ABSTRACT

KING, ETHAN ANDREW. Nonlinear Feedback Controllers and State Estimators: Theory, Applications, and Real-Time Implementation. (Under the direction of Hien Tran and Tien Khai Nguyen.)

This work presents research within three topics: state estimation, feedback control, and mathematical biology.

Proximal Point Moving Horizon Estimation

This work explores use of the proximity operator for constructing moving horizon estimators, which successively fit model trajectories to recent system outputs in order to construct state estimates of dynamical systems. In the presence of both modeling and measurement noise a general convergence result for state estimates using a proximity operator is given for nonlinear systems. Stronger convergence results are shown for linear systems using both least squares and modified least squares fitting functionals. Use of linearization with proximal point moving horizon estimation for nonlinear systems is explored and shown to compare well to the extended Kalman filter on a numerical example. The approach is also found to perform well compared to a low pass derivative filter for supplying state estimates for online stabilization of a double inverted pendulum on a moving cart, in laboratory experiments.

Relaxed Projection Feedback Control

For affine input stabilizing feedback control of nonlinear systems, a family of feedback controls is proposed. The controls are parameterized by a symmetric positive definite (SPD) matrix P, and for discrete dynamics they can be understood as a projection with respect to the P norm. If the projection control is stabilizing for a system, then it is shown that a relaxation for appropriate choices of a parameter γ is also stable. Therefore if a stabilizing P can be identified, weights γ can tune the relaxed projection control for a particular implementation. An analogous control is also developed for continuous nonlinear systems. To construct controls a control synthesis methodology is proposed using an ensemble Kalman search procedure to find stabilizing P over selected subsets of the SPD cone. On numerical examples, the control is shown to perform well in comparison to LQR control for both linear and nonlinear dynamics. The control is also shown to perform well for online stabilization of a double inverted pendulum on a cart in laboratory experiments.

An Optimal Innate Immune Response At The Onset Of Infection

Optimal control of viral infection in a domain of host cells is explored as a means of comparison to experimental observations of the dynamics of the immune system. The immune response and control action studied can induce an antiviral state in cells which protects them from the infecting virus. Optimality of a response is approached in a framework common

to the study of vaccination, using the measurement R^*, the expected number of infection progeny of an infected cell in a fixed population under the intervention regime, to quantify the intervention efficacy. This work defines a protection control that achieves a target value for R^* while protecting the least number of cells for the least amount of time, as optimal. It is shown that a cell autonomous response where protection is initiated when viral density is above a threshold in the neighborhood of the cell, can coordinate optimal regions of cell protection.

Nonlinear Feedback Controllers and State Estimators: Theory, Applications, and Real-Time Implementation

by
Ethan Andrew King

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Applied Mathematics

Raleigh, North Carolina

2020

APPROVED BY:

Kazufumi Ito	Kevin Flores
Hien Tran Co-chair of Advisory Committee	Tien Khai Nguyen Co-chair of Advisory Committee

BIOGRAPHY

Ethan King was born and raised in Salt Lake City Utah. He initially studied biology as an undergraduate before developing an interest in mathematics. In 2015 he graduated from the University of Utah with a Bachelors of Science in both Mathematics and Biology, as well as a minor in Philosophy. He then went on to North Carolina State University for his graduate studies in Applied Mathematics.

ACKNOWLEDGEMENTS

I would like to acknowledge first and foremost my advisors Hien Tran and Khai Nguyen for keeping me pointed in me in good directions and supporting me to pursue them. For their time and input, I would like to acknowledge my committee members Kazufumi Ito and Kevin Flores. I would also like to extend my thanks back to the individuals who first engaged me in research Ann-Marie Torregrossa, Denise Dearing, and Fred Adler.

Thank you to my family who have supported me through thick and thin, for ready advice and constant encouragement.

My graduate research was supported by the Center for Research in Scientific Computation which provided me with the time and equipment necessary to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

PROXIMAL POINT MOVING HORIZON ESTIMATION

Given a system of interest, a model of the dynamics can help inform control and management decisions. Models often summarize the state of the system and dynamics in terms of a finite number of variable quantities. In many cases only a subset of the quantities needed to specify the system state can be measured directly. Model based state estimation approaches use system outputs and the model dynamics to recover the unmeasured system states.

This chapter presents methods for state estimation of discrete time dynamical systems. State estimates are constructed by fitting the model dynamics to the most recent system measurements in a moving horizon approach. Within the framework of the proximal point minimization algorithm a family of moving horizon state estimators are given, and in the presence of both modeling and measurement noise a general convergence result for state estimates of nonlinear systems is established. Stronger results are then given for estimators utilizing least squares functionals on linear dynamics. The use of linearization in conjunction with proximal point moving horizon estimation on nonlinear systems is explored on a numerical example.

1.1 State Estimation

This chapter considers dynamical systems in the discrete time setting. Available system measurements for state estimation are generally discrete and many continuous time model dynamics can be treated in a discrete manner, allowing the discrete setting a broad applicability. The state estimation problem will be introduced for linear systems first.

Suppose that the states $\{x_k\}_{k\in\mathbb{N}} \in \mathbb{R}^d$ at times k of a system satisfy the linear time invariant dynamics

$$\begin{aligned} x_{k+1} &= Ax_k \\ y_{k+1} &= Cx_{k+1} \end{aligned}$$

with matrices $A \in \mathbb{R}^{d\times d}$ and $C \in \mathbb{R}^{m\times d}$. Given the past measurements from the system $\{y_i\}_{i=0}^k$ up to the current time k the state estimation problem is to find the current state of the system x_k. Note that after k time steps the problem can be formulated as solving the linear system

$$\begin{bmatrix} C \\ CA \\ \dots \\ CA^k \end{bmatrix} x = \begin{bmatrix} y_k \\ y_{k-1} \\ \dots \\ y_0 \end{bmatrix} . \tag{1.1}$$

In particular, if $k \cdot m \geq d$ and the matrix $\begin{bmatrix} C & CA & \dots & CA^k \end{bmatrix}^T$ has full column rank then the true state of the system can be uniquely specified by the solution of (1.1). In this case the system is called observable [9]. Similar observability criteria for guaranteeing that a finite number of measurements can be used to determine the system state can be developed for nonlinear continuous and discrete time systems also [20, 43]. Direct solution of (1.1) will yield the exact system state and is an example of a deadbeat observer. Deadbeat observers yielding the exact system state can also be constructed in continuous time settings [19,57]. While methods for exact construction of the system state seem ideal, note that in general for applications with both model error and measurement noise, the linear system (1.1) using all of the available measurements, will be over-determined and have no solution. State estimation methods are needed that can produce good approximations and that are robust to noise, possible perturbations of the system, and erroneous measurements.

Let $\{\eta_k\}_{k\in\mathbb{N}}$ and $\{\epsilon_k\}_{k\in\mathbb{N}}$ be measurement and model noise respectively, such that the system dynamics are given by

$$\begin{aligned} x_{k+1} &= Ax_k + \eta_k \\ y_{k+1} &= Cx_{k+1} + \epsilon_k \, . \end{aligned} \tag{1.2}$$

Suppose further that the noise terms are known to be normally distributed with mean zero and

covariances given by the positive definite matrices $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{m\times m}$, that is

$$\eta_k \sim \mathcal{N}(0, Q) \qquad \epsilon_k \sim \mathcal{N}(0, R)\,.$$

A natural idea for producing an estimate of the state at a time k is to find a weighted fit of the model to the available measurements $\{y_i\}_{i=1}^k$. For example a fit weighted by the covariance of the noise terms. In particular, for $x \in \mathbb{R}^d$ let

$$||x||^2_{Q^{-1}} = x^T Q^{-1} x$$

the norm associated with the positive definite matrix Q^{-1}, and consider the state estimate given by a solution to the problem

$$\begin{aligned} &\min_{\{\hat{x}_i,\eta_i\}_{i=0}^k} \left\{ \sum_{i=1}^k ||C\hat{x}_i - y_i||^2_{Q^{-1}} + \sum_{i=1}^k ||\eta_i||^2_{R^{-1}} \right\} \\ &\text{subject to } \hat{x}_{i+1} = A\hat{x}_i + \eta_i \text{ for all } i \in \{1, 2 \dots k-1\}\,. \end{aligned} \tag{1.3}$$

An immediate issue with this approach is that the size of the problem will grow with time and can quickly become unmanageable. The Kalman filter as introduced in [49] develops a recursion which updates the current state estimate using only the most recent measurement such that the state estimate remains optimal with respect to (1.3).

The Kalman filter can be broken up into two steps, a prediction step and an update step. The prediction step uses the past state estimate and model dynamics to predict the state at time k according to

$$\begin{aligned} \hat{x}_k^- &= A\hat{x}_{k-1} \\ P_k^- &= AP_{k-1}A^T + Q\,. \end{aligned}$$

The estimate is then updated using the current system measurement as follows

$$\begin{aligned} K_k &= P_k^- C^T (CP_k^- C^T + R)^{-1} \\ \hat{x}_k &= \hat{x}_k^- + K_k(z_k - C\hat{x}_k^-) \\ P_k &= (I - K_k C)P_k^-\,. \end{aligned}$$

The matrix P_k in the recursion is a positive definite matrix which can be interpreted as an estimate for the covariance of the state estimation error $(\hat{x}_k - x_k)$ [17]. The Kalman filter is often derived in a probabilistic setting using Baye's rule, though as is shown in [82] a least squares formulation, as is presented here, provides an equivalent state estimation procedure.

The state estimate $\hat{x}_k$ computed by the Kalman filter is equivalent to the estimate generated by solving the minimization problem (1.3) but avoids the need to successively solve a growing

optimization problem at each time point. The assumption of linear dynamics are critical to formulating the Kalman filter recursion, and in general, for systems governed by nonlinear dynamics, such a recursion cannot be constructed.

State estimation for nonlinear systems has been approached by using approximations to compute Kalman filter updates. The extended Kalman filter (EKF) uses a linearization of the system dynamics about the most recent state estimate to compute the Kalman update recursion. The EKF has been widely used for many applications, some industry examples are reviewed in [29]. The unscented Kalman filter (UKF) and the ensemble Kalman filter (EnKF) use an evaluation of the dynamics at a set of points to produce a state estimate. The UKF as presented in [48], can be understood to be performing a linear regression at a set of regression points to estimate the system dynamics [53]. On the other hand, the EnKF as introduced in [32] evaluates the dynamics at an ensemble of inputs to directly estimate the error covariance and error output cross covariance to compute a Kalman filter update, and can be particularly useful for large dimensional systems.

Rather than seeking to approximate a Kalman filter type recursion, moving horizon estimation (MHE) looks to solve a truncated minimization problem, for example of the form (1.3), to directly fit the model to the recent measurement outputs up to a horizon time, an approach which can directly incorporate nonlinear dynamics. This chapter studies a family of moving horizon state estimators and introduces the moving horizon framework in the next section.

1.2 Moving Horizon State Estimation

Moving horizon estimation (MHE) constructs state estimates for a system by fitting model dynamics to the most recent measurement outputs. Only the most recent measurements are used often out to a fixed time called the horizon, it is sometimes described as moving a sliding window within which the model is fit to the data. The approach is analogous to that of model predictive control for computing system inputs. The functional used to fit the model to the data is often composed of two terms, a term which explicitly penalizes the model fit to the most recent measurements, and a term which implicitly fits the model to the excluded past data.

One of the fist moving horizon type approaches was developed by Jazwinski in [47]. Jazwinski proposed resetting the Kalman filter after every N time steps to address an issue with the Kalman filter when used for long periods of time, for example in applications to space vehicle trajectories, where the Kalman filter can become insensitive to new measurements and diverge. The first moving horizon approach for estimation of nonlinear systems was proposed by Michalska and Mayne in [58], and similar approaches have since been developed by multiple authors.

Consider a nonlinear discrete time system with dynamics at each time step k given by an $f_k : \mathbb{R}^d \to \mathbb{R}^d$ and measurements of the system given by a $g_k : \mathbb{R}^d \to \mathbb{R}^m$, such that the system is

modeled by the recursion

$$\begin{aligned} x_{k+1} &= f_k(x_k) + \eta_k \\ y_{k+1} &= g_{k+1}(x_{k+1}) + \epsilon_k \,, \end{aligned} \tag{1.4}$$

where $\{\eta_k\}_{k\in\mathbb{N}}$ and $\{\epsilon_k\}_{k\in\mathbb{N}}$ are model and measurement noise with distributions

$$\eta_k \sim \mathcal{N}(0, Q) \qquad \epsilon_k \sim \mathcal{N}(0, R)\,,$$

for $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{m\times m}$ positive definite matrices defining the covariances. A common approach for moving horizon state estimation as in [69] constructs state estimates as solutions to a problem of the form

$$\begin{aligned} &\min_{\{z_i,\eta_i\}_{i=k-N+1}^{k}} \left\{ \Theta_{k-N+1}(z_{k-N+1}) \; + \sum_{i=k-N+1}^{k} ||g(z_i) - y_i||^2_{Q^{-1}} \; + \sum_{i=k-N+1}^{k} ||\eta_i||^2_{R^{-1}} \right\} \\ &\text{subject to } z_{i+1} = f(z_i) + \eta_i \text{ for all } i \in \{k-N+1, k-N+2, \dots k-1\}\,. \end{aligned} \tag{1.5}$$

The cost functional to be minimized fits the model directly to the most recent N measurements and the functional Θ summarizes a fit to the previous data. For some approaches, as in [69], Θ is called the arrival cost and defined such that

$$\begin{aligned} \Theta_{k-N+1}(z) = \min_{\{z_i,\eta_i\}_{i=0}^{k-N}} &\sum_{i=1}^{k-N} ||g(z_i) - y_i||^2_{Q^{-1}} + \sum_{i=1}^{k-N} ||\eta_i||^2_{R^{-1}}\,, \\ &\text{subject to } z_{k-N+1} = z \text{ and } z_{i+1} = f(z_i) + \eta_i \text{ for all } i \in \{1, 2 \dots k-N\}\,. \end{aligned} \tag{1.6}$$

If the arrival cost (1.6) is used, then minimizing (1.5) is equivalent to fitting the model to the full output history. For example if the dynamics are linear, then positive definite matrices $P_{k-N+1} \in \mathbb{R}^{d\times d}$ can be found such that

$$\Theta_{k-N+1}(z) = ||z - \hat{x}_{k-N+1}||^2_{P_{k-N+1}} \tag{1.7}$$

where $\hat{x}_{k-N+1}$ is given by the previous state trajectory estimate. In this case, the MHE estimates are equivalent to using the Kalman filter estimates and in fact by choosing the horizon size to be $N = 1$, the Kalman filter can be recovered [69]. In general, for nonlinear systems and constrained linear systems a simple form for the arrival cost cannot be derived. Methods have been explored for approximating the arrival cost in these cases, for example, those in [69, 88].

Summarizing the model fit to the past data has also been approached from other perspectives, including a probabilistic perspective as in [22, 87]. Another approach, considers a term which is interpreted as a confidence in the previous estimate, which penalises the norm difference between the new and previous estimate at the horizon of the estimation window [4]. Whatever

the method, the inclusion of a summary of the past data with an implicit penalty term has been found to be important to ensure convergence and stability properties of MHE algorithms as shown in [6, 61, 88].

Beyond just the available measurement data for a system, other information may also be available that can be incorporated as constraints on state variables. For example, it may be impossible for a measurement to be negative. A major advantage of MHE methods is that explicit minimization at each iteration makes the inclusion of additional inequality constraints on state variables and measurement errors straightforward, and MHE has been found to perform well on constrained systems. In a review of methods for incorporating constraints into Kalman filtering approaches [78], Simon found that in a comparison of Kalman methods to an MHE method, the MHE had lower estimation error, though at increased computational cost.

MHE methods have been found to perform well in comparison to extended Kalman filters (EKF) for state estimation of nonlinear systems in some cases as well. Haseltine and Rawlings found a MHE to converge faster than an EKF from poor initial estimates [38], a result also observed by others [5, 51, 88]. Alessandri et al. found also for their MHE implementation that when subject to noise, the root mean square error of the MHE trajectory was smaller than for an EKF [5]. Similar results have been observed in online applications; Shen et al. found faster convergence of a MHE than EKF from poor initial estimates in charge estimation of batteries [77], as did Abdollahpouri et al. for state estimation of a vibrating active cantilever [2]. Further, Abpdollahpouri et al. observed robust estimation by an MHE in the presence of disturbance and measurement noise [2].

One of the biggest disadvantages of MHE methods is the computational cost of solving a minimization problem at every iteration. For state estimation of a large scale water treatment plant, Busch et al. found an EKF and MHE to perform similarly but the MHE was much more computationally expensive [30]. Methods to reduce the computational burden of MHE focus on approximation of the minimization at each MHE iteration, for example with Newton methods [3, 51, 59, 92], conjugate gradient [3] and gradient methods [3, 60]. Another potential issue for MHE methods for nonlinear dynamics is that the minimization problem may not be well posed, and minimization algorithms may converge to only local minimums resulting in poor state estimates.

This chapter presents a family of MHE methods constructed within the framework of the proximal point minimization algorithm. The proximal point minimization algorithm as in [71] naturally gives rise to a quadratic regulating term similar to the implicit penalty terms shown to be effective for MHE in practice. Proximal operator methods have been effective in signal processing, with operator splitting methods allowing for applications to a wide range of cost functionals and state constraints [12, 21, 25]. While splitting methods may allow for construction of MHE algorithms which can incorporate additional state constraints, this chapter focuses on

the unconstrained case. The use of linearization with proximal point MHE for nonlinear state estimation is explored as a means for real time computation, and implementation of this method for online stabilization control of a double inverted pendulum on a cart is presented in Chapter 3.

1.3 Proximal Point Moving Horizon Estimation

A family of moving horizon estimation methods will be constructed with functionals fitting a model trajectory to recent measurement data, in the class $\Gamma_0(\mathbb{R}^d)$, the functionals on $\mathbb{R}^d$ which are convex, proper, and lower semi-continuous. The minimization at each iteration of MHE, will be approached as an application of the proximity operator of the trajectory fitting functional to the previous state estimate.

Definition 1.3.1. *Let the function $\phi : \mathbb{R}^d \to [-\infty, \infty]$ and $\gamma \in]0, \infty[$. The proximity operator of $\gamma\phi$ is defined by*

$$\textit{Prox}_{\gamma\phi} : \mathbb{R}^d \to \mathbb{R}^d : x \to \textit{argmin}_{z\in\mathbb{R}^d}\ \phi(z) + \frac{1}{2}\frac{1}{\gamma}||z - x||^2 .$$

The weight γ will allow for shifting the balance of the proximity operator between minimizing the fitting functional ϕ and minimizing the distance to a given prior state estimate.

For time steps $k \in \mathbb{N}$, let $f_k : \mathbb{R}^d \to \mathbb{R}^d$ be the state transition maps and $g_k : \mathbb{R}^d \to \mathbb{R}^m$ be the state to output maps for the discrete dynamical system

$$\begin{aligned} x_{k+1} &= f_k(x_k) + \eta_k \\ y_{k+1} &= g_{k+1}(x_{k+1}) + \epsilon_k \,, \end{aligned} \tag{1.8}$$

where $\{\eta_k\}_{k\in\mathbb{N}}$ and $\{\epsilon_k\}_{k\in\mathbb{N}}$ are sequences of unknown model and measurement noise, respectively, and will be treated in a deterministic setting. The system is further taken to satisfy the assumptions **(A)**:

(A1) The $\{f_k\}_{k\in\mathbb{N}}$ are Lipschitz continuous functions with a K_{min} and K_{max} in $]0, \infty[$ such that the Lipschitz constants $\{K_k\}_{k\in\mathbb{N}}$ satisfy $K_{min} \leq K_k \leq K_{max}$ for all $k \in \mathbb{N}$.

(A2) The noise terms are bounded, with $\bar{\eta}$ and $\bar{\epsilon}$ in $]0, \infty[$, such that for all $k \in \mathbb{N}$, $||\eta_k|| \leq \bar{\eta}$ and $||\epsilon_k|| \leq \bar{\epsilon}$.

Suppose that no information is available about the distribution of the noise terms for system (1.8), then a simple least squares fit may be a good approach. Consider a model fit to only the

two most recent measurement outputs $\{y_k, y_{k+1}\}$. For example the fitting functional defined by

$$\phi_k(z) \doteq ||g(z) - y_k||^2 + ||g \circ f(z) - y_{k+1}||^2 . \tag{1.9}$$

With a previous estimate for the state at time k given by $\hat{x}_k \in \mathbb{R}^d$, let

$$\begin{aligned} p_k &= Prox_{\gamma\phi_k}(\hat{x}_k) \\ &= \text{argmin}_{z\in\mathbb{R}^d} \left\{ ||g(z) - y_k||^2 + ||g \circ f(z) - y_{k+1}||^2 + \frac{1}{2\gamma}||z - \hat{x}_k|| \right\} . \end{aligned}$$

The output of the proximal operator p_k provides an MHE type estimate for the state at time k using the two most recent measurements for the system, and is indirectly fit to the past data by the term $\frac{1}{2\gamma}||z - \hat{x}_k||$. The estimate p_k can be propagated forward in time by the model dynamics to provide an estimate for the state at time $k + 1$ or a prediction of the state at future times. Iterative application of the proximity operator of a model fitting functional provides a framework for constructing MHE methods. However, note that in general, for ϕ_k of the form in (1.9), the proximal operator is not well defined.

For a given cost functional, iterative application of the proximity operator can generate a minimizing sequence and is called the proximal point minimization algorithm. In particular, if $\phi \in \Gamma_0(\mathbb{R}^d)$ and argmin $\phi \neq \{\emptyset\}$ then for all $\gamma \in]0, \infty[$, Prox$_{\gamma\phi}$ is well defined. Moreover, for any initial value $z_0 \in \mathbb{R}^d$ with sequence $\{\gamma_k\}_{k\in\mathbb{N}}$ in $]0, \infty[$ such that $\sum_{k\in\mathbb{N}} \gamma_k = \infty$ the proximal point iteration

$$z_{k+1} = \text{Prox}_{\gamma_k\phi} z_k$$

generates a minimizing sequence of ϕ [12].

Let $\{x_k\}_{k\in\mathbb{N}}$, a given state trajectory for system (1.4). Discrete time state estimators can be designed by constructing functionals $\{\phi_k\}_{k\in\mathbb{N}}$ such that $(x_k + \zeta_k) \in \text{argmin}_{z\in\mathbb{R}^d}\phi_k(z)$, for some unknown discrepancy term ζ_k, dependent on the system noise and available model outputs. Further, let the sequence of functionals $\{\phi_k\}_{k\in\mathbb{N}}$ satisfy the assumptions **(B)**:

(B1) $\phi_k \in \Gamma_0(\mathbb{R}^d)$ for all $k \in \mathbb{N}$

(B2) $\min_{z\in\mathbb{R}^n} \phi_k(z) = \phi_k(x_k + \zeta_k) = 0$

(B3) There exists $\bar{\zeta} > 0$ such that for all $k \in \mathbb{N}$, $||\zeta_k|| \leq \bar{\zeta}$

It may be assumed without loss of generality that $\min_{z\in\mathbb{R}^n} \phi_k(z) = 0$ since only the minimizer is of interest.

Given an initial state estimate $\hat{x}_0 \in \mathbb{R}^n$ and sequence of weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$, the corresponding proximal point MHE observer is then defined as the state estimates $\{\hat{x}_k\}_{k\in\mathbb{N}}$ and

$\{p_k\}_{k\in\mathbb{N}}$ constructed according to the recursion

$$\begin{aligned} p_k &= \text{Prox}_{\gamma_k\phi_k}\hat{x}_k \\ \hat{x}_{k+1} &= f_k(p_k)\,. \end{aligned} \tag{1.10}$$

Note that at each iteration p_k gives a new estimate for the state at time k which in general is not the current system time, p_k may need to be propagated with the system model several steps forward in time to provide an estimate of the current system state. For simplicity only the updated estimate $\hat{x}_{k+1}$, which will serve as the prior state estimate for the next MHE iteration, will be kept track of here.

Given fitting functionals $\{\phi_k\}_{k\in\mathbb{N}}$ constructed for system (1.8) satisfying the assumptions **(B)**, the following convergence result for the proximal point MHE (1.10) can be established.

Theorem 1.3.2. *If sequence* $\{\gamma_k\}_{k\in\mathbb{N}}$ *in* $]0,\infty[$ *satisfies,*$\gamma_{k+1} \geq \max\left\{K_k^2\gamma_k, \frac{\beta_{k+1}\gamma_k K_k^2}{K_{k+1}}\right\}$*, and* $\sum_{k=1}^{\infty}\frac{1}{\gamma_k} < \infty$*, then*

$$\phi_k(p_k) \to 0\,,$$

where $\beta_k = 2K_k||\zeta_k|| + 2||\eta_k + \zeta_{k+1}|| + 1$ *for all* $k \in \mathbb{N}$.

Theorem 1.3.2 can be shown by a proof similar to that in [12] for the convergence of the proximal point minimization algorithm. Before giving the proof it is useful to establish a relation for the proximity operator from [12].

Proposition 1.3.3. *Let* $\phi \in \Gamma_0(\mathbb{R}^d)$*,* $x \in \mathbb{R}^d$ *and* $\gamma > 0$*. If* $p = Prox_{\gamma\phi}(x)$ *then for any* $y \in \mathbb{R}^d$

$$||y-p||^2 \leq ||x-y||^2 - 2\gamma(\phi(p)-\phi(y))\,.$$

Proof. Let $\phi \in \Gamma_0(\mathbb{R}^d)$, the subgradient of ϕ at $x \in \mathbb{R}^d$ is defined by

$$\partial\phi : \mathbb{R}^d \to 2^{\mathbb{R}^d} : x \to \{u \;:\; (\forall y \in \mathbb{R}^d)\;\; \langle y-x|u\rangle + \phi(x) \leq \phi(y)\}\,.$$

Using Fermat's rule, for any $x \in \mathbb{R}^d$, $p \in \mathbb{R}^d$, $\gamma \in]0,\infty]$

$$p = \text{Prox}_{\gamma\phi}(x) \;\text{ iff }\; (x-p) \in \partial\gamma\phi(p)\,, \tag{1.11}$$

which follows since $p = \text{argmin}_{y\in\mathbb{R}^d}\ \gamma\phi(y) + \frac{1}{2}||y-x||^2$ iff $0 \in \partial\gamma\phi(p) + (p-x)$ therefore iff $(x-p) \in \partial\gamma\phi(p)$.

Let $x \in \mathbb{R}^d$, and let $p = \text{Prox}_{\gamma\phi}(x)$, then from (1.11) for any $y \in \mathbb{R}^d$

$$\langle y-p|x-p\rangle + \gamma\phi(p) \leq \gamma\phi(y)$$

therefore also

$$\langle y-p|x-p\rangle \leq -\gamma(\phi(p)-\phi(y))\,. \tag{1.12}$$

Moreover, note that

$$||y-p||^2 = ||y-x+x-p||^2 = ||y-x||^2 + ||x-p||^2 + 2\langle y-x|x-p\rangle$$

and

$$2\langle y-x|x-p\rangle = 2\langle p-x \ + \ y-p|x-p\rangle = -2||x-p||^2 + 2\langle y-p|x-p\rangle\,.$$

Taken together,

$$\begin{aligned}||y-p||^2 &= ||y-x||^2 - ||x-p||^2 + 2\langle y-p|x-p\rangle\\ ||y-p||^2 &\leq ||y-x||^2 + 2\langle y-p|x-p\rangle\,,\end{aligned}$$

which in conjunction with (1.12) yields the result

$$||y-p||^2 \leq ||y-x||^2 - 2\gamma(\ \phi(p)-\phi(y)\ )\,.$$

□

Theorem 1.3.2, can then be established using Proposition 1.3.3.

Proof. From the definition of the proximal point MHE (1.10), for each $k \in \mathbb{N}$ the state estimates satisfy

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1}+\zeta_{k+1})||^2 =& ||f_k(p_k) - f_k(x_k+\zeta_k) + f_k(x_k+\zeta_k) - (f_k(x_k)+\eta_k) - \zeta_{k+1}||^2\\ \leq& (||f_k(p_k) - f_k(x_k+\zeta_k)|| + ||f_k(x_k+\zeta_k) - f_k(x_k)|| + ||\eta_k+\zeta_k||)^2\,.\end{aligned}$$

Using the Lipschitz continuity of f_k and expanding gives

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1}+\zeta_{k+1})||^2 \leq& K_k^2||p_k-(x_k+\zeta_k)||^2 + 2K_k||p_k-(x_k+\zeta_k)||(\ K_k||\zeta_k|| + ||\eta_k+\zeta_{k+1}||\ ) +\\ & K_k^2||\zeta_k||^2 + ||\eta_k+\zeta_{k+1}||^2 + 2K_k||\zeta_k||\ ||\eta_k+\zeta_{k+1}||\,.\end{aligned} \tag{1.13}$$

Using Proposition 1.3.3 with $y = x_k+\zeta_k$ and the assumption $\phi_k(x_k+\zeta_k)=0$,

$$||p_k-(x_k+\zeta_k)||^2 \leq ||\hat{x}_k-(x_k+\zeta_k)||^2 - 2\gamma_k\phi_k(p_k)\,,$$

and since $\phi_k \geq 0$ we obtain

$$||p_k-(x_k+\zeta_k)|| \leq ||\hat{x}_k-(x_k+\zeta_k)||\,.$$

Using the two relations above with (1.13) gives,

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})||^2 \leq \\ K_k^2||\hat{x}_k - (x_k + \zeta_k)||^2 - 2\gamma_k K_k^2 \phi_k(p_k) + 2K_k||\hat{x}_k - (x_k + \zeta_k)||(K_k||\zeta_k|| + ||\eta_k + \zeta_{k+1}||)+ \\ K_k^2||\zeta_k||^2 + ||\eta_k + \zeta_{k+1}||^2 + 2K_k||\zeta_k||\, ||\eta_k + \zeta_{k+1}||\,.\end{aligned} \tag{1.14}$$

Note also that

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})|| =& ||f_k(p_k) - f_k(x_k + \zeta_k) + f_k(x_k + \zeta_k) - (f_k(x_k) + \eta_k) - \zeta_{k+1}|| \\ \leq& K_k||p_k - (x_k + \zeta_k)|| + K_k||\zeta_k|| + ||\eta_k + \zeta_{k+1}||\,,\end{aligned}$$

then

$$||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})|| \leq K_k||\hat{x}_k - (x_k + \zeta_k)|| + K_k||\zeta_k|| + ||\eta_k + \zeta_{k+1}||\,. \tag{1.15}$$

Adding (1.14) and (1.15) yields

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})||^2 + ||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})|| \leq \\ K_k^2||\hat{x}_k - (x_k + \zeta_k)||^2 - 2\gamma_k K_k^2 \phi_k(p_k) + K_k||\hat{x}_k - (x_k + \zeta_k)||(2K_k||\zeta_k|| + 2||\eta_k + \zeta_{k+1}|| + 1)+ \\ K_k^2||\zeta_k||^2 + ||\eta_k + \zeta_{k+1}||^2 + 2K_k||\zeta_k||\, ||\eta_k + \zeta_{k+1}|| + K_k||\zeta_k|| + ||\eta_k + \zeta_{k+1}||\,.\end{aligned} \tag{1.16}$$

Let

$$\alpha_k = K_k^2||\zeta_k||^2 + ||\eta_k + \zeta_{k+1}||^2 + 2K_k||\zeta_k||\, ||\eta_k + \zeta_{k+1}|| + K_k||\zeta_k|| + ||\eta_k + \zeta_{k+1}||$$

and

$$\beta_k = 2K_k||\zeta_k|| + 2||\eta_k + \zeta_{k+1}|| + 1\,.$$

Then (1.16) becomes

$$\begin{aligned}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})||^2 + ||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})|| \leq& K_k^2||\hat{x}_k - (x_k + \zeta_k)||^2 - \\ & 2\gamma_k K_k^2 \phi_k(p_k) + K_k\beta_k||\hat{x}_k - (x_k + \zeta_k)|| + \alpha_k\,,\end{aligned}$$

and rearranging gives

$$\begin{aligned}2\phi_k(p_k) \leq& \frac{1}{\gamma_k}||\hat{x}_k - (x_k + \zeta_k)||^2 - \frac{1}{K_k^2\gamma_k}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})||^2 + \\ & \frac{\beta_k}{\gamma_k K_k}||\hat{x}_k - (x_k + \zeta_k)|| - \frac{1}{K_k^2\gamma_k}||\hat{x}_{k+1} - (x_{k+1} + \zeta_{k+1})|| + \frac{\alpha_k}{\gamma_k K_k}\,.\end{aligned}$$

Therefore for any $M \in \mathbb{N}$

$$2\sum_{k=0}^{M}\phi_k(p_k) \leq \frac{1}{\gamma_0}||\hat{x}_0-(x_0+\zeta_0)||^2+\sum_{k=1}^{M}\left[(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}K_{k-1}^2})||\hat{x}_k-(x_k+\zeta_k)||^2\right]-\frac{1}{K_M^2\gamma_M}||\hat{x}_{M+1}-(x_{M+1}+\zeta_{M+1})||^2+$$
$$\frac{\beta_0}{\gamma_0 K_0}||\hat{x}_0-(x_0+\zeta_0)||+\sum_{k=1}^{M}\left[(\frac{\beta_k}{\gamma_k K_k}-\frac{1}{\gamma_{k-1}K_{k-1}^2})||\hat{x}_k-(x_k+\zeta_k)||\right]-\frac{1}{K_M^2\gamma_M}||\hat{x}_{M+1}-(x_{M+1}+\zeta_{M+1})||+$$
$$\sum_{k=0}^{M}\frac{\alpha_k}{\gamma_k K_k}\,,$$

and since $\gamma_{k+1} \geq \max\left\{K_k^2\gamma_k, \frac{\beta_{k+1}\gamma_k K_k^2}{K_{k+1}}\right\}$ for all $k \in \mathbb{N}$, the summation terms on the right hand side are negative, hence

$$2\sum_{k=0}^{M}\phi_k(p_k) \leq \frac{1}{\gamma_0}||\hat{x}_0-(x_0+\zeta_0)||^2+\frac{\beta_0}{\gamma_0 K_0}||\hat{x}_0-(x_0+\zeta_0)||+\sum_{k=0}^{M}\frac{\alpha_k}{\gamma_k K_k}\,.$$

Moreover $\sum_{k=1}^{\infty}\frac{1}{\gamma_k} < \infty$ and the noise and Lipschitz constants are bounded, whence

$$\sum_{k=0}^{\infty}\phi_k(p_k) < \infty\,,$$

therefore also

$$\phi_k(p_k) \to 0\,.$$

□

Theorem 1.3.2 states that for any fitting functionals ϕ_k which can be constructed satisfying the assumptions **(B)** a sequence of weighting parameters can be chosen such that the proximal point state estimates (1.10) will converge with respect to the fitting functionals and so have the potential to be used to construct a system state observer. The ϕ_k though, must be constructed such that $\phi_k(p_k) \to 0$ implies also that $||\hat{x}_k - x_k||$ will be small in the limit in order for the state estimates to be accurate. Note also that, in general, the weighting parameters must grow exponentially to ensure convergence.

For nonlinear systems, constructing ϕ_k which satisfy the assumptions of Theorem 1.3.2 is difficult. Construction of state estimators for nonlinear systems is approached here by using linearization of the dynamics. In the next sections, proximal point MHE methods for linear dynamics are presented for which stronger convergence results under weaker conditions on the weighting parameters can be shown.

1.4 Least Squares Proximal Point Moving Horizon Estimation for Linear Systems

Consider a linear system of the form (1.4) under the assumptions **(A)**, with matrices $\{\Phi_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{d\times d}$ and $\{C_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{m\times d}$ such that

$$\begin{aligned} x_{k+1} &= \Phi_k x_k + \eta_k \\ y_{k+1} &= C_{k+1}x_{k+1} + \epsilon_{k+1}\,. \end{aligned} \tag{1.17}$$

State estimates for the system can be constructed using a least squares fit to the measurement outputs. For example, let $\{y_{k+i}\}_{i=0}^{N-1}$ be N recent measurement outputs and consider a least squares measure of a model fit to the outputs of the form

$$\begin{aligned} &\sum_{i=0}^{N-1} ||\ y_{k+i} - C_{k+i}z_{k+i}||^2\,, \\ &\text{where} \quad z_{k+i+1} \ = \ \Phi_k z_{k+i} \quad \forall i \in \{0,1,..N-1\}\,. \end{aligned} \tag{1.18}$$

The fit to the model can be written more succinctly with a matrix G_k giving the model to output map, and vector v_k incorporating the model and measurement noise. In particular, let

$$G_k = \begin{bmatrix} C_{k+1}\Phi_k \\ \cdots \\ C_{k+N}\left(\prod_{i=0}^{N-1}\Phi_{k+i}\right) \end{bmatrix}, \qquad v_k = \begin{bmatrix} C_{k+1}\eta_k + \epsilon_k \\ \cdots \\ C_{k+N}\sum_{j=0}^{N-2}(\prod_{i=j}^{N-1}\Phi_{k+i})\eta_{k+j} + C_{k+N}\eta_{(k+(N-1))} + \epsilon_{k+N} \end{bmatrix}.$$

Then if $\{x_k\}_{k\in\mathbb{N}}$ is the true state trajectory satisfying the linear system (1.17), the least squares fit (1.18) can be written equivalently as

$$\frac{1}{2}||G_k(x_k - z) + v_k||^2\,. \tag{1.19}$$

This section will develop results for proximal point MHE using least squares fitting functionals of this general form.

For functionals of the form (1.19), let the matrices $\{G_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{\ell\times d}$ and vectors $\{v_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{\ell}$ satisfy the assumptions **(C):**

(C1) $G_k^TG_k$ is positive definite for all $k \in \mathbb{N}$.

(C2) For $\lambda_{k_{min}}$ the smallest eigenvlaue of $G_k^TG_k$, there exists a $\bar{\lambda}_{min}$ such that for all $(k \in \mathbb{N})$, $\lambda_{k_{min}} \geq \bar{\lambda}_{min}$.

(C3) For a $\bar{v}$ in $]0,\infty[$, $||v_k|| \leq \bar{v}$ for all $(k \in \mathbb{N})$.

Then for every k, the functional (1.19) has a unique minimizer z_k^*, that is given by

$$z_k^* = x_k + (G^T G)^{-1} G^T v_k \,,$$

with the minimum value equal to $\frac{1}{2}||(I - G_k(G_k^T G_k)^{-1} G_k^T) v_k||^2$. For convenience and to fit into the framework of Theorem 1.3.2 least squares fitting functionals for proximal point state estimation of (1.17) are defined as follows

$$\phi_k(z) = \frac{1}{2}(\, ||G_k(x_k - z) + v_k||^2 - ||(I - G_k(G_k^T G_k)^{-1} G_k^T) v_k||^2 \,) \,,$$

which may be written with the discrepancy term $\zeta_k = (G^T G)^{-1} G^T v_k$ more conveniently as

$$\phi_k(z) = \frac{1}{2}||G_k((x_k + \zeta_k) - z)||^2 \,. \tag{1.20}$$

Note that the assumption **(C1)** is equivalent to the statement that for the number of observations used in a cost functional of the form (1.18) the system is observable.

From an initial estimate $\hat{x}_0$, let the sequences $\{\hat{x}_k\}_{k\in\mathbb{N}}$ and $\{p_k\}_{k\in\mathbb{N}}$ be generated according to the proximal point MHE (1.10) using the cost functionals (1.20), with sequence of weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ in $]0, \infty[$, then the following proposition holds.

Proposition 1.4.1. *If $\{\gamma_k\}_{k\in\mathbb{N}}$ are chosen to satisfy the conditions of Theorem 1.3.2 then*

$$||p_k - (x_k + \zeta_k)|| \to 0 \,.$$

Proof. Each ϕ_k is in $\Gamma_0(\mathbb{R}^d)$ and by assumption **(C3)** the terms $\{\zeta_k\}_{k\in\mathbb{N}}$ are bounded, therefore the assumptions of Theorem 1.3.2 are satisfied. Then, if weighting terms $\{\gamma_k\}$ are chosen to satisfy Theorem 1.3.2, the proximal point observer (1.10) generates a sequence $\{p_k\}_{k\in\mathbb{N}}$ such that $\phi_k(p_k) \to 0$. Moreover, using assumption **(C2)**, $\phi_k(p_k) \geq \bar{\lambda}_{min} \frac{1}{2}||p_k - (x_k + \zeta_k)||^2$ whence the result follows $||p_k - (x_k + \zeta_k)|| \to 0$. □

Weaker conditions on the weighting terms than those given by Theorem 1.3.2 can also be found to guarantee stability of the state estimates. For each $k \in \mathbb{N}$, let $U_k \in \mathbb{R}^{d\times d}$ a unitary matrix, and $\Lambda_k \in \mathbb{R}^{d\times d}$ diagonal such that $G_k^T G_k = U_k \Lambda_k U_k^T$. The eigenvalues of $G_k^T G_k$ will be denoted by $\{\lambda_{k_i}\}_{i=1}^d$.

Proposition 1.4.2. *The error terms $e_k = (\hat{x}_k - x_k)$ satisfy the recursion*

$$e_{k+1} = \Phi_k U_k \bar{\Lambda}_k U_k^T e_k \;+\; \Phi_k U_k \ddot{\Lambda}_k U_k^T \zeta_k \;-\; \eta_k \,,$$

where the matrices $\bar{\Lambda}_k$ and $\ddot{\Lambda}_k$ are diagonal and have the entries

$$(\bar{\Lambda}_k)_{i,i} = \frac{1}{1+\gamma_k\lambda_{k_i}}\,, \qquad \textit{and} \qquad (\ddot{\Lambda}_k)_{i,i} = \frac{\gamma_k\lambda_{k_i}}{1+\gamma_k\lambda_{k_i}}$$

respectively.

Proof. Note from (1.10)

$$p_k = \text{Prox}_{\gamma_k\phi_k}(\hat{x}_k) = argmin_{z\in\mathbb{R}^d}\left\{\frac{1}{2}||G_k(z-(x_k+\zeta_k))||^2 + \frac{1}{2\gamma_k}||z-\hat{x}_k||^2\right\}.$$

Then computing the gradient and setting it equal to zero yields

$$p_k = (G_k^TG_k + \frac{1}{\gamma_k}I)^{-1}G_k^TG_k(x_k+\zeta_k) \;+\; \frac{1}{\gamma_k}(G_k^TG_k+\frac{1}{\gamma_k}I)^{-1}\hat{x}_k\,.$$

Using the fact $G_k^TG_k = U_k\Lambda_kU_k^T$,

$$p_k = U_k(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_kU_k^Tx_k \;+\; \frac{1}{\gamma_k}U_k(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}U_k^T\hat{x}_k \;+\; U_k(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_kU_k^T\zeta_k\,.$$

Therefore,

$$(p_k-x_k) = U_k((\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_k - I)U_k^Tx_k \;+\; \frac{1}{\gamma_k}U_k(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}U_k^T\hat{x}_k \;+\; U_k(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_kU_k^T\zeta_k\,.$$

Note that

$$(\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_k - I = -\bar{\Lambda}_k, \qquad \frac{1}{\gamma_k}(\Lambda_k+\frac{1}{\gamma_k}I)^{-1} = \bar{\Lambda}_k, \qquad (\Lambda_k+\frac{1}{\gamma_k}I)^{-1}\Lambda_k = \ddot{\Lambda}_k$$

then

$$(p_k-x_k) = U_k\bar{\Lambda}_kU_k^Te_k + U_k\ddot{\Lambda}_kU_k^T\zeta_k\,.$$

Therefore,

$$e_{k+1} = (\hat{x}_{k+1}-x_{k+1}) = \Phi_k(p_k-x_k) - \eta_k = \Phi_kU_k\bar{\Lambda}_kU_k^Te_k + \Phi_kU_k\ddot{\Lambda}_kU_k^T\zeta_k - \eta_k\,.$$

□

The error recursion of proposition 1.4.2, can be used to choose weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ to ensure the error is small relative to the noise. In particular,

Proposition 1.4.3. *Let $\gamma \in \mathbb{R}$ such that $\gamma > max\left\{\frac{K_{max}-1}{\bar{\lambda}_{min}}, 0\right\}$, if $\gamma_k > \gamma$ for all $k \in \mathbb{N}$ then*

$$\lim_{k\to\infty} ||e_k|| \leq \frac{\kappa}{1-r}\,,$$

where $\kappa = K_{max}\bar{\zeta} + \bar{\eta}$ and $r = \dfrac{K_{max}}{1+\gamma\bar{\lambda}_{min}}$.

Proof. Using Proposition 1.4.2 and assumption **(B2)** for all $k \in \mathbb{N}$

$$||e_{k+1}|| \leq ||\Phi_k||\frac{1}{1+\gamma\bar{\lambda}_{min}}||e_k|| \;+\; ||\Phi_k||\frac{\gamma\lambda_{k_{max}}}{1+\gamma\lambda_{k_{max}}}||\zeta_k|| \;+\; ||\eta_k||\,, \tag{1.21}$$

and with assumption **(A1)**

$$||e_{k+1}|| \leq K_{max}\frac{1}{1+\gamma\bar{\lambda}_{min}}||e_k|| \;+\; K_{max}||\zeta_k|| \;+\; ||\eta_k||\,.$$

Then $||e_{k+1}|| \leq r||e_k|| + \kappa$. Therefore, $||e_k|| \leq ||e_0||r^k + \kappa\sum_{m=0}^{k} r^m$, and since $r < 1$

$$\lim_{k\to\infty} ||e_k|| \leq \frac{\kappa}{1-r}\,.$$

□

Note from the error recursion of Proposition 1.4.2 and corresponding inequality relation (1.21), if the error is small and expected to not grow under the dynamics, then it is advantageous to choose weighting terms γ_k small, such that the contribution of the noise at each iteration is small. Correspondingly, if the error is relatively large compared to the noise terms then large γ_k can be used to reduce the error, though such a choice will also increase the potential contribution of the noise. For good performance then, the choice of weighting parameters must balance reduction in error with introduction of noise. In the next section, a modification of the least squares cost functionals is considered as a means to allow reduction of the noise contributions.

1.5 Subspace Solutions for Least Squares Proximal Point Moving Horizon Estimation

Let $\{x_k\}_{k\in\mathbb{N}}$ a state trajectory for the linear system (1.17) and let the matrices $\{G_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{\ell\times d}$ and vectors $\{v_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{\ell}$ satisfy the assumptions **(C)**. For symmetric positive semidefinite matrices (PSD) $\{P_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}^{\ell\times\ell}$, consider fitting functionals to construct proximal point observers of the form

$$\hat{\psi}_k(z) = (G_k(z-x) - v_k)^T P(G_k(z-x) - v_k)\,.$$

Note that $\hat{\psi}_k$ is in $\Gamma_0(\mathbb{R}^d)$ and therefore can be used to construct proximal point MHE observers satisfying the conditions of theorem 1.3.2. In particular, we consider P_k constructed to weight only the component of the state in the span of the eigenvectors corresponding to the largest eigenvalues of $G_k^T G_k$.

Let the diagonal matrix $\Lambda_k \in \mathbb{R}^{d\times d}$ and orthonormal matrix $U_k \in \mathbb{R}^d$ denote a singular value decomposition of $G_k^T G_k$ and let $\{\lambda_{k_i}\}_{i=1}^d$ denote the eigenvalues of $G_k^T G_k$ ordered such that $\lambda_{k_1} \geq \lambda_{k_2} \geq \ldots \lambda_{k_d} > 0$. For a $p \in 1, 2 \ldots d$ let $\Sigma_k \in \mathbb{R}^{d\times d}$ be the diagonal matrix with $\left\{\frac{\lambda_{k_1}}{\lambda_{k_1}^2}, \ldots \frac{\lambda_{k_p}}{\lambda_{k_p}^2}, 0 \ldots 0\right\}$ on the diagonal and set $V_k \in \mathbb{R}^{\ell\times\ell}$ the matrix with columns an orthornormal basis of the form $\{G_k u_{k_1}, \ldots G_k u_{k_d}\} \bigcup \{v_{k_j}\}_{j=d+1}^{\ell}$, where u_{k_i} is the i^{th} column of the orthonormal matrix U_k.

Let P_k be the PSD matrix that satisfies the singular value decomposition

$$P_k = \begin{bmatrix} G_k U_k & V_k \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_k^T G_k^T \\ V_k^T \end{bmatrix},$$

where V_k is the matrix with columns $\{v_{k_j}\}_{j=d+1}^{\ell}$. Then

$$\begin{aligned}
\min_{z\in\mathbb{R}^d} \hat{\psi}_k(z) =& \frac{1}{2} \min_{z\in\mathbb{R}^d} \left\{(G_k(x_k - z) - v_k)^T P_k (G_k(x_k - z) - v_k)\right\} \\
=& \frac{1}{2} \min_{z\in\mathbb{R}^d} \left\{(x_k - z)^T G_k^T G_k U_k \Sigma_k U_k^T G_k^T G_k (x_k - z) - \right. \\
& \left. 2(x_k - z)^T G_k^T G_k U_k \Sigma_k U_k^T G_k^T v_k + v_k^T P_k v_k\right\} \\
=& \frac{1}{2} \min_{z\in\mathbb{R}^d} \left\{(x_k - z)^T U_k \tilde{\Lambda}_k U_k^T (x_k - z) - 2(x_k - z)^T \tilde{I} G_k^T v_k + v_k^T P v_k\right\},
\end{aligned}$$

where $\tilde{\Lambda}_k \in \mathbb{R}^{d\times d}$ is the diagonal matrix with $\left\{\lambda_{k_1}, \ldots \lambda_{k_p}, 0 \ldots 0\right\}$ on the diagonal, and $\tilde{I} \in \mathbb{R}^{d\times d}$ is block diagonal with the $(p \times p)$ identity in the upper block and zeros elsewhere. Therefore for $z_k^* = x_k + (G_k^T G_k)^{-1} G_k^T v_k$,

$$z_k^* \in \text{argmin}_{z\in\mathbb{R}^d} \hat{\psi}_k(z)$$

Let $\zeta_k = (G_k^T G_k)^{-1} G_k^T v_k$. Fitting functionals to construct proximal point state estimates are defined as follows

$$\psi_k(z) = \frac{1}{2}(G_k(x_k - z) + v_k)^T P_k (G_k(x_k - z) + v_k) - \frac{1}{2}(v_k - G_k\zeta_k)^T P_k (v_k - G_k\zeta_k),$$

which may be written in the equivalent and more convenient form,

$$\psi_k(z) = \frac{1}{2}(G_k(x_k + \zeta_k) - z))^T P_k (G_k((x_k + \zeta_k) - z)). \tag{1.22}$$

1.5.1 Computation of the proximal point update

Given a $\hat{x}_k \in \mathbb{R}^d$ and $\gamma_k \in]0, \infty[$ the proximal operator of ψ_k can be computed as follows

$$\begin{aligned} p_k = \text{Prox}_{\psi_k}(\hat{x}_k) &= \arg\min_{z\in\mathbb{R}^d} \left\{ (G(z-(x_k+\zeta_k)))^T P_k (G(z-(x_k+\zeta_k))) + \frac{1}{2}\frac{1}{\gamma}(z-\hat{x}_k)^T(z-\hat{x}_k) \right\} \\ &= \arg\min_{z\in\mathbb{R}^d} \left\{ (z-(x_k+\zeta_k))^T U \tilde{\Lambda}_k U^T (z-(x_k+\zeta_k)) + \frac{1}{2}\frac{1}{\gamma}(z-\hat{x}_k)^T(z-\hat{x}_k) \right\} . \end{aligned}$$

Let $z', x'_k, \zeta'_k, \hat{x}'_k$ be vectors in $\mathbb{R}^d$ such that

$$z = U_k z' , \qquad x_k = U_k x'_k , \qquad \zeta_k = U_k \zeta'_k , \qquad \hat{x}_k = U_k \hat{x}'_k .$$

Then the following minimization problem is equivalent

$$\arg\min_{z'\in\mathbb{R}^d} \left\{ (z'-(x'_k+\zeta'_k))^T \tilde{\Lambda} (z'-(x'_k+\zeta'_k)) + \frac{1}{2}\frac{1}{\gamma}(z'-\hat{x}'_k)^T(z'-\hat{x}'_k) \right\} , \tag{1.23}$$

and it can be decomposed into two independent minimization problems.

For $x \in \mathbb{R}^d$ let $(x)_p \in \mathbb{R}^p$ and $(x)_{p^c} \in \mathbb{R}^{d-p}$ be such that

$$x = \begin{bmatrix} (x)_p \\ (x)_{p^c} \end{bmatrix} .$$

Denote by $\tilde{\Lambda}_{k_p}$ the diagonal matrix in $\mathbb{R}^{p\times p}$ with diagonal entries $\{\lambda_{k_1} \dots \lambda_{k_p}\}$ then (1.23) is composed of the two independent minimization problems

$$\begin{aligned} (p'_k)_p = \text{argmin}_{(z')_p\in\mathbb{R}^p} \Big\{ &((z')_p - ((x'_k)_p + (\zeta'_k)_p))^T \tilde{\Lambda}_{k_p} ((z')_p - ((x'_k)_p + (\zeta'_k)_p)) + \\ &\frac{1}{2}\frac{1}{\gamma_k}((z')_p - (\hat{x}'_k)_p)^T((z')_p - (\hat{x}'_k)_p) \Big\} \end{aligned} \tag{1.24}$$

$$(p'_k)_{p^c} = \arg\min_{(z')_{p^c}\in\mathbb{R}^{d-p}} \left\{ \frac{1}{2}\frac{1}{\gamma_k}((z')_{p^c} - (\hat{x}'_k)_{p^c})^T((z')_{p^c} - (\hat{x}'_k)_{p^c}) \right\} . \tag{1.25}$$

By computing the gradient and setting it equal to zero the unique minimizer of (1.24) is given by

$$(p'_k)_p = (\tilde{\Lambda}_{k_p} + \frac{1}{\gamma} I)^{-1} (\tilde{\Lambda}_{k_p}((x'_k)_p + (\zeta'_k)_p) + \frac{1}{\gamma}(\hat{x}'_k)_p) ,$$

and the unique minimizer of (1.25) is trivially

$$(p'_k)_{p^c} = (\hat{x}'_k)_{p^c} .$$

Then, for $U_{k_{p^c}} \in \mathbb{R}^{(d-p)\times(d-p)}$ the matrix with columns $\{u_{k_i}\}_{i=p+1}^{d}$ and $U_{k_p} \in \mathbb{R}^{p\times p}$ the matrix with columns $\{u_{k_i}\}_{i=1}^{p}$, p_k is given by

$$p_k = U_{k_p}(\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1} U_{k_p}^T (U_{k_p}\tilde{\Lambda}_{k_p} U_{k_p}^T (x_k + \zeta_k) + \frac{1}{\gamma_k}\hat{x}_k) + U_{k_{p^c}} U_{k_{p^c}}^T \hat{x}_k \,. \tag{1.26}$$

Alternatively, if $(\tilde{\Lambda}_k + \frac{1}{\gamma}\tilde{I})^\dagger$ denotes the pseudo inverse of $(\tilde{\Lambda}_k + \frac{1}{\gamma}\tilde{I})$ then p_k may be written equivalently as follows

$$p_k = U_k(\tilde{\Lambda}_k + \frac{1}{\gamma}\tilde{I})^\dagger U_k^T (U_k\tilde{\Lambda}_k U_k^T (x_k + \zeta_k) + \frac{1}{\gamma_k}\hat{x}_k) + U_{k_{p^c}} U_{k_{p^c}}^T \hat{x}_k \,.$$

1.5.2 Stability of the state estimates

Following a similar analysis to that given for the least squares cost functionals, weaker conditions on the weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ can be found than those of Theorem 1.3.2 to achieve stability of the state estimates.

From initial estimate $\hat{x}_0$, let the sequences $\{\hat{x}_k\}_{k\in\mathbb{N}}$ and $\{p_k\}_{k\in\mathbb{N}}$ be generated according to (1.10) using the cost functionals (1.22), and the sequence of weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ in $[0,\infty)$.

Proposition 1.5.1. *The error terms $e_k = (\hat{x}_k - x_k)$ satisfy the recursion*

$$e_{k+1} = \Phi_k U_{k_p}\bar{\Lambda}_{k_p} U_{k_p}^T e_k + \Phi_k U_{k_{p^c}} U_{k_{p^c}}^T e_k + \Phi_k U_{k_p}\ddot{\Lambda}_{k_p} U_{k_p}^T \zeta_k - \eta_k \,, \tag{1.27}$$

where the diagonal matrices $\bar{\Lambda}_{k_p}$ and $\ddot{\Lambda}_{k_p}$ in $\mathbb{R}^{p\times p}$ have entries

$$(\bar{\Lambda}_{k_p})_{i,i} = \frac{1}{1+\gamma_k\lambda_{k_i}} \,, \qquad \textit{and} \qquad (\ddot{\Lambda}_{k_p})_{i,i} = \frac{\gamma_k\lambda_{k_i}}{1+\gamma_k\lambda_{k_i}} \,,$$

respectively.

Proof. From (1.26),

$$\begin{aligned}(p_k - x_k) = U_{k_p}((\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1}\tilde{\Lambda}_{k_p} - I_p) U_{k_p}^T x_k + \frac{1}{\gamma_k} U_{k_p}(\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1} U_{k_p}^T \hat{x}_k + \\ U_{k_{p^c}} U_{k_{p^c}}^T (\hat{x}_k - x_k) + U_{k_p}(\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1}\tilde{\Lambda}_{k_p} U_{k_p}^T \zeta_k \,.\end{aligned}$$

Note that

$$(\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1}\tilde{\Lambda}_{k_p} - I_p = -\bar{\Lambda}_{k_p}, \qquad \frac{1}{\gamma_k}(\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1} = \bar{\Lambda}_{k_p}, \qquad (\tilde{\Lambda}_{k_p} + \frac{1}{\gamma_k} I_p)^{-1}\tilde{\Lambda}_{k_p} = \ddot{\Lambda}_{k_p} \,,$$

then

$$(p_k - x_k) = U_{k_p}\bar{\Lambda}_{k_p}U_{k_p}^T e_k + U_{k_{p^c}}U_{k_{p^c}}^T e_k + U_{k_p}\ddot{\Lambda}_{k_p}U_{k_p}^T\zeta_k\,.$$

Therefore,

$$e_{k+1} = (\hat{x}_{k+1} - x_{k+1}) = \Phi_k(p_k - x_k) - \eta_k = \Phi_k U_{k_p}\bar{\Lambda}_{k_p}U_{k_p}^T e_k + \Phi_k U_{k_{p^c}}U_{k_{p^c}}^T e_k + \Phi_k U_{k_p}\ddot{\Lambda}_{k_p}U_{k_p}^T\zeta_k - \eta_k\,.$$

□

The term $\Phi_k U_{k_{p^c}}U_{k_{p^c}}^T e_k$ in the error recursion (1.27) can not be influenced by the choice of weighting parameters. However, if for all $k \in \mathbb{N}$, $||\Phi_k U_{k_{p^c}}U_{k_{p^c}}^T|| < 1$ then the error can be made stable.

Proposition 1.5.2. *Let* $(0 < \alpha < 1)$ *such that for all* $k \in \mathbb{N}$, $||\Phi_k U_{k_{p^c}}U_{k_{p^c}}^T|| < \alpha$ *and let* $\bar{\lambda}_{p_{min}} = \min_{k\in\mathbb{N}} \lambda_{k_p}$ *. If* $\gamma > max\left\{\dfrac{K_{max} - (1-\alpha)}{\bar{\lambda}_{p_{min}}(1-\alpha)}, 0\right\}$*, then*

$$\lim_{k\to\infty} ||e_k|| \leq \frac{\kappa}{1-r}\,,$$

where $\kappa = K_{max}\bar{\zeta} + \bar{\eta}$ *and* $r = \dfrac{K_{max}}{1+\gamma\bar{\lambda}_{p_{min}}} + \alpha$*.*

Proof. Using proposition 1.5.1 for all $k \in \mathbb{N}$

$$||e_{k+1}|| \leq (||\Phi_k||\frac{1}{1+\gamma\bar{\lambda}_{p_{min}}} + \alpha)||e_k|| \;+\; ||\Phi_k||\frac{\gamma\lambda_{k_1}}{1+\gamma\lambda_{k_1}}||\zeta_k|| \;+\; ||\eta_k||\,,$$

and with assumption **(A1)**,

$$||e_{k+1}|| \leq (K_{max}\frac{1}{1+\gamma\bar{\lambda}_{p_{min}}} + \alpha)||e_k|| \;+\; K_{max}\frac{\gamma\lambda_{k_1}}{1+\gamma\lambda_{k_1}}||\zeta_k|| \;+\; ||\eta_k||\,.$$

Therefore, $||e_{k+1}|| \leq r||e_k|| + \kappa$, so also $||e_k|| \leq ||e_0||r^k + \kappa\sum_{m=0}^{k} r^m$ and $r < 1$, hence

$$\lim_{k\to\infty} ||e_k|| \leq \frac{\kappa}{1-r}\,.$$

□

1.6 Numerical Results

In this section, two proximal point MHE methods for state estimation are implemented for the nonlinear Lorentz dynamical system by utilizing linearization about the current state estimate

similar to the implementation of the extended Kalman filter (EKF). The dynamics of the Lorentz system are given by

$$\begin{aligned} \dot{x_L} &= \sigma(y_L - x_L) \\ \dot{y_L} &= x_L(\rho - z_L) - y_L \\ \dot{z_L} &= x_L y_L - \beta z_L \,. \end{aligned} \tag{1.28}$$

The parameters for the system were taken to be $\{\sigma = 10,\ \beta = 8/3,\ \rho = 28\}$ for which the dynamics are well known to be chaotic. From a given initial condition $x_0 = [x_L(0), y_L(0), z_L(0)]^T$, the dynamics were simulated using matlab's ode45 solver. Measurements y_k of the system state were computed every h simulated time units, with

$$y_k = Cx(t_k) + \epsilon_k \,,$$

for $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\epsilon_k \sim \mathcal{N}(0, \sigma_\epsilon I)$. At each measurement time point the state of the system was also perturbed by adding a Gaussian noise term $\eta_k \sim \mathcal{N}(0, \sigma_\eta I)$.

To perform state estimation, at each measurement time the dynamics of the system were linearized around the current state estimate and a forward linear state transition map Φ_k, and backward map Φ_k^{-1}, over time intervals of size h, were approximated with matlab's expm command, in order to construct a linear approximation of the dynamics of the form (1.17).

State estimates were constructed using a centered proximal point MHE (CPX), utilizing the three most recent system outputs with cost functionals $\phi_{cntr_k} : \mathbb{R}^d \to \mathbb{R}$ defined by

$$\phi_{cntr_k}(z) \doteq \frac{1}{2}||C\Phi_k^{-1}z - y_{k-1}||^2 + \frac{1}{2}||Cz - y_k||^2 + \frac{1}{2}||C\Phi_k z - y_{k+1}||^2 \,, \tag{1.29}$$

which are of the form (1.19) for

$$G_k = \begin{bmatrix} C\Phi_k^{-1} \\ C \\ C\Phi_k \end{bmatrix} \quad \text{and} \quad v_k = \begin{bmatrix} -C\Phi_k^{-1}\eta_{k-1} + \epsilon_{k-1} \\ \epsilon_k \\ C\eta_k + \epsilon_{k+1} \end{bmatrix} . \tag{1.30}$$

Note that if A gives the linearized dynamics and the pair (C, A) is observable then $G_k^T G_k$ is positive definite, hence assumption **C1** is satisfied.

At each time step t_{k+1}, with measurement vector $q_k = [y_{k-1}, y_k, y_{k+1}]^T$ and weighting term γ the CPX update (1.10) was computed by

$$\begin{aligned} p_k &= U_k(\Lambda_k + \frac{1}{\gamma}I)^{-1}U_k^T(G_k^T q_k + \frac{1}{\gamma}\hat{x}_k) \\ \hat{x}_{k+1} &= \Phi_k p_k \,, \end{aligned}$$

where U_k and Λ_k were computed with matlab's svd command. The returned state estimate for the system $\tilde{x}$ at each time step was taken to be the model prediction $\tilde{x}_{k+2} = \Phi_k \hat{x}_{k+1}$.

The restricted subspace functionals ψ_{cntr_k} of the form (1.22), constructed with the matrices and vectors (1.30) were also used to compute a proximal point observer (SCPX) using $p = 2$. At each time step t_{k+1}, with measurement vector $q_k = [y_{k-1}, y_k, y_{k+1}]^T$ and weighting term γ the SCPX update (1.10) was computed by

$$\begin{aligned} p_k &= U_k(\tilde{\Lambda}_k + \frac{1}{\gamma}\tilde{I})^{\dagger} U_k^T (G_k^T q_k + \frac{1}{\gamma}\hat{x}_k) + U_{k_{p^c}} U_{k_{p^c}}^T \hat{x}_k \\ \hat{x}_{k+1} &= \Phi_k p_k \,, \end{aligned}$$

where again the returned state estimate at each time step was taken to be the model prediction $\tilde{x}_{k+2} = \Phi_k \hat{x}_{k+1}$.

Data was generated from the Lorentz system using a measurement time step of size $h = .001$ over a time interval of length 20. Model noise was generated with standard deviation $\sigma_\eta = .1$ and measurement noise with standard deviation $\sigma_\epsilon = 5$. State estimates were then computed with a CPX observer using a fixed weighting parameter $\gamma = .5$ and a SCPX observer using fixed $\gamma = 50$. For comparison two EKF estimators were also used. One (EKF) using the the true noise parameters $\sigma_{\text{EKF}_\eta} = \sigma_\eta$ and $\sigma_{\text{EKF}_\epsilon} = \sigma_\epsilon$. For the other (EKF2), the noise parameters were tuned for a better fit of the unmeasured state variable x_L, the best parameters found were $\sigma_{\text{EKF2}_\eta} = .005$ and $\sigma_{\text{EKF2}_\epsilon} = 10$. The average norm error of the full state estimates $\tilde{x}_k$ over the time interval, $e = \frac{1}{N}\sum_{k=1}^{N} ||x_k - \tilde{x}_k||$ and the average error of only the unmeasured state variable $e_{x_L} = \frac{1}{N}\sum_{k=1}^{N} |x_{L_k} - \tilde{x}_{L_k}|$ were computed for each method. The results of 50 trials are plotted in Figure 1.1.

Both proximal point methods had smaller estimation error than the EKF, and the SCPX had the smallest estimation error of the unmeasured state variable x_L. A comparison of the SCPX and EKF fits to the state variable x_L are shown in Figure 1.2, which illustrates that the SCPX was able to more closely recover the extrema of the x_L state variable's oscillations than the other methods.

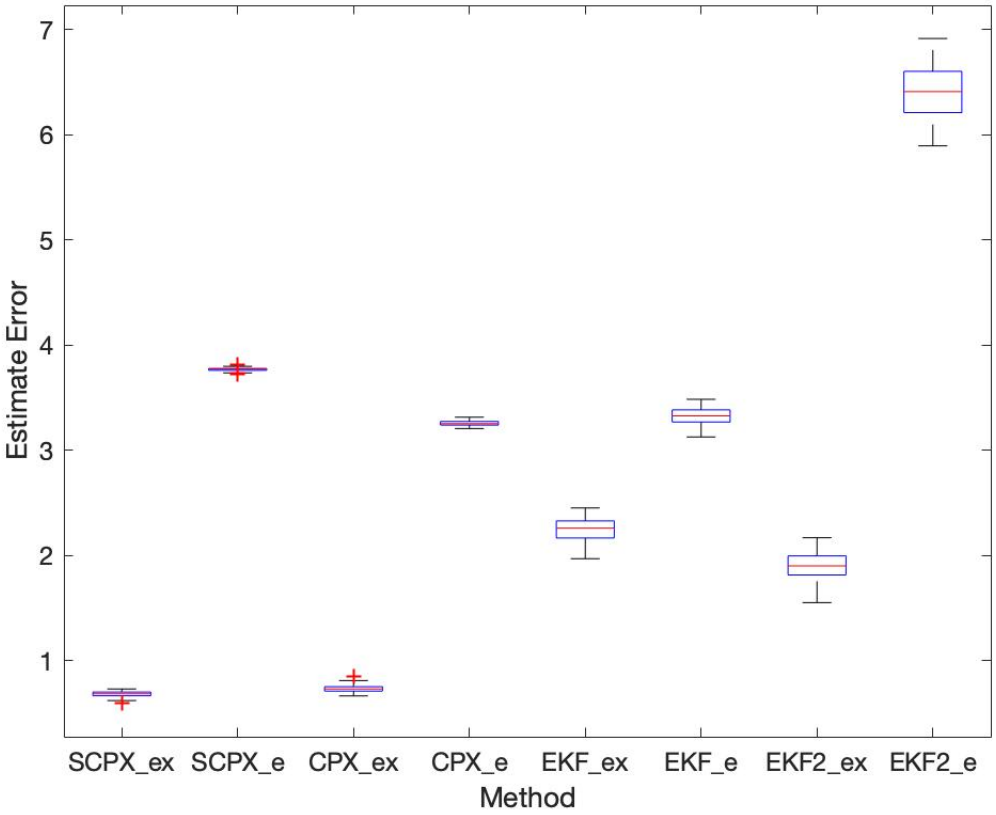

Figure 1.1: Plot of the average norm error ('MTHD'_e) and average x_L estimate error ('MTHD'_ex) for each state estimation method on the Lorentz system over 50 trials.

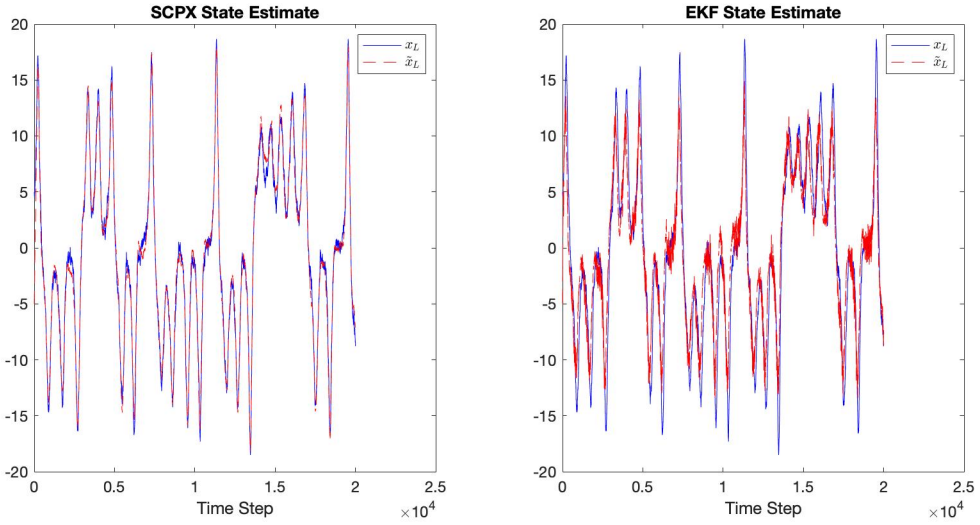

Figure 1.2: Comparison of the SCPX and EKF estimates $\tilde{x}_L$, of the Lorentz state variable x_L.

CHAPTER

2

RELAXED PROJECTION CONTROL

Given knowledge of the dynamics for a system and a mechanism for influencing the system through a set of inputs, a control specifies inputs which achieve a desired system response. In particular, feedback controls give inputs as a function of the current system state. By computing inputs based on the current state of the system feedback controls can be robust to noise and disturbances.

This chapter introduces a method for affine input stabilizing feedback control of nonlinear discrete and continuous time systems. Stabilizing control is a general framework, in which the control objective is to stabilize, drive to the origin, the state of the system. The presented control can be understood as a projection of the system state with respect to a norm parameterized by a positive definite matrix relaxed with a tuning parameter. Convergence results are established for both discrete and continuous nonlinear systems, and stronger results are given for discrete time linear systems. To synthesize controls, a methodology using an ensemble Kalman search procedure is introduced. Both the control design and implementation are shown on numerical examples.

2.1 Optimal Feedback Control

Feedback control will be introduced in the discrete time settings. Methods using optimal control design will be presented first for reference and comparison to the controls which will be

introduced in later sections.

Consider an affine control input system where for $x_k \in \mathbb{R}^d$ the state of the system at time k the system dynamics are given by

$$x_{k+1} = Ax_k + Bu_k \tag{2.1}$$

for matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times m}$. The control input to the system at time k is specified by the vector $u_k \in \mathbb{R}^m$.

This chapter considers stabilizing control problems, where the objective is to identify control inputs $\{u_k\}_{k\in\mathbb{N}}$ such that when applied to the system (2.1) the state of the system is driven to the origin, that is

$$x_k \to 0\,.$$

In particular, feedback control is sought, that is a function $h : \mathbb{R}^d \to \mathbb{R}^m$ such that the control, given at each time k by $u_k = h(x_k)$, is stabilizing for the system.

A common approach for control design is to construct a control which is optimal with respect to a performance measure, where the response of the system under the application of the control is measured by a cost functional which quantifies how well the control achieves a desired response. For stabilizing feedback control a common cost to use is a quadratic of the form

$$\begin{aligned} J(x, \{u_k\}_{k\in\mathbb{N}}) &\doteq \sum_{k=0}^{\infty} x_k^T Q x_k + u_k^T R u_k \\ &\text{with } x_0 = x\,, \text{and } x_{k+1} = Ax_k + Bu_k \text{ for all } k \in \mathbb{N}\,. \end{aligned} \tag{2.2}$$

The matrices $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{m\times m}$ are symmetric positive definite and weight the quadratic penalty for the distance of the state from the origin and the penalty for the magnitude of control effort used at each time step.

If the system (2.1) is such that for all $x \in \mathbb{R}^d$ there exists a control $\{u_k\}_{k\in\mathbb{N}}$ such that

$$J(x, \{u_k\}_{k\in\mathbb{N}}) < \infty$$

then the optimal control which minimizes the cost (2.2) is called the linear quadratic regulator (LQR) and as given in [9] has a feedback form given by

$$u_k = -(R + B^T P B)^{-1} B^T P A x_k\,, \tag{2.3}$$

where the matrix $P \in \mathbb{R}^{d\times d}$ is the unique symmetric positive definite matrix which is a solution

of the algebraic Riccati equation, given by

$$P = A^T(P - PB(R + B^T PB)^{-1}B^T P)A + Q\,.$$

The matrices Q and R can be tuned in the cost (2.2) such that the corresponding LQR feedback control appropriately stabilizes the system while using a feasible amount of control effort. A similar cost and corresponding feedback control can also be formulated for continuous time linear dynamical systems. Controls that are optimal with respect to meaningful cost functions often have other desirable properties such as robustness to noise and the avoidance of unnecessary or counterproductive control effort.

If a systems dynamics are nonlinear, then solving for an optimal control is difficult. For quadratic cost functionals of the form (2.2), no general solution is known for nonlinear systems. Methods for approximating the optimal control have been proposed, for instance by using linearization of the dynamics, several such methods are reviewed and compared in [13]. Given the difficulty in constructing optimal controls for nonlinear systems, more direct methods of control design have also been explored, and will be discussed in the next section.

2.2 Inverse Optimal Control Design

One alternative approach to optimal control design uses control Lyapunov functions (CLF). Design of control using CLFs was first proposed by Artstein in [11] for continuous time systems, and similar strategies can also be applied in the discrete time setting.

For an $f : \mathbb{R}^d \to \mathbb{R}^d$ and a $g : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ with control inputs $u_k \in \mathbb{R}^m$, consider the discrete time system with dynamics

$$x_{k+1} = f(x_k) + g(x_k)u_k\,. \tag{2.4}$$

A positive definite function $V : \mathbb{R}^d \to \mathbb{R}$ is a CLF for system (2.4) if for all $x \in \mathbb{R}^d\backslash\{0\}$ there exists a $u \in \mathbb{R}^m$ such that

$$V(f(x) + g(x)u) - V(x) \;<\; 0\,.$$

That is, for all system states not the origin there is a control input which can reduce the value of the Lyapunov function sometimes referred to as the 'energy' of the system, which implies a control can be constructed which will bring the state of the system to the origin. The definition for a CLF for continuous time dynamics is analogous and requires that, for each state, a control exists such that the gradient of the CLF is negative.

As shown in [33, 80], given an appropriate CLF for a continuous time dynamical system, a stabilizing continuous feedback control can be constructed. Moreover, feedback controls can be designed such that a corresponding cost functional can also be constructed that they are

optimal with respect to [33, 46]. This suggests a control design approach sometimes called inverse optimal control design, in which a control is constructed for a system and then certified as optimal with respect to a constructed cost functional. The approach avoids the need to solve for an optimal control directly, but if a meaningful cost functional can be constructed for which a given control is optimal, then that control may still have beneficial properties of optimality. Finding and verifying a CLF for a nonlinear system is ,in general, still quite difficult. For nonlinear discrete time systems, Ornelas-Tellez et al. propose a method for systematically constructing controls in [62].

For systems of the form (2.4) Ornelas-Tellez et al. propose constructing feedback controls by searching over controls of the form

$$u_k = -(g(x_k)^T P g(x_k) + E)^{-1} g(x_k)^T P f(x_k)\,, \tag{2.5}$$

by searching over the symmetric positive definite matrices (SPD) $P \in \mathbb{R}^{d\times d}$ and $E \in \mathbb{R}^{m\times m}$ which parameterize the control. Controls are defined as stabilizing for a system if the corresponding quadratic functional given by

$$V_P(x) \doteq x^T P x\,. \tag{2.6}$$

is verified as a CLF for the system when using the control [62].

Note that the controls (2.5) have the same structure as the optimal LQR controls (2.3) for linear systems. Further an LQR control strictly decreases the corresponding functional (2.6) which is in fact the value function for the optimal cost from a given state [54]. Ornelas-Tellez propose expanding this linear control structure for nonlinear discrete systems, and they show further that if a P and E exist such that the control (2.5) is stabilizing for the system (2.4) then a cost functional can be constructed such that the control is also optimal with respect to that cost functional [62].

Several methods for searching over the SPD matrices to synthesize a control of the form (2.5) have been explored including, speed gradient descent [62], particle swarm optimization [63, 73], the extended Kalman filter [7], and an ensemble Kalman filter [86].

This chapter presents a control synthesis methodology similar to that given by Ornelas-Tellez et al. A related family of controls is proposed which is also parameterized by the SPD matrices for nonlinear discrete time systems and an analogous control is presented for continuous time nonlinear systems.

To search for an SPD matrix parameterizing a stabilizing control for a given system, an ensemble Kalman procedure is presented. The ensemble Kalman filter was introduced for state estimation in noisy systems by Evensen in [32] and has since been adapted for other inverse problems including parameter estimation [10, 31] and training of neural networks [37, 50]. Use of ensemble statistics and a Kalman filter update allow for derivative free parallelizable

methods, in particular the method proposed here does not require an explicit model for the system dynamics. The control and control synthesis methodology are shown on several numerical examples and an implementation is presented for stabilization of a double inverted pendulum on a cart in Chapter 3.

2.3 Control For Nonlinear Discrete Time Systems

Let the discrete dynamics for the states $x_k \in \mathbb{R}^d$ of a system at times k, be described by an $f : \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ with control inputs $u_k \in \mathbb{R}^m$, according to

$$x_{k+1} = f(x_k) + g(x_k)u_k \,. \tag{2.7}$$

The system is further taken to satisfy the assumption:

(A) *for all $x \in \mathbb{R}^d$ the columns of $g(x)$ are nonzero and linearly independent,*

Note that the assumption **(A)** amounts to requiring that each control has an independent affect on the system dynamics, that is to say, no control inputs are redundant. If that is not the case it may be possible to reduce the system such that it holds.

Let $P \in \mathbb{R}^{d\times d}$ a symmetric positive definite (SPD) matrix. For x and y in $\mathbb{R}^d$, let

$$\langle x \ , \ y\rangle_P \doteq x^T P y \,, \qquad \text{and} \qquad ||x||_P \doteq \sqrt{x^T P x}$$

the P inner product and norm, respectively, on $\mathbb{R}^d$.

The stabilizing controls (2.5) proposed by Ornelas-Tellez et al. can also be understood to be solutions to a minimization problem at each time step. In particular, for E and P symmetric positive definite (SPD) matrices, the control

$$u_k = -(g(x_k)^T P g(x_k) + E)^{-1} g(x_k)^T P f(x_k)$$

is also the solution to the minimization problem

$$u_k = \text{argmin}_{u\in\mathbb{R}^m} \frac{1}{2}||f(x_k) + g(x_k)u||_P^2 + \frac{1}{2}||u||_E^2 \,.$$

2.3.1 Projection Control

This chapter considers a related control for stablization, given by

$$u_k = \text{argmin}_{u\in\mathbb{R}^m} \frac{1}{2}||f(x_k) + g(x_k)u||_P^2 \,. \tag{2.8}$$

The control is well defined for all inputs x_k, and can be written in the explicit feedback form

$$u_k = -(g(x_k)^T P g(x_k))^{-1} g(x_k)^T P f(x_k)\,. \tag{2.9}$$

In particular, $\nabla_u[\frac{1}{2}||f(x_k)+g(x_k)u||_P^2] = 0$ if and only if $g(x_k)^T P g(x_k)u = -g(x_k)^T P f(x_k)$. Further, note that $g(x_k)^T P g(x_k)$ is the Gramian matrix of the columns of $g(x_k)$ with respect to the P inner product, hence by the assumption **(A)** it is positive definite and therefore invertible.

Proposition 2.3.1. *The feedback control* (2.9) *at each time step k is the projection with respect to the P norm of the future state in the absence of control $f(x_k)$, onto the linear subspace*

$$g(x_k)^{P_\perp} \doteq \{y \in \mathbb{R}^d \ : g(x_k)^T P y \ = 0\}\,.$$

Proof. Let $x_k \in \mathbb{R}^d$ and to simplify notation let $G \in \mathbb{R}^{d\times m}$ and $\vec{f} \in \mathbb{R}^d$ be

$$G = g(x_k)\,, \qquad \vec{f} = f(x_k)\,.$$

Then

$$u_k = \text{argmin}_{u\in\mathbb{R}^m} \frac{1}{2}||\vec{f}+Gu||_P^2\,.$$

By the orthogonal decomposition theorem, there exists a $\zeta \in N(G^T P^{\frac{1}{2}}) := \{z \in \mathbb{R}^d \ : \ G^T P^{\frac{1}{2}} z = 0\}$ and $\eta \in R(P^{\frac{1}{2}}G) := \{z \in \mathbb{R}^d \ : \ \exists v \in \mathbb{R}^m \text{ such that } \ P^{\frac{1}{2}}Gv = z\}$ such that $P^{\frac{1}{2}}\vec{f} = \zeta + \eta$. Therefore

$$\begin{aligned}
\frac{1}{2}||\vec{f}+Gu||_P^2 = \frac{1}{2}||P^{\frac{1}{2}}\vec{f}+P^{\frac{1}{2}}Gu||^2 &= \frac{1}{2}\langle \zeta+\eta+P^{\frac{1}{2}}Gu \mid \zeta+\eta+P^{\frac{1}{2}}Gu\rangle \\
&= \frac{1}{2}\langle \zeta \mid \zeta\rangle + \frac{1}{2}\langle \eta+P^{\frac{1}{2}}Gu \mid \eta+P^{\frac{1}{2}}Gu\rangle + \langle \zeta \mid \eta+P^{\frac{1}{2}}Gu\rangle \\
&= \frac{1}{2}||\zeta||^2 + \frac{1}{2}||\eta+P^{\frac{1}{2}}Gu||^2
\end{aligned}$$

hence

$$\min_{u\in\mathbb{R}^m} \frac{1}{2}||\vec{f}+Gu||_P^2 = \frac{1}{2}||\zeta||^2 + \min_{u\in\mathbb{R}^m} \frac{1}{2}||\eta+P^{\frac{1}{2}}Gu||^2$$

but $\eta \in R(P^{\frac{1}{2}}G)$, therefore

$$\min_{u\in\mathbb{R}^m} \frac{1}{2}||\vec{f}+Gu||_P^2 = \frac{1}{2}||\zeta||^2$$

with minimizer u^* the vector such that $P^{\frac{1}{2}}Gu^* = -\eta$. Then by applying the minimizer as the control the next state can be found. Consider that

$$P^{\frac{1}{2}}x_{k+1} = P^{\frac{1}{2}}\vec{f} + P^{\frac{1}{2}}Gu^* = \zeta$$

therefore

$$x_{k+1} = P^{-\frac{1}{2}}\zeta\,.$$

Note that $x_{k+1} \in G^{P_\perp}$ and further it will be shown that x_{k+1} is the projection of $\vec{f}$ onto $G^{P_\perp}$ with respect to the P norm.
Suppose it is not, then there exists $h \in G^{P_\perp}$ such that $||h - x_{k+1}||_P > 0$ and $||h - \vec{f}||_P^2 < ||x_{k+1} - \vec{f}||_P^2$. Therefore,

$$||h - \vec{f}||_P^2 < ||P^{-\frac{1}{2}}\zeta - \vec{f}||_P^2$$

$$||P^{\frac{1}{2}}h - P^{\frac{1}{2}}\vec{f}||^2 < ||\zeta - P^{\frac{1}{2}}\vec{f}||^2$$

that is

$$||P^{\frac{1}{2}}h - \zeta - \eta||^2 < ||\eta||^2$$

in which case

$$||P^{\frac{1}{2}}h - \zeta||^2 + 2\langle P^{\frac{1}{2}}h - \zeta \mid \eta\rangle + ||\eta||^2 < ||\eta||^2\,.$$

Note that $P^{\frac{1}{2}}h \in N(G^T P^{\frac{1}{2}})$ since $h \in G^{P_\perp}$ therefore

$$||h - x_{k+1}||_P^2 + ||\eta||^2 < ||\eta||^2$$

a contradiction, hence the proposition holds. □

The goal is to find a control of the form (2.9), in particular an SPD matrix P such that the control is stabilizing for the system (2.7).

Definition 2.3.2. *P is globally exponentially stabilizing for system* (2.7) *if there exists $0 < \alpha < 1$ such that for all $x_k \in \mathbb{R}^d$ and u_k given by* (2.9) *the system dynamics satisfy*

$$||x_{k+1}||_P^2 - ||x_k||_P^2 < -\alpha||x_k||_P^2\,.$$

While P may be stabilizing, if $f(x_k)$ is far from the subspace $g(x_k)^{P_\perp}$ the control magnitude necessary for the projection may be extremely large. The control must be modulated for practical implementations. A relaxation of the projection is introduced in the next section to allow for tuning of the control.

2.3.2 Relaxed Projection Control

We consider a relaxation to the control (2.8) weighted by positive parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ of the following form

$$u_k = \min_{u\in\mathbb{R}^m} \frac{1}{2}\left(||x_{k+1}||_P^2 + \frac{1}{\gamma_k}||x_{k+1} - x_k||_P^2 \right)\,.$$

The control magnitude is modulated through the implicit penalty $||x_{k+1} - x_k||_P^2$. Under the assumption (A), the control has the explicit feedback form

$$u_k = -(g(x_k)^T P g(x_k))^{-1} g(x_k)^T P (f(x_k) - \frac{1}{\gamma_k + 1} x_k)\,. \tag{2.10}$$

Let $P \in \mathbb{R}^{d\times d}$ a globally exponentially stabilizing matrix for (2.7) with parameter $0 < \alpha < 1$. For a sequence of weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ used with P in the feedback control (2.10) the following proposition holds

Proposition 2.3.3. *If for all $k \in \mathbb{N}$, $\gamma_k > \frac{1-\sqrt{\alpha}}{\sqrt{\alpha}}$ then the feedback control* (2.10) *is globally exponentially stable.*

Proof. We show that there exists a $0 < \zeta < 1$ such that for all $x_k \in \mathbb{R}^d$, $||x_{k+1}||_P^2 - ||x_k||_P^2 < -\zeta||x_k||_P^2$ and therefore the control is globally exponentially stable.

Let $x_k \in \mathbb{R}^d$ and let

$$u_{p_k} = -(g(x_k)^T P g(x_k))^{-1} g(x_k)^T P f(x_k) \qquad \text{and} \qquad u_{r_k} = (g(x_k)^T P g(x_k))^{-1} g(x_k)^T P x_k\,.$$

Then the control is given by

$$u_k = u_{p_k} + \frac{1}{\gamma_k + 1} u_{r_k}\,,$$

and

$$\begin{aligned}
||x_{k+1}||_P^2 =& \left\langle f(x_k) + g(x_k)(u_{p_k} + \frac{1}{\gamma_k+1}u_{r_k})\ ,\ f(x_k) + g(x_k)(u_{p_k} + \frac{1}{\gamma_k+1}u_{r_k}) \right\rangle_P \\
=& \langle f(x_k) + g(x_k)u_{p_k}\ ,\ f(x_k) + g(x_k)u_{p_k}\rangle_P + \\
& 2\frac{1}{\gamma_k+1}\langle f(x_k) + g(x_k)u_{p_k}\ ,\ g(x_k)u_{r_k}\rangle_P + \frac{1}{(\gamma_k+1)^2}\langle g(x_k)u_{r_k}\ ,\ g(x_k)u_{r_k}\rangle_P\,.
\end{aligned}$$

Note that $\langle f(x_k) + g(x_k)u_{p_k}\ ,\ g(x_k)u_{r_k}\rangle_P = 0$ since $(f(x_k) + g(x_k)u_{p_k}) \in g(x_k)^{P_\perp}$. Therefore,

$$||x_{k+1}||_P^2 = ||f(x_k) + g(x_k)u_{p_k}||_P^2 + \frac{1}{(\gamma_k+1)^2}||g(x_k)u_{r_k}||_P^2\,,$$

and

$$||x_{k+1}||_P^2 - ||x_k||_P^2 = ||f(x_k) + g(x_k)u_{p_k}||_P^2 - ||x_k||_P^2 + \frac{1}{(\gamma_k+1)^2}||g(x_k)u_{r_k}||_P^2\,.$$

Moreover $f(x_k) + g(x_k)u_{p_k}$ is the future state using the projection control and P is stabilizing with constant α, hence

$$||x_{k+1}||_P^2 - ||x_k||_P^2 < -\alpha||x_k||_P^2 + \frac{1}{(\gamma_k+1)^2}||g(x_k)u_{r_k}||_P^2\,.$$

Note also that

$$||g(x_k)u_{r_k}||_P^2 \leq ||x_k||_P^2$$

since

$$\begin{aligned} ||x_k||_P^2 &= \langle x_k - g(x_k)u_{r_k}\ ,\ x_k - g(x_k)u_{r_k}\rangle_P + 2\langle x_k - g(x_k)u_{r_k}\ ,\ g(x_k)u_{r_k}\rangle_P + \langle g(x_k)u_{r_k}\ ,\ g(x_k)u_{r_k}\rangle_P \\ &= ||x_k - g(x_k)u_{r_k}||_P^2 + ||g(x_k)u_{r_k}||_P^2 \,. \end{aligned}$$

In particular, $(x_k - g(x_k)u_{r_k})$ is the projection of x_k onto $g(x_k)^{P_\perp}$ with respect to the P norm, therefore $\langle x_k - g(x_k)u_{r_k}\ ,\ g(x_k)u_{r_k}\rangle_P = 0$.

Then for all $x_k \in \mathbb{R}^d$,

$$||x_{k+1}||_P^2 - ||x_k||_P^2 < -\left(\alpha - \frac{1}{(\gamma_k+1)^2}\right)||x_k||_P^2 \,.$$

Let $\gamma_{min} = min_{k\in\mathbb{N}}\gamma_k$ and set $\zeta = \left(\alpha - \frac{1}{(\gamma_{min}+1)^2}\right)$, then $0 < \zeta < 1$ since $\gamma_{min} > \frac{1-\sqrt{\alpha}}{\sqrt{\alpha}}$ which completes the proof.

□

A control design strategy for discrete time systems using this framework is to first identify a stabilizing P for the system using the projection control (2.9), and then to use Proposition 2.3.3 to construct an appropriate choice of tuning parameters for implementation with the relaxed projection control (2.10). This strategy will be shown for some numerical examples in section 2.7. Weaker conditions for Stabilizing P and choice of weighting parameters can be found for discrete time linear systems and are presented in Section 2.5.

2.4 Continuous Time Systems

In many cases the dynamics for a system are described by continuous time differential equations. While true continuous time feedback control cannot be implemented in practice, it can be useful to formulate a control in the continuous time setting as will be done for the application example presented in Chapter 3.

An analogous control to the relaxed projection feedback control (2.9) for discrete time systems can be constructed for continuous time dynamics. Let the continuous dynamics for the state $x \in \mathbb{R}^d$ of a system be given by a $f : \mathbb{R}^d \to \mathbb{R}^d$ and a $g : \mathbb{R}^d \to \mathbb{R}^{d\times m}$, with control input $u \in \mathbb{R}^m$ according to the ordinary differential equation

$$\dot{x} = f(x) + g(x)u \,, \tag{2.11}$$

with g such that the assumption **(A)** is satisfied.

Let $h > 0$ a small time step, consider the feedback control at state $x \in \mathbb{R}^d$ given by

$$u_h(x) \doteq \text{argmin}_{u \in \mathbb{R}^m} \left\{ \frac{1}{2}||x + hf(x) + hg(x)u||_P^2 + \frac{1}{2}\frac{1}{\zeta}||x(t) + hf(x) + hg(x)u - x||_P^2 \right\}. \tag{2.12}$$

Under the assumption **(A)** the control (2.12) has the explicit feedback form

$$u_h(x) = -(g(x)^T P g(x))^{-1}(\frac{\zeta}{h(1+\zeta)} g(x)^T P x + g(x)^T P f(x)) .$$

Note that as the time step h is decreased the control magnitude will increase. Suppose that the weighting term ζ is scaled with h to modulate the control effort, in particular the weighting term value is chosen as the function $\zeta(h) \doteq h\gamma$ for some fixed $\gamma > 0$, then

$$u_h(x) = -(g(x)^T P g(x))^{-1}(\frac{\gamma h}{h(1+\gamma h)} g(x)^T P x + g(x)^T P f(x)) .$$

The continuous time relaxed projection control is defined as

$$u(x) \doteq \lim_{h \to 0} u_h(x) , \tag{2.13}$$

which has the explicit feedback form

$$u(x) = -(g(x)^T P g(x))^{-1}(\gamma g(x)^T P x + g(x)^T P f(x)) . \tag{2.14}$$

The control functions similarly to the discrete control (2.10), the first term moves the current state of the system towards the subspace $g(x)^{P_\perp}$, while the second term counteracts the system dynamics moving away from the subspace. A similar result for tuning the control as found for the discrete time case can also be found in the continuous setting.

Definition 2.4.1. *An SPD matrix P is globally exponentially stabilizing for the system* (2.11) *if there exists a $\zeta > 0$ and $\alpha > 0$, such that for all $x \in \mathbb{R}^d$ using the feedback control* (2.14)*, the system dynamics satisfy*

$$\frac{d}{dt}[\ \frac{1}{2}||x||_P^2\ ] < -\alpha||x||_P^2$$

Let $P \in \mathbb{R}^{d \times d}$ be an SPD matrix globally exponentially stable for system (2.11) with parameters $\zeta > 0$ and $\alpha > 0$. Suppose that the weighting terms for the feedback control (2.14) are given by the function $\gamma : [0, \infty) \to [0, \infty)$.

Proposition 2.4.2. *If $\gamma(t) > \zeta$ for all $t > 0$ then the feedback control* (2.14) *is globally exponentially stable.*

Proof. From initial condition $x_0 \in \mathbb{R}^d$ let $x : [0, \infty) \to \mathbb{R}^d$ the solution of (2.11) using control

(2.14) with weighting parameters $\gamma(t)$. Let $t > 0$,

$$\frac{1}{2}\frac{d}{dt}[x^T P x] = x(t)^T P f(x) - x^T P g(x)(g(x)^T P g(x))^{-1} g(x)^T P f(x) - \gamma(t) x^T P g(x)(g(x)^T P g(x))^{-1} g(x)^T P x \,.$$

Note, $(g(x)^T P g(x))^{-1}$ is positive definite by assumption (A) . Then, since $\gamma(t) > \zeta$,

$$\gamma(t) x^T P g(x)(g(x)^T P g(x))^{-1} g(x)^T P x \geq \zeta x^T P g(x)(g(x)^T P g(x))^{-1} g(x)^T P x \,.$$

Therefore, since P is globally exponentially stable for the parameters (α, ζ)

$$\frac{1}{2}\frac{d}{dt}[x^T P x] < -\alpha ||x||_P^2 \,.$$

Hence the control is globally exponentially stable. □

As in the discrete time setting a stabilizing P can be found for a system and then tuned by choice of weighting parameters to fit to an implementation. A control of this form is designed and implemented for stabilization control of a double inverted pendulum on a cart in Chapter 3.

2.5 Relaxed Projection Control For Discrete Time Linear Systems

Let $(A \in \mathbb{R}^{d \times d})$ and $(B \in \mathbb{R}^{d \times m})$ with controls $u \in \mathbb{R}^m$ describe the discrete dynamics

$$x_{k+1} = A x_k + B u_k \,, \tag{2.15}$$

where it is further assumed that the columns of B are linearly independent.

For a given SPD matrix P, using the discrete time feedback projection control (2.9), leads to the state update recursion

$$x_{k+1} = (I - B(B^T P B)^{-1} B^T P) A x_k \,.$$

Let

$$\mathbb{P}_{B^{P_\perp}} \doteq I - B(B^T P B)^{-1} B^T P$$

denote the P norm projection operator onto $B^{P_\perp}$.

Definition 2.5.1. *Let $P \in \mathbb{R}^{d \times d}$ an SPD matrix, P is stabilizing for (2.15) if with $\{\lambda_i\}_1^d$ the eigenvalues of $(\mathbb{P}_{B^{P_\perp}} A)$, for all $i \in \{1, 2 \ldots, d\}$, $|\lambda_i| < 1$.*

Given a stabilizing P, the discrete time relaxed projection control (2.10), with sequence

$\{\gamma_k\}_{k\in\mathbb{N}}$, will give the state update recursion

$$x_{k+1} = \mathbb{P}_{B^{P_\perp}} A x_k + \frac{1}{1+\gamma_k} B(B^T P B)^{-1} B^T P x_k \,, \tag{2.16}$$

and will converge for only mild assumptions on the sequence $\{\gamma_k\}_{k\in\mathbb{N}}$.

An explicit form for the state at each iteration in terms of the initial condition can be derived for the system when the relaxed projection control (2.16) is used , from which conditions for convergence easily follow.

Let $x_0 \in \mathbb{R}^d$ be the initial condition for the system (2.15), and let $v_0 \in B^{P_\perp}$ and $z_0 \in R(B)$ give the unique P orthogonal representation, $x_0 = v_0 + z_0$. For positive weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$ let the control inputs be given by the feedback control (2.16), then the following proposition holds.

Proposition 2.5.2. *For all $k \geq 1$ the state of the system is given by*

$$x_k = (\mathbb{P}_{B^{P_\perp}} A)^k v_0 + \sum_{m=0}^{k-1} \Big(\prod_{i=0}^{(k-1)-m} \frac{1}{1+\gamma_i} \Big) (\mathbb{P}_{B^{P_\perp}} A)^m z_0 \,. \tag{2.17}$$

Proof. Employing proof by induction, suppose that

$$x_k = (\mathbb{P}_{B^{P_\perp}} A)^k v_0 + \sum_{m=1}^{k-1} \Big(\prod_{i=1}^{(k-1)-m} \frac{1}{1+\gamma_i} \Big) (\mathbb{P}_{B^{P_\perp}} A)^m z_0 + \prod_{i=0}^{(k-1)} \frac{1}{1+\gamma_i} z_0 \,.$$

Note

$$(\mathbb{P}_{B^{P_\perp}} A)^k v_0 + \sum_{m=1}^{k-1} \Big(\prod_{i=1}^{(k-1)-m} \frac{1}{1+\gamma_i} \Big) (\mathbb{P}_{B^{P_\perp}} A)^m z_0 \in B^{P_\perp} \,,$$

therefore,

$$B(B^T P B)^{-1} B^T P \left((\mathbb{P}_{B^{P_\perp}} A)^k v_0 + \sum_{m=1}^{k-1} \Big(\prod_{i=1}^{(k-1)-m} \frac{1}{1+\gamma_i} \Big) (\mathbb{P}_{B^{P_\perp}} A)^m z_0 \right) = 0$$

and $(z_0 \in R(B))$, so

$$B(B^T P B)^{-1} B^T P z_0 = z_0 \,.$$

Hence, it follows

$$\begin{aligned} x_{k+1} &= \mathbb{P}_{B^{P_\perp}} A x_k + \frac{1}{1+\gamma_k} B(B^T P B)^{-1} B^T P x_k \\ &= (\mathbb{P}_{B^{P_\perp}} A)^{k+1} v_0 + \sum_{m=1}^{k} \Big(\prod_{i=1}^{(k)-m} \frac{1}{1+\gamma_i} \Big) (\mathbb{P}_{B^{P_\perp}} A)^m z_0 + \prod_{i=0}^{(k)} \frac{1}{1+\gamma_i} z_0 \,. \end{aligned}$$

□

Suppose further that P is a stabilizing matrix for (2.15), and let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of $\mathbb{P}_{B^{P_\perp}} A$ with $\lambda_{max} = max\{|\lambda_i| \ : \ i \in \{1, \dots n\}\}$. Let $\sigma > 0$ with $\lambda_{max} < \sigma < 1$ and $M > 0$ be such that for any system trajectory $\{y_k\}_{k\in\mathbb{N}}$ from initial condition $y_0 \in \mathbb{R}^d$ under the projection control, the following holds

$$||y_k|| < M\sigma^k ||y_0|| \,.$$

Such a choice of σ and M exist since P is stabilizing, therefore the projection control update is globally exponentially stable.

With P stabilizing an upper bound on the state can be established for any relaxed projection control. For weighting parameters $\{\gamma_k\}_{k\in\mathbb{N}}$, let $\{x_k\}_{k\in\mathbb{N}}$ be the system trajectory from x_0 under the relaxed projection control (2.16).

Proposition 2.5.3. *If $\{\gamma_k\}_{k\in\mathbb{N}}$ in $]0, \infty[$, then*

$$||x_k||_2 \leq M \left(||v_0||_2 + \frac{1}{1-\sigma} ||z_0||_2 \right) .$$

Proof. From (2.17),

$$||x_k||_2 \leq ||(\mathbb{P}_{B^{P_\perp}} A)^k v_0||_2 + \sum_{m=0}^{k-1} (\prod_{i=0}^{(k-1)-m} \frac{1}{1+\gamma_i}) ||(\mathbb{P}_{B^{P_\perp}} A)^m z_0||_2 \,.$$

Then since $(\forall k \in \mathbb{N})$, $(\frac{1}{1+\gamma_k} < 1)$ it follows that

$$||x_k||_2 \leq M(\sigma)^k ||v_0||_2 + M \sum_{m=0}^{k} (\sigma)^m ||z_0||_2 \,,$$

and $0 < \sigma < 1$, hence the result. □

With a slight restriction to the weighting parameters, the relaxed projection controls will be stabilizing as is given in the following proposition.

Proposition 2.5.4. *Let sequence $(\{\gamma_k\}_{k\in\mathbb{N}}$ in $]0, \infty[)$, if there exists $\eta \in]0, 1[$ such that $(\forall k \in \mathbb{N})$, $\frac{1}{1+\gamma_k} < \eta$, then the relaxed projection control is globally exponentially stable*

Proof. Let $r = \max\{\sigma, \eta\}$, then using (2.17),

$$||x_k||_2 \leq M\sigma^k ||v_0||_2 + M \sum_{m=0}^{k-1} \eta^{(k-1)-m} \sigma^m ||z_0|| \,,$$

therefore,

$$||x_k||_2 \leq M\sigma^k||v_0||_2 + Mkr^{k-1}||z_0||_2\,.$$

Since $(0 < \sigma < 1)$ and $(0 < r < 1)$, then for a $1 > \kappa > \max\{\sigma, r\}$, there exists an K such that for all $(x_0 \in \mathbb{R}^d)$ and $(k \in \mathbb{N})$

$$||x_k|| \leq K||x_0||\kappa^k\,.$$

Hence the relaxed projection control is globally exponentially stable. □

The weighting parameters to tune the control may also be chosen as a function of the state. It may be useful to modulate the magnitude of the control more when the state is far from the equilibrium, but allow for faster convergence once the state is near the origin. In particular, consider the choice

$$\gamma_k = \frac{\gamma}{||x_k||_2^\alpha}\,, \tag{2.18}$$

for a $(\gamma > 0)$ and $(\alpha > 0)$.

Proposition 2.5.5. *The relaxed projection control using weighting parameters* (2.18) *is globally exponentially stable.*

Proof. The proof follows immediately from Proposition 2.5.3 and Proposition 2.5.4. In particular, by Proposition 2.5.3, with $\ell = \frac{1}{1-\sigma}$ for all $(k \in \mathbb{N})$

$$\frac{1}{1+\gamma_k} \leq \frac{(M||v_0||_2 + M||z_0||_2\ell)^\alpha}{\gamma + (M||v_0||_2 + M||z_0||_2\ell)^\alpha}\,,$$

hence by Proposition 2.5.4 the control is globally exponentially stable. □

2.6 Synthesizing a control

Synthesizing a relaxed projection control for a discrete or continuous time system requires identifying a stabilizing P and an appropriate relaxation to produce the desired system response. Identifying stabilizing P for a system by searching over the positive definite cone can be challenging, particularly as the dimension of the system grows large. This section explores searching over subsets of the positive definite cone by searching over matrices with select singular value decompositions (SVD).

2.6.1 Discrete Time Scalar Control Case

Suppose that a system has dynamics given by

$$x_{k+1} = f(x_k) + g(x_k)u\,,$$

where f and g are functions $\mathbb{R}^d \to \mathbb{R}^d$, and u is a scalar control input. For an SPD matrix $P \in \mathbb{R}^{d\times d}$, the subspace $g(x_k)^{P_\perp}$ is a hyperplane with normal vector $Pg(x_k)$. Suppose that there is a fixed hyperplane with normal vector $v \in \mathbb{R}^d$ for which the projection and relaxed projection control exhibit fast convergence for the system. Then it may be effective to construct a P which fixes the hyperplane, that is for all $x \in \mathbb{R}^d$, construct P such that $Pg(x) \approx v$. Methods to compute such P can be constructed by designing singular value decompositions with the desired property.

Constructing an SVD for an SPD matrix requires specifying an orthonormal basis and set of corresponding singular values. Let $v \in \mathbb{R}^d$ and $\{e_i\}_{i=1}^d$ the standard orthonormal basis. Let R_v be a rotation matrix that rotates e_1 to v, for example R_v constructed with two Householder reflections as,

$$R_v = (I - 2ww^T)(I - 2e_1e_1^T)\,,$$

where $w = \frac{(-e_1-v)}{||-e_1-v||}$. Then R_v can be used to construct an orthonormal basis with the first element given by v. The basis will serve as singular vectors for the constructed SPD matrix with corresponding singular values $[s, 1\ldots, 1]$ where s has a large magnitude, for example $s = 1\times 10^6$. In particular, let

$$U_v = \begin{bmatrix} | & | & | \\ v & R_ve_2\ldots & R_ve_d \\ | & | & | \end{bmatrix}$$

and let P_v be constructed with the SVD,

$$P_v \doteq U_v\Lambda U_v^T\,, \tag{2.19}$$

where $\Lambda = diag([s, 1\ldots, 1])$.

If for all $x \in \mathbb{R}^d$, $sv^Tg(x)$ is sufficiently large then $P_vg(x) \approx v$. A stabilizing P for the system is then searched for over the set

$$\mathcal{P} = \{P_v \;:\; v \in \mathbb{R}^d, ||v|| = 1\}\,,$$

that is over the unit normal vectors in $\mathbb{R}^d$, instead of searching over the whole positive definite cone.

2.6.2 General Stabilizing P Construction

Consider that in general specifying stabilizing P for both discrete and continuous time systems may require only specifying the leading singular vectors and singular values. This section proposes a general form for designing stabilizing P by constructing appropriate SVD decompositions.

The first N orthonormal vectors $V = \{v_1, v_2, \dots v_N\}$ of the SVD basis are taken as variables. The remaining basis vectors are constructed using Gram Schmidt on the set of vectors

$$\{v_1, v_2, \dots v_N, e_{N+1}, \dots e_d\},$$

where e_i is the i^{th} standard normal basis vector. The resulting basis is denoted

$$\{v_1, v_2, \dots v_N, e_{g_{N+1}}, \dots e_{g_d}\}$$

and forms the orthonormal matrix

$$U_V = \begin{bmatrix} | & | & | & | \\ v_1 \dots & v_N & e_{g_{N+1}} \dots & e_{g_d} \\ | & | & | & | \end{bmatrix}.$$

The first N singular values $s = \{\sigma_1, \sigma_2, \dots \sigma_N\}$ corresponding to the first N basis vectors are also taken as variables while the remainder are set to the value one. The matrix of singular values is denoted $\Lambda_s = diag([\sigma_1 \dots, \sigma_N, 1 \dots, 1])$.

Stabilizing P are searched for over the SPD matrices $P_{V,s}$ of the form,

$$P_{V,s} = U_V \Lambda_s U_V^T, \tag{2.20}$$

by searching over the N basis vectors and singular values.

2.6.3 Ensemble Kalman algorithm

In this section, a general ensemble Kalman algorithm is outlined following a similar construction to [42, 76]. Then an implementation to synthesize relaxed projection controls is presented. Note that the intent of this section is only to present a general approach, future work is needed to evaluate construction of algorithms of this form to reliably achieve good performance for control synthesis particularly for high dimensional systems.

Let $y \in \mathbb{R}^m$ be outputs or performance measures for a system and let $\hat{y} \in \mathbb{R}^m$ a desired performance or output objective, with $\mathcal{G} : \mathcal{V} \to \mathbb{R}^m : v \to y$ the map from the input space $\mathcal{V}$ to the performance measures. An ensemble Kalman algorithm is constructed to find a minimizer of the difference between the observed performance and desired performance over the input space, by solving a problem of the form

$$\min_{v \in \mathcal{V}} \frac{1}{2} ||\mathcal{G}(v) - \hat{y}||^2. \tag{2.21}$$

The procedure evaluates the performance map $\mathcal{G}$ at an ensemble of input points, then each

point in the ensemble is updated using summary statistics of the output map over the ensemble according to the ensemble Kalman filter rule. The procedure is iterated until the mean of the ensemble achieves a desired performance.

For iteration n, the ensemble of size J will be denoted by $\{v_n^j\}_{j=1}^J$ and the computed performance measures for each ensemble point j of generation n will be denoted by $\{y_n^j\}_{j=1}^J$, where for all $j \in \{1, 2 \ldots, J\}$

$$y_n^j = \mathcal{G}(v_n^j)\,.$$

For iteration n the mean ensemble estimate and mean performance measure are given by

$$\bar{v}_n = \frac{1}{J}\sum_{j=1}^{J} v_n^j\,, \qquad \bar{y}_n = \frac{1}{J}\sum_{j=1}^{J} y_n^j\,.$$

Let $\Sigma_n^{v,y}$ denote the ensemble performance cross co-variance matrix, that is for $v_n^{j,i}$ element i of ensemble point j, and $y_n^{j,i}$ performance measure i of ensemble point j in iteration n, $\Sigma_n^{v,y}$ is the $(d \times m)$ matrix with entries (ℓ, i) given by

$$(\Sigma_n^{v,y})_{\ell,i} = \frac{1}{J-1}\sum_{j=1}^{J}(v_n^{j,\ell} - \bar{v}_n^{\ell})(y_n^{j,i} - \bar{y}_n^{i})\,.$$

Let $\Sigma_n^{y,y}$ the performance measure co-variance, the $(m \times m)$ matrix with entries (ℓ, i) given by

$$(\Sigma_n^{y,y})_{\ell,i} = \frac{1}{J-1}\sum_{j=1}^{J}(y_n^{j,\ell} - \bar{y}_n^{\ell})(y_n^{j,i} - \bar{y}_n^{i})\,.$$

The target performance objective is used in the role of the measurements of the ensemble Kalman filter. Random perturbation of the objective has been found to be needed for ensemble Kalman type schemes in order to keep the ensemble from collapsing towards the mean value [18].

Random performance measure objectives $\{\hat{y}_n^j\}_{j=1}^J$ are generated for each ensemble point j at each iteration n, by random normal perturbation of a given target objective $\hat{y}$ with co-variance R, that is

$$\hat{y}_n^j \sim \mathcal{N}(\hat{y}, R)\,.$$

At each iteration the ensemble points are updated by the Kalman rule, for all $j \in \{1, 2 \ldots, J\}$

$$v_{n+1}^j = v_n^j + \Sigma_n^{v,y}(\Sigma_n^{y,y} + R)^{-1}(y_n^j - \hat{y}_n^j) + w_n^j\,, \tag{2.22}$$

where $\{w_n^j\}_{j=1}^J$ are small random multi-normal perturbations with co-variance Σ_w, that is for all

$j \in \{1, 2 \ldots, J\}$

$$w_n^j \sim \mathcal{N}(0, \Sigma_w)\,.$$

2.6.4 Stabilizing P Synthesis Procedure

Stabilizing P for both continuous and discrete time systems are searched for using an ensemble Kalman search methodology over basis vectors $V \in \mathbb{R}^{d\times N}$ and singular values $s \in [0,\infty)^N$. The performance map $\mathcal{G} : \mathbb{R}^d \times \mathbb{R}^{d\times N} \times [0,\infty)^d \to \mathbb{R} : (x_0, \gamma, V, s) \to y$, constructs matrix $P_{V,s}$ according to (2.20) and computes the state trajectory from initial condition x_0, over a fixed time interval, under the appropriate control using matrix $P_{V,s}$. A measure of the performance of the control is then computed using the state trajectory.

The ensemble is initialized by random normal perturbation from a starting estimate for V and s. At each search iteration the following procedure is implemented,

Procedure 1 Ensemble Kalman Search Update

1: **for** $j = 1, 2 \ldots J$ **do**
2: generate random initial state x_0, $x_0 \sim \mathcal{N}(0, \Sigma_{x_0})$
3: Set $V = V_n^j$, and $s = s_n^j$, compute $P_{V,s}$ with (2.20).
4: Compute the state trajectory from x_0 under feedback control using $P_{V,s}$ to terminal time T.
5: Compute performance measures for the control
6: **end for**
7: Update the ensemble with rule (2.22).
8: Re-orthoganalize and normalize the ensemble basis elements.

The procedure tests controls from only one initial state at each iteration, while a P is likely sought which will produce good convergence over a neighborhood of the origin if not globally. While multiple initial states may be used at each iteration in the procedure, the computational cost is increased significantly. Using one initial state was found to be sufficient for the applications presented here.

Implementation For Discrete Time Systems

For discrete time systems a stabilizing P is found by evaluating the associated projection feedback controls (2.9) using the ensemble Kalman procedure. For the numerical examples which will be presented, control performance was evaluated after a fixed number of time steps K, using two

measures of the final state x_K given by

$$y^1 = ||x_K||_2 \qquad \text{and} \qquad y^2 = \frac{\langle f(x_K) - x_K \mid x_K \rangle_P}{||f(x_K) - x_K||_P ||x_K||_P} \,. \tag{2.23}$$

The criterion y^1 measures how close the control has brought the state to the origin. The angle criterion y^2 measures the component of the change in the state due to the system dynamics, $f(x_K) - x_K$, which is in the subspace $g(x_{K-1})^{P_\perp}$, and directed towards the origin with respect to the P inner product. Then, if the subspace is fixed, for example the hyperplane $v^{P_\perp}$, that is the component of the change which will be preserved by the control and move the state towards the origin. Inclusion of this term was found to improve performance for the numerical examples which will be presented in the next section.

Implementation For Continuous Time Systems

In the continuous time setting to search for stabilizing P requires using relaxed projection feedback controls (2.14) which require specifying weighting parameters. In the application example which is presented in Chapter 3 a fixed weighting parameter γ is used and selected to be small while still achieving convergence in the simulated time interval, such that Proposition 2.4.2 guarantees a large range of parameters for tuning the control.

The state trajectory is computed over a fixed time interval $[0, T]$ and then evaluated with the performance measure

$$y = \frac{1}{T - t^*} \int_{t^*}^{T} x(t)^T Q x(t) \; dt \,, \tag{2.24}$$

where $Q \in \mathbb{R}^{d \times d}$ is a positive definite matrix, and $0 < t^* < T$. The value y measures how close with respect to Q the control brings the state of the system to the origin in some interval from a t^* up to the final time T.

2.7 Numerical Examples For Discrete Time Dynamics

This section presents two examples of control synthesis and implementation using the ensemble Kalman search procedure to identify relaxed projection feedback controls for discrete time systems. A linear example is given first and then a nonlinear example is shown.

2.7.1 Linear Example

Consider a linear discrete time dynamical system of the form (2.15) with

$$A = \begin{bmatrix} .9974 & .0539 \\ -.1078 & 1.1591 \end{bmatrix}, \qquad B = \begin{bmatrix} .0013 \\ .0539 \end{bmatrix}.$$

The ensemble Kalman procedure was used to search for a stabilizing P of the form (2.19). An ensemble of 50 vectors in $\mathbb{R}^2$ was randomly initialized from the distribution $\mathcal{N}^2(0, 0.2)$. At each search iteration solutions under the control of each ensemble point were simulated for $K = 8$ steps from a random intial state $x_0 \sim \mathcal{N}(0, 1)$. Then the performance measures (2.23) were computed, with the performance objectives given by

$$\hat{y}^1 \sim \mathcal{N}(0, .0002)\,, \qquad \hat{y}^2 \sim \mathcal{N}(-.2, .01)\,.$$

The search procedure was iterated until satisfactory performance measures were returned, after about 150 iterations the ensemble mean was,

$$v = [-.9975, -.07]^T\,.$$

A reference optimal LQR feedback control was computed using

$$Q = \begin{bmatrix} 350 & 0 \\ 0 & 5 \end{bmatrix}, \qquad R = \begin{bmatrix} 3.5 \end{bmatrix}.$$

A relaxed projection control (2.10) using weighting parameters of the form (2.18) was then chosen such that the control had roughly the same maximal magnitude, and converged in approximately the same time as the reference LQR control. A suitable choice was found to be

$$\gamma_k = \frac{.45}{||x_k||_2^3}\,.$$

A comparison of the controls is shown in Fig. 2.1

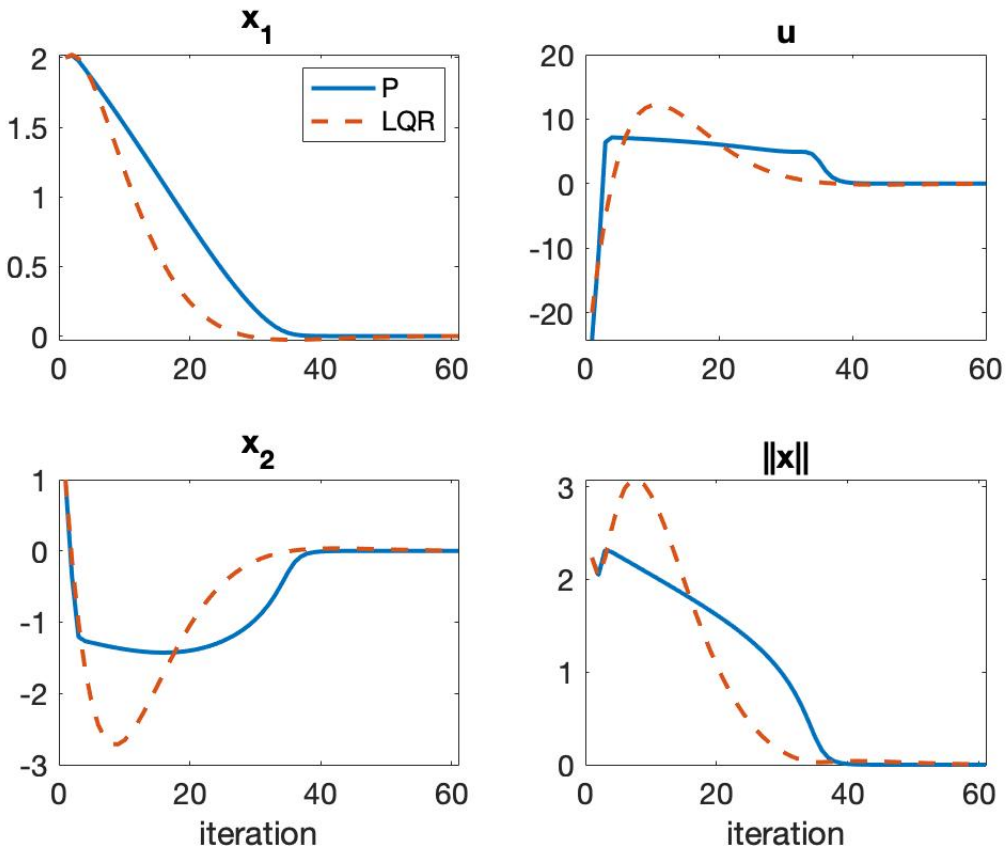

Figure 2.1: Comparison of LQR and relaxed P control from initial state $x_0 = [2, 1]^T$.

Note that the relaxed projection control convergence rate increases as the state approaches the origin, as can be more clearly seen in Fig. 2.2.

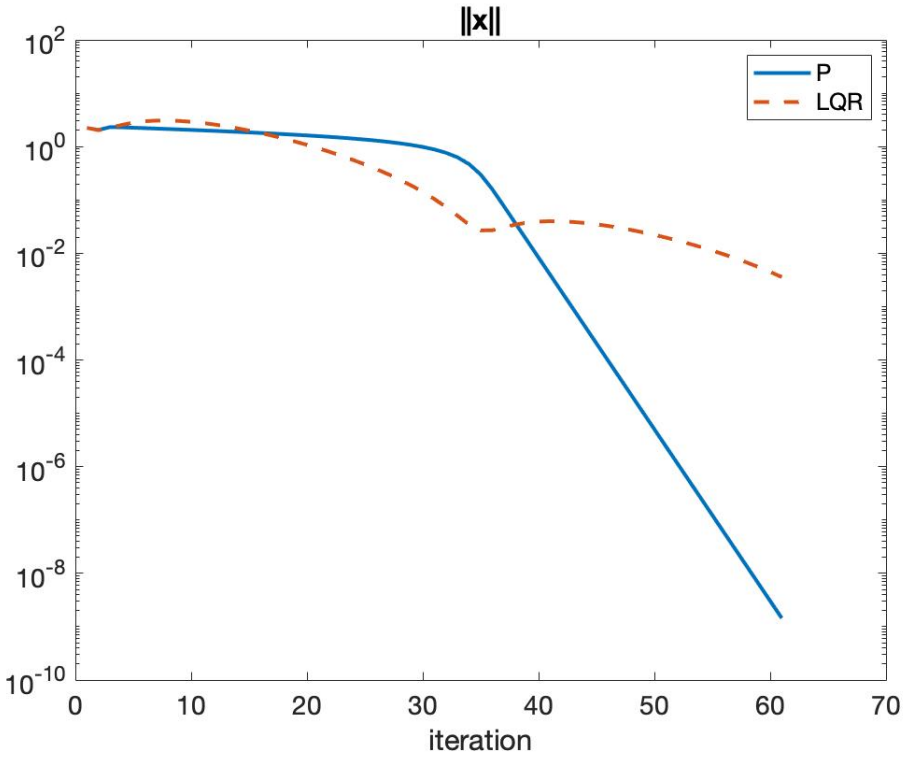

Figure 2.2: Comparison of convergence of LQR and relaxed projection control on a Log scale.

The relaxed projection control is slower than the LQR control to bring the system near the origin, but has a reduced overshoot, and transitions to a faster convergence rate once near the origin.

2.7.2 Nonlinear Example

Consider a system of the form (2.7) with

$$f(x) \doteq \begin{bmatrix} 2.2\sin(.5x_1) + .1x_2 \\ .1x_1^2 + 1.8x_2 \end{bmatrix}, \qquad g(x) \doteq \begin{bmatrix} 0 \\ 2 + .1\cos(x_2) \end{bmatrix}.$$

The Ensemble Kalman procedure was used to search for a stabilizing P of the form (2.19). An ensemble of 100 vectors in $\mathbb{R}^2$ was randomly initialized according to $\mathcal{N}^2(0, 0.2)$. At each search iteration solutions under the projection control of each ensemble point were simulated for $K = 4$ steps from a random initial state $x_0 \sim \mathcal{N}(0, 1)$. Then the performance measures (2.23)

were computed, with the performance objectives given by

$$\hat{y}^1 \sim \mathcal{N}(0, .002)\,, \qquad \hat{y}^2 \sim \mathcal{N}(-.2, .01)\,.$$

The search procedure was iterated until satisfactory performance measures were returned, after about 250 iterations the ensemble mean was,

$$v = [.9950, .1003]^T\,.$$

A relaxed projection control was then implemented with weighting parameters

$$\gamma_k = \frac{.1}{||x_k||_2^2}\,.$$

The result is shown in Fig. 2.3

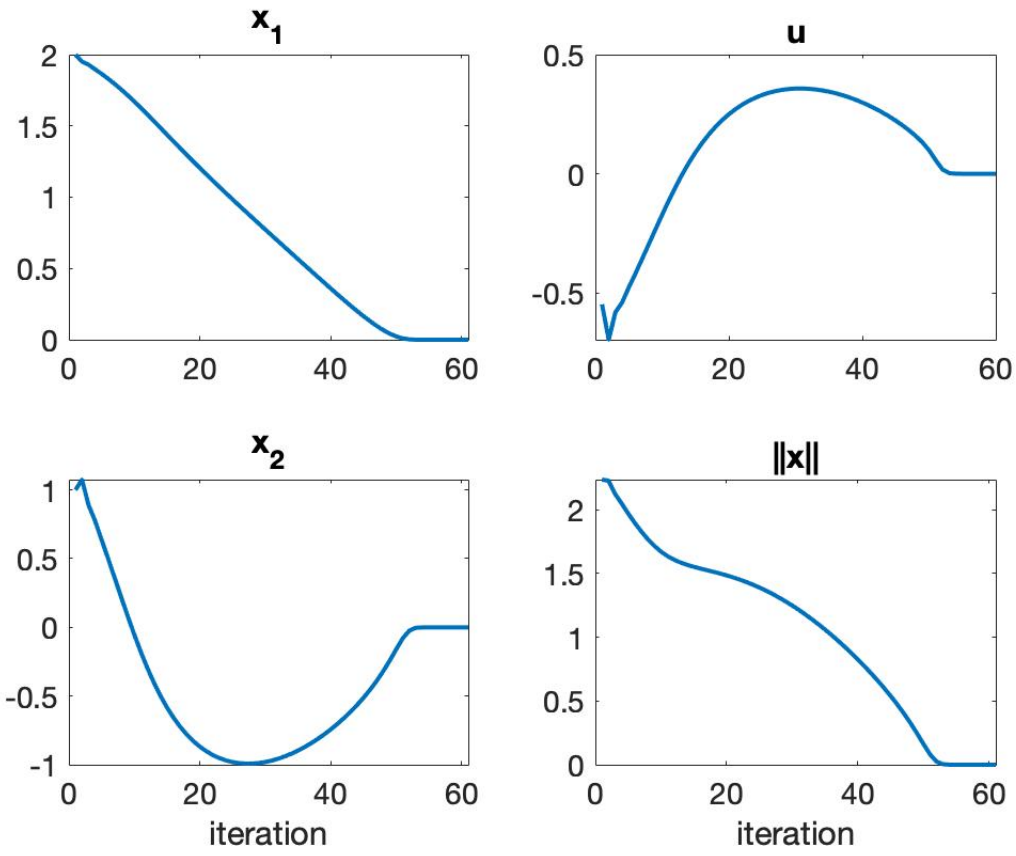

Figure 2.3: Nonlinear example with relaxed projection control implementation.

The parameter choice (2.18) for the relaxed P projection control resulted in smooth system responses for many choices of γ and α tested.

CHAPTER

3

REAL-TIME IMPLEMENTATIONS FOR STABILIZATION OF A DOUBLE INVERTED PENDULUM

Methods for both state estimation and control of dynamical systems may work in numerical simulation but fail to be effective for implementation in practice. The requirement for real-time computation of controls and state estimates can present a major hurdle, particularly as computational resources for online implementation may be limited while values may be needed quickly for a method to be feasible. Model based design for both control and state estimation also require that the model is a close representation of the true system dynamics. For a method to be effective in practice it must be robust to model errors, additional noise, and potential disturbances of the system.

This chapter addresses the feasiblity of the state estimation method presented in Chapter 1 and the control method presented in Chapter 2 for practical application. Each method is implemented online for stabilization of a double inverted pendulum (DIP) on a cart in laboratory experiments, a benchmark problem in nonlinear control. Their performance is then compared to methods implemented previously on the DIP system.

3.1 Double Inverted Pendulum

The double inverted pendulum on a cart is a benchmark problem in nonlinear control. The systems are relatively inexpensive and simple to operate while sharing important properties common to many other control systems. Properties which can present challenges for control and state estimation including; highly nonlinear dynamics, under-actuated control input, system and measurement noise, and model error. Thus a controls efficacy on the DIP system may be indicative of how it will perform on other systems of interest. Control of the DIP can also provide a model for systems such as robotic limbs [74, 81], human posture,balance, and gymnast motion [75, 83, 90].

The DIP on a cart, consists of two pendulums in tandem connected on a hinge to a cart which moves on a linear track as shown in Figure 3.1. The methods presented here are implemented for stabilization control, which refers to moving the cart along the track, such that the pendulums are balanced vertically over the cart in an upright unstable equilibrium.

3.1.1 Model Of The DIP Dynamics

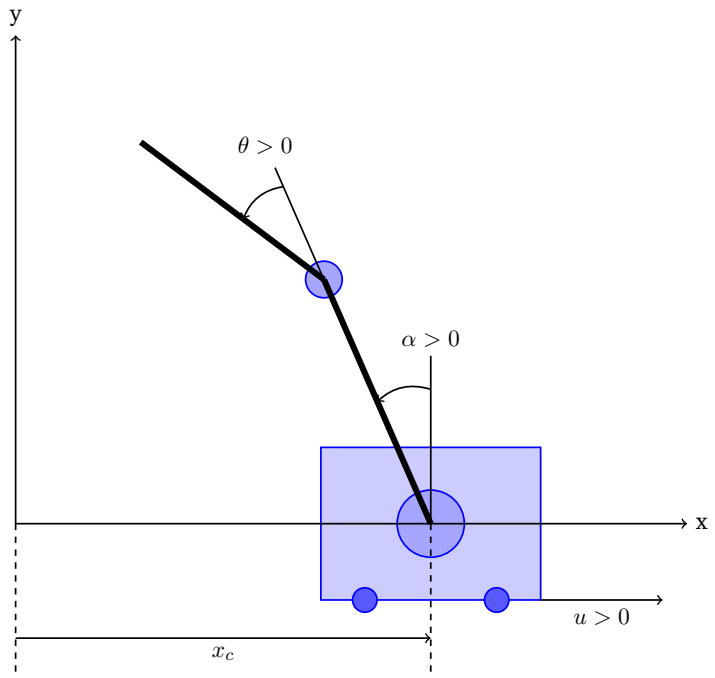

Figure 3.1: Diagram the DIP system.

A model is used for the DIP system derived using Lagrange's energy method as is commonly done in the literature [16,23,36]. In particular, the model derived and given by Bernstein in [14] is used here.

The state of the system is modeled by six variables: the position of the cart (x_c), the angle between the lower pendulum and normal vector vertical to the cart (α), the angle between the lower and upper pendulum (θ), and their derivatives $(\dot{x}_c, \dot{\alpha}, \dot{\theta})$. The measurements are defined such that the upright unstable equilibrium is the origin and counterclockwise rotation is positive.

The equations of motions for the system derived using Lagrange's energy method are

$$\begin{aligned}
&\left[M_c + \frac{J_m K_g^2}{r_{mp}^2} + M_1 + M_2 + M_h\right] x_c''(t) - \Big[\big(M_1\ell_1 + M_2 L_1 + M_h L_1\big)\cos(\alpha(t)) \\
&\qquad + M_2\ell_2 \cos(\alpha(t) + \theta(t))\Big]\alpha''(t) - M_2\ell_2\cos(\alpha(t) + \theta(t))\theta''(t) + B_c x_c'(t) \\
&\qquad \Big[\Big[\big(M_1\ell_1 + M_2 L_1 + M_h L_1\big)\sin(\alpha(t)) + M_2\ell_2\sin(\alpha(t) + \theta(t))\Big]\alpha'(t) + \\
&\qquad\qquad 2M_2\ell_2\sin(\alpha(t) + \theta(t))\theta'(t)\Big]\alpha'(t) + M_2\ell_2\sin(\alpha(t) + \theta(t))\big(\theta'(t)\big)^2 = F_c(t)\,,
\end{aligned}$$

$$\begin{aligned}
&-\Big[\big(M_1\ell_1 + M_2 L_1 + M_h L_1\big)\cos(\alpha(t)) + M_2\ell_2\cos(\alpha(t) + \theta(t))\Big]x_c''(t) \\
&+\Big[M_1\ell_1^2 + I_1 + M_2 L_1^2 + M_h L_1^2 + M_2\ell_2^2 + 2M_2 L_1\ell_2\cos(2\alpha(t) + \theta(t))\Big]\alpha''(t) \\
&+\Big[M_2 L_1\ell_2\cos(2\alpha(t) + \theta(t)) + M_2\ell_2^2\Big]\theta''(t) \\
&+\Big[B_1 - 2M_2 L_1\ell_2\sin(2\alpha(t) + \theta(t))\big(\alpha'(t) + \theta'(t)\big)\Big]\alpha'(t) \\
&-M_2 L_1\ell_2\sin(2\alpha(t) + \theta(t))\big(\theta'(t)\big)^2 \\
&-g\Big[M_1\ell_1 + M_2 L_1 + M_h L_1\Big]\sin(\alpha(t)) - gM_2\ell_2\sin(\alpha(t) + \theta(t)) = 0\,,
\end{aligned}$$

$$\begin{aligned}
&-M_2\ell_2\cos(\alpha(t) + \theta(t))x_c''(t) + \Big[M_2 L_1\ell_2\cos(2\alpha(t) + \theta(t)) + M_2\ell_2^2\Big]\alpha''(t) \\
&+\Big[M_2\ell_2^2 + I_2\Big]\theta''(t) - M_2 L_1\ell_2\sin(2\alpha(t) + \theta(t))\big(\alpha'(t)\big)^2 + B_2\theta'(t) \\
&-gM_2\ell_2\sin(\alpha(t) + \theta(t)) = 0\,.
\end{aligned}$$

The control affects the system by supplying a voltage V_m to a DC motor on the cart, which applies the driving force on the cart F_c in the first equation of motion. The driving force can be

written explicitly in terms of the voltage supplied as

$$F_c(t) = \frac{K_g K_t (r_{mp} V_m(t) - K_g K_m x_c'(t))}{R_m r_{mp}^2}.$$

Details for the derivation of the equations of motion can be found in [14, 15]. Descriptions of the parameters and values used for the computations in this paper are supplied in Table 3.3.

Let $f_d : \mathbb{R}^6 \to \mathbb{R}^6$ and $g_d : \mathbb{R}^6 \to \mathbb{R}^6$ give the solution of the equations of motion for the derivative of the system states,and let the control u be the voltage V_m supplied to the cart's DC motor, giving the state-space DIP model dynamics

$$\dot{x} = f_d(x) + g_d(x)u \qquad \text{where } x = [x_c, \alpha, \theta, \dot{x}_c, \dot{\alpha}, \dot{\theta}]^T. \tag{3.1}$$

Both a relaxed projection control and a moving horizon proximal point state estimator were implemented on a laboratory DIP system using this model.

3.1.2 Experimental Apparatus

The DIP laboratory setup was provided by Quanser Consulting Inc. and consists of an upper (12 in.) aluminium rod connected on a hinge to a lower (7 in.) rod which is in turn connected on a hinge to a cart (an IPO2 linear servo unit) that moves on a track. Encoders measure the position of the cart on the track and the pendulum angles, while the corresponding derivatives must be estimated for a full state description. Control is applied by changing the voltage supplied to a DC motor that moves the cart on the track.

State estimation and control were implemented in real time through MATLAB Simulink interfaced with Quanser's Quarc software on a desktop computer with a 3.20 GHz Intel Core i5 650 processor and 4 GB of RAM, connected to the DIP system by two Q2-USB DAQ control boards, with the control voltage applied to the cart by a VoltPAQ amplifier.

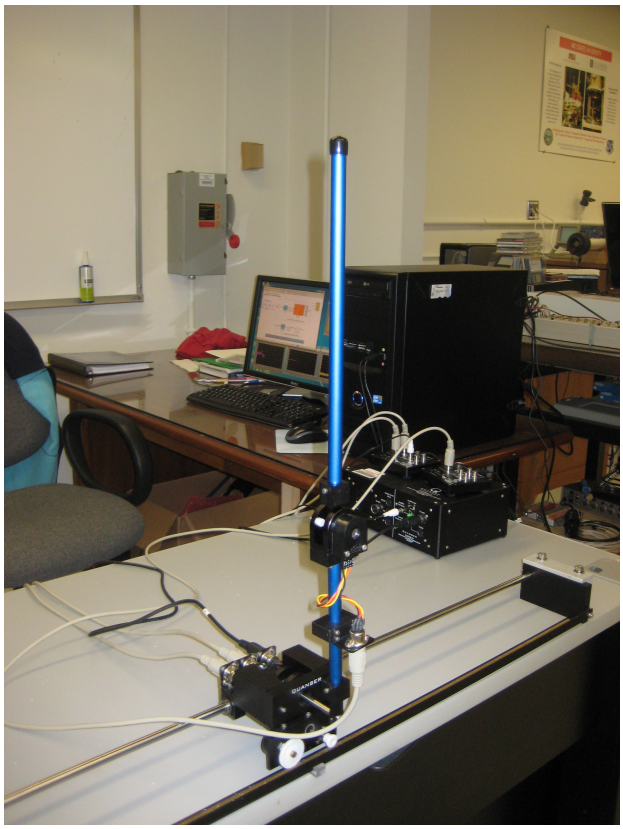

Figure 3.2: Laboratory DIP apparatus.

3.2 Implementation of the Relaxed Projection Control

Many control techniques have been applied for stabilization of the DIP system including the linear quadratic regulator (LQR) [16, 23], state dependent Riccati equation control [16], and neural network control [16, 89], though often the methods are only tested in numerical simulations. This section implements the relaxed projection control and control synthesis methodology presented in Chapter 2 for stabilization of the DIP system. The control is shown to compare well in experimental trials to previous methods that were implemented on the DIP laboratory system in [14].

Previous Stabilization Controls

The performance of the relaxed projection control is compared with two other feedback controls which have been previously implemented on the DIP laboratory system. Both controls are

computed for the DIP model dynamics

$$\dot{x} = f_d(x) + B_d u\,, \tag{3.2}$$

where $B_d \in \mathbb{R}^6$ is the linearization of g_d about the origin.

Let $x : [0, \infty) \to \mathbb{R}^6$ the solution of (3.2) from initial condition $x_0 \in \mathbb{R}^6$ under a control a measurable function $u : [0, \infty) \to \mathbb{R}$. Consider the cost functional

$$J(x_0, u) = \frac{1}{2} \int_0^\infty x(t)^T Q x(t) + R u(t)^2 dt\,, \tag{3.3}$$

where $Q \in \mathbb{R}^{6\times 6}$ and $R \in \mathbb{R}$ are symmetric positive definite matrices. The optimal control that minimizes the cost functional (3.3) is a feedback control given by

$$u(x) \doteq -R^{-1} B^T V_x(x)\,,$$

where V is the solution to the corresponding Hamilton Jacobi Bellman equation [8]. Solving for V is intractable for the DIP system as is true in general for nonlinear systems. Instead, approximations to the optimal control are used.

Let the expansion of f_d about the origin be given by

$$f_d(x) = A_d x + \sum_{n=2}^{\infty} f_{d_n}(x), \quad \text{where } f_{d_n}(x) = O(|x|^n)\,.$$

Linear Quadratic Regulator control (LQR) The dynamics of the system are approximated with the linear model $\dot{x} = A_d x + B_d u$, for which V is a quadratic and can be solved explicitly with the algebraic Riccatti equation [8]. Let the positive definite matrix $P \in \mathbb{R}^6$ be the corresponding solution, then the linear quadratic regulator (LQR) feedback control is given by

$$u(x) = -R^{-1} B_d^T P x\,. \tag{3.4}$$

While linearization is often an effective approach for constructing feedback controls for nonlinear systems, better approximations to the optimal control can also be made.

Power Series Control (PS) For the system dynamics (3.2) and cost functional (3.3) the solution V of the corresponding Hamilton Jacobi Bellman equation is approximated using power series expansions following [35]. In particular, a third order expansion is used which gives the

feedback control

$$u(x) = -R^{-1}B_d^T \left[Px - \left(A_d^T - PB_dR^{-1}B_d^T \right)^{-1} Pf_{d_3}(x) \right], \tag{3.5}$$

where P is the solution of the algebraic Riccatti equation corresponding to the linearized system. Full details for the derivation of this control for the DIP model and its implementation on the laboratory system can be found in [14].

3.2.1 DIP relaxed projection control synthesis

A relaxed projection control as presented in Chapter 2 was implemented for stabilization of the DIP system. Since the physical model for the system is in continuous time, the continuous time control formulation was used. Note that the laboratory system can only update the control in discrete time. While the discrete update map for the system could be approximated using the continuous time model it would be computationally expensive to do so, therefore to ensure real-time implementation the control was computed using the continuous time dynamics.

A stabilizing P for the DIP model (3.1) was identified using the ensemble Kalman methodology presented in Chapter 2, by searching for a matrix of the form (2.20). Orthonormal bases V were constructed iteratively. A search was conducted first with only one free basis vector and the corresponding singular value fixed at 1×10^6. The procedure was found to generally be very sensitive to the initial estimate and initialization of the ensemble. Some success was found initializing v_1 by computing the solution of the algebraic Ricatti equation corresponding to an LQR control for the linearized system, then taking an initial estimate for v_1 as the eigenvector of the solution corresponding to the largest eigenvalue. While P using one free basis vector could be found which were stabilizing in a neighborhood of the origin for simulations, none were effective on the laboratory system. A search was then conducted over two free basis vectors and free corresponding singular values. The ensembles were initialized by perturbing stabilizing v_1 and constructing v_2 by small random normal perturbations of vectors from the set $v_1^\perp \cap B_d^\perp$, where B_d is the linearization of g_d about the origin. The singular values were initialized as random normal perturbations from $s = [1 \times 10^6, 1 \times 10^5]$.

The ensemble Kalman procedure was iterated using an ensemble of size $J = 80$ for a maximum of 1000 iterations. For each ensemble point the control (2.14) was simulated with weighting parameter $\gamma = 10$. Simulations were computed with MATLAB ode45. Controls were updated at $.001$ second intervals, which is the frequency of state measurements for the laboratory system. The system state in the simulations was also perturbed every $.001$ seconds by the addition of small random noise $\mathcal{N}(0, 5 \times 10^{-5})$ which was found to improve robustness of the control. To achieve feasible computation times control simulations for the ensemble points were run in parallel on twelve cores using the North Carolina State Math Department's High Performance

Computing Cluster. Performance of the control was evaluated with the functional (2.23) using $Q = diag([80, 300, 100])$. Search Trials were run successively while increasing the total simulation time up to $T = 3(s)$, and adjusting the performance objective parameters $(\hat{y}, R)$, until good control performance could be achieved. Note that a control could fail to stabilize the system, if the norm of the state surpassed the threshold value 10, then the simulation was terminated and the performance measure was set to the value 1000.

Of the stabilizing P identified for the model, the best performance on the laboratory DIP system was achieved using the basis and singular values

$$\begin{aligned} v_1 &= \begin{bmatrix} .0758 & -.2888 & -.9414 & .0699 & -.1077 & -.0904 \end{bmatrix}^T \\ v_2 &= \begin{bmatrix} .4917 & .5334 & -.2177 & -.5440 & .3085 & .1877 \end{bmatrix}^T \\ s &= \begin{bmatrix} 9.99 \times 10^5 & 3.37 \times 10^4 \end{bmatrix} . \end{aligned} \tag{3.6}$$

3.2.2 DIP stabilization control implementation

The relaxed projection (RP) control (2.14) using P constructed with (3.6) and fixed weighting parameter $\gamma = 40$ was compared to an LQR control (3.4) and power-series (PS) control (3.5). The LQR and PS control methodologies were implemented on the DIP laboratory system by Bernstein in [14]. Out of those tested, for both methods, using $Q = diag([30, 350, 100])$ and $R = .1$ produced feedback control which maintained the smallest average magnitude for the pendulum angles during stabilization and are the implementations which the RP control is compared against here.

Numerical simulation The controls were simulated on the DIP model (3.1) using MATLAB ode45, with the control value updated every $.001$ seconds from the initial condition

$$x_0 = [.0192, .0025, .0287, -.0392, -.004, -.0242]^T .$$

Figure 3.3 shows that the control and state trajectory for the RP and PS controls appear more similar to each-other than the LQR control, though for some initial conditions the PS control was more similar to the LQR than the RP control. The RP control generally induced more oscillatory state behavior than either the LQR or PS control and had the largest magnitude control efforts.

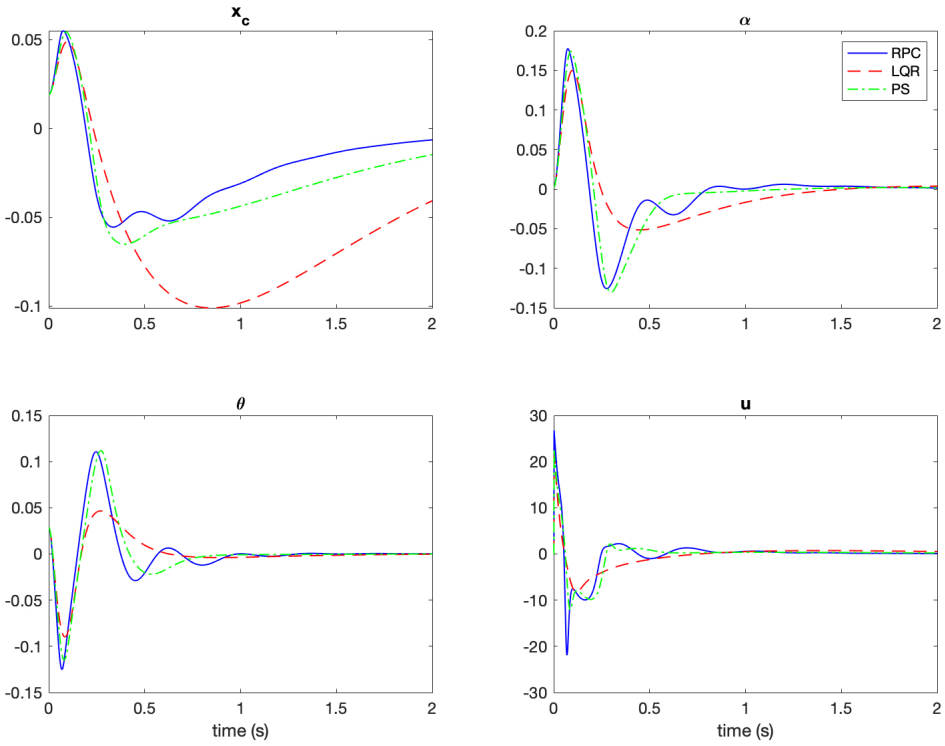

Figure 3.3: Comparison of relaxed projection (RP), power-series (PS), and LQR feedback controls in a numerical simulation for stabilization of the DIP model

Online implementation The controls were implemented on the laboratory DIP system. The cart started in the center of the track and the pendulums were manually rotated into the upright position, stabilization control was initiated once measurement values were within $.01$ of the balanced state.

For each control, measurement data was collected over a typical twenty second interval. The average absolute value of the measured states and control voltage are reported in Table 3.1. The RP control maintained the smallest average magnitude and the smallest variance for all state values. The average control magnitude for the RP control was almost twice that of the LQR and PS controls, and the variance more than twice as large. Figure 3.4 illustrates the much more oscillatory behavior of the RP control, over a typical $.5$ second interval, than either the LQR or PS controls. The figure also shows that the angle between the pendulums θ was kept closer to the balanced state by the RP control.

Table 3.1: A comparison of values for the mean and variance, on a typical twenty second interval, of the cart position (x_c), pendulum angles (α, θ), and control voltage (u) for the double inverted pendulum system under stabilization control by each of: LQR control, powerseries control, and relaxed projection control.

	LQR		Power-series		Relaxed projection	
	mean	variance	mean	variance	mean	variance
$\lvert\alpha\rvert$	3.4	3.03	1.8	1.1	1.4	.85
$\lvert\theta\rvert$	1.9	.81	.9	.35	.46	.09
$\lvert x_c\rvert$	29.4	410.7	36.5	639.4	23.4	285.1
$\lvert u\rvert$	1.7	1.5	1.8	2.9	2.5	5.0

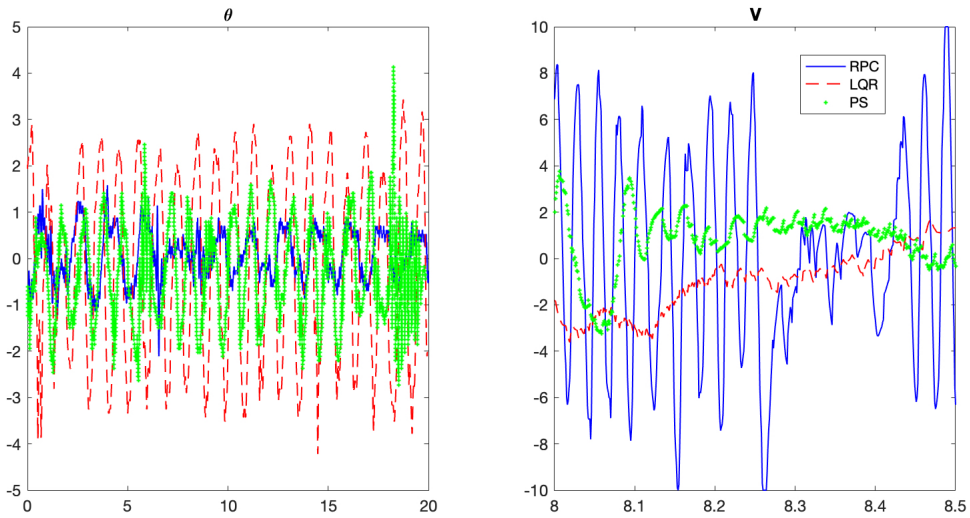

Figure 3.4: Comparison of relaxed projection (RP), power-series (PS), and LQR feedback controls for online stabilization of the laboratory system.

3.3 Implementation of Proximal Point Moving Horizon Estimation

Feedback control of the DIP system requires full state information for computation. The laboratory system has encoders which directly measure the pendulum angles and cart position, however

their velocities must be estimated.

State estimation for stabilization control of the DIP from an upright start is a non-trivial problem. Rapid convergence of state estimates, from a potentially poor initial estimate, is required in order for the feedback control to rescue the system. Moreover, the computation of the estimates must be fast such that the control can be frequently updated. The estimation method must also achieve robust performance in the presence of significant model error, measurement error, and disturbances to be effective. State estimation for DIP stabilization control has been found to be a challenging task for some methods. In particular, Bernstein found that an extended Kalman filter was not able to supply state estimates to a power-series controller such that the system could be stabilized even in simulations, if the full nonlinear model dynamics with white noise added to the model and measurements was used. While other methods were found to be simply too computationally intensive to be feasible [14].

This section presents an implementation of a moving horizon proximal point state estimator as developed in Chapter 1. Linearization of the DIP dynamics about the unstable equilibrium is used to construct a state estimator which can be computed quickly.

A centered proximal point moving horizon estimator of the form (1.29) was applied to the DIP system by first constructing a discrete system of the form

$$\begin{aligned} x_{k+1} &= \Phi x_k + W u_k \eta_k \\ y_{k+1} &= C x_{k+1} + \epsilon_{k+1}\,, \end{aligned}$$

using a linearization of the nonlinear DIP model about the unstable equilibrium. The state transition matrices Φ, Φ^{-1}, W, and W^{-1} for the linearized system were approximated using Matlab's expm command. A fixed $(\gamma > 0)$ was used in the computation of state estimates according to proximal point observer (1.10) using the cost functionals (1.29), which at each iteration requires a solution to

$$p_k = \text{argmin}_{z\in\mathbb{R}^6} \frac{1}{2}||Hz - q_k||^2\,, \tag{3.7}$$

for

$$H = \begin{bmatrix} \frac{1}{\gamma} I \\ C\Phi^{-1} \\ C \\ C\Phi \end{bmatrix} \quad \text{and} \quad q_k = \begin{bmatrix} \frac{1}{\gamma}\hat{x}_k \\ y_{k-1} - CW^{-1}Bu_{n-1} \\ y_k \\ y_{k+1} - CWBu_n \end{bmatrix}\,.$$

To compute (3.7), an offline QR factorization for $H = Q_H R_H$ was computed with the Matlab qr command. Then the centered proximal point moving horizon estimation (CPX) with a fixed γ

was iterated from initial estimate $\hat{x}_0 = 0$ according to

$$\begin{aligned} p_k &= R_H^{-1} Q_H^T q_k \\ \hat{x}_{k+1} &= \Phi p_k + W u_k \,. \end{aligned} \tag{3.8}$$

When implemented for the DIP feedback control in real time, the estimates supplied to compute the control were the model predictions $\tilde{x}_{k+2} = \Phi \hat{x}_{k+1} + u_{k+1}$.

The power series feedback stabilization control was computed with

$$Q = diag([80, 300, 100, 0, 0, 0]) \text{ and } R = .5\,,$$

where the state of the system is $x = [x_c, \theta, \alpha, \dot{x_c}, \dot{\theta}, \dot{\alpha}]^T$.

For a comparison study, both a CPX and a second order low pass derivative filter (LDF) were used to supply state estimates for stabilization control. The CPX estimator (3.8) was applied with $\gamma = 150$, while the LDF was used with Quanser's supplied parameters: cutoff frequency $\omega = 100\pi$ for the cart, $\omega = 20\pi$ for the pendulum angles, and damping ratios $.9$. Stabilization control was initiated once measurement values were brought to within $.01$ of the balanced state, the average value and variance for the measured DIP states when under stabilization control with CPX and LDF are reported in Table 3.2. Feedback control using the CPX estimates maintained the system closer to the balanced state and with less variance than with LDF estimates.

Table 3.2: Output of DIP stabilization over (6.5 sec) interval using either centered proximal point MHE (CPX) or low pass derivative filter (LDF) to compute the feedback control, the stabilized state is the origin

	x_c (cm)		α (degrees)		θ (degrees)	
	mean	variance	mean	variance	mean	variance
CPX	-.002	5.52	.178	35.2	-.45	3.89
LDF	-.119	5.98	1.21	41.8	-.867	6.78

The CPX and LDF differed most for the estimates of the rate of change for θ, the angle between the pendulums. Figure 3.5 shows a comparison between CPX and LDF angle velocity estimates using measurement data from the physical DIP system under stabilization control.

The criteria for γ to guarantee CPX estimate convergence given in Proposition 1.4.3 is $\gamma \geq 24400$ for the DIP system. When values of γ satisfying the condition were used, the CPX angle velocity estimates had large amplitude high frequency oscillations unsuitable for computing

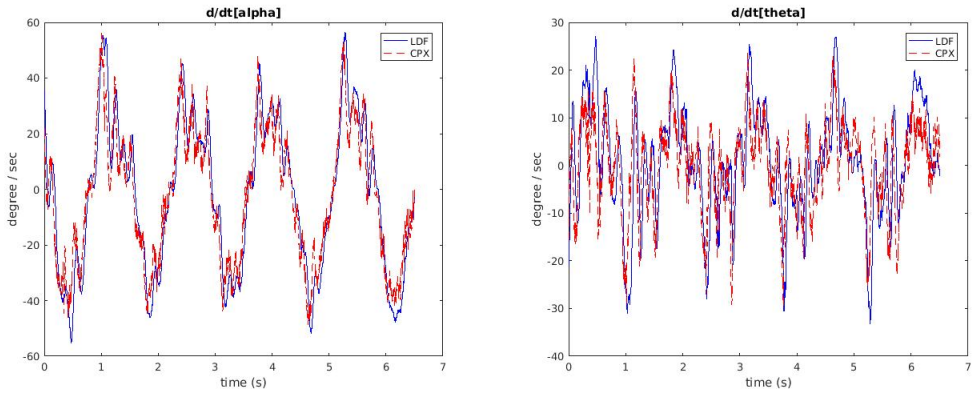

Figure 3.5: Comparison of CPX and LDF angle velocity estimates for the real time DIP system under stabilization control.

a control, γ was reduced two orders of magnitude from the Proposition 1.4.3 criteria before a reasonable control could be computed. The need for a smaller γ value is likely due to H in (3.7) becoming more ill conditioned for larger γ and the solutions of (3.7) more sensitive to noise.

Table 3.3: Description of the parameters and values used in the double inverted pendulum model for all computations

Symbol	Description	Value
B_c	Viscous Damping at the Motor Pinion	5.4 N.m.s/rad
B_1	Viscous Damping at the Lower Pendulum Axis	0.0024 N.m.s/rad
B_2	Viscous Damping at the Upper Pendulum Axis	0.0024 N.m.s/rad
g	Gravitational Constant	9.81 m/s^2
I_1	Moment of Inertia of the Lower Pendulum	2.6347E-004 kg.m^2
I_2	Moment of Inertia of the Upper Pendulum	1.1987E-003 kg.m^2
J_m	Rotational Moment of Inertia of the DC Motor	3.9E-007 kg.m^2
K_g	Planetary Gearbox Gear Ratio	3.71
K_m	Back-ElectroMotive-Force Constant	0.00767 V.s/rad
K_t	Motor Torque Constant	0.00767 N.m/A
ℓ_1	Length of Lower Pendulum from Pivot to Center of Gravity	0.1143 m
ℓ_2	Length of Upper Pendulum from Hinge to Center of Gravity	0.1778 m
L_1	Total Length of Lower Pendulum	0.2096 m
L_2	Total Length of Upper Pendulum	0.3365 m
M_c	Cart Mass	0.57 kg
M_h	Hinge Mass	0.170 kg
M_w	Extra Weight Mass	0.37 kg
M_1	Lower Pendulum Mass	0.072 kg
M_2	Upper Pendulum Mass	0.127 kg
R_m	Motor Armature Resistance	2.6 Ω
r_{mp}	Motor Pinion Radius	6.35E-003 m

CHAPTER

4

CHARACTERIZATION OF AN OPTIMAL INNATE IMMUNE RESPONSE AT THE ONSET OF INFECTION

At the onset of viral infection in a human host, host responses to slow the spread of infection are generally categorized within two broad classes, the innate immune system and the adaptive immune system. The innate immune system comprises those responses which are non-specific to the infecting virus, and are generally understood to be a first line of defense which reduces the spread of the virus, allowing time for adaptive immune responses, with more specificity, to become engaged and help to clear the infection [1]. The innate immune system can also clear many infections on its own and is involved in controlling the adaptive response [45, 85].

This chapter considers a component of the innate immune response mediated by Interferon-β, a signalling molecule that can initiate an antiviral state in cells surrounding sites of infection. The protection of cells from viral infection to slow the infection spread is approached from an optimal control perspective. The initial spread of infection is modeled as a stochastic branching process and the performance of cell protection control strategies are measured with respect to reducing the expected number of secondary infected cells while limiting the size of the protected region and total time cells spend protected. Optimal antiviral protection control strategies in the presented framework are then compared to experimental observations.

4.1 The Interferon-β Response

Upon infection of a host cell by a virus, host cell pattern recognition receptors can bind conserved viral-associated molecular patterns initiating a signaling cascade with downstream responses that can interfere with viral replication, recruit immune effector cells, and initiate adaptive immune responses [1,84]. For example, the receptor retinoic acid-inducible gene I (RIG-I), binds double stranded RNA and can induce production of type I interferons [84].

Type I interferons (IFN) are a class of cytokines commonly observed to be produced by infected cells that can 'interfere' with viral reproduction through up-regulation of the interferon stimulated genes [40,44,56]. In particular, interferon-β can be produced by a large variety of cell types, is found to be produced by cells when infected, and has been shown to function as an inter-cellular signal that can induce an antiviral physiological state in uninfected cells; providing a mechanism for protection of cells around sites of infection [44,56,68].

The effectiveness of the IFN response for inhibiting infection is underscored by the fact that a part of most viral genomes is dedicated to disrupting the production and signalling pathways of IFN [56]. However, IFN also has the potential to be damaging, and can be a very powerful signal since most cells can respond to it [44]. High expression of IFN is associated with poor outcomes of influenza infection in humans [27] and disease states can be exacerbated in animal models that lack negative regulators for the IFN response [66]. Porritt et al. suggest that an extensive negative regulatory network within IFN signaling works to strike a balance which ensures sufficient action to slow viral spread, while limiting damage from immune responses [66].

Both initiation of the production of IFN-β in infected cells and initiation of an antiviral response in uninfected cells upon exposure to IFN-β have been observed to exhibit what is called cell intrinsic stochasticity [64,68,93]. A response is a cell intrinsic stochastic response if when a population of identical cells are exposed to a homogeneous signal, only a random subset respond. For instance the initiation of IFN-β production by infected cells does not appear to be linked to the amount of virus in the cell [64,93]. Similarly, Rand et al. determined that heterogeneous initiation of an anti-viral state in cells was not due to limiting concentrations of IFN-β. In a detailed study of the initiation of IFN-β production upon infection, Zhao et al. hypothesize that the stochastic response is due to nearly every constituent part of the signalling pathway being present in rate limiting quantities [93]. Zhao et al. also tested the response of cells to IFN-β and in contrast to Rand et al. observed a homogeneous rather than stochastic response. However, they did not verify that the downstream product they measured corresponded to an antiviral response as Rand et al. did in their study [68,93].

There has been some speculation about how the observed dynamics of both the production and response to IFN-β may be optimized to slow the infection [64,68,93]. Less though has been formally said about what may qualify a response as optimal. The focus of this chapter is on

constructing a rigorous framework and definition of optimality for the antiviral protection of uninfected cells.

4.2 Optimal Control Approach

It should be stated first that there is not an expectation that the dynamics of the immune response are in fact optimal for some measure of immune performance. Rather, it is the intent that by specifying quantitatively some measure of what is hypothesized to capture good performance, the discrepancy between what is optimal with respect to that performance measure and observations of the true system, is revealing. A similar approach was undertaken by Perelson for the adaptive immune system in [65].

Based on the observation of Rand et al. in [68] that the initiation of an antiviral state in cells was 'all or nothing', this work approaches antiviral protection of cells in a simplified setting where protection in cells is either on or off, and protected cells can not become infected. The antiviral protection of cells shares much in common with vaccination control at the population level.

The efficacy of vaccination strategies for preventing the spread of infection have been commonly and effectively studied using the quantity R_0, the expected number of infection progeny of a single infected individual in a fixed susceptible population [39, 70]. The approach models the initial spread of infection as a stochastic branching process and if a control strategy brings the native value of R_0 to an R^* with $R^* < 1$ then the infection will not be able to invade or become established within the population [39, 70].

In a homogeneously mixing population with potential transmission contacts between individuals modeled by mass action, a vaccination strategy to reduce the native value of the initial expected number of infection progeny R_0 for an infection, to an R^*, requires that the proportion $(1 - \frac{R^*}{R_0})$ of the population be vaccinated [39]. Often R^* is taken to be one, and this proportion is referred to as the vaccination fraction. Note that for an infection where each infected is initially expected to infect several individuals, an $R_0 > 2$, to achieve an R^* near one may require a large majority of the population to be vaccinated. Within the context of an immune response which initiates an antiviral state in cells, a majority of the cell population would need to enter into an antiviral state, which could have severe physiological penalties. This suggests such an immune strategy may have limited feasibility in a free mixing system and is perhaps instead more suited for controlling spread of infection within bodies of cells with fixed spatial structure. Indeed, studies of vaccination control have consistently found underlying structure and dynamics of infection spread to be important considerations [26, 28, 67, 72, 91]. Unlike vaccination control, an antiviral immune protection response must also consider the timing of cell protection, to ensure that cells are protected fast enough and long enough to slow spread, but also return to a

physiologically normal state with limited delay.

The importance of understanding the spatial structure of viral infection in a host is highlighted by Funk et al. in [34] and both experiments and modeling have been conducted to understand the role of cell spatial structure and infection dynamics for the IFN response [24, 41, 52]. Models of the IFN response incorporating spatial structure have been studied in both discrete [41] and continuous [52] settings. Detailed modeling of the IFN signaling pathway within cells has also been explored to understand the dynamics of the IFN response [55, 79, 85, 93].

This work models the spread of viral infection through a body of cells with a fixed spatial structure using a stochastic branching process. An explicit mechanistic model for the IFN response is not used, instead an optimal antiviral cell protection response is constructed to serve as a comparison to the observed dynamics for the true system. The performance of the antiviral protection response is measured with respect to a given reduction in R_0 the expected number of infection progeny of an initial infected cell.

4.3 Infection Model

The initial spread of infection is modeled by a single infected cell located in a domain D of susceptible cells, where D is a Lebesgue measurable subset of $\mathbb{R}^d$. The infection is modeled as a stochastic process with concurrent release of viral progeny from the infected cell upon cell death. The probability measure space is denoted with the probability triple $(\Omega, \mathcal{A}, \mathbb{P})$.

Let $\tau_\eta : \Omega \to [0, \infty)$, the random variable for the time to death of the infected cell, and $Z : \Omega \times [0, \infty) \to \mathbb{N}$ be the random variable for the number of virus produced by the infected cell given time of cell death $t > 0$, with distributions following the assumptions **(A)**:

(A1) The time to infected cell death τ_η is exponentially distributed with parameter $\eta > 0$, that is

$$\mathbb{P}(\{\omega : \tau_\eta(\omega) \leq t\}) = \int_0^t \eta e^{-\eta t} dt \,.$$

(A2) The number of virus produced Z has a Poisson distribution with birth rate parameter $\beta > 0$, then given time of cell death $t > 0$, for $(n \in \mathbb{N})$

$$\mathbb{P}(\{\omega : Z(\omega, t) = n\}) = \frac{(\beta t)^n e^{-\beta t}}{n!} \,.$$

For each virus j produced by the infected cell, let $X_j : \Omega \to \{0, 1\}$ be the random variable defined

$$X_j(\omega) \;=\; \begin{cases} 1 & \text{virus } j \text{ infects a new cell} \\ 0 & \text{virus } j \text{ decays before infecting a new cell} \end{cases}$$

and it is assumed that

(A3) The $\{X_j\}_{j\in\mathbb{N}}$ are independent and identically distributed.

The number of infection progeny of the original infected cell $I : \Omega \to \mathbb{N}$ is given by

$$I \;=\; \sum_{j=1}^{Z} X_j \;=\; \sum_{k=1}^{\infty}\sum_{j=1}^{k} 1_{\{Z=k\}} \cdot X_j \,.$$

In this model, a control can be implemented to reduce the expected number of new infections $\mathbb{E}[I]$ by reducing the expectation of infection for each virus $\{\mathbb{E}[X_j]\}_{j\in\mathbb{N}}$. The cell protection control impacts the $\{\mathbb{E}[X_j]\}_{j\in\mathbb{N}}$, by changing the amount of time each virus is in contact with susceptible cells. A virus j can only infect a new cell at time t if the virus position is in the set $S(t)$, the subset of the domain D which is susceptible at time t. It is assumed that the whole of the domain D is susceptible unless occluded by the control, with the control a set valued function

$$A : [0,\infty[\;\to\; \mathcal{M}(D) \;:= \{B \subseteq D \mid B \text{ is Lebesgue measurable}\},$$

where $A(t)$ gives the region of the domain D in which cells are protected at time $t \geq 0$.

Given a control A, we define the susceptible set

$$S(t) \doteq D \backslash A(t)\,.$$

The expectation for infection by each virus $\{\mathbb{E}[X_j]\}_{j\in\mathbb{N}}$ depends on the infection process which is described by the following three random variables for each viral progeny j:

- the position of the viral particle in the domain $W_j : [0,\infty[\times\Omega \to \mathbb{R}^n$;
- the time to decay $\tau_{\delta_j} : \Omega \to [0,\infty]$;
- the time to entering a cell $\tau_{\gamma_j} : \Omega \to [0,\infty[$.

Virus may either decay or successfully enter a cell. If the cell the virus enters is susceptible then an infection will result, otherwise the virus is assumed to decay on entry. The control $A(t)$ acts by changing the probability that a cell that is entered by a virus is susceptible at the time of entry.

In terms of these random variables, X_j can be written

$$X_j \;=\; \begin{cases} 1 & \text{if} \qquad \tau_{\gamma_j} \leq \tau_{\delta_j} \text{ and } \; W_j(\tau_{\gamma_j}) \in A^c(\tau_{\gamma_j}) \\ \\ 0 & \text{otherwise.} \end{cases}$$

The affect of the protection control on the expected number of infection progeny is studied with the distributions for the infection model random variables satisfying the following assumptions **(A)**:

(A4) *Given $j \in \mathbb{N}$, the time to decay τ_{δ_j} of virus j is exponentially distributed with parameter δ, that is*

$$\mathbb{P}(\{\omega : \tau_{\delta_j}(\omega) \leq t\}) = \int_0^t \delta e^{-\delta t} dt\,.$$

(A5) *Given $j \in \mathbb{N}$, the time to cell entry τ_{γ_j} of virus j is exponentially distributed with parameter γ, that is*

$$\mathbb{P}(\{\omega : \tau_{\gamma_j}(\omega) \leq t\}) = \int_0^t \gamma e^{-\gamma t} dt\,.$$

(A6) *The viral position W_j follows a Brownian type motion, with a given transition density function $p : [0,\infty) \times [0,\infty) \times \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ that is for a measurable set $C \in \mathcal{M}(D)$, from a given starting position $(x_0 \in D)$ and starting time $t_0 > 0$, for final time $t > t_0$,*

$$\mathbb{P}(\{\omega : W_j(\omega, t) \in C\}) = \int_C p(t_0, t, x_0, x)\; dx\,.$$

4.3.1 Affect of the control on the expected number of infection progeny

Under assumptions **(A1)**-**(A3)** the expected number of new infections is computed by

$$\begin{aligned} \mathbb{E}[I] &= \sum_{k=1}^{\infty}\sum_{j=1}^{k} \mathbb{E}[1_{\{Z=k\}} X_j] \;=\; \sum_{k=1}^{\infty}\sum_{j=1}^{k} \mathbb{P}(\{Z=k\})\mathbb{E}[X_j] \\ &= \sum_{k=1}^{\infty} k\mathbb{P}(\{Z=k\})\mathbb{E}[X_j] \;=\; \mathbb{E}[X_j]\mathbb{E}[Z] \end{aligned} \tag{4.1}$$

and using **(A1)**-**(A2)**,

$$\mathbb{E}[Z] = \int_0^{\infty} \sum_{k=1}^{\infty} k\mathbb{P}(\{Z(t)=k\})\eta e^{-\eta t} dt = \int_0^{\infty} \beta t\; \eta e^{-\eta t} dt = \frac{\beta}{\eta}\,.$$

It then remains to compute $\mathbb{E}[X_j]$ to find $\mathbb{E}[I]$.

Given a control A, and under the assumptions **(A4)-(A6)**

$$\begin{aligned} \mathbb{E}[X_j] &= \mathbb{P}(X_j = 1) \\ &= \int_0^{\infty} \mathbb{P}(\{\tau_{\delta_j} > t\;,\; W_j(t) \in A^c(t)\} | \tau_{\gamma_j} = t)\; d\mathbb{P}(\tau_{\gamma_j} = t) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \mathbb{P}(\tau_{\delta_j} > t)\, \mathbb{P}(W_j(t) \in A^c(t))\, d\mathbb{P}(\tau_{\gamma_j} = t) \\
&= \int_0^\infty \gamma e^{-(\gamma+\delta)t} \left(1 - \int_{A(t)} p(0,t,0,x)dx\right) dt \\
&= \frac{\gamma}{\delta+\gamma} - \int_0^\infty \int_{A(t)} \gamma e^{-(\gamma+\delta)t} p(0,t0,x)dx\, dt\,. \qquad (4.2)
\end{aligned}
$$

Combining (4.1) and (4.2), the expected number of new infections is

$$
\begin{aligned}
\mathbb{E}[I] &= \mathbb{E}[Z]\mathbb{E}[X_j] \\
&= \frac{\beta\gamma}{\eta(\gamma+\delta)} - \int_0^\infty \int_{A(t)} \frac{\beta\gamma}{\eta} e^{-(\gamma+\delta)t} p(0,t,0,x)dx\, dt \qquad (4.3)
\end{aligned}
$$

from which it is clear that in the absence of control, $A(t) \doteq \{\emptyset\}$, the expected number of infection progeny, often called R_0 is given by

$$
R_0 = \mathbb{E}[I] = \frac{\beta\gamma}{\eta(\gamma+\delta)}\,.
$$

Then given a control $A(t)$, the reduction to R_0 achieved by the control is given by

$$
\mathbb{E}[I] = R_0 \; - \int_0^\infty \int_{A(t)} \frac{\beta\gamma}{\eta} e^{-(\gamma+\delta)t} p(0,t,0,x)dx\, dt\,. \qquad (4.4)
$$

The objective of the cell protection immune response to slow the spread of infection will be quantified as a target reduction to the expected number of infection progeny (4.4). An optimal cell protection control will be defined such that it contributes a given reduction to the expected number of infection progeny, while minimizing the area of protection used. The control problem is framed in a more general setting in the following section.

4.4 Optimal control problem

The control gives the region of protection $A(t)$, a subset of the domain D, at each time t from initial infection and is therefore a set valued function, whereas much theory is primarily developed for controls measurable functions taking values in $\mathbb{R}^d$. Directly studying how the region of protection should evolve in time is difficult, instead note that for a control to be feasible for (4.4) requires that $\bigcup_{t\geq 0} A(t)$ be Lebesgue measurable, therefore the feasible controls $\mathscr{A}$ are contained in the Lebesgue measurable subsets of the time and space domain, that is for

$T_D = [0, \infty[\times D$,

$$\mathscr{A} \ \subset \ \mathcal{M}(T_D)\,.$$

For a control $A \in \mathscr{A}$ the region of protection can be recovered for all $(t \in [0, \infty[)$ by

$$A(t) \ \doteq \ A \bigcap \left(t \times \mathbb{R}^d \right).$$

Let $g : T_D :\to [0, \infty[$ a function defined by the infection model, with

$$g \in L^1(T_D, [0, \infty[) \ := \ \{ f : T_D \to [0, \infty[\mid \int_{T_D} f < \infty \}\,,$$

and let $(V \in \mathbb{R})$ a given target value satisfying $0 \leq V \leq \int_{T_D} g \; dz$. Controls are considered which satisfy a constraint

$$\int_A g \; dz \ = \ V\,.$$

For example from (4.4), given a target $\bar{R}$ for the expected number of infection progeny $\mathbb{E}[I]$,

$$g(x,t) \doteq \frac{\beta\gamma}{\eta} e^{-(\delta+\gamma)t} p(0,t,0,x) \qquad \text{and} \qquad V = R_0 - \bar{R}\,. \tag{4.5}$$

The set of feasible controls is then

$$\mathscr{A} \doteq \left\{ A \in \mathcal{M}(T_D) \ : \ \int_A g \; dz \ = \ V \right\}.$$

The goal is to find an control $A \in \mathscr{A}$ which minimizes a measure of its area. More precisely, let $h \in \mathbf{L}^1(T_D,]0, +\infty[)$ be the cost for including each point of the domain in the control set. The optimal control problem studied here is

$$\min_{A \in \mathscr{A}} \int_A h \; dz\,. \tag{$\mathbf{P}$}$$

Within the context of the infection control, h measures the relative penalty for protecting cells within a given region or of a given type.

4.4.1 A characterization of optimal controls

An existence result for the optimal control problem $(\mathbf{P})$ and characterization of the solutions can be found by following the intuition that the set of least cost that achieves the target value, includes the points of the domain which have the greatest ratio of value to cost. Assume that the function $\frac{g}{h}$ is integrable over the domain T_D, i.e. $\frac{g}{h} \in \mathbf{L}^1(T_D, [0, \infty[)$. For any $r > 0$, let L_r be the

corresponding upper level set of the ratio $\frac{g}{h}$, that is

$$L_r \doteq \left\{ z \in T_D \ : \ \frac{g(z)}{h(z)} \geq r \right\}. \tag{4.6}$$

Let the function $F :]0, \infty[\rightarrow]0, \infty[$ be a measure of the sets L_r with respect to g, given by

$$F(r) \ \doteq \ \int_{L_r} g(z) \, dz \, . \tag{4.7}$$

Since g and $\frac{g}{h}$ are in $\mathbf{L}^1(T_D,]0, +\infty[)$, one has that F is upper semi-continuous and monotone decreasing with

$$F(0) \ = \ M \qquad \text{and} \qquad \lim_{r \rightarrow \infty} F(r) \ = \ 0.$$

In particular, the sublevel set of F

$$F^{-1}([V, +\infty[) = \{ r \in [0, \infty[\ : \ F(r) \geq V \}$$

is closed and bounded in $\mathbb{R}$. Hence, the following is well defined

$$\bar{r} \ \doteq \ \max \{ r \in [0, \infty[\ : \ F(r) \geq V \} \ < \ + \infty \tag{4.8}$$

The main result is stated as follows:

Theorem 4.4.1. *Under the given assumptions on g and h, the optimization problem $(\mathbf{P})$ admits at least one solution. Moreover, $A \in \mathscr{A}$ is an optimal solution of $(\mathbf{P})$ if and only if the Lebesgue measure of each of the sets, $A \bigcap L_{\bar{r}}^c$ and $A^c \bigcap \left(\bigcup_{r > \bar{r}} \ L_r \right)$, is zero.*

Proof. For simplicity, denote the Lebesgue measure of a set $A \in \mathcal{M}(T_D)$ by $\mu(A)$. The proof is divided into two main steps:

Step 1. Assume that A is a solution of $(\mathbf{P})$, then the claim is that

$$\mu \left(A \bigcap L_{\bar{r}}^c \right) \ = \ 0 \qquad \text{and} \qquad \mu \left(A^c \bigcap \left(\bigcup_{r > \bar{r}} \ L_r \right) \right) \ = \ 0 \, . \tag{4.9}$$

That the first equality of (4.9) holds will be shown first. Assume for contradiction that

$$\mu \left(A \bigcap L_{\bar{r}}^c \right) > 0$$

an admissible set $\tilde{A} \in \mathscr{A}$ will be constructed such that

$$\int_{\tilde{A}} h \, dz \ < \ \int_A h \, dz \, , \tag{4.10}$$

and therefore the set A can not be a minimizer of **(P)**.

From (4.7)-(4.8), one has

$$\int_A g\, d\mu \;=\; V \;\leq\; F(\bar{r}) \;=\; \int_{L_{\bar{r}}} g\, d\mu,$$

which implies

$$\int_{A\cap L^c_{\bar{r}}} g\, d\mu - \int_{A^c\cap L_{\bar{r}}} g\, d\mu \;=\; \int_A g\, d\mu - \int_{L_{\bar{r}}} g\, d\mu \;\leq\; 0.$$

By the continuity of Lebesgue integration, there exists a Lebesgue measurable set $H \subset A^c \cap L_{\bar{r}}$ such that

$$\int_H g\, d\mu \;=\; \int_{A\cap L^c_{\bar{r}}} g\, d\mu\,.$$

Recalling (4.6), note that

$$\begin{cases} g(x) & \geq\; \bar{r}\cdot h(x) \qquad a.e.\ x\in H \\ g(x) & <\; \bar{r}\cdot h(x) \qquad a.e.\ x\in A\cap L^c_{\bar{r}}\,, \end{cases}$$

whence

$$\bar{r}\int_H h\, d\mu \;\leq\; \int_H g\, d\mu \;=\; \int_{A\cap L^c_{\bar{r}}} g\, d\mu \;<\; \bar{r}\int_{A\cap L^c_{\bar{r}}} h\, d\mu\,.$$

Thus, the admissible set $\tilde{A} = (A\cap L_{\bar{r}})\cup H$ satisfies (4.10).

To complete this step, it will be shown also that the second equality of (4.9) holds. Assume for contradiction that there exists $r > \bar{r}$ such that $\mu\,(A^c\cap L_r) > 0$. Again an admissible set $\tilde{A}$ will be constructed such that

$$\int_{\tilde{A}} h\, dz \;<\; \int_A h\, dz\,.$$

From (4.7)-(4.8), it follows that

$$\int_A g\, d\mu \;=\; V \;>\; F(r) \;=\; \int_{L_r} g\, d\mu,$$

therefore also

$$\int_{A\cap L^c_r} g\, d\mu - \int_{A^c\cap L_r} g \;=\; \int_A g\, d\mu - \int_{L_r} g\, d\mu \;>\; 0\,.$$

By the continuity of Lebesgue integration there exists $H\subset (A\cap L^c_r)$ such that

$$\int_H g = \int_{A^c\cap L_r} g\,.$$

Again recalling (4.6),

$$\begin{cases} g(x) & < \; r \cdot h(x) & a.e.\; x \in H \\ g(x) & \geq \; r \cdot h(x) & a.e.\; x \in A^c \cap L_r \,, \end{cases}$$

it follows that

$$r \int_H h \; d\mu \; > \; \int_H g \; d\mu \; = \; \int_{A^c \cap L_r} g \; d\mu \; \geq \; r \int_{A^c \cap L_r} h \; d\mu \,.$$

Therefore, the admissible set $\tilde{A} = (A \backslash H) \cup (A^c \cap L_r)$ satisfies (4.10). This concludes step one.

Step 2. Assume that A is an admissible set that satisfies (4.9). The claim is that A is an optimal solution of (P). It will be shown that for any other admissible set B satisfying (4.9)

$$\int_A h = \int_B h \,,$$

therefore each must be an optimal solution.

From (4.9), there exists subsets E_A and E_B of $\{z \in D : \frac{g(z)}{h(z)} = \bar{r}\}$, such that

$$\int_A h \; d\mu \; = \; \int_{(\bigcup_{r \geq \bar{r}} L_r) \bigcup E_A} h \; d\mu \qquad \text{and} \qquad \int_B h \; d\mu \; = \; \int_{(\bigcup_{r \geq \bar{r}} L_r) \bigcup E_B} h \; d\mu \,.$$

Therefore,

$$\int_A h \; d\mu - \int_B h \; d\mu = \int_{E_A} h \; d\mu - \int_{E_B} h \; d\mu = \int_{E_A} \bar{r} \cdot g \; d\mu - \int_{E_B} \bar{r} \cdot g \; d\mu = \bar{r} \left(\int_A g \; d\mu - \int_B g \; d\mu \right) = 0$$

which concludes step two.

A choice of an admissible set which satisfies (4.9) must exist by the continuity of Lebesgue integration, which completes the proof. □

From Theorem 4.4.1, the case where the solution to $(\mathbf{P})$ is unique follows.

Corollary 4.4.2. *If $F(\bar{r}) = V$ then $L_{\bar{r}}$ is the unique, up to Lesbegue measure zero, solution of $(\mathbf{P})$.*

Proof. It will be shown that any set A that is an optimal solution of $(\mathbf{P})$ must satisfy

$$\mu \left(A \bigcap L_{\bar{r}}^c \right) \; = \; 0 \qquad \text{and} \qquad \mu \left(A^c \bigcap L_{\bar{r}} \right) \; = \; 0 \,, \tag{4.11}$$

and is therefore equal to $L_{\bar{r}}$ up to measure zero.

The first equality of (4.11) follows immediately from Theorem 4.4.1 and implies

$$0 \; = \; \int_A g \; d\mu - F(\bar{r}) \; = \; \int_A g \; d\mu - \int_{L_{\bar{r}}} g \; d\mu \; = \; \int_{A^c \cap L_{\bar{r}}} g \; d\mu \,.$$

Recalling (4.6), $g > 0$ on $L_{\bar{r}}$. Therefore, $\mu\left(A^c \bigcap L_{\bar{r}}\right) = 0$ which completes the proof. □

Note that in general $F(\bar{r}) > V$ since F is upper semi-continuous rather than continuous. In which case $(\mathbf{P})$ admits many solutions, though all are characterized by Theorem 4.4.1.

4.5 Biologically Feasible Control

The conditions for an area of cell protection given by Theorem 4.4.1 to be optimal for achieving a given reduction in the initial expected number of new infections, suggest a potentially biologically feasible feedback control.

If the cost of cell protection to the host is the same at all points in the domain around an initial infection, then in the framework of problem $(\mathbf{P})$, h is identically one and g may be defined as in (4.5). Theorem 4.4.1 then states that the optimal region of cell protection follows an upper level set of g, where g is directly proportional to the viral density on the domain. Therefore, a cell autonomous feedback control in which an antiviral state is initiated in each cell whenever the viral density surpasses a threshold in the neighborhood of the cell can produce an optimal protection control with respect to the framework presented here. If the cost of protection varies over the domain, then an optimal strategy will vary the viral density threshold required to initiate protection for each cell in proportion to its cost of protection.

In hosts, anti-viral states in cells are induced by the inter-cellular signaling molecule Interferon-β. This framework suggests that the role of Interferon-β is to provide an approximation of the viral distribution, and scale the approximation such that protection is initiated at an appropriate threshold viral density for a particular infection, perhaps adaptively as the progression of infection reveals its virility.

If the viral density is nearly homogenous in a large region above the optimal threshold, a threshold switch will respond too aggressively and fail to be optimal. This is the case of Theorem 4.4.1 in which the solution is not unique, as there is freedom of choice in the viral density level set, any subset choice of sufficient area produces an equivalent solution. Rand et al. found cell intrinsic stochastic initiation of an antiviral state in response to homogeneous Interferon-β exposure [68]. Given the homogeneity of input in the non-unique case, a stochastic response may best construct an optimal subset selection, and prevent over response. They also report a higher proportion of cells enter into an antiviral state as the concentration of IFN-β increases. It many be that such a response can approximate an optimal control in the setting studied here.

Several studies have observed stochastic initiation of the release of Interferon-β by cells upon infection [64, 68, 93]. Stochastic initiation of production and release may allow the Interferon-β distribution to approximate the viral distribution while scaling the Interferon-β concentration such that antiviral states are initiated at an appropriate viral threshold for the particular infecting

virus. Exploring additional cost functionals and control frameworks may provide further insights into how observed patterns of immune responses contribute to control of infections within hosts.

REFERENCES

[1] Abul Abbas, Andrew H. Lichtman, and Shiv Pillai. *Cellular and Molecular Immunology*. Elsevier, 9 edition, 2017.

[2] Mohammad Abdollahpouri, Gergely Takács, and Boris Rohal Ilkiv. Real-time moving horizon estimation for a vibrating active cantilever. *Mechanical Systems and Signal Processing*, 86:1–15, 2017.

[3] A. Alessandri and M. Gaggero. Fast moving horizon state estimation for discrete-time systems using single and multi iteration descent methods. *IEEE Transactions on Automatic Control*, 62(9):4499–4511, Sep. 2017.

[4] Angelo Alessandri, Marco Baglietto, and Giorgio Battistelli. Receding-horizon estimation for discrete-time linear systems. *IEEE Transactions on Automatic Control*, 48(3):473–478, 2003.

[5] Angelo Alessandri, Marco Baglietto, and Giorgio Battistelli. Moving-horizon state estimation for nonlinear discrete-time systems: New stability results and approximation schemes. *Automatica*, 44(7):1753 – 1765, 2008.

[6] Angelo Alessandri, Marco Baglietto, Giorgio Battistelli, and Victor Zavala. Advances in moving horizon estimation for nonlinear systems. *49th IEEE Conference on Decision and Control (CDC)*, pages 5681–5688, 2010.

[7] Moayed Almobaied, Ibrahim Eksin, and Mujde Guzelkaya. Inverse optimal controller based on extended kalman filter for discrete-time nonlinear systems. *Optimal Control Applications and Methods*, 39(1):19–34, 2018.

[8] B.D.O. Anderson and J.B. Moore. *Optimal Control:Linear Quadratic Methods*. Prentice-Hall, Englewook Cliffs, NJ, 1990.

[9] Panos J. Antsaklis and Anthony N. Michel. *Linear Systems*. Birkhauser, 2006.

[10] Andrea Arnold, Daniela Calvetti, and Erkki Somersalo. Parameter estimation for stiff deterministic dynamical systems via ensemble kalman filter. *Inverse Problems*, 30(10):105008, 2014.

[11] Zvi Artstein. Stabilization with relaxed controls. *Nonlinear Analysis: Theory, Methods & Applications*, 7:1163–1173, 1983.

[12] Combettes Patrick L. Bauschke, Heinz H. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

[13] Sven Beeler, Hien Tran, and H Banks. Feedback control methodologies for nonlinear systems. *Journal of Optimization Theory and Applications*, 107, 2000.

[14] Amanda Bernstein. Modeling and control: Applications to a double inverted pendulum and radio frequency interference. *Phd Thesis, North Carolina State University, Raleigh, NC*, 2018.

[15] Amanda Bernstein and Hien T. Tran. Real-time implementation of a lqr-based controller for the stabilization of a double inverted pendulum. In *Proceedings of The International MultiConference of Engineers and Computer Scientists*, pages 245–250, 2017.

[16] A. Bogdanov. Optimal control of double inverted pendulum on a cart. *Oregon Health and Science University, Tech. Rep. CSE-04-006, OGI School of Science and Engineering, Beaverton, OR*, 2004.

[17] R. G. Brown and P. Y. C. Hwang. *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, Inc., 2nd edition edition, 1992.

[18] Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble kalman filter. *Monthly Weather Review*, 126(6):1719–1724, 1998.

[19] Dong H. Chyung. State variable reconstruction. *International Journal of Control*, 39(5):955–963, 1984.

[20] G. Ciccarella, M. Dalla Mora, and A. Germani. A Luenberger-like observer for nonlinear systems. *International Journal of Control*, 57(3):537–556, 1993.

[21] Patrick Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 49, 2009.

[22] Ramon Delgado and Graham C. Goodwin. A combined map and bayesian scheme for finite data and/or moving horizon estimation. *Automatica*, 50, 2014.

[23] Mustafa Demirci. Design of feedback controllers for linear system with applications to control of a double-inverted pendulum. *International Journal of Computational Cognition*, 2(1):65–84, 2004.

[24] Karen A. Duca, Vy Lam, Iris Keren, Elizabeth E. Endler, Geoffrey J. Letchworth, Isabel S. Novella, and John Yin. Quantifying viral propagation in vitro: Toward a method for characterization of complex phenotypes. *Biotechnology Progress*, 17(6):1156–1165, 2001.

[25] Jonathan Eckstein and Dimitri P. Bertsekas. On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[26] Adil El-Alami Laaroussi, Mostafa Rachik, and Mohamed Elhia. An optimal control problem for a spatiotemporal SIR model. *International Journal of Dynamics and Control*, 6(1):384–397, 2018.

[27] Carole R. Baskin et al. Early and sustained innate immune response defines pathology and death in nonhuman primates infected by highly pathogenic influenza virus. *PNAS*, 106:3455–3460, 2009.

[28] Elaine A. Ferguson et al. Heterogeneity in the spread and control of infectious disease: Consequences for the elimination of canine rabies. *Scientific Reports*, 5:1–13, 2015.

[29] Francois Auger et al. Industrial applications of the kalman filter: A review. *IEEE Transactions On Industrial Electronics*, 60(12), 2013.

[30] Jan Busch et al. State estimation for large-scale wastewater treatment plants. *Water Research*, 47, 2013.

[31] G. Evensen. The ensemble kalman filter for combined state and parameter estimation. *IEEE Control Systems Magazine*, 29(3):83–104, 2009.

[32] Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5):10143–10162, 1994.

[33] R. Freeman and P. Kokotovic. Inverse optimality in robust stabilization. *SIAM Journal on Control and Optimization*, 34(4):1365–1391, 1996.

[34] Georg A. Funk, Vincent A. A. Jansen, Sebastian Bonhoeffer, and Timothy Killingback. Spatial models of virus-immune dynamics. *Journal of Theoretical Biology*, 233(2):221–236, 2005.

[35] W. L. Garrard. Suboptimal feedback control for nonlinear systems. *Automatica*, 8(2):219–221, 1972.

[36] Knut Graichen, Michael Treuer, and Michael Zeitz. Swing-up of the double pendulum on a cart by feedforward and feedback control with experimental validation. *Automatica*, 43(1):63–71, 2007.

[37] Eldad Haber, Felix Lucka, and Lars Ruthotto. Never look back - a modified EnKF method and its application to the training of neural networks without back propagation. *arXiv:1805.08034v2*, 2018.

[38] Eric L. Haseltine and James B. Rawlings. Critical evaluation of extended kalman filtering and moving-horizon estimation. *Industrial & Engineering Chemistry Research*, 44(8):2451–2460, 2005.

[39] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[40] Hans Heinrich Hoffmann, William M. Schneider, and Charles M. Rice. Interferons and viruses: An evolutionary arms race of molecular interactions. *Trends in Immunology*, 36(3):124–138, 2015.

[41] Tom J. Howat, Cristina Barreca, Peter O'Hare, Julia R. Gog, and Bryan T. Grenfell. Modelling dynamics of the type I interferon response to *in vitro* viral infection. *Journal of The Royal Society Interface*, 3(10):699–709, 2006.

[42] Marco A. Iglesias, Kody J. H. Law, and Andrew M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 2013.

[43] Alberto Isidori. *Nonlinear Control Systems: An Introduction*. Springer-Verlag, 2nd edition, 1989.

[44] Lionel B. Ivashkiv and Laura T. Donlin. Regulation of type I interferon responses. *Nature Reviews Immunology*, 14(1):36–49, 2015.

[45] Akiko Iwasaki and Ruslan Medzhitov. Control of adaptive immunity by the innate immune system. *Nature Immunology*, 16(4):343–353, 2015.

[46] Mrdjan Jankovic, Rodolphe Sepulchre, and Petar V. Kokotovic. CLF based designs with robustness to dynamic input uncertainties. *Systems & Control Letters*, 37:45–54, 1999.

[47] Andrew H. Jazwinski. Limited memory optimal filtering. *IEEE Transactions On Automatic Control*, 1968.

[48] Simon J. Julier and Jeffery K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3), 2004.

[49] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Transactions of the ASME. Series D, Journal of basic engineering*, 1961.

[50] Nikola Borislavov Kovachki and Andrew M Stuart. Ensemble Kalman inversion: A derivative-free technique for machine learning tasks. *Inverse Problems*, 2019.

[51] Peter Kühl, Moritz Diehl, Tom Kraus, Johannes P. Schlöder, and Hans Bock. A real-time algorithm for moving horizon state and parameter estimation. *Computers & Chemical Engineering*, 35:71–83, 2011.

[52] Mauricio Labadie and Anna Marciniak-czochra. A reaction-diffusion model for viral infection and immune response. *hal-00546034*, 2010.

[53] Tine Lefebvre, Herman Bruyninckx, and Joris De Schutter. Comment on "a new method for the nonlinear transformation of means and covariances in filters and estimators". *IEEE Transaction On Automatic Control*, 47(8), 2002.

[54] Frank L. Lewis, Draguna L. Vrabie, and Vassilis L. Syrmos. *Optimal Control*. John Wiley & Sons, 2012.

[55] Tim Maiwald, Annette Schneider, Hauke Busch, Sven Sahle, Norbert Gretz, Thomas S. Weiss, Ursula Kummer, and Ursula Klingmüller. Combining theoretical analysis and experimental data generation reveals IRF9 as a crucial factor for accelerating interferon-α induced early antiviral signalling. *FEBS Journal*, 277(22):4741–4754, 2010.

[56] Finlay McNab, Katrin Mayer-Barber, Alan Sher, Andreas Wack, and Anne O'Garra. Type I interferons in infectious disease. *Nature Reviews Immunology*, 15(2):87–103, 2015.

[57] Alexander Medvedev and Hannu Toivonen. Continuous-time deadbeat observation problem with application to predictive control of systems with delay. *Kybernetica*, 30(6):669–688, 1994.

[58] Hannah Michalska and D.Q. Mayne. Moving horizon observers and observer-based control. *IEEE Transactions on Automatic Control*, 40:995 – 1006, 1995.

[59] P. E. Moraal and J. W. Grizzle. Observer design for nonlinear systems with discrete-time measurements. *IEEE Transactions on Automatic Control*, 40(3):395–404, 1995.

[60] Bruno Morabito, Markus Kögel, Eric Bullinger, Gabriele Pannocchia, and Rolf Findeisen. Simple and efficient moving horizon estimation based on the fast gradient method. *IFAC-PapersOnLine*, 48(23):428–433, 2015.

[61] Matthias A Müller. Nonlinear moving horizon estimation in the presence of bounded disturbances. *Automatica*, 79:306–314, 2017.

[62] F. Ornelas-Tellez, E. N. Sanchez, A. G. Loukianov, and E. M. Navarro-López. Speed-gradient inverse optimal control for discrete-time nonlinear systems. *Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference*, pages 290–295, 2011.

[63] Fernando Ornelas-Tellez, Edgar Sanchez, Alexander Loukianov, and J.J. Rico. Robust inverse optimal control for discrete-time nonlinear system stabilization. *European Journal of Control*, 20, 2013.

[64] Sonali Patil, Miguel Fribourg, Yongchao Ge, Mona Batish, Sanjay Tyagi, Fernand Hayot, and Stuart C Sealfon. Single-cell analysis shows that paracrine signaling by first responder cells shapes the interferon-beta response to viral infection. *Science signaling*, 8(363), 2015.

[65] Alan S. Perelson. Applications of optimal control theory to immunology. *Proceedings of the US.-Italy Seminar on Variable Structure Systems*, 1977.

[66] Rebecca A. Porritt and Paul J. Hertzog. Dynamic control of type I IFN signalling by an integrated network of negative regulators. *Trends in Immunology*, 36(3):150–160, 2015.

[67] Eduardo Ramirez-Llanos and Sonia Martinez. Distributed discrete-time optimization algorithms with applications to resource allocation in epidemics control. *Optimal Control Applications and Methods*, 39(1):160–180, 2018.

[68] Ulfert Rand, Melanie Rinas, Johannes Seh Werk, Gesa Nöhren, Melanie Linnes, Andrea Kröger, Michael Flossdort, Kristóf Kály-Kullai, Hansjörg Hauser, Thomas Höfer, and Mario Köster. Multi-layered stochasticity and paracrine signal propagation shape the type-l interferon response. *Molecular Systems Biology*, 8(584):1–13, 2012.

[69] Christopher V Rao, James B. Rawlings, and Jay H. Lee. Constrained linear state estimation - a moving horizon approach. *Automatica*, 37(10):1619–1628, 2001.

[70] RM May RM Anderson. *Infections diseases of humans: Dynamics and control*. Wiley, 1992.

[71] R Rockafellar. Monotone operators and the proximal point algorithm. *Siam Journal on Control and Optimization*, 14, 1976.

[72] Robert E Rowthorn, Ramanan Laxminarayan, and Christopher A Gilligan. Optimal control of epidemics in metapopulations. *Journal of the Royal Society Interface*, 6:1135–1144, 2009.

[73] R. Ruiz-Cruz, E. N. Sanchez, F. Ornelas-Tellez, A. G. Loukianov, and R. G. Harley. Particle swarm optimization for discrete-time inverse optimal control of a doubly fed induction generator. *IEEE Transactions on Cybernetics*, 43(6):1698–1709, 2013.

[74] Fuminori Saito, Toshio Fukuda, and Fumihito Arai. Swing and locomotion control for two-link brachiation robot. *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 719–724, 1993.

[75] Shun Sasagawa, Masahiro Shinya, and Kimitaka Nakazawa. Interjoint dynamic interaction during constrained human quiet standing examined by induced acceleration analysis. *Journal of Neurophysiology*, 111(2):313–322, 2014.

[76] Claudia Schillings and Andrew M. Stuart. Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55, 2016.

[77] Jia-ni Shen, Yi-jun He, Zi-feng Ma, Hong-bin Luo, and Zi-feng Zhang. Online state of charge estimation of lithium-ion batteries : A moving horizon estimation approach. *Chemical Engineering Science*, 154:42–53, 2016.

[78] D. Simon. Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory and Applications*, 2010.

[79] Jaroslaw Smieja, Mohammad Jamaluddin, Allan R. Brasier, and Marek Kimmel. Model-based analysis of interferon-β induced signaling pathway. *Bioinformatics*, 24(20):2363–2369, 2008.

[80] Eduardo D Sontag. A universal construction of artstein's theorem on nonlinear stabilization. *Systems & Control Letters*, 13(2):117–123, 1989.

[81] Mark W. Spong. The swing up control problem for the acrobot. *IEEE control systems*, 15(1):49–55, 1995.

[82] Peter Swerling. Modern state estimation methods from the viewpoint of the method of least squares. *IEEE Transactions On Automatic Control*, ac-16(6), 1971.

[83] S. Takashima. Control of gymnast on a high bar. In *Proceedings of the IEEE/RSJ International Workshop on Intelligent Robots and Systems*, pages 1424–1429, 1991.

[84] Osamu Takeuchi and Shizuo Akira. Innate immunity to virus infection. *Immunological Reviews*, 227:75–86, 2009.

[85] Jinying Tan, Ruangang Pan, Lei Qiao, Xiufen Zou, and Zishu Pan. Modeling and Dynamical Analysis of Virus-Triggered Innate Immune Signaling Pathways. *PLoS ONE*, 7(10), 2012.

[86] Hien T. Tran and Andrea M Arnold. Ensemble Kalman filtering for inverse optimal control. In *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists*, pages 526–530, Hong Kong, March 2018.

[87] Sridhar Ungarala. Computing arrival cost parameters in moving horizon estimation using sampling based filters. *Journal of Process Control*, 2009.

[88] Christopher V. Rao, James B. Rawlings, and D.Q. Mayne. Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations. *IEEE Transactions on Automatic Control*, 48:246 – 258, 2003.

[89] F. Vicentini. Stability analysis of evolved continuous time recurrent neural networks that balance a double inverted pendulum on a cart. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2689–2694, Aug 2007.

[90] D. A. Winter. Human balance and posture control during standing and walking. *Gait and Posture*, 3(4):193–214, 1995.

[91] Hamed Yarmand, Julie S. Ivy, Brian Denton, and Alun L. Lloyd. Optimal two-phase vaccine allocation to geographically different regions under uncertainty. *European Journal of Operational Research*, 233(1):208–219, 2014.

[92] Victor M. Zavala, Carl D. Laird, and Lorenz T. Biegler. A fast moving horizon estimation algorithm based on nonlinear programming sensitivity. *Journal of Process Control*, pages 876–884, 2008.

[93] Mingwei Zhao, Jiangwen Zhang, Hemali Phatnani, Stefanie Scheu, and Tom Maniatis. Stochastic expression of the interferon-β gene. *PLoS Biology*, 10(1), 2012.