Numerical method and application of optimal control problem

Thesis committee:
Prof. dr. Barry Koren, Eindhoven University of Technology
Prof. dr. Kees Vuik, Delft University of Technology
Prof. dr. Rob H. Bisseling, Utrecht University
Prof. dr. Daan Crommelin, CWI Amsterdam
Dr. Sevetlana Dubinkina, VU University Amsterdam

ISBN: 978-94-6458-019-8
Printed By: Ridderprint. *www.ridderprint.nl.*

Numerical method and application of optimal control problem

Numerieke methode en toepassing van optimaal regelprobleem

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, Prof. dr. H. R. B. M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
woensdag 19 januari 2022 des ochtend te 10.15 uur

door

Xin Liu

geboren op 18 September 1990 te Henan, China

Promotor: Prof. dr. ir. J. E. Frank

This thesis was accomplished with financial support from the CSC (China Scholarship Council).

Contents

Chapter 1

Introduction

Optimal control problems are a classical topic whose wide range of applicability continues to grow [12, 116, 76, 16, 10, 104, 21, 75, 125]. In economic markets, business people try to control the cost of products to maximize their benefits. For the issue of natural resources protection, governments have to make optimal policies on how to utilize natural resources to relieve the huge demand for consumption with increasing population, which is becoming very urgent [80, 47]. In robotics technology, engineers try to reduce the uncertainties arising from limited knowledge to direct robot behaviors [87, 13]. In multi-agent systems (e.g Cucker-Smale model), scientists try to find sparse controls to nudge the dynamics into a coordinated state [124, 7] when a system does not self-organise spontaneously. Even in weather prediction, when the weather models or the observations are imperfect, geoscientists apply optimal control techniques to improve state estimates through minimising the error covariance cost [94, 63]. However, in general, solving optimal control problems is a complex computation issue as many times there are no analytical solutions. In this thesis, we study a numerical algorithm to solve optimal control problems and apply it in two distinct applications: sparse control of Cucker-Smale dynamics and data assimilation via a particle filter.

Optimal control theory

An optimal control problem consists of a cost functional related to the state and control functions (of time) and a set of differential equations describing the controlled dynamics. And the optimal control problem is to find a best control functional to minimize the cost functional under the constrained dynamics.

Since the solutions to many optimal control problems cannot be found by analytical means, many numerical methods have been developed to solve general optimal control problems. The numerical methods are classified into two categories, which are direct methods and indirect methods. Direct methods use the discretize-then-optimize approach in which one discretizes the original optimal control problem to obtain a nonlinear discrete problem to be solved numerically by well-established optimization methods. Especially, either the state or the control, or both, will be approximated by using some appropriate functions such as piecewise linear functions.

Indirect methods (optimize-then-discretize) employ the calculus of variations to obtain

the first-order optimality conditions [104], [27], [85]. These involve applying Pontryagin's maximum principle (a necessary condition for optimality) or the Hamilton-Jacobi-Bellman equation (a sufficient and necessary condition). For the Pontryagin principle, this yields a two-point boundary-value problem which arises from the derivative of a Hamiltonian function. In the end, the resulting dynamical system becomes a Hamiltonian system, which includes a state equation and co-state equation, as well as an optimization problem or a constraint. The resulting system of forward-backward equations is subsequently discretized and solved.

There are many numerical algorithms to solve the two-point boundary-value problem. Among the algorithms, the forward-backward sweep (FBS) method is easy to be implemented and advantageous with respect to memory use. However, the forward-backward sweep method for the Pontryagin maximum principle is not generically convergent for nonlinear dynamics [82].

Cucker-Smale model

It is a common phenomenon that some birds such as starlings will aggregate into huge flocks with hundreds to thousands of individuals. After flying for a period, they will fly in the same direction even the initial state is chaotic. This phenomenon is called "flocking" in which birds (more generally "agents") are self-organised into an ordered behavior from a disordered state effected by their neighbours's motion. The flocking phenomenon is very common in nature, for example the shoaling behavior of fish [96], the swarming behavior of insects.

Recently, especially with the computer technology highly developed, many agent-based models are proposed to study the flocking behavior. In 2007, Cucker and Smale proposed a mathematical model named Cucker-Smale model (C-S model) [32, 33] which concentrates on the alignment of agents' velocities. The C-S model makes use of a communication weight that depends on the metric distance between agents. Cucker and Smale indicated that the model exhibits a kind of phase transition phenomenon between the local flocking and global flocking depending on the decay rate of the communication weight. The main result in [32] about the Cucker-Smale model can be summarized that under certain parameter conditions, the agents reach uniform velocity regardless of the initial conditions, whereas under other parameter conditions, the initial velocities and the positions of the flock have to satisfy certain compatible conditions so that all agents can converge to uniform velocity asymptotically.

In recent years, the Cucker-Smale model has attracted much attention as a toy model for attempts at influencing or controlling self-organization in complex systems. The paper [106] investigated the flocking behavior of an extended Cucker-Smale model with hierarchical leadership. The paper [56] provided a simple model for Cucker-Smale model and derived some conditions for reaching exponential flocking. In [91] the authors proposed an augmented Cucker-Smale model by introducing inter-agent bounding forces. The paper [107] proposed two Cucker-Smale models by introducing cohesive and repulsive forces. In [46] the authors studied sparse control in the Cucker-Smale model. Similiarly, [20] consider consensus stabilization for the Cucker-Smale model by two kinds of controls: feedback control and open-loop, sparse optimal control. The paper [14] studies different

variations of the feedback structure for consensus stabilization. In our thesis, we apply the regularized forward-backward sweep method to studying the optimal control for the Cucker-Smale model for different norms in the cost function.

In particular, sparse control is designed to model a minimal amount of intervention of an external policy maker. To get optimal sparse control for the Cucker-Smale model, the ℓ_1 norm is employed in the cost function to penalize the control. Applying the ℓ_1 norm to minimize the control was introduced in [31] which studied models of linear fuel consumption. The paper [117] studied the sparsity character with ℓ_1 norm in optimal control problems. Employing ℓ_1 norm to enforce the sparsity of the controls is studied in many other research works related to optimal control problems [24, 28, 29, 118]. However, the ℓ_1 norm is non-differentiable, which presents a challenge to finding the sparse solution.

Data assimilation

Data assimilation has been heavily developed in the field of numerical weather forecasting (among others). Meteorologists try to employ all the available measurement data from the atmosphere, e.g. temperature, pressure, velocity field observations, to attain a good estimate of current weather conditions, which are in turn used as initial conditions in weather prediction models. The initial step in weather forecasting is very crucial since most weather prediction problems are chaotic dynamics and they are very sensitive to the initial conditions.

In general, to estimate or predict the state of dynamical systems, we could just employ *state evolution models* described by some initial value problems. However, in many cases, the dynamic models are subject to structural errors from simplifications, assumptions, or contain unknown parameters (including the unknown initial conditions). Alternatively we could just use the *observations* which are usually made of a real-world system, to estimate or predict the whole state. However, in many situations, observations are sparse, incomplete (i.e. of lower dimension than the full state) and imperfect versions of reality. Hence, two different kinds of errors will be caused [94]: (i) "model error" which is the difference between the computational model and the real dynamic system. (ii) "measurement error" which is unavoidable in the observation measure process. Therefore, how to minimise the impact of the error to obtain a best estimation or prediction is the main purpose of data assimilation.

Data assimilation is a technique that combines dynamic models and observational data to obtain an optimal estimation or prediction of the real state of a system. Since it can improve the accuracy of the estimate of the state, data assimilation has been widely used in many areas [70, 45, 95, 99].

Data assimilation methods utilize a forecast (also known as the first guess, or background information) based on a dynamic model, then adjust the forecast value based on a set of observed data and estimated errors. The difference between the forecast state and observation is called innovation or departure. Data assimilation techniques are classified into two categories, [15]: (i) *sequential data assimilation*, which involves an analysis process through combining the dynamic model and observations at successive times and updating the estimation when new observations are obtained, for example the Kalman filter [121], extended Kalman filter, ensemble Kalman filter [43], and particle filters[23]. (ii) *Varia-*

tional data assimilation minimize the covariance error cost over a fixed time interval, for example three dimensional variational data assimilation (3D-Var) [30], four dimensional variational data assimilation (4D-Var) [63]. We will explain these in more detail in the next chapter.

In this thesis, in Chapter 2, first, we present some background material for the succeeding chapters. We start with basic theory about optimal control theory and symplectic integrators. Subsequently, we introduce the Cucker-Smale model in detail. Then, some basic data assimilation algorithms will be presented. In Chapter 3, we prove the convergence of a regularised forward-backward sweep method when discretized with a symplectic Runge-Kutta method. In Chapter 4, the augmented forward-backward sweep method is applied to the Cucker-Smale model. As is known, the Cucker-Smale model is convergent (i.e. the agent velocities converge to a uniform state) automatically under some parameter conditions. In this chapter, we focus on the situation that the model is not convergent, such that external forces, e.g. a control, are added in this model. We observe that the forward-backward iteration converges rapidly when the 2-norm is employed in the cost, which is convex and differentiable. However, the costs employing the 1-norm are more complex to solve, but can be alleviated to some degree using a further regularization. In Chapter 5, we construct a new particle filter for data assimilation. The particle filter combines an optimal control structure with a Wasserstein cost function. Hence, the augmented forward-backward sweep algorithm from Chapter 3 is adapted to solve the particle filter. The new particle filter is applied to estimate uncertainty due to noisy dynamics or noisy measurements. In Chapter 6, we will present the conclusion for this thesis.

Chapter 2

Preliminaries

In this chapter, some background knowledge is provided for the material in the following chapters. In section 2.1, we discuss optimal control theory. Since the Hamiltonian structure of the optimal control problem is relevant for its iterative solution, symplectic integrators are explained in section 2.2. In section 2.3, the basic details of the Cucker-Smale model are introduced. Finally, some background on data assimilation algorithms is presented in section 2.4.

2.1 Optimal control problem

The state of the continuous system to be controlled is described by a vector $x(t) : \mathcal{T} \to \mathcal{R}^d$, where $\mathcal{T} = [0, T]$ represents a time interval. The control function $u(t)$ is an element of the set of admissible controls $u \in \mathcal{U} \subset \mathbf{R}^m, m \leq d$ at time t. The motion of the system is described by a differential equation

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = \xi \tag{2.1.1}$$

where $\dot{x}(t)$ is a commonly used notation for $\frac{dx(t)}{dt}$, $f : \mathcal{R}^d \times \mathcal{U} \to \mathcal{R}^d$ is a given function, and $\xi \in \mathcal{R}^d$ is the initial value for the state.

Considering the set of admissible controls $\mathcal{U}$, we define the cost functional J: $\mathcal{U} \to \mathbf{R}$ by

$$J[u] = \Phi(x(T)) + \int_0^T h(x(t), u(t))dt \tag{2.1.2}$$

where $\Phi : \mathcal{R}^d \to \mathcal{R}$ is the cost at the terminal time and $h : \mathcal{R}^d \times \mathcal{U} \to \mathcal{R}$ is the running cost. Furthermore, the functions Φ and h are also assumed to be continuously differentiable. Since $x(t)$ is the unique trajectory driven by a given control u which satisfies the initial condition, the cost functional J depends on u. The optimal control problem is associated with finding a control through minimizing (or maximizing) the cost functional J.

The optimal control $u(t)$ may not exist, which means it may be impossible to find an admissible control and associated admissible trajectory in (2.1.1). In this thesis, we try to

find the optimal control rather than prove its existence. To ensure a unique state equation $x(t)$ given an admissible optimal control $u(t)$ we assume the function f to be continuous in the variables $x(t)$ and $u(t)$ and continuously differentiable with respect to x, i.e. the functions $f(x, u)$ and

$$f_x = \frac{\partial f}{\partial x}(x(t), u(t))$$

are continuous [19, 67].

Hence under the above assumptions, if we know the initial condition and the control trajectory $u(t)$ over the whole time interval $[0, T]$, then we can integrate the differential equation (2.1.1) to get the state trajectory $x(t)$.

Considering simple optimal control problem examples [19]:

Example 1. *A factory produces a good which could be sold or reinvested. At the very beginning, the factory productive capacity is $\xi > 0$. At time t, the producing good is $x(t)$. To maximize the sales of the good, one part of the good is sold with fixed price $P > 0$ and the rest of the good is reinvested. We introduce the fraction $u(t)$ as the share of good sold. Hence the cost function is*

$$\int_0^T u(t)x(t)Pdt$$

and the state dynamic is

$$\dot{x} = (1 - u)x, \quad x(0) = \xi.$$

Example 2. *Supposing we know the initial point $x(0) = a$ and the terminal time is T, we are trying to find a curve $x(t) : [0, T] \to \mathcal{R}$ for which length is minimal. The curve function satisfies*

$$\dot{x}(t) = u(t)$$

The length of the curve is calculated as

$$\int_0^T \sqrt{1 + u(t)^2}dt$$

In the following, to simplify the expressions, sometimes, we omit the time notation (t); thus, $x(t)$ will be written simply as x, $u(t)$ will simply written as u. We mainly talk about the case of *minimizing* the cost functional, which is similar to the maximum case. Combining (2.1.1) with (2.1.2), we restate the optimal control problem as :

$$\begin{cases} \min\limits_{u \in \mathcal{U}} \Phi(x(T)) + \displaystyle\int_0^T h(x(t), u(t))dt, \\ \text{subject to} \\ \dot{x}(t) = f(x(t), u(t)), \quad x(0) = \xi. \end{cases} \tag{2.1.3}$$

The admissible control $u^* \in \mathcal{U}$ is called an *optimal control* if it satisfies

$$J[u^*] \leq J[u], \quad \text{for all} \quad u \in \mathcal{U}$$

The associated x^* is called the *optimal trajectory* or the *optimal path*. The optimal control problem (2.1.3) is referred to as the *Bolza form*. If $\Phi = 0$ in the cost functional, we say the optimal control problem is in *Lagrange form*. We say the problem is in *Mayer form* if $h = 0$ in the cost functional.

For the optimal control problems, there are some more complex forms like adding some constraints on state or control [105, 76]. In this thesis, we are concerned with the basic optimal control problem in *Bolza form*, only involving (2.1.1)-(2.1.2) without extra constraints on the states or the control and with initial time and final time fixed.

2.1.1 Calculus of variations and Pontryagin's maximum principle

Optimal control theory follows from the calculus of variations [50, 52, 40] which is concerned with the optimization of functionals and is a tool to derive necessary conditions for the optimum. In this section, we will apply the calculus of variations on the cost function of (2.1.2) subject to equation (2.1.1). To this end, the problem can be reformulated as an unconstrained optimization problem by introducing the Lagrange multiplier function $\lambda(t) : \mathcal{T} \to \mathcal{R}^d$ and the Lagrangian functional:

$$\mathcal{L} = \Phi(x(T)) + \lambda_0^T(x(0) - \xi) + \int_0^T h(x(t), u(t)) + \lambda^T(t)(\dot{x}(t) - f(x(t), u(t)))\, dt. \quad (2.1.4)$$

The variation of (2.1.2) is given by taking independent variations in $\delta u, \delta x, \delta\lambda$, and

$$\begin{aligned}\delta\mathcal{L} = D^x\Phi \cdot \delta x|_T + \int_0^T & [D^x(h(x(t), u(t)) - \lambda^T D^x f(x(t), u(t))]\delta x \\ & + [D^u(h(x(t), u(t)) - \lambda^T D^u f(x(t), u(t))]\delta u \\ & + \lambda^T\delta\dot{x} + [\dot{x} - f(x(t), u(t))]^T\delta\lambda\, dt. \end{aligned} \quad (2.1.5)$$

The notation D^x, D^u stand for the derivative with respect to x and u, resp. Integrating by parts for $\int \lambda^T\delta\dot{x}dt$ yields

$$\begin{aligned}\delta\mathcal{L} = (D^x\Phi + \lambda) \cdot \delta x|_T + \int_0^T & [-\dot{\lambda} + D^x(h(x(t), u(t)) - \lambda^T D^x f(x(t), u(t))]\delta x \\ & + [D^u(h(x(t), u(t)) - \lambda^T D^u f(x(t), u(t))]\delta u \\ & + [\dot{x} - f(x(t), u(t))]^T\delta\lambda\, dt. \end{aligned} \quad (2.1.6)$$

At an extremum of $\mathcal{L}$ it holds that $\delta\mathcal{L} = 0$ for all independent variations δx, $\delta\lambda$, δu. The variational derivatives of the functional $\mathcal{L}$ with respect to the functions $x(t)$, $\lambda(t)$ and $u(t)$, denoted $\mathcal{L}_x$, $\mathcal{L}_\lambda$, $\mathcal{L}_u$, are defined with respect to the L^2 inner product. The first order necessary conditions for an optimum of (2.1.4) are given by the Euler-Lagrange equations ($\mathcal{L}_x \equiv \mathcal{L}_\lambda \equiv \mathcal{L}_u \equiv 0$):

We write $D^x f(x(t), u(t)) = f_x(x(t), u(t)), D^u f(x(t), u(t)) = f_u(x(t), u(t))$, the same with

$h(x(t), u(t))$. Hence the first order necessary conditions are

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = \xi \tag{2.1.7}$$

$$\dot{\lambda}(t) = -f_x(x(t), u(t))\lambda + h_x(x(t), u(t)), \quad \lambda(T) = -\Phi_x(x(T)), \tag{2.1.8}$$

$$0 = f_u(x, u)^T \lambda(t) - h_u(x, u). \tag{2.1.9}$$

The function $\lambda(t)$ is called the adjoint or costate variable and the equation (2.1.8) is called the adjoint equation (or the costate equation). If u is an optimal control in the interior of $\mathcal{U}$, then it satisfies (2.1.7)-(2.1.9). It is convenient to define the **Hamiltonian function**

$$H(x, \lambda, u) = \lambda^T(t) f(x, u) - h(x, u), \tag{2.1.10}$$

which combines the objective function and state equations much like a Lagrangian in a static optimization problem and the multiplier $\lambda(t)$ as the costate variable. The optimal control policy function $u^*(t)$ with the optimal trajectory of the state variable $x^*(t)$ and the adjoint variable $\lambda^*(t)$ is

$$H(x^*(t), u^*(t), \lambda^*(t)) \geq H(x(t), u(t), \lambda(t))$$

for all $u(t) \in \mathcal{U}$. With the Hamiltonian function, we rewrite first-order necessary conditions (2.1.7)–(2.1.9)

$$\dot{x}(t) = \partial H/\partial \lambda, \quad x(0) = \xi \tag{2.1.11}$$

$$\dot{\lambda}(t) = -\partial H/\partial x, \quad \lambda(T) = -\Phi_x(x(T)) \tag{2.1.12}$$

$$0 = H_u(x(t), \lambda(t), u(t)). \tag{2.1.13}$$

The triple (x^*, λ^*, u^*) is the local minimum of the cost functional J and u^* is the stationary point of the Hamiltonian function with x^* and λ^* at each time $t \in [0, T]$. Note that minimizing the objective functional J corresponds to maximizing the Hamiltonian with respect to u. The condition (2.1.13) above can be generalized to apply to controls $u(t)$ constrained to lie in $\mathcal{U}$ by replacing (2.1.13) with **Pontryagin's Maximum principle**

$$\dot{x}(t) = \partial H/\partial \lambda, \quad x(0) = \xi \tag{2.1.14}$$

$$\dot{\lambda}(t) = -\partial H/\partial x, \quad \lambda(T) = -\Phi_x(x(T)) \tag{2.1.15}$$

$$u^* = \arg\max_{u(t) \in \mathcal{U}} H(x, \lambda, u), \quad \forall t \in \mathcal{T}. \tag{2.1.16}$$

For general optimal control problems, Pontryagin's maximum principle gives necessary optimality conditions which are in the form of Hamiltonian differential equation. If the Hamiltonian function satisfies the concavity condition, then the maximum principle condition is also a sufficient condition. Pontryagin's maximum principle, which defines a two point boundary value problem, is very useful as it allows to find analytical solutions to special types of optimal control problems [9] and to define numerical algorithms to search for solutions in general cases. Since Pontryagin's maximum principle leads to a Hamiltonian system with a constraint or maximality condition on the control Hamiltonian, it is natural to consider symplectic methods for its numerical integration.

2.2 Variational integrators and symplectic integrators

Symplectic integrators are numerical schemes for Hamiltonian systems that preserve the symplectic property inherent in the solution operator of the Hamiltonian problems. They are widely used in nonlinear dynamics, molecular dynamics, discrete element methods, accelerator physics, plasma physics and quantum physics. Historically, symplectic integrators were firstly developed by De Vogelaere [36]. Then they were further developed by Ruth [98], Channell [25], Menyuk [84]. Meanwhile, Lasagni [72], Sanz-Serna [103] and Suris [109] showed that implicit Runge-Kutta methods are symplectic for an appropriate choice of parameters. Next, we will discuss definitions of symplecticity, symplectic Runge-Kutta method and symplectic partitioned Runge-Kutta method. We summarize a few results given in [59].

2.2.1 Symplectic maps

To describe the problem clearly, we introduce the conjugate variables $(p, q) \in \mathbf{R}^d \times \mathbf{R}^d$ consistent with most literature. Consider the following Hamiltonian system:

$$\begin{aligned} \frac{dp}{dt} &= -\frac{\partial H}{\partial q}, \\ \frac{dq}{dt} &= \frac{\partial H}{\partial p}, \end{aligned} \tag{2.2.1}$$

where H is the Hamiltonian function, which is sufficiently differentiable. If the total energy of a Hamiltonian problem is conserved, it will be satisfied

$$H(p(t), q(t)) = H(p(0), q(0)), \quad \text{for all time } t$$

along the exact solution of the problem. To see that this is so, take the derivative of the Hamiltonian function, we will have

$$\frac{d}{dt}H(p(t), q(t)) = \frac{\partial H}{\partial p}^T \dot{p} + \frac{\partial H}{\partial q}^T \dot{q} = \frac{\partial H}{\partial p}^T (-\frac{\partial H}{\partial q}) + \frac{\partial H}{\partial q}^T (-\frac{\partial H}{\partial p}) = 0.$$

Consider the parallelogram $\mathcal{P} \subset \mathcal{R}^{2d}$ spanned by the vectors $\xi = \binom{\xi^p}{\xi^q}$ and $\eta = \binom{\eta^p}{\eta^q}$ and the operator $\omega : \mathcal{R}^{2d} \times \mathcal{R}^{2d} \to \mathcal{R}$

$$\omega(\xi, \eta) = \xi^T \mathbf{J} \eta$$

where $\mathbf{J} = \left(\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right)$, I is the identity matrix of dimension d, $\xi^p, \xi^q, \eta^p, \eta^q \in \mathbf{R}^d$.

Definition 2.2.1. *If a linear map $A : \mathcal{R}^{2d} \to \mathcal{R}^{2d}$ satisfies $\omega(A\xi, A\eta) = \omega(\xi, \eta)$ for all $\xi, \eta \in \mathcal{R}^{2d}$, then the linear map A is called symplectic, equivalently $A^T \mathbf{J} A = \mathbf{J}$*

In the case $d = 1$, one can check that the expression $\omega(\xi, \eta)$ is equivalent to the area of the parallelogram spanned by ξ and η. Since ω is unchanged under symplectic A, we see that in the case $d = 1$, area is preserved under a symplectic transformation.

Definition 2.2.2. *A differentiable map $f : \mathcal{X} \to \mathbf{R}^d \times \mathbf{R}^d$ (where $\mathcal{X} \subset \mathbf{R}^d \times \mathbf{R}^d$ is an open set) is said to be a symplectic map if its Jacobian matrix $F = \frac{\partial f(p,q)}{\partial (p,q)}$ is symplectic at any point $(p, q) \in \mathcal{X}$, i.e.,*

$$F^T \mathbf{J} F = \mathbf{J}.$$

Definition 2.2.3. *We define the flow of the Hamiltonian system $\varphi_t : \mathcal{X} \to \mathbf{R}^d \times \mathbf{R}^d$ as*

$$(p(t), q(t)) = \varphi_t(p(0), q(0))$$

with the initial condition $(p(0), q(0))$, where $(p(t), q(t))$ is the solution of the Hamiltonian system.

Theorem 2.2.1. *(***Poincare 1899***) Let the Hamiltonian function $H(p, q)$ be twice continuously differentiable on $\mathcal{X} \subset \mathbf{R}^{2d}$. Then for each fixed t, the flow φ_t is a symplectic transformation wherever it is defined, i.e. the Jacobian $D\phi_t = \frac{\partial (p(t),q(t))}{\partial (p(0),q(0))}$ satisfies*

$$D\varphi_t^T \mathbf{J} D\varphi_t = \mathbf{J}$$

Considering a one-step numerical integrator, if τ denotes the step length and (p_n, q_n) denotes the numerical approximations at time $t_n = n\tau$ to $(p(t_n), q(t_n))$ of the solution of the (2.2.1)

$$(p_{n+1}, q_{n+1}) = \varphi_\tau(p_n, q_n),$$

where the transformation φ_τ is assumed to depend smoothly on τ and the Hamiltonian function H. Given an initial condition (p_0, q_0), the numerical approximation at time t_{n+1} is obtained by iterating the mapping φ_τ $n + 1$ times, which is

$$(p_{n+1}, q_{n+1}) = \varphi_\tau^{n+1}(p_0, q_0).$$

Definition 2.2.4. *A numerical one-step method $(p_{n+1}, q_{n+1}) = \varphi_\tau(p_n, q_n)$ is called symplectic if, when applied to a Hamiltonian system, the discrete flow $(p, q) \to \varphi_\tau(p, q)$ is a symplectic transformation for all sufficiently small step size τ.*

2.2.2 Symplectic Runge-Kutta methods

In the following section, we set $y = (p, q)^T$, then the equations (2.2.1) could be rewritten

$$\dot{y}(t) = \mathbf{J}^{-1} \nabla H =: f(y). \tag{2.2.2}$$

An s-stage Runge-Kutta method for (2.2.2) is given by the formulas

$$\begin{aligned} y_{n+1} &= y_n + \tau \sum_{i=1}^{s} b_i f(Y_i), \\ Y_i &= y_n + \tau \sum_{j=1}^{s} a_{ij} f(Y_j), \quad i = 1, \cdots, s \end{aligned} \tag{2.2.3}$$

where τ is the time interval, $b_i \geq 0, a_{ij}$ are real parameters, and $\sum_{i=1}^{s} b_i = 1$. Butcher [18] proposed a coefficient tableau to represent the Runge-Kutta method

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

where mostly (not all) $c_i = \sum_{j=1}^{s} a_{ij} (i = 1, \cdots, s)$. This kind of expression is often called the Butcher form. If $a_{ij} = 0$, $j \geq i$, such a Runge-Kutta method is called an explicit Runge-Kutta scheme, otherwise it is called an implicit Runge-Kutta scheme.

Definition 2.2.5. *A Runge-Kutta method is symplectic if the Jacobian matrix of its transformation* (2.2.3) *is symplectic, i.e.,* $\frac{\partial y_{n+1}}{\partial y_n}$ *is a symplectic map.*

To figure out the coefficient character of a symplectic Runge-Kutta method, we set $M = (m_{ij})_{i,j=1}^{s}$ to be the real $s \times s$ matrix given by

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j$$

for $i, j = 1, ..., s$. Lasagni [72], Sanz-Serna [103], Suris [109] and Hairer [59] showed that

Theorem 2.2.2. *If* $M = 0$*, the corresponding Runge-Kutta method is symplectic*

It is straightforward to see that explicit Runge-Kutta methods cannot satisfy the condition $M = 0$.

- For the backward Euler method,

  $$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

  $M = 1$, hence it is not symplectic.

- For the implicit midpoint rule,

  $$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

  $M = 0$, hence it is symplectic.

- For the 2-stage Gauss-Legendre Runge-Kutta method,

  $$\begin{array}{c|cc} 1/2 - \sqrt{3}/6 & 1/4 & 1/4 - \sqrt{3}/6 \\ 1/2 + \sqrt{3}/6 & 1/4 + \sqrt{3}/6 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}$$

  $M = 0$, hence, it is symplectic.

- For the 3-stage, fourth-order method of Lobatto IIIA methods,

  $$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

  $M \neq 0$ (for instance, $M_{11} = -1/36$), and the method is not symplectic.

- For the explicit 4-stage classical method,

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

  $M \neq 0$ (again, $M_{11} = -1/36$), and the method is not symplectic.

2.2.3 Symplectic partitioned Runge-Kutta method

In the differential equations, it is possible for us to integrate some components of the unknown vector with one Runge-Kutta method and integrate the remaining components with a different Runge-Kutta method, assuming the internal stages are collocated. For example, in the Hamiltonian system (2.2.1), the component p and the component q may be integrated by two different Runge-Kutta schemes, which is called partitioned-Runge-Kutta method. To be more specify, the partitioned Runge-Kutta with tableaux

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

and

$$\begin{array}{c|ccc} C_1 & A_{11} & \cdots & A_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ C_s & A_{s1} & \cdots & A_{ss} \\ \hline & B_1 & \cdots & B_s \end{array}$$

applied to the Hamiltonian system, becomes

$$p_{n+1} = p_n - \tau \sum_{i=1}^{s} b_i \frac{\partial H}{\partial q}(Q_i, P_i), \quad n = 1, \cdots, N \tag{2.2.4}$$

$$q_{n+1} = q_n + \tau \sum_{i=1}^{s} B_i \frac{\partial H}{\partial p}(Q_i, P_i), \quad n = 1, \cdots, N \tag{2.2.5}$$

$$P_i = p_n - \tau \sum_{j=1}^{s} a_{ij} \frac{\partial H}{\partial q}(Q_j, P_j), \quad i = 1, \cdots, s \tag{2.2.6}$$

$$Q_i = q_n + \tau \sum_{j=1}^{s} A_{ij} \frac{\partial H}{\partial p}(Q_j, P_j), \quad i = 1, \cdots, s \tag{2.2.7}$$

There is a similar result regarding symplecticity of partitioned Runge-Kutta schemes.

Theorem 2.2.3. *If the coefficients of a partitioned Runge-Kutta method satisfy the following condition:*

$$b_i A_{ij} + B_j a_{ji} - b_i B_j = 0, \quad i, j = 1, \cdots s \tag{2.2.8}$$

then the partitioned Runge-Kutta method is symplectic [1, 110, 51, 100].

Partitioned Runge-Kutta methods are of particular interest because they may be made explicit in some case, e.g., when the Hamiltonian is separable $H(p,q) = H_1(p) + H_2(q)$. Also, they arise naturally from discrete variational principles, as we illustrate in Chapter 3. Specifically, for any Runge-Kutta method, there exists a symplectic conjugate Runge-Kutta method such that the pair constitutes a partitioned Runge-Kutta method. However, the conjugate Runge-Kutta method may be of a different order of accuracy than the original one.

Next, we will present two simple symplectic partitioned Runge-Kutta algorithms.

There are two first order **symplectic Euler** methods that combine one explicit Euler method and one implicit Euler method, specifically as follows (implicit in p)

$$\begin{aligned} p_{n+1} &= p_n - \tau \frac{\partial H}{\partial q}(p_{n+1}, q_n) \\ q_{n+1} &= q_n + \tau \frac{\partial H}{\partial p}(p_{n+1}, q_n), \end{aligned}$$

or alternatively, (implicit in q)

$$\begin{aligned} p_{n+1} &= p_n - \tau \frac{\partial H}{\partial q}(p_n, q_{n+1}) \\ q_{n+1} &= q_n + \tau \frac{\partial H}{\partial p}(p_n, q_{n+1}). \end{aligned}$$

Störmer-Verlet methods are second order symplectic methods, which combine the trapezoidal method and the implicit midpoint method. The Butcher tableau for these methods is

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} 0 & 1/2 & 0 \\ 1 & 1/2 & 0 \\ \hline & 1/2 & 1/2 \end{array}$$

Some manipulation yields the method

$$\begin{aligned} p_{n+1/2} &= p_n - \tau/2 \frac{\partial H}{\partial q}(p_{n+1/2}, q_n) \\ q_{n+1} &= q_n + \tau/2 \left(\frac{\partial H}{\partial p}(p_{n+1/2}, q_n) + \frac{\partial H}{\partial p}(p_{n+1/2}, q_{n+1})\right) \\ p_{n+1} &= p_{n+1/2} - \tau/2 \frac{\partial H}{\partial q}(p_{n+1/2}, q_{n+1}), \end{aligned}$$

or, reversing the application to the variables,

$$\begin{aligned} q_{n+1/2} &= p_n + \tau/2 \frac{\partial H}{\partial p}(p_n, q_{n+1/2}) \\ p_{n+1} &= p_n - \tau/2 \left(\frac{\partial H}{\partial q}(p_n, q_{n+1/2}) + \frac{\partial H}{\partial q}(p_{n+1}, q_{n+1/2})\right) \\ q_{n+1} &= q_{n+1/2} + \tau/2 \frac{\partial H}{\partial p}(p_{n+1}, q_{n+1/2}). \end{aligned}$$

Comparing with non-symplectic integrators, symplectic integrators have some advantages as follows (see [86, 68]).

1. Symplectic integrators conserve the Hamiltonian (total energy) H to within bounded oscillations of amplitude proportional to τ^k, where k is the order of accuracy of the method. This property makes symplectic integrators advantageous for long integrations of Hamiltonian systems.
2. Symplectic integrators are volume preserving maps for Hamiltonian fields, making them effective for statistical mechanical calculations.

Neither of the above properties is relevant for the application of symplectic integrators to optimal control problems. However, in Chapter 3 we show that symplectic integrators have another advantageous property in that context.

2.3 Cucker-Smale model

In this section, we review background information about the Cucker-Smale model. The Cucker-Smale model describes how a group of agents interact when influenced by their neighbours. In general, it depends on a simple rule that the individuals tend to align their motion vector with that of their neighbours according to the distance between the agents. The closer they are, the more they are influenced by their neighbours. In the model, a system of M interacting agents is considered. For each agent, the state is described by a pair $(x_i(t), v_i(t))$ of vectors in $\mathbf{R}^d \times \mathbf{R}^d$, where $x_i(t)$ represents the position state of the i-th agent at time t and the $v_i(t)$ represents its velocity. Therefore the state of the group of M agents is given by $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_M(t))$ and the velocity of the group of M agents is $\mathbf{v}(t) = (v_1(t), \cdots, v_M(t))$. The space of the position states is $(\mathbf{R}^d)^M$, the same with the space of the velocity. The dynamics of the agents is governed by the Cucker-Smale equations

$$\begin{cases} \dot{x}_i(t) = v_i(t), i = 1, \cdots, M, \\ \dot{v}_i(t) = \sum_{j=1}^{M} \phi(\|x_j(t) - x_i(t)\|)(v_j(t) - v_i(t)), i = 1, \cdots, M, \\ \mathbf{x}(0) = \mathbf{x}_0, \quad \mathbf{v}(0) = \mathbf{v}_0, \end{cases} \tag{2.3.1}$$

where $\|\cdot\|$ denotes the ℓ_2-Euclidean norm,

$$\|x_i\| = \big(\sum_{j=1}^{d} (x_i^j)^2\big)^{1/2}, \quad \|\mathbf{x}\| = (\sum_{i=1}^{M} \|x_i\|^2)^{1/2}.$$

In the following, to express the problem easily, we write x and $x(t)$ interchangeably, the same with other functions related to time t. In the equation (2.3.1), the function $\phi \in C^1([0, +\infty])$, called the *influence function*, is non-increasing and positive. It is a function of the distance between agents. And it quantifies the weight with which agent i and agent j influence each other.

Typically, the function ϕ is given by [33, 57, 56].

$$\phi(\|x_i - x_j\|) = \frac{K}{(\sigma + \|x_i - x_j\|^2)^\beta}. \tag{2.3.2}$$

where $K > 0$, $\beta \geq 0$, $\sigma > 0$ are constants modelling the social properties of the group of agents.

Some other forms of ϕ are discussed in other papers. For example, [56] consider the function

$$\phi(\|x_i - x_j\|) = \frac{K}{\|x_i - x_j\|^\beta}.$$

However, for this case, the distance between two agents may not approach zero. In this thesis, we use the general form (2.3.2).

The question of interest is whether a group of agents converge to a common velocity vector, at which point they will move as a solid body. The mean state and mean velocity are

$$\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{2.3.3}$$

$$\bar{v} = \frac{1}{M}\sum_{i=1}^{M} v_i, \tag{2.3.4}$$

and the fluctuations are defined by

$$\breve{x}_i := x_i - \bar{x}, \quad \breve{v}_i := v_i - \bar{v}. \tag{2.3.5}$$

In the following, we will simply present some definitions and theorems [20, 32, 56, 55].

It is not difficult to find that the Cucker-Smale model is symmetric in the sense that the coefficient matrix ϕ_{ij} satisfies

$$\phi_{ij} = \phi_{ji},$$

namely, agent i and agent j have the same influence on the alignment of each other. The symmetry implies that the total momentum in the Cucker-Smale model is conserved,

$$\frac{d}{dt}\bar{v}(t) = 0$$

which means

$$\bar{v}(t) = \bar{v}(0).$$

To illustrate the problem easily, we define a symmetric bilinear form B on $(\mathbf{R}^d)^M \times (\mathbf{R}^d)^M$ by

$$B(\mathbf{w}, \mathbf{v}) = \frac{1}{2M^2}\sum_{i,j=1}^{M} \|w_i - v_j\|^2$$

for any $\mathbf{v}, \mathbf{w} \in (\mathbf{R}^d)^M$. In order to characterize consensus emergence in terms of the solution of the Cucker-Smale model (2.3.1), we define the following quantities

$$X(t) = B(\mathbf{x}(t), \mathbf{x}(t)), \quad V(t) = B(\mathbf{v}(t), \mathbf{v}(t)).$$

Definition 2.3.1. *(Consensus). The solution* $(\mathbf{x}(t), \mathbf{v}(t))$ *of the Cucker-Smale* (2.3.1) *is said to tend to consensus if the velocity vectors* $v_i(t)$ *tend to the mean velocity* $\bar{v}$*, i.e.,*

$$\lim_{t\to+\infty} \|v_i(t) - \bar{v}(t)\| = 0,$$

for every $i = 1, \cdots M$*, equivalently*

$$\lim_{t\to+\infty} V(t) = 0,$$

and the position fluctuations are uniformly bounded in time t

$$\sup_{0\le t<\infty} \sum_{i=1}^{M} \|x_i(t) - \bar{x}(t)\| < \infty,$$

for every $i = 1, \cdots M$*.*

Remark . *Because of uniqueness, a solution of* (2.3.1) *cannot reach consensus within finite time, unless the initial datum is already a consensus point.*

The following theorems are from the paper [56]

Theorem 2.3.1. *(Unconditional consensus emergence) Assume that the parameter* $0 \le \beta \le \frac{1}{2}$*. Let* $(\mathbf{x}, \mathbf{v}) \in \mathbf{R}^{dM} \times \mathbf{R}^{dM}$ *be the solution of the equation* (2.3.1)*, and let* $(\breve{\mathbf{x}}, \breve{\mathbf{v}}) \in \mathbf{R}^{dM} \times \mathbf{R}^{dM}$ *denote the fluctuations* (2.3.5)*, then there exist positive constants* x_{c1} *and* x_{c2} *independent of* t *satisfying*

$$x_{c1} \le \|\breve{\mathbf{x}}\| \le x_{c2}, \quad \|\breve{\mathbf{v}}\| \le \|\breve{\mathbf{v}}_0\| \exp^{-\phi(x_{c2})t},$$

i.e, the Cucker-Smale dynamic (2.3.1) *converges asymptotically to consensus.*

Theorem 2.3.2. *(Conditional flocking) Let* $(\mathbf{x}, \mathbf{v})$ *be a solution to* (2.3.1) *with* $\beta > \frac{1}{2}$*, supposing the initial configuration* $(\mathbf{x}_0, \mathbf{v}_0)$ *satisfies*

$$\frac{2\beta - 1}{K} \|\mathbf{v}_0\| < (1 + \|\mathbf{x}_0\|^2)^{\frac{1-2\beta}{2}}$$

or

$$\sqrt{V_0} \le \int_{\sqrt{X_0}}^{\infty} \phi(\sqrt{2M} r) dr.$$

Then there exist positive constants $\mathbf{x}_{c1}$ *and* $\mathbf{x}_{c2}$ *independent of* t *satisfying*

$$\mathbf{x}_{c1} \le \|\breve{\mathbf{x}}(t)\| \le \mathbf{x}_{c2}, \qquad \|\breve{\mathbf{v}}(t)\| \le \|\breve{\mathbf{v}}_0\| e^{-\phi(x_{c2})t},$$

i.e, the Cucker-Smale dynamic (2.3.1) *converges asymptotically to consensus.*

From the theorems, we find that when the initial fluctuation of the velocities is small enough and the initial positions of the agents are sufficiently close to consensus, the dynamics of the Cucker-Smale will tend to consensus exponentially.

Next we will give a simple example to express the theorems.

Two-agent Cucker-Smale model [20, 32] Two agents move on $\mathbf{R}$ with position and velocity at time t, $(x_1(t), v_1(t))$ and $(x_2(t), v_2(t))$, respectively. To simplify the problem,

we set the parameters $\beta = 1 > \frac{1}{2}$, $K = 2$, $\sigma = 1$, and let $x(t) = x_1(t) - x_2(t)$, $v(t) = v_1(t) - v_2(t)$, then the Cucker-Smale Equation (2.3.1) reduces to a simple formula

$$\begin{cases} \dot{x}(t) = v \\ \dot{v}(t) = -\dfrac{2v}{1+x^2} \\ v(0) > 0. \end{cases} \tag{2.3.6}$$

Putting the first equation into the second equation in (2.3.6), we have

$$\dot{v} = -\frac{2\dot{x}}{1+x^2},$$

and the solution is easily found to be

$$v(t) = -2\arctan x(t) + 2\arctan x(0) + v(0).$$

Analysing the solution, we found:

When the initial conditions satisfy $2\arctan x(0) + v(0) < \pi$, the relative main state $x(t)$ is globally bounded by $\frac{1}{2}\tan(2\arctan x(0) + v(0))$ which is sufficient for consensus.

When the initial conditions satisfy $2\arctan x(0) + v(0) = \pi$, then $v(t) = \pi - 2\arctan x(t)$, hence the system tends to consensus as well.

However, when $2\arctan x(0) + v(0) > \pi$, which means there exists $\epsilon > 0$ such that $|2\arctan x(0) + v(0)| \geq \pi + \epsilon$, then the consensus parameter $v(t)$ remains far away from 0 at every time, since

$$v(t) = -2\arctan x(t) + 2\arctan x(0) + v(0) \geq -2\arctan x(t) + \pi + \epsilon > \epsilon,$$

Namely, in this situation the uncontrolled solution does not tend to consensus. In Chapter 4 we consider the problem of reaching consensus via optimal control.

2.4 Basic algorithms for data assimilation

In this section, we review standard data assimilation algorithms. Since data assimilation is usually applied to discretized systems, we will use a formulation in terms of discrete maps (discrete time).

2.4.1 Variational data assimilation: 3D-Var and 4D-Var

Variational data assimilation techniques are based on minimising appropriate cost functions which are subject to model constraints. In general, there are two popular variational data assimilation algorithms, i.e. three dimensional variational data assimilation (3D-Var), four dimensional variational data assimilation (4D-Var).

3D-Var

3D-Var is a relatively simple variational data assimilation method. It generates a corrected state at each time, independent of data or solution at other times. 3D-Var proceeds under

the assumption that the model error (in each time step) is distributed as $\mathcal{N}(0, B)$ and the observation error is distributed as $\mathcal{N}(0, R)$. To obtain the optimum estimation of the state x at the current time, we try to minimise the cost function given by

$$J(x) = (x - x_b)^T B^{-1}(x - x_b) + (z - \mathcal{H}(x))^T R^{-1}(z - \mathcal{H}(x)),$$

where x_b is the background state obtained by propagating the model from the previous time, $\mathcal{H}$ is the observation operator(Notice: The $\mathcal{H}$ is different from the Hamiltonian function), z is the observation. The cost function includes two parts, the distance between the state x to the background state x_b and the distance between the model trajectory and the observations over the assimilation time window. To find the optimum state x^* which minimises the cost function J, we calculate the gradient of J, which is

$$\nabla J(x) = 2B^{-1}(x - x^b) - 2\mathcal{H}^T R^{-1}(z - \mathcal{H}(x)) \tag{2.4.1}$$

and choose x^* such that $\nabla J(x^*) = 0$. The 3D-Var method is limited by the use of only local information, as well as by its assumption of normally distributed errors.

4D-Var

4D-Var is a generalization of the 3D-Var that combines the observations with the time domain, which means that all observations obtained within a time window should be taken into account when we define the cost function. Usually, 4D-Var is a popular method to seek the best estimation of the initial value for the state so that the prediction is consistent with the observations within the assimilation interval $[t_0, t_N]$. The cost function is

$$J(x_0) = \frac{1}{2}(x_0 - x_0^b)^T B_0^{-1}(x_0 - x_0^b) + \frac{1}{2}\sum_{n=0}^{N}(z_n - \mathcal{H}(x_n))^T R_n^{-1}(z_n - \mathcal{H}(x_n)),$$

where x_0^b is the initial predict or background value at t_0, z_n is the observation at t_n . N is the time step number within $[t_0, t_N]$. B_0 is the background error covariance at time t_0 and the R_n is the observation error covariance at time t_n.

In general, the 4D-var depends on perfect model, which means we assume the model given by

$$x_{n+1} = f(x_n).$$

describes the system exactly over the assimilation period, thus we can neglect the model error. Minimising the cost function in 4D-Var is usually associated with adjoint variables λ_n, which is a little complex to compute.

4D-Var can be regard as optimal control problem with control x_0 and the Lagrangian formulation

$$L = \frac{1}{2}(x_0 - x_0^b)^T B_0^{-1}(x_0 - x_0^b) + \frac{1}{2}\sum_{n=0}^{N}(z_n - \mathcal{H}(x_n))^T R_n^{-1}(z_n - \mathcal{H}(x_n)) + \sum_{n=0}^{N-1}\lambda_{n+1}(x_{n+1} - f(x_n)).$$

The method in Chapter 3 of this thesis could be applied to this formulation, but has not yet been done so.

2.4.2 Sequential data assimilation

Sequential data assimilation algorithms are applied step-by-step as opposed to over a window as with 3D-Var.

2.4.3 Kalman filter

The Kalman filter [66] is one of the most important and commonly used estimation algorithms in data assimilation. It provides estimates for some unknown variables when observed measurements are given. The Kalman filter is formulated for a linear dynamical system and linear measurement process. The state evolution from time n to time $n+1$ is given as

$$x_n = Fx_{n-1} + W_{n-1}$$

where the state $x_n \in \mathbf{R}^d$ and F is the state transition matrix. The initial condition is x_0. W_n is the process noise (state noise), which is assumed to be drawn from a zero mean multivariate normal distribution with covariance Q, i.e. $W_n \sim \mathcal{N}(0, Q)$.
At time n an observation (or measurement) $z_n \in \mathbf{R}^m$ of the true state x_{n+1} is made according to

$$z_n = Hx_n + v_n$$

where z_n is the observation vector, H is the observation matrix (Notice the matrix H is different from Hamiltonian function), and v_n is the observation noise that is also assumed to be normally distributed with zero mean and covariance R, i.e. $v_n \sim \mathcal{N}(0, R)$.

In Algorithm 1 the Kalman filter iterates two steps: prediction (propagation) and update (correction). In the following, we will use the superscript b to denote the predicted (prior or background) estimate and superscript a to denote the update (posterior or "analysis") estimate. The Kalman filter propagates an estimate x_n of the mean state as well as an estimate P_n of the error covariance.

Algorithm 1 Kalman Filter

procedure INITIALIZATION
Set $x_0^a = x^0$ and $P_a^0 = B$, where B is the initial guess covariance matrix.
for $n = 1, \ldots, N-1$ **do**
procedure PREDICTION
Predicted state $x_n^b = Fx_{n-1}^a$,
Predicted error covariance $P_n^b = FP_{n-1}^a F^T + Q$,
procedure UPDATE
Measurement residual (innovation) $\tilde{y}_n = z_n - Hx_n^b$
Kalman gain: $K_n = P_n^b H^T (R + HP_n^b H^T)^{-1}$,
Updated state estimate $x_n^a = x_n^b + K_n \tilde{y}_n$,
Updated error covariance $P_n^a = (I - K_n H) P_n^b$

As mentioned, the Kalman filter is only suited to linear and Gaussian systems. Next we review other common algorithms, the extended Kalman filter and ensemble Kalman filter, which can be applied to nonlinear systems.

2.4.4 Extended Kalman filter

The extended Kalman filter [64, 49] is generalized to nonlinear dynamics, which means the discrete dynamics are

$$x_n = f(x_{n-1}) + W_{n-1}$$

where the map f is nonlinear. Its Jacobian is denoted Df. The noise is assumed the same as for the Kalman filter.

Algorithm 2 Extended Kalman Filter

procedure INITIALIZATION
Set $x_0^a = x_0$ and $P_0^a = B$, where B is the initial guess covariance matrix.
for $n = 1, \ldots, N-1$ **do**
procedure PREDICTION
Predicted state $x_n^b = f(x_{n-1}^a)$,
Predicted error covariance $P_n^b = Df_n P_{n-1}^a \, Df_n^T + Q$, where $Df_n = \frac{\partial f}{\partial x}|_{x_n^a}$
procedure UPDATE
Measurement residual (innovation): $\tilde{y}_n = z_n - H x_n^b$
Kalman gain: $K_n = P_n^b H^T (R + H P_n^b H^T)^{-1}$,
Updated state estimate $x_n^a = x_n^b + K_n \tilde{y}_n$,
Updated error covariance $P_n^a = (I - K_n H) P_n^b$

For the extended Kalman filter, the measurement operator could also be nonlinear, and the matrix H would be the Jacobian of the nonlinear operator, as in the case of the ensemble Kalman filter described next

2.4.5 Ensemble Kalman filter

The ensemble Kalman filter (EnKF) introduced by Evensen(1994) [42] can be viewed as a Monte Carlo approximation of the Kalman filter. Comparing with the standard Kalman filter and the extended Kalman filter, the state distribution of the EnKF is represented by an ensemble or sample from the distribution, and the covariance matrix is approximated by the sample covariance. Since the ensemble representation is a form of dimension reduction, it is computationally feasible for high-dimensional systems. In other words, the EnKF is suitable for systems with a large number of variables, such as discretizations of partial differential equations in geophysical models.

We consider again a discrete non-linear system

$$x_n = f(x_{n-1}) + W_{n-1},$$

The EnKf does not need to compute the covariance matrix of the probability density function of the state, instead using an M-member ensemble $\{x_1, x_2, \cdots, x_M\}$ to approximate the density. The assumption about the noise distribution is identical to that of the Kalman filter.

Similar to the Kalman filter, the ensemble Kalman filter consists of a prediction step and an update step at every time n. The ensemble Kalman filter obtains a sample from the

forecast distribution by simply applying the non-linear evolution equation

$$x_n^{b(i)} = f(x_{n-1}^{a(i)}) + W_{n-1}^i,$$

where $i = 1, \cdots, M$. The forecast ensemble $x_n^{b(1)}, \cdots, x_n^{b(M)}$ at time n is updated based on new observed data $z_n^{b(1)}, \cdots, z_n^{b(M)}$.

The observations are

$$z_n^i = z_n + v_n^i, \qquad i = 1, \ldots, M$$

An ensemble Kalman filter corrects the forecast ensemble $x_i^b, i = 1, \cdots, M$, to yield an analysis ensemble $x_i^a, i = 1, \cdots, M$. Therefore it provides a coupling between the underlying forecast and analysis random variables. The basic algorithmic steps of the EnKF can be summarised as follows:

Generate an M-member ensemble at initial time with state x_0, and the prior covariance guess B

$$x_0^{a(i)} = x_0 + \eta_i, \qquad i = 1, \ldots, M$$

where the $\eta_i \sim \mathcal{N}(0, B)$ are drawn from a normal distribution with zero mean and covariance B.

(i) **Prediction**

The forecast step is used to define the empirical mean

$$x_n^{b(i)} = f(x_{n-1}^{a(i)}), \qquad i = 1, \ldots, M,$$
$$\bar{x}_n^b = \frac{1}{M} \sum_{i=1}^{M} x_n^{b(i)}$$

and the covariance matrix

$$P_n^b = \frac{1}{M-1} \sum_{i=1}^{M} (x_n^{b(i)} - \bar{x}_n^b)(x_n^{b(i)} - \bar{x}_n^b)^T$$

(ii) **Updating**:

the Kalman gain: $K_n = P_n^b H^T (H P_n^b H^T + R)^{-1}$,

analysis: $x_n^{a(i)} = x_n^{b(i)} + K_n(z_n^i - H x_n^{b(i)})$,

The analysis can be calculated as the mean of the analysis

$$x_n^a = \frac{1}{M} \sum_{i=1}^{N} x_n^{a(i)}$$

and

$$P_n^a = \frac{1}{M-1} \sum_{i=1}^{M} (x_n^{a(i)} - x_n^a)(x_n^{a(i)} - x_n^a)^T.$$

2.4.6 Particle filter

Particle filter is a kind of Monte Carlo-based data assimilation methods, which use a set of particles to represent probabilities. The probability density over the system state is expressed as an empirical distribution, by randomly extracting the particle states from the posterior probability. Like ensemble Kalman filter, particle filter is a good way to track the state of a high-dimensional dynamical system given a model related to the state evolution in time and observations of particular states. However, the advantage of applying particle filter compared with Kalman filter is that for high-dimensional, nonlinear problems, particle filter methods do not rely on an assumption of Gaussianity of the posterior distribution.

In the following, we will explain the basic particle filter algorithm. We assume a noisy dynamic model with state vector x_n which comprises all the variables at given time step n. It is supposed to be a Markov process. Additionally, we also have the measurement information which represent observations of the system. The observation z_n takes value from some state x_n and is conditionally independent with the previous states. In other words, z_n only depends on x_n.

To describe the algorithm, we denote by $p(x_0)$ the initial distribution of the process; by $p(x_n|x_{n-1})$ the Markov transition probability density, which we usually get from the dynamic model; and by $p(z_n|x_n)$ the measurement likelihood.

The goal is to estimate the posterior probability distributions

$$p(x_n|z_{1:n}),$$

where we employ the notation $z_{1:n}$ as shorthand for $z_1, \ldots, z_n$.

Prediction

First, depending on the previous observation data $z_{1:n-1}$, we can predict the state distribution through the Chapman-Kolmogorov equation:

$$p(x_n|z_{1:n-1}) = \int p(x_{n-1}|z_{1:n-1})p(x_n|x_{n-1})dx_{n-1}.$$

Update

We will apply Bayes' rule to calculate the prior distribution after receiving the new observation z_n

$$p(x_n|z_{1:n}) = \frac{p(x_n|z_{1:n-1})p(z_n|x_n)}{\int p(x'_n|z_{1:n-1})p(z_n|x'_n)dx'_n} \tag{2.4.2}$$

Next, the sequential Monte Carlo method (SMC) may be employed to approximate the integrals above.

The basic Monte Carlo method

Computing the expected value of function $f(x)$, we would sample M independent random variables from the probability distribution $p(x)$, Hence, the probability density $p(x)$ could

be approached by $p(x) \approx P^M(x) = \frac{1}{M}\sum_{i=1}^{M}\delta(x - x^i)$, and the function's expected value becomes

$$E_{P^M}[f] \approx \int f(x)\frac{1}{M}\sum_{i=1}^{M}\delta(x - x^{(i)})dx = \frac{1}{M}\sum_{i=1}^{M} f(x^{(i)})$$

This estimate is unbiased and converges to the expected value of the original distribution with large particle number M. However, the probability $p(x)$, which we need to compute, is unknown. Hence, importance sampling is introduced next.

Importance sampling

Choosing an alternative probability density $q(x)$, named the importance probability density, we sample M independent particles $x_n^i \sim q(x_n)$.

$$\begin{aligned} E[f(x_n)] = \int f(x_n)p(x_n|z_{1:n})dx_n &= \frac{\int f(x_n)q(x_n|z_{1:n})\tilde{\omega}(x_n)dx_n}{\int q(x_n|z_{1:n})\tilde{\omega}(x_n)dx_n} \\ &\approx \frac{\frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_n^{(i)}f(x_n^{(i)})}{\frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_n^{(i)}} = \sum_{i=1}^{M}\omega_n^{(i)}f(x_n^{(i)}) \end{aligned}$$

with weights

$$\tilde{\omega}(x_n) = \frac{p(x_n|z_{1:n})}{q(x_n|z_{1:n})}, \qquad \omega^{(i)} = \frac{\tilde{\omega}^{(i)}}{\sum_{i=1}^{M}\tilde{\omega}^{(i)}}.$$

However, especially for high dimensional problems, it is difficult to choose an appropriate function $q(x)$, which should be as close to our desired probability distribution as possible in order to obtain good quality estimation. We introduce the sequential importance sampling.

Sequential importance sampling (SIS)

The importance sampling functions should be selected so that they are successively conditional to avoid the computational cost of recomputing the weights over the whole state sequence each time a new measurement is received:

$$q(x_{0:n}|z_{1:n}) = q(x_0)\prod_{k=1}^{n} q(x_k|x_{0:k-1}, z_{1:k}).$$

The importance weight is

$$\omega_n^{(i)} = \frac{p(x_{0:n}^{(i)}|z_{1:n})}{q(x_{0:n}^{(i)}|z_{1:n})}.$$

Since the state trajectories are preserved, this allows us to update recursively the importance weights:

$$\omega_n^{(i)} \propto \omega_{n-1}^{(i)}\frac{p(z_n|x_n^{(i)})p(x_n^{(i)}|x_{n-1}^{(i)})}{q(x_n^{(i)}|x_{n-1}^{(i)}, z_n)}.$$

where ω_{n-1}^i represents the weight at time step $n-1$ for particle i. In application, we simply sample choose the transition probability as the importance sampling function, which is

$$q(x_n) = p(x_n|x_{n-1})$$

Then the importance weights at time step n takes the form:

$$\omega_n^{(i)} \propto \omega_{n-1}^{(i)} p(z_n|x_n^{(i)})$$

However, this algorithm has a drawback during the filtering process. A small number of the particles will usually accumulate most of the weight of the sample, which leads to weight unbalance. This situation is problematic because the particles with very low weights will have negligible effects on the filtering distributions. This can be avoided by introducing a resampling step from the discrete distribution of the particles before propagation to replace the set of particles with equally weighted distribution.

Resampling

Starting with the weighted approximation, the posterior probability is

$$\hat{p}(x_n|z_{1:n}) = \hat{P}^M(x_n) = \sum_{j=1}^{M} \omega^{(i)} \delta(x_n - \hat{x}_n^{(i)})$$

we replace the above measure with a uniformly weighted measure

$$p(x_n|z_{1:n}) = P^M(x_n) = \frac{1}{M} \sum_{i=1}^{M} \delta(x_n - x_n^{(i)}).$$

such that the probability of selecting each sample in the new approximation is equal to its weight in the original sample:

$$p(x_n^i = \hat{x}_n^j) = \omega^{(j)}, \quad i, j = 1, \cdots, M.$$

The traditional particle filter illustrated above is a Bayesian particle method. In Chapter 5, we introduce a new particle filter method for problems in which a large amount of measurement data is available, either due to repeated (noisy) experiments or due to redundant observations. The first case provides us with samples from $p(x_n|x_{n-1})$ as determined by a stochastic process. The second case provides us with samples of the conditional expectation $p(z|x)$. We address the resampling problem by adding an optimal control to minimize Wasserstein distance of the empirical measure in the observation space.

In this section we have reviewed the basic and common algorithms for data assimilations. Many variants and improved methods have been developed [94].

Chapter 3

Symplectic Runge-Kutta discretization of a regularized forward-backward sweep iteration for optimal control problems

Abstract

Li, Chen, Tai & E. (*J. Machine Learning Research*, 2018) have proposed a regularization of the forward-backward sweep iteration for solving the Pontryagin maximum principle in optimal control problems. The authors prove the global convergence of the iteration in the continuous time case. In this thesis we show that their proof can be extended to the case of numerical discretization by symplectic Runge-Kutta pairs. We demonstrate the convergence with a simple numerical experiment.

This chapter is transcripted from the paper "Symplectic Runge-Kutta discretization of regularized forward-backward sweep iteration for optimal control problems" published in the Journal of Computational and Applied Mathematics.

Recently, Li et al. [77] proposed a new indirect iteration for optimal control problems in the context of deep neural networks, that utilizes the 'method of successive approximations', i.e. forward and backward integrations, combined with an 'augmented Lagrangian' regularization that ensures global convergence. The authors argue that this approach is particularly suitable for high-dimensional optimal control problems as encountered in deep learning. Large scale optimal control problems figure centrally in a number of modern applications such as deep neural networks [77], reinforcement learning [111, 12], filtering and data assimilation methods [8, 126] and mean field and stochastic differential games [22]. In this thesis we describe how the iteration of Li et al. combines naturally with symplectic/variational integrators to yield a convergent numerical scheme.

Optimal control problems possess a natural variational structure that gives rise to Hamiltonian dynamics which may be exploited in a numerical treatment [65]. Symplectic methods for Hamiltonian *initial* value problems have been much studied since the mid-1990s due to their demonstrated superiority for conserving energy and other first integrals [102, 59, 74]. In contrast, optimal control problems lead to *boundary* value problems, and it is unclear that the advantages of symplectic integrators for IVPs should translate to the BVP setting. Recent papers that address the use of symplectic Runge-Kutta methods for optimal control stress the conservation of quadratic invariants [101, 48] and the persistence of critical orbits in modified equation expansions [26]. See also recent work on the preservation of bifurcations under symplectic discretization of boundary value problems [83].

In the first three sections of this thesis we review the Hamiltonian structure of optimal control problems (§3.1), the regularized forward-backward sweep iteration proposed by Li et al. [77] (§3.1.2) and the discrete variational approach to constructing symplectic Runge-Kutta methods (§3.2). In Section 3.3 we prove the convergence of the discrete regularized forward-backward sweep iteration, which follows closely the proof of [77] for the continuous case. It is the symplectic structure of the discretization that facilitates this proof. Finally, in Section 3.4 we demonstrate the convergence of the method for a simple example using two symplectic discretizations.

3.1 Background

In this section we define continuous optimal control of differential equations and discuss their Hamiltonian structure, and we review the regularized forward-backward sweep iteration of Li et al. [77].

3.1.1 Hamiltonian structure of optimal control problems

The state of the system to be controlled is described by a vector $x(t) : \mathcal{T} \to \mathbf{R}^d$, where $\mathcal{T} = [0, T]$ represents a time interval. The control function $u(t)$ is for each t an element of the set of admissable controls $\mathcal{U} \subset \mathbf{R}^m$. The motion of the system is described by a differential equation

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(0) = \xi, \tag{3.1.1}$$

where $f : \mathbf{R}^d \times \mathcal{U} \to \mathbf{R}^d$ and $\xi \in \mathbf{R}^d$ is the initial state. The control $u(t)$ is chosen to minimize the objective functional

$$J[u] = \Phi(x(T)) + \int_0^T h(x(t), u(t))\, dt, \tag{3.1.2}$$

where $\Phi : \mathbf{R}^d \to \mathbf{R}$ is the end cost and $h : \mathbf{R}^d \times \mathcal{U} \to \mathbf{R}$ is the running cost.

In [77] no running cost h is considered. We include it here because it is present in many applications and its treatment is straightforward. As in [77] (cf. equations (A1) and (A2) of that article) we assume that Φ and f are twice continuously differentiable with respect to x and satisfy Lipschitz conditions for all x, $x' \in \mathbf{R}^d$, $u \in \mathcal{U}$ and $t \in \mathcal{T}$. We require similar assumptions on h:

$$\begin{aligned} &|\Phi(x) - \Phi(x')| + \|\Phi_x(x) - \Phi_x(x')\| \le K\|x - x'\|, \\ &\|f(x,u) - f(x',u)\| + \|f_x(x,u) - f_x(x',u)\| \le K\|x - x'\|, \\ &|h(x,u) - h(x',u)| + \|h_x(x,u) - h_x(x',u)\| \le K\|x - x'\|, \end{aligned} \tag{3.1.3}$$

where h_x denotes the vector of partial derivatives of h with respect to x and f_x denotes the Jacobian matrix of partial derivatives of f with respect to x. Here and throughout this chapter, we denote by $\|\cdot\|$ the Euclidean norm on vector spaces. Note that the solution $x(t)$ of (3.1.1) is well-defined for appropriate $u(t)$ so that we may think of J as a functional essentially depending only on $u(t)$.

The problem can be reformulated as a constrained optimization problem by introducing the Lagrange multiplier function $\lambda(t) : \mathcal{T} \to \mathbf{R}^d$ and the Lagrangian functional

$$\mathcal{L}[x, \lambda, u] = \Phi(x(T)) + \lambda_0^T(x(0) - \xi) + \int_0^T h(x,u) + \lambda^T\left(\dot{x} - f(x,u)\right)\, dt. \tag{3.1.4}$$

(Throughout the paper we use the transpose and dot product notation interchangeably, whichever is more convenient.) The variational derivatives of the functional $\mathcal{L}$ with respect to the functions $x(t)$, $\lambda(t)$ and $u(t)$, denoted $\mathcal{L}_x$, $\mathcal{L}_\lambda$ and $\mathcal{L}_u$, are defined with respect to the L^2 inner product. The first order necessary conditions for an optimum of (3.1.4) are given by the Euler-Lagrange equations ($\mathcal{L}_x \equiv \mathcal{L}_\lambda \equiv \mathcal{L}_u \equiv 0$):

$$\dot{x} = f(x,u), \quad x(0) = \xi, \tag{3.1.5}$$

$$\dot{\lambda} = -f_x(x,u)^T\lambda + h_x(x,u), \quad \lambda(T) = -\Phi_x(x(T)), \tag{3.1.6}$$

$$0 = f_u(x,u)^T\lambda - h_u(x,u). \tag{3.1.7}$$

In particular, if f and h are smooth and u is an optimal control in the interior of $\mathcal{U}$, then it satisfies (3.1.5)–(3.1.7). It is convenient to define a function $g(x, \lambda, u)$ for the right side of (3.1.6):

$$g(x, \lambda, u) = -f_x(x,u)^T\lambda + h_x(x,u). \tag{3.1.8}$$

A Legendre transform yields the Hamiltonian function

$$H(x, \lambda, u) = \lambda^T f(x,u) - h(x,u), \tag{3.1.9}$$

and Hamilton's equations are

$$\dot{x} = H_\lambda(x, \lambda, u), \tag{3.1.10}$$
$$\dot{\lambda} = -H_x(x, \lambda, u), \tag{3.1.11}$$
$$0 = H_u(x, \lambda, u). \tag{3.1.12}$$

Note that minimizing the objective functional J corresponds to maximizing the Hamiltonian with respect to u. The condition (3.1.12) above can be generalized to apply to controls $u(t)$ constrained to lie in $\mathcal{U}$ by replacing (3.1.12) with Pontryagin's maximum principle

$$\dot{x} = f(x, u^*), \quad x(0) = \xi, \tag{3.1.13}$$
$$\dot{\lambda} = g(x, \lambda, u^*), \quad \lambda(T) = -\Phi_x(x(T)) \tag{3.1.14}$$
$$u^*(t) = \arg\max_{u(t)\in\mathcal{U}} H(x, \lambda, u), \quad \forall t \in \mathcal{T} \tag{3.1.15}$$

3.1.2 Regularized forward-backward sweep iteration

Solution of (3.1.13)–(3.1.15) is challenging due to the boundary conditions. One approach is to solve in succession (3.1.13) for $x(t)$, (3.1.14) for $\lambda(t)$ and (3.1.15) for $u^*(t)$ and iterate. Such a forward-backward sweep iteration typically diverges unless the Lipschitz constant K and the time interval T are small [82]. In a recent article, Li et al. [77] proposed a modified iteration based on a regularized Lagrangian approach. They introduce the augmented Hamiltonian function

$$\tilde{H}(x, \lambda, u, p, q) = H(x, \lambda, u) - \frac{\rho}{2}\left(\|p - H_\lambda(x, \lambda, u)\|^2 + \|q + H_x(x, \lambda, u)\|^2\right), \tag{3.1.16}$$

where $\rho > 0$ is a regularization parameter. Subsequently, the forward-backward sweep iteration is modified to solve consecutively:

$$\dot{x}^{(k+1)} = \tilde{H}_\lambda(x^{(k+1)}, \lambda^{(k)}, u^{(k)}, \dot{x}^{(k+1)}, \dot{\lambda}^{(k)}), \tag{3.1.17}$$
$$\dot{\lambda}^{(k+1)} = -\tilde{H}_x(x^{(k+1)}, \lambda^{(k+1)}, u^{(k)}, \dot{x}^{(k+1)}, \dot{\lambda}^{(k+1)}), \tag{3.1.18}$$
$$u^{(k+1)} = \arg\max_{u(t)\in\mathcal{U}} \tilde{H}(x^{(k+1)}, \lambda^{(k+1)}, u, \dot{x}^{(k+1)}, \dot{\lambda}^{(k+1)}). \tag{3.1.19}$$

It is important to note that along solutions to (3.1.13) and (3.1.14), the right two terms of (3.1.16) are zero. Consequently, only (3.1.19) is modified with respect to (3.1.15). However, Li et al. show that this modification is sufficient to ensure convergence [77].

Li et al. introduce the regularized forward-backward sweep iteration to train deep neural networks [77] and argue that an advantage of this approach is that it is suitable for application to high dimensional systems.

The analysis of [77] addresses only the continuous time case. Li et al. point out that the question of whether Pontryagin's principle holds under numerical discretization is 'a delicate one' and refer to counterexamples. In this paper we show that for variational/symplectic RK methods, an analysis analogous to that of Li et al. holds. In particular, their proof of convergence may be translated directly to discrete form.

3.2 Variational integrators and symplectic Runge–Kutta pairs

Symplectic Runge-Kutta methods possess two properties that make them attractive for numerical integration of Hamiltonian *initial* value problems: they conserve certain quadratic first integrals and they conserve a modified Hamiltonian function over exponentially long time intervals. See the monographs [102, 59, 74] for a complete discussion. Symplectic Runge-Kutta methods can be derived using a discrete variational formalism, see [81].

Variational methods are also well known in the optimal control literature see e.g. the work of Marsden, Leok and Ober-Blöbaum [88] and references therein. In a recent review, Sanz-Serna [101] argues that it is the property of conservation of quadratic integrals that it is most relevant in the adjoint context.

For optimal control, the use of the variational integrator framework may have additional advantages: first, by discretizing the integral before optimizing, one constructs a discrete problem for which an optimum may be established, whereas directly discretizing the Euler-Lagrange equations relies on the approximation property in the limit $\tau \to 0$, where $\tau > 0$ is the step size, to guarantee an optimum. Second, backward error analysis implies the existence of a modified Hamiltonian, near the continuous Hamiltonian, which may have consequences for optimality in the presence of nonunique minima. Backward error analysis may also be applicable for control problems on long time intervals, or for problems with multiple time scales for which the time interval is long on a fast time scale.

We discretize the interval $\mathcal{T}$ into $N > 0$ equal steps of size $\tau = T/N$. An s-stage Runge-Kutta method for the state equation (3.1.1) is

$$x_{n+1} = x_n + \tau \sum_{i=1}^{s} b_i f(X_{i,n}, U_{i,n}), \tag{3.2.1}$$

$$X_{i,n} = x_n + \tau \sum_{j=1}^{s} a_{ij} f(X_{j,n}, U_{j,n}), \quad i = 1, \ldots, s, \tag{3.2.2}$$

where $n = 0, \ldots, N-1$ denotes the time step index and the coefficients b_i and a_{ij}, $i, j = 1, \ldots, s$, are chosen to ensure accuracy, stability, and additional properties. See the monographs [60, 61] for a thorough treatment. Numerical consistency requires the coefficients b_i satisfy $\sum_i b_i = 1$. In this paper we will also assume that $b_i \geq 0$, $i = 1, \ldots, s$.

To simplify notation we will frequently suppress the time step index n in the internal stage variables $X_{i,n}$ and $U_{i,n}$. In all formulas the stage variables are evaluated at time level n, so there should be no ambiguity.

A variational integrator for the Lagrangian (2.1.4) is a quadrature formula consistent with the above RK method. Enforcing the internal stage relations (3.2.2) requires the

introduction of additional Lagrange multipliers. The discrete Lagrangian becomes

$$\mathcal{L}[\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{X},\boldsymbol{u},\boldsymbol{G}] = \Phi(x_N) + \lambda_0^T(x_0-\xi) + \tau\sum_{n=0}^{N-1}\Bigg\{ \sum_{i=1}^{s} b_i h(X_i,U_i) + \lambda_{n+1}^T\left(\frac{x_{n+1}-x_n}{\tau} - \sum_{i=1}^{s} b_i f(X_i,U_i)\right) - \sum_{i=1}^{s} b_i G_i\cdot\left(X_i - x_n - \tau\sum_{j=1}^{s} a_{ij} f(X_j,U_j)\right)\Bigg\}. \quad (3.2.3)$$

Here and henceforth we denote $\boldsymbol{x} = \{x_n \,|\, n = 0,\dots,N\}$, $\boldsymbol{X} = \{X_{i,n} \,|\, i = 1,\dots,s; n = 0,\dots,N-1\}$, etc. An exception is the control variable, which only appears at internal stage values. Consequently we may denote $\boldsymbol{u} = \{U_{i,n} \,|\, i = 1,\dots,s; n = 0,\dots,N-1\}$ without ambiguity. We also denote $u_n = \{U_{i,n} \,|\, i = 1,\dots,s\}$.

The associated discretization of the cost function (3.1.2) is

$$J^\tau[\boldsymbol{u}] = \Phi(x_N) + \tau\sum_{n=0}^{N-1}\sum_{i=1}^{s} b_i h(X_i,U_i). \quad (3.2.4)$$

One can formally construct a discrete variational derivative of (3.2.3) with respect to discrete function spaces and a discrete inner product. However for uniform time step τ it is sufficient to consider just partial derivatives of $\mathcal{L}$. The Euler-Lagrange equations become:

$$\frac{\partial\mathcal{L}}{\partial\lambda_n} = 0 = x_{n+1} - x_n - \tau\sum_{i=1}^{s} b_i f(X_i,U_i), \quad x_0 = \xi, \quad (3.2.5)$$

$$\frac{\partial\mathcal{L}}{\partial G_i} = 0 = X_i - x_n - \tau\sum_{j=1}^{s} a_{ij} f(X_j,U_j), \quad (3.2.6)$$

$$\frac{\partial\mathcal{L}}{\partial x_n} = 0 = -\lambda_{n+1} + \lambda_n + \tau\sum_{i=1}^{s} b_i G_i, \quad \lambda_N = -\Phi_x(x_N), \quad (3.2.7)$$

$$\frac{\partial\mathcal{L}}{\partial X_k} = 0 = b_k h_x(X_k,U_k) - b_k f_x(X_k,U_k)^T\lambda_{n+1} - b_k G_k + \tau\sum_{i=1}^{s} b_i a_{ik} f_x(X_k,U_k)^T G_i, \quad (3.2.8)$$

$$\frac{\partial\mathcal{L}}{\partial U_k} = 0 = b_k h_u(X_k,U_k) - b_k f_u(X_k,U_k)^T\lambda_{n+1} - \tau\sum_{i=1}^{s} b_i a_{ik} f_u(X_k,U_k)^T G_i. \quad (3.2.9)$$

The relations (3.2.5)–(3.2.6) are clearly equivalent to (3.2.1)–(3.2.2). Solving (3.2.7) for λ_{n+1}, substituting into (3.2.8) and defining the coefficients $\tilde{a}_{ij} = b_j - b_j a_{ji}/b_i$, one finds

$$G_i = -f_x(X_i,U_i)^T\left[\lambda_n + \tau\sum_{j=1}^{s}\tilde{a}_{ij}G_j\right] + h_x(X_i,U_i).$$

Similarly (3.2.9) is written

$$0 = h_u(X_i, U_i) - f_u(X_i, U_i)^T \left[\lambda_n + \tau \sum_{j=1}^{s} \tilde{a}_{ij} G_j \right]. \tag{3.2.10}$$

It is useful to introduce the auxiliary stage variable Λ_i to represent the term in square brackets in the previous two expressions:

$$\Lambda_i = \lambda_n + \tau \sum_{i=1}^{s} \tilde{a}_{ij} G_j,$$

such that (cf. (3.1.8))

$$G_i = g(X_i, \Lambda_i, U_i) = -f_x(X_i, U_i)^T \Lambda_i + h_x(X_i, U_i)$$

and the condition (3.2.10) becomes

$$0 = h_u(X_i, U_i) - f_u(X_i, U_i)^T \Lambda_i.$$

In terms of the new variable, the variational Runge-Kutta discretization of Pontryagin's maximum principle is

$$x_{n+1} = x_n + \tau \sum_{i=1}^{s} b_i f(X_i, U_i), \quad x_0 = \xi, \tag{3.2.11}$$

$$X_i = x_n + \tau \sum_{j=1}^{s} a_{ij} f(X_j, U_j), \quad i = 1, \ldots, s, \tag{3.2.12}$$

$$\lambda_{n+1} = \lambda_n + \tau \sum_{i=1}^{s} b_i g(X_i, \Lambda_i, U_i), \quad \lambda_N = -\Phi_x(x_N), \tag{3.2.13}$$

$$\Lambda_i = \lambda_n + \tau \sum_{j=1}^{s} \tilde{a}_{ij} g(X_j, \Lambda_j, U_j), \quad i = 1, \ldots, s, \tag{3.2.14}$$

$$0 = h_u(X_i, U_i) - f_u(X_i, U_i)^T \Lambda_i, \quad i = 1, \ldots, s. \tag{3.2.15}$$

This system consists of the state equations (3.2.11) and (3.2.12), the adjoint equations (3.2.13) and (3.2.14), and the optimality condition (3.2.15).

Recalling the Hamiltonian (3.1.9), we can also write the above relations in a form that emphasizes the Hamiltonian structure:

$$x_{n+1} = x_n + \tau \sum_{i=1}^{s} b_i H_\lambda(X_i, \Lambda_i, U_i), \quad x_0 = \xi, \tag{3.2.16}$$

$$X_i = x_n + \tau \sum_{j=1}^{s} a_{ij} H_\lambda(X_j, \Lambda_j, U_j), \quad i = 1, \ldots, s, \tag{3.2.17}$$

$$\lambda_{n+1} = \lambda_n - \tau \sum_{i=1}^{s} b_i H_x(X_i, \Lambda_i, U_i), \quad \lambda_N = -\Phi_x(x_N), \tag{3.2.18}$$

$$\Lambda_i = \lambda_n - \tau \sum_{j=1}^{s} \tilde{a}_{ij} H_x(X_j, \Lambda_j, U_j), \quad i = 1, \ldots, s, \tag{3.2.19}$$

$$0 = H_u(X_i, \Lambda_i, U_i), \quad i = 1, \ldots, s. \tag{3.2.20}$$

In some cases, it is appropriate to replace the latter condition by the more general

$$U_i = \arg\max_{u\in\mathcal{U}} H(X_i, \Lambda_i, u), \quad i = 1, \ldots, s. \tag{3.2.21}$$

As noted in [101], a pair of RK methods defined by coefficients $\{b_i, a_{ij}\}$ and $\{b_i, \tilde{a}_{ij}\}$, where $\tilde{a}_{ij} = b_j - b_j a_{ij}/b_i$, constitute a symplectic partitioned RK pair. That is, if these methods are applied to a pair of differential equations $\dot{x} = H_\lambda(x, \lambda)$, $\dot{\lambda} = -H_x(x, \lambda)$, then the resulting map from t_n to t_{n+1} is a symplectic map. Hence, we obtain the well-known result that the discrete variational approach automatically produces a symplectic integrator for the Euler-Lagrange equations.

3.2.1 Symplectic Euler method

The elementary example of a symplectic variational integrator is the symplectic Euler method, which corresponds to the RK pair with $s = 1$, $b_1 = 1$, $a_{11} = 0 = 1 - \tilde{a}_{11}$. In this case all the internal stage relations can be eliminated, leaving the discrete Lagrangian

$$\mathcal{L}[\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{u}] = \Phi(x_N) + \lambda_0^T(x_0 - \xi) + \tau \sum_{n=0}^{N-1} h(x_n, u_n) + \lambda_{n+1}^T \left(\frac{x_{n+1} - x_n}{\tau} - f(x_n, u_n) \right). \tag{3.2.22}$$

The discrete Pontryagin maximum principle is

$$x_{n+1} = x_n + \tau f(x_n, u_n), \tag{3.2.23}$$
$$\lambda_{n+1} = \lambda_n - \tau f_x(x_n, u_n)^T \lambda_{n+1} + \tau h_x(x_n, u_n), \tag{3.2.24}$$
$$0 = f_u(x_n, u_n)^T \lambda_{n+1} - h_u(x_n, u_n), \tag{3.2.25}$$

with boundary conditions $x_0 = \xi$, $\lambda_N = -\Phi_x(x_N)$.

Note that (3.2.23)–(3.2.25) can also be written in terms of the Hamiltonian H:

$$\frac{x_{n+1} - x_n}{\tau} = H_\lambda(x_n, \lambda_{n+1}, u_n), \tag{3.2.26}$$
$$\frac{\lambda_{n+1} - \lambda_n}{\tau} = -H_x(x_n, \lambda_{n+1}, u_n), \tag{3.2.27}$$
$$0 = H_u(x_n, \lambda_{n+1}, u_n). \tag{3.2.28}$$

3.2.2 Reduced notation for Runge-Kutta methods

Hager [58] introduced notation that casts general symplectic Runge-Kutta methods (3.2.16)-(3.2.20) in a form consistent with the symplectic Euler method. Define

$$f^\tau(x, u) = \sum_{i=1}^{s} b_i f(X_i(x, u), U_i(u)), \quad h^\tau(x, u) = \sum_{i=1}^{s} b_i h(X_i(x, u), U_i(u)), \tag{3.2.29}$$

where we view the stage values X_i and U_i as functions of grid point value x and discrete control $u = \{U_1, \ldots, U_s\}$ according to

$$X_i(x, u) = x + \tau \sum_{j=1}^{s} a_{ij} f(X_j(x, u), U_j(u)), \quad i = 1, \ldots, s. \tag{3.2.30}$$

Similarly, define the Hamiltonian

$$H^\tau(x,\lambda,u) = \lambda^T f^\tau(x,u) - h^\tau(x,u). \tag{3.2.31}$$

With this notation, the discretization of Pontryagin's maximum principle with any symplectic Runge-Kutta pair can be written as

$$\frac{x_{n+1}-x_n}{\tau} = H^\tau_\lambda(x_n,\lambda_{n+1},u_n), \tag{3.2.32}$$

$$\frac{\lambda_{n+1}-\lambda_n}{\tau} = -H^\tau_x(x_n,\lambda_{n+1},u_n), \tag{3.2.33}$$

$$0 = H^\tau_u(x_n,\lambda_{n+1},u_n). \tag{3.2.34}$$

To see the equivalence, note that evaluating (3.2.30) at x_n yields the implicit relations (3.2.12). Taking the derivative of (3.2.31) with respect to λ and substituting (3.2.29) shows (3.2.32) to be equivalent to (3.2.11). The proof of the relation (3.2.33) is more involved. We adapt the proof from [58] to our notation.

Let $\Psi_i(x) = \partial_x X_i(x,u)$ and denote $\Psi_i = \Psi_i(x_n)$. Then computing the derivative of (3.2.30) at x_n yields the linear system

$$\Psi_i = I + \tau \sum_j a_{ij} f_x(X_i,U_i)\Psi_j. \tag{3.2.35}$$

The derivative on the right side of (3.2.33) is

$$H^\tau_x(x_n,\lambda_{n+1},u_n) = \sum_{j=1}^s b_j \Psi_j^T f_x(X_j,U_j)^T \lambda_{n+1} - b_j \Psi_j^T h_x(X_j,U_j). \tag{3.2.36}$$

Rearranging (3.2.8) gives

$$b_j G_j - \tau \sum_{i=1}^s b_i a_{ij} f_x(X_j,U_j)^T G_i = b_j h_x(X_j,U_j) - b_j f_x(X_j,U_j)^T \lambda_{n+1}.$$

Premultiplying by Ψ_j^T and summing over j gives

$$\begin{aligned}\sum_{j=1}^s b_j \Psi_j^T G_j - \tau \sum_{i,j=1}^s b_i a_{ij} \Psi_j^T f_x(X_j,U_j)^T G_i \\ = \sum_{j=1}^s b_j \Psi_j^T h_x(X_j,U_j) - b_j \Psi_j^T f_x(X_j,U_j)^T \lambda_{n+1} = -H^\tau_x(x_n,\lambda_{n+1},u_n),\end{aligned} \tag{3.2.37}$$

where the last equality follows from (3.2.36). Now changing the index of summation in the first sum on the left, we obtain

$$\begin{aligned}-H^\tau_x(x_n,\lambda_{n+1},u_n) &= \sum_{i=1}^s b_i \Psi_i^T G_i - \tau \sum_{i=1}^s \left(\sum_{j=1}^s a_{ij} \Psi_j^T f_x(X_j,U_j)^T \right) b_i G_i, \\ &= \sum_{i=1}^s b_i G_i, \\ &= \frac{\lambda_{n+1}-\lambda_n}{\tau},\end{aligned}$$

where the second equality follows from (3.2.35), thus confirming (3.2.33).

The proof of (3.2.20) follows similar arguments, see [58]. Note the analogy between the relations (3.2.16)–(3.2.20) and (3.2.26)–(3.2.27) for the symplectic Euler method.

3.3 Convergence analysis

In this section we prove the convergence of the regularized forward-backward sweep iteration (3.1.17)–(3.1.19) for symplectic Runge-Kutta methods. The proof here follows closely that of Li et al. for the continuous case [77]. It is the symplectic/variational structure that facilitates this analogy.

Using the compact notation (3.2.29) and (3.2.31), we define the discrete regularized Hamiltonian function

$$\tilde{H}^\tau(x,\lambda,u,q,p) = H^\tau(x,\lambda,u) - \frac{\rho}{2}\left(\|q - H^\tau_\lambda(x,\lambda,u)\|^2 + \|p + H^\tau_x(x,\lambda,u)\|^2\right). \tag{3.3.1}$$

In iterate k, the symplectic Runge-Kutta discretization of the regularized forward-backward sweep iteration (3.1.17)–(3.1.19) solves, in sequence,

$$x_{n+1}^{(k+1)} = x_n^{(k+1)} + \tau \tilde{H}^\tau_\lambda\left(x_n^{(k+1)}, \lambda_{n+1}^{(k)}, u_n^{(k)}, \frac{x_{n+1}^{(k+1)} - x_n^{(k+1)}}{\tau}, \frac{\lambda_{n+1}^{(k)} - \lambda_n^{(k)}}{\tau}\right), \tag{3.3.2}$$

$$\lambda_{n+1}^{(k+1)} = \lambda_n^{(k+1)} - \tau \tilde{H}^\tau_x\left(x_n^{(k+1)}, \lambda_{n+1}^{(k+1)}, u_n^{(k)}, \frac{x_{n+1}^{(k+1)} - x_n^{(k+1)}}{\tau}, \frac{\lambda_{n+1}^{(k+1)} - \lambda_n^{(k+1)}}{\tau}\right), \tag{3.3.3}$$

$$u_n^{(k+1)} = \arg\max_{u\in\mathcal{U}} \tilde{H}^\tau\left(x_n^{(k+1)}, \lambda_{n+1}^{(k+1)}, u, \frac{x_{n+1}^{(k+1)} - x_n^{(k+1)}}{\tau}, \frac{\lambda_{n+1}^{(k+1)} - \lambda_n^{(k+1)}}{\tau}\right), \tag{3.3.4}$$

proceeding as follows: (3.3.2) by forward integration with $\boldsymbol{u}$ and $\boldsymbol{\lambda}$ fixed, then (3.3.3) by backward integration with $\boldsymbol{x}$ and $\boldsymbol{u}$ fixed, and finally (3.3.4) solved for each time step independently (e.g. in parallel), with $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ fixed.

It is important to recall that with u fixed, along solutions of (3.3.2) and (3.3.3) the extra regularization terms in the extended Hamiltonian $\tilde{H}^\tau$ are identically zero and

$$\tilde{H}^\tau_\lambda\left(x_n, \lambda_{n+1}, u_n, \frac{x_{n+1} - x_n}{\tau}, \frac{\lambda_{n+1} - \lambda_n}{\tau}\right) = H^\tau_\lambda(x_n, \lambda_{n+1}, u_n),$$

$$\tilde{H}^\tau_\lambda\left(x_n, \lambda_{n+1}, u_n, \frac{x_{n+1} - x_n}{\tau}, \frac{\lambda_{n+1} - \lambda_n}{\tau}\right) = H^\tau_x(x_n, \lambda_{n+1}, u_n),$$

i.e., the regularization terms only affect the maximization step (3.3.4).

Notation and identities

In the following we consider a single iteration of (3.3.2)–(3.3.4). We think of H^τ, $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ as functions of $\boldsymbol{u}$. Consequently we denote by x_n^u and λ_n^u the numerical solutions to (3.2.32) and (3.2.33) given a candidate control $\boldsymbol{u}$.

It is convenient to define the composite notation

$$z_n = \begin{pmatrix} x_n \\ \lambda_{n+1} \end{pmatrix}, \quad H_z^\tau(z_n, u_n) = \begin{pmatrix} H_x^\tau(x_n, \lambda_{n+1}, u_n) \\ H_\lambda^\tau(x_n, \lambda_{n+1}, u_n) \end{pmatrix}.$$

We consider two control sequences $\boldsymbol{u}$ and $\boldsymbol{v}$, and we are interested in bounding the change in $\tilde{H}^\tau$ when $\boldsymbol{u}$ is replaced by $\boldsymbol{v}$. To that end we define an operator that denotes the difference between quantities dependent on $\boldsymbol{u}$ and $\boldsymbol{v}$:

$$\delta_u x_n = x_n^v - x_n^u.$$

We use this notation also for functions, e.g.

$$\delta_u H^\tau|_n = H^\tau(z_n^v, v_n) - H^\tau(z_n^u, u_n).$$

We denote by $\bar{\delta}_u H^\tau$ the change due to an update in $\boldsymbol{u}$ with $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ fixed as functions of u:

$$\bar{\delta}_u H^\tau|_n = H^\tau(x_n^u, \lambda_{n+1}^u, v_n) - H^\tau(x_n^u, \lambda_{n+1}^u, u_n). \tag{3.3.5}$$

We denote the temporal forward difference operator by δ_t:

$$\delta_t x_n = \frac{x_{n+1} - x_n}{\tau},$$

and remark that δ_t commutes with δ_u when applied to variables, i.e. $\delta_u \delta_t x_n = \delta_t \delta_u x_n$.

Next we note the discrete integration by parts formula:

$$\begin{aligned} \tau \sum_{n=0}^{N-1} \lambda_{n+1}^T \delta_t x_n &= \sum_{n=0}^{N-1} \lambda_{n+1}^T (x_{n+1} - x_n) \\ &= -\lambda_0^T x_0 + \lambda_0^T x_0 - \lambda_1^T x_0 + \lambda_1^T x_1 + \cdots - \lambda_N^T x_{N-1} + \lambda_N^T x_N \\ &= \lambda_n^T x_n \big|_0^N - \tau \sum_{n=0}^{N-1} (\delta_t \lambda_n)^T x_n. \end{aligned}$$

This formula holds for any discrete functions defined for $n = 0, \ldots, N$, and in particular we may insert the difference operator δ_u to obtain two useful alternatives:

$$\tau \sum_{n=0}^{N-1} \lambda_{n+1}^u \cdot \delta_t \delta_u x_n = \lambda_n^u \cdot \delta_u x_n \big|_0^N - \tau \sum_{n=0}^{N-1} \delta_t \lambda_n^u \cdot \delta_u x_n. \tag{3.3.6}$$

$$\tau \sum_{n=0}^{N-1} \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n = \delta_u \lambda_n \cdot \delta_u x_n \big|_0^N - \tau \sum_{n=0}^{N-1} \delta_t \delta_u \lambda_n \cdot \delta_u x_n. \tag{3.3.7}$$

Estimates

In the Appendix of this chapter we show that—possibly with a restriction on step size—the Lipschitz conditions (3.1.3) on f and h translate into related Lipschitz conditions

on f^τ and h^τ. Henceforth choosing K to be a generic Lipschitz constant we obtain the bounds

$$\begin{aligned} &\|f^\tau(x,u) - f^\tau(x',u)\| + \|f^\tau_x(x,u) - f^\tau_x(x',u)\| \le K\|x - x'\|, \\ &|h^\tau(x,u) - h^\tau(x',u)| + \|h^\tau_x(x,u) - h^\tau_x(x',u)\| \le K\|x - x'\|. \end{aligned} \tag{3.3.8}$$

Note also that the leftmost terms in the above inequalities as well as the analogous ones of (3.1.3) imply global bounds on the derivatives (which may be relaxed, see [77])

$$\|\Phi_x(x)\| \le K, \quad \|f_x(x,u)\| \le K, \quad \|h_x(x,u)\| \le K, \quad \|f^\tau_x(x,u)\| \le K, \quad \|h^\tau_x(x,u)\| \le K. \tag{3.3.9}$$

We use two discrete forms of Grönwall's lemma [41]. Let $\{b_n\}$ be a given, monotone sequence and $\tau, K > 0$. Then the following implication holds:

$$a_{n+1} \le (1 + \tau K)a_n + \tau b_n, \forall n \quad \Rightarrow \quad a_n \le e^{\tau n K} a_0 + K^{-1} e^{\tau n K} b_{n-1}. \tag{3.3.10}$$

Under the same conditions, the following implication holds:

$$a_{n+1} \le b_{n+1} + \tau K \sum_{m=0}^{n} a_m, \forall n \quad \Rightarrow \quad a_n \le e^{\tau n K} b_n. \tag{3.3.11}$$

From (3.2.36) and (3.5.5), and using the bounds (3.3.9) on f_x and h_x,

$$\|\lambda_n\| \le \|\lambda_{n+1}\| + \tau\|H^\tau_x(x_n\lambda_{n+1}, u_n)\| \le (1 + \tau K)\|\lambda_{n+1}\| + \tau K,$$

where we have absorbed the constant from (3.5.5) into K. Further using Grönwall bound (3.3.10) and the bound (3.3.9) on $\Phi_x(x)$,

$$\|\lambda_n\| \le K_1 := (K+1)e^{\tau K N} = (K+1)e^{KT}. \tag{3.3.12}$$

From $\delta_u x_{n+1} = \delta_u x_n + \tau \delta_u f^\tau|_n$ and $\delta_u x_0 = 0$ we calculate

$$\begin{aligned} \|\delta_u x_n\| &\le \tau \sum_{m=0}^{n-1} \|\delta_u f^\tau|_m\| \\ &\le \tau \sum_{m=0}^{n-1} \|\bar{\delta}_u f^\tau|_m\| + \|f^\tau(x^v_m, v_m) - f^\tau(x^u_m, v_m)\| \\ &\le \tau \sum_{m=0}^{n-1} \|\bar{\delta}_u f^\tau|_m\| + K\|\delta_u x_m\|, \end{aligned}$$

and using Grönwall bound (3.3.11),

$$\|\delta_u x_n\| \le \tau e^{KT} \sum_{m=0}^{N-1} \|\bar{\delta}_u f^\tau|_m\|. \tag{3.3.13}$$

Similarly, from $\delta_u\lambda_n = \delta_u\lambda_{n+1} + \tau\delta_u H_x^\tau(x_n, \lambda_{n+1}, u_n)$ we obtain

$$\begin{aligned}\|\delta_u\lambda_n\| &\leq \|\delta_u\lambda_N\| + \tau\sum_{m=n}^{N-1}\|\delta_u H_x^\tau|_m\| \\ &\leq K\|\delta_u x_N\| + \tau\sum_{m=n}^{N-1}\|\bar{\delta}_u H_x^\tau|_m\| + \tau K\sum_{m=n}^{N-1}\|\delta_u\lambda_{m+1}\| + \tau K(K_1+1)\sum_{m=n}^{N-1}\|\delta_u x_m\|,\end{aligned}$$

where the last term uses (3.1.3) and the Lipschitz condition (3.3.8) on H_x^τ. The discrete Grönwall's lemma gives

$$\|\delta_u\lambda_n\| \leq Ke^{KT}\left(\|\delta_u x_N\| + \tau(K_1+1)\sum_{m=0}^{N-1}\|\delta_u x_m\|\right) + \tau e^{KT}\sum_{m=0}^{N-1}\|\bar{\delta}_u H_x^\tau|_m\|.$$

Finally, making use of (3.3.13) gives

$$\|\delta_u\lambda_n\| \leq \tau K_2\sum_{m=0}^{N-1}\|\bar{\delta}_u f^\tau|_m\| + \tau e^{KT}\sum_{m=0}^{N-1}\|\bar{\delta}_u H_x^\tau|_m\|, \quad K_2 = Ke^{2KT}(1+(K_1+1)T). \tag{3.3.14}$$

The following estimates make use of Taylor's theorem in the mean value form:

$$\delta_u H_z^\tau|_n \cdot \delta_u z_n = \bar{\delta}_u H_z^\tau|_n \cdot \delta_u z_n + \delta_u z_n \cdot H_{zz}^\tau(z_n^u + r_1\delta_u z_n, u_n)\cdot\delta_u z_n, \tag{3.3.15}$$

for some $r_1 \in [0,1]$, where H_{zz}^τ denotes the Hessian matrix of second partial derivatives of H^τ.

$$\delta_u\Phi_x(x_N)\cdot\delta_u x_N = \delta_u x_N \cdot \Phi_{xx}(x_N^u + r_2\delta_u x_N)\cdot\delta_u x_N, \tag{3.3.16}$$

for some $r_2 \in [0,1]$. Similarly,

$$\Phi_x(x_N^u)\cdot\delta_u x_N = \Phi(x_N^v) - \Phi(x_N^u) - \frac{1}{2}\delta_u x_N \cdot \Phi_{xx}(x_N^u + r_3\delta_u x_N)\cdot\delta_u x_N, \tag{3.3.17}$$

for some $r_3 \in [0,1]$.

$$\delta_u H^\tau = \bar{\delta}_u H^\tau + H_z^\tau(z_n^u, v)\cdot\delta_u z_n + \frac{1}{2}\delta_u z_n \cdot H_{zz}^\tau(z_n^u + r_4\delta_u z_n, v_n)\cdot\delta_u z_n, \tag{3.3.18}$$

for some $r_4 \in [0,1]$.

Convergence of the iteration

Convergence of the regularized forward-backward sweep iteration relies on Lemma 2 of [77], the proof of which we adapt for the symplectic RK method here. The result we want states that under the assumptions (3.1.3), there exists a constant $C > 0$ such that for any

two discrete controls $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$, the discrete cost function (3.2.4) satisfies

$$\begin{aligned} J^\tau(\boldsymbol{v}) \leq J^\tau(\boldsymbol{u}) - \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n + C\tau \sum_{n=0}^{N-1} \|f^\tau(x_n^u, v_n) - f^\tau(x_n^u, u_n)\|^2 \\ + C\tau \sum_{n=0}^{N-1} \|H_x^\tau(x_n^u, \lambda_{n+1}^u, v_n) - H_x^\tau(x_n^u, \lambda_{n+1}^u, u_n)\|^2 \\ = J^\tau(\boldsymbol{u}) - \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n + C\tau \sum_{n=0}^{N-1} \|\bar{\delta}_u H_z^\tau|_n\|^2 \end{aligned} \tag{3.3.19}$$

Define the discrete functional

$$\mathcal{I}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{u}) = \tau \sum_{n=0}^{N-1} \lambda_{n+1}^T \delta_t x_n - H^\tau(x_n, \lambda_{n+1}, u_n) - h^\tau(x_n, u_n) \equiv 0. \tag{3.3.20}$$

The functional $\mathcal{I}$ is identically zero for sequences $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ satisfying (3.2.32)–(3.2.33). Note the identity

$$\delta_u(\lambda_{n+1} \cdot \delta_t x_n) = \lambda_{n+1}^u \cdot \delta_t \delta_u x_n + \delta_u \lambda_{n+1} \cdot \delta_t x_n^u + \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n. \tag{3.3.21}$$

We find

$$\begin{aligned} 0 \equiv \mathcal{I}(\boldsymbol{x}^v, \boldsymbol{\lambda}^v, \boldsymbol{v}) - \mathcal{I}(\boldsymbol{x}^u, \boldsymbol{\lambda}^u, \boldsymbol{u}) = \\ \tau \sum_{n=0}^{N-1} \lambda_{n+1}^u \cdot \delta_t \delta_u x_n + \delta_u \lambda_{n+1} \cdot \delta_t x_n^u + \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n \\ - \tau \sum_{n=0}^{N-1} \left(H^\tau(x_n^v, \lambda_{n+1}^v, v_n) - H^\tau(x_n^u, \lambda_{n+1}^u, u_n) \right) \\ - \tau \sum_{n=0}^{N-1} \left(h^\tau(x_n^v, v_n) - h^\tau(x_n^u, u_n) \right) \end{aligned}$$

In our notation this is

$$0 \equiv \delta_u \mathcal{I} = \tau \sum_{n=0}^{N-1} \lambda_{n+1}^u \cdot \delta_t \delta_u x_n + \delta_u \lambda_{n+1} \cdot \delta_t x_n^u + \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n - \delta_u H^\tau|_n - \delta_u h^\tau|_n. \tag{3.3.22}$$

Remark. *This is the point where the symplectic/variational property of the symplectic RK method is important. Since x_n and λ_n are discretized by a symplectic partitioned Runge-Kutta method, we see that $\mathcal{I}$ is also equivalent to the constraint part of the discrete Lagrangian:*

$$\mathcal{I} = \tau \sum_{N=0}^{N-1} \lambda_{n+1}^T \left(\frac{x_{n+1} - x_n}{\tau} - f^\tau(x_n, u_n) \right),$$

which is identically zero along a solution to the state dynamics (3.2.32)*. Of course, one could define $\mathcal{I}$ as above for an arbitrary choice of the λ_n. Then $\mathcal{I}$ would be identically zero, but one would not be able to translate this into a statement about the Hamiltonian.*

Using (3.3.6) the first two terms on the right side of (3.3.22) are equal to

$$\tau \sum_{n=0}^{N-1} \lambda_{n+1}^u \cdot \delta_t \delta_u x_n + \delta_u \lambda_{n+1} \cdot \delta_t x_n^u$$
$$= \lambda_n^u \cdot \delta_u x_n \big|_0^N + \tau \sum_{n=0}^{N-1} f^\tau(x_n^u, u_n) \cdot \delta_u \lambda_{n+1} + H_x^\tau(x_n^u, \lambda_{n+1}^u, u_n) \cdot \delta_u x_n,$$

or in compact notation

$$\tau \sum_{n=0}^{N-1} \lambda_{n+1}^u \cdot \delta_t \delta_u x_n + \delta_u \lambda_{n+1} \cdot \delta_t x_n^u = \lambda_n^u \cdot \delta_u x_n \big|_0^N + \tau \sum_{n=0}^{N-1} H_z^\tau(z_n^u, u_n) \cdot \delta_u z_n. \tag{3.3.23}$$

Similarly, using (3.3.7) the third term on the right side of (3.3.22) is equal to

$$\tau \sum_{n=0}^{N-1} \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n = \frac{1}{2}\tau \sum_{n=0}^{N-1} \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n + \frac{1}{2}\tau \sum_{n=0}^{N-1} \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n$$
$$= \frac{1}{2}\delta_u \lambda_n \cdot \delta_u x_n \big|_0^N + \frac{1}{2}\tau \sum_{n=0}^{N-1} \left(H_x^\tau(x_n^v, \lambda_{n+1}^v, v_n) - H_x^\tau(x_n^u, \lambda_{n+1}^u, u_n)\right) \cdot \delta_u x_n$$
$$+ \left(H_\lambda^\tau(x_n^v, \lambda_{n+1}^v, v_n) - H_\lambda^\tau(x_n^u, \lambda_{n+1}^u, u_n)\right) \cdot \delta_u \lambda_{n+1},$$

or,

$$\tau \sum_{n=0}^{N-1} \delta_u \lambda_{n+1} \cdot \delta_t \delta_u x_n = \frac{1}{2}\delta_u \lambda_n \cdot \delta_u x_n \big|_0^N + \frac{1}{2}\tau \sum_{n=0}^{N-1} \delta_u H_z^\tau|_n \cdot \delta_u z_n. \tag{3.3.24}$$

Remark. *Again the symplectic property of the discretization allows us to express this as the gradient of the Hamiltonian collocated at the numerical solution of the forward and backward equations, which in turn will allow cancellation with the second term of the Taylor expansion in* (3.3.27).

Combining (3.3.22), (3.3.23) and (3.3.24) gives

$$0 \equiv \delta_u \mathcal{I} = (\lambda_n^u + \frac{1}{2}\delta_u \lambda_n) \cdot \delta_u x_n \big|_0^N +$$
$$\tau \sum_{n=0}^{N-1} H_z^\tau(z_n^u, u_n) \cdot \delta_u z_n + \frac{1}{2}\delta_u H_z^\tau|_n \cdot \delta_u z_n - \delta_u H^\tau|_n - \delta_u h^\tau|_n. \tag{3.3.25}$$

Given that $\delta_u x_0 = 0$, the boundary term in (3.3.25) reduces to

$$(\lambda_N^u + \frac{1}{2}\delta_u \lambda_N) \cdot \delta_u x_N = -\Phi_x(x_N) \cdot \delta_u x_N - \frac{1}{2}\left(\Phi_x(x_N^v) - \Phi_x(x_N^u)\right) \cdot \delta_u x_N. \tag{3.3.26}$$

We substitute (3.3.15) and (3.3.18) into the second and third summand of (3.3.25), (3.3.26) into the boundary term, and subsequently the estimates (3.3.16) and (3.3.17)

to yield:

$$
\begin{aligned}
0 \equiv \delta_u \mathcal{I} = &- \left(\Phi(x_N^v) - \Phi(x_N^u) - \frac{1}{2}\delta_u x_N \cdot \Phi_{xx}(x_N^u + r_3 \delta_u x_N) \cdot \delta_u x_N \right) \\
&- \frac{1}{2} \left(\delta_u x_N \cdot \Phi_{xx}(x_N^u + r_2 \delta_u x_N) \cdot \delta_u x_N \right) + \tau \sum_{n=0}^{N-1} -\delta_u h^\tau|_n + H_z^\tau(z_n^u, u_n) \cdot \delta_u z_n \\
&\quad + \frac{1}{2} \left(\bar{\delta}_u H_z^\tau|_n \cdot \delta_u z_n + \delta_u z_n \cdot H_{zz}^\tau(z_n^u + r_1 \delta_u z_n, u_n) \cdot \delta_u z_n \right) \\
&\quad - \left(\bar{\delta}_u H^\tau|_n + H_z^\tau(z_n^u, v_n) \cdot \delta_u z_n + \frac{1}{2}\delta_u z_n \cdot H_{zz}^\tau(z_n^u + r_4 \delta_u z_n, v_n) \cdot \delta_u z_n \right),
\end{aligned}
$$

or,

$$
\begin{aligned}
\delta_u \Phi(x_N) + \tau \sum_{n=0}^{N-1} \delta_u h^\tau(x_n, u_n) = & \\
&- \frac{1}{2}\delta_u x_N \cdot (\Phi_{xx}(x_N^u + r_2 \delta_u x_N) - \Phi_{xx}(x_N^u + r_3 \delta_u x_N)) \cdot \delta_u x_N \\
&\qquad - \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n + \frac{1}{2}\tau \sum_{n=0}^{N-1} \bar{\delta}_u H_z^\tau|_n \cdot \delta_u z_n \\
&+ \frac{1}{2}\tau \sum_{n=0}^{N-1} \delta_u z_n \cdot (H_{zz}^\tau(z_n^u + r_1 \delta_u z_n, v_n) - H_{zz}^\tau(z_n^u + r_4 \delta_u z_n, v_n)) \cdot \delta_u z_n. \quad (3.3.27)
\end{aligned}
$$

Next, we use the estimates (3.3.13) and (3.3.14) and the fact that the quadratic terms are bounded by some constant K_3 to calculate

$$
\begin{aligned}
J^\tau[\boldsymbol{v}] - J^\tau[\boldsymbol{u}] \leq &- \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n, \\
&+ K_3 \|\delta_u x_N\|^2 + K_3 \tau \sum_{n=0}^{N-1} \left(\|\delta_u x_n\|^2 + \|\delta_u \lambda_{n+1}\|^2 \right), \\
&+ \frac{1}{2}\tau \sum_{n=0}^{N-1} \|\delta_u x_n\| \|\bar{\delta}_u f^\tau|_n\| + \frac{1}{2}\tau \sum_{n=0}^{N-1} \|\delta_u \lambda_{n+1}\| \|\bar{\delta}_u H_x^\tau|_n\|, \\
\leq &- \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n + C \left(\tau \sum_{n=0}^{N-1} \|\bar{\delta}_u f^\tau|_n\| \right)^2 + C \left(\tau \sum_{n=0}^{N-1} \|\bar{\delta}_u H_x^\tau|_n\| \right)^2, \\
\leq &- \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n + C\tau \sum_{n=0}^{N-1} \|\bar{\delta}_u f^\tau|_n\|^2 + C\tau \sum_{n=0}^{N-1} \|\bar{\delta}_u H_x^\tau|_n\|^2,
\end{aligned}
$$

which is the result sought (cf. (3.3.19)).

It now remains to show that the regularized forward-backward sweep iteration converges. We first show that an estimate of the same form as (3.3.19) holds for $\delta_u H^\tau$ when the regularized Hamiltonian is maximized. These can be combined to show monotone decay

of the objective function $J^\tau[\boldsymbol{u}]$. Thereafter, it is shown that the sum of the decrements is finite, which implies convergence of the differences.

Let $\boldsymbol{v}$ denote the improved control obtained by solving (3.3.4). The resulting change in $\tilde{H}^\tau$ must be nonnegative, hence

$$\begin{aligned} 0 \le \tau \sum_{n=0}^{N-1} \bar{\delta}_u \tilde{H}^\tau|_n = \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n \\ - \frac{\rho}{2}\left[\|\frac{x^u_{n+1}-x^u_n}{\tau} - f^\tau(x^u_n, v_n)\|^2 + \|\frac{\lambda^u_{n+1}-\lambda^u_n}{\tau} + H^\tau_x(x^u_n, \lambda^u_{n+1} v_n)\|^2\right] \\ + \frac{\rho}{2}\left[\|\frac{x^u_{n+1}-x^u_n}{\tau} - f^\tau(x^u_n, u_n)\|^2 + \|\frac{\lambda^u_{n+1}-\lambda^u_n}{\tau} + H^\tau_x(x^u_n, \lambda^u_{n+1}, u_n)\|^2\right]. \end{aligned} \tag{3.3.28}$$

The last term in square brackets vanishes since x^u_n and λ^u_n satisfy (3.2.32)–(3.2.33). Consequently, the above expression is equivalent to

$$0 \le \tau \sum_{n=0}^{N-1} \bar{\delta}_u \tilde{H}^\tau|_n = \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n - \frac{\rho}{2}\left[\|\bar{\delta}_u f^\tau|_n\|^2 + \|\bar{\delta}_u H^\tau_x|_n)\|^2\right]. \tag{3.3.29}$$

Combining this with Lemma 2 gives

$$J^\tau[\boldsymbol{v}] - J^\tau[\boldsymbol{u}] \le -(1 - \frac{2C}{\rho})\tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n. \tag{3.3.30}$$

The summation on the right side is nonnegative, as a consequence of (3.3.29) . Therefore, choosing $\rho > 2C$ ensures that J^τ is nonincreasing. Next suppose we iterate (3.3.2)–(3.3.4). Let $\boldsymbol{u}^{(k)}$ denote the control variable in iteration k. Then it holds that

$$\sum_{k=0}^{M} \tau \sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n^{(k)} \le D^{-1}(J^\tau[\boldsymbol{u}^{(0)}] - J^\tau[\boldsymbol{u}^{(M+1)}]) \le D^{-1}(J^\tau[\boldsymbol{u}^{(0)}] - \inf_{\boldsymbol{u}\in\mathcal{U}} J^\tau[\boldsymbol{u}]),$$

where $D = (1 - 2C/\rho) > 0$. Consequently, in the limit $M \to \infty$ this sum is bounded, which implies

$$\sum_{n=0}^{N-1} \bar{\delta}_u H^\tau|_n \to 0,$$

proving convergence of the iteration.

3.4 Numerical illustration

In this section we study numerically the convergence of the discrete regularized forward-backward sweep iteration. As a test problem we control the motion of a damped oscillator in a double well potential. The controlled motion is given by

$$x = \begin{pmatrix} q \\ p \end{pmatrix}, \qquad f(x,u) = \begin{pmatrix} p \\ q - q^3 - \nu p + u \end{pmatrix}, \tag{3.4.1}$$

where $\nu > 0$ is a damping parameter. The control $u(t)$ acts only on the velocity. As initial condition we choose $\xi = (-1, 0)$ in the left potential well, and we seek to minimize the cost function

$$J[u] = \frac{\alpha}{2}\|x(T) - x_f\|^2 + \int_0^T \frac{1}{2}u(t)^2\, dt, \tag{3.4.2}$$

where the target final position is $x_f = (1, 0)$, in the right potential well. For the numerical computations we take $T = 6$, $\nu = 1$, and $\alpha = 10$.

We solve the optimal control problem using the discrete regularized forward-backward sweep iteration (3.3.2)–(3.3.4) and the symplectic Euler scheme (3.2.23)–(3.2.25). We iterate until the update to the control variable u is less than a prescribed tolerance

$$\sum_{n=0}^{N-1} \|u_n^{(k)} - u_n^{(k-1)}\| < \varepsilon,$$

where $\varepsilon = 1e^{-8}$. The computed optimal path $x(t) = (q(t), p(t))$ is shown as a solid blue curve on the left plot of Figure 3.1. The background contours are level sets of the total energy function $E = \frac{1}{2}p^2 + \frac{1}{4}q^4 - \frac{1}{2}q^2$. The optimal control must accelerate the motion of the particle to reach an energy level above the saddle point, allowing it to cross to the potential well on the right.

For this computation we chose $\rho = 100$ for the regularization parameter. Convergence occurs in 4206 iterations. Figure 3.2 shows the discrete cost function (3.2.4) during the first 2000 iterations for values $\rho = 50$, $\rho = 100$ and $\rho = 200$. For $\rho = 100$, the convergence is monotone as predicted by the theory of the previous section (cf. (3.3.30)). For $\rho = 50$, we observe an initial reduction in cost, which eventually oscillates and does not converge. For $\rho = 200$, the iteration converges but at a slower rate than for $\rho = 100$. Hence, our experience suggests there is a critical value of ρ below which there is no convergence of the regularized forward-backward sweep iteration, and above which the convergence becomes steadily slower.

The minimal cost obtained using the symplectic Euler method and $N = 160$ was $J = 0.7712$. We also computed the optimal solution for $N = 20$ time steps, shown as the red dash-dot line in the left plot of Figure 3.1. As noted in Section 3.2, by discretizing the Lagrangian we obtain a discrete optimal control problem for each N. For the case $N = 20$ the optimal path deviates significantly from that for $N = 160$. Because the Lipschitz constant is larger for this solution, it was necessary to take $\rho = 400$ for convergence. The optimal cost in the case $N = 20$ is $J = 0.7006$, which is less than the optimal cost obtained in the case $N = 160$.

We also solved the optimal control problem using the implicit midpoint rule, a second order symplectic Runge-Kutta method with $s = 1$ and coefficients $a_{11} = b_1 = 1/2$. The solutions for $N = 20$ and $N = 160$ are shown in the right plot of Figure 3.1. Here we see that the discrete optimum at low resolution is much closer to that at high resolution. The optimal costs were computed $J = 0.7837$ for $N = 20$ and $J = 0.7769$ for $N = 160$. Both resolutions converged with $\rho = 100$.

Although the convergence is monotone in the cost J for large enough ρ, the forward-backward sweep iteration may require a large number of iterations to attain a sufficiently

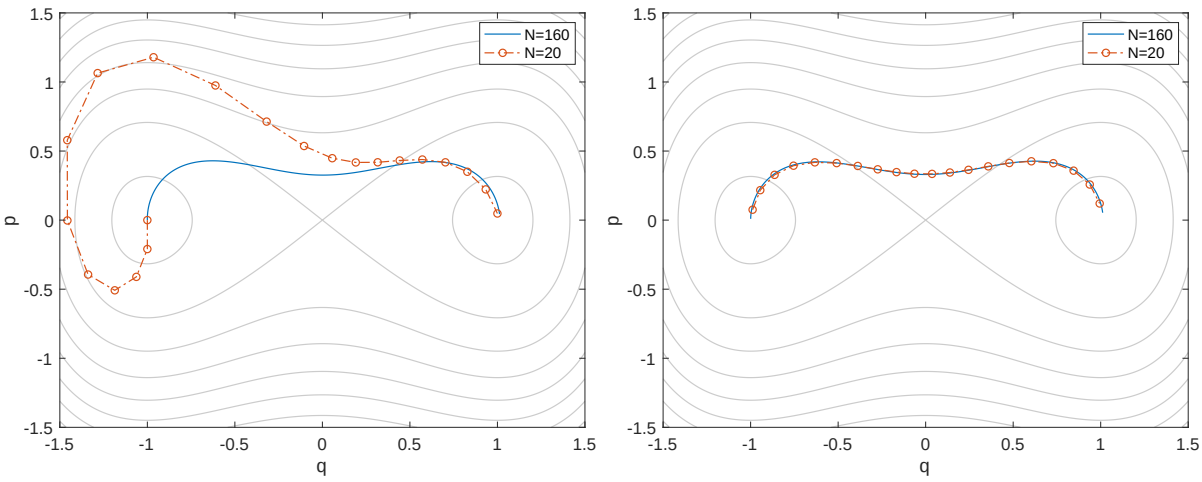


FIGURE 3.1: Optimal motion in q–p plane, computed with the symplectic Euler method (left) and implicit midpoint method (right), for $N = 160$ (solid blue line) and $N = 20$ (dash-dot red line).

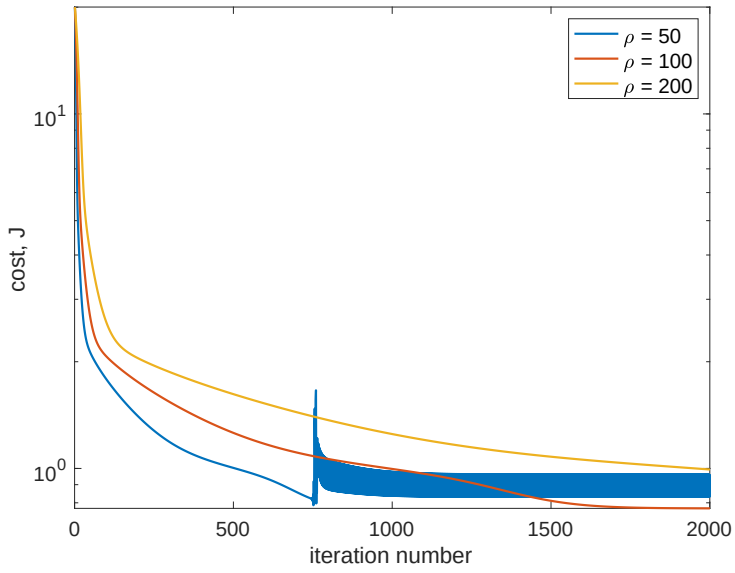


FIGURE 3.2: Convergence of the cost function for the regularized forward-backward sweep iteration using the symplectic Euler method (3.2.23)–(3.2.25), with $\rho = 50$ (blue), $\rho = 100$ (red) and $\rho = 200$ (yellow).

small cost. Acceleration techniques such as Anderson acceleration [119] may be employed to improve the convergence rate. We implement (3.3.2)–(3.3.4) as a fixed point iteration on the control function $\boldsymbol{u}$, i.e. $\boldsymbol{u}^{(k+1)} = \mathcal{F}(\boldsymbol{u}^{(k)})$. Subsequently we apply Anderson acceleration with restarts every three iterations. In Figure 3.3 we see that the cost function converges in 221 iterations (nearly a factor 20 fewer), but the cost no longer decays monotonically. See [62] for a more sophisticated strategy with adaptive damping and preserving monotonicity. In our experience the choice of a good acceleration algorithm depends heavily on the problem.

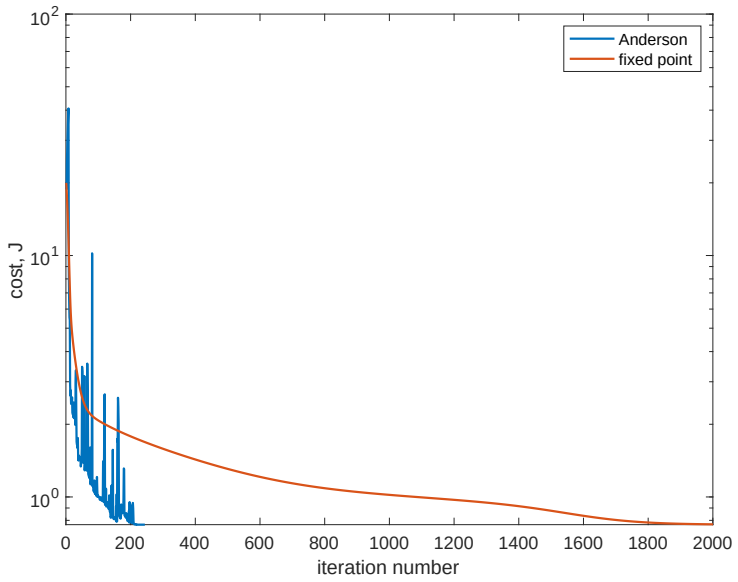


FIGURE 3.3: Comparison of the Anderson accelerated (blue) and fixed point (red) iterations. Shown are the cost functions using the symplectic Euler method (3.2.23)–(3.2.25), with $\rho = 100$.

3.5 Summary

In this chapter we have extended the convergence proof of a regularized forward-backward sweep iteration [77] for solving optimal control problems to the discrete setting. We showed that if the continuous problem is discretized by a symplectic partitioned Runge-Kutta pair (using a variational integrator approach), then the convergence proof of [77] may be easily adapted. Numerical experiments with the first order, explicit symplectic Euler method and the second order implicit midpoint rule demonstrate monotonic convergence of the cost function if the regularization parameter ρ is chosen large enough. For insufficiently large ρ the cost undergoes bounded oscillations; whereas for excessively large ρ the convergence is slower. In our experiments, convergence was observed even with large step sizes, however the resulting discrete optimization problem is an inaccurate approximation of the continuous problem. In an efficient implementation, the regularized forward-backward sweep iteration may be combined with an acceleration techniques for nonlinear iterations such as Anderson acceleration [119].

Appendix

In this appendix we prove that the bounds (3.3.8) follow from (3.1.3).

Since $b_i \geq 0$, $i = 1, \ldots, s$,

$$\|f^\tau(x', u) - f^\tau(x, u)\| \leq \sum_{i=1}^{s} b_i \|f(X_i, U_i) - f(X_i', U_i)\|, \tag{3.5.1}$$

where X_i' satisfies

$$X_i' = x' + \tau \sum_{j=1}^{s} a_{ij} f(X_j', U_j).$$

Denoting $\Delta X_i = X_i - X_i'$ and using the Lipschitz condition on f (cf. (3.1.3)), we find

$$\|\Delta X_i\| \leq \|x - x'\| + \tau \sum_{j=1}^{s} |a_{ij}| \cdot K \|\Delta X_j\|.$$

Denote by $|A|$ the matrix with elements $|a_{ij}|$, by $|\Delta X|$ the vector with elements $\|\Delta X_i\|$, and let $\mathbb{1}$ be the vector of dimension s with all elements equal to 1. Then the above inequality becomes

$$(I - \tau K |A|)|\Delta X| \leq \|x - x'\| \mathbb{1}. \tag{3.5.2}$$

For *explicit* Runge-Kutta methods, the matrix on the left always has positive inverse given by

$$(I - \tau K |A|)^{-1} = \sum_{i=0}^{s-1} (\tau K |A|)^i.$$

For *implicit* Runge-Kutta methods, the matrix on the left of (3.5.2) is an M-matrix with positive inverse if we impose the step size restriction

$$\tau \leq (K \max_{ij} |a_{ij}|)^{-1}. \tag{3.5.3}$$

In either of the above cases we find

$$\|X_i - X_i'\| \leq K^\tau \|x - x'\|, \qquad K^\tau = \|(I - \tau K |A|)^{-1} \mathbb{1}\|_\infty. \tag{3.5.4}$$

Returning to (3.5.1) we obtain

$$\|f^\tau(x', u) - f^\tau(x, u)\| \leq \sum_{i=0}^{s} b_i K K^\tau \|x - x'\| = K K^\tau \|x - x'\|.$$

proving the first bound in (3.3.8).

To prove the second bound, recall (3.2.35). Taking norms, and using the bound (3.1.3),

$$\|\Psi_i\| \leq 1 + \tau \sum_{j=1}^{s} |a_{ij}| K \|\Psi_j\|,$$

from which we conclude that

$$\|\Psi_i\| \le K^\tau. \tag{3.5.5}$$

We also find

$$\begin{aligned}
\|\Psi_i - \Psi_i'\| &\le \tau \sum_{j=1}^{s} |a_{ij}| \|f_x(X_j, U_j)\Psi_j - f_x(X_j', U_j)\Psi_j'\| \\
&= \tau \sum_{j=1}^{s} |a_{ij}| \|f_x(X_j, U_j)(\Psi_j - \Psi_j') + (f_x(X_j, U_j) - f_x(X_j', U_j))\Psi_j'\| \\
&\le \tau \sum_{j=1}^{s} |a_{ij}| (K\|\Psi_j - \Psi_j'\| + KK^\tau \|X_j - X_j'\|) \\
&\le \tau \sum_{j=1}^{s} |a_{ij}| (K\|\Psi_j - \Psi_j'\| + K(K^\tau)^2 \|x - x'\|) \\
&\le \tau (\max_i \sum_{j=1}^{s} |a_{ij}|) K(K^\tau)^3 \|x - x'\|,
\end{aligned}$$

where the last inequality follows by inverting the matrix of (3.5.2)—in the case of implicit RK methods under the step size restriction (3.5.3). Similarly, we compute

$$\begin{aligned}
\|f_x^\tau(x, u) - f_x^\tau(x', u)\| &\le \sum_{i=1}^{s} b_i \|f_x(X_i, U_i)\Psi_i - f_x(X_i', U_i)\Psi_i'\| \\
&= \sum_{i=1}^{s} b_i \|f_x(X_i, U_i)(\Psi_i - \Psi_i') + (f_x(X_i, U_i) - f_x(X_i', U_i))\Psi_i'\| \\
&\le \sum_{i=1}^{s} b_i (K\|\Psi_i - \Psi_i'\| + KK^\tau \|X_i - X_i'\|) \\
&\le (\tau \max_i \sum_{j=1}^{s} |a_{ij}|) K^2 (K^\tau)^3 + K(K^\tau)^2) \|x - x'\|,
\end{aligned}$$

proving the second bound in (3.3.8).

The bounds on h^τ and h_x^τ in (3.3.8) follow the same reasoning.

Chapter 4

Accelerated convergence of the regularized maximum principle for optimal control of the Cucker-Smale model

Abstract

In this chapter, we investigate numerically the convergence properties of the regularized forward-backward sweep method in the context of consensus forming in the Cucker-Smale model. Using Anderson acceleration, we observe that the fast convergence is possible, but depends on the norm used in the cost function. For sparse control in the ℓ_1 norm, convergence may be very slow. Regularization of the norm alleviates the slow convergence to some degree.

This chapter is transcripted from the paper "Accelerated convergence of the regularized maximum principle for optimal control of the Cucker-Smale model" submitted to the Journal of Optimal Control Applications and Methods.

4.1 Background

In recent article [78], we discussed the numerical implementation of a regularized forward-backward sweep iteration for solving optimal control problems, originally proposed by Li, Chen, Tai & E [77]. The method makes use of only local in time information in the form of forward and backward time integration sweeps, can be parallelized in the optimization step, and the convergence can be proved under the condition that a symplectic Runge-Kutta pair (or variational integrator) is employed to discretize the regularized Euler-Lagrange equations [78]. We also observed that the rate of convergence could be improved by complementing the iteration with an acceleration technique for nonlinear iterations, such as Anderson acceleration [113].

In this chapter we report on further experiments with the regularized forward-backward sweep method in the context of sparse optimal control of multi-agent systems.

Self-organization is an interesting phenomenon in multi-agent systems, which commonly appears in biological, economical and social groups. In recent years some mathematical models, see e.g. [5], are proposed to simulate this kind of behavior. Of particular interest is the Cucker-Smale model proposed in [32, 33], which represents synchronized motion as observed in flocks of birds and schools of fish. Within the model, agents in d-dimensional space adapt their velocity vectors towards that of their neighbors[56, 55] naturally. The Cucker-Smale model has also been suggested as a model of consensus forming.

However, the Cucker-Smale dynamics does not converge to consensus for all initial conditions. When consensus is not reached under the free dynamics, an extra intervention, e.g., a control function, may be used to impose consensus, see [20]. Since the focus on multi-agent systems typically assumes large populations, it is difficult or undesirable in practice to attempt to continuously control all agents. Hence, the recent work has addressed sparsity constraints on the control, striving for most components of the control to be zero most of the time. One means of achieving sparse optimal control is through the use of ℓ_1-norms in the cost function, which in turn leads to discontinuous controls and the need for upper bound constraints on the admissibility set. Consequently we study the convergence of the regularized forward-backward sweep iteration for constrained and discontinuous controls.

The model consists of M interacting agents, each characterized by its position $x_i(t) \in \mathbf{R}^d$ and velocity $v_i \in \mathbf{R}^d$, $i = 1, \ldots, M$. The dynamics is governed by

$$\dot{x}_i = v_i, \quad i = 1, \ldots, M \tag{4.1.1}$$

$$\dot{v}_i = \sum_{j \neq i} \phi(\|x_j - x_i\|)(v_j - v_i), \quad i, j = 1, \ldots, M \tag{4.1.2}$$

$$x_i(0) = x_{i0}, \quad v_i(0) = v_(i0), \tag{4.1.3}$$

where $\phi : \mathbf{R}^+ \to \mathbf{R}^+$ is a bounded, non-increasing, continuous function that represents the influence of an agent. Generally, ϕ is taken to be

$$\phi(r) = \frac{1}{(1 + r^2)^\beta}, \tag{4.1.4}$$

where $\beta > 0$ is a parameter that affects the decay rate of the radius of influence of agents. The system is said to reach 'consensus' if the velocities v_i of individual agents converge to a common vector. Specifically, defining

$$V(t) = \frac{1}{2M^2} \sum_{i,j} \|v_i(t) - v_j(t)\|^2, \qquad X(t) = \frac{1}{2M^2} \sum_{i,j} \|x_i(t) - x_j(t)\|^2. \tag{4.1.5}$$

it has been shown that the system converges asymptotically to the consensus state $V = 0$, i.e.,

$$V(t) \leq V(0)e^{-\kappa t}, \quad \kappa > 0$$

in two situations. First, under significantly slow decay of influence $\beta \leq \frac{1}{2}$, consensus will be reached from arbitrary initial condition, with the rate κ depending on the initial condition [32, 33, 56]. Second, for $\beta > \frac{1}{2}$, consensus conditionally will be reached if the initial condition satisfies:

$$\|v(0)\| < \frac{1}{2\beta - 1}(1 + \|x(0)\|^2)^{\frac{1}{2}-\beta}$$

When the conditions for consensus given above are not met, the Cucker-Smale model generally does not reach consensus, and several authors have employed the model to study how consensus can be influenced by an outside agent. Of particular interest is the possibility of a sparse controller who exerts influence only locally within the population or for short intervals of time. To this end, we consider a system in which each agent is additionally subjected to an external control $u_i(t) \in \mathbf{R}^d$. The time evolution of the state is governed by

$$\dot{x}_i = v_i, \tag{4.1.6}$$

$$\dot{v}_i = \sum_{j \neq i} \phi(\|x_j - x_i\|)(v_j - v_i) + u_i, \tag{4.1.7}$$

The objective is to find admissible controls to steer the system into the consensus region in finite time. The paper [20] considers consensus stabilization for the Cucker-Smale system by means of both feedback-based controllers and sparse optimal control. The paper [14] studied different variations of feedback control structure for consensus stabilization. Local control based on instantaneous feedback models the more realistic situation where the policymaker is not omniscient. The optimal control problem presented in the next section describes a model where the policy maker is allowed to see how the dynamics can develop. Balio et al. [7] investigated numerical realization of optimal consensus control for the Cucker-Smale model.

This chapter is organized as follows: In Section 2, we formulate the maximum principle optimal control problem for the Cucker-Smale model and recall the regularized forward-backward sweep iteration method proposed in [77], whose numerical implementation is discussed in [78]. In Section 3, we discuss the maximization of the regularized Hamiltonian in three different norms on the control function. Section 4 discusses a simple example to illustrate the need for upper bound constraints on the set of admissible controls. In Section 4 we study the convergence of the regularized forward-backward sweep iteration with Anderson acceleration, including dependence on norm in the cost function and method parameters.

4.2 Optimal control of the Cucker-Smale model

Control sparsity can be attained by minimizing the cost of the control with respect to the ℓ_1-norm [31]. Recalling the vector notation $\mathbf{x}(t) = (x_1(t), x_2(t), \cdots, x_M(t)) \in \mathbf{R}^{dM}$, $\mathbf{v}(t) = (v_1(t), v_2(t), \cdots, v_M(t)) \in \mathbf{R}^{dM}$, $\boldsymbol{u}(t) = (u_1(t), u_2(t), \cdots, u_M(t)) \in \mathbf{R}^{dM}$ in pre Cucker-Smale model interpretion in preliminaries part, we introduce the norms $\|\cdot\|_{q,p}$ to indicate a mixed ℓ_q-ℓ_p-norm that is ℓ_q in $\mathbf{R}^d$ and ℓ_p in $\mathbf{R}^M$:

$$\|\mathbf{x}\|_{q,p} = \left(\sum_{i=1}^{M} (\|x_i\|_q)^p \right)^{1/p}. \tag{4.2.1}$$

(When not stated explicitly, $\|\cdot\|$ denotes the 2-norm in this article.)

We consider the optimal control problem of determining a trajectory of (4.1.6)–(4.1.7), with initial condition $(\mathbf{x}(0), \mathbf{v}(0)) = (\mathbf{x}_0, \mathbf{v}_0)$, which minimizes a cost functional that penalizes distance to consensus and magnitude of the control in the mixed ℓ_q-ℓ_p-norm. More precisely, the cost functional considered here is, for a given $\gamma > 0$,

$$J[u] = \int_0^T \sum_{i=1}^{M} \frac{1}{2} \|v_i - \bar{v}\|^2 + \frac{\gamma}{p} \sum_{i=1}^{M} \|u_i\|_q^p dt. \tag{4.2.2}$$

(We emphasize that the second term is equivalent to $\|\boldsymbol{u}(t)\|_{q,p}^p$.)

The Pontryagin Maximum Principle (see [92]) provides a necessary condition for the existence of an optimal control. We make use of the Hamiltonian formulation, with Hamiltonian functional

$$H(\mathbf{x}, \mathbf{v}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{u}) = \sum_{i=1}^{M} \left[\lambda_i^T v_i + \mu_i^T \left(\sum_{j=1}^{M} \phi(\|x_j - x_i\|)(v_j - v_i) + u_i \right) \right] - \sum_{i=1}^{M} \frac{1}{2} \|v_i - \bar{v}\|^2 - \frac{\gamma}{p} \sum_{i=1}^{M} \|u_i\|_q^p, \tag{4.2.3}$$

where $\boldsymbol{\lambda}(\cdot) = (\lambda_1, \cdots, \lambda_M) \in \mathbf{R}^{Md}$ and $\boldsymbol{\mu}(\cdot) = (\mu_1, \cdots, \mu_M) \in \mathbf{R}^{Md}$ are adjoint variables associated to $\mathbf{x}(t)$ and $\mathbf{v}(t)$, respectively, and satisfying

$$\dot{\lambda}_i = -\sum_{j=1}^{M} \frac{\phi'(\|x_j - x_i\|)}{\|x_j - x_i\|} \langle x_j - x_i, v_j - v_i \rangle (\mu_j - \mu_i), \tag{4.2.4}$$

$$\dot{\mu}_i = -\lambda_i - \sum_{j=1}^{M} \phi(\|x_j - x_i\|)(\mu_j - \mu_i) + (v_i - \bar{v}), \tag{4.2.5}$$

$$\lambda_i(T) = 0, \quad \mu_i(T) = 0, \quad i = 1, \cdots, M. \tag{4.2.6}$$

The maximum principle states that an optimal control $\boldsymbol{u}$ maximizes the Hamiltonian H, among solutions of Hamilton's equations. For convenience of notation, let us define augmented vectors $z_i = (x_i, v_i) \in \mathbf{R}^{2d}$, $\nu_i = (\lambda_i, \mu_i) \in \mathbf{R}^{2d}$, $\boldsymbol{z} = (z_i, i = 1 \ldots, M) \in \mathbf{R}^{2dM}$

and $\boldsymbol{\nu} = (\nu_i, i = 1, \ldots, M) \in \mathbf{R}^{2dM}$. In these variables, the maximum principle states that an optimal control $\boldsymbol{u}(t)$ satisfies

$$\dot{\boldsymbol{z}} = \frac{\partial H}{\partial \boldsymbol{\nu}}(\boldsymbol{z}, \boldsymbol{\nu}, \boldsymbol{u}), \qquad \boldsymbol{z}(0) = \boldsymbol{z}_0, \tag{4.2.7}$$

$$\dot{\boldsymbol{\nu}} = -\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{z}, \boldsymbol{\nu}, \boldsymbol{u}), \qquad \boldsymbol{\nu}(T) = 0, \tag{4.2.8}$$

$$\boldsymbol{u}(t) = \arg\max_{\boldsymbol{u} \in \mathcal{U}^n} H(\boldsymbol{z}(t), \boldsymbol{\nu}(t), \boldsymbol{u}), \qquad \forall t \in (0, T). \tag{4.2.9}$$

Solution of (4.2.7)–(4.2.9) is a boundary value optimization problem due to the initial and terminal conditions (4.1.3) and (4.2.6). One approach is to iterate, successively solving (4.2.7) for the $z_i(t)$ forward in time, (4.2.8) for the $\nu_i(t)$ backward in time, and (4.2.9) for the $u_i(t)$. Such a method of successive approximations will typically diverge unless an initial guess near the optimal $\boldsymbol{u}$ is known. In [77] the authors proposed a regularized forward-backward sweep for solving the Pontryagin equations, and they proved the global convergence of the iteration in the continuous case. In [78] we extended the proof to numerical discretization by symplectic Runge-Kutta pairs. Li et al. [77] introduced the extended Hamiltonian functional

$$\tilde{H}(\boldsymbol{z}, \boldsymbol{\nu}, \boldsymbol{u}, \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\nu}}) = H(\boldsymbol{z}, \boldsymbol{\nu}, \boldsymbol{u}) - \frac{\rho}{2}\left(\left\|\tilde{\boldsymbol{z}} - \frac{\partial H}{\partial \boldsymbol{\nu}}\right\|^2 + \left\|\tilde{\boldsymbol{\nu}} + \frac{\partial H}{\partial \boldsymbol{z}}\right\|^2\right), \tag{4.2.10}$$

where $\rho > 0$ is a regularization parameter. Solutions to equations (4.2.7)–(4.2.9) are approximated by successively solving, in the kth iteration,

$$\dot{\boldsymbol{z}}^{(k+1)} = \frac{\partial \tilde{H}}{\partial \boldsymbol{\nu}}(\boldsymbol{z}^{(k+1)}, \boldsymbol{\nu}^{(k)}, \boldsymbol{u}^{(k)}, \dot{\boldsymbol{z}}^{(k+1)}, \dot{\boldsymbol{\nu}}^{(k)}), \qquad \boldsymbol{z}^{(k+1)}(0) = \boldsymbol{z}_0, \tag{4.2.11}$$

$$\dot{\boldsymbol{\nu}}^{(k+1)} = -\frac{\partial \tilde{H}}{\partial \boldsymbol{z}}(\boldsymbol{z}^{(k+1)}, \boldsymbol{\nu}^{(k+1)}, \boldsymbol{u}^{(k)}, \dot{\boldsymbol{z}}^{(k+1)}, \dot{\boldsymbol{\nu}}^{(k+1)}), \qquad \boldsymbol{\nu}^{(k+1)}(T) = 0, \tag{4.2.12}$$

$$\boldsymbol{u}^{(k+1)}(t) = \arg\max_{\boldsymbol{u} \in \mathcal{U}^n} \tilde{H}(\boldsymbol{z}^{(k+1)}(t), \boldsymbol{\nu}^{(k+1)}(t), \boldsymbol{u}, \dot{\boldsymbol{z}}^{(k+1)}(t), \dot{\boldsymbol{\nu}}^{(k+1)}(t)), \quad \forall t \in (0, T). \tag{4.2.13}$$

It can be checked that the added term to the augmented Hamiltonian (4.2.10) is identically zero along solutions to (4.2.7)–(4.2.8). Hence (4.2.11)–(4.2.12) are unmodified, whereas the maximization (4.2.13) ensures that updates to the control function remain close to solutions of Hamilton's equations [77]. For significantly large ρ, and appropriate Lipschitz conditions, Li et al. prove convergence of the iteration (4.2.11)–(4.2.13) in the continuous case.

In the kth iteration of (4.2.11)–(4.2.13), having performed forward and backward integrations for fixed $\boldsymbol{u}$, the discrete optimization step (4.2.13) is carried out for fixed $\boldsymbol{z}$ and $\boldsymbol{\nu}$. Consequently, the Hamiltonian (4.2.3) can be written

$$H(\boldsymbol{u}) = \sum_i \mu_i^T u_i - \frac{\gamma}{p}\|u_i\|_q^p + \bar{H},$$

where $\bar{H}$ collects terms that are constant during the optimization step. Similarly, since $\boldsymbol{z}^{(k+1)}$ and $\boldsymbol{\nu}^{(k+1)}$ are solutions of (4.2.11) and (4.2.12) for $\boldsymbol{u} = \boldsymbol{u}^{(k)}$, the optimization step (4.2.13) can be expressed as

$$\boldsymbol{u}^{(k+1)}(t) = \arg\max_{\boldsymbol{u}\in\mathcal{U}^n} H(\boldsymbol{u}) - \frac{\rho}{2}\left(\left\|\frac{\partial H}{\partial \boldsymbol{\nu}}(\boldsymbol{u}^{(k)}) - \frac{\partial H}{\partial \boldsymbol{\nu}}(\boldsymbol{u})\right\|^2 + \left\|\frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{u}^{(k)}) - \frac{\partial H}{\partial \boldsymbol{z}}(\boldsymbol{u})\right\|^2\right).$$

Explicitly, and ignoring constant terms, the optimization step becomes

$$\boldsymbol{u}^{(k+1)}(t) = \arg\max_{\boldsymbol{u}\in\mathcal{U}^n} (\boldsymbol{\mu}^{(k+1)})^T\boldsymbol{u} - \frac{\gamma}{p}\|\boldsymbol{u}\|_{q,p}^p - \frac{\rho}{2}\|\boldsymbol{u}^{(k)} - \boldsymbol{u}\|^2. \tag{4.2.14}$$

In [78] we showed that the convergence proof of [77] extends in a straightforward manner to the numerical discretization, *if* (4.2.11)–(4.2.13) *are discretized using a symplectic Runge-Kutta method or symplectic partitioned Runge-Kutta pair.*

Discretization yields a set of variables x_i^j, v_i^j, λ_i^j, $\mu_i^j \in \mathbf{R}^d$, $i = 1, \ldots, M$, $j = 0, \ldots, J$, where j denotes the time index. Within each time step, the Runge-Kutta method makes use of a set of internal stage variables $X_i^{j,\ell}$, $V_i^{j,\ell}$, $\Lambda_i^{j,\ell}$, $M_i^{j,\ell}$, $U_i^{j,\ell}$, where $\ell = 1, \ldots, L$ is the stage index of an L-stage Runge-Kutta method. The control variables $U_i^{j,\ell}$ appear only as internal stage variables. For details of the symplectic Runge-Kutta implementation, we refer the reader to [78].

The discrete form of (4.2.14) becomes

$$U_i^{j,\ell} = \arg\max_{U\in\mathcal{U}} (M_i^{j,\ell})^T U - \frac{\gamma}{p}\|U\|_q^p - \frac{\rho}{2}\|U - (U_i^{j,\ell})^{(k)}\|^2, \qquad \forall i, j, \ell, \tag{4.2.15}$$

where, due to the additive nature, this optimization may be performed independently for all i, j, and ℓ.

4.3 Considerations for solving the Hamiltonian optimization step

In this section we discuss the exact solutions of the decoupled optimization problems (4.2.15), depending on the choices of ℓ_q-ℓ_p-norm. When considering sparse control, different norms have been employed in the literature. The ℓ_1-ℓ_1-norm, considered in [6, 31, 20], which acts on a few agents over a finite time frame, presents a challenge due to the lack of smoothness of the cost functional. The mixed norm ℓ_2-ℓ_1 has been employed in [20], where the authors present a full analysis of optimal sparse control solutions, distinguishing five regions.

4.3.1 Constrained control functions

The Hamiltonian maximization step (4.2.15) decouples into a collection of independent optimization problems on $\mathbf{R}^d$, in each of which we seek to maximize

$$A[u] = (\mu^T u - \frac{\gamma}{p}\|u\|_q^p - \frac{\rho}{2}\|u - \bar{u}\|_2^2), \tag{4.3.1}$$

where γ, $\mu \in \mathbf{R}^d$ and $\bar{u} \in \mathbf{R}^d$ are treated as parameters. In the following we discuss each of the ℓ_2-ℓ_2, ℓ_1-ℓ_1 and ℓ_2-ℓ_1-norms separately.

Maximum in the ℓ_2-ℓ_2-norm.

For $p = q = 2$, the function $A[u]$ is convex and differentiable. The maximum of (4.3.1) is attained at u^* solving $\nabla A[u^*] = 0$, i.e.,

$$u^* = \frac{\mu + \rho\bar{u}}{\gamma + \rho}.$$

Maximum in the ℓ_1-ℓ_1-norm

For $q = p = 1$, the function (4.3.1) becomes

$$A[u] = \mu u - \gamma\|u\|_1 - \frac{\rho}{2}\|u - \bar{u}\|_2^2. \tag{4.3.2}$$

Let $u^{(j)}$ denote the jth component of u, $1 \le j \le d$. Then maximizing $A[u]$ in (4.3.2) amounts to maximizing a sum of scalar functions

$$\bar{A}[u^{(j)}], \quad j = 1, \dots, d,$$

where each of the $\bar{A}$ takes the form

$$\bar{A}[\eta] = \kappa\eta - \gamma|\eta| - \frac{\rho}{2}(\eta - \bar{\eta})^2, \tag{4.3.3}$$

for scalar η and known scalar parameters κ and $\bar{\eta}$.

For the particular case $\rho = 0$ we have

$$\bar{A}[\eta] = \kappa\eta - \gamma|\eta|. \tag{4.3.4}$$

If $\kappa > \gamma > 0$, then $\bar{A}[\eta]$ has no maximum. Hence, it is necessary to bound the control from above. In this paper we add a constraint on each component of the control, i.e., we choose the admissible control set $\mathcal{U} = \{u \in \mathbf{R}^d \,|\, \|u\|_\infty \le u_{\max}\}$.

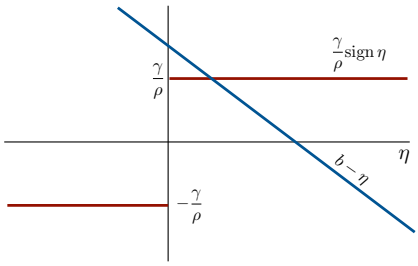
For general ρ, the function (4.3.3) is not differentiable at $\eta = 0$. For $\eta \neq 0$, its gradient is

$$\bar{A}'[\eta] = \kappa - \gamma\,\mathrm{sign}(\eta) - \rho(\eta - \bar{\eta})$$

If the optimum η^* satisfies $\bar{A}'[\eta^*] = 0$, then

$$\frac{\gamma}{\rho}\mathrm{sign}(\eta^*) = (\frac{\kappa}{\rho} + \bar{\eta}) - \eta^*. \tag{4.3.5}$$

Let $b = \frac{\kappa}{\rho} + \bar{\eta}$. The functions on the left and right sides of (4.3.5) are shown in the following plot.



The optimum η^* is given by

$$\eta^* = \begin{cases} b - \frac{\gamma}{\rho}, & b > \frac{\gamma}{\rho}, \\ b + \frac{\gamma}{\rho}, & b < -\frac{\gamma}{\rho}, \\ 0, & \text{otherwise}, \end{cases} \tag{4.3.6}$$

where the last condition follows because for $|b| < \frac{\gamma}{\rho}$, it can be checked that $\bar{A}[\eta] < \bar{A}[0]$ for all nonzero η. The optimal control is the argument maximizing $\bar{A}[\eta]$ over the set $\{\eta^*, M, -M\}$.

Maximum in the ℓ_2-ℓ_1-norm

For $\rho = 0$, and similar to the ℓ_1-ℓ_1 case, if for some j it holds that $|\mu^j| > \gamma$, the optimal control is unbounded. Therefore, we need to impose the same constraint on u as in the ℓ_1-ℓ_1 case.

For general ρ, when $u \neq 0$ the gradient condition $\nabla A[u] = 0$ yields

$$\mu - \gamma \frac{u}{\|u\|} - \rho(u - \bar{u}) = 0,$$

or

$$\left(1 + \frac{\gamma}{\rho\|u\|}\right) u = \alpha := \bar{u} + \frac{\mu}{\rho}.$$

If this equation has a solution, then u and α have the same direction. Taking the 2-norm of both sides yields

$$\|u\| + \frac{\gamma}{\rho} = \|\alpha\|,$$

which has a solution only if $\|\alpha\| \geq \frac{\gamma}{\rho}$. In this case, we get the candidate optimal control

$$u^* = \alpha(1 - \frac{\gamma}{\rho\|\alpha\|}), \quad \alpha = \bar{u} + \frac{\mu}{\rho}, \quad \|\alpha\| \geq \frac{\gamma}{\rho}. \tag{4.3.7}$$

If $\|\alpha\| < \frac{\gamma}{\rho}$, we can rewrite (4.3.1) as

$$A[u] = \rho\alpha^T u - \gamma\|u\| - \frac{\rho}{2}\left(\|u\|^2 + \|\bar{u}^2\|\right).$$

The first term on the right can be bounded by

$$\rho\alpha^T u \leq \rho\|\alpha\|\|u\| < \gamma\|u\|.$$

Substituting, we find

$$A[u] < -\frac{\rho}{2}\|u\|^2 - \frac{\rho}{2}\|\bar{u}\|^2 \leq A(0).$$

Consequently, if $\|\alpha\| < \frac{\gamma}{\rho}$, we set $u^* = 0$.

Furthermore, the optimal control u^* needs to satisfy $u^* \in \mathcal{U}$.

4.3.2 Soft constraints

Since the sparse control is constrained, the (augmented)Hamiltonian is not differentiable at the boundary of the admissible control set $\mathcal{U}$. We introduce soft constraints using the approach of Wang & Li [120] to study the role of smoothness in the convergence of the regularized forward-backward sweep method. Suppose $u_{\min} \leq u \leq u_{\max}$, then the soft constraint is imposed by replacing u with

$$\tilde{u} = \frac{1}{2}(\sqrt{(u-u_{\min})^2+\delta^2} - \sqrt{(u-u_{\max})^2+\delta^2} + u_{\min} + u_{\max}). \tag{4.3.8}$$

Our numerical experiments show that the use of soft constraints can have a significant impact on convergence.

4.3.3 Splitting approach for ℓ_1-ℓ_1 optimization

Vossen & Maurer [117] have proposed a splitting approach that converts the ℓ_1 optimization problem into an optimization problem with constrained control. In their approach they split the controls into two parts to ensure differentiability. For $i = 1, \ldots, M$, $j = 1, \ldots, d$, define the new control functions corresponding to the positive and negative branches of the original control:

$$u_i^+ = \max\{0, u_i\}, \qquad u_i^- = \max\{0, -u_i\}, \tag{4.3.9}$$

where the maximum is applied element-wise to the vector $u_i \in \mathbf{R}^d$. Then we have the relations

$$u_i = u_i^+ - u_i^-, \qquad |u_i| = u_i^+ + u_i^-,$$

where also the absolute value is applied element-wise. In this notation, we find

$$\|u_i\|_1 = \mathbf{1}^T|u_i| = \mathbf{1}^T(u_i^+ + u_i^-),$$

where $\mathbf{1} = (1, \ldots, 1)^T \in \mathbf{R}^d$.

This results in an optimal control problem involving the extended control variable $\boldsymbol{u} = (u_1^+, \cdots, u_n^+, u_1^-, \ldots, u_M^-) \in \mathbf{R}^{2dM}$. The cost function (4.2.2) is replaced by:

$$J[\omega] = \int_0^T \sum_{i=1}^M \frac{1}{2}\|v_i - \bar{v}\|^2 + \gamma \sum_{i=1}^M \mathbf{1}^T(u_i^+ + u_i^-). \tag{4.3.10}$$

with the added (element-wise) constraints $0 \leq u_i^+ \leq u_{\max}$, $0 \leq u_i^- \leq -u_{\min}$, $i = 1, \ldots, M$, (where we assume $u_{\min} \leq 0$). Note that in this formulation, the cost function is linear in the extended control variable.

4.4 A simple two-agent symmetric problem

In this section, we discuss a difficulty with the use of unconstrained sparse controls using a simple two-agent system with symmetric initial condition $x_1(0) = (1, 0) = -x_2(0)$,

$v_1(0) = (1, 1), v_2(0) = (-1, 1)$. The motion can be expressed in terms of scalar quantities $x(t)$, $y(t)$, $v(t)$, $u(t)$, $w \in \mathbf{R}$ such that

$$x_1 = \begin{pmatrix} x \\ y \end{pmatrix}, \quad x_2 = \begin{pmatrix} -x \\ y \end{pmatrix}, \quad v_1 = \begin{pmatrix} v \\ w \end{pmatrix}, \quad v_2 = \begin{pmatrix} -v \\ w \end{pmatrix}, \quad u_1 = \begin{pmatrix} u \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -u \\ 0 \end{pmatrix}.$$

Neglecting the irrelevant motion in the y-direction, the reduced system is

$$\dot{x} = v, \tag{4.4.1}$$
$$\dot{v} = -2\phi(2x)v + u, \tag{4.4.2}$$

with initial conditions $x(0) = v(0) = 1$. We seek to minimize the ℓ_1 cost functional

$$J = \int_0^T \frac{1}{2}v(t)^2 + \gamma|u(t)|\, dt. \tag{4.4.3}$$

If we do not constrain the maximum value of the control u we find that the optimal control assumes the form of a singular adjustment of the initial velocity. For the numerical solution, the control is nonzero only during the first time step, the velocity is adjusted instantaneously to an optimal value, and the system evolves further without control. This behavior persists upon reducing the step size or using higher order methods.

Discretizing the problem using Euler's method with step size τ, the velocity equation in the first time step is

$$v_1 = v_0 - 2\tau\, \phi(2x_0)v_0 + \tau\, u_0.$$

Substituting the initial conditions $x_0 = v_0 = 1$, the necessary control for reaching velocity $v_1 = \xi$ in the first time step is

$$u_0 = \frac{\xi - (1 - 2\tau\, \phi(2))}{\tau}$$

The first step contributes to the total cost by an amount:

$$J_0(\xi) = \frac{1}{2}\tau + \gamma|\xi - 1 + 2\tau\, \phi(2)|. \tag{4.4.4}$$

In the left graph of Figure 4.1 we plot J_0 as a function of ξ for $\gamma = \{1, 0.5, 0.25\}$ using time step $\tau = 1/80$. We also plot the discretized cost of a non-controlled trajectory of (4.4.1)–(4.4.2) with initial condition $v_1 = \xi$,

$$J_{u=0}(\xi) = \tau \sum_{n=1}^{N-1} \frac{1}{2}v_n^2, \qquad v_1 = \xi,$$

as well as the sum of these two: $\tilde{J}(\xi) = J_0(\xi) + J_{u=0}(\xi)$. The optimal solutions are shown in the right graph of Figure 4.1. The minima in the graph on the left in Fig. 4.1 are obtained, for the cases $\gamma = \{1, 0.5, 0.25\}$, respectively, at $\xi = \{0.339, 0.228, 0.143\}$, whereas the velocities of the optimal solutions after one time step in the graph on the right of Fig. 4.1 are $v_1 = \{0.352, 0.230, 0.142\}$. We note that the optimal velocity at time $t = \tau$ is well approximated by the values of ξ at the minima $\tilde{J}(\xi)$. We observe that

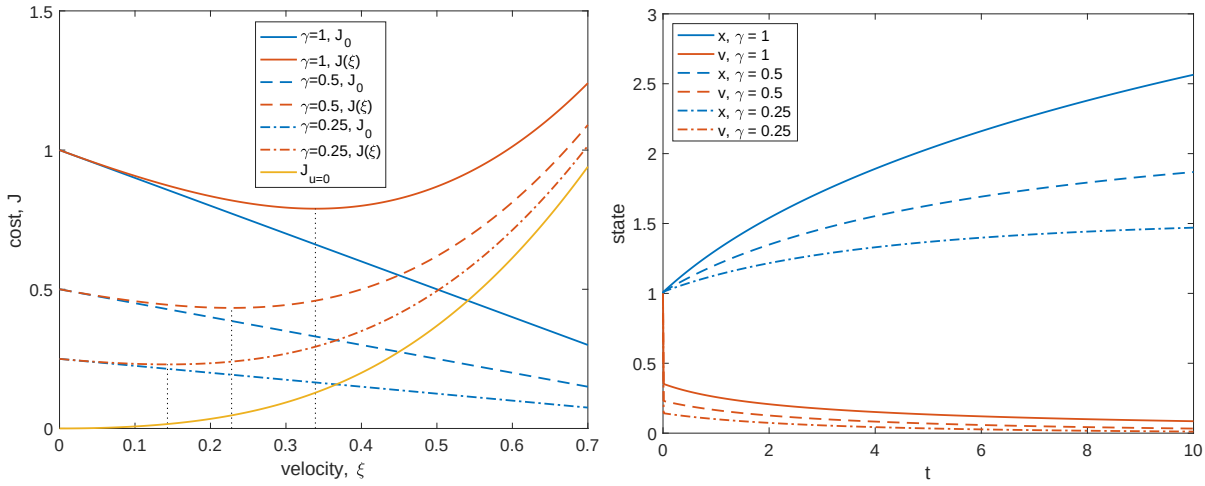


FIGURE 4.1: With unconstrained controls, the optimal solution of the 2-agent model with ℓ_1 cost (4.4.3) is a discontinuous jump to a new initial velocity. In the graph on the left, the yellow curve is the cost of a non-controlled solution starting from initial condition $(x_0 = 1, v_0 = \xi)$, the blue lines are the cost of adjusting velocity from $v_0 = 1$ to $v_0 = \xi$ in a single time step (4.4.4), and the red lines are the sums of these two costs for, respectively, $\gamma = 1$ (solid), $\gamma = 0.5$ (dashed), $\gamma = 0.25$ (dash-dotted). The graph on the right shows the optimal solutions. In each case the control is nonzero only in the first time step.

the approximation improves when the step size is reduced, suggesting that the optimal solution is a discontinuous adjustment of the initial velocity.

For comparison, in Figure 4.2 we illustrate the optimal solution obtained with control constraint $|u| \leq 1$ with $\gamma = 1$. With constraints on the control, the state converges to a continuous function. The control is still 'sparse' in the sense that it is nonzero only on a relatively short interval.

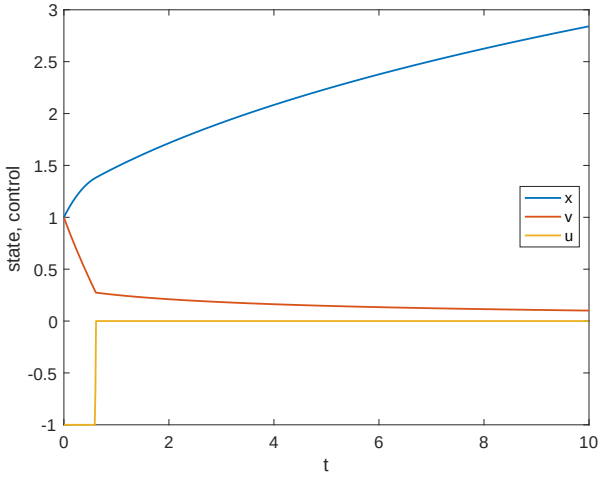


FIGURE 4.2: With constraint $|u| < 1$, the solution of the 2-agent problem becomes continuous in the state space ($\gamma = 1$). The control is still sparse in the sense that it is nonzero only on an interval.

4.5 Numerical results for n-agent systems

In this section, we report on numerical experiments to test the convergence behavior of the regularized forward-backward sweep iteration for optimal control of consensus in M-agent systems. In all experiments, the agents move in the plane ($d = 2$). We will consider optimal control in the three norms ℓ_2-ℓ_2, ℓ_1-ℓ_1 and ℓ_2-ℓ_1. We choose parameter values $\beta = 2$ in (4.1.4) and $\gamma = 1/(2M)$ in (4.2.2). For the initial condition we (uniformly) randomly place agents on the unit circle centered at the origin $\|x_i(0)\| = 1$, $i = 1, \ldots, M$. Moreover we assume a radial velocity configuration with $v_i(0) = x_i(0)$, chosen such that consensus would not be reached without a control.

In all experiments we apply the optimal control over a time window of length $T = 10$. The optimal control problem is discretized using the first order symplectic Euler pair [78] using step size $\tau = 0.125$. The method is explicit in the forward sweep and linearly implicit in the backward sweep. To accelerate convergence of the iteration we apply Anderson acceleration using the implementation described in Henderson & Varadan [62]. The results we report were computed using a fixed restart every m iterations, or following an increase in the cost function. We did not apply damping. The iteration was carried out until the update between the kth and $(k + 1)$th control iterates was smaller than a specified tolerance:

$$\|\boldsymbol{u}^{(k+1)} - \boldsymbol{u}^{(k)}\| < tol. \tag{4.5.1}$$

4.5.1 Optimal control in the ℓ_2-ℓ_2-norm

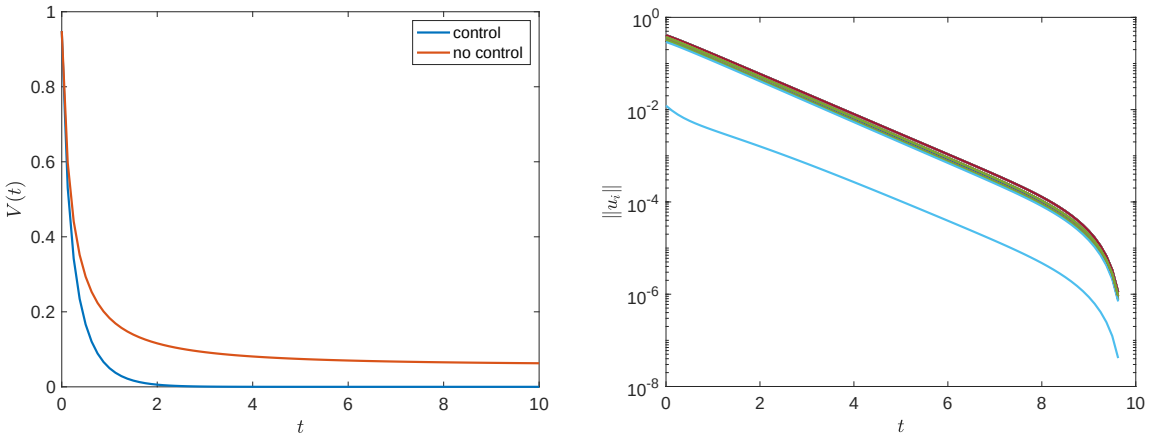
We first consider the convergence of the regularized forward-backward sweep algorithm when the ℓ_2-ℓ_2-norm is used in (4.2.2) ($q = p = 2$). The resulting cost function is convex in the control $\boldsymbol{u}$. In this case, the set of admissible controls $\mathcal{U}$ can be taken unconstrainted, so soft constraints (4.3.8) are unneeded. We solve the optimal control problem for a population size $M = 20$.

Figure 4.3(a) shows the consensus function $V(t)$ given by (4.1.5) as a function of t for the uncontrolled and controlled cases, illustrating that control is needed to reach consensus. Figure 4.3(b) shows the optimal control u. In the ℓ_2-ℓ_2-norm, the control is 'non-sparse', i.e., active for all time t.

Without regularization ($\rho = 0$) the forward-backward sweep method does not converge. Setting $\rho = 3$, convergence to the specified tolerance $tol = 10^{-6}$ occurs in $k = 503$ iterations. Applying Anderson acceleration with restart every $m = 5$ iterations reduces the number of iterations to $k = 44$, as shown in Figure 4.4.

4.5.2 Optimal control in the ℓ_1-ℓ_1-norm

For the choice $q = p = 1$ in (4.2.2), the optimal control becomes sparse in the sense that the control is nonzero only for a short adjustment period at the beginning of the time interval. Thereafter the system evolves under its own dynamics, with no control. As we saw in Section 4.4 it is necessary to place constraints on the admissible control space to avoid a singular control. In this section we place upper and lower bounds on each



(a) a (b) b

FIGURE 4.3: Convergence to consensus with ℓ_2-ℓ_2-norm cost function (4.2.2) with $p = q = 2$. (a) The consensus function $V(t)$ given in (4.1.5) with (blue) and without (red) control. (b) the optimal control is not sparse, but rather active on all agents over the entire interval.

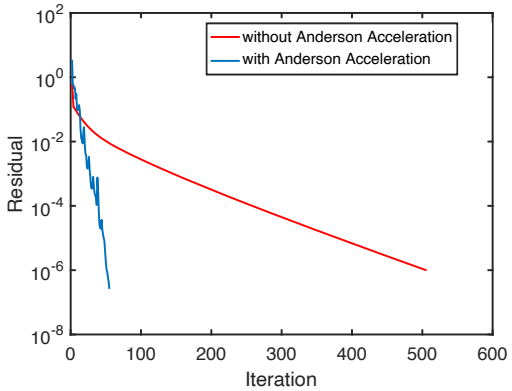


FIGURE 4.4: Convergence of the regularized forward-backward sweep algorithm (4.2.11)–(4.2.13) for $\rho = 3$ (red) is monotone but relatively slow. Applying restarted Anderson acceleration (blue) significantly speeds convergence.

component of the control: $u_{\max} = 0.5 = -u_{\min}$ and study the effect on convergence of the regularization parameter δ in the soft-constraint function (4.3.8).

For the case of $M = 20$ agents, Figure 4.5 illustrates the effect of soft constraints on the optimal control. In the left panel of Figure 4.5 the choice $\delta = 0.01$ shows a visible effect of the regularization compared to the right panel, for which $\delta = 0.001$. For these computations, the regularization parameter in the forward-backward sweep method (4.2.11)–(4.2.13) was chosen to be $\rho = 1$, and the Anderson acceleration was restarted every 6 iterations.

In the (mathematically equivalent) Vossen & Maurer formulation of the ℓ_1 control discussed in Section 4.3.3, it can be seen that the augmented controls ω_i appear linearly in the cost function. Such a formulation leads to a bang-bang controller that attains its extreme values ($\omega_i = 0$ or $\omega_i = u_{\max}$). Note that if the constraint $u_{\max} = 0.5$ is active for both components of a control $u \in \mathbf{R}^2$, then $\|u\|_1 = 1$. Similarly if the constraint is active on one component and the other component is zero (due to sparsity), then we find $\|u\|_1 = 0.5$. In Figure 4.5 we observe that for most agents, only one component of the control is active. This explains the plateau observed in Fig. 4.5 at $u = 0.5$.

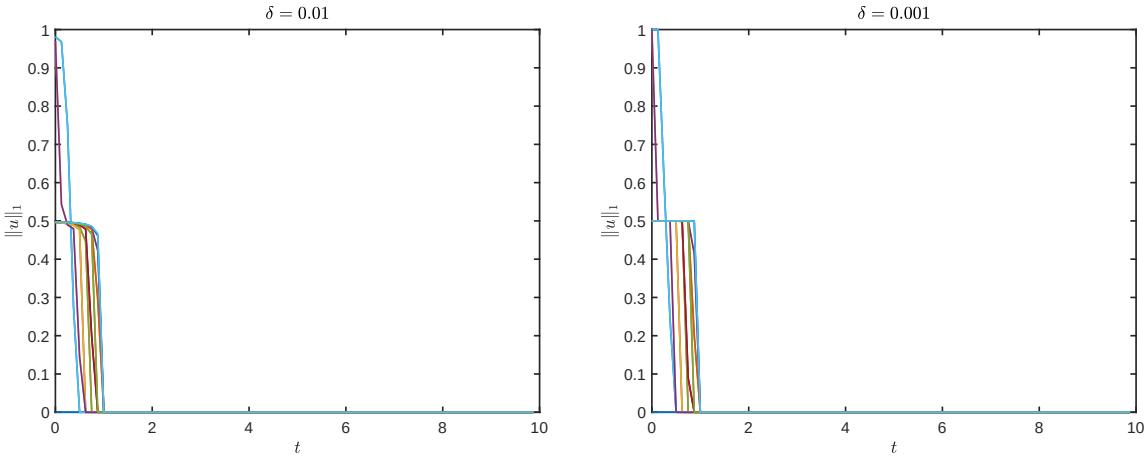


FIGURE 4.5: Left: control u when $\delta = 0.01$. Right: control u when $\delta = 0.001$,

We have introduced the soft constraint (4.3.8) to study the effect of smoothness of the control on convergence of the regularized forward-backward sweep iteration with Anderson acceleration. Figure 4.6 shows the number of iterations needed for convergence to $tol = 10^{-6}$ for values of $\delta = \{1, 2, 5\} \times 10^{-l}$, for $l = 1, \ldots, 4$. Except for very small values $\delta < 0.001$, the number of iterations increases with decreasing δ. (At the smallest values of δ the convergence dependence is irregular, due to lack of resolution using time step $\tau = 0.125$.) In these computations, $\rho = 8$ is employed. We choose $\delta = 0.01$ as a compromise between accuracy and computational cost.

Using $\delta = 0.01$ we investigated the effect of periodically restarting Anderson acceleration for different population sizes. We restart Anderson acceleration every m iterations or whenever the cost J increases within an iteration. Table 4.1 indicates the number of iterations required to satisfy (4.5.1) for $tol = 10^{-6}$ for restart numbers $4 \leq m \leq 9$. We

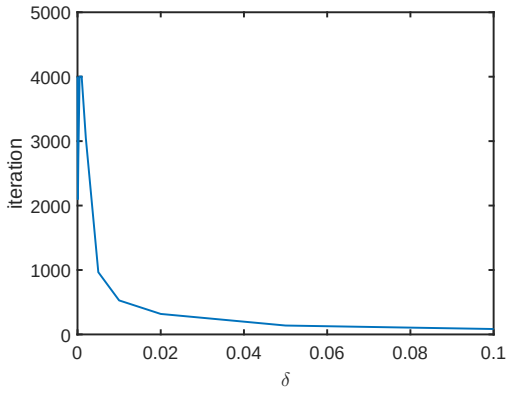


FIGURE 4.6: Number of iterations using cost (4.2.2) with $q = p = 1$ needed to satisfy (4.5.1) for $tol = 10^{-6}$ as a function of regularization parameter $10^{-4} \leq \delta \leq 10^{-1}$. The dependence is monotone for $\delta \geq 10^{-3}$.

found the optimal convergence rate to be near $m = 8$, with more pronounced effect for larger populations.

TABLE 4.1: Number of iterations for convergence (4.5.1) to $tol = 10^{-6}$ with ℓ_1-ℓ_1 cost (4.2.2) as a function of restart number of Anderson acceleration for increasing population size.

iteration count / restart number m / population size M	4	5	6	7	8	9
10	180	245	222	196	192	189
20	688	410	456	427	360	378
30	312	340	330	322	304	342

4.5.3 Optimal control in the ℓ_2-ℓ_1-norm

Also for the ℓ_2-ℓ_1-norm we compute the sparse control of $M = 20$ agents under the parameter $\delta = 0.01$ and $\delta = 0.001$. In the ℓ_2-ℓ_1-norm, the control only becomes bang-bang type when one component of the individual control ($u_i \in \mathbf{R}^2$) is zero (inactive). Figure 4.7 shows the converged optimal control satisfying (4.5.1) for $tol = 10^{-6}$, with $\rho = 3$, restart number $m = 7$. For both values of δ the control is sparse in the sense that beyond a short adjustment time, the control is inactive ($\boldsymbol{u} = 0$). For $\delta = 0.01$, the individual controls vary smoothly while active, whereas for $\delta = 0.001$, the bang-bang nature is noticeable once all controls have become single-component.

Table 4.2 again shows iteration counts m for different restart numbers of Anderson acceleration and increasing population size. Also in the $\ell_2 - \ell_1$-norm, $m = 8$ is optimal with pronounced improvement.

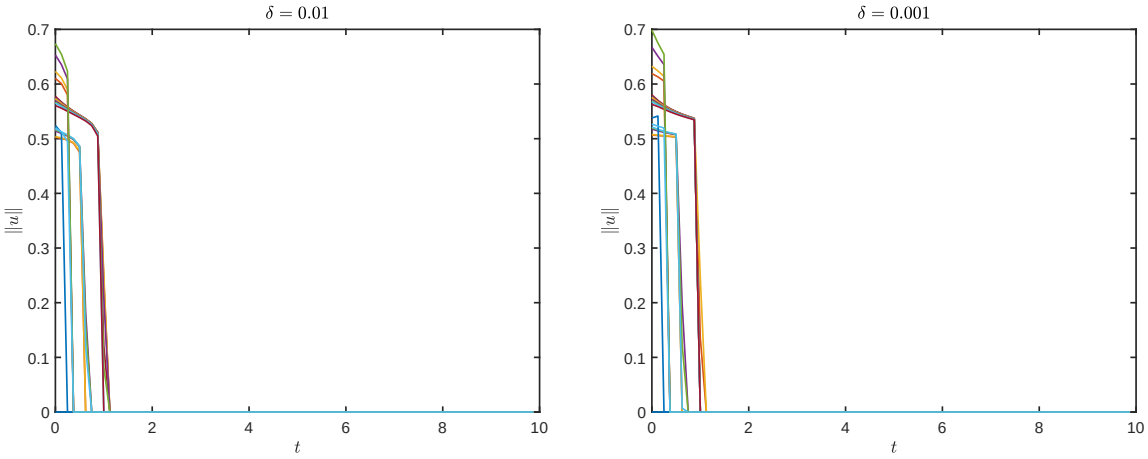


FIGURE 4.7: Number of iterations using cost (4.2.2) with $q = 2$ and $p = 1$ needed to satisfy (4.5.1) for $tol = 10^{-6}$ as a function of regularization parameter $10^{-4} \leq \delta \leq 10^{-1}$. The dependence is monotone for $\delta \geq 10^{-3}$.

TABLE 4.2: Number of iterations for convergence (4.5.1) to $tol = 10^{-6}$ with ℓ_2-ℓ_1 cost (4.2.2) as a function of restart number of Anderson acceleration for increasing population size.

iteration count / restart number m / population size M	4	5	6	7	8	9
10	236	185	210	231	184	207
20	416	335	348	350	328	297
30	396	415	498	497	320	441

4.6 Summary

In this chapter, we employed regularized forward-backward sweep iteration (4.2.11)–(4.2.13) and restarted Anderson acceleration to compute optimal control of consensus in the Cucker-Smale model. We tested the convergence of the algorithm for different choice of norm in the cost functional. The ℓ_1-ℓ_1-norm and ℓ_2-ℓ_1-norm cases lead to non-smooth sparse optimal controls, whose formulation require placing upper bound constraints on the admissible controls. To improve convergence of the algorithm we employed soft constraints (4.3.8) with regularization parameter δ to smooth the transition near the boundary of the admissible control region. Numerical experiments indicate that a significant improvement in convergence rate can be achieved with moderate values of the smoothing paramter, however with noticeable effect on the optimal solution. We also investigated the efficiency of dependence of Anderson acceleration on restart number, which we found to be robust with respect to norm and population size.

Acknowledgments

The authors express gratitude to Dante Kalise for detailing his experiences with in ℓ_1 optimal control of the Cucker-Smale model.

Chapter 5

Ensemble data assimilation using optimal control in the Wasserstein metric

Abstract

A data assimilation method is proposed that is based on optimal control minimizing the cost of mismatch in the Wasserstein metric on the observation space. The method is appropriate for systems in which multiple, noisy, partial observations are available (e.g. citizen weather stations or smart phones). The method is demonstrated for: (i) deterministic dynamics with uncertain initial conditions, (ii) multiple noisy observations of a randomly forced ODE, (iii) observations from multiple sample paths from an SDE. A bi-modal measure and a measure supported on a strange attractor are tested.

This chapter is transcripted from the paper "Ensemble data assimilation using optimal control in the Wasserstein metric" Submitted in the Journal of Computational Science.

5.1 Introduction

Data assimilation is a commonly used computational method for combining dynamic model simulations and observational data to estimate a state or trajectory of a dynamical system in fields as diverse as weather forecasting, computer vision, robotics and navigation. In uncertainty quantification, data assimilation may be used to approximate an evolving probability measure expressing uncertainty in the model, initial conditions or observations. Standard references on data assimilation include [44, 73, 94, 8].

Since particle filters were introduced in [54], they have become a very popular class of method that solve estimation problems in a recursive way depending on the observation data [4, 38]. Particle filter algorithms use a set of particles to represent the posterior distribution of the stochastic process and they update their prediction in an approximate way. A common method of particle filters is sequential importance sampling (SIS), which relates all particles generated according to their importance weight at every stage [17]. However, The disadvantage of particle filters is that SIS will have a significant weight-degeneracy after a large number of iterations, i.e. all but one particle will be eliminated due to the low weight [108, 71]. To avoid this problem, Sampling Importance Resampling (SIR) was introduced. The difference with SIS is that SIR resamples particles at every time stage, Specifically, it replicates the high-weighted particles and eliminates low-weighted particles. This approach is very useful and applied to solve many different kinds of problems. However the main difficulty is that weights always become unbalanced, hence most resampled particles coincide which leads to lower particle diversity especially for deterministic dynamics.

In [122, 123], the authors proposed a method to derive a particle filter by applying optimal control techniques. The method has a self-oriented formulation that provides a self-correcting feedback mechanism to stabilize the particles around the posterior. Inspired by [122, 123], in this paper, we introduce an alternative method to the construction of a particle filter by adding a control in the particle states. In contrast to [122, 123], we evolve particle states deterministically. We obtain the optimal control without calculating the posterior distribution of the particle states. We employ a Wasserstein metric [69, 115] in the cost function to measure the distance between probability distributions in the observation space. Reich [93] introduced optimal transport into particle methods as a means of resampling.

Although computationally complex, the Wasserstein distance is more robust than, e.g., the Kullback-Leibler divergence [97, 3]. Since it relies on a metric equipped in a metric space, the Wasserstein distance can be employed for two measures even if their supports are mutually exclusive. As a result, the Wasserstein approach is applicable to alternative measures besides absolutely continuous ones, e.g. empirical measures or measures supported on strange attractors [37].

5.2 Data assimilation problem

In this chapter we study an optimal control-based data assimilation method for modelling uncertainty of a partially observed process. Our starting point is an ensemble of possibly

noisy observations given in the form of K discrete time series

$$\hat{Z}_n^k = \hat{Z}^k(n\tau) \in \mathbf{R}^\ell, \quad n = 0, \ldots, N, \quad k = 1, \ldots, K,$$

where k denotes the ensemble index and n the time index over an interval $T = N\tau$.

We assume the underlying process $X(t)$, $t \in [0, T]$ is time-continuous and is described by either a deterministic (case $\sigma \equiv 0$) or stochastic differential equation

$$dX = a(X)\, dt + \Sigma\, dW, \tag{5.2.1}$$

where $X(t) \in \mathcal{D} \subset \mathbf{R}^d$ is the state at time t, $a(X) : \mathcal{D} \to \mathbf{R}^d$, $\Sigma \in \mathbf{R}^{d\times s}$ and $W(t)$ is an s-dimensional Wiener process.

Let $h(X) : \mathcal{D} \to \mathbf{R}^\ell$ be an observation function. Usually, one needs to deal with partial observations: $\ell < d$. An underlying assumption is that the state $X(t)$ is detectable by the observation function h. We assume the state $X(t)$ is unknown, due to uncertainty in initial condition, model error, or noise in the dynamics or measurements.

We distinguish three scenarios:

In the first scenario, we consider a deterministic system (i.e. $\Sigma \equiv 0$ in (5.2.1)) with uncertain initial condition and partial observations. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}^k(n\tau)), \quad n = 0, \ldots, N, \quad k = 1, \ldots, K, \tag{5.2.2}$$

where $\hat{X}^k(t)$, $k = 1, \ldots, K$, denotes an ensemble of solutions of the deterministic differential equation (5.2.1), $\sigma \equiv 0$, with initial conditions $\hat{X}^k(0)$ drawn from a probability distribution.

In the second scenario, $\hat{X}(t)$ corresponds to a single sample path of the SDE (5.2.1) for which multiple noisy observations are available. This scenario models the case of weather measurements using a scattering of imperfect personal devices such as smart phones or private weather stations. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}(n\tau)) + \eta_n^k, \quad n = 0, \ldots, N, \quad k = 1, \ldots, K, \tag{5.2.3}$$

where the kth time series $\{\eta_n^k\}_{n=0}^N$ denotes the kth realization of the discrete noise process, and $\eta_n^k \sim \mathcal{N}(0, R)$, where $R \in \mathbf{R}^{\ell\times\ell}$ is the covariance matrix of the observational noise.

In the third scenario we assume we are given K sample paths of (5.2.1), i.e., $\hat{X}^k(t)$, $k = 1, \ldots, K$, and the kth sequence $\{\hat{Z}_n^k\}_{n=0}^N$ is observed from $X^k(t)$, for $k = 1, \ldots, K$. This scenario models the case of (possibly noisy) measurements of a repeated experiment with random forcing. The observations are given by

$$\hat{Z}_n^k = h(\hat{X}^k(n\tau)) + \eta_n^k, \quad n = 0, \ldots, N, \quad k = 1, \ldots, K.$$

In all three scenarios, our objective is to estimate the uncertainty in our knowledge of the underlying process $\hat{X}(t)$ by approximating an evolving probability measure $\mu(x, t)$ such that for measurable $A \subset \mathcal{D}$,

$$\int_A \mu(x, n\tau)\, dx = \text{Prob}\{X(n\tau) \in A\}.$$

The measure μ will be approximated by an empirical measure $\nu_n(x)$ supported on an ensemble of J particles:

$$\nu_n(x) = \frac{1}{J}\sum_{j=1}^{J} \delta(x - X_n^j), \tag{5.2.4}$$

where δ denotes the Dirac distribution. The motion of the jth particle is governed by the drift vector field $a(X)$ and an optimal control via the differential equation

$$\begin{aligned} \frac{dX^j}{dt} &= a(X^j) + Bu^j(t), && j = 1, \ldots, J, \\ Z^j(t) &= h(X^j(t)), && j = 1, \ldots, J, \end{aligned}$$

where $B \in \mathbf{R}^{d\times m}$ and $u^j(t) \in \mathbf{R}^m$ is the control input for jth particle at time t, chosen to minimize a cost function that penalizes mismatch (in the observation space) with respect to a Wasserstein metric. Of course, optimizing the mismatch does not guarantee the convergence of the measure ν to μ. Nevertheless such a strategy is common in variational data assimilation methods such as 4D-Var. The convergence question is related to concepts such as the synchronization of chaos, detectability, and Lyapunov stability theory [48, 53, 35, 114, 90]. By comparison with variational data assimilation, we can view the controls $u^j(t)$ as representing the unknown model error required to explain the observations.

The particle motion is discretized in time using Euler's method to obtain the discrete dynamics

$$X_{n+1}^j = X_n^j + \tau\, a(X_n^j) + \tau\, Bu_{n+1}^j, \qquad n = 0, \ldots, N-1, \quad j = 1, \ldots, J, \tag{5.2.5}$$

$$Z_n^j = h(X_n^j), \qquad n = 0, \ldots, N, \quad j = 1, \ldots, J. \tag{5.2.6}$$

5.2.1 Wasserstein cost function.

In this chapter we study numerically the use of a Wasserstein metric to measure the mismatch in empirical distributions defined by the measurement ensemble and particle filter. The Wasserstein metric has found increased application in data assimilation, machine learning and data science in general, due to a number of attractive properties. For instance, the Wasserstein distance is well defined for singular measures and distributions, e.g. for measuring distance between empirical distributions or measures supported on strange attractors. Also, in the Wasserstein metric, the geodesic path between two distributions is the optimal transport path, along which the deformation of a density is minimal. Consequently, in the context of data assimilation when observations may be sparse, the probability density function will deform in a minimal way between observation times.

Our goal is to choose the controls u_n^j in (5.2.5) so that the particle distribution ν_n approximates $\mu(x, n\Delta t)$. Given that we only have access to the sample observations $\{\hat{Z}_n^k\}$ we minimize a cost function that penalizes mismatch in the Wasserstein metric. Let

$$\hat{\zeta}_n(z) = \frac{1}{K}\sum_{k=1}^{K} \delta(z - \hat{Z}_n^k), \qquad \zeta_n(z) = \frac{1}{J}\sum_{j=1}^{J} \delta(z - Z_n^j) \tag{5.2.7}$$

The cost function is defined as

$$C = \tau \sum_{n=0}^{N-1} \left[\sum_{j=1}^{J} \frac{1}{2} \|u_{n+1}^j\|^2 + \frac{\beta}{2} \mathcal{W}_2^2(\hat{\zeta}_{n+1}, \zeta_{n+1}) \right], \tag{5.2.8}$$

where $\mathcal{W}_2$ denotes the 2-Wasserstein distance (see below) and the constant $\beta \geq 0$ is a weight parameter.

5.2.2 Calculation of Wasserstein distance.

The Wasserstein distance is a metric on the space of probability measures. The p-Wasserstein distance between two probability measures μ and ν on a metric space $(\mathcal{X}, d)$ is given by

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}$$

where Π denotes the set of transport couplings of μ and ν, that is, $\Pi = \{\pi(x, y) | \int \pi(dx, y) = \mu(y), \int \pi(x, dy) = \nu(x)\}$.

For empirical measures such as (5.2.4), computing the Wasserstein distance reduces to solving an optimal transportation problem of weighted point sets, a special case of the minimum cost flow problem [11, 2]. Given empirical measures

$$\mu(x) = \frac{1}{J} \sum_{j=1}^{J} \delta(x - \hat{X}^j), \quad \nu(x) = \frac{1}{K} \sum_{k=1}^{K} \delta(x - X^k),$$

consider the space of finite transport maps:

$$\mathcal{F} = \left\{ F = (f_{jk}) \in \mathbf{R}^{J \times K} \middle| f_{jk} \geq 0, \sum_j f_{jk} = \frac{1}{K}, \sum_k f_{jk} = \frac{1}{J} \right\}.$$

The 2-Wasserstein distance is equal to

$$\mathcal{W}_2(\mu, \nu) = \left(\min_{F \in \mathcal{F}} \sum_{jk} f_{jk} d_{jk}^2 \right)^{1/2}, \tag{5.2.9}$$

where we use a weighted norm

$$d_{jk}^2 = \|\hat{X}^j - X^k\|_M^2 = (\hat{X}^j - X^k)^T M (\hat{X}^j - X^k). \tag{5.2.10}$$

For instance, for noisy observations (5.2.3) with covariance matrix R, we choose $M = R^{-1}$ to reflect our confidence/uncertainty in the observations.

The minimization (5.2.9) constitutes a linear program. The Wasserstein distance can be efficiently estimated using the Sinkhorn algorithm [34].

5.3 Optimal control

To determine the optimal control $\{u_j^n\}$ in (5.2.5), we minimize the cost function (5.2.8) under constraints (5.2.5)–(5.2.6). We introduce Lagrange multipliers $\{\lambda_n^j\}$ and $\{\Lambda_n^j\}$ and define a discrete Lagrangian functional:

$$L = C + L_0 + L_\lambda,$$

where C is the cost function (5.2.8), L_0 enforces the constraint on the initial conditions, $X_0^j = \xi_0^j$, presumed known (or sampled from a known initial distribution),

$$L_0 = \sum_{j=1}^{J} \left[\lambda_0^j (X_0^j - \xi_0^j) + \Lambda_0^j (Z_0^j - h(\xi_0^j)) \right], \tag{5.3.1}$$

and L_λ defines the constraint relations:

$$L_\lambda = \sum_{n=0}^{N-1} \sum_{j=1}^{J} \left[(\lambda_{n+1}^j)^T (X_{n+1}^j - X_n^j - \tau a(X_n^j) - \tau B u_{n+1}^j) + (\Lambda_{n+1}^j)^T (Z_{n+1}^j - h(X_{n+1}^j)) \right]. \tag{5.3.2}$$

Note that we include the observation function (5.2.6) as a constraint, as the observations appear implicitly in the cost function (5.2.8).

We demand that the Lagrangian be stationary under variations with respect to X_n^j , Z_n^j, λ_n^j and Λ_n^j. In addition we minimize L with respect to u_n^j. Assuming sufficient differentiability, we set derivatives of L with respect to these variables equal to zero. This approach is known to yield a variational integrator [81] that defines a symplectic map. In the context of optimal control, see for example [88, 101].

Assuming the cost C is differentiable with respect to u at a (local) minimum, from $\partial L / \partial u_n^j = 0$ follows

$$u_n^j = B^T \lambda_n^j, \quad n = 1, \ldots, N, \quad j = 1, \ldots, J, \tag{5.3.3}$$

Enforcing $\partial L / \partial \lambda_n^j = 0$ and $\partial L / \partial \Lambda_n^j = 0$, and making use of (5.3.3), we obtain the filter relations (5.2.5)–(5.2.6):

$$X_{n+1}^j = X_n^j + \tau\, a(X_n^j) + \tau B B^T \lambda_{n+1}^j, \qquad n = 0, \ldots, N-1, \tag{5.3.4}$$

$$Z_n^j = h(X_n^j), \qquad n = 0, \ldots, N, \tag{5.3.5}$$

$$X_0^j = \xi_0^j. \tag{5.3.6}$$

From $\partial L / \partial Z_n^j = 0$, we obtain the definition

$$\Lambda_n^j = -\tau \frac{\beta}{2} \frac{\partial}{\partial Z_n^j} \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n), \quad n = 1, \ldots, N, \quad j = 1, \ldots, J, \tag{5.3.7}$$

where ζ_n and $\hat{\zeta}_n$ are given by (5.2.7).

Finally, from the condition $\partial L/\partial X_n^j = 0$, and making use of (5.3.7), we obtain the adjoint relations:

$$\lambda_1^j = \lambda_0^j - \tau\, \nabla a(X_0^j)^T \lambda_1^j, \tag{5.3.8}$$

$$\lambda_{n+1}^j = \lambda_n^j - \tau\, \nabla a(X_n^j)^T \lambda_{n+1}^j + \tau \frac{\beta}{2} \nabla h(X_n^j)^T \frac{\partial}{\partial Z_n^j} \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n), \quad n = 1, \ldots, N-1, \tag{5.3.9}$$

$$\lambda_N^j = \tau \frac{\beta}{2} \nabla h(X_N^j)^T \frac{\partial}{\partial Z_N^j} \mathcal{W}_2^2(\zeta_N, \hat{\zeta}_N). \tag{5.3.10}$$

Numerical evaluation of the gradient of Wasserstein distance. To evaluate the second term on the right of (5.3.9), we represent $\nabla h(X_n^j)$ as a matrix of dimension $\ell \times d$. Denote the columns of this matrix by the vectors $\hat{h}_1, \ldots, \hat{h}_d \in \mathbf{R}^\ell$. We approximate the Λ_n^j in (5.3.7) numerically using a finite difference formula:

$$\left(\nabla h(X_n^j)^T \frac{\partial \mathcal{W}_2^2}{\partial Z_n^j} \right)_i \approx \frac{1}{\varepsilon} \left[\mathcal{W}_2^2(\zeta_n^{(j,i)}(\varepsilon), \hat{\zeta}_n) - \mathcal{W}_2^2(\zeta_n, \hat{\zeta}_n) \right], \tag{5.3.11}$$

where

$$\zeta_n^{(j,i)}(\varepsilon) = \frac{1}{J} \left[\delta(z - (Z_n^j + \varepsilon \hat{h}_i)) + \sum_{k \neq j} \delta(z - Z_n^k) \right],$$

and ε can be chosen to be the square root of machine precision. Consequently, the second term on the right of (5.3.9) can be approximated using $d+1$ evaluations of the Wasserstein distance.

The complete set of equations that define the filter can be expressed in terms of the variables X_n^j, Z_n^j and λ_n^j given by (5.3.4)–(5.3.6) and (5.3.8)–(5.3.10). Forward-backward sweep iteration proceeds by solving (5.3.4)–(5.3.6) forward in time, followed by (5.3.8)–(5.3.10) backward in time, and repeating. However, such iteration is not convergent in general, especially for nonlinear dynamics.

Instead, the regularized forward-backward sweep method [78] proposed to augment the optimal control (5.3.3)

$$u_n^j = \frac{1}{1+\rho} \left[\lambda_n^j + \rho \left(\frac{X_{n+1}^j - X_n^j}{\tau} - a(X_n^j) \right) \right], \quad n = 1, \ldots, N, \quad j = 1, \ldots, J. \tag{5.3.12}$$

where $\rho > 0$ is the regularization parameter. Convergence of the resulting iteration for suficiently large ρ is proven for continuous dynamics in [78]. The proof is confirmed for the discrete case with symplectic discretization in [78]. In practice the convergence can be greatly accelerated using Anderson acceleration with restart [62]. The Wasserstein distance is Lipschitz continuous with respect to the state variable [3, 39]. Consequently, it satisfies the criterion for convergence of the regularized forward-backward sweep algorithm [77, 78].

5.4 Numerical experiments

In this section, we study numerically the properties of the proposed filter for quantifying uncertainty in some simple differential equations. We first study the propagation of uncertainty in the initial condition of a deterministic differential equation, the Lorenz attractor

model [79]. Subsequently, we consider stochastically forced motion in a double-well potential, for which the equilibrium distribution is bi-modal. We study both the case of a single sample path with noisy measurements and the case of multiple samples. In all numerical experiments we directly observe one dependent variable. Hence, the observations are partial ($\ell < d$), but the observation operator is linear (corresponding to a row of the identity).

For all experiments we computed the Wasserstein distance by solving the linear program (5.2.9), for which the complexity is unfavorable for large ensemble size [112, 89]. Improved performance could possibly be achieved using the Sinkhorn iteration [34], especially given that the many Wasserstein distances that need to be computed via (5.3.11). Note that the transport paths in (5.3.11) are expected to be very similar, providing good starting values for the iterations. We have not investigated this further.

5.4.1 Uncertainty in initial condition: deterministic Lorenz 63 model

In this section we study the behavior of the particle filter to approximate a probability measure relaxing onto the attractor of the Lorenz 63 system [79]. The invariant measure of the Lorenz system is a Sinai-Ruelle-Bowen measure, supported on a strange attractor of fractal dimension. The dynamics is deterministic, but we introduce uncertainty in the initial conditions by drawing an ensemble from a normal distribution. We compare the particle filter to the ensemble Kalman filter (EnKF, [44]) to study the potential advantage of optimizing with respect to mismatch in the Wasserstein metric. The EnKF method focuses on properties of evolving Gaussian distributions (approximating the mean and covariance matrices), which are smooth absolutely continuous measures. The Wasserstein metric does not require differentiability of the evolving density. Consequently, it is useful for comparing measures that evolve on a strange attractor. Applying Euler's method, the discrete Lorenz system is given by

$$\hat{x}_{n+1} = \hat{x}_n + \tau\, c_1(\hat{y}_n - \hat{x}_n), \tag{5.4.1}$$
$$\hat{y}_{n+1} = \hat{y}_n + \tau\, (\hat{x}_n(c_2 - \hat{z}_n) - \hat{y}_n), \tag{5.4.2}$$
$$\hat{z}_{n+1} = \hat{z}_n + \tau\, (\hat{x}_n\hat{y}_n - c_3\hat{z}_n), \tag{5.4.3}$$

with the parameters $c_1 = 10, c_2 = 28, c_3 = 8/3$ as originally studied by Lorenz. We employ step-size $\tau = 0.001$.

To generate observations, we simulate an ensemble of $K = 30$ trajectories over the time interval $t \in [0, 6]$, with initial conditions $X_0^k = (x_0^k, y_0^k, z_0^k)$ drawn from

$$x_0^k \sim \mathcal{N}(1, 0.5^2), \quad y_0^k \sim \mathcal{N}(-1, 0.5^2), \quad z_0^k \sim \mathcal{N}(25, 0.5^2).$$

This initial condition was chosen with a small variance but rapidly spreading ensemble that splits across the two lobes of the Lorenz attractor. The (partial) observable is the x-component

$$\hat{Z}_n^k = \hat{x}_n^k, \qquad k = 1, \ldots, K, \quad n = 1, \cdots, N.$$

The control is applied only to the y-component. The particle dynamics satisfy

$$x_{n+1} = x_n + \tau\, c_1(y_n - x_n), \tag{5.4.4}$$
$$y_{n+1} = y_n + \tau\,(x_n(c_2 - z_n) - y_n) + \tau u_{n+1}, \tag{5.4.5}$$
$$z_{n+1} = z_n + \tau\,(x_n y_n - c_3 z_n). \tag{5.4.6}$$

We select a particle filter ensemble size $J = 100$. For the Wasserstein metric (5.2.9)-(5.2.10) we choose $M = I$, and in the cost function $\beta = 60$.

The states of the two methods are shown in Figure 5.1. The particle filter appears to provide better coverage of the empirical measure. In particular it can be seen that some EnKF ensemble members appear in the low probability region between the two lobes of the attractor.

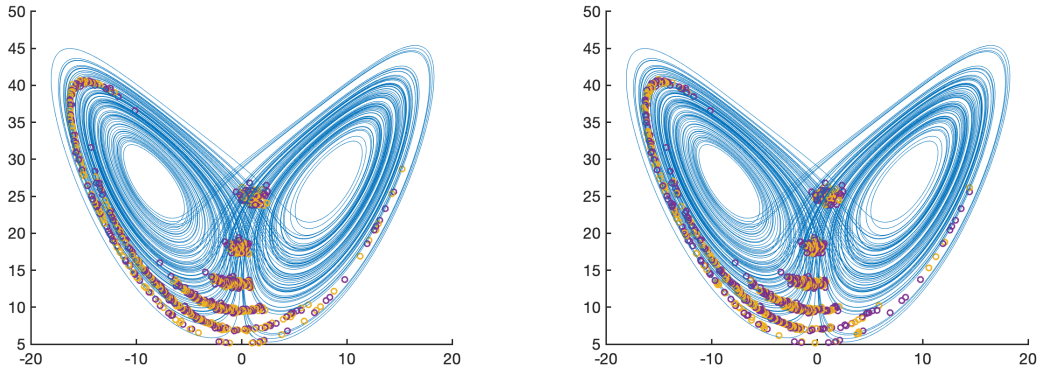


FIGURE 5.1: Comparison of particle filter (left) and ensemble Kalman filter (right) for the Lorenz attractor at time $t = 6$. Partial observations (x-component only) were generated from a sample ensemble of trajectories whose final states are indicated by purple circles. The final states of the filters are indicated with yellow circles.

.

The better approximation of the evolving measure by the particle filter is confirmed in Figure 5.2, where we compare the Wasserstein distances between the sample ensemble μ_n based on the full states $\{\hat{X}_n^k\}$ and the filter ensembles $\nu_n(X)$ computed using the particle filter and EnKF. We see that Wasserstein distances of the empirical measures to that of the sample ensemble $\mathcal{W}_2(\nu_0, \mu_0) < 1$ for both filters, the final distance $\mathcal{W}_2(\nu_N, \mu_N)$ is approximately 0.5 for the particle filter and 5 for the EnKF.

5.4.2 Noisy observations: a randomly forced ODE

For the experiments in this and the next section we consider stochastically forced motion in a double-well potential:

$$dq = p\, dt + \sigma_q\, dW_q,$$
$$dp = (q - q^3 - rp)\, dt + \sigma_q\, dW_p,$$

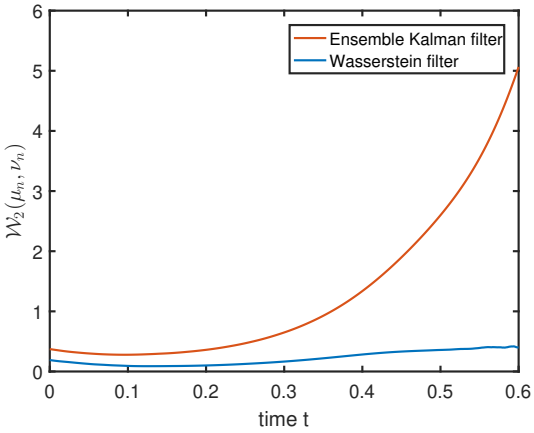


FIGURE 5.2: The Wasserstein distance in full state (x, y, z)

where $r > 0$ is a damping parameter. For the numerical experiments we choose $r = 1$, $\sigma_q = \sigma_p = 0.1$. Probability distributions transported by this system converge to a bimodal equilibrium state with peaks centered at the stable equilibria $(q^*, p^*) = (\pm 1, 0)$ of the drift vector field.

To generate samples of this system we discretize using the Euler-Maruyama method

$$\hat{q}_{n+1} = \hat{q}_n + \tau\, \hat{p}_n + \sigma_q \Delta W_{q,n}, \tag{5.4.7}$$

$$\hat{p}_{n+1} = \hat{p}_n + \tau\, (\hat{q}_n - \hat{q}_n^3 - r\hat{p}_n) + \sigma_p \Delta W_{p,n}, \tag{5.4.8}$$

where $\Delta W_{p,n}$, $\Delta W_{q,n} \sim \mathcal{N}(0, \tau)$ are independent and normally distributed. Noisy observations of the variable q are obtained from

$$\hat{Z}_n = \hat{q}_n + \eta_n, \quad n = 0, 1, \ldots, N, \tag{5.4.9}$$

where $\eta_n \sim \mathcal{N}(0, \sigma_n^2)$.

For the particle filter, the motion of the jth controlled particle is given by

$$q_{n+1}^j = q_n^j + \tau\, p_n^j + \tau\, u_{q,n+1}^j, \tag{5.4.10}$$

$$p_{n+1}^j = p_n^j - \tau\, ((q_n^j)^3 - q_n^j + r p_n^j) + \tau\, u_{p,n+1}^j, \tag{5.4.11}$$

for $j = 1, \ldots, J$, and the observation function applied to the jth particle yields

$$Z_n^j = q_n^j, \quad n = 0, \ldots, N. \tag{5.4.12}$$

For all experiments we choose stepsize $\tau = 0.01$. In each computation, the regularization parameter ρ in (5.3.12) was experimentally determined as small as possible to still observe convergence of the forward-backward sweep iteration.

We first investigate the scenario of noisy observations of a single sample path of (5.4.7)–(5.4.8). We choose initial conditions $\hat{q}_0 = 0.2$, $\hat{p}_0 = 0.5$ and integrate to time $t = 5$. The particle filter positions were sampled from initial distribution $\nu_0(q, p)$ given by the product measure

$$q_0 \sim \mathcal{N}(0.2, 0.04^2), \qquad p_0 \sim \mathcal{N}(0.5, 0.06^2). \tag{5.4.13}$$

Observational noise was generated with standard deviation $\sigma_n = 0.1$, and for each time step we sample $K = 30$ noisy observations. Wasserstein metric is given by (5.2.9)–(5.2.10) with $M = R^{-1}$ and $R = \sigma_n^2 I$, where I is the identity matrix.

We apply the particle filter with particle number $J = 30$ and $J = 10$. The results for $J = 30$ are shown in Figures 5.3. For these simulations, we chose $\beta = 4$ in the cost function (5.2.8). The red curves in 5.3 show the sample path $\hat{q}_n$ (upper plots) and $\hat{p}_n$ (lower plots). The noisy measurement data $\{\hat{Z}_n^k\}_{k=1}^K$ is plotted in yellow in 5.3(a) the sample mean is plotted as yellow circles in 5.3(b). The particle trajectories are plotted as blue curves in Figures 5.3(a) and (c). The particle mean trajectory is plotted in blue in Figures 5.3(b) and (d). As expected, the pdf of the observed q-component is approximately normally distributed about the sample path. This is not the case for the unobserved p-component, for which the marginal pdf is time-dependent. We see that the mean particle motion is much smoother than the sample path. It also appears as if the q-component, which is directly observed, is better estimated than the p-component.

In Figure 5.4 we repeat the above experiment, but for a smaller particle size $J = 10$ for the particle filter. The conclusions are similar. The particle mean trajectory is of similar accuracy to the higher resolution simulation in Figure 5.3.

The parameter β in the cost function (5.2.8) determines the relative weight of the observations compared of the cost of controlling the particle motion. In Figure 5.5 we choose a much larger value $\beta = 200$ and repeat the experiment. We observe that the particle filter paths are much less smooth in the q-component in Figure 5.5(a) and that the particle ensemble mean closely follows the sample path in Figure 5.5(b). There is no noticable improvement in the trajectories of the unobserved component p

5.4.3 Multiple sample paths of a stochastic system

In this section we generate observations by simulating an ensemble of sample paths of the stochastic double well potential (5.4.7)–(5.4.8). All parameters are identical to those in the previous section unless stated otherwise.

We first study the approximation of the bimodal distribution at high resolution. For this example, we choose a deterministic (Dirac distribution) initial condition $q_0 = -1$, $p_0 = 0$ for both the samples and the filter particles. We generated a large number $K = 20000$ of sample paths to approximate the time evolving pdf, which is exhibited at time $t = 5$ by the yellow curve in Figure 5.6. We then generated observations using an ensemble of size $K = 200$ without noise (i.e. $\sigma_n = 0$ in (5.4.9)) and applied the particle filter (5.3.4)–(5.3.10) with $J = 200$ particles. Histograms of the samples and particle filter pdfs are shown in Figure 5.6 for parameter values $\beta = 4$ (left plot) and $\beta = 200$ (right plot). The bi-modality of the pdf is clearly noticable, and the approximation more closely matches the observations for $\beta = 200$ as expected.

Figure 5.7 shows the time evolution of the Wasserstein distance $\mathcal{W}_2$ over the full state empirical measures $\nu_n(q,p)$ $\mu_n(q,p)$ for $\beta = 4$ and $\beta = 200$. For $\beta = 200$, the Wasserstein distance is bounded below $\mathcal{W}_2 < 0.007$ for most of the interval. For $\beta = 4$ the distance is somewhat greater at around $\mathcal{W}_2 < 0.02$ but decreases over time.

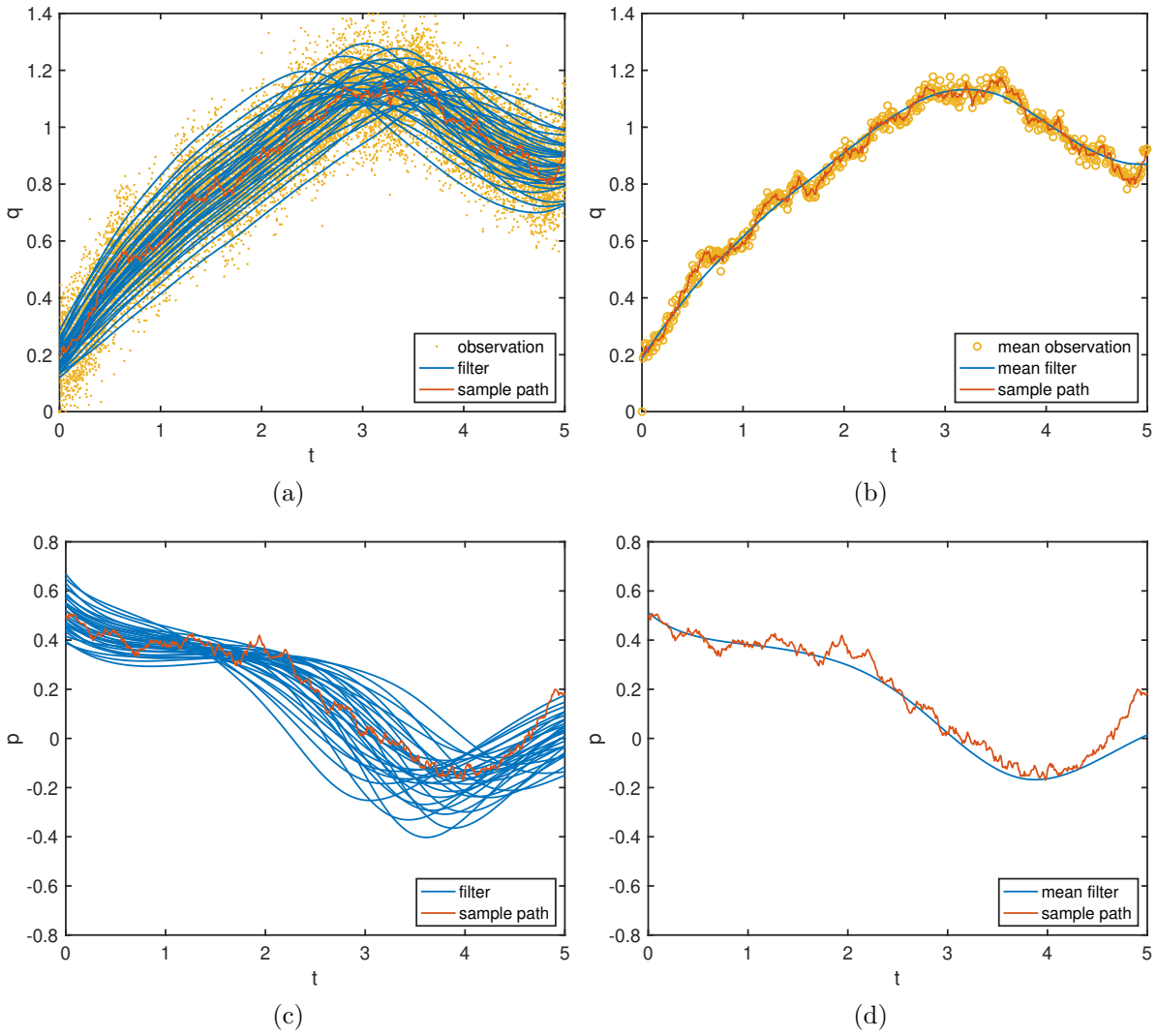


FIGURE 5.3: Noisy observations of a single sample path, particle number $J = K = 30$. The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d). For this simulation $\beta = 4$ was used.

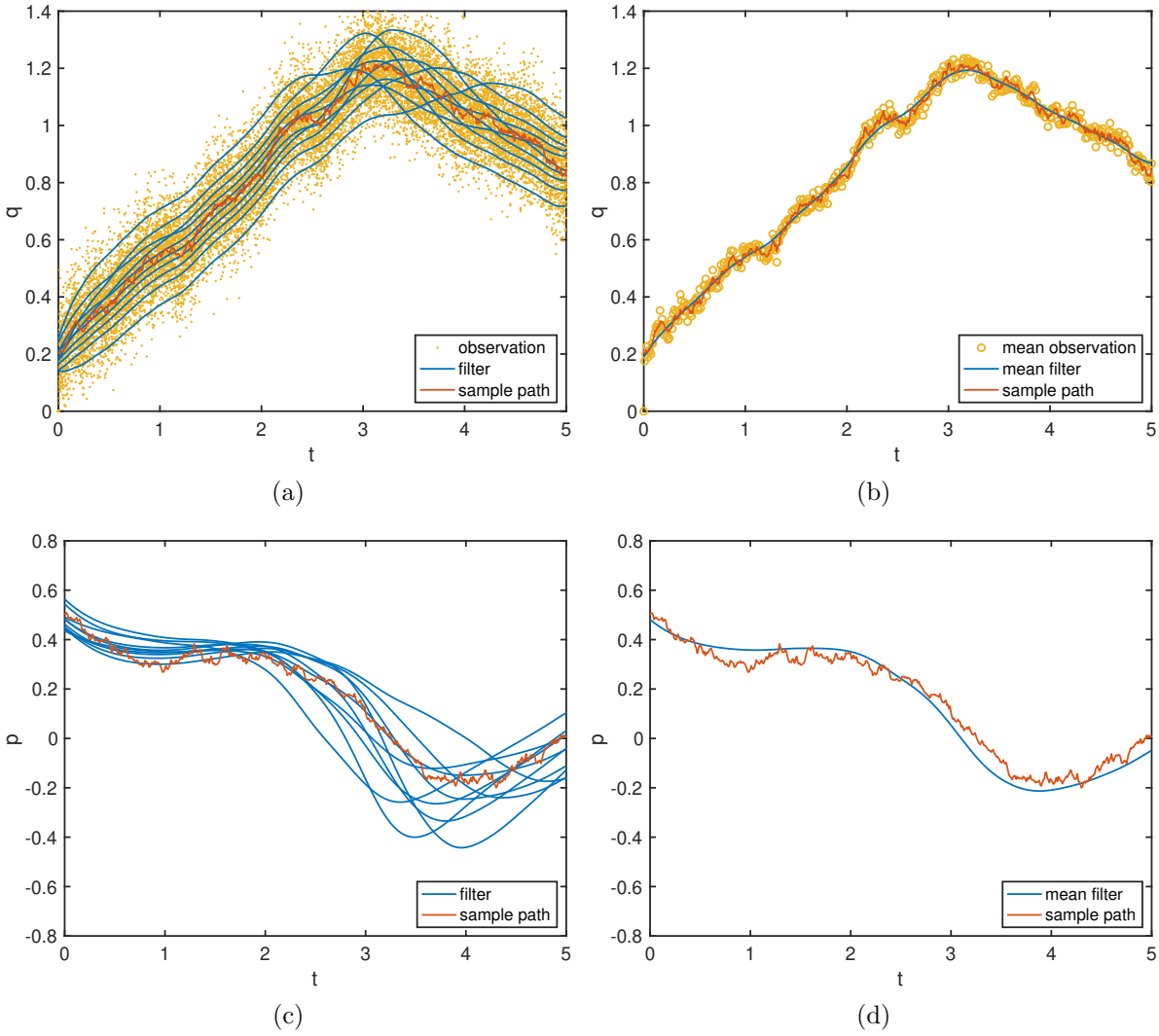


FIGURE 5.4: Same as Figure 5.3, but with particle number $J = 10$. The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d). For this simulation $\beta = 4$ was used in the cost function (5.2.8).

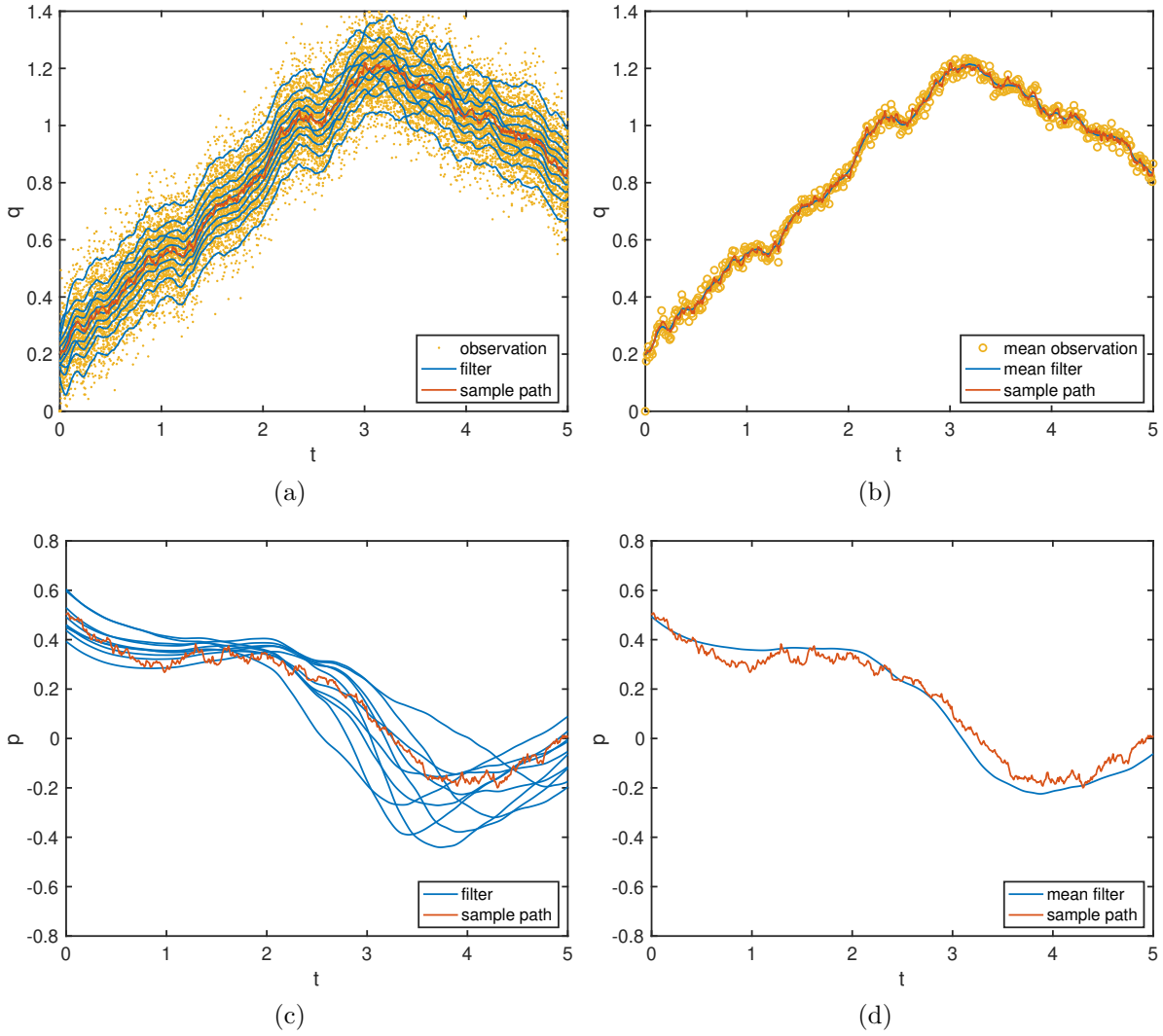


FIGURE 5.5: Same as Figure 5.4, but computed with $\beta = 200$ in the cost function (5.2.8). The sample path is shown in red: $\hat{q}(t)$ (upper two plots), $\hat{p}(t)$ (lower two plots). The observations are indicated by the yellow dots in (a) and the observation sample mean by the yellow circles in (b). The particle filter trajectories are indicated by blue curves in (a) and (c), and the particle ensemble mean by the blue curves in (b) and (d).

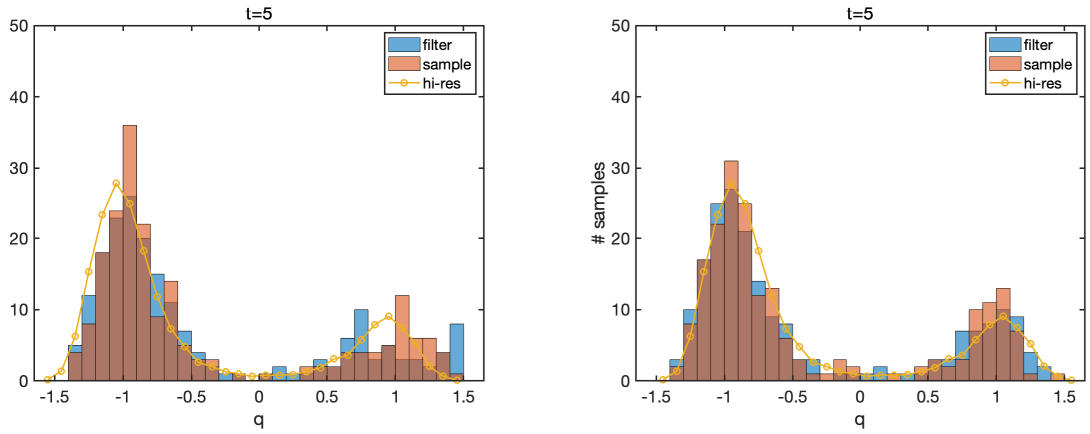


FIGURE 5.6: Histograms of coordinate $q(t)$ of the sample ensemble (red) and filter (blue) at time $t = 5$ for weight parameters $\beta = 4$ (left) and $\beta = 200$ (right) for a 200-member ensemble. The yellow curve indicates the expected bin size based on a high-resolution sample of 20000 members. The figure shows that the proposed method is effective at approximating a bi-modal probability density function.

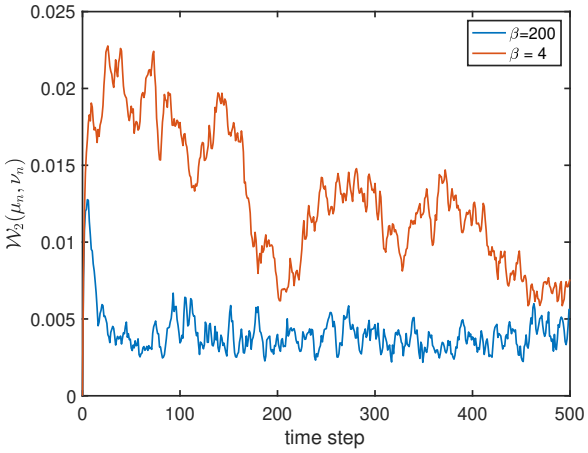


FIGURE 5.7: Time-evolution of Wasserstein distance between the particle ensemble $\nu_n(q, p)$ and the sample ensemble $\mu_n(q, p)$ for $\beta = 4$ and $\beta = 200$.

In Figures 5.8 and 5.9 we compare the particle filter approximation of an evolving measure with particle number $J = 20$ and sample path ensemble sizes $K = 20$ and $K = 30$. For both simulations we draw initial distributions for both sample paths and particles from (5.4.13). We use $\beta = 4$ and $M = (0.1)^{-2}I$ in (5.2.10) (consistent with experiments in the previous section). We see that the sample path means of both the q- and p-components are well approximated.

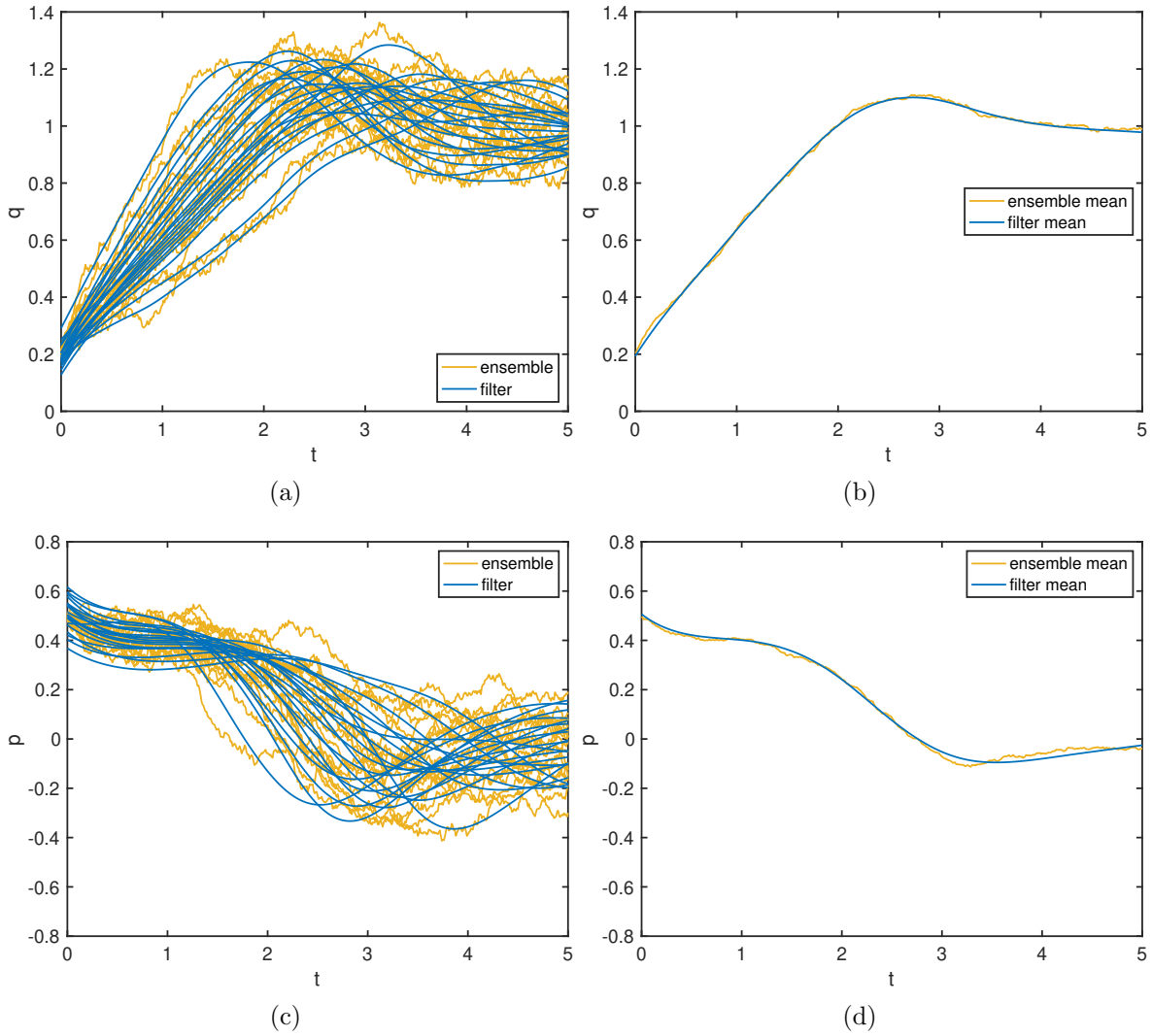


FIGURE 5.8: Multiple sample paths of an SDE. An ensemble of $K = 20$ sample paths of the system (5.4.7)–(5.4.8) are plotted as yellow curves in (a) (q-component) and (c) (p-component). The corresponding particle filter paths ($J = 20$) are plotted as blue curves in (a) and (c). The ensemble means and particle filter means are compared in (b) and (d). For these simulations, $\beta = 4$.

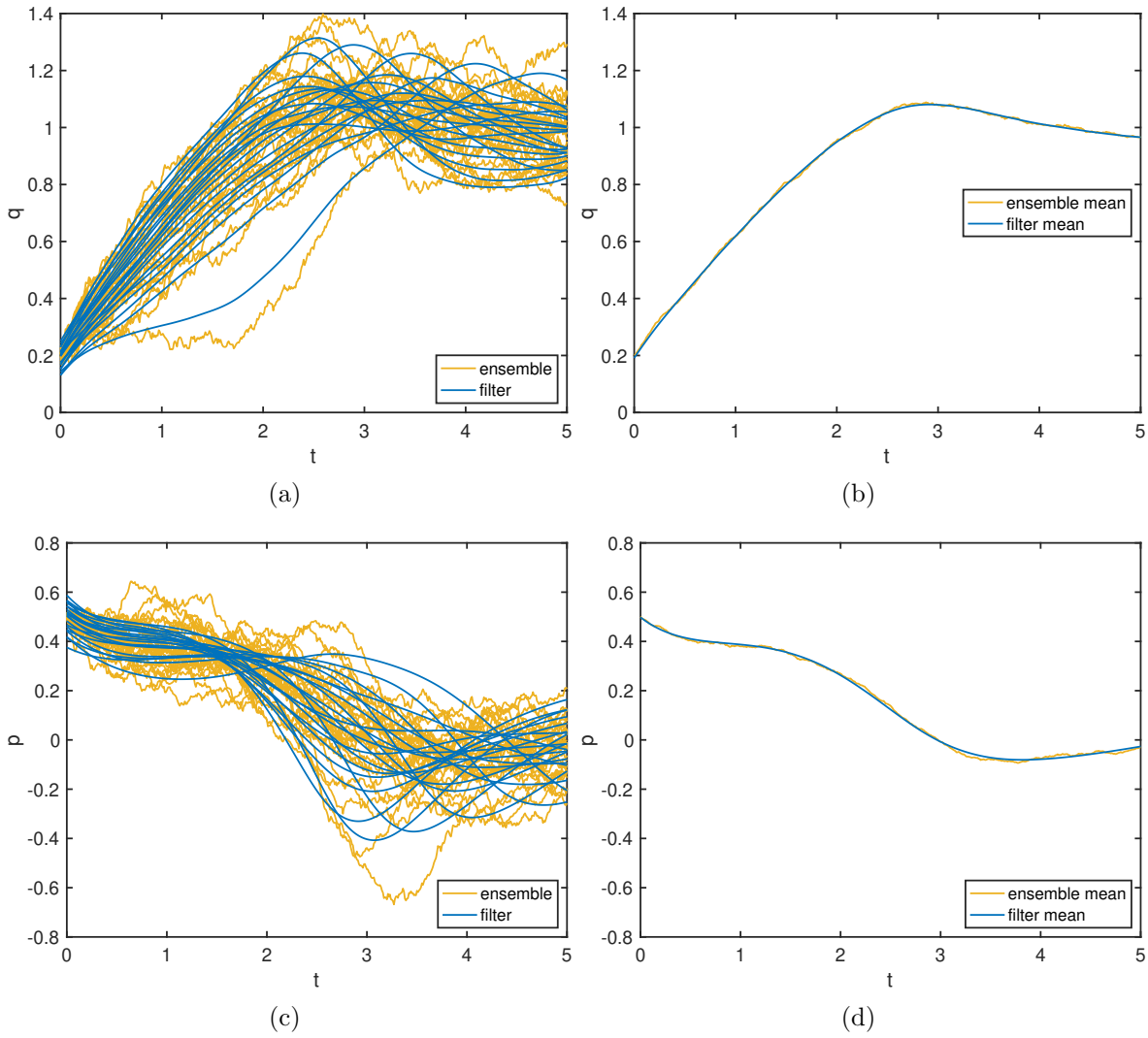


FIGURE 5.9: Same as Figure 5.9, but with $K = 30$. An ensemble of sample paths of the system (5.4.7)–(5.4.8) are plotted as yellow curves in (a) (q-component) and (c) (p-component). The corresponding particle filter paths with particle number $J = 20$ are plotted as blue curves in (a) and (c). The ensemble means and particle filter means are compared in (b) and (d). For these simulations, $\beta = 4$.

5.5 Conclusion

In this chapter, we construct a particle filter in the form of an optimal control that minimizes mismatch in the Wasserstein distance on observation space. Numerical examples show that the Wasserstein distance between the empirical measure on the whole state space is well bounded over the assimilation window. We compared scenarios with (i) deterministic (chaotic) dynamics with uncertainty in initial conditions, (ii) a single sample path of an SDE with multiple uncertain observations, and (iii) multiple sample paths of an SDE with accurate (partial) observations. The method was shown to recover bi-modal probability measures, compare favorably to the ensemble Kalman filter for an SRB measure, and accurately reproduce the sample path mean for latter scenario. The numerical implementation used was suboptimal, as is a topic for future research.

Chapter 6

Summary

In this thesis, we mainly study a numerical method for solving optimal control problems and apply it to Cucker-Smale dynamics and data assimilation.

Pontryagin's maximum principle, which provided the necessary condition for optimal control problems, results in two-point boundary problem. One numerical method that is easy to employ for such problems is the so-called "forward-backward sweep" method. However this method is not always convergent especially when applied to non-linear systems. In this thesis, in Chapter 3, we extend the "regularised forward-backward sweep iteration method" from the continuous setting in paper [77] to the discrete setting for solving optimal control problems. The continuous problem is discretized by using a variational integrator which yields a symplectic method. The regularised forward-backward iteration method depends on a regularization parameter ρ. We provide the proof that when ρ is large enough, the forward-backward sweep method is convergent all the time. The proof is firstly given for the first order symplectic Euler method, then it is extended to general symplectic Runge-Kutta methods. According to the proof, the parameter ρ depends on the length of the time window and the Lipschitz constant. Numerical experiments illustrate convergence, which may still be slow, especially for large ρ. However, significant speed up can be realized using Anderson acceleration.

The Cucker-Smale model is a conceptual model of flocking, in which a group of agents attempt to synchronize into uniform motion. The conditions for synchronization to occur depend on the initial condition and on the force. When these conditions are not met, some extra control may be added to the Cucker-Smale model to make the dynamic tend to consensus. In this contex, it is interesting to consider 'sparse control', in which steering by the external controller is limited to a small number of finite actions. We apply the method of Chapter 3 to this problem in Chapter 4. The optimal control cost functional combines distance to velocity consensus and the magnitude of the control in a class of so-called ℓ_p-ℓ_q-norms. In this chapter, we focus on discussing the ℓ_2-ℓ_2-norm, ℓ_1-ℓ_1-norm, ℓ_2-ℓ_1-norm on the control. The results in these three different cases show that the optimal controls become 0 or asymptotically tend to 0, after a finite time period. We calculate the optimal control in these three cases, and find the optimal control is unbounded in ℓ_1-norm, therefore we implemented constrained controls. Meanwhile, to avoid slow convergence due to discontinuity of the control, we add a soft-constraint δ and study the effect of

the smoothness of the control on the convergence of the regularized forward-backward sweep iteration. Under the condition of ℓ_1-ℓ_1 norm, the experiment shows that the control is a bang-bang controller, either zero or sharply constrained, before the control totally goes to 0. The optimal control under the ℓ_2-ℓ_1 norm is also sparse. After studying the effect of the soft-constraint parameter δ on the iteration convergence, we find that the convergence is faster with smaller parameter δ, however the solution is more smooth and less accurate.

In Chapter 5, we proposed a new data assimilation algorithm, to utilizes the probability distribution of an ensemble of controlled particles to quantify uncertainty in stochastic systems. Most importantly, the controlled dynamical system for the particles is deterministic. In the end, the method is defined as an optimal control problem. The cost function is composed of the norm of the control function and the Wasserstein distance on the observation space. To solve this optimal control problem, we apply the regularised forward-backward sweep iteration mentioned in chapter 3. Two different sampling processes were studied. With the first situation, we take many noisy samples of the observable along a distinct sample path. Alternatively, we take single observations of an ensemble of sample paths. The state and observation error distributions are assumed to be Gaussian. We compared cases with more particles than observations and fewer particles than observations. Experiments with a (bi-modal) double well potential indicate good results, with no evidence of ensemble collapse. We also compare to the ensmble Kalman filter for deterministic ensemble simulation of the Lorenz-63 model to illustrate the advantage of the Wasserstein distance for dynamics on a strange attractor. We find the cost of the method to be high, but improvement may be possible with more efficient implementation of the Wasserstein distance (e.g. using the Sinkhorn algorithm).

Chapter 7

Nederlandse samenvatting

Dit proefschrift is voornamelijk gewijd aan een oplossingsmethode voor optimale regelproblemen en de toepassing hiervan op Cucker-Smale dynamische systemen en data-assimilatie.

Het maximumprincipe van Pontryagin verschaft noodzakelijke voorwaarden voor oplossingen van optimale regelproblemen voor tweepunts-randwaardeproblemen die het mogelijk maken om numerieke oplossingsmethoden als de "forward-backward sweep" toe te passen. Convergentie van deze methode is echter niet gegarandeerd, wat met name een probleem is voor niet-lineaire systemen. In hoofdstuk 3 van dit proefschrift breiden we de zogeheten "geregulariseerde forward-backward sweep" methode uit het artikel [77] uit naar een discrete context voor toepassing op discrete optimale regelproblemen. Het oorspronkelijke continue regelprobleem wordt gediscretiseerd door middel van een variationele integrator die symplectisch is. We tonen aan dat er altijd convergentie optreedt voor voldoende grote waarden van de regularisatieparameter ρ die gebruikt wordt bij de geregulariseerde forward-backward iteratiemethode. Het bewijs hiervoor geven we eerst alleen voor de eerste orde symplectische Eulermethode, om het resultaat vervolgens uit te kunnen breiden naar Runge-Kutta methoden in het algemeen door deze tot de Eulermethode te reduceren. Uit het bewijs blijkt dat de parameter ρ afhankelijk is van zowel de tijdsduur als van de Lipschitz-constante. Experimenteel kan eenvoudig worden vastgesteld dat de regularisatieparameter ρ en de convergentiesnelheid negatief gecorreleerd zijn, wat de vraag hoe de convergentiesnelheid verbeterd kan worden bijzonder interessant maakt. De Anderson versnellingsmethode is toegepast in het beschreven experiment.

Aangezien de convergentie van de Cucker–Smale dynamica bepaalde beginvoorwaarden vereist wordt een extra regelgrootheid toegevoegd aan het Cucker–Smale model om te garanderen dat de dynamica naar een consensus convergeert. Dit wordt in hoofdstuk 4 beschreven. Het probleem wordt vervolgens omgezet in een optimaal regelprobleem waarbij de kostenfunctie de afstand tot consensus combineert met de grootte van het regelsignaal in de $\ell_p - \ell_q$-norm. In dit hoofstuk kijken we voornamelijk naar de $\ell_2 - \ell_2$-norm, de $\ell_2 - \ell_1$-norm en de $\ell_1 - \ell_1$-norm op het regelsignaal. Het resultaat voor alledrie deze gevallen is dat de optimale regeling 0 wordt of uiteindelijk asymptotisch naar 0 neigt. We berekenen de optimale regeling in deze gevallen en vinden zo dat er een optimale regeling bestaat op een ∞-punt in de ℓ_1-norm, en hiermee dat begrenzing van het regelsignaal

noodzakelijk is. Ondertussen zorgen we ervoor dat de regeling op de rand differentieerbaar is door een zachte beperking δ op de rand toe te voegen. We bestuderen het effect van de gladheid van de regeling op de convergentie van de forward-backward sweep iteratie. Voor de $\ell_1 - \ell_1$-norm toont het experiment aan dat de regeling een aan-uit regeling is, die ofwel nul is ofwel op de rand zit, totdat deze geheel naar 0 gaat. Het optimale regelsignaal voor de $\ell_2 - \ell_1$-norm is ook ijl. Wanneer we het effect van de parameter δ op de convergentie van de iteratie bestuderen zien we een snellere convergentie voor kleinere δ, maar ook dat de gevonden oplossingen minder glad en minder nauwkeurig worden.

In hoofdstuk 5 dragen we een nieuw algoritme aan voor data-assimilatie dat gebruik maakt van de kansverdeling van de gecontroleerde dynamica van deeltjes om een stochastisch systeem te benaderen. Wat hierbij vooral belangrijk is is dat het dynamische systeem deterministisch is. Uiteindelijk wordt het algoritme geconverteed naar een optimaal regelprobleem, waarbij de kostenfunctie bestaat uit regelfuncties en een Wasserstein-matrix. Om dit optimale regelprobleem op te lossen passen we de geregulariseerde backward-forward sweep uit hoofdstuk 3 toe. Er kleven wel een paar nadelen aan deze methode, zoals het feit dat de Wasserstein-afstand niet eenvoudig te bepalen valt omdat deze voor elk moment in de tijd berekend moet worden.

Omdat in dit hoofdstuk wordt aangenomen dat de toestands- en waarnemingsfouten beide Gaussisch verdeeld zijn hebben we twee verschillende bemonsteringsmethoden bekeken. Bij de eerste methode bemonsteren we de waarneming veelvuldig en beschouwen we de resultaten als echte waarnemingen. In het tweede geval bemonsteren we zowel de toestand als de waarneming herhaaldelijk en gebruiken we de gevonden waarden in plaats van de echte toestand en waarneming. Tijdens het experiment hebben we de relatie tussen het aantal deeltjes en het aantal bemonsteringen onderzocht en hebben we geprobeerd de cumulatieve verdeling van deeltjes en monsters te vergelijken. In alle gevallen blijkt het algoritme goede schattingen te geven voor de ware toestand van het systeem. In het Lorenz 63 experiment maken we een vergelijking met het ensemble Kalman filter en zien we dat onze methode soms beter werkt.

Al met al zien we dat de geregulariseerde forward-backward sweep iteratiemethode convergentie kan garanderen en zo kan helpen om optimale regelproblemen op te lossen. In sommige gevallen hebben we deze methode kunnen combineren met versnellingsmethoden zoals de Anderson versnellingsmethode. Er zijn nog wel een aantal problemen die nader bekeken moeten worden. Wat is de relatie tussen de regularisatieparameter ρ en de convergentiesnelheid? Wanneer we de geregulariseerde forward-backward methode toepassen op het Cucker–Smale model introduceren we ook de split-methode. Er zal verder onderzocht moeten worden of deze methode ook effectief gecombineerd kan worden met de geregulariseerde forward-backward methode.

Bibliography

[1] L. Abia and J.M. Sanz-Serna. Partitioned Runge-Kutta methods for separable Hamiltonian problems. *Mathematics of Computation*, 60(202):617–634, 1993.

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory Algorithms and Applications*. Prentice Hall, 1993.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[5] A. Aydoğdu, M. Caponigro, S. McQuade, B. Piccoli, N. P. Duteil, F. Rossi, and E. Trélat. Interaction network, state space, and control in social dynamics. In *Active Particles, Volume 1*, pages 99–140. Springer, 2017.

[6] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5:19–53, 2011.

[7] Rafael Bailo, Mattia Bongini, José A Carrillo, and Dante Kalise. Optimal consensus control of the Cucker-Smale model. *IFAC-PapersOnLine*, 51(13):1–6, 2018.

[8] Alan Bain and Dan Crisan. *Fundamentals of Stochastic Filtering*. Springer New York, 2009.

[9] V. M. Becerra. Optimal control. *Scholarpedia*, 3(1):5354, 2008.

[10] L. D. Berkovitz. *Optimal control theory*, volume 12. Springer Science & Business Media, 2013.

[11] Dimitri P Bertsekas. *Linear Network Optimization: Algorithms and Codes*. MIT Press, 1991.

[12] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control I and II*. Athena Scientific, Belmont, MA, fourth edition, 2005.

[13] J. E. Bobrow, S. Dubowsky, and J. S. Gibson. Time-optimal control of robotic manipulators along specified paths. *The International Journal of Robotics Research*, 4(3):3–17, 1985.

[14] M. Bongini, M. Fornasier, and D. Kalise. (Un)conditional consensus emergence under perturbed and decentralized feedback controls. *Discrete and Continuous Dynamical Systems*, 35:4071–4094, 2015.

[15] F. Bouttier and P. Courtier. Data assimilation concepts and methods March 1999. *Meteorological training course lecture series. ECMWF*, 718:59, 2002.

[16] A. E. Bryson. *Applied optimal control: optimization, estimation and control*. CRC Press, 1975.

[17] Amarjit Budhiraja, Lingji Chen, and Chihoon Lee. A survey of numerical methods for nonlinear filtering problems. *Physica D: Nonlinear Phenomena*, 230(1-2):27–36, 2007.

[18] J. C. Butcher. *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods*. Wiley-Interscience, 1987.

[19] A. Calogero. Notes on optimal control theory. Lecture notes, 2020.

[20] M. Caponigro, M. Fornasier, B. Piccoli, and E. Trélat. Sparse stabilization and optimal control of the Cucker-Smale model. *Mathematical Control and Related Fields*, 3(4):447–466, 2013.

[21] M. R. Caputo and M. R. Caputo. *Foundations of dynamic economic analysis: optimal control theory and applications*. Cambridge University Press, 2005.

[22] René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I AND II*, volume 83 and 84. Springer, 2018.

[23] James Carpenter, Peter Clifford, and Paul Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7, 1999.

[24] Eduardo Casas, Christian Clason, and Karl Kunisch. Approximation of elliptic control problems in measure spaces with sparse solutions. *SIAM Journal on Control and Optimization*, 50(4):1735–1752, 2012.

[25] P. J. Channell and C. Scovel. Symplectic integration of Hamiltonian systems. *Nonlinearity*, 3(2):231, 1990.

[26] M. Chyba, E. Hairer, and G. Vilmart. The role of symplectic integrators in optimal control. *Optimal Control Applications and Methods*, 30(4):367–382, 2009.

[27] F. Clarke. *Functional analysis, calculus of variations and optimal control*, volume 264. Springer Science & Business Media, 2013.

[28] Christian Clason and Karl Kunisch. A duality-based approach to elliptic control problems in non-reflexive banach spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(1):243–266, 2011.

[29] Christian Clason and Karl Kunisch. A measure space approach to optimal source placement. *Computational Optimization and Applications*, 53(1):155–171, 2012.

[30] Philippe Courtier, E Andersson, W Heckley, D Vasiljevic, M Hamrud, A Hollingsworth, F Rabier, M Fisher, and J Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3d-var). i: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

[31] A.J. Craig and I. Flügge-Lotz. Investigation of optimal control with a minimum-fuel consumption criterion for a fourth-order plant with two control inputs; synthesis of an efficient suboptimal control. *Journal of Fluids Engineering*, 87:39–57, 1965.

[32] F. Cucker and S. Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007.

[33] F. Cucker and S. Smale. On the mathematics of emergence. *Japanese Journal of Mathematics*, 2(1):197–227, 2007.

[34] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems, NIPS 26*, pages 2292–2300, 2013.

[35] Bart De Leeuw, Svetlana Dubinkina, Jason Frank, Andrew Steyer, Xuemin Tu, and Erik Van Vleck. Projected shadowing-based data assimilation. *SIAM Journal on Applied Dynamical Systems*, 17(4):2446–2477, 2018.

[36] Rene De Vogelaere. Methods of integration which preserve the contact transformation property of the Hamilton equations. *Technical report (University of Notre Dame. Dept. of Mathematics)*, 1956.

[37] Eustasio del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27(2):1009–1071, 1999.

[38] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12(656-704):3, 2009.

[39] Weinan E, Jiequn Han, and Qianxiao Li. A mean-field optimal control formulation of deep learning. *Research in Mathematical Sciences*, 6(10), 2018.

[40] L. D. Elsgolc. *Calculus of variations*. Courier Corporation, 2012.

[41] Etienne Emmrich. Discrete versions of Grönwall's lemma and their application to the numerical analysis of parabolic problems. Technical Report 637, T.U. Berlin, 1999.

[42] G. 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.

[43] Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.

[44] Geir Evensen. *Data assimilation: the ensemble Kalman filter*. Springer, 2009.

[45] S.J Fletcher. *Data assimilation for the geosciences: From theory to application*. Elsevier, 2017.

[46] Massimo Fornasier. Learning and sparse control of multiagent systems. In *European Congress of Mathematics*, volume 7, 2016.

[47] B. A. Forster. Optimal consumption planning in a polluted environment. *Economic Record*, 49(4):534–545, 1973.

[48] Jason Frank and Sergiy Zhuk. Symplectic Möbius integrators for LQ optimal control problems. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 6377–6382. IEEE, 2014.

[49] Keisuke Fujii. Extended Kalman filter. *Refernce Manual*, pages 14–22, 2013.

[50] I. M. Gelfand and R. A. Silverman. *Calculus of variations*. Courier Corporation, 2000.

[51] Sun Geng. Symplectic partitioned Runge-Kutta methods. *Journal of Computational Mathematics*, pages 365–372, 1993.

[52] M Giaquinta and S Hildebrandt. *Calculus of variations II*, volume 311. Springer Science & Business Media, 2013.

[53] Cecilia González-Tokman and Brian R Hunt. Ensemble data assimilation for hyperbolic systems. *Physica D: Nonlinear Phenomena*, 243(1):128–142, 2013.

[54] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.

[55] S. Y. Ha, T. Ha, and J. H. Kim. Emergent behavior of a Cucker-Smale type particle model with nonlinear velocity couplings. *IEEE transactions on automatic control*, 55(7):1679–1683, 2010.

[56] S. Y. Ha and J. G. Liu. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Communications in Mathematical Sciences*, 7(2):297–325, 2009.

[57] Seung-Yeal Ha and Eitan Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *arXiv preprint arXiv:0806.2182*, 2008.

[58] William W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87:247–282, 2000.

[59] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2006.

[60] Ernst Hairer, Syvert P Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin,, second edition, 1993.

[61] Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, second edition, 1996.

[62] Nicholas C. Henderson and Ravi Varadhan. Damped anderson acceleration with restarts and monotonicity control for accelerating em and em-like algorithms. *Journal of Computational and Graphical Statistics*, 28(4):834–846, 2019.

[63] Xiang-Yu Huang, Qingnong Xiao, Dale M Barker, Xin Zhang, John Michalakes, Wei Huang, Tom Henderson, John Bray, Yongsheng Chen, and Zaizhong Ma. Four-dimensional variational data assimilation for wrf: Formulation and preliminary results. *Monthly Weather Review*, 137(1):299–314, 2009.

[64] Maria Isabel Maria. Kalman and extended Kalman filters: Concept, derivation and properties. *Institute for Systems and Robotics*, 43:46, 2004.

[65] Oliver Junge, Jerrold E Marsden, and Sina Ober-Blöbaum. Discrete mechanics and optimal control. *IFAC Proceedings Volumes*, 38(1):538–543, 2005.

[66] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82:35–45, 1960.

[67] H. K. Khalil and J. W. Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

[68] H. Kinoshita, H. Yoshida, and H. Nakai. Symplectic integrators and their application to dynamical astronomy. *Celestial Mechanics and Dynamical Astronomy*, 50:59–71, 1991.

[69] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[70] I. Kitiashvili, A.G Kosovichev. Application of data assimilation method for predicting solar cycles. *The Astrophysical Journal Letters*, 688(1):L49, 2008.

[71] H. R. Künsch. Particle filters. *Bernoulli*, 19(4):1391–1403, 2013.

[72] F.M. Lasagni. Canonical Runge-Kutta methods. *ZAMP*, 39(6):952–953, 1988.

[73] Kody Law, Andrew Stuart, and Kostas Zygalakis. Data assimilation. *Cham, Switzerland: Springer*, 214, 2015.

[74] Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2005.

[75] S. Lenhart and J. T. Workman. *Optimal control applied to biological models*. CRC press, 2007.

[76] F. L. Lewis, D. Vrabie, and V. L. Syrmos. *Optimal control*. John Wiley & Sons, 2012.

[77] Qianxiao Li, Long Chen, Cheng Tai, and E Weinan. Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026, 2017.

[78] X. Liu and J. Frank. Symplectic Runge-Kutta discretization of a regularized forward-backward sweep iteration for optimal control problems. *Journal of Computational and Applied Mathematics*, 383:113113, 2021.

[79] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

[80] T. Manzoor, S. Aseev, E. Rovenskaya, and A. Muhammad. Optimal control for sustainable consumption of natural resources. *IFAC Proceedings Volumes*, 47(3):10725–10730, 2014.

[81] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numerica 2001*, 10:357—514, May 2001.

[82] Michael McAsey, Mou Libin, and Han Weimin. Convergence of the forward-backward sweep method in optimal control. *Computational Optimization and Applications*, 53:207–226, 2012.

[83] R. I. McLachlan and C. Offen. Symplectic integration of boundary value problems. *Numerical Algorithms*, 81:1219–1233, 2019.

[84] C. R. Menyuk. Some properties of the discrete Hamiltonian method. *Physica D: Nonlinear Phenomena*, 11(1-2):109–129, 1984.

[85] A.A. Milyutin and N.P. Osmolovskii. Calculus of variations and optimal control. *Chapman and Hall CRC Research Notes in Mathematics*, pages 159–172, 1999.

[86] H. Nakai, H. Kinoshita, and H. Yoshida. Applications of the symplectic integrator to celestial mechanics. In *Twenty-Third Symposium on Celestial Mechanics*, pages 7–15, 1990.

[87] V. Nenchev, C. G. Cassandras, and J Raisch. Optimal control for a robotic exploration, pick-up and delivery problem. *arXiv preprint arXiv:1607.01202*, 2016.

[88] Sina Ober-Blöbaum. *Discrete mechanics and optimal control*. PhD thesis, Universtät Paderborn, 2004.

[89] A. M. Oberman and Y. L. Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.

[90] Luigi Palatella, Alberto Carrassi, and Anna Trevisan. Lyapunov vectors and assimilation in the unstable subspace: theory and applications. *Journal of Physics A: Mathematical and Theoretical*, 46(25):254020, 2013.

[91] Jaemann Park, H. Jin Kim, and Seung-Yeal Ha. Cucker-Smale flocking with inter-particle bonding forces. *IEEE Transactions on Automatic Control*, 55(11):2617–2623, 2010.

[92] Lev Semenovich Pontryagin. *Mathematical Theory of Optimal Processes*. Routledge, 2018.

[93] S. Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[94] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[95] Rolf H Reichle. Data assimilation methods in the earth sciences. *Advances in Water Resources*, 31(11):1411–1418, 2008.

[96] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, 1987.

[97] Ludger Rüschendorf. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

[98] R. D. Ruth. A canonical integration technique. *IEEE Trans. Nucl. Sci.*, 30(CERN-LEP-TH-83-14):2669–2671, 1983.

[99] H. Riazi S Kim, D.J. Seo and C. Shin. Improving water quality forecasting via data assimilation–application of maximum likelihood ensemble filter to HSPF. *Journal of Hydrology*, 519:2797–2809, 2014.

[100] Satoshi Saito, Hiroshi Sugiura, and Taketomo Mitsui. Family of symplectic implicit Runge-Kutta formulae. *BIT Numerical Mathematics*, 32(3):539–543, 1992.

[101] J. M. Sanz-Serna. Symplectic Runge-Kutta schemes for adjoint equations, automatic differentiation, optimal control, and more. *SIAM Review*, 58(1):3–33, 2016.

[102] Jesus-Maria Sanz-Serna and Mari-Paz Calvo. *Numerical Hamiltonian Problems*, volume 7 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1994.

[103] J.M. Sanz-Serna. Runge-Kutta schemes for Hamiltonian systems. *BIT Numerical Mathematics*, 28(4):877–883, 1988.

[104] R.W.H Sargent. Optimal control. *Journal of Computational and Applied Mathematics*, 124(1-2):361–371, 2000.

[105] S. P. Sethi and G. L. Thompson. *What is optimal control theory?* Springer, 2000.

[106] Jackie Shen. Cucker–Smale flocking under hierarchical leadership. *SIAM Journal on Applied Mathematics*, 68(3):694–719, 2008.

[107] Qiang Song, Fang Liu, Jinde Cao, and Jianlong Qiu. Cucker-Smale flocking with bounded cohesive and repulsive forces. In *Abstract and Applied Analysis*, volume 2013. Hindawi, 2013.

[108] M. Speekenbrink. A tutorial on particle filters. *Journal of Mathematical Psychology*, 73:140–152, 2016.

[109] Y.B. Suris. On the canonicity of mappings that can be generalized by methods of Runge–Kutta type for integrating systems $x'' = -\partial U/\partial x$. *U.S.S.R. Comput. Math. and Math. Phys.*, 29(1):138–144, 1989.

[110] Y.B. Suris. Hamiltonian methods of Runge–Kutta type and their variational interpretation. *Math. Simulation*, 2(4):78–87, 1990.

[111] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambrdige MA., second edition, 2018.

[112] C. Tameling, M. Sommerfeld, and A. Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781, 2019.

[113] A. Toth and C. T. Kelley. Convergence analysis for Anderson acceleration. *SIAM Journal on Numerical Analysis*, 53(2):805–819, 2015.

[114] Anna Trevisan, Massimo D'Isidoro, and Olivier Talagrand. Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 136(647):487–496, 2010.

[115] Cédric Villani. *Topics in Optimal Transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Soc., 2003.

[116] R. Vinter. *Optimal control*. Springer Science & Business Media, 2010.

[117] G. Vossen and H. Maurer. On L^1-minimization in optimal control and applications to robotics. *Optimal Control Applications and Methods*, 27(6):301–321, 2006.

[118] Gerd Wachsmuth and Daniel Wachsmuth. Convergence and regularization results for optimal control problems with sparsity functional. *ESAIM: Control, Optimisation and Calculus of Variations*, 17(3):858–886, 2011.

[119] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49:1715–1735, 2011.

[120] Z. Wang and Y. Li. An indirect method for inequality constrained optimal control problems. *IFAC-PapersOnLine*, 50(1):4070–4075, 2017.

[121] Greg Welch and Gary Bishop. An introduction to the Kalman filter. Technical Report TR 95-041, Department of Computer Science, University of North Carolina, Chapel Hill, 1995.

[122] Tao Yang, Prashant G Mehta, and Sean P Meyn. A mean-field control-oriented approach to particle filtering. In *Proceedings of the 2011 American Control Conference*, pages 2037–2043. IEEE, 2011.

[123] Tao Yang, Prashant G Mehta, and Sean P Meyn. Feedback particle filter. *IEEE transactions on Automatic control*, 58(10):2465–2480, 2013.

[124] Xi-Xin Yang, Gong-You Tang, Yang Li, and Pei-Dong Wang. Optimal formation control for multi-agents systems with external disturbances. In *Proceedings of the 31st Chinese Control Conference*, pages 2291–2295. IEEE, 2012.

[125] Y. N. Yu, K. Vongsuriya, and L. N. Wedman. Application of an optimal control theory to a power system. *IEEE Transactions on Power Apparatus and Systems*, PAS-89:55–62, 1970.

[126] Sergiy Zhuk, Jason Frank, Isabelle Herlin, and Robert Shorten. Data assimilation for linear parabolic equations: minimax projection method. *SIAM Journal on Scientific Computing*, 37(3):A1174–A1196, 2015.

Acknowledgements

Time goes fast, the Covid-19 has already lasted more than one year. During this special time, I am going to finish my PhD journey. The scene when I arrived at this amazing country seems to have happened yesterday. At that time, even though the totally different culture and environment made me a little nervous, the warm-hearted people made me feel happy and release the anxiety feeling. This country is full of magic to make people slow down to enjoy the life. Thanks for the doctors working in the first line to fight the virus and hoping the world comes back to the normal.

During my PhD process, first and foremost, I would like to express my deep gratitude towards my supervisor Jason Frank for his expert guidance. Dear Jason, thank you so much for supervising me in the last four years. You have always put a lot of effort in guiding and educating me in patience, as well as other students. I enjoyed working together with you because no matter what kind of questions about mathematics, you always gave me the most professional and detailed answers. And you are so kind and warm-hearted, Especially at times when the lack of progress made me somewhat pessimistic, you always gave me warm encouragement which was an important stimulant towards the completion of the PhD program. And your abundant mathematical knowledge always inspires me to be a mathematician like you.

I also feel grateful to the committee members for spending time reading my thesis and providing helpful feedback.

I feel grateful to have this opportunity to show my thanks to my colleges, Arjen and Fabian, thank you for inviting me to eat lunch together and introducing the Dutch culture to me, which helped me a lot to live happily in Netherlands. Especially, I want to show my greatest thanks to Arjen who spent his precious time to help me to translate my summary into dutch. I would also like to thank my other former and current colleges, Lasse, Leandro, Felix, Hong, Han. Ralph, David, Thomas, Giacomo for your kindness and help.

I am thankful to all the current and previous members of the Mathematical Institute for providing me the comfort environment to carry out my research. Thanks for the secretary staff: Cecile, Linda, Jean and Carin for their sincere assistance and kindness during my stay in a fantastic mathematical institute.

I want to show my thanks to my best friend, Bohui. I would never forget that you show the greatest kindness to me when I was depressed and helpless. You are almost my family member and we accompany each other to finish the PhD journey. To another best friend

Tao shi, I would never forget the enjoyable talking time with you and Bohui. Without you and Bohui, I couldn't imagine the life in the Netherlands. I also show my thanks to my other Chinese friends, Weiyang, Xiao Gang, Jingwen, Chuan for sharing beautiful life moments.

At the same time, I am deeply indebted and grateful to my parents: You are the best and most marvelous parents in the world. Without the love, help, understanding and freedom from you, I would not be where I am today. To my older brother and my sister-in-law, thank you for supporting me and encouraging me to do what I want to. To my lovely cousins, you guys always listened to my complaints with patience.

In the end, I want to show my thanks to the Chinese Scholarship Council, which supported my whole four year life in Netherlands.

Curriculum Vitae

Xin Liu was born on 18th September (Lunar Calendar), 1990 in Nanyang, a small city of Henan Province in China. After finishing high school, she went to Zhengzhou University studying Applied Mathematics in September, 2009.

Four years later, she was recommended for postgraduate study without taking the entrance examinations in Zhengzhou University and obtained National Scholarship to do her Master. In the years 2013 - 2016, she did research project in the field of building financial market models. In 2016, She performed her master thesis on 'A Quantum Model And Data Mining Technology for Financial Market' under the guidance of Prof. Haijun Liu

In 2016, she obtained the Chinese Scholarship Council Scholarships and started a Ph.D project in the Mathematical Institute in November at Utrecht University in the Netherlands under the supervision of Prof. Jason Frank. During this time she was a member of SIAM student Chapter.