Fall 12-3-2018

Control Theory: The Double Pendulum Inverted on a Cart

Ian J P Crowe-Wright
University of New Mexico

Ian Crowe-Wright

Candidate

Mathematics and Statistics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Jens Lorenz , Chairperson

Stephen Lau

Monika Nitsche

Control Theory: Double Pendulum Inverted on a Cart

by

Ian Crowe-Wright

B.Sc. Mathematics, University of Birmingham, 2017

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Mathematics

The University of New Mexico

Albuquerque, New Mexico

December, 2018

Dedication

To my family, my parents Josephine and Peter and my sister Alice, for their constant support and encouragement.

Acknowledgments

I heartily acknowledge my advisor and thesis committee chair, Dr. Lorenz for encouraging me through this project and giving me all his support. His valuable teachings will stay with me as I continue my career. I also thank committee members, Professor Nitsche and Professor Lau for their valuable feedback pertaining to this thesis and assistance in my professional development.

Control Theory: Double Pendulum Inverted on a Cart

by

Ian Crowe-Wright

B.Sc. Mathematics, University of Birmingham, 2017

M.S. Mathematics, University of New Mexico, 2018

Abstract

In this thesis the Double Pendulum Inverted on a Cart (DPIC) system is modeled using the Euler-Lagrange equation for the chosen Lagrangian, giving a second-order nonlinear system. This system can be approximated by a linear first-order system in which linear control theory can be used. The important definitions and theorems of linear control theory are stated and proved to allow them to be utilized on a linear version of the DPIC system. Controllability and eigenvalue placement for the linear system are shown using MATLAB. Linear Optimal control theory is likewise explained in this section and its uses are applied to the DPIC system to derive a Linear Quadratic Regulator (LQR). Two different LQR controllers are then applied to the full nonlinear DPIC system, which is concurrently modeled in MATLAB. Also, an in-depth look is taken at the Riccati equation and its solutions. Finally, results from various MATLAB simulations are shown.

Contents

List of Figures

List of Tables

Glossary

θ_0	horizontal position of the cart
θ_1	angle between the first pendulum link and the vertical
θ_2	angle between the second pendulum link and the vertical
m_0	mass of the cart
m_1	mass of the first pendulum link
m_2	mass of the second pendulum link
L_1	length of the first pendulum link
L_2	length of the second pendulum link
l_1	distance between the base of the first pendulum link and its center of mass
l_2	distance between the base of the first pendulum link and its center of mass
I_1	moment of inertia for the first pendulum link
I_2	moment of inertia for the second pendulum link
g	constant of acceleration due to gravity

Glossary

$u(t)$	control force
L	Lagrangian
E_{kin}	kinetic energy
E_{pot}	potential energy
(x_0, y_0)	Cartesian coordinates of the cart
(x_1, y_1)	Cartesian coordinates of the top of the first pendulum link
(x_2, y_2)	Cartesian coordinates of the top of the second pendulum link
$\mathbf{x}(t)$	state vector
p	determinant of $\mathbf{D}(\mathbf{0})$
$\mathbf{M}_n$	controllability matrix
$\mathbf{S}_n$	observability matrix
$J(u)$	cost
$\boldsymbol{\lambda}$	adjoint trajectory
$\mathbf{P}_+$	stabilizing solution to the algebraic Riccati equation
$\mathbf{P}_-$	anti-stabilizing solution to the algebraic Riccati equation
$\mathbf{P}_\theta$	a mixed solution to the algebraic Riccati equation
$\mathbf{x}_0$	initial state
t_f	time taken to get to the pendulum-up position
τ	computation time
α	maximum angle of deflection

Chapter 1

Introduction

Dynamical Systems extensively represent natural phenomena occurring anywhere [1]. These systems of ordinary differential equations in terms of the 'state' of the system are modeled by taking in real world data. A model is then built around this data. These models can then be used for prediction [2]. However, it is often required that one goes beyond describing the system of interest and actually actively manipulates the system to change its fundamental behaviour. This is achieved by *Control Theory*.

Control, or how we affect the world around us, has been human nature since civilization began. The earliest works alluding to concepts of control theory are the writings of Aristotle. In book 1 of *Aristotle's Politics* [3], Aristotle states

> "...if every instrument could accomplish its own work, obeying or anticipating the will of others ...if the shuttle weaved and the pick touched the lyre without a hand to guide them, chief workmen would not need servants, nor masters slaves."

This quote describes a motivation for control. A need for automotive processes that ease habitation for mankind.

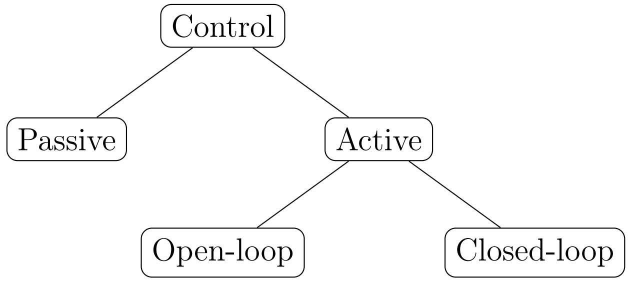

Figure 1.1: Types of Control

1.1 Concepts of Control

Manipulating dynamical systems requires control. Control helps attain a desired outcome for a system. There are many different types of control [4], some of these control types are shown in Figure 1.1. Here a few common controls are explained along with a specific control type that will be used multiple times in this thesis.

1.1.1 Types of Control

Passive Control

The most basic and common type of control is *passive control*. Passive control is a control law that passively effects the system, it cannot be changed or altered as it is constant. For example, in automobile design, the aerodynamic body of a car is placed upon the chassis in such a way that the air moves favourably around the vehicle and allows for a smoother and faster ride by minimizing drag. A significant advantage of passive control is that it is predetermined, there is no energy expenditure.

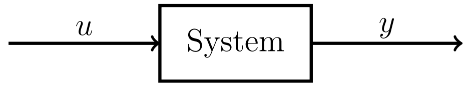

Figure 1.2: Open-loop Control

Active Control

However, passive control alone is typically not enough. Often, to get a desired outcome or effect, *active control* is also required. Active control is essentially control where energy is actively pumped into the system to actively manipulate its behaviour. Active control has many forms. The most recurrent type of active control is *open-loop control*.

Open-loop Control

Open-loop control, on a system with inputs u and outputs y, aims to Figure out exactly what is the perfect u to get a specific desired y. Toasting bread is a simple but perfect example of this control type, time toasted is the input u and preferred shade of toast, say 'light brown and slightly crisp', is the output y. The downside to this specific type of active control is that the energy cost is constant. It can sometimes be very expensive.

Closed-loop Control

The next form of active control builds on open-loop control. It uses sensors to measure the current output y, this measurement is fed into a 'controller' and fed back into the input signal u. This is called *closed-loop control*, or *closed-loop feedback control*. Essentially, the output is measured and 'feeds back' through a controller into the input and the system. In this thesis, this method of control will eventually

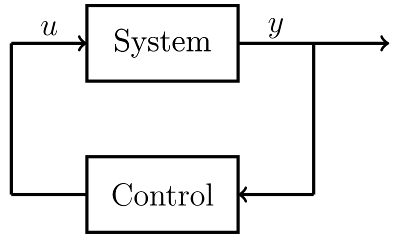

Figure 1.3: Closed-loop Control

be used on the *Double Pendulum Inverted on a Cart* (DPIC) system derived in the following sections.

1.1.2 Closed-loop Control

Why Closed-loop Control?

Closed-loop feedback control has many benefits over open-loop. Any pre-planned trajectory for an input u will be sub-optimal if there is any inherent uncertainty in the system. Feedback measures this and appropriately adjusts the controller, even if the system itself is not perfect. Moreover, open-loop control cannot change the fundamental behaviour of the system itself, an unstable system will stay unstable. However, as this thesis will show, feedback control can directly manipulate the dynamics of the system and stabilize the eigenvalues. Furthermore, feedback control immediately accounts for external disturbances to the system, a gust of wind to a pendulum will be measured and corrected in the feedback loop. The final advantage is energy cost or efficiency. This concept will be explored more in the *Optimal Control* section, section 3.

Review of Linear Systems

Before diving further into control theory it is necessary to review the required mathematical theory. Throughout this thesis, select topics from linear systems and linear algebra will be reviewed in order to fully understand that specific section. At this point, solutions of linear systems and the stability of eigenvalues are recapped to help show why closed-loop feedback can eradicate instability in a system, see [5] for more. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$. Recall that the linear system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}, \quad \mathbf{x}(0) = \mathbf{x}_0$$

has solution

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0$$

for $\mathbf{x} \in \mathbb{R}^n$. Assume that $\mathbf{A}$ can be expressed in diagonal form

$$\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}$$

where $\mathbf{T}, \mathbf{D} \in \mathbb{R}^{n \times n}$. Let λ_i be eigenvalues and $\mathbf{v}_i$ be the eigenvectors of $\mathbf{A}$, here $i = 1, \ldots, n$. Here

$$\mathbf{T} = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_n], \quad \mathbf{D} = Diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Using the series representation of $e^{\mathbf{A}t}$ implies

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_0 = \mathbf{T}e^{\mathbf{D}t}\mathbf{T}^{-1}\mathbf{x}_0$$

If all the eigenvalues of $\mathbf{A}$ are in the negative half plane, $Re(\lambda) < 0$ for all λ, the system is stable. If just one of the eigenvalues has non-negative real part, $Re(\lambda) \geq 0$ for some λ, then the system is unstable. Figure 1.4 shows this. Essentially, control theory seeks to stabilize system eigenvalues.

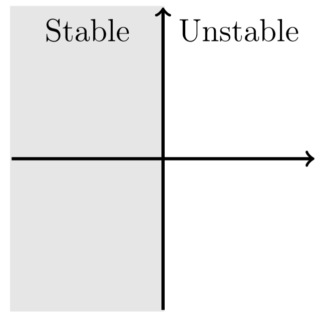

Figure 1.4: Eigenvalue Stability

Motivation for Closed-loop Control

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$. Consider the dynamical system, this system will be explained in section 2.1.

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

for $\mathbf{x}(t) \in \mathbb{R}^n$ and $\mathbf{u}(t) \in \mathbb{R}^m$. Suppose a controller is applied

$$\mathbf{u} = -\mathbf{K}\mathbf{x}$$

where $\mathbf{K} \in \mathbb{R}^{m\times n}$. Ways to find $\mathbf{K}$ will be shown in sections 2.2.4, 3.1.5 and 3.2.1. The system becomes

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\mathbf{x}$$

Closed-loop Control seeks to give all the eigenvalues of $(\mathbf{A} - \mathbf{B}\mathbf{K})$ negative real parts such that the controlled system is a stable system. $(\mathbf{A}, \mathbf{B})$ is called a *stabilizable pair* if there exists such a $\mathbf{K}$.

Stabilizability

Definition 1.1.1 (Stabizable Pair)**.** $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$. The pair of matrices $(\mathbf{A}, \mathbf{B})$ is *stabilizable* if there exists $\mathbf{K} \in \mathbb{R}^{m\times n}$ such that all the eigenvalues of $\mathbf{A} - \mathbf{B}\mathbf{K}$ have negative real part.

1.2 System

The Double Pendulum Inverted on a Cart (DPIC) system is a pendulum system with two joints, which typically start suspended vertically in the air, the pendulum-up position. The pendulums are connected to a cart which can be powered externally to move horizontally and manipulate the system. This is our controller. See Figure 1.5 for a diagram. This problem has been used extensively to test linear and nonlinear control laws [6], see some of this research in [6],[7] and [8].

1.2.1 Modeling

The system is modeled so that the center of the cart begins at $(0, 0)$, in Cartesian coordinates. Following the notation in [9], m_0 is the mass of the cart, m_1, m_2 are the masses of the first and second pendulum link. θ_0 is the horizontal position of the cart, θ_1, θ_2 are the angles between each pendulum link and the vertical. l_1, l_2 are the distances from the base of each pendulum link to the center of mass for that pendulum link, L_1, L_2 are the lengths of the each pendulum link. I_1, I_2 are the moments of inertia of each pendulum link with respect to its center of mass. g is the gravitational constant and the force $u(t)$ is the control. Each pendulum link can move through all 360°. When released from suspension the DPIC will collapse, oscillating indefinitely. If friction is present, the DPIC will decelerate over time and converge to the only stable fixed point, the pendulum-down position, $\theta_1 = \theta_2 = \pi$. This occurs when released from any combination of initial angles unless control is used. However, the DPIC system provides a difficult problem with regard to control theory, there is only a single controlling force $u(t)$ and three degrees of freedom $\theta_0, \theta_1, \theta_2$. The control force $u(t)$ acts only horizontally on the cart through a motor.

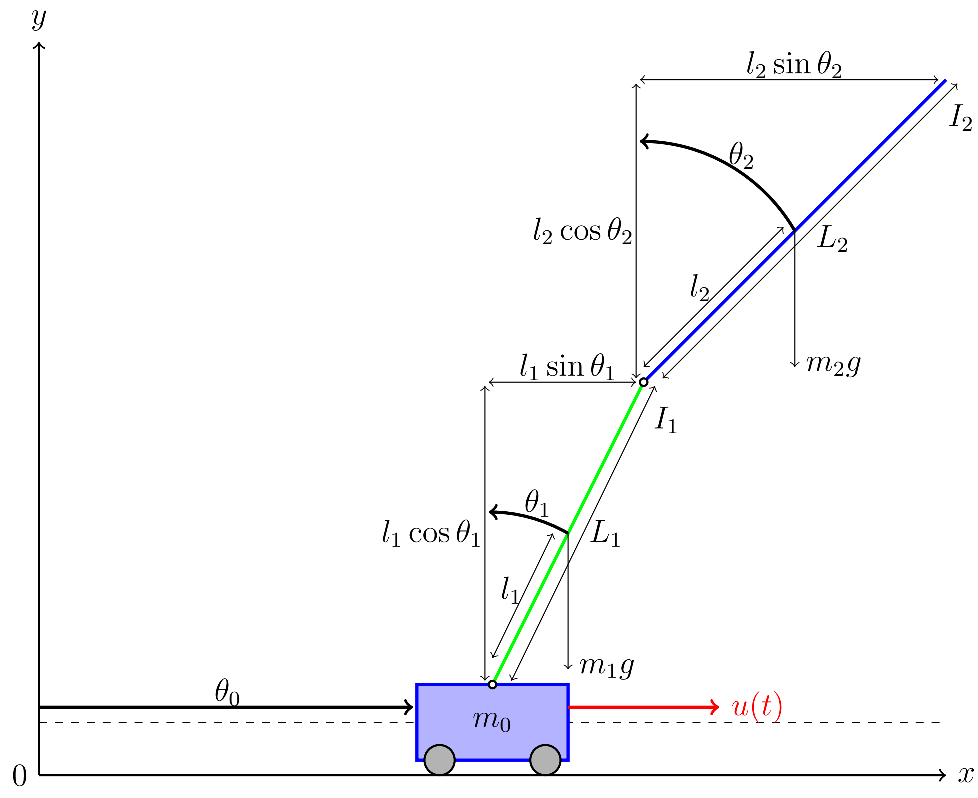

Figure 1.5: Double Pendulum Inverted on a Cart (DPIC)

1.2.2 Derivation via Lagrange

Next, it is necessary to calculate the equations of motion for the DPIC system. There are multiple methods to derive the system of equations modeling the DPIC system. For example, a Newtonian approach regarding forces would eventually calculate the system equations. However, it is most elegant and efficient to use the Lagrange equations for a particular Lagrangian [10].

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}}\right) - \frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{q} \tag{1.1}$$

L is the Lagrangian, $\mathbf{q}$ is a vector of generalized forces which act in the direction of each component in θ, these forces are not included in the kinetic and potential energies of the cart and each pendulum link. The control force $u(t)$ is one of these forces. This method of derivation is similar to that of [9] but explains it in full. Let

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$$

Recall that θ_0 is the horizontal position of the cart and θ_1 and θ_2 are the angles of the first and second pendulum links to the vertical. Because of these components, the control force $u(t)$, which is the force acting on the cart, only acts horizontally and does not affect the angles directly. Negating other external forces, $\mathbf{q}$ is simply

$$\mathbf{q} = \begin{pmatrix} u(t) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} u(t) = \mathbf{H}u(t)$$

Thus, looking at the Lagrange equation (1.1) for each component of θ the following system of equations is obtained.

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_0}\right) - \frac{\partial L}{\partial \theta_0} = u(t)$$
$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} = 0$$
$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_2}\right) - \frac{\partial L}{\partial \theta_2} = 0$$

The Lagrangian, L is the difference between the total kinetic energy, E_{kin}, and the total gravitational potential energy, E_{pot} of the system.

$$L = E_{kin} - E_{pot} \tag{1.2}$$

These energies can be broken down into energies of the specific components. The cart, the first pendulum link, and the second pendulum link.

$$E_{kin} = E_{kin}^{(0)} + E_{kin}^{(1)} + E_{kin}^{(2)}$$
$$E_{pot} = E_{pot}^{(0)} + E_{pot}^{(1)} + E_{pot}^{(2)}$$

Here the subscript notation denotes either kinetic or potential energy and the superscript indicates which component of the system is being referred to. $E_{kin}^{(0)}$ is the kinetic energy of the cart and $E_{pot}^{(2)}$ is the gravitational potential energy of the second, or top, pendulum link. Using Figure 1.5, the specific coordinates of each component in the DPIC system can be calculated, this will help calculate the Lagrangian (1.2). For the position of the cart one has

$$x_0 = \theta_0$$
$$y_0 = 0$$

The position of the midpoint of the first pendulum link

$$x_1 = \theta_0 + l_1 \sin\theta_1$$
$$y_1 = l_1 \cos\theta_1$$

The position of the midpoint of the second pendulum link

$$x_2 = \theta_0 + L_1 \sin\theta_1 + l_2 \sin\theta_2$$
$$y_2 = L_1 \cos\theta_1 + l_2 \cos\theta_2$$

Using these coordinates the energy components can be calculated. First, using standard results from Newtonian mechanics [11], calculate the kinetic and potential energy, $E_{kin}^{(0)}$ and $E_{pot}^{(0)}$, for the cart

$$E_{kin}^{(0)} = \frac{1}{2} m_0 \dot{\theta}_0^2$$
$$E_{pot}^{(0)} = 0$$

Next look at the bottom pendulum link. Due to modeling each pendulum link with the center of mass at the midpoint of that link, the kinetic energy has two components. These are translational kinetic energy and rotational kinetic energy [11].

$$
\begin{aligned}
E_{kin}^{(1)} &= E_{kin}^{(1)}(trans) + E_{kin}^{(1)}(rot) \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\big(\dot{x}_1^2 + \dot{y}_1^2\big) + \frac{1}{2}I_1\dot{\theta}_1^2 \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\Bigg\{\Big(\frac{d}{dt}\big[\theta_0 + l_1\sin\theta_1\big]\Big)^2 + \Big(\frac{d}{dt}\big[l_1\cos\theta_1\big]\Big)^2\Bigg\} + \frac{1}{2}I_1\dot{\theta}_1^2 \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\Big[\big(\dot{\theta}_0 + l_1\dot{\theta}_1\cos\theta_1\big)^2 + \big(-l_1\dot{\theta}_1\sin\theta_1\big)^2\Big] + \frac{1}{2}I_1\dot{\theta}_1^2 \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\big[\dot{\theta}_0^2 + 2l_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1 + l_1^2\dot{\theta}_1^2\cos^2\theta_1 + l_1^2\dot{\theta}_1^2\sin^2\theta_1\big] + \frac{1}{2}I_1\dot{\theta}_1^2 \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\big[\dot{\theta}_0^2 + 2l_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1 + l_1^2\dot{\theta}_1^2\big] + \frac{1}{2}I_1\dot{\theta}_1^2 \\
E_{kin}^{(1)} &= \frac{1}{2}m_1\dot{\theta}_0^2 + \frac{1}{2}\big(m_1l_1^2 + I_1\big)\dot{\theta}_1^2 + m_1l_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1
\end{aligned}
$$

Furthermore

$$
\begin{aligned}
E_{pot}^{(1)} &= m_1gy_1 \\
E_{pot}^{(1)} &= m_1gl_1\cos\theta_1
\end{aligned}
$$

Finally, calculate the energies for the top pendulum link. As before, for the same reasons, the kinetic energy has two components, translational and rotational.

$$\begin{aligned}
E_{kin}^{(2)} &= \frac{1}{2}m_2\big(\dot{x}_2^2+\dot{y}_2^2\big)+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\left\{\Big(\frac{d}{dt}\big[\theta_0+L_1\sin\theta_1+l_2\sin\theta_2\big]\Big)^2+\Big(\frac{d}{dt}\big[L_1\cos\theta_1+l_2\cos\theta_2\big]\Big)^2\right\}\\
&\quad+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\Big[\big(\dot{\theta}_0+L_1\dot{\theta}_1\cos\theta_1+l_2\dot{\theta}_2\cos\theta_2\big)^2+\big(-L_1\dot{\theta}_1\sin\theta_1-l_2\dot{\theta}_2\sin\theta_2\big)^2\Big]\\
&\quad+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\Big[\dot{\theta}_0^2+2L_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1+2l_2\dot{\theta}_0\dot{\theta}_2\cos\theta_2+L_1^2\dot{\theta}_1^2\cos^2\theta_1\\
&\quad+2L_1l_2\dot{\theta}_1\dot{\theta}_2\cos\theta_1\cos\theta_2+l_2^2\dot{\theta}_2^2\cos^2\theta_2+L_1^2\dot{\theta}_1^2\sin^2\theta_1\\
&\quad+2L_1l_2\dot{\theta}_1\dot{\theta}_2\sin\theta_1\sin\theta_2+l_2^2\dot{\theta}_2^2\cos^2\theta_2\Big]+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\big[\dot{\theta}_0^2+2L_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1+2l_2\dot{\theta}_0\dot{\theta}_2\cos\theta_2+L_1^2\dot{\theta}_1^2\big(\cos^2\theta_1+\sin^2\theta_1\big)\\
&\quad+2L_1l_2\dot{\theta}_1\dot{\theta}_2\big(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2\big)+l_2^2\dot{\theta}_2^2\big(\cos^2\theta_2+\sin^2\theta_2\big)\big]\\
&\quad+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\big[\dot{\theta}_0^2+2L_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1+2l_2\dot{\theta}_0\dot{\theta}_2\cos\theta_2+L_1^2\dot{\theta}_1^2\\
&\quad+2L_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)+l_2^2\dot{\theta}_2^2\big]+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\big[\dot{\theta}_0^2+2L_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1+2l_2\dot{\theta}_0\dot{\theta}_2\cos\theta_2+L_1^2\dot{\theta}_1^2\\
&\quad+2L_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)+l_2^2\dot{\theta}_2^2\big]+\frac{1}{2}I_2\dot{\theta}_2^2\\
&= \frac{1}{2}m_2\dot{\theta}_0^2+\frac{1}{2}m_2L_1^2\dot{\theta}_1^2+\frac{1}{2}\left(m_2l_2^2+I_2\right)\dot{\theta}_2^2+m_2L_1\dot{\theta}_0\dot{\theta}_1\cos\theta_1\\
&\quad+m_2l_2\dot{\theta}_0\dot{\theta}_2\cos\theta_2+m_2L_1l_2\dot{\theta}_1\dot{\theta}_2\cos(\theta_1-\theta_2)
\end{aligned}$$

Also

$$\begin{aligned}
E_{pot}^{(2)} &= m_1gy_2\\
&= m_1g(L_1\cos\theta_1+l_2\cos\theta_2)
\end{aligned}$$

Adding all of these kinetic and potential energies together gives the overall energies for the system.

$$
\begin{aligned}
E_{kin} &= E_{kin}^{(0)} + E_{kin}^{(1)} + E_{kin}^{(2)} \\
&= \frac{1}{2}m_0\dot{\theta}_0^2 + \frac{1}{2}m_1\dot{\theta}_0^2 + \frac{1}{2}(m_1 l_1^2 + I_1)\dot{\theta}_1^2 + m_1 l_1 \dot{\theta}_0\dot{\theta}_1 \cos\theta_1 \\
&\quad + \frac{1}{2}m_2\dot{\theta}_0^2 + \frac{1}{2}m_2 L_1^2\dot{\theta}_1^2 + \frac{1}{2}(m_2 l_2^2 + I_2)\dot{\theta}_2^2 + m_2 L_1 \dot{\theta}_0\dot{\theta}_1 \cos\theta_1 \\
&\quad + m_2 l_2 \dot{\theta}_0\dot{\theta}_2 \cos\theta_2 + m_2 L_1 l_2 \dot{\theta}_1\dot{\theta}_2 \cos(\theta_1 - \theta_2) \\
&= \frac{1}{2}(m_0 + m_1 + m_2)\dot{\theta}_0^2 + \frac{1}{2}(m_1 l_1^2 + m_2 L_1^2 + I_1)\dot{\theta}_1^2 + \frac{1}{2}(m_2 l_2^2 + I_2)\dot{\theta}_1^2 \\
&\quad + (m_1 l_1 + m_2 L_1)\dot{\theta}_0\dot{\theta}_1 \cos\theta_1 + m_2 l_2 \dot{\theta}_0\dot{\theta}_2 \cos\theta_2 + m_2 L_1 l_2 \dot{\theta}_1\dot{\theta}_2 \cos(\theta_1 - \theta_2)
\end{aligned}
$$

Also

$$
\begin{aligned}
E_{pot} &= E_{pot}^{(0)} + E_{pot}^{(1)} + E_{pot}^{(2)} \\
&= 0 + m_1 g l_1 \cos\theta_1 + m_1 g(L_1 \cos\theta_1 + l_2 \cos\theta_2) \\
&= g(m_1 l_1 + m_2 L_1)\cos\theta_1 + m_2 g l_2 \cos\theta_2
\end{aligned}
$$

The Lagrangian (1.2) is

$$
\begin{aligned}
L &= E_{kin} - E_{pot} \\
&= \frac{1}{2}(m_0 + m_1 + m_2)\dot{\theta}_0^2 + \frac{1}{2}(m_1 l_1^2 + m_2 L_1^2 + I_1)\dot{\theta}_1^2 + \frac{1}{2}(m_2 l_2^2 + I_2)\dot{\theta}_2^2 \\
&\quad + (m_1 l_1 + m_2 L_1)\cos\theta_1 \dot{\theta}_0\dot{\theta}_1 + m_2 l_2 \cos\theta_2 \dot{\theta}_0\dot{\theta}_2 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\dot{\theta}_1\dot{\theta}_2 \\
&\quad - g(m_1 l_1 + m_2 L_1)\cos\theta_1 - m_2 g l_2 \cos\theta_2
\end{aligned}
$$

Now that the Lagrangian is known, explicitly calculate the partial, and full derivatives for the system of equations (1.1).

$$\begin{aligned}
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_0}\right) - \frac{\partial L}{\partial \theta_0} &= u(t) \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} &= 0 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_2}\right) - \frac{\partial L}{\partial \theta_2} &= 0
\end{aligned}$$

Now, calculating the derivatives

$$\begin{aligned}
\frac{\partial L}{\partial \dot{\theta}_0} &= (m_0 + m_1 + m_2)\dot{\theta}_0 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \dot{\theta}_1 + m_2 l_2 \cos\theta_2 \dot{\theta}_2 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_0}\right) &= (m_0 + m_1 + m_2)\ddot{\theta}_0 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot{\theta}_1 + m_2 l_2 \cos\theta_2 \ddot{\theta}_2 \\
&\quad - (m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_1^2 - m_2 l_2 \sin\theta_2 \dot{\theta}_2^2 \\
\frac{\partial L}{\partial \theta_0} &= 0 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_0}\right) - \frac{\partial L}{\partial \theta_0} &= (m_0 + m_1 + m_2)\ddot{\theta}_0 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot{\theta}_1 + m_2 l_2 \cos\theta_2 \ddot{\theta}_2 \\
&\quad - (m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_1^2 - m_2 l_2 \sin\theta_2 \dot{\theta}_2^2 \\
&= u(t)
\end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial L}{\partial \dot{\theta}_1} &= (m_1 l_1^2 + m_2 L_1^2 + I_1)\dot{\theta}_1 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \dot{\theta}_0 \\
&\quad + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\dot{\theta}_2 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) &= (m_1 l_1^2 + m_2 L_1^2 + I_1)\ddot{\theta}_1 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot{\theta}_0 \\
&\quad + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot{\theta}_2 - (m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_0 \dot{\theta}_1 \\
&\quad - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)(\dot{\theta}_1 - \dot{\theta}_2)\dot{\theta}_2 \\
\frac{\partial L}{\partial \theta_1} &= -(m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_0 \dot{\theta}_1 - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_1 \dot{\theta}_2 \\
&\quad + g(m_1 l_1 + m_2 L_1)\sin\theta_1 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} &= (m_1 l_1^2 + m_2 L_1^2 + I_1)\ddot{\theta}_1 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot{\theta}_0 \\
&\quad + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot{\theta}_2 - (m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_0 \dot{\theta}_1 \\
&\quad - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)(\dot{\theta}_1 - \dot{\theta}_2)\dot{\theta}_2 \\
&\quad + (m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_0 \dot{\theta}_1 + m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_1 \dot{\theta}_2 \\
&\quad - g(m_1 l_1 + m_2 L_1)\sin\theta_1 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_1}\right) - \frac{\partial L}{\partial \theta_1} &= (m_1 l_1^2 + m_2 L_1^2 + I_1)\ddot{\theta}_1 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot{\theta}_0 \\
&\quad + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot{\theta}_2 + m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_2^2 \\
&\quad - g(m_1 l_1 + m_2 L_1)\sin\theta_1 \\
&= 0
\end{aligned}
$$

finally,

$$
\begin{aligned}
\frac{\partial L}{\partial \dot\theta_2} &= m_2 l_2 \cos\theta_2 \dot\theta_0 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\dot\theta_1 + (m_2 l_2^2 + I_2)\dot\theta_2 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot\theta_2}\right) &= m_2 l_2 \cos\theta_2 \ddot\theta_0 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot\theta_1 + (m_2 l_2^2 + I_2)\ddot\theta_2 \\
&\quad - m_2 l_2 \sin\theta_2 \dot\theta_0 \dot\theta_2 - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)(\dot\theta_1 - \dot\theta_2)\dot\theta_1 \\
\frac{\partial L}{\partial \theta_2} &= -m_2 l_2 \sin\theta_2 \dot\theta_0 \dot\theta_2 + m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot\theta_1 \dot\theta_2 \\
&\quad + m_2 g l_2 \sin\theta_2 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot\theta_2}\right) - \frac{\partial L}{\partial \theta_2} &= m_2 l_2 \cos\theta_2 \ddot\theta_0 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot\theta_1 + (m_2 l_2^2 + I_2)\ddot\theta_2 \\
&\quad - m_2 l_2 \sin\theta_2 \dot\theta_0 \dot\theta_2 - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)(\dot\theta_1 - \dot\theta_2)\dot\theta_1 \\
&\quad + m_2 l_2 \sin\theta_2 \dot\theta_0 \dot\theta_2 - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot\theta_1 \dot\theta_2 \\
&\quad - m_2 g l_2 \sin\theta_2 \\
\frac{d}{dt}\left(\frac{\partial L}{\partial \dot\theta_2}\right) - \frac{\partial L}{\partial \theta_2} &= m_2 l_2 \cos\theta_2 \ddot\theta_0 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot\theta_1 + (m_2 l_2^2 + I_2)\ddot\theta_2 \\
&\quad - m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot\theta_1^2 - m_2 g l_2 \sin\theta_2 \\
&= 0
\end{aligned}
$$

This gives the calculated system, in full

$$
\begin{aligned}
(m_0 + m_1 + m_2)\ddot\theta_0 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot\theta_1 + m_2 l_2 \cos\theta_2 \ddot\theta_2 \\
-(m_1 l_1 + m_2 L_1)\sin\theta_1 \dot\theta_1^2 - m_2 l_2 \sin\theta_2 \dot\theta_2^2 = u(t)
\end{aligned} \tag{1.3}
$$

$$
\begin{aligned}
(m_1 l_1^2 + m_2 L_1^2 + I_1)\ddot\theta_1 + (m_1 l_1 + m_2 L_1)\cos\theta_1 \ddot\theta_0 \\
+ m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot\theta_2 + m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot\theta_2^2 \\
-g(m_1 l_1 + m_2 L_1)\sin\theta_1 = 0
\end{aligned} \tag{1.4}
$$

$$
\begin{aligned}
m_2 l_2 \cos\theta_2 \ddot\theta_0 + m_2 L_1 l_2 \cos(\theta_1 - \theta_2)\ddot\theta_1 + (m_2 l_2^2 + I_2)\ddot\theta_2 \\
-m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot\theta_1^2 - m_2 g l_2 \sin\theta_2 = 0
\end{aligned} \tag{1.5}
$$

2nd-Order System

The system (1.3-1.5) is a nonlinear second-order system of the form

$$\mathbf{D}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\dot{\boldsymbol{\theta}} + \mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}u \tag{1.6}$$

where

$$\mathbf{D}(\boldsymbol{\theta}) = \begin{bmatrix} m_0 + m_1 + m_2 & (m_1 l_1 + m_2 L_1)\cos\theta_1 & m_2 l_2 \cos\theta_2 \\ (m_1 l_1 + m_2 L_1)\cos\theta_1 & m_1 l_1^2 + m_2 L_1^2 + I_1 & m_2 L_1 l_2 \cos(\theta_1 - \theta_2) \\ m_2 l_2 \cos\theta_2 & m_2 L_1 l_2 \cos(\theta_1 - \theta_2) & m_2 l_2^2 + I_2 \end{bmatrix}$$

$$\mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = \begin{bmatrix} 0 & -(m_1 l_1 + m_2 L_1)\sin\theta_1 \dot{\theta}_1 & -m_2 l_2 \sin\theta_2 \dot{\theta}_2 \\ 0 & 0 & m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_2 \\ 0 & -m_2 L_1 l_2 \sin(\theta_1 - \theta_2)\dot{\theta}_1 & 0 \end{bmatrix}$$

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ -(m_1 l_1 + m_2 L_1) g \sin\theta_1 \\ -m_2 g l_2 \sin\theta_2 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

To simplify the system, we adopt the choices from [9]. Here

$$\implies l_1 = \frac{1}{2}L_1 \qquad l_2 = \frac{1}{2}L_2$$
$$I_1 = \frac{1}{12}m_1L_1^2 \qquad I_2 = \frac{1}{12}m_2L_2^2$$

this updates the system with $\mathbf{D}(\boldsymbol{\theta})$

$$= \begin{bmatrix} m_0 + m_1 + m_2 & (\frac{1}{2}m_1 + m_2)L_1\cos\theta_1 & \frac{1}{2}m_2L_2\cos\theta_2 \\ (\frac{1}{2}m_1 + m_2)L_1\cos\theta_1 & (\frac{1}{3}m_1 + m_2)L_1^2 & \frac{1}{2}m_2L_1L_2\cos(\theta_1 - \theta_2) \\ \frac{1}{2}m_2L_2\cos\theta_2 & \frac{1}{2}m_2L_1L_2\cos(\theta_1 - \theta_2) & \frac{1}{3}m_2L_2^2 \end{bmatrix}$$

and

$$\mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = \begin{bmatrix} 0 & -(\frac{1}{2}m_1 + m_2)L_1\sin\theta_1\dot{\theta}_1 & -\frac{1}{2}m_2L_2\sin\theta_2\dot{\theta}_2 \\ 0 & 0 & \frac{1}{2}m_2L_1L_2\sin(\theta_1 - \theta_2)\dot{\theta}_2 \\ 0 & -\frac{1}{2}m_2L_1L_2\sin(\theta_1 - \theta_2)\dot{\theta}_1 & 0 \end{bmatrix}$$

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ -\frac{1}{2}(m_1 + m_2)L_1g\sin\theta_1 \\ -\frac{1}{2}m_2gL_2\sin\theta_2 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

Note. $\mathbf{D}(\boldsymbol{\theta})$ is symmetric and nonsingular, $\implies$ $\mathbf{D}^{-1}(\boldsymbol{\theta})$ exists and is also symmetric [9].

1st-Order System

In order to use control theory on this system, it makes most sense to convert the system to first-order by manipulating the system and employing a few 'tricks'. Recall the system

$$\mathbf{D}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\dot{\boldsymbol{\theta}} + \mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}u$$

disregarding the arguments for now in the notation

$$\mathbf{D}\ddot{\boldsymbol{\theta}} + \mathbf{C}\dot{\boldsymbol{\theta}} + \mathbf{G} = \mathbf{H}u$$

$$\mathbf{D}\ddot{\boldsymbol{\theta}} = -\mathbf{C}\dot{\boldsymbol{\theta}} - \mathbf{G} + \mathbf{H}u$$

$$\ddot{\boldsymbol{\theta}} = -\mathbf{D}^{-1}\mathbf{C}\dot{\boldsymbol{\theta}} - \mathbf{D}^{-1}\mathbf{G} + \mathbf{D}^{-1}\mathbf{H}u$$

which is

$$\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix} - \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

Let $\mathbf{x} = \begin{bmatrix} \boldsymbol{\theta} \\ \dot{\boldsymbol{\theta}} \end{bmatrix}, \quad \dot{\mathbf{x}} = \begin{bmatrix} \dot{\boldsymbol{\theta}} \\ \ddot{\boldsymbol{\theta}} \end{bmatrix}$

$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix} u$$

gives the first-order system

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x})\mathbf{x} + \mathbf{B}(\mathbf{x})u + \mathbf{L}(\mathbf{x}) \tag{1.7}$$

where

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} 0 & I \\ 0 & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix}, \quad \mathbf{B}(\mathbf{x}) = \begin{bmatrix} 0 \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix}, \quad \mathbf{L}(\mathbf{x}) = \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix}.$$

Here $\mathbf{C}, \mathbf{D}, \mathbf{G}$ depend on $\mathbf{x}$. This form is close to that of a typical control problem which will be explained in section 2.1.

1.2.3 Code in this thesis

Throughout this thesis, MATLAB was used to model and solve this system. Codes are not provided in this thesis but are available upon request. The built-in MATLAB function ODE45 [12] is a numerical integration function that uses Runge-Kutta methods [13] to solve a first-order ordinary differential system [14]. In this thesis, ODE45 will be used multiple times to solve and simulate the controlled nonlinear system, as well as solving the differential Riccati equation. These solutions are then used to model the system with animations and results.

Chapter 2

Control

Dynamical systems represent many phenomena in our universe, the double pendulum inverted on a cart (DPIC) being one of them, the analysis of these systems can definitely enhance our understanding. However it is influencing these systems that is the evolved objective [1]. The field of *control theory* is concerned with influencing and controlling these dynamical systems by a preconceived design, otherwise known as a control. This control can often yield a desired result [2].

The earliest formal use of control theory, was in 1868 when James Clerk Maxwell analyzed the dynamical systems of the centrifugal governor [15]. These results were abstracted by Maxwell's classmate Edward John Routh and applied to general linear systems [16]. Over the following century many important scientific breakthroughs were enabled by control theory. The Wright brothers made their first successful manned flight in 1903, controlling the amount of lift. Control theory was central to many military projects during the second world war. The Space Race also relied heavily upon control theory. It was not until 1960 that Rudolf Kalman formally defined *controllability*, *observability* and formulated the *Kalman Filter* [17]. In this thesis, all of Kalman's results will be used on the DPIC.

2.1 Control Theory

A control problem, in continuous time, is a given system defined on a fixed time interval $0 \leq t \leq t_1$, an initial condition and a set of allowable controls $\mathbf{u}(t) \in \mathbb{R}^m$ where

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{2.1}$$

However, to apply significant results in control theory, the nonlinear system (2.1) must be simplified into a more workable linear format. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$ and consider the following initial value problem (IVP).

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{0}. \tag{2.2}$$

Here $\mathbf{x}(t) \in \mathbb{R}^n$ is the state vector for the system and $\mathbf{u}(t) \in \mathbb{R}^m$ the control function.

Note. It turns out that the assumption $\mathbf{x}(0) = \mathbf{0}$ is not restrictive so that the state $\mathbf{x}(t)$ and the length of the time interval t_1 is unimportant.

Informally, the aim of control theory is choosing the control $\mathbf{u}(t)$ to influence, as desired, the evolution of the state vector function $\mathbf{x}(t)$. Given any time $t_1 > 0$ and any state $\mathbf{x}^{(1)} \in \mathbb{R}^n$, under what assumptions on the pair of matrices $\mathbf{A}$ and $\mathbf{B}$ does there exist a control function $\mathbf{u}(t)$ such that the state $\mathbf{x}(t)$ moves to any given state $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$? This question leads to the notion of *Controllability*.

2.1.1 Controllability

Definition 2.1.1 (Controllability)**.** Fix any $t_1 > 0$. The system (2.2) is called *controllable* in the interval $0 \leq t \leq t_1$ if for any $x^{(1)} \in \mathbb{R}^n$ there exists a control function

$$u : [0, t_1] \to \mathbb{R}^m, \quad u \in C$$

so that the solution $\mathbf{x}(t)$ of the system (2.2) satisfies $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$.

If $\mathbf{B} = \mathbf{0}$ then the system (2.2) is obviously not controllable. Therefore, from now on, assume $\mathbf{B} \neq \mathbf{0}$. Next look at the conditions necessary in order for system (2.2) to be controllable. What follows is a theorem showing the simple set of conditions on the $n \times n$ matrix $\mathbf{A}$ and the $n \times m$ matrix $\mathbf{B}$ that implies controllability of the system.

Theorem 2.1.1 (Controllability Condition)**.** *Define the matrix*

$$\mathbf{M}_n = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}] \in \mathbb{R}^{n \times (mn)}.$$

The system (2.2) is controllable in $0 \leq t \leq t_1$ *if and only if rank*$(\mathbf{M}_n) = n$.

Note. Every part $\mathbf{A}^j\mathbf{B}$ has the same dimensions as $\mathbf{B}$. i.e. $\mathbf{A}^j\mathbf{B}$ has n rows and m columns.

In order to prove Theorem 2.1.1 it helps to state and prove the following lemma about the rank of the controllability $\mathbf{M}_n$ matrix and the rank of matrices with more columns.

Lemma 2.1.2. *For* $k = 1, 2, \ldots$ *set*

$$\mathbf{M}_k = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{k-1}\mathbf{B}] \in \mathbb{R}^{n \times (nk)}$$

and let

$$r_k = rank(\mathbf{M}_k).$$

If $r_k = r_{k+1}$ *then* $r_{k+1} = r_{k+2}$.

Proof. First, recall that

$$r_k = rank(\mathbf{M}_k) = dim(range(\mathbf{M}_k)) \quad \forall k = 1, 2, \ldots$$

Note that

$$\mathbf{M}_{k+1} = (\mathbf{M}_k, \mathbf{A}^k\mathbf{B}),$$

along with the assumption that $r_k = r_{k+1}$, this implies that every column of $\mathbf{A}^k\mathbf{B}$ lies in the range of $\mathbf{M}_k$. If $(\mathbf{A}^k\mathbf{B})_j$ denotes the j-th column of $\mathbf{A}^k\mathbf{B}$ then there exists a vector $\mathbf{c}_j \in \mathbb{R}^{mk}$ with

$$(\mathbf{A}^k\mathbf{B})_j = \mathbf{M}_k\mathbf{c}_j \quad \text{for} \quad j = 1, \ldots, m.$$

Set

$$\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m) \in \mathbb{R}^{(mk)\times m}$$

and obtain that

$$\begin{aligned}
\mathbf{A}^k\mathbf{B} &= \mathbf{M}_k\mathbf{C}\\
\mathbf{A}^{k+1}\mathbf{B} &= \mathbf{A}\mathbf{M}_k\mathbf{C}\\
\mathbf{A}^{k+1}\mathbf{B} &= \mathbf{A}(\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{k-1}\mathbf{B})\mathbf{C}\\
\mathbf{A}^{k+1}\mathbf{B} &= (\mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \mathbf{A}^3\mathbf{B}, \ldots, \mathbf{A}^k\mathbf{B})C\\
\mathbf{A}^{k+1}\mathbf{B} &= (\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{k-1}\mathbf{B})\tilde{\mathbf{C}} \quad \text{where } \tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{C} \end{pmatrix} \text{ with } \mathbf{0} \in \mathbb{R}^{m\times m}\\
\mathbf{A}^{k+1}\mathbf{B} &= \mathbf{M}_{k+1}\tilde{\mathbf{C}}
\end{aligned}$$

Note $\tilde{\mathbf{C}} \in \mathbb{R}^{m(k+1)\times m}$, thus the columns of $\mathbf{A}^{k+1}\mathbf{B}$ lie in

$$range(\mathbf{M}_{k+1}) = range(\mathbf{M}_k).$$

therefore

$$r_k = r_{k+1} = r_{k+2}.$$

□

Proof. ***Theorem 1.1***. Assume that $rank(\mathbf{M}_n) = n$. To prove the theorem, a control $\mathbf{u}(t)$ is constructed for which the solution $\mathbf{x}(t)$ of the IVP (2.2) satisfies $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$. Since every matrix $\mathbf{M}_k$ has n rows $rank(\mathbf{M}_k) \leq n$ for all k. Therefore, the assumption $rank(\mathbf{M}_n) = n$ implies that $rank(\mathbf{M}_k) = n$ for all $k \geq n$. Define the matrix

$$\mathbf{K} = \int_0^{t_1} e^{-\mathbf{A}t}\mathbf{B}\mathbf{B}^T e^{-\mathbf{A}^T t} dt \in \mathbb{R}^{n\times n}$$

It is clear that $\mathbf{K} = \mathbf{K}^T$. It will be shown that $\mathbf{K}$ is positive definite. Set

$$\mathbf{C}(t) = \mathbf{B}^T e^{-\mathbf{A}^T t} \in \mathbb{R}^{m\times n}.$$

If $\mathbf{a} \in \mathbb{R}^n$ is arbitrary

$$\mathbf{a}^T\mathbf{K}\mathbf{a} = \int_0^{t_1} \mathbf{a}^T\mathbf{C}(t)^T\mathbf{C}(t)\mathbf{a}dt$$
$$\mathbf{a}^T\mathbf{K}\mathbf{a} = \int_0^{t_1} |\mathbf{C}(t)\mathbf{a}|^2 dt$$

This shows that $\mathbf{a}^T\mathbf{K}\mathbf{a} \geq 0$ and if $\mathbf{a}^T\mathbf{K}\mathbf{a} = 0$ then

$$\mathbf{a}^T\mathbf{C}(t)^T = \mathbf{a}^T e^{-\mathbf{A}t}\mathbf{B} = \mathbf{0} \quad \text{for} \quad 0 \leq t \leq t_1.$$

Now let $\mathbf{a} \in \mathbb{R}^n$ be arbitrary and define the vector function

$$\boldsymbol{\phi}(t) = \mathbf{a}^T e^{-\mathbf{A}t}\mathbf{B} \quad \text{for} \quad 0 \leq t \leq t_1.$$

Note that $\phi(t)$ is a row vector of dimension m. If one assumes that $\mathbf{a}^T\mathbf{K}\mathbf{a} = 0$ then

$$\boldsymbol{\phi}(t) = \mathbf{a}^T e^{-\mathbf{A}t}\mathbf{B} = \mathbf{0} \quad \text{for} \quad 0 \leq t \leq t_1.$$

Therefore,

$$\begin{aligned}
\boldsymbol{\phi}(t) &= \mathbf{a}^T e^{-\mathbf{A}t}\mathbf{B} \\
\boldsymbol{\phi}(0) &= \mathbf{a}^T\mathbf{B} = \mathbf{0} \\
\dot{\boldsymbol{\phi}}(t) &= -\mathbf{a}^T e^{-\mathbf{A}t}\mathbf{A}\mathbf{B} \\
\dot{\boldsymbol{\phi}}(0) &= -\mathbf{a}^T\mathbf{A}\mathbf{B} = \mathbf{0} \\
\ddot{\boldsymbol{\phi}}(t) &= \mathbf{a}^T e^{-\mathbf{A}t}\mathbf{A}^2\mathbf{B} \\
\ddot{\boldsymbol{\phi}}(0) &= \mathbf{a}^T\mathbf{A}^2\mathbf{B} = \mathbf{0} \\
&\vdots
\end{aligned}$$

One obtains that $\mathbf{a}^T\mathbf{K}\mathbf{a} = 0$, this implies

$$\begin{aligned}
\mathbf{a}^T\mathbf{B} &= \mathbf{0} \\
\mathbf{a}^T\mathbf{A}\mathbf{B} &= \mathbf{0} \\
\mathbf{a}^T\mathbf{A}^2\mathbf{B} &= \mathbf{0} \\
&\vdots
\end{aligned}$$

Therefore

$$\mathbf{a}^T\mathbf{M}_n = \mathbf{a}^T(\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}]) = \mathbf{0}$$

Since $\mathbf{M}_n$ has n linearly independent columns it follows that $\mathbf{a} = \mathbf{0}$. Thus it is shown that $\mathbf{a}^T\mathbf{K}\mathbf{a} > 0$ if $\mathbf{a} \neq \mathbf{0}$ i.e.

$$\mathbf{K} = \mathbf{K}^T > 0.$$

Set

$$\mathbf{u}(t) = \mathbf{B}^T e^{-\mathbf{A}^T t}\mathbf{K}^{-1}e^{-\mathbf{A}t_1}\mathbf{x}^{(1)} \in \mathbb{R}^m \tag{2.3}$$

Then the solution of the IVP (2.2) satisfies

$$\mathbf{x}(t_1) = \int_0^{t_1} e^{\mathbf{A}(t_1-t)}\mathbf{B}\mathbf{u}(t)dt$$
$$\mathbf{x}(t_1) = e^{\mathbf{A}t_1}\int_0^{t_1} e^{-\mathbf{A}t}\mathbf{B}\mathbf{B}^T e^{-\mathbf{A}^T t}\mathbf{K}^{-1}e^{-\mathbf{A}t_1}\mathbf{x}^{(1)}dt$$
$$\mathbf{x}(t_1) = e^{\mathbf{A}t_1}\mathbf{K}\mathbf{K}^{-1}e^{-\mathbf{A}t_1}\mathbf{x}^{(1)}$$
$$\mathbf{x}(t_1) = \mathbf{x}^{(1)}$$

This proves the control $\mathbf{u}(t)$ given in (2.3) leads to a solution $\mathbf{x}(t)$ of (2.2) with $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$. Assume that $rank(\mathbf{M}_n) < n$, i.e., $r_n < n$. Since $1 \leq r_1 \leq r_2 \leq ... \leq r_n < n$ there exists $k \in [1, 2, ..., n-1]$ with $r_k = r_{k+1} < n$. Using lemma 2.1.2 it can be concluded that

$$range\mathbf{M}_j = range\mathbf{M}_n \neq \mathbb{R}^n \qquad \forall j \geq n.$$

For the solution $\mathbf{x}(t)$ of the IVP (2.2) one has

$$\mathbf{x}(t_1) = \int_0^{t_1} e^{\mathbf{A}(t_1-t)}\mathbf{B}\mathbf{u}(t)dt$$
$$\mathbf{x}(t_1) = \int_0^{t_1} \sum_{j=0}^{\infty} \frac{1}{j!}\mathbf{A}^j\mathbf{B}(t_1-t)^j\mathbf{u}(t)dt$$
$$\mathbf{x}(t_1) = \sum_{j=0}^{\infty} \mathbf{A}^j\mathbf{B}\alpha_j$$

with

$$\boldsymbol{\alpha}^j = \frac{1}{j!}\int_0^{t_1} (t_1-t)^j\mathbf{u}(t)dt \in \mathbb{R}^m.$$

Obtain that

$$\mathbf{x}(t_1) = \lim_{J\to\infty}\sum_{j=0}^{J} \mathbf{A}^j\mathbf{B}\alpha_j \in \mathbf{M}_n$$

for every control function $\mathbf{u}(t)$. This proves that if $range(\mathbf{M}_n)$ is a strict subspace of $\mathbb{R}^n$, then the system (2.2) is not controllable in $0 \leq t \leq t_1$. This proves the theorem. □

General Initial Data

Consider a similar initial value problem with a more general initial condition.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}^{(0)} \tag{2.4}$$

where $\mathbf{x}^{(0)} \in \mathbb{R}^n$ is given. Also, let $\mathbf{x}^{(1)} \in \mathbb{R}^n$ be given.

Theorem 2.1.3. *Assume that the system (2.2) is controllable on the interval* $0 \leq t \leq t_1$*, i.e. rank*$(\mathbf{M}_n) = n$*. Then there exists a control function* $\mathbf{u}(t)$ *such that the solution of (2.4) satisfies* $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$.

Proof. By Theorem 2.1.1 there exists a control function $\mathbf{y}(t)$ so that the solution of the following initial value problem

$$\dot{\mathbf{y}}(t) = \mathbf{A}\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{y}(0) = \mathbf{0},$$

satisfies

$$\mathbf{y}(t_1) = \mathbf{x}^{(1)} - e^{\mathbf{A}t_1}\mathbf{x}^{(0)}.$$

Setting

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}^{(0)} + \mathbf{y}(t)$$

gives

$$\mathbf{x}(0) = \mathbf{x}^{(0)} \quad \text{and} \quad \mathbf{x}(t_1) = \mathbf{x}^{(1)}.$$

Also

$$\dot{\mathbf{x}}(t) = \mathbf{A}e^{\mathbf{A}t}\mathbf{x}^{(0)} + \dot{\mathbf{y}}(t)$$
$$\dot{\mathbf{x}}(t) = \mathbf{A}e^{\mathbf{A}t}\mathbf{x}^{(0)} + \mathbf{A}\mathbf{y}(t) + \mathbf{B}\mathbf{u}(t)$$
$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t).$$

Therefore, $\mathbf{x}(t)$ satisfies the differential equation $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, the initial condition $\mathbf{x}(0) = \mathbf{x}^{(0)}$ and the end condition $\mathbf{x}(t_1) = \mathbf{x}^{(1)}$. □

2.1.2 Observability

Observability is a concept similar to controllability. The conditions for controllability and observability have related requirements. They are often referred to as mathematical 'duals' [1]. Informally, observability is a measure of how well the state vector of a system can be inferred from particular knowledge of system outputs [2]. Let $\mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{D} \in \mathbb{R}^{p \times m}$, with $\mathbf{A}$ and $\mathbf{B}$ as before. Consider the system

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y} &= \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \end{aligned} \tag{2.5}$$

with outputs $\mathbf{y} \in \mathbb{R}^p$. $\mathbf{C}$ is a matrix that determines which components of the state are measurable. $\mathbf{D}$ dictates how the control affects this output. Typically, these matrices have only diagonal components of either 0 or 1 [2].

Definition 2.1.2 (Observability)**.** The system (2.5) is called *completely observable* if there is a $t_1 > 0$ such that knowledge of $\mathbf{y}(t)$ for the interval $0 \leq t \leq t_1$ is sufficient to determine $\mathbf{x}(0)$.

Theorem 2.1.4 (Observability Condition)**.** *Define the matrix*

$$\mathbf{S}_n = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} \in \mathbb{R}^{(pn)\times(n)}.$$

The system (2.5) is completely observable if and only if $rank(\mathbf{S}_n) = n$.

Proof. The proof of this theorem is very similar to that of Theorem 2.1.1. Which does not need to be restated here. Omitted [1]. □

2.1.3 Eigenvalue Placement

In control theory, it is often important that the controller increases or creates stability for the system. In other words, to what extent can the eigenvalues of the system be affected by the control function $\mathbf{u}(t)$?

Theorem 2.1.5. *Let the system (2.2) be controllable. Then for any nth order polynomial*

$$p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0$$

there exists a real-valued matrix $\mathbf{K}$ such that $\mathbf{A} - \mathbf{BK}$ has $p(\lambda)$ as its characteristic polynomial.

Proof. Omitted [1]. □

If the linear system (2.2) is controllable, we can control the state for this system $\mathbf{x}(t)$ to any point in $\mathbb{R}^n$, at any time. We can also choose any eigenvalues for the matrix $\mathbf{A} - \mathbf{BK}$.

2.2 Control Theory and the DPIC System

For the DPIC problem, the equations of motion are represented in system (1.6), the parameter $u(t)$ is the control. $u(t)$ is the horizontal force exerted on the cart. This force can be implemented in either direction to move the cart either left or right. In other words, the system has only a single positive or negative scalar input $u(t)$. Recall equation (1.7)

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x})\mathbf{x} + \mathbf{B}(\mathbf{x})u + \mathbf{L}(\mathbf{x})$$

where

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & -\mathbf{D}^{-1}\mathbf{C} \end{bmatrix}, \quad \mathbf{B}(\mathbf{x}) = \begin{bmatrix} \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{H} \end{bmatrix}, \quad \mathbf{L}(\mathbf{x}) = \begin{bmatrix} \mathbf{0} \\ -\mathbf{D}^{-1}\mathbf{G} \end{bmatrix}.$$

which is a nonlinear problem of the form (2.1)

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0.$$

This is a very complex and difficult nonlinear differential system, one in which only certain numerical methods could calculate an approximate solution. To apply the basic concepts of linear control from the previous section, one must use the linear version of the system of the form (2.2), where $\mathbf{A}, \mathbf{B}$ are constant,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u.$$

2.2.1 Extracting the Linear System

In order to reduce the system to a linear form, series expansions about $\boldsymbol{\theta} = \mathbf{0}$ are employed [9], it is assumed that the pendulum is close to the vertical. The $\mathbf{L}$ term is absorbed into the $\mathbf{x}$ coefficient. This eliminates theta dependence. Technically, this is a linearization about $\mathbf{x} = \mathbf{0}$. The linear system could also be found by computing the Jacobian and evaluating at a fixed point [2].

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ 0 & -\mathbf{D}^{-1}(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}(\boldsymbol{\theta})\mathbf{G}(\boldsymbol{\theta}) \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}(\boldsymbol{\theta})\mathbf{H} \end{bmatrix} u$$

where $\mathbf{D}(\boldsymbol{\theta})$ is

$$\begin{bmatrix} m_0 + m_1 + m_2 & (\frac{1}{2}m_1 + m_2)L_1 \cos\theta_1 & \frac{1}{2}m_2 L_2 \cos\theta_2 \\ (\frac{1}{2}m_1 + m_2)L_1 \cos\theta_1 & (\frac{1}{3}m_1 + m_2)L_1^2 & \frac{1}{2}m_2 L_1 L_2 \cos(\theta_1 - \theta_2) \\ \frac{1}{2}m_2 L_2 \cos\theta_2 & \frac{1}{2}m_2 L_1 L_2 \cos(\theta_1 - \theta_2) & \frac{1}{3}m_2 L_2^2 \end{bmatrix}$$

and

$$\mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = \begin{bmatrix} 0 & -(\frac{1}{2}m_1 + m_2)L_1 \sin\theta_1 \dot{\theta}_1 & -\frac{1}{2}m_2 L_2 \sin\theta_2 \dot{\theta}_2 \\ 0 & 0 & \frac{1}{2}m_2 L_1 L_2 \sin(\theta_1 - \theta_2)\dot{\theta}_2 \\ 0 & -\frac{1}{2}m_2 L_1 L_2 \sin(\theta_1 - \theta_2)\dot{\theta}_1 & 0 \end{bmatrix}$$

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{bmatrix} 0 \\ -\frac{1}{2}(m_1 + m_2)L_1 g \sin\theta_1 \\ -\frac{1}{2}m_2 g L_2 \sin\theta_2 \end{bmatrix}$$

Using multi-variable Taylor expansions

$$\mathbf{D}^{-1}(\boldsymbol{\theta}) \sim \mathbf{D}^{-1}(\mathbf{0}), \quad \mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) \sim \mathbf{C}(\mathbf{0}, \mathbf{0}) \quad \text{and} \quad \mathbf{G}(\boldsymbol{\theta}) \sim \frac{\partial \mathbf{G}(\mathbf{0})}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}$$

$$\mathbf{C}(\mathbf{0},\mathbf{0}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \implies \mathbf{A}(0) = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \text{ one now has}$$

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}(\mathbf{0})\mathbf{H} \end{bmatrix} u + \begin{bmatrix} 0 \\ -\mathbf{D}^{-1}(\mathbf{0})\frac{\partial \mathbf{G}(\mathbf{0})}{\partial \boldsymbol{\theta}}\boldsymbol{\theta} \end{bmatrix}$$

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}(\mathbf{0})\mathbf{H} \end{bmatrix} u + \begin{bmatrix} 0 & 0 \\ -\mathbf{D}^{-1}(\mathbf{0})\frac{\partial \mathbf{G}(\mathbf{0})}{\partial \boldsymbol{\theta}} & 0 \end{bmatrix} \mathbf{x}$$

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & I \\ -\mathbf{D}^{-1}(\mathbf{0})\frac{\partial \mathbf{G}(\mathbf{0})}{\partial \boldsymbol{\theta}} & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ \mathbf{D}^{-1}(\mathbf{0})\mathbf{H} \end{bmatrix} u$$

This system is now of the form (2.2), with the particular linear system as follows

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{2.6}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & I \\ -\mathbf{D}^{-1}(\mathbf{0})\frac{\partial \mathbf{G}(\mathbf{0})}{\partial \boldsymbol{\theta}} & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 \\ \mathbf{D}^{-1}(\mathbf{0})\mathbf{H} \end{bmatrix}$$

A and **B** are now new constant matrices. Here

$$\mathbf{D}(\mathbf{0}) = \begin{bmatrix} m_0 + m_1 + m_2 & (\frac{1}{2}m_1 + m_2)L_1 & \frac{1}{2}m_2L_2 \\ (\frac{1}{2}m_1 + m_2)L_1 & (\frac{1}{3}m_1 + m_2)L_1^2 & \frac{1}{2}m_2L_1L_2 \\ \frac{1}{2}m_2L_2 & \frac{1}{2}m_2L_1L_2 & \frac{1}{3}m_2L_2^2 \end{bmatrix}$$

and $\mathbf{D}(\mathbf{0})^{-1}$ is

$$\begin{bmatrix} 4pm_1 + 3pm_2 & -\frac{3p}{L_1}(2m_1 + m_2) & \frac{3p}{L_2}m_1 \\ -\frac{3p}{L_1}(2m_1 + m_2) & \frac{3p}{L_1^2}(4m_0 + 4m_1 + m_2) & -\frac{9p}{L_1 L_2}(2m_0 + m_1) \\ \frac{3p}{L_2}m_1 & -\frac{9p}{L_1 L_2}(2m_0 + m_1) & \frac{3p}{m_2 L_2^2}(m_1^2 + 4m_0 m_1 + 4m_1 m_2 + 12m_0 m_2) \end{bmatrix}$$

where

$$p = \det \mathbf{D}(\mathbf{0}) = \frac{1}{4m_0 m_1 + 3m_0 m_2 + m_1^2 + m_1 m_2}$$

Furthermore

$$\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}}(\mathbf{0}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -\frac{L_1 g}{2}(m_1 + 2m_2) & 0 \\ 0 & 0 & -\frac{L_2 g}{2}m_2 \end{bmatrix}$$

therefore $-\mathbf{D}(\mathbf{0})^{-1}\frac{\partial \mathbf{G}}{\partial \boldsymbol{\theta}}(\mathbf{0})$ is

$$\begin{bmatrix} 0 & \frac{3}{2}(2m_1 + m_2)(m_1 + 2m_2) & -\frac{3}{2}m_1 m_2 \\ 0 & -\frac{3}{2L_1}(4m_0 + 4m_1 + m_2)(m_1 + 2m_2) & \frac{9}{2L_1}(2m_0 + m_1)m_2 \\ 0 & \frac{9}{2L_2}(2m_0 + m_1)(m_1 + 2m_2) & -\frac{3}{2L_2}(m_1^2 + 4m_0 m_1 + 4m_1 m_2 + 12m_0 m_2) \end{bmatrix}$$

also

$$\mathbf{H} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \implies -\mathbf{D}(\mathbf{0})^{-1}\mathbf{H} = \begin{bmatrix} p(4m_1 + 3m_2) \\ -\frac{3p}{L_1}(2m_1 + m_2) \\ \frac{3pm_2}{L_2} \end{bmatrix}$$

Finally, the linear system for the DPIC is as follows.

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{2.7}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & a_{42} & a_{43} & 0 & 0 & 0 \\ 0 & a_{52} & a_{53} & 0 & 0 & 0 \\ 0 & a_{62} & a_{63} & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix}$$

where

$$\begin{aligned} a_{42} &= -\frac{3}{2}p(2m_1^2 + 5m_1m_2 + 2m_2^2)g \\ a_{43} &= \frac{3}{2}pm_1m_2g \\ a_{52} &= \frac{3}{2}\frac{p}{L_1}(4m_0m_1 + 8m_0m_2 + 4m_1^2 + 9m_1m_2 + 2m_2^2)g \\ a_{53} &= -\frac{9}{2}\frac{p}{L_1}(2m_0m_2 + m_1m_2)g \\ a_{62} &= -\frac{9}{2}\frac{p}{L_2}(2m_0m_1 + 4m_0m_2 + m_1^2 + 2m_1m_2) \\ a_{63} &= \frac{3}{2}\frac{p}{L_2}(m_1^2 + 4m_0m_1 + 12m_0m_2 + 4m_1m_2) \end{aligned}$$

and

$$\begin{aligned} b_4 &= p(4m_1 + 3m_2) \\ b_5 &= -\frac{3p}{L_1}(2m_1 + m_2) \\ b_6 &= \frac{3pm_2}{L_2} \end{aligned}$$

here p is the determinant of $\mathbf{D}(\mathbf{0})$.

$$p = \frac{1}{4m_0m_1 + 3m_0m_2 + m_1^2 + m_1m_2}$$

2.2.2 Controllability of the DPIC

For the DPIC and the constant matrices, $\mathbf{A}$ and $\mathbf{B}$ are of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \tilde{\mathbf{A}} & \mathbf{0} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0} \\ \tilde{\mathbf{B}} \end{bmatrix}$$

where $\tilde{\mathbf{A}} \in \mathbb{R}^{3\times 3}$ and $\tilde{\mathbf{B}} \in \mathbb{R}^{3}$. This implies that the controllability matrix for the linear system

$$\mathbf{M} = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \dots, \mathbf{A}^5\mathbf{B}]$$

is actually as follows

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{B}} & \mathbf{0} & \tilde{\mathbf{A}}\tilde{\mathbf{B}} & \mathbf{0} & \tilde{\mathbf{A}}^2\tilde{\mathbf{B}} \\ \tilde{\mathbf{B}} & \mathbf{0} & \tilde{\mathbf{A}}\tilde{\mathbf{B}} & \mathbf{0} & \tilde{\mathbf{A}}^2\tilde{\mathbf{B}} & \mathbf{0} \end{bmatrix}$$

Recall that $\tilde{\mathbf{B}}$ is a vector, let

$$\mathbf{M} = \begin{bmatrix} \mathbf{c}^{(1)} & \mathbf{c}^{(2)} & \mathbf{c}^{(3)} & \mathbf{c}^{(3)} & \mathbf{c}^{(5)} & \mathbf{c}^{(6)} \end{bmatrix}$$

Using Theorem 2.1.1, the linear system is controllable if the matrix $\mathbf{M}$ has full rank, $rank(\mathbf{M}) = 6$. For $\mathbf{M}$, it is clear that the 'odd' columns, columns with nonzero terms in the top of the partition $\mathbf{c}^{(1)}, \mathbf{c}^{(3)}, \mathbf{c}^{(5)}$, are linearly independent to the even columns, $\mathbf{c}^{(2)}, \mathbf{c}^{(4)}, \mathbf{c}^{(6)}$. As the rank of a matrix is the number of linearly independent columns in the matrix, the structure of $\mathbf{M}$ enables simplifying the criteria to $rank(\tilde{\mathbf{M}}) = 3$ where

$$\tilde{\mathbf{M}} = \begin{bmatrix} \tilde{\mathbf{B}} & \tilde{\mathbf{A}}\tilde{\mathbf{B}} & \tilde{\mathbf{A}}^2\tilde{\mathbf{B}} \end{bmatrix}$$

Recall

$$\tilde{\mathbf{A}} = \begin{bmatrix} 0 & a_{42} & a_{43} \\ 0 & a_{52} & a_{53} \\ 0 & a_{62} & a_{63} \end{bmatrix} \quad \tilde{\mathbf{B}} = \begin{bmatrix} b_4 \\ b_5 \\ b_6 \end{bmatrix}$$

label these columns

$$\tilde{\mathbf{M}} = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \mathbf{v}^{(3)} \end{bmatrix}$$

where

$$\mathbf{v}^{(1)} = \tilde{\mathbf{B}} = \begin{bmatrix} b_4 \\ b_5 \\ b_6 \end{bmatrix} \quad \mathbf{v}^{(2)} = \tilde{\mathbf{A}}\tilde{\mathbf{B}} = \begin{bmatrix} a_{42}b_5 + a_{43}b_6 \\ a_{52}b_5 + a_{53}b_6 \\ a_{62}b_5 + a_{63}b_6 \end{bmatrix}$$

$$\mathbf{v}^{(3)} = \tilde{\mathbf{A}}^2\tilde{\mathbf{B}} = \begin{bmatrix} (a_{42}a_{52} + a_{43}a_{62})b_5 + (a_{42}a_{53} + a_{43}a_{63})b_6 \\ (a_{52}^2 + a_{53}a_{62})b_5 + (a_{52}a_{53} + a_{53}a_{63})b_6 \\ (a_{52}a_{62} + a_{62}a_{63})b_5 + (a_{53}a_{62} + a_{63}^2)b_6 \end{bmatrix}$$

Therefore the DPIC system will be controllable for parameters m_0, m_1, m_2, L_1, L_2, if the vectors $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}$ are linearly independent.

2.2.3 Parameters

We use the results in [9] as a benchmark. Henceforth the system will be designed with the following parameters; $m_0 = 1.5kg$, $m_1 = 0.5kg$, $m_2 = 0.75kg$, $L_1 = 0.5m$, $L_2 = 0.75m$ and the constant of acceleration due to gravity $g = 9.8ms^{-2}$. These parameters will now be fixed for the rest of this thesis. This gives

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -7.35 & 0.7875 & 0 & 0 & 0 \\ 0 & 73.5 & -33.0750 & 0 & 0 & 0 \\ 0 & -58.8 & 51.1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.6071 \\ -1.5 \\ 0.2857 \end{bmatrix}$$

Controllability

In the linear system $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$. Using the theory of control on this system gives a particular controllability matrix $\mathbf{M} \in \mathbb{R}^{6\times 6}$ which, using MATLAB, has rank 6. Using the singular value decomposition [18], the distance from this matrix to deficient rank is 0.355, the smallest singular value σ_6. By Theorem 2.1.1, the system is controllable. So for any state $\mathbf{x}^{(1)}$ in $\mathbb{R}^6$ and time t_1 there exists a control function $\mathbf{u}(t)$ such that the solution of the system satisfies $\mathbf{x}(t) = \mathbf{x}^{(1)}$ at time t_1. Now that the system is shown to be controllable, by Theorem 2.1.5 we can experiment with eigenvalue placement.

Observability

For the system with these parameters, we assume that we can directly measure each component of the state vector. $\mathbf{C} = \mathbf{I}$, so $\mathbf{y} = \mathbf{x}$. It follows trivially that the system is fully observable.

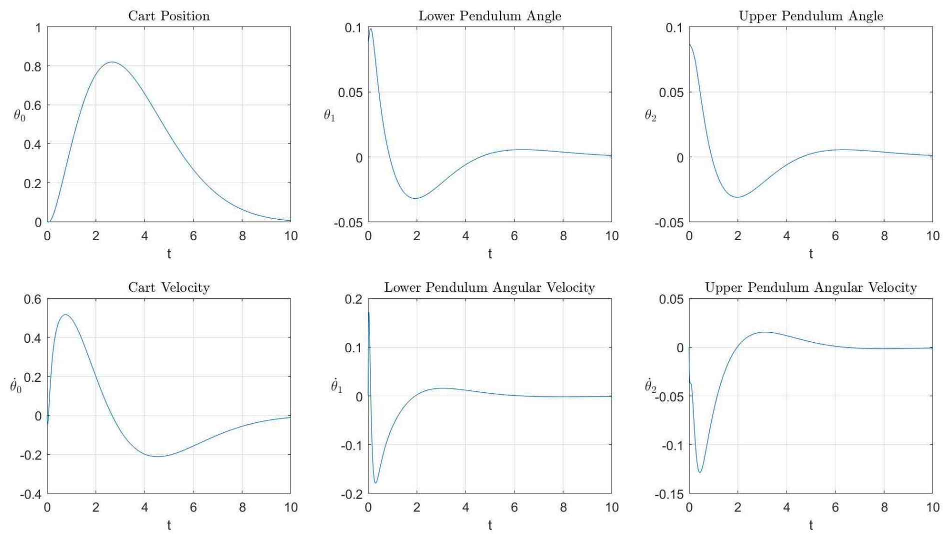

Figure 2.1: State Component Evolution with Weak Eigenvalue Placement Control

2.2.4 Eigenvalue Placement

In MATLAB one can find a controller of the form $u(t) = -\mathbf{Kx}(t)$. Finding the matrix $\mathbf{K} \in \mathbb{R}^{1\times 6}$ by using the built-in MATLAB function PLACE [19]. Input $\mathbf{A}$, $\mathbf{B}$ and the desired eigenvalues of the controlled system matrix $(\mathbf{A} - \mathbf{BK})$, into PLACE and the code outputs $\mathbf{K}$. The linear control, $u(t) = -\mathbf{Kx}(t)$ derived by using eigenvalue placement on the linear system, will be tested by use in the nonlinear system and observing the result. When testing this control, the pendulum begins at an almost vertical position, $\theta_1 = \theta_2 = 5°$ at $\theta_0 = 0$, and is tasked with moving to the straight pendulum-up position at $\theta_1 = \theta_2 = 0$. Note that the angles are referred to in degrees but in the Figures, due to use of MATLAB, they appear in radians. The horizontal variable is time in seconds.

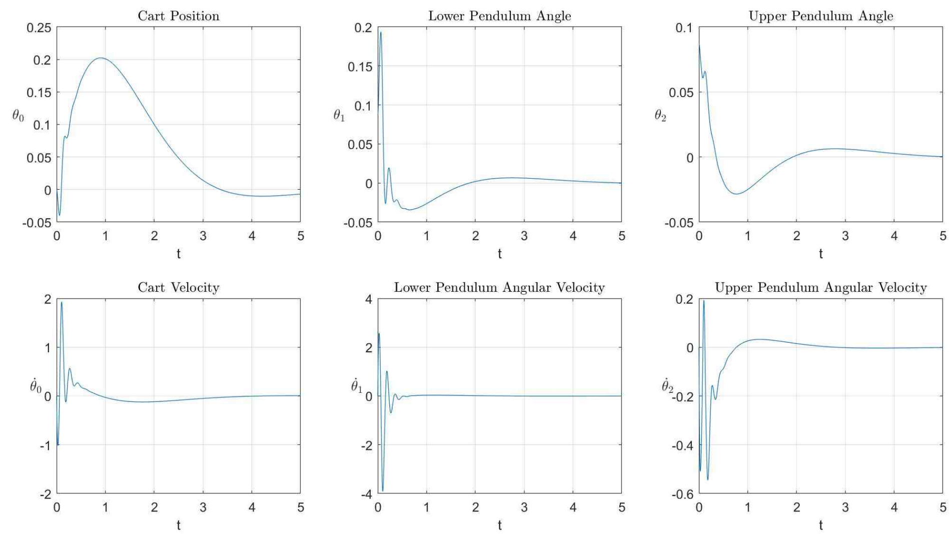

Figure 2.2: State Component Evolution with Strong Eigenvalue Placement Control

The first set of results, shown in Figure 2.1, is created by choosing the eigenvalues of $\mathbf{A} - \mathbf{BK}$ as $(73, -0.11 + i, -0.11 - i, -0.3, -6, -9)$, note here that the position of the cart is a free variable, so the eigenvalue corresponding to the horizontal position does not need to be stable, The other variables are just stable, smaller negative values closer to zero. This control works, it successfully moves the state in the nonlinear system to the pendulum-up position, $\theta_1 = \theta_2 = 0°$. However, it is slow, it takes 10 seconds to control the state to the pendulum-up position. The terminal requirement is that the components of the state vector $\mathbf{x}$ lie between -0.01 and 0.01.

The second set of results, shown in Figure 2.2, is created by choosing the eigenvalues at $(73, -1+i, -1-i, -3, -10, -20)$. The eigenvalues here are more stable than before, they have greater negative real part. This control is also successful, it is twice as fast as the controller in Figure 2.1, taking 5 seconds to get the the same terminal requirement as above. However controlling this system through a force on the cart has a cost, the energy expended by a motor in the cart. The velocities have incredibly high initial oscillatory behaviour, this will have a huge energy cost.

Experimentation with the eigenvalues also leads to results that break down, work slowly, work incredibly fast and diverge [20]. Whilst these controls continue to work successfully for the linear system as the eigenvalues are made more stable, there comes a point where the inherent dynamics of the nonlinear system take over and the linear control fails. Controls like Figure 2.1 are too slow to be useful. Whilst control $u(t)$ in Figure 2.2 is too costly. Both are sub-optimal, optimality will be explored in section 3. One requires a way to reduce cost whilst maintaining performance, this problem now also requires optimization. The optimization is finding a $\mathbf{K}$ corresponding to an ideal set of eigenvalues which has the best trade off between performance, the speed of the control process, and cost, the energy requirement of the control.

Chapter 3

Optimal Control

An optimal control problem generally has a straightforward structure. A given dynamical system for which input control functions can be specified. Along with an objective function which quantifies the cost of the control function and its effect on the system. The goal of optimal control theory is to optimize this objective function and, in so doing, to calculate the optimal control [1].

Resulting from the calculus of variations, Optimal control theory has a 360 year history [21]. However, interest in Optimal Control theory snowballed once computers were made readily available. In 1961, there were particular successes in aerospace applications where optimal trajectory prediction was used [21].

In 1958, Russian mathematician Lev Pontryagin created the *Maximum Principle*, a problem which is equivalent to finding an optimal control for a system given constraints and input controls [22]. This theory will be used and built upon for systems with quadratic cost, namely the DPIC.

3.1 Optimal Control Theory

3.1.1 The Basic Optimal Control Problem

The *basic optimal control problem* is a system of differential equations, defined on $0 \leq t \leq t_1$

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \tag{3.1}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{3.2}$$

$$\mathbf{u}(t) \in U \tag{3.3}$$

U is the set of allowable controls dependent on the system [1]. For simplicity, assume $U = C([0, t_1], \mathbb{R}^m)$. Alongside this control problem there is an objective function to be maximized

$$J(\mathbf{u}) = \phi(\mathbf{x}(t_1)) + \int_0^{t_1} l(\mathbf{x}(t), \mathbf{u}(t))dt \tag{3.4}$$

Assume that the state vector $\mathbf{x}(t) \in \mathbb{R}^n$, the control vector $\mathbf{u}(t) \in \mathbb{R}^m$ and the function system $\mathbf{f} \in \mathbb{F}^n$ have well-behaved components such that the problem above, (3.1-3.3) has a unique solution $\mathbf{x}(t)$. Furthermore, assume that J, ϕ and l in (3.4) are real-valued functions. $\phi(\mathbf{x}(t_1)$ is a function from $\mathbb{R}^m$ to $\mathbb{R}$ that measures the contribution of the objective to the final state, for example achieving maximum velocity at a given time. The function in the integral $l(\mathbf{x}(t), \mathbf{u}(t))$, from $\mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}$, corresponds to expenditure over time, for example minimizing fuel expenditure throughout. In most cases either ϕ or l will be zero, but not both. Once the control $\mathbf{u}(t)$ on $0 \leq t \leq t_1$ is specified, it determines, in conjunction with the system equation and the initial condition, a unique state trajectory $\mathbf{x}(t)$ for $0 \leq t \leq t_1$. This trajectory and the control function then determine the value of J according to the objective function.The *basic optimal control problem* is essentially, to find the control $\mathbf{u}(t)$ from a set of allowable controls U which leads to the largest possible value of J. To summarize all these conditions for optimality the *Maximum principle* is used.

3.1.2 Derivation of the Maximum Principle

The maximum principle will not be explicitly proved here, however the derivation is shown as an aside in Appendix A. Essentially, this is done by considering the effects of small changes near the maximum point. It also defines a modified objective function to obtain an added helpful adjoint trajectory $\boldsymbol{\lambda}(t)$.

3.1.3 The Maximum Principle

Theorem 3.1.1 (Maximum Principle)**.** *Suppose* $\mathbf{u}(t) \in \mathbb{R}^m$ *is the optimal control and* $\mathbf{x}(t)$ *is the state trajectory for the optimal control problem. Then there is an adjoint trajectory* $\boldsymbol{\lambda}(t)$ *such that* $\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}$ *satisfy*

$$
\begin{aligned}
\dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \\
\mathbf{x}(0) &= \mathbf{x}_0 \\
-\dot{\boldsymbol{\lambda}}(t)^T &= \boldsymbol{\lambda}(t)^T \mathbf{f}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t)) + l_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t)) \\
\boldsymbol{\lambda}(t_1)^T &= \phi_{\mathbf{x}}(\mathbf{x}(t_1))
\end{aligned}
$$

$$\forall\ 0 \leq t \leq t_1$$

$$H_{\mathbf{u}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) = 0$$

Where H is a Hamiltonian function $H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) = \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{u}) + l(\mathbf{x}, \mathbf{u})$. The system above is complete; there are $2n + m$ unknowns $\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}$ and $2n$ differential equations with $2n$ boundary conditions and m static equations. One has everything necessary to determine the $2n + m$ functions. Avoiding singular solutions the optimal $\mathbf{u} \in \mathbb{R}^m$ can be found. Solving this system is synonymous with maximizing the objective function [1].

3.1.4 Systems with Quadratic Cost

These systems are ones in which the dynamic system is linear and the objective is quadratic. In the standard *linear-quadratic* optimal control problem one has the following n-th order system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{3.5}$$

with cost function J to be minimized

$$J = \frac{1}{2}\int_0^{t_1} \left[\mathbf{x}(t)^T\mathbf{Q}(t)\mathbf{x}(t) + \mathbf{u}(t)^T\mathbf{R}(t)\mathbf{u}(t)\right] dt$$

where $\mathbf{u}(t) \in \mathbb{R}^m$, $\mathbf{Q}(t) \in \mathbb{R}^{n\times n}$, $\mathbf{R}(t) \in \mathbb{R}^{m\times m}$. The cost function J is quadratic in both state and control. The quadratic functions are defined by both $\mathbf{Q}$ and $\mathbf{R}$. The matrix $\mathbf{Q}(t)$ is symmetric and positive semi-definite, $\mathbf{x}(t)^T\mathbf{Q}(t)\mathbf{x}(t) \geq 0 \forall \mathbf{x}(t)$, $\mathbf{Q}$ quantifies the penalty for the state if it is not in the correct position [2]. $\mathbf{R}(t)$ is symmetric positive definite which implies $\mathbf{R}(t)$ is non-singular, hence invertible. $\mathbf{R}$ quantifies the cost of the control [2]. Finally, assume that all functions are continuous in time. The adjoint equation is

$$-\dot{\boldsymbol{\lambda}}(t)^T = \boldsymbol{\lambda}(t)^T\mathbf{A}(t) - \mathbf{x}(t)^T\mathbf{Q}(t)$$

transposing this adjoint equation gives

$$\dot{\boldsymbol{\lambda}}(t) = \mathbf{Q}(t)\mathbf{x}(t) - \mathbf{A}(t)^T\boldsymbol{\lambda}(t)$$

with terminal condition $\boldsymbol{\lambda}(t_1) = \mathbf{0}$ the Hamiltonian becomes

$$H = \boldsymbol{\lambda}(t)^T\mathbf{A}(t)\mathbf{x}(t) + \boldsymbol{\lambda}(t)^T\mathbf{B}(t)\mathbf{u}(t) - \frac{1}{2}\mathbf{x}(t)^T\mathbf{Q}(t)\mathbf{x}(t) - \frac{1}{2}\mathbf{u}(t)^T\mathbf{R}(t)\mathbf{u}(t)$$

with

$$H_{\mathbf{u}} = \boldsymbol{\lambda}(t)^T\mathbf{B}(t) - \mathbf{u}(t)^T\mathbf{R}(t)$$

To maximize the Hamiltonian with respect to $\mathbf{u}(t)$ it is required that $H_{\mathbf{u}} = 0$ or

$$\begin{aligned} \boldsymbol{\lambda}(t)^T\mathbf{B}(t) - \mathbf{u}(t)^T\mathbf{R}(t) &= 0 \\ \implies \mathbf{u}(\mathbf{t}) &= \mathbf{R}(t)^{-1}\mathbf{B}(t)^T\boldsymbol{\lambda}(t) \end{aligned}$$

Substituting this $\mathbf{u}(t)$ into the differential equation gives

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\boldsymbol{\lambda}(t)$$

Use of this new equation and the transposed adjoint equation along with the original initial and final conditions gives a new maximum principle, which is

Theorem 3.1.2 (Maximum Principle for systems with quadratic cost)**.** *Suppose* $\mathbf{u}(t) \in U$ *is the optimal control and* $\mathbf{x}(t)$ *is the state trajectory. Then there is an adjoint trajectory* $\boldsymbol{\lambda}(t)$ *such that* $\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}$ *satisfy*

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\boldsymbol{\lambda}(t) \\ \dot{\boldsymbol{\lambda}}(t) &= \mathbf{Q}(t)\mathbf{x}(t) - \mathbf{A}(t)^T\boldsymbol{\lambda}(t) \\ \mathbf{x}(0) &= \mathbf{x}_0 \\ \boldsymbol{\lambda}(t_1) &= \mathbf{0} \end{aligned}$$

In this system, there are $2n$ differential equations, $2n$ endpoint conditions and $2n$ unknown functions. A problem arises when solving this system as n endpoint conditions occur at $t = 0$ and the other n happen at time $t = t_1$. This problem can be solved numerically [14]. However, it is easier to use something called a *Riccati equation*, section 3.2.

3.1.5 Derivation of the Differential Riccati Equation

The system in Theorem 3.1.2 is linear. $\mathbf{x}(t)$ and $\boldsymbol{\lambda}(t)$ depend linearly on $\mathbf{x}_0$. Also, $\boldsymbol{\lambda}(t)$ linearly depends on $\mathbf{x}(t)$. Knowing this, we try a solution of the following form

$$\boldsymbol{\lambda}(t) = -\mathbf{P}(t)\mathbf{x}(t)$$

[1] where $\mathbf{P}(t)$ is an unknown $n \times n$ matrix, this expression can be substituted into the system. Here

$$\dot{\boldsymbol{\lambda}}(t) = -\mathbf{P}(t)\dot{\mathbf{x}}(t) - \dot{\mathbf{P}}(t)\mathbf{x}(t)$$

which, when substituted into the system, gives

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t) - \mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t)\mathbf{x}(t) \\ -\mathbf{P}(t)\dot{\mathbf{x}}(t) - \dot{\mathbf{P}}(t)\mathbf{x}(t) &= \mathbf{Q}(t)\mathbf{x}(t) + \mathbf{A}(t)^T\mathbf{P}(t)\mathbf{x}(t) \end{aligned}$$

and simplifies to

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \left[\mathbf{A}(t) - \mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t)\right]\mathbf{x}(t) \\ -\mathbf{P}(t)\dot{\mathbf{x}}(t) - \dot{\mathbf{P}}(t)\mathbf{x}(t) &= \left[\mathbf{Q}(t) + \mathbf{A}(t)^T\mathbf{P}(t)\right]\mathbf{x}(t) \end{aligned}$$

Multiplying the first by $\mathbf{P}(t)$ gives

$$\begin{aligned} \mathbf{P}(t)\dot{\mathbf{x}}(t) &= \left[\mathbf{P}(t)\mathbf{A}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t)\right]\mathbf{x}(t) \\ -\mathbf{P}(t)\dot{\mathbf{x}}(t) - \dot{\mathbf{P}}(t)\mathbf{x}(t) &= \left[\mathbf{Q}(t) + \mathbf{A}(t)^T\mathbf{P}(t)\right]\mathbf{x}(t) \end{aligned}$$

Now, adding the two equations together gives

$$\left[\dot{\mathbf{P}}(t) + \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}(t)^T\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t) + \mathbf{Q}(t)\right]\mathbf{x}(t) = \mathbf{0}$$

The above expression will be satisfied $\forall\, \mathbf{x}(t)$ if

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}(t)^T\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t) + \mathbf{Q}(t)$$

and

$$\mathbf{P}(t_1) = \mathbf{0}$$

then $\boldsymbol{\lambda}(t_1) = \mathbf{0}$ holds for all $\mathbf{x}(t)$.

The differential equation above is quadratic in the unknown $\mathbf{P}(t)$. It is called a *differential Riccati equation* [23], which is a system of the form

$$-\dot{\mathbf{P}} = \mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} - \mathbf{P}\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{Q}$$

with final condition $\mathbf{P}(t_1) = \mathbf{0}$. It will be shown later that $\mathbf{P}$ is symmetric for all t. The solution to the *Riccati* differential equation can be found by backward integration starting at $t = t_1$ with the condition $\mathbf{P}(t_1) = \mathbf{0}$, [1]. Once $\mathbf{P}(t)$ is known, one has an expression for $\boldsymbol{\lambda}(t)$ given any state $\mathbf{x}(t)$. Which, in turn, gives an expression for $\mathbf{u}(t)$. Which is

$$\mathbf{u}(t) = -\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t)\mathbf{x}(t) \tag{3.6}$$

or

$$\mathbf{u}(t) = -\mathbf{K}(t)\mathbf{x}(t)$$

where

$$\mathbf{K} = \mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t)$$

3.2 Riccati Equation

The term *Riccati equation* is generally used to refer to matrix equations with an analogous quadratic term [24]. Named after Venetian mathematician Jacopo Riccati. Riccati equations occur in both continuous-time and discrete-time linear-quadratic control problems. In this thesis only the continuous-time problem will be studied. The *Basic Riccati differential equation* for these control problems is an equation of the form

$$-\dot{\mathbf{P}} = \mathbf{E}^T\mathbf{P}\mathbf{A} + \mathbf{A}^T\mathbf{P}\mathbf{E} - (\mathbf{E}^T\mathbf{P}\mathbf{B} + \mathbf{S})\mathbf{R}^{-1}(\mathbf{B}^T\mathbf{P}\mathbf{E} + \mathbf{S}^T) + \mathbf{C}^T\mathbf{Q}\mathbf{C} \tag{3.7}$$

where the matrices above are functions of t. This system can still be solved numerically, but it is quite difficult [24]. A more familiar form occurs when $\mathbf{S} = \mathbf{0}$, $\mathbf{E} = \mathbf{I}$ and $\mathbf{C} = \mathbf{I}$, this simplifies the required theory for the solution [25]. Recall that the *differential Riccati equation* from the previous equation is

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A}(t) + \mathbf{A}(t)^T\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}(t)\mathbf{R}(t)^{-1}\mathbf{B}(t)^T\mathbf{P}(t) + \mathbf{Q}(t) \tag{3.8}$$

This is the simplified version of the basic Riccati differential equation. Recall again that the solution is found by backward integration starting at $t = t_1$ with the condition $\mathbf{P}(t_1) = \mathbf{0}$. This continuous solution is effective but sometimes expensive to calculate, in most cases it can only be calculated numerically [23]. This will be done for the DPIC later in this thesis.

3.2.1 Time-Independent Case

Now suppose that $\mathbf{A}(t), \mathbf{B}(t), \mathbf{Q}(t), \mathbf{R}(t)$ are constant matrices, $\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{R}$, independent of time. Here t_1 goes to infinity. Then, since the Riccati equation is integrated backward in time, the solution $\mathbf{P}(t)$ can be expected to approach a constant matrix $\mathbf{P}$ for t near 0 [1]. Accordingly, $\dot{\mathbf{P}}(t)$ approaches $\mathbf{0}$ [1]. This gives the following *Algebraic Riccati equation* (ARE) that the limiting constant matrix $\mathbf{P}$ solves.

$$\mathbf{PA} + \mathbf{A}^T\mathbf{P} - \mathbf{PBR}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{Q} = \mathbf{0} \tag{3.9}$$

There are three types of solutions to the ARE [26]. The *stabilizing* solution $\mathbf{P}_+$, which, when substituted into the control, and the system, implies that the controlled system matrix $\mathbf{A}-\mathbf{BK}$ has eigenvalues with negative real parts. The *anti-stabilizing* solution $\mathbf{P}_-$, this implies eigenvalues with positive real part for the system matrix. Finally the *mixed* solutions $\mathbf{P}_\theta$, mixed solutions imply that the controlled system matrix has both negative and positive eigenvalue real parts. There is only one stabilizing and one anti-stabilizing solution whereas there are as many as 2n choose n mixed solutions [18]. $\mathbf{P}_+$ and $\mathbf{P}_-$ are structurally stable, the controlled system matrix $\mathbf{A} - \mathbf{BK}$ maintains the same stability properties under data perturbation. $\mathbf{P}_\theta$ may or may not be structurally stable, depending on the different combinations of eigenvalues. $\mathbf{P}_+$, $\mathbf{P}_-$ and $\mathbf{P}_\theta$ are all needed to understand the complete phase portrait of the Riccati differential equations [26]. However, for control purposes the only solution needed is the unique stabilizing solution $\mathbf{P}_+$, this will help make the eigenvalues of $(\mathbf{A} - \mathbf{BK})$ stable [27]. Here the control is of the form

$$\mathbf{u}(t) = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_+\mathbf{x}(t)$$

or

$$\mathbf{u}(t) = -\mathbf{K}_+\mathbf{x} \quad \textit{where} \quad \mathbf{K}_+ = \mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_+$$

giving

$$\dot{\mathbf{x}}(t) = (\mathbf{A} - \mathbf{BK}_+)\mathbf{x}(t)$$

which is the closed-loop system with stable eigenvalues. The ARE is one of the most deeply studied problems in both mathematics and engineering, its use in control theory is well known and of particular interest to this thesis. Finding the stabilizing solution to the algebraic Riccati equation is the topic for the following pages.

3.2.2 A Note on Solution/Numerical Methods

There are many different theoretical approaches and corresponding algorithms to solve the ARE [24]. The classical eigenvalue approach is typically used for most systems [28]. Reliable algorithms for the Riccati equations require sufficient attention to the condition of the underlying problem, the numerical stability of the algorithm and the robustness of the actual software implementation. Each of these help build reliable stabilizing solutions to the ARE [24]. In this thesis, an optimal LQR controller will be derived from the linear DPIC and used in the nonlinear system. The built-in MATLAB function CARE [29], which solves the ARE and outputs the stabilizing solution $\mathbf{P}_+$, to find this controller. CARE, in loose terms, stands for Continuous-time Algebraic Riccati Equation. The code enacts algorithms [24] to numerically calculate the solution to the *basic algebraic Riccati equation* which is the algebraic version of (3.7).

$$\mathbf{E}^T\mathbf{PA} + \mathbf{A}^T\mathbf{PE} - (\mathbf{E}^T\mathbf{PB} + \mathbf{S})\mathbf{R}^{-1}(\mathbf{B}^T\mathbf{PE} + \mathbf{S}^T) + \mathbf{C}^T\mathbf{QC} = \mathbf{0}$$

However, a more familiar form occurs when $\mathbf{S} = \mathbf{0}$, $\mathbf{E} = \mathbf{I}$ and $\mathbf{C} = \mathbf{I}$ to give the *Algebraic Riccati equation* (3.9) as before

$$\mathbf{PA} + \mathbf{A}^T\mathbf{P} - \mathbf{PBR}^{-1}\mathbf{B}^T\mathbf{P} + \mathbf{Q} = \mathbf{0}$$

This simplified system facilitates the theory and the corresponding algorithms [24]. Essentially, the theory behind the solution is as follows. This particular method focuses on a specific set of Schur vectors that span the stable invariant subspace of $\mathbf{M}$, where $\mathbf{M} \in \mathbb{R}^{2n\times 2n}$ is a Hamiltonian matrix partitioned into $\mathbb{R}^{n\times n}$ blocks [25].

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & -\mathbf{G} \\ -\mathbf{Q} & -\mathbf{A}^T \end{bmatrix}$$

where $\mathbf{G} = \mathbf{BR}^{-1}\mathbf{B}^T$. We will see in section 3.2.4 that the stabilizing solution can be obtained by finding an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{2n\times 2n}$ such that

$$\mathbf{U}^T\mathbf{MU} = \mathbf{S}.$$

Note here the relation between the Hamiltonian matrix and the Riccati equation

$$
\begin{aligned}
-\begin{bmatrix}\mathbf{I} & \mathbf{P}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\mathbf{M}\begin{bmatrix}\mathbf{I}\\ \mathbf{P}\end{bmatrix} &= -\begin{bmatrix}\mathbf{I} & \mathbf{P}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{I}\\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{A} & -\mathbf{G}\\ -\mathbf{Q} & -\mathbf{A}^T\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{P}\end{bmatrix}\\
&= -\begin{bmatrix}\mathbf{I} & \mathbf{P}\end{bmatrix}\begin{bmatrix}-\mathbf{Q} & -\mathbf{A}^T\\ -\mathbf{A} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{P}\end{bmatrix}\\
&= -\begin{bmatrix}-\mathbf{Q}-\mathbf{P}\mathbf{A} & -\mathbf{A}^T+\mathbf{P}\mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{I}\\ \mathbf{P}\end{bmatrix}\\
&= -(-\mathbf{Q}-\mathbf{P}\mathbf{A}-\mathbf{A}^T\mathbf{P}+\mathbf{P}\mathbf{G}\mathbf{P})\\
&= \mathbf{Q}+\mathbf{P}\mathbf{A}+\mathbf{A}^T\mathbf{P}-\mathbf{P}\mathbf{G}\mathbf{P}\\
&= \mathbf{P}\mathbf{A}+\mathbf{A}^T\mathbf{P}-\mathbf{P}\mathbf{G}\mathbf{P}+\mathbf{Q}
\end{aligned}
$$

It turns out that, when the Riccati equation is simplified, this method saves valuable computation time over other traditional methods. This method has been proved to be reliable for 'modest' sized problems $n < 100$ [24], as the state is a vector in $\mathbb{R}^6$ this process can be trusted and computed efficiently. Whilst it is the best method for solving the simplified ARE (3.9), its use cannot be extended to a Riccati equation of the form (3.7) or to non-symmetric Riccati problems [24]. To show that the stabilizing solution $\mathbf{P}_+$ can be calculated this way, and to see the specific requirements for it to work, a theorem is formulated. However, it is important to review some necessary definitions and theorems from linear algebra.

3.2.3 Linear Algebra Review

The following definitions, taken from [30], are necessary to formulate the required lemmas.

Definition 3.2.1. $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal if $\mathbf{A}^T = \mathbf{A}^{-1}$

Let $\mathbf{J} = \begin{bmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{bmatrix} \in \mathbb{R}^{2n\times 2n}$ where $\mathbf{I}$ denotes the $n \times n$ identity matrix. Note $\mathbf{J}^T = \mathbf{J}^{-1} = -\mathbf{J}$.

Definition 3.2.2. $\mathbf{A} \in \mathbb{R}^{2n\times 2n}$ is Hamiltonian if $\mathbf{J}^{-1}\mathbf{A}^T\mathbf{J} = -\mathbf{A}$

Definition 3.2.3. $\mathbf{A} \in \mathbb{R}^{2n\times 2n}$ is symplectic if $\mathbf{J}^{-1}\mathbf{A}^T\mathbf{J} = \mathbf{A}^{-1}$

The following lemmas are drawn from known theory in linear algebra [25]. These lemmas will help prove the theorem on the Riccati solution, which is stated in section 3.2.4.

Lemma 3.2.1. *Let $\mathbf{A} \in \mathbb{R}^{2n\times 2n}$ be Hamiltonian. If λ is an eigenvalue of $\mathbf{A}$ with algebraic multiplicity α and geometric multiplicity d, $-\lambda$ is also an eigenvalue of $\mathbf{A}$ with algebraic multiplicity α and geometric multiplicity d.*

Proof. $\mathbf{A}$ is Hamiltonian if $\mathbf{J}^{-1}\mathbf{A}^T\mathbf{J} = -\mathbf{A}$. In other words, $\mathbf{A}$ and $-\mathbf{A}^T$ are similar. If λ is an eigenvalue of $\mathbf{A}$ then it is also an eigenvalue of $-\mathbf{A}^T$. Since the eigenvalues of $-\mathbf{A}^T$ are the negatives of the eigenvalues of $\mathbf{A}$, the lemma is proved. □

Lemma 3.2.2. *Let $\mathbf{A} \in \mathbb{R}^{2n\times 2n}$ be Hamiltonian. Let $\mathbf{U} \in \mathbb{R}^{2n\times 2n}$ be symplectic. Then $\mathbf{U}^{-1}\mathbf{A}\mathbf{U}$ is also Hamiltonian.*

Proof. Omitted. [30] □

Lemma 3.2.3. *Let* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *have eigenvalues* $\lambda_1, \lambda_2, ..., \lambda_n$*, listed according to their algebraic multiplicity. Then there exists an orthogonal similarity transformation* $\mathbf{U}$ *such that* $\mathbf{U}^T\mathbf{A}\mathbf{U}$ *is upper-triangular with diagonal elements* $\lambda_1, \lambda_2, ..., \lambda_n$ *in that order.*

Proof. Omitted. [30] □

Lemma 3.2.3 can also work in block form. By reducing to quasi-upper-triangular form with 2×2 blocks on the diagonal for complex conjugate eigenvalues and single one-dimensional blocks corresponding to the real eigenvalues. One works only over $\mathbb{R}$. This canonical form is known as the *Murnaghan-Wintner canonical form* [31], for this thesis denote it the *real Schur form* (RSF).

Lemma 3.2.4. *Let* $\mathbf{A} \in \mathbb{R}^{n \times n}$*. Then there exists an orthogonal similarity transformation* $\mathbf{U}$ *such that* $\mathbf{U}^T\mathbf{A}\mathbf{U}$ *is quasi-upper-triangular. Moreover,* $\mathbf{U}$ *can be chosen so that* 2×2 *and* 1×1 *diagonal blocks appear in any order.*

Proof. Omitted. [30] □

In lemma 3.2.4 take $\mathbf{S} = \mathbf{U}^T\mathbf{A}\mathbf{U}$ partitioned into blocks such that

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ 0 & \mathbf{S}_{22} \end{bmatrix}$$

with $\mathbf{S}_{11} \in \mathbb{R}^{k \times k}$ with $0 \leq k \leq n$. The first k column vectors of $\mathbf{U}$ are referred to as the Schur vectors corresponding to the eigenvalues of $\mathbf{S}_{11}$. These vectors span the sum of the eigenspaces of $\mathbf{S}_{11}$ even when some of the eigenvalues are multiple. This property will prove to be useful in the following section.

3.2.4 Theorem and Proof

This specific method used to solve the algebraic Riccati equation involves using a specific set of Schur vectors. This method is one that will be used later in this thesis for the DPIC system. The *Algebraic Riccati equation* is

$$\mathbf{PA} + \mathbf{A}^T\mathbf{P} - \mathbf{PGP} + \mathbf{Q} = \mathbf{0} \tag{3.10}$$

Assumptions

where $\mathbf{P}, \mathbf{A}, \mathbf{G}, \mathbf{Q} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$ and $\mathbf{R} \in \mathbb{R}^{m \times m}$. Assume $\mathbf{G} = \mathbf{G}^T$ is positive semi-definite. Recall $\mathbf{Q} = \mathbf{Q}^T$ is assumed to be positive semi-definite. Recall $\mathbf{G} = \mathbf{BR}^{-1}\mathbf{B}^T \in \mathbb{R}^{n \times n}$ and note that $\mathbf{G} = \mathbf{G}^T \geq \mathbf{0}$. $(\mathbf{A}, \mathbf{B})$ is assumed to be a stabilizable pair, see section 1.1.2. Finally, assume full observability, section 2.1.2. Note this was previously assumed when it was set that $\mathbf{C} = \mathbf{I}$. Under these assumptions, the ARE can be shown to have a unique stabilizing solution $\mathbf{P}_+$ of the form (3.6), [27]. From now on $\mathbf{P} = \mathbf{P}_+$. Now consider the Hamiltonian matrix $\mathbf{M} \in \mathbb{R}^{2n \times 2n}$.

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & -\mathbf{G} \\ -\mathbf{Q} & -\mathbf{A}^T \end{bmatrix}$$

$(\mathbf{A}, \mathbf{B})$ is assumed to be a stabilizable pair, this implies that there exists a $\mathbf{K}$ such that the eigenvalues of $(\mathbf{A} - \mathbf{BK})$ lie strictly in the left-half plane. In other words, there are n stable eigenvalues. Let $\mathbf{P}$ be a symmetric solution to the ARE and consider the matrix

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P} & \mathbf{I} \end{bmatrix}$$

$$\begin{aligned}\mathbf{T}^{-1}\mathbf{M}\mathbf{T} &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{P} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{A} & -\mathbf{G}\\ -\mathbf{Q} & -\mathbf{A}^T\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{P} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{A} & -\mathbf{G}\\ -\mathbf{PA}-\mathbf{Q} & \mathbf{PG}-\mathbf{A}^T\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{P} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{A}-\mathbf{GP} & -\mathbf{G}\\ -(\mathbf{A}^T\mathbf{P}+\mathbf{PA}-\mathbf{PGP}+\mathbf{Q}) & -(\mathbf{A}^T-\mathbf{PG})\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{A}-\mathbf{GP} & -\mathbf{G}\\ \mathbf{0} & -(\mathbf{A}-\mathbf{GP})^T\end{bmatrix}\end{aligned}$$

Thus, the eigenvalues of $\mathbf{M}$ are the eigenvalues of $\mathbf{A}-\mathbf{GP}$ together with the eigenvalues of $-(\mathbf{A}-\mathbf{GP})^T$. Because the eigenvalues of $\mathbf{A}-\mathbf{GP}$ equal the eigenvalues of $\mathbf{A}-\mathbf{BK}$, where n eigenvalues are stable, there are also n unstable eigenvalues for $-(\mathbf{A}-\mathbf{GP})^T$. This follows from

$$\mathbf{A}-\mathbf{GP} = \mathbf{A}-\mathbf{BR}^{-1}\mathbf{B}^T\mathbf{P} = \mathbf{A}-\mathbf{BK}$$

So the spectrum of $\mathbf{M}$ includes no purely imaginary eigenvalues [18]. Now apply the lemma 3.2.4, there exists an orthogonal transformation $\mathbf{U} \in \mathbb{R}^{2n\times 2n}$ such that $\mathbf{M}$ is brought to the real Schur form

$$\mathbf{U}^T\mathbf{M}\mathbf{U} = \mathbf{S}.$$

where

$$\mathbf{U} = \begin{bmatrix}\mathbf{U}_{11} & \mathbf{U}_{12}\\ \mathbf{U}_{21} & \mathbf{U}_{22}\end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix}\mathbf{S}_{11} & \mathbf{S}_{12}\\ \mathbf{0} & \mathbf{S}_{22}\end{bmatrix}$$

here $\mathbf{U}_{ij}, \mathbf{S}_{ij} \in \mathbb{R}^{n\times n}$. For $\mathbf{S} \in \mathbb{R}^{2n\times 2n}$, it is possible to arrange $\mathbf{U}$ such that every eigenvalue of $\mathbf{S}_{11}$ has negative real part and all eigenvalues of $\mathbf{S}_{22}$ have positive real part. It is now possible to state and prove the theorem on the stabilizing solution of the ARE.

Stabilizing Solution to the ARE

Theorem 3.2.5. *Using the same notation and under the same assumptions as above. This yields*

(*i*) $\mathbf{U}_{11}$ *is invertible, and* $\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$ *solves the ARE, equation (3.10).*

(*ii*) $\mathbf{S}_{11}$ *is similar to* $(\mathbf{A} - \mathbf{GP})$.

(*iii*) $\mathbf{P} = \mathbf{P}^T$, *the solution is symmetric.*

(*iv*) $\mathbf{P} \geq \mathbf{0}$, *the solution is positive.*

Proof of Stabilizing Solution to the ARE

The proof of this theorem is long, however necessary in order to understand how a workable solution to the algebraic Riccati equation is formulated. Each point will be proven sequentially.

Proof. (i) $\mathbf{U}_{11}$ is invertible, and $\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$ solves the ARE, equation (3.10).

(a) $\mathbf{U}_{11}$ is invertible.

For simplicity assume $\mathbf{S}$ is upper-triangular and $\mathbf{U}$ is orthogonal. Note that

$$\begin{aligned} \mathbf{U}^T\mathbf{M}\mathbf{U} &= \mathbf{S} \\ \mathbf{M}\mathbf{U} &= \mathbf{U}\mathbf{S} \end{aligned}$$

Suppose $\mathbf{U}_{11}$ is singular. Without loss of generality assume that $\mathbf{U}_{11} = \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{U}}_{11} \end{bmatrix}$ where $\mathbf{0} \in \mathbb{R}^n$ and $\tilde{\mathbf{U}}_{11} \in \mathbb{R}^{n\times(n-1)}$ thus

$$\mathbf{M} \begin{bmatrix} \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{U}}_{11} \end{bmatrix} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{U}}_{11} \end{bmatrix} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix} \mathbf{S}$$

$$\begin{bmatrix} \mathbf{A} & -\mathbf{G} \\ -\mathbf{Q} & -\mathbf{A}^T \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{U}}_{11} \end{bmatrix} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \mathbf{0} & \tilde{\mathbf{U}}_{11} \end{bmatrix} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}$$

Thus, the first row of $\mathbf{U}$ yields

$$\begin{bmatrix} \mathbf{A} & -\mathbf{G} \\ -\mathbf{Q} & -\mathbf{A}^T \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{u} \end{bmatrix} \cdot (-\lambda)$$

where $\mathbf{u} \in \mathbb{R}^n$ and $(-\lambda)$ is the top left value of $\mathbf{S}$, $Re(\lambda) > 0$. Expanding gives

$$\begin{bmatrix} -\mathbf{G}\mathbf{u} \\ -\mathbf{A}^T\mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ -\lambda\mathbf{u} \end{bmatrix}$$

This gives

$$\mathbf{A}^T\mathbf{u} = \lambda\mathbf{u} \quad \& \quad -\mathbf{G}\mathbf{u} = \mathbf{0}$$

but for any $\mathbf{P}$

$$\begin{aligned} (\mathbf{A} - \mathbf{G}\mathbf{P})^T\mathbf{u} &= (\mathbf{A}^T - \mathbf{P}^T\mathbf{G})\mathbf{u} \\ &= \mathbf{A}^T\mathbf{u} - \mathbf{P}^T\mathbf{G}\mathbf{u} \\ &= \lambda\mathbf{u} + \mathbf{0} \\ (\mathbf{A} - \mathbf{G}\mathbf{P})^T\mathbf{u} &= \lambda\mathbf{u} \end{aligned}$$

therefore there exists an eigenvalue of $\mathbf{A}$ with positive real part which is uncontrollable. This contradicts the assumption of stabilizability. So, by contradiction, $\mathbf{U}_{11}$ is invertible.

(b) $\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$ solves the algebraic Riccati equation.

To prove this, sub this $\mathbf{P}$ into the Riccati equation. But first recall that the Riccati equation can be expressed in the following form

$$\mathbf{PA} + \mathbf{A}^T\mathbf{P} - \mathbf{PGP} + \mathbf{Q} = -\begin{bmatrix}\mathbf{I} & \mathbf{P}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{A} & -\mathbf{G} \\ -\mathbf{Q} & -\mathbf{A}^T\end{bmatrix}\begin{bmatrix}\mathbf{I} \\ \mathbf{P}\end{bmatrix}$$

Now take the first line above and eventually sub in $\mathbf{P}$

$$\begin{aligned}
&= -\begin{bmatrix}\mathbf{I} & \mathbf{P}\end{bmatrix}\begin{bmatrix}\mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0}\end{bmatrix}\mathrm{M}\begin{bmatrix}\mathbf{I} \\ \mathbf{P}\end{bmatrix} \\
&= -\begin{bmatrix}-\mathbf{P} & \mathbf{I}\end{bmatrix}\mathrm{M}\begin{bmatrix}\mathbf{I} \\ \mathbf{P}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{P} & -\mathbf{I}\end{bmatrix}\mathrm{M}\begin{bmatrix}\mathbf{I} \\ \mathbf{P}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{U}_{21}\mathbf{U}_{11}^{-1} & -\mathbf{I}\end{bmatrix}\mathrm{M}\begin{bmatrix}\mathbf{I} \\ \mathbf{U}_{21}\mathbf{U}_{11}^{-1}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{U}_{21}\mathbf{U}_{11}^{-1} & -\mathbf{I}\end{bmatrix}\mathrm{M}\begin{bmatrix}\mathbf{U}_{11} \\ \mathbf{U}_{21}\end{bmatrix}\mathbf{U}_{11}^{-1} \\
&= \begin{bmatrix}\mathbf{U}_{21}\mathbf{U}_{11}^{-1} & -\mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{11} \\ \mathbf{U}_{21}\end{bmatrix}\mathbf{S}^{11}\mathbf{U}_{11}^{-1} \\
&= (\mathbf{U}_{21}\mathbf{U}_{11}^{-1}\mathbf{U}_{11} - \mathbf{U}_{21})\mathbf{S}_{11}\mathbf{U}_{11}^{-1} \\
&= (\mathbf{U}_{21} - \mathbf{U}_{21})\mathbf{S}_{11}\mathbf{U}_{11}^{-1} \\
&= \mathbf{0}
\end{aligned}$$

$\implies$ $\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$ solves the Riccati equation.

(ii) $\mathbf{S}_{11}$ is similar to $(\mathbf{A} - \mathbf{GP})$.

Recall

$$\mathbf{M}\begin{bmatrix}\mathbf{U}_{11}\\ \mathbf{U}_{21}\end{bmatrix} = \begin{bmatrix}\mathbf{U}_{11}\\ \mathbf{U}_{21}\end{bmatrix}\mathbf{S}_{11}$$

$$\begin{bmatrix}\mathbf{A} & -\mathbf{G}\\ -\mathbf{Q} & -\mathbf{A}^T\end{bmatrix}\begin{bmatrix}\mathbf{U}_{11}\\ \mathbf{U}_{21}\end{bmatrix} = \begin{bmatrix}\mathbf{U}_{11}\\ \mathbf{U}_{21}\end{bmatrix}\mathbf{S}_{11}$$

Recall $\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$ and thus $\mathbf{U}_{21} = \mathbf{P}\mathbf{U}_{11}$. For the upper block part of the above system,

$$\begin{aligned}\mathbf{A}\mathbf{U}_{11} - \mathbf{G}\mathbf{U}_{21} &= \mathbf{U}_{11}\mathbf{S}_{11}\\ \mathbf{A}\mathbf{U}_{11} - \mathbf{G}\mathbf{P}\mathbf{U}_{11} &= \mathbf{U}_{11}\mathbf{S}_{11}\\ (\mathbf{A} - \mathbf{GP})\mathbf{U}_{11} &= \mathbf{U}_{11}\mathbf{S}_{11}\\ \mathbf{U}_{11}^{-1}(\mathbf{A} - \mathbf{GP})\mathbf{U}_{11} &= \mathbf{S}_{11}\end{aligned}$$

Therefore $\mathbf{S}_{11}$ is similar to $(\mathbf{A} - \mathbf{GP})$. They have the same eigensystems.

(iii) $\mathbf{P} = \mathbf{P}^T$, the solution is symmetric.

Let $\mathbf{X} = \mathbf{U}_{11}^T\mathbf{U}_{21}$, thus $\mathbf{U}_{21} = (\mathbf{U}_{11}^{-1})^T\mathbf{X}$ therefore

$$\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$$
$$\mathbf{P} = (\mathbf{U}_{11}^{-1})^T\mathbf{X}\mathbf{U}_{11}^{-1}$$

and

$$\mathbf{P}^T = (\mathbf{U}_{11}^{-1})^T\mathbf{X}\mathbf{U}_{11}^{-1}$$

Thus for $\mathbf{P} = \mathbf{P}^T$ it suffices that

$$\mathbf{X} = \mathbf{X}^T$$

Now consider skew-symmetric, orthogonal $\mathbf{Y} = \mathbf{U}^T\mathbf{J}\mathbf{U}$.

$$\mathbf{Y} = \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix}^T \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix}$$
$$\mathbf{Y} = \begin{bmatrix} \mathbf{U}_{11}^T & \mathbf{U}_{21}^T \\ \mathbf{U}_{12}^T & \mathbf{U}_{22}^T \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix}$$
$$\mathbf{Y} = \begin{bmatrix} -\mathbf{U}_{21}^T & \mathbf{U}_{11}^T \\ -\mathbf{U}_{22}^T & \mathbf{U}_{12}^T \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix}$$
$$\mathbf{Y} = \begin{bmatrix} -\mathbf{U}_{21}^T\mathbf{U}_{11} + \mathbf{U}_{11}^T\mathbf{U}_{21} & -\mathbf{U}_{21}^T\mathbf{U}_{12} + \mathbf{U}_{11}^T\mathbf{U}_{22} \\ -\mathbf{U}_{22}^T\mathbf{U}_{11} + \mathbf{U}_{12}^T\mathbf{U}_{21} & -\mathbf{U}_{22}^T\mathbf{U}_{12} + \mathbf{U}_{12}^T\mathbf{U}_{22} \end{bmatrix}$$

Using the fact that $\mathbf{M}$ is Hamiltonian, i.e. $\mathbf{M} = -\mathbf{J}^{-1}\mathbf{M}^T\mathbf{J} = \mathbf{J}\mathbf{M}^T\mathbf{J}$ and $\mathbf{M}^T = \mathbf{U}\mathbf{S}^T\mathbf{U}^T$ it can be shown that

$$\begin{aligned}
\mathbf{U}^T\mathbf{M}\mathbf{U} &= \mathbf{S} \\
\mathbf{U}^T\mathbf{J}\mathbf{M}^T\mathbf{J}\mathbf{U} &= \mathbf{S} \\
\mathbf{U}^T\mathbf{J}\mathbf{U}\mathbf{S}^T\mathbf{U}^T\mathbf{J}\mathbf{U} &= \mathbf{S} \\
\mathbf{S}^T\mathbf{U}^T\mathbf{J}\mathbf{U} &= \mathbf{U}^T\mathbf{J}^{-1}\mathbf{U}\mathbf{S} \\
\mathbf{S}^T\mathbf{U}^T\mathbf{J}\mathbf{U} &= -\mathbf{U}^T\mathbf{J}\mathbf{U}\mathbf{S} \\
\mathbf{S}^T\mathbf{Y} &= -\mathbf{Y}\mathbf{S}
\end{aligned}$$

Thus

$$\begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}^T \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix} = - \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}$$
$$\begin{bmatrix} \mathbf{S}_{11}^T & \mathbf{0} \\ \mathbf{S}_{12}^T & \mathbf{S}_{22}^T \end{bmatrix} \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix} = - \begin{bmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{0} & \mathbf{S}_{22} \end{bmatrix}$$

For the upper block

$$\mathbf{S}_{11}^T\mathbf{Y}_{11} = -\mathbf{Y}_{11}\mathbf{S}_{11}$$

Via classical Lyapunov theory [32], since $\mathbf{S}_{11}$ is stable, $\mathbf{Y}_{11} = \mathbf{0}$. Therefore

$$\begin{aligned}
-\mathbf{U}_{21}^T\mathbf{U}_{11} + \mathbf{U}_{11}^T\mathbf{U}_{21} &= \mathbf{0} \\
\mathbf{U}_{11}^T\mathbf{U}_{21} &= \mathbf{U}_{21}^T\mathbf{U}_{11} \\
\mathbf{U}_{11}^T\mathbf{U}_{21} &= (\mathbf{U}_{11}^T\mathbf{U}_{21})^T
\end{aligned}$$

So $\mathbf{X} = \mathbf{X}^T$, therefore $\mathbf{P} = \mathbf{P}^T$, the solution is symmetric.

(iv) $\mathbf{P} \geq \mathbf{0}$, the solution is positive.

$$\mathbf{P} = \mathbf{U}_{21}\mathbf{U}_{11}^{-1}$$
$$\mathbf{X} = \mathbf{U}_{11}^{T}\mathbf{U}_{21}$$
$$\mathbf{P} = (\mathbf{U}_{11}^{-1})^{T}\mathbf{X}\mathbf{U}_{11}^{-1}$$

Therefore $\mathbf{P} \geq \mathbf{0}$ if $\mathbf{X} \geq \mathbf{0}$. Define

$$\mathbf{F}(t) = \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} e^{t\mathbf{S}_{11}}$$

Note

$$\mathbf{F}(0) = \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} \quad \& \quad \lim_{t\to\infty^+} \mathbf{F}(t) = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

since $\mathbf{S}_{11}$ is stable.

$$\dot{\mathbf{F}}(t) = \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} \mathbf{S}_{11} e^{t\mathbf{S}_{11}}$$
$$\dot{\mathbf{F}}(t) = \mathbf{M} \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} e^{t\mathbf{S}_{11}}$$
$$\dot{\mathbf{F}}(t) = \mathbf{M}\mathbf{F}(t)$$

Now, let

$$\mathbf{G}(t) = \mathbf{F}^T(0)\mathbf{L}\mathbf{F}(0) - \mathbf{F}^T(t)\mathbf{L}\mathbf{F}(t)$$

where $\mathbf{L} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ here

$$\mathbf{G}(t) = -\int_0^t \frac{d}{ds}[\mathbf{F}^T(s)\mathbf{L}\mathbf{F}(s)]ds$$
$$\mathbf{G}(t) = -\int_0^t \mathbf{F}^T(s)[\mathbf{M}^T\mathbf{L} + \mathbf{L}\mathbf{M}]\mathbf{F}(s)ds$$
$$\mathbf{G}(t) = -\int_0^t \mathbf{F}^T(s) \begin{bmatrix} -\mathbf{Q} & \mathbf{0} \\ \mathbf{0} & -\mathbf{G} \end{bmatrix} \mathbf{F}(s)ds \leq \mathbf{0}$$

Therefore the integral is greater than $\mathbf{0}$ because $\mathbf{Q}$ and $\mathbf{G}$ are positive semi-definite,

$$\mathbf{G}(t) \geq \mathbf{0} \quad \forall t$$

$$\begin{aligned} \lim_{t\to\infty^+} \mathbf{G}(t) &= \mathbf{F}^T(0)\mathbf{L}\mathbf{F}(0) \\ &= \begin{bmatrix} \mathbf{U}_{11}^T & \mathbf{U}_{21}^T \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{U}_{11}^T \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} \\ \mathbf{U}_{21} \end{bmatrix} \\ &= \mathbf{U}_{11}^T\mathbf{U}_{21} \geq \mathbf{0} \end{aligned}$$

Therefore $\mathbf{X} \geq \mathbf{0} \implies \mathbf{P} \geq \mathbf{0}$. This concludes the proof. □

Consequences

Thus we can find a unique stabilizing solution using Theorem 3.2.5, this is what the built-in MATLAB function CARE uses. We can now apply this to the linear DPIC system (2.7).

3.3 Optimal Control Theory and the DPIC System

To test the different controls on the DPIC system from the different Riccati equations, we must first build a cost function. For this DPIC system we use a *Linear Quadratic Regulator* (LQR).

3.3.1 Linear Quadratic Regulator

Recall the linear DPIC system (2.7), with the specific parameters from [9], defined by

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{2.7}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -7.35 & 0.7875 & 0 & 0 & 0 \\ 0 & 73.5 & -33.075 & 0 & 0 & 0 \\ 0 & -58.8 & 51.1 & 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.6071 \\ -1.5 \\ 0.2857 \end{bmatrix}$$

The linear quadratic regulator for the linear system above is an optimal control that minimizes the cost function

$$J(\mathbf{u}) = \int_0^{t_1} (\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}) dt \tag{3.11}$$

For the DPIC system, the control force acts horizontally only on the cart, affecting only one of the degrees of freedom, see Figure 1.5. Therefore the allowable controls are positive or negative one-dimensional values, $u(t) \in \mathbb{R}$. This implies $\mathbf{R}$ will be one-dimensional, $R \in \mathbb{R}$. Recall R measures the cost of control and $\mathbf{Q}$ measures the penalty for each component of the state being in the incorrect position [1]. Also from [9], take

$$\mathbf{Q} = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 50 & 0 & 0 & 0 & 0 \\ 0 & 0 & 50 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & 0 \\ 0 & 0 & 0 & 0 & 700 & 0 \\ 0 & 0 & 0 & 0 & 0 & 700 \end{bmatrix}, \quad R = 1 \tag{3.12}$$

3.3.2 Different Riccati Controls

Two different controllers which minimize (3.11) will be derived. The first will use the stabilizing solution to the ARE (3.9) and the second will use a time-dependent solution to the Riccati differential equation (3.8). The first method will use the method discussed in the previous section to compute the constant stabilizing solution $\mathbf{P}_+$ to the algebraic Riccati equation (3.9).

$$\mathbf{PA} + \mathbf{A}^T\mathbf{P} - \mathbf{PGP} + \mathbf{Q} = \mathbf{0} \tag{3.9}$$

The second method computes a time-dependent Riccati solution $\mathbf{P}(t)$ to the differential Riccati equation (3.8) using Runge-Kutta numerical methods [13]. However here we use constant matrices $\mathbf{A}$ and $\mathbf{B}$ from the linearized DPIC system (2.7) and $\mathbf{Q}$ and R from (3.12).

$$-\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A} + \mathbf{A}^T\mathbf{P}(t) - \mathbf{P}(t)\mathbf{B}\mathbf{B}^T\mathbf{P}(t) + \mathbf{Q} \tag{3.13}$$

with final condition $\mathbf{P}(t_1) = \mathbf{0}$. Note that the first solution at time $t = 0$, $\mathbf{P}(0)$ should, and does, approximately equal $\mathbf{P}_+$ [23]. These controls should be effective locally for the nonlinear problem, close to the vertical where the system (1.7) was linearized. Here the linear dynamics of the system overpower the nonlinear dynamics.

Chapter 4

System Results

In order to test the performance of the two controllers, three different simulations are run. The first looks at the performance of both controllers in the nonlinear system and compares them to the results in Figures 2.1 and 2.2. The second looks at the set of angles at which the controllers are successful, according to the accuracy required in 2.2.4. This test finds the maximum angle of deflection from the vertical at which the controller will work. Finally, the third looks at the effect of drag on controller performance and its effect on the first two tests for the DPIC system. The first control type will be obtained by finding the constant stabilizing solution $\mathbf{P}_+$ to the algebraic Riccati equation (3.9). Denote this control the *constant-K control*

$$u(t) = -\mathbf{K}_+\mathbf{x}(t) = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_+ \tag{4.1}$$

The second control type will be found by solving the differential Riccati equation (3.13) for a time-dependent solution $\mathbf{P}(t)$. The control here will be

$$u(t) = -\mathbf{K}(t)\mathbf{x}(t) = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}(t) \tag{4.2}$$

Denote this the *time-K control*. From now on, any result in red has a constant-K control of the form (4.1) and any result in blue denotes time-K control of the form (4.2).

4.1 Different Riccati Controls

The first test of the controls derived from the linear system is simply utilizing them in the nonlinear system and observing the results.

Objective

For these simulations the nonlinear system is placed at initial condition

$$\mathbf{x}_0 = (0, 5^\circ, 5^\circ, 0, 0, 0)^T \tag{4.3}$$

at time $t = 0$. The objective is to use the controllers to move the DPIC state vector to the pendulum-up position

$$\mathbf{x}(t) = (0, 0, 0, 0, 0, 0)^T \tag{4.4}$$

using the controllers (4.1) and (4.2). We model this process using MATLAB and extract Figures for the evolution of each controlled state $\mathbf{x}(t) \in \mathbb{R}^{6\times 6}$. We record the time when every component of the state vector has a sufficiently small magnitude,

$$|x_i| < 0.01 \quad \text{for} \quad i = 1, \ldots, 6. \tag{4.5}$$

and denote this time t_f, the time taken to get to the pendulum-up position. In these Figures, each subgraph corresponds to a component of the state vector. Note that all angles appear in radians.

4.1.1 Constant-K Control $\mathbf{K}_+$

The evolution of the state when controlled using the constant-K control (4.1) is shown in Figure 4.1. The controller here takes 11 seconds to get to the pendulum-up

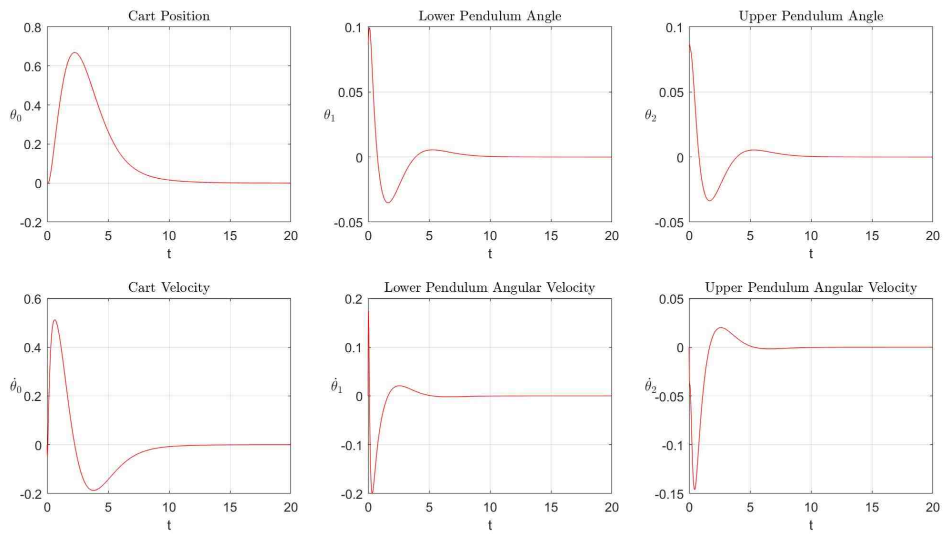

Figure 4.1: State Evolution for Constant-K control

position (4.4) at accuracy given in (4.5), $t_f = 11$. In all state components the controller (4.1), after the first few major corrections, causes each component to converge asymptotically to zero. Consider the cost function (3.11) with $\mathbf{Q}$ and R from (3.12)

$$J(u) = \int_0^{t_1} (\mathbf{x}^T \mathbf{Q} \mathbf{x} + u^2) dt \tag{4.6}$$

For these parameters, and using time $t_1 = 11$ for the integral limit, the value of cost function (4.6) can be computed numerically. The constant-K control (4.1) has cost $J(-\mathbf{K}_+ \mathbf{x}(t)) = 43$ in Joules. When compared to the results from eigenvalue placement in section 2.2.2, the control is not as fast as shown in Figure 2.2, and not as cheap as shown in Figure 2.1. However, this is the optimal control that minimizes the cost function 4.6.

4.1.2 Time-K Control $\mathbf{K}(t)$

The evolution of the controlled state using the time-K control (4.2) is shown in Figure 4.2. Here, for the controlled system, the objective (4.5) is achieved at $t_f =$

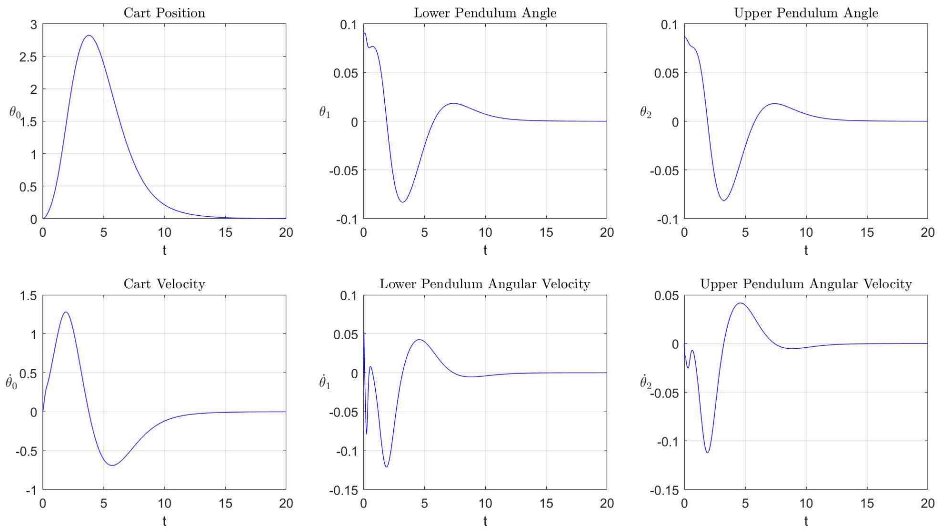

Figure 4.2: State Evolution for Time-K control

16. All components of the state converge asymptotically to zero. When evaluating the cost function (4.6) numerically over this time interval from 0 to t_f, the cost is $J(-\mathbf{K}(t)\mathbf{x}(t)) = 263$. For $\theta_1 = \theta_2 = 5°$ the constant-K controller (4.1) is cheaper than the time-K controller (4.2). Table 4.1 shows the costs $J(u)$ at larger angles β. The time-K control (4.2) is more expensive than the constant-K control (4.1) for the angles β. However, the scale decreases for the bigger angles, but even then the constant-K control is smaller than a third of the cost of the time-K controller. The time-K control is not as 'optimal' considering the objective of minimizing the cost function. However, it may be more encompassing.

β	5°	10°	15°	20°	25°	30°	35°
$J(-\mathbf{K}_+\mathbf{x}(t))$	43	174	395	712	1148	1780	3017
$J(-\mathbf{K}(t)\mathbf{x}(t))$	263	1043	2311	3999	5995	8079	9763

Table 4.1: Quadratic Cost (Joules) for both Controllers

4.2 Initial Conditions

The second test of the different linear controls tests which initial conditions for the state variable $\mathbf{x}$ will yield 'successful' results. The simulation run in this section tests either method of control at different initial angles. The controller could also be tested at different initial velocities. However, velocities are not always directly measurable, it is more reliable to give data on the angles only.

4.2.1 Complete Initial Conditions

Objective

In this test the nonlinear system is simulated with initial conditions

$$\mathbf{x}_0 = (0, \theta_1, \theta_2, 0, 0, 0)^T \tag{4.7}$$

for every possible combination of $\theta_1, \theta_2 \in [0, 360°]$ there are 129,600 of these combinations. For every initial condition (4.7) the control is tasked with moving the DPIC state vector to the pendulum-up position (4.4) from section 4.1

$$\mathbf{x}(t) = (0, 0, 0, 0, 0, 0)^T \tag{4.4}$$

and accuracy

$$|x_i| < 0.01 \quad \text{for} \quad i = 1, \ldots, 6. \tag{4.5}$$

If the control satisfies (4.5), we call it successful. The initial state corresponding to the successful control is stored and plotted in Figures 4.3 and 4.4.

Result

In Figure 4.3 all initial conditions at which the controllers (4.1) and (4.2) are successful (4.5) are plotted. If the controller is successful the initial state $\mathbf{x}_0$ (4.7) is

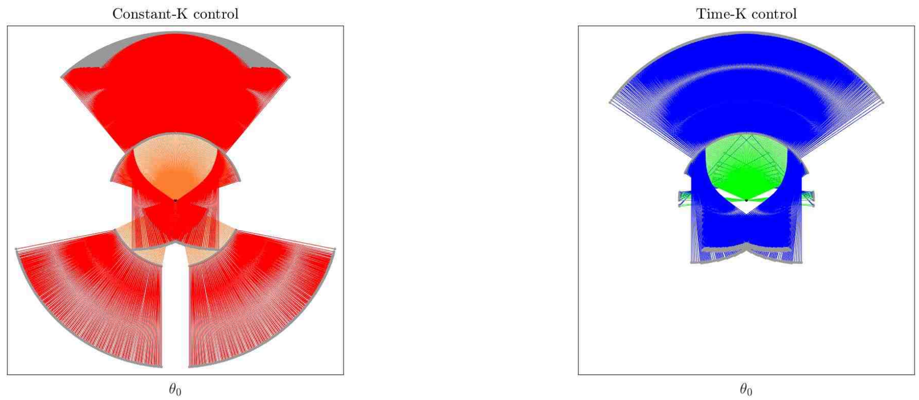

Figure 4.3: Successful Starting DPIC Positions for both Controllers

plotted in Cartesian coordinates in the Figure. The left plot of Figure 4.3 shows the successful initial conditions with the constant-K control (4.1). The right plot of Figure 4.3 shows the successful initial conditions with the time-K control (4.2). Note that the constant-K control (4.1) covers a greater percentage of initial conditions than the time-K control (4.2). For any random pair of angles corresponding to an initial condition (4.7), it more likely that the constant-K control (4.1) successfully controls the system than the time-K control (4.2). However, the regions further from the pendulum-up position (4.4), are less reliable and considered coincidental. In Figure 4.3 there is a clear cone showing how large a deflection from the pendulum-up position (4.4) is possible. We call this the *maximal angle of deflection.* The results with successful initial conditions with either $|\theta_1|, |\theta_2| > \frac{\pi}{2}$ are disregarded.

4.2.2 Maximum Angle of Deflection

Modified objective

Here the objective is the same as in section 4.2.1 except the initial condition is

$$\mathbf{x}_0 = (0, \theta, \theta, 0, 0, 0)^T \tag{4.8}$$

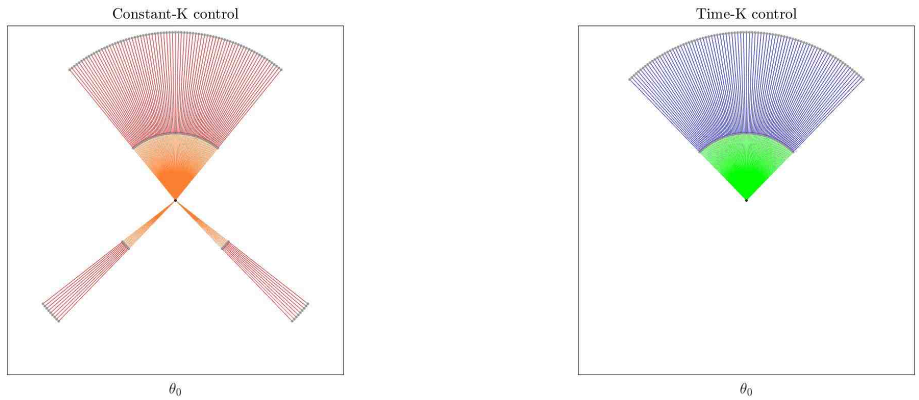

Figure 4.4: Successful Straight Starting DPIC Positions for both Controllers

We test straight starting positions for the DPIC where $\theta_1 = \theta_2$. The successful straight initial states for both controllers are plotted in Figure 4.4. This cone of initial conditions that are successful is of particular importance because, if the controllers (4.1) and (4.2) work for these initial angles, the controlled state trajectory will stay within a valid linear region [33] whilst moving back to the pendulum-up position (4.4). The state vector moves through other starting positions that also work when tested with linear controls (4.1) and (4.2). Anything from the results outside the cone will pass through other states with angles θ_1, θ_2 that are unsuccessful when used as initial conditions.

Result

The left plot of Figure 4.4 shows the successful straight initial conditions with the constant-K control (4.1). The cone encompasses the maximum angle of deflection α of the DPIC to the vertical at which the controller is successful in bringing the state to the pendulum-up position. For the constant-K control (4.1) the maximum angle of deflection is $\alpha = 39°$ in either direction. The successful initial conditions where $\theta > \frac{\pi}{2}$ are ignored. The right plot of Figure 4.4 shows the successful straight initial

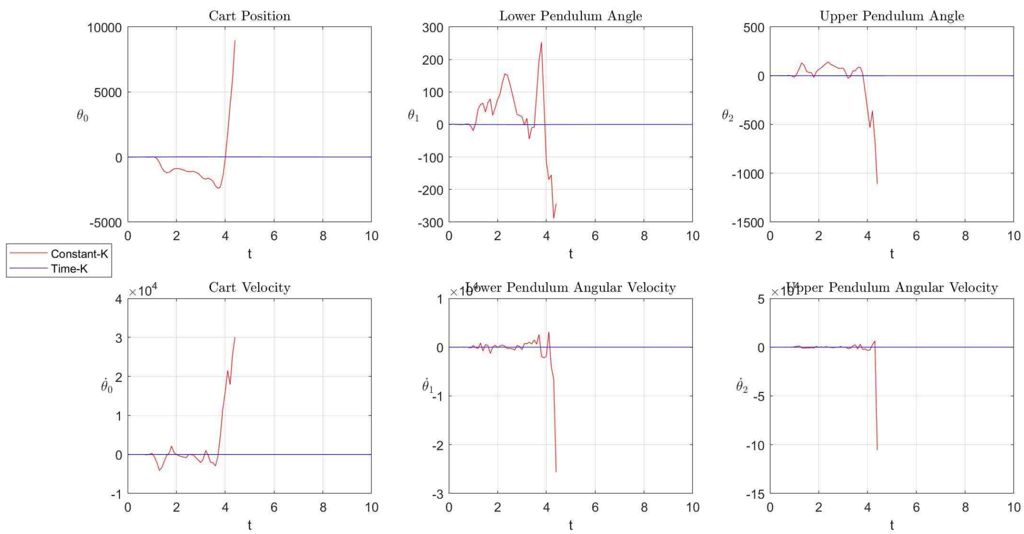

Figure 4.5: State Evolution for Time-K and Constant-K controllers at 40°

conditions with the time-K control (4.2). The maximum angle of deflection here is $\alpha = 44°$ in either direction. See what happens at 40° in Figure 4.5, the first straight initial condition at which the constant-K control (4.1) fails. The constant-K control (4.1) in red attempts control but collapses on itself, the nonlinear mechanics quickly spiral the pendula out of control, accelerating rapidly. The step size eventually becomes smaller than MATLAB's machine epsilon ϵ_M [14], causing the numerical ordinary differential equation solver to fail just after time $t = 4$. The straight blue line uses the time-K control (4.2), this is successful at initial condition (4.8) with $\theta = 40°$. One cannot see the exact movement of the pendulum in this Figure, appearing as a straight line due to the sheer size of the divergence of the constant-K control. Figure 4.6 shows the state evolution under this control at (4.8) with $\theta = 40°$ in more detail. Recall in all Figures with angles, the angles appear in radians. For more extreme values, initial conditions with $\theta \in [40, 41, 42, 43, 44]$, the time-K control (4.2) is successful. The constant-K control (4.1) is unsuccessful.

In section 3.2.4, a specific method was used to save computation time when solving

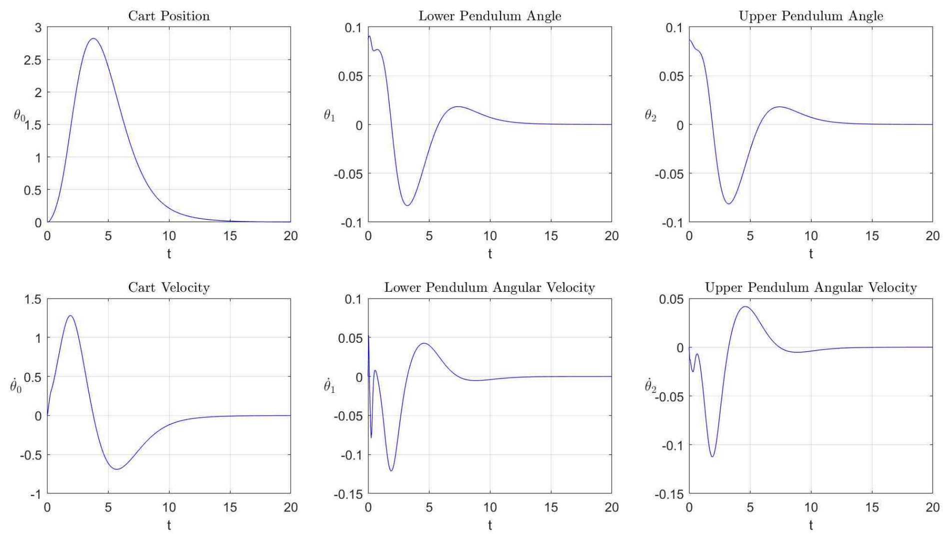

Figure 4.6: State Evolution for Time-K controller at 40°

the algebraic Riccati equation (3.9). Table 4.2 shows the computation time τ taken in seconds, for each method to compute controls at different angles β. τ_+ is the computation time for constant-K control (4.1) and τ_t the computation time for the time-K control (4.2).

β	5°	10°	15°	20°	25°	30°	35°	40°	45°
τ_+	0.0493	0.0132	0.0354	0.0500	0.0422	0.0213	0.0312	Fails	-
τ_t	0.2891	0.2551	0.2365	0.2362	0.2934	0.2422	0.2499	0.2632	Fails

Table 4.2: Computation Time (seconds) for both Controllers

The computation time does not change much with the size of the angle. The computation time for constant-K control (4.1) is smaller than 0.05s, the computation time for time-K control (4.2) is roughly 0.24s. Therefore, computing the time-K control (4.2) is significantly slower than constant-K control (4.1). So as well as the time-K controller (4.2) being slower in time taken to get to the pendulum-up position (4.4) with accuracy (4.5), the time-K controller (4.2) is also slower computationally. These computational and practical observations help explain why most modern control problems solve the algebraic Riccati equation (3.9) instead of computing a time-dependent solution to (3.8) [24].

4.3 Resistance Effect

In reality all oscillators have some form of drag, it can be very small but it can also affect the dynamics of the system. For the DPIC 1.5, the drag force could be air around the pendulum or friction within the joints. Now consider the DPIC system with these forces. The drag forces, $\boldsymbol{\mu}$ on DPIC are approximated to be proportional to the velocities of the various components of $\boldsymbol{\theta}$, here

$$\boldsymbol{\mu} = \begin{bmatrix} d_0\dot{\theta}_0 \\ d_1\dot{\theta}_1 \\ d_2\dot{\theta}_2 \end{bmatrix} \tag{4.9}$$

The resisting forces on the cart is roughly proportional to the velocity of the cart $\boldsymbol{\theta}$. The drag force on each pendulum link is approximated to be proportional to the angular velocity of the respective pendulum link. Modifying the nonlinear 2nd-order system (1.6) to accommodate these new forces gives

$$\mathbf{D}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathbf{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\dot{\boldsymbol{\theta}} + \mathbf{G}(\boldsymbol{\theta}) = \mathbf{H}u + \boldsymbol{\mu}$$

The drag terms can absorbed into the diagonal of the $\mathbf{C}$ matrix. This gives the nonlinear first-order system as before (1.7) but with a different $\mathbf{A}$ due to the change in $\mathbf{C}$. This is slightly more nonlinear than previously.

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x})\mathbf{x} + \mathbf{B}(\mathbf{x})u + \mathbf{L}(\mathbf{x}) \tag{4.10}$$

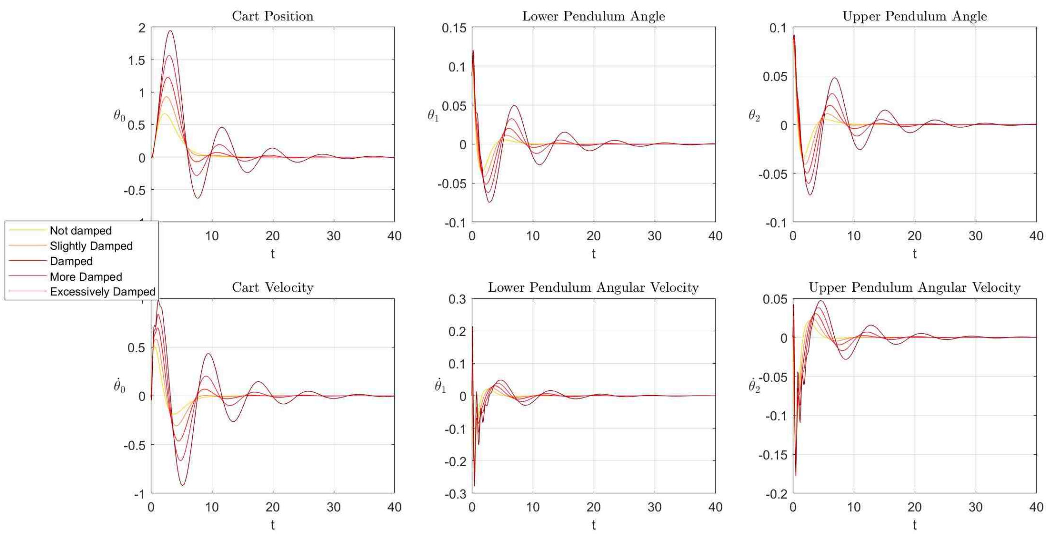

Figure 4.7: State Evolution for Constant-K controller for Different Resistances

Objective

In these tests the DPIC begins at initial condition (4.3) as in section 4.1 and the controllers are tasked with moving the state to the pendulum-up position (4.4). Because the horizontal position of the cart is directly controlled by $u(t)$, set $d_0 = 0$. Set $d_1 = d_2 = d$ in (4.9). For each test the resistance coefficient d is gradually increased from 0 to 1 in the system (4.10). Figure 4.7 shows the effect of the different resistance coefficients on constant-K control (4.1) and Figure 4.8 shows the effect of the different resistance coefficients on time-K control (4.2). The darker the line, the stronger the resistance. Recall angles in Figures are in radians.

4.3.1 Constant-K Control $\mathbf{K}_+$

Figure 4.7 shows the state evolution for the controlled DPIC using constant-K control (4.1) at increasing values of drag. Different values of d in (4.9) are tested. As drag

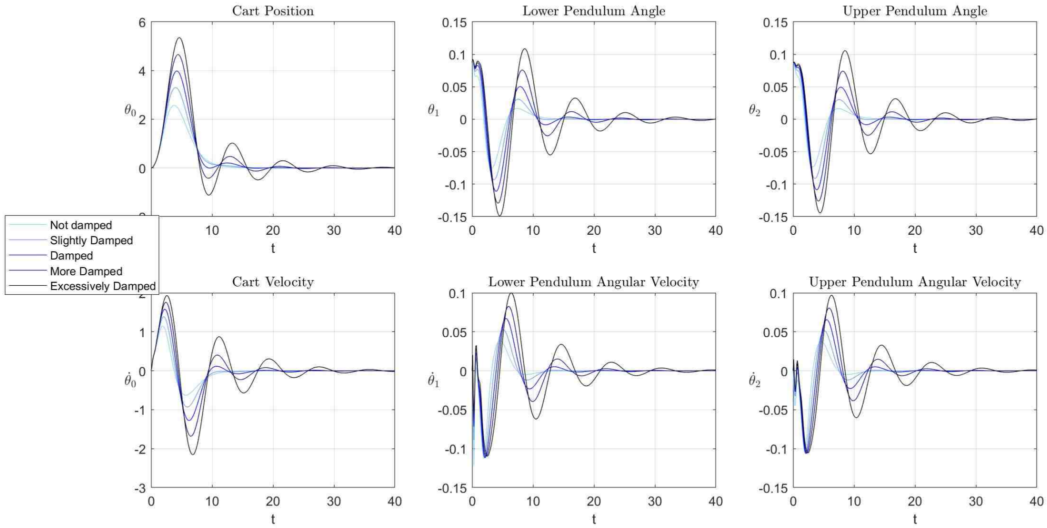

Figure 4.8: State Evolution for Time-K controller for Different Resistances

increases the movements of each component in the state trajectory are stretched. Drag increases the time taken to get the desired result, and increases the amount of control needed to get to the pendulum-up position (4.4).

4.3.2 Time-K Control $\mathbf{K}(t)$

Figure 4.8 shows the effect of resistance on the state trajectory of the time-K control (4.2) controlled system. Fluctuations of the state vector become increasingly stretched as the drag gets larger. The time taken t_f to get the pendulum-up position (4.4) also increases. Comparing Figures 4.7 and 4.8, the effect of drag is similar for both controls. Drag stretches movements vertically and extends the time taken to get to the pendulum-up position (4.4). The differences between Figure 4.7 and Figure 4.8 are similar to the differences between Figures 4.1 and 4.2, where there is no resistance present. With resistance, constant-K control (4.1) in Figure 4.7 takes a shorter amount of time than the time-K control (4.2) in Figure 4.8.

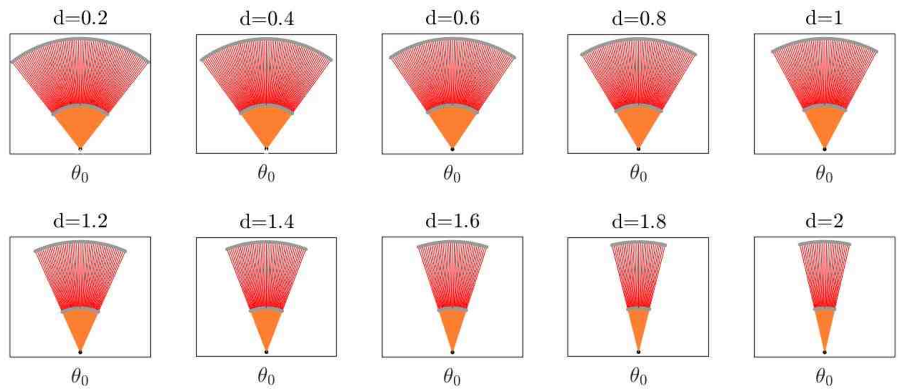

Figure 4.9: Maximum Angles of Deflection for Resistances with Constant-K control.

4.4 Resistance Effect on the Maximum Angle of Deflection

Drag also affects the maximum angle of deflection α from section 4.2.2. The air resistance or the friction in the joints both help and hinder the controllers in certain cases.

Objective

The objective here is the same as in 4.2.2. This time the maximum angle of deflection α is found for varying levels of resistance. The goal is to get to the pendulum-up position (4.4) at each starting position for each value of resistance $d_1 = d_2 = d$ in (4.9). Figure 4.9 plots all successful straight initial conditions (4.7) for the constant-K control (4.1). Figure 4.10 plots all successful straight initial conditions (4.7) for the time-K control (4.2). Table 4.3 shows the maximum angle of deflection α for each controller at the different levels of resistance.

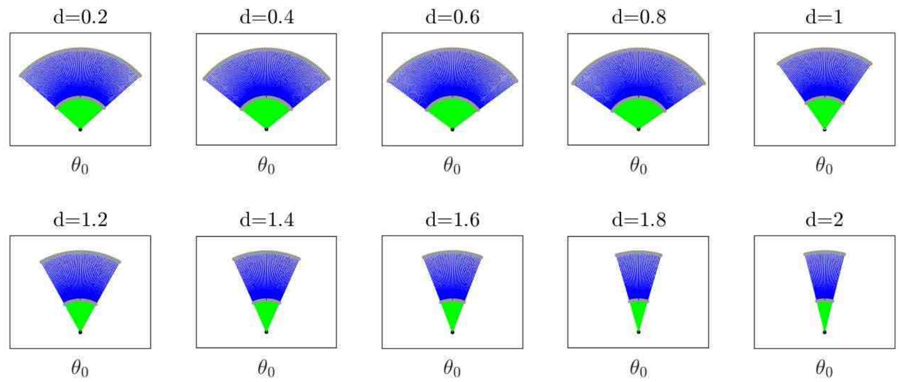

Figure 4.10: Maximum Angles of Deflection for Resistances with Time-K control.

4.4.1 Constant-K Control $\mathbf{K}_+$

For constant-K control (4.1) in Figure 4.9, the maximum angle of deflection α decreases as the forces disrupting the system increase. At $d = 2$ the maximum angle of deflection decreases to $\alpha = 13°$.

4.4.2 Time-K Control $\mathbf{K}(t)$

For time-K controller (4.2) in Figure 4.10, the maximum angle of deflection α increases for resistance values up to $d = 0.8$. At $d = 0.8$ the maximum angle of deflection $\alpha = 55°$. However, after this point the size of the maximum angle of deflection begins to decrease, at a drag coefficient of $d = 2$ the maximum angle of deflection is only $\alpha = 14°$. All results from Figures 4.9 and 4.10 are summarized in table 4.3. α_+ denotes the maximum angle of deflection for the constant-K control (4.1), α_t denotes the maximum angle of deflection for the time-K control (4.2).

d	0	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
α_+	39	38	36	34	31	28	24	21	18	14	13
α_t	44	48	51	53	55	35	29	24	21	16	14

Table 4.3: Maximum Angle of Deflection α for both Controllers with Varying Drag

The time-K control (4.2) has a larger cone of straight initial conditions (4.7) that are successful (4.5) for every different drag coefficient. It even improves under certain smaller forces of drag $d = [0, 0.8]$, before the coefficients get too large and the cone's size decreases $d > 0.8$. If the goal is to attain the largest valid region of these successful straight initial conditions, in all of which the DPIC is successfully controlled to the pendulum-up position (4.4), the time-K control (4.2) is a better controller than the constant-K control (4.1).

Chapter 5

Conclusions

5.1 Conclusion

This thesis has shown that, for parameters from [9], the nonlinear Double Pendulum Inverted on a cart 1.6 can be controlled by linear controllers $u(t) = \mathbf{K}_{+}\mathbf{x}(t)$ and $u(t) = \mathbf{K}(t)\mathbf{x}(t)$ from (4.1) and (4.2) in section 4. Using results in section 2.1 the linear version of the nonlinear DPIC system is controllable. The time-K control (4.2), derived from solving the Riccati differential equation (3.8), performs better under resistance forces d, and has a larger maximum angle of deflection α. However, it is not fast, neither objectively nor computationally. Time taken to get to the pendulum-up position (4.4) with accuracy (4.5) t_f is slower than the t_f for the constant-K control (4.1). Computation time τ_t for the time-K control is also larger than the computation time for the constant-K control τ_{+}, see table 4.2. Solving the Riccati differential equation (3.8) numerically is slower than using Theorem 3.2.5 to solve the algebraic Riccati equation (3.9). The only time it would be beneficial to use the time-K control (4.2) would be if the initial angle was large $\theta \geq 40°$ in (4.7) or the resistance values where time-K control performance increases $0.2 \leq d \leq 0.8$.

Even then one would have to accept huge delays in computation time τ and objective time t_f. Therefore, the algebraic Riccati equation (3.9) and the algorithm in which it is solved, MATLAB's CARE function in section 3.2.2, is a fast and reliable solution for finding a Linear Quadratic Regulator for the DPIC control problem (1.6). When used in the nonlinear system it can be relied upon to produce quick efficient results. This result helps explain why the ARE is still used today in many control problems [9][24].

5.2 Future Work

To enhance this thesis and the particular control methods used, it would be interesting to use the maximum angle of deflection α result to determine a new state to linearize about and return to, say $\theta = 30^\circ$, deriving an LQR control and a respective new maximum angle of deflection for $\theta = 30^\circ$. This process could be completed multiple times creating a chain of linear controls for the entire state, moving through each region and the corresponding controllers. However this process may not be optimal as there is only optimal control within each linear region and its corresponding LQR controller, there may be a control which is sub-optimal for each region but optimal when considering the whole trajectory.

Furthermore, the time-K control (4.2) could be improved by solving the Riccati equation with time-dependent $\mathbf{A}$ and $\mathbf{B}$ giving a full *State-Dependent Riccati Equation Control* (SDRE), which would perhaps perform better when compared to constant-K control (4.1). There are also other types of control than the linear closed-loop controls used in this thesis. Neural Networks are also used in modern control problems [9], these Neural Networks can even be integrated into LQR and SDRE methods to improve performance [9].

Another consideration that was not explored in this thesis was observability 2.1.2. Dual to controllability [1], full observability is the notion that all of the state components can be determined based on the information in the output [17]. It was assumed in this thesis that we can directly measure each component of the state 2.2.3. If not, we would have to prove that system (2.5) is observable using the observability matrix $\mathbf{S}$ from theorem 2.1.4. Observability also effects the optimal control, using the basic Riccati differential equation (3.6) instead of the ARE (3.9) or the differential Riccati equation (3.8), complicating the solution process.

Appendices

Appendix A

Derivation of the Maximum Principle, Theorem 3.1.1

Firstly assume $\mathbf{u}(t)$ is optimal. Make a small change in $\mathbf{u}(t)$ and determine the corresponding change in the objective J. This change should be negative, non-improving as the original $\mathbf{u}(t)$ is optimal. However, complications arise here because changing the control $\mathbf{u}(t)$ will change the state vector $\mathbf{x}(t)$. To overcome this problem employ a 'trick' to the objective function. Define a new modified objective function

$$\tilde{J} = J - \int_0^{t_1} \boldsymbol{\lambda}(t)^T \left[\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t))\right] dt$$

This function gives more freedom because the term within the square brackets is zero for any state trajectory $\mathbf{x}(t)$. Therefore, for any choice of $\boldsymbol{\lambda}(t)$, $\tilde{J} = J$ where $\mathbf{x}(t), \mathbf{u}(t)$ satisfy the differential equations and initial conditions in the *basic optimal control problem.* Now consider the problem maximizing $\tilde{J}$ instead of J. The added choice of $\boldsymbol{\lambda}(t)$ helps simplify the problem. For convenience, define the Hamiltonian function, note $\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}$ are all functions of t,

$$H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) = \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{u}) + l(\mathbf{x}, \mathbf{u})$$

Now express $\tilde{J}$ in terms of H

$$\tilde{J} = J - \int_0^{t_1} \boldsymbol{\lambda}^T \big[\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}, \mathbf{u})\big] dt$$

$$\tilde{J} = \left(\phi(\mathbf{x}(t_1)) + \int_0^{t_1} l(\mathbf{x}, \mathbf{u}) dt\right) - \int_0^{t_1} \boldsymbol{\lambda}^T \dot{\mathbf{x}} - \boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{u}) dt$$

$$\tilde{J} = \phi(\mathbf{x}(t_1)) + \int_0^{t_1} \big[\boldsymbol{\lambda}^T \mathbf{f}(\mathbf{x}, \mathbf{u}) + l(\mathbf{x}, \mathbf{u}) - \boldsymbol{\lambda}^T \dot{\mathbf{x}}\big] dt$$

$$\tilde{J} = \phi(\mathbf{x}(t_1)) + \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) - \boldsymbol{\lambda}^T \dot{\mathbf{x}}\big] dt$$

Now suppose that one specifies a nominal control $\mathbf{u} \in U$, and thus determines a corresponding state trajectory $\mathbf{x}$. Now consider a small change to a new control $\mathbf{v}(t) \in U$. The absolute value of difference is small in the L_1-norm for each component of control.

$$\int_0^{t_1} |u_i(t) - v_i(t)| dt < \epsilon \qquad \forall i \in \{1, 2, ..., m\}$$

However, the actual change can be large over a very short time interval. The new control $\mathbf{v}(t)$ leads to a new state trajectory, denote as $\mathbf{x}(t) + \delta\mathbf{x}(t)$. $\delta\mathbf{x}(t)$ is small for all t because the state depends on the integral of the control function. There is also a corresponding change in the modified objective, denote this with $\delta\tilde{J}$.

$$\delta\tilde{J} = \phi(\mathbf{x}(t_1) + \delta\mathbf{x}(t_1)) - \phi(\mathbf{x}(t_1)) + \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x} + \delta\mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) - \boldsymbol{\lambda}^T \delta\dot{\mathbf{x}}\big] dt$$

Calculating the final term by integrating by parts yields

$$\int_0^{t_1} \boldsymbol{\lambda}^T \delta\dot{\mathbf{x}} dt = \boldsymbol{\lambda}^T(t_1)\delta\mathbf{x}(t_1) - \boldsymbol{\lambda}^T(0)\delta\mathbf{x}(0) - \int_0^{t_1} \dot{\boldsymbol{\lambda}}^T \delta\mathbf{x}$$

Therefore

$$\delta\tilde{J} = \phi(\mathbf{x}(t_1) + \delta\mathbf{x}(t_1)) - \phi(\mathbf{x}(t_1)) - \boldsymbol{\lambda}^T(t_1)\delta\mathbf{x}(t_1) + \boldsymbol{\lambda}^T(0)\delta\mathbf{x}(0)$$
$$+ \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x} + \delta\mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) + \dot{\boldsymbol{\lambda}}^T \delta\mathbf{x}\big] dt$$

Now approximate the Hamiltonian expressions within the integral to first-order using Taylor expansion.

$$
\begin{aligned}
&= \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x} + \delta\mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt \\
&= \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x} + \delta\mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) + H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt \\
&\approx \int_0^{t_1} \big[H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v})\delta\mathbf{x} + H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt \\
&\approx \int_0^{t_1} \big[H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\delta\mathbf{x} + (H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}))\delta\mathbf{x} \\
&\quad + H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt \\
&\approx \int_0^{t_1} \big[H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\delta\mathbf{x} + H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt
\end{aligned}
$$

Substituting this into the expression for $\delta\tilde{J}$ and using a differential approximation to the first two terms in this expression yields

$$
\begin{aligned}
\delta\tilde{J} = &\big[\phi_{\mathbf{x}}(\mathbf{x}(t_1)) - \boldsymbol{\lambda}^T(t_1)\big]\delta\mathbf{x}(t_1) + \boldsymbol{\lambda}^T(0)\delta\mathbf{x}(0) + \int_0^{t_1} \big[H_{\mathbf{x}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) + \dot{\boldsymbol{\lambda}}^T\big]\delta\mathbf{x} dt \\
&+ \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big] dt + O(\epsilon^2)
\end{aligned}
$$

Now simplify the above by selecting $\boldsymbol{\lambda}(t)$ properly. Note here that $\delta\mathbf{x}(0) = \mathbf{0}$ since the control function change cannot alter the initial state. Choose $\boldsymbol{\lambda}(t)$ as the solution to the adjoint differential equation.

$$-\dot{\boldsymbol{\lambda}}(t)^T = H_{\mathbf{x}}(\boldsymbol{\lambda}(t), \mathbf{x}(t), \mathbf{u}(t))$$

Or, more explicitly

$$-\dot{\boldsymbol{\lambda}}(t)^T = \boldsymbol{\lambda}(t)^T \mathbf{f}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t)) + l_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t))$$

with final condition

$$\boldsymbol{\lambda}(t_1)^T = \phi_{\mathbf{x}}(\mathbf{x}(t_1))$$

$\mathbf{f}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t))$ is a $n \times n$ matrix which is time-varying. However, it is known for nominal $\mathbf{x}, \mathbf{u}$. $\mathbf{l}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t))$ is a known, time-varying, n-dimensional row vector. Therefore the system above is a linear, time-varying, differential equation in the unknown row vector $\boldsymbol{\lambda}(t)^T$, with a final condition on $\boldsymbol{\lambda}(t_1)^T$. Thus consider solving the adjoint equation by moving backward in time from t_1 to 0. This determines a unique solution $\boldsymbol{\lambda}(t)$. With this new $\boldsymbol{\lambda}(t)$, $\delta\tilde{J}$ becomes

$$\delta\tilde{J} = \int_0^{t_1} \big[H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) - H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u})\big]\, dt + O(\epsilon^2)$$

If the original control function $\mathbf{u}$ is optimal, then for all t,

$$H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{v}) \leq H(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) \qquad \forall \mathbf{v} \in U$$

Theorem A.0.1 (Maximum Principle)**.** *Suppose* $\mathbf{u}(t) \in U$ *is the optimal control and* $\mathbf{x}(t)$ *is the state trajectory for the optimal control problem. Then there is an adjoint trajectory* $\boldsymbol{\lambda}(t)$ *such that* $\mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}$ *satisfy*

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \\
\mathbf{x}(0) &= \mathbf{x}_0 \\
-\dot{\boldsymbol{\lambda}}(t)^T &= \boldsymbol{\lambda}(t)^T \mathbf{f}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t)) + \mathbf{l}_{\mathbf{x}}(\mathbf{x}(t), \mathbf{u}(t)) \\
\boldsymbol{\lambda}(t_1)^T &= \boldsymbol{\phi}_{\mathbf{x}}(\mathbf{x}(t_1))
\end{aligned}$$

$$\forall\ 0 \leq t \leq t_1$$

$$H_{\mathbf{u}}(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{u}) = 0$$

References

[1] D. Luenberger. *Introduction to Dynamic Systems - Theory, Models and Applications.* John Wiley & Sons, New York, 1979.

[2] T. Duriez, S. Brunton and B. Noack. *Machine Learning Control - Taming Nonlinear Dynamics and Turbulence.* Springer International Publishing, Switzerland, 2017.

[3] Aristotle. *Aristotle's Politics.* Clarendon Press, Oxford, 1905.

[4] J. Antsaklis and Z. Gao. Control Systems. *The Electronics Engineers' Handbook, 5th Edition*, pages 19.1–19.30, 2005.

[5] M. Meyer. *Matrix Analysis and Applied Linear Algebra.* SIAM: Society for Industrial and Applied Mathematics, Philadelphia, 2001.

[6] R. Brockett and H. Li. A Light Weight Rotary Double Pendulum: Maximizing the Domain of Attraction. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, volume 42, pages 3299–3304, Hawaii, 2003.

[7] K. Furata, T. Okutani and H. Sone. Computer Control of a Double Inverted Pendulum. *Computer and Electrical Engineering*, 5:67–84, 1978.

[8] R. Brockett and H. Li. A Light Weight Rotary Double Pendulum: Maximizing the Domain of Attraction. In *Proceedings of the 2001 IEEE International Conference on Control Applications*, pages 896–901, Mexico City, Mexico, 2001.

[9] A. Bogdanov. *Optimal Control of a Double Inverted Pendulum on a Cart*. Technical report, OGI School of Science and Engineering, OHSU, 2004.

[10] M. Hazewinkel. *Lagrange equations (in mechanics), Encyclopedia of Mathematics.* Springer, New York, 2001.

[11] R. Shankar. *Fundamentals of Physics, Mechanics, Relativity and Thermodynamics*. Yale University Press, New Haven, 2014.

[12] Mathworks. Solve Non-stiff Differential Equations - MATLAB ode45 [online].

[13] J. Dormand and P. Prince. A Family of Embedded Runge-Kutta Formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980.

[14] L. Shampine and M. Reichelt. The MATLAB ODE Suite. *SIAM Journal of Scientific Computing*, 18(1):1–22, 1997.

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

[16] E. Routh and A. Fuller. *A Treatise on the Stability of a Given State of Motion*. University of Cambridge, Cambridge, 1877.

[17] R. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82:35–45, 1960.

[18] B. Datta. *Numerical Methods for Linear Control Systems - Design and Analysis*. Elsevier Academic Press, San Diego, 2004.

[19] Mathworks. Continuous-time Algebraic Riccati equation solution - MATLAB care [online].

[20] N. Shah and M. Yeolekar. On a Matrix Riccati Equation of Stochastic control. *Global Journal of Researches in Engineering Mechanical and MAchanics Engineering*, 13(2):17–23, 2013.

[21] R. Sargent. Optimal Control. *Journal of Computational and Applied Mathematics*, 124(1):361–371, 2000.

[22] R. Gamkrelidze L. Pontryagin, V. Boltyanskii and E. Mischenko. *The Mathematical Theory of Optimal Processes*. Wiley, New York, 1962.

[23] T. Nguyen and Z. Gajic. Solving the Matrix Differential Riccati Equation: A Lyapunov Equation Approach. *IEEE Transactions on Automatic Control*, 55(1):191–194, 2010.

[24] W. Arnold and A. Laub. Generalized Eigenproblem Algorithms and Software for Algebraic Riccati Equations. *Proceedings of the IEEE*, 72(12):1746–1754, 1984.

[25] A. Laub. A Schur Method for Solving the Algebraic Riccati Equations. Technical report, Massachusetts Institute of Technology, 1978.

[26] J. Rodriguez-Canabal. *The Geometry of the Riccati Equation.* PhD thesis, University of Southern California, 1972.

[27] W. Wonham. On a Matrix Riccati Equation of Stochastic control. *SIAM Journal of Control*, 6(4):681–697, 1968.

[28] P. Lancaster and L. Rodman. *Algebraic Riccati Equations.* Clarendon Press, Oxford, United Kingdom, 1995.

[29] Mathworks. Pole Placement Design - MATLAB place [online].

[30] K. Meyer, G. Hall and D. Offin. *Introduction to Hamiltonian Dynamical Systems and the N-body Problem.* Springer, New York, 2009.

[31] F. Murnaghan and A. Winter. A Canonical Form for Real Matrices under Orthogonal Transformations. In JSTOR, editor, *Proceedings of the National Academy of Sciences of the United States of America*, volume 17, pages 417–420, 1931.

[32] F. Gantmacher. *The Theory of Matrices.* Chelsea Publishing Company, New York, 1959.

[33] J. Sun and A. Motter. Controllability Transition and Nonlocality in Network Control. *American Physical Society*, 110, 2013.