Extended Proximal Primal-Dual Online Feedback Controller

by

Spencer Kelly

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science

Edward S. Rogers Department of Electrical and Computer Engineering
University of Toronto

Extended Proximal Primal-Dual Online Feedback Controller

Spencer Kelly
Master of Applied Science
Edward S. Rogers Department of Electrical and Computer Engineering
University of Toronto
2023

Abstract

This thesis studies the convergence rate of the primal-dual method for convex equality-constrained optimization problems. Three different approaches are used to determine the convergence rate, and the rates guaranteed are compared. Further, an online feedback-based optimization controller for a non-linear plant is developed. The control algorithm uses an extension of the primal-dual algorithm, the extended proximal primal-dual algorithm, to drive the system toward a critical point of a cost function. The cost function is able to consider differentiable and non-differentiable costs on the inputs and outputs of the system. Finally, the algorithm is applied to a simulated distribution feeder system. The controller is able to successfully optimize the power inputs while satisfying the physical constraints of the system.

Acknowledgements

I would like to thank my supervisor Professor John Simpson-Porco for helping me throughout my graduate school journey. His knowledge and enthusiasm to teach were imperative in my research. He provided the guidance and encouragement required to complete this thesis, and, for that, I am extremely grateful.

I would also like to thank the committee members Professor Lacra Pavel and Professor Margaret Chapman. I appreciate you both for taking the time to read my thesis and providing feedback on my results.

I thank my family for their continuous support. Without my parents, I would not be in the position I am today. Their unwavering love and encouragement have given me endless opportunities. Throughout my life, they created an environment in which learning was encouraged, and I was free to explore my interests. My brother, Dayton, has always been a source of sound advice that I have leaned on. Since I can remember, his work ethic and determination have motivated me to be my best. Thank you.

I thank my girlfriend, Sabrina, for supporting me throughout my journey. Your love and support have helped me power through some difficult times. I am forever grateful to have you in my life.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Motivation

The landscape of electricity consumption, generation, and distribution is undergoing a consequential shift. The implementation of renewable energy sources and energy storage devices are changing the way electricity is used and produced. In order to not only manage but embrace this shift, the electrical grid needs to work as intelligently as possible.

Traditionally, the power grid follows a centralized framework. Large plants produce electricity that the transmission and distribution networks channel to consumers [1]. These networks were originally designed for unidirectional power flow; however, with the rise of distributed energy resources, this architecture is decentralizing [2]. Nodes on the distribution grid can now dual as both a consumer and producer of electricity. A distributed generation framework provides numerous benefits for the overall quality and reliability of the grid but poses unique challenges in optimizing and managing the network [3],[4].

With environmental and economic consequence, optimal power flow is an important problem in power grid operation. Optimal power flow is concerned with calculating the output power at each generator that will satisfy power demands as well as minimize a cost function [3]. The cost function can be designed to minimize some undesirable quality of the system, such as power loss, non-renewable usage, or operating cost. The optimization problem incorporates the physical constraints of the system, including maximal line voltage and generator limits. With high-dimensional inputs and constraints, some of which are non-convex, optimal power flow is a challenging problem.

Optimal power flow lies at the intersection of control and optimization theory. Because the power grid is a physical system, control algorithms are needed to ensure line frequencies and voltages are safely regulated. To optimize the dispatching of generator set points, optimization theory needs to be leveraged. The traditional approach is to use offline predictions over a given time horizon and use a grid model to determine the optimal set points over the horizon. The trajectory generated is then provided to the real-time controllers. Due to the offline nature of the optimization algorithm, the system is unable to respond to unexpected disturbances. Further, the optimization method requires

an accurate system model. With more distributed energy resources entering the grid, these models will become increasingly complex.

This provides the motivation for the control algorithm that will be developed in this thesis. An online feedback-based optimization (OFBO) approach will be explored. Using OFBO in control has become of interest in recent years due to its applicability in large-scale system control and optimization [5]. The overarching goal of OFBO is to leverage real-time measurements from the system of interest to determine the optimal controller set points. Because it is not attempting to predict the future states of a system and optimize over different possibilities, OFBO algorithms are generally efficient and do not require a model. Due to the feedback mechanism used in OFBO, this approach boasts many of the same advantages as using a traditional feedback loop in a control system. The algorithm is robust to disturbances to the system; therefore, it is able to effectively adapt to unexpected system behaviour. Further, typically OFBO implementations are relatively simple to implement. These benefits do come at a cost. OFBO approaches are unable to optimize over the transient states of the system and only consider costs on the steady-state values. Applications of OFBO algorithms include machine learning, telecommunications, and power grid control [5], [6].

1.2 Literature Review

In [7], a general framework for online optimization with feedback was introduced. The approach used a projected primal-dual algorithm to drive the optimization. The system model considered was a linear input-output mapping and did not consider the dynamics of the system. The cost function used in [7] handled differentiable costs on both the input and output. Further, the algorithm used a projection operator to ensure the inputs remained in the feasible input set. This paper was expanded in [5] where a general non-linear input-output mapping was considered. Further, the input-output mapping was considered unknown and an approximate Jacobian was used in the optimization algorithm. Because of this, the notion of an online approximate solution was introduced. The optimization algorithm was an online version of projected gradient descent. The cost function considered differentiable cost functions on the inputs and outputs of the system. The projection operator also ensured that the inputs stay within the feasible set.

In [8], the input-output mapping considered was the steady-state transfer function for an LTI system. Therefore, the mapping was linear. The optimization approach used was a primal-dual method based on the proximal augmented Lagrangian. This primal-dual algorithm allowed for non-differentiable costs to be placed on the output. Non-differentiable cost functions allow users to place hard constraints on the output of the system or use the l_1 norm to induce sparsity in the output variable. This approach did not include a projection operator on the input, thus there is no mechanism for ensuring the inputs stay within the constraint set.

The convergence of the primal-dual method based on the proximal augmented Lagrangian has been researched in a number of papers. Outside of its use in control, it is an effective approach for nonsmooth composite minimization. In [9], the approach used for convergence is adapted from [10]. The non-linearities of the algorithm are rewritten as disturbances to the system and a pointwise IQC

is used to bound the effect of the disturbances. A linear matrix inequality can then be solved to determine the feasibility of the algorithm for a given rate and step size. In [11] the continuous-time algorithm was considered and a Lyapunov approach was used to prove convergence to a fixed point. A quadratic Lyapunov function was proposed and a convergence rate was provided as a function of system parameters. This was the only Lyapunov approach I found for this algorithm.

1.3 Contributions

The most significant results in this thesis are as follows.

- In Chapter 3, three approaches for determining the convergence rate of the primal-dual algorithm are proposed and analyzed. The rates are compared as problem parameters vary. The analysis provides insights into the benefits of each approach.
- An alternative to the primal-dual algorithm is proposed. This algorithm introduces a tunable time-scale separation between the primal and dual updates. We show that the rate can be determined depending on the time-scale separation chosen. By changing the time-scale separation, we can tune the rate between the results achieved in [12] and a rate proportional to the optimal rate for the dual method.
- In Chapter 4, an online feedback-based optimization controller is proposed. The control algorithm considers both differentiable and non-differentiable costs on the inputs and outputs of the system. Further, a non-linear input-output mapping is considered. This improves upon the results in [8] by allowing non-differentiable costs on the inputs and considering a non-linear plant. A Lyapunov approach is used to determine the convergence of the algorithm to a unique fixed point when the input-output mapping is linear and known. The Lyapunov function provides greater rates than the Lyapunov function used in [11]. Finally, when the input-output mapping is unknown and non-linear, a linear matrix inequality is proposed. Solving the linear matrix inequality for a given rate and step size guarantees exponential convergence to a fixed point.
- In Chapter 5, the control algorithm is implemented to control a simulated 36-bus distribution feeder system. The controller successfully minimizes the power curtailed while enforcing hard constraints on the inputs and outputs of the system.

1.4 Thesis Layout

The thesis is organized in the following structure:

- In Chapter 2, a summary of the necessary mathematical background is provided. The background goes over results in optimization and introduces some algorithms of interest.
- In Chapter 3, the rate guarantees for the primal-dual algorithm for a convex equality-constrained optimization problem are studied. Three different approaches are discussed and used to determine convergence rates. These rates are then compared to determine which approach provides the optimal rate. Further, the time-scale separated primal-dual algorithm is motivated and proposed.

- In Chapter 4, an online feedback-based optimization controller is proposed. The system is introduced and the control algorithm is derived. The algorithm is then proven to converge under two different assumptions: when the input-output mapping is linear and known and when the mapping is non-linear and unknown.
- In Chapter 5, the algorithm is adapted to control a 36-bus distribution feeder system. The implementation details are outlined and the results are given.
- In Chapter 6, a conclusion is provided, and the future research directions are listed.

Chapter 2

Background

2.1 Matrices and Norms

We begin by presenting some results in matrix algebra. The following will provide the mathematical pretext for the methods and results of the thesis. This background will include a brief overview of symmetric matrices and matrix norms. We denote the space of real $n \times m$ matrices as $\mathbb{R}^{n\times m}$ and the space of real symmetric $n \times n$ matrices as $\mathbb{S}^n \subset \mathbb{R}^{n\times n}$. We will use $\mathbf{0}_{n\times m}$ to denote an $n \times m$ matrix with all zero entries. Further, $\mathbf{I}_n$ will be used to represent the n-dimensional identity matrix. If the indices are omitted, the dimensions of both the zero matrix and identity matrix can be inferred. For a matrix $A \in \mathbb{R}^{n\times m}$, the notation $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ will be used to denote the maximum and minimum singular values of A, respectively. The condition number of A will be denoted as $\text{cond}(A)$ and defined as $\text{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$. Properties of symmetric matrices will be leveraged throughout the thesis; Lemma 2.1.1 will highlight some of these properties.

Lemma 2.1.1. *[13] Given a matrix $Q \in S^n$, the following properties hold*

(i) $Q = Q^\top$;

(ii) Let λ_i denote the i^{th} eigenvalue of Q, then $\lambda_i \in \mathbb{R}$ for all $i \in \{1, ..., n\}$;

(iii) Let σ_i denote the i^{th} singular value of Q and λ_i denote the i^{th} eigenvalue of Q, then $\sigma_i = |\lambda_i|$ for all $i \in \{1, ..., n\}$;

(iv) There exists an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ such that $Q = V^\top \Lambda V$ where Λ is a diagonal $n \times n$ matrix.

For a symmetric matrix $Q \in \mathbb{S}^n$, we say that Q is positive (semi)definite if all of the eigenvalues of Q are greater than (or equal to) zero. We will use $Q \succ 0$ ($Q \succeq 0$) to represent Q being positive (semi)definite. We say that Q is negative (semi)definite if $-Q$ is positive (semi)definite. We will use $Q \prec 0$ ($Q \preceq 0$) to represent Q being negative (semi)definite. If a matrix is denoted as a definite matrix, this implies that it is also symmetric, as definite asymmetric matrices are out of the scope of the thesis. Lemma 2.1.2 lists some useful properties about definite matrices.

Lemma 2.1.2. *Consider the matrices $Q \in S^n$ and $A \in \mathbb{R}^{n\times m}$. The following properties hold*

(i) $Q \succ 0$ *if and only if* $Q^{-1} \succ 0$;

(ii) $AA^\top \succ 0$ *if and only if* A *is full row-rank;*

(iii) If A *is full row-rank and* $Q \succ 0$*, then* $AQA^\top \succ 0$*;*

(iv) If $Q \succeq 0$ *then* $A^\top QA \succeq 0$.

Proof. (i) and (iii) are given in [14]. For property (ii) assume that $AA^\top \succ 0$ and assume that A is not full-row rank. Then there exists a non-zero vector $x \in \mathbb{R}^n$ such that $A^\top x = \mathbf{0}$. Thus, for some $x \neq \mathbf{0}$ we have $x^\top AA^\top x = 0$. By contradiction, A must be full row-rank. To prove the reverse direction, we know that for all non-zero $x \in \mathbb{R}^n$, there exists non-zero $v \in \mathbb{R}^m$ such that $v = A^\top x$. Thus,

$$x^\top AA^\top x = ||v||_2^2 > 0.$$

Thus, $AA^\top \succ 0$. To prove (iv), notice that for $v = A^\top x$,

$$x^\top AQA^\top x = v^\top Qv \geq 0,$$

where the final inequality follows from the positive semidefiniteness of Q □

Definite matrices will be useful throughout the algorithm analysis. Throughout the process, scalar-valued functions with convexity and continuity assumptions are considered. Definite matrices can act as linear analogs to these functions and their behaviour can be easily bounded. To determine these bounds, the notion of a matrix norm is required. The Euclidean norm $||x||_2 = \sqrt{x^\top x}$ is used as a vector norm throughout the thesis. The P-norm, $||x||_{P,2} = \sqrt{x^\top Px}$, where P is positive definite, will also be used. The matrix 2-norm will be the matrix norm used throughout this thesis unless otherwise explicitly stated.

Definition 2.1.3. *Consider a matrix* $A \in \mathbb{R}^{m\times n}$*. The* ***matrix*** 2***-norm*** *of* A*,* $||A||_2$*, is defined as*

$$||A||_2 := \sup_{x\in\mathbb{R}^n} \frac{||Ax||_2}{||x||_2} = \sup_{||x||_2=1} ||Ax||_2 = \sigma_{\max}(A) \tag{2.1}$$

To transition between the P-norm and Euclidean norm or bound positive definite matrices, Lemma 2.1.4 will be used.

Lemma 2.1.4. *[13] Given a matrix* $Q \in S^n$ *and a vector* $x \in \mathbb{R}^n$*, the following inequality holds*

$$\sigma_{\min}(Q)||x||_2^2 \leq x^\top Qx \leq \sigma_{\max}(Q)||x||_2^2$$

2.2 Functions

Along with linear functions, more general non-linear functions will be considered. Throughout the thesis, we will consider the extended real space $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. For a differentiable function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, we will say the gradient ∇f is given by

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

where $\frac{\partial f}{\partial x_i}$ are the partial derivatives of f with respect to x_i. The Hessian of a twice differentiable function, f, is given by

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) & \dots & \frac{\partial^2 f}{\partial x_2 x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1}(x) & \frac{\partial^2 f}{\partial x_n x_2}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{bmatrix}$$

where $\frac{\partial^2 f}{\partial x_j x_i}$ are the second partial derivatives with respect to x_i and x_j. Non-differentiable functions will also be analyzed. To consider non-differentiable functions in an optimization framework, the subdifferential is useful.

Definition 2.2.1 (Subdifferential)**.** *[15, Section 1.2] Given a function $g : \mathbb{R}^n \to \bar{\mathbb{R}}$, a **subgradient** of g at $x \in \mathbb{R}^n$ is any vector $d \in \mathbb{R}^n$ such that*

$$g(y) \geq g(x) + d^\top (y - x), \quad \forall y \in \mathbb{R}^n.$$

*The set of all subgradients for a point $x \in \mathbb{R}^n$ is called the **subdifferential** of g at x and is denoted by*

$$\partial g(x) := \{d \in \mathbb{R}^n \mid g(y) \geq g(x) + d^\top (y - x),\ \forall y \in \mathbb{R}^n\}.$$

The subdifferential of a differentiable function contains one element, which is the gradient of the function [15, Section 1.2]. The subdifferential of a vector is always well-defined but may be the empty set [15, Section 1.2]. To motivate this concept, consider the function $g(x) = ||x||_1$ for $x \in \mathbb{R}$. We know that f is differentiable at each point in $\mathbb{R}$ except 0. To find the subdifferential of g at 0, we must find all d that satisfies $||y||_1 \geq dy$. Clearly, any $d \in [-1, 1]$ satisfies this inequality. Therefore, $\partial g(0) = \{d \in \mathbb{R} \mid -1 \leq d \leq 1\}$.

2.3 Optimization Theory

The algorithms explored in the thesis will be designed to find values of interest for a defined objective. This section will formally introduce optimization theory and relevant results.

An optimization problem involves minimizing or maximizing a cost function within a constraint set. The standard form of an optimization problem is given as

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \\ &\text{s.t. } g_i(x) \leq 0,\ i \in \{1, ..., l\}, \\ &\quad h_i(x) = 0,\ i \in \{1, ..., m\}, \end{aligned} \tag{2.2}$$

where $x \in \mathbb{R}^n$ is the optimization variable, $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the cost or objective function, $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i \in \{1, ..., l\}$ are the inequality constraints, and $h_i : \mathbb{R}^n \to \mathbb{R}$ for $i \in \{1, ..., m\}$ are the equality constraints [16, Section 4.1.1]. The intersection of the equality constraints and the inequality constraints form the constraint set or feasibility set. Notice that problem (2.2) is formatted as a

minimization problem rather than a maximization problem. Any maximization problem can be equivalently formulated in the standard form by considering the negative cost function. In this thesis, we will consider optimization problems without inequality constraints. Therefore, we assume $g_i = 0$ for all $i \in \{1, .., l\}$ throughout this background. The optimization problems of interest will be in the form of

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} f(x) \\ &\text{s.t. } h_i(x) = 0,\ i \in \{1, ..., m\}. \end{aligned} \tag{2.3}$$

For the general problem, the goal is to find the point in the constraint set which minimizes the cost function. This point is called the global minimizer or optimal point.

Definition 2.3.1. *[16, Section 4.1.1] Consider the optimization problem (2.3). We say that x^* is a **global minimizer** or an **optimal point** if*

$$f(x^*) = \inf_{x \in \mathbb{R}^n} \{f(x) \mid h_i(x) = 0\ \forall i \in \{1, ..., m\}\}.$$

The global minimizer is not necessarily unique. Searching over the feasible set, one may find other optimal points, but they will never have a lower value than $f(x^*)$. Consequently, once a vector that satisfies Definition 2.3.1 is found, problem (2.3) is solved. A weaker property can also be stated for a family of points that share some similarities to global minima.

Definition 2.3.2. *[16, Section 4.1.1] Consider the optimization problem (2.3). We say that x^* is a **local minimizer** or **critical point** if there exists $R > 0$ such that*

$$f(x^*) = \inf_{x \in P(x^*)} \{f(x) |\ h_i(x) = 0\ \forall i \in \{1, ..., m\}\},$$

where $P(x^*) = \{x \in \mathbb{R}^n \mid ||x - x^*||_2 < R\}$.

The presence of local minima can cause issues for traditional optimization methods because they share similar properties to global minima. Optimization algorithms typically have a stopping criterion that determines that the algorithm has identified a potential solution. If a local minimizer satisfies this stopping criterion, the algorithm may terminate at this sub-optimal point.

Depending on the properties of the objective function and constraints, the ease of finding solutions and determining their optimality varies. We will begin by considering the simplest case: unconstrained optimization. An unconstrained optimization problem takes the form of problem (2.3) when $h_i(x) = 0$ for all $x \in \mathbb{R}^n$, $i \in \{1, ..., m\}$,

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2.4}$$

For the unconstrained optimization problem, the following necessary optimality condition can be considered.

Corollary 2.3.2.1 (First Order Necessary Optimality Condition)**.** *[15, Proposition 1.2.17] Consider the optimization problem (2.4). For $x^* \in \mathbb{R}^n$, $0 \in \partial f(x^*)$ if x^* is a global minimum.*

Corollary 2.3.2.1 encodes the notion that the global minimum will lie at a point where the gradient is zero. In the differentiable scalar case, thinking of the gradient as the slope of the tangent line,

the tangent at an inflection point must be a flat line. Returning to the example of the subdifferential of the l_1 norm, we see that $0 \in \partial g(0)$, as expected. The condition is necessary and not sufficient because local minima or maxima may also satisfy this condition. Thus, it provides the mechanics to identify critical points, but without further assumptions, global optimality cannot be guaranteed.

A similar condition can be stated for equality-constrained optimization problems, problem (2.3). To handle the equality constraints, an equivalent unconstrained optimization problem is proposed. To do this, we introduce the Lagrangian, L, for problem (2.3),

$$L(x, \lambda) := f(x) + \sum_{i=0}^{m} \lambda_i h_i(x). \tag{2.5}$$

For equation (2.5), $\lambda = [\lambda_1, \ldots, \lambda_m] \in \mathbb{R}^m$ is the dual variable or Lagrange Multiplier and $x \in \mathbb{R}^n$ is the primal variable. An important property of the Lagrangian is that it can provide a lower bound on the minimum value of problem (2.3). Suppose x^* is an optimal point of problem (2.3), then from [16, Section 5.2], we know that

$$g(\lambda) := \inf_{x \in \mathbb{R}^n} \left(f(x) + \sum_{i=0}^{m} \lambda_i h_i(x) \right) \leq f(x^*). \tag{2.6}$$

The function $g(\lambda)$ is known as the dual function. Because the dual function acts as a lower bound for the optimal value of problem (2.3), it is desirable to determine the tightest lower bound that can be achieved. This is exactly the motivation for the dual problem,

$$\max_{\lambda \in \mathbb{R}^m} \; g(\lambda). \tag{2.7}$$

Suppose that λ^* is a global maximum of problem (2.7). We know from (2.6) that $g(\lambda^*) \leq f(x^*)$. We define the optimal duality gap as $d(x^*, \lambda^*) = f(x^*) - g(\lambda^*)$ [16, Section 5.2.2]. The optimal duality gap is clearly always positive and encodes the distance between the true optimal value and the tightest lower bound provided by the Lagrangian. When the optimal duality gap is zero, $d(x^*, \lambda^*) = 0$, we say that strong duality holds. Under certain conditions, strong duality can be guaranteed. These conditions will be highlighted in Section 2.4. Similar to the necessary optimality conditions defined for general unconstrained optimization problems, we introduce the necessary conditions for equality-constrained optimization problems.

Theorem 2.3.3 (KKT Conditions For Equality-Constrained Problems)**.** *[15, Theorem 4.1.2] Consider the optimization problem (2.3) and its associated dual problem, problem (2.7). Suppose $x^* \in \mathbb{R}^n$ and $\lambda^* \in \mathbb{R}^m$ are optimal points for the primal and dual problems, respectively. If strong duality holds for x^* and λ^*, then the following conditions hold*

$$\begin{aligned} h_i(x^*) &= 0, \; i \in \{1, ..., m\} \\ \partial f(x^*) + \sum_{i=0}^{m} \lambda_i \partial h_i(x^*) &\ni \mathbf{0} \end{aligned} \tag{2.8}$$

Conditions (2.8) are known as the Karush–Kuhn–Tucker (KKT) conditions. The KKT conditions are derived using Corollary 2.3.2.1. To see this, consider that the gradient of the dual problem with

respect to λ_i is $\nabla_{\lambda_i} g(\lambda) = h_i(x)$ for $i \in \{1, ..., m\}$. Because λ^* maximizes the dual problem, the gradient must be zero, which matches the first KKT condition. Similarly, x^* minimizes the Lagrangian with respect to x. Thus, $\nabla_x L(x^*, \lambda^*) \ni \mathbf{0}$, which exactly matches the second condition.

2.4 Convex Optimization Theory

Convex functions are a class of functions that possess qualities that simplify optimization problems. For this reason, introductory optimization often starts with considering convex problems, and convex cost functions are often desired. Convex optimization problems have stronger optimality guarantees than more general problems. Before formally defining a convex function, we introduce convex sets.

Definition 2.4.1. *[16, Section 2.1.1] A **convex set** $C \subseteq \mathbb{R}^n$ is a set where all $x, y \in C$ and $\theta \in [0, 1]$ satisfy*

$$\theta x + (1 - \theta) y \in C.$$

From this definition, a convex set is a set in which all line segments connecting two points in the set are also contained in the set. With the notion of a convex set defined, we can formally introduce the definition for a convex function.

Definition 2.4.2. *[16, Section 3.1.1] Consider a function $f : C \to \bar{\mathbb{R}}$ where $C \subseteq \mathbb{R}^n$. A function f is a **convex function** if C is a convex set and for all $x, y \in C$ and $\theta \in [0, 1]$, the following inequality holds*

$$f(\theta x + (1 - \theta) y) \leq \theta f(x) + (1 - \theta) f(y).$$

Intuitively, if f is convex, then the value of f evaluated at points along a line segment between x and y in the domain must be less than or equal to the value of the line segment between $f(x)$ and $f(y)$. There is also a dual concept of concavity. A function f is concave if $-f$ is convex. These ideas are illustrated in Figure 2.1. A convex optimization problem is an optimization problem in the form

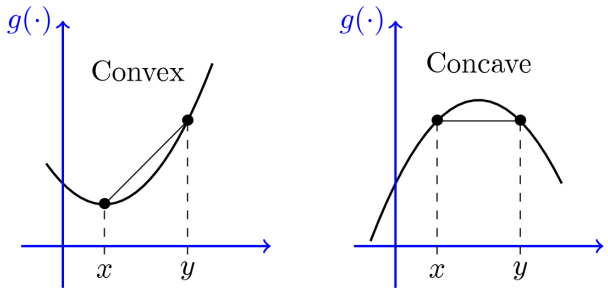

Figure 2.1: Illustration of convex and concave functions [17]

of problem (2.3) where f is convex and h is affine. Thus, when the problem is convex, we represent $h(x) = Ax - b$ where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Considering a convex optimization problem provides the following benefits:

(i) The gradient condition in Corollary 2.3.2.1 becomes a sufficient condition for global optimality [15, Proposition 1.2.17].

(ii) If there exists $x^* \in \mathbb{R}^n$ such that $Ax^* = b$, then strong duality holds (Slater's Condition) [16, Section 5.5.3].

(iii) If strong duality holds, then the KKT conditions in Theorem 2.3.3 become sufficient conditions for global optimality [16, Section 5.5.3].

Convex optimization problems allow us to ensure the optimality of a vector by guaranteeing all local minimizers are global minimizers. Throughout this thesis, we will assume that Slater's Condition is satisfied, ensuring strong duality.

Along with convexity, we will regularly assume that cost functions meet stronger assumptions. Specifically, we may require that a function be strongly convex and have a Lipschitz continuous gradient.

Definition 2.4.3. *[18] Consider a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$. The function f is m-**strongly convex** if for all $x \in \mathbb{R}^n$, the function $f(x) - \frac{m}{2}||x||_2^2$ is convex.*

Strong convexity is closely related to the notion of strong monotonicity. Precisely, if a function, f, is m-strongly convex and differentiable, then ∇f is m-strongly monotone. A function F is m-strongly monotone if it satisfies the inequality

$$(F(x) - F(y))^\top (x - y) \geq m||x - y||_2^2, \quad \forall\, x, y \in \mathbb{R}^n.$$

Further, if f is m-strongly convex and twice differentiable, then $m\mathbf{I} \preceq \nabla^2 f(x)$ for all $x \in \mathbb{R}^n$.

Definition 2.4.4. *[18] Consider a function $F : \mathbb{R}^n \to \mathbb{R}^n$. The function is L-Lipschitz continuous if for all $x, y \in \mathbb{R}^n$*

$$||F(x) - F(y)||_2 \leq L||x - y||_2.$$

Suppose a function f is twice differentiable and has a Lipschitz continuous gradient.Then we have $\nabla^2 f \preceq L\mathbf{I}$. Functions that are both strongly convex and have a Lipschitz continuous gradient will be considered throughout the thesis. Let $S(m, L)$ denote the set of functions that are m-strongly convex and have an L-Lipschitz continuous gradient. A useful property of these functions is the ability to place upper and lower bounds on the gradient's behavior. For any $f \in S(m, L)$ and $x, y \in \mathbb{R}^n$, the following inequality is satisfied [18]

$$m||x - y||_2^2 \leq (\nabla f(x) - \nabla f(y))^\top (x - y) \leq L||x - y||_2^2.$$

2.5 Optimization Algorithms

To find a solution to an optimization problem, an optimization algorithm can be employed. In this thesis, we will explore discrete-time fixed point algorithms that can be written as

$$x^{k+1} = T(x^k), \tag{2.9}$$

where $k \in \mathbf{Z}_{\geq 0}$ is the time index, $T : \mathbb{R}^n \to \mathbb{R}^n$ is a function, and x^0 is some arbitrary initial point. The sequence $\{x^i\}_{i=0}^k$ generated by algorithm (2.9) from the starting point x^0 is known as the trajectory of the algorithm. To understand why this method is called a fixed point algorithm, we must define a fixed point.

Definition 2.5.1. *[18] Consider a function $T : \mathbb{R}^n \to \mathbb{R}^n$. A point $x^* \in \mathbb{R}^n$ is a fixed point of T if $x^* = T(x^*)$.*

Thus, any equilibrium point of algorithm (2.9) must be a fixed point of T. To solve a problem using a fixed point iteration, one needs to determine a mapping, T, such that the iteration converges to a fixed point and the fixed points of T are solutions to the problem in question.

2.5.1 Convergence Analysis

Throughout the thesis, we will be interested in proving the algorithms of interest converge exponentially.

Definition 2.5.2 (Exponential Convergence)**.** *Consider a trajectory defined by algorithm (2.9) with initial point $x^0 \in \mathbb{R}^n$. Suppose there is a unique fixed point of T, $x^* = T(x^*)$. The algorithm is exponentially convergent if there exists $\beta > 0$ and $\rho \in [0, 1)$ such that for all $k \in \mathbf{Z}_{\geq 0}$ and $x^0 \in \mathbb{R}^n$,*

$$||x^k - x^*||_2 \leq \beta \rho^k ||x^0 - x^*||_2.$$

Exponential convergence to an equilibrium point is a powerful notion. Given the time step, initial condition, and convergence rate, the current iteration's distance from the fixed point can be upper bounded. Note that in optimization theory, exponential convergence is often called geometric convergence. Looking at algorithm (2.9) through a control theorist's lens, we see that the algorithm can be viewed as a discrete-time dynamical system. Further, our definition of exponential convergence is equivalent to the notion of exponential stability for a dynamical system. Thus, to show that an algorithm converges, we will borrow a popular technique from control and dynamical systems.

Theorem 2.5.3 (Lyapunov Stability Theory)**.** *[19, Theorem 13.2] Consider the fixed point iteration (2.9) and assume that $T(\mathbf{0}) = \mathbf{0}$. If $V : \mathbb{R}^n \to \mathbb{R}$ satisfies*

$$\begin{aligned} V(x) &> 0, && \forall x \in \mathbb{R}^n \setminus \{0\} \\ V(0) &= 0 \\ V(T(x)) - V(x) &< 0 \\ ||x||_2 \to \infty &\implies V(x) \to \infty \end{aligned}$$

*then V is a **Lyapunov function** for (2.9). Further, if there exists $\gamma_1, \gamma_2 > 0$, $p \geq 1$, and $\rho \in (0, 1)$ such that*

$$\begin{aligned} \gamma_1 ||x||_2^p \leq V(x) \leq \gamma_2 ||x||_2^p, &\qquad \forall x \in \mathbb{R}^n \\ V(T(x)) - V(x) \leq (\rho - 1) V(x), &\qquad \forall x \in \mathbb{R}^n \setminus \{0\} \end{aligned}$$

then the trajectories generated by Algorithm 2.9 satisfy

$$||x^k||_2 \leq \left(\tfrac{\gamma_2}{\gamma_1}\right)^{1/p} \rho^{k/p} ||x^0||_2$$

Often, the Lyapunov function is thought of as an energy function for the system. It is a function that compiles all of the information about the current state into a scalar. The goal is then to determine how this scalar changes over system trajectories. If it is consistently decreasing, we are able to conclude the state is converging. For some of the algorithms we will analyze, we will require

a slightly stronger property than exponential convergence. We may wish to show that an algorithm is contractive.

Definition 2.5.4. *[18] A function $T : \mathbb{R}^n \to \mathbb{R}^n$ is a **contraction** with rate $c \in [0, 1)$ with respect to a norm $||\cdot||$ if*

$$||T(x) - T(y)|| \leq c||x - y||, \quad \forall\, x, y \in \mathbb{R}^n.$$

Thinking of the norm as a measure of distance, applying a contraction to any two points shortens the distance between them on each iteration. Therefore, if F is a contraction, applying the composition of F with itself multiple times to two points will shrink the distance between the two points exponentially. This notion is formalized in the following corollary.

Corollary 2.5.4.1. *Consider the fixed point iteration defined by algorithm (2.9). Suppose that T is a contraction with rate c with respect to the norm $||\cdot||_2$. Then the following inequality holds for all $x^0, y^0 \in \mathbb{R}^n$*

$$||x^k - y^k||_2 \leq c^k||x^0 - y^0||_2$$

where x^k and y^k are the trajectories at iteration k defined by (2.9) with initial conditions x^0 and y^0, respectively.

Proof. Take any $x^0, y^0 \in \mathbb{R}^n$ and $k \in \mathbf{Z}_{\geq 0}$. Then

$$\begin{aligned} ||x^k - y^k|| &= ||T(x^{k-1}) - T(y^{k-1})|| \\ &\leq c||x^{k-1} - y^{k-1}|| \end{aligned}$$

Applying this inequality repeatedly, we arrive at

$$\begin{aligned} ||x^k - y^k|| &\leq c^n||x^{k-n} - y^{k-n}||, \quad n \leq k \\ &\leq c^k||x^0 - y^0|| \end{aligned}$$

□

Thus, a fixed point iteration where the mapping is contractive, guarantees that all trajectories generated by the iteration converge exponentially to a fixed point. Further, using Banach Contraction Theorem, it is known that if $T : \mathbb{R}^n \to \mathbb{R}$ is contractive with respect to a norm $||\cdot||$, then (2.9) converges to a unique fixed point if $(\mathbb{R}^n, ||\cdot||)$ is a complete metric space [20, Theorem 1.6]. Therefore, for any initial point x^0, the distance between the k^{th} iteration and the fixed point can be bounded,

$$||x^k - x^*|| \leq c^k||x^0 - x^*||.$$

Notice that the results guaranteed using a Lyapunov analysis are not as strong as proving that an algorithm is contractive with respect to $||\cdot||_2$. Lyapunov theory allows us to determine an exponential bound on the convergence of any trajectory to an equilibrium point, while a contraction is the exponential convergence of all trajectories towards one another.

2.5.2 Optimization Algorithms

Fixed point iterations can be used to find the critical points of optimization problems. When the cost function and equality constraints are differentiable, gradient descent can be an effective algorithm.

Gradient descent takes advantage of the fact that the gradient of a function points in the direction of the steepest increase of the function's value. Thus, moving in the direction of the negative gradient should provide the steepest descent to a critical point. Gradient descent for a function $f : \mathbb{R}^n \to \mathbb{R}$ can be written formally as

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

where $k \in \mathbf{Z}_{\geq 0}$ is the time step and $\alpha > 0$ is the step size. Because the gradient is $\mathbf{0}$ at local minima, gradient descent is a fixed point iteration with the equilibria corresponding to critical points of the optimization problem. Placing assumptions on the cost function, we can provide convergence results for a family of algorithms that includes gradient descent.

Corollary 2.5.4.2. *[12] Consider $F : \mathbb{R}^n \to \mathbb{R}^n$ where F is m_f-strongly monotone and L_f-Lipschitz continuous. Then, if $\alpha \leq \frac{2}{m_f + L_f}$, for all $x, y \in \mathbb{R}^n$*

$$||x - \alpha F(x) - (y - \alpha F(y))||_2 \leq (1 - \alpha m_f)||x - y||_2.$$

Corollary 2.5.4.2 provides sufficient conditions to determine an algorithm in the form of $x^{k+1} = x^k - \alpha F(x^k)$ is contractive. Gradient descent belongs to this family of algorithms.

Throughout this thesis, we will analyze equality-constrained optimization problems, problem (2.4). A number of optimization problems, from machine learning to signal processing, have this structure [21]. Network flow optimization problems also take this form, where A encodes the network graph structure, b represents the incoming and outgoing flow rates at each node, and x represents the flow rate at each edge [22]. Problem (2.4) generalizes a large number of key engineering problems; therefore, developing and analyzing algorithms that can identify the minimizers of these problems is crucial.

There are two different gradient methods that can be used to find the critical points of problem (2.4). The first algorithm is the primal-dual method. This method is a specific implementation of gradient descent. The algorithm involves descending along the gradient of the Lagrangian with respect to the primal variable while ascending along the gradient of the Lagrangian with respect to the dual variable. The general form for the primal-dual method is given by

$$\begin{aligned} x^{k+1} &= x^k - \alpha \left(\nabla f(x^k) + \nabla h(x^k)^\top \lambda^k \right) \\ \lambda^{k+1} &= \lambda^k + \beta h(x^k), \end{aligned} \tag{2.10}$$

where $\alpha, \beta > 0$, $\lambda = [\lambda_1, \ \ldots, \ \lambda_m]^\top \in \mathbb{R}^m$, and $h(x) = [h_1(x), \ \ldots, \ h_m(x)]$. In Chapter 2, we will explore a number of methods to determine the convergence rate for the primal-dual algorithm and compare them.

The dual method can also be used to find the critical points of problem (2.4). The dual method is simply gradient descent of the dual function. To use the dual method, the minimizer of the

Lagrangian for a given λ needs to be determined for each time step. The dual method is given by

$$\begin{aligned}\bar{x} &= \text{argmin}_{x\in\mathbb{R}^n}\ L(x,\lambda^k)\\ \lambda^{k+1} &= \lambda^k + \beta h(\bar{x}),\end{aligned}$$

for $\beta > 0$. The dual method works well when $\bar{x}$ can be calculated easily and can provide significant speed increases compared to the primal-dual method. For problem (2.4), when $f \in S(m, L)$ and A is full row-rank, the dual method can be implemented as

$$\begin{aligned}\bar{x} &= \nabla f^{-1}(-A^\top \lambda^k)\\ \lambda^{k+1} &= \lambda^k + \beta\left(A\bar{x} - b\right).\end{aligned}$$

The assumption $f \in S(m, L)$ guarantees that ∇f^{-1} exists [12]. Further, since A is full row rank, $-A\nabla f^{-1}(-A^\top x)$ is $\frac{\sigma^2_{\min}}{L_f}$-strongly monotone and $\frac{\sigma^2_{\max}(A)}{m_f}$-Lipschitz continuous [12]. Using Corollary 2.5.4.2, we know that this algorithm converges for step-size

$$\beta \leq \frac{2}{\frac{\sigma^2_{\max}(A)}{m_f} + \frac{\sigma^2_{\min}(A)}{L_f}} = \frac{2m_f L_f}{L_f\sigma^2_{\max}(A) + m_f\sigma^2_{\min}(A)} \tag{2.11}$$

with rate $(1 - \beta\frac{\sigma^2_{\min}(A)}{L_f})$.

The final method considered will be used for non-differentiable cost functions. Non-differentiable cost functions introduce challenges when using gradient methods because their gradient is not well-defined over $\mathbb{R}^n$. Some functions that can be useful in optimization but suffer from a lack of differentiablity are the indicator function and l_1 norm. The indicator function for a set $C \subset \mathbb{R}^n$, $I_C(x) : \mathbb{R}^n \to \bar{\mathbb{R}}$, is given by

$$I_C(x) = \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}.$$

The indicator function can be used to enforce hard constraints on the problem and ensure the solution is within the set C. The l_1 norm, $||x||_1 : \mathbb{R}^n \to \mathbb{R}$ is given by

$$||x||_1 = \sum_{i=1}^{n} |x_i|.$$

The l_1 norm is used in optimization problems to encourage sparsity in the variable. The weighted l_1 norm can also be considered which allows the user to assign different weights to each element in the vector,

$$||x||_{a,1} = \sum_{i=1}^{n} |a_i||x_i|, \quad a \in \mathbb{R}^n.$$

Before proposing the problem and algorithm, the proximal operator and Moreau envelope will be defined. To do this, the notion of a closed proper convex function needs to be understood.

Definition 2.5.5. *[23] Consider a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$. The function f is a **closed proper convex function** if the set,*

$$C = \{(x, t) \in \mathbb{R}^{n+1} \mid f(x) \leq t\}$$

is nonempty, convex, and closed.

The set C in the definition is known as the epigraph of f [16, Section 3.1.7]. Also, note that a set C is closed if its complement is open. With the necessary background defined, the proximal operator and Moreau envelope can be defined.

Definition 2.5.6. *[23] Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a closed proper convex function. The **proximal operator** of f, $\boldsymbol{prox}_{\lambda f} : \mathbb{R}^n \to \mathbb{R}^n$, at $u \in \mathbb{R}^n$ for $\lambda > 0$ is defined as*

$$\boldsymbol{prox}_{\lambda f}(u) := \text{argmin}_{x \in \mathbb{R}^n} \left(f(x) + \tfrac{1}{2\lambda} ||x - u||_2^2 \right).$$

*Further, the **Moreau envelope** of f, $M_{\lambda f}$, at $u \in \mathbb{R}^n$ for $\lambda > 0$ is defined as*

$$M_{\lambda f}(u) := \inf_{x \in \mathbb{R}^n} \left(f(x) + \tfrac{1}{2\lambda} ||x - u||_2^2 \right). \tag{2.12}$$

For non-differentiable functions, the Moreau envelope acts as a smoothed version of the function [23]. The Moreau envelope is also convex, continuous, and differentiable on all of $\mathbb{R}^n$ [23]. The connection between the proximal operator and Moreau envelope becomes apparent when looking at the following equality

$$M_{\lambda f}(u) = f(\mathbf{prox}_{\lambda f}(u)) + \tfrac{1}{2\lambda} ||u - \mathbf{prox}_{\lambda f}(u)||_2^2. \tag{2.13}$$

Further, the gradient of the Moreau envelope is given by $\nabla M_{\lambda f}(u) = \frac{1}{\lambda} \left(u - \mathbf{prox}_{\lambda f}(u) \right)$ [23].

Consider the problem of finding $x \in \mathbb{R}^n$ such that

$$\min_{x \in \mathbb{R}} f(x) + g(x), \tag{2.14}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and satisfies $f \in S(m, L)$ and $g : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a closed proper convex function. Then the problem can be solved using the proximal gradient method

$$x^{k+1} = \mathbf{prox}_{\alpha g}(x^k - \alpha \nabla f(x^k)), \tag{2.15}$$

for some $\alpha > 0$ [18]. The proximal gradient method allows for the minimization of cost functions that consist of both differentiable and non-differentiable components. Proximal gradient descent is especially effective if the proximal operator of g is easy to compute. A number of useful non-differentiable functions such as the indicator function and l_1 norm have easily computed proximal operators. The convergence results of proximal methods will be explored further in Chapter 4.3.

Chapter 3

Primal-Dual Algorithm Rate Comparison

In the last chapter, we provided a brief background of the necessary topics required for the thesis. Gradient descent and the primal-dual algorithm were introduced as methods to solve unconstrained and constrained optimization problems. Consider a convex equality-constrained optimization problem

$$\begin{aligned} &\min_{x\in\mathbb{R}^n} \ f(x) \\ &\text{s.t. } Ax = b, \end{aligned} \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is unknown, differentiable, and $f \in S(m_f, L_f)$. Further, assume $A \in \mathbb{R}^{m\times n}$ has full-row rank and there exists $x \in \mathbb{R}^n$ such that $Ax = b$. By assuming that $f \in S(m_f, L_f)$ and the feasible set is non-empty, we know that the minimizer exists and is unique [16]. By assuming A is full row-rank, we will be able to use the fact that $AA^\top \succ 0$ in the analysis. With these assumptions, the primal-dual algorithm for (3.1) is given by

$$\begin{aligned} x^{k+1} &= x^k - \alpha \left(\nabla f(x^k) + A^\top \lambda^k\right) \\ \lambda^{k+1} &= \lambda^k + \beta \left(Ax^k - b\right). \end{aligned} \tag{3.2}$$

The primal-dual method introduces unique challenges in its convergence analysis due to the lack of explicit strong convexity in the dual update. Fixing the dual variable, the primal update clearly converges exponentially to a fixed point with $\alpha \leq \frac{2}{m_f+L_f}$; however, the dual update does not converge when fixing x. The dual update is solely determined by a linear function of x. Thus, it is the primal dynamics that drive the convergence of this method.

The choice to analyze the primal-dual method for the equality-constrained optimization problem was motivated by the controller which will be discussed in the next chapter. The feedback controller takes the form of an equality-constrained problem, which is why inequality constraints were not considered. Implementing inequality constraints would add a second dual variable to the analysis that would be constrained to non-negative values. This addition would require a significant change in the analysis approach.

In this chapter, three different approaches will be used to analyze the convergence rate of algorithm (3.2). The first method will determine the contraction rate of the algorithm with respect to different norms. Three different norms of the form $||\cdot||_{P,2}$, where $P \succ 0$, will be considered. We will then determine if algorithm (3.2) is contractive with respect to the norm and the rate at which it contracts. The second approach will involve restructuring algorithm (3.2) as the interconnection of two exponentially convergent systems. A Lyapunov function for each system will be constructed and combined to determine the convergence of the connected system. This approach simplifies to the rate analysis performed in [12]. Finally, a robust control approach adapted from [10] will be used. In this approach, a linear matrix inequality is constructed. If the linear matrix inequality is feasible for a given rate, ρ, and step size, we are able to conclude that the algorithm contracts in some norm with rate ρ. A modification of the primal-dual algorithm will also be proposed and analyzed. The modification will introduce a tunable time-scale separation between the primal and dual updates. A rate for the system will be guaranteed.

The rates for each analysis of the original primal-dual algorithm will be compared. These rates will be compared as m_f, L_f, $\sigma_{\min}(A)$, and $\sigma_{\max}(A)$ are varied.

3.1 Contractive Approach

In this section, we will determine the convergence of algorithm (3.2) using a contractive approach. Specifically, we will be analyzing (3.2) using a number of different norms of the form $||\cdot||_{P,2}$, where $P \succ 0$. For each norm matrix, P, a contraction rate will be determined. From Chapter 1, we know that if algorithm (3.2) is contractive with rate ρ, then it is exponentially convergent to a unique equilibrium with rate ρ. Further, proving an algorithm is contractive with rate ρ with respect to a norm $||\cdot||_{P,2}$ immediately provides a Lyapunov function for the system. Taking $V(z) = ||z - z^*||_{P,2}$, where $z^* \in \mathbb{R}^{n+m}$ is the fixed point of the contractive algorithm, we have

$$V(z^{k+1}) = ||z^{k+1} - z^*||_{P,2} \leq \rho||z^k - z^*||_{P,2} = \rho V(z^k), \tag{3.3}$$

where $z^{k+1}, z^k \in \mathbb{R}^{n+m}$ are generated by the algorithm of interest. Thus, showing that (3.2) is contractive provides a Lyapunov function and rate of convergence. To determine that the algorithm is contractive for each norm, Lemma 3.1.1 will be used.

Lemma 3.1.1. *Consider the discrete-time algorithm in the form of*

$$x^{k+1} = x^k + \alpha F(x^k) \tag{3.4}$$

where $\alpha > 0$ and $F : \mathbb{R}^n \to \mathbb{R}^n$. Assume that

(i) F is L_F-Lipschitz continuous with respect to $\|\cdot\|_2$, and

(ii) there exists a scalar $b > 0$ and a positive definite matrix $P \succ 0$ such that

$$(F(x) - F(y))^\top P(x - y) \leq -b||x - y||^2_{P,2}, \quad \forall x, y \in \mathbb{R}^n. \tag{3.5}$$

Then (3.4) is contractive with respect to the norm $||\cdot||_{P,2}$ for step-size $0 < \alpha < \frac{2b}{L_F^2 \text{cond}(P)}$. Further, the step-size $\alpha = \frac{b}{\text{cond}(P)L_F^2}$ provides the lowest contraction rate $\rho = \sqrt{1 - \frac{b^2}{\text{cond}(P)L_F^2}}$.

Proof. To prove the lemma, assume F is L_F-Lipschitz continuous with respect to $||\cdot||_2$ and inequality (3.5) holds. Then for any $x^k, y^k \in \mathbb{R}^n$

$$\begin{aligned} ||x^{k+1} - y^{k+1}||_{P,2}^2 &= ||x^k + \alpha F(x^k) - (y^k + \alpha F(y^k))||_{P,2}^2 \\ &\leq ||x^k - y^k||_{P,2}^2 + \alpha^2 \text{cond}(P) L_F^2 ||x^k - y^k||_{P,2}^2 \\ &+ 2\alpha (F(x^k) - F(y^k))^\top P(x^k - y^k) \\ &\leq (1 + \alpha^2 \text{cond}(P) L_F^2 - 2\alpha b) ||x^k - y^k||_{P,2}^2. \end{aligned} \tag{3.6}$$

Thus, setting $\alpha \in \left(0, \frac{2b}{L_F^2 \text{cond}(P)}\right]$ ensures that the update is contractive. To find the step size that optimizes the contraction rate, notice that the rate is a positive quadratic function. Finding the minimum value is equivalent to finding the zero of the derivative with respect to α

$$2\alpha \text{cond}(P) L_F^2 - 2b = 0 \implies \alpha = \frac{b}{\text{cond}(P)L_F^2}. \tag{3.7}$$

Using the step size given in (3.7) with (3.6), we see that

$$||x^{k+1} - y^{k+1}||_{P,2} \leq \sqrt{1 - \frac{b^2}{\text{cond}(P)L_F^2}} ||x^k - y^k||_{P,2}.$$

This concludes the proof. □

To fit the structure of Lemma 3.1.1, algorithm (3.2) can be written as

$$\begin{bmatrix} x^{k+1} \\ \lambda^{k+1} \end{bmatrix} = \begin{bmatrix} x^k \\ \lambda^k \end{bmatrix} + \alpha \underbrace{\begin{bmatrix} -\nabla f(x^k) - A^\top \lambda^k \\ Ax^k - b \end{bmatrix}}_{:= F(x^k, \lambda^k)}, \tag{3.8}$$

where we set $\beta = \alpha$ to simplify the calculations. Lemma 3.1.2 will also be used.

Lemma 3.1.2. *[11] Suppose that a function $F : \mathbb{R}^n \to \mathbb{R}^n$ is m-strongly monotone and L-Lipschitz continuous. Then for any $x, \bar{x} \in \mathbb{R}^n$, there exists a matrix $B_{x,\bar{x}}$ such that $B_{x,\bar{x}}(x - \bar{x}) = (F(x) - F(\bar{x}))$ and $m\mathbf{I} \preceq B_{x,\bar{x}} \preceq L\mathbf{I}$.*

Consider $x, y \in \mathbb{R}^n$ and $\lambda, \gamma \in \mathbb{R}^m$ where $z = [x - y,\ \lambda - \gamma]^\top$. Using Lemma 3.1.2, let $Q_{x,y}(x - y) = \nabla f(x) - \nabla f(y)$. Using this notation, inequality (3.5) can be simplified to

$$\begin{aligned} &(F(x, \lambda) - F(y, \gamma))^\top P \begin{bmatrix} x - y \\ \lambda - \gamma \end{bmatrix} \leq -b||z||_{P,2}^2 \\ \iff &\begin{bmatrix} -(\nabla f(x) - \nabla f(y)) - A^\top(\lambda - \gamma) \\ A(x - y) \end{bmatrix}^\top P \begin{bmatrix} x - y \\ \lambda - \gamma \end{bmatrix} \leq -b||z||_{P,2}^2 \\ \iff &z^\top \begin{bmatrix} -Q & -A \\ A & 0 \end{bmatrix}^\top Pz \ \leq -b||z||_{P,2}^2 \end{aligned} \tag{3.9}$$

where the subscript on Q is removed for notational simplicity. We see that showing inequality (3.9) holds is equivalent to showing inequality (3.5) holds.

We will begin the analysis by using Lemma 3.1.1 with $P = \mathbf{I}$. This approach does not prove the algorithm is exponentially convergent, but it is a quick and illustrative exercise. Consider $x, y \in \mathbb{R}^n$ and $\lambda, \gamma \in \mathbb{R}^m$ where $z = [x - y,\ \lambda - \gamma]^\top$.

$$\begin{aligned} z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} z &= -(x-y)^\top Q(x-y) + (x-y)^\top A^\top(\lambda - \gamma) - (\lambda-\gamma)^\top A(x-y) \\ &= -x^\top Q x \end{aligned}$$

We see that the quadratic term is negative for all x and y, but it is only a product of the primal variable. All of the strong monotonicity is captured in the Q matrix, with none of the convergence being expressed in the λ variable. Alternative norms can be used to highlight the transfer of convergence to the dual variable.

Contractive Approach using P

The first norm matrix that will be considered is given by

$$P = \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & \mathbf{I} \end{bmatrix}, \tag{3.10}$$

for $\epsilon > 0$. For P to be positive definite, $\epsilon < \frac{1}{\sigma_{\max}(A)}$ must hold. Using matrix (3.10) and $x, y \in \mathbb{R}^n$, $\lambda, \gamma \in \mathbb{R}^m$ where $z = [x - y,\ \lambda - \gamma]^\top$, inequality (3.9) is given by

$$\begin{aligned} & z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & \mathbf{I} \end{bmatrix} z \\ &= (x-y)^\top \left(-Q + \epsilon A^\top A\right)(x-y) - \epsilon (x-y)^\top Q A^\top (\lambda - \gamma) - \epsilon A A^\top ||\lambda - \gamma||_2^2 \\ &\leq -(m_f - \epsilon \sigma_{max}^2(A))||x-y||_2^2 + \epsilon L_f \sigma_{\max}(A)||\lambda-\gamma||_2 ||x-y||_2 - \epsilon \sigma_{\min}(A)||\lambda-\gamma||_2^2 \\ &= - \begin{bmatrix} ||x-y||_2 \\ ||\lambda-\gamma||_2 \end{bmatrix}^\top \underbrace{\begin{bmatrix} m_f - \epsilon \sigma_{\max}^2(A) & -\frac{\epsilon}{2} L_f \sigma_{\max}(A) \\ -\frac{\epsilon}{2} L_f \sigma_{\max}(A) & \epsilon \sigma_{\min}^2(A) \end{bmatrix}}_{:= T(\epsilon)} \begin{bmatrix} ||x-y||_2 \\ ||\lambda-\gamma||_2 \end{bmatrix}, \end{aligned}$$

where the inequality follows from using the triangle inequality and the bounds from Lemma 3.1.2. Using the Schur complement, we see that

$$\epsilon < \min\left(\frac{m_f}{\sigma_{\max}^2(A) - \frac{1}{4} L_F^2 \text{cond}(A)}, \frac{1}{\sigma_{\max}(A)} \right) \tag{3.11}$$

guarantees that $T(\epsilon)$ and P are positive definite [24, Theorem 1.12]. This allows us to bound inequality (3.5),

$$z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & \mathbf{I} \end{bmatrix} z \leq -\sigma_{\min}(T(\epsilon))||z||_2^2 \leq -\frac{\sigma_{\min}(T(\epsilon))}{\sigma_{\max}(P)} ||z||_{P,2}^2.$$

Therefore, Inequality 3.5 is satisfied and Lemma 3.1.1 guarantees the following convergence rate

$$\rho = \sqrt{1 - \frac{\sigma_{\min}^2(T(\epsilon))}{\sigma_{\max}^2(P)\text{cond}(P)L_F^2}},$$

for $\alpha = \frac{\sigma_{\min}(T(\epsilon))}{\sigma_{\max}(P)\text{cond}(P)L_F^2}$. In the matrix $T(\epsilon)$, we can see that the dual variable's convergence is characterized by $\epsilon AA^\top$. Because ϵ is relatively small to keep P positive definite and $T(\epsilon)$ negative definite, the rate at which λ is decreasing is small. It is clear that the minimum eigenvalue of $T(\epsilon)$ is directly linked to the value of ϵ. Because the rate is a function of the eigenvalues of $T(\epsilon)$ and P, which both depend on ϵ, a computational approach is used to find the optimal ϵ for a given problem. This approach is discussed further in the results section.

Contractive Approach using $\tilde{P}$

The same method is repeated using a slightly different matrix.

$$\tilde{P} = \begin{bmatrix} \mathbf{I} & \epsilon A^\top (AA^\top)^{-1} \\ \epsilon (AA^\top)^{-1} A & \mathbf{I} \end{bmatrix} \tag{3.12}$$

where $\epsilon > 0$. We see that $\epsilon < \sigma_{\min}(A)$ ensures the positive definiteness of (3.12). The $(AA^\top)^{-1}$ term was added to cancel with some of the terms present in the last analysis and change the feasible ϵ values. Consider $x, y \in \mathbb{R}^n$, $\lambda, \gamma \in \mathbb{R}^m$, and $z = [x - y,\ \lambda - \gamma]^\top$, then (3.9) can be bounded by,

$$\begin{aligned}
&z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top (AA^\top)^{-1} \\ \epsilon (AA^\top)^{-1} A & \mathbf{I} \end{bmatrix} z \\
&= -(x-y)^\top \left(Q - \epsilon A^\top (AA^\top)^{-1} A\right)(x-y) - \epsilon (x-y)^\top Q A^\top (AA^\top)^{-1}(\lambda - \gamma) - \epsilon ||\lambda - \gamma||_2^2 \\
&\leq -(m_f - \epsilon)||x-y||_2^2 + \epsilon \tfrac{L_f \text{cond}(A)}{\sigma_{\min}(A)} ||\lambda - \gamma||_2 ||x-y||_2 - \epsilon ||\lambda - \gamma||_2^2 \\
&\leq -z^\top \underbrace{\begin{bmatrix} m_f - \epsilon & -\frac{\epsilon}{2}\frac{L_f \text{cond}(A)}{\sigma_{\min}(A)} \\ -\frac{\epsilon}{2}\frac{L_f \text{cond}(A)}{\sigma_{\min}(A)} & \epsilon \end{bmatrix}}_{:=\tilde{T}(\epsilon)} z
\end{aligned}$$

Again, using the Schur complement we can determine that ϵ must satisfy

$$\epsilon < \min\left(\sigma_{\min}(A), \tfrac{4 m_f \sigma_{\min}^2(A)}{4\sigma_{\min}^2 + L_F^2 \text{cond}(A)^2}\right) \tag{3.13}$$

so that $\tilde{T}(\epsilon)$ and P are positive definite [24, Theorem 1.12]. For ϵ satisfying (3.13), inequality (3.9) can be bounded by

$$z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top (AA^\top)^{-1} \\ \epsilon (AA^\top)^{-1} A & \mathbf{I} \end{bmatrix} z \leq -\sigma_{\min}(\tilde{T}(\epsilon))||z||_2^2 \leq -\frac{\sigma_{\min}(\tilde{T}(\epsilon))}{\sigma_{\max}(\tilde{P})} ||z||_{\tilde{P},2}^2.$$

Using Lemma 3.1.1 the convergence rate

$$\rho = \sqrt{1 - \frac{\sigma_{\min}^2(\tilde{T}(\epsilon))}{\sigma_{\max}^2(\tilde{P})\text{cond}(\tilde{P})L_F^2}}$$

is guaranteed for $\alpha = \frac{\sigma_{\min}(\tilde{T}(\epsilon))}{\sigma_{\max}(\tilde{P})\text{cond}(\tilde{P})L_F^2}$. Again, the optimal ϵ value can be determined computationally.

Contractive Approach using $\hat{P}$

For the matrices P and $\tilde{P}$, the value cond(P) is kept relatively low. This is because ϵ is relatively small and appears on the off diagonal of the matrix. Therefore, the eigenvalues of P remain close to 1. The next approach will consider a matrix that allows ϵ to have more of an effect on cond(P). The matrix considered is given by

$$\hat{P} = \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & (1+\epsilon m_f)\mathbf{I} \end{bmatrix} \tag{3.14}$$

To keep (3.14) positive definite, it is required that $\epsilon < \frac{\sqrt{1+m_f}}{\sigma_{\max}(A)}$ is satisfied. Using matrix (3.14), $x, y \in \mathbb{R}^n$, $\lambda, \gamma \in \mathbb{R}^m$, and $z = [x-y,\ \lambda-\gamma]^\top$, inequality (3.9) is given by

$$\begin{aligned}
&z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & (1+\epsilon m_f)\mathbf{I} \end{bmatrix} z \\
&= -(x-y)^\top \left(Q - \epsilon A^\top A\right)(x-y) - \epsilon(x-y)^\top (Q - m_f)A^\top(\lambda-\gamma) - \epsilon(\lambda-\gamma)^\top AA^\top(\lambda-\gamma) \\
&\leq -(m_f - \epsilon\sigma_{\max}^2(A))||x-y||_2^2 + (L_f - m_f)||x-y||_2||\lambda-\gamma||_2 - \epsilon\sigma_{\min}^2(A)||\lambda-\gamma||_2^2 \\
&= -\begin{bmatrix} ||x-y||_2 \\ ||\lambda-\gamma||_2 \end{bmatrix} \underbrace{\begin{bmatrix} m_f - \epsilon\sigma_{\max}^2 & \frac{\epsilon}{2}(L_f - m_f)\sigma_{\max}(A) \\ \frac{\epsilon}{2}(L_f - m_f)\sigma_{\max}(A) & \epsilon\sigma_{\min}^2(A) \end{bmatrix}}_{:=\hat{T}(\epsilon)} \begin{bmatrix} ||x-y||_2 \\ ||\lambda-\gamma||_2 \end{bmatrix}
\end{aligned}$$

Using the Schur complement, we see that

$$\epsilon < \min\left(\frac{\sqrt{1+m_f}}{\sigma_{\max}(A)}, \frac{4m_f}{4\sigma_{\max}(A) + (L_f - m_f)^2\text{cond}(A)^2}\right) \tag{3.15}$$

guarantees $\hat{T}(\epsilon)$ and $\hat{P}$ are positive definite. For epsilon satisfying (3.15), inequality (3.9) can be bounded by

$$z^\top \begin{bmatrix} -Q & A^\top \\ -A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \epsilon A^\top \\ \epsilon A & (1+\epsilon m_f)\mathbf{I} \end{bmatrix} z \leq -\sigma_{\min}(\hat{T}(\epsilon))||z||_2^2 \leq \frac{\sigma(\hat{T}(\epsilon))}{\sigma_{\max}(\tilde{P})}. \tag{3.16}$$

Using Lemma 3.1.1 with the bound described in (3.16), algorithm (3.2) is contractive with the rate

$$\rho = \sqrt{1 - \frac{\sigma_{\min}^2(\hat{T}(\epsilon))}{\sigma_{\max}^2(\hat{P})\text{cond}(\hat{P})L_F^2}},$$

for $\alpha = \frac{\sigma_{\min}(\hat{T}(\epsilon))}{\sigma_{\max}(\hat{P})\text{cond}(\hat{P})L_F^2}$. The optimal ϵ value can be computed computationally.

3.2 Interconnected Systems Approach

The Lyapunov approach yields results that confirm convergence; however, suggesting and trying different Lyapunov functions can feel like a guessing game. In this approach, we look at the dual and primal update as the interconnection of two exponentially convergent systems. Figure 3.1 is a block diagram representing this approach. There are different ways to derive the interconnection,

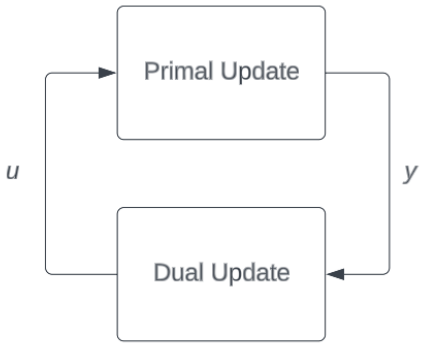

Figure 3.1: Block diagram for the suggested approach.

but we will use the dual method to motivate it. For problem (3.1), the dual method is given by

$$\begin{aligned} \bar{x} &= \operatorname{argmin}_{x\in\mathbb{R}^n} L(x,\lambda^k) \\ \lambda^{k+1} &= \lambda^k + \beta(A\bar{x} - b). \end{aligned}$$

Since $f \in S(m_f, L_f)$, we know that ∇f^{-1} exists [12], so we can use the equivalent formulation proposed in Section 2.5.2,

$$\lambda^{k+1} = \lambda^k + \beta(A\nabla f^{-1}(-A^\top\lambda^k) - b). \tag{3.17}$$

Using Corollary 2.5.4.2, we know that this algorithm contracts with rate $1-\beta\frac{\sigma^2_{\min}(A)}{L_f}$ with an optimal step-size of $\beta = \frac{2m_fL_f}{L_f\sigma_{\max}(A)+m_f\sigma_{\min}(A)}$. With the desire to inject strong convergence properties into the dual update, we decide to look at the λ update in algorithm (3.2) as a perturbation of algorithm (3.17),

$$\begin{aligned} \lambda^{k+1} &= \lambda^k + \beta(A\nabla f^{-1}(-A^\top\lambda^k) - b) + \beta Ay^k \\ y^k &= x^k - \nabla f^{-1}(-A^\top\lambda^k). \end{aligned}$$

Notice that this system simplifies to the dual update for algorithm (3.2), and when $y^k = 0$, the iteration simplifies to algorithm (3.17). The primal update can be seen as a perturbed gradient descent algorithm for an unconstrained optimization problem.

$$\begin{aligned} x^{k+1} &= x^k - \alpha\nabla f(x^k) - \alpha u^k \\ u^k &= A^\top\lambda^k \end{aligned}$$

Again, we see that this is equivalent to the primal update for algorithm (3.2). We will call the combination of these two systems the Interconnected System (IS) representation for (3.2).

$$\begin{aligned} x^{k+1} &= x^k - \alpha\left(\nabla f(x^k)\right) - \alpha u^k \\ y^k &= x^k - \nabla f^{-1}(-u^k) \\ \lambda^{k+1} &= \lambda^k + \beta\left(A\nabla f^{-1}(-A^\top\lambda^k) - b\right) + \beta A y^k \\ u^k &= A^\top\lambda^k \end{aligned} \tag{3.18}$$

The block diagram of the system is represented in Figure 3.2. With the system rearranged into IS

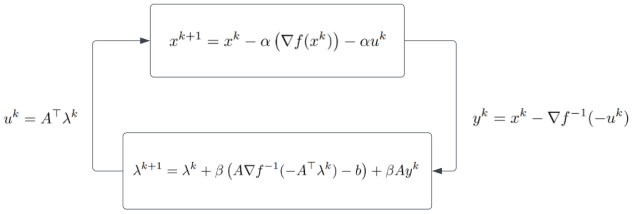

Figure 3.2: Block diagram for the IS representation for (3.2).

representation, the main theorem of this section can be stated. The proof approach is adapted from [12]. The IS representation provides a new intuition for the solution.

Theorem 3.2.1 (Convergence of the Primal-Dual Algorithm using IS Representation)**.** *Consider algorithm (3.18). Suppose that f is differentiable, satisfies $f \in S(m_f, L_f)$, and $A \in \mathbb{R}^{m\times n}$ is full-row rank. Then the following inequality is satisfied*

$$||y^{k+1}||_2 + \omega||\lambda^{k+1} - \lambda^*||_2 \leq \left(1 - \frac{1}{12\kappa^3(f)\text{cond}^4(A)}\right)\left(||y^k||_2 + \omega||\lambda^k - \lambda^*||_2\right) \tag{3.19}$$

for

$$\begin{aligned} \omega &= \frac{2L_f\sigma_{\max}^3(A)}{m_f^2\sigma_{\min}(A)} \\ \alpha &= \frac{2}{L_f+m_f} \\ \beta &= \frac{m_f}{(m_f+L_f)\left(\frac{\sigma_{\max}^2(A)}{m_f}+\omega\sigma_{\max}(A)\right)}. \end{aligned} \tag{3.20}$$

Proof. To prove the theorem, a Lyapunov approach will be used. The two systems will be considered separately, and then combined to determine the behaviour of the interconnected system. We begin

by considering the primal update. The Lyapunov function $V_y(y) = ||y||_2$ will be used.

$$\begin{aligned}
||y^{k+1}||_2 - ||y^k||_2 =& ||x^{k+1} - \nabla f^{-1}(-u^{k+1})||_2 - ||y^k||_2 \\
\leq& ||x^{k+1} - \nabla f^{-1}(-u^k)||_2 + ||\nabla f^{-1}(-u^{k+1}) - \nabla f^{-1}(-u^k)||_2 - ||y^k||_2 \\
\leq& ||x^k - \nabla f^{-1}(-u^k) - \alpha \left(\nabla f(x^k) - \nabla f(\nabla f^{-1}(-u^k))\right)||_2 \\
&+ \frac{\sigma_{\max}(A)}{m_f}||\lambda^{k+1} - \lambda^k||_2 - ||y^k|| \\
\leq& -\alpha m_f||y^k|| + \beta\frac{\sigma_{\max}(A)}{m_f}||A\nabla f^{-1}(-A^\top\lambda) - b + Ay^k|| \\
\leq& \left(\beta\frac{\sigma^2_{\max}(A)}{m_f} - \alpha m_f\right)||y^k|| + \beta\frac{\sigma^3_{\max}(A)}{m_f^2}||\lambda^k - \lambda^*||_2
\end{aligned} \tag{3.21}$$

for $\alpha \leq \frac{2}{m_f + L_f}$ due to Corollary 2.5.4.2. The first inequality is calculated by adding and subtracting $\nabla f(-u^k)$ and using the triangle inequality. Clearly, for small β, the Lyapunov function is strictly negative and the system is exponentially convergent with a rate that approaches $1 - \alpha m_f$. For the dual system, we consider the Lyapunov function $V_\lambda(\lambda) = ||\lambda - \lambda^*||_2$,

$$\begin{aligned}
||\lambda^{k+1} - \lambda^*||_2 - ||\lambda^k - \lambda^*||_2 =& ||\lambda^k - \lambda^* + \beta(A\nabla f^{-1}(-A^\top\lambda^k) - A\nabla f^{-1}(-A^\top\lambda^*)) + \beta Ay^k||_2 \\
&- ||\lambda^k - \lambda^*||_2 \\
\leq& ||\lambda^k - \lambda^* + \beta(A\nabla f^{-1}(-A^\top\lambda^k) - A\nabla f^{-1}(-A^\top\lambda^*))||_2 \\
&+ \beta\sigma_{\max}(A)||y^k||_2 - ||\lambda^k - \lambda^*||_2 \\
\leq& -\beta\frac{\sigma^2_{\max}}{L_f}||\lambda^k - \lambda^*||_2 + \beta\sigma_{\max}(A)||y^k||_2
\end{aligned} \tag{3.22}$$

for $\beta < \frac{2m_f L_f}{L_f\sigma_{\max}(A) + m_f\sigma_{\min}(A)}$ due to Corollary 2.5.4.2. Adding together (3.21) and (3.22) with $\omega > 0$, we get

$$\begin{aligned}
&||y^{k+1}||_2 - ||y^k||_2 + \omega(||\lambda^{k+1} - \lambda^*||_2 - ||\lambda^k - \lambda^*||_2) \\
&\leq \left(\beta\frac{\sigma^2_{\max}(A)}{m_f} - \alpha m_f\right)||y^k|| + \beta\frac{\sigma^3_{\max}(A)}{m_f^2}||\lambda^k - \lambda^*||_2 \\
&+ \omega\left(-\beta\frac{\sigma^2_{\max}}{L_f}||\lambda^k - \lambda^*||_2 + \beta\sigma_{\max}(A)||y_k||_2\right) \\
&= \left(\beta\sigma_{\max}(A)\left(\frac{\sigma_{\max}(A)}{m_f} + \omega\right) - \alpha m_f\right)||y^k||_2 + \beta\left(\frac{\sigma^3_{\max}(A)}{m_f^2} - \omega\frac{\sigma^2_{\max}}{L_f}\right)||\lambda^k - \lambda^*||_2.
\end{aligned} \tag{3.23}$$

For small β and properly selected ω, the Lyapunov function is negative. From [12], using the step sizes and ω given in (3.20), inequality (3.23) simplifies to,

$$\begin{aligned}
&||y^{k+1}||_2 - ||y^k||_2 + \omega(||\lambda^{k+1} - \lambda^*||_2 - ||\lambda^k - \lambda^*||_2) \leq (\tfrac{1}{12\kappa^3(f)\text{cond}^4(A)})(||y^k||_2 + \omega||\lambda^k||_2) \\
\iff& ||y^{k+1}||_2 + \omega||\lambda^{k+1} - \lambda^*||_2 \leq (1 - \tfrac{1}{12\kappa^3(f)\text{cond}^4(A)})(||y^k||_2 + \omega||\lambda^k||_2).
\end{aligned}$$

This is exactly the rate given in (3.19). □

From Theorem 3.2.1, we see that using the Lyapunov function $V(x, \lambda) = ||x - \nabla f^{-1}(-A^\top\lambda)||_2 +$

$\omega||\lambda - \lambda^*||$ for (3.2) ensures

$$V(x^{k+1}, \lambda^{k+1}) \leq \rho V(x^k, \lambda^k) \quad \forall\, k \in \mathbf{Z}_{\geq 0}, \tag{3.24}$$

where

$$\rho = 1 - \tfrac{1}{12\kappa^3(f)\text{cond}^4(A)}.$$

3.3 IQC Approach

The following approach uses a method described in [10]. The method uses techniques from robust control theory to determine the convergence of discrete-time optimization algorithms. Rather than forming a Lyapunov function, this approach involves determining the feasibility of a linear matrix inequality (LMI). If the LMI is feasible for a given rate ρ, then it guarantees there exists $P \succ 0$ such that (3.2) is contractive with rate ρ with respect to $||\cdot||_{P,2}$. Integral Quadratic Constraints (IQCs) are used in the formulation to bound the input and output behaviour. Before introducing the main theorem, a lemma will be introduced that allows for an easy transition from the continuity and convexity assumptions we have placed on our cost function to the IQC framework.

Lemma 3.3.1. *[10] Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies $f \in S(m, L)$, then for all $x, y \in \mathbb{R}^n$*

$$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^\top \begin{bmatrix} -2mL\mathbf{I}_n & (L+m)\mathbf{I}_n \\ (L+m)\mathbf{I}_n & -2\mathbf{I}_n \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$

The matrix in Lemma 3.3.1 is known as a pointwise IQC. To use this approach, the non-linearities of the algorithm need to be considered as a disturbance to an LTI system and bounded using a pointwise IQC. Thus, the algorithm has to be restructured again. The Perturbed LTI (PLTI) representation of the primal-dual algorithm is given by

$$\begin{aligned} z^{k+1} &= \overbrace{\begin{bmatrix} (1 - \alpha m_f)\mathbf{I}_n & -\alpha A^\top \\ \alpha A & \mathbf{I}_m \end{bmatrix}}^{G} z^k + \overbrace{\begin{bmatrix} -1 \\ 0 \end{bmatrix}}^{B} u^k \\ u^k &= \Delta(z^k) \\ \Delta(z^k) &= \nabla f(z_1^k) - m_f z_1^k, \end{aligned} \tag{3.25}$$

where $z^k = [x^k,\ \lambda^k]^\top$ and $\Delta : \mathbb{R}^{n+m} \to \mathbb{R}^n$. This representation removes the non-linearity from the system equations and stores it in the input. We subtract $m_f z_1^k$ from $\nabla f(x^k)$ in the input so that we have $\Delta \in S(0, L_f - m_f)$. This simplifies the quadratic constraint by canceling the top term. Using Lemma 3.3.1, the pointwise IQC for the PLTI representation is given by

$$\begin{bmatrix} z^k - y^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix}^\top \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{0}_{n\times m} & (L_f - m_f)\mathbf{I}_n \\ \mathbf{0}_{m\times n} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times n} \\ (L_f - m_f)\mathbf{I}_n & \mathbf{0}_{n\times m} & -2\mathbf{I}_n \end{bmatrix} \begin{bmatrix} z^k - y^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix} \geq 0, \quad \forall z^k, y^k \in \mathbb{R}^{n+m}.$$

With this in place, we can state the main theorem of this section.

Theorem 3.3.2. *Consider the algorithm (3.25). Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differen-*

tiable and $f \in S(m_f, L_f)$*. If there exists* $P \succ 0$ *and* $\gamma \geq 0$ *such that*

$$\begin{bmatrix} G^\top PG - \rho^2 P & G^\top PB \\ B^\top PG & B^\top PB \end{bmatrix} + \gamma \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{0}_{n\times m} & (L_f - m_f)\mathbf{I}_n \\ \mathbf{0}_{m\times n} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times n} \\ (L_f - m_f)\mathbf{I}_n & \mathbf{0}_{n\times m} & -2\mathbf{I}_n \end{bmatrix} \preceq 0 \tag{3.26}$$

then for all trajectories generated by (3.25)*,* $z^k, y^k \in \mathbb{R}^n$*, for* $k \in \mathbf{Z}_{\geq 0}$*, we have*

$$||z^{k+1} - y^{k+1}||_{P,2} \leq \rho ||z^k - y^k||_{P,2}. \tag{3.27}$$

Proof. The proof approach is an adaptation of the proof from [10]. This proof is specifically modified for the pointwise IQCs defined in Lemma 3.3.1. This allows us to guarantee the algorithm is contractive rather than just exponentially convergent.

Let $z^k, y^k \in \mathbb{R}^{n+m}$ be any two points generated by (3.25). Multipling the left and right hand sides of (3.26) by $[z^k - y^k,\ \Delta(z^k) - \Delta(y^k)]$, (3.26) simplifies to the following inequality

$$\begin{aligned} &(z^{k+1} - y^{k+1})^\top P(z^{k+1} - y^{k+1}) - \rho^2 (z^k - y^k)^\top P(z^k - y^k) \\ &+ \gamma \begin{bmatrix} z^k - y^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix}^\top \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{0}_{n\times m} & (L_f - m_f)\mathbf{I}_n \\ \mathbf{0}_{m\times n} & \mathbf{0}_{m\times m} & \mathbf{0}_{m\times n} \\ (L_f - m_f)\mathbf{I}_n & \mathbf{0}_{n\times m} & -2\mathbf{I}_n \end{bmatrix} \begin{bmatrix} z^k - y^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix} \leq 0. \end{aligned} \tag{3.28}$$

Using Lemma 3.3.1, we know that the final quadratic term is positive for all $z^k, y^k \in \mathbb{R}^{n+m}$. Thus, we are able to remove it from (3.28). With the quadratic term removed, (3.28) can be written as

$$\begin{aligned} &||z^{k+1} - y^{k+1}||_{P,2}^2 \leq \rho^2 ||z^k - y^k||_{P,2}^2 \\ \implies &||z^{k+1} - y^{k+1}||_{P,2} \leq \rho ||z^k - y^k||_{P,2} \end{aligned}$$

□

The LMI approach can be powerful as it enables you to find the optimal convergence rate for a given step size. To do this, determine if the LMI is feasible for a certain rate, and if it is, lower the rate and repeat until it is infeasible. This method allows the user to triangulate the best convergence rate. The downside of this approach is that it guarantees convergence for a single α value rather than a range of potential step sizes.

3.4 Time-Scale Separated Primal-Dual Algorithm

The previous approaches are all effective in proving the convergence of the original primal-dual algorithm. In this section, we aim to improve the results by adjusting the algorithm. The motivation for this adjustment, again, comes from the dual method, (3.17). The calculation of $\bar{x}$ in (3.17) involves finding the optimal x to minimize $f(x) + \lambda^\top (Ax - b)$ for some λ. Instead of using the approach, $\bar{x} = \nabla f^{-1}(-A^\top \lambda)$, gradient descent can be used to find the minimum value. Thinking of (3.17) in this way, we can represent the algorithm as a two-step process. For $\bar{x} = \text{argmin}_{x\in\mathbb{R}^n}\ L(x, \lambda)$

and $\delta > 0$:

$$\begin{aligned}
&\textbf{Step 1}\text{: Iterate over } n \text{ until at time } n_f \text{ we have } ||x^{n_f} - \bar{x}||_2 \leq \delta \\
&x^{n+1} = x^n - \alpha\left(\nabla f(x^n) + A^\top \lambda^k\right) \\
&\textbf{Step 2}\text{: Update the dual variable} \\
&\lambda^{k+1} = \lambda^k + \beta\left(Ax^{n_f} - b\right).
\end{aligned} \tag{3.29}$$

The value δ is small and determines the stopping criterion for the first step. As $\delta \to 0$, we have $x^{n_f} \to \bar{x}$ and $n_f \to \infty$. Thus, as $\delta \to 0$, (3.29) is equivalent to (3.17). When written as an explicit two-step process, we see that (3.17) takes the form of (3.2), with the primal update running infinitely faster than the dual update. An equivalent formulation can be written using IS representation.

$$\begin{aligned}
&\textbf{Step 1}\text{: Iterate over } n \text{ until at time } n_f \text{ we have } ||y||_2 \leq \delta \\
&x^{n+1} = x^n - \alpha\left(\nabla f(x^n) + u^k\right) \\
&y^n = x^n - \nabla f^{-1}(-u^k) \\
&\textbf{Step 2}\text{: Update the dual variable} \\
&\lambda^{k+1} = \lambda^k + \beta\left(A\nabla f^{-1}(-A^\top \lambda^k) - b\right) + \beta A y^{n_f} \\
&u^k = A^\top \lambda^k
\end{aligned}$$

This sets up the motivation for the new algorithm. For the IS representation, if a tunable time-scale separation is introduced between the two systems, we should be able to adjust it to achieve a rate that approaches $1 - \beta\frac{\sigma_{\min}^2(A)}{L_f}$ where $\beta = \frac{2m_f L_f}{m_f\sigma_{\min}^2(A) + L_f\sigma_{\max}^2}$. When the equations are operating at the same speed, the rate should be equivalent to the rate guaranteed in Theorem 3.2.1. With this motivation in mind, we introduce the time-scale separated primal-dual algorithm (TSSPD).

$$\begin{aligned}
x^{k+1} &= x^k - \alpha\nabla f(x^k) - \alpha u^k \\
y^k &= x^k - \nabla f^{-1}(-u^k) \\
\lambda^{k+1} &= \begin{cases} \lambda^k + \beta\left(A\nabla f^{-1}(-A^\top\lambda^k) - b\right) + \beta A y^k, & k \bmod n = 0 \\ \lambda^k, & \text{otherwise} \end{cases} \\
u^k &= A^\top \lambda^k
\end{aligned} \tag{3.30}$$

for some $n \in \mathbf{Z}_{\geq 1}$. We see that as n increases, the primal update gets executed more often than the dual update. When $n = 1$, (3.30) is equivalent to the primal-dual algorithm.

Theorem 3.4.1. *Consider Algorithm (3.30). Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ where f is differentiable and $f \in S(m_f, L_f)$. Further, assume that A is full-row rank, $n \in \mathbf{Z}_{\geq 1}$, and $t \in \mathbf{Z}_{>0}$, then the algorithm satisfies the following inequality*

$$\begin{aligned}
&a^{n(t+1)} + \omega b^{n(t+1)} \\
&\leq \min\left(\epsilon(1 - \alpha m_f) + (1-\epsilon)\left(1 - \beta\frac{\sigma_{\min}^2}{L_f}\right), 1 - \frac{(1-\epsilon)m_f\sigma_{\min}^4 + \epsilon L_f\sigma_{\min}^2\sigma_{\max}^2}{(1-\epsilon)L_f m_f\sigma_{\min}^2 + 2\epsilon\sigma_{\max}^2 L_f^2}\beta\right)\left(a^{nt} + \omega b^{nt}\right),
\end{aligned} \tag{3.31}$$

where $a^{n(t+1)} = ||y^{n(t+1)}||_2$, $b^{n(t+1)} = ||\lambda^{n(t+1)} - \lambda^*||_2$, $\sigma_{\max} = \sigma_{\max}(A)$, $\sigma_{\min} = \sigma_{\min}(A)$ *and*

$$\begin{aligned}
\alpha &= \tfrac{2}{m_f+L_f} \\
\beta &= \tfrac{m_f+(1-\epsilon)L_f}{(m_f+L_f)(\epsilon\frac{\sigma_{\max}^2}{m_f}+(1-\epsilon)\frac{\sigma_{\min}^2}{L_f}+\omega\sigma_{\max})} \\
\omega &= \tfrac{2\epsilon L_f\sigma_{\max}^3+(1-\epsilon)m_f\sigma_{\max}\sigma_{\min}^2}{\sigma_{\min}^2 m_f^2} \\
\epsilon &= (1-\alpha m_f)^{n-1}.
\end{aligned} \tag{3.32}$$

Further, for $n = 1$, *the inequality* (3.31) *is equivalent to the inequality in Theorem 3.2.1. Also, as* $n \to \infty$, *the inequality tends towards*

$$a^{n(t+1)} + \omega b^{n(t+1)} \leq \left(1 - \tfrac{2m_f L_f}{\sigma_{\max}L_f+\sigma_{\min}m_f}\tfrac{\sigma_{\min}^2}{2L_f}\right)\left(a^{nt} + \omega b^{nt}\right) \tag{3.33}$$

Proof. A similar proof approach as Theorem 3.2.1 will be used. Again, we consider the Lyapunov function $V_y(y) = ||y||_2$. Take $t \in \mathbf{Z}_{\geq 0}$. Because we have $u^{n(t+1)} = \ldots = u^{nt+1}$, we have

$$\begin{aligned}
||y^{n(t+1)}||_2 &= ||x^{n(t+1)} - \nabla f^{-1}(-u^{n(t+1)})||_2 \\
&= ||x^{n(t+1)-1} - \alpha\nabla f(x^{n(t+1)-1}) - \alpha u^{n(t+1)-1} - \nabla f^{-1}(-u^{n(t+1)-1})||_2 \\
&= \left|\left|x^{n(t+1)-1} - \nabla f^{-1}(-u^{n(t+1)-1}) - \alpha\left(\nabla f(x^{n(t+1)-1}) - \nabla f(-\nabla f^{-1}(-u^{n(t+1)-1}))\right)\right|\right|_2 \\
&\leq (1-\alpha m_f)\,||y^{n(t+1)-1}||_2 \\
&\leq (1-\alpha m_f)^{n-1}\,||y^{nt+1}||_2 \\
&= \epsilon||x^{nt} - \alpha(\nabla f(x^{nt}) + u^{nt}) - \nabla f^{-1}(u^{nt+1})||_2 \\
&\leq \epsilon(1-\alpha m_f)||y^{nt}||_2 + \epsilon\beta\tfrac{\sigma_{\max}}{m_f}||A\nabla f^{-1}(-A^\top\lambda^{nt}) - A\nabla f^{-1}(-A^\top\lambda^*) + Ay^{nt}||_2 \\
&\leq \epsilon(1-\alpha m_f+\beta\tfrac{\sigma_{\max}^2}{m_f})||y^{nt}||_2 + \epsilon\beta\tfrac{\sigma_{\max}^3}{m_f^2}||\lambda^{nt} - \lambda^*||_2,
\end{aligned} \tag{3.34}$$

for $\alpha < \frac{2}{m_f+L_f}$ due to Corollary 2.5.4.2. The second last inequality is a result of adding and subtracting $\nabla f^{-1}(u^{nt})$ and using the triangle inequality. For the dual variable, consider $V_\lambda(\lambda) = ||\lambda - \lambda^*||_2$. Again, because $\lambda^{n(t+1)} = \ldots = \lambda^{nt+1}$, we have

$$\begin{aligned}
||\lambda^{n(t+1)} - \lambda^*||_2 &= ||\lambda^{nt+1} - \lambda^*||_2 \\
&= ||\lambda^{nt} + \beta\left(A\nabla f^{-1}(-A^\top\lambda^{nt}) - A\nabla f^{-1}(-A^\top\lambda^*)\right) + \beta Ay^{nt}||_2 \\
&\leq (1-\beta\tfrac{\sigma_{\min}^2}{L_f})||\lambda^{nt} - \lambda^*||_2 + \beta\sigma_{\max}||y^{nt}||_2.
\end{aligned} \tag{3.35}$$

for $\beta \leq \frac{2m_f L_f}{\sigma_{\min}^2 m_f+\sigma_{\max}^2 L_f}$. Taking a linear combination of (3.34) and (3.35), we get

$$\begin{aligned}
&a^{n(t+1)} + \omega b^{n(t+1)} \\
&\leq \left(\epsilon\left(1-\alpha m_f+\beta\tfrac{\sigma_{\max}^2}{m_f}\right)+\omega\beta\sigma_{max}\right)||y^{nt}||_2 + \left(1-\beta\left(\tfrac{\sigma_{\min}^2}{L_f} - \epsilon\tfrac{\sigma_{\max}^3}{\omega m_f^2}\right)\right)\omega||\lambda^{nt} - \lambda^*||_2.
\end{aligned}$$

To show that the inequality is satisfied for the parameters (3.32), we first must show that $\alpha \leq \frac{2}{m_f+L_f}$ and $\beta \leq \frac{2m_f L_f}{\sigma_{\min}^2 m_f+\sigma_{\max}^2 L_f}$. Clearly, α satisfies this bound. To prove that β meets the requirements, note that $\epsilon \in (0, 1]$. As $\epsilon \to 0$, β is maximized. Therefore, we can bound β by setting $\epsilon = 0$. With

this bound, we see that

$$\begin{aligned}\beta &\leq \frac{m_f + L_f}{(m_f + L_f)(\frac{\sigma_{min}^2}{L_f} + \frac{\sigma_{\max}^2}{m_f})} \\ &\leq \frac{2}{\frac{\sigma_{\min}^2}{L_f} + \frac{\sigma_{\max}^2}{m_f}} \\ &= \frac{2m_f L_f}{\sigma_{\min}^2 m_f + \sigma_{\max}^2 L_f}.\end{aligned}$$

Thus, both β and α satisfy the necessary requirements. We can now analyze the term associated with a^{nt}. Using the values of β and α from (3.32), we have

$$\begin{aligned}&\left(\epsilon(1 - \alpha m_f + \beta\tfrac{\sigma_{\max}^2}{m_f}) + \omega\beta\sigma_{\max}\right) a^{nt} \\ =&\left(\epsilon(1 - \tfrac{2m_f}{m_f + L_f}) + \beta(\epsilon\tfrac{\sigma_{\max}^2}{m_f} + (1-\epsilon)\tfrac{\sigma_{\min}^2}{L_f} + \omega\sigma_{\max}) - (1-\epsilon)\beta\tfrac{\sigma_{\min}^2}{L_f}\right) a^{nt} \\ =&\left(\epsilon(1 - \tfrac{2m_f}{m_f + L_f}) + \tfrac{m_f + (1-\epsilon)L_f}{m_f + L_f} - (1-\epsilon)\beta\tfrac{\sigma_{\min}^2}{L_f}\right) a^{nt} \\ =&\left(\epsilon(1 - \tfrac{m_f}{m_f + L_f}) + \tfrac{(1-\epsilon)m_f + (1-\epsilon)L_f}{m_f + L_f} - (1-\epsilon)\beta\tfrac{\sigma_{\min}^2}{L_f}\right) a^{nt} \\ =&\left(\epsilon(1 - \tfrac{m_f}{m_f + L_f}) + (1-\epsilon)(1 - \beta\tfrac{\sigma_{\min}^2}{L_f})\right) a^{nt}.\end{aligned}$$

The first equality is achieved by adding and subtracting $(1-\epsilon)\beta\frac{\sigma_{\min}^2}{L_f}$. Because $\alpha \leq \frac{2}{m_f + L_f}$ and $\beta \leq \frac{2m_f L_f}{\sigma_{\min}^2 m_f + \sigma_{\max}^2 L_f}$ for all ϵ, both $(1 - \frac{m_f}{m_f + L_f}) < 1$ and $(1 - \beta\frac{\sigma_{\min}^2}{L_f}) < 1$. The rate guaranteed is a convex combination of these two terms. Consequently, it is always bounded above by 1. For the term in front of ωb^{nt}, we have

$$\begin{aligned}&\left(1 - \beta\left(\frac{\sigma_{\min}^2}{L_f} + \epsilon\frac{\sigma_{\max}^3}{\omega m_f^2}\right)\right)\omega b^{nt} \\ =&\left(1 - \beta\left(\frac{\sigma_{\min}^2}{L_f} + \frac{\epsilon\sigma_{\max}^3\sigma_{\min}^2}{2\epsilon L_f\sigma_{\max}^3 + (1-\epsilon)m_f\sigma_{\max}\sigma_{\min}^2}\right)\right)\omega b^{nt} \\ =&\left(1 - \beta\left(\frac{\epsilon L_f\sigma_{\min}^2\sigma_{\max}^3 + (1-\epsilon)m_f\sigma_{\max}\sigma_{\min}^4}{2\epsilon L_f^2\sigma_{\max}^3 + (1-\epsilon)L_f m_f\sigma_{\max}\sigma_{\min}^2}\right)\right)\omega b^{nt} \\ =&\left(1 - \beta\left(\frac{\epsilon L_f\sigma_{\min}^2\sigma_{\max}^2 + (1-\epsilon)m_f\sigma_{\min}^4}{2\epsilon L_f^2\sigma_{\max}^2 + (1-\epsilon)L_f m_f\sigma_{\min}^2}\right)\right)\omega b^{nt}.\end{aligned}$$

Therefore, (3.31) is proved for general n. For $n = 1$, we have $\epsilon = 1$ which simplifies to the same inequality in Theorem 3.2.1. Thus,

$$\begin{aligned}a^{n(t+1)} + \omega b^{n(t+1)} &\leq \min\left(1 - \alpha\tfrac{m_f}{2}, 1 - \beta\tfrac{\sigma_{\min}^2}{2L_f}\right)\left(a^{nt} + \omega b^{nt}\right) \\ &= \left(1 - \tfrac{1}{12\kappa^3(f)\text{cond}^4(A)}\right)\left(a^{nt} + \omega b^{nt}\right)\end{aligned}$$

where the final inequality holds from [12]. The final step is to show that (3.33) holds as $n \to \infty$. As $n \to \infty$ we have $\epsilon \to 0$. To identify the limit, consider $\epsilon = 0$. We have,

$$\begin{aligned}a^{n(t+1)} + \omega b^{n(t+1)} &\leq \min\left(1 - \beta\tfrac{\sigma_{\min}^2}{L_f}, 1 - \beta\tfrac{\sigma_{\min}^2}{L_f}\right)\left(a^{nt} + \omega b^{nt}\right) \\ &= \left(1 - \tfrac{m_f L_f}{\sigma_{\max} L_f + \sigma_{\min} m_f}\tfrac{\sigma_{\min}^2}{L_f}\right)\left(a^{nt} + \omega b^{nt}\right).\end{aligned}$$

Thus, (3.33) is satisfied. This completes the proof. □

3.5 Results

In the final section of this chapter, we compare the different analyses and the rates that they guarantee. The contractive approach, the IS approach, and the IQC approach will be compared. The TSSPD algorithm will not be compared to these rates as it runs on two different time scales; consequently, its rate is conveying different information than the others. We will begin the analysis by plotting the three different contractive rates. The IS rate will then be compared against the contractive rates. Finally, the IQC approach will be compared against the IS approach. The plots of the analyses are being staggered in this way to improve the visibility of the results. Each rate from the contractive approaches are relatively close, while the IS approach and IQC approach significantly outperform them. Because of this, separating the graphs improves legibility. As a reminder, the rate, ρ, that is being plotted represents the speed at which a Lyapunov function, V, approaches the equilibrium.

$$V(z^{k+1}) \leq \rho V(z^k)$$

To test the rates guaranteed by each approach, an unknown function $f : \mathbb{R}^3 \to \mathbb{R}$ where $f \in S(m_f, L_f)$ and constraint matrix $A \in \mathbb{R}^{2\times 3}$ where

$$A = \begin{bmatrix} \sqrt{\gamma_{\max}^2 - 1} & 0 & 1 \\ 0 & \gamma_{\min} & 0 \end{bmatrix}$$

was considered. This A matrix was used so that the singular values could be easily adjusted. We see that

$$AA^\top = \begin{bmatrix} \gamma_{\max}^2 & 0 \\ 0 & \gamma_{\min}^2 \end{bmatrix}. \tag{3.36}$$

Thus, $\sigma_{\min}(A) = \gamma_{\min}$ and $\sigma_{\max}(A) = \gamma_{\max}$.

The first test involved fixing $\gamma_{\min} = 1$ and $\gamma_{\max} = 2$ and adjusting the m_f and L_f parameters. The parameters were adjusted such that $\kappa(f) = \frac{L_f}{m_f}$ ranged from 1 to 100. The second test performed involved varying $\gamma_{\max}$ and $\gamma_{\min}$ while fixing $m_f = 1$ and $L_f = 2$. The values of $\text{cond}(A)$ were evaluated from 1 to 100 using the same process.

3.5.1 Contractive Approach

The rates guaranteed for each norm had the same structure. It involved determining the maximum and minimum eigenvalues of the norm matrices, P, $\tilde{P}$, and $\hat{P}$, and the bounding matrices, $T(\epsilon)$, $\tilde{T}(\epsilon)$, and $\hat{T}(\epsilon)$, given some ϵ value. The P matrices require specific knowledge of A, while the T matrices only require knowledge of the singular values of A. For each analysis, code was developed to iterate over 500 linearly spaced ϵ values in the allowable range that was determined in the analysis. For each $\kappa(f)$ and $\text{cond}(A)$, the code identifies the ϵ value that determines the best rate. These rates were then plotted using a logarithmic x-axis for better visibility. The results for the first test are illustrated in Figure 3.3. We see that the rate guaranteed by $\hat{P}$ provides the best results. The optimal rate was achieved when $\kappa(f) = 1$ with a value of $\rho = 0.9998$. The results for the second test

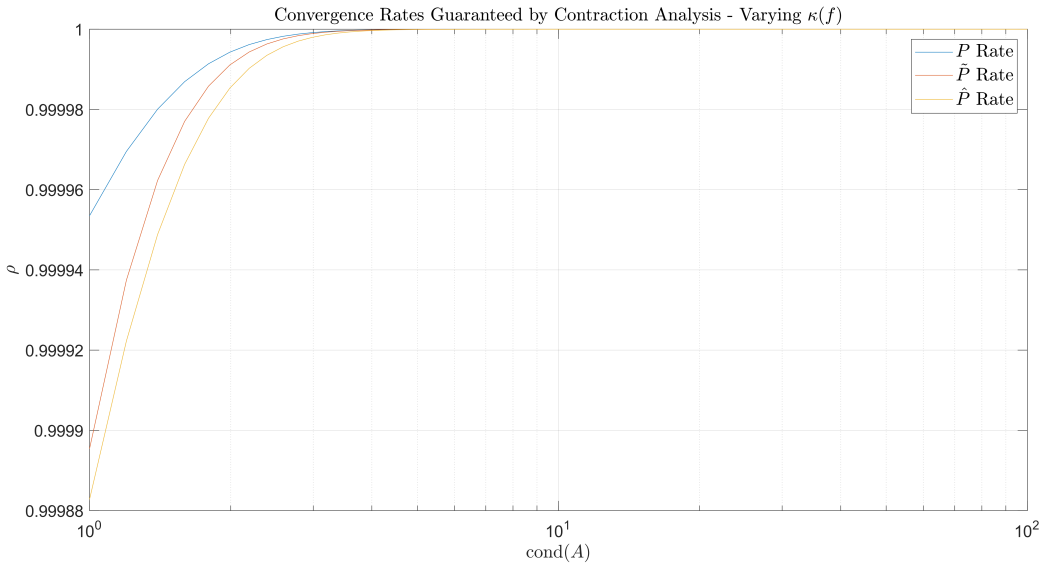

Figure 3.3: Convergence rates guaranteed by the three different contractive approaches as $\kappa(f)$ is varied.

are graphed in Figure 3.4. We see again that the rate produced by $\hat{P}$ outperforms the other two

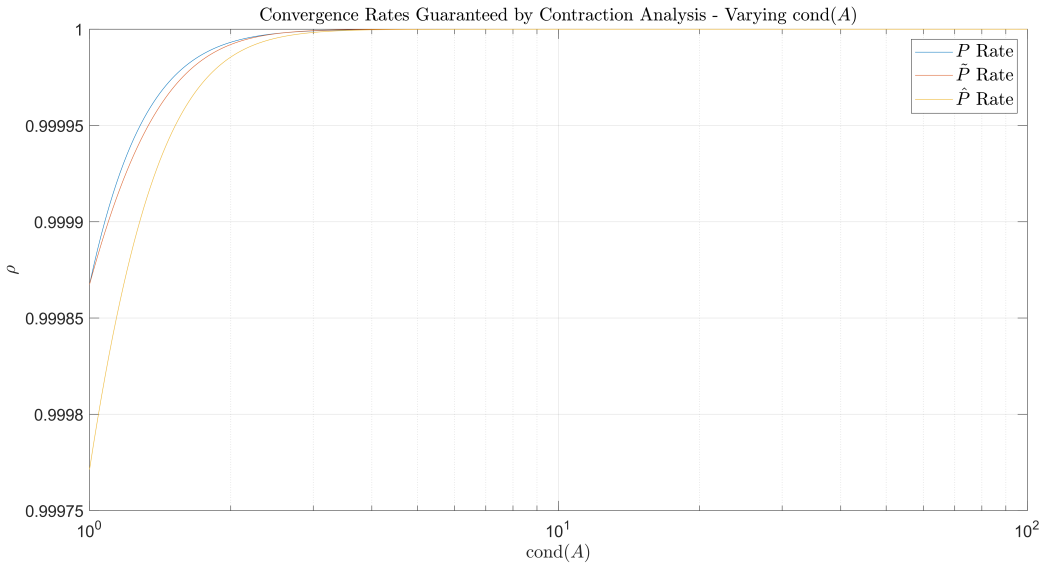

Figure 3.4: Convergence rates guaranteed by the three different contractive approaches as cond(A) is varied.

options. The optimal rate of $\rho = 0.9997$ was achieved when cond$(A) = 1$. Also, notice the effect of changing cond(A) compared to $\kappa(f)$. In the first test, we have a fixed cond$(A) = 2$ and the best achievable rate is slightly below 0.9998. In the second test, we fix $\kappa(f) = 2$ and the best rate is 0.9997. This result suggests that the condition number has a more significant impact on the rate of this analysis than $\kappa(f)$.

3.5.2 Interconnected Systems Approach

Calculating the rate for each $\kappa(f)$ and cond(A) was straightforward for this method. The rate is simply given by the function

$$\rho = 1 - \frac{1}{12\kappa^3(f)\text{cond}^4(A)}$$

Therefore, this approach is agnostic to the construction of A and is concerned solely with the singular values of A. The results for the first test are given in Figure 3.5. We see that the IS approach achieves

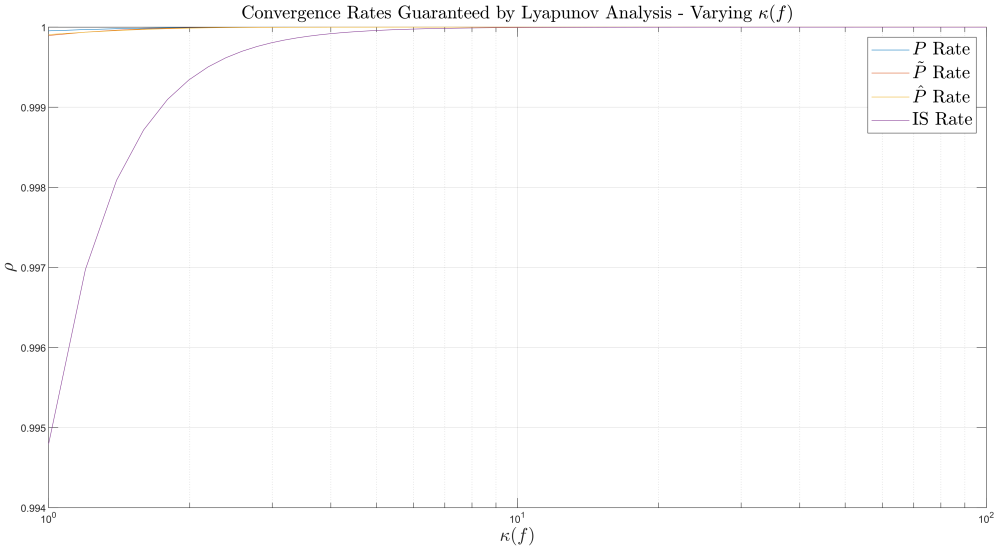

Figure 3.5: Convergence rates guaranteed by contractive analyses and IS Analysis as $\kappa(f)$ is varied.

a far better rate for each value of $\kappa(f)$. The optimal rate achieved was $\rho = 0.9947$ when $\kappa(f) = 1$. The same trend is seen for the second test as illustrated in Figure 3.6. The optimal rate is again achieved by the IS approach. The rate was achieved when cond$(A) = 1$ and had a value of $\rho = 0.9895$. Again, we see that cond(A) more significantly impacts the rate.

3.5.3 IQC Approach

To determine the optimal rate for the IQC approach, the LMI of interest was implemented in MATLAB. For each value of $\kappa(f)$ and cond(A), code was developed to iterate over a number of potential rate values and step sizes to determine the smallest rate for which the LMI was feasible. The computation time for this procedure took significantly longer than the other methods, thus the resolution of the plot is lower than that of the other tests. The other tests considered 500 unique κ and cond(A) values between 1 and 100. This approach considered 150 values. The results are as expected. The rate guaranteed by the IQC approach significantly outperforms all other approaches as seen in Figure 3.7 and Figure 3.8. As mentioned, the contractive approach rates are not plotted in these graphs as they would look essentially flat at this scale. We see that the IQC approach guarantees the best rates, but the approach considers specific step sizes rather than ranges. The IS approach outperformed the Lyapunov analyses; however, it only guaranteed exponential convergence.

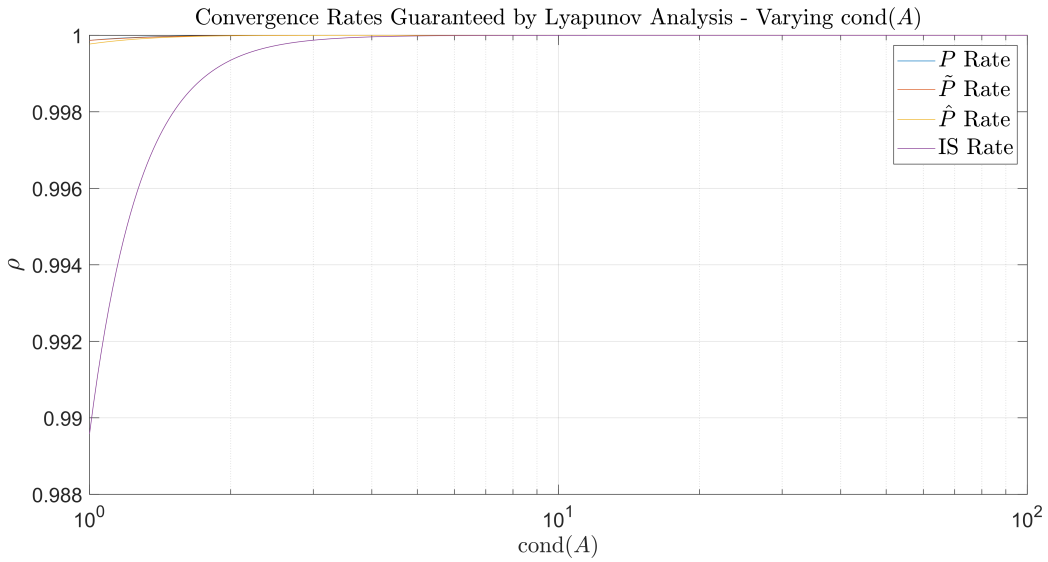

Figure 3.6: Convergence rates guaranteed by contractive and IS Analyses as cond(A) is varied.

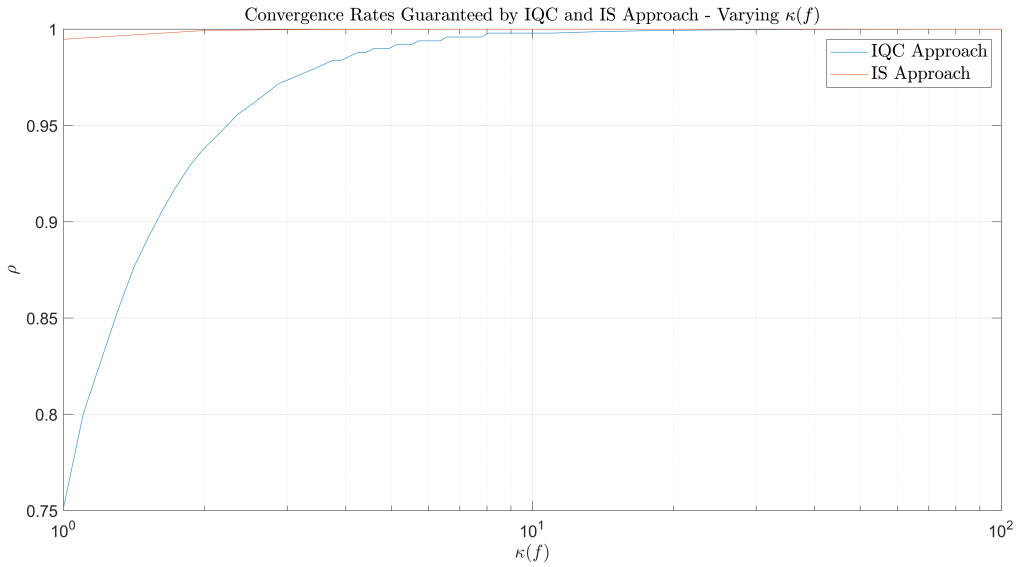

Figure 3.7: Convergence rates guaranteed by IS analysis and IQC analysis as $\kappa(f)$ is varied.

The Lyapunov approaches got the worst results, but they are able to guarantee contractivity for a range of step sizes.

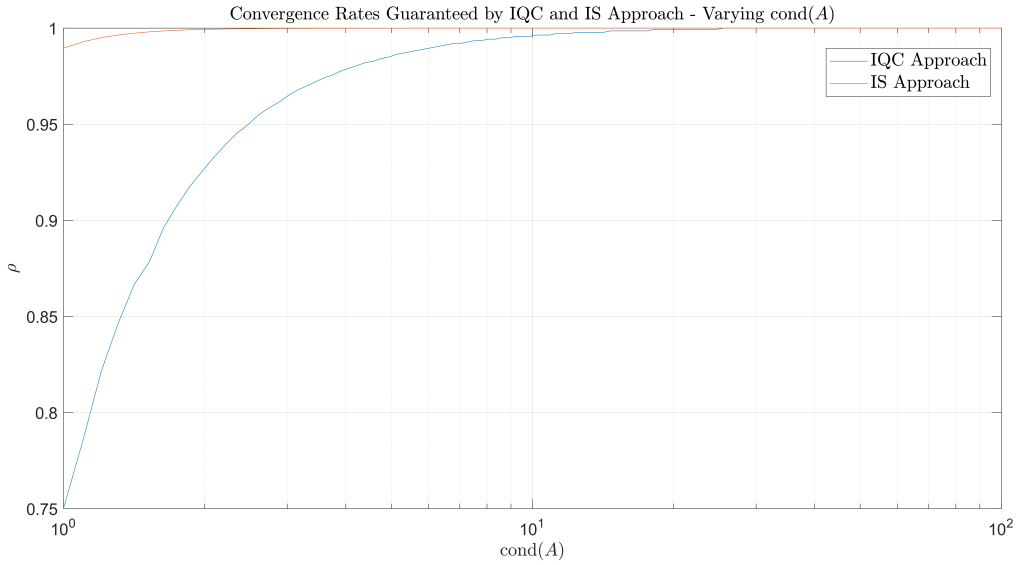

Figure 3.8: Convergence rates guaranteed by IS analysis and IQC analysis as cond(A) is varied.

Chapter 4

Extended Proximal Primal-Dual Feedback Controller

In this chapter, an online feedback-based optimization control method will be developed. The goal of the controller will be to drive the system's inputs and outputs to a critical point of the cost function. The cost function that will be considered in the optimization algorithm will be able to handle non-differentiable and differentiable costs on both the input and output. This setup provides users with considerable flexibility when designing the cost function. The plant considered will be a discrete-time non-linear system. It will be assumed that the plant is exponentially stable.

We will begin by introducing the system of interest and defining the input-output mapping. We will then define the cost function and derive the feedback algorithm. With the algorithm defined, two convergence proofs will be provided. The first proof will consider the algorithm when the plant is known and linear with respect to the input. This approach will provide sufficient conditions for the algorithm to converge exponentially to a unique global minimizer. The second approach will assume the plant is unknown and non-linear. This proof will guarantee that the algorithm converges to a critical point of the cost function.

4.1 Control Algorithm Derivation

Consider an exponentially stable discrete-time dynamical system of the form

$$\begin{aligned} x^{k+1} &= \phi(x^k, u^k, \omega^k) \\ y^k &= \theta(x^k, \omega^k) \\ u^{k+1} &= \Psi(y^k), \end{aligned} \tag{4.1}$$

where $x \in \mathbb{R}^p$ is the system's state, $u \in \mathbb{R}^n$ is the input, $y \in \mathbb{R}^m$ is the output, and $\omega \in \mathbb{R}^w$ is the disturbance. The function $\phi : \mathbb{R}^{\cdot} \times \mathbb{R}^n \times \mathbb{R}^w \to \mathbb{R}^{\cdot}$ determines the system's dynamics, $\theta : \mathbb{R}^{\cdot} \times \mathbb{R}^w \to \mathbb{R}^m$ is the measurement function, and $\Psi : \mathbb{R}^m \times \mathbb{R}^w \to \mathbb{R}^n$ is the feedback controller.

To define an input-output mapping, we will use an approach inspired by [25]. Assume that f is con-

tinuously differentiable and there exists a continuously differentiable function $\pi_x(u,\omega) : \mathbb{R}^n \times \mathbb{R}^w \to \mathbb{R}^p$ such that

$$\pi_x(u,\omega) = f(\pi_x(u,\omega), u, \omega), \quad \forall\ (u,\omega) \in \mathbb{R}^n \times \mathbb{R}^w.$$

Further, it is assumed that $\pi_x(u,\omega)$ is uniformly exponentially stable for all $(u,\omega) \in \mathbb{R}^n \times \mathbb{R}^w$. We will consider the steady-state input-output mapping $\pi : \mathbb{R}^n \times \mathbb{R}^\omega \to \mathbb{R}^m$ as

$$\pi(u,\omega) = \theta(\pi_x(u,\omega), \omega).$$

The input-output mapping is a non-linear analog to the DC gain for LTI systems. We will further assume that the state dynamics in system (4.1) operate much faster than the disturbance, ω^k, and the feedback controller, Ψ. With this assumption, all measurements, y, used in the controller are steady-state measurements. We will assume that Ψ is slow enough such that the interconnection of the system and the controller is exponentially stable if the controller exponentially converges to a fixed point. Thus, our goal is to derive Ψ and prove that it is exponentially convergent.

The goal of the feedback controller is to steer the system's inputs and steady-state outputs toward the critical point of some cost function. The optimization problem considered can be represented as

$$\begin{aligned} \min_{u\in\mathbb{R}^n, y\in\mathbb{R}^m}\ & f(u)+g(u) + h(y) + c(y) \\ \text{s.t. } & y = \pi(u,\omega). \end{aligned} \tag{4.2}$$

Both f and h are differentiable convex functions where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$. Also, g and c are closed proper convex functions where $g : \mathbb{R}^n \to \bar{\mathbb{R}}$ and $c : \mathbb{R}^m \to \bar{\mathbb{R}}$. Notice that this optimization problem is only convex if $\pi(u,\omega)$ is linear in u. For general non-linear π, we can not guarantee the optimality of a fixed point. Thus, the goal of the algorithm will be to find the critical points of the problem. To do this, the Lagrangian of the optimization problem (4.2) is constructed,

$$L(u,y;\lambda) = f(u) + g(u) + h(y) + c(y) + \langle \lambda, y - \pi(u,\omega)\rangle, \tag{4.3}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product. From Theorem 2.3.3, we know that the critical points of (4.2) are characterized by

$$\begin{aligned} y - \pi(u,\omega) &= 0 \\ \frac{\partial f}{\partial u}(u) + \partial g(u) - \lambda^\top \frac{\partial \pi}{\partial u}(u,\omega) &\ni 0 \\ \frac{\partial h}{\partial y}(y) + \partial c(y) + \lambda^\top &\ni 0. \end{aligned} \tag{4.4}$$

To develop an algorithm that converges to (4.4), the augmented proximal Lagrangian will be developed. This approach is adapted from [8]. To develop the proximal augmented Lagrangian, we first need to introduce the augmented Lagrangian. The augmented Lagrangian, L_μ can be constructed by adding a strictly convex term with respect to the equality constraint. Note that the critical points of the augmented Lagrangian are exactly the same as (4.4). The augmented Lagrangian for (4.2) is

$$L_\mu(u,y;\lambda) = f(u) + g(u) + h(y) + c(y) - \langle \lambda, y - \pi(u,\omega)\rangle + \frac{1}{2\mu}||y - \pi(u,\omega)||^2. \tag{4.5}$$

The linear term can be removed by completing the square. Further, we replace y with $\pi(u,\omega)$ in the argument of h,

$$L_\mu(u,y;\lambda) = f(u) + g(u) + h(\pi(u,\omega)) + c(y) + \frac{1}{2\mu}||y - (\pi(u,\omega) + \mu\lambda)||^2 - \frac{\mu}{2}||\lambda||^2. \tag{4.6}$$

To find the critical points of the system, we maximize equation (4.6) with respect to the dual variable, and minimize with respect to the primal variables,

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} L_\mu(u,y;\lambda) \\
&= \sup_{\lambda\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m}\{f(u) + g(u) + h(\pi(u,\omega)) + c(y) + \frac{1}{2\mu}||y - (\pi(u,\omega) + \mu\lambda)||^2 - \frac{\mu}{2}||\lambda||^2\} \\
&= \sup_{\lambda\in\mathbb{R}^m}\inf_{u\in\mathbb{R}^n}\{f(u) + g(u) + h(\pi(u,\omega)) + \inf_{y\in\mathbb{R}}\{c(y) + \frac{1}{2\mu}||y - (\pi(u,\omega) + \mu\lambda)||^2\} - \frac{\mu}{2}||\lambda||^2\}.
\end{aligned} \tag{4.7}$$

We will denote the minimizer with respect to y as y^*. Notice that y^* is given by

$$y^* = \inf_y \left\{ c(y) + \frac{1}{2\mu}||y - (\pi(u,\omega) + \mu\lambda)||^2 \right\} = \text{prox}_{\mu c}(\pi(u,\omega) + \mu\lambda).$$

Using the definition of the Moreau envelope, (4.7) simplifies to

$$\begin{aligned}
\sup_{\lambda\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} L_\mu(u,y;\lambda) &= \sup_{\lambda\in\mathbb{R}^m}\inf_{u\in\mathbb{R}^n} L_\mu(u,y^*;\lambda) \\
&= \sup_{\lambda\in\mathbb{R}^m}\inf_{u\in\mathbb{R}^n}\{f(u) + g(u) + h(\pi(u,\omega)) \\
&\quad + M_{\mu r}(\pi(u,\omega) + \mu\lambda) - \frac{\mu}{2}||\lambda||^2\} \\
&:= \sup_{\lambda\in\mathbb{R}^m}\inf_{u\in\mathbb{R}^n}\{\underbrace{\mathcal{L}_\mu(u;\lambda) + g(u)}_{\mathcal{L}^{\text{e}}_\mu(u;\lambda)}\},
\end{aligned} \tag{4.8}$$

where we call $\mathcal{L}_\mu(u;\lambda)$ the proximal augmented Lagrangian, and $\mathcal{L}^{\text{e}}_\mu(u;\lambda)$ the extended proximal augmented Lagrangian (e-PAL).

From Theorem 2.3.3, we know that $(u^*,\lambda^*)\mathbb{R}^{n+m}$ that solves (4.8) will satisfy (4.4). Thus, the challenge becomes developing an algorithm that converges to the saddle point defined by (4.8). Because the e-Pal is continuously differentiable with respect to λ, gradient ascent can be used for the dual variable. For the primal variable, we see that (4.8) is the sum of a differentiable and non-differentiable cost function. Thus, the proximal gradient method can be used. Writing the algorithm explicitly, we have

$$\begin{aligned}
u^{k+1} &= \mathbf{prox}_{\alpha g}\left(u^k - \alpha\nabla_u\mathcal{L}_\mu(u^k,\lambda^k)\right) \\
\lambda^{k+1} &= \lambda^k + \alpha\nabla_\lambda\mathcal{L}_\mu(u^k,\lambda^k).
\end{aligned} \tag{4.9}$$

The algorithm can be written explicitly as

$$\begin{aligned}
u^{k+1} &= \mathbf{prox}_{\mu g}\left(u^k - \alpha\nabla f(u^k) - \alpha\frac{\partial\pi}{\partial u}(u^k,\omega^k)^\top(\nabla h(\pi(u^k,\omega^k)) + \nabla M_{\mu c}(\pi(u^k,\omega^k) + \mu\lambda^k))\right) \\
\lambda^{k+1} &= \lambda^k + \alpha\mu(\nabla M_{\mu c}(\pi(u^k,\omega^k) + \mu\lambda^k) - \lambda^k)
\end{aligned}$$

We note that the algorithm requires the evaluation of $\frac{\partial \pi}{\partial u}(u, \omega)$. Rather than requiring an exact formulation for $\frac{\partial \pi}{\partial u}(u, \omega)$, which would imply full knowledge of $\pi(u, \omega)$, an approximation in the operating range of interest is used, $\Pi u + \Pi_\omega \omega = \pi(u, \omega)$ where $\Pi \in \mathbb{R}^{m \times n}$ and $\Pi_\omega \in \mathbb{R}^{m \times w}$. Also, we know that $\pi(u^k, \omega^k) = y^k$ by definition. We will adjust the algorithm to include the approximation, and replace the input-output mapping with the measurement y^k to arrive at the final algorithm

$$\begin{aligned} u^{k+1} &= \Psi(y^k) = \mathbf{prox}_{\alpha g}\left(u^k - \alpha \nabla f(u^k) - \alpha \Pi^\top (\nabla h(y^k) + \nabla M_{\mu r}(y^k + \mu \lambda^k))\right) \\ \lambda^{k+1} &= \lambda^k + \alpha \mu (\nabla M_{\mu r}(y^k + \mu \lambda^k) - \lambda^k). \end{aligned} \tag{4.10}$$

Because an approximation is used, we cannot guarantee that the fixed point of (4.10) meets the necessary conditions in (4.4). To accommodate this, we redefine what the fixed points of the algorithm encode.

Definition 4.1.1. *Consider the optimization problem (4.2). Given a matrix $\Pi \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, and $\omega \in \mathbb{R}^w$. A vector $(u^*, \lambda^*) \in \mathbb{R}^{n+m}$ is an online approximate solution if the following conditions hold*

$$\begin{aligned} y &= \pi(u^*, \omega) \\ u^* &= \boldsymbol{prox}_{\alpha g}\left(u^* - \alpha \left(\nabla f(u^*) + \Pi^\top \left(\nabla h(y) + \nabla M_{\mu c}(y + \mu \lambda^*)\right)\right)\right) \\ \lambda^* &= \lambda^* + \alpha \mu (\nabla M_{\mu c}(y + \mu \lambda^*) - \lambda^*). \end{aligned} \tag{4.11}$$

When $\Pi = \frac{\partial \pi}{\partial u}(u)$, the online approximate solution is an equivalent condition to (4.4). When the approximation is not equivalent to the true Jacobian, the online approximate solution represents a vector that satisfies the KKT conditions for the approximated system. Also, notice that the algorithm uses a proximal operator to handle the non-differentiable cost functions. The simplicity of implementing this algorithm depends on the simplicity of evaluating the proximal operators of these functions. For example, the proximal operator of the indicator function and l_1 norm can easily be evaluated.

4.2 Convergence under Convexity Assumption

In this section, we will consider the algorithm (4.10) when the input-output mapping is known and linear. This is given by $\pi(u, \omega) = \Pi u + \Pi_\omega \omega$. Replacing y in the algorithm with the known mapping gives the following algorithm representation

$$\begin{aligned} u^{k+1} &= \mathbf{prox}_{\alpha g}\left(u^k - \alpha \nabla f(u^k) - \alpha \Pi^\top (\nabla h(\Pi u^k + \Pi_\omega \omega^k) + \nabla M_{\mu r}(\Pi u^k + \Pi_\omega \omega^k + \mu \lambda^k))\right) \\ \lambda^{k+1} &= \lambda^k + \alpha \mu \left(\nabla M_{\mu r}(\Pi u^k + \Pi_\omega \omega^k + \mu \lambda^k) - \lambda^k\right). \end{aligned} \tag{4.12}$$

Considering the algorithm with this assumption provides a number of benefits. As mentioned in the algorithm derivation, under this assumption, the online approximate solution is equivalent to the necessary conditions for the true system. Further, for convex f, h, c, and g, (4.2) becomes a convex optimization problem; therefore, the online approximate solution is a global minimizer of (4.2).

To show that the algorithm is convergent under these assumptions, the following lemma will be used.

Lemma 4.2.1. *Consider the function $F : \mathbb{R}^n \to \mathbb{R}^n$ and a fixed-point algorithm of the form*

$$z^{k+1} = z^k + \alpha F(z^k). \tag{4.13}$$

If (4.13) is contractive with respect to $||\cdot||_2$ with rate $\rho \in (0,1)$, then the associated proximal update

$$\hat{z}^{k+1} = \boldsymbol{prox}_{\mu g}(\hat{z}^k + \alpha F(\hat{z}^k)) \tag{4.14}$$

is also contractive with respect to $||\cdot||_2$ with rate ρ where $\mu > 0$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a closed proper convex function.

Proof. Consider $\hat{z}^{k+1}, \hat{y}^{k+1} \in \mathbb{R}^n$ defined by the update law (4.14). We have

$$\begin{aligned} ||\hat{z}^{k+1} - \hat{y}^{k+1}||_2 &= ||\mathbf{prox}\left(\hat{z}^k - \alpha F(\hat{z}^k)\right) - \mathbf{prox}\left(\hat{y}^k - \alpha F(\hat{y}^k)\right)||_2 \\ &\leq ||\hat{x}^k - \hat{y}^k - \alpha\left(F(\hat{x}^k) - F(\hat{y}^k)\right)||_2 \\ &\leq \rho||\hat{x}^k - \hat{y}^k||_2, \end{aligned}$$

where the first inequality holds due to the non-expansiveness of the proximal operator. □

We will be using Lemma 4.2.1 to simplify the proof approach. The lemma allows us to analyze the algorithm without the proximal operator and use the results to determine the algorithm with the proximal operator converges. The algorithm without the proximal operator is given by

$$\begin{aligned} u^{k+1} &= u_k - \alpha\nabla f(u^k) - \alpha\Pi^\top(\nabla h(\Pi u^k + \Pi_\omega \omega^k) + \nabla M_{\mu c}(\Pi u^k + \Pi_\omega \omega^k + \mu\lambda^k)) \\ \lambda_{k+1} &= \lambda^k + \alpha\mu(\nabla M_{\mu r}(\Pi u^k + \Pi_\omega \omega^k + \mu\lambda^k) - \lambda^k). \end{aligned} \tag{4.15}$$

We will call (4.15) the stripped algorithm and (4.12) the full algorithm. Notice that showing that (4.15) exponentially converges to a fixed point is not sufficient for proving the convergence of (4.12). We must show that the algorithm is globally contractive with respect to $||\cdot||_2$, which is a stronger condition.

Before introducing the convergence results, we will highlight the connections between the primal-dual algorithm and the algorithm of interest. Although they are both derived using a Lagrangian approach, the structural similarities are not obvious. From Lemma 3.1.2, it is known there exists $Q_z \in S^n$, $B_z \in S^m$, and $D_z \in S^m$ such that (4.15) can be written as

$$z^{k+1} = z^k + \alpha \begin{bmatrix} -Q_z - \frac{1}{\mu}\Pi^\top(B_z + D_z)\Pi & -\Pi^\top D_z \\ D_z\Pi & \mu(D_z - \mathbf{I}) \end{bmatrix} z^k \tag{4.16}$$

where $z^k = [u^k,\ \lambda^k]^\top$. As a reminder, the primal-dual algorithm for a convex equality-constrained optimization problem has the form

$$z^k = z^k + \alpha \begin{bmatrix} -T_z & -A^\top \\ A & \mathbf{0} \end{bmatrix} z^k \tag{4.17}$$

for some $T_z \succ 0$ and full row-rank $A \in \mathbb{R}^{m\times n}$. When written as linear functions, the connection

between the (4.17) and (4.16) becomes much clearer. Notice that

$$-Q_z - \frac{1}{\mu}\Pi^\top (B_z + D_z) \succ 0, \tag{4.18}$$

because $Q_z \succ 0$ and $\frac{1}{\mu}\Pi^\top (B_z + D_z) \succeq 0$. Thus, both algorithms have strongly monotone terms in the top left corner of the matrices. Further both algorithms have off diagonals that are related through a negation and transpose,

$$-(D_z\Pi)^\top = -\Pi^\top D_z \quad \text{and} \quad -(A)^\top = -A^\top$$

Despite the similarities, there are key differences between (4.17) and (4.16). Notice that $D\Pi$ is not full row-rank and the dual update in (4.16) has a positive definite λ term. With the similarities between the two algorithms, it is only natural that a proof method used in Chapter 3 was considered. The seemingly obvious choice may be the Interconnected Systems approach as it yielded the best rates out of all the analytical approaches; however, $D\Pi$ not being full row-rank presents an issue. The dual update for this algorithm in IS representation is given by

$$\lambda^{k+1} = \lambda^k - \beta \underbrace{\left(D_z\Pi \left(Q_z + \Pi^\top (B_z + D_z)\,\Pi\right)^{-1} \Pi^\top D_z + \mu(\mathbf{I} - D_z)\right)}_{:=K} \lambda^k + \beta D_z \Pi y^k$$

It can be shown that K is positive definite, but it is unknown how to bound this matrix. Without a lower bound, the IS approach cannot be used. Further, we must show that the algorithm is contractive, and the IS approach only guarantees exponential convergence. For these reasons, the contractive approach will be used.

Theorem 4.2.2. *Consider the algorithm (4.15). Assume that $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$ are strongly convex with parameters m_f and m_h, respectively, and have Lipschitz continuous gradients with parameters L_f and L_h, respectively. Further, assume that $g : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^m \to \mathbb{R}$ are closed proper convex functions and $\Pi \in \mathbb{R}^{m\times n}$ is full-row rank. Then there exists α^* and $P \succ 0$ such that (4.15) is contractive with respect to $||\cdot||_{P,2}$ for all $\alpha \in (0, \alpha^*)$.*

Proof. The structure of the proof approach is inspired by the proof used in [11]. To start the proof, the algorithm is simplified using Lemma 3.1.2. For $x, y \in \mathbb{R}^n$ and $\lambda, \gamma \in \mathbb{R}^m$, let $\hat{u} = u - v$, $\hat{\lambda} = \lambda - \gamma$, and

$$\begin{aligned} Q_{\hat{u}}\hat{u} &= (\nabla f(u) - \nabla f(v)) \\ B_{\hat{u}}\Pi\hat{u} &= \nabla h(\Pi u + \Pi_\omega \omega) - \nabla h(\Pi v + \Pi_\omega \omega) \\ D_{\hat{u},\hat{\lambda}}(\Pi\hat{u} + \mu\hat{\lambda}) &= \mu\nabla M_{\mu c}(\Pi u + \Pi_\omega\omega + \mu\lambda) - \mu\nabla M_{\mu c}(\Pi v + \Pi_\omega\omega + \mu\gamma) \end{aligned}$$

Using the matrix notation, the algorithm can be rewritten as

$$\hat{z}^{k+1} = \hat{z}^k + \alpha \underbrace{\begin{bmatrix} -\hat{Q} - \frac{1}{\mu}\Pi^\top D\Pi & -\Pi^\top D \\ D\Pi & \mu(D - \mathbf{I}) \end{bmatrix}}_{:=F(\hat{z^k})} \hat{z}^k \tag{4.19}$$

Where the matrix subscripts have been dropped for notational simplicity, $\hat{Q} = Q + \Pi^\top B\Pi$, and $\hat{z}^k = [\hat{u}^k,\ \hat{\lambda}^k]$. Notice that $m_f\mathbf{I} \prec \hat{Q} \prec \left(L_f + \sigma^2_{\max}(\Pi)L_h\right)\mathbf{I}$. To prove the algorithm is contractive,

Lemma 3.4 will be used. As a reminder, this involves showing that F is Lipschitz continuous and

$$F(\hat{z})^\top P\hat{z} \leq -b||\hat{z}||_{P,2}^2 \tag{4.20}$$

holds for all $\hat{z} \in \mathbb{R}^{n+m}$. It is trivial to show that F is Lipschitz continuous. Let L_F be the Lipschitz parameter of F. Thus, showing inequality (4.20) holds for some b proves the algorithm is contractive. The matrix P that will be used in the analysis is inspired by the norms analyzed in Chapter 2.

$$P = \begin{bmatrix} \mathbf{I} & \epsilon\Pi^\top\bar{\Pi} \\ \epsilon\bar{\Pi}\Pi & (1+\epsilon\beta)\mathbf{I} \end{bmatrix} \tag{4.21}$$

where $\bar{\Pi} = (\Pi\Pi^\top)^{-1}$, $\epsilon > 0$ and $\beta > 0$. We know that (4.21) is positive definite if $\epsilon < \sigma_{\min}(\Pi)$. Using Lemma 2.1.2, we know that $\bar{\Pi}$ is well-defined because Π is full-row rank. With the algorithm rewritten in the form of an LTI system, showing that (4.20) is satisfied is equivalent to

$$\begin{aligned}
&F^\top(\hat{z}^k)P\hat{z}^k \leq -b||\hat{z}^k||_{P,2}^2 \\
\iff &\left(\begin{bmatrix} -\hat{Q} - \frac{1}{\mu}\Pi^\top D\Pi & -\Pi^\top D \\ D\Pi & \mu(D-\mathbf{I}) \end{bmatrix}\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix}\right)^\top P\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix} \leq -b||\hat{z}^k||_{P,2}^2 \\
\iff &\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix}^\top \left(\begin{bmatrix} -\hat{Q} - \frac{1}{\mu}\Pi^\top D\Pi & \Pi^\top D \\ -D\Pi & \mu(D-\mathbf{I}) \end{bmatrix}\begin{bmatrix} \mathbf{I} & \epsilon\Pi^\top\bar{\Pi} \\ \epsilon\bar{\Pi}\Pi & (1+\epsilon\beta)\mathbf{I} \end{bmatrix}\right)\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix} \leq -b||\hat{z}^k||_{P,2}^2 \\
\iff &\frac{1}{2}\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix}^\top \left(\begin{bmatrix} -\hat{Q} - \frac{1}{\mu}\Pi^\top D\Pi & \Pi^\top D \\ -D\Pi & \mu(D-\mathbf{I}) \end{bmatrix}\begin{bmatrix} \mathbf{I} & \epsilon\Pi^\top\bar{\Pi} \\ \epsilon\bar{\Pi}\Pi & (1+\epsilon\beta)\mathbf{I} \end{bmatrix}\right. \\
&\quad \left. + \begin{bmatrix} \mathbf{I} & \epsilon\Pi^\top\bar{\Pi} \\ \epsilon\bar{\Pi}\Pi & (1+\epsilon\beta)\mathbf{I} \end{bmatrix}\begin{bmatrix} -\hat{Q} - \frac{1}{\mu}\Pi^\top D\Pi & -\Pi^\top D \\ D\Pi & \mu(D-\mathbf{I}) \end{bmatrix}\right)\begin{bmatrix} \hat{u}^k \\ \hat{\lambda}^k \end{bmatrix} \leq -b||\hat{z}^k||_{P,2}^2 \\
\iff &-\begin{bmatrix} W_0 & \frac{\epsilon}{2}\left(W_1\Pi^\top\bar{\Pi} + W_2\right) \\ \frac{\epsilon}{2}\left(\bar{\Pi}\Pi W_1 + W_2^\top\right) & W_3 \end{bmatrix} \preceq 0
\end{aligned}$$

where the matrices are defined by

$$\begin{aligned}
W_0 &= \hat{Q} - \tfrac{\epsilon}{2}\Pi^\top(D\bar{\Pi} + \bar{\Pi}D)\Pi - b\mathbf{I} \\
W_1 &= Q + (\tfrac{1}{\mu} - \beta)\Pi^\top D\Pi - 2b\mathbf{I} \\
W_2 &= \mu\Pi^\top\bar{\Pi}(\mathbf{I} - D) \\
W_3 &= \epsilon D + \mu(1+\epsilon\beta)(\mathbf{I} - D) - (1+\epsilon\beta)b\mathbf{I}.
\end{aligned} \tag{4.22}$$

Therefore, showing that

$$\begin{bmatrix} W_0 & \frac{\epsilon}{2}\left(W_1\Pi^\top\bar{\Pi} + W_2\right) \\ \frac{\epsilon}{2}\left(\bar{\Pi}\Pi W_1 + W_2^\top\right) & W_3 \end{bmatrix} \succeq 0 \tag{4.23}$$

holds ensures algorithm (4.15) is contractive. To prove that (4.23) holds, the Schur complement is used. The Schur complement states that if $W_3 \succ 0$ and

$$W_0 - \tfrac{\epsilon^2}{4}(W_1\Pi^\top\bar{\Pi} + W_2)(W_3)^{-1}(\bar{\Pi}\Pi W_1^\top + W_2^\top) \succeq 0 \tag{4.24}$$

then (4.23) is satisfied [24, Theorem 1.12]. Taking $b = \frac{\epsilon}{2(1+\epsilon\beta)}$, $\epsilon \leq \frac{\mu}{1-\mu\beta}$, and $\beta \in [0, \frac{1}{\mu}]$, we know that $W_3 \succeq \frac{\epsilon}{2}\mathbf{I}$. To show this we use the fact that $D \preceq \mathbf{I}$ and

$$\begin{aligned} W_3 - \tfrac{\epsilon}{2}\mathbf{I} &= \epsilon D + \mu(1+\epsilon\beta)(\mathbf{I} - D) - \tfrac{\epsilon}{2}\mathbf{I} - (1+\epsilon\beta)b\mathbf{I} \\ &= (\epsilon - \mu(1+\epsilon\beta))D + \mu(1+\epsilon\beta)\mathbf{I} - \tfrac{\epsilon}{2}\mathbf{I} - \tfrac{\epsilon}{2}\mathbf{I} \\ &\succeq (\epsilon - \mu(1+\epsilon\beta))\mathbf{I} + \mu(1+\epsilon\beta)\mathbf{I} - \epsilon\mathbf{I} \\ &= \mathbf{0}. \end{aligned} \tag{4.25}$$

Using (4.25), statement (4.24) can be reduced to

$$W_0 - \tfrac{\epsilon}{2}(W_1\Pi^\top\bar{\Pi} + W_2)(\bar{\Pi}\Pi W_1^\top + W_2^\top) \succeq 0. \tag{4.26}$$

Expanding (4.26) and using the triangle inequality, we see that (4.26) is satisfied if

$$W_0 - \tfrac{\epsilon}{2}(W_1\Pi^\top\bar{\Pi}^2\Pi W_1 + 2||W_1\Pi^\top\bar{\Pi}W_2^\top||_2 + W_2W_2^\top) \succeq 0. \tag{4.27}$$

We now aim to bound each element of (4.27). Starting with W_0, we use the fact that $\frac{\epsilon}{2}\Pi^\top(D\bar{\Pi} + \bar{\Pi}D)\Pi \preceq \epsilon\mathbf{I}$. This bound is proven in Appendix A. This gives us,

$$W_0 \succeq \left(m_f - \epsilon - \tfrac{\epsilon}{2(1+\beta)}\right)\mathbf{I}. \tag{4.28}$$

For the $W_1\Pi^\top\bar{\Pi}^2\Pi W_1$ term, we use the fact that $\Pi\bar{\Pi}^2\Pi \succeq \frac{1}{\sigma_{\min}^2(\Pi)}\mathbf{I}$. The proof of this can be found in Appendix A. Using this,

$$\begin{aligned} W_1\Pi^\top\bar{\Pi}^2\Pi W_1 &\preceq \tfrac{1}{\sigma_{\min}^2(\Pi)}(\hat{Q} + (\tfrac{1}{\mu} - \beta)\Pi^\top D\Pi - 2b\mathbf{I})^2 \\ &\preceq \tfrac{1}{\sigma_{\min}^2(\Pi)}\big(\hat{Q}^2 + 2(\tfrac{1}{\mu} - \beta)||\hat{Q}||_2||\Pi^\top D\Pi||_2 \\ &\quad - 4b\hat{Q} + (\tfrac{1}{\mu} - \beta)^2\Pi^\top D\Pi\Pi^\top D\Pi - 4b(\tfrac{1}{\mu} - \beta)\Pi^\top D\Pi + 4b^2\mathbf{I}\big) \\ &\preceq \tfrac{1}{\sigma_{\min}^2(\Pi)}\left(\hat{L}_f^2 + 2(\tfrac{1}{\mu} - \beta)\hat{L}_f\sigma_{\max}^2(\Pi) + (\tfrac{1}{\mu} - \beta)^2\sigma_{\max}^4(\Pi) + \tfrac{\epsilon^2}{(1+\epsilon\beta)^2}\right)\mathbf{I} \\ &:= w_1(\epsilon, \beta) \end{aligned} \tag{4.29}$$

when $\beta \in [0, \frac{1}{\mu}]$. We can bound the $2||W_1\Pi^\top\bar{\Pi}W_2^\top||_2$ term using the triangle inequality, the fact that $\Pi^\top\bar{\Pi}\Pi \preceq \mathbf{I}$, and $I - D \preceq \mathbf{I}$. Using these properties,

$$\begin{aligned} 2||W_1\Pi^\top\bar{\Pi}W_2^\top||_2 &= 2\mu||(\hat{Q} + (\tfrac{1}{\mu} - \beta)\Pi^\top D\Pi - 2b)\Pi^\top\bar{\Pi}(\mathbf{I} - D)\bar{\Pi}\Pi||_2 \\ &\preceq 2\mu||\hat{Q} + (\tfrac{1}{\mu} - \beta)\Pi^\top D\Pi - 2b||_2||\Pi^\top\bar{\Pi}(\mathbf{I} - D)\bar{\Pi}\Pi||_2\mathbf{I} \\ &\preceq 2\mu(\hat{L}_f + \sigma_{\max}^2(\Pi)(\tfrac{1}{\mu} - \beta) + 2b)||\Pi^\top\bar{\Pi}^2\Pi||_2\mathbf{I} \\ &\preceq 2\tfrac{\mu}{\sigma_{\min}^2(\Pi)}(\hat{L}_f + \sigma_{\max}^2(\Pi)(\tfrac{1}{\mu} - \beta) + \tfrac{\epsilon}{1+\beta})\mathbf{I} \\ &:= w_2(\epsilon, \beta). \end{aligned} \tag{4.30}$$

Finally, the $W_2W_2^\top$ term can be bounded as

$$\begin{aligned} W_2W_2^\top &\preceq \mu^2\Pi^\top\bar{\Pi}(\mathbf{I}-D)\bar{\Pi}\Pi \\ &\preceq \tfrac{\mu^2}{\sigma^2_{\min}(\Pi)}\mathbf{I} \\ &:= w_3(\epsilon,\beta). \end{aligned} \tag{4.31}$$

Using the bounds described in (4.28), (4.29), (4.30), and (4.31), the following holds

$$W_0 - \tfrac{\epsilon}{2}(W_1\Pi^\top\bar{\Pi}+W_2)(\bar{\Pi}\Pi W_1^\top + W_2^\top) \succeq m_f\mathbf{I} - \epsilon\left(1+\tfrac{1}{2(1+\beta)}+\tfrac{1}{2}\sum_{i=1}^{3}w_i(\epsilon,\beta)\right). \tag{4.32}$$

Notice that $\sum_{i=1}^{3} w_i(\epsilon,\beta)$ monotonically increases with respect to ϵ. Because $\epsilon < \min\left(\sigma_{\min}(\Pi), \tfrac{\mu}{(1+\epsilon\beta)}\right)$, set $\bar{\epsilon} = \min\left(\sigma^2_{\min}(\Pi), \tfrac{\mu}{(1+\epsilon\beta)}\right)$. Then,

$$\begin{aligned} \epsilon\left(1+\tfrac{1}{2(1+\beta)}+\tfrac{1}{2}\sum_{i=1}^{3}w_i(\epsilon,\beta)\right) &\le \epsilon\left(1+\tfrac{1}{2(1+\beta)}+\tfrac{1}{2}\sum_{i=1}^{3}w_i(\bar{\epsilon},\beta)\right) \\ &=: \epsilon w_0(\bar{\epsilon},\beta). \end{aligned} \tag{4.33}$$

Combining (4.32) and (4.33), we see that the Schur complement is positive definite and, thus, (4.20) is satisfied for

$$\epsilon < \min\left(\sigma^2_{\min}(\Pi), \tfrac{\mu}{(1+\epsilon\beta)}, \frac{m_f}{w_0(\bar{\epsilon},\beta)}\right)$$

with $b = \tfrac{\epsilon}{2(1+\beta)}$ and $\beta \in [0, \tfrac{1}{\mu}]$. Using Lemma 3.1.1, algorithm (4.15) converges for $\alpha < \tfrac{\epsilon}{(1+\epsilon\beta)L_F^2\text{cond}(P)}$. Further, algorithm (4.15) converges with rate

$$\rho = \sqrt{1 - \frac{\epsilon^2}{4L_F^2(1+\epsilon\beta)^2\text{cond}(P)}}.$$

for $\alpha = \tfrac{\epsilon}{2(1+\beta)L_F^2\text{cond}(P)}$. □

Theorem 4.2.2 guarantees that (4.15) is contractive for properly selected ϵ, but this does not guarantee that (4.12) is exponentially convergent. The algorithm (4.15) has to be contractive enough that switching from the $||\cdot||^2_{P,2}$ norm to the $||\cdot||_2^2$ norm keeps the contraction rate below 1. To use Lemma 4.2.1, we need (4.15) to be contractive with parameter $\rho < \tfrac{1}{\text{cond}(P)}$ with respect to $||\cdot||_{P,2}$ for convergence.

4.3 Convergence under General Conditions

In this section, we assume that the mapping $\pi(u,\omega)$ is unknown and non-linear. The contractive approach cannot be used under these assumptions. To prove convergence, the IQC method discussed in Chapter 3 will be used. To do this, we again consider (4.10) without the proximal operator,

$$\begin{aligned} u^{k+1} &= u^k - \alpha\nabla f(u^k) - \alpha\Pi^\top(\nabla h(y^k + \nabla M_{\mu r}(y^k+\mu\lambda^k)) \\ \lambda^{k+1} &= \lambda^k + \alpha\mu(\nabla M_{\mu r}(y^k+\mu\lambda^k) - \lambda^k). \end{aligned} \tag{4.34}$$

To use the IQC approach, (4.34) must be reconfigured such that it is in the form of an LTI system with the non-linearities represented as disturbances to the system. This representation is given by

$$\begin{aligned} z^{k+1} &= Az^k + B\gamma^k \\ \zeta^k &= Cz^k + D\gamma^k \\ \gamma^k &= \Delta(\zeta^k) \end{aligned} \tag{4.35}$$

where z, ζ, q, and Δ are given by

$$\begin{aligned} z^k &= \begin{bmatrix} u^k & \lambda^k \end{bmatrix}^\top \\ \zeta_1^k &= z_1^k \\ \zeta_2^k &= \Pi z_1^k + \mu z_2^k + \gamma_4^k \\ \zeta_3^k &= \Pi z_1^k + \gamma_4^k \\ \gamma_1^k &= \Delta_1(\zeta_1^k) = \nabla f(\zeta_1^k) - m_f \zeta_1^k \\ \gamma_2^k &= \Delta_2(\zeta_2^k) = \mu \nabla M_{\mu c}(\zeta_2^k) \\ \gamma_3^k &= \Delta_3(\zeta_3^k) = \nabla h(\zeta_3^k) - m_h \zeta_3^k \\ \gamma_4^k &= \Delta_4(\zeta_1^k) = \pi(u^k, \omega^k) - \Pi \zeta_1^k \end{aligned} \tag{4.36}$$

and A, B, C, and D are given by

$$\begin{aligned} A &:= \begin{bmatrix} 1 - \alpha\left(m_f + m_h \Pi^\top \Pi\right) & 0 \\ 0 & (1 - \alpha\mu)\mathbf{I} \end{bmatrix} \\ B &:= \alpha \begin{bmatrix} -\mathbf{I} & -\frac{1}{\mu}\Pi^\top & -\Pi^\top & -m_h \Pi^\top \\ 0 & \mathbf{I} & 0 & 0 \end{bmatrix} \\ C &:= \begin{bmatrix} \mathbf{I} & 0 \\ \Pi & \mu\mathbf{I} \\ \Pi & 0 \end{bmatrix} \\ D &:= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I} \\ 0 & 0 & 0 & \mathbf{I} \end{bmatrix}. \end{aligned}$$

The block matrix for this system is illustrated in Figure 4.1. We see that for a range of α values, the A matrix is Schur stable. Notice that $\Delta_i(\zeta_i)$ for $i \in \{1, 2, 3\}$ has been designed such that they are Lipschitz continuous and 0-strongly convex. For $i \in \{1, 2, 3\}$, let Δ_i be L_i-Lipschitz continuous where $L_1 = L_f - m_f$, $L_2 = 1$, and $L_3 = L_h - m_h$. Using these results, an IQC can be constructed for the disturbance channels. Using Lemma 4.37, for all outputs, ζ_i and $\bar{\zeta}_i$, and disturbances, $\gamma_i = \Delta_i(\zeta_i)$ and $\bar{\gamma}_i = \Delta_i(\bar{\zeta}_i)$, the following inequality holds

$$\begin{bmatrix} \zeta_i - \bar{\zeta}_i \\ \gamma_i - \bar{\gamma}_i \end{bmatrix}^\top \begin{bmatrix} 0 & L_i\mathbf{I} \\ L_i\mathbf{I} & -2 \end{bmatrix} \begin{bmatrix} \zeta_i - \bar{\zeta}_i \\ \gamma_i - \bar{\gamma}_i \end{bmatrix} \geq 0.$$

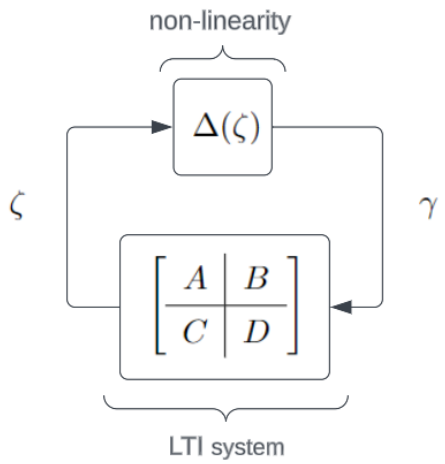

Figure 4.1: Illustration of the block diagram for (4.35)

The final disturbance channel is not in the family of strongly monotone and Lipschitz continuous functions. The output γ_4 encodes the difference between the true input-output function and the undisturbed approximate function. Assuming that this disturbance is strongly monotone is an unnecessarily restrictive property. Only a Lipschitz assumption will be placed on the disturbance channel. To motivate the bound, consider the Taylor series approximation around ζ evaluated at $\bar{\zeta}$,

$$\begin{aligned} &\pi(\bar{\zeta}_1, \omega) = \pi(\zeta_1, \omega) + \frac{\partial \pi}{\partial \zeta}(\zeta_1)(\bar{\zeta}_1 - \zeta_1) + o(\zeta_1, \bar{\zeta}_1) \\ \iff &\pi(\bar{\zeta}_1, \omega) - \Pi\bar{\zeta}_1 - (\pi(\zeta_1, \omega) - \Pi\zeta_1) = (\Pi - \frac{\partial \pi}{\partial \zeta}(\zeta_1))(u - \bar{u}) + o(\zeta, \bar{\zeta}_1) \end{aligned}$$

where $o(\zeta_1, \bar{\zeta}_1)$ are the higher order components. Assume there exists L_{π_1} and L_{π_2} such that $||\Pi - \frac{\partial \pi}{\partial \zeta}(\zeta_1)||_2 \leq L_{\pi_1}$ and $||o(\zeta, \bar{\zeta})||_2 \leq L_{\pi_2}||\zeta - \bar{\zeta}||$, then

$$\begin{aligned} ||\pi(\bar{\zeta}_1, \omega) - \Pi\bar{\zeta}_1 - (\pi(\zeta_1, \omega) - \Pi\zeta_1)|| &= ||(\Pi - \frac{\partial \pi}{\partial \zeta_1}(\zeta_1))(u - \bar{u}) + o(\zeta, \bar{\zeta}_1)|| \\ &\leq (L_{\pi_1} + L_{\pi_2})||\zeta_1 - \bar{\zeta}_1||_2. \end{aligned}$$

This Lipschitz bound encodes the error of the approximate Jacobian and the higher-order terms from the Taylor series expansion. Let $L_{\pi_1} + L_{\pi_2} = L_\pi$, then the inequality can be written as

$$||\gamma_4(\zeta_1) - \gamma_4(\bar{\zeta}_1)||_2 \leq L_\pi ||\zeta^k - \bar{\zeta^k}||_2.$$

This construction can be rewritten in a similar form to the other IQCs,

$$\begin{bmatrix} \zeta_1 - \bar{\zeta}_1 \\ \gamma_4 - \bar{\gamma}_4 \end{bmatrix} \begin{bmatrix} L_\pi^2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \zeta_1 - \bar{\zeta}_1 \\ \gamma_4 - \bar{\gamma}_4 \end{bmatrix} \geq 0.$$

Stacking the constraints, we define the constraint matrix as

$$K := \begin{bmatrix} L_\pi^2 & 0 & 0 & L_1\mathbf{I} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & L_3\mathbf{I} & 0 \\ L_1\mathbf{I} & 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & \mathbf{I} & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & L_3\mathbf{I} & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}. \tag{4.37}$$

With the system matrices and constraint matrix defined, we are able to state the main result.

Theorem 4.3.1. *Consider Algorithm (4.10), where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$ belong to a family of functions that are strongly convex with parameters m_f and m_h, respectively, and gradient Lipschitz continuous with parameters L_f and L_h, respectively. Further, assume that $\Pi \in \mathbb{R}^{m\times n}$ is full row-rank. Let $\hat{z}^k, \hat{y}^k \in \mathbb{R}^{n+m}$ be any trajectories generated by (4.10) for initial conditions $\hat{z}^0, \hat{y}^0 \in \mathbb{R}^{n+m}$. For step size $\alpha > 0$, if there exists $P \in S^n$, $P \succ 0$, $\tau \geq 0$, and $\rho \in (0, \frac{1}{\sqrt{cond(P)}})$ such that*

$$\begin{bmatrix} A^\top PA - \rho^2 P & A^\top PB \\ B^\top PA & B^\top PB \end{bmatrix} + \tau \begin{bmatrix} C^\top & 0 \\ D^\top & \mathbf{I} \end{bmatrix} K \begin{bmatrix} C & D \\ 0 & \mathbf{I} \end{bmatrix} \prec 0$$

where A, B, C, and D are defined in (4.3) and K is defined in (4.37), then

$$||\hat{z}^k - \hat{y}^k||_2 \leq \left(\sqrt{\text{cond}(P)}\right)^k \rho ||\hat{z}^0 - \hat{y}^0||. \tag{4.38}$$

Proof. To prove the theorem, we adapt the proof of Theorem 4 in [10]. We first will prove that (4.35) is contractive which will allow us to conclude that (4.10) is contractive. Consider any $z^k, y^k \in \mathbb{R}^{n+m}$ and Δ as defined in (4.36). Multiply both sides of the LMI by $[z^k - y^k,\ \Delta(z^k) - \Delta(y^k)]$,

$$\begin{aligned}(z^{k+1} - y^{k+1})^\top P(z^{k+1} - y^{k+1}) - \rho^2 (z^k - y^k)^\top P(z^k - y^k) \\ + \tau \begin{bmatrix} \zeta^k - \phi^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix}^\top K \begin{bmatrix} \zeta^k - \phi^k \\ \Delta(z^k) - \Delta(y^k) \end{bmatrix} < 0\end{aligned} \tag{4.39}$$

where ζ^k and ϕ^k are the outputs of (4.35) generated by $(z^k, \Delta(z^k))$ and $(y^k, \Delta(y^k))$, respectively, and z^{k+1}, y^{k+1} are the updates of (4.35). Because the final quadratic term is always positive due to Lemma 3.3.1, (4.39) can be simplified to

$$||z^{k+1} - y^{k+1}||_{P,2}^2 \leq \rho^2 ||z^k - y^k||_{P,2}^2. \tag{4.40}$$

To prove that (4.10) converges, (4.35) must be contractive with respect to the Euclidean norm. Using the properties of matrix norms, we know that (4.40) is equivalent to

$$\begin{aligned} &||z^{k+1} - y^{k+1}||_2^2 \leq \text{cond}(P)\rho^2 ||z^k - y^k||_2^2 \\ \iff &||z^{k+1} - y^{k+1}||_2 \leq \sqrt{\text{cond}(P)}\rho ||z^k - y^k||_2 \end{aligned}$$

Because $\sqrt{\text{cond}(P)}\rho < 1$ by assumption, we see that (4.35) is contractive with respect to the norm

$||\cdot||_2$. Using Lemma 4.2.1, we can conclude that (4.10) is contractive with rate $\sqrt{\text{cond}(P)}\rho$. Thus, take any $\hat{z}^k, \hat{y}^k \in \mathbb{R}^{n+m}$ generated by (4.10) with initial conditions $\hat{z}^0, \hat{y}^0 \in \mathbb{R}^{n+m}$, from Banach's Contraction Theorem [20], we have

$$||\hat{z}^k - \hat{y}^k||_2 \leq \sqrt{\text{cond}(P)}\rho||\hat{z}^0 - \hat{y}^0||_2. \tag{4.41}$$

□

The following theorem provides a numerical method for determining the feasibility of the algorithm for a class of functions and a given step size. Algorithm users can determine the feasibility of the LMI to ensure that the algorithm can be used with their system for a given α. This method provides guarantees for a single step size rather than a range. Notice that if $L_\pi = 0$ and $L_3 = 0$, the algorithm can be simplified to the algorithm in [9]. Thus, by continuity, for enough L_3 and L_π, the step size guarantees in [9] will guarantee convergence for (4.10).

Chapter 5

Distribution Feeder Simulation

5.1 Simulation Introduction

The algorithm will be tested on a simulated distribution network with a large number of photovoltaic generators. The implemented algorithm will aim to keep the relevant system states within their physical safety constraints while optimizing the system's power usage. The system considered is a variation of the IEEE 37-node test feeder system as proposed in [6]. The system is a single-phase AC system and uses 10 hours of real solar irradiance and load data collected from Anatolia, California. A diagram of the system of interest is given in Figure 5.1.

The input of the algorithm to the distribution feeder system is the reactive and active power injections for the controllable nodes. Let $u \in \mathbb{R}^{36}$ denote the input vector. Each node has an associated input pair $[p_i, q_i] \in \mathbb{R}^2$ where p_i is the active power injection at node i and q_i is the reactive power injection. At each controllable node, the inverter has a maximum apparent power rating and an available amount of power that can be injected. These constraints shape the input set $\mathcal{U}$ as

$$\mathcal{U}_i = \{[p_i,\ q_i] \in \mathbb{R}^2 |\ 0 \leq p_i \leq p_i^{\max},\ q_i^2 + p_i^2 \leq s_i^{\text{rated}}\}$$
$$\mathcal{U} = \times_{i=1}^{18} \mathcal{U}_i$$

where s_i^{rated} is the apparent power rating for inverter i and $p_i^{\max}$ is the maximum available power for inverter i. Note that s_i^{rated} will remain constant throughout the process, but $p_i^{\max}$ is a function of the solar irradiance data. The measurements of the system are the voltage magnitudes at each node except the point of common coupling. Let $y \in \mathbb{R}^{35}$ represent the output vector. The voltage limits are uniform across nodes and given by $v_l = 0.95$ p.u and $v_u = 1.05$ p.u. The output set is represented as

$$\mathcal{Y} = \{y \in \mathbb{R}^{35} | v_l \mathbf{1} \leq y \leq v_u \mathbf{1}\}.$$

The approximate Jacobian is calculated in [6] and will be represented by $\Pi \in \mathbb{R}^{35 \times 36}$. The uncontrollable reactive and active loads and power injections are designated as disturbances to the system. Let $\omega \in \mathbb{R}^{70}$ represent the disturbances. The algorithm will aim to control and optimize the system subject to these changing loads. The controller will constrain the inputs and outputs of the system to $\mathcal{U}$ and $\mathcal{Y}$, respectively, while minimizing the amount of power curtailed. Formally, this can be

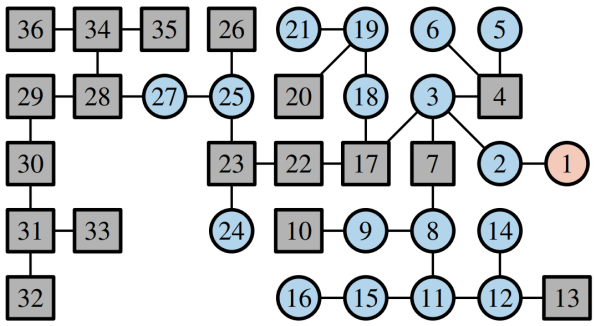

Figure 5.1: IEEE 37-node feeder [5]. Node 1 is the point of common coupling. All other nodes have sensors which provide load and voltage data. The circular blue nodes have controllable power injections while the other nodes are uncontrollable.

written as

$$\min_{u\in\mathbb{R}36, y\in\mathbb{R}35} \; ||u - u^{\text{ref}}||_2^2 + I_{\mathcal{U}}(u) + I_{\mathcal{Y}}(y)$$

where $u_i^{\text{ref}} = p_i^{\max}$ if u_i is an active power input and $u_i^{\text{ref}} = 0$ if u_i is a reactive power input. $I_C(x)$ is the indicator for the set C which is defined as

$$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}.$$

5.2 Simulation Implementation and Results

To implement the algorithm, the proximal operator of the indicator function must be determined,

$$\begin{aligned} \mathbf{prox}_{\mu I_C}(v) &= \text{argmin}_{x\in\mathbb{R}^n} \left(I_C(x) + \frac{1}{2\mu}||x - v||_2^2 \right) \\ &= \text{argmin}_{x\in C}||x - v||_2^2. \end{aligned}$$

Thus, the proximal operator of the indicator function on the set C is the projection operator on the set C. The efficiency of implementing the projection operator into the algorithm depends on the complexity of the set. For $\mathcal{Y}$, it is extremely easy to implement. If $y_i < v_l$, then $\mathbf{prox}_{\mu I_C}(y_i) = v_l$ and if $y_i > v_u$, then $\mathbf{prox}_{\mu I_C}(y_i) = v_u$. Implementing the projection operator for $\mathcal{U}$ is slightly more complicated and is described in [6].

The results of the simulation can be seen in Figure 5.2. Figure 5.2 tracks the voltage for each node in the system, $y \in \mathbb{R}^{35}$. The top graph shows the evolution of the system when using the control algorithm. The bottom graph reveals the natural evolution of the system. As one can see, the algorithm effectively suppresses the output voltage below the maximum value. When the voltage does not need to be limited, the algorithm tracks the uncontrolled system and uses all of the available power. This is all accomplished while constraining the active and reactive power injections.

We can compare this algorithm against other similar algorithms. We see that the algorithm proposed in [8] does not offer hard constraints on the inputs of the system. This requires the user to

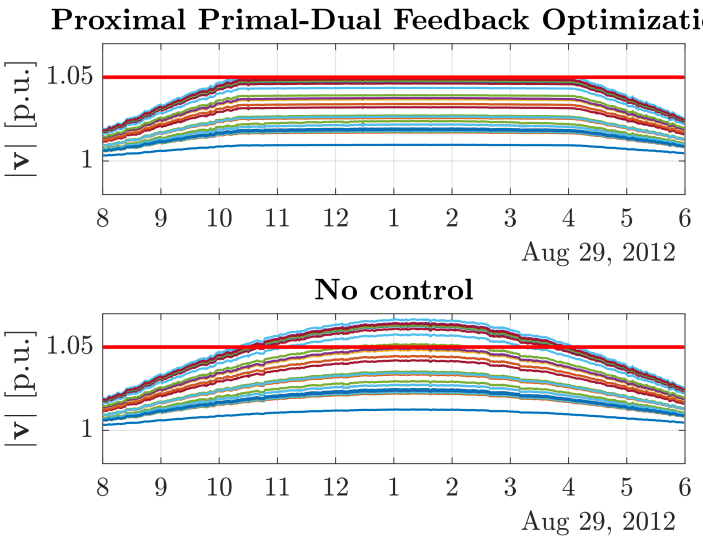

Figure 5.2: Simulation results using $\mu = 0.1$ and $\alpha = 0.004$

introduce soft constraints on the input to encourage the algorithm to stay within the input set. This can lead to potential constraint violations. Similarly, the algorithm proposed in [5] provides hard constraints on the inputs, but not on the outputs. Thus, soft constraints must be used, introducing potential violations of the output set. The algorithm proposed in this thesis has hard constraints on both, ensuring the algorithm keeps the input and output within their respective sets. To illustrate this, the same simulation was performed with the algorithm from [5] using the implementation in the paper. The total output constraint violation for a single node throughout the simulation is plotted in Figure 5.3. As we can see, during the period when the power is being curtailed, the soft

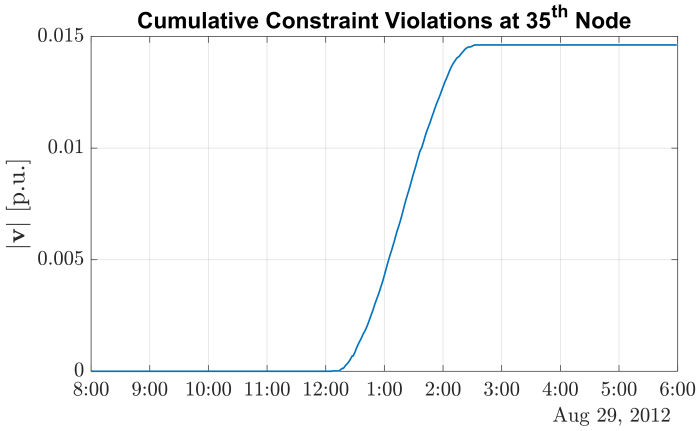

Figure 5.3: Cumulative constraint violations of the output set at node 35 using $\alpha = 0.004$ with the algorithm from [5]

constraints are not able to keep the output within the output set. The algorithm proposed in this thesis has no constraint violations at node 35 throughout the simulation.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we study the convergence rates for the primal-dual algorithm for convex equality-constrained optimization problems. Further, we develop an online feedback-based optimization controller for a general non-linear system.

There are three key contributions in this thesis. We first provide three different methods for analyzing the convergence rate of the primal-dual algorithm. These rates were compared and the benefits of each approach were discussed. Further, the TSSPD algorithm was introduced. The motivation behind the algorithm was discussed and a convergence rate was provided.

The second contribution was the OFBO controller that was developed. We proposed a dynamic OFBO controller that aimed to optimize a general cost function. The cost function was able to consider differentiable and non-differentiable costs on the inputs and outputs of the system. This framework allows users to enforce hard constraints on the inputs and outputs. Two convergence proofs were provided. The first considered a known, linear plant and determined sufficient conditions to ensure the algorithm is exponentially convergent. The second approach allowed for the plant to be unknown and non-linear. This proof involved developing a linear matrix inequality to determine if the algorithm is exponentially convergent for a given step size.

The final contribution involved testing the OFBO controller in a simulated environment. The controller was configured to control a simulated distribution feeder. The simulation was implemented and tested in MATLAB.

6.2 Future Research

The following are potential future research directions:

- For the time-scale separated primal-dual algorithm proposed in Chapter 3, other proof approaches should be considered. The convergence rate approached half of the optimal rate of the dual method as the time-scale separation approached infinity. Developing a proof in which

the rate approaches the optimal rate of the dual method would be an area of interest. The motivation behind the algorithm, as discussed in Chapter 3, suggests this should be possible.

- The stability analysis of the interconnected controller and system is an area of future interest. In this thesis, we assume that if the step size of the controller is small enough, the interconnection between the exponentially stable plant and the controller remains exponentially stable. Formalizing this idea mathematically and providing explicit bounds would be helpful for the algorithm users and remove the need for this assumption.

- A convergence analysis for a range of step size values for the feedback controller when the mapping is non-linear would be an interesting extension to the results in this thesis. In [9], the KYP Lemma is used to guarantee convergence to a fixed point for a range of step sizes. Performing a similar analysis for the algorithm in this thesis would be more difficult, due to the increased number of parameters, but would provide interesting results.

Appendix A

Chapter 4 Inequalities

Proof of $\frac{\epsilon}{2}\Pi^\top(D\bar{\Pi}+\bar{\Pi})\Pi \preceq \epsilon\mathbf{I}$

To show this inequality, note that the eigenvalues of D are between 0 and 1. Because D is symmetric, there exists V such that $V^\top V = VV^\top = \mathbf{I}$ and $\Sigma = V^\top DV$ where Σ is a diagonal matrix composed of the eigenvalues of D. Let $\Gamma = V^\top\bar{\Pi}V$, then

$$\begin{aligned}\Pi^\top(D\bar{\Pi}+\bar{\Pi})\Pi =& \Pi^\top V(V^\top DVV^\top\bar{\Pi}V + V^\top\bar{\Pi}VV^\top DV)\Pi^\top V\\ =& \Pi^\top V(\Sigma\Gamma+\Gamma\Sigma)\Pi^\top V\end{aligned}$$

Since Σ is a convex combination of diagonal matrices with 0 and 1 entries, it is sufficient to consider $\Gamma R + R\Sigma$ where R is diagonal with the entries $r_i = 0$ or $r_i = 1$ [11]. Without loss of generality, consider $r_1 = \ldots = r_j = 1$ and $r_{j+1} = \ldots = r_m = 0$ and

$$\Gamma = \begin{bmatrix}\Gamma_1 & \Gamma_0^\top\\ \Gamma_0 & \Gamma_2\end{bmatrix}$$

Where $\Gamma_1 \in S^j$ and $\Gamma_2 \in S^{m-j}$. Then

$$\begin{aligned}R\Gamma + \Gamma R - 2\Gamma = \begin{bmatrix}0 & -\Gamma_0^\top\\ -\Gamma_0 & -2\Gamma_2\end{bmatrix}\\ \preceq \mathbf{0}\end{aligned}$$

Thus $\Gamma\Sigma + \Sigma\Gamma \preceq 2\Gamma$. So

$$\begin{aligned}\Pi^\top V(\Sigma\Gamma+\Gamma\Sigma)V^\top Pi &\preceq 2\Pi^\top V(\Gamma)V^\top\Pi\\ &= 2\Pi^\top\bar{\Pi}\Pi\\ &\preceq 2\mathbf{I}\end{aligned}$$

Therefore, we have shown that

$$\frac{\epsilon}{2}\Pi^\top(D\bar{\Pi}+\bar{\Pi})\Pi \preceq \epsilon\mathbf{I}$$

Proof that $\epsilon D + \mu(D - \mathbf{I}) \succeq \epsilon$

To show this, notice that $\mu \geq \epsilon$ and $\mathbf{0} \preceq D \preceq \mathbf{I}$, then

$$\begin{aligned}\epsilon D + \mu(\mathbf{I} - D) &= \mu\mathbf{I} - (\mu - \epsilon)D \\ &\succeq \mu\mathbf{I} - (\mu - \epsilon)\mathbf{I} \\ &= -\epsilon\mathbf{I}\end{aligned}$$

This completes the proof.

Proof that $\Pi^\top \bar{\Pi}^2 \Pi \preceq \frac{1}{\sigma^2_{\min}(\Pi)}\mathbf{I}$

Consider the singular value decomposition of $\Pi = U\Sigma V^\top$. Then

$$\begin{aligned}\Pi^\top(\bar{\Pi})^2\Pi &= V\Sigma^\top U^\top (U(\Sigma^\top\Sigma)^{-1}U^\top U(\Sigma^\top\Sigma)^{-1}U^\top)U\Sigma V^\top \\ &= V\Sigma^\top((\Sigma\Sigma^\top)^{-1})^2\Sigma V^\top\end{aligned}$$

We know that these matrices are constructed as

$$\Sigma = \begin{bmatrix}\text{Diag}(\sigma_1(\Pi), \ldots, \sigma_m(\Pi)) & \mathbf{0}_{m\times n-m}\end{bmatrix}((\Sigma\Sigma^\top)^{-1})^2 = \begin{bmatrix}\frac{1}{\sigma_1^4(\Pi)} & \cdots & 0 \\ 0 & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_m^4(\Pi)}\end{bmatrix}$$

Evaluating $\Sigma^\top((\Sigma\Sigma^\top)^{-1})^2\Sigma$ simplifies to a block matrix with $(\Pi\Pi^\top)^{-1}$ in the top left corner and zero everywhere else.

$$\Sigma^\top((\Sigma\Sigma^\top)^{-1})^2\Sigma = \begin{bmatrix}(\Sigma\Sigma^\top)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

This block matrix clearly has the same eigenvalues as $(\Pi\Pi^\top)^{-1}$ with an additional eigenvalue of 0. Thus, we can bound the matrix of interest.

$$\Pi^\top(\bar{\Pi})^2\Pi = V\Sigma^\top((\Sigma\Sigma^\top)^{-1})^2\Sigma V^\top \preceq \frac{1}{\sigma^2_{\min}(\Pi)}\mathbf{I}$$

Bibliography

[1] Patricio Salmerón Revuelta, Salvador Pérez Litrán, and Jaime Prieto Thomas. "8 - Distributed Generation". In: *Active Power Line Conditioners*. Ed. by Patricio Salmerón Revuelta, Salvador Pérez Litrán, and Jaime Prieto Thomas. San Diego: Academic Press, 2016, pp. 285–322. ISBN: 978-0-12-803216-9. DOI: https://doi.org/10.1016/B978-0-12-803216-9.00008-0. URL: https://www.sciencedirect.com/science/article/pii/B9780128032169000080.

[2] Toshihisa Funabashi. "Chapter 1 - Introduction". In: *Integration of Distributed Energy Resources in Power Systems*. Ed. by Toshihisa Funabashi. Academic Press, 2016, pp. 1–14. ISBN: 978-0-12-803212-1. DOI: https://doi.org/10.1016/B978-0-12-803212-1.00001-5. URL: https://www.sciencedirect.com/science/article/pii/B9780128032121000015.

[3] Barry Hayes. "Distribution generation optimization and energy management". In: *Distributed Generation Systems* (2017), 415–451. DOI: 10.1016/b978-0-12-804208-3.00009-1.

[4] P.P. Barker and R.W. De Mello. "Determining the impact of distributed generation on Power Systems. I. Radial Distribution Systems". In: *2000 Power Engineering Society Summer Meeting (Cat. No.00CH37134)* (). DOI: 10.1109/pess.2000.868775.

[5] Marcello Colombino, John W. Simpson-Porco, and Andrey Bernstein. "Towards robustness guarantees for feedback-based optimization". In: *2019 IEEE 58th Conference on Decision and Control (CDC)*. 2019, pp. 6207–6214. DOI: 10.1109/CDC40024.2019.9029953.

[6] Emiliano Dall'Anese and Andrea Simonetto. "Optimal Power Flow Pursuit". In: *IEEE Transactions on Smart Grid* 9.2 (Mar. 2018). ISSN: 1949-3053. DOI: 10.1109/TSG.2016.2571982. URL: https://www.osti.gov/biblio/1423552.

[7] Andrey Bernstein, Emiliano Dall'Anese, and Andrea Simonetto. *Online Primal-Dual Methods with Measurement Feedback for Time-Varying Convex Optimization*. 2019. arXiv: 1804.05159 [math.OC].

[8] Marcello Colombino, Emiliano Dall'Anese, and Andrey Bernstein. "Online optimization as a feedback controller: Stability and tracking". In: *IEEE Transactions on Control of Network Systems* 7.1 (2020), 422–432. DOI: 10.1109/tcns.2019.2906916.

[9] Dongsheng Ding et al. "An Exponentially Convergent Primal-Dual Algorithm for Nonsmooth Composite Minimization". In: *2018 IEEE Conference on Decision and Control (CDC)*. 2018, pp. 4927–4932. DOI: 10.1109/CDC.2018.8619760.

[10] Laurent Lessard, Benjamin Recht, and Andrew Packard. "Analysis and design of optimization algorithms via integral quadratic constraints". In: *SIAM Journal on Optimization* 26.1 (2016), 57–95. DOI: 10.1137/15m1009597.

[11] Dongsheng Ding and Mihailo R. Jovanović. "Global exponential stability of primal-dual gradient flow dynamics based on the proximal augmented Lagrangian". In: *2019 American Control Conference (ACC)*. 2019, pp. 3414–3419. DOI: 10.23919/ACC.2019.8815153.

[12] Simon S. Du and Wei Hu. *Linear Convergence of the Primal-Dual Gradient Method for Convex-Concave Saddle Point Problems without Strong Convexity*. 2019. arXiv: 1802.01504 [math.OC].

[13] David C. Lay, Steven R. Lay, and Judith McDonald. *Linear algebra and its applications*. 5th ed. Pearson Education Limited, 2022.

[14] John W. Simpson-Porco. *Lecture notes in Robust and Optimal Control*. https://www.control.utoronto.ca/~jwsimpson/robust/ECE1659H-Instructor-1x1.pdf/. 2023.

[15] A. Auslender and M. Teboulle. *Asymptotic cones and functions in optimization and variational inequalities*. Springer, 2011.

[16] Stephen Boyd and Lieven Vandenberghe. "Convex Optimization". In: *Applied and Computational Mathematics* 15.1 (2016), pp. 3–43.

[17] Hossein Pishro-Nik. *Jensen's Inequality*. URL: https://www.probabilitycourse.com/chapter6/6_2_5_jensen.

[18] Ernest K. Ryu and Stephen Boyd. "A Primer on Monotone Operator Methods". In: *Applied and Computational Mathematics* 15.1 (2016), pp. 3–43.

[19] Wassim M. Haddad and VijaySekhar Chellaboina. *Nonlinear Dynamical Systems and control: A Lyapunov-based approach*. Princeton University, 2008.

[20] F. Bullo. *Contraction Theory for Dynamical Systems*. 1.1. Kindle Direct Publishing, 2023. ISBN: 979-8836646806. URL: https://fbullo.github.io/ctds.

[21] Adil Salim et al. *An Optimal Algorithm for Strongly Convex Minimization under Affine Constraints*. 2022. arXiv: 2102.11079 [math.OC].

[22] Michael Zargham et al. "Accelerated Dual Descent for Network Flow Optimization". In: *IEEE Transactions on Automatic Control* 59.4 (2014), pp. 905–920. DOI: 10.1109/TAC.2013.2293221.

[23] Stephen Boyd and Neal Parikh. "Proximal Algorithms". In: *Foundations in Trends in Optimization* 1.3 (2013), pp. 123–231.

[24] Roger A. Horn and Fuzhen Zhang. "Basic Properties of the Schur Complement". In: *The Schur Complement and Its Applications*. Ed. by Fuzhen Zhang. Boston, MA: Springer US, 2005, pp. 17–46. ISBN: 978-0-387-24273-6. DOI: 10.1007/0-387-24273-2_2. URL: https://doi.org/10.1007/0-387-24273-2_2.

[25] J. W. Simpson-Porco. "Analysis and Synthesis of Low-Gain Integral Controllers for Nonlinear Systems". In: *IEEE Transactions on Automatic Control* 66.9 (Sept. 2021), pp. 4148–4159. DOI: 10.1109/TAC.2020.3035569.