method should converge to $\xi^* = \left[\begin{smallmatrix} x^* \\ x^* \end{smallmatrix}\right]$. Now the linear convergence of an optimization method $F_u(\bar{G}, \bar{\Delta}_g)$ is defined as follows.

Definition 7. *The optimization method $F_u(\bar{G}, \bar{\Delta})$ converges linearly at rate ρ if it is well-posed and if $\exists\ c \geq 0$ such that $\|\bar{\xi}^k - \xi^*\| \leq c\rho^k \|\bar{\xi}^0 - \xi^*\|$ for all $\bar{\xi}^0$ and $k \geq 0$.*

In the above definition, we require the constant c to be independent of $\bar{\xi}^0$. This dissertation focuses on the this strong notion of linear convergence. A weaker version of linear convergence can be defined by allowing c to depend on $\bar{\xi}^0$. We do not specifically study this weaker notion of linear convergence in this dissertation.

Now we state the equivalence between linear rate analysis of an optimization method and ρ-exponential stability analysis of a related interconnection. The optimization method $F_u(\bar{G}, \bar{\Delta}_g)$ can be written as

$$\begin{aligned} \bar{\xi}^{k+1} &= A\bar{\xi}^k + B\bar{w}^k \\ \bar{v}^k &= C\bar{\xi}^k + D\bar{w}^k \\ \bar{w}^k &= \nabla g(\bar{v}^k) \end{aligned} \tag{2.51}$$

where A and C are assumed to satisfy $A\xi^* = \xi^*$ and $C\xi^* = x^*$, respectively. One can easily verify the assumptions on A and C for various optimization methods (the gradient descent method, Nesterov's accelerated method, etc). Now set $\xi^k = \bar{\xi}^k - \xi^*$, $w^k = \bar{w}^k$, and $v^k = \bar{v}^k - x^*$. Since $\nabla g(x^*) = 0$, $A\xi^* = \xi^*$, and $C\xi^* = x^*$, it is straightforward to rewrite (2.51) as

$$\begin{aligned} \xi^{k+1} &= A\xi^k + Bw^k \\ v^k &= C\xi^k + Dw^k \\ w^k &= \nabla g(v^k + x^*) - \nabla g(x^*) \end{aligned} \tag{2.52}$$

which is equivalent to $F_u(G, \Delta_g)$. Here G and $\bar{G}$ have the same state matrices. The operator Δ_g has been defined in (2.17). The state of G and the state of $\bar{G}$ satisfy $\xi^k = \bar{\xi}^k - \xi^*$. Therefore, the interconnection $F_u(G, \Delta_g)$ is ρ-exponentially stable if and only if the optimization method $F_u(\bar{G}, \bar{\Delta})$ converges to ξ^* at a linear rate ρ. In [12], time domain ρ-hard IQCs on Δ_g were derived and applied to study the linear convergence rates of several optimization methods. Notice the derivations in [12] were slightly different from our arguments here. In the original work of [12], the concept of time domain ρ-hard IQCs was modified such that the IQC analysis was directly applied to $F_u(\bar{G}, \bar{\Delta}_g)$. However, the modification in [12] is essentially equivalent to the state shifting argument in this section.

2.11 Related Operators and Existing IQCs

Now we present several operators which are useful for analysis of optimization methods. In addition, we also briefly review some known IQCs on these operators.

2.11.1 Memoryless Nonlinearity in a Sector

Suppose the operator $\Delta : \ell_{2e}^p \to \ell_{2e}^p$ maps v to $w = \Delta(v)$ as $w^k = \phi(v^k, k)$, where $\phi : \mathbb{R}^p \times \mathbb{Z}^+ \to \mathbb{R}^p$ is in a sector $[m, L]$, i.e. the following inequality holds for all k:

$$\Big(\phi(v^k, k) - Lv^k\Big)^T \Big(\phi(v^k, k) - mv^k\Big) \leq 0 \tag{2.53}$$

Then this sector condition directly gives us a quadratic constraint:

$$\begin{bmatrix} v^k \\ w^k \end{bmatrix}^T \begin{bmatrix} -2mLI_p & (L+m)I_p \\ (L+m)I_p & -2I_p \end{bmatrix} \begin{bmatrix} v^k \\ w^k \end{bmatrix} \geq 0 \tag{2.54}$$

Hence Δ satisfies the sector IQC: $\Delta \in \text{HardIQC}(\Psi, M)$ and $\Delta \in \rho\text{-HardIQC}(\Psi, M, \rho)$ where $0 < \rho \leq 1$ and the pair (Ψ, M) is given as

$$\Psi = \begin{bmatrix} LI_p & -I_p \\ -mI_p & I_p \end{bmatrix} \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{2.55}$$

Notice (2.54) holds for any k. Hence the sector IQC is a time domain hard IQC as well as a time domain ρ-hard IQC. This is the most commonly-used IQC in the analysis of a feedback interconnection of a linear system and a nonlinearity.

When $L > 0$, (2.54) can be rewritten as

$$\begin{bmatrix} v^k \\ w^k \end{bmatrix}^T \begin{bmatrix} -2mI_p & (1+\frac{m}{L})I_p \\ (1+\frac{m}{L})I_p & -\frac{2}{L}I_p \end{bmatrix} \begin{bmatrix} v^k \\ w^k \end{bmatrix} \geq 0.$$

In the limiting case $L \to \infty$, the above inequality becomes

$$\begin{bmatrix} v^k \\ w^k \end{bmatrix}^T \begin{bmatrix} -2mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} v^k \\ w^k \end{bmatrix} \geq 0. \tag{2.56}$$

Consequently, Ψ and M should also be modified accordingly.

2.11.2 Static Nonlinearity

Suppose $\Delta : \ell_{2e}^p \to \ell_{2e}^p$ maps v to $w = \Delta(v)$ as $w^k = \phi(v^k)$, where $\phi : \mathbb{R}^p \to \mathbb{R}^p$ is a continuous function. This nonlinearity is static in the sense that the function ϕ does not change with k. When ϕ lies within a sector $[m, L]$ for finite m and L, the IQC in Section 2.11.1 can be applied to Δ. Under certain circumstances, the Zames-Falb IQCs [23,44] can be constructed for the static nonlinearity Δ. The nonlinear function ϕ is bounded and monotone nondecreasing if $\phi(0) = 0$, $[\phi(y_1) - \phi(y_2)]^T(y_1 - y_2) \geq 0$, and $\|\phi(y_1)\| \leq c\|y_1\|$ for some c and all y_1, y_2. When ϕ is bounded, the operator Δ is also bounded. The Zames-Falb IQCs can be specified in either frequency or time domain.

Lemma 6. *Let $\phi : \mathbb{R}^p \to \mathbb{R}^p$ be bounded and monotone nondecreasing. Suppose ϕ is a gradient of some potential function which maps from $\mathbb{R}^p$ to $\mathbb{R}$. Then*

1. *(Off-by-τ Hard IQC): For any $v = \{v^0, v^1, \ldots\} \in \ell_{2e}^p$, and $\tau \geq 0$, we set $w^k = \phi(v^k)$ for $k \geq 0$, and define $v^k = 0$ for $-\tau \leq k < 0$. Then for any $N \geq 0$, the following inequality always holds*

$$\sum_{k=0}^{N} (w^k)^T \left(v^k - v^{k-\tau} \right) \geq 0. \tag{2.57}$$

Hence $\Delta \in HardIQC(\Psi, M)$ with

$$\Psi = \left[\begin{array}{cccc|cc} 0_p & 0_p & \ldots & 0_p & -I_p & 0_p \\ I_p & 0_p & \ldots & 0_p & 0_p & 0_p \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0_p & \ldots & I_p & 0_p & 0_p & 0_p \\ \hline 0_p & 0_p & \ldots & I_p & I_p & 0_p \\ 0_p & 0_p & \ldots & 0_p & 0_p & I_p \end{array}\right], \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{2.58}$$

Here the state dimension of Ψ is $p\tau \times 1$.

2. *(Frequency Domain Zames-Falb IQC): Let $h \in \ell_{2e}$ satisfy $\sum_{k=0}^{\infty} h^k \leq 1$ and $h^k \geq 0$ for all k. Then $\Delta \in IQC(\Pi)$ with $\Pi = \left[\begin{smallmatrix} 0 & 1-H^* \\ 1-H & 0 \end{smallmatrix}\right] \otimes I_p$ where H denotes the Laplace transform of h.*

3. *(Off-by-τ ρ-hard IQC): For any $v = \{v^0, v^1, \ldots\} \in \ell_{2e}^p$, and $\tau \geq 0$, we set $w^k = \phi(v^k)$ for $k \geq 0$, and define $v^k = 0$ for $-\tau \leq k < 0$. Then for any $N \geq 0$, the following inequality always holds*

$$\sum_{k=0}^{N} \rho^{-2k} (w^k)^T \left(v^k - \rho^{2\tau} v^{k-\tau} \right) \geq 0. \tag{2.59}$$

Hence $\Delta \in \rho\text{-}HardIQC(\Psi, M, \rho)$ with

$$\Psi = \left[\begin{array}{cccc|cc} 0_p & 0_p & \ldots & 0_p & -I_p & 0_p \\ I_p & 0_p & \ldots & 0_p & 0_p & 0_p \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0_p & \ldots & I_p & 0_p & 0_p & 0_p \\ \hline 0_p & 0_p & \ldots & \rho^{2\tau} I_p & I_p & 0_p \\ 0_p & 0_p & \ldots & 0_p & 0_p & I_p \end{array}\right], \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{2.60}$$

Here the state dimension of Ψ is $p\tau \times 1$.

Proof. Statement 1 was originally proved as [44, Lemma 1]. Notice Statement 1 is a special case of Statement 3 with the choice of $\rho = 1$. Statement 3 can be proved using (18) and (19) in [60]. See [44, Corollary 1] for a proof of Statement 2. □

The bounded and monotone properties are of fundamental importance in the constructions of Zames-Falb IQCs. These properties are quite common in convex optimization. For example, the subdifferential of a convex function is monotone nondecreasing [61, Proposition 6.1.1]. A general connection between convexity and monotonicity is stated in [61, Theorem 4.1.4].

A bounded and monotone-nondecreasing function ϕ is further said to be slope-restricted to the interval $[m, L]$ if $[\phi(x) - \phi(y) - m(x-y)]^T [\phi(x) - \phi(y) - L(x-y)] \leq 0$ for all $x, y \in \mathbb{R}^p$. Zames-Falb IQCs characterized by m and L have been developed for such nonlinearities, e.g. see [44, Corollary 1]. The details of these IQCs are omitted here.

2.11.3 Gradients of Smooth Strongly-Convex Functions

Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$. Recall that the operator Δ_g maps $v \in \ell_{2e}^p$ to $w = \Delta_g(v)$ as $w^k = \nabla g(v^k + v^*) - \nabla g(v^*)$. Several commonly-used IQCs on Δ_g are

listed as follows.

Lemma 7. *Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$, $v = \{v^0, v^1, \ldots\} \in \ell_{2e}^p$, $v^* \in \mathbb{R}^p$, and $w^k = \nabla g(v^k + v^*) - \nabla g(v^*)$ for $k \geq 0$. Then*

1. *(Sector IQC): The pair (v^k, w^k) satisfies the sector constraint (2.54). Hence Δ_g satisfies the time domain hard and ρ-hard IQCs defined by (2.55).*

2. *(Off-by-One Hard IQC): Given any $N \geq 0$, one has*

$$\sum_{k=0}^{N} (w^k - mv^k)^T \left(Lv^k - w^k - Lv^{k-1} + w^{k-1} \right) \geq 0 \tag{2.61}$$

where v^k and w^k are set to be 0 for $k = -1$. Hence $\Delta_g \in HardIQC(\Psi, M)$ with

$$\Psi = \left[\begin{array}{c|cc} 0_p & -LI_p & I_p \\ \hline I_p & LI_p & -I_p \\ 0_p & -mI_p & I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{2.62}$$

3. *(Off-by-One ρ-hard IQC): Given any $0 < \rho \leq 1$, and $N \geq 0$, one has*

$$\sum_{k=0}^{N} \rho^{-2k} (w^k - mv^k)^T \left(Lv^k - w^k - \rho^2 Lv^{k-1} + \rho^2 w^{k-1} \right) \geq 0 \tag{2.63}$$

where v^k and w^k are set as 0 for $k = -1$. Hence $\Delta_g \in \rho\text{-}HardIQC(\Psi, M, \rho)$ with

$$\Psi = \left[\begin{array}{c|cc} 0_p & -LI_p & I_p \\ \hline \rho^2 I_p & LI_p & -I_p \\ 0_p & -mI_p & I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{2.64}$$

Proof. Statement 1 directly follows from Lemma 1. Statement 2 has been proved as [12, Lemma 8]. See [12, Lemma 10] for a proof of Statement 3. It is worth mentioning that the proof techniques in [12] work for arbitrary $v^* \in \mathbb{R}^p$, although the original proofs for [12, Lemma 8, Lemma 10] were written in a more tutorial style such that the discussions mainly focused on the case where v^* is an optimal point for g such that $w^* = \nabla g(v^*) = 0$. For readers' convenience, we sketch the main idea of the proof here. In the original proofs for [12, Lemma 8, Lemma 10], a new function $h(x) :=$

$g(x) - g(v^*) - \frac{m}{2}\|x - v^*\|^2$ was introduced. When $\nabla g(v^*) = 0$, one has $h \in \mathcal{F}(0, L-m)$, $\nabla h(v^*) = 0$, $h(x) \geq h(v^*) = 0\ \forall x \in \mathbb{R}^p$, and $(L-m)h(x) - \frac{1}{2}\|\nabla h(x)\|^2 \geq 0\ \forall x \in \mathbb{R}^p$. These facts were sufficient for constructing the proof in [12, Lemma 8, Lemma 10]. When $v^* \in \mathbb{R}^p$ is arbitrary, one only needs to slightly modify the introduced function as $h(x) := g(x) - g(v^*) - \nabla g(v^*)^T(x - v^*) - \frac{m}{2}\|x - v^*\|^2$. Then the rest of the proof is almost identical. □

Here we only include the IQCs which are most commonly-used for optimization analysis. It is worth mentioning that many other Zames-Falb IQCs on Δ_g can be derived using the slope-restricted property. By Lemma 1, it is clear that Δ_g is given by a nonlinearity which is slope-restricted to $[m, L]$. Hence the existing Zames-Falb IQCs for slope-restricted nonlinearities in $[44, 60]$ can be directly applied to Δ_g.

Remark 2. *Recall that linear convergence rate analysis of a deterministic first-order optimization method is equivalent to ρ-exponential stability analysis of an interconnection $F_u(G, \Delta)$. Notice all (Ψ, M) in Lemma 7 have the repeated block diagonal structure. For example, the state-space realization for Ψ in the off-by-one ρ-hard IQC is determined by $\left(0 \otimes I_p, \left[\begin{smallmatrix} -L & 1 \end{smallmatrix}\right] \otimes I_p, \left[\begin{smallmatrix} \rho^2 \\ 0 \end{smallmatrix}\right] \otimes I_p, \left[\begin{smallmatrix} L & -1 \\ -m & 1 \end{smallmatrix}\right] \otimes I_p\right)$. In addition, $M = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right] \otimes I_p$. Similarly, from Section 2.10, we can see that G also has this repeatable block diagonal structure. When applying Theorems 1 and 2, we can make use of this repeated block diagonal structure and Lemma 5 to formulate LMIs whose sizes are relatively small and do not depend on the parameter p. More explanations can be found in [12, Section 4.2].*

2.11.4 Multiplication with an Uncertain Parameter

In robustness analysis of optimization methods, a large class of perturbations Δ have a multiplicative form $w^k = \delta^k v^k$, where δ^k is the uncertain source term. Some examples of δ^k include, but are not limited to:

- Constant real scalar: $\delta^k = \delta \in [-1, 1]$
- Time-varying real scalar: $\delta^k \in [-1, 1]$
- Time-varying real matrix: $\delta^k \in \mathbb{R}^{n_w \times n_v}$

- Coefficients from a polytope: δ^k is a measurable matrix in a polytope of matrices with the extremal points $\delta_1, \ldots, \delta_N$
- Periodic real scalar: δ^k is a scalar function with period T, i.e. $\delta^{k+T} = \delta^k$
- Rate-bounded, time-varying scalar: δ^k satisfies $|\delta^{k+1} - \delta^k| \leq d$

A library of IQCs has been provided for the continuous-time counterparts of the above operators [19]. Many of these continuous-time IQCs have discrete-time counterparts. One example is provided by [12, Section 5.2], where IQCs for time-varying multiplicative uncertainty have been applied to study optimization methods subject to relative deterministic noise in the gradient computation.

The above multiplicative uncertainty can also be used in the worst-case analysis of an optimization method when the constant stepsize is only known to be in an interval $[\alpha_1, \alpha_2]$. Consider the gradient descent iteration $x^{k+1} = x^k - \alpha \nabla g(x^k)$ where $\alpha_1 \leq \alpha \leq \alpha_2$. Then this iteration can be rewritten as $x^{k+1} = x^k - (\frac{\alpha_1+\alpha_2}{2} + \delta\frac{\alpha_2-\alpha_1}{2})\nabla g(x^k)$ where δ is a constant in the interval $[-1, 1]$. Define the operator $\Delta_\delta : \ell_{2e}^p \to \ell_{2e}^p$ which maps $v \in \ell_{2e}^p$ to $w = \Delta_\delta(v)$ as $w^k = \delta v^k$. Then the gradient descent iteration can be cast as $F_u(G, \Delta)$ where G is determined by $(I_p, -I_p, I_p, 0_p)$, and $\Delta = (\frac{\alpha_1+\alpha_2}{2} + \frac{\alpha_2-\alpha_1}{2}\Delta_\delta) \circ \Delta_g$. Since δ is an unknown constant which does not change with k, hence Δ_δ satisfies the frequency domain IQC defined by the multiplier $\Pi = \left[\begin{smallmatrix} X(e^{j\omega}) & Y(e^{j\omega}) \\ Y(e^{j\omega})^* & -X(e^{j\omega}) \end{smallmatrix}\right]$, where $X(e^{j\omega}) = X(e^{j\omega})^* \geq 0$ and $Y(e^{j\omega}) = -Y(e^{j\omega})^*$ are bounded measurable matrix functions [19, Section VI.B].

2.11.5 Time Delay

In some distributed optimization problems, the gradient computations are subject to time delays [62]. The feedback representations of optimization methods with time delays involve a time delay operator Δ_τ, which is defined to map $v \in \ell_{2e}^p$ to $w = \Delta_\tau(v)$ as $w^k = 0$ for $k < \tau^k$ and $w^k = v^{k-\tau^k}$ for $k \geq \tau^k$, where $0 \leq \tau^k \leq \tau_{\max}$.

Various IQCs for Δ_τ are well documented in [46, Section III]. We briefly review one simplest IQC on Δ_τ as follows.

Lemma 8. *$\Delta_\tau \in HardIQC(\Psi, M)$ with $\Psi = diag(I_p, I_p)$ and $M = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -(\tau_{\max}+1)I_p \end{smallmatrix}\right]$.*

Proof. Suppose $v \in \ell_{2e}^p$ and $w = \Delta_\tau(v)$. For any $N \geq 0$, we have

$$\sum_{k=0}^{N} \|w^k\|^2 = \sum_{k=0}^{N} \|v^{k-\tau^k}\|^2 \leq \sum_{k=0}^{N} \left(\sum_{t=k-\tau_{\max}}^{k} \|v^k\|^2 \right) \leq (\tau_{\max}+1) \sum_{k=0}^{N} \|v^k\|^2 \qquad (2.65)$$

This leads to the desired conclusion. □

The above bound on the gain of Δ_τ is actually tight. In [46], it is stated that the operator Δ_τ is a bounded operator with $\|\Delta_\tau\| = \sqrt{\tau_{\max}+1}$.

We omit the details of many other related IQCs on Δ_τ. See [46, Section III] for a comprehensive treatment on related IQCs. Most of these IQCs involve loop transformation of the original feedback interconnection. Therefore, the application of these IQCs to uniform stability analysis requires careful justifications. The ρ-hard IQCs on Δ_τ have not been well developed. In Chapter 4, we will present one way to construct ρ-hard IQCs on Δ_τ.

Chapter 3

Hard Factorizations of Frequency Domain IQCs

As discussed previously, time domain hard IQCs are required in the dissipation inequality approach. This chapter focuses on the constructions of hard IQCs from frequency domain IQCs for bounded causal operators. The main result is a special J-spectral factorization [63,64] (Lemma 12 in Section 3.1). In particular, it is possible to factorize frequency domain multipliers Π as $\Psi^{\sim} M \Psi$. This factorization is not unique, and a time domain hard IQC can be specified by (Ψ, M) obtained by the J-spectral factorization. More specifically, $(\hat{\Psi}, \hat{J})$ is called a *J-spectral factorization* of $\Pi = \Pi^{\sim}$ if: (i) $\Pi = \hat{\Psi}^{\sim} \hat{J} \hat{\Psi}$, (ii) $\hat{J} = diag(I_{n_v}, -I_{n_w})$ and (iii) $\hat{\Psi}, \hat{\Psi}^{-1} \in \mathbb{RH}_{\infty}^{(n_v+n_w)\times(n_v+n_w)}$ [63]. In other words, the factorization yields a square, stable filter $\hat{\Psi}$ with a stable inverse and $\hat{J}$ is a signature matrix. A simple condition for the existence of a J-spectral factorization can be stated using the following definition.

Definition 8. *Let* $\Pi = \Pi^{\sim} \in \mathbb{RL}_{\infty}^{(n_v+n_w)\times(n_v+n_w)}$ *be partitioned as* $\begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^{\sim} & \Pi_{22} \end{bmatrix}$ *where* $\Pi_{11} \in \mathbb{RL}_{\infty}^{n_v \times n_v}$ *and* $\Pi_{22} \in \mathbb{RL}_{\infty}^{n_w \times n_w}$. Π *is a* Strict Positive-Negative (PN) multiplier *if there exists* $\epsilon > 0$ *such that*

(a) $\Pi_{11}(e^{j\omega}) \geq \epsilon I$ *for all* $\omega \in [0, 2\pi]$.

(b) $\Pi_{22}(e^{j\omega}) \leq -\epsilon I$ *for all* $\omega \in [0, 2\pi]$.

Π *is simply called a* PN multiplier *if (a) and (b) hold with* $\epsilon = 0$.

The PN terminology refers to the Positive semidefinite and Negative semidefinite properties specified by conditions (a) and (b) with $\epsilon = 0$. Strict-PN multipliers strictly satisfy (a) and (b) over all frequencies. It will be shown (Lemma 12 in Section 3.1.3) that if Π is a Strict-PN multiplier then it has a J-spectral factorization. This result is a variation of the canonical factorization theorem in [63]. Condition (a) with $\epsilon = 0$ is necessary and sufficient for the zero operator $\Delta \equiv 0$ to satisfy the frequency domain IQC defined by Π. Condition (b) with $\epsilon > 0$ implies that if $\Delta \in \mathrm{IQC}(\Pi)$ then Δ maps zero input to zero output. Bounded operators automatically have this zero input-zero output property. Condition (b) with $\epsilon = 0$ further implies that the set of all $\Delta \in \mathrm{IQC}(\Pi)$ is a convex set [65, 66]. The class of PN multipliers is quite general and covers the most typical multipliers used in IQC analysis. In fact, all of the IQCs listed in [19] satisfy Conditions (a) and (b) with $\epsilon = 0$ except for the IQCs for certain sector bounded nonlinearities and polytopic uncertainties.

A detailed discussion on the J-spectral factorization of Strict-PN multipliers is presented in Section 3.1. A game-theoretic interpretation is used to prove several important properties of this J-spectral factorization.

The J-spectral factorization result not only provides a systematic approach for hard IQC constructions, but also plays an important role in the proof of the modified dissipation inequality (Theorem 3) which differs in two respects from the standard dissipation/IQC result. Theorem 3 involves a dissipation inequality but it does not enforce the storage function to be non-negative. In addition, Theorem 3 allows for more general IQC parameterizations. In particular, the IQC need not be hard, i.e. it need not specify a valid finite-horizon integral constraint. Instead, the modified dissipation inequality replaces the constraints of non-negative storage functions and hard IQCs with a milder technical assumption on the combined multiplier. This technical assumption essentially implies that the combined multiplier has some hidden stored energy. As a result, the analysis condition can be reformulated into a valid dissipation inequality with a single hard IQC and a non-negative storage function. The modified dissipation inequality is presented in Section 3.2. Then some related work is reviewed in Section 3.3.

The J-spectral factorization result (Lemma 12) may also have other applications, e.g. formulating topological separation theorems [67]. A detailed discussion on these further applications is beyond the scope of this dissertation.

3.1 J-Spectral Factorizations and Related Games

This section presents a J-spectral factorization lemma (Lemma 12), which can be used to construct time domain hard IQCs from frequency domain IQCs (or equivalently time domain soft IQCs) for bounded causal operators. The J-spectral factorization lemma can also be used to prove a modified dissipation inequality theorem which is later presented in Section 3.2. This section first presents several intermediate game theory results which are required in the proof of the main J-spectral factorization lemma.

3.1.1 Open-loop Dynamic Games and IQC Factorizations

Suppose $\Pi = \Psi^{\sim} M \Psi$ is an arbitrary (not necessarily hard) factorization of the frequency domain IQC multiplier Π. This section connects properties of the factorization (Ψ, M) to the upper and lower values of an open-loop linear quadratic (LQ) discrete-time game. There is a large body of literature on linear quadratic discrete-time games [68, 69]. The results here build on previous results connecting discrete-time IQCs to min/max games [70]. Consider a two-player, zero-sum, linear quadratic difference game with the following cost defined by Ψ (with state space representation in Equation (2.33)) and matrix $M = M^T$:

$$
\begin{aligned}
J_{\Psi,M}(v, w, \psi^0) := \sum_{k=0}^{\infty} (r^k)^T M r^k \qquad (3.1)\\
\text{subject to:}\\
\psi^{k+1} = A_\psi \psi^k + B_{\psi 1} v^k + B_{\psi 2} w^k,\ \psi^0 \in \mathbb{R}^{n_\psi}\\
r^k = C_\psi \psi^k + D_{\psi 1} v^k + D_{\psi 2} w^k
\end{aligned}
$$

The infinite horizon cost function $J_{\Psi,M}$ is defined on $v \in \ell_2^{n_v}$, $w \in \ell_2^{n_w}$, and $\psi^0 \in \mathbb{R}^{n_\psi}$. Player 1 uses the "control variable" v to minimize $J_{\Psi,M}$ while Player 2 uses w to maximize $J_{\Psi,M}$. The game has an open-loop information structure and neither player can adapt their action during the game. The upper value of the game is defined as:

$$
\bar{J}_{\Psi,M}(\psi^0) := \inf_{v \in \ell_2^{n_v}} \sup_{w \in \ell_2^{n_w}} J_{\Psi,M}(v, w, \psi^0) \qquad (3.2)
$$

The lower value of the game is defined as

$$\underline{J}_{\Psi,M}(\psi^0) := \sup_{w\in\ell_2^{n_w}} \inf_{v\in\ell_2^{n_v}} J_{\Psi,M}(v,w,\psi^0) \tag{3.3}$$

The next two lemmas relate the upper and lower values of this open-loop game to the properties of the IQC factorization (Ψ, M). The proofs are only sketched as they are similar to those used in the continuous-time counterparts [26, Lemma 2, Lemma 3].

Lemma 9. *Let $\Pi = \Psi^{\sim} M \Psi \in \mathbb{RL}_\infty^{(n_v+n_w)\times(n_v+n_w)}$ be any factorization with $\Psi \in \mathbb{RH}_\infty^{n_r\times(n_v+n_w)}$. Let Δ be a bounded, casual operator with $\Delta \in IQC(\Pi)$. Then the following inequality holds for all $v \in \ell_{2e}^{n_v}$, $w = \Delta(v)$ and $N \geq 0$:*

$$\sum_{k=0}^{N} (r^k)^T M r^k \geq -\bar{J}_{\Psi,M}(\psi^{N+1}) \tag{3.4}$$

where r and ψ are the output and state of Ψ, respectively, driven by inputs (v,w) with initial condition $\psi^0 = 0$. Moreover, if $\bar{J}_{\Psi,M}(\psi) \leq 0\ \forall \psi \in \mathbb{R}^{n_\psi}$ then $\Delta \in HardIQC(\Psi, M)$.

Proof. Let $v \in \ell_{2e}^{n_v}$ and $w = \Delta(v)$ generate the state ψ and output r of Ψ from $\psi^0 = 0$. By Lemma 4, $\Delta \in \text{SoftIQC}(\Psi, M)$. The time domain soft IQC inequality (2.29) holds, by definition, for signals in ℓ_2. Hence it need not hold for r as the signals (v,w) are not necessarily in ℓ_2. Instead the causality of Δ is required to lower bound $\sum_{k=0}^{N}(r^k)^T M r^k$.

Let $\tilde{v} \in \ell_2^{n_v}$ be any signal that matches v up to time N, i.e. $(\tilde{v})_N = (v)_N$, and define $\tilde{w} := \Delta(\tilde{v})$. Let $\tilde{r}$ and $\tilde{\psi}$ denote the resulting output and state of Ψ starting from $\tilde{\psi}^0 = 0$. By causality of Δ and Ψ, if $(\tilde{v})_N = (v)_N$ then $(\tilde{w})_N = (w)_N$ and $(\tilde{r})_N = (r)_N$. This leads to the following bound:

$$\sum_{k=0}^{N} (r^k)^T M r^k = \sum_{k=0}^{N} (\tilde{r}^k)^T M \tilde{r}^k \geq -\sum_{k=N+1}^{\infty} (\tilde{r}^k)^T M \tilde{r}^k \tag{3.5}$$

The inequality follows from the time domain soft IQC (Ψ, M) applied to the signals $(\tilde{v}, \tilde{w})$. Thus any $\tilde{v}$ satisfying $(\tilde{v})_N = (v)_N$ can be used to lower bound the sum $\sum_{k=0}^{N}(r^k)^T M r^k$ for v. Maximizing over all feasible $\tilde{v}$ yields the following lower bound on $\sum_{k=0}^{N}(r^k)^T M r^k$:

$$\begin{aligned} &\sup_{\tilde{v}\in\ell_2^{n_v}} -\sum_{k=N+1}^{\infty} (\tilde{r}^k)^T M \tilde{r}^k \\ &\text{subject to: } (\tilde{v})_N = (v)_N,\ \tilde{w} = \Delta(\tilde{v}),\ \tilde{r} = \Psi \left[\begin{smallmatrix} \tilde{v} \\ \tilde{w} \end{smallmatrix}\right],\ \tilde{\psi}^0 = 0 \end{aligned} \tag{3.6}$$

The cost in this optimization only depends on the state of Ψ at $k = N+1$ and the signals $(\tilde{v}, \tilde{w})$ for $k \geq N+1$. Note that $\tilde{\psi}^{N+1}$ is the same for all feasible $\tilde{v}$. In particular, $((\tilde{v})_N, (\tilde{w})_N) = ((v)_N, (w)_N)$ for any feasible $\tilde{v}$. Hence Ψ evolves from $\tilde{\psi}^0 = 0$ to the state $\tilde{\psi}^{N+1} = \psi^{N+1}$ given by the inputs (v, w). Thus the lower bound can be expressed as:

$$\begin{aligned} &\sup_{\tilde{v}_l \in \ell_2^{n_v}} -\sum_{k=N+1}^{\infty} (\tilde{r}^k)^T M \tilde{r}^k \qquad (3.7) \\ &\text{subject to: } \tilde{w} = \Delta(\tilde{v}) \text{ where } \tilde{v}^k = \begin{cases} v^k & k \leq N \\ \tilde{v}_l^k & k > N \end{cases}, \\ &\tilde{\psi}^{k+1} = A_\psi \tilde{\psi}^k + B_{\psi 1} \tilde{v}^k + B_{\psi 2} \tilde{w}^k, \quad \tilde{\psi}^{N+1} = \psi^{N+1} \\ &\tilde{r}^k = C_\psi \tilde{\psi}^k + D_{\psi 1} \tilde{v}^k + D_{\psi 2} \tilde{w}^k \end{aligned}$$

In this bound, the relation $\tilde{w} = \Delta(\tilde{v})$ is the only constraint that connects the past $(k \leq N)$ to the future $(k > N)$. This connection is removed by replacing the true future output of Δ with a minimization over all possible output signals. This leads to the following lower bound on $\sum_{k=0}^{N} (r^k)^T M r^k$:

$$\begin{aligned} &\sup_{\tilde{v} \in \ell_2^{n_v}} \inf_{\tilde{w} \in \ell_2^{n_w}} -\sum_{k=N+1}^{\infty} (\tilde{r}^k)^T M \tilde{r}^k \qquad (3.8) \\ &\text{subject to:} \\ &\tilde{\psi}^{k+1} = A_\psi \tilde{\psi}^k + B_{\psi 1} \tilde{v}^k + B_{\psi 2} \tilde{w}^k, \quad \tilde{\psi}^{N+1} = \psi^{N+1} \\ &\tilde{r}^k = C_\psi \tilde{\psi}^k + D_{\psi 1} \tilde{v}^k + D_{\psi 2} \tilde{w}^k \end{aligned}$$

This removes the dependence on Δ but introduces some conservatism, i.e. the bound in Equation (3.8) is no greater than the bound in Equation (3.7). The time-invariance of Ψ is used to equivalently write Equation (3.8) as $-\bar{J}(\psi^{N+1})$. Hence (3.4) holds as desired.

If $\bar{J}_{\Psi,M}(\psi) \leq 0 \ \forall \psi \in \mathbb{R}^{n_\psi}$, then $\Delta \in \text{HardIQC}(\Psi, M)$ follows immediately from (3.4). □

Lemma 10. *Let $\Pi = \Psi^\sim M \Psi \in \mathbb{RL}_\infty^{(n_v+n_w)\times(n_v+n_w)}$ be any factorization with $\Psi \in \mathbb{RH}_\infty^{n_r \times (n_v+n_w)}$. Let $G \in \mathbb{RH}_\infty^{(n_v+n_e)\times(n_w+n_d)}$ be given. If $P = P^T$ satisfies the condition*

$LMI_{(G,\Psi)}(P,M) < 0$, then

$$V(\eta^0) := (\eta^0)^T P \eta^0 \geq \underline{J}_{\Psi,M}(\psi^0) \;\; \forall \eta^0 := \begin{bmatrix} \xi^0 \\ \psi^0 \end{bmatrix} \in \mathbb{R}^{n_\xi + n_\psi} \tag{3.9}$$

Moreover, if $\underline{J}_{\Psi,M}(\psi^0) \geq 0 \;\forall \psi^0 \in \mathbb{R}^{n_\psi}$ then $P \geq 0$.

Proof. $\underline{J}_{\Psi,M}\left(\psi^0\right)$ involves a min over $v \in \ell_2^{n_v}$. The choice of v may depend on w as long as v is an ℓ_2 signal. Choose v to be the output of G generated by $w \in \ell_2^{n_w}$ with any initial condition ξ^0. G is stable by assumption and hence this choice for v belongs to $\ell_2^{n_v}$. This yields a value that is no lower than the infimum over all possible $v \in \ell_2^{n_v}$. Hence $V^\star\left(\eta^0\right) \geq \underline{J}\left(\psi^0\right)$ where $V^\star$ is defined as:

$$V^\star\left(\eta^0\right) := \sup_{w \in \ell_2^{n_w}} \sum_{k=0}^{\infty} (r^k)^T M r^k \tag{3.10}$$

subject to:

$$\begin{bmatrix} \eta^{k+1} \\ r^k \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} \eta^k \\ w^k \end{bmatrix}, \quad \eta^0 := \begin{bmatrix} \xi^0 \\ \psi^0 \end{bmatrix} \in \mathbb{R}^{n_\xi + n_\psi} \tag{3.11}$$

The extended system state matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ are defined in Equations (2.25) and (2.26). The proof is completed by showing $V(\eta^0) \geq V^\star(\eta^0)$ for all η^0. This follows along the lines of Theorems 2 and 3 in [71] and hence the proof is only sketched. Let η and r be the resulting solutions of $\Psi \circ \left[\begin{smallmatrix} G \\ I \end{smallmatrix}\right]$ for a given input $w \in \ell_2^{n_w}$ and the initial condition η^0. Multiply $LMI_{(G,\Psi)}(P,M) < 0$ on the left/right by $\left[\begin{smallmatrix} (\eta^k)^T & (w^k)^T \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} (\eta^k)^T & (w^k)^T \end{smallmatrix}\right]^T$ to show $V(\eta^{k+1}) - V(\eta^k) + (r^k)^T M r^k \leq 0$. Sum this inequality from $k = 0$ to $k = N$ to obtain

$$V(\eta^0) \geq V(\eta^{N+1}) + \sum_{k=0}^{N} (r^k)^T M r^k \tag{3.12}$$

$\mathcal{A}$ is stable and hence $\lim_{N\to\infty} \eta^N = 0$ for any $w \in \ell_2^{n_w}$ (this is the so-called input-to-state stability). Maximizing the right side of Equation (3.12) over $w \in \ell_2^{n_w}$ for $N = \infty$ thus yields $V(\eta^0) \geq V^\star(\eta^0)$. Hence $V(\eta^0) \geq V^\star(\eta^0) \geq \underline{J}\left(\psi^0\right)$, and this proves (3.9).

If $\underline{J}_{\Psi,M}(\psi^0) \geq 0 \;\forall \psi^0 \in \mathbb{R}^{n_\psi}$, then $(\eta^0)^T P \eta^0 \geq 0 \;\forall \eta^0 \in \mathbb{R}^{n_\xi + n_\psi}$ and consequently $P \geq 0$. □

By Lemma 9, $\bar{J}_{\Psi,M}(\psi) \leq 0$ ensures the factorization (Ψ, M) leads to a time domain hard IQC. By Lemma 10, $\underline{J}_{\Psi,M}(\psi^0) \geq 0$ ensures the storage matrix satisfies $P \geq 0$. It

is easily shown that the two costs satisfy $\underline{J}_{\Psi,M}(\psi^0) \leq \bar{J}_{\Psi,M}(\psi^0)$ [68,69]. Hence the two conditions in Lemmas 9 and 10 can only be satisfied if $\bar{J}_{\Psi,M}(\psi^0) = \underline{J}_{\Psi,M}(\psi^0) = 0$ for all $\psi^0 \in \mathbb{R}^{n_\psi}$. It will be shown in Section 3.1.3 that the lower and upper values of the game are both equal to zero if (Ψ, M) is a J-spectral factorization.

3.1.2 Nash Equilibrium for the Two-Player Game

This section provides explicit values for $\bar{J}_{\Psi,M}(\psi^0)$ and $\underline{J}_{\Psi,M}(\psi^0)$ using the stabilizing solution of a related discrete-time algebraic Riccati equation. It is known that the upper and lower values can be effectively computed if a Nash equilibrium for the game exists [72, Theorem 3.26]. The basic intuition is provided before formally stating the result. Let $\Pi = \Psi^{\sim} M \Psi$ be the frequency domain multiplier associated with (Ψ, M). If $v \in \ell_2^{n_v}$ and $\psi^0 = 0$ then Parseval's theorem can be used to write $J_{\Psi,M}(v, 0, 0)$ in the frequency domain as:

$$J_{\Psi,M}(v,0,0) = \frac{1}{2\pi} \int_0^{2\pi} V(e^{j\omega})^* \, \Pi_{11}(e^{j\omega}) \, V(e^{j\omega}) d\omega \tag{3.13}$$

where $V(e^{j\omega})$ is the discrete-time Fourier transform of v. If $\Pi_{11}(e^{j\omega}) \geq \epsilon I$ for all $\omega \in [0, 2\pi]$ then $J_{\Psi,M}(v, 0, 0) \geq \epsilon \|v\|^2$. Similarly if $\Pi_{22}(e^{j\omega}) \leq -\epsilon I$ then $J_{\Psi,M}(0, w, 0) \leq -\epsilon \|w\|^2$ for all $w \in \ell_2^{n_w}$. Moreover, the Strict-PN condition actually implies that $J_{\Psi,M}$ is strictly convex in v and strictly concave in w. The following lemma constructs a Nash Equilibrium using the Strict-PN assumption.

Lemma 11. *Let $\Pi = \Psi^{\sim} M \Psi \in \mathbb{RL}_\infty^{(n_v+n_w)\times(n_v+n_w)}$ be any factorization with $\Psi \in \mathbb{RH}_\infty^{n_r \times (n_v+n_w)}$ and $M = M^T \in \mathbb{R}^{n_r \times n_r}$. Define $Q := C_\psi^T M C_\psi$, $S := C_\psi^T M D_\psi$ and $R := D_\psi^T M D_\psi$ where $(A_\psi, B_\psi, C_\psi, D_\psi)$ are the state matrices of Ψ. If Π is a Strict-PN multiplier then*

1. *There exists a unique, real, stabilizing solution $X = X^T$ to $DARE(A_\psi, B_\psi, Q, R, S)$. In addition, $R + B_\psi^T X B_\psi$ is nonsingular.*

2. *For $\psi^0 \in \mathbb{R}^{n_\psi}$ define $\mathbf{v} \in \ell_2^{n_v}$ and $\mathbf{w} \in \ell_2^{n_w}$ by*

$$\begin{bmatrix} \mathbf{v}^k \\ \mathbf{w}^k \end{bmatrix} := -K(A_\psi - B_\psi K)^k \psi^0 \tag{3.14}$$

where $K := (R + B_\psi^T X B_\psi)^{-1}(A_\psi^T X B_\psi + S)^T$ *is the stabilizing DARE gain. This input pair yields a value* $J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) = \psi^0 X \psi^0$ *for the two-player, LQ game in Equation* (3.1)*. In addition,* $(\mathbf{v}, \mathbf{w})$ *provides an open loop Nash equilibrium for this game, i.e.*

$$J_{\Psi,M}(\mathbf{v}, w, \psi^0) \le J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) \le J_{\Psi,M}(v, \mathbf{w}, \psi^0),\ \forall v \in \ell_2^{n_v},\ w \in \ell_2^{n_w} \tag{3.15}$$

3. $\bar{J}_{\Psi,M}(\psi^0) = \underline{J}_{\Psi,M}(\psi^0) = (\psi^0)^T X \psi^0$.

Proof. Statement 1 is a restatement of Lemma 26 in the appendix. If A_ψ is singular then Π has poles at $z = \infty$ and hence Π is non-proper. As a result, the proof of Lemma 26 requires the use of the descriptor system notation and results. This is the only technical lemma that requires descriptor systems and hence the proof is given in the appendix for readability.

To prove Statement 2, first note that $A_\psi - B_\psi K$ is a Schur stable matrix since X is the stabilizing solution of $DARE(A_\psi, B_\psi, Q, R, S)$. Hence $\mathbf{v}$ and $\mathbf{w}$ are ℓ_2 signals as claimed. The output of Ψ resulting from the inputs $(\mathbf{v}, \mathbf{w})$ and initial condition ψ^0 is

$$\mathbf{r}^k := C_\psi \boldsymbol{\psi}^k + D_\psi \begin{bmatrix} \mathbf{v}^k \\ \mathbf{w}^k \end{bmatrix} \tag{3.16}$$

where $\boldsymbol{\psi}^k := (A_\psi - B_\psi K)^k \psi^0$ is the state. This yields the following cost for the game:

$$J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) = \sum_{k=0}^{\infty} \begin{bmatrix} \boldsymbol{\psi}^k \\ \begin{bmatrix} \mathbf{v}^k \\ \mathbf{w}^k \end{bmatrix} \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi}^k \\ \begin{bmatrix} \mathbf{v}^k \\ \mathbf{w}^k \end{bmatrix} \end{bmatrix} \tag{3.17}$$

Substitute for Q using the DARE and use $\left[\begin{smallmatrix} \mathbf{v} \\ \mathbf{w} \end{smallmatrix}\right] = -K\boldsymbol{\psi}$. After completing the square the cost is written as

$$J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) = \sum_{k=0}^{\infty} \left((\boldsymbol{\psi}^k)^T X \boldsymbol{\psi}^k - (\boldsymbol{\psi}^{k+1})^T X \boldsymbol{\psi}^{k+1} \right) \tag{3.18}$$

This is a telescoping sum which yields $J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) = (\psi^0)^T X \psi^0$.

Next let $\psi \in \ell_2^{n_\psi}$ denote the state of Ψ for initial condition ψ^0 and arbitrary inputs $v \in \ell_2^{n_v}$ and $w \in \ell_2^{n_w}$. Define deviation signals as:

$$\delta_\psi := \psi - \boldsymbol{\psi}, \quad \delta_v := v - \mathbf{v}, \quad \delta_w := w - \mathbf{w} \tag{3.19}$$

Note that δ_v belongs to ℓ_2 since it is a difference of ℓ_2 signals. Similarly, δ_w and δ_ψ are also in ℓ_2. By linearity, δ_ψ is the state of Ψ driven by inputs (δ_v, δ_w) from zero initial conditions ($\delta_\psi^0 = 0$). The cost for the game with inputs (v, w) and initial condition ψ^0 is:

$$J_{\Psi,M}(v, w, \psi^0) = \sum_{k=0}^{\infty} \begin{bmatrix} \boldsymbol{\psi}^k + \delta_\psi^k \\ \begin{bmatrix} \mathbf{v}^k + \delta_v^k \\ \mathbf{w}^k + \delta_w^k \end{bmatrix} \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} \boldsymbol{\psi}^k + \delta_\psi^k \\ \begin{bmatrix} \mathbf{v}^k + \delta_v^k \\ \mathbf{w}^k + \delta_w^k \end{bmatrix} \end{bmatrix} \tag{3.20}$$

This can be expanded into four quadratic terms involving $(\boldsymbol{\psi}, \mathbf{v}, \mathbf{w})$ and $(\delta_\psi, \delta_v, \delta_w)$. Simplify using a similar completion of square and telescoping sum argument as above:

$$J_{\Psi,M}(v, w, \psi^0) = (\psi^0)^T X \psi^0 + (\psi^0)^T X \delta_\psi^0 + (\delta_\psi^0)^T X \psi^0 + \sum_{k=0}^{\infty} \begin{bmatrix} \delta_\psi^k \\ \begin{bmatrix} \delta_v^k \\ \delta_w^k \end{bmatrix} \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} \delta_\psi^k \\ \begin{bmatrix} \delta_v^k \\ \delta_w^k \end{bmatrix} \end{bmatrix} \tag{3.21}$$

The second and third terms are zero because $\delta_\psi^0 = 0$. The fourth term is equal to $J_{\Psi,M}(\delta_v, \delta_w, 0)$. Therefore, Equation (3.21) can be rewritten as

$$J_{\Psi,M}(v, w, \psi^0) = J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) + J_{\Psi,M}(\delta_v, \delta_w, 0) \tag{3.22}$$

This relation can be used to demonstrate that $(\mathbf{v}, \mathbf{w})$ provides an open loop Nash equilibrium. Specifically, Equation (3.22) directly leads to

$$J_{\Psi,M}(v, \mathbf{w}, \psi^0) - J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) = J_{\Psi,M}(\delta_v, 0, 0) \tag{3.23}$$

As discussed before the lemma, the Strict-PN assumption implies that $J_{\Psi,M}(\delta_v, 0, 0) \geq 0$. Hence Equation (3.23) implies

$$J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0) \leq J_{\Psi,M}(v, \mathbf{w}, \psi^0), \ \forall v \in \ell_2^{n_v} \tag{3.24}$$

The Strict-PN assumption and Equation (3.22) similarly imply that

$$J_{\Psi,M}(\mathbf{v}, w, \psi^0) \leq J_{\Psi,M}(\mathbf{v}, \mathbf{w}, \psi^0), \ \forall w \in \ell_2^{n_w} \tag{3.25}$$

This completes the proof of Statement 2.

Statement 3 follows from [72, Theorem 3.26]. The upper and lower values of the discrete-time linear quadratic game are both equal to the game value at the Nash equilibrium. □

Theorem 3.3 in [73] provides a related Nash equilibrium result for the continuous-time LQ game. The continuous-time result is more general in that it only requires (A_ψ, B_ψ) to be stabilizable. Lemma 11 requires the stronger assumption that A_ψ is stable. To the best of our knowledge, the discrete-time counterpart of Theorem 3.3 in [73] has not been established. However, the assumption that A_ψ is stable is sufficient for the IQC analysis considered in this paper. The proof for Statement 1 in Lemma 11 has some subtleties that do not appear in the continuous-time counterpart. In continuous-time, Π is assumed to be bounded on the closed imaginary axis and this implies that Π is proper. In discrete-time, Π is required to be bounded on the unit circle and hence Π can be improper. As a consequence, the discrete-time proof for Statement 1 in Lemma 11 cannot simply mimic its continuous-time counterpart. Instead a descriptor system representation of Π is required as is done for the proof of Lemma 26 in the appendix. Finally, notice that Statements 1 and 3 in Lemma 11 can also be proved by tailoring the operator-theoretic results in [64]. The operator-theoretic framework is more general while the linear algebra approach in this chapter is more closely aligned with possible numerical implementations.

3.1.3 J-Spectral Factorization for Strict-PN Multipliers

Lemma 12 provides a simple frequency domain condition on Π that is sufficient for the existence of a J-spectral factor. In addition, this lemma provides several useful properties of the J-spectral factorization. The Strict-PN assumption again plays a key role in the result.

Lemma 12. *Let $\Pi = \Psi^\sim M \Psi \in \mathbb{RL}_\infty^{(n_v+n_w)\times(n_v+n_w)}$ be any factorization with $\Psi \in \mathbb{RH}_\infty^{n_r\times(n_v+n_w)}$ and $M = M^T \in \mathbb{R}^{n_r\times n_r}$. Define $Q := C_\psi^T M C_\psi$, $S := C_\psi^T M D_\psi$ and $R := D_\psi^T M D_\psi$ where $(A_\psi, B_\psi, C_\psi, D_\psi)$ are the state matrices of Ψ. If Π is a Strict-PN multiplier then*

1. *Π has a J-spectral factorization $(\hat{\Psi}, \hat{J})$ with $\hat{J} := diag(I_{n_v}, -I_{n_w})$. Moreover, this J-spectral factorization can be constructed from the unique stabilizing solution X of $DARE(A_\psi, B_\psi, Q, R, S)$. Let $\hat{D}_\psi$ satisfy $\hat{D}_\psi^T \hat{J} \hat{D}_\psi = R + B_\psi^T X B_\psi$ and define*

$\hat{C}_\psi := \hat{J}\hat{D}_\psi^{-T}(B_\psi^T X A_\psi + S^T)$. *Then* $(\hat{\Psi}, \hat{J})$ *is a* J*-spectral factorization of* Π *with*

$$\hat{\Psi} := \left[\begin{array}{c|c} A_\psi & B_\psi \\ \hline \hat{C}_\psi & \hat{D}_\psi \end{array}\right] \tag{3.26}$$

2. $\hat{X} = 0$ *is the unique stabilizing solution of* $DARE(A_\psi, B_\psi, \hat{Q}, \hat{R}, \hat{S})$ *where* $\hat{Q} := \hat{C}_\psi^T \hat{J} \hat{C}_\psi$, $\hat{S} := \hat{C}_\psi^T \hat{J} \hat{D}_\psi$, *and* $\hat{R} := \hat{D}_\psi^T \hat{J} \hat{D}_\psi$.

3. $\bar{J}_{\hat{\Psi},\hat{J}}(\psi^0) = \underline{J}_{\hat{\Psi},\hat{J}}(\psi^0) = 0$, $\forall \psi^0 \in \mathbb{R}^{n_\psi}$.

4. $\Delta \in$ *HardIQC*$(\hat{\Psi}, \hat{J})$ *for any bounded, casual operator* $\Delta \in IQC(\Pi)$.

5. *For any* $G \in \mathbb{RH}_\infty^{n_v \times n_w}$, *and* $P = P^T$,

$$LMI_{(G,\Psi)}(P, M) = LMI_{(G,\hat{\Psi})}(\hat{P}, \hat{J}) \tag{3.27}$$

where $\hat{P} := P - \left[\begin{smallmatrix} 0 & 0 \\ 0 & X \end{smallmatrix}\right]$. *Moreover, if* $LMI_{(G,\hat{\Psi})}(\hat{P}, \hat{J}) < 0$ *then* $\hat{P} \geq 0$.

Proof. The existence of the stabilizing solution X follows from Lemma 11. Recall the stabilizing gain is given by $K := (R + B_\psi^T X B_\psi)^{-1}(A_\psi^T X B_\psi + S)^T$. A J-spectral factorization of Π can be constructed from X using a standard expansion technique [74]. First express Π as:

$$\Pi(z) = \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix}^\sim \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix} \tag{3.28}$$

Use the DARE and the definition of K to show:

$$\begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} = \begin{bmatrix} K^T \\ I \end{bmatrix} (R + B_\psi^T X B_\psi) \begin{bmatrix} K & I \end{bmatrix} - \begin{bmatrix} A_\psi^T X A_\psi - X & A_\psi^T X B_\psi \\ B_\psi^T X A_\psi & B_\psi^T X B_\psi \end{bmatrix} \tag{3.29}$$

Substitute Equation (3.29) into the expression for Π to obtain

$$\Pi(z) = \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix}^\sim \left(\begin{bmatrix} K^T \\ I \end{bmatrix} (R + B_\psi^T X B_\psi) \begin{bmatrix} K & I \end{bmatrix} \right) \begin{bmatrix} (zI - A_\psi)^{-1} B_\psi \\ I \end{bmatrix} \tag{3.30}$$

The Strict-PN conditions imply that $\Pi(e^{j\omega})$ has n_v positive eigenvalues and n_w negative eigenvalues for all $\omega \in [0, 2\pi]$. This follows from the Courant-Fischer minimax theorem

[75]. Moreover, $(R + B_\psi^T X B_\psi)$ must have the same signature as Π by Equation (3.30). Thus there exists a nonsingular matrix $\hat{D}_\psi$ such that $\hat{D}_\psi^T \hat{J} \hat{D}_\psi = R + B_\psi^T X B_\psi$ with $\hat{J} := diag(I_{n_v}, -I_{n_w})$. Finally, it can be verified from Equation (3.30) that $\hat{\Psi}$ as defined in the lemma satisfies $\Pi = \hat{\Psi}^\sim \hat{J} \hat{\Psi}$. It remains to show that $\hat{\Psi}^{-1}$ is stable. A realization for the inverse is

$$\hat{\Psi}^{-1} := \left[\begin{array}{c|c} A_\psi - B_\psi \hat{D}_\psi^{-1} \hat{C}_\psi & B_\psi \hat{D}_\psi^{-1} \\ \hline -\hat{D}_\psi^{-1} \hat{C}_\psi & \hat{D}_\psi^{-1} \end{array}\right] \tag{3.31}$$

The state matrix is $A_\psi - B_\psi \hat{D}_\psi^{-1} \hat{C}_\psi = A_\psi - B_\psi K$. This is a Schur stable matrix because K is the stabilizing gain. Hence $\hat{\Psi}^{-1}$ is a stable system and this completes the proof of Statement 1.

To prove Statement 2, first note that $(\hat{Q}, \hat{S}, \hat{R})$ as defined can be written as:

$$\begin{bmatrix} \hat{Q} & \hat{S} \\ \hat{S}^T & \hat{R} \end{bmatrix} = \begin{bmatrix} K^T \\ I \end{bmatrix} (R + B_\psi^T X B_\psi) \begin{bmatrix} K & I \end{bmatrix} \tag{3.32}$$

$\hat{R} = R + B_\psi^T X B_\psi$ is nonsingular as shown above and $\hat{Q} - \hat{S} \hat{R}^{-1} \hat{S}^T = 0$. Hence $\hat{X} = 0$ is a solution of $DARE(A_\psi, B_\psi, \hat{Q}, \hat{R}, \hat{S})$. The corresponding gain $\hat{K} = \hat{R}^{-1} \hat{S}^T = K$ is stabilizing since $A_\psi - B_\psi \hat{K} = A_\psi - B_\psi K$ is a Schur stable matrix. Thus $\hat{X} = 0$ is the unique stabilizing solution of $DARE(A_\psi, B_\psi, \hat{Q}, \hat{R}, \hat{S})$.

Next, note that $\bar{J}_{\hat{\Psi},\hat{J}}(\psi^0) = \underline{J}_{\hat{\Psi},\hat{J}}(\psi^0) = (\psi^0)^T \hat{X} \psi^0$ by Lemma 11. Hence Statement 3 follows from the fact $\hat{X} = 0$.

To prove Statement 4, note $\bar{J}_{\hat{\Psi},\hat{J}}(\psi^0) = 0$ for all $\psi^0 \in \mathbb{R}^{n_\psi}$. The factorization is hard if $\bar{J}_{\hat{\Psi},\hat{J}}(\psi^0) \leq 0$ for all $\psi^0 \in \mathbb{R}^{n_\psi}$ by Lemma 9. Hence $(\hat{\Psi}, \hat{J})$ is a hard factorization of Π.

To show Statement 5, first express the "M" term of $LMI_{(G,\Psi)}(P, M)$ as follows:

$$\begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix}^T = L^T \begin{bmatrix} C_\psi^T \\ D_\psi^T \end{bmatrix} M \begin{bmatrix} C_\psi & D_\psi \end{bmatrix} L \tag{3.33}$$

where L is given by

$$L = \left[\begin{array}{ccc|c} 0 & I & 0 & \\ C & 0 & D & \\ 0 & 0 & I & \end{array}\right] \tag{3.34}$$

Next use (3.29), (3.32) and the definitions of Q, S, R, $\hat{Q}$, $\hat{S}$, $\hat{R}$ to show

$$\begin{bmatrix} C_\psi^T \\ D_\psi^T \end{bmatrix} M \begin{bmatrix} C_\psi & D_\psi \end{bmatrix} = \begin{bmatrix} \hat{C}_\psi^T \\ \hat{D}_\psi^T \end{bmatrix} \hat{J} \begin{bmatrix} \hat{C}_\psi & \hat{D}_\psi \end{bmatrix} - \begin{bmatrix} A_\psi^T X A_\psi - X & A_\psi^T X B_\psi \\ B_\psi^T X A_\psi & B_\psi^T X B_\psi \end{bmatrix} \tag{3.35}$$

Substitute this expression into the "M" term of $LMI_{(G,\Psi)}(P, M, \gamma)$ (Equation (3.33)). Some lengthy but straightforward algebraic manipulations yield $LMI_{(G,\Psi)}(P, M) = LMI_{(G,\hat{\Psi})}(\hat{P}, \hat{J})$. Finally, it remains to show that the assumption $LMI_{(G,\hat{\Psi})}(\hat{P}, \hat{J}) < 0$ implies $\hat{P} \geq 0$. By Lemma 11, $\underline{J}_{\hat{\Psi},\hat{J}}(\psi^0) = (\psi^0)^T \hat{X} \psi^0$. By Lemma 10, $\hat{P} \geq 0$ if $\underline{J}_{\hat{\Psi},\hat{J}}(\psi^0) \geq 0$ for all $\psi^0 \in \mathbb{R}^{n_\psi}$. Thus $\hat{P} \geq 0$ since $\hat{X} = 0$ as already shown. □

The above result complements the minimax theorems in [70]. In particular, [70, Theorem 2.1] states a sufficient condition to ensure $\bar{J}_{\Psi,M}(\psi^0) = \underline{J}_{\Psi,M}(\psi^0)$. Statement 3 in Lemma 12 states the J-spectral factorization ensures the upper and lower game values are, in fact, both equal to zero. Moreover, [70, Theorem 2.2] states a sufficient condition to ensure a hard IQC factorization. The above result shows that the J-spectral factorization is hard and satisfies the extra "storage function" property mentioned in Statement 5 of Lemma 12.

Based on Lemma 12, time domain hard IQCs can be efficiently constructed from the frequency domain IQCs with Strict-PN multipliers. As a matter of fact, the hard IQC constructions can be done for non-strict PN multipliers using a perturbation argument. Notice only bounded operators are considered. Hence it is natural to prove Δ satisfies $\|\Delta\| \leq \gamma$ for some $\gamma > 0$. Hence, Δ satisfies the multiplier $\Pi_0 = \left[\begin{smallmatrix} \gamma^2 I_{n_v} & 0 \\ 0 & -I_{n_w} \end{smallmatrix}\right]$. Any PN multiplier Π can be perturbed to $\Pi + \epsilon\Pi_0$ for sufficiently small $\epsilon > 0$. The perturbed multipliers are Strict-PN such that the J-spectral factorization approach can be applied.

3.2 Modified Dissipation Inequality

As commented before, there are two key features of the standard dissipation inequality results from the last chapter. First, the IQCs are specified by conic combinations of time domain hard IQCs. The proof explicitly uses the time domain hard IQC inequalities $\sum_{k=0}^{N} (r_j^k)^T M_j r_j^k \geq 0$ $(j = 1, \ldots, N_J)$. Second, the constraint $P \geq 0$ is required for the construction of a valid storage function. The J-spectral factorization result is now

applied to prove a modified dissipation inequality that removes the constraint $P \geq 0$ and allows for the use of time domain soft IQCs.

The modified dissipation inequality result is stated as Theorem 3 below. The proof of this theorem relies on Lemma 12.

Theorem 3. *Let $G \in \mathbb{RH}_\infty^{n_v \times n_w}$ be a stable LTI system defined by (2.18) and $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a bounded, causal operator. Assume $F_u(G, \Delta)$ is well-posed and $\Delta \in$ SoftIQC$(\Psi, M(\lambda))$ for all λ in some set Λ. If $\exists$ a matrix $P = P^T$ and vector $\lambda \in \Lambda$ such that $\Psi^\sim M(\lambda)\Psi$ is a PN multiplier and $LMI_{(G,\Psi)}(P, M(\lambda)) < 0$, then $F_u(G, \Delta)$ is uniformly stable.*

Proof. First assume that $\Psi^\sim M(\lambda)\Psi$ is a Strict-PN multiplier. By Statement 1 of Lemma 12, this multiplier has a J-spectral factorization $(\hat{\Psi}, \hat{J})$ constructed from the stabilizing solution X of a related DARE. By Statement 4 of Lemma 12, if $\Delta \in$ SoftIQC$(\Psi, M(\lambda))$ then $\Delta \in$ HardIQC$(\hat{\Psi}, \hat{J})$. In other words, the J-spectral factorization provides a time domain hard IQC for Δ. By Statement 5 of Lemma 12, $LMI_{(G,\hat{\Psi})}(\hat{P}, \hat{J}) = LMI_{(G,\Psi)}(P, M(\lambda)) < 0$ where $\hat{P} := P - \left[\begin{smallmatrix} 0 & 0 \\ 0 & X \end{smallmatrix}\right]$. Thus the LMI condition can be rewritten using the J-spectral factorization. Finally, Statement 5 of Lemma 12 also implies that $\hat{P} \geq 0$. Hence the dissipation conditions hold using the hard IQC $(\hat{\Psi}, \hat{J})$ and storage matrix $\hat{P} \geq 0$. The analysis conclusions follow from the standard dissipation result in Theorem 1.

A perturbation argument is needed if $\Psi^\sim M(\lambda)\Psi$ is a PN multiplier. Δ is a bounded operator, by assumption. Hence it satisfies the multiplier $\Pi_0 := diag(\|\Delta\| I_{n_v}, -I_{n_w})$. For all $\epsilon > 0$, the perturbed multiplier $\Psi^\sim M(\lambda)\Psi + \epsilon\Pi_0$ is a Strict-PN multiplier that defines a valid frequency domain IQC for Δ. In addition, it can be factorized as:

$$\Psi_{pert}^\sim M_{pert}(\lambda, \epsilon)\Psi_{pert} := \begin{bmatrix} \Psi \\ I \end{bmatrix}^\sim \begin{bmatrix} M(\lambda) & 0 \\ 0 & \epsilon\Pi_0 \end{bmatrix} \begin{bmatrix} \Psi \\ I \end{bmatrix} \tag{3.36}$$

By Lemma 4, $\Delta \in$ SoftIQC$(\Psi_{pert}, M_{pert}(\lambda, \epsilon))$. Moreover, $LMI_{(G,\Psi)}(P, M(\lambda)) < 0$ implies that $LMI_{(G,\Psi_{pert})}(P, M_{pert}(\lambda, \epsilon)) < 0$ holds for sufficiently small $\epsilon > 0$. The result now follows using the arguments above with the Strict-PN multiplier given in Equation (3.36). $\square$

3.3 Related Work

The results in this chapter complement several existing results in the literature. First, the results here provide a discrete-time counterpart to the continuous-time results in [25, 26, 76, 77]. The intermediate results regarding discrete-time IQC factorizations and a related open-loop LQ difference game in this chapter parallel existing continuous-time results for J-spectral factorizations [78] and open loop LQ differential games [72, 73]. The generalization in this chapter is not immediate since descriptor systems are needed to handle non-proper multipliers that appear in some proofs. Similar discrete-time technical results on factorizations and LQ games are provided in [70] and [64] using operator theoretic methods. This chapter provides alternative linear algebra proofs for completeness. In particular, the minimax theorems in [70] were used to demonstrate the existence of hard IQCs for both discrete-time and continuous-time systems. This paper extends the game theoretic results to show several desired properties of a specific J-spectral factorization.

The benefit of the time domain dissipation theory is that it enables generalization to cases where the known system in the feedback connection is not necessarily LTI. For example, the approach enables the analysis of uncertain linear parameter varying systems or uncertain time-varying systems over finite horizons. The standard IQC homotopy theory developed for both continuous and discrete-time systems [19, 46, 48, 79, 80] can also be generalized for systems which do not have frequency domain interpretations [58]. The homotopy approach emphasizes input-output properties while internal states are incorporated more transparently in the dissipativity approach. Directly handling internal states can be potentially beneficial for the analysis of optimization methods since internal states are of interest in this setup [12]. In general, the two approaches are complementary and both are useful.

Chapter 4

Linear Rate Analysis Using Internal Uniform Stability Tests

We have demonstrated that the uniform stability analysis of $F_u(G, \Delta)$ can be performed for a large class of Δ. The main reason for this is that there exist a large library of known IQCs. In addition, both frequency and time domain IQCs can be directly incorporated into the uniform stability analysis. On the other hand, the library of ρ-hard IQCs is still under development. It is beneficial to connect the recently developed ρ-hard IQC approach with the standard IQC approach. In Section 2.10, we have explained that linear convergence rate analysis of a first-order optimization method is equivalent to ρ-exponential stability analysis of a related interconnection. In this chapter, we show that ρ-exponential stability of an interconnection $F_u(G, \Delta)$ is equivalent to uniform stability of a related scaled interconnection (Section 4.1). This enables derivation of linear rate testing conditions from uniform stability conditions using standard IQCs. This connection requires IQCs to be constructed for a scaled perturbation operator. A library of IQCs for this scaled perturbation operator is derived in Section 4.2. Section 4.3 discusses the connections between the proposed framework and the ρ-hard IQC approach. Section 4.4 builds upon our proposed framework and presents a GEVP formulation for linear rate analysis of deterministic first-order methods. In Section 4.5, we illustrate the utility of the derived GEVP condition via a case study of Nesterov's accelerated method.

4.1 Equivalence between ρ-Exponential Stability Analysis and Uniform Stability Analysis

This section establishes the connections between linear rate analysis and uniform stability analysis. The connections are built upon a specific loop transformation, as shown in Figure 4.1. For any fixed $\rho \in (0, 1]$, define the scaling operator $\mathcal{S}_{\rho+} : \ell_{2e}^{n_v} \to \ell_{2e}^{n_v}$ that maps v_ρ to $v = \mathcal{S}_{\rho+}(v_\rho)$ as follows: $v^k := \rho^k v_\rho^k$. Notice ρ^k denotes the k-th power of ρ while v_ρ^k is the k-th entry of the sequence v_ρ. Similarly, define another scaling operator $\mathcal{S}_{\rho-} : \ell_{2e}^{n_w} \to \ell_{2e}^{n_w}$ that maps w to $w_\rho = \mathcal{S}_{\rho-}(w)$ by setting $w_\rho^k := \rho^{-k} w^k$. The operators $\mathcal{S}_{\rho+}$ and $\mathcal{S}_{\rho-}$ have well-defined inverse operators denoted by $\mathcal{S}_{\rho+}^{-1}$ and $\mathcal{S}_{\rho-}^{-1}$, respectively. Notice $\mathcal{S}_{\rho+}^{-1} = \mathcal{S}_{\rho-}$ and $\mathcal{S}_{\rho-}^{-1} = \mathcal{S}_{\rho+}$ if and only if $n_v = n_w$. The connections between $F_u(G, \Delta)$ and $F_u(\mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}, \mathcal{S}_{\rho-} \circ \Delta \circ \mathcal{S}_{\rho+})$ are important for the results in this chapter. An almost identical loop transformation has been used in [29], which defines the scaled plant $\mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}$ in the frequency domain and relates the ρ-exponential stability of $F_u(G, \Delta)$ to the input-output stability of $F_u(\mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}, \mathcal{S}_{\rho-} \circ \Delta \circ \mathcal{S}_{\rho+})$. We will relate the ρ-exponential stability of $F_u(G, \Delta)$ to the uniform stability of $F_u(\mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}, \mathcal{S}_{\rho-} \circ \Delta \circ \mathcal{S}_{\rho+})$. This requires a specific time domain state space definition for $\mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}$, which leads to useful relationships between the states of the original and transformed interconnections.

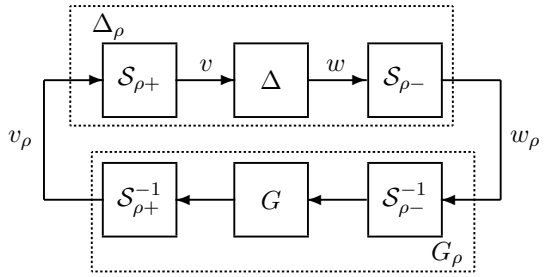

Figure 4.1: Transformed Interconnection

Define the scaled systems $G_\rho := \mathcal{S}_{\rho+}^{-1} \circ G \circ \mathcal{S}_{\rho-}^{-1}$ and $\Delta_\rho := \mathcal{S}_{\rho-} \circ \Delta \circ \mathcal{S}_{\rho+}$. These are input/ouput definitions for the scaled systems. A specific, state-space realization for G_ρ can be obtained from shifting the state-space model of G in Equation (2.18). Define

$\xi_\rho^k := \rho^{-k}\xi^k$. A state-space realization for G_ρ is then given by:

$$\begin{aligned} \xi_\rho^{k+1} &= \rho^{-1}A\xi_\rho^k + \rho^{-1}Bw_\rho^k \\ v_\rho^k &= C\xi_\rho^k + Dw_\rho^k \end{aligned} \tag{4.1}$$

As a slight abuse of notation, the scaled system G_ρ will always refer to this specific state-space realization. The main loop transformation result is now stated.

Theorem 4. *Assume $0 < \rho \leq 1$. $F_u(G, \Delta)$ is well-posed if and only if $F_u(G_\rho, \Delta_\rho)$ is well-posed. Moreover, $F_u(G, \Delta)$ is exponentially stable with rate ρ if and only if $F_u(G_\rho, \Delta_\rho)$ is uniformly stable.*

Proof. It is straightforward to prove that $\xi \in \ell_{2e}^{n_\xi}$, $v \in \ell_{2e}^{n_v}$, and $w \in \ell_{2e}^{n_w}$ is a solution for Equation (2.18) and $w = \Delta(v)$ with initial condition $\xi^0 \in \mathbb{R}^{n_\xi}$ if and only if ($\xi^k\rho^{-k}$, $v^k\rho^{-k}$, $w^k\rho^{-k}$) provides an ℓ_{2e} solution for Equation (4.1) and $w_\rho = \Delta_\rho(v_\rho)$ with initial condition $\xi_\rho^0 = \xi^0$. Therefore, $F_u(G, \Delta)$ is well-posed if and only if $F_u(G_\rho, \Delta_\rho)$ is well-posed. Next suppose $F_u(G, \Delta)$ and $F_u(G_\rho, \Delta_\rho)$ are well-posed and have the same initial condition $\xi^0 = \xi_\rho^0$. The following holds

$$\begin{aligned} \xi_\rho^k &= \rho^{-k}\xi^k \\ w_\rho^k &= \rho^{-k}w^k \\ v_\rho^k &= \rho^{-k}v^k \end{aligned} \tag{4.2}$$

where (ξ, v, w) and $(\xi_\rho, v_\rho, w_\rho)$ are the resultant ℓ_{2e} solutions for $F_u(G, \Delta)$ and $F_u(G_\rho, \Delta_\rho)$, respectively. Moreover, $\|\xi^k\| \leq c\|\xi^0\|\rho^k \Leftrightarrow \|\xi_\rho^k\| \leq c\|\xi_\rho^0\|$. Therefore, $F_u(G, \Delta)$ is exponentially stable with rate ρ if and only if $F_u(G_\rho, \Delta_\rho)$ is uniformly stable. □

Remark 3. *Proposition 5 in [29] states that input-output stability of the transformed loop is a sufficient condition for ρ -exponential stability of the original loop. Theorem 4 here states that uniform stability of the transformed loop is a necessary and sufficient condition for ρ-exponential stability of the original loop.*

Theorem 4 states that a uniform stability test for $F_u(G_\rho, \Delta_\rho)$ is equivalent to a ρ-exponential stability test for $F_u(G, \Delta)$. Thus LMI conditions formulated for uniform stability of the scaled interconnection $F_u(G_\rho, \Delta_\rho)$ can be used to test ρ-exponential

stability of the original loop. This approach requires IQCs to be specified for Δ_ρ. Most existing IQCs were specified for the unscaled operator Δ. One contribution of this chapter is that a library of IQCs for Δ_ρ is derived in Section 4.2 for a large class of operators. Note that this IQC construction step requires the operator Δ_ρ to be causal. It is easily shown that causality of Δ_ρ is equivalent to causality of Δ. This follows because $\mathcal{S}_{\rho-}$ and $\mathcal{S}_{\rho+}$ are memoryless, pointwise-in-time multiplication operators. The frequency domain construction of IQC multipliers for Δ_ρ requires its boundedness, which is not as straightforward. Since $\mathcal{S}_{\rho-}$ is an unbounded operator, it is possible for a bounded operator Δ to yield an unbounded scaled operator Δ_ρ. The boundedness of Δ_ρ needs to be proven for each specific Δ. This issue is addressed in Section 4.2.

LMI conditions for ρ-exponential stability of $F_u(G,\Delta)$ are now formulated using the loop transformation result in Theorem 4 and the IQC-based uniform stability tests (Theorems 1 and 3).

Theorem 5. *Let G be an LTI system defined by (2.18) and $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a causal operator such that $F_u(G,\Delta)$ is well-posed. If one of the following conditions holds*

1. *The operator Δ_ρ satisfies the time domain hard IQCs defined by $\{(\Psi_j, M_j)\}_{j=1}^{N_J}$, and $\exists$ a matrix $P = P^T > 0$ and scalars $\lambda_j \geq 0$ such that $LMI_{(G_\rho,\Psi)}(P, M_\lambda) \leq 0$.*

2. *The operator Δ_ρ is bounded, and $\Delta_\rho \in SoftIQC(\Psi, M(\lambda))$ for all λ in some set Λ. In addition, $G_\rho \in \mathbb{RH}_\infty^{n_v \times n_w}$ and there exists a matrix $P = P^T$ and vector $\lambda \in \Lambda$ such that $\Psi^\sim M(\lambda)\Psi$ is a PN multiplier and $LMI_{(G_\rho,\Psi)}(P, M(\lambda)) < 0$.*

then $F_u(G,\Delta)$ is exponentially stable with rate ρ.

Proof. By Theorem 4, the well-posedness of $F_u(G,\Delta)$ implies that $F_u(G_\rho,\Delta_\rho)$ is well-posed. Moreover, the causality of Δ implies the causality of Δ_ρ. Clearly, Δ_ρ maps ℓ_{2e} signals to ℓ_{2e} signals. It follows from Theorems 1 and 3 that $F_u(G_\rho,\Delta_\rho)$ is uniformly stable. Based on Theorem 4, $F_u(G,\Delta)$ is exponentially stable with rate ρ. □

4.2 Boundedness and IQCs for Scaled Perturbation

This section provides a list of IQCs for the scaled operator Δ_ρ. The results are developed for several important (unscaled) components Δ, which have been reviewed in

Section 2.11.

Some IQCs developed in this section are specified as frequency domain multipliers. The frequency domain IQC constructions for Δ_ρ require its boundedness. Hence we also check the boundedness of Δ_ρ for each specific Δ. If the boundedness of Δ_ρ is checked and the specified frequency domain IQC multiplier is PN, then the J-spectral factorization results and related perturbation arguments in the last chapter can be used to construct corresponding time domain hard IQCs.

4.2.1 Scaled Operator for Memoryless Nonlinearity in a Sector

Consider the operator Δ defined in Section 2.11.1. Suppose the operator $\Delta : \ell_{2e}^p \to \ell_{2e}^p$ maps v to $w = \Delta(v)$ as $w^k = \phi(v^k, k)$, where $\phi : \mathbb{R}^p \times \mathbb{Z}^+ \to \mathbb{R}^p$ is in a sector: $\left(\phi(v^k,k) - Lv^k\right)^T \left(\phi(v^k,k) - mv^k\right) \le 0$. Then Δ_ρ maps v_ρ to $w_\rho = \Delta_\rho(v_\rho)$ as $w_\rho^k = \rho^{-k}\phi(v_\rho^k\rho^k, k)$. It is straightforward to verify

$$\left(\rho^{-k}\phi(v_\rho^k\rho^k,k) - Lv_\rho^k\right)^T \left(\rho^{-k}\phi(v_\rho^k\rho^k,k) - mv_\rho^k\right) \le 0 \tag{4.3}$$

Therefore, Δ_ρ is a bounded operator and $\|\Delta_\rho\| \le \max(|m|, |L|)$. Moreover, $\Delta_\rho \in \text{HardIQC}(\Psi, M)$ with $\Psi = \begin{bmatrix} LI_p & -I_p \\ -mI_p & I_p \end{bmatrix}$ and $M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}$.

4.2.2 Scaled Operator for Static Nonlinearity

Consider the static nonlinearity in Section 2.11.2. Suppose $\Delta : \ell_{2e}^p \to \ell_{2e}^p$ maps v to $w = \Delta(v)$ as $w^k = \phi(v^k)$, where $\phi : \mathbb{R}^p \to \mathbb{R}^p$ is a continuous function. In addition, ϕ is assumed to be bounded, i.e. $\exists c$ s.t. $\|\phi(x)\| \le c\|x\|$, $\forall x \in \mathbb{R}^p$. Then Δ_ρ maps v_ρ to $w_\rho = \Delta_\rho(v_\rho)$ as $w_\rho^k = \rho^{-k}\phi(\rho^k v_\rho^k)$. The boundedness of ϕ implies that $\|w_\rho^k\| \le c\rho^{-k}\|\rho^k v_\rho^k\| = c\|v_\rho^k\|$ for some c. Hence Δ_ρ is bounded. When ϕ lies within a sector $[m, L]$ for finite m and L, the multiplier in Section 4.2.1 can be directly applied. When ϕ is bounded and monotone nondecreasing, Zames-Falb IQCs can be constructed for Δ_ρ. The following lemma is useful for such IQC constructions.

Lemma 13. *Let $\phi : \mathbb{R}^p \to \mathbb{R}^p$ be bounded and monotone nondecreasing. Suppose ϕ is a gradient of some potential function which maps from $\mathbb{R}^p$ to $\mathbb{R}$. Then*

1. *For any* $v = \{v^0, v^1, \ldots\} \in \ell_2^p$, $\tau \geq 0$, *and* $k_0 \geq 0$, *one has*

$$\sum_{k=k_0}^{\infty} (v^k)^T \left(w^k - w^{k+\tau} \right) \geq 0 \tag{4.4}$$

where $w^k = \phi(v^k)$.

2. *If* $0 < \rho \leq 1$, $\{v^k \rho^{-k} : k = 0, 1, \ldots\} \in \ell_2^p$, $w^k = \phi(v^k)$, *and* $\tau \geq 0$, *then*

$$\sum_{k=0}^{\infty} \rho^{-2k} (v^k)^T (w^k - w^{k+\tau}) \geq 0 \tag{4.5}$$

In addition, set $v^k = 0$ *for* $k < 0$. *One has*

$$\sum_{k=0}^{\infty} \rho^{-2(k-\tau)} (w^k)^T v^{k-\tau} \leq \sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T v^k \tag{4.6}$$

Proof. To prove Statement 1, first let $k_0 = 0$. It follows from (2.57) that:

$$\sum_{k=0}^{\infty} (w^k)^T v^k \geq \sum_{k=0}^{\infty} (w^k)^T v^{k-\tau} \tag{4.7}$$

where $v^k = 0$ for $k < 0$. The right side can be re-written with a change of variables as:

$$\sum_{k=0}^{\infty} (w^k)^T v^k \geq \sum_{k=0}^{\infty} (w^{k+\tau})^T v^k \tag{4.8}$$

Notice $(w^k)^T v^k = (v^k)^T w^k$, and hence (4.4) holds for $k_0 = 0$. For $k_0 \geq 1$, set $\tilde{v} := v - P_{k_0-1}(v)$. Then

$$\sum_{k=0}^{\infty} (\tilde{v}^k)^T \left(\tilde{w}^k - \tilde{w}^{k+\tau} \right) \geq 0, \forall \tau \geq 0 \tag{4.9}$$

This proves Statement 1 for any $k_0 \geq 0$.

To prove Statement 2, first notice $\{w^k \rho^{-k} : k = 0, 1, \ldots\} \in \ell_2^p$ since $\{v^k \rho^{-k} : k = 0, 1, \ldots\} \in \ell_2^p$ and ϕ is bounded. Therefore, the integral on the left side of (4.5) is finite (Cauchy-Schwartz). Since $\rho^{-2k} = 1 + \sum_{k_0=1}^{k} (1 - \rho^2) \rho^{-2k_0}$, the left side of (4.5) equals

$$\begin{aligned} &\sum_{k=0}^{\infty} (v^k)^T \left(w^k - w^{k+\tau} \right) \\ &+ \sum_{k=0}^{\infty} \left(\sum_{k_0=1}^{k} (1 - \rho^2) \rho^{-2k_0} \right) (v^k)^T \left(w^k - w^{k+\tau} \right) \end{aligned} \tag{4.10}$$

The first summation on the left side is finite because $v, w \in \ell_2^p$ (Cauchy-Schwartz) and hence the double integral is also finite. From Fubini's theorem, this double summation can be re-arranged as

$$\sum_{k_0=1}^{\infty} \left(\sum_{k=k_0}^{\infty} (v^k)^T \left(w^k - w^{k+\tau} \right) \right) (1-\rho^2)\rho^{-2k_0} \tag{4.11}$$

Statement 1 implies the inner summation in (4.11) is ≥ 0 $\forall k_0 \geq 0$. Thus the double summation in (4.10) is ≥ 0. By Statement 1, the first term in (4.10) is also ≥ 0. Hence (4.5) holds. Finally, rewrite the left side of (4.6) with a change of variables as $\sum_{k=0}^{\infty} \rho^{-2k} (w^{k+\tau})^T v^k$. Thus (4.6) is equivalent to (4.5), and Statement 2 is true. □

Now Zames-Falb IQCs for Δ_ρ can be constructed as follows.

Lemma 14. *Let $\phi : \mathbb{R}^p \to \mathbb{R}^p$ be bounded and monotone nondecreasing. Suppose ϕ is a gradient of some potential function which maps from $\mathbb{R}^p$ to $\mathbb{R}$. Then*

1. *(Off-by-τ Hard IQC): For any $v_\rho = \{v_\rho^0, v_\rho^1, \ldots\} \in \ell_{2e}^p$, $\tau \geq 0$, $0 < \rho \leq 1$, and $N \geq 0$, one has*

$$\sum_{k=0}^{N} (w_\rho^k)^T \left(v_\rho^k - \rho^\tau v_\rho^{k-\tau} \right) \geq 0 \tag{4.12}$$

where $v_\rho^k = 0$ for $k < 0$ and $w_\rho^k = \rho^{-k}\phi(\rho^k v_\rho^k)$. Hence $\Delta_\rho \in HardIQC(\Psi, M)$ with

$$\Psi = \left[\begin{array}{cccc|cc} 0_p & 0_p & \ldots & 0_p & -I_p & 0_p \\ I_p & 0_p & \ldots & 0_p & 0_p & 0_p \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0_p & \ldots & I_p & 0_p & 0_p & 0_p \\ \hline 0_p & 0_p & \ldots & \rho^\tau I_p & I_p & 0_p \\ 0_p & 0_p & \ldots & 0_p & 0_p & I_p \end{array}\right], \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{4.13}$$

Here the state dimension of Ψ is $p\tau \times 1$.

2. *(Forward Off-by-τ Soft IQC): For any $v_\rho \in \ell_2^p$, $0 < \rho \leq 1$, and $\tau \geq 0$, one has*

$$\sum_{k=0}^{\infty} (\rho^{-\tau} v_\rho^k)^T \left(\rho^{-\tau} w_\rho^k - w_\rho^{k+\tau} \right) \geq 0 \tag{4.14}$$

where $w^k = \rho^{-k}\phi(\rho^k v_\rho^k)$.

3. *(Frequency Domain Zames-Falb IQC): Let* $h \in \ell_{2e}$ *satisfy* $\sum_{k=0}^{\infty} \rho^{-k} h^k \leq 1$ *for some* $0 < \rho \leq 1$ *and* $h^k \geq 0$ *for all* k*. Then,* $\Delta_\rho \in IQC(\Pi)$ *with* $\Pi = \left[\begin{smallmatrix} 0 & 1-H^* \\ 1-H & 0 \end{smallmatrix}\right] \otimes I_p$ *where* H *denotes the Laplace transform of* h*.*

Proof. To prove Statement 1, we define $\tilde{w}_\rho = P_N(w_\rho)$, $\tilde{v}_\rho = P_N(v_\rho)$, $\tilde{v}^k = \rho^k \tilde{v}_\rho^k$, and $\tilde{w}^k = \rho^k \tilde{w}_\rho^k$. By Lemma 13, we have

$$\sum_{k=0}^{\infty} \rho^{-2(k-\tau)} (\tilde{w}^k)^T \tilde{v}^{k-\tau} \leq \sum_{k=0}^{\infty} \rho^{-2k} (\tilde{w}^k)^T \tilde{v}^k \tag{4.15}$$

The above inequality immediately leads to the off-by-τ IQC for Δ_ρ.

To prove Statement 2, we define $v^k = \rho^k v_\rho^k$, and $w^k = \rho^k w_\rho^k$. Then by Lemma 13, we know (4.5) holds. Notice (4.5) is equivalent to (4.14). Then Statement 2 is true.

To prove Statement 3, let v_ρ be in ℓ_2^p. It suffices to show that

$$\sum_{k=0}^{\infty} (w_\rho^k)^T (h * v_\rho)^k \leq \sum_{k=0}^{\infty} (w_\rho^k)^T v_\rho^k \tag{4.16}$$

where $(h * v_\rho)^k$ denotes the k-th entry of the sequence $h * v_\rho$, which is the convolution of h and v_ρ. Notice (4.16) is equivalent to

$$\sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T (\bar{h} * v)^k \leq \sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T v^k \tag{4.17}$$

where $\bar{h}^\tau := \rho^\tau h^\tau \in \ell_1$. Since $v_\rho \in \ell_2^p$ and Δ_ρ is bounded, one has $w_\rho \in \ell_2^p$. Moreover $h \in \ell_1$ implies $h * v_\rho \in \ell_2^p$. It follows from Cauchy-Schwartz inequality that the left side of (4.16) (and hence the left side of (4.17)) is finite. Hence, Fubini's Theorem can be used to rewrite the left side of (4.17) as

$$\sum_{k=0}^{\infty} \sum_{\tau=0}^{\infty} \rho^{-2k} (w^k)^T \bar{h}^\tau v^{k-\tau} = \sum_{\tau=0}^{\infty} \rho^{-2\tau} \bar{h}^\tau \left(\sum_{k=0}^{\infty} \rho^{-2(k-\tau)} (w^k)^T v^{k-\tau} \right) \tag{4.18}$$

Since $v_\rho, w_\rho \in \ell_2^p$, Statement (2) of Lemma 13 can be directly applied to show the first inequality below:

$$\sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T (\bar{h} * v)^k \leq \left(\sum_{\tau=0}^{\infty} \rho^{-2\tau} \bar{h}^\tau \right) \left(\sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T v^k \right) \leq \sum_{k=0}^{\infty} \rho^{-2k} (w^k)^T v^k \tag{4.19}$$

The second inequality follows from the definition of $\bar{h}$ and the assumptions on h. Thus (4.17) holds. This completes the proof. □

The forward off-by-τ soft IQC (4.14) is important since it can be used to formulate a GEVP for rate analysis. We will discuss this more carefully in Section 4.4.

Remark 4. *Following the procedure in [44], the above result can be extended to odd or slope-restricted nonlinearities. Another important related result is the frequency domain ρ-IQC construction of Zames-Falb multipliers for the original operaotr Δ [29].*

4.2.3 Scaled Operator for Gradients of Smooth Strongly-Convex Functions

Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$. Recall that the operator Δ_g maps $v \in \ell_{2e}^p$ to $w = \Delta_g(v)$ as $w^k = \nabla g(v^k + v^*) - \nabla g(v^*)$. Then Δ_ρ maps v_ρ to $w_\rho = \Delta_\rho(v_\rho)$ as $w_\rho^k = \rho^{-k}\left(\nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*)\right)$. It is straightforward to show that $\|w_\rho^k\| \leq L\rho^{-k}\|\rho^k v_\rho^k\| = L\|v_\rho^k\|$. Hence Δ_ρ is bounded. Now we present the following Zames-Falb IQCs on Δ_ρ.

Lemma 15. *Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$, $v_\rho = \{v_\rho^0, v_\rho^1, \ldots\} \in \ell_{2e}$, $v^* \in \mathbb{R}^p$, and $w_\rho^k = \rho^{-k}\left(\nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*)\right)$ for $k \geq 0$. Then*

1. *(Sector IQC): The pair (v_ρ^k, w_ρ^k) satisfies the sector constraint (2.54). Hence Δ_ρ satisfies the time domain hard and ρ-hard IQCs defined by (2.55).*

2. *(Off-by-One Hard IQC): For any $N \geq 0$, one has*

$$\sum_{k=0}^{N}(w_\rho^k - mv_\rho^k)^T\left(Lv_\rho^k - w_\rho^k - \rho(Lv_\rho^{k-1} - w_\rho^{k-1})\right) \geq 0 \tag{4.20}$$

   where $v_\rho^{-1} = 0$ and $w_\rho^{-1} = 0$. Hence $\Delta_\rho \in HardIQC(\Psi, M)$ with

$$\Psi = \left[\begin{array}{c|cc} 0_p & -LI_p & I_p \\ \hline \rho I_p & LI_p & -I_p \\ 0_p & -mI_p & I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{4.21}$$

3. *(Forward Off-by-One Soft IQC): If $v_\rho = \{v_\rho^0, v_\rho^1, \ldots\} \in \ell_2^p$, and $0 < \rho \leq 1$, one has*

$$\sum_{k=0}^{\infty}\rho^{-1}(Lv_\rho^k - w_\rho^k)^T\left(\rho^{-1}(w_\rho^k - mv_\rho^k) - (w_\rho^{k+1} - mv_\rho^{k+1})\right) \geq 0 \tag{4.22}$$

Hence $\Delta_\rho \in SoftIQC(\Psi, M)$ *with*

$$\Psi = \left[\begin{array}{cc|cc} 0_p & 0_p & \rho^{-1}LI_p & -\rho^{-1}I_p \\ 0_p & 0_p & -\rho^{-1}mI_p & \rho^{-1}I_p \\ \hline I_p & 0_p & 0_p & 0_p \\ 0_p & I_p & mI_p & -I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{4.23}$$

Proof. By Lemma 1, we have

$$\begin{bmatrix} \rho^k v_\rho^k \\ \nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*) \end{bmatrix}^T \begin{bmatrix} -2mLI_p & (m+L)I_p \\ (m+L)I_p & -2I_p \end{bmatrix} \begin{bmatrix} \rho^k v_\rho^k \\ \nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*) \end{bmatrix} \geq 0$$

We can extract the factor ρ^k out of the above inequality to prove Statement 1.

Statements 2 and 3 can be proved using the same technique in Lemma 14. Specifically, one can combine the expanding trick in Lemma 13 with Inequality (2.61) to prove these statements. The details are omitted. □

Notice the forward off-by-one soft IQC has a very special Ψ whose state matrices have the form $(\rho^{-1}A, \rho^{-1}B, C, D)$ where A, B, C, D are constant matrices which do not depend on ρ. Moreover, G_ρ also has this property. Hence a GEVP can be formulated when we use this forward off-by-one soft IQC to derive LMIs for the uniform stability analysis of $F_u(G_\rho, \Delta_\rho)$. We will further discuss this in Section 4.4.

4.2.4 Scaled Operator for Multiplicative Perturbation

As discussed in Section 2.11.4, a large class of perturbations Δ have a multiplicative form $w^k = \delta^k v^k$, where δ^k is the uncertain source term. Some examples of δ have been reviewed in Section 2.11.4. In this case, Δ and the scaling operator $\mathcal{S}_{\rho\pm}$ commute: $\Delta \circ \mathcal{S}_{\rho\pm} = \mathcal{S}_{\rho\pm} \circ \Delta$. Therefore, the scaling relationship directly leads to $w_\rho^k = \delta^k v_\rho^k$, and $\Delta_\rho = \Delta$. The boundedness of Δ guarantees that Δ_ρ is a bounded operator, and any IQCs on Δ are directly IQCs on Δ_ρ. The IQCs on Δ are well documented in [19, Section VI]. All these IQCs can be directly applied to describe the input/output behavior of Δ_ρ.

4.2.5 Scaled Operator for Delay

In Section 2.11.5, a delay operator Δ is defined as $w^k = 0$ for $k < \tau^k$ and $w^k = v^{k-\tau^k}$ for $k \geq \tau^k$, where $\tau^k \in [0, \tau_{\max}]$. When $k \geq \tau^k$, one can use the scaling relationship to get:

$$w_\rho^k = w^k \rho^{-k} = v^{k-\tau^k} \rho^{-k} = v_\rho^{k-\tau^k} \rho^{-\tau^k} \tag{4.24}$$

When $k < \tau^k$, one trivially gets $w_\rho^k = 0$. Therefore, $w_\rho^k = 0$ for $k < \tau^k$ and $w_\rho^k = v_\rho^{k-\tau^k} \rho^{-\tau^k}$ for $k \geq \tau^k$. It is straightforward to verify that Δ_ρ is bounded and $\|\Delta_\rho\| \leq \sqrt{\tau_{\max}+1}\rho^{-\tau_{\max}}$.

Δ_ρ is the product of the original delay Δ and a multiplicative perturbation $\delta = \rho^{-\tau^k}$. The scaled system $F_u(G_\rho, \Delta_\rho)$ can be transformed into a system with block diagonal uncertainty $diag(\Delta, \delta)$. There exist standard IQCs for time delays Δ [19, 41–43] and uncertain real parameters [19]. This approach decouples Δ_ρ into two operators and constructs separate IQCs for Δ and δ.

4.3 Equivalence between ρ-Hard IQCs on Δ and Hard IQCs on Δ_ρ

This section discusses the connections between our proposed approach (Theorem 4) and the ρ-hard IQC approach in Section 2.9.

The next lemma provides a connection between time domain hard IQCs for the scaled operator Δ_ρ and time domain ρ-hard IQCs for the original operator Δ. The lemma statement involves the scaled filter $\Psi_\rho = \mathcal{S}_{\rho+}^{-1} \circ \Psi \circ \mathcal{S}_{\rho-}^{-1}$. As discussed in Section 4.1, Ψ_ρ will denote the specific LTI state-space realization $\left(\rho^{-1}A_\psi, \rho^{-1}\left[\begin{smallmatrix} B_{\psi 1} & B_{\psi 2} \end{smallmatrix}\right], C_\psi, \left[\begin{smallmatrix} D_{\psi 1} & D_{\psi 2} \end{smallmatrix}\right]\right)$. Similarly, $\mathcal{G}_\rho^{(G,\Psi)}$ denotes the specific state-space realization for $\mathcal{S}_{\rho+}^{-1} \circ \mathcal{G}^{(G,\Psi)} \circ \mathcal{S}_{\rho-}^{-1}$ based on shifting the state matrices of $\mathcal{G}^{(G,\Psi)}$. The use of $\mathcal{S}_{\rho\pm}$ here involves a slight abuse of notation because Ψ and $\mathcal{G}^{(G,\Psi)}$ have different input/output dimensions than G.

Lemma 16. *Let G be an $n_v \times n_w$ LTI system described by Equation (2.18). Δ satisfies the time domain ρ-hard IQC defined by (Ψ, M) if and only if Δ_ρ satisfies the time*

domain hard IQC defined by (Ψ_ρ, M)*. Moreover,* $\mathcal{G}_\rho^{(G,\Psi)} = \mathcal{G}^{(G_\rho,\Psi_\rho)}$*, and*

$$\rho^2 LMI_{(G_\rho,\Psi_\rho)}(P, M_\lambda) = \begin{bmatrix} \mathcal{A}^T P \mathcal{A} - \rho^2 P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \rho^2 \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \tag{4.25}$$

where $\mathcal{A}$*,* $\mathcal{B}$*,* $\mathcal{C}$*, and* $\mathcal{D}$ *are the state matrices of* $\mathcal{G}^{(G,\Psi)}$*.*

Proof. The proof follows by simply tracking the various signal definitions. The key of the proof is the following fact. Let r be the output of Ψ driven by (v, w) with zero initial condition. Set $r_\rho^k := \rho^{-k} r^k$. Then r_ρ will be the output of Ψ_ρ driven by (v_ρ, w_ρ) with zero initial condition. The details of the proof are omitted. □

Remark 5. *The frequency domain* ρ*-IQCs introduced in [29] can be connected to the frequency domain IQCs on* Δ_ρ *in a similar manner.*

Lemma 16 states that Theorem 2 and Statement 1 in Theorem 5 are equivalent. Both theorems use time domain proofs and can be extended to other linear systems which do not have frequency domain interpretations. Note the non-negativity constraint on P has been dropped in Statement 2 of Theorem 5 using the modified dissipation inequality developed in the last chapter.

Based on Lemma 16, one can also efficiently construct time domain ρ-hard IQCs for various Δ using the IQCs in Section 4.2. For example, consider the static nonlinearity Δ_g where $g \in \mathcal{F}(m, L)$ with $m \geq 0$. Based on Lemma 16 and (4.21), it is straightforward to verify that $\Delta_g \in \rho\text{-HardIQC}(\Psi, M, \rho)$ with

$$\Psi = \left[\begin{array}{c|cc} 0_p & -\rho L I_p & \rho I_p \\ \hline \rho I_p & L I_p & -I_p \\ 0_p & -m I_p & I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}, \tag{4.26}$$

which is just another state-space realization of the off-by-one ρ-hard IQC (2.63).

In the original work of [12], there are cases where the specified $\{(\Psi_j, M_j)\}_{j=1}^{N_J}$ do not depend on ρ. In this case, the state space matrices $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ in Theorem 2 do not depend on ρ, and LMI (2.41) leads to a GEVP. However, a direct application of the off-by-one ρ-hard IQC does not lead to a GEVP. To find the best (i.e. smallest) exponential rate bound for $F_u(G, \Delta)$ in this case, a bisection algorithm is required. In

next section, we show how to obtain a GEVP formulation for linear rate analysis of first-order optimization methods by combining the forward off-by-one soft IQC in Lemma 15 with the LMI condition in Statement 2 of Theorem 5.

4.4 A GEVP Formulation for Linear Rate Analysis of Deterministic First-Order Optimization Methods

Feedback representations for optimization methods are not unique. To apply Statement 2 of Theorem 5, we need to obtain a scaled feedback interconnection $F_u(G_\rho, \Delta_\rho)$ with $G_\rho \in \mathbb{RH}_\infty^{n_v \times n_w}$ and Δ_ρ described by some PN multipliers. As commented in Chapter 3, the zero operator should be included in the perturbation set. Now we present new feedback representations for several optimization methods such that Statement 2 of Theorem 5 can be applied. The gradient descent method can be rewritten as:

$$\begin{aligned} \bar{\xi}^{k+1} &= (1-\alpha m)\bar{\xi}^k - \alpha \bar{w}^k \\ \bar{v}^k &= \bar{\xi}^k \\ \bar{w}^k &= \nabla g(\bar{v}^k) - m\bar{v}^k \end{aligned} \tag{4.27}$$

This is a feedback form $F_u(\bar{G}, \bar{\Delta})$ where $\bar{G}$ is determined by $((1-\alpha m)I_p, -\alpha I_p, I_p, 0_p)$, and $\bar{\Delta}$ is defined to map $\bar{v} \in \ell_{2e}^p$ to $\bar{w} = \bar{\Delta}(\bar{v})$ as $\bar{w}^k = \nabla g(\bar{v}^k) - m\bar{v}^k$. Now, assume $g \in \mathcal{F}(m, L)$ with $m > 0$. We can shift the gradient descent iteration as

$$\begin{aligned} \bar{\xi}^{k+1} - x^* &= (1-\alpha m)(\bar{\xi}^k - x^*) - \alpha(\bar{w}^k + mx^*) \\ \bar{v}^k - x^* &= \bar{\xi}^k - x^* \\ \bar{w}^k + mx^* &= \nabla g(\bar{v}^k) - m(\bar{v}^k - x^*) \end{aligned} \tag{4.28}$$

where x^* is the unique point satisfying $\nabla g(x^*) = 0$. We set $\xi^k = \bar{\xi}^k - x^*$, $w^k = \bar{w}^k + mx^*$, and $v^k = \bar{v}^k - x^*$. Then we get the shifted feedback interconnection $F_u(G, \Delta)$:

$$\begin{aligned} \xi^{k+1} &= (1-\alpha m)\xi^k - \alpha w^k \\ v^k &= \xi^k \\ w^k &= \nabla g(v^k + x^*) - mv^k \end{aligned} \tag{4.29}$$

Eventually, we need to analyze the ρ-exponential stability of $F_u(G, \Delta)$ where G is determined by $((1-\alpha m)I_p, -\alpha I_p, I_p, 0_p)$, and Δ is defined to map $v \in \ell_{2e}^p$ to $w = \Delta(v)$ as

$w^k = \nabla g(v^k + x^*) - mv^k$. Here $\Delta(0) = 0$, and Δ is bounded. More importantly, we can pose PN multipliers on Δ as described later. Define the scaled operator Δ_ρ which maps $v_\rho \in \ell_{2e}^p$ to $w_\rho = \Delta_\rho(v_\rho)$ as $w_\rho^k = \rho^{-k}\left(\nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*)\right) - mv_\rho^k$ with $v^* = x^*$. Based on the triangle inequality, the above scaled operator Δ_ρ is also bounded.

Nesterov's accelerated method and the Heavy-ball method can be rewritten in a similar manner. With the same choice of Δ, Nesterov's accelerated method can be cast as $F_u(G, \Delta)$ with

$$G = \left[\begin{array}{cc|c} (1-\alpha m)(1+\beta)I_p & -(1-\alpha m)\beta I_p & -\alpha I_p \\ I_p & 0_p & 0_p \\ \hline (1+\beta)I_p & -\beta I_p & 0_p \end{array}\right]. \tag{4.30}$$

With the same choice of Δ, the Heavy-ball method can be cast as $F_u(G, \Delta)$ with

$$G = \left[\begin{array}{cc|c} (1+\beta-\alpha m)I_p & -\beta I_p & -\alpha I_p \\ I_p & 0_p & 0_p \\ \hline I_p & 0_p & 0_p \end{array}\right]. \tag{4.31}$$

PN IQC multipliers on the resultant Δ_ρ can be constructed as follows.

Lemma 17. *Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$, $v_\rho = \{v_\rho^0, v_\rho^1, \ldots\} \in \ell_{2e}^p$, $v^* \in \mathbb{R}^p$, and Δ_ρ maps v_ρ to $w_\rho = \Delta_\rho(v_\rho)$ as $w_\rho^k = \rho^{-k}\left(\nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*)\right) - mv_\rho^k$. Then*

1. *(Sector IQC): Δ_ρ satisfies the time domain hard and ρ-hard IQCs defined by*

$$\Psi = \begin{bmatrix} (L-m)I_p & -I_p \\ 0_p & I_p \end{bmatrix} \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{4.32}$$

   In addition, the frequency domain representation of $\Psi^\sim M \Psi$ is a PN multiplier.

2. *(Forward Off-by-One Soft IQC): Given any $0 < \rho \leq 1$, we define*

$$\left[\begin{array}{c|c} A_\psi & B_\psi \\ \hline C_\psi & D_\psi \end{array}\right] = \left[\begin{array}{cc|cc} 0_p & 0_p & (L-m)I_p & -I_p \\ 0_p & 0_p & 0_p & I_p \\ \hline I_p & 0_p & 0_p & 0_p \\ 0_p & I_p & 0_p & -I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{4.33}$$

   Then $\Delta_\rho \in$ SoftIQC(Ψ, M) with $\Psi = (\rho^{-1}A_\psi, \rho^{-1}B_\psi, C_\psi, D_\psi)$. Moreover, the frequency domain representation of $\Psi^\sim M \Psi$ is a PN multiplier.

Proof. Notice we have

$$\begin{bmatrix} v_\rho^k \\ \rho^{-k}\left(\nabla g(\rho^k v_\rho^k + v^*) - \nabla g(v^*)\right) \end{bmatrix} = \begin{bmatrix} I_p & 0_p \\ mI_p & I_p \end{bmatrix} \begin{bmatrix} v_\rho^k \\ w_\rho^k \end{bmatrix}. \tag{4.34}$$

We can combine the above relationship with Lemma 15 to show that Δ_ρ satisfies the IQCs specified by (4.32) and (4.33). Now we need to check that (4.32) and (4.33) both lead to PN multipliers. For (4.32), we have $\Psi^\sim M\Psi = \left[\begin{smallmatrix} 0 & L-m \\ L-m & -2 \end{smallmatrix}\right] \otimes I_p$, which is clearly PN. For (4.33), it is straightforward to verify that $\Psi^\sim M\Psi$ has an zero $(1,1)$-entry and a non-positive $(2,2)$-entry. Hence $\Psi^\sim M\Psi$ is also a PN multiplier in this case. This completes the proof. □

Finally, we can combine Statement 2 of Theorem 5 with the above IQCs to obtain LMI conditions for linear rate analysis of first-order methods. The resultant LMI condition can be rewritten as a GEVP for linear rate analysis of first-order methods due to (4.25) and the specific state matrices of Ψ in Lemma 17. This is formalized by the following result.

Theorem 6. *Let G be an LTI system defined by (2.18) and $\Delta : \ell_{2e}^{n_v} \to \ell_{2e}^{n_w}$ be a causal operator such that $F_u(G,\Delta)$ is well-posed. Assume $G_\rho \in \mathbb{RH}_\infty^{n_v \times n_w}$, and $0 < \rho \leq 1$. Suppose Ψ is governed by $(A_\psi, B_\psi, C_\psi, D_\psi)$, and Ψ_ρ is governed by $(\rho^{-1}A_\psi, \rho^{-1}B_\psi, C_\psi, D_\psi)$. Here $(A_\psi, B_\psi, C_\psi, D_\psi)$ are known matrices which do not depend on ρ. Let $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be the state matrices of $\mathcal{G}^{(G,\Psi)}$. Suppose the operator Δ_ρ is bounded. In addition, assume $\Delta_\rho \in$ SoftIQC(Ψ_ρ, M_λ) given $\lambda_j \geq 0$ for all j. If there exists a matrix $P = P^T$ and $\lambda_j \geq 0$ such that $\Psi_\rho^\sim M_\lambda \Psi_\rho$ is a PN multiplier and*

$$\begin{bmatrix} \mathcal{A}^T P \mathcal{A} - \rho^2 P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} < 0 \tag{4.35}$$

then $F_u(G,\Delta)$ is exponentially stable with rate ρ.

Proof. If (4.35) is feasible with $P = P^T$ and $\lambda_j \geq 0$, then the following inequality holds with this P and $\tilde{\lambda}_j = \rho^{-2}\lambda_j \geq 0$

$$\begin{bmatrix} \mathcal{A}^T P \mathcal{A} - \rho^2 P & \mathcal{A}^T P \mathcal{B} \\ \mathcal{B}^T P \mathcal{A} & \mathcal{B}^T P \mathcal{B} \end{bmatrix} + \rho^2 \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_{\tilde{\lambda}} \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} < 0 \tag{4.36}$$

Based on (4.25), we have $LMI_{(G_\rho,\Psi_\rho)}(P, M_{\tilde{\lambda}}) < 0$. Then, Statement 2 of Theorem 5 can be used to conclude that $F_u(G,\Delta)$ is exponentially stable with rate ρ. □

Notice (4.35) provides a GEVP formulation for ρ-exponential stability analysis, since the state matrices $(\mathcal{A},\mathcal{B},\mathcal{C},\mathcal{D})$ do not depend on ρ. To apply this GEVP formulation, one needs to justify two issues. First, one needs to ensure that $\Psi_\rho^\sim M_\lambda \Psi_\rho$ is a PN multiplier. Notice the conic combinations of PN multipliers will still be PN. Hence the combined multiplier $\Psi_\rho^\sim M_\lambda \Psi_\rho$ is guaranteed to be PN if we use a conic combination of the sector IQC (4.32) and the forward off-by-one soft IQC (4.33) to formulate the GEVP (4.35). Second, it is also necessary to check the condition $G_\rho \in \mathbb{RH}_\infty^{n_v \times n_w}$. This can be easily done using well-known results from linear system theory. Recall that the state-space realization of G is (A,B,C,D), and consequently the state-space realization of G_ρ is $(\rho^{-1}A, \rho^{-1}B, C, D)$. If the eigenvalues of A have simple analytical expressions, one can use these expressions to check whether $\rho^{-1}A$ is Schur stable. Otherwise, one can use an LMI approach. If there exists a matrix $P > 0$ such that

$$A^T P A - \rho^2 P < 0, \tag{4.37}$$

then $\rho^{-1}A$ is Schur stable and $G_\rho \in \mathbb{RH}_\infty^{n_v \times n_w}$. We will illustrate the utility of the GEVP (4.35) in next section.

Remark 6. *When ρ is given,* (4.37) *is an LMI and can be solved in a numerically efficient manner. However, a typical task is finding the smallest ρ such that* (4.35) *and* (4.37) *are both feasible (with potentially different choices of P). In this case,* (4.35) *and* (4.37) *both become GEVPs. One can get two rate bounds by solving these two GEVPs separately, and then take the larger one from the two resultant rate bounds as the final convergence rate of the feedback interconnection.*

4.5 Numerical Example: Analysis of Nesterov's Accelerated Method

We will demonstrate the application of (4.35) and Lemma 17 by a case study on Nesterov's accelerated method. Suppose $g \in \mathcal{F}(m,L)$ with $m > 0$. Theorem 2.2.3 in [11] states that in this case one can apply Nesterov's accelerated method (2.11) with $\alpha = \frac{1}{L}$

and $\beta = \frac{\sqrt{L}-\sqrt{m}}{\sqrt{L}+\sqrt{m}}$ to guarantee a linear convergence rate $\rho = \sqrt{1-\sqrt{\frac{m}{L}}}$. A better numerical rate bound has been obtained in [12, Section 4.5] by applying Theorem 2 with a conic combination[1] of the sector IQC and the ρ-hard off-by-one IQC. Currently there is no analytical expression for the numerical rate bounds in [12, Section 4.5]. Notice this formulation does not lead to a GEVP, and a bisection on ρ is required to make (2.41) an LMI. Lemma 5 was used to reduce the dimension of the testing LMI (2.41). The details of this ρ-hard IQC approach are now presented. Specifically, set $\tilde{M}_1 = \tilde{M}_2 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. We further specify the following matrices

$$\tilde{A} = \begin{bmatrix} 1+\beta & -\beta \\ 1 & 0 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} -\alpha \\ 0 \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} 1+\beta & -\beta \end{bmatrix}, \quad \tilde{D} = 0,$$

$$\tilde{A}_\psi = 0, \quad \tilde{B}_{\psi 1} = -L, \quad \tilde{B}_{\psi 2} = 1, \quad \tilde{C}_\psi = \begin{bmatrix} 0 \\ 0 \\ \rho^2 \\ 0 \end{bmatrix}, \quad \tilde{D}_{\psi 1} = \begin{bmatrix} L \\ -m \\ L \\ -m \end{bmatrix}, \quad \tilde{D}_{\psi 2} = \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}. \tag{4.38}$$

Let $(\tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}, \tilde{\mathcal{D}})$ be calculated from (2.43) and (2.44). Set $\tilde{M}_\lambda = diag(\lambda_1 \tilde{M}_1, \lambda_2 \tilde{M}_2)$. If $\exists \tilde{P} > 0$, $\lambda_1 \geq 0$, and $\lambda_2 \geq 0$ such that

$$\begin{bmatrix} \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{A}} - \rho^2 \tilde{P} & \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{B}} \\ \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{A}} & \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{B}} \end{bmatrix} + \begin{bmatrix} \tilde{\mathcal{C}}^T \\ \tilde{\mathcal{D}}^T \end{bmatrix} \tilde{M}_\lambda \begin{bmatrix} \tilde{\mathcal{C}} & \tilde{\mathcal{D}} \end{bmatrix} \leq 0, \tag{4.39}$$

then

$$\left(\begin{bmatrix} \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{A}} - \rho^2 \tilde{P} & \tilde{\mathcal{A}}^T \tilde{P} \tilde{\mathcal{B}} \\ \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{A}} & \tilde{\mathcal{B}}^T \tilde{P} \tilde{\mathcal{B}} \end{bmatrix} + \begin{bmatrix} \tilde{\mathcal{C}}^T \\ \tilde{\mathcal{D}}^T \end{bmatrix} \tilde{M}_\lambda \begin{bmatrix} \tilde{\mathcal{C}} & \tilde{\mathcal{D}} \end{bmatrix} \right) \otimes I_p \leq 0. \tag{4.40}$$

Consequently, LMI (2.41) is feasible with $P = \tilde{P} \otimes I_p > 0$, and Nesterov's accelerated method is guaranteed to have the linear convergence rate ρ. Hence one only needs to check the feasibility of LMI (4.39) for any given ρ. Notice (4.39) is a 4×4 LMI with the above specified state matrices for any fixed ρ. In addition, $\tilde{C}_\psi$ directly depends on ρ^2, and hence (4.39) does not lead to a GEVP.

[1] This conic combination was originally termed as "weighted off-by-one IQC" in [12].

On the other hand, we can also apply the soft-IQC approach to analyze Nesterov's accelerated method. Specifically, we can use a conic combination of the forward off-by-one soft IQC and the sector IQC in Lemma 17. Since a soft IQC is involved, Theorem 6 is required to formulate the testing condition. The forward off-by-one IQC depends on ρ in a special way such that the resultant testing condition is a GEVP. Lemma 5 was also used to reduce the dimension of the testing LMI (4.35). Hence we still set $\tilde{M}_1 = \tilde{M}_2 = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. Now the state matrices are set as

$$
\begin{aligned}
&\tilde{A} = \begin{bmatrix} (1-\alpha m)(1+\beta) & -(1-\alpha m)\beta \\ 1 & 0 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} -\alpha \\ 0 \end{bmatrix}, \\
&\tilde{C} = \begin{bmatrix} 1+\beta & -\beta \end{bmatrix}, \quad \tilde{D} = 0, \\
&\tilde{A}_\psi = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{B}_{\psi 1} = \begin{bmatrix} L-m \\ 0 \end{bmatrix}, \quad \tilde{B}_{\psi 2} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \\
&\tilde{C}_\psi = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{D}_{\psi 1} = \begin{bmatrix} L-m \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \tilde{D}_{\psi 2} = \begin{bmatrix} -1 \\ 1 \\ 0 \\ -1 \end{bmatrix}.
\end{aligned}
\tag{4.41}
$$

Let $(\tilde{\mathcal{A}}, \tilde{\mathcal{B}}, \tilde{\mathcal{C}}, \tilde{\mathcal{D}})$ be calculated from (2.43) and (2.44). Suppose ρ is given. If $\exists \tilde{P} = \tilde{P}^T$, $\lambda_1 \geq 0$, and $\lambda_2 \geq 0$ such that LMI (4.39) is feasible, then Theorem 6 can be used to conclude that Nesterov's accelerated method has the linear convergence rate ρ. Notice $\tilde{P}$ is no longer required to be positive definite. For any fixed ρ, the LMI (4.39) is 5×5. Notice the LMI size here is slightly larger compared with the ρ-hard IQC approach. However, the resultant testing condition potentially leads to a GEVP. To apply Theorem 6, one needs to ensure the stability of G_ρ. Since we have a simple analytical expression of $\tilde{A}$, there is no need to solve the LMI (4.37). We can write out an analytical formula for the eigenvalues of $\tilde{A}$ and show that the stability of G_ρ is guaranteed if $\rho \geq 1 - \sqrt{\frac{m}{L}}$.

The numerical rate results obtained by the above two approaches are summarized in Table 4.1. It is an important fact that Theorem 6 drops the positivity constraint on P. To highlight this, Table 4.1 also presents the rate results obtained by the soft-IQC approach but with the positivity constraint enforced. The notation "$-$" means "infeasible". For comparison, the theoretical rate bound for Nesterov's accelerated method

$\left(\rho = \sqrt{1-\sqrt{\frac{m}{L}}}\right)$ and the rate value guaranteeing the stability of G_ρ $\left(\rho = 1-\sqrt{\frac{m}{L}}\right)$ are also presented. A practical issue in numerically solving the LMIs is how to break homogeneity. There are multiple ways to address this issue. We replaced the zero matrix on the right side of (4.39) with the multiplication of an identity matrix and $(-\epsilon)$ where ϵ is a small positive number. When formulating the LMIs, we also set $m = 1$ and vary the value of L. Theoretically, the rate results should only depend on the ratio L/m and there is no need to fix the value of m. However, a very large value of L will lead to an extremely small value of α. This potentially leads to ill-conditioning issues. We fix $m = 1$ and choose $\epsilon = 10^{-5}$. The obtained results suggested that the ill-conditioning issues are avoided with such choices of parameters. One can try other values of m. Based on our numerical observations, the resultant rates only depend on L/m when there is no ill-conditioning issue. It is emphasized that the selections of m and ϵ are not general. One has to carefully address this issue when applying the LMI methods to analyze other optimization algorithms.

L/m	5	10	100	500	10^3	10^4
ρ-Hard IQC Approach	0.633	0.752	0.928	0.969	0.978	0.994
Soft-IQC Approach with $\tilde{P} = \tilde{P}^T$	0.633	0.752	0.928	0.969	0.979	0.994
Soft-IQC Approach with $\tilde{P} > 0$	0.666	0.810	-	-	-	-
$\rho = \sqrt{1-\sqrt{\frac{m}{L}}}$	0.744	0.827	0.949	0.978	0.984	0.995
$\rho = 1-\sqrt{\frac{m}{L}}$	0.553	0.684	0.900	0.956	0.969	0.990

Table 4.1: Various Numerical Rate Results for Nesterov's Accelerated Method

Table 4.1 shows that the ρ-hard IQC approach and the soft-IQC approach agree with each other. Although we use a bisection on ρ in both approaches, the soft-IQC approach provides a GEVP formulation which can be potentially solved by more efficient algorithms. Notice the standard GEVP algorithm in [59] requires the positivity constraint on $\tilde{P}$. To fully take the advantage of the soft-IQC approach, one needs to develop more general algorithms for the case where such positivity constraints are dropped.

Notice the rate bounds obtained by the soft-IQC approach are always larger than $\left(1-\sqrt{\frac{m}{L}}\right)$. This justifies the application of Theorem 6. Moreover, Table 4.1 also highlights the importance of dropping the positivity constraint on P in Theorem 6. Without dropping this constraint, the soft-IQC approach may even be infeasible.

Chapter 5

Analysis for SAG and Related Variants Using IQCs and Jump System Formulations

In many machine learning problems, the objective function is a sum of smooth convex functions, i.e. $g(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$. The minimization problem becomes:

$$\min_{x\in\mathbb{R}^p} \; g(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{5.1}$$

where $f_i \in \mathcal{F}(m_i, L_i)$ and $g \in \mathcal{F}(m, L)$. In practice, f_i can be a regularizer or the loss function evaluated at one data block. The framework of (2.8) is useful for the ℓ_2-regularized empirical risk minimization problems, e.g. ℓ_2-regularized logistic regression, ridge regression, smooth support vector machine, etc. The key feature of (5.1) is that n is typically quite large. This is related to the recent development of big data science.

The deterministic first-order methods, e.g. the gradient descent method (2.9), Nesterov's accelerated method (2.11), etc, can be applied to solve (5.1). However, this type of deterministic methods requires the computation of the full gradient and may have a high iteration cost when the size of the training set is large.

A widely-used alternative approach for solving (5.1) is the stochastic gradient (SG) method [81, 82], which uses the following iteration:

$$x^{k+1} = x^k - \alpha\nabla f_{i_k}(x^k)$$

where for each step k, the index i_k is sampled uniformly from the set $\mathcal{N} := \{1, 2, \ldots, n\}$. The SG method has an iteration cost independent of n, and hence has been widely applied for large-scale empirical risk minimization problems. However, the SG method only linearly converges to some tolerance of the optimum of (5.1) given a well-chosen constant stepsize. If a diminishing stepsize is used, the SG method will converge to the optimum but at a sublinear rate. Hence the SG method is efficient for obtaining a rough approximated solution for (5.1) but inefficient when an accurate solution is desired.

More recently, the stochastic average gradient (SAG) method [14, 15] has been proposed to combine the advantages of the deterministic full gradient descent method and the SG method. The SAG method converges linearly to the optimum point while preserving the iteration cost of the SG method. SAG uses the following iteration rule:

$$x^{k+1} = x^k - \frac{\alpha}{n} \sum_{i=1}^{n} y_i^k \tag{5.2}$$

where at each step k, a random training example i_k is drawn uniformly from the set $\mathcal{N}$ and

$$y_i^k := \begin{cases} \nabla f_i(x^k) & \text{if } i = i_k \\ y_i^{k-1} & \text{otherwise} \end{cases} . \tag{5.3}$$

The development of SAG is inspired by the incremental aggregated gradient (IAG) method [83], which draws the index i_k cyclically based on a deterministic order and also applies the iteration rule (5.2) (5.3). SAG has many variants, e.g. SAGA [16]. Although convergence rate guarantees have been proved for SAG, SAGA, and IAG with certain stepsizes, there is a need to develop a unified analysis for variants of SAG with arbitrary stepsizes and more complicated sampling strategies.

The objective of this chapter is to present a unified IQC-based framework to analyze the convergence rates of SAG and its variants with arbitrary stepsizes and possibly non-uniform sampling strategies. Our approach can be viewed as a stochastic analog of the analysis in Sections 2.9 and 2.10. The key insight here is that SAG can be viewed as a feedback interconnection of a dynamic jump system and a static nonlinearity. Notice that a jump system is described by a linear state space model whose state matrices are functions of a jump parameter sampled from a given distribution. Instead of modeling the randomness in SAG as additive noises, we capture the randomness in these methods

using the jump system model. Since Lyapunov theory for jump systems has been well established in the controls field, we can incorporate IQCs to obtain semidefinite programs for convergence rate analysis of SAG and its variants. The jump system viewpoint on SAG and its variants plays a key role in our analysis. The IQC analysis of the SG method with a constant stepsize can also be done. However, the SG method with a constant stepsize converges linearly only up to some tolerance. The IQC framework has to be modified significantly to capture this phenomenon. We will present the IQC analysis of the SG method in Chapter 6.

Our analysis and the existing theoretical rate analysis in [14–16] provide complementary benefits. The main advantage of the IQC framework is that the IQC analysis can be automated for many variants of SAG and can even handle the non-uniform sampling strategies. However, our approach relies on numerically solving semidefinite programs and is subject to numerical errors. Our approach is most useful for two cases. First, our approach is useful in providing numerical confirmations of existing theoretical rate results for SAG and its variants. Second, our approach is also useful when one wants to get some initial analysis results for a new variant of SAG, e.g. SAG with non-uniform sampling strategies or Nesterov's version of SAG.

Section 5.1 reviews the basic concepts of stochastic jump systems. In Section 5.2, we provide a jump system formulation for SAG and develop an IQC framework to analyze the convergence rates of SAG. Then we present how the proposed IQC approach automates the convergence rate analysis for variants of SAG (Section 5.3).

5.1 Background of Dynamic Jump Systems

The underlying probability space for the sampling index i_k is denoted as $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_k$ be the σ-algebra generated by $(i_1, i_2, \ldots, i_k)$. Clearly, i_k is $\mathcal{F}_k$-adapted and we obtain a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$ which the SAG iteration is defined on.

Now we briefly review some required concepts in jump system theory. When A, B, C, and D in (2.18) are functions of a random process $\{i_k : k = 1, 2, \ldots\}$, the system G is then referred to as a linear dynamic jump system with the jump parameter i_k. Specifically, a dynamic jump system is typically described by the following set of

recursive equations:

$$\begin{aligned} \xi^{k+1} &= A_{i_k}\xi^k + B_{i_k}w^k \\ v^k &= C_{i_k}\xi^k + D_{i_k}w^k. \end{aligned} \tag{5.4}$$

At each step k, the jump parameter i_k is a random variable taking value in a finite set $\mathcal{N} = \{1, \cdots, n\}$. In addition, $A_{i_k} : \mathcal{N} \to \mathbb{R}^{n_\xi \times n_\xi}$, $B_{i_k} : \mathcal{N} \to \mathbb{R}^{n_\xi \times n_w}$, $C_{i_k} : \mathcal{N} \to \mathbb{R}^{n_v \times n_\xi}$, and $D_{i_k} : \mathcal{N} \to \mathbb{R}^{n_v \times n_w}$ are functions of i_k. When $i_k = i \in \mathcal{N}$, clearly we have $A_{i_k} = A_i$, $B_{i_k} = B_i$, $C_{i_k} = C_i$, and $D_{i_k} = D_i$. When the process $\{i_k : k = 1, 2, \ldots\}$ is a Markov chain, the resultant jump system is termed as a discrete-time Markovian jump linear system (MJLS). There is a large body of literature on MJLS in the controls field [84,85]. We confine our scope to the special case where i_k is an identically and independently distributed (IID) process. When i_k is sampled from a uniform distribution, we have the following assumption:

Assumption 1. $\mathbb{P}(i_k = i \,|\, \mathcal{F}_{k-1}) = \mathbb{P}(i_k = i) = \frac{1}{n}$ *for all* $k \in \mathbb{Z}^+$, *and* $i \in \mathcal{N}$.

When i_k is generated cyclically based on a deterministic order, the model (5.4) is no longer a jump system. In this case, the system is termed as a linear periodic system.

5.2 IQC-Based Analysis for SAG

5.2.1 A Jump System Formulation for SAG

To rewrite the SAG iteration as a feedback interconnection, we first need to define the operator $\bar{\Delta}_f : \ell_{2e}^p \to \ell_{2e}^{np}$ that maps $\bar{v} \in \ell_{2e}^p$ to $\bar{w} = \bar{\Delta}_f(\bar{v})$ as

$$\bar{w}^k = \begin{bmatrix} \nabla f_1(\bar{v}^k) \\ \nabla f_2(\bar{v}^k) \\ \vdots \\ \nabla f_n(\bar{v}^k) \end{bmatrix}. \tag{5.5}$$

Next, we show how to cast SAG into a feedback interconnection $F_u(\bar{G}, \bar{\Delta}_f)$ where $\bar{G}$ is a linear jump system.

Notice a key iteration rule for SAG is (5.3). Now we use the following notation:

$$y^k = \begin{bmatrix} y_1^k \\ y_2^k \\ \vdots \\ y_n^k \end{bmatrix} \tag{5.6}$$

In addition, set $\bar{v}^k = x^k$. Then the iteration rule (5.3) can be rewritten as:

$$y^k = \left((I_n - e_{i_k} e_{i_k}^T) \otimes I_p\right) y^{k-1} + \left((e_{i_k} e_{i_k}^T) \otimes I_p\right) \bar{w}^k \tag{5.7}$$

where i_k is uniformly sampled from $\mathcal{N}$ at each step k, and $\bar{w}^k$ is given by Equation (5.5). The iteration (5.2) can be rewritten as:

$$\begin{aligned} x^{k+1} &= x^k - \frac{\alpha}{n} \sum_{i=1}^{n} y_i^k \\ &= x^k - \frac{\alpha}{n} (e^T \otimes I_p) y^k \\ &= x^k - \frac{\alpha}{n} (e^T \otimes I_p) \left((I_n - e_{i_k} e_{i_k}^T) \otimes I_p\right) y^{k-1} - \frac{\alpha}{n} (e^T \otimes I_p) \left((e_{i_k} e_{i_k}^T) \otimes I_p\right) \bar{w}^k \\ &= x^k - \frac{\alpha}{n} \left((e - e_{i_k})^T \otimes I_p\right) y^{k-1} - \frac{\alpha}{n} (e_{i_k}^T \otimes I_p) \bar{w}^k \end{aligned} \tag{5.8}$$

Since $\bar{v}^k = x^k$, we can combine (5.7) and (5.8) to obtain the following jump system model mapping from $\bar{w}$ to $\bar{v}$:

$$\begin{aligned} \begin{bmatrix} y^k \\ x^{k+1} \end{bmatrix} &= \begin{bmatrix} (I_n - e_{i_k} e_{i_k}^T) \otimes I_p & \tilde{0} \otimes I_p \\ -\frac{\alpha}{n} (e - e_{i_k})^T \otimes I_p & I_p \end{bmatrix} \begin{bmatrix} y^{k-1} \\ x^k \end{bmatrix} + \begin{bmatrix} (e_{i_k} e_{i_k}^T) \otimes I_p \\ (-\frac{\alpha}{n} e_{i_k}^T) \otimes I_p \end{bmatrix} \bar{w}^k \\ \bar{v}^k &= \begin{bmatrix} \tilde{0}^T \otimes I_p & I_p \end{bmatrix} \begin{bmatrix} y^{k-1} \\ x^k \end{bmatrix} \end{aligned} \tag{5.9}$$

The above jump system model can be denoted as $\bar{v} = \bar{G}(\bar{w})$. Notice that the state for $\bar{G}$ at the step k is $\bar{\xi}^k = \begin{bmatrix} y^{k-1} \\ x^k \end{bmatrix}$. In addition, we already have $\bar{w} = \bar{\Delta}_f(\bar{v})$. Hence we can combine $\bar{v} = \bar{G}(\bar{w})$ with $\bar{w} = \bar{\Delta}_f(v)$ to conclude that the feedback representation for SAG is $F_u(\bar{G}, \bar{\Delta}_f)$. Denote the state matrices of $\bar{G}$ as (A_{i_k}, B_{i_k}, C, D). It is obvious

that $D = \tilde{0}^T \otimes I_p$. Therefore, the feedback form $F_u(\bar{G}, \bar{\Delta}_f)$ can be cast as:

$$\begin{aligned} \bar{\xi}^{k+1} &= A_{i_k}\bar{\xi}^k + B_{i_k}\bar{w}^k \\ \bar{v}^k &= C\bar{\xi}^k \\ \bar{w}^k &= \begin{bmatrix} \nabla f_1(\bar{v}^k) \\ \nabla f_2(\bar{v}^k) \\ \vdots \\ \nabla f_n(\bar{v}^k) \end{bmatrix} \end{aligned} \tag{5.10}$$

Later we will see similar jump system formulations for variants of SAG can be generated in a highly automated manner by applying the series connection rule of state-space models.

We can shift $F_u(\bar{G}, \bar{\Delta}_f)$ to another interconnection $F_u(G, \Delta)$ with Δ being bounded. Let x^* be the unique point satisfying $\nabla g(x^*) = 0$. First, based on the facts $A_{i_k} + B_{i_k} = I_n \otimes I_p$ and $\sum_{i=1}^n \nabla f_i(x^*) = n\nabla g(x^*) = 0$, we can rewrite the feedback interconnection of SAG as

$$\begin{aligned} \bar{\xi}^{k+1} - \xi^* &= A_{i_k}(\bar{\xi}^k - \xi^*) + B_{i_k}(\bar{w}^k - w^*) \\ \bar{v}^k - v^* &= C(\bar{\xi}^k - \xi^*) \\ \bar{w}^k - w^* &= \begin{bmatrix} \nabla f_1(\bar{v}^k) - \nabla f_1(x^*) \\ \nabla f_2(\bar{v}^k) - \nabla f_2(x^*) \\ \vdots \\ \nabla f_n(\bar{v}^k) - \nabla f_n(x^*) \end{bmatrix} \end{aligned} \tag{5.11}$$

with $w^* = \begin{bmatrix} \nabla f_1(x^*)^T & \dots & \nabla f_n(x^*)^T \end{bmatrix}^T$, $v^* = x^*$, and $\xi^* = \begin{bmatrix} w^* \\ x^* \end{bmatrix}$. Now we denote $\xi^k = \bar{\xi}^k - \xi^*$, $w^k = \bar{w}^k - w^*$, and $v^k = \bar{v}^k - v^*$. In addition, let the jump system G be governed by the state-space model:

$$\begin{aligned} \xi^{k+1} &= A_{i_k}\xi^k + B_{i_k}w^k \\ v^k &= C\xi^k \end{aligned} \tag{5.12}$$

And define the operator $\Delta_f : \ell_{2e}^p \to \ell_{2e}^{np}$ that maps $v \in \ell_{2e}^p$ to $w = \Delta_f(v)$ as

$$w^k = \begin{bmatrix} \nabla f_1(v^k + x^*) - \nabla f_1(x^*) \\ \nabla f_2(v^k + x^*) - \nabla f_2(x^*) \\ \vdots \\ \nabla f_n(v^k + x^*) - \nabla f_n(x^*) \end{bmatrix} \tag{5.13}$$

Then the SAG iteration (5.11) can be rewritten as $F_u(G, \Delta_f)$

$$\begin{aligned} \xi^{k+1} &= A_{i_k}\xi^k + B_{i_k}w^k \\ v^k &= C\xi^k \\ w^k &= \begin{bmatrix} \nabla f_1(v^k + x^*) - \nabla f_1(x^*) \\ \nabla f_2(v^k + x^*) - \nabla f_2(x^*) \\ \vdots \\ \nabla f_n(v^k + x^*) - \nabla f_n(x^*) \end{bmatrix} \end{aligned} \tag{5.14}$$

We are interesting in analyzing $\|\bar{\xi}^k - \xi^*\|^2$. Equivalently, we only need to analyze $F_u(G, \Delta_f)$ and draw conclusions on $\|\xi^k\|^2$. Next, we will construct ρ-hard IQCs for Δ_f, and develop a jump system dissipation inequality to analyze the linear convergence rate of $F_u(G, \Delta_f)$.

It is worth mentioning that there is a well-posedness issue implicitly embedded with the feedback setup $F_u(G, \Delta_f)$. It is straightforward to check that $F_u(G, \Delta_f)$ is well-posed for every sample path of i_k. More specifically, the feedback interconnection $F_u(G, \Delta_f)$ adopts a unique ℓ_{2e} solution from any initial conditions in the almost sure sense. Therefore, we do not need to worry about this issue, and we will not explicitly mention this well-posedness issue from now on.

5.2.2 ρ-Hard IQCs on Δ_f

We can easily obtain ρ-hard IQCs on Δ_f by manipulating the ρ-hard IQCs introduced in Section 2.11.3. To see this, define the operator $\Delta_f^{(i)} : \ell_{2e}^p \to \ell_{2e}^p$ that maps $v \in \ell_{2e}^p$ to $w_i = \Delta_f^{(i)}(v)$ as

$$w_i^k = \nabla f_i(v^k + x^*) - \nabla f_i(x^*) \tag{5.15}$$

The condition $g \in \mathcal{F}(m, L)$ implies Δ_g satisfies the sector IQC and the off-by-one ρ-hard IQC in Lemma 7. Similarly, the condition $f_i \in \mathcal{F}(m_i, L_i)$ also poses a sector IQC and an off-by-one ρ-hard IQC for $\Delta_f^{(i)}$. Since $\Delta_g = (\frac{1}{n}e^T \otimes I_p) \circ \Delta_f$ and $\Delta_f^{(i)} = (e_i^T \otimes I_p) \circ \Delta_f$, we can easily obtain the following ρ-hard IQC result for Δ_f.

Lemma 18. *Suppose $g \in \mathcal{F}(m, L)$ with $m \geq 0$, and $f_i \in \mathcal{F}(m_i, L_i)$ with $m_i \geq 0$. Assume $0 < \rho \leq 1$ is given. Then $\Delta_f \in \rho\text{-HardIQC}(\Psi, M, \rho)$ with (Ψ, M) defined by*

1. *(Sector IQC Related to g):*

$$\Psi = \begin{bmatrix} LI_p & -\frac{1}{n}e^T \otimes I_p \\ -mI_p & \frac{1}{n}e^T \otimes I_p \end{bmatrix} \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{5.16}$$

2. *(Off-by-One ρ-hard IQC Related to g):*

$$\Psi = \left[\begin{array}{c|cc} 0_p & -LI_p & \frac{1}{n}e^T \otimes I_p \\ \hline \rho^2 I_p & LI_p & -\frac{1}{n}e^T \otimes I_p \\ 0_p & -mI_p & \frac{1}{n}e^T \otimes I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{5.17}$$

3. *(Sector IQC Related to f_i):*

$$\Psi = \begin{bmatrix} L_i I_p & -e_i^T \otimes I_p \\ -m_i I_p & e_i^T \otimes I_p \end{bmatrix} \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{5.18}$$

4. *(Off-by-One ρ-hard IQC Related to f_i):*

$$\Psi = \left[\begin{array}{c|cc} 0_p & -L_i I_p & e_i^T \otimes I_p \\ \hline \rho^2 I_p & L_i I_p & -e_i^T \otimes I_p \\ 0_p & -m_i I_p & e_i^T \otimes I_p \end{array}\right] \quad , \quad M = \begin{bmatrix} 0_p & I_p \\ I_p & 0_p \end{bmatrix}. \tag{5.19}$$

Proof. Since $\Delta_g = (\frac{1}{n}e^T \otimes I_p) \circ \Delta_f$ and $\Delta_f^{(i)} = (e_i^T \otimes I_p) \circ \Delta_f$, we can directly rewrite the IQCs in Lemma 7 to prove the above lemma. The details are omitted. □

We can see that ρ-hard IQCs can be flexibly constructed under various assumptions on g and f_i. The existing analysis for SAG in [15] is analytical and assumes $f_i \in \mathcal{F}(0, L)$. It is very difficult to extend the analytical approach in [15] to the cases where we have different conditions on each f_i, i.e. $m_i \neq 0$. On the other hand, the IQC analysis can be easily automated when the conditions on f_i vary with i.

5.2.3 Convergence Rate Analysis of SAG Using Semidefinite Programs

Suppose $\Delta_f \in \rho\text{-HardIQC}(\Psi_j, M_j, \rho)$ for $j = 1, \ldots, N_J$. All $\{\Psi_j\}_{j=1}^{N_J}$ are aggregated into a filter Ψ governed by Equation (2.33). Similar to the deterministic case, the IQC analysis of $F_u(G, \Delta_f)$ is based on the extended system $\mathcal{G}^{(G,\Psi)}$ shown in Figure 5.1.

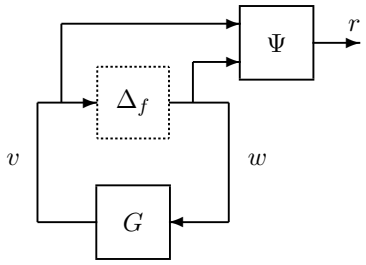

Figure 5.1: Removing Δ by Enforcing a Constraint on the Output of Ψ

From (2.24), we can see that the state space realization of $\mathcal{G}^{(G,\Psi)}$ has the form $(\mathcal{A}_{i_k}, \mathcal{B}_{i_k}, \mathcal{C}, \mathcal{D})$ where $\mathcal{C}$ and $\mathcal{D}$ do not depend on the sampling index i_k. Hence $\mathcal{G}^{(G,\Psi)}$ has the following state-space model:

$$\begin{bmatrix} \eta^{k+1} \\ r^k \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{i_k} & \mathcal{B}_{i_k} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} \begin{bmatrix} \eta^k \\ w^k \end{bmatrix} \tag{5.20}$$

The extended state vector is $\eta^k := \begin{bmatrix} \xi^k \\ \psi^k \end{bmatrix} \in \mathbb{R}^{n_\xi + n_\psi}$. Based on (2.25) and (2.26), the state matrices for the extended system $\mathcal{G}^{(G,\Psi)}$ can be computed from the state matrices of G and Ψ.

Still define $M_\lambda = diag(\lambda_1 M_1, \ldots, \lambda_{N_J} M_{N_J})$. The next theorem presents an LMI condition for linear rate analysis of SAG using time domain ρ-hard IQCs and a jump system dissipation inequality.

Theorem 7. *Let G be a jump system defined by (5.12) and $\Delta \in \rho\text{-HardIQC}(\Psi_j, M_j, \rho)$ for $j = 1, \ldots, N_J$. Suppose $F_u(G, \Delta_f)$ is well-posed almost surely, and ξ^k is generated by the interconnection $F_u(G, \Delta)$ with the initial condition ξ^0. Assume i_k is sampled in an IID manner and $\mathbb{P}(i_k = i) = p_i$. If $\exists$ a matrix $P = P^T > 0$ and scalars $\lambda_j \geq 0$ s.t.*

$$\begin{bmatrix} \sum_{i=1}^n p_i \mathcal{A}_i^T P \mathcal{A}_i - \rho^2 P & \sum_{i=1}^n p_i \mathcal{A}_i^T P \mathcal{B}_i \\ \sum_{i=1}^n p_i \mathcal{B}_i^T P \mathcal{A}_i & \sum_{i=1}^n p_i \mathcal{B}_i^T P \mathcal{B}_i \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \leq 0 \tag{5.21}$$

Then $\mathbb{E}\left[\|\xi^k\|^2\right] \le \rho^{2k}\left(\mathrm{cond}(P)\|\xi^0\|^2\right)$ *holds for all* $k \ge 1$ *and* $\xi^0 \in \mathbb{R}^{n_\xi}$.

Proof. Let (ξ, v, w) be generated by $F_u(G, \Delta)$ with the initial condition ξ^0. Let r be generated by (2.22) with initial condition $\psi^0 = 0$. Let $\eta^k = \begin{bmatrix}(\xi^k)^T & (\psi^k)^T\end{bmatrix}^T$. We can verify that (5.20) holds with the current choice of (η, w, r) with the initial condition $\eta^0 = \begin{bmatrix}(\xi^0)^T & 0\end{bmatrix}^T$. Moreover, the solution (η, w, r) is in ℓ_{2e} almost surely.

Define the storage function by $V(\eta^k) = (\eta^k)^T P \eta^k$. Based on (5.20), we have the following key relation:

$$\begin{aligned}
&\mathbb{E}[V(\eta^{k+1}) \,|\, \mathcal{F}_{k-1}] \\
&= \mathbb{E}[(\eta^{k+1})^T P \eta^{k+1} \,|\, \mathcal{F}_{k-1}] \\
&= \sum_{i=1}^{n} \mathbb{P}(i_k = i) \left(\mathcal{A}_i \eta^k + \mathcal{B}_i w^k\right)^T P \left(\mathcal{A}_i \eta^k + \mathcal{B}_i w^k\right) \\
&= \begin{bmatrix}\eta^k \\ w^k\end{bmatrix}^T \begin{bmatrix}\sum_{i=1}^n p_i \mathcal{A}_i^T P \mathcal{A}_i & \sum_{i=1}^n p_i \mathcal{A}_i^T P \mathcal{B}_i \\ \sum_{i=1}^n p_i \mathcal{B}_i^T P \mathcal{A}_i & \sum_{i=1}^n p_i \mathcal{B}_i^T P \mathcal{B}_i\end{bmatrix} \begin{bmatrix}\eta^k \\ w^k\end{bmatrix}
\end{aligned} \tag{5.22}$$

Therefore, left and right multiply the LMI condition (5.21) by $[\eta^T, w^T]$ and $[\eta^T, w^T]^T$ to show that V satisfies:

$$\mathbb{E}[V(\eta^{k+1}) \,|\, \mathcal{F}_{k-1}] - \rho^2 V(\eta^k) + \sum_{j=1}^{N_J} \lambda_j (r_j^k)^T M_j r_j^k \le 0 \tag{5.23}$$

We can take full expectation to get

$$\mathbb{E}V(\eta^{k+1}) - \rho^2 \mathbb{E}V(\eta^k) + \mathbb{E}\left[\sum_{j=1}^{N_J} \lambda_j (r_j^k)^T M_j r_j^k\right] \le 0 \tag{5.24}$$

Multiply the above inequality by ρ^{-2k} and sum over k to yield:

$$\rho^{2-2k}\mathbb{E}V(\eta^k) - \rho^2 V(\eta^0) + \sum_{j=1}^{N_J} \mathbb{E}\left[\sum_{t=0}^{k-1} \rho^{-2t} \lambda_j (r_j^t)^T M_j r_j^t\right] \le 0 \tag{5.25}$$

Since $\sum_{t=0}^{k-1} \rho^{-2t} \lambda_j (r_j^t)^T M_j r_j^t \ge 0$ for every sample path of i_k, we have the expected quadratic constraint $\mathbb{E}\left[\sum_{t=0}^{k-1} \rho^{-2t} \lambda_j (r_j^t)^T M_j r_j^t\right] \ge 0$. Therefore, (5.25) implies:

$$\rho^{2-2k}\mathbb{E}V(\eta^k) - \rho^2 V(\eta^0) \le 0 \tag{5.26}$$

Hence $\mathbb{E}V(\eta^k) \leq \rho^{2k} V(\eta^0)$, and we immediately get

$$\mathbb{E}[\|\xi^k\|^2] \leq \mathbb{E}[\|\eta^k\|^2] \leq \rho^{2k} \left(\text{cond}(P)\|\eta^0\|^2\right) = \rho^{2k} \left(\text{cond}(P)\|\xi^0\|^2\right) \tag{5.27}$$

This completes the proof. □

When $y^0 = 0$, we have $\|\xi^0\|^2 = \|\bar{\xi}^0 - \xi^*\|^2 = \|x^0 - x^*\|^2 + \sum_{i=1}^n \|\nabla f_i(x^*)\|^2$. This term depends on $\sum_{i=1}^n \|\nabla f_i(x^*)\|^2$ as expected. We can see that the proof of the jump system dissipation inequality is very similar to its deterministic counterpart. For any fixed $0 < \rho \leq 1$, the testing condition (5.21) is an LMI. In principle, we can solve the semidefinite program (5.21) to certify the linear convergence rate of SAG even if the sampling strategy is not uniform, i.e. $p_i \neq \frac{1}{n}$. However, numerically checking the feasibility of (5.21) is practical only if the size of (5.21) is not too large. Notice the size of (5.21) is roughly proportional to both n and p. We can easily extend Lemma 5 to reduce (5.21) to a smaller LMI whose size does not depend on p. Similar arguments can be found in Remark 2 (Section 2.11.3) and [12, Section 4.2]. The real challenge is how to make use of (5.21) given the fact that the resultant LMI can only be numerically checked for intermediate values of n (i.e. $n \leq 400$).

5.2.4 Numerical Results on SAG

Now we use Theorem 7 to analyze the convergence rate of SAG. Suppose i_k is sampled from a uniform distribution, i.e. $p_i = \frac{1}{n}$ for all i. In addition, assume $m_i = 0$ and $L_i = L$. Hence $g \in \mathcal{F}(m, L)$ and $f_i \in \mathcal{F}(0, L)$. The most relevant existing result for this case was presented in [15, Theorem 1] and states the following fact. If one chooses $\alpha = \frac{1}{16L}$, then the SAG iteration converges at a linear rate $\rho = \sqrt{1 - \min\{\frac{m}{16L}, \frac{1}{8n}\}}$ in the mean square sense. This result demonstrates the power of the SAG method. Notice the full gradient descent method accesses the oracle n times at one iteration, since the computation of ∇g requires the computation of ∇f_i for all i. Roughly speaking, the full gradient descent iteration shrinks at a certain factor which depends on the ratio L/m after accessing the oracle n times. On the other hand, the SAG method in big data applications shrinks at a factor which is independent of L/m after accessing the oracle n times. To see this, notice n is typically quite large for big data applications. Hence SAG converges at a rate $\rho^2 = 1 - \frac{1}{8n}$. After accessing the oracle n times, the SAG iteration shrinks at a factor $(1 - \frac{1}{8n})^n \approx 0.8825$.

We can apply the LMI method to verify and strengthen the above result. In big data applications, n is typically very large. Evaluating the LMI with that large n is impractical. Based on the result in [15, Theorem 1], it is reasonable to guess the convergence rate of SAG takes the form of $\rho^2 = 1 - \frac{1}{cn}$ for some c when n is large. Hence it makes sense to study the case where n is large enough such that $(1 - \frac{1}{cn})^n$ approaches its limit and $\frac{1}{cn} \leq \frac{m}{16L}$. To determine the range of n for our LMI tests, we compute the values of $(1 - \frac{1}{cn})^n$ for different c. The result is summarized in Table 5.1. From Table 5.1, we can see $(1 - \frac{1}{cn})^n$ usually already approaches its limit even when n is 20. Hence we can select n based on the value of L/m. For example, if $L/m = 10$, we can choose n to be larger than $160/c$ where c is related to the tested rate.

n	5	20	50	100	10^3	10^4
$(1 - \frac{1}{8n})^n$	0.8811	0.8822	0.8824	0.8824	0.8825	0.8825
$(1 - \frac{1}{6n})^n$	0.8441	0.8459	0.8462	0.8464	0.8465	0.8465
$(1 - \frac{1}{2n})^n$	0.5905	0.6027	0.6050	0.6058	0.6065	0.6065
$(1 - \frac{2}{3n})^n$	0.4889	0.5076	0.5111	0.5123	0.5133	0.5134
$(1 - \frac{1}{n})^n$	0.3277	0.3585	0.3642	0.3660	0.3677	0.3679

Table 5.1: Values of $\left(1 - \frac{1}{cn}\right)^n$ for $c \in \{8, 6, 2, 1.5, 1\}$

Now we present the details of our LMI analysis. We use the sector IQCs (5.16) (5.18) and the off-by-one ρ-hard IQC in (5.17) to formulate the LMI condition. Since $p_i = \frac{1}{n}$, the LMI condition (5.21) becomes

$$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \mathcal{A}_i^T P \mathcal{A}_i - \rho^2 P & \frac{1}{n}\sum_{i=1}^n \mathcal{A}_i^T P \mathcal{B}_i \\ \frac{1}{n}\sum_{i=1}^n \mathcal{B}_i^T P \mathcal{A}_i & \frac{1}{n}\sum_{i=1}^n \mathcal{B}_i^T P \mathcal{B}_i \end{bmatrix} + \begin{bmatrix} \mathcal{C}^T \\ \mathcal{D}^T \end{bmatrix} M_\lambda \begin{bmatrix} \mathcal{C} & \mathcal{D} \end{bmatrix} \leq 0 \tag{5.28}$$

Since totally $(n + 2)$ ρ-hard IQCs are used, we have $N_J = n + 2$. Similar to the deterministic case, we can set $P = \tilde{P} \otimes I_p$ and reduce the above LMI to a smaller LMI whose size does not depend on p. The resultant LMI can be further simplified by exploiting the symmetry in the problem setup. There is some symmetry in A_i and B_i. In addition, there is also some symmetry between the sector IQCs on different f_i. Hence

we set λ_j for all the sector IQCs on f_i to be one parameter, and parameterize $\tilde{P}$ as

$$\tilde{P} = \begin{bmatrix} p_1 I_n + p_2(e^T e) & p_3 e & p_4 e \\ p_3 e^T & p_5 & p_6 \\ p_4 e^T & p_6 & p_7 \end{bmatrix} \tag{5.29}$$

where p_j $(j = 1, \ldots, 7)$ are scalar decision variables. Without this parameterization, $\tilde{P}$ is an $(n+2) \times (n+2)$ matrix. With this parameterization, the number of the decision variables is significantly reduced. The number of the decision variables of the resultant LMI becomes 10, and the LMI can be solved when n is several hundred. The decision variable reductions here are quite intuitive. Although we do not have a proof to show that this step does not introduce any further conservatism, we did some numerical tests and the results seemed to confirm the applicability of such variable reduction.

Another practical issue is how to break homogeneity in the LMI. Instead of enforcing $\tilde{P} > 0$, we actually use the condition $\tilde{P} \geq \epsilon I$ with $\epsilon = 10^{-3}$. The choice of ϵ can affect the result especially when n is large. Some infeasible points may become feasible if smaller ϵ is used. Larger ϵ leads to slightly more conservative ρ^2.

Now we check the feasibility of the resultant LMI with various choices of ρ^2. The size of the LMI is large such that a bisection on ρ^2 becomes inefficient. Hence we test the feasibility of the LMI with several prescribed ρ^2. The result for the case where $m = 1$ and $L = 10$ is summarized in Table 5.2. The notation "Y" means the LMI is feasible while "-" means the LMI is not feasible.

n	20	50	100	200	300
$\rho^2 = 1 - \frac{1}{8n}$	Y	Y	Y	Y	Y
$\rho^2 = 1 - \frac{1}{6n}$	Y	Y	Y	Y	Y
$\rho^2 = 1 - \frac{1}{2n}$	-	Y	Y	Y	Y
$\rho^2 = 1 - \frac{2}{3n}$	-	-	-	Y	Y
$\rho^2 = 1 - \frac{1}{n}$	-	-	-	-	-

Table 5.2: Numerical Rate Results for SAG with $m = 1$, $m_i = 0$, $L_i = L = 10$, and $\alpha = \frac{1}{16L}$

The results in Table 5.2 confirm [15, Theorem 1] since $\frac{1}{8n} \leq \frac{1}{160}$ for $n \geq 20$ and the LMI is feasible with $\rho^2 = 1 - \frac{1}{8n}$ for all tested n. However, Table 5.2 also suggests that [15, Theorem 1] can be conservative. The LMI method implies that SAG may

converge at a faster rate. In addition, Table 5.2 also implies that faster rates will be obtained when SAG is applied to problems with large n. Hence a rate bound which does not hold for small n could potentially hold for large n. An important future task is to develop less conservative rate bounds which only hold for large n. We also present the LMI results for $L/m = 100$ (Table 5.3). The same trend has been observed, i.e. better rate bounds may hold only for large n. Another interesting observation is that the LMI is not feasible for the rate $\rho^2 = 1 - \frac{2}{3n}$ when $L/m = 100$. There are three possibilities here. One possibility is that n is not large enough to make the LMI feasible with this rate. The second possibility is that the rate bound of the SAG may still get worse for larger values of L/m even for big data applications. The third possibility is that this is a numerical error. One evidence for this is that the LMI with $\rho^2 = 1 - \frac{2}{3n}$ becomes feasible for $n = 300$ and $L/m = 100$ if we choose $\epsilon = 10^{-5}$ instead of 10^{-3}. It is emphasized that the LMI method is subject to numerical errors. The numerical errors can even be more significant when n is large. Hence the LMI method is only used to complement the theoretical rate results at this point.

n	200	250	300	400
$\rho^2 = 1 - \frac{1}{4n}$	Y	Y	Y	Y
$\rho^2 = 1 - \frac{1}{2n}$	-	-	Y	Y
$\rho^2 = 1 - \frac{2}{3n}$	-	-	-	-

Table 5.3: Numerical Rate Results for SAG with $m = 1$, $m_i = 0$, $L_i = L = 100$, and $\alpha = \frac{1}{16L}$

5.3 Generalizations for Variants of SAG

5.3.1 Variants of SAG

SAG has many possible variants. An important variant for the SAG method is the Nesterov's version of stochastic average gradient method [15]. This variant applies the

following iteration:

$$\begin{aligned} x^{k+1} &= r^k - \frac{\alpha}{n}\sum_{i=1}^{n} y_i^k \\ r^k &= (1+\beta)x^k - \beta x^{k-1} \end{aligned} \tag{5.30}$$

where at each iteration k, the index i_k is drawn uniformly from the set $\mathcal{N}$ and

$$y_i^k := \begin{cases} \nabla f_i(r^k) & \text{if } i = i_k \\ y_i^{k-1} & \text{otherwise} \end{cases} \tag{5.31}$$

Theoretically, it is not clear whether the above method has better worst-case guarantees than SAG or not.

Another important variant for the SAG method is the SAGA method [16], which uses the following iteration:

$$x^{k+1} = x^k - \alpha\left(\nabla f_{i_k}(x^k) - y_{i_k}^{k-1} + \frac{1}{n}\sum_{i=1}^{n} y_i^{k-1}\right) \tag{5.32}$$

where at each iteration k, a random training example i_k is drawn uniformly from the set $\mathcal{N}$ and y_i^k is updated by (5.3).

Potentially, one can also obtain a Nesterov's version and a Heavy-ball version for the SAGA method.

5.3.2 Jump System Formulations for Variants of SAG

In this section, we summarize a general jump system formulation for variants of SAG. We represent variants of SAG into the feedback form $F_u(G, \Delta_f)$ as shown in Figure 5.2.

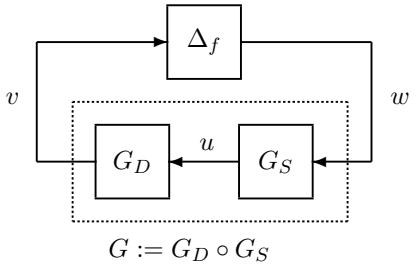

Figure 5.2: The Block-Diagram for Variants of SAG

In Figure 5.2, G is the series connection of a deterministic LTI system G_D (the subscript "D" stands for "deterministic") and a jump system G_S (the subscript "S" stands for "stochastic") which involves the iteration of y^k. Choose $G_D = (I_p, -\alpha I_p, I_p, 0_p)$ and

$$G_S = \left[\begin{array}{c|c} (I_n - e_{i_k} e_{i_k}^T) \otimes I_p & (e_{i_k} e_{i_k}^T) \otimes I_p \\ \hline \frac{1}{n}(e^T \otimes I_p)\left((I_n - e_{i_k} e_{i_k}^T) \otimes I_p\right) & \frac{1}{n}(e^T \otimes I_p)\left((e_{i_k} e_{i_k}^T) \otimes I_p\right) \end{array}\right], \tag{5.33}$$

Then we recover the SAG iteration as $F_u(G_D \circ G_S, \Delta_f)$.

Choose $G_D = (I_p, -\alpha I_p, I_p, 0_p)$ and

$$G_S = \left[\begin{array}{c|c} (I_n - e_{i_k} e_{i_k}^T) \otimes I_p & (e_{i_k} e_{i_k}^T) \otimes I_p \\ \hline \left((\frac{1}{n} e^T - e_{i_k}^T) \otimes I_p\right) & e_{i_k}^T \otimes I_p \end{array}\right]. \tag{5.34}$$

Then we recover the SAGA iteration as $F_u(G_D \circ G_S, \Delta_f)$.

We can choose G_D as (2.49) and G_S as (5.33) to recover the Nesterov's accelerated variant of SAG. Similarly, we can choose G_D as (4.31) and G_S as (5.33) to recover the Heavy-ball variant of SAG. The Nesterov's variant and the Heavy-ball variant of SAGA can be obtained using the same technique.

Now we highlight the key feature of $F_u(G_D \circ G_S, \Delta_f)$. Since the D matrix for G_D is a zero matrix, we can directly use the series connection rule (2.20) to show that the D matrix for $G_D \circ G_S$ is also a zero matrix, and the C matrix for $G_D \circ G_S$ does not depend on the sampling index i_k. Without loss of generality, the feedback form $F_u(G, \Delta_f)$ can always be cast as (5.14). Hence Theorem 7 provides a general tool for analysis of variants of SAG.

5.3.3 Numerical Results on SAGA

Now we use Theorem 7 to analyze the convergence rate of SAGA. Suppose i_k is sampled from a uniform distribution, i.e. $p_i = \frac{1}{n}$ for all i. In addition, assume $m_i = 0$ and $L_i = L$. Hence $g \in \mathcal{F}(m, L)$ and $f_i \in \mathcal{F}(0, L)$. The key convergence result in [16] states that SAGA with the stepsize $\alpha = \frac{1}{3L}$ achieves a rate $\rho = \sqrt{1 - \min\{\frac{m}{3L}, \frac{1}{4n}\}}$ in this case.

We formulate the testing LMI using the same technique in Section 5.2.4 and check the feasibility of the resultant LMI with various choices of ρ^2. The result for the case where $m = 1$ and $L = 10$ is summarized in Table 5.2.

n	20	50	100	200	300
$\rho^2 = 1 - \frac{1}{4n}$	Y	Y	Y	Y	Y
$\rho^2 = 1 - \frac{1}{2n}$	-	-	-	Y	Y
$\rho^2 = 1 - \frac{2}{3n}$	-	-	-	-	-

Table 5.4: Numerical Rate Results for SAGA with $m = 1$, $m_i = 0$, $L_i = L = 10$, and $\alpha = \frac{1}{3L}$

Since $L/m = 10$, hence $\frac{1}{4n} \leq \frac{m}{3L}$ for $n \geq 8$. Hence $1 - \min\{\frac{m}{3L}, \frac{1}{4n}\} = 1 - \frac{1}{4n}$. The results in Table 5.2 confirm the rate result in [16] since the LMI is feasible with $\rho^2 = 1 - \frac{1}{4n}$ for all tested n. The trends in the numerical rate results for SAG and SAGA are similar. This is consistent with the fact that SAG and SAGA have similar practical performances. Although Table 5.4 gives slightly worse results compared with Table 5.2, this difference could be caused by numerical errors. Notice we replaced $\tilde{P} > 0$ with $\tilde{P} \geq 10^{-3} I$ when we did the computation. If we use $\tilde{P} \geq 10^{-5} I$, then the difference becomes much smaller. Hence the more useful information from Table 5.4 is the trend in the numerical rate results. Table 5.4 also suggests that there may exist less conservative SAGA rate bounds which only hold for large n. We believe this is the most important message from the LMI results.

5.3.4 Further Discussions

The LMI approach proposed in Theorem 7 has the following advantages:

1. The result includes a linear rate certification for y^k. If (5.21) is feasible, we not only know x^k converge to x^* at a linear rate ρ, but also can conclude that y^k converges at the same rate.

2. The full information of the randomness in the optimization method is captured by the jump system model G_S. The linear state-space structures of these jump systems are used in the analysis to reduce the conservatism.

3. IQCs on the mapping Δ_f have been well established. Δ_f itself is a deterministic element, which has been studied thoroughly in the controls literature. When we try to analyze other variants of SAG, we do not need to derive new ρ-hard IQCs.

However, despite all the above advantages, Theorem 7 has a significant drawback. It is impractical to solve the resultant LMI numerically when n is very large. One way to address this issue is to construct analytical solutions for LMI (5.21). Then we no longer need to numerically solve the semidefinite programs. However, the construction of P becomes case-dependent again, and one cannot automate this proof process for various algorithms. On the other hand, in many situations, it is sufficient to prove that x^k converges to x^* in the mean square sense. We do not need to draw conclusion on the convergence rate of y^k. If this is the case, we may absorb G_S into the troublesome element as shown in Figure 5.3 and derive a semidefinite program whose size does not depend on n.

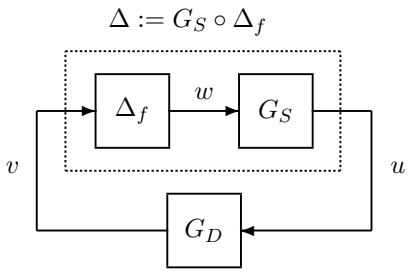

Figure 5.3: The Block-Diagram for Reducing-Order Modeling

Now we need to directly pose ρ-hard IQCs between v and u. The drawback is that this new composite troublesome element has not been extensively studied in the controls literature, and there is a need to develop new ρ-hard IQCs. Moreover, since G_S is a stochastic jump system, it is very possible that we need stochastic averaged IQCs on Δ in this case. We will investigate this approach more carefully in the future.

5.3.5 Remarks on Increment Aggregated Gradient

SAG has a deterministic counterpart, which is the so-called incremental aggregated gradient (IAG) method [83]. The iteration rule for IAG is identical to SAG except that in IAG, the index i_k is drawn cyclically based on a prescribed deterministic order. Actually, IAG was first proposed and inspired the development of SAG. All the stochastic variants of SAG will have incremental counterparts by changing the sampling rule of i_k to a known cyclic order. The existing results on IAG only give convergence guarantees for small stepsize [86]. Clearly, we can combine the periodic system theory with IQCs to

obtain LMI conditions for IAG and its variants. However, the dissipation inequality for periodic systems usually lead to n coupled LMIs. This makes the LMI approach less practical for such incremental methods. It is possible to study IAG using the idea introduced in the end of last section. We may absorb the periodic dynamics into the troublesome element and develops periodic constraints on the resultant troublesome element. This may improve the applicability of the LMI approach for incremental optimization methods.

Chapter 6

IQC Analysis of Stochastic Gradient with a Constant Stepsize

In this chapter, we revisit the analysis of the stochastic gradient method with a constant stepsize, and develop an IQC-based framework with the aim of automating the analysis for such optimization algorithms. Consider the empirical risk minimization problem

$$\min_{x\in\mathbb{R}^p} g(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

where $g : \mathbb{R}^p \to \mathbb{R}$ is the objective function. The SG method uses the following iteration rule:

$$x^{k+1} = x^k - \alpha \nabla f_{i_k}(x^k) \tag{6.1}$$

where at each k, the index i_k is uniformly sampled from $\mathcal{N} = \{1, 2, \ldots, n\}$ in an IID manner. It is well known that the convergence of the above SG method with a fixed choice of α requires very strong assumptions on the relations between different f_i [87]. In a general setup, the SG method with a constant stepsize typically achieves a linear convergence rate only up to some fixed tolerance [88, Proposition 3.4]. This well-known result is formally stated as follows.

Proposition 1. *Assume* $f_i \in \mathcal{F}(0,\infty)$ *and* $\|\nabla f_i(x^k)\| \leq c$ *for all* x^k *generated by the SG iteration* (6.1) *with any* $1 \leq i \leq n$. *In addition, assume there exists* $x^* \in \mathbb{R}^p$ *satisfying* $\nabla g(x^*) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*) = 0$ *and*

$$g(x) - g(x^*) \geq \frac{m\|x - x^*\|^2}{2} \tag{6.2}$$

for some $m > 0$ *and all* $x \in \mathbb{R}^p$. *Then the sequence* x^k *generated by the SG method* (6.1) *with* $0 < \alpha \leq \frac{1}{m}$ *satisfies*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - \alpha m)^k \|x^0 - x^*\|^2 + \frac{\alpha c^2}{m}. \tag{6.3}$$

The original version of the above proposition and related proof were presented in [88], although slightly different notation is used there. More specifically, the cost function considered in [88] is $\sum_{i=1}^n f_i(x)$ while we consider the averaged sum $\frac{1}{n}\sum_{i=1}^n f_i(x)$. Hence μ in (38) of [88] is equivalent to $\frac{mn}{2}$ in our setup, and the statement of Proposition 1 is consistent with [88, Proposition 3.4].

Now we give some interpretations of Proposition 1. The error term $\mathbb{E}[\|x^k - x^*\|^2]$ is bounded by the sum of two terms. The first term converges to 0 at a linear rate while the second term is a constant for a given fixed stepsize. Therefore, given a constant stepsize, the SG method converges linearly only up to the tolerance level $\frac{\alpha c^2}{m}$. It is also interesting to notice the trade-off between the convergence rate and the computation accuracy. The tolerance level $\frac{\alpha c^2}{m}$ is proportional to the stepsize α. Thus the optimization requires small α to obtain high accuracy, but this comes at the expense of a slow convergence rate. On the other hand, selecting α to be its maximum $\frac{1}{m}$ causes the convergence rate to be zero. But the tolerance level is $\frac{c^2}{m^2}$. This means the SG iteration with such a stepsize is not leading to any benefits in the case $g \in \mathcal{F}(m,\infty)$ since the strong convexity condition on g directly guarantees $\|x - x^*\| \leq \frac{\|\nabla g(x)\|^2}{m^2} \leq \frac{c^2}{m^2}$ for all $x \in \mathbb{R}^p$.

The above proposition highlights the trade-off between the convergence rate and the computation accuracy. For machine learning problems, there is typically no need to optimize below the so-called estimation error and obtain an extremely accurate solution [89]. Hence, the SG method with a constant stepsize has been used quite frequently in machine learning [13]. Notice SAG and SG have similar iteration costs. The accuracy of SAG is much higher than the SG method while the memory size required by SAG

is much larger than the SG method. It is possible to balance this trade-off between computation accuracy and the required memory size by designing new optimization methods which lie in the middle of the SG method and the SAG method. There is a need to develop a unified framework to analyze these generalizations of the SG method. The goal of this chapter is to develop an IQC approach which automates the analysis of the SG method and its variants.

In this chapter, i_k is always assumed to be sampled from a uniform distribution in an IID manner. The underlying probability space for the sampling index i_k is still denoted as $(\Omega, \mathcal{F}, \mathbb{P})$. And $\mathcal{F}_k$ denotes the σ-algebra generated by $(i_1, i_2, \ldots, i_k)$. Clearly, i_k is $\mathcal{F}_k$-adapted and we obtain a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}, \mathbb{P})$ which the SG iteration is defined on. Given x^0, the SG iteration (6.1) defines an $\mathcal{F}_n$-predictable process x^k whose sample path is almost surely in ℓ_{2e}.

6.1 An IQC-Based Proof for Proposition 1

6.1.1 Stochastic Quadratic Constraints

We rewrite the SG method (6.1) as the following feedback interconnection:

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \alpha w^k \\ w^k &= \nabla f_{i_k}(x^k) \end{aligned} \tag{6.4}$$

where x^* is a point satisfying $\nabla g(x^*) = 0$. We want to analyze the term $\|x^k - x^*\|^2$ for increasing k. It is crucial to construct quadratic constraints between $(x^k - x^*)$ and w^k. The assumption $f_i \in \mathcal{F}(0, \infty)$ provides known constraints between $(x^k - x^*)$ and $(w^k - \nabla f_{i_k}(x^*))$ (Section 2.2). Moreover, the assumption on g and the fact $\nabla g(x^*) = 0$ provide known constraints between $(x^k - x^*)$ and $\nabla g(x^k)$. We can use the statistical properties of i_k to manipulate these constraints into the required quadratic constraints between $(x^k - x^*)$ and w^k. This is formally stated in the next lemma. These quadratic constraints hold in the mean sense, and should be sufficiently useful when we want to study $\mathbb{E}[\|x^k - x^*\|^2]$.

Lemma 19. *Assume $f_i \in \mathcal{F}(0, \infty)$ for $1 \leq i \leq n$. In addition, assume there exists $x^* \in \mathbb{R}^p$ satisfying $\nabla g(x^*) = 0$ and Condition (6.2) holds. Suppose x^k is an $\mathcal{F}_n$-*

predictable process whose sample path is almost surely in ℓ_{2e}, and $w^k = \nabla f_{i_k}(x^k)$. Then for all $k \geq 0$, we have

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} -mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq 0 \tag{6.5}$$

Proof. Since $f_i \in \mathcal{F}(0, \infty)$, the following inequality holds in the almost sure sense:

$$\begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} 0_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} \geq f_i(x^k) - f_i(x^*) \tag{6.6}$$

Notice i_k and x^k are independent. Moreover, x^k is $\mathcal{F}_{k-1}$-measurable. We have

$$\begin{aligned}
&\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 0_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \\
=&\sum_{i=1}^n \mathbb{P}(i_k = i) \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} 0_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} \\
=&\frac{1}{n}\sum_{i=1}^n \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} 0_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} \\
\geq&\frac{1}{n}\sum_{i=1}^n (f_i(x^k) - f_i(x^*)) = g(x^k) - g(x^*)
\end{aligned} \tag{6.7}$$

This can be combined with Condition (6.2) to conclude:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 0_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \geq \frac{m}{2}\|x^k - x^*\|^2 \tag{6.8}$$

Notice

$$\begin{aligned}
\frac{m}{2}\|x^k - x^*\|^2 &= \mathbb{E}\left[\frac{m}{2}\|x^k - x^*\|^2 \,|\mathcal{F}_{k-1}\right] \\
&= \mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} \frac{m}{2}I_p & 0_p \\ 0_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right]
\end{aligned}$$

Therefore, (6.8) can be rewritten as

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} -\frac{m}{2}I_p & \frac{1}{2}I_p \\ \frac{1}{2}I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \geq 0 \tag{6.9}$$

Taking the full expectation of the above inequality leads to the desired conclusion (6.5). □

Another useful quadratic constraint is posed by the condition $\|\nabla f_i(x^k)\| \leq c$. This condition leads to the fact that $-\|\nabla f_{i_k}(x^k)\|^2 \geq -c^2$, and hence we have the following quadratic constraint:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 0_p & 0_p \\ 0_p & -I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -c^2 \tag{6.10}$$

There is a hidden energy term ($-c^2$) in this constraint. Later we will see this hidden energy can be viewed as a disturbance added into the dissipation inequality. The SG method with a constant stepsize does not converge due to this disturbance term.

6.1.2 Recovery of Proposition 1

We will first state an IQC-based lemma and then use this lemma to recover the proof of Proposition 1. In an IQC-based analysis, we always try to obtain some testing conditions in the form of matrix inequalities. For analysis of the SG method, we derive the following testing condition.

Lemma 20. *Let x^k be generated by the SG method (6.1). Assume $f_i \in \mathcal{F}(0,\infty)$ and $\|\nabla f_i(x^k)\| \leq c$ for all $1 \leq i \leq n$. In addition, assume there exists $x^* \in \mathbb{R}^p$ satisfying $\nabla g(x^*) = 0$ and Condition (6.2) holds. If there exists $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $0 < \rho^2 \leq 1$ such that*

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \leq 0 \tag{6.11}$$

Then

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \rho^{2k}\|x^0 - x^*\|^2 + \frac{\lambda_2 c^2}{1-\rho^2} \tag{6.12}$$

Proof. It (6.11) holds, then we can use the property of the Kronecker product [12, Section 4.2] to obtain the following inequality:

$$\left(\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}\right) \otimes I_p \leq 0 \tag{6.13}$$

which is equivalent to

$$\begin{bmatrix}(1-\rho^2)I_p & -\alpha I_p \\ -\alpha I_p & \alpha^2 I_p\end{bmatrix} + \lambda_1 \begin{bmatrix}-mI_p & I_p \\ I_p & 0_p\end{bmatrix} + \lambda_2 \begin{bmatrix}0_p & 0_p \\ 0_p & -I_p\end{bmatrix} \leq 0 \tag{6.14}$$

We also have

$$\|x^{k+1}-x^*\|^2 = \|x^k - x^* - \alpha w^k\|^2 = \begin{bmatrix}x^k - x^* \\ w^k\end{bmatrix}^T \begin{bmatrix} I_p & -\alpha I_p \\ -\alpha I_p & \alpha^2 I_p\end{bmatrix} \begin{bmatrix}x^k - x^* \\ w^k\end{bmatrix} \tag{6.15}$$

Therefore, left and right multiply (6.14) by $[(x^k - x^*)^T, (w^k)^T]$ and $[(x^k - x^*)^T, (w^k)^T]^T$, and take full expectation to get

$$\begin{aligned}&\mathbb{E}[\|x^{k+1}-x^*\|^2] - \rho^2 \mathbb{E}[\|x^k - x^*\|^2] \\ &+ \lambda_1 \mathbb{E}\left[\begin{bmatrix}x^k - x^* \\ \nabla f_{i_k}(x^k)\end{bmatrix}^T \begin{bmatrix}-mI_p & I_p \\ I_p & 0_p\end{bmatrix} \begin{bmatrix}x^k - x^* \\ \nabla f_{i_k}(x^k)\end{bmatrix}\right] - \lambda_2 \mathbb{E}[\|w^k\|^2] \leq 0\end{aligned} \tag{6.16}$$

Applying the constraints (6.5) and (6.10) to the above inequality, we can get:

$$\mathbb{E}[\|x^{k+1}-x^*\|^2] - \rho^2 \mathbb{E}[\|x^k - x^*\|^2] \leq \lambda_2 c^2 \tag{6.17}$$

Equation (6.17) yields the following relation which completes the proof:

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \rho^{2k}\|x^0 - x^*\|^2 + \lambda_2 c^2 \sum_{k_0=0}^{k-1} \rho^{2k_0} \leq \rho^{2k}\|x^0 - x^*\|^2 + \frac{\lambda_2 c^2}{1-\rho^2}$$

□

Notice the matrix inequality (6.11) is equivalent to:

$$\begin{bmatrix}1-\rho^2-\lambda_1 m & -\alpha + \lambda_1 \\ -\alpha + \lambda_1 & \alpha^2 - \lambda_2\end{bmatrix} \leq 0 \tag{6.18}$$

Therefore, Proposition 1 follows as a Corollary to Lemma 20 with the special choices of $\lambda_2 = \alpha^2$, $\lambda_1 = \alpha$, and $\rho^2 = 1 - \alpha m$.

The SG iteration with a constant stepsize does not converge due to the hidden energy term $(-c^2)$ in the constraint (6.10). Similarly, we should expect to use similar terms in the study of other optimization methods which only converge to some tolerance level. As we will demonstrate later, the IQC approach is powerful in automating analysis for variants of the SG method.

6.2 IQC Analysis for SG with General Cost Functions

Proposition 1 requires a strong boundedness condition $\|\nabla f_i(x^k)\| \leq c$. This assumption is considered unrealistic for many machine learning problems. In this section, we will adapt the IQC analysis for a bigger class of cost functions. For example, we will consider the case where $f_i \in \mathcal{F}(0, L)$ and $g \in \mathcal{F}(m, L)$ with $m > 0$. This is the case for most ℓ_2-regularized empirical risk minimization problems. We still consider the SG iteration (6.1) with a constant α. We will develop quadratic constraints between $(x^k - x^*)$ and w^k under several relaxed assumptions on g and f_i. Then we apply these constraints to construct matrix inequalities for the analysis of the SG iterations.

6.2.1 A General Construction of Stochastic Quadratic Constraints

As commented in last section, the assumption on f_i typically provides known constraints between $(x^k - x^*)$ and $(\nabla f_i(x^k) - \nabla f_i(x^*))$, while the assumption on g provides known constraints between $(x^k - x^*)$ and $\nabla g(x^k)$. Now we show how to convert these known constraints into the required quadratic constraints between $(x^k - x^*)$ and w^k.

Lemma 21. *Suppose x^k is an $\mathcal{F}_n$- predictable process whose sample path is almost surely in ℓ_{2e}, and $w^k = \nabla f_{i_k}(x^k)$. Let the symmetric matrices $Q \in \mathbb{R}^{p\times p}$, $S \in \mathbb{R}^{p\times p}$, and $R \in \mathbb{R}^{p\times p}$ be given. Moreover, $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$.*

1. *Assume the following constraint between $(x^k - x^*)$ and $\nabla g(x^k)$ holds almost surely:*

$$\begin{bmatrix} x^k - x^* \\ \nabla g(x^k) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla g(x^k) \end{bmatrix} \geq 0 \tag{6.19}$$

   Then for all $k \geq 0$, the following constraint holds:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \right] \geq 0 \tag{6.20}$$

2. *Assume the following constraint between $(x^k - x^*)$ and $\big(\nabla f_i(x^k) - \nabla f_i(x^*)\big)$ holds almost surely:*

$$\begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) - \nabla f_i(x^*) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & -R \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) - \nabla f_i(x^*) \end{bmatrix} \geq 0 \tag{6.21}$$

where R is positive semidefinite. Then for all $k \geq 0$, we have

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 2Q & 2S \\ 2S^T & -R \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -\frac{2}{n}\sum_{i=1}^{n} (\nabla f_i(x^*))^T R \nabla f_i(x^*) \quad (6.22)$$

Proof. To prove Statement 1, first notice i_k and x^k are independent. Moreover, x^k is $\mathcal{F}_{k-1}$-measurable. We have

$$\begin{aligned} &\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \\ =&\sum_{i=1}^{n} \mathbb{P}(i_k = i) \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} \\ =&\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} \\ =&\begin{bmatrix} x^k - x^* \\ \nabla g(x^k) \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla g(x^k) \end{bmatrix} \end{aligned} \quad (6.23)$$

The last step relies on the following fact:

$$\frac{1}{n}\sum_{i=1}^{n} (x^k - x^*)^T S \nabla f_i(x^k) = (x^k - x^*)^T S \left(\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^k)\right) = (x^k - x^*)^T S \nabla g(x^k) \quad (6.24)$$

Hence the constraint (6.19) can be combined with (6.23) to show:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} Q & S \\ S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \geq 0 \quad (6.25)$$

We can complete the proof of Statement 1 by taking the full expectation of the above inequality.

To prove Statement 2, we first obtain:

$$\begin{aligned} &\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 2Q & 2S \\ 2S^T & -R \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix} \;\middle|\; \mathcal{F}_{k-1}\right] \\ =&\frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix}^T \begin{bmatrix} 2Q & 2S \\ 2S^T & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) \end{bmatrix} - \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^k)\|_R^2 \end{aligned} \quad (6.26)$$

where $\|x\|_R^2 = x^T R x$, which is the R-weighted ℓ_2 seminorm. Notice

$$\frac{1}{n}\sum_{i=1}^{n}(x^k - x^*)^T S \nabla f_i(x^*) = (x^k - x^*)^T S \left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x^*)\right) = 0 \tag{6.27}$$

Hence the first term on the right side of (6.26) is equal to the following term

$$\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) - \nabla f_i(x^*)\end{bmatrix}\begin{bmatrix} 2Q & 2S \\ 2S^T & 0_p\end{bmatrix}\begin{bmatrix} x^k - x^* \\ \nabla f_i(x^k) - \nabla f_i(x^*)\end{bmatrix} \tag{6.28}$$

By the constraint condition (6.21), we know the above term is greater than or equal to $\frac{2}{n}\sum_{i=1}^{n}\|(\nabla f_i(x^k) - \nabla f_i(x^*)\|_R^2$. Hence (6.26) leads to the following inequality:

$$\begin{aligned}&\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k\end{bmatrix}^T\begin{bmatrix} 2Q & 2S \\ 2S^T & -R\end{bmatrix}\begin{bmatrix} x^k - x^* \\ w^k\end{bmatrix}\,\middle|\,\mathcal{F}_{k-1}\right] \\ &\geq \frac{1}{n}\sum_{i=1}^{n}\left(2\|(\nabla f_i(x^k) - \nabla f_i(x^*)\|_R^2 - \|\nabla f_i(x^k)\|_R^2\right)\end{aligned} \tag{6.29}$$

By the triangle inequality, we have

$$\|\nabla f_i(x^k)\|_R \leq \|\nabla f_i(x^k) - \nabla f_i(x^*)\|_R + \|\nabla f_i(x^*)\|_R \tag{6.30}$$

Therefore, the following holds

$$\|\nabla f_i(x^k)\|_R^2 \leq 2(\|\nabla f_i(x^k) - \nabla f_i(x^*)\|_R^2 + \|\nabla f_i(x^*)\|_R^2) \tag{6.31}$$

The above inequality can be combined with (6.29) to show

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k\end{bmatrix}^T\begin{bmatrix} 2Q & 2S \\ 2S^T & -R\end{bmatrix}\begin{bmatrix} x^k - x^* \\ w^k\end{bmatrix}\,\middle|\,\mathcal{F}_{k-1}\right] \geq -\frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|_R^2 \tag{6.32}$$

Taking the full expectation leads to the desired conclusion. □

Now we can use the above result to construct corresponding stochastic quadratic constraints between $(x^k - x^*)$ and w^k under various assumptions on f_i and g. We summarize several useful constraints between $(x^k - x^*)$ and w^k in next lemma.

Lemma 22. *Assume $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$. Suppose x^k is an $\mathcal{F}_n$- predictable process whose sample path is almost surely in ℓ_{2e}, and $w^k = \nabla f_{i_k}(x^k)$.*

1. *If $g \in \mathcal{F}(m, \infty)$, then*

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} -2mI_p & I_p \\ I_p & 0_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq 0 \tag{6.33}$$

2. *If f_i has L-smooth gradients, then*

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 2L^2 I_p & 0_p \\ 0_p & -I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \tag{6.34}$$

3. *If $f_i \in \mathcal{F}(0, L)$, then*

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} 0_p & LI_p \\ LI_p & -I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \tag{6.35}$$

4. *If $f_i \in \mathcal{F}(m, L)$, then*

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T \begin{bmatrix} -2mLI_p & (L+m)I_p \\ (L+m)I_p & -I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \tag{6.36}$$

Proof. This lemma follows as a consequence of Lemma 21 and known quadratic inequalities in Chapter 2. The details are omitted. □

It is worth mentioning that Statement 2 in the above lemma works for even non-convex f_i. In addition, Statement 1 of the above lemma also holds for non-convex g with gradients in the sector $[m, \infty)$, although we stated the result assuming the strong convexity on g.

6.2.2 Analysis Results

Now we formulate an LMI condition for general analysis of the SG method.

Lemma 23. *Suppose $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$. Let x^k be generated by the SG method* (6.1) *with a constant stepsize α. Assume for each k, $(x^k - x^*)$ and w^k satisfy*

the quadratic constraints:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \geq -c_j^2 \tag{6.37}$$

where $M_j = M_j^T \in \mathbb{R}^{2\times 2}$ *and* $c_j \in \mathbb{R}$ *for* $j = 1, \ldots, N_J$. *If there exists* $\lambda_j \geq 0$ *for all* j *and* $0 < \rho^2 \leq 1$ *such that*

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j M_j \leq 0 \tag{6.38}$$

Then

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \rho^{2k}\|x^0 - x^*\|^2 + \frac{1}{1-\rho^2}\sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.39}$$

Proof. It (6.38) holds, then we can use the property of the Kronecker product [12, Section 4.2] to obtain the following inequality:

$$\left(\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j M_j\right) \otimes I_p \leq 0 \tag{6.40}$$

which is equivalent to

$$\begin{bmatrix} (1-\rho^2)I_p & -\alpha I_p \\ -\alpha I_p & \alpha^2 I_p \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j (M_j \otimes I_p) \leq 0 \tag{6.41}$$

Left and right multiply (6.41) by $[(x^k - x^*)^T, (w^k)^T]$ and $[(x^k - x^*)^T, (w^k)^T]^T$, and take full expectation to get

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] - \rho^2\mathbb{E}[\|x^k - x^*\|^2] + \sum_{j=1}^{N_J} \lambda_j \mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} x^k - x^* \\ w^k \end{bmatrix}\right] \leq 0 \tag{6.42}$$

Applying the constraint conditions (6.37) to the above inequality, we can get:

$$\mathbb{E}[\|x^{k+1} - x^*\|^2] - \rho^2\mathbb{E}[\|x^k - x^*\|^2] \leq \sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.43}$$

Equation (6.43) yields the following relation which completes the proof:

$$\begin{aligned}\mathbb{E}[\|x^k - x^*\|^2] &\leq \rho^{2k}\|x^0 - x^*\|^2 + \left(\sum_{k_0=0}^{k-1} \rho^{2k_0}\right)\left(\sum_{j=1}^{N_J} \lambda_j c_j^2\right) \\ &\leq \rho^{2k}\|x^0 - x^*\|^2 + \frac{\sum_{j=1}^{N_J} \lambda_j c_j^2}{1-\rho^2}\end{aligned}$$

□

Notice (6.38) is linear in λ_j and ρ^2, yielding an LMI condition when α is given. This LMI has a simple form and can be solved analytically. This leads to the following result.

Theorem 8. *Assume $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$.*

1. *If $g \in \mathcal{F}(m, \infty)$ with $m > 0$, and $\|\nabla f_i(x^k)\| \leq c$ for all x^k generated by (6.1) and all i, then the sequence x^k generated by the SG method (6.1) with $0 \leq \alpha \leq \frac{1}{2m}$ satisfies*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - 2\alpha m)^k \|x^0 - x^*\|^2 + \frac{\alpha c^2}{2m}. \tag{6.44}$$

2. *If the gradients of f_i are L-smooth, and $g \in \mathcal{F}(m, L)$ with $m > 0$, then the sequence x^k generated by the SG method (6.1) with $0 \leq \alpha \leq \frac{m}{L^2}$ satisfies*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - 2m\alpha + 2L^2\alpha^2)^k \|x^0 - x^*\|^2 + \frac{\alpha}{(m - L^2\alpha)n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 \tag{6.45}$$

3. *If $f_i \in \mathcal{F}(0, L)$ and $g \in \mathcal{F}(m, L)$ with $m > 0$, then the sequence x^k generated by the SG method (6.1) with $0 \leq \alpha \leq \frac{1}{L}$ satisfies*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - 2m\alpha + 2mL\alpha^2)^k \|x^0 - x^*\|^2 + \frac{\alpha}{m(1 - L\alpha)n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 \tag{6.46}$$

4. *If $f_i \in \mathcal{F}(m, L)$ and $g \in \mathcal{F}(m, L)$ with $m > 0$, then the sequence x^k generated by the SG method (6.1) with $0 \leq \alpha \leq \frac{1}{L+m}$ satisfies*

$$\mathbb{E}[\|x^k - x^*\|^2] \leq (1 - 2m\alpha + 2m^2\alpha^2)^k \|x^0 - x^*\|^2 + \frac{\alpha}{m(1 - m\alpha)n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 \tag{6.47}$$

Proof. The proof of the above theorem is based on a direction combination of Lemma 22 and Lemma 23. To prove Statement 1, we are using the constraints (6.33) and (6.10). Hence we set $M_1 = \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix}$, $c_1 = 0$, $M_2 = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$, and $c_2 = c$. The matrix inequality (6.38) becomes

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \leq 0 \tag{6.48}$$

We can choose $\lambda_1 = \alpha$, $\lambda_2 = \alpha^2$, and $\rho^2 = 1 - 2m\alpha$ to make the left side of the above inequality to be the zero matrix. This choice of parameters leads to the desired conclusion.

To prove Statement 2, we require the constraints (6.33) and (6.34). Hence we set $M_1 = \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix}$, $c_1 = 0$, $M_2 = \begin{bmatrix} 2L^2 & 0 \\ 0 & -1 \end{bmatrix}$, and $c_2^2 = \frac{2}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$. The matrix inequality (6.38) becomes

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2L^2 & 0 \\ 0 & -1 \end{bmatrix} \leq 0 \tag{6.49}$$

We can choose $\lambda_1 = \alpha$, $\lambda_2 = \alpha^2$, and $\rho^2 = 1 - 2m\alpha + 2L^2\alpha^2$ to make the left side of the above inequality to be the zero matrix. This choice of parameters leads to the desired conclusion.

To prove Statement 3, we require the constraints (6.33) and (6.35). Hence we set $M_1 = \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix}$, $c_1 = 0$, $M_2 = \begin{bmatrix} 0 & L \\ L & -1 \end{bmatrix}$, and $c_2^2 = \frac{2}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$. The matrix inequality (6.38) becomes

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 0 & L \\ L & -1 \end{bmatrix} \leq 0 \tag{6.50}$$

We can choose $\lambda_1 = \alpha - \alpha^2 L$, $\lambda_2 = \alpha^2$, and $\rho^2 = 1 - 2m\alpha + 2mL\alpha^2$ to make the left side of the above inequality to be the zero matrix. This choice of parameters leads to the desired conclusion.

To prove Statement 4, we require the constraints (6.33) and (6.36). Hence we set $M_1 = \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix}$, $c_1 = 0$, $M_2 = \begin{bmatrix} -2mL & L+m \\ L+m & -1 \end{bmatrix}$, and $c_2^2 = \frac{2}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$. The

matrix inequality (6.38) becomes

$$\begin{bmatrix} (1-\rho^2) & -\alpha \\ -\alpha & \alpha^2 \end{bmatrix} + \lambda_1 \begin{bmatrix} -2m & 1 \\ 1 & 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2mL & L+m \\ L+m & -1 \end{bmatrix} \leq 0 \tag{6.51}$$

We can choose $\lambda_1 = \alpha - \alpha^2(L+m)$, $\lambda_2 = \alpha^2$, and $\rho^2 = 1 - 2m\alpha + 2m^2\alpha^2$ to make the left side of the above inequality to be the zero matrix. Since $\lambda_1 \geq 0$, we have $\alpha \leq \frac{1}{L+m}$. This choice of parameters leads to the desired conclusion. □

Now we compare the above results to existing results. Statement 1 of the above theorem does not require f_i to be convex, and hence can be viewed as an extension of Proposition 1. We compare Statements 2 and 3 with [62, Theorem 2]. One can set $\tau_{\max} = 0$ and $\theta = 1$ in [62, Theorem 2] and show that the SG iteration with the conditions $f_i \in \mathcal{F}(0, L)$ and $g \in \mathcal{F}(m, L)$ satisfies:

$$\mathbb{E}[g(x^k)] - g(x^*) \leq (1 - 2m\alpha + 2L^2\alpha^2)^k (g(x^0) - g(x^*)) + \frac{\alpha L}{2(m - L^2\alpha)n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$$

Notice $g(x^k) - g(x^*) \leq \frac{L}{2}\|x^k - x^*\|^2$. Hence Statement 2 in Theorem 8 leads to

$$\mathbb{E}[g(x^k)] - g(x^*) \leq (1 - 2m\alpha + 2L^2\alpha^2)^k \frac{\|x^0 - x^*\|^2 L}{2} + \frac{\alpha L}{2(m - L^2\alpha)n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2$$

Hence both the convergence rate and the tolerance level in [62, Theorem 2] and Statement 2 of Theorem 8 match up. However, Statement 2 of Theorem 8 does not require f_i to be convex while the proof of [62, Theorem 2] relies on the convexity of f_i (see Section 6.1 in [62]). When f_i is further assumed to be convex, we can get Statement 3 in the above theorem, which allows a larger step size compared with Statement 2. Finally, when f_i is assumed to be strongly-convex, the result is strengthened as Statement 4 in Theorem 8, which gives better rates and smaller error terms for $\alpha \leq \frac{1}{L+m}$. We can now see the benefit of the quadratic constraint approach. Specifically, this approach clarifies the precise assumptions required to obtain a certain result. Moreover, the analysis is highly automated. Once the stochastic quadratic constraints between $(x^k - x^*)$ and w^k are derived, we can apply Lemma 23 to formulate a related matrix inequality in an automated manner. In Section 6.3, we will show how this analysis can be automated

for the case where the SG method uses a gradient computation subject to relative deterministic noise. In Section 6.4, we further show how to perform similar analysis for variants of the SG method.

Notice the testing condition (6.38) is not linear in α. Actually Shur complements can be used to rewrite (6.38) as an equivalent LMI condition which is linear in ρ^2, λ_j and α at the same time. See Equation (4.8) in [12] for the related transformation tricks.

6.3 Robustness of SG with respect to Deterministic Noise

Now we consider the following noisy SG iteration:

$$x^{k+1} = x^k - \alpha(I_p + \delta^k)\nabla f_{i_k}(x^k) \tag{6.52}$$

where i_k is uniformly sampled from $\mathcal{N}$, and $\delta^k = diag(\delta_1^k, \ldots, \delta_p^k)$ with $\delta_j^k \in \mathbb{R}$ represents the relative deterministic noise ratio at the j-th direction of the gradient. We assume $|\delta_j^k| \leq \delta$ where δ quantifies the noise level. When the gradient computation is accurate, we expect to have small noise level, i.e. $\delta \leq 0.01$. If the gradient computation is inaccurate, then we have large δ. Now we define $\epsilon^k := \delta^k \nabla f_{i_k}(x^k)$ and obtain the following feedback interconnection:

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \alpha w^k - \alpha \epsilon^k \\ w^k &= \nabla f_{i_k}(x^k) \\ \epsilon^k &= \delta^k w^k \end{aligned} \tag{6.53}$$

where $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$. Notice $\|\epsilon\|^2 \leq \delta^2 \|w^k\|^2$. This can be rewritten as a constraint:

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}^T \begin{bmatrix} 0_p & 0_p & 0_p \\ 0_p & \delta^2 I_p & 0_p \\ 0_p & 0_p & -I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}\right] \geq 0 \tag{6.54}$$

The above type of constraints are very useful in the analysis of noisy SG iterations. Now we are ready to derive the analysis condition for the noisy SG iteration (6.52).

Lemma 24. *Suppose $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$. Let x^k be generated by the noisy SG iteration* (6.52)*. Assume for each k, the following quadratic constraints hold:*

$$\mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}\right] \geq -c_j^2 \tag{6.55}$$

where $M_j = M_j^T \in \mathbb{R}^{3\times 3}$ and $c_j \in \mathbb{R}$ for $j = 1, \ldots, N_J$. If there exists $\lambda_j \geq 0$ for all j and $0 < \rho^2 \leq 1$ such that

$$\begin{bmatrix} 1-\rho^2 & -\alpha & -\alpha \\ -\alpha & \alpha^2 & \alpha^2 \\ -\alpha & \alpha^2 & \alpha^2 \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j M_j \leq 0 \tag{6.56}$$

then

$$\mathbb{E}[\|x^k - x^*\|^2] \leq \rho^{2k}\|x^0 - x^*\|^2 + \frac{1}{1-\rho^2}\sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.57}$$

Proof. Define the Lyapunov function $V^k = \|x^k - x^*\|^2$. Notice the following holds

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^* - \alpha w^k - \alpha\epsilon^k\|^2 \\ &= \begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}^T \begin{bmatrix} I_p & -\alpha I_p & -\alpha I_p \\ -\alpha I_p & \alpha^2 I_p & \alpha^2 I_p \\ -\alpha I_p & \alpha^2 I_p & \alpha^2 I_p \end{bmatrix} \begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix} \end{aligned}$$

Since (6.56) holds, we have

$$\left(\begin{bmatrix} 1-\rho^2 & -\alpha & -\alpha \\ -\alpha & \alpha^2 & \alpha^2 \\ -\alpha & \alpha^2 & \alpha^2 \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j M_j\right) \otimes I_p \leq 0 \tag{6.58}$$

Left and right multiply the above inequality by $[(x^k - x^*)^T, (w^k)^T, (\epsilon^k)^T]$ and $[(x^k - x^*)^T, (w^k)^T, (\epsilon^k)]^T$, and take full expectation to get

$$\mathbb{E}[V^{k+1}] - \rho^2\mathbb{E}[V^k] + \sum_{j=1}^{N_J} \lambda_j \mathbb{E}\left[\begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} x^k - x^* \\ w^k \\ \epsilon^k \end{bmatrix}\right] \leq 0 \tag{6.59}$$

Applying the constraint conditions (6.55) to the above inequality, we can get:

$$\mathbb{E}[V^{k+1}] - \rho^2 \mathbb{E}[V^k] \leq \sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.60}$$

Equation (6.60) yields the following relation:

$$\mathbb{E}[V^k] \leq \rho^{2k} V^0 + \left(\sum_{k_0=0}^{k-1} \rho^{2k_0} \right) \left(\sum_{j=1}^{N_J} \lambda_j c_j^2 \right) \leq \rho^{2k} V^0 + \frac{\sum_{j=1}^{N_J} \lambda_j c_j^2}{1-\rho^2}$$

This completes the proof. □

Notice any quadratic constraint on $(x^k - x^*)$ and w^k can be rewritten as a constraint in the form of (6.55) by lifting the M_j matrix with extra zeros. Hence Lemma 22 can be used to construct (6.55) under various assumptions. The constraint (6.54) is always required in the analysis since the noise level δ is embedded into the analysis by this constraint. It is not straightforward to derive analytical bounds using the matrix inequality (6.56). For a fixed α, the testing condition (6.56) is an LMI which can be numerically checked. Hence the above lemma provides a computable condition which quantifies the convergence properties of the noisy SG iterations.

Again, (6.56) is only linear in ρ^2 and λ_j. However, it is not linear in α. The trick in [12, Section 4.4] can be used to transform (6.56) into an equivalent LMI which is linear in $(\rho^2, \lambda_j, \alpha)$ at the same time.

6.4 A General Analysis Framework for Variants of SG

This section presents a general analysis condition for variants of the SG method. In principle, every optimization method with the feedback interconnection form (2.51) has a stochastic variant:

$$\begin{aligned} \bar{\xi}^{k+1} &= A\bar{\xi}^k + B\bar{w}^k \\ \bar{v}^k &= C\bar{\xi}^k + D\bar{w}^k \\ \bar{w}^k &= \nabla f_{i_k}(\bar{v}^k) \end{aligned} \tag{6.61}$$

where i_k is uniformly sampled from $\mathcal{N}$ in an IID manner. For example, the stochastic variant of Nesterov's accelerated method uses the iteration:

$$\begin{aligned} x^{k+1} &= \zeta^k - \alpha \nabla f_{i_k}(\zeta^k) \\ \zeta^k &= (1+\beta)x^k - \beta x^{k-1} \end{aligned} \tag{6.62}$$

This iteration can be cast in the form of (6.61) by setting $A = \left[\begin{smallmatrix} 1+\beta & -\beta \\ 1 & 0 \end{smallmatrix}\right] \otimes I_p$, $B = \left[\begin{smallmatrix} -\alpha \\ 0 \end{smallmatrix}\right] \otimes I_p$, $C = \left[\begin{smallmatrix} 1+\beta & -\beta \end{smallmatrix}\right] \otimes I_p$, $D = 0_p$, $\bar{\xi}^k = \left[\begin{smallmatrix} x^k \\ x^{k-1} \end{smallmatrix}\right]$, and $\bar{v} = \zeta^k$.

Again, let x^* be one point satisfying $\nabla g(x^*) = 0$. Similar to the discussion in Section 2.10, the equilibrium for $\bar{v}^k$ is set to be x^*. The equilibrium of the state $\bar{\xi}^k$ is set as $\xi^* = e_{n_\xi/p} \otimes x^*$. Hence $A\xi^* = \xi^*$ and $C\xi^* = x^*$. Then, (6.61) can be written as

$$\begin{aligned} \bar{\xi}^{k+1} - \xi^* &= A(\bar{\xi}^k - \xi^*) + B\bar{w}^k \\ \bar{v}^k - x^* &= C(\bar{\xi}^k - \xi^*) + D\bar{w}^k \\ \bar{w}^k &= \nabla f_{i_k}(\bar{v}^k) \end{aligned} \tag{6.63}$$

Lemma 22 can be used to construct quadratic constraints on $(\bar{v}^k - x^*)$ and $\bar{w}^k$. Then we can extend Lemma 23 to the following general analysis condition.

Lemma 25. *Suppose $x^* \in \mathbb{R}^p$ satisfies $\nabla g(x^*) = 0$. Suppose the stochastic optimization method is described by the feedback model (6.61). Assume for each k, $(\bar{v}^k - x^*)$ and $\bar{w}^k$ satisfy the quadratic constraints:*

$$\mathbb{E}\left[\begin{bmatrix} \bar{v}^k - x^* \\ \bar{w}^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} \bar{v}^k - x^* \\ \bar{w}^k \end{bmatrix}\right] \geq -c_j^2 \tag{6.64}$$

where $M_j = M_j^T$ and $c_j \in \mathbb{R}$ for $j = 1, \ldots, N_J$.

1. *If there exists $\lambda_j \geq 0$ for all j and $0 < \rho^2 \leq 1$ such that*

$$\begin{bmatrix} A^TPA - \rho^2 P & A^TPB \\ B^TPA & B^TPB \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j \begin{bmatrix} C^T & 0_p \\ D^T & I_p \end{bmatrix} (M_j \otimes I_p) \begin{bmatrix} C & D \\ 0_p & I_p \end{bmatrix} \leq 0 \tag{6.65}$$

Then

$$\mathbb{E}[\|\bar{\xi}^k - \xi^*\|^2] \leq \mathrm{cond}(P)\rho^{2k}\|\bar{\xi}^0 - \xi^*\|^2 + \frac{1}{c_L(1-\rho^2)} \sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.66}$$

where c_L is the smallest eigenvalue of P.

2. Suppose $D = 0$. If there exists $\lambda_j \geq 0$ for all j and $0 < \rho^2 \leq 1$ such that

$$\begin{bmatrix} A^T C^T C A - \rho^2 C^T C & A^T C^T C B \\ B^T C^T C A & B^T C^T C B \end{bmatrix} + \sum_{j=1}^{N_J} \lambda_j \begin{bmatrix} C^T & 0_p \\ D^T & I_p \end{bmatrix} (M_j \otimes I_p) \begin{bmatrix} C & D \\ 0_p & I_p \end{bmatrix} \leq 0 \tag{6.67}$$

Then

$$\mathbb{E}[\|\bar{v}^k - x^*\|^2] \leq \rho^{2k} \|\bar{v}^0 - x^*\|^2 + \frac{1}{1-\rho^2} \sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.68}$$

Proof. To prove Statement 1, define the Lyapunov function $V^k = (\bar{\xi}^k - \xi^*)^T P (\bar{\xi}^k - \xi^*)$. Left and right multiply (6.65) by $[(\bar{\xi}^k - \xi^*)^T, (\bar{w}^k)^T]$ and $[(\bar{\xi}^k - \xi^*)^T, (\bar{w}^k)^T]^T$, and take full expectation to get

$$\mathbb{E}[V^{k+1}] - \rho^2 \mathbb{E}[V^k] + \sum_{j=1}^{N_J} \lambda_j \mathbb{E}\left[\begin{bmatrix} \bar{v}^k - x^* \\ \bar{w}^k \end{bmatrix}^T (M_j \otimes I_p) \begin{bmatrix} \bar{v}^k - x^* \\ \bar{w}^k \end{bmatrix} \right] \leq 0 \tag{6.69}$$

Applying the constraint conditions (6.64) to the above inequality, we can get:

$$\mathbb{E}[V^{k+1}] - \rho^2 \mathbb{E}[V^k] \leq \sum_{j=1}^{N_J} \lambda_j c_j^2 \tag{6.70}$$

Hence, we have

$$\mathbb{E}[V^k] \leq \rho^{2k} V^0 + \left(\sum_{k_0=0}^{k-1} \rho^{2k_0} \right) \left(\sum_{j=1}^{N_J} \lambda_j c_j^2 \right) \leq \rho^{2k} V^0 + \frac{\sum_{j=1}^{N_J} \lambda_j c_j^2}{1-\rho^2}$$

Statement 1 follows as a direct consequence of the above inequality.

The proof of Statement 2 is very similar. We can define a Lyapunov function $V^k = \|\bar{v}^k - x^*\|^2 = (\bar{\xi}^k - \xi^*)^T C^T C (\bar{\xi}^k - \xi^*)$ and then construct a dissipation inequality. The details are omitted here. □

Statement 2 has the advantage that the error term does not depend on the matrix P. The dimensions of the above analysis conditions can be reduced if A, B, C, and D are Kronecker products of some small matrices and I_p. See Lemma 5, Remark 2 and [12, Section 4.2] for similar dimension reduction arguments.

Chapter 7

Conclusions and Future Work

In this dissertation, we tailor the integral quadratic constraint approach from the robust control theory to formulate numerically tractable LMI conditions for linear rate analysis of various deterministic and stochastic first-order optimization methods. Chapter 3 presents a J-spectral factorization approach for hard IQC constructions. We also apply this J-spectral factorization to prove a modified dissipation inequality that requires neither non-negative storage functions nor hard IQCs. The key result of Chapter 4 is a GEVP formulation for analysis of deterministic optimization methods (gradient descent, Nesterov's accelerated method, etc). The GEVP is derived using a new soft Zames-Falb IQC and the modified dissipation inequality. In Chapter 5, we combine IQCs with jump system theory to formulate LMI conditions for linear rate analysis of the SAG method and its variants (SAGA, etc) with uniform or non-uniform sampling strategies. Finally, we develop averaged quadratic constraints and formulate LMI testing conditions to analyze the SG method under different conditions (Chapter 6).

The key idea of this dissertation is that IQCs provide a unified framework to automate analysis of various optimization methods whose iterations rules may look quite different. We briefly comment on several future directions along the path of this dissertation.

Sublinear rate analysis and non-quadratic Lyapunov functions: This dissertation focuses on the case where the objective function is strongly-convex, and hence only considers linear rate analysis. The IQC approach can be extended to formulate

LMI conditions for sublinear rate analysis of optimization methods when the objective function is not strongly convex. This extension requires including the objective function as hidden energy in the IQCs. Using this type of constraints, the resultant Lyapunov function in the dissipation inequality is a sum of two terms. The first term is quadratic, and the second term is related to the objective function itself. A detailed treatment of the IQC-based sublinear rate analysis will be presented in future work.

Averaged Zames-Falb multipliers: In Chapter 6, we extend the sector bound constraints to stochastically averaged quadratic constraints which hold pointwise in time. It is possible to extend the Zames-Falb IQCs in a similar manner. Such extensions could be important for the analysis of the stochastic variant of Nesterov's accelerated method.

Control synthesis perspectives on stochastic optimization: SAG and SG have similar iteration costs. However, SAG requires a large memory size while SG converges linearly only up to a tolerance level. It is interesting to investigate how to design optimization methods which require less memory than SAG and are more accurate than the SG method. It is possible that the stochastic optimization design problems can be cast as control synthesis problems.

Other nonlinear control tools for machine learning applications: This dissertation adapts tools from robust control theory. However, there exist many other nonlinear control tools (adaptive control [90, 91], sliding mode control [92–94], model predictive control [95, 96], etc), which may be tailored for machine learning problems. Notice there are several types of loss functions which have been frequently used in machine learning problems. It can be beneficial to exploit the detailed information of these loss functions [97]. Nonlinear control tools may be suitable for this task.

ADMM with multiple blocks: The alternating direction method of multipliers (ADMM) [98] is an important distributed optimization algorithm. There are some initial convergence results on ADMM with multiple blocks [99–101]. The quantification of the convergence rates of ADMM with multiple blocks remains an open topic. IQCs have been successfully applied to analyze ADMM with two blocks [18]. The extension of IQC analysis for ADMM with multiple blocks is an important future task. It is

also interesting to investigate how to adapt IQC analysis for block ordinate descent methods [102, 103].

Analytic proofs: One drawback of the LMI method in this dissertation is that there is numerical error embedded in the results if the LMIs are checked numerically. It will be useful if one can provide mathematical proofs for the convergence rates by constructing analytical solutions to the semidefinite programs in this dissertation. We present such proofs in the analysis of the SG method. However, it remains an open question how to systematically construct mathematical proofs for more complicated optimization schemes.

Other stepsize rules: Only constant stepsize is considered in this dissertation. Tailoring IQCs for other stepsize (backtracking line search [104], etc) is an interesting future task.

Proximal gradient: Non-smooth regularizers have been widely used in machine learning problems. To address the non-smoothness issue, proximal gradient methods have been proposed in both deterministic [97, 105] and stochastic [106] cases. In the deterministic setup, the proximal operator has been successfully included in the IQC analysis [12]. It will be interesting to modify the IQC analysis for proximal variants of stochastic optimization methods.

Analysis for SDCA and SVRG: It will be interesting to generalize the IQC analysis for the stochastic dual coordinate ascent (SDCA) method [107] and the stochastic variance reduced gradient (SVRG) method [108]. Such generalizations may require more advanced system theory other than the jump system theory.

Non-convex objective functions: The IQC approach may be used to analyze the convergence properties of optimization methods when the objective function is non-convex. One particular interesting direction is the IQC analysis of the SG method when applied to deep learning problems [109] The constructions of IQCs for non-convex functions can be case dependent. One may need to develop a new IQC library for the gradients of various non-convex functions used in practice.

Asynchronous settings: In distributed optimization, the algorithm performance will typically be impacted by the communication delay and memory contention. In this

case, it is necessary to assess the robustness of the optimization methods with respect to the time-varying delays in the gradient update. As reviewed in Section 2.11.5, there exist many IQCs for time-varying delays in the controls literature [42, 43, 45, 46]. These IQCs can be potentially used to analyze asynchronous optimization schemes [110, 111].

Derivative Free Method: It is interesting to investigate how to apply IQCs for analysis of zero-order or derivative free methods [112].

Expected Risk Minimization: In this dissertation, we mainly focus on the empirical risk minimization problems. An important class of problems that we have not considered is the expected risk minimization. For such type of problems, each data block can only be accessed once, and the optimization objective is to minimize the expected risk under further statistical assumptions on how the empirical data is sampled. It is interesting to investigate the applicability of dynamic system theory for such problems.

Connection between Continuous-time Systems and Optimization Schemes: A continuous-time viewpoint may also be valuable for optimization research, although it is natural to study optimization schemes as discrete-time systems. Recently there is an interesting paper which studies the acceleration of optimization methods using the discretization theory of continuous-time systems [113]. It is also possible that one can tailor the existing results from the optimization field to obtain new solutions for research problems related to nonlinear continuous-time systems.

References

[1] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[2] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[3] B. Schölkopf and C. Burges. *Advances in kernel methods: support vector learning*. MIT press, 1999.

[4] A. Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[5] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[6] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[7] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[8] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] C. Teo, A. Smola, S. Vishwanathan, and Q. Le. A scalable modular convex solver for regularized risk minimization. In *Proceedings of the 13th ACM SIGKDD*

international conference on Knowledge discovery and data mining, pages 727–736, 2007.

[10] Y. Lee and O. Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational optimization and Applications*, 20(1):5–22, 2001.

[11] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2003.

[12] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[13] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. 2010.

[14] N. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. In *Advances in Neural Information Processing Systems*, 2012.

[15] M. Schmidt, N. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *ArXiv preprint*, 2013.

[16] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

[17] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1-2):451–482, 2014.

[18] R. Nishihara, L. Lessard, B. Recht, A. Packard, and M. Jordan. A general analysis of the convergence of ADMM. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 343–352, 2015.

[19] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42:819–830, 1997.

[20] J. Carrasco, M.C. Turner, and W.P. Heath. Zames-falb multipliers for absolute stability: from osheas contribution to convex searches. In *European Control Conference*, pages 1261–178, 2015.

[21] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, 1996.

[22] S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control*. John Wiley and Sons, Chichester, 2005.

[23] G. Zames and P.L. Falb. Stability conditions for systems with monotone and slope-restricted nonlinearities. *SIAM Journal of Control*, 6(1):89–108, 1968.

[24] U. Jönsson. A nonlinear Popov criterion. In *IEEE Conf. on Decision and Control*, pages 3523–3527, 1997.

[25] J. Veenman and C. Scherer. Stability analysis with integral quadratic constraints: A dissipativity based proof. In *IEEE Conf. on Decision and Control*, pages 3770–3775, 2013.

[26] P. Seiler. Stability analysis with dissipation inequalities and integral quadratic constraints. *IEEE Transactions on Automatic Control*, 60(6):1704–1709, 2015.

[27] J.C. Willems. Dissipative dynamical systems part i: General theory. *Archive for Rational Mech. and Analysis*, 45(5):321–351, 1972.

[28] J.C. Willems. Dissipative dynamical systems part ii: Linear systems with quadratic supply rates. *Archive for Rational Mech. and Analysis*, 45(5):352–393, 1972.

[29] R. Boczar, L. Lessard, and B. Recht. Exponential convergence bounds using integral quadratic constraints. In *IEEE Conf. on Decision and Control*, pages 7516–7521, 2015.

[30] M. Corless and G. Leitmann. Bounded controllers for robust exponential convergence. *Journal of Optimization Theory and Applications*, 76(1):1–12, 1993.

[31] C.W. Scherer and I.E. Köse. Robustness with dynamic IQCs: An exact state-space characterization of nominal stability with applications to robust estimation. *Automatica*, 44:1666–1675, 2008.

[32] J. Veenman and C. Scherer. IQC-synthesis with general dynamic multipliers. *International Journal of Robust and Nonlinear Control*, 24(17):3027–3056, 2014.

[33] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, 1994.

[34] L. Dai. *Singular Control Systems*, volume 118 of *Lecture Notes in Control and Information Sciences*. Springer, 1989.

[35] J. Willems. *The analysis of feedback systems*. The MIT Press, 1971.

[36] U Jönsson. Lecture notes on integral quadratic constraints. 2001.

[37] E.Kreyszig. *Introductory Functional Analysis with Application*. Wiley, 1978.

[38] N. Young. *An Introduction to Hilbert Space*. Cambridge University Press, 1988.

[39] J. Hespanha. *Linear systems theory*. Princeton university press, 2009.

[40] H.K. Khalil. *Nonlinear Systems*. Prentice Hall, third edition, 2001.

[41] K. Gu, V. Kharitonov, and J. Chen. *Stability of Time-Delay Systems*. Birkhäuser, 2002.

[42] C. Kao and A. Rantzer. Stability analysis of systems with uncertain time-varying delays. *Automatica*, 43(6):959–970, 2007.

[43] H. Pfifer and P. Seiler. Integral quadratic constraints for delayed nonlinear and parameter-varying systems. *Automatica*, 56:36 – 43, 2015.

[44] W. Heath and A. Wills. Zames-Falb multipliers for quadratic programming. In *IEEE Conf. on Decision and Control*, pages 963–968, 2005.

[45] C.Y. Kao and B. Lincoln. Simple stability criteria for systems with time-varying delays. *Automatica*, 40:1429–1434, 2004.

[46] C. Kao. On stability of discrete-time LTI systems with varying time delays. *IEEE Transactions on Automatic Control*, 57:1243–1248, 2012.

[47] B. Francis. *A Course in H_∞ Control Theory*. Springer-Verlag, 1987.

[48] S. Wang, W. Heath, and J. Carrasco. A complete and convex search for discrete-time noncausal FIR Zames-Falb multipliers. In *IEEE Conf. on Decision and Control*, pages 3918 – 3923, 2014.

[49] B.L. Jones, E.C. Kerrigan, and J.F. Morrison. A modeling and filtering framework for the semi-discretised Navier-Stokes equations. In *European Control Conference*, pages 1215–1220, 2009.

[50] A. Varga. Task I.A.1 - Selection of basic software tools for standard and generalized state-space systems and transfer matrix factorizations. Technical report, Subroutine Library in Systems and Control Theory (SLICOT), 1998.

[51] B. Kågström and P. Poromaa. Computing eigenspaces with specified eigenvalues of a regular matrix pair (A, B) and condition estimation: theory, algorithms and software. *Numerical Algorithms*, 12(2):369–407, 1996.

[52] A.V.D. Schaft and A.J.Schaft. *L_2-gain and passivity in nonlinear control*. Springer-Verlag New York, Inc., 1999.

[53] Inc. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0. `http://cvxr.com/cvx`, August 2012.

[54] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

[55] R.H Tutuncu, K.C. Toh, and M.J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming Ser. B*, 95:189–217, 2003.

[56] K.C. Toh, M.J. Todd, and R.H. Tutuncu. SDPT3 - a matlab software package for semidefinite programming. *Optimization Methods and Software*, 11:545–581, 1999.

[57] J. Veenman and C.W. Scherer. IQC-synthesis with general dynamic multipliers. *International Journal of Robust and Nonlinear Control*, 24:3027–3056, 2012.

[58] M. Cantoni, U. Jönsson, and S.Z. Khong. Robust stability analysis for feedback interconnections of time-varying linear systems. *SIAM J. of Control Optim.*, 51(1):353–379, 2013.

[59] S. Boyd and L. El Ghaoui. Method of centers for minimizing generalized eigenvalues. *Linear algebra and its applications*, 188:63–111, 1993.

[60] R. Boczar, L. Lessard, and B. Recht. Exponential convergence bounds using integral quadratic constraints. `http://arxiv.org/pdf/1503.07222v4.pdf`, 2015.

[61] J. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer Science & Business Media, 1996.

[62] H. Feyzmahdavian, A. Aytekin, and M. Johansson. A delayed proximal gradient method with linear convergence rate. In *Machine Learning for Signal Processing, 2014 IEEE International Workshop on*, pages 1–6, 2014.

[63] H. Bart, I. Gohberg, and M.A. Kaashoek. *Minimal Factorization of Matrix and Operator Functions*. Birkhäuser, 1979.

[64] V. Ionescu, C. Oară, and M. Weiss. *Generalized Riccati Theory and Robust Control: A Popov Function Approach*. Wiley, 1999.

[65] A. Helmersson. An IQC-based stability criterion for systems with slowly varying parameters. Technical Report LiTH-ISY-R-1979, Linköping University, 1997.

[66] M. Fu, S. Dasgupta, and Y.C. Soh. Integral quadratic constraint approach vs. multiplier approach. *Automatica*, 41:281–287, 2005.

[67] J. Carrasco and P. Seiler. Integral quadratic constraint theorem: A topological separation approach. In *Accepted to IEEE Conf. on Decision and Control*, 2015.

[68] T. Başar and G.J. Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 2nd edition, 1999.

[69] T. Başar and P. Bernhard. *H^∞-Optimal Control and Related Minimax Design Problems*. Birkhäuser, 2nd edition, 1995.

[70] A. Megretski. KYP lemma for non-strict inequalities and the associated minimax theorem. Arxiv, 2010.

[71] J.C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 16:621–634, 1971.

[72] J. Engwerda. *LQ Dynamic Optimization and Differential Games*. Wiley, 1st edition, 2005.

[73] J. Engwerda. Uniqueness conditions for the affine open-loop linear quadratic differential game. *Automatica*, 44:504–511, 2008.

[74] B. Molinari. The stabilizing solution of the discrete algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 20(3):396–399, 1975.

[75] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[76] H. Pfifer and P. Seiler. Robustness analysis of linear parameter varying systems using integral quadratic constraints. *International Journal of Robust and Nonlinear Control*, 2014.

[77] H. Pfifer and P. Seiler. Less conservative robustness analysis of linear parameter varying systems using integral quadratic constraints. *To appear in International Journal of Robust and Nonlinear Control*, 2016.

[78] G. Meinsma. J-spectral factorization and equalizing vectors. *Systems and Control Letters*, 25:243–249, 1995.

[79] M.C. Turner and M.L. Kerr. L_2 gain bounds for systems with sector bounded and slope-restricted nonlinearities. *International Journal of Robust and Nonlinear Control*, 22(13):1505–1521, 2012.

[80] C. Kao and M. Chen. Robust estimation with dynamic integral quadratic constraints: the discrete-time case. *IET Control Theory and Applications*, 7:15991608, 2013.

[81] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[82] L. Bottou and Y. LeCun. Large scale online learning. *Advances in neural information processing systems*, 16:217, 2004.

[83] D. Blatt, A. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

[84] O. Costa, M. Fragoso, and R. Marques. *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

[85] V. Dragan, T. Morozan, and A. Stoica. *Mathematical methods in robust control of discrete-time linear stochastic systems*. Springer, 2010.

[86] M. Gurbuzbalaban, A. Ozdaglar, and P. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *ArXiv preprint*, 2015.

[87] M. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11(1):23–35, 1998.

[88] A. Nedić and D. Bertsekas. Convergence rate of incremental subgradient algorithms. pages 223–264, 2001.

[89] Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[90] K. Åström and B. Wittenmark. *Adaptive control*. Courier Corporation, 2013.

[91] P. Ioannou and J. Sun. *Robust adaptive control*. Courier Corporation, 2012.

[92] C. Edwards and S. Spurgeon. *Sliding mode control: theory and applications*. CRC Press, 1998.

[93] K. Furuta. Sliding mode control of a discrete system. *Systems & Control Letters*, 14(2):145–152, 1990.

[94] K. Young, V. Utkin, and U. Ozguner. A control engineer's guide to sliding mode control. *IEEE transactions on control systems technology*, 7(3):328–342, 1999.

[95] E. Camacho and C. Alba. *Model predictive control*. Springer Science & Business Media, 2013.

[96] C. Garcia, D. Prett, and M. Morari. Model predictive control: theory and practice-a survey. *Automatica*, 25(3):335–348, 1989.

[97] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[98] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[99] M. Hong and Z. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

[100] C. Chen, B. He, Y. Ye, and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155(1-2):57–79, 2016.

[101] R. Sun, Z. Luo, and Y. Ye. On the expected convergence of randomly permuted ADMM. *arXiv preprint arXiv:1503.06387*, 2015.

[102] M. Razaviyayn, M. Hong, and Z. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[103] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

[105] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[106] J. Mairal. Optimization with first-order surrogate functions. In *Proceedings of The 30th International Conference on Machine Learning*, pages 783–791, 2013.

[107] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

[108] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[109] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. Le, and A. Ng. On optimization methods for deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 265–272, 2011.

[110] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[111] R. Zhang and J. Kwok. Asynchronous distributed ADMM for consensus optimization. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1701–1709, 2014.

[112] A. Conn, K. Scheinberg, and L. Vicente. *Introduction to derivative-free optimization*, volume 8. SIAM, 2009.

[113] A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. *arXiv preprint arXiv:1603.04245*, 2016.

[114] T. Katayama. (J, J')-Spectral factorization and conjugation for discrete-time descriptor system. *Circuits, Systems and Signal Processing*, 15(5):649–669, 1996.

[115] V. Ionescu, C. Oara, and M. Weiss. General matrix pencil techniques for the solution of algebraic Riccati equations: a unified approach. *IEEE Transactions on Automatic Control*, 42(8):1085–1097, 1997.

[116] Vlad Ionescu and Martin Weiss. On computing the stabilizing solution of the discrete-time Riccati equation. *Linear Algebra and its Applications*, 174:229 – 238, 1992.

Appendix A

IQC Multipliers and DARE Stabilizing Solutions

This appendix presents one key lemma stating that if Π satisfies the Strict-PN condition then there exists a stabilizing solution to a related DARE. The multiplier Π is assumed to be bounded on the unit circle but can, in general, be non-proper. Moreover, the feedthrough matrix can be singular. Hence the proof requires descriptor system notation and matrix pencil techniques to resolve both these issues. Some background on descriptor form and matrix pencil techniques can be found in [34, 114–116]. The proof in this appendix also relies on the connection between the invariant subspace of a Hamiltonian matrix and its related Riccati equation. A summary of the developments on this connection can be found in [115, Section III].

A few basic facts regarding descriptor form are provided before stating and proving the lemma. Consider a discrete-time system H in descriptor form:

$$\begin{aligned} E\xi^{k+1} &= A\xi^k + Bu^k \\ y^k &= C\xi^k + Du^k \end{aligned} \tag{A.1}$$

This system has the transfer function $H(z) := C(zE-A)^{-1}B+D$. The matrix inversion lemma can be used to show $H^{\sim}(z) = -zB^T(zA^T - E^T)^{-1}C^T + D^T$. Thus $H^{\sim}$ has the

following descriptor representation:

$$\begin{aligned} A^T\xi^{k+1} &= E^T\xi^k - C^T u^k \\ y^k &= B^T\xi^{k+1} + D^T u^k \end{aligned} \tag{A.2}$$

Next, a descriptor realization for $H^{-1}(z)$ is

$$\begin{aligned} \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} \xi_{in}^{k+1} &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \xi_{in}^k + \begin{bmatrix} 0 \\ -I \end{bmatrix} y^k \\ u^k &= \begin{bmatrix} 0 & I \end{bmatrix} \xi_{in}^k \end{aligned} \tag{A.3}$$

where $\xi_{in} := [\xi^T \;\; u^T]^T$ is the state of the inverse system H^{-1}.

These facts are used to construct the required descriptor representations for the multiplier Π and its inverse Π^{-1}. Let $\Pi = \Psi^{\sim} M \Psi$ be any factorization with Ψ statble. Define $Q := C_\psi^T M C_\psi$, $S := C_\psi^T M D_\psi$ and $R := D_\psi^T M D_\psi$ where $(A_\psi, B_\psi, C_\psi, D_\psi)$ are the state matrices of Ψ. A descriptor representation for Π is given by:

$$\begin{aligned} E_\pi \xi_\pi^{k+1} &= A_\pi \xi_\pi^k + B_\pi u^k \\ y^k &= C_\pi \xi_\pi^k + D_\pi u^k \end{aligned} \tag{A.4}$$

where $\xi_\pi \in \mathbb{R}^{2n_\psi}$ is the state of Π and the descriptor matrices are defined as:

$$E_\pi := \begin{bmatrix} I & 0 \\ 0 & A_\psi^T \end{bmatrix}, \quad \left[\begin{array}{c|c} A_\pi & B_\pi \\ \hline C_\pi & D_\pi \end{array}\right] := \left[\begin{array}{cc|c} A_\psi & 0 & B_\psi \\ -Q & I & -S \\ \hline S^T & zB_\psi^T & R \end{array}\right] \tag{A.5}$$

Notice $C_\pi = [\, S^T \;\; zB_\psi^T \,]$ and hence y^k in (A.4) partially depends on ξ_π^{k+1}. This is similar to Equation (A.2).

A descriptor representation for Π^{-1} is given by:

$$\begin{aligned} E_{in}\xi_{in}^{k+1} &= A_{in}\xi_{in}^k + B_{in} y^k \\ u^k &= C_{in}\xi_{in}^k + D_{in} y^k \end{aligned} \tag{A.6}$$

where $\xi_{in} := [\xi_\pi^T, \;\; u^T]^T \in \mathbb{R}^{2n_\psi + (n_v + n_w)}$ is the state of Π^{-1} and the matrices are defined as:

$$E_{in} := \begin{bmatrix} I & 0 & 0 \\ 0 & A_\psi^T & 0 \\ 0 & -B_\psi^T & 0 \end{bmatrix}, \quad \left[\begin{array}{c|c} A_{in} & B_{in} \\ \hline C_{in} & D_{in} \end{array}\right] := \left[\begin{array}{ccc|c} A_\psi & 0 & B_\psi & 0 \\ -Q & I & -S & 0 \\ S^T & 0 & R & -I \\ \hline 0 & 0 & I & 0 \end{array}\right] \tag{A.7}$$

It is emphasized that the filter Ψ is proper but the descriptor notation is required because A_ψ and/or R may be singular. In particular, if A_ψ is singular then $\Psi^\sim$ (and hence Π) is non-proper. In addition, if R is singular then Π^{-1} is non-proper. The lemma is now stated.

Lemma 26. *Let $\Pi = \Psi^\sim M \Psi \in \mathbb{RL}_\infty^{(n_v+n_w)\times(n_v+n_w)}$ be any factorization with $\Psi \in \mathbb{RH}_\infty^{n_r\times(n_v+n_w)}$ and $M = M^T \in \mathbb{R}^{n_r\times n_r}$. Define $Q := C_\psi^T M C_\psi$, $S := C_\psi^T M D_\psi$ and $R := D_\psi^T M D_\psi$ where $(A_\psi, B_\psi, C_\psi, D_\psi)$ are the state matrices of the filter Ψ. If Π is a Strict-PN multiplier then there exists a unique, real, stabilizing solution $X = X^T$ to $DARE(A_\psi, B_\psi, Q, R, S)$. In addition, $R + B_\psi^T X B_\psi$ is nonsingular.*

Proof. The multiplier Π has $2n_\psi$ zeros (possibly at $z = \infty$) where n_ψ is the state dimension of Ψ. These zeros are symmetric about the unit disk because $\Pi = \Pi^\sim$. The block-determinant formula yields

$$\det\left(\Pi\left(e^{j\omega}\right)\right) = \det\left(\Pi_{22}\left(e^{j\omega}\right)\right)\det\left(\Pi_{11}\left(e^{j\omega}\right) - \Pi_{12}\left(e^{j\omega}\right)\Pi_{22}^{-1}\left(e^{j\omega}\right)\Pi_{12}^*\left(e^{j\omega}\right)\right).$$

Then the Strict-PN conditions imply that Π is nonsingular, i.e. contains no zeros, on the unit circle. Therefore Π has n_ψ zeros strictly inside the unit circle. The poles of Π^{-1} are the zeros of Π and thus the matrix pencil $\lambda E_{in} - A_{in}$ has n_ψ generalized eigenvalues inside the unit disk. The generalized stable eigenspace of (E_{in}, A_{in}) is spanned by the columns of some matrix $X_s \in \mathbb{R}^{(2n_\psi+n_v+n_w)\times n_\psi}$. Hence there exists a Schur stable matrix $\Lambda \in \mathbb{R}^{n_\psi\times n_\psi}$ such that

$$A_{in} X_s = E_{in} X_s \Lambda. \tag{A.8}$$

Partition $X_s = \begin{bmatrix} X_1^T, & X_2^T, & X_3^T \end{bmatrix}^T$ compatibly with the blocks of A_{in} so that $X_1, X_2 \in \mathbb{R}^{n_\psi\times n_\psi}$ and $X_3 \in \mathbb{R}^{(n_v+n_w)\times n_\psi}$.

Next it is shown by contradiction that X_1 is nonsingular. Assume that X_1 is singular and let $\psi^0 \in \mathbb{R}^{n_\psi}$ denote a non-trivial vector in the null space of X_1. This vector cannot lie in the null space of $\begin{bmatrix} X_2 \\ X_3 \end{bmatrix}$ otherwise X_s would not span an n_ψ-dimensional space.

Define the signals u, y, ξ_π as follows:

$$u^k = \begin{cases} 0 & \text{for } k < 0 \\ X_3 \Lambda^k \psi^0 & \text{for } k \geq 0 \end{cases} \tag{A.9}$$

$$y^k = \begin{cases} B_\psi^T (A_\psi^T)^{-k-1} X_2 \psi^0 & \text{for } k < 0 \\ 0 & \text{for } k \geq 0 \end{cases} \tag{A.10}$$

$$\xi_\pi^k = \begin{cases} \begin{bmatrix} 0 \\ (A_\psi^T)^{-k} X_2 \psi^0 \end{bmatrix} & \text{for } k < 0 \\ \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \Lambda^k \psi^0 & \text{for } k \geq 0 \end{cases} \tag{A.11}$$

The signals u, y, and ξ_π are all in ℓ_2, [1] since A_ψ and Λ are Schur stable matrices. In addition, u, ξ_π, and y are input, state, and output solutions for Π (Equation (A.4)) with boundary condition $\xi_\pi^0 = \begin{bmatrix} 0 \\ X_2 \psi^0 \end{bmatrix}$. This can be directly verified for $k < 0$. For $k \geq 0$, define $\xi_{in}^k := X_s \Lambda^k \psi^0$. Use Equation (A.8) to show that ξ_{in}^k is a forward solution of Π^{-1} (Equation (A.6)) with initial condition $\xi_{in}^0 = X_s \psi^0$ and input $y^k = 0$. This verifies that the signals u, y, and ξ_π defined above are also a solution to Π for $k \geq 0$. Therefore, the Fourier transforms of u and y, denoted as U and Y, satisfy

$$Y(e^{j\omega}) = \Pi(e^{j\omega}) U(e^{j\omega}) \ \ \forall \omega \in [0, 2\pi]$$

Partition the signals as $u = \begin{bmatrix} u_1^T & u_2^T \end{bmatrix}^T$ and $y = \begin{bmatrix} y_1^T & y_2^T \end{bmatrix}^T$ such that $u_1, y_1 \in \ell_2^{n_v}$ and $u_2, y_2 \in \ell_2^{n_w}$. By construction, the inner products satisfy $\langle u_1, y_1 \rangle = \langle u_2, y_2 \rangle = 0$. Use Parseval's theorem and the Strict-PN sign-definiteness conditions to show[2] :

$$0 = \langle u_1, y_1 \rangle = \langle u_1, \Pi_{11} u_1 + \Pi_{12} u_2 \rangle \geq \langle u_1, \Pi_{12} u_2 \rangle$$
$$0 = \langle u_2, y_2 \rangle = \langle u_2, \Pi_{21} u_1 + \Pi_{22} u_2 \rangle \leq \langle u_2, \Pi_{21} u_1 \rangle$$

This immediately implies $\langle u_1, \Pi_{11} u_1 \rangle = \langle u_2, \Pi_{22} u_2 \rangle = 0$ because $\langle u_1, \Pi_{12} u_2 \rangle = \langle u_2, \Pi_{21} u_1 \rangle$. The Strict-PN conditions then yield $u_1 = u_2 = 0$ and hence $u = y = 0$. As a consequence $0 = u^0 := X_3 \psi^0$ and it must be that $X_2 \psi^0$ is non-trivial. In addition, $u = 0$ implies that ξ_π^k for $k \geq 0$ satisfies

$$E_\pi \xi_\pi^{k+1} = A_\pi \xi_\pi^k + B_\pi u^k = A_\pi \xi_\pi^k$$

[1] A slight abuse of notation is used here as these are two-sided signals.

[2] The inner product $\langle u_1, \Pi_{11} u_1 \rangle$ can be interpreted, via Parseval's theorem, in the frequency domain. For example, $\langle u_1, \Pi_{11} u_1 \rangle = \frac{1}{2\pi} \int_0^{2\pi} U_1(e^{j\omega})^* \Pi_{11}(e^{j\omega}) U_1(e^{j\omega}) d\omega$.

This is impossible since the nontrivial initial condition $\xi_\pi^0 = \begin{bmatrix} 0 \\ X_2\psi^0 \end{bmatrix}$ is in the antistable eigenspace of the pair (E_π, A_π) and this initial condition cannot yield a forward ℓ_2 solution $\xi_\pi^k = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \Lambda^k \psi^0$ for $k \geq 0$.

By contradiction, X_1 is nonsingular. Define $X := X_2 X_1^{-1}$. It follows from [116, Section 4] that $R + B_\psi^T X B_\psi$ is nonsingular and X is the unique stabilizing solution to $DARE(A_\psi, B_\psi, Q, R, S)$. This is a standard result and the remainder of the proof is only sketched. Define $K := -X_3 X_1^{-1}$ and $\tilde{X}_s := [\, I \; X^T \; -K^T \,]^T$. Equation (A.8) is equivalent to

$$A_{in}\tilde{X}_s = E_{in}\tilde{X}_s\tilde{\Lambda} \tag{A.12}$$

where $\tilde{\Lambda} := X_1 \Lambda X_1^{-1}$ is a Schur stable matrix. This leads to the following three equations:

$$A_\psi - B_\psi K = \tilde{\Lambda} \tag{A.13}$$

$$-Q + X + SK = A_\psi^T X \tilde{\Lambda} \tag{A.14}$$

$$S^T - RK = -B_\psi^T X \tilde{\Lambda} \tag{A.15}$$

Substituting the expression for $\tilde{\Lambda}$ (Equation (A.13)) into Equation (A.15) yields $K = (R + B_\psi^T X B_\psi)^{-1} \left(A_\psi^T X B_\psi + S \right)^T$. This expression along with (A.13) and (A.14) can be used to show, via standard manipulations, that X satisfies the $DARE(A_\psi, B_\psi, Q, R, S)$. Based on (A.13), $A_\psi - B_\psi K$ is a Schur stable matrix. Therefore, X is a stabilizing solution to the $DARE$. The above steps require a few additional facts to be demonstrated, e.g. X is symmetric and $R + B_\psi^T X B_\psi$ is nonsingular. These details can be found in [116]. □

As mentioned before, an alternative proof of the above lemma can be constructed using operator-theoretic arguments. Specifically, one can justify the applicability of [64, Theorem 4.12.8] using the Strict-PN condition and then prove the above lemma as a consequence.