Markov Decision Processes and Discrete-Time Mean-Field Games Constrained with Costly Observations

Jonathan Yick Yeung Tam

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

Acknowledgements

Firstly, I would like to express my profound gratitude towards my supervisor, Prof. Christoph Reisinger, for his invaluable guidance and positive encouragement throughout my doctoral studies. His expertise and constructive feedback has consistently pushed me to strive for excellence. His utmost professionalism is something that I aspire towards as I move further along the paths of academia.

I would like to express my gratefulness to Prof. Dirk Becherer from Humboldt University of Berlin, who has co-supervised me for the last two years, and has been the most welcoming host during my time in Berlin. I appreciate his comprehensive and practical advice, and our frequent discussions have often lead to many thoughtful ideas and fruitful results.

I would like to thank Prof. Sam Cohen, Prof. Michael Monoyios, and Prof. Renyuan Xu for their feedback for my transfer and confirmation of status examinations. I would also like to extend my appreciation to the committee of the Mathematics of Random Systems CDT at Oxford. Without their hard efforts, I would not be able to embark on this exceptional journey.

Last but not least, I would like to thank all my friends and family. In particular:

To Alain, Felix, Julian, Regan and Siobhan, thank you for riding the ebbs and flows with me through the last four years. There has been so many unforgettable memories and I am glad to have you all as my closest friends. To Sheh and Gaia, we managed to wade through an unprecedented pandemic and I am most thankful for the company and sense of normality you provided during the harshest lockdown period. To Justin and Ethan, thank you for keeping in touch regularly, even though I have moved away from Hong Kong for more than a decade now, it is certainly not something I take for granted lightly.

To Leanne, thank you for always being there no matter the circumstances and giving me the strength to always push on. I am grateful to have you as my partner and I cherish your love and support. Most importantly, to Mum and Dad, thank you for your unwavering support behind the scenes, you have been my anchor throughout this journey. I hope I have made you proud with the culmination of this thesis!

Statement of Originality

I declare that this thesis contains no material which has been accepted or is currently being submitted for any other qualifications at the University of Oxford or elsewhere. This thesis contains no material previously published and precise references are made when a previously published result is used or discussed.

This thesis includes the contents of two preprints. Chapter 2 is based on

C. Reisinger and J. Tam. *Markov decision processes with observation costs: framework and computation with a penalty scheme.* arXiv.2201.07908,

and Chapter 3 is based on

D. Becherer, C. Reisinger and J. Tam. *Mean-field games of speedy information access with observation costs.* arXiv:2309.07877.

The contents of the thesis, including both preprints above, are written by myself, under the supervision of Prof. Christoph Reisinger and Prof. Dirk Becherer.

Abstract

In this thesis, we consider Markov decision processes with actively controlled observations. Optimal strategies involve the optimisation of observation times as well as the subsequent action values. We first consider an observation cost model, where the underlying state is observed only at chosen observation times at a cost. By including the time elapsed from observations as part of the augmented Markov system, the value function satisfies a system of quasi-variational inequalities (QVIs). Such a class of QVIs can be seen as an extension to the interconnected obstacle problem. We prove a comparison principle for this class of QVIs, which implies uniqueness of solutions to our proposed problem. Penalty methods are then utilised to obtain arbitrarily accurate solutions. Finally, we perform numerical experiments on three applications which illustrate this model.

We then consider a model where agents can exercise control actions that affect their speed of access to information. The agents can dynamically decide to receive observations with less delay by paying higher observation costs. Agents seek to exploit their active information gathering by making further decisions to influence their state dynamics to maximize rewards. We also extend this notion to a corresponding mean-field game (MFG). In the mean field equilibrium, each generic agent solves individually a partially observed Markov decision problem in which the way partial observations are obtained is itself also subject to dynamic control actions by the agent. Based on a finite characterisation of the agents' belief states, we show how the mean field game with controlled costly information access can be formulated as an equivalent standard mean field game on a suitably augmented but finite state space. We prove that with sufficient entropy regularisation, a fixed point iteration converges to the unique MFG equilibrium and yields an approximate ε-Nash equilibrium for a large but finite population size. We illustrate our MFG by an example from epidemiology, where medical testing results at different speeds and costs can be chosen by the agents.

Contents

List of Figures

Chapter 1

Introduction

The thesis concerns the control of costly observations in stochastic control models. We shall work in discrete time, under the framework of Markov decision processes (MDPs). We will primarily focus on two instances of information structures: the observation cost model, which toggles between no or full observation, and speed of information access, which involves the control of one's observation delay.

The acquisition of information often requires costly effort and time, due to various factors such as resource capacity [73,75], physical constraints [49] and scarce attention [1]. Making optimal decisions whilst balancing costs for information is a general problem that appears in applications including data collection [72, 74] and medical treatment scheduling [67].

Inferring latent states through imperfect information falls under the classical topic of filtering and control under partial observations [10, 34]. However, the information stream there is typically assumed to be fixed and exogeneously given. The models of consideration in this thesis involve a dynamically controlled information stream. Specifically, an agent has to first decide on the quality of their observations, upon which the decision of the next action is based. Thus, the agent aims to exploit their accurate observations in return for better future rewards, at the expense of a higher observation cost.

The absence of observations and/or presence of observation delays are circumstances that frequently occur in models featuring data sampling [29], signal sensing [49,64,70] and network communications [2,3,33]. The specific information structures we choose also allow us to retain a finite structure when solving for a suitably augmented belief

state MDP, so that numerical schemes for discrete frameworks can be employed for obtaining approximate solutions.

Contributions of the thesis

We present the observation cost model (OCM) in Chapter 2, where agents must pay a cost to observe the current state. The sequential nature of the controls (the decision on observing, followed by the change in actions) lead to a quasi-variational inequality (QVI) in dynamic programming. These are structurally different to the Bellman-type equations in standard models. We prove a comparison principle for this class of QVIs, which can be seen as a generalisation of monotone systems with interconnected obstacles [56]. We establish the existence of solutions constructively via a penalty scheme and demonstrate the monotone convergence of the penalised solutions towards the solutions of said QVI. We apply our model to a time-discretised version of the HIV-treatment model [67]. Our results show qualitatively different optimal behaviour when dealing with large observation gaps.

In Chapter 3 we present a Markov Controllable Delay model (MCDM), where an individual agent can exercise dynamic control over the latency of their observations, with less information delay being more costly. We show that this partial information problem is equivalent to solving a finite MDP on an augmented finite state space. We introduce the corresponding Mean Field Game (MFG) where speedy information access is subject to the agents' strategic control decisions. A challenge here is to define the mean-field Nash equilibrium (MFNE) for this problem: although it is defined in terms of the augmented space, the underlying dynamics and rewards still depend on the underlying state distribution. Moreover, due to the finite parametrisation, the barycenter approach for measure-valued belief states in [60] do not apply here. Instead, we exploit this finite parametrisation to explicitly construct a measure flow on the underlying space, given that of the augmented space.

By using a sufficiently strong entropy regularisation in the reward functional, we prove that the regularised MFG has a unique MFNE which is described by a fixed point, and can serve as an approximate Nash equilibrium for a large but finite population size. This extends the results in [6, 23] to partially observable MFGs formulated on infinite horizon with time-dependent measure flows. We demonstrate our model by an epidemiology example, in which we compute both qualitative effects of information delay and cost to the equilibrium, and also the quantitative properties of convergence relating to the entropy regularisation.

For the rest of this chapter, we shall give a brief overview of the introductory material related to the thesis. We first introduce the canonical construction of MDPs, as well as the related POMDP construction, and its equivalence to the belief state MDP. This is important, particularly for the observation cost model in Chapter 2, as we deal with the sequential nature of observations followed by actions. We then give an overview of MFGs in discrete time, together with some of the issues related to the numerical convergence towards MFNE.

1.1 Markov decision processes

We present here the canonical construction of the MDP model, the partially observable model, and its equivalence to an augmented fully observable MDP model, also known as the belief state MDP. The material in this section, as well as Section 1.2, can largely be found in [34, 35].

Definition 1.1. A Markov control model is given by a tuple $\langle \mathcal{X}, A, p, r \rangle$, where

- $\mathcal{X}$ is the state space,
- A is the action space, and for each $x \in \mathcal{X}$ there associates a non-empty $A(x) \subset A$, known as the admissible actions for the state x. The set of state-action admissible pairs is denoted by

$$K \coloneqq \{(x, a) \mid x \in \mathcal{X} \text{ and } a \in A(x)\},$$

- $p : \mathcal{X} \times A \to \mathcal{P}(\mathcal{X})$ is the transition kernel, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability measures on the space $\mathcal{X}$,
- $r : \mathcal{X} \times A \to \mathbb{R}$ is the one-step reward function.

For this thesis, we shall assume throughout that both the state space $\mathcal{X}$ and the action space A are finite. Therefore, the space of measures on $\mathcal{X}$ is equivalent to the simplex on $\mathcal{X}$, and we shall also write $\Delta_{\mathcal{X}}$ for $\mathcal{P}(\mathcal{X})$. In general, the statements in this chapter readily extends to the case of Borel spaces for both $\mathcal{X}$ and A with sufficient continuity and compactness conditions, and can be found in more detail in [35].

At each time $n = 0, 1, \ldots$, the system is observed to be in state $x_n \in \mathcal{X}$, and an action $a_n \in A$ is applied. A reward $r(x_n, a_n)$ is received and the system moves to a new state $x_{n+1} \in \mathcal{X}$ with probability $p(x_{n+1} \mid x_n, a_n)$. A new action $a_{n+1} \in A$ is chosen and the process is repeated. To specify the selection of actions at each time,

the notion of policies is required. To this end, define the history sets $H_0 \coloneqq \mathcal{X}$ and $H_t \coloneqq K^t \times \mathcal{X} = K \times H_{t-1}$ for $t \geq 1$, so that an element $h_t \in H_t$ is of the form

$$h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t),$$

where $(x_n, a_n)_{n=0}^{t-1} \in K^t$ and $x_t \in \mathcal{X}$. A policy can then be considered as a probability distribution over A, dependent on the history $h_t \in H_t$.

Definition 1.2. A policy is a sequence of kernels $\pi = (\pi_t)_{t \geq 0}$, $\pi_t : H_t \to \Delta_A$, such that

$$\pi_t(A \mid h_t) = 1 \quad \text{for all } h_t \in H_t,\ t \geq 0.$$

The set of policies is denoted by Π. Furthermore,

1. A policy π is deterministic if there exists a sequence of functions $\phi_t : H_t \to A$ such that

$$\pi_t(\cdot \mid h_t) = \delta_{\phi_t(h_t)}(\cdot).$$

2. A policy π is Markovian if for all $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t$,

$$\pi_t(\cdot \mid h_t) = \pi_t(\cdot \mid x_t).$$

An MDP can then be defined to be the set of associated canonical probability spaces. Take $(\Omega, \mathcal{F})$, where Ω is the product space K^∞, and $\mathcal{F}$ the corresponding σ-algebra. An element of Ω is then

$$\omega = (x_0, a_0, x_1, a_1, \ldots),$$

so that the state and control variables x_t and a_t can be regarded as projections from Ω to $\mathcal{X}$ and A respectively. The Ionescu-Tulcea Theorem ([34, Appendix C]) states that for any policy $\pi \in \Pi$, and initial condition $x \in \mathcal{X}$, there exists a unique probability measure $\mathbb{P}_x^\pi$ on $(\Omega, \mathcal{F})$ such that

$$\begin{aligned} &\mathbb{P}_x^\pi(x_0, a_0, x_1, a_1, \ldots) \\ &= \delta_x(x_0)\pi_0(a_0 \mid x_0)p(x_1 \mid x_0, a_0)\pi_1(a_1 \mid x_0, a_0, x_1)\ldots \end{aligned}$$

Definition 1.3. A Markov decision process (MDP) is the stochastic process $(\Omega, \mathcal{F}, \mathbb{P}_x^\pi, (x_t))$.

In general, there is not much need to distinguish between a Markov control model and a Markov decision process. Therefore without loss of generality we shall simply use the term MDP to refer to both.

For the optimisation problem, we shall focus on the two variants of finite horizon and discounted infinite horizon problems. For the finite horizon problem, given an MDP, a policy $\pi \in \Pi$, a terminal time $T \geq 0$, and an initial condition $x \in \mathcal{X}$, define the reward functional

$$J(t, x, \pi) = \mathbb{E}_x^\pi \left[\sum_{n=t}^{T-1} r(x_n, a_n) + g(x_T) \right],$$

where $\mathbb{E}_x^\pi$ is the expectation under the measure $\mathbb{P}_x^\pi$, and $g : \mathcal{X} \to \mathbb{R}$ is the terminal reward function. The objective is to maximise the reward functional over the set of policies, that is solving for the value function

$$v(t, x) \coloneqq \sup_{\pi \in \Pi} J(t, x, \pi).$$

The value function can be solved recursively from the terminal condition via the dynamic programming equation, which is also sometimes referred to as the Bellman equation. Intuitively, this states that solving for the optimal policy is the same as solving for the optimal action at each step, assuming that the optimal policy is taken afterwards.

Theorem 1.4. *The value function v is the unique solution to the dynamic programming equation*

$$v(t, x) = \max_{a \in A} \left(r(x, a) + \sum_{y \in \mathcal{X}} v(t+1, y) p(y \mid x, a) \right), \; v(T, x) = g(x).$$

Moreover, a deterministic Markovian policy $\pi^* \in \Pi$ is optimal if and only if for each $0 \leq t < T$, $\pi_t^*(x)$ maximises the RHS of the dynamic programming equation for all $x \in \mathcal{X}$, i.e.

$$v(t, x) = r(x, \pi_t^*(x)) + \gamma \sum_{y \in \mathcal{X}} v(t+1, y) p(y \mid x, \pi_t^*(x)).$$

Therefore, for fully observable MDPs, it is sufficient to consider deterministic Markovian policies (with respect to $\mathcal{X}$) when seeking the optimal policy.

The discounted infinite horizon case proceeds similarly. For any policy $\pi \in \Pi$ and initial state $x \in \mathcal{X}$, define the reward functional

$$J(x, \pi) = \mathbb{E}_x^\pi \left[\sum_{n=0}^{\infty} \gamma^n r(x_n, a_n) \right],$$

where $\gamma \in (0, 1)$ is the discount factor. The value function is given by

$$v(x) \coloneqq \sup_{\pi \in \Pi} J(x, \pi).$$

The corresponding dynamic programming equation is given as follows.

Theorem 1.5. *The value function v is the unique solution to the dynamic programming equation*

$$v(x) = \max_{a \in A} \left(r(x, a) + \gamma \sum_{y \in \mathcal{X}} v(y) p(y \mid x, a) \right).$$

Moreover, a stationary, deterministic and Markovian policy $\pi^* \in \Pi$ is optimal if and only if $\pi^*(x)$ maximises the RHS of the dynamic programming equation for all $x \in \mathcal{X}$, i.e.

$$v(x) = r(x, \pi^*(x)) + \gamma \sum_{y \in \mathcal{X}} v(y) p(y \mid x, \pi^*(x)).$$

1.2 Partially observable MDPs

It is not always the case that one has full knowledge of the underlying process $(x_t)_t$, therefore we have to consider instead a partially observable model.

Definition 1.6. A partially observable control model is defined by the tuple $\langle \mathcal{X}, \mathcal{Y}, A, p, q, q_0, p_0, r \rangle$, where

- $\mathcal{X}$ is the (finite) state space,
- $\mathcal{Y}$ is the observation space (a Borel set),
- A is the (finite) action set,
- $p : \mathcal{X} \times A \to \Delta_{\mathcal{X}}$ is the transition kernel, which gives the transition probabilities of the underlying process,
- $q : A \times \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is the observation kernel,
- $q_0 : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is the initial observation kernel,

– $p_0 \in \Delta_{\mathcal{X}}$ is the initial distribution,

– $r : \mathcal{X} \times A \to \mathbb{R}$ is the one-step reward function.

For the partially observable model, an observation $y_0 \in \mathcal{Y}$ is first generated with probability $q_0(y_0 \mid x_0)$, an action $a_0 \in A$ is applied, so that a reward $r(x_0, a_0)$ is received. The system then moves to $x_1 \in \mathcal{X}$ with probability $p(x_1 \mid x_0, a_0)$ and generates a new observation $y_1 \sim q(\cdot \mid a_0, x_1)$, after which a new action $a_1 \in A$ is applied and the new reward $r(x_1, a_1)$ is received. The process then repeats.

In this setting, policies should only depend on observations, rather than the underlying unobserved states. Therefore, define the (observable) history set $H_0 = \Delta_{\mathcal{X}} \times \mathcal{Y}$ and

$$H_t = H_{t-1} \times A \times Y, \quad t \geq 1.$$

Then a policy $\pi = (\pi_t)_t$ is a sequence of stochastic kernels on A given H_t, the set of which we denote by Π as before. Then, the canonical space is now $\Omega \coloneqq (\mathcal{X} \times \mathcal{Y} \times A)^\infty$. Then once again by the Ionescu-Tulcea theorem, given a policy $\pi \in \Pi$ and initial distribution p_0, there exists a unique probability measure $\mathbb{P}^\pi_{p_0}$ on $(\Omega, \mathcal{F})$ such that

$$\begin{aligned} &\mathbb{P}^\pi_{p_0}(x_0, y_0, a_0, x_1, y_1, a_1, \ldots) \\ &= p_0(x_0) q_0(y_0 \mid x_0) \pi_0(a_0 \; p_0, y_0) p(x_1 \mid x_0, a_0) q(y_1 \mid a_0, x_1) \pi_1(a_1 \mid p, y_0, a_0, y_1) \ldots \end{aligned}$$

Here the value function for, e.g. the discounted infinite horizon problem, reads

$$v(p) \coloneqq \sup_{\pi \in \Pi} \; \mathbb{E}^\pi_p \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \right], \; p \in \Delta_{\mathcal{X}}.$$

Due to the presence of incomplete observations, the problem is non-Markovian in nature, and dynamic programming cannot be directly applied. However, it is possible to consider an equivalent fully observable MDP, by considering the *belief state* as the new underlying state. The belief state $z = (z_t)_t$ can be considered as the conditional distribution of the underlying state, given the observations, so that formally $z_t = \mathbb{P}^\pi_p(x_t \mid h_t)$. The idea is that the belief state contains sufficient information such that the lifted problem now becomes Markovian. The dynamics for $z = (z_t)_t$ is of the form

$$z_{t+1} = H(z_t, a_t, y_{t+1}),$$

where H can be interpreted as

$$z_{t+1}(x) = H(z_t, a, y_{t+1})(x) = \mathbb{P}_p^\pi(x_{t+1} = x \mid h_{t+1}).$$

The existence and construction of such a map H is given in [34, Lemma 3.2].

A policy for the belief MDP is then a sequence of stochastic kernels $\delta = (\delta_t)_t$ on A given I_t, where $I_t \coloneqq Z \times (A \times Z)^t$, and denote the set of admissible belief MDP policies as Π'. Then we can define the value function for the discounted infinite horizon problem for the belief MDP as

$$v'(z) \coloneqq \sup_{\delta \in \Pi'} \mathbb{E}_z^\delta \left[\sum_{t=0}^{\infty} \gamma^t r'(z_t, a_t) \right]$$

where

$$r'(z, a) \coloneqq \sum_{x \in \mathcal{X}} r(x, a) z(x).$$

Note that each $h_t \in H_t$ induces a corresponding $i_t = (z_0, a_0, \ldots, z_{t-1}, a_{t-1}, z_t) \in I_t$. Moreover, Π and Π' can be seen as equivalent, in that any $\delta \in \Pi'$ defines a $\pi^\delta \in \Pi$ by

$$\pi_t^\delta(\cdot \mid h_t) \coloneqq \delta_t(\cdot \mid i_t(h_t)) \quad \text{for all } h_t \in H_t \text{ and } t \geq 0.$$

Furthermore, π_t^δ assigns the same conditional probability on A as that assigned by δ_t for any observable history h_t. That the POMDP and the belief MDP are two equivalent problems is established by the following proposition.

Proposition 1.7 ([34, Theorem 3.11, 3.13])**.** *For any policy $\pi \in \Pi$ there exists an augmented policy $\delta \in \Delta$ such that*

$$J'(\delta, p) = J(\pi, p), \quad \textit{for all } p \in Z = \Delta_{\mathcal{X}}.$$

and moreover

$$v'(p) = v(p), \quad \textit{for all } p \in Z.$$

1.2.1 MDPs with information delay

The information delay model for MDPs first appeared in [4], and can be described by the tuple $\langle \mathcal{X}, A, p, r, d \rangle$, where d is a fixed constant. Suppose that the state of the system at time t is not known until time $t + d$. Assume that once applied, all past actions are known to the user. This differs slightly from the POMDP construction,

in that the user does not receive an observation about the current state of the system (possibly corrupted by noise). Instead, the user observes a past state of the system (thus the notion of the observation kernel q does not apply here). Nevertheless, one can still augment the state space suitably by incorporating the historical observations and action sequence to obtain an equivalent Markovian problem as before.

Let $\mathcal{Y} \coloneqq \mathcal{X} \times A^d$. The suitable notion of policies is denoted as follows.

Definition 1.8. A d-delay policy $\pi = (\pi_t)_t$ is defined as a sequence of transition kernels $\pi : K_t \to \Delta_A$, with $K_d \coloneqq \mathcal{Y}$ and $K_{t+1} \coloneqq K_t \times (A \times \mathcal{X})$ for all $t > d$, such that $\pi(A \mid k_t) = 1$ for all $k_t \in K_t$, $t \geq d$.

Remark 1.9. *We use the terminology d-delay policy here to keep the notation consistent with the rest of the thesis. In [4], these are referred to as N-SDSI-policies.*

The reward functional for the infinite horizon problem here is given by

$$J(y, \pi) = \mathbb{E}_y^\pi \left[\sum_{t=d}^{\infty} \gamma^{t-d} r(x_t, a_t) \right]$$

for all $y \in \mathcal{Y}$, where

$$y \coloneqq (x_0, a_0, \ldots, a_{d-1}).$$

The delay MDP is equivalent to an augmented problem, by considering an augmented process on the space $\mathcal{Y}$. Define the augmented transition kernel $q : \mathcal{Y} \times A \to \Delta_{\mathcal{Y}}$ by

$$q(y' \mid y, a) = \mathbb{1}_{\{(a'_1, a'_2, \ldots, a'_d) = (a_2, \ldots, a_d, a)\}} p(x' \mid x, a_1),$$

for all $y = (x, a_1, a_2, \ldots, a_d)$, $y' = (x', a'_1, a'_2, \ldots, a'_d) \in \mathcal{Y}$ and $a \in A$. Let $\pi' = (\pi'_t)_t$ with $\pi'_t : \mathcal{Y} \to \Delta_A$ be a policy for the augmented MDP. The augmented reward functional is

$$J'(y, \pi') \coloneqq \mathbb{E}_y^{\pi'} \left[\sum_{t=0}^{\infty} \gamma^t r'(y_t, a_t) \right],$$

where

$$r'(y_n, a_n) = \mathbb{E}[r(x_n, a_n) | y_n].$$

Proposition 1.10 ([4, Proposition 2.1])**.** *The d-delay MDP $\langle \mathcal{X}, A, p, r, d \rangle$ is reducible to a fully observable MDP without delays, given by the tuple $\langle \mathcal{Y}, A, q, r' \rangle$,*

Therefore, the information necessary to choose the optimal action is the most recently observed history, alongside with the actions taken up to the current time.

1.3 Finite N-player stochastic game

The notion of an N-player stochastic game generalises the single player game, which is modelled by an MDP in the previous subsection. In the game situation, players are assumed to act rationally and in their own interest, therefore competing against the rest of all the players. Throughout this thesis, we will use the terms players and agents interchangeably.

We shall follow the presentation of [59] for the discrete-time N-player stochastic game. Much like the single agent MDP, this can also be described by a tuple $\langle \mathcal{X}, A, p, r \rangle$. Here both the transition kernel $p : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{X}}$ and the one-stage reward function $r : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \mathbb{R}$ now depend on the state distribution of all the players. For each time $t \geq 0$ and each agent i, let $x_{i,t}^N \in \mathcal{X}$ and $a_i^N \in A$ denote the state and action of agent i at time t respectively, and denote the empirical distribution by

$$e_t^{(N)}(\cdot) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \delta_{x_{i,t}^N}(\cdot) \in \Delta_{\mathcal{X}},$$

where δ_x is the Dirac measure at $x \in \mathcal{X}$. Let $\mu_0 \in \Delta_{\mathcal{X}}$ be the specified initial state distribution. The initial states $x_{i,0}^N$ are distributed i.i.d. according to μ_0. At each time t, agent i receives a reward

$$r(x_{i,t}^N,\ a_{i,t}^N,\ e_t^{(N)}).$$

The N agents then move to their new states $(x_{1,t+1}^N, \ldots, x_{N,t+1}^N)$ at time $t+1$, with probability

$$\prod_{i=1}^{N} p(x_{i,t+1}^N \mid x_{i,t}^N\ , a_{i,t}^N\ , e_t^{(N)}).$$

For the agents' policies, denote the history spaces $H_0 \coloneqq \mathcal{X} \times \Delta_{\mathcal{X}}$ and

$$H_t \coloneqq H_{t-1} \times \mathcal{X} \times A \times \Delta_{\mathcal{X}}, \quad t \geq 1.$$

A policy for a generic agent is a sequence of kernels $\pi = (\pi_t)_t$ on A given H_t, and a policy is Markovian if each π_t is a kernel on A given $\mathcal{X}$. The set of policies for each agent i is given by Π_i, and the set of Markovian policies for each agent i is denoted by M_i. Let also $\mathbf{\Pi}^{(N)} = \prod_{i=1}^N \Pi_i$ and $M^{(N)} = \prod_{i=1}^N M_i$, which represent the corresponding N-tuple of policies for all the agents.

For agent i, the infinite horizon discounted cost functional with discount factor $\gamma \in (0,1)$ and policy $\boldsymbol{\pi}^{(N)} \in \boldsymbol{\Pi}^{(N)}$ is given by

$$J_i^{(N)}(\boldsymbol{\pi}^{(N)}) \coloneqq \mathbb{E}^{\boldsymbol{\pi}^{(N)}}\left[\sum_{t=0}^{\infty} \gamma^t r(x_{i,t}^N, a_{i,t}^N, e_t^{(N)})\right],$$

where the expectation $\mathbb{E}^{\boldsymbol{\pi}^{(N)}}$ is constructed analogously as in the single agent MDP. The notion of optimality is given by the Nash equilibrium, which intuitively states that at equilibrium, no player can gain further by deviating from the optimal policy.

Definition 1.11. A policy $\boldsymbol{\pi}^{(N*)}$ is a Nash equilibrium if

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) = \sup_{\pi^i \in \Pi_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i),$$

for each $i = 1, \dots, N$, where $\boldsymbol{\pi}_{-i}^{(N*)} \coloneqq (\pi^{j*})_{j \neq i}$.

In general, seeking Nash equilibria across all policies is challenging, as agents only have access to their local states and the empirical distribution. Therefore it is more reasonable to consider the concept of a Nash equilibrium in terms of Markovian strategies instead.

Definition 1.12. A policy $\boldsymbol{\pi}^{(N*)} \in M^{(N)}$ is a Markov-Nash equilibrium if

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) = \sup_{\pi^i \in M_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i),$$

for each $i = 1, \dots, N$, where $\boldsymbol{\pi}_{-i}^{(N*)} \coloneqq (\pi^{j*})_{j \neq i}$. For a given $\varepsilon > 0$, $\boldsymbol{\pi}^{(N*)} \in M^{(N)}$ is an ε-Markov-Nash equilibrium if

$$J_i^{(N)}(\boldsymbol{\pi}^{(N*)}) \geq \sup_{\pi^i \in M_i} J_i^{(N)}(\boldsymbol{\pi}_{-i}^{(N*)}, \pi^i) - \varepsilon, \quad \text{for each } i = 1, \dots, N.$$

Despite the simplification, the problem of searching Nash equilibria is still intractable and suffers from the curse of dimensionality. This leads to the notion of mean-field games, which considers the infinite population limit, such that mean-field Nash equilibria serve as approximate Nash equilibria for finite but large N-player games.

1.4 Mean-field games

Mean-field games were pioneered by the works of [45] and [19], in the setting of continuous time. The idea is to significantly simplify the analysis required for large finite N-player games by considering the infinite population limit. The empirical

distribution of the agents, in the infinite limit regime, is replaced by a measure flow. A mean-field Nash equilibrium can then be characterised by a forward equation for a generic agent's distribution, coupled with a backwards optimality equation. As the distribution of the players is replaced by a fixed measure flow in the coupled equations, the problem of finding an MFNE becomes much more tractable.

We state here the premise of the MFG in discrete-time [59]. A mean-field game is specified by a tuple $\langle \mathcal{X}, A, p, r, \mu_0 \rangle$, where $\mu_0 \in \Delta_{\mathcal{X}}$ is the initial state distribution of the agents. Assume as before that both $\mathcal{X}$ and A are finite. Let $\mathcal{M}$ be the set of measure flows on $\mathcal{X}$, and let $\boldsymbol{\mu} = (\mu_t)_t \in \mathcal{M}$ be an exogeneously given measure flow, representing the postulated law of the population. At each state, a representative generic agent has dynamics according to

$$x_{t+1} \sim p(\cdot \mid x_t, a_t, \mu_t),$$

and receives a reward $r(x_t, a_t, \mu_t)$ at each time t. For a policy $\pi \in \Pi$, the reward functional is then defined as

$$J_{\boldsymbol{\mu}}(\pi) = \mathbb{E}^{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t, \mu_t) \right].$$

A mean-field Nash equilibrium (MFNE) is then characterised by a fixed point as follows.

Definition 1.13. Defined the set-valued best-response map $\Phi : \mathcal{M} \to 2^{\Pi}$ by

$$\Phi(\boldsymbol{\mu}) = \left\{ \pi^* \in M : J_{\boldsymbol{\mu}}(\pi^*) = \sup_{\pi \in M} J_{\boldsymbol{\mu}}(\pi) \right\}.$$

Next, define the measure flow map $\Psi : \Pi \to \mathcal{M}$, where for a policy $\pi \in \Pi$, $\boldsymbol{\mu} = \Psi(\pi)$ is defined recursively by $\Psi(\pi)_0 = \mu_0$ and

$$\Psi(\pi)_{t+1}(\cdot) = \sum_{x \in \mathcal{X}} \sum_{a \in A} p(\cdot \mid x, a, \mu_t) \pi_t(a \mid x) \Psi(\pi)_t(x).$$

A pair $(\pi, \boldsymbol{\mu}) \in \Pi \times \mathcal{M}$ is a mean-field Nash equilibrium (MFNE) if $\pi \in \Phi(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \Psi(\pi)$.

Thus, searching for an MFNE amounts to first optimising for a generic agent under the fixed postulated law of the population given by the measure flow $\boldsymbol{\mu} \in \mathcal{M}$, and then ensuring that the state distribution of this agent under the optimal policy is consistent with $\boldsymbol{\mu}$. Note when $\boldsymbol{\mu}$ is fixed, the optimisation problem reduces back to a single agent MDP.

The existence of MFNE in discrete-time MFGs is shown in [59]. For finite state and action spaces, the required assumptions are as follows.

Assumption 1.14. Assume that:

– The stochastic kernel $p : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{X}}$ is weakly continuous, i.e. if $(x_n, a_n, \mu_n) \to (x, a, \mu)$, then $p(\cdot \mid x_n, a_n, \mu_n)$ converges to $p(\cdot \mid x, a, \mu)$ weakly in the sense of measures, i.e., denoting $\mathbb{E}^n$ and $\mathbb{E}$ for the expectation for the respective measures, then for all bounded and continuous functions $f : \mathcal{X} \to \mathbb{R}$:

$$\mathbb{E}^n[f] \to \mathbb{E}[f].$$

– The reward function $r : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \mathbb{R}$ is continuous.

Theorem 1.15 ([59, Theorem 3.3])**.** *Under Assumption 1.14, the MFG* $\langle \mathcal{X}, A, p, r, \mu_0 \rangle$ *admits an MFNE* $(\pi^*, \boldsymbol{\mu}^*) \in M \times \mathcal{M}$.

An MFNE serves as a valid approximation to Nash equilibria of large but finite N-player games, given by the result below:

Theorem 1.16 ([59, Theorem 4.1])**.** *Under Assumption 1.14, assume in addition that for an MFNE* $(\pi, \boldsymbol{\mu})$ *(which exists by Theorem 1.15),* π_t *is weakly continuous for each* $t \geq 0$*. Then, for any* $\varepsilon > 0$*, there exists a positive integer* $M(\varepsilon)$*, such that, for each* $N \geq M(\varepsilon)$*, the policy* $\pi^{(N)} = \{\pi, \pi, \ldots, \pi\}$ *is an* ε*-Markov–Nash equilibrium for the game with* N *agents.*

1.4.1 Entropy regularisation for MFGs

As the optimal policy for a fixed measure flow is not unique in general, $\Phi(\boldsymbol{\mu})$ is set valued in the definition of the MFNE. Therefore the proof of Theorem 1.15 utilises the Kakutani fixed-point theorem for set-valued functions, and is non-constructive. In order to compute for a fixed point, a first approach would be to define any single-valued map $\hat{\Phi}$ for an optimal policy, then attempt to apply $\hat{\Phi}$ and Ψ repeatedly in hopes of converging towards an MFNE. To guarantee such a convergence, one would have to appeal to the Banach fixed point theorem. However, it is shown in [23] that for finite MFGs, the map $\Psi \circ \hat{\Phi}$ is non-contractive in general.

Theorem 1.17 ([23, Theorem 2])**.** *If the image of* $\hat{\Phi}$ *is finite, then either the MFNE operator* $\Psi \circ \hat{\Phi}$ *is constant, or it is not Lipschitz continuous. Therefore the MFE operator does not form a contraction.*

A possible workaround is to assume that the Q-function is strongly concave with respect to a and has a Lipschitz continuous gradient in a, with respect to all other arguments [5]. This is however a very restrictive condition (the LR problem in [23, Section 3.1] is a simple counterexample that violates this condition).

Recent focus has turned towards the use of entropy regularisation to aid convergence. This has been considered in [6] for stationary measure flows, and further extended to the non-stationary case for finite horizon problems in [23]. Let $\Omega : \Delta_A \to \mathbb{R}$ be a differentiable ρ-strongly convex function, that is, for all $u, v \in \Delta_A$,

$$\Omega(u) \geq \Omega(v) + \langle \nabla\Omega(v), u - v\rangle + \frac{1}{2}\rho\|u - v\|^2, \quad \rho > 0,$$

where $\|\cdot\|$ is the 1-norm. Here Ω is considered as a regulariser and is added as an additional term into the reward functional. The duality between smoothness and convexity can be exploited to achieve the desired contractiveness. The presence of the regulariser leads to higher entropy policies across iterations, which encourages policy exploration. This also overcomes oscillation issues that arises from the hard maximisation during each optimisation step.

In the case of [23], for a regularised MFG, we consider the regularised value function

$$J^*_{\eta,\boldsymbol{\mu}}(t, x) = \sup_{\pi\in\Pi} \sum_{n=t}^{T} \Big(\mathbb{E}^{\pi}[r(x_n, a_n, \mu_n)] - \eta\ \Omega(\pi_n)\Big),$$

where η is the regularisation parameter, also referred to as the temperature. Define also the associated optimal Q-function, given by

$$Q^*_{\eta,\boldsymbol{\mu}}(t, x, a) = r(x, a, \mu_t) + \sum_{x'\in\mathcal{X}} J^*_{\eta,\boldsymbol{\mu}}(t + 1, x')p(x' \mid x, a, \mu_t).$$

When the regulariser is of the from of the KL divergence with respect to some policy $q \in \Delta_A$,

$$\Omega(\pi) = \sum_{a\in A} \pi(a) \ln \frac{\pi(a)}{q(a)},$$

then the maximising policy is the softmax policy, given by

$$\pi^{\text{soft}}_t(a \mid x) = \frac{q(a \mid x) \exp(Q^{\text{reg},*}_{\eta,\boldsymbol{\mu}}(t, x, a)/\eta)}{\sum_{a'\in A} q(a \mid x) \exp(Q^{\text{reg},*}_{\eta,\boldsymbol{\nu}}(t, x, a')/\eta)},$$

where $Q^{\text{reg},*}$ is the corresponding Q-function. When q is the uniform distribution, Ω reduces to the negative entropy. Such classes of regularisers leads to the following contraction theorem for regularised MFGs.

Theorem 1.18 ([23, Theorem 3]). *Assume that both the transition kernel p and one-step reward function r are Lipschitz continuous. Let T be the finite time horizon, $\boldsymbol{\mu} = (\mu_t)_t \in \Delta_{\mathcal{X}}^T$ and $\eta > 0$. Let $q = (q_t)_t$ be a reference policy. Define the following maps:*

(i) The best-response map $\Phi_\eta^{\mathrm{reg}} : \Delta_{\mathcal{X}}^T \to \Pi$, given by

$$\Phi_\eta^{\mathrm{reg}}(\boldsymbol{\mu})_t(a \mid x) = \frac{q_t(a \mid x) \exp(Q_{\eta,\boldsymbol{\mu}}^{\mathrm{reg},*}(t,x,a)/\eta)}{\sum_{a' \in A} q_t(a \mid x) \exp(Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t,x,a')/\eta)}.$$

(ii) The measure flow map $\Psi^{\mathrm{aug}} : \Pi \to \Delta_{\mathcal{X}}^T$, where $\Psi^{\mathrm{aug}}(\pi)_0 = \mu_0$ and for $t \geq 0$,

$$\Psi^{\mathrm{aug}}(\pi)_{t+1}(\cdot) = \sum_{x \in \mathcal{X}} \sum_{a \in A} p\Big(\cdot \mid x, a, \Psi^{\mathrm{aug}}(\pi)\Big) \pi_t(a \mid x) \Psi^{\mathrm{aug}}(\pi)_t(x).$$

Then for sufficiently large η, there exists a unique fixed point $\boldsymbol{\mu}^*$ for the map $\Psi^{\mathrm{aug}} \circ \Phi_\eta^{\mathrm{reg}}$. In particular, $(\pi^*, \boldsymbol{\mu})$ is the regularised MFNE, given by $\boldsymbol{\mu}^*$ of $\Psi^{\mathrm{aug}} \circ \Phi_\eta^{\mathrm{reg}}$, for which $\pi^* = \Phi_\eta^{\mathrm{reg}}(\boldsymbol{\nu}^*)$ (best response map) and $\boldsymbol{\mu}^* = \Psi^{\mathrm{aug}}(\pi^*)$ (measure flow induced by policy) holds.

The selection of an optimal η that allows a fixed point contraction, as well as achieving a good approximation towards the MFNE remains an open question, with various heuristics proposed to dynamically change η during the algorithm [23]. We note also that there are alternatives for the computation of algorithms towards the MFNE, including fictitious play, online mirror descent [53] and reformulation to constrained optimisation (MF-OMO) [32]. We refer to the survey [46] for a comprehensive review of all the related methods.

Chapter 2

Markov decision processes with observation costs: framework and penalty scheme

2.1 Introduction

In this chapter, we examine the observation cost model (OCM) for Markov Decision Processes (MDPs). A cost must be paid in order to observe the state of the underlying MDP, and only then can adjustments be made to the action which influences the dynamics of the MDP. We propose a penalty scheme for efficient numerical computation for the resulting system of equations.

MDPs are mathematical tools that model the optimisation of a random process, in order to maximise the expected profit over time. Applications are common in maintenance, portfolio optimisation, sensor detection, reinforcement learning and more. Most setups implicitly assume a fixed source of information upon which the user relies to select an optimal action. However, such a steady stream of information might not be available in situations where resources are constrained, either by the expensive cost of measurements, or by the impracticality of frequent sampling. This calls for an extra layer of optimisation, where the user has to decide on the optimal observation times of the information source, as well as the optimal sequence of actions to maximise the expected returns.

The literature involving observation control appear across several different fields, and appear under terms such as 'optimal inspections', 'costly observations' or 'controllable observations'. To the best of our knowledge, the earliest works appear in [43,47], which

concerns the linear quadratic Gaussian (LQG) problem over a finite horizon with fixed number of measurements, as well as the papers [7,8], which examines a costly optimal stopping problem in continuous time. Numerous applications have emerged in the literature over the years, which we list (non-exhaustively) below, broadly categorising into the following areas:

- environmental management control models [72–75],
- optimal sampling rates in communications [29, 33],
- optimal sensing problems [49, 64, 70],
- medical treatment cycles [66, 67],
- detection of drift in Brownian motion [13, 14, 24, 25],
- empirical works in reinforcement learning [16, 17, 42].

The standard approach is to formulate the problem in terms of a partially observable Markov decision process (POMDP). Dynamic programming for the value function leads to a search for the optimal observation time after the currently observed state, as well as the optimal action sequence between the observation times. The formulation of the OCM in such generality, however, suffers from the curse of dimensionality: as the time between observations is unbounded, the number of actions to be optimised also grows unbounded. Indeed, non-constant controls between observations are mostly only treated under the LQG framework [22,64,70]. We will therefore restrict ourselves in this chapter to consider only *constant* actions between observations. Such an assumption applies to models where actions cannot be feasibly changed without an accompanying observation, such as the medical treatment applications in [66, 67] or the environmental management control models [72–75].

Due to the nature of the action being fixed upon an observation until the next, the passage of time has a lingering effect on the optimal control. For example, it might be optimal to diagnose and repair certain machinery when performance is subpar, but it might be more favourable to directly purchase new equipment with new technology if said piece of machinery was left unfixed and unobserved for a prolonged period of time. When considering the belief MDP for the OCM, we show that the amount of time elapsed since the last observation becomes a part of the augmented Markov system. To our knowledge, only the works of [7, 8] and [33] model the OCM in this specific formulation, but the problems considered were restricted to fixed dynamics for the underlying Markov chain.

As in the other formulations, the OCM with constant actions between observations can also be represented by a non-standard version of a POMDP. We assume that the Markov chain takes values in a finite state space $\mathcal{X}$ and that its dynamics are known and are given by the transition matrices $\{P_a\}_{a\in A}$, where A is a finite action set. We also assume that the actions can only be adjusted at the observation times. The one-step reward function is given by $r(x,a) = r_{a,x}$ and the observation cost is given by a constant $c_{\text{obs}} > 0$. The inclusion of time elapsed as a variable in the Markov system leads to a system of discrete quasi-variational inequalities (QVI), which for the discounted infinite horizon problem, reads:

$$\min\left\{v^n_{a,x} - \gamma v^{n+1}_{a,x} - \left(P^n_a \; r_a\right)_x \,, \right.$$
$$\left. v^n_{a,x} - \left(P^n_a \; \overline{\gamma v^1 + r}\right)_x + c_{\text{obs}}\right\} = 0, \tag{2.1}$$

where v is the value function, indexed by: $x \in \mathcal{X}$, the state of the chain at the previous observation; $n \in \mathbb{N}_{\geq 1}$, the time elapsed since the previous observation; and $a \in A$, the action applied at the previous observation. The vector $\left(\overline{\gamma v^1 + r}\right)_x = \max_{a\in A}(\gamma v^1_{a,x} + r_{a,x})$ represents the 'inner loop' optimisation over the space of actions after an observation is made.

As seen above, the inclusion of the variable of time elapsed n leads to a structurally different set of optimality equations, in the form of a quasi-variational inequality (QVI). More generally, we consider the following class of QVIs below, which includes the specific form of (2.1).

Problem 2.1. Find $u = (u_1, \ldots, u_d) \in \mathbb{R}^{N\times L\times d}$ such that

$$\min\{F_a(u), u_a - \mathcal{M}u\} = 0, \quad a \in \{1,\ldots,d\} =: A, \tag{2.2}$$

where

- $\mathcal{M} : \mathbb{R}^{N\times L\times d} \to \mathbb{R}^{N\times L}$ is defined by

$$(\mathcal{M}u)^n_l = \left(Q_n\overline{u^1 - c}\right)_l, \quad \left(\overline{u^1 - c}\right)_l = \max_{a\in A}\left(u^1_{a,l} - c_{a,l}\right), \tag{2.3}$$

  for a given vector $c \in \mathbb{R}^{L\times d}$ and $\{Q_n\} \subset \mathbb{R}^{L\times L}$ is a sequence of strictly substochastic matrices;

- $F_a : \mathbb{R}^{N\times L} \to \mathbb{R}^{N\times L}$ is a continuous function that satisfies the following property: there exists a constant $\beta > 0$ such that for any $u, v \in \mathbb{R}^{N\times L\times d}$ with $u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}} = \max_{n,a,l}(u^n_{a,l} - v^n_{a,l}) \geq 0$, we have

$$F_{\bar{a}}(u)^{\bar{n}}_{\bar{l}} - F_{\bar{a}}(v)^{\bar{n}}_{\bar{l}} \geq \beta(u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}}). \tag{2.4}$$

The QVI (2.2) is a generalisation of a monotone system with interconnected obstacles [56], which can arise from the discretisation of optimal switching problems in continuous time. In our case, we shall refer to the operator $\mathcal{M}$ as the inspection operator. Much like the systems with interconnected obstacles, the QVI (2.2) is typically not amenable to the use of policy iteration, as the matrices arising from the inspection operator do not necessarily satisfy the M-matrix or weakly chain diagonally dominant conditions [9]. We propose instead a penalty scheme, which sees use on variational inequalities [26,36,38] and extensions to HJB VIs [57,68,69]/ QVIs [56], as an approximation. Penalty schemes have seen comparable computational performance to policy iteration in HJB QVIs , and is robust to the choice of initial estimates [68, 69]. An adaptation of the penalty scheme to the QVI (2.2) circumvents issues with numerical instabilities arising from computing iterates of the policy update, and the penalised equation can be solved with semismooth Newton methods.

Finally, we note a closely related work to ours in [37], which uses the term 'self-triggered MDPs' to refer to the OCM with constant action between observations. There, however, the time elapsed variable is not considered as part of the Markov system. Here we also provide the penalty scheme as a viable alternative to the value iteration scheme given by the authors in [37]. We demonstrate in Section 2.4 that the penalty method achieves quick convergence within a few iterations on a large system whilst also mapping out accurately the optimal policy.

The main contributions of this chapter are as follows:

- We formulate the observation cost model (OCM) for Markov decision processes where the time elapsed after an observation is considered as part of the augmented Markov system. We present the optimality equations obtained from dynamic programming for the finite horizon problem, discounted infinite horizon problem, and the respective problems with parameter uncertainty. In all cases the optimality equations are in the form of a QVI, which are structurally different to the Bellman-type equations from existing approaches in the literature.
- We establish a comparison principle for the class of QVIs (2.2), of which the solution to the OCM belongs to. The class of QVIs are a generalisation of monotone systems with interconnected obstacles as seen in [56]. We propose a penalty scheme for this class of QVIs (2.2), and demonstrate the monotone

convergence of the penalised solutions towards the solutions of said QVI, thereby establishing constructively the existence of solutions.

– We demonstrate the numerical performance of our model by applying it to the time-discretised version of the HIV-treatment model [67]. Our framework is compatible with the original results, and also shows qualitatively different optimal behaviour when dealing with large observation gaps.

The remainder of this chapter is organised as follows. Section 2.2 sets out the framework for the OCM and establishes the corresponding set of discrete QVIs. A model problem with an explicit solution is also provided to illustrate the setup. We also outline the case of parameter uncertainty in Section 2.2.3. In Section 2.3 we prove a comparison principle for a class of discrete QVIs which subsumes the QVIs obtained in Section 2.2, as well as outlining the penalty method as a numerical scheme for the QVI. Finally the numerical experiments are presented in detail in Section 2.4.

2.1.1 Notation for MDPs and POMDPs

As the goal of the next section is to layout the OCM precisely by formulating the model in terms of a POMDP, we shall quote here some of the standard notation for MDPs, POMDPs and a brief overview of its construction. These are largely taken from [34, Ch 4] and we refer the reader to the references within for further detail.

We will generally be considering Markov decision processes on finite state spaces. Whilst most arguments extend naturally to more general state spaces, we shall focus on the finite setting here to retain a simplified presentation. Let $\mathcal{P}(\mathcal{X})$ denote the space of probability measures over a set $\mathcal{X}$. If $\mathcal{X}$ is finite, we will also identify $\mathcal{P}(\mathcal{X})$ with the simplex $\Delta_{\mathcal{X}}$.

Definition 2.2. A Markov control model is a tuple $\langle \mathcal{X}, A, p, r \rangle$, where

– $\mathcal{X}$ is the finite *underlying state space*;

– A is the finite *action space*;

– $p : \mathcal{X} \times A \to \Delta_{\mathcal{X}}$ is the *transition kernel*;

– $r : \mathcal{X} \times A \to \mathbb{R}$ is the *one-step reward function*.

At each time t, a state $x_t \in \mathcal{X}$ is observed. The controller chooses an action $a_t \in A$ and receives a reward $r(x_t, a_t)$. The system then moves to a new state $x_{t+1} \in \mathcal{X}$ with probability $p(x_{t+1}|x_t, a_t)$ and the process repeats at time $t + 1$. Actions are chosen

according to a *policy* $\pi = (\pi_t)_t$, which is a sequence of kernels $\pi_t : H_t \to A$, where $H_0 \coloneqq \mathcal{X}$ and $H_t \coloneqq (\mathcal{X} \times A)^t \times \mathcal{X}$ for $t \geq 1$, known as the *history set* at time t. The set of all policies is denoted by Π. Given an initial state $x_0 \in \mathcal{X}$ and policy $\pi \in \Pi$, by the Ionescu-Tulcea theorem (see [34, Appendix C]), there exists a unique probability measure $\mathbb{P}^\pi_{x_0}$ on the canonical sample space $\Omega \coloneqq H_\infty \coloneqq (\mathcal{X} \times A)^\infty$, such that given $\omega = (x_0, a_0, x_1, a_1, \ldots) \in \Omega$,

$$\mathbb{P}^\pi_{x_0}(\omega) = \delta(x_0)\ \pi_0(a_0 \mid x_0)\ p(x_1 \mid x_0, a_0)\ \pi(a_1 \mid x_0, a_0, x_1) \ldots$$

The objective is to maximise an objective function over the set of policies Π, for example, in the finite horizon case,

$$J(\pi, x_0) = \mathbb{E}^\pi_{x_0}\left[\sum_{n=0}^{N} r(x_n, a_n)\right], \quad \pi \in \Pi,\ x_0 \in \mathcal{X},$$

where $N \in \mathbb{N}$ is the time horizon, and $\mathbb{E}^\pi_{x_0}$ is the expectation under the measure $\mathbb{P}^\pi_{x_0}$. The value function is given by

$$v(x) = \sup_{\pi \in \Pi} J(\pi, x), \quad x \in \mathcal{X}.$$

It is well known that an optimal policy π^* for an MDP is deterministic and Markovian, i.e. there exists deterministic functions $\{\phi_t\}_{t \geq 0}$ such that for $h_t = (x_0, a_0, \ldots, x_t) \in H_t$, $\pi^*_t(h_t) = \phi_t(x_t)$, and $v(x) = J(\pi^*, x)$.

In many cases, rather than having full information of the MDP, one instead has access to noisy observations correlated to the underlying MDP. This gives rise to the notion of partially observable Markov decision processes (POMDPs), which can also be described by a given tuple as follows.

Definition 2.3. A partially observable control model is a tuple $\langle \mathcal{X}, \mathcal{O}, A, p, p_0, q, q_0, r \rangle$, where

- $\mathcal{X}$ is the finite *state space*;
- $\mathcal{O}$ is the finite *observation space*;
- A is the finite *action space*;
- $p : \mathcal{X} \times A \to \Delta_{\mathcal{X}}$ is the *transition kernel*;
- $p_0 \in \Delta_{\mathcal{X}}$ is the *initial distribution*;
- $q : A \times \mathcal{X} \to \Delta_{\mathcal{O}}$ is the *observation kernel*;

– $q_0 : \mathcal{X} \to \Delta_{\mathcal{O}}$ is the *initial observation kernel*;

– $r : \mathcal{X} \times A \to [0, \infty)$ is the *one-step reward function.*

In this setting, given an underlying state $x_t \in \mathcal{X}$, an observation $\bar{x}_t \in \mathcal{O}$ is generated according to the observation kernel $q(\bar{x}_t | a_{t-1}, x_t)$. The controller chooses an action $a_t \in A$ based on their observations, rather than the values of the underlying states. For this, define the *observable history sets*

$$\mathcal{H}_0 \coloneqq \mathcal{O}, \ \mathcal{H}_t \coloneqq (\mathcal{O} \times A)^t \times \mathcal{O}, \quad t \geq 1.$$

A policy for a POMDP is now a sequence of kernels $\pi_t : \mathcal{H}_t \to A$. Denote the set of policies for the POMDP as Π_{po}. By the Ionescu-Tulcea theorem again, given an initial distribution $p_0 \in \Delta_{\mathcal{X}}$ and policy $\pi \in \Pi_{\text{po}}$, there exists a unique probability measure $\mathbb{P}^{\pi}_{p_0}$ on the canonical space $\Omega = (\mathcal{X} \times \mathcal{O} \times A)^{\infty}$ such that for $\omega = (x_0, \bar{x}_0, a_0, x_1, \bar{x}_1, a_1, \ldots) \in \Omega$,

$$\mathbb{P}^{\pi}_{p_0}(\omega) = p_0(x_0) \ q_0(\bar{x}_0 \mid x_0) \ \pi_0(a_0 \mid \bar{x}_0) \ p(x_1 \mid x_0, a_0) \ q(\bar{x}_1 \mid x_1, a_0) \ \pi(x_1 \mid \bar{x}_0, a_0, \bar{x}_1) \ldots$$

The maximisation is now performed over $\pi \in \Pi_{\text{po}}$, with the objective and value function

$$J(\pi, p_0) = \mathbb{E}^{\pi}_{p_0}\left[\sum_{n=0}^{N} r(x_n, a_n)\right], \ v(p) = \sup_{\pi \in \Pi_{\text{po}}} J(\pi, p), \quad \pi \in \Pi_{\text{po}}, \ p, p_0 \in \Delta_{\mathcal{X}}.$$

Without knowledge of the underlying states, the POMDP is a non-Markovian problem. The standard approach to solve a POMDP is to consider an equivalent (fully observable) problem, known as the *belief MDP*, on the space $Z \coloneqq \Delta_{\mathcal{X}}$. The Markovian strucutre is recovered when lifted to the belief MDP, so that classical dynamic programming techniques can be applied. The transition kernel for the *belief state* $z = (z_t)_t$ can be constructed as follows: given the POMDP, construct a kernel $R : Z \times A \to \mathcal{X} \times \mathcal{O}$ such that

$$R(x, \bar{x} \mid z, a) = q(\bar{x} \mid a, x')p(x' \mid x, a)z(x), \quad (x, \bar{x}) \in \mathcal{X} \times \mathcal{O}, \ (z, a) \in Z \times A.$$

It can be shown that there exists a kernel $H' : Z \times A \to \mathcal{X}$, such that R can be disintegrated into

$$R(x, \bar{x} \mid z, a) = H'(x \mid z, a, \bar{x})R'(\bar{x} \mid z, a), \quad (x, \bar{x}) \in \mathcal{X} \times \mathcal{O}, \ (z, a) \in Z \times A.$$

where R' is the marginal of R on $\mathcal{O}$. Then, letting $H_{z,a,\bar{x}} = H'(\cdot \mid z, a, \bar{x}) \in Z$, define the kernel $q' : Z \times A \to Z$ by

$$q'(z' \mid z, a) = \sum_{\bar{x} \in \mathcal{O}} \delta_{H_{z,a,\bar{x}}}(z') R'(\bar{x} \mid z, a), \quad z, z' \in Z,\ a \in A.$$

Then one takes q' as the transition kernel for the belief state, and construct the initial kernel q'_0 analogously. The belief MDP is then $\langle Z, A, q', q'_0, r_z \rangle$, where $r_z : Z \times A \to \mathbb{R}$ as $r_z(z, a) = \sum_{x \in \mathcal{X}} r(x, a) z(x)$. The belief state can be interpreted as the conditional distribution of the underlying state x_t, given the observed history $(\bar{x}_0, a_0, \ldots, \bar{x}_t)$. Let Π^z be the set of policies for the belief MDP, which are now a sequence of kernels on A, given the history sets $\mathcal{H}_t^z = (Z \times A)^t \times Z$. The objective and value function for the belief MDP are given by

$$J_z(\pi^z, z_0) \coloneqq \mathbb{E}_{z_0}^{\pi^z} \left[\sum_{n=0}^{N} r_z(z_n, a_n) \right],\ v_z(z) = \sup_{\pi^z \in \Pi^z} J_z(\pi^z, z), \quad \pi^z \in \Pi^z,\ z, z_0 \in Z,$$

where $\mathbb{E}_{z_0}^{\pi^z}$ is the expectation over the canonical space $(Z \times A)^\infty$ under the policy π^z and initial condition $z_0 \in Z$. It can then be shown that policies in Π_{po} are equivalent to policies in Π^z, in the sense that any $h_t \in \mathcal{H}_t$ can be mapped to a corresponding $h_t^z \in \mathcal{H}_t^z$, so that given $\pi^z \in \Pi^z$, one can construct a corresponding $\pi \in \Pi_{\text{po}}$ via

$$\pi(\cdot \mid h_t) \coloneqq \pi^z(\cdot \mid h_t^z),$$

and the conditional probabilities assigned on the action set A are the same. Moreover, the POMDP $\langle \mathcal{X}, \mathcal{O}, A, p, p_0, q, q_0, r \rangle$ and the belief MDP $\langle Z, A, q', q'_0, r_z \rangle$ are equivalent [34, Ch 4]: $\pi^* \in \Pi_{\text{po}}$ is optimal for J if and only if $\pi^{*,z} \in \Pi^z$ is optimal for J_z, and it holds that

$$v_z(z) = J_z(\pi^{*,z}, z) = J(\pi^*, z) = v(z), \quad z \in Z = \Delta_{\mathcal{X}}.$$

Thus, when considering a POMDP, it is sufficient to consider its equivalent MDP in the belief state, of which the optimal policy is Markovian with respect to $z = (z_t)_t$.

2.2 Problem formulation

In the Observation Cost Model, the process evolves sequentially as follows. At each time t, the controller decides if they would like to pay *an observation cost* $c_{\text{obs}} > 0$ to observe the state $x_t \in \mathcal{X}$. If they decide to do so, then the controller applies the action $a_t \in A$ according to some suitable policy π, and receives a reward $r(x_t, a_t)$.

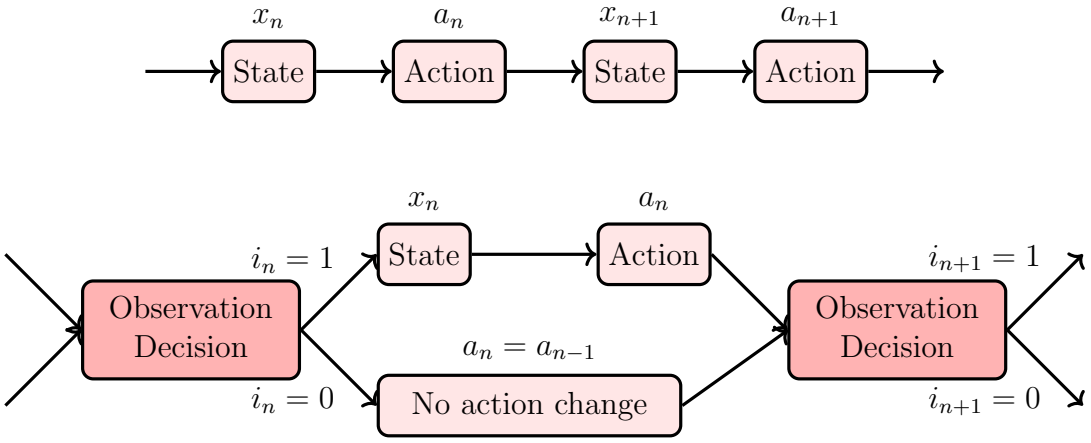

Figure 2.1: *Top: Standard formulation with full observations; Bottom: The inclusion of observation costs leads to an extra decision step.*

If they decide not to observe, then no cost is paid, but the controller cannot change the action value, so that $a_t = a_{t-1}$. We assume that the reward $r(x_t, a_t)$ is collected and 'locked in' at time t, but is not observable to the user if x_t is not observable. In both instances, the system moves to a new state $x_{t+1} \in \mathcal{X}$ according to the transition kernel $p(x_{t+1} \mid x_t, a_t)$.

We now formally write down the objective function, and shall make precise the terms appearing within in the rest of this section. In view of the description above, the controller wishes to maximise (for example, in the finite horizon case)

$$\mathbb{E}^{\pi}\left[\sum_{n=0}^{N}\left(r(x_n, a_n) - i_n \cdot c_{\text{obs}}\right)\right], \tag{2.5}$$

where π a suitably admissible control policy to be made precise later. The sequence $i = (i_n)_n$ takes values in $\mathcal{I} := \{0, 1\}$ and will be referred to as the *inspection values*. A value of $i_n = 1$ represents an observation made at time n, so that observation cost c_{obs} is deducted from the total reward in (2.5). Conversely no observations are made if $i_n = 0$. Figure 2.1 illustrates the sequential flow of a standard MDP, compared to that of the OCM.

We now proceed to establish the OCM as a non-standard form of a POMDP, in order to fully make sense of (2.5). A policy should output an action value $a_n \in A$, as well as an inspection value $i_n \in \mathcal{I}$. The observation space is represented by $\mathcal{X} \cup \{\varnothing\}$: either the underlying chain with values in $\mathcal{X}$ is observed, or the dummy variable $\varnothing$ nothing is observed, which represents the case of no observations. A final but subtle point is the difference in the sequential structure of the OCM compared to a POMDP. In

the case of a POMDP, first a state x_n is generated, follow by the observation $\bar{x}_n$, and then the action a_n. Thus a realisation on the canonical space of a POMDP looks like:

$$(x_0, \bar{x}_0, a_0, x_1, \bar{x}_1, a_1, \ldots). \tag{2.6}$$

In the OCM, as depicted in Figure 2.1, the observation occurs *after* the inspection value, after which the action value follows. Thus a realisation of the system of an OCM will instead look like

$$(x_0, i_0, \bar{x}_0, a_0, x_1, i_1, \bar{x}_1, a_1, \ldots). \tag{2.7}$$

In order to obtain the sequential structure of 'state - observation - action' for the OCM, we augment the sequence (2.7) with fictitious state and observation values, and treat both a_n and i_n as an 'action'. By suitably augmenting the transition and observation kernels, the OCM takes the form of a POMDP, over the timescale of $\frac{1}{2}\mathbb{N}$. This augmented sequence then takes the form of

$$\begin{aligned} &(x_0, \bar{x}_0, \pi_0, x_{1/2}, \bar{x}_{1/2}, \pi_{1/2}, x_1, \bar{x}_1, \pi_1, \ldots) \\ \coloneqq\ &(x_0, \bar{x}_0, i_0, x_0, \bar{x}_0, a_0, x_1, \bar{x}_1, i_1, \ldots). \end{aligned} \tag{2.8}$$

This leads us to the following definition.

Definition 2.4. Given an MDP $\langle \mathcal{X}, A, p, r \rangle$, the associated observation cost model (OCM) is defined as the POMDP $\langle \mathcal{X}, \mathcal{X}_\varnothing, \mathcal{A}, p, p_0, q, q_0, r_{\text{obs}} \rangle$ (see Definition 2.3), on the time scale $\frac{1}{2}\mathbb{N} = \{0, 1/2, 1, \ldots\}$, where

- $\mathcal{X}_\varnothing = \mathcal{X} \cup \{\varnothing\}$ is the *observation space*, with $\varnothing$ a dummy variable representing no observations.
- $\mathcal{A}$ is the *disjoint* union of A and $\mathcal{I}$, with time dependent admissible sets given by $\mathcal{A}(n) = \mathcal{I}$ and $\mathcal{A}(n + \frac{1}{2}) = A$;
- $p : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$ is the transition kernel, with its domain extended to $\mathcal{X} \times \mathcal{A}$ by defining

$$p(\cdot \mid x, i) = \delta_x(\cdot), \quad i \in \mathcal{I}.$$

- $q : \mathcal{A} \times \mathcal{X} \to \Delta_{\mathcal{X}_\varnothing}$ is the *observation kernel*, given by

$$\begin{aligned} q(\cdot \mid a, x) &= \delta_\varnothing(\cdot), \quad a \in A, \\ q(\cdot \mid i, x) &= i \cdot \delta_x(\cdot) + (1 - i)\delta_\varnothing(\cdot), \quad i \in \mathcal{I}. \end{aligned}$$

– $r_{\text{obs}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the *one-step reward function* given by

$$r_{\text{obs}}(x, a) = r(x, a), \quad a \in A,$$
$$r_{\text{obs}}(x, i) = -i \cdot c_{\text{obs}}, \quad i \in \mathcal{I}.$$

The kernels p and q above are defined such that transitions of the underlying chain only occurs at the integer steps, and new observations at the half steps. Let us write the observable history sets here as

$$\mathcal{H}_0 = \mathcal{X}_\varnothing, \ \mathcal{H}_t = (\mathcal{X}_\varnothing \times \mathcal{A})^{2t} \times \mathcal{X}_\varnothing, \quad t \in \frac{1}{2}\mathbb{N}.$$

To be precise, we should only consider *admissible* history sets, that is state-action pairs that satisfy the constraints $\mathcal{A}(n) = \mathcal{I}$ and $\mathcal{A}(n + \frac{1}{2}) = A$, but we shall take this assumption as implicit for ease of notation. A policy π is hence a sequence of kernels $\pi_t : \mathcal{H}_t \to \mathcal{A}$.

As before, a policy π as defined above and an initial distribution $p_0 \in \Delta_\mathcal{X}$ induces a unique measure $\mathbb{P}^\pi_{p_0}$ on the canonical sample space $\Omega = (\mathcal{X} \times \mathcal{X}_\varnothing \times \mathcal{A})^\infty$. With slight abuse of notation, we also write π for the values applied by the policy. Then, the action values $a = (a_n)_n$ and inspection values $i = (i_n)_n$ appearing in (2.5) can be recovered by defining

$$i_n = \pi_n, \quad a_n = \pi_{n+\frac{1}{2}}.$$

As we are assuming in the OCM that actions remain constant between new observations. We will have to consider a smaller class of admissible policies.

Definition 2.5. An **admissible policy** $\pi = (\pi_t)_t$ for the OCM is a sequence of kernels $\pi_t : \mathcal{H}_t \to \Delta_\mathcal{A}$ which satisfies the following: for $n \in \mathbb{N}$, if

$$h_{n+\frac{1}{2}} = (\bar{x}_0, \pi_0, \ldots, \bar{x}_{n-\frac{1}{2}}, \pi_{n-\frac{1}{2}}, \bar{x}_n, \pi_n, \bar{x}_{n+\frac{1}{2}}),$$

with $\pi_n = i_n = 0$, then

$$\pi_{n+\frac{1}{2}}\left(\cdot \,\middle|\, h_{n+\frac{1}{2}} \right) = \delta_{\pi_{n-\frac{1}{2}}} = \delta_{a_{n-1}}$$

The set of admissible policies for the observation cost model is denoted Π_{obs}.

With the above setup, we can give a full meaning to the expression (2.5) by writing for $\pi \in \Pi_{\text{obs}}$ and $p_0 \in \Delta_\mathcal{X}$,

$$J(\pi, p_0) \coloneqq \mathbb{E}^\pi_{p_0}\left[\sum_{n=0}^{2N} r_{\text{obs}}(x_{n/2}, \pi_{n/2})\right] = \mathbb{E}^\pi_{p_0}\left[\sum_{n=0}^{N} \left(r(x_n, a_n) - i_n \cdot c_{\text{obs}}\right)\right].$$

Remark 2.6. ***Regarding the initial distribution*** p_0 ***and observation kernel*** q_0***:*** *for the purposes of this chapter, we will assume that an observation (made at some previous time) is always available. This allows a consistent characterisation of the belief state by a finite tuple, which leads to a system of finite-dimensional QVIs in Sections 2.2.1 and 2.2.2. Thus, we will only consider initial kernels* p_0 *that is in the form of some* n*-step transition probabilities of the kernel* p*, and the initial observation kernel* q_0 *will be taken as the Dirac measure* $q_0(\cdot \mid x) = \delta_x(\cdot)$*.*

Remark 2.7. *The formulation with the half time-steps and the inclusion of fictitious state/ observation variables above are strictly a theoretical construct, such that the OCM can be reframed as a POMDP, and thus allowing us to directly appeal to standard results to formulate dynamic programming. It will be shown later that the half steps become redundant once dynamic programming is established, and the value function will only need to be considered over the integer steps.*

2.2.1 Finite horizon problem

For the finite horizon problem, let $N \in \mathbb{N}$ be the time horizon. As we have introduced the POMDP for the OCM to be defined on the timescale $\frac{1}{2}\mathbb{N}$, we now have a total of $2N$ 'half steps', where at each step t, either a value of i_t or a_t is applied. Therefore, for a policy $\pi \in \Pi_{\text{obs}}$, we write the objective function as

$$\mathbb{E}^{\pi}_{p_0}\left[\sum_{n=0}^{2N} r_{\text{obs}}(x_{n/2}, \pi_{n/2})\right] = \mathbb{E}^{\pi}_{p_0}\left[\sum_{n=0}^{N} \left(r(x_n, a_n) - i_n \cdot c_{\text{obs}}\right)\right].$$

With the OCM problem characterised as a POMDP, we consider the belief MDP, with the belief state given by $\mathbb{P}^{\pi}_{p_0}(x_t \mid h_t)$, where $h_t \in \mathcal{H}_t$. By the Markov property, the belief state is fully characterised by the controller's most recently observed information, in that

$$\mathbb{P}^{\pi}_{p_0}(x_t \mid h_t) = p^{(\lfloor t \rfloor - k)}(x_t \mid x_k, a_k), \tag{2.9}$$

where $p^{(\lfloor t \rfloor - k)}$ is the $(\lfloor t \rfloor - k)$-step transition kernel of the underlying process X, and $k \in \mathbb{N}$ is the last occurrence such that $i_k = 1$. Note that $k = t$ implies an immediate observation, and the belief state reduces trivially to a Dirac measure at x_t. The belief state therefore has a finite dimensional parametrisation, and we can consider instead the **augmented state** $y = (y_t)_{t \in \frac{1}{2}\mathbb{N}}$:

$$y_t \coloneqq \begin{cases} (k, x_k, a_k), & \text{if } \mathbb{N} \ni k = \arg\max_{n \le t}\{i_n = 1\} \neq t\ , \\ (t, x_t, \varnothing), & \text{otherwise,} \end{cases} \tag{2.10}$$

where $\varnothing$ also acts as a dummy variable here. The three components of y_t represent the most recent observation time k, with the correspondingly observed state $x_k \in \mathcal{X}$, and applied action $a_k \in A$.

Given this equivalence in representation, we can consider the OCM as an MDP with the tuple $\langle \mathcal{Y}, \mathcal{A}, p_y, r_y \rangle$ on the timescale $\frac{1}{2}\mathbb{N}$, where

- $\mathcal{Y} \coloneqq \mathbb{N} \times \mathcal{X} \times A_\varnothing$ is the *augmented state space*;
- $\mathcal{A}$ is the *disjoint* union of A and $\mathcal{I}$, with admissible sets $\mathcal{A}(n) = \mathcal{I}$ and $\mathcal{A}(n+\frac{1}{2}) = A$;
- $p_y = (p_{y,t})_{t \in \frac{1}{2}\mathbb{N}}$ is a (time-inhomogeneous) transition kernel on $\mathcal{Y}$ given $\mathcal{Y} \times \mathcal{A}$: for $n \in \mathbb{N}$, $y = (k, x, a)$, $\hat{y} \in \mathcal{Y}$, $i \in \mathcal{I}$, and $a' \in A$,

$$p_{y,n}(\hat{y} \mid y, i) = i \cdot p^{(n-k)}(\hat{x} \mid x, a)\mathbb{1}_{\{\hat{y}=(n,\hat{x},\varnothing)\}} + (1-i)\mathbb{1}_{\{\hat{y}=y\}}, \tag{2.11}$$

$$p_{y,n+\frac{1}{2}}(\hat{y} \mid y, a') = \mathbb{1}_{\{\hat{y}=(k,x,a'),\ a'=a\}}, \tag{2.12}$$

  and for $y = (n, x, \varnothing)$,

$$p_{y,n+\frac{1}{2}}(\hat{y} \mid y, a') = \mathbb{1}_{\{\hat{y}=(n,x,a')\}}. \tag{2.13}$$

- $r_y = (r_{y,t})_{t \in \frac{1}{2}\mathbb{N}}$ is a time-dependent one-step reward function on $\mathcal{Y} \times \mathcal{A}$: for $y = (k, x, a)$, $i \in \mathcal{I}$, $a' \in A$,

$$r_{y,n}(y, i) = -i \cdot c_{\text{obs}}, \tag{2.14}$$

$$r_{y,n+\frac{1}{2}}(y, a') = \sum_{x' \in \mathcal{X}} r(x', a')p^{(n-k)}(x' \mid x, a), \tag{2.15}$$

  and for $y = (n, x, \varnothing)$,

$$r_{y,n+\frac{1}{2}}(y, a') = r(x, a'). \tag{2.16}$$

In this augmented problem, define its observable history sets as $\mathcal{H}_t^y = (\mathcal{Y} \times \mathcal{A})^{2t} \times \mathcal{Y}$, $t \in \frac{1}{2}\mathbb{N}$. Policies π^y are then a sequence of kernels $\pi_t^y : \mathcal{Y} \to \Delta_{\mathcal{A}}$. For the set of admissible policies of this augmented MDP, we will have to consider the corresponding 'image' of Π_{obs}. This in turn, is equivalent to imposing a state constraint on the admissible action sets:

$$\mathcal{A}(n + 1/2, (k, x, a)) = \{a\}, \quad n \in \mathbb{N},\ a \in A.$$

With the above constraints noted, we shall not distinguish between a policy $\pi \in \Pi_{\text{obs}}$ and its corresponding policy π^y in the augmented MDP, and write π for both.

For the finite horizon problem, the objective function for the augmented state MDP is

$$J(t, y, \pi) = \mathbb{E}_y^{\pi}\left[\sum_{n=t}^{2N} r_{y,\frac{n}{2}}\left(y_{\frac{n}{2}}, \pi_{\frac{n}{2}}\right)\right], \quad 0 \le t \le 2N,\ y \in \mathcal{Y},\ \pi \in \Pi_{\text{obs}}, \tag{2.17}$$

with value function

$$v(t, y) = \sup_{\pi \in \Pi_{\text{obs}}} J(t, y, \pi). \tag{2.18}$$

The proposition below shows the dynamic programming equation in the form of a quasi-variational inequality, for which the value function for the OCM satisfies.

Proposition 2.8. *For $n \in \mathbb{N}$, $y = (k, x, a) \in \mathcal{Y}$, define $v_{a,x}^{n,k} = v(n, y)$ as in (2.18). Then the value function satisfies the following quasi-variational inequality (QVI): for all $0 \le k < n \le N-1$, $x \in \mathcal{X}$, and $a \in A$,*

$$\begin{aligned} \min\Bigg\{ & v_{a,x}^{n,k} - v_{a,x}^{n+1,k} - \left(P_a^{n-k}\ r_a\right)_x, \\ & v_{a,x}^{n,k} - \left(P_a^{n-k}\ \overline{v^{n+1,n} + r}\right)_x + c_{\text{obs}} \Bigg\} = 0, \end{aligned} \tag{2.19}$$

with the terminal condition

$$v_{a,x}^{N,k} = \left(P^{N-k} r_a\right)_x, \tag{2.20}$$

where P_a^n is the n-step transition matrix with constant action a, and

$$(r_a)_x = r(x, a), \quad \bar{r}_x = \max_{a \in A} r(x, a), \tag{2.21}$$

$$\left(\overline{v^{n+1,n} + r}\right)_x = \max_{a \in A}(v_{a,x}^{n+1,n} + r_{a,x}). \tag{2.22}$$

Proof. A standard application of dynamic programming gives us

$$v(t, y) = \sup_{\pi \in \Pi_{\text{obs}}} \left\{ r_{y,t}(y, \pi_t) + \mathbb{E}^{\pi}\left[v\left(t + \frac{1}{2}, y_{t+\frac{1}{2}}\right)\right]\right\}.$$

Expanding explicitly, for $n \in \mathbb{N}$ and $y = (k, x, a) \in \mathcal{Y}$,

$$v(n, (k, x, a)) = \max\left\{-c_{\text{obs}} + \max_{a \in A}\ \mathbb{E}^a\left[v\left(n + \frac{1}{2}, (n, x_n, \varnothing)\right)\right],\ v\left(n + \frac{1}{2}, (k, x, a)\right)\right\},$$

where $\mathbb{E}^a$ is the expectation taken with respect to a constant $a \in A$. Furthermore

$$v\left(n+\frac{1}{2},(k,x,a)\right) = \mathbb{E}^a\left[r(x_n,a)\right] + v(n+1,(k,x,a)),$$
$$v\left(n+\frac{1}{2},(n,x,\varnothing)\right) = \max_{a\in A}\left\{r(x,a) + v(n+1,(n,x,a))\right\}$$

Thus, by combining the inspection stage and the action stage together, and rearranging the terms accordingly, we obtain the QVI in the desired form. □

The optimal policy π^* is then Markovian with respect to the augmented state y, i.e. the optimal policy depends on the most recent observation. Given a solution to the QVI (2.19), one can retrieve the optimal policy at time n, by first finding the region where the minimum is achieved, which determines if an inspection is optimal. If an inspection is optimal, one observes the latest state, say x, and the optimal action is given by $\arg\max_{a\in A}(v_{a,x}^{n+1,n} + r_{a,x})$.

2.2.2 Infinite horizon problem

For the discounted infinite horizon problem of the OCM, we will have to consider the appropriate stationary formulations. This can easily be obtained by further considering (n, y_n) as an augmented state. Now recall that the transition kernel of y in (2.11) to (2.13) depends on n and k strictly through the difference $n-k$. Hence, after relabelling, it is sufficient to consider $y = (n, x, a)$ as the augmented state, where here n now represents the time elapsed from the previous observation, rather than the standard linear passage of time. The objective function of this equivalent MDP is

$$J(y,\pi) = \mathbb{E}^\pi\left[\sum_{n=0}^{\infty}\gamma^{\lfloor\frac{n}{2}\rfloor} r_y\left(y_{\frac{n}{2}},\pi_{\frac{n}{2}}\right)\right], \quad y\in\mathcal{Y},\ \pi\in\Pi_{\text{obs}}, \tag{2.23}$$

where $\gamma \in (0,1)$ is the discount factor, the value function is

$$v(y) = \sup_{\pi\in\Pi_{\text{obs}}} J(y,\pi). \tag{2.24}$$

This gives us the following QVI for the value function, which we shall state here without proof.

Proposition 2.9. *For $y = (n,x,a) \in \mathcal{Y}$, define $v_{a,x}^n = v(y)$. Then the value function (2.24) satisfies the following quasi-variational inequality (QVI): for all $n \geq 1$, $x \in \mathcal{X}$, and $a \in A$,*

$$\min\left\{v_{a,x}^n - \gamma v_{a,x}^{n+1} - \left(P_a^n\ r_a\right)_x,\right.$$

$$v^n_{a,x} - \left(P^n_a \ \overline{\gamma v^1 + r} \right)_x + c_{\text{obs}} \Big\} = 0, \tag{2.25}$$

where P^n_a is the n-step transition matrix with constant action a, and

$$(r_a)_x = r(x,a), \quad \bar{r}_x = \max_{a \in A} r(x,a) \tag{2.26}$$

$$\left(\overline{\gamma v^1 + r} \right)_x = \max_{a \in A} (\gamma v^1_{a,x} + r_{a,x}). \tag{2.27}$$

Note that the QVI (2.25) is defined on the infinite domain $\mathbb{N}_{\geq 1} \times \mathcal{X} \times A$. In practice, we will have to truncate this domain for the time variable. A natural boundary condition is to enforce an inspection of the underlying chain after some large time N has elapsed. This is equivalent to further restricting the admissible policies in Π_{obs} to those such that $i_N = 1$.

2.2.3 Observation cost with parameter uncertainty

We now consider the case of the OCM with parameter uncertainty in the dynamics of the Markov chain. We shall adopt the approach of Bayesian adaptive control. Suppose that the transition kernel p now depends on an unknown parameter $\theta \in \Theta$, where Θ denotes a finite parameter space. We write $p_\theta(\cdot \mid x, a)$ for a fixed value of θ. To consider a Markov system for the problem, we take $\mathcal{X} \times \Theta$ as our underlying space, with transition kernel

$$\mathfrak{p}((x', \theta') \mid (x, \theta), a) \coloneqq \mathbb{1}_{\{\theta = \theta'\}} p_\theta(x' \mid x, a), \quad (x, \theta), (x', \theta') \in \mathcal{X} \times \Theta, \ a \in A. \tag{2.28}$$

For each value of $\theta \in \Theta$, we associate a probability measure $p^\theta_0 \in \Delta_{\mathcal{X}}$, so that the initial distribution is given by

$$\mathfrak{p}_0(x, \theta) = \sum_{\theta \in \Theta} p^\theta_0(x) \rho_0(\theta)$$

for some $\rho_0 \in \Delta_\Theta$. Given the half-step construction as laid out in the beginning of this section, the OCM with parameter uncertainty can be written as a POMDP $\langle \mathcal{X} \times \Theta, \mathcal{X}_\varnothing, \mathcal{A}, \mathfrak{p}, \mathfrak{p}_0, \mathfrak{q}, \mathfrak{q}_0, r_{\text{obs}} \rangle$ over the timescale $\frac{1}{2}\mathbb{N}$, where the domain of the transition kernel $\mathfrak{p}$ is extended to $(\mathcal{X} \times \Theta) \times \mathcal{A}$ by defining

$$\mathfrak{p}(\cdot \mid x, \theta, i) = \delta_{(x,\theta)}(\cdot), \quad i \in \mathcal{I},$$

and the observation kernel $\mathfrak{q}$ is now defined on $\mathcal{A} \times (\mathcal{X} \times \Theta)$, with

$$\mathfrak{q}(\cdot \mid a, x, \theta) = \delta_\varnothing(\cdot), \quad a \in A,$$

$$\mathfrak{q}(\cdot \mid i, x, \theta) = i \cdot \delta_x(\cdot) + (1-i)\delta_{\varnothing}(\cdot), \quad i \in \mathcal{I}.$$

The initial observation kernel $\mathfrak{q}_0$ will be taken as $\mathfrak{q}_0(\cdot \mid x, \theta) = \delta_x(\cdot)$ (see Remark 2.6). The set of admissible policies, denoted by $\Pi^{\Theta}_{\text{obs}}$ in this case, can be established analogously as in Definition 2.5. Denote the canonical measure here by $\mathbb{P}^{\pi}_{\mathfrak{p}_0}$, under which θ can be considered as a constant process, i.e. $\theta_{n+1} \equiv \theta_n$ with $\theta_0 \sim \rho_0 \in \Delta_\Theta$, and ρ_0 can be interpreted as a prior estimate for θ.

When considering the belief MDP for this problem, the observable sequence at time t remains as previously,

$$h_t = (\bar{x}_0, \pi_0, \ldots, \bar{x}_{t-1/2}, \pi_{t-1/2}, \bar{x}_t) \in \mathcal{H}_t.$$

Then, the belief state $\mathbb{P}^{\pi}_{\mathfrak{p}_0}(x_t, \theta_t \mid h_t)$ can be decomposed as

$$\mathbb{P}^{\pi}_{\mathfrak{p}_0}(x_t, \theta_t \mid h_t) = \sum_{\theta \in \Theta} \mathbb{P}^{\pi}_{\mathfrak{p}_0}(x_t \mid \theta_t, h_t)\ \mathbb{P}^{\pi}_{p_0}(\theta_t \mid h_t). \tag{2.29}$$

For a fixed value of θ, $\mathbb{P}^{\pi}_{\mathfrak{p}_0}(x_t \mid \theta_t, h_t)$ has a finite dimensional characterisation by the Markov property. As in the previous section, we denote this characterisation by $y = (y_t)_t$, which is a tuple given by

$$y_t \coloneqq \begin{cases} (k, x_k, a_k), & \text{if } k = \arg\max_{n \le t}\{i_n = 1\} \neq t\ , \\ (t, x_t, \varnothing), & \text{otherwise.} \end{cases} \tag{2.30}$$

The second term on the right hand side of (2.29), $\mathbb{P}^{\pi}_{\mathfrak{p}_0}(\theta_t \mid h_t)$ can be interpreted as the posterior distribution of θ at time t. Denote this term by $\rho_t(\theta_t)$. Note that ρ_t can be computed online via the classical Bayes' Theorem,

$$\rho_t(\theta) = \frac{\mathbb{P}^{\pi}_{\mathfrak{p}_0}(\bar{x}_t \mid \theta, h_{t-\frac{1}{2}}, \pi_{t-\frac{1}{2}})}{\sum_{\theta' \in \Theta} \mathbb{P}^{\pi}_{\mathfrak{p}_0}(\bar{x}_t \mid \theta', h_{t-\frac{1}{2}}, \pi_{t-\frac{1}{2}})\ \rho_{t-\frac{1}{2}}(\theta')}\ \rho_{t-\frac{1}{2}}(\theta). \tag{2.31}$$

In the OCM, observations only occur at the half steps, therefore we have in fact $\rho_n \equiv \rho_{n-1/2}$ for $n \in \mathbb{N}$. Thus, the update (2.31) can be reduced to

$$\rho_n(\theta) = \frac{\mathbb{P}^{\pi}_{\mathfrak{p}_0}(\bar{x}_{n-\frac{1}{2}} \mid \theta, y_{n-1}, i_{n-1})}{\sum_{\theta' \in \Theta} \mathbb{P}^{\pi}_{\mathfrak{p}_0}(\bar{x}_{n-\frac{1}{2}} \mid \theta', y_{n-1}, i_{n-1})\ \rho_{n-1}(\theta')}\ \rho_{n-1}(\theta) \tag{2.32}$$

$$=: U(\rho_{n-1}, y_{n-1}, i_{n-1}), \quad n \in \mathbb{N}. \tag{2.33}$$

The belief state at time t can now be represented by $y_t \in \mathcal{Y}$ and $\rho_t \in \Delta_\Theta$, with its transitions given by the kernel $\mathfrak{p}' = (\mathfrak{p}'_t)_{t \in \frac{1}{2}\mathbb{N}}$ on $\mathcal{Y} \times \Delta_\Theta$, given $(\mathcal{Y} \times \Delta_\Theta) \times \mathcal{A}$:

$$\mathfrak{p}'_n(y', \rho' \mid y, \rho, i) = i \cdot \sum_{\theta \in \Theta} p^{(n-k)}_{\theta}(\hat{x} \mid x, a)\rho(\theta)\ \mathbb{1}_{\{y' = (n, \hat{x}, \varnothing),\ \rho' = U(\rho, y, i)\}}$$

$$
\begin{aligned}
&+ (1-i)\mathbb{1}_{\{\hat{y}=y,\ \rho'=\rho\}}, && y=(k,x,a),\\
\mathfrak{p}'_{n+\frac{1}{2}}(y',\rho' \mid y,\rho,a') &= \mathbb{1}_{\{y'=(k,x,a'),\ a'=a,\ \rho'=\rho\}}, && y=(k,x,a),\\
\mathfrak{p}'_{n+\frac{1}{2}}(y',\rho' \mid y,\rho,a') &= \mathbb{1}_{\{\hat{y}=(n,x,a'),\ \rho'=\rho\}}, && y=(n,x,\varnothing).
\end{aligned}
$$

As before, we shall not distinguish between policies for the POMDP and policies for the belief state MDP. For the finite horizon problem, let $y=(k,x,a)\in\mathcal{Y}$, $\rho\in\Delta_\Theta$, and consider

$$
J(t,y,\rho,\pi)=\mathbb{E}^\pi_{y,\rho}\left[\sum_{n=t}^{2N} r^\rho_{n/2}\left(y_{\frac{n}{2}},\pi_{\frac{n}{2}}\right)\right],\quad 0\le t\le 2N,\ \pi\in\Pi_{\text{obs}}, \tag{2.34}
$$

where for $y=(k,x,a)$, $i\in\mathcal{I}$, $a'\in A$, $n\in\mathbb{N}$,

$$
\begin{aligned}
r^\rho_n(y,i) &= -i\cdot c_{\text{obs}},\\
r^\rho_{n+\frac{1}{2}}(y,a') &= \sum_{\theta\in\Theta}\sum_{x'\in\mathcal{X}} r(x',a')p^{(n-k)}_\theta(x'\mid x,a)\rho(\theta),
\end{aligned}
$$

and for $y=(n,x,\varnothing)$,

$$
r^\rho_{n+\frac{1}{2}}(y,a')=r(x,a').
$$

The value function is

$$
v(t,y,\rho)=\sup_{\pi\in\Pi_{\text{obs}}} J(t,y,\rho,\pi). \tag{2.35}
$$

As in the previous case without parameter uncertainty, the dynamic programming equation can be reduced to only the integer time steps. We state the optimality equation for the observation cost model under parameter uncertainty below.

Proposition 2.10. *For $y=(k,x,a)\in\mathcal{Y}$ and $\rho\in\Delta_\Theta$, the value function (2.35) satisfies the following equation:*

$$
v(n,(k,x,a),\rho) \tag{2.36}
$$

$$
\begin{aligned}
=\ \max\Bigg\{ & v(n+1,(k,x,a),\rho)+\sum_{\substack{\theta\in\Theta\\ x'\in\mathcal{X}}} p^{(n-k)}_\theta(x'\mid x,a)r(x',a)\rho(\theta),\\
& \sum_{\substack{\theta\in\Theta\\ x'\in\mathcal{X}}} p^{(n-k)}_\theta(x'\mid x,a)\rho(\theta)\ \left[\max_{a'\in A}\Big(v(n+1,(n,x',a'),\rho')+r(x',a')\Big)\right]-c_{\text{obs}}\Bigg\},
\end{aligned} \tag{2.37}
$$

where $\rho'=U(\rho,y,1)$ as in (2.32).

For the infinite horizon case, a similar stationary argument leads us to the objective function and value function:

$$J(y,\rho,\pi) = \mathbb{E}^{\pi}\left[\sum_{n=0}^{\infty}\gamma^{\lfloor\frac{n}{2}\rfloor}r^{\rho}\left(y_{\frac{n}{2}},\pi_{\frac{n}{2}}\right)\right],\ y\in\mathcal{Y},\ \rho\in\Delta_{\Theta},\ \pi\in\Pi_{\text{obs}}, \tag{2.38}$$

$$v(y,\rho) = \sup_{\pi\in\Pi_{\text{obs}}} J(y,\rho,\pi). \tag{2.39}$$

Proposition 2.11. *For $y=(n,x,a)\in\mathcal{Y}$ and $\rho\in\Delta_{\Theta}$, the value function* (2.39) *satisfies the following equation:*

$$\begin{aligned} v((n,x,a),\rho) = \max\Bigg\{&\gamma v((n+1,x,a),\rho) + \sum_{\substack{\theta\in\Theta\\ x'\in\mathcal{X}}} p_{\theta}^{(n)}(x'\mid x,a)r(x',a)\rho(\theta),\\ &\sum_{\substack{\theta\in\Theta\\ x'\in\mathcal{X}}} p_{\theta}^{(n)}(x'\mid x,a)\rho(\theta)\ \left[\max_{a'\in A}\Big(\gamma v((1,x',a'),\rho') + r(x',a')\Big)\right] - c_{\text{obs}}\Bigg\}, \end{aligned} \tag{2.40}$$

where $\rho' = U(\rho,y,1)$ as in (2.32).

It is worth noting that both (2.36) and (2.40) are MDPs over the augmented space $\mathcal{Y}\times\Delta_{\Theta}$. The inclusion of the simplex Δ_{Θ} makes the MDP non-discrete. For computation, one would have to approximate the solution, either via computing a discrete MDP on a finite grid on $\mathcal{Y}\times\Delta_{\Theta}$, or via functional approximation methods such as the use of neural networks on larger scale problems. We refer the reader to the textbook [44] and survey paper [58] for a comprehensive review of choosing appropriate approximating grids. Then, given a finite grid $\mathbb{G}=\{s_1,\ldots,s_G\}$ on the simplex Δ_{Θ}, one can define the approximating transition kernels by

$$\mathfrak{p}_{\mathbb{G}}(y',s_i\mid y,s_j,\pi) = \frac{\mathfrak{p}'(y',s_i\mid y,s_j,a)}{\sum_{i=1}^{G}\mathfrak{p}'(y',s_i\mid y,s_j,a)},\quad y,y'\in\mathcal{Y},\ s_i,s_j\in\mathbb{G},\ \pi\in\mathcal{A},$$

so that one solves the approximating finite MDP on $\mathcal{Y}\times\mathbb{G}$ instead. For the QVIs resulting from the observation cost problems, we propose to solve the MDPs by a penalty method, detailed in Section 2.3. In Section 2.4.2, we consider a random walk with parameter uncertainty, imposing conjugate distributions for the unknown parameter, such that the distributions $\{\rho_n\}$ can be described by a finite number of values over a finite time horizon.

2.2.4 Toy problem

To illustrate the framework, we present a model problem involving a two-state Markov chain and give an explicit solution. We assume the following setup:

- the state space $\mathcal{X} = \{0, 1\}$;
- the action space $A = \{0, 1\}$;
- the reward function $r(x, a) = a \cdot x + (1 - a)(a - x)$;
- the transition matrix

$$P_a = \begin{array}{c} \\ a \\ 1-a \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{bmatrix} a + p(1-a) & (1-p)(1-a) \\ (1-p)a & p \cdot a + (1-a) \end{bmatrix} \end{array} \begin{array}{c} \\ 0 \\ 1 \end{array} \tag{2.41}$$

  where $p \in (0, 1)$.

This can be seen as a model for a maintenance problem, to find an optimal interval for inspecting equipment to avoid wear and tear over time. The reward function r gives a reward of 1 when the state and action values are the same, and zero otherwise. If no changes are made to the action, the chain eventually arrives at the absorbing state which does not incur any reward. Figure 2.2 illustrates the chain for the case $p = 0.9$.

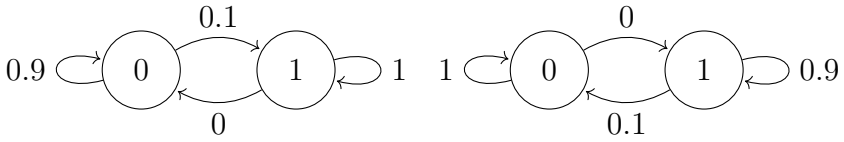

Figure 2.2: *Illustration of the two-state Markov chain. Left: $a = 0$; Right: $a = 1$.*

Consider the infinite horizon problem. The QVI for $x = 0$ and $a = 0$ is

$$\min\left\{ v_{0,0}^n - \gamma v_{0,0}^{n+1} - \left(P_0^n \; r_0\right)_x \,, \right.$$
$$\left. v_{0,0}^n - \left(P_0^n \overline{\gamma v^1 + r}\right)_x + c_{\text{obs}} \right\} = 0. \tag{2.42}$$

Due to the symmetry of the problem, we have $v_{0,0}^n = v_{1,1}^n$. Moreover, it is clear that given the knowledge of x_n, the optimal action is to set $a_n = x_n$. Hence we can write (2.42) as

$$\min\{v_{0,0}^n - \gamma v_{0,0}^{n+1} - p^n,\; v_{0,0}^n - \gamma v_{0,0}^1 - 1 + c_{\text{obs}}\} = 0$$

or simply by writing $v(n) = v^n_{0,0}$:

$$v(n) = \max\{p^n + \gamma v(n+1),\ 1 - c_{\text{obs}} + \gamma v(1)\}. \tag{2.43}$$

Let T be the first optimal inspection time (where by convention $T = \infty$ if it is optimal to never inspect). The value function is given recursively by

$$v(n) = \begin{cases} 1 - c_{\text{obs}} + \gamma v(1), & \text{if } n \geq T; \\ p^n + \gamma v(n+1), & \text{otherwise.} \end{cases} \tag{2.44}$$

Solving the above for $v(1)$, we obtain the explicit solution

$$v(1) = \max\left\{ \sup_{m \geq 2} \left(\frac{p \sum_{k=0}^{m-2} (\gamma p)^k + \gamma^{m-1}(1 - c_{\text{obs}})}{1 - \gamma^m} \right), \frac{1 - c_{\text{obs}}}{1 - \gamma} \right\}, \tag{2.45}$$

from which $v(n)$ for $n \geq 1$ can be calculated from (2.44). The first term in (2.45) is a geometric series, where $p \sum_{k=0}^{m-2} (\gamma p)^k + \gamma^{m-1}(1 - c_{\text{obs}})$ is the expected returns across the optimal inspection interval, and γ^m is the discount factor over the whole interval. We can interpret (2.45) as searching for the optimal inspection interval to maximise the sum of the rewards, minus the observation cost.

2.3 Comparison principle and penalisation

In this section, we consider a class of discrete QVIs given by Problem 2.12 below.

Problem 2.12. Find $u = (u_1, \ldots, u_d) \in \mathbb{R}^{N \times L \times d}$ such that

$$\min\{F_a(u), u_a - \mathcal{M}u\} = 0, \quad a \in \{1, \ldots, d\} =: A, \tag{2.46}$$

where

- $\mathcal{M} : \mathbb{R}^{N \times L \times d} \to \mathbb{R}^{N \times L}$ is defined by

$$(\mathcal{M}u)^n_l = \left(Q_n \overline{u^1 - c}\right)_l, \quad \left(\overline{u^1 - c}\right)_l = \max_{a \in A} \left(u^1_{a,l} - c_{a,l}\right), \tag{2.47}$$

  for a given vector $c \in \mathbb{R}^{L \times d}$ and $\{Q_n\} \subset \mathbb{R}^{L \times L}$ is a sequence of strictly sub-stochastic matrices;

- $F_a : \mathbb{R}^{N \times L} \to \mathbb{R}^{N \times L}$ is a continuous function that satisfies the following property: there exists a constant $\beta > 0$ such that for any $u, v \in \mathbb{R}^{N \times L \times d}$ with $u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}} = \max_{n,a,l}(u^n_{a,l} - v^n_{a,l}) \geq 0$, we have

$$F_{\bar{a}}(u)^{\bar{n}}_{\bar{l}} - F_{\bar{a}}(v)^{\bar{n}}_{\bar{l}} \geq \beta(u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}}). \tag{2.48}$$

We shall refer to (2.48) as the monotonicity condition. In general the vector c in (2.47) can also depend on n, and the proofs for such cases extends naturally. We interpret the indices $n \in \{1, \dots, N\}$ as the time domain, $l \in \{1, \dots, L\}$ as the spatial domain, and $a \in A$ as the action space. For example, for the infinite horizon problem (2.25), one has N is the size of the truncated time domain (see comments after Proposition 2.9), $L = |\mathcal{X}|$, with

$$\begin{aligned} F_a(u)_l^n &= u_{a,l}^n - \gamma u_{a,l}^{n+1} - (P_a^n r_a)_l, \\ Q_n &= \gamma P_a^n, \\ c_{a,l} &= c_{\text{obs}} - \frac{1}{\gamma} r_{a,l}. \end{aligned}$$

Moreover, it is straightforward to see that F_a satisfies the monotonicity condition (2.48) with parameter $\beta = 1 - \gamma$. For the case with parameter uncertainty, if the measures $\rho_n(d\theta)$ can be parametrised by a finite number of parameters w, this can also be considered as part of the spatial domain. In this case $L = |\mathcal{X}| \cdot |w|$.

The monotonicity condition arises naturally, for example, from the discretisation of QVIs in continuous time involving switching controls. Indeed, one can view, for example, the terms $u_{a,l}^n - \gamma u_{a,l}^{n+1}$ as suitably rescaled finite difference terms. The operator $\mathcal{M}$ will be referred to as the *inspection operator*, and is a non-local term in the QVI that couples the solution u across the different action values. In the case where the Q_n's are the identity matrix, then (2.46) reduces to a QVI with interconnected obstacles, see [56] for a more detailed analysis for such classes of QVIs.

In the following, we adapt the argument in [56] to prove a comparison principle of the QVI (2.46).

Proposition 2.13. *Suppose $u = (u_a)_{a \in A}$ (resp. $v = (v_a)_{a \in A}$) satisfies*

$$\min\{F_a(u),\ u_a - \mathcal{M}u\} \leq 0 \quad (\text{resp.} \geq 0), \quad a \in A; \tag{2.49}$$

then $u \leq v$.

Proof. Let $M \coloneqq u_{\bar{a},\bar{l}}^{\bar{n}} - v_{\bar{a},\bar{l}}^{\bar{n}} = \max_{n,a,l}(u_{a,l}^n - v_{a,l}^n)$. Suppose for a contradiction that $M > 0$. Since u is a subsolution, we have $F_{\bar{a}}(u) \leq 0$ or $u_{\bar{a}} - \mathcal{M}u \leq 0$. First suppose that $u_{\bar{a}} \leq (\mathcal{M}u)$. Then

$$u_{\bar{a},\bar{l}}^{\bar{n}} \leq \left(Q_{\bar{n}} \overline{u^1 - c}\right)_{\bar{l}}. \tag{2.50}$$

Similarly, as v is a supersolution, $v_{\bar{a}} - \mathcal{M}v \geq 0$ and

$$v^{\bar{n}}_{\bar{a},\bar{l}} \geq \left(Q_{\bar{n}}\overline{v^1 - c}\right)_{\bar{l}}. \tag{2.51}$$

Let $\gamma < 1$ be the maximum of the row sums of $Q_{\bar{n}}$. Then by combining both inequalities above we obtain

$$\begin{aligned} u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}} &\leq \left(Q_{\bar{n}}\overline{u^1 - c}\right)_{\bar{l}} - \left(Q_{\bar{n}}\overline{v^1 - c}\right)_{\bar{l}} \\ &\leq \left(Q_{\bar{n}}\overline{u^1 - v^1}\right)_{\bar{l}} \\ &\leq \gamma \left(\overline{u^1 - v^1}\right)_{l^*} && \text{(for some } l^* \in \{1, \ldots, L\}) \\ &< u^1_{a^*,l^*} - v^1_{a^*,l^*} && \text{(for some } a^* \in A) \end{aligned}$$

which is a contradiction by the maximality of M. Hence, we must have $F_{\bar{a}}(u)^{\bar{n}}_{\bar{l}} \leq 0$, but then since v is a supersolution and $M > 0$, we have by the monotonicity property

$$\beta(u^{\bar{n}}_{\bar{a},\bar{l}} - v^{\bar{n}}_{\bar{a},\bar{l}}) \leq F_{\bar{a}}(u)^{\bar{n}}_{\bar{l}} - F_{\bar{a}}(v)^{\bar{n}}_{\bar{l}} \leq 0 \tag{2.52}$$

which is again a contradiction. Hence we must have $M \leq 0$ as required. □

We now present a penalty approximation to the QVI (2.46). The motivations behind this are twofold. Firstly, it presents itself as an alternative numerical scheme to policy iteration, which can encounter instability issues due to potential singularities at the matrix iterates. Secondly, it provides naturally a constructive proof of existence to the solutions of the QVI (2.46). Consider the following penalised problem.

Problem 2.14. Let $\rho \geq 0$ be the penalty parameter. Find $u^\rho = (u^\rho_a)_{a\in A} \in \mathbb{R}^{N\times L\times d}$ such that

$$F_a(u^\rho) - \rho\, \pi\left(\mathcal{M}u^\rho - u^\rho_a\right) = 0,\ a \in A, \tag{2.53}$$

where the penalisation function $\pi : \mathbb{R} \to \mathbb{R}$ is continuous, non-decreasing with $\pi|_{(-\infty,0]} = 0$ and $\pi|_{(0,\infty)} > 0$, and is applied elementwise.

Thus in the penalised problem, a penalty ρ is applied whenever the condition $u^\rho_a - \mathcal{M}u^\rho \geq 0$ is violated. As $\rho \uparrow \infty$, the penalised solution should then converge to the solution of the discrete QVI (2.46). We first show below that for each fixed ρ, (2.53) satisfies a comparison principle. This implies uniqueness for Problem 2.14. The argument follows similarly to the approach in [56] and Proposition 2.13.

Proposition 2.15. *For any penalty parameter $\rho \geq 0$, if $u^\rho = (u_a^\rho)_{a\in A}$ (resp., $v^\rho = (v_a^\rho)_{a\in A}$) satisfies*

$$F_a(u^\rho) - \rho\, \pi\left(\mathcal{M}u^\rho - u_a^\rho\right) \leq 0 \quad (\text{resp.}, \geq 0), \tag{2.54}$$

then $u^\rho \leq v^\rho$.

Proof. Let $M \coloneqq u^{\rho,\bar{n}}_{\bar{a},\bar{l}} - v^{\rho,\bar{n}}_{\bar{a},\bar{l}} = \max_{n,a,l}(u^{\rho,n}_{a,l} - v^{\rho,n}_{a,l})$. Suppose for a contradiction that $M > 0$. From the previous proposition, we have that

$$\left(Q_{\bar{n}}\overline{u^{\rho,1} - c}\right)_{\bar{l}} - \left(Q_{\bar{n}}\overline{v^{\rho,1} - c}\right)_{\bar{l}} < u^{\rho,\bar{n}}_{\bar{a},\bar{l}} - v^{\rho,\bar{n}}_{\bar{a},\bar{l}}. \tag{2.55}$$

As π is non-decreasing,

$$\pi\left(\left(Q_{\bar{n}}\overline{u^{\rho,1} - c}\right)_{\bar{l}} - u^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right) \leq \pi\left(\left(Q_{\bar{n}}\overline{v^{\rho,1} - c}\right)_{\bar{l}} - v^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right). \tag{2.56}$$

As u^ρ and v^ρ are respectively sub and super solutions of (2.53), we have

$$\begin{aligned}
&F_{\bar{a}}(u^\rho)^{\bar{n}}_{\bar{l}} - \rho\, \pi\left((\mathcal{M}u^\rho)^{\bar{n}}_{\bar{l}} - u^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right) - \left(F_{\bar{a}}(v^\rho)^{\bar{n}}_{\bar{l}} - \rho\, \pi\left((\mathcal{M}v^\rho)^{\bar{n}}_{\bar{l}} - v^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right)\right)\\
\leq\ & F_{\bar{a}}(u^\rho)^{\bar{n}}_{\bar{l}} - F_{\bar{a}}(v^\rho)^{\bar{n}}_{\bar{l}} - \rho\left(\pi\left(\left(Q_{\bar{n}}\overline{u^{\rho,1} - c}\right)_{\bar{l}} - u^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right) - \pi\left(\left(Q_{\bar{n}}\overline{v^{\rho,1} - c}\right)_{\bar{l}} - v^{\rho,\bar{n}}_{\bar{a},\bar{l}}\right)\right)\\
\leq\ & 0
\end{aligned}$$

Hence by (2.56), $F_{\bar{a}}(u^\rho)^{\bar{n}}_{\bar{l}} - F_{\bar{a}}(v^\rho)^{\bar{n}}_{\bar{l}} \leq 0$. The monotonicity assumption of F then leads to a contradiction, so that $M \leq 0$ as required. □

The existence of solutions to the penalised equation (2.53) and its convergence to the QVI (2.46) is a straightforward adaptation of the results in [56], which we shall state here without proof.

Theorem 2.16 ([56, Theorem 2.5, 2.6])**.** *For any penalty parameter ρ, there exists a unique solution u^ρ to the penalised equation* (2.53)*, satisfying the bound $\|u^\rho\| \leq \|F(0)\|/\beta$. For a fixed $c \geq 0$, u^ρ converges monotonically from below to a function $u \in \mathbb{R}^{N\times L\times d}$ as $\rho \to \infty$. Moreover, u solves the discrete QVI* (2.46)*.*

Thus, we have a straightforward computation scheme to solve for the OCM. We first set up the discrete QVIs arising from the problem, which is then approximated by the penalised problem. The solution of the penalised problem is in turn approximated iteratively with semismooth Newton methods [69]. Formally speaking, starting with an initialisation $v^{(0)}$ to the penalised problem

$$G^\rho(v) \coloneqq F_a(v) - \rho\, \pi\left(\mathcal{M}v - v\right) = 0, \tag{2.57}$$

we obtain the next iterate by solving for

$$v^{(k+1)} = v^{(k)} - \mathcal{L}^{\rho}(v^{(k)})^{-1} G^{\rho}(v^{(k)}), \tag{2.58}$$

where $\mathcal{L}^{\rho}$ denotes the generalised derivative of the function G^{ρ}.

2.4 Numerical experiments

In this section, we apply our observation cost framework to three numerical experiments. Sections 2.4.1 and 2.4.3 analyse two infinite horizon problems. For these examples, we examine the numerical performance of the penalty method and Newton iterations, as well as the effects of the observation cost on the qualitative behavior of the solutions. For the penalised equations, we will employ the penalty function $\pi(x) = x^+$ as in [56]. Section 2.4.2 considers the parameter uncertainty formulation over a finite horizon. The solutions are obtained through backwards recursion from the terminal conditions. We examine the impact that the extra parameter uncertainty has on the optimal trajectories.

2.4.1 Random walk with drift

Consider an integer-valued random walk whose drift depends on the action space $A = \{+1, -1\}$. The probability of each step is parametrised by θ. Specifically, for any $x \in \mathcal{X} = \mathbb{N}$,

$$\begin{aligned} p(x+1 \mid x, +1) = \theta, \quad p(x-1 \mid x, +1) = 1-\theta; \\ p(x+1 \mid x, -1) = 1-\theta, \quad p(x-1 \mid x, -1) = \theta. \end{aligned} \tag{2.59}$$

We also adopt the following reward function:

$$r(x, a) = \frac{1}{|x|+1}. \tag{2.60}$$

The mass of this reward function r is concentrated around the origin, so naturally, the optimal action is one that reverts the process back towards the origin.

For this example, we consider the infinite horizon problem. Recall that the discrete QVI (2.25) reads: for all $n \geq 0$, $x \in \mathcal{X}$, and $a \in A$,

$$\min\left\{ v^n_{a,x} - \gamma v^{n+1}_{a,x} - \left(P^n_a \; r_a\right)_x \;,\; v^n_{a,x} - \left(P^n_a \; \overline{\gamma v^1 + r}\right)_x + c_{\text{obs}} \right\} = 0, \tag{2.61}$$

Note that there exists a path from x to y over m units of time if and only if $m \geq |y-x|$ and $m \equiv y \pmod 2$. If $\mathcal{S}_m^x$ denotes the set of states that can be reached from x after m units of time, then for a constant action, the n-step transition probabilties are given by

$$p^{(n)}(x' \mid x, +1) = \begin{cases} \binom{n}{k}\theta^k(1-\theta)^{n-k} & , \ x' \in \mathcal{S}_n^x; \\ 0 & , \ x' \notin \mathcal{S}_n^x, \end{cases} \tag{2.62}$$

$$p^{(n)}(x' \mid x, -1) = \begin{cases} \binom{n}{k}\theta^{n-k}(1-\theta)^{k} & , \ x' \in \mathcal{S}_n^x; \\ 0 & , \ x' \notin \mathcal{S}_n^x, \end{cases} \tag{2.63}$$

where $k = (n + x' - x)/2$. Hence, in full, the QVI reads:

$$\begin{aligned} \min \Bigg\{ & v_{+1,x}^n - \sum_{x' \in \mathcal{S}_n^x} \binom{n}{k} \theta^k (1-\theta)^{n-k} \left(\frac{1}{|x'|+1} + \gamma \left(\theta \overline{v}_{x'+1}^1 + (1-\theta)\overline{v}_{x'-1}^1 \right) \right) + c_{\text{obs}}, \\ & v_{+1,x}^n - \gamma v_{+1,x}^{n+1} - \sum_{x' \in \mathcal{S}_n^x} \frac{1}{|x'|+1} \binom{n}{k} \theta^k (1-\theta)^{n-k} \Bigg\} = 0, \\ \min \Bigg\{ & v_{-1,x}^n - \sum_{x' \in \mathcal{S}_n^x} \binom{n}{k} \theta^{n-k} (1-\theta)^{k} \left(\frac{1}{|x'|+1} + \gamma \left(\theta \overline{v}_{x'-1}^1 + (1-\theta)\overline{v}_{x'+1}^1 \right) \right) + c_{\text{obs}}, \\ & v_{-1,x}^n - \gamma v_{-1,x}^{n+1} - \sum_{x' \in \mathcal{S}_n^x} \frac{1}{|x'|+1} \binom{n}{k} \theta^{n-k} (1-\theta)^{k} \Bigg\} = 0. \end{aligned} \tag{2.64}$$

To close the system to ensure a unique solution, we enforce the following time and spatial boundary conditions. We impose a reflecting boundary at $x = \pm L$, where L is suitably large. In particular,

$$\begin{aligned} & p(L \mid L, +1) = \theta, && p(L-1 \mid L, +1) = 1-\theta, \\ & p(L \mid L, -1) = 1-\theta, && p(L-1 \mid L, -1) = \theta, \\ & p(-L \mid -L, +1) = 1-\theta, && p(-L+1 \mid -L, +1) = \theta, \\ & p(-L \mid -L, -1) = \theta, && p(-L+1 \mid -L, -1) = 1-\theta, \end{aligned} \tag{2.65}$$

so that the QVI (2.61) for the states $x = \pm L$ will use the transition probabilities (2.65) instead.

For the time boundary, we enforce an observation at some large time $N > 0$. The

terminal condition then reads (for $-L < x < L$):

$$\begin{cases} v^N_{+1,x} - \displaystyle\sum_{x' \in \mathcal{S}^x_N} \binom{N}{k'} \theta^k (1-\theta)^{N-k'} \left(\frac{1}{|x'|+1} + \gamma \left(\theta \overline{v}^1_{x'+1} + (1-\theta) \overline{v}^1_{x'-1} \right) \right) + c_{\text{obs}} = 0, \\ v^N_{-1,x} - \displaystyle\sum_{x' \in \mathcal{S}^x_N} \binom{N}{k'} \theta^{N-k'} (1-\theta)^{k'} \left(\frac{1}{|x'|+1} + \gamma \left(\theta \overline{v}^1_{x'-1} + (1-\theta) \overline{v}^1_{x'+1} \right) \right) + c_{\text{obs}} = 0. \end{cases} \tag{2.66}$$

where $k' = (N + x' - x)/2$. The analogous equations hold for the spatial boundary $x = \pm L$, but with the transition probabilities (2.65). These terminal conditions can be interpreted as the largest possible interval between two observations.

We now proceed to solve the penalised problem for the system (2.64), with boundary conditions (2.65) and (2.66), through the use of semismooth Newton methods. To initialise the iteration, we solve for the uncoupled system

$$\begin{cases} v^n_{+1,x} - \gamma v^{n+1}_{+1,x} - \displaystyle\sum_{x' \in \mathcal{S}^x_n} \frac{1}{|x'|+1} \binom{n}{k} \theta^k (1-\theta)^{n-k} = 0, \\ v^n_{-1,x} - \gamma v^{n+1}_{-1,x} - \displaystyle\sum_{x' \in \mathcal{S}^x_n} \frac{1}{|x'|+1} \binom{n}{k} \theta^{n-k} (1-\theta)^{k} = 0, \\ 0 \le n < N, \ -L < x < L, \end{cases} \tag{2.67}$$

with the spatial boundary transition probabilities (2.65) and time boundary

$$\begin{cases} v^N_{+1,x} = \displaystyle\sum_{x' \in \mathcal{S}^x_N} \binom{N}{k'} \theta^k (1-\theta)^{N-k'} \left(\frac{1}{|x'|+1} + \gamma \left(\theta v^1_{+1,x'+1} + (1-\theta) v^1_{+1,x'-1} \right) \right) - c_{\text{obs}}, \\ v^N_{-1,x} = \displaystyle\sum_{x' \in \mathcal{S}^x_N} \binom{N}{k'} \theta^{N-k'} (1-\theta)^{k'} \left(\frac{1}{|x'|+1} + \gamma \left(\theta v^1_{-1,x'-1} + (1-\theta) v^1_{-1,x'+1} \right) \right) - c_{\text{obs}}, \\ -L < x < L. \end{cases} \tag{2.68}$$

The system (2.67) corresponds to the penalised equation with penalty parameter $\rho = 0$. The uncoupled time boundary condition is equivalent to enforcing an observation but with no switching (i.e., assuming that $\overline{v} = v_a$ in each equation for v_a). The iteration terminates once a relative tolerance threshold of 10^{-8} is reached.

We investigate the numerical performance of our described methods for the case $\theta = 0.75$, $\gamma = 0.99$, $L = 50$ and $N = 500$, across different cost parameters c_{obs}. Computations are performed using MATLAB R2019b. The numerical solutions are shown in Table 2.3. Row (a) shows that the number of Newton iterations required to

	ρ	10^3	2×10^3	4×10^3	8×10^3	16×10^3	32×10^3
$c_{\text{obs}} = 0$	(a)	2	2	2	2	2	2
	(b)	$6.33e^{-3}$	$3.17e^{-3}$	$1.58e^{-3}$	$7.92e^{-4}$	$3.96e^{-4}$	$1.98e^{-4}$
$c_{\text{obs}} = 1/8$	(a)	5	5	5	5	5	5
	(b)	$4.85e^{-3}$	$2.42e^{-3}$	$1.21e^{-3}$	$6.06e^{-4}$	$3.03e^{-4}$	$1.52e^{-4}$
$c_{\text{obs}} = 1/4$	(a)	6	6	6	6	6	6
	(b)	$3.38e^{-3}$	$1.69e^{-3}$	$8.47e^{-4}$	$4.23e^{-4}$	$2.12e^{-4}$	$1.06e^{-4}$
$c_{\text{obs}} = 1/2$	(a)	6	6	6	6	6	6
	(b)	$1.54e^{-3}$	$7.69e^{-4}$	$3.85e^{-4}$	$1.92e^{-4}$	$9.62e^{-5}$	$4.81e^{-5}$
$c_{\text{obs}} = 1$	(a)	7	7	7	7	7	7
	(b)	$6.21e^{-4}$	$3.11e^{-4}$	$1.55e^{-4}$	$7.76e^{-5}$	$3.88e^{-5}$	$1.94e^{-5}$
$c_{\text{obs}} = 2$	(a)	8	8	8	8	8	8
	(b)	$2.08e^{-4}$	$1.04e^{-4}$	$5.19e^{-5}$	$2.60e^{-5}$	$1.30e^{-5}$	$6.5e^{-6}$
$c_{\text{obs}} = 4$	(a)	7	7	7	7	7	7
	(b)	$8.52e^{-5}$	$4.26e^{-5}$	$2.13e^{-5}$	$1.53e^{-5}$	$5.3e^{-6}$	$2.7e^{-6}$
$c_{\text{obs}} = 6$	(a)	6	6	6	6	6	6
	(b)	$3.07e^{-5}$	$1.54e^{-5}$	$7.7e^{-6}$	$3.8e^{-6}$	$1.9e^{-6}$	$1.0e^{-6}$

Table 2.3: *Numerical results for the random walk with drift problem. Line (a): number of Newton iterations to reach the relative tolerance threshold of* $1e^{-8}$*. Line (b): the increment sizes* $\|v^\rho - v^{2\rho}\|_\infty$*.*

reach the tolerance threshold is independent from the size of the penalty parameter ρ. Fewer iterations are required for more extreme values of c_{obs}, but the overall number of iterations remains low across different observation costs. Row (b) shows the increments $\|v^\rho - v^{2\rho}\|_\infty$. The values suggests a first-order convergence of the penalisation error with respect to the penalty parameter ρ, which is in line with the analogous theoretical results in [56, Theorem 3.9, 4.2].

c_{obs}	0	1/8	1/4	1/2	1	2	4	6
$x = 5$	5	5	7	9	11	15	25	37
$x = 10$	10	12	12	16	20	26	36	48
$x = 30$	30	40	42	46	54	62	78	110

Table 2.4: *List of optimal observation times across various states* x *and costs* c_{obs}*.*

We now discuss the qualitative behaviour of the solution. It is clear that if the chain is observed to be at a positive state, then the control should be switched to $a = -1$ for a negative drift and vice versa. Table 2.4 lists the optimal observation time gap for selected states across different observation costs c_{obs}. As the problem is symmetric by construction, it is sufficient to only examine the behavior for the positive states. In general, the optimal observation time increases as c_{obs} increases. A longer unobserved

period of time then leads to a lower average reward. This is illustrated in Figure 2.5, where the function $n \mapsto v^n_{-1,30}$ is plotted for various values of c_{obs}. In the absence of an observation cost, i.e., for $c_{\text{obs}} = 0$, the optimal observation time equals the magnitude of the last observed state, as there is no need to observe until it is possible for the walk to cross the origin again.

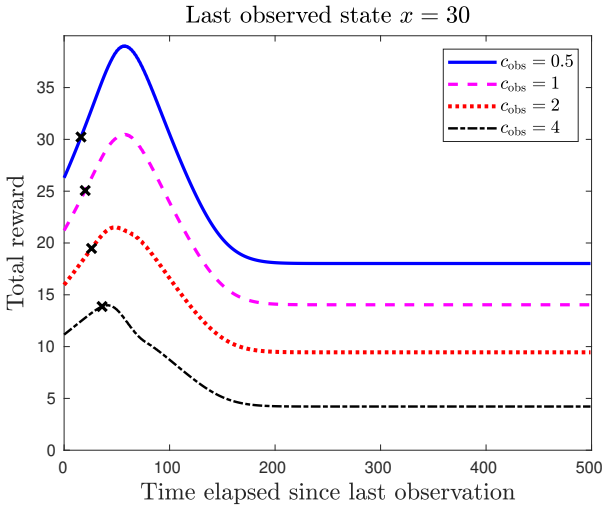

Figure 2.5: *Difference in total reward obtained when altering the observation cost c_{obs}. Each line shows the graph of $n \mapsto v^n_{-1,30}$. The cross indicates the optimal observation time.*

2.4.2 Random walk with drift with parameter uncertainty

In this subsection, we consider a random walk with drift, as set up in Section 2.4.1, but with the additional assumption that the true value of the drift parameter θ is unknown to the user. To avoid complications with boundary conditions and infinite domains, we shall only consider the finite horizon problem. Recall that for a fixed value of θ and constant action, the n-step transition probabilities are given by

$$p^{(n)}_\theta(x' \mid x, +1) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \; p^{(n)}_\theta(x' \mid x, -1) = \binom{n}{k}\theta^{n-k}(1-\theta)^k, \quad x' \in \mathcal{S}^x_n, \tag{2.69}$$

where $\mathcal{S}^x_n$ is the set of states that can be reached from x after n units of time, and $k = \frac{1}{2}(n + x' - x)$. As remarked at the end of Section 2.2.3, we choose the prior from a family of beta distributions to obtain conjugacy in the parameter distributions. This reduces (2.36) to a finite QVI. The posterior can be updated as follows. Suppose the prior $\rho_0 \sim \text{Beta}(\alpha, \beta)$ and the next observation occurs at time n at a state $x' \in \mathcal{X}$.

Then a standard calculation shows that

$$\int_\Theta p_\theta^{(n)}(x' \mid x, +1)\ \rho_0(d\theta) = g(k \mid n, \alpha, \beta),$$
$$\int_\Theta p_\theta^{(n)}(x' \mid x, -1)\ \rho_0(d\theta) = g(n-k \mid n, \alpha, \beta),$$

where

$$g(k \mid n, \alpha, \beta) = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha, \beta)},\ k = \frac{1}{2}(n + x' - x), \tag{2.70}$$

and g is the probability mass function of the Beta-binomial distribution, $B(\alpha, \beta)$ is the Beta function. The posterior distribution is then given by

$$\rho_n \sim \begin{cases} \text{Beta}(\alpha + k, \beta + n - k), & a_0 = +1, \\ \text{Beta}(\alpha + n - k, \beta + k), & a_0 = -1. \end{cases} \tag{2.71}$$

Since $\{\rho_n\}_n$ can now be characterised by the parameters of the Beta distribution (α, β), we write $v(n, (k, x, a), (\alpha, \beta))$ for $v(n, y, \rho)$. Let us first consider the same reward function $r(x, a) = r(x) = \frac{1}{|x|+1}$ as in the previous section. We can write the QVI (2.36) as

$$\begin{aligned} & v(n, (k, x, a), (\alpha, \beta)) \\ = \ & \max \Bigg\{ v(n+1, (k, x, a), (\alpha, \beta)) + \sum_{x' \in \mathcal{X}} g'(x', x \mid n-k, \alpha, \beta)\ r(x'), \\ & \sum_{x' \in \mathcal{X}} g'(x', x \mid n-k, \alpha, \beta) \Big[\max_{a' \in A} \Big(v(n+1, (n, x', a'), (\alpha', \beta')) + r(x') \Big) \Big] - c_{\text{obs}} \Bigg\}, \end{aligned} \tag{2.72}$$

where we define $g'(x', x \mid n, \alpha, \beta) \coloneqq g((n - x' + x)/2 \mid k, \alpha, \beta)$. The terminal conditions are

$$v(N, (N-k, x, a), (\alpha, \beta)) = \sum_{x' \in \mathcal{S}_k^x} g'(x', x \mid k, \alpha, \beta)\ r(x'), \quad k < N. \tag{2.73}$$

For our experiment, we set the true value of $\theta = 0.3$ and a time horizon of $N = 50$. Figure 2.6 illustrates a sample realisation of an optimal trajectory, given a prior of $\text{Beta}(5, 2)$, as well as the evolution of the estimate over θ over time.

We consider three different choices of (α, β) for the prior ρ_0 as well as varying the observation cost. For each parameter combination we compute the optimal policy and

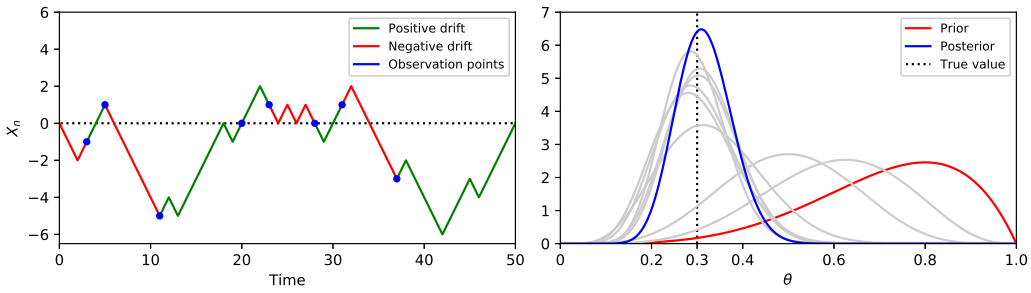

Figure 2.6: *Left: sample realisation of the controlled random walk along the optimal trajectory. Right: prior and posterior distribution of θ; the grey lines indicate 'intermediate posteriors' obtained from earlier observations.*

compare their respective performances across 5000 sampled trajectories. A typical criteria of measuring the performance of the policy is to examine its regret, defined as

$$\mathrm{reg}(N, \pi) = J_0^{\theta^*}(\pi^*) - J^N(\pi),$$

where $J_0^{\theta^*}(\pi^*) = \mathbb{E}^{\pi^*}[\sum r(x, a)]$ is the reward functional with no observation cost, with known parameter θ^* and optimal policy π^*, and $J^N(\pi)$ is the reward functional (2.34) under a policy $\pi \in \Pi_{\mathrm{obs}}$. In this case we consider π to be the optimal policy under observation costs and parameter uncertainty. The regret is therefore the cumulative sum of the suboptimal gap from the optimal policy. For the observation cost problem, the control between observations are constant and therefore suboptimal in general, as such we do not expect the regret to achieve asymptotically sublinear growth. Instead, we consider the following alternative criteria:

$$\mathrm{reg}_{c_{\mathrm{obs}}}(N, \pi) = J_{c_{\mathrm{obs}}}^{\theta^*}(\pi^*) - J^N(\pi),$$

where here $J_{c_{\mathrm{obs}}}^{\theta^*}$ is the reward functional with known parameter θ^* and observation cost c_{obs}, so that $\mathrm{reg}_{c_{\mathrm{obs}}}$ measures the contribution of the regret that arises from parameter uncertainty. On the left of Figure 2.7, we show the overall regret for varying the observation cost for a fixed prior, and on the right, $\mathrm{reg}_{c_{\mathrm{obs}}}$ is plotted with a fixed observation cost of $c_{\mathrm{obs}} = 0.1$ across different initial priors ρ_0. As expected, the regret is generally higher when the prior estimate ρ_0 is less accurate, or when a larger c_{obs} value is used. Moreover the regret grows in a rather linear fashion. However, when examining the graph involving $\mathrm{reg}_{c_{\mathrm{obs}}}$ on the right side, we empirically observe sublinear growth. This can be interpreted as a gradual learning of the unknown

parameters, despite the fact that observations only arrive in intervals. The results suggests that $\text{reg}_{c_{\text{obs}}}$ can be used as an alternative notion to capture the learning rate in problems involving observation costs, which we see as a possible direction for future analysis.

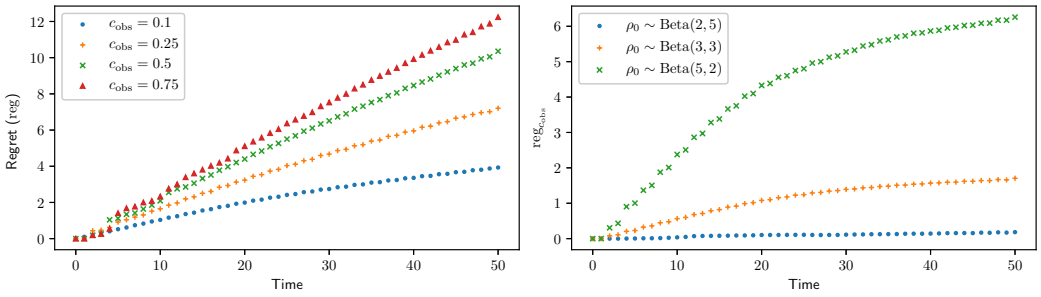

Figure 2.7: *Left: regret over time for* $\rho_0 \sim \text{Beta}(3,3)$ *for different values of* c_{obs}*. Right: the growth of* $\text{reg}_{c_{\text{obs}}}$ *for fixed* $c_{\text{obs}} = 0.1$ *and different initial priors* ρ_0*.*

To demonstrate the effects of observation cost and prior estimates on the number of observations, we consider an alternative reward function, given by

$$r(x,a) = r(x) = \begin{cases} 2 & x = 0 \\ -1 & x = \pm 2 \\ 0 & \text{otherwise} \end{cases}$$

In the absence of observation cost and parameter uncertainty, the controller aims to keep the process at the origin as often as possible, whilst avoiding the penalising boundary at $x \pm 2$. Table 2.8 lists the performance of the optimal policies under each combination of observation cost and prior estimate. As the true value of $\theta = 0.3$, a prior of $\rho_0 \sim \text{Beta}(2,5)$ acts a good estimate, and $\rho_0 \sim \text{Beta}(5,2)$ acts as a poor estimate. In general, we see that the value of the observation cost c_{obs} has a more dominating effect on the resulting optimal policies and rewards obtained, as seen in the big drop-off in the number of observations when $c_{\text{obs}} = 0.75$ in row (a), at which each observation comes at the cost of a significant proportion of the potential reward. Its effect on the sub-optimality is compounded with a bad prior estimate, with a negative reward and a 95% credible interval width of 0.5 in the extreme case in the bottom-right entry of Table 2.8.

2.4.3 Extension of an HIV-treatment model

In this subsection, we implement our formulation of the OCM to an HIV-treatment scheduling problem in [67]. There, the authors modelled the problem with a

		$c_{\text{obs}} = 0.1$	$c_{\text{obs}} = 0.25$	$c_{\text{obs}} = 0.5$	$c_{\text{obs}} = 0.75$
$\rho_0 \sim \text{Beta}(2,5)$	(a)	22.48	22.2	21.2	17.55
	(b)	20.622	17.15	11.26	6.0375
	(c)	0.2341	0.2360	0.2455	0.2844
$\rho_0 \sim \text{Beta}(3,3)$	(a)	21.4	20.97	18.36	11.27
	(b)	17.99	14.6475	8.92	2.5775
	(c)	0.2437	0.2459	0.2696	0.3624
$\rho_0 \sim \text{Beta}(5,2)$	(a)	19.22	17.3	11.21	3.34
	(b)	10.628	7.55	1.825	-0.835
	(c)	0.2488	0.2583	0.3302	0.5034

Table 2.8: *Numerical results for the parameter uncertainty problem. Line (a): average number of observations. Line (b): average profit ($N = 50$). Line (c): average credible interval width (HDI 95%).*

continuous-time MDP with observation costs, but does not include the time elapsed variable in dynamic programming. This can be interpreted as an implicit assumption that the observer is given the state of the underlying process at initialisation. We shall implement a discretised version of their model under our formulation with the time elapsed variable. As alluded to in the introduction, this allows in addition initial conditions that are outdated or sub-optimal relative to the objective. We demonstrate the qualitative difference in the optimal policies when varying the initial conditions, whilst replicating the results in the original paper when the initial conditions coincide. We also examine the numerical performance of the penalty method when applied to the system of QVIs for this larger system, compared to that in Section 2.4.1.

We now proceed to describe the original problem in [67]. A continuous-time MDP is used to model virus levels of HIV-positive patients over time. With two types of treatment available, the action space is $A = \{0, 1, 2\}$ (where 0 represents no treatment given). Four virus strains are considered: WT denotes the wild type (susceptible to both treatments), R1 and R2 denotes strains that are each resistant to Treatment 1 and Treatment 2 respectively, and HR denotes the strain that is highly resistant to both. The level of each strain is represented by the states 'none' (0), 'low' (l), 'medium' (m), and 'high' (h). Therefore, the state space for the Markov chain is $\mathcal{X} = \{0, l, m, h\}^4 \cup \{*\}$, where the asterisk represents patient death. Note in particular that $*$ is an absorbing state. The goal in the original model is to then minimise a cost functional $J : \mathcal{X} \times A \to \mathbb{R}$ of the form:

$$J(x, \alpha) = \mathbb{E}\left[\sum_{j=0}^{\infty}\left(\int_{\tau_j}^{\tau_{j+1}} e^{-\gamma s} c(X_s, \iota(X_{\tau_j}))\, ds + e^{-\gamma \tau_{j+1}} c_{\text{obs}}\right)\right], \tag{2.74}$$

where $\{\tau_j\}_{j=0}^{\infty}$ are the observation times, and the cost function $c : \mathcal{X} \times A \to \mathbb{R}$ is a linear combination of the productivity loss resulting from each patient's condition and their received treatment.

To adapt the model above for our formulation, we first discretise the MDP, taking each step to represent one day. We then take the model parameters from the original article [67, Section 3], which provides the transition rate matrices $\{Q_a\}_{i\in A}$ and the cost function $c(x, a)$. The transition matrices $\{P_a\}_{a\in A}$ are then given by $P_a = e^{Q_a}$ (as the time unit in [67] is one day). For illustration purposes, a sparse plot of the transition matrix P_0 is shown in Figure 2.9. As we are considering maximisation problems in this chapter, we take $r = -c$ for the reward function. We can now formulate our problem in terms of the following QVI:

$$\min\left\{v_{a,x}^n - \gamma v_{a,x}^{n+1} + \left(e^{nQ_a}c_a\right)_x\ ,\ v_{a,x}^n - \left(e^{nQ_a}\overline{\gamma v^1 + c}\right)_x + c_{\text{obs}}\right\} = 0. \tag{2.75}$$

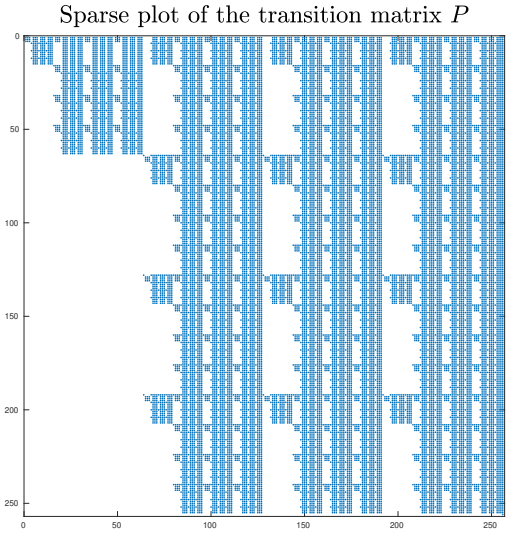

Figure 2.9: *Sparsity pattern of the transition matrix P_0 (the pattern is the same across all control states). The state space is encoded as $\{1, \ldots, 256\}$, by considering the state vectors [WT, R1, R2, HR] as a base-4 string in reverse order (for example, $[h, 0, l, l]$ corresponds to 83). The death state $*$ is represented by 256.*

We now follow the same procedure in Section 2.4.1 to obtain a numerical solution. Note that for this problem, the spatial domain is finite and we also have a natural spatial boundary arising from the absorbing death state $*$, that is, for all $n \geq 0$ and

$a \in A$,

$$v^n_{a,*} = \sum_{k=0}^{\infty} l\gamma^k = \frac{l}{1-\gamma}, \tag{2.76}$$

where l is a constant representing the average GDP loss due to patient death [66,67]. A time boundary is once again enforced at some large time $N > 0$, which can be interpreted as a mandatory observation at time N. Explicitly, this reads

$$v^N_{a,x} - \left(e^{NQ_a}\overline{\gamma v^1 + c}\right)_x + c_{\text{obs}} = 0, \quad x \in \mathcal{X} \setminus \{*\},\ a \in A. \tag{2.77}$$

We now solve the associated penalised problem with semismooth Newton methods. As in Section 2.4.1, we choose the initial guess to be the solution to the penalised problem with $\rho = 0$, with uncoupled time boundary conditions. The iterations terminate once a relative tolerance threshold of 10^{-8} is reached. The numerical experiments are performed on MATLAB R2019b.

Table 2.11 shows the numerical solution for different values of the truncation time N and c_{obs} across different penalty parameters ρ. Row (a) shows that the number of iterations remains constant with respect to ρ, much like the random walk experiment in Section 2.4.1. For this problem, the number of Newton iterations required to reach the $1e^{-8}$ threshold is higher at approximately 20 iterations. However, we find that convergence to the optimal policy is typically achieved within the first 2 iterations. This is depicted in Figure 2.10, which graphs the first two iterates as well as the final solution for the value function. Row (b) in Table 2.11 shows the successive increments $\|v^\rho - v^{2\rho}\|_\infty$ between doubling penalty values. Reassuringly, for this more complicated system, we still see a clear first-order convergence of the penalisation error with respect to the penalty parameter ρ. Even for small values of ρ, the successive increments were within $O(1)$ (in comparison to the magnitude of the solution which is of $O(10^6)$). This shows that the penalty approximation is very effective for small penalty parameters, and that it works well when extended to the class of QVIs that we introduced in Section 2.3.

We now analyse the behaviour of the value function when plotted as a function against time. The top-left graph of Figure 2.12 depicts an instance where the patient is under a stable condition. Here the observation region is $[15, N]$. There are limited benefits of frequently paying a high observation cost when it is unlikely that the patient's condition will deteriorate over a short period of time. On the other hand, the top-right graph has an observation region of $[0, 53]$. The mathematical intuition behind this is that beyond the observation region, the MDP is expected to enter the

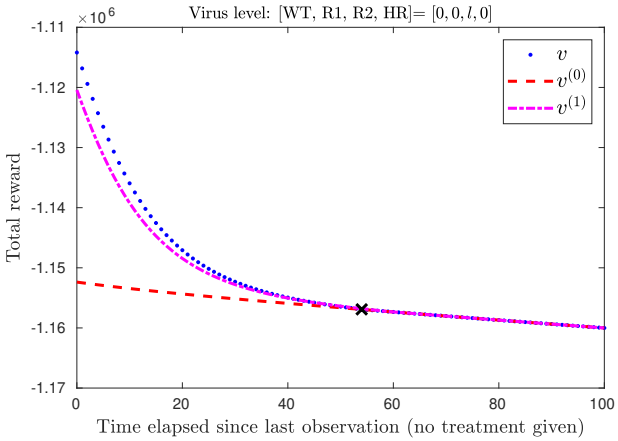

Figure 2.10: *Convergence of the Newton iterates towards the solution. The lines show the graphs of* $n \mapsto v_{0,4}^n$ *for the initial guess* $v^{(0)}$*, first iterate* $v^{(1)}$ *and true solution* v*, where the state* $[WT, R1, R2, HR] = [0, 0, l, 0]$ *is encoded as* 4 *in base* 4*. The cross indicates the boundary between the observation regions.*

	ρ	10^3	2×10^3	4×10^3	8×10^3	16×10^3	32×10^3
$N = 150,\ c_{\text{obs}} = 200$	(a)	18	18	18	18	18	18
	(b)	1.6141	0.8071	0.4036	0.2018	0.1009	0.0504
$N = 150,\ c_{\text{obs}} = 400$	(a)	21	21	21	21	21	21
	(b)	1.5147	0.7577	0.3790	0.1895	0.0948	0.0474
$N = 150,\ c_{\text{obs}} = 800$	(a)	20	20	20	20	20	20
	(b)	1.4087	0.7047	0.3524	0.1762	0.0881	0.0441
$N = 300,\ c_{\text{obs}} = 200$	(a)	20	20	20	20	20	20
	(b)	1.6122	0.8061	0.4031	0.2015	0.1008	0.0504
$N = 300,\ c_{\text{obs}} = 400$	(a)	19	19	19	19	19	19
	(b)	1.5131	0.7569	0.3785	0.1893	0.0947	0.0473
$N = 300,\ c_{\text{obs}} = 800$	(a)	20	20	20	20	20	20
	(b)	1.4102	0.7055	0.3528	0.1764	0.0882	0.0441
$N = 600,\ c_{\text{obs}} = 200$	(a)	19	19	19	19	19	19
	(b)	1.6111	0.8056	0.4028	0.2014	0.1007	0.0504
$N = 600,\ c_{\text{obs}} = 400$	(a)	17	17	17	17	17	17
	(b)	1.5114	0.7561	0.3781	0.1891	0.0945	0.0473
$N = 600,\ c_{\text{obs}} = 800$	(a)	18	18	18	18	18	18
	(b)	1.4065	0.7036	0.3519	0.1760	0.0880	0.0440

Table 2.11: *Numerical results for the HIV-treatment problem. Line (a): number of Newton iterations. Line (b): the increments* $\|v^\rho - v^{2\rho}\|$*.*

absorbing state $*$ with high probability, and the negative reward associated with this absorbing state outweighs any potential benefits of paying the observation cost c_{obs} for information. In the original model in [67], one determines the optimal policy based on an immediate observation in hand. Putting this in the context of our formulation,

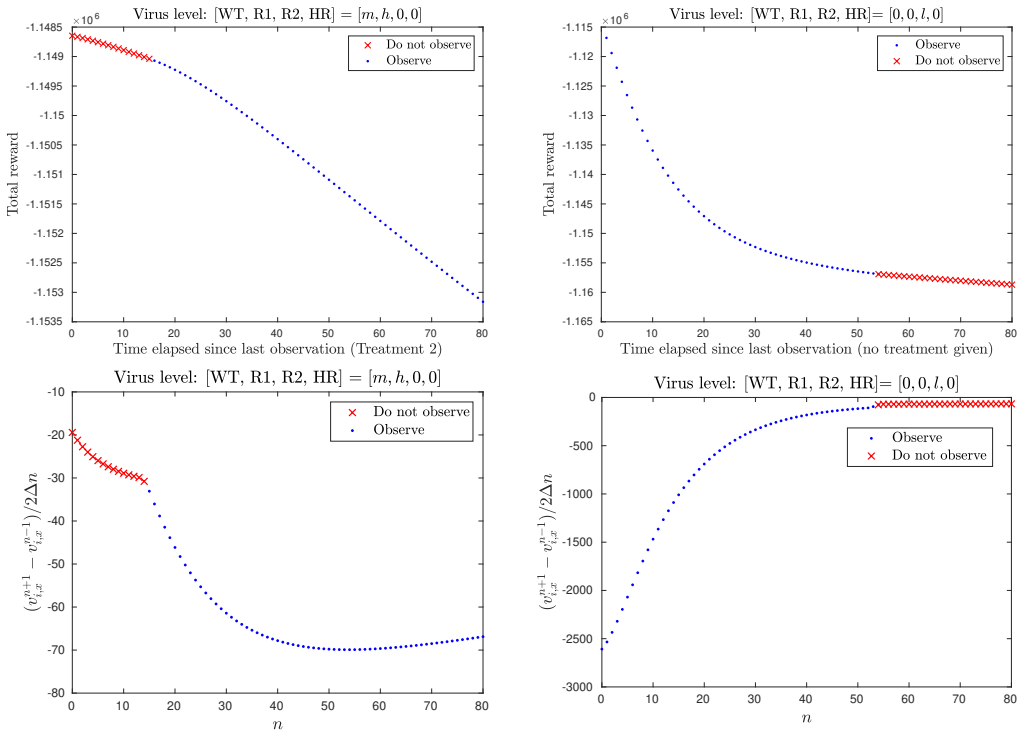

Figure 2.12: *The value function exhibits two qualitatively different decay modes depending on the starting states x. Left: a stable condition with the correct treatment. Right: a worse condition with no treatment. The top row shows the mappings $n \mapsto v^n_{i,x}$. The bottom row plots the corresponding central finite difference terms.*

this amounts to fixing an initial condition in the form of $y = (1, x, a) \in \mathcal{Y}$, and looking forward ahead in time to find the first observation time. This overlooks situations such as that occurring in the top-right graph of Figure 2.12: by initialising at the time origin, one immediately 'loops back' for an immediate observation, and therefore does not see the effect of the passage of time on the optimal observation policy.

To examine the behaviour around the decision boundaries, we plot the central finite difference terms $(v^{n+1}_{a,x} - v^{n-1}_{a,x})/2\Delta n$ in the bottom row of Figure 2.12, underneath their respective graphs of the value function. If we consider the plots as a discretisation of a continuous value function, we see that there is much bigger variation within the observation region. Critically, there is non-smoothness across the boundary in the bottom-left graph. This suggests that the solution in continuous-time is C^2 in time within each decision region, but only C^1 across the boundary. This is in line with theoretical results on the regularity of viscosity solutions in optimal stopping and switching problems [54, Chapter 5], which is a potential direction for future analysis.

Chapter 3

Mean-field games of speedy information access with observation costs

3.1 Introduction

In decision making, one often has the opportunity to improve the quality of one's observations by expending extra resources. For example, medical laboratories can invest in infrastructure to reduce waiting times for testing results, to enable faster diagnosis and treatment for patients. Balancing such trade-off between information acquisition and the associated costs may be as important as selecting the course of further actions which optimise one's rewards.

We introduce a novel mean-field game (MFG) model in discrete-time, in which agents actively control their speed of access to information. The MFG considered can be viewed as a partial observation problem, in which the information stream is not exogeneously given but rather dynamically controlled by the agents. In the game, agents can adjust their speed of information access with suitable costly efforts, and exploit their dynamic information stream to inform the choice of controls for the state dynamics so as to maximise their rewards. We utilise the information structure to construct a suitable augmentation of the state space, which includes the belief state as well as past actions taken within the dynamic delay period, that serves as the finite state space of an equivalent mean field game of standard form. Thereby, numerical schemes for discrete MFGs can be employed to compute approximate mean-field Nash equilibria (MFNE) for our MFG of speedy information access with controlled observations.

This chapter covers three themes: (1) actively controlled observation delay, (2) observation costs, and (3) the analysis of an associated MFG incorporating the combination of those two features. Standard Markov decision process (MDP) frameworks assume that state observations are received instantaneously, with corresponding actions in response being also applied instantaneously. This limits the applicability of such models in many real-life situations. It is often the case that observation delay arises due to inherent features of a system, or practical limitations from data collection. For example, the times to receive medical diagnosis test results depend on the processing time required for laboratory analysis. In high-frequency trading, observation delay occurs in the form of latency, aggregated over the multiple stages of communication with the exchange [20].

There has been a large amount of literature involving the modelling of observation delays, with applications in (but not limited to) network communications [2,3], quantitative finance [18,20,50] and reinforcement learning [21,48,63]. Most models involve an MDP framework with either a constant or random observation delay, both of which are exogenously given by the system. Both constant and random observation delay MDPs can be modelled as a partially observable MDP (POMDP) via state augmentation [4,11,41,48]. It has also been shown that action delays can be considered as a form of observation delay, under a suitable transformation of the MDP problem [41]. The continuous-time counterpart with an associated HJB-type master equation has been studied in [62].

In many formulations of optimisation problems in MDPs, the information source is fixed *a priori*. However, it is often desirable to control the observations that one receives, in addition to the dynamics of the underlying process. This frequently occurs in resource-constrained environments where frequent measurements or sampling are either too expensive or impractical. Applications include efficient medical treatment allocation [67], environmental management [72–75], communications sampling [29,33], optimal sensing [49, 64, 70], reinforcement learning [16, 17, 42], and much more. We shall refer to these as observation cost models (OCMs). In OCM problems, the user can opt to receive an observation of the current state of the process, at the price of an observation cost which is included in the reward functional to be optimised.

The OCM can equivalently be characterised as a POMDP, by including the time elapsed, together with the last observed states and actions applied to form an augmented Markov system. In many cases, a reasonable simplification is to assume constant actions between observations [37, 55]. This leads to a finite dimensional

characterisation of the augmented state, and allows efficient computation of the resulting system of quasi-variational inequalities via a penalty scheme [55]. Analysis for the more general non-constant action case has generally been restricted to the linear-quadratic Gaussian case [22, 64, 70].

In stochastic games, the computation of Nash equilibria is often intractable for large number of players. Mean-field games (MFGs), first introduced in [45] and [19], provide a way of seeking approximate Nash equilibria, by assuming symmetric interactions between agents that can be modelled by a mean-field term, in the form of a measure flow. MFGs can be treated as an asymptotic approximation of a game with large number of interacting players. Finding a mean-field Nash equilibrium (MFNE) amounts to a search for an optimal policy for a representative player, and ensuring that the state distribution of said player under such a policy is consistent with the postulated law of the other players, given by the measure flow. In discrete time, the existence of MFNE has been established in [59]. Analysis has also appeared for several model variants such as risk-sensitive criteria [61], partially observable systems [60, 61] and unknown reward/transition dynamics [31].

In general, finite MFGs suffer from non-uniqueness of MFNE and non-contractivity of the naively iterated fixed point algorithm [23]. Several algorithms have emerged to address the efficient computation of MFNEs. Entropy regularisation exploits the duality between convexity and smoothness to achieve contractivity, by either incorporating the entropy term directly into the reward functional, or imposing softmax policies during the optimisation step [6, 23, 27]. Fictitious play schemes aim to smooth the mean-field updates by averaging new iterates over the past mean-field terms, effectively damping the update to aid numerical convergence [53]. Online mirror descent further decreases computational complexity by replacing best response updates with direct Q-function computations [52]. In contrast, [32] reformulates the problem of searching an MFNE to an equivalent optimisation problem, allowing a possible search for multiple MFNE with standard gradient descent algorithms. We refer to the survey [46] for a comprehensive overview of the above algorithms.

Our work. We model agents' strategic choices for speed of information access in the game, by studying a novel MFG where the speed of access is in itself also a part of the costly control. Throughout this chapter, we assume that both the state and action spaces are finite. The agents participating in the game have control over two aspects: the time period of their observation delay, and their actions that influence

their rewards and transition dynamics. The agent can choose over a given finite set of delay periods, with each value being associated to an observation cost. A higher observation cost corresponds to a shorter delay period, and vice versa.

Our framework here differs from existing works, in that the delay period is not exogeneously given as in the constant case [4, 11], nor is it a random variable with given dynamics as in the stochastic case [20, 21, 48]. Instead, the length of the delay is dynamically and actively decided by the agent, based on the trade-off between the extra cost versus the accuracy of more speedy observations, the latter of which can be exploited though better informed control of the dynamics and hence higher rewards. The choice of the delay period becomes an extra part of the control in the optimisation problem in tandem with the agent's actions. When considering this as a single agent problem, which occurs during the optimisation step when the measure flow is fixed, we refer to it as a Markov Controllable Delay model (MCDM). The MCDM can be reformulated in terms of a POMDP, by augmenting the state with the most recent observation and actions taken since, to form a Markovian system. This allows the formulation of dynamic programming to obtain the Bellman equation.

When viewed as part of the overall MFG, the partial information structure of the problem implies that the measure flow should be specified on the augmented space for the fixed point characterisation of the mean field Nash equilibira (MFNE). However, the underlying transition dynamics and reward structure would depend on the distribution of the states at the present time. In the models of [60, 61], the mapping from measures on the augmented space to measures on the underlying state is given by taking the barycenter of the measure. However, our model here differs in two aspects. Firstly, although the belief state is an element of the simplex on the underlying state space, we find a finite parameter description (the state last observed and actions taken thereafter) to establish the MFG on a finitely augmented space. Secondly, due to the delayed structure, the observation kernel depends on the distribution of the states throughout each moment in time across the delay period. Thus, taking an average of a distribution over the augmented space of parameters, as a barycenter map would do, is not applicable here. Instead, we explicitly map a measure flow on the augmented space to a sequence of measures on the underlying states. Intuitively, this corresponds to an agent estimating the distribution of the current states of the population, given the observations he/she possesses (i.e., the distribution of the delay period amongst agents, and the states and actions given such a delay). We detail

the construction of the MCDM in Section 3.2 and the corresponding MFG formulation, which we will also refer to as the MFG-MCDM, with its MFNE definition in Section 3.3.

The second part of this chapter focuses on the computation of an MFNE for the MFG of control of information speed. We employ the popular entropy regularisation technique, which aids convergence of the classical iterative scheme: computing an optimal policy for a fixed measure flow, followed by computing the law of the player under said policy. In the standard MFG model, it is shown that the fixed point operator for the regularised problem is contractive under mild conditions [6, 23]. This forms the basis of the prior descent algorithm, which is one of the current state-of-the-art algorithms for the computation of approximate Nash equilibria for MFGs [23, 30]. We prove that for our MFG model of control of information speed, the corresponding fixed point operator also converges when it is sufficiently regularised by an entropy term. This can be summarised in the following theorem, which is a condensed version of Theorem 3.28.

Theorem 3.1. *Let Φ_η^{reg} be the regularised best-response map, with regulariser parameter η, and let Ψ^{aug} be the measure-flow map. Then for $\eta > c_\eta$, there exists a unique fixed point for $\Psi^{\text{aug}} \circ \Phi_\eta^{\text{reg}}$, i.e. there exists a unique regularised MFNE for the MFG-MCDM problem. Here c_η is a constant that only depends on the Lipschitz constants and bounds of the transition kernels and rewards functions.*

We defer the precise definitions of the operators Φ_η^{reg} and Ψ^{aug}, as well as the constant c_η to Section 3.4. We investigate the infinite horizon discounted cost problem with time-dependent measure flows. This extends the result in [23] for finite horizon problems, and the result in [6] for infinite horizon problems with stationary measure flows. As the MFG-MCDM is a partially observable problem, the proof also requires a crucial extra step to demonstrate that the aforementioned mapping of the measure flow on the augmented space to that on the underlying space is Lipschitz, in other to prove the required contraction.

The contributions of this chapter can be summarised as follows.

1. We show dynamical programming for a Markov Controllable Delay model (MCDM), an MDP model where an individual agent can exercise dynamic control over the latency of their observations, with less information delay being more costly. The problem is cast in terms of a partially observed MDP

(POMDP) with controlled but costly partial observations, for which the belief state can be described by a finite parametrization. Solving this POMDP is shown to be equivalent to solving a finite MDP on an augmented finite state space, whose extension also involves past actions taken during the (non-constant but dynamically controlled) delay period.

2. We introduce a corresponding Mean Field Game (MFG) where speedy information access is subject to the agents' strategic control decisions. For a fixed measure flow, which describes the statistical population evolution, the ensuing single agent control problem becomes an MCDM. Although a mean-field Nash equilibrium (MFNE) is defined in terms of the augmented space, the underlying dynamics and rewards still depend on the underlying state distribution. We show how a measure flow on the underlying space is determined and computed from that of the augmented space. This construction exploits the finite parameterization of the belief state; whereas the barycenter approach for belief state which are measure-valued as in [60] does not apply here.

3. By using a sufficiently strong entropy regularisation in the reward functional, we prove that the regularised MFG-MCDM has a unique MFNE which is described by a fixed point, and can serve as an approximate Nash equilibrium for a large but finite population size. The characterisation of the MCDM as a finite MDP enables to compute the Nash equilibrium of the corresponding MFG, by using methods from [6, 23]. The results also extend to a MFG formulated on infinite horizon with time-dependent measure flows.

4. We demonstrate our model by an epidemiology example, in which we compute both qualitative effects of information delay and cost to the equilibrium, and also the quantitative properties of convergence relating to the entropy regularisation. For computation, we employ the Prior Descent algorithm [23], applying the new `mfglib` Python package [30] to our partially observable model.

The remainder of the chapter is organized as follows. Section 3.2 develops the formulation of the MCDM as a POMDP and establishes dynamic programming on the augmented space. Section 3.3 explains the corresponding MFG setup, the fixed point characterisation of the equilibrium, and shows some basic properties. Section 3.4 establishes the contraction property of the fixed point iteration map for the regularised mean field game and its ability to yield an approximate equilibrium for the finite

player game. Finally, Section 3.5 demonstrates a numerical example from epidemiology for illustrtration.

3.1.1 Notation and preliminaries

For any finite set E, we identify the space of probability measures on E with the simplex Δ_E. We equip Δ_E with the metric δ_{TV} induced by the total variation norm on the space of signed measures. That is,

$$\delta_{TV}(p, \hat{p}) = \|p - \hat{p}\|_{TV} = \frac{1}{2}\sum_{e \in E} |p(e) - \hat{p}(e)|, \qquad p, \hat{p} \in \Delta_E.$$

We will generally be considering Markovian policies in this chapter. A Markovian policy $\pi = (\pi_t)_t$ is then a sequence of maps $\pi_t : E \to \Delta_{E'}$, mapping a finite set E to the simplex on another finite set E'. Since a policy is bounded, we equip it with the sup norm

$$\delta_{\Pi}(\pi, \hat{\pi}) = \sup_{t \geq 0} \max_{e \in E} \delta_{TV}(\pi_t(\cdot \mid e), \hat{\pi}_t(\cdot \mid e)).$$

Let Δ_E^T denote the space of measure flows on E, with $T \in \mathbb{N} \cup \{\infty\}$. If T is finite, we equip Δ_E^T with the sup metric

$$\delta_{\max}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \max_{0 \leq t \leq T} \delta_{TV}(\mu_t, \hat{\mu}_t), \quad \boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \Delta_E^T.$$

If $T = \infty$, we instead use the metric

$$\delta_{\infty}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{t=1}^{\infty} \zeta^{-t} \delta_{TV}(\mu_t, \hat{\mu}_t), \quad \boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \Delta_E^{\infty}, \tag{3.1}$$

where $\zeta > 1$. Note that the choice of ζ is not canonical, and as long as $\zeta > 1$, δ_∞ induces the product topology on Δ_E^∞, which is compact by Tychonoff's theorem, since each individual simplex Δ_E is also compact. Hence $(\Delta_E^\infty, \delta_\infty)$ is a complete metric space. This allows us to appeal to Banach's fixed point theorem when considering the contraction mapping arguments later.

We will often consider a sequence of actions taken, e.g. $a_1, \ldots, a_n \in A$. In these cases we will use the shorthand notation $(a)_1^n = (a_1, \ldots, a_n)$. We will use both notations interchangebly throughout the rest of this chapter.

We will frequently make use of the following proposition in our analysis.

Proposition 3.2 ([28, p.141]). *For any real valued function F on a finite set E, given $\nu, \nu' \in \Delta_E$ we have the inequality (see also proof of [6, Proposition 1])*

$$\left| \sum_{e \in E} F(e)\nu(e) - \sum_{e \in E} F(e)\nu'(e) \right| \leq \lambda(F)\ \delta_{TV}(\nu, \nu'),$$

where $\lambda(F) \coloneqq \max_{e \in E} F(e) - \min_{e \in E} F(e)$.

3.2 MDPs with controllable information delay

We first state the definition of a Markov controllable delay model (MCDM) below, which characterises the scenarios where agents can control their information delay.

Definition 3.3. A Markov controllable delay model (MCDM) is a tuple $\langle \mathcal{X}, A, \mathcal{D}, \mathcal{C}, p, r \rangle$, where

- $\mathcal{X}$ is the finite *state space*;
- A is the finite *action space*;
- $\mathcal{D} = \{d_0, \ldots, d_K\}$, with $0 \leq d_K < \ldots < d_0$, for some given value K, is the set of *delay values*;
- $\mathcal{C} = \{c_0, c_1, \ldots, c_K\}$, with $0 = c_0 < c_1 \ldots < c_K$, is the set of *cost values*;
- $p : \mathcal{X} \times A \to \Delta_{\mathcal{X}}$ is the *transition kernel*;
- $r : \mathcal{X} \times A \to [0, \infty)$ is the *one-step reward function.*

Let us also denote the n-step transition probabilities by $p^{(n)}(\ \cdot \mid x, (a)_1^n)$, where we use the notation $(a)_1^n \coloneqq (a_1, \ldots, a_n) \in A^n$. For a given set of delay values $\mathcal{D}$, define also

- $\overline{\mathcal{D}} \coloneqq [d_K, d_0] = \{d_K, d_K + 1, \ldots, d_0\}$.
- The intervention variables $(i_t)_t$, taking values in the *intervention set* $\mathcal{I} \coloneqq \{0, 1, \ldots, K\}$.

$\mathcal{D}$ represents the delay values that an agent can choose from, with $(i_t)_t$ representing the decision on the choice of delay, and $\overline{\mathcal{D}}$ represents the range of delay values of the system at any given point in time. A value of $i_n = k$ indicates that at time n the agent wishes to pay a cost of c_k to change their delay to d_k units. To ensure that the setup is well-defined, if $i_n = k$ and the current delay d is shorter than d_k, then the delay at time $n + 1$ will simply be extended to $d + 1$ units (in reality, paying a

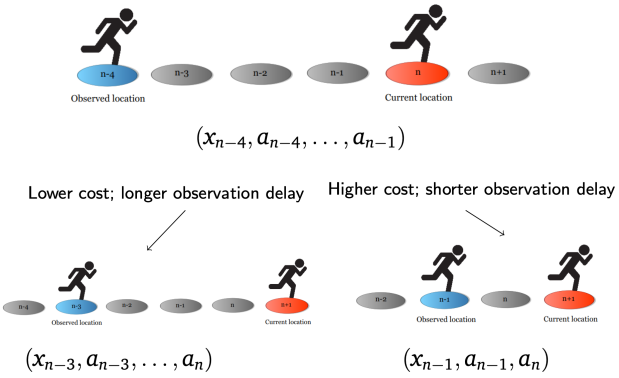

Figure 3.1: *Control of information speed*

higher cost for a longer delay is clearly sub-optimal, so such a choice of i_n would not practically occur).

Formally, the MCDM evolves sequentially as follows. Suppose at time t, the controller observes the underlying state $x_{t-d_0} \in \mathcal{X}$, with knowledge of their actions $a_{t-d_0}, \ldots, a_{t-1} \in A$ applied since. Based on this information, the controller applies an action a_t and receives a reward $r(x_t, a_t)$ (which we assume not to be observable until x_t becomes observable). The controller then decides on the choice of cost c_i, which determines their next delay period of d_i units, i.e. observing x_{t+1-d_i} at time $t+1$. This process then repeats at the next time. If no cost is paid, then no new observations occur, until the delay reaches d_0 units again. Figure 3.1 depicts a typical evolution of an MCDM.

The precise construction can be set up as follows. We assume that the problem initiates at time $t = 0$, and denote prior observations with negative indices.

Definition 3.4. Define the *history sets* $\{H_t\}_{t\geq 0}$ by

$$H_t \coloneqq H_{t-1} \times A \times \mathcal{I} \times \mathcal{X}, \quad t \geq 1.$$

where $H_0 \coloneqq \overline{\mathcal{D}} \times (\mathcal{X} \times A)^{d_0} \times \mathcal{X}$, denoting its elements in the form

$$h_0 = (d, x_{-d_0}, a_{-d_0}, \ldots, x_{-1}, a_{-1}, x_0).$$

A *policy* $\pi = (\pi_t)_{t\geq 0}$ is a sequence of kernels $\pi_t : H_t \to \Delta_{A\times\mathcal{I}}$. The corresponding canonical sample space is then

$$\Omega \coloneqq H_\infty \coloneqq H_0 \times (A \times \mathcal{I} \times \mathcal{X})^\infty.$$

Given an initial distribution $q_0 \in \Delta_{H_0}$ and a policy π, the Ionescu-Tulcea theorem [34, Appendix C] gives a unique probability measure $\mathbb{P}^{\pi}_{q_0}$ such that for any

$$\omega = (d, x_{-d_0}, a_{-d_0}, \ldots, x_0, a_0, i_0, x_1, a_1, i_1, \ldots),$$

$\mathbb{P}^{\pi}_{q_0}$ satisfies

$$\begin{aligned}\mathbb{P}^{\pi}_{q_0}(\omega) = q_0(h_0)\ p(x_{-d_0+1} \mid x_{-d_0}, a_{-d_0}) \ldots p(x_0 \mid x_{-1}, a_{-1}) \\ \pi_0(a_0, i_0 \mid d, x_{-d_0}, a_{-d_0}, \ldots, x_0)\ p(x_1 \mid x_0, a_0) \ldots\end{aligned}$$

The value d appearing in H_0 represents the initial delay period. Given a history sequence $h_t \in H_t$, subsequent delay periods at time $t > 0$ can be deduced from the values of d and $(i_n)_{n=0}^{t-1}$. Denote this value by $d(t) \in \overline{\mathcal{D}}$. This leads to the following definition for the set of admissible policies.

Definition 3.5. A policy π is admissible for an MCDM if at each time t, there exists a sequence of kernels $\phi_t^d : \mathcal{X} \times A^d \to \Delta_{A \times \mathcal{I}}$, $d \in \overline{\mathcal{D}}$, such that for each $h_t \in H_t$,

$$\pi_t(\cdot \mid h_t) = \phi_t^{d(t)}(\cdot \mid x_{-d_0}, \ldots, x_{t-d_{h_t}}, a_{-d_0}, \ldots, a_{t-1}),$$

where $d(t) \in \overline{\mathcal{D}}$ is the delay period at time t for a corresponding history sequence $h_t \in H_t$. The set of admissible policies for the MCDM is denoted by Π_{DM}.

Given the MCDM $\langle \mathcal{X}, A, \mathcal{D}, \mathcal{C}, p, r \rangle$ and an admissible policy $\pi \in \Pi_{DM}$, the objective function for the infinite horizon problem with discounted cost is

$$J(\pi) \coloneqq \mathbb{E}^{\pi}_{q_0}\left[\sum_{n=0}^{\infty} \gamma^n \left(r(x_n, a_n) - \sum_{k=1}^{K} c_k \mathbb{1}_{\{i_n = k\}} \right)\right],$$

where $\mathbb{E}^{\pi}_{q_0}$ is the expectation over the measure $\mathbb{P}^{\pi}_{q_0}$, and $\gamma \in (0, 1)$ is the discount factor.

The search for an optimal $\pi \in \Pi_{DM}$ can be solved by considering an equivalent MDP on an augmented state, which contains all the information that occurred between the current time and the delayed time. As noted in [4] for the constant delay case, the lifting is akin to the classical POMDP approach on constructing an equivalent MDP on the belief state, but in this case the 'observations' do not come from an exogenous information stream, but from a past occurred state instead.

For a full Markovian system, the augmented variable will include the delay of the system at the current time, the underlying state that is observed with that delay, and the actions applied from that moment until the present. This will be presented as the following.

Definition 3.6. Given the delay values $\mathcal{D} = \{d_0, \dots, d_K\}$, define the **augmented space** $\mathcal{Y}$ by

$$\mathcal{Y} \coloneqq \bigcup_{d=d_K}^{d_0} \{d\} \times \mathcal{X} \times A^d.$$

Then an element $y \in \mathcal{Y}$ can be written in the form

$$(d, x, a_{-d}, \dots, a_{-1}), \text{ or } (d, x, (a)_{-d}^{-1}),$$

where negative indices are used to indicate that the actions had occurred in the past. If specific indices are not required, we will also use the notation $y = (d, x, \boldsymbol{a})$.

Although the length of the delay is implicit from the number of elements in $\boldsymbol{a}$, we explicitly include d in $\mathcal{Y}$ for simpler comprehension.

Remark 3.7. *As the length of the delay is variable and dependent on the control, the dimension of the augmented state is also variable. In practice, during computation, we can keep the dimensions consistent by introducing dummy variables $\varnothing$. This follows the treatment of stochastic delays in [48]. Specifically. for any set E we write $E_\varnothing = E \cup \{\varnothing\}$. Then an element $y \in \mathcal{Y}$ can be embedded into the space $\mathcal{X} \times A_\varnothing^{d_0}$ via the mapping*

$$(d, x, (a)_{-d}^{-1}) \mapsto (x, (a)_{-d}^{-1}, \underbrace{\varnothing, \dots, \varnothing}_{d_0 - d \text{ units}}).$$

We can now construct the MDP on the augmented space. For $y = (d, x, \boldsymbol{a})$, $\hat{y} = (\hat{d}, \hat{x}, \hat{\boldsymbol{a}}) \in \mathcal{Y}$, $a' \in A$, and $i \in \mathcal{I}$, let $p_y : \mathcal{Y} \times (A \times \mathcal{I}) \to \Delta_{\mathcal{Y}}$ be the augmented kernel, where

$$p_y(\hat{y} \mid y, a', i) = \begin{cases} p^{(d-d_i+1)}\big(\hat{x} \mid x, (a)_{-d}^{-d_i}\big)\ \mathbb{1}_{\left\{\hat{d}=d_i,\ \hat{\boldsymbol{a}}=\left((a)_{-d_i+1}^{-1}, a'\right)\right\}}, & d_i \le d \le d_0; \\ \mathbb{1}_{\left\{\hat{x}=x,\ \hat{d}=d+1,\ \hat{\boldsymbol{a}}=\left((a)_{-d}^{-1}, a'\right)\right\}}, & d_K \le d < d_i. \end{cases} \tag{3.2}$$

Let Π' denote the set of policies for this augmented MDP. That is, $\pi' = (\pi'_t) \in \Pi'$ is such that $\pi'_t : H'_t \to \Delta_{A \times \mathcal{I}}$, where $H'_0 \coloneqq \mathcal{Y}$ and $H'_t \coloneqq H'_{t-1} \times A \times \mathcal{I} \times \mathcal{Y}$ for $t \ge 1$. By the Ionescu-Tulcea theorem again, for an initial distribution $q_{y_0} \in \Delta_{\mathcal{Y}}$ and a policy $\pi' \in \Pi'$, there exists a unique probability measure $\mathbb{P}^{\pi'}_{q_{y_0}}$ such that

$$\mathbb{P}^{\pi'}_{q_{y_0}}(y_0, a_0, i_0, y_1, \dots) = q_{y_0}(y_0)\ \pi'_0(a_0, i_0 \mid y_0)\ p_y(y_1 \mid y_0, a_0, i_0) \dots \tag{3.3}$$

It is then straightforward to see that there is a one-to-one correspondence between policies in the original MCDM and policies in the augmented MDP. This follows analogously from the case of a fixed information delay [4], and we summarise the argument here: each $h_t \in H_t$ can be mapped to a corresponding $h'_t \in H'_t$ via

$$(d, x_{-d_0}, a_{-d_0}, \dots, x_0, a_0, i_0, \dots, x_{t-1}, a_{t-1}, i_{t-1}, x_t) \mapsto (y_0, a_0, i_0, \dots, y_{t-1}, a_{t-1}, i_{t-1}, y_t),$$

where $y_t = \left(d(t), x_{t-d(t)}, (a)_{t-d(t)}^{t-1}\right)$ for $t \geq 0$. Then, given a policy $\pi \in \Pi_{DM}$, one can define a policy $\pi' \in \Pi'$ via

$$\pi'_t(\cdot \mid h'_t) = \pi_t(\cdot \mid h_t), \quad t \geq 0.$$

Moreover, the policies π and π' assign the same joint law to $(y_t, a_t, i_t)_{t\geq 0}$ (when viewed as the canonical coordinate projection). One can then consider the objective function in the augmented space $\mathcal{Y}$, which is now a fully observable problem:

$$J'(\pi') := \mathbb{E}^{\pi'}_{q_{y_0}}\left[\sum_{n=0}^{\infty} \gamma^n r_y(y_n, a_n, i_n)\right], \quad \pi' \in \Pi',$$

where $r_y : \mathcal{Y} \times (A \times \mathcal{I}) \to [0, \infty)$ and

$$r_y(y, a', i) = \sum_{x' \in \mathcal{X}} r(x', a') p^{(d)}(x' \mid x, \boldsymbol{a}) - \sum_{k=1}^{K} c_k \mathbb{1}_{\{i=k\}}, \quad y = (x, d, \boldsymbol{a}). \tag{3.4}$$

The two problems are equivalent in that $\pi_* \in \Pi_{DM}$ is optimal for J if and only if $\pi'_* \in \Pi'$ is optimal for J', and it holds that

$$\sup_{\pi \in \Pi_{DM}} J(\pi) = J(\pi_*) = J'(\pi'_*) = \sup_{\pi' \in \Pi'} J'(\pi')$$

Given the equivalence, we shall use Π_{DM} to represent the set of admissible policies without loss of generality. This allows us to establish dynamic programming for the MCDM as follows.

Proposition 3.8. *Let $v : \mathcal{Y} \to \mathbb{R}$ be the value function*

$$v(y) = \sup_{\pi \in \Pi_{DM}} \mathbb{E}^{\pi}\left[\sum_{n=0}^{\infty} \gamma^n r_y(y_n, a_n, i_n) \middle| y_0 = y\right],$$

Then v satisfies the dynamic programming equation

$$v(y) = \max_{(a', i) \in A \times \mathcal{I}} \left\{ \sum_{x' \in \mathcal{X}} r(x', a') p^{(d)}(x' \mid x, \boldsymbol{a}) - \sum_{k=1}^{K} c_k \mathbb{1}_{\{i=k\}} + \gamma \sum_{y' \in \mathcal{Y}} p_y(y' \mid y, a', i) v(y') \right\},$$

where p_y is the augmented kernel as in (3.2)*. Moreover, the optimal policy is given in feedback form, so that $\pi_{*,n} = \phi_n(y_n)$ for some feedback function ϕ_n.*

Proof. This is a standard application of dynamic programming for a fully observable MDP, see e.g. [35, Theorem 4.2.3]. □

Remark 3.9. *When considering the MFG in the next section, a deterministic measure flow representing the population distribution introduces an implicit time dependence within the transition kernel and reward. The generic single agent problem in the definition of the MFNE then becomes time-inhomogeneous. The time-homogeneous setup in this section readily generalises directly to a setup with time-inhomogeneous transition kernels, rewards and dynamic programming equations. However, for ease of exposition we choose to present the MCDM under the time-homogeneous setting here.*

3.3 MFG formulation with control of information speed

To ease notation, in the remainder of this chapter we write $U \coloneqq A \times \mathcal{I}$ and $u = (a, i) \in U$.

3.3.1 Finite agent game with observation delay

Consider an N-player game with mean-field interaction, where each agent can control their observation delay. We shall start with incorporating the measure dependence into the MCDM in Definition 3.3.

Definition 3.10. An MCDM with measure dependence is a tuple $\langle \mathcal{X}, A, \mathcal{D}, \mathcal{C}, p, r \rangle$, where

- $\mathcal{X}$ is the finite *state space*;
- A is the finite *action space*;
- $\mathcal{D} = \{d_0, \ldots, d_K\}$, with $0 \leq d_K < \ldots < d_0$, is the set of *delay values*;
- $\mathcal{C} = \{c_1, \ldots, c_K\}$, with $0 = c_0 < c_1 < \ldots < c_K$, is the set of *cost values*;
- $p : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{X}}$ is the *transition kernel*;
- $r : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to [0, \infty)$ is the *one-step reward function.*

Denote by $x_t^j \in \mathcal{X}$ the state of the j-th player at time t, and $a_t^j \in A$ the corresponding action. Assume that the mean-field interaction occurs in the reward and the transition probabilities of the players, and is identically distributed for each player. The

transition kernel is given by $p : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to \Delta_{\mathcal{X}}$ so that the j-th player moves from state x_t^j to x_{t+1}^j with probability

$$p(x_{t+1}^j \mid x_t^j, a_t^j, e_t^N), \quad e_t^N(\cdot) = \frac{1}{N}\sum_{k=1}^{N} \delta_{x_t^k}(\cdot).$$

Here e_t^N is the empirical distribution of the states of the agents. Similarly, the one-step reward function is given by $r : \mathcal{X} \times A \times \Delta_{\mathcal{X}} \to [0, \infty)$ so that player j receives a reward of $r(x_t^j, a_t^j, e_t^N)$ at time t.

Recall that for a fully Markovian system, we consider the lifted problem in the augmented space

$$\mathcal{Y} \coloneqq \bigcup_{d=d_K}^{d_0} \{d\} \times \mathcal{X} \times A^d.$$

For the N-player model, consider the history sets $H_0 \coloneqq \mathcal{Y} \times \Delta_{\mathcal{Y}}$ and $H_t \coloneqq H_{t-1} \times U \times \mathcal{Y} \times \Delta_{\mathcal{Y}}$ for $t \geq 1$. A policy $\pi = (\pi_t)_{t \geq 0}$ is a sequence of maps $\pi_t : H_t \to \Delta_U$.

Definition 3.11. A policy π is admissible for the N-player MCDM if at each time t, there exists a sequence of kernels $\phi_t^d : \mathcal{X} \times A^d \times \Delta_{\mathcal{Y}}^t \to \Delta_{A \times \mathcal{I}}$, $d \in \overline{\mathcal{D}}$, such that for each $h_t \in H_t$,

$$\pi_t(\cdot \mid h_t) = \phi_t^{d(t)}(\cdot \mid x_{-d_0}, \ldots, x_{t-d_{h_t}}, a_{-d_0}, \ldots, a_{t-1}, \tilde{e}_0^N, \ldots, \tilde{e}_t^N),$$

where $d(t) \in \overline{\mathcal{D}}$ is the delay period at time t for a corresponding history sequence $h_t \in H_t$, and $\tilde{e}_t^N$ is the empirical distribution of augmented state, i.e.

$$\tilde{e}_t^N = \frac{1}{N}\sum_{k=1}^{N} \delta_{y_t^k}(\cdot), \quad y_t^k = (d^k(t), x_{t-d^k(t)}, a_{t-d^k(t)}, \ldots, a_{t-1}).$$

The set of admissible policies for player j is denoted by Π^j.

Let $\Pi = \prod_{j=1}^N \Pi^j$. Player j's objective function is given by

$$J_j^N(\pi^{(N)}) = \mathbb{E}^{\pi^{(N)}}\left[\sum_{n=0}^{\infty} \gamma^n r(x_n^j, a_n^j, e_n^N)\right]$$

where $\pi^{(N)} = (\pi^1, \ldots, \pi^N) \in \Pi$. The notion of optimality in the N-player game can be captured by the Nash equilibrium, which intuitively says at equilibrium, no player can make gains by deviating from their current strategy, provided that all other players remain at their strategy.

Definition 3.12 (Nash equilibrium)**.** $\pi_*^{(N)} \in \Pi$ is a Nash equilibrium for the N-player MCDM if for each $j \in \{1, \dots, N\}$,

$$J_j^N(\pi_*^{(N)}) = \sup_{\pi \in \Pi^j} J_j^N(\pi, \pi_*^{-j}),$$

where $\pi_*^{-j} = (\pi_*^1, \dots, \pi_*^{j-1}, \pi_*^{j+1}, \dots, \pi_*^N)$.

Definition 3.13 (ε-Nash equilibrium)**.** For $\varepsilon > 0$, a policy $\pi^{(N)} \in \Pi$ is an ε-Nash equilibrium for the MCDM if for each $j \in \{1, \dots, N\}$,

$$J_j^N(\pi) \geq \sup_{\pi \in \Pi^j} J_j^N(\pi, \pi^{-j}) - \varepsilon.$$

In general, the Nash equilibrium is hard to characterise and computationally intractable. It is also impractical to search over policies that depend on the distribution of all players. Therefore it is more useful to consider a search over Markovian policies for each player, and formulate the equilibrium condition with respect to such policies. Indeed the common approach for modelling partially observable games is to consider Markovian policies as above [60]. This is a reasonable assumption as in practice it will be hard for each agent to keep track of the movement of all other players, when the number of players grow increasingly large.

A policy is *Markovian* if $\pi = (\pi_t)_{t \geq 0}$ is such that $\pi_t : \mathcal{Y} \to \Delta_U$. Let Π^j_{mrkv} denote the set of Markov policies for the player j, with $\Pi_{\text{mrkv}} = \prod_{j=1}^N \Pi^j_{\text{mrkv}}$.

Definition 3.14 (Markov–Nash equilibrium)**.** $\pi_*^{(N)} \in \Pi_{\text{mrkv}}$ is a Markov–Nash equilibrium for the N-player MCDM if for each $j \in \{1, \dots, N\}$,

$$J_j^N(\pi_*^{(N)}) = \sup_{\pi \in \Pi^j_{\text{mrkv}}} J_j^N(\pi, \pi_*^{-j}),$$

where $\pi_*^{-j} = (\pi_*^1, \dots, \pi_*^{j-1}, \pi_*^{j+1}, \dots, \pi_*^N)$.

Definition 3.15 (ε-Markov–Nash equilibrium)**.** For $\varepsilon > 0$, a policy $\pi^{(N)} \in \Pi_{\text{mrkv}}$ is an ε-Nash equilibrium for the MCDM if for each $j \in \{1, \dots, N\}$,

$$J_j^N(\pi) \geq \sup_{\pi \in \Pi^j_{\text{mrkv}}} J_j^N(\pi, \pi^{-j}) - \varepsilon.$$

3.3.2 MFNE for the MFG-MCDM

The computation and characterisation of Nash equilibria is typically intractable due to the curse of dimensionality and the coupled dynamics across the different agents.

Therefore, as an approximation, we consider the infinite population limit by sending the number of players $N \to \infty$, and replacing the empirical distribution of the agents by a measure flow $\boldsymbol{\mu} = (\mu_n)_n \in \Delta_{\mathcal{X}}^{\infty}$. In the mean-field setting, we consider the viewpoint of one representative agent, and assume that its interactions with members of the population, modelled by the measure flow $\boldsymbol{\mu}$, are symmetric. As in the N-player game, we consider a tuple $\langle \mathcal{X}, A, \mathcal{D}, \mathcal{C}, p, r \rangle$ (see Definition 3.10). For a given measure flow $\boldsymbol{\mu} \in \Delta_{\mathcal{X}}^{\infty}$, at time n, a representative agent transitions from the state x_n to a new state x_{n+1} with probability

$$p(x_{n+1} \mid x_n, a_n, \mu_n),$$

and collects a reward of $r(x_n, a_n, \mu_n)$. As each transition of the underlying state now depends on the given measure, the d-step transition kernel $p^{(d)} : \mathcal{X} \times A^d \times \Delta_{\mathcal{X}}^d \to \Delta_{\mathcal{X}}$ now depends on the measure flow across the d time steps, so that we have

$$x_{n+d} \sim p^{(d)}(\cdot \mid x_n, (a)_n^{n+d-1}; (\mu)_n^{n+d-1}). \tag{3.5}$$

We impose the follow Lipschitz assumptions on the transition kernels and reward function.

Assumption 3.16.

(a) The one-step reward function r satisfies a Lipschitz bound: there exists a constant L_r such that for all $x, \hat{x} \in \mathcal{X}$, $a, \hat{a} \in A$, $\mu, \hat{\mu} \in \Delta_{\mathcal{X}}$,

$$|r(x, a, \mu) - r(\hat{x}, \hat{a}, \hat{\mu})| \leq L_r \left(\mathbb{1}_{\{x \neq \hat{x}\}} + \mathbb{1}_{\{a \neq \hat{a}\}} + \delta_{TV}(\mu, \hat{\mu}) \right).$$

(b) For $1 \leq n \leq d_0$, the n-step transition kernels satisfy a uniform Lipschitz bound: there exists a constant L_p such that for all $x, \hat{x} \in \mathcal{X}$, $\boldsymbol{a}, \hat{\boldsymbol{a}} \in A^n$, $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \Delta_{\mathcal{X}}^n$,

$$\delta_{TV}\left(p^{(n)}(\cdot \mid x, \boldsymbol{a}, \boldsymbol{\mu}),\ p^{(n)}(\cdot \mid \hat{x}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\mu}}) \right) \leq L_p \left(\mathbb{1}_{\{x \neq \hat{x}\}} + \mathbb{1}_{\{\boldsymbol{a} \neq \hat{\boldsymbol{a}}\}} + \delta_{\max}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \right).$$

In particular, as both $\mathcal{X}$ and A are assumed to be finite, and the simplex $\Delta_{\mathcal{X}}$ is compact, both the reward function r and transition kernel p are bounded by some constants M_r and M_p respectively.

Once again, we shall consider the lifted problem on the augmented space

$$\mathcal{Y} := \bigcup_{d=d_K}^{d_0} \{d\} \times \mathcal{X} \times A^d.$$

Under this augmented space $\mathcal{Y}$, now with the inclusion of the measure dependence, the counterparts to p_y and r_y in (3.2) and (3.4) are given as follows. Let $y = \left(d, x, (a)_{n-d}^{n-1}\right)$, $\hat{y} = (\hat{d}, \hat{x}, \hat{\boldsymbol{a}})$, $u = (a_n, i)$, and $\boldsymbol{\mu} = (\mu)_{n-d_0}^{n-d_K}$,

– $p_y : \mathcal{Y} \times U \times \Delta_{\mathcal{X}}^{d_0-d_K+1} \to \Delta_{\mathcal{Y}}$ is given by

$$
\begin{aligned}
&p_y(\hat{y} \mid y, u, \boldsymbol{\mu}) \\
&= \begin{cases} p^{(d-d_i+1)}\big(\hat{x} \mid x, (a)_{n-d}^{n-d_i}, (\mu)_{n-d}^{n-d_i}\big)\ \mathbb{1}_{\left\{\hat{d}=d_i,\ \hat{\boldsymbol{a}}=(a)_{n-d_i+1}^{n}\right\}}, & d_i \le d \le d_0; \\ \mathbb{1}_{\left\{\hat{x}=x,\ \hat{d}=d+1,\ \hat{\boldsymbol{a}}=(a)_{n-d}^{n}\right\}}, & d_K \le d < d_i. \end{cases}
\end{aligned} \tag{3.6}
$$

– The reward function $r_y : \mathcal{Y} \times U \times \Delta_{\mathcal{X}}^{d_0+1} \to [0, \infty)$ is

$$
r_y(y, u, \boldsymbol{\mu}) = \sum_{x' \in \mathcal{X}} r(x', a_n, \mu_n) p^{(d)}(x' \mid x, \boldsymbol{a}, (\mu)_{n-d}^{n-1}) - \sum_{k=1}^{K} c_k \mathbb{1}_{\{i=k\}}, .
$$

Given Assumption 3.16, we have the following bounds in p_y and r_y.

Proposition 3.17. *Under Assumption 3.16:*

(a) For all $y, \hat{y} \in \mathcal{Y}$, $u, \hat{u} \in U$, $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \Delta_{\mathcal{X}}^{d_0-d_K+1}$, the augmented kernel p_y satisfies the Lipschitz bound

$$
\delta_{TV}\left(p_y(\cdot \mid y, u, \boldsymbol{\mu}),\ p_y(\cdot \mid \hat{y}, \hat{u}, \hat{\boldsymbol{\mu}})\right) \le L_P \left(\mathbb{1}_{\{y \neq \hat{y}\}} + \mathbb{1}_{\{u \neq \hat{u}\}} + \delta_{\max}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\right).
$$

where $L_P = \max\{2M_p, L_p\}$.

(b) For all $y, \hat{y} \in \mathcal{Y}$, $u \in U$, $\boldsymbol{\mu}, \hat{\boldsymbol{\mu}} \in \Delta_{\mathcal{X}}^{d_0+1}$, the augmented reward function r_y is in $\boldsymbol{\mu}$ and satisfies the bound:

$$
\begin{aligned}
&|r_y(y, u, \boldsymbol{\mu}) - r_y(\hat{y}, u, \hat{\boldsymbol{\mu}})| \\
&\le 2L_r M_p \mathbb{1}_{\{y \neq \hat{y}\}} + (L_r + c_K - c_0)\mathbb{1}_{\{u \neq \hat{u}\}} + L_R\ \delta_{\max}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}),
\end{aligned}
$$

where $L_R = L_r + L_r L_p$. Also r_y is bounded by $M_r + c_K =: M_R$.

Proof. (a) We have from the triangle inequality

$$
\begin{aligned}
&\delta_{TV}\left(p_y(\cdot \mid y, u, \boldsymbol{\mu}),\ p_y(\cdot \mid \hat{y}, \hat{u}, \hat{\boldsymbol{\mu}})\right) \\
&\le \delta_{TV}\left(p_y(\cdot \mid y, u, \boldsymbol{\mu}),\ p_y(\cdot \mid \hat{y}, u, \boldsymbol{\mu})\right) + \delta_{TV}\left(p_y(\cdot \mid \hat{y}, u, \boldsymbol{\mu}),\ p_y(\cdot \mid \hat{y}, \hat{u}, \hat{\boldsymbol{\mu}})\right) \\
&\le 2M_p \mathbb{1}_{\{y \neq \hat{y}\}} + L_p\big(\mathbb{1}_{\{u \neq \hat{u}\}} + \delta_{\max}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\big)
\end{aligned}
$$

where the first term in the last inequality follows from the uniformly bounded M_p for all d-step transition kernels $p^{(d)}$, and the the other terms follow from the Lipschitz assumption of $p^{(d)}$.

(b) For consistency sake in notation, we index $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ from time $n-d_0$ to n. Let $y=(d,x,\boldsymbol{a})$, $\hat{y}=(\hat{d},\hat{x},\hat{\boldsymbol{a}})$, $u=(a,i)$ and $u'=(a',i')$, then

$$
\begin{aligned}
&|r_y(y,u,\boldsymbol{\mu})-r_y(\hat{y},\hat{u},\hat{\boldsymbol{\mu}})| \\
\leq\ &\left|\sum_{x'\in\mathcal{X}} r(x',a,\mu_n)\ p^{(d)}(x'\mid x,\boldsymbol{a},(\mu)_{n-d}^{n-1})-\sum_{x'\in\mathcal{X}} r(x',\hat{a},\hat{\mu}_n)\ p^{(\hat{d})}(x'\mid \hat{x},\hat{\boldsymbol{a}},(\hat{\mu})_{n-\hat{d}}^{n-1})\right| \\
&\quad +|c_i-c_{i'}| \\
\leq\ &\left|\sum_{x'\in\mathcal{X}} r(x',a,\mu_n)\left(p^{(d)}(x'\mid x,\boldsymbol{a},(\mu)_{n-d}^{n-1})-\ p^{(\hat{d})}(x\mid \hat{x},\hat{\boldsymbol{a}},(\hat{\mu})_{n-\hat{d}}^{n-1})\right)\right| \\
&\quad +\left|\sum_{x'\in\mathcal{X}} (r(x',a,\mu_n)-r(x',\hat{a},\hat{\mu}_n))\,p^{(\hat{d})}(x'\mid \hat{x},\hat{\boldsymbol{a}},(\hat{\mu})_{n-\hat{d}}^{n-1})\right|+(c_K-c_0)\mathbb{1}_{\{i\neq i'\}} \\
\leq\ &(r(x_{\max},a,\mu_n)-r(x_{\min},a,\mu_n)) \\
&\quad \cdot\delta_{TV}\left(p^{(d)}\big(\cdot\mid x,\boldsymbol{a},(\mu)_{n-d}^{n-1}\big),\ p^{(\hat{d})}\big(\cdot\mid \hat{x},\hat{\boldsymbol{a}},(\hat{\mu})_{n-\hat{d}}^{n-1}\big)\right) \\
&\quad +\max_{x\in\mathcal{X}}|r(x,a,\mu_n)-r(x,\hat{a},\hat{\mu}_n)|+(c_K-c_0)\mathbb{1}_{\{i\neq i'\}} \\
\leq\ &L_r\left(2M_P\mathbb{1}_{\{y\neq\hat{y}\}}+L_p\ \delta_{\max}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})+\mathbb{1}_{\{a\neq a'\}}+\delta_{\max}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}})\right)+(c_K-c_0)\mathbb{1}_{\{i\neq i'\}} \\
\leq\ &2L_rM_P\mathbb{1}_{\{y\neq\hat{y}\}}+(L_r+c_K-c_0)\mathbb{1}_{\{u\neq u'\}}+(L_r+L_rL_p)\ \delta_{\max}(\boldsymbol{\mu},\hat{\boldsymbol{\mu}}),
\end{aligned}
$$

where Proposition 3.2 is used for the third inequality. The second part is immediate from the definition of r_y.

□

We now proceed to establish the mean-field Nash equilibrium (MFNE) condition for agents operating under the MCDM formulation. This is characterised by a fixed point of the composition of the best response map and the measure flow map (e.g. [59]). As the presence of observation delays leads to a non-Markovian problem, the fixed point characterisation will be established in terms of the augmented space. However, both p_y and r_y in general depend on the various d-step transition kernels (3.5), which in turn depend on measures on the underlying space $\mathcal{X}$. Thus, given a distribution $\nu_t\in\Delta_{\mathcal{Y}}$ on the augmented space, one would like to construct a sequence of measures $(\mu_{t,d})_{d=0}^{d_0}\in\Delta_{\mathcal{X}}^{d_0+1}$ for the transition kernel p_y and augmented reward r_y. This can be seen as analogous to players in the N-player game estimating the distribution of the underlying states of all players, given the belief state. In order to construct such a sequence of measures described above, **we shall have to further enlarge $\mathcal{Y}$ and**

consider the space

$$\tilde{\mathcal{Y}} \coloneqq \bigcup_{d=d_K}^{d_0} \{d\} \times \mathcal{X}^{d_0-d+1} \times A^d. \tag{3.7}$$

In this instance, an element $\tilde{y}_n \in \tilde{\mathcal{Y}}$ can now be understood as

$$\tilde{y}_n = (d, x_{n-d_0}, \ldots, x_{n-d}, a_{n-d}, \ldots, a_{n-1}),$$

where once again, negative indices are used to indicate that the relevant states and actions occurred in the past. Now, given $\nu_t \in \Delta_{\tilde{\mathcal{Y}}}$, we successively compute a sequence of distributions for the states $(x_{n-d_0}, \ldots, x_0)$, starting with x_{n-d_0}. The inclusion of the entire sequence of $(x_{n-d_0}, \ldots, x_{n-d})$ for the space $\tilde{\mathcal{Y}}$ is essential, as for each $0 < t \leq d_0$, we require a distribution of the state x_{n-t} in order to compute a distribution for the next state x_{n-t+1}.

The construction of the map $\Delta_{\tilde{\mathcal{Y}}} \ni \nu_t \mapsto \boldsymbol{\mu}_t^\nu = (\mu_{t,d})_{d=0}^{d_0} \in \Delta_{\mathcal{X}}^{d_0+1}$ is then given by the following. We use superscripts to denote the corresponding marginal and conditional distributions on the coordinates. For example, ν_t^d as the marginal of ν_t on the delay coordinate, and $\nu_t^{\boldsymbol{x},\boldsymbol{a}|\bar{d}}$ is the conditional distribution of ν_t on the x and $\boldsymbol{a}$ coordinates, given a delay of $\bar{d}$, so that we have

$$\nu_t(\bar{d}, \boldsymbol{x}, \boldsymbol{a}) = \nu_t^d(\bar{d})\ \nu_t^{\boldsymbol{x},\boldsymbol{a}|\bar{d}}(\boldsymbol{x}, \boldsymbol{a}).$$

Now, starting with $d' = d_0$, take

$$\mu_{t,d_0} = \nu_t^{x_{t-d_0}} \in \Delta_{\mathcal{X}},$$

the marginal of ν_t on the x_{t-d_0} coordinate. Next, define recursively for each $0 \leq d' < d_0$,

$$\mu_{t,d'}(x) = \sum_{\bar{d} \in \overline{\mathcal{D}}} \nu_t^d(\bar{d}) \xi_{t,d'}^{\bar{d}}(x)$$

where

$$\xi_{t,d'}^{\bar{d}}(\cdot) = \begin{cases} \nu_t^{x_{t-d'}|\bar{d}}(\cdot), & \bar{d} \leq d'; \\ \sum_{x,\boldsymbol{a}} p^{(\bar{d}-d')}(\cdot \mid x, \boldsymbol{a}, (\mu_{t,d})_{d=d'+1}^{\bar{d}})\ \nu_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}), & \bar{d} > d'. \end{cases}$$

Intuitively, the measures $\boldsymbol{\mu}_t^\nu = (\mu_{t,d})_{d=0}^{d_0}$ represent the distribution of the underlying states of the agents from time $t - d_0$ to time t based on ν_t, which can be interpreted as the distribution of the information states of the population at time t. Since the

information state varies with the delay period, the conditional distributions $\nu_t^{\boldsymbol{x},\boldsymbol{a}|\bar{d}}$ have to be considered separately for each $\bar{d} \in \overline{\mathcal{D}}$. The following lemma shows that this mapping is also Lipschitz, and will be useful later when establishing a contraction in the regularised regime.

Lemma 3.18. *The mapping $\nu_t \mapsto \boldsymbol{\mu}_t^{\nu}$ is Lipschitz with constant $L_M = \sum_{k=0}^{d_0}(2L_p)^k$.*

Proof. Let $\nu_t, \hat{\nu}_t \in \Delta_{\tilde{\mathcal{Y}}}$, with respective images $\boldsymbol{\mu}_t^{\nu} = (\mu_{t,d})_{d=0}^{d_0}$ and $\boldsymbol{\mu}_t^{\hat{\nu}} = (\hat{\mu}_{t,d})_{d=0}^{d_0}$. First, by definition we have

$$\delta_{TV}(\mu_{t,d_0}, \hat{\mu}_{t,d_0}) = \delta_{TV}(\nu_t^{x_{t-d_0}}, \hat{\nu}_t^{x_{t-d_0}}) \leq \delta_{TV}(\nu_t, \hat{\nu}_t).$$

Now fix $0 \leq d' \leq d_0$. Then

$$\delta_{TV}(\mu_{t,d'}, \hat{\mu}_{t,d'}) = \sum_{x' \in \mathcal{X}} \Bigg| \underbrace{\sum_{\bar{d}=0}^{d_0} \nu_t^d(\bar{d}) \xi_{t,d'}^{\bar{d}}(x')}_{I_1} - \underbrace{\sum_{\bar{d}=0}^{d_0} \hat{\nu}_t^d(\bar{d}) \hat{\xi}_{t,d'}^{\bar{d}}(x')}_{I_2} \Bigg| \tag{3.8}$$

We can write I_1 as

$$\begin{aligned} I_1 &= \sum_{\bar{d}=0}^{d'} \nu_t^d(\bar{d}) \nu_t^{x_{t-d'|\bar{d}}}(x') + \sum_{\bar{d}=d'+1}^{d_0} \nu_t^d(\bar{d}) \sum_{x,\boldsymbol{a}} p^{(\bar{d}-d')}(x' \mid x, \boldsymbol{a}, (\mu_{t,d})_{d=d'+1}^{\bar{d}})\ \nu_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}) \\ &=: \sum_{\bar{d}=0}^{d'} \nu_t^{d,x_{t-d'}}(\bar{d}, x') + J_1 \end{aligned}$$

Similarly, we can write I_2 as

$$I_2 = \sum_{\bar{d}=0}^{d'} \hat{\nu}_t^{d,x_{t-d'}}(\bar{d}, x') + J_2,$$

with J_2 defined analogously. Then

$$\begin{aligned} &\sum_{x' \in \mathcal{X}} |J_1 - J_2| \\ &\leq \sum_{x' \in \mathcal{X}} \sum_{\bar{d}=d'+1}^{d_0} \sum_{x,\boldsymbol{a}} \Big| \nu_t^d(\bar{d})\ p^{(\bar{d}-d')}(x' \mid x, \boldsymbol{a}, (\mu_{t,d})_{d=d'+1}^{\bar{d}})\ \nu_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}) \\ &\qquad\qquad - \hat{\nu}_t^d(\bar{d})\ p^{(\bar{d}-d')}(x' \mid x, \boldsymbol{a}, (\hat{\mu}_{t,d})_{d=d'+1}^{\bar{d}})\ \hat{\nu}_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}) \Big| \\ &\leq \sum_{\bar{d}=d'+1}^{d_0} \Bigg(\sum_{x,\boldsymbol{a}} \Big| \nu_t^d(\bar{d}) \nu_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}) - \hat{\nu}_t^d(\bar{d}) \hat{\nu}_t^{x,\boldsymbol{a}|\bar{d}}(x, \boldsymbol{a}) \Big| \sum_{x' \in \mathcal{X}} p^{(\bar{d}-d')}(x' \mid x, \boldsymbol{a}, (\hat{\mu}_{t,d})_{d=d'+1}^{\bar{d}}) \end{aligned}$$

$$+\sum_{x,\boldsymbol{a}}\left|\nu_t^d(\bar{d})\nu_t^{x,\boldsymbol{a}|\bar{d}}(x,\boldsymbol{a})\right|\sum_{x'\in\mathcal{X}}\Big|p^{(\bar{d}-d')}(x'\mid x,\boldsymbol{a},(\mu_{t,d})_{d=d'+1}^{\bar{d}})$$
$$-p^{(\bar{d}-d')}(x'\mid x,\boldsymbol{a},(\hat{\mu}_{t,d})_{d=d'+1}^{\bar{d}})\Big|\Bigg)$$
$$\leq\sum_{\bar{d}=d'+1}^{d_0}\sum_{x,\boldsymbol{a}}\left|\nu_t^{d,x,\boldsymbol{a}}(\bar{d},x,\boldsymbol{a})-\hat{\nu}_t^{d,x,\boldsymbol{a}}(\bar{d},x,\boldsymbol{a})\right|+2L_p\ \delta_{\max}\big((\mu_{t,d})_{d=d'+1}^{\bar{d}},(\hat{\mu}_{t,d})_{d=d'+1}^{\bar{d}}\big).$$

Returning to (3.8), we have

$$\begin{aligned}\delta_{TV}(\mu_{t,d'},\hat{\mu}_{t,d'})&\leq\sum_{\bar{d}=0}^{d'}\sum_{x'\in\mathcal{X}}\left|\nu_t^{d,x_{t-d'}}(\bar{d},x')-\hat{\nu}_t^{d,x_{t-d'}}(\bar{d},x')\right|+\sum_{x'\in\mathcal{X}}|J_1-J_2|\\&\leq 2L_p\ \delta_{\max}\big((\mu_{t,d})_{d=d'+1}^{\bar{d}},(\hat{\mu}_{t,d})_{d=d'+1}^{\bar{d}}\big)+\delta_{TV}(\nu_t,\hat{\nu}_t),\end{aligned}$$

so that

$$\delta_{\max}(\boldsymbol{\mu}_t^{\nu},\boldsymbol{\mu}_t^{\hat{\nu}})=\max_{0\leq d\leq d_0}\delta_{TV}(\mu_{t,d},\hat{\mu}_{t,d})=L_M\ \delta_{TV}(\nu_t,\hat{\nu}_t),$$

as required. □

Now let $\boldsymbol{\nu}\in\Delta_{\tilde{\mathcal{Y}}}^{\infty}$. Given this fixed $\boldsymbol{\nu}$, optimising the objective function becomes the single agent problem in Section 3.2. Hence, for a policy $\pi\in\Pi_{DM}$, define the objective function

$$J_{\boldsymbol{\nu}}(\pi)\coloneqq\mathbb{E}^{\pi}\left[\sum_{n=0}^{\infty}\gamma^n r_y(y_n,u_n,\boldsymbol{\mu}_n^{\nu})\right],$$

where $\mathbb{E}^{\pi}$ is the expectation induced by the transition kernel p_y and policy π. Then, the MFNE for the MCDM is defined as the following.

Definition 3.19. Let $\boldsymbol{\nu}=(\nu_t)_t\in\Delta_{\tilde{\mathcal{Y}}}^{\infty}$. Define:

(i) The best-response map $\Phi^{\text{aug}}:\Delta_{\tilde{\mathcal{Y}}}^{\infty}\to\Pi_{DM}$, given by

$$\Phi^{\text{aug}}(\boldsymbol{\nu})=\left\{\hat{\pi}\in\Pi_{DM}:J_{\boldsymbol{\nu}}(\hat{\pi})=\sup_{\pi\in\Pi_{DM}}J_{\boldsymbol{\nu}}(\pi)\right\},$$

(ii) The measure flow map $\Psi^{\text{aug}}:\Pi_{DM}\to\Delta_{\tilde{\mathcal{Y}}}^{\infty}$, defined recursively by $\Psi^{\text{aug}}(\pi)_0=\nu_0$ and

$$\Psi^{\text{aug}}(\pi)_{t+1}(\cdot)=\sum_{y\in\tilde{\mathcal{Y}}}\sum_{u\in U}p_y\Big(\cdot\mid y,u,\boldsymbol{\mu}_t^{\Psi^{\text{aug}}(\pi)}\Big)\pi_t(u\mid y)\Psi^{\text{aug}}(\pi)_t(y).$$

(iii) The mean-field Nash equilibrium (MFNE) for the MCDM problem $(\pi^*, \boldsymbol{\nu}^*) \in \Pi_{DM} \times \Delta_{\tilde{\mathcal{Y}}}^{\infty}$ is given by the fixed point $\boldsymbol{\nu}^*$ of $\Psi^{\text{aug}} \circ \Phi^{\text{aug}}$, for which $\pi^* \in \Phi^{\text{aug}}(\boldsymbol{\nu}^*)$ (best response map) and $\boldsymbol{\nu}^* = \Psi^{\text{aug}}(\pi^*)$ (measure flow induced by policy) holds.

The existence of MFNE for discrete MFGs in the fully observable case is shown in [59], by utilising the Kakutani fixed point theorem, and further extended to MFGs with partial information in [60]. As we see above, Definition 3.19 is analogous to classical MFNE characterisations in discrete MFG setups [6, 23, 46, 59], with the extra step of incorporating the maps $\nu_t \mapsto \boldsymbol{\mu}_t^{\nu}$. This is different to the barycenter approach in [60, p.9]: when the belief state is measure-valued, taking the barycenter of a measure on the augmented state is effectively 'taking the average' to give a measure on the underlying state. Here, the belief state is parameterised by a finite set given by past observations, so the notion of taking the barycenter does not apply here. Moreover, both p_y and r_y depend on the distribution of the underlying state across multiple time points in the past. Therefore an explicit construction of $\boldsymbol{\mu}^{\nu}$ here is required.

Remark 3.20. *The extra enlargement of the space $\mathcal{Y}$ to $\mathcal{Y}'$ is necessary to formulate the MFNE fixed point condition and to compute $\boldsymbol{\mu}^{\nu}$ from ν. This enlargement is not required for the best response update, as the extra states are irrelevant when solving the MDP for a fixed measure flow. One can view an element of $\mathcal{Y}$ as an equivalence class on $\tilde{\mathcal{Y}}$, defined by the relation that two elements are equivalent if and only if the values of $(d, x_{n-d}, a_{n-d}, \ldots, a_{n-1})$ are identical.*

3.4 Regularised MFG for the MCDM

It is known that for finite-state MFGs, the MFNE need not be unique, and the fixed point operator given by $\Psi^{\text{aug}} \circ \Phi^{\text{aug}}$ does not form a contraction in general [23]. In order to compute for an approximate MFNE, we mirror the approaches of [6, 23] and consider a closely related game with a regulariser. This regulariser is an additive term to the reward in the objective function, and is given by a strongly convex function $\Omega : \Delta_U \to \mathbb{R}$. Then, we consider the regularised objective function

$$J_{\eta,\boldsymbol{\nu}}^{\text{reg}}(\pi) = \sum_{n=0}^{\infty} \gamma^n \left(\mathbb{E}^{\pi}[r\big(y_n, u_n, \boldsymbol{\mu}_n^{\nu}\big)] - \eta\, \Omega(\pi_n) \right),$$

where η is the regularisation parameter. The regularisation allows for a smoothed maximum to be obtained for the value function, and is often applied in reinforcement

learning problems to improve policy exploration [27]. Specifically, if Ω is strongly convex, then its Legendre-Fenchel transform $\Omega^* : \mathbb{R}^U \to \mathbb{R}$, defined as

$$\Omega^*(f) = \max_{\pi \in \Delta_U} \left(\langle \pi, f \rangle - \Omega(\pi) \right),$$

has the property that $\nabla\Omega^*$ is Lipschitz and satisfies

$$\nabla\Omega^*(f) = \arg\max_{\pi \in \Delta_U} \left(\langle \pi, f \rangle - \Omega(\pi) \right).$$

In view of the above, one can interpret the $\Omega^*(f)$ term as the optimal value for f across the set of admissible policies, with the optimal policy given by $\nabla\Omega^*(f)$. Commonly, Ω will be a KL divergence: $\Omega(\pi_n) = D_{KL}(\pi_n \| q)$, for some reference measure $q \in \Delta_U$. Then, the objective function reads

$$J^{\mathrm{reg}}_{\eta,\boldsymbol{\nu}}(\pi) = \mathbb{E}^{\pi} \left[\sum_{n=0}^{\infty} \gamma^n R^{\eta}\big(y_n, u_n, \boldsymbol{\mu}^{\nu}_n\big) \right],$$

where $R^{\eta}(y_n, u_n, \boldsymbol{\mu}^{\nu}_n) = r_y(y_n, u_n, \boldsymbol{\mu}^{\nu}_n) - \eta \log \frac{\pi(u_n|y_n)}{q(u_n)}$. To simplify the analysis, we shall consider q as the uniform distribution, i.e. $q(u) = 1/|U|$ for the rest of this section. Our statements readily extend to the case of arbitrary reference measures q, so long as q is bounded. We shall state the corresponding results for arbitrary q at the end of this section.

Following the notation of [6], we also consider the following quantities:

– the regularised value function $J^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}} : \mathbb{N} \times \tilde{\mathcal{Y}} \to \mathbb{R}$, where

$$J^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t, y) = \sup_{\pi \in \Pi_{DM}} \mathbb{E}^{\pi} \left[\sum_{n=t}^{\infty} \gamma^{n-t} R^{\eta}\big(y_n, u_n, \boldsymbol{\mu}^{\nu}_n\big) \middle| y_t = y \right].$$

– the optimal regularised Q-function $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}} : \mathbb{N} \times \tilde{\mathcal{Y}} \times U \to \mathbb{R}$, where

$$Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t, y, u) = \sup_{\pi \in \Pi_{DM}} \mathbb{E}^{\pi} \left[\sum_{n=t}^{\infty} \gamma^{n-t} R^{\eta}\big(y_n, u_n, \boldsymbol{\mu}^{\nu}_n\big) \middle| y_t = y, u_t = u \right].$$

Similarly to the metric δ_∞ for measure flows, for the Q-functions we shall use the metric

$$\delta_Q(f, g) = \sum_{t=0}^{\infty} \zeta^{-t} \max_{\substack{y \in \tilde{\mathcal{Y}} \\ u \in U}} \big(f(t, y, u), g(t, y, u) \big).$$

Intuitively, we are giving more weight to the values closer to the current time. The optimal regularised Q-function satisfies the dynamic programming relation

$$Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u) = R^{\eta}\big(y,u,\boldsymbol{\mu}^{\nu}_{t}\big) + \gamma \sum_{y'\in\tilde{\mathcal{Y}}} J^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t+1,y')p_y\big(y' \mid y,u,\boldsymbol{\mu}^{\nu}_{t}\big),$$

Note that although the transition kernel and reward are time-homogeneous, the inclusion of the time-dependent measure flow leads to a time-inhomogeneous dynamic programming relation (see Remark 3.9). It is well known that, when the regulariser Ω is given as relative entropy, the policy that maximises the regularised value function $J^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ is the softmax policy π^{soft}, where

$$\pi^{\mathrm{soft}}_{t}(u \mid y) = \frac{\exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u)/\eta)}{\sum_{u'\in U}\exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u')/\eta)}.$$

Then, the optimal regularised Q-function can be written in the form

$$\begin{aligned} Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u) &= r_y\big(y,u,\boldsymbol{\mu}^{\nu}_{t}\big) \\ &\quad + \gamma \sum_{y'\in\tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}^{\nu}_{t})\ \eta \log\left(\frac{1}{|U|}\sum_{u'\in U}\exp\frac{Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t+1,y',u')}{\eta}\right). \end{aligned}$$

We can thus define the regularised MCDM-MFNE by the analogous fixed point criteria.

Definition 3.21. Let $\boldsymbol{\nu} = (\nu_t)_t \in \Delta^{\infty}_{\tilde{\mathcal{Y}}}$ and $\eta > 0$. Define:

(i) The best-response map $\Phi^{\mathrm{reg}}_{\eta} : \Delta^{\infty}_{\tilde{\mathcal{Y}}} \to \Pi_{DM}$, given by

$$\Phi^{\mathrm{reg}}_{\eta}(\boldsymbol{\nu})_t(u \mid y) = \frac{\exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u)/\eta)}{\sum_{u'\in U}\exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u')/\eta)}.$$

(ii) $\Psi^{\mathrm{aug}} : \Pi_{DM} \to \Delta^{\infty}_{\tilde{\mathcal{Y}}}$, the measure flow map as defined previously, where $\Psi^{\mathrm{aug}}(\pi)_0 = \nu_0$ and for $t \geq 0$,

$$\Psi^{\mathrm{aug}}(\pi)_{t+1}(\cdot) = \sum_{y\in\tilde{\mathcal{Y}}}\sum_{u\in U} p_y\Big(\cdot \mid y,u,\boldsymbol{\mu}^{\Psi^{\mathrm{aug}}(\pi)}_{t}\Big)\pi_t(u \mid y)\Psi^{\mathrm{aug}}(\pi)_t(y).$$

(iii) The regularised MFNE for the MCDM problem $(\pi^*,\boldsymbol{\nu}^*) \in \Pi_{DM} \times \Delta^{\infty}_{\tilde{\mathcal{Y}}}$, given by the fixed point $\boldsymbol{\nu}^*$ of $\Psi^{\mathrm{aug}} \circ \Phi^{\mathrm{reg}}_{\eta}$, for which $\pi^* = \Phi^{\mathrm{reg}}_{\eta}(\boldsymbol{\nu}^*)$ (best response map) and $\boldsymbol{\nu}^* = \Psi^{\mathrm{aug}}(\pi^*)$ (measure flow induced by policy) holds.

The next step is to show that the fixed point operator $\Psi^{\text{aug}} \circ \Phi_\eta^{\text{reg}}$, under a suitable choice of metric and regulariser parameter η, forms a contraction mapping, such that the iteration of these maps will converge towards the fixed point, which is the regularised MFNE. We combine the approaches of [6, 23], extending their proof to the case of the infinite horizon problem with time-dependent measure flows, as well as the inclusion of the map $\nu_t \mapsto \boldsymbol{\mu}_t^\nu$ within the definitions of Φ_η^{reg} and Ψ^{aug}.

In order to demonstrate contraction of the regularised iterations, we defer the full statement and first show the following series of propositions regarding the Lipschitz continuity of the individual mappings. When treating the infinite horizon problem, we can approximate the optimal regularised Q-functions by considering its truncation at some finite time N, that is, first define

$$J_{\eta,\boldsymbol{\nu}}^{N,*}(t,y) = \sup_{\pi\in\Pi} \mathbb{E}^\pi \left[\sum_{n=t}^{N} \gamma^{n-t} R^\eta(y_n, u_n, \boldsymbol{\mu}_n^\nu) \middle| y_t = y \right], \quad t \le N,\ y \in \tilde{\mathcal{Y}}.$$

Then, extend $J_{\eta,\boldsymbol{\nu}}^{N,*}$ to $\mathbb{N}\times\tilde{\mathcal{Y}}$ by defining $J_{\eta,\boldsymbol{\nu}}^{N,*}(t,y) = 0$ for all $t > N$, $y \in \tilde{\mathcal{Y}}$. Similarly, define the truncated versions of the optimal regularised Q-function: for $t \le N$, $y \in \tilde{\mathcal{Y}}$ and $u \in U$,

$$Q_{\eta,\boldsymbol{\nu}}^{N,*}(t,y,u) = \sup_{\pi\in\Pi_{DM}} \mathbb{E}^\pi \left[\sum_{n=t}^{N} \gamma^{n-t} R^\eta\big(y_n, u_n, \boldsymbol{\mu}_n^\nu\big) \middle| y_t = y, u_t = u \right],$$

and once again extend $Q_{\eta,\boldsymbol{\nu}}^{N,*}$ to $\mathbb{N}\times\tilde{\mathcal{Y}}\times U$ by defining $Q_{\eta,\boldsymbol{\nu}}^{N,*}(t,\cdot,\cdot) = 0$ for all $t > N$. Then, the truncated optimal regularised Q-functions satisfy the following: for $t < N$,

$$\begin{aligned} Q_{\eta,\boldsymbol{\nu}}^{N,*}(t,y,u) &= r_y\big(y,u,\boldsymbol{\mu}_t^\nu\big) \\ &\quad + \gamma \sum_{y'\in\tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}_t^\nu)\ \eta \log\left(\frac{1}{|U|} \sum_{u'\in U} \exp \frac{Q_{\eta,\boldsymbol{\nu}}^{N,*}(t+1,y',u')}{\eta} \right), \quad (3.9) \\ Q_{\eta,\boldsymbol{\nu}}^{N,*}(t,y,u) &= r_y\big(y,u,\boldsymbol{\mu}_t^\nu\big) \\ &\quad + \gamma \sum_{y'\in\tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}_t^\nu)\ \eta \log\left(\frac{1}{|U|} \sum_{u'\in U} \exp \frac{Q_{\eta,\boldsymbol{\nu}}^{N-1,*}(t,y',u')}{\eta} \right). \quad (3.10) \end{aligned}$$

It is a standard result via successive approximations that $J_{\eta,\boldsymbol{\nu}}^{N,*} \to J_{\eta,\boldsymbol{\nu}}^{\text{reg},*}$ and $Q_{\eta,\boldsymbol{\nu}}^{N,*} \to Q_{\eta,\boldsymbol{\nu}}^{\text{reg},*}$ pointwise [35, Section 4.2]. We will utilise this pointwise convergence repeatedly in our analysis for the rest of this section.

Lemma 3.22. *For any $\boldsymbol{\nu}, \hat{\boldsymbol{\nu}} \in \Delta_{\tilde{\mathcal{Y}}}^\infty$ and any pair (t,N) such that $t \le N$, the truncated Q-functions $Q_{\eta,\nu}^{N,*}$ are uniformly bounded by $q^* := M_R/(1-\gamma)$.*

Proof. First note that

$$|Q^{N,*}_{\eta,\boldsymbol{\nu}}(N,y,u)| = |r_y\big(y,u,\boldsymbol{\mu}^{\nu}_N\big)| \le M_R =: q_{N,N}.$$

Then, for each $t < N$,

$$\begin{aligned}
|Q^{N,*}_{\eta,\boldsymbol{\nu}}(t,y,u)| &\le M_R + \gamma\eta \max_{y'\in\tilde{\mathcal{Y}}} \left|\log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{N,*}_{\eta,\boldsymbol{\nu}}(t+1,y',u')}{\eta}\right)\right| \\
&\le M_R + \gamma\eta\left(\frac{q_{N,t+1}}{\eta}\right) \\
&= M_R + \gamma q_{N,t+1} =: q_{N,t} \qquad (3.11)
\end{aligned}$$

where $q_{N,t+1}$ is the bound for $Q^{N,*}_{\eta,\boldsymbol{\nu}}(t+1,y,u)$. As $N\to\infty$, $q_{N,t}$ converges to the fixed point q^* of the map $x\mapsto\gamma x + M_R$, i.e. $q^* = M_R/(1-\gamma)$. Moreover, $q^* > M_R$ and is independent of t. Hence, for each t, we have $q_{N,t}\uparrow q^*$ as $N\to\infty$. Together with the fact that $Q^{N,*}_{\eta,\boldsymbol{\nu}}\to Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ pointwise, sending $N\to\infty$ in (3.11) gives as the uniform bound

$$|Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u)| \le q^*.$$

□

Now we shall prove by induction the following statement:

Lemma 3.23. *Let $\boldsymbol{\nu},\hat{\boldsymbol{\nu}}\in\Delta^{\infty}_{\tilde{\mathcal{Y}}}$. Then for each N, the truncated Q-functions satisfies*

$$\left|Q^{N,*}_{\eta,\boldsymbol{\nu}}(t,y,u) - Q^{N,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y,u)\right| \le l_{n,t}\ \delta_{TV}(\nu_t,\hat{\nu}_t),\quad 0\le t\le N,\ y\in\tilde{\mathcal{Y}},\ u\in U,\ \eta\ge 0,$$

where $(l_{n,t})_{0\le t\le n}$ satisfy the recurrence relation

$$l_{n,n} = L_R L_M,\quad l_{n,t} = \left(L_R L_M + \gamma\exp\left(\frac{2q^*}{\eta}\right) l_{n-1,t} + 2\gamma q^* L_p L_M\right).$$

Proof. For the base case $N=0$, we have

$$|Q^{0,*}_{\eta,\boldsymbol{\nu}}(0,y,u) - Q^{0,*}_{\eta,\hat{\boldsymbol{\nu}}}(0,y,u)| = |r_y\big(y,u,\boldsymbol{\mu}^{\nu}_0\big) - r_y\big(y,u,\boldsymbol{\mu}^{\hat{\nu}}_0\big)| \le L_R L_M \delta_{TV}(\nu_0,\hat{\nu}_0)$$

Now assume the hypothesis holds up to some $N=n$. Let $N=n+1$, the case $t=n+1$ is as above. Otherwise for $0\le t\le n$,

$$\begin{aligned}
&\left|Q^{n+1,*}_{\eta,\boldsymbol{\nu}}(t,y,u) - Q^{n+1,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y,u)\right| \\
\le\ &\left|r_y(y,u,\boldsymbol{\mu}^{\nu}_t) - r_y(y,u,\boldsymbol{\mu}^{\hat{\nu}}_t)\right|
\end{aligned}$$

$$
\begin{aligned}
&+\gamma \left| \sum_{y'\in\tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}_t^{\nu})\ \eta \log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\boldsymbol{\nu}}(t,y',u')}{\eta}\right)\right.\\
&\qquad \left. - \sum_{y'\in\tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}_t^{\hat{\nu}})\ \eta \log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y',u')}{\eta}\right)\right|\\
&\leq L_R\ \delta_{\max}(\boldsymbol{\mu}_t^{\nu},\boldsymbol{\mu}_t^{\hat{\nu}})\\
&\quad +\gamma\eta \max_{y'\in\tilde{\mathcal{Y}}}\left|\log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\boldsymbol{\nu}}(t,y',u')}{\eta}\right)-\log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y',u')}{\eta}\right)\right|\\
&\quad +\gamma\eta\ \delta_{TV}\left(p_y(\cdot \mid y,u,\boldsymbol{\mu}_t^{\nu}),\ p_y(\cdot \mid y,u,\boldsymbol{\mu}_t^{\hat{\nu}})\right)\\
&\qquad \left(\log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\boldsymbol{\nu}}(t,y_{\max},u')}{\eta}\right)-\log\left(\sum_{u'\in U}\frac{1}{|U|}\exp\frac{Q^{n,*}_{\eta,\boldsymbol{\nu}}(t,y_{\min},u')}{\eta}\right)\right)\\
&\leq L_R L_M \delta_{TV}(\nu_t,\hat{\nu}_t) + I_1 + I_2
\end{aligned}
$$

For I_2, noting the bound on $Q^{n,*}_{\eta,\boldsymbol{\nu}}$, we have

$$
I_2 \leq 2\gamma q^* L_p\ \delta_{\max}(\boldsymbol{\mu}_t^{\nu},\boldsymbol{\mu}_t^{\hat{\nu}}) \leq 2\gamma q^* L_p L_M \delta_{TV}(\nu_t,\hat{\nu}_t).
$$

For I_1 we use the mean value theorem and the Lipschitz property from induction to obtain

$$
\begin{aligned}
I_1 &\leq \gamma\eta \max_{y'\in\tilde{\mathcal{Y}}}\sum_{u'\in U}\left|\frac{\frac{1}{\eta}\exp(\frac{\zeta_{u'}}{\eta})}{\sum_{u''\in U}\exp(\frac{\zeta_{u''}}{\eta})}\right|\left|Q^{n,*}_{\eta,\boldsymbol{\nu}}(t,y',u')-Q^{n,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y',u')\right|\\
&\leq \gamma\exp\left(\frac{2q^*}{\eta}\right) l_{n,t}\ \delta_{TV}(\nu_t,\hat{\nu}_t).
\end{aligned}
$$

Combining all the above, we have

$$
\begin{aligned}
&\left|Q^{n+1,*}_{\eta,\boldsymbol{\nu}}(t,y,u)-Q^{n+1,*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y,u)\right|\\
&\leq\ \leq \left(L_R L_M+\gamma\exp\left(\frac{2q^*}{\eta}\right) l_{n,t}+2\gamma q^* L_p L_M\right)\delta_{TV}(\nu_t,\hat{\nu}_t),
\end{aligned}
$$

which completes the induction step as required. □

Proposition 3.24. *For $\eta > \frac{2M_R}{-(1-\gamma)\log\gamma}$, where M_R is the uniform bound of r_y, $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ is Lipschitz continuous with respect to $\boldsymbol{\nu}$ with Lipschitz constant*

$$
l_\eta = \frac{L_M(L_R+2\gamma q^* L_p)}{1-\gamma\exp\left(\frac{2q^*}{\eta}\right)}.
$$

Proof. By the pointwise convergence $Q^{N,*}_{\eta,\boldsymbol{\nu}} \to Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$, and the assumption that $\eta > \frac{2M_R}{-(1-\gamma)\log\gamma}$, for each t, $l_{N,t} \uparrow l_\eta$ as $N \to \infty$, to the fixed point of the map

$$x \mapsto L_R L_M + \gamma \exp\left(\frac{2q^*}{\eta}\right) x + 2\gamma q^* L_p L_M,$$

so that for each t,

$$\left|Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u) - Q^{\mathrm{reg},*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y,u)\right| \leq l_\eta\ \delta_{TV}(\nu_t, \hat{\nu}_t).$$

Note that l_η is independent of t. Therefore,

$$\begin{aligned}
\delta_Q(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}, Q^{\mathrm{reg},*}_{\eta,\hat{\boldsymbol{\nu}}}) &= \sum_{t=0}^{\infty} \zeta^{-t} \max_{\substack{y\in\tilde{\mathcal{Y}}\\ u\in U}} \left|Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u) - Q^{\mathrm{reg},*}_{\eta,\hat{\boldsymbol{\nu}}}(t,y,u)\right| \\
&\leq \sum_{t=0}^{\infty} \zeta^{-t} l_\eta\ \delta_{TV}(\nu_t, \hat{\nu}_t) \\
&= l_\eta\ \delta_\infty(\boldsymbol{\nu}, \hat{\boldsymbol{\nu}}).
\end{aligned}$$

as required. □

In particular, let η^* be a constant with $\eta^* > \frac{2M_R}{-(1-\gamma)\log\gamma}$. Then for all $\eta \geq \eta^*$, $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ has a uniform Lipschitz bound of l_{η^*}. The Lipschitz continuity of $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ allows us to obtain the Lipschitz continuity of Φ^{reg}_η. This relies on the following lemma from [23], which we restate here.

Lemma 3.25 ([23, Lemma B.7.5])**.** *Let $\eta > 0$ and $f_u : \Delta^\infty_{\tilde{\mathcal{Y}}} \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant K_f for any $u \in U$. Then the function*

$$\boldsymbol{\nu} \mapsto \frac{\exp\left(\frac{f_u(\boldsymbol{\nu})}{\eta}\right)}{\sum_{u'\in U} \exp\left(\frac{f_{u'}(\boldsymbol{\nu})}{\eta}\right)}$$

is Lipschitz with Lipschitz constant $K = \frac{(|U|-1)K_f}{2\eta}$ for any $u \in U$.

Corollary 3.26. *For $\eta \geq \eta^*$, the map Φ^{reg}_η is Lipschitz continuous with Lipschitz constant $K^\eta_{\mathrm{soft}} = \frac{|U|(|U|-1)l_{\eta^*}}{2\eta}$.*

Proof. Given any $\boldsymbol{\nu} \in \Delta^\infty_{\tilde{\mathcal{Y}}}$, Φ^{reg}_η maps $\boldsymbol{\nu}$ to the softmax policy

$$\pi^{\mathrm{soft}}_{\boldsymbol{\nu},t}(u \mid y) \coloneqq \frac{\exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u)/\eta)}{\sum_{u'\in U} \exp(Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(t,y,u')/\eta)}.$$

Then we simply note that for any $\boldsymbol{\nu}, \hat{\boldsymbol{\nu}} \in \Delta_{\tilde{\mathcal{Y}}}^{\infty}$,

$$\begin{aligned}\delta_{\Pi}(\Phi_{\eta}^{\mathrm{reg}}(\boldsymbol{\nu}), \Phi_{\eta}^{\mathrm{reg}}(\hat{\boldsymbol{\nu}})) &= \sup_{t\geq 0} \max_{y\in\tilde{\mathcal{Y}}} \ \delta_{\Pi}\left(\pi_{\boldsymbol{\nu},t}^{\mathrm{soft}}(\cdot \mid y),\ \pi_{\hat{\boldsymbol{\nu}},t}^{\mathrm{soft}}(\cdot \mid y)\right) \\ &= \sup_{t\geq 0} \max_{y\in\tilde{\mathcal{Y}}} \sum_{u\in U} |\pi_{\boldsymbol{\nu},t}^{\mathrm{soft}}(u \mid y) - \pi_{\hat{\boldsymbol{\nu}},t}^{\mathrm{soft}}(u \mid y)|\end{aligned}$$

Applying Lemma 3.25, together with the uniform Lipschitz constant l_{η^*} for $Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}$, gives us the desired result. □

We now show that the measure flow map Ψ^{aug} is Lipschitz, under a suitable choice of the constant ζ in the metric δ_{∞}. Intuitively, given two similar policies (in the sense of the metric δ_{Π}), the corresponding measure flows will gradually drift apart at a constant rate. The choice of ζ amounts to the weighting one gives to the current time over the distant future.

Proposition 3.27. *For $\zeta \in \mathbb{N}$ such that $\zeta > 2L_P L_M + 2$ in the metric δ_{∞} (3.1), the map Ψ^{aug} is Lipschitz with constant*

$$L_{\Psi} = \frac{2L_P}{2L_P L_M + 1}\left(\frac{\zeta}{\zeta - 2L_P L_M - 2} + \frac{1}{\zeta - 1}\right).$$

Proof. We will show inductively that

$$\delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_t, \Psi^{\mathrm{aug}}(\hat{\pi})_t) \leq S_t\ \delta_{\Pi}(\pi, \hat{\pi})$$

for constants where $S_{t+1} = 2L_P + 2S_t(L_P L_M + 1)$, $S_0 = 0$. Clearly at $t = 0$ we have

$$\delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_0, \Psi^{\mathrm{aug}}(\hat{\pi})_0) = \delta_{TV}(\nu_0, \nu_0) = 0$$

Then for the induction step, for $t \geq 0$,

$$\begin{aligned}&\delta_{TV}\left(\Psi^{\mathrm{aug}}(\pi)_{t+1}, \Psi^{\mathrm{aug}}(\hat{\pi})_{t+1}\right) \\ &= \sum_{y'\in\tilde{\mathcal{Y}}} \left| \sum_{y\in\tilde{\mathcal{Y}}} \sum_{u\in U} \left(p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)\pi_t(u \mid y)\Psi^{\mathrm{aug}}(\pi)_t(y) \right.\right. \\ &\qquad\qquad \left.\left. - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\hat{\pi}_t(u \mid y)\Psi^{\mathrm{aug}}(\hat{\pi})_t(y)\right)\right| \\ &\leq J_1 + J_2,\end{aligned}$$

where

$$J_1 \coloneqq \sum_{y'\in\tilde{\mathcal{Y}}} \left| \sum_{y\in\tilde{\mathcal{Y}}} \sum_{u\in U} \left(p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)\pi_t(u \mid y)\right.\right.$$

$$
\begin{aligned}
&\quad - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\hat{\pi}_t(u \mid y)\Big)\Psi^{\mathrm{aug}}(\pi)_t(y) \;\Big| \\
&\leq \sum_{y\in\tilde{\mathcal{Y}}}\sum_{y'\in\tilde{\mathcal{Y}}}\sum_{u\in U}\Big| p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)\pi_t(u \mid y) \\
&\qquad - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\hat{\pi}_t(u \mid y)\Big|\Psi^{\mathrm{aug}}(\pi)_t(y),
\end{aligned}
$$

and

$$
\begin{aligned}
J_2 &\coloneqq \sum_{y\in\tilde{\mathcal{Y}}}\sum_{y'\in\tilde{\mathcal{Y}}}\sum_{u\in U} p_y\left(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\right)\hat{\pi}_t(u \mid y)\,|\Psi^{\mathrm{aug}}(\pi)_t(y) - \Psi^{\mathrm{aug}}(\hat{\pi})_t(y)| \\
&\leq 2\ \delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_t, \Psi^{\mathrm{aug}}(\hat{\pi})_t).
\end{aligned}
$$

The summation over $u \in U$ in J_1 can be simplified to

$$
\begin{aligned}
&\sum_{u\in U}\Big|p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)\pi_t(u \mid y) - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\hat{\pi}_t(u \mid y)\Big| \\
\leq\ &\sum_{u\in U} p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)|\pi_t(u \mid y) - \hat{\pi}_t(u \mid y)| \\
&\quad + \sum_{u\in U}\Big|p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big) - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\Big|\,\hat{\pi}_t(u \mid y) \\
\leq\ &\left(p_y\Big(y' \mid y, u_{\max}, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big) - p_y\Big(y' \mid y, u_{\min}, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big)\right)\delta_{TV}(\pi_t(\cdot \mid y), \hat{\pi}_t(\cdot \mid y)) \\
&\quad + \max_{u\in U}\Big|p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big) - p_y\Big(y' \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\Big|,
\end{aligned}
$$

so that

$$
\begin{aligned}
J_1 &\leq 2L_P\ \delta_\Pi(\pi, \hat{\pi}) + 2\max_{y\in\tilde{\mathcal{Y}}}\max_{u\in U}\delta_{TV}\left(p_y\Big(\cdot \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\pi)}\Big) - p_y\Big(\cdot \mid y, u, \boldsymbol{\mu}_t^{\Psi^{\mathrm{aug}}(\hat{\pi})}\Big)\right) \\
&\leq 2L_P\ \delta_\Pi(\pi, \hat{\pi}) + 2L_P L_M\ \delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_t, \Psi^{\mathrm{aug}}(\hat{\pi})_t).
\end{aligned}
$$

Then, applying the inductive step,

$$
\begin{aligned}
&\delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_{t+1}, \Psi^{\mathrm{aug}}(\hat{\pi})_{t+1}) \\
&\leq 2L_P\ \delta_\Pi(\pi, \hat{\pi}) + (2L_P L_M + 2)\ \delta_{TV}(\Psi^{\mathrm{aug}}(\pi)_t, \Psi^{\mathrm{aug}}(\hat{\pi})_t) \\
&\leq \big(2L_P + 2S_t(L_P L_M + 1)\big)\delta_\Pi(\pi, \hat{\pi}),
\end{aligned}
$$

which proves the claim. Next, we see that S_t satisfies a first-order linear recurrence relation, and more generally has the explicit formula

$$
S_t = \frac{2L_P}{2L_P L_M + 1}(2L_P L_M + 2)^t - \frac{2L_P}{2L_P L_M + 1}.
$$

Therefore, fix some $\zeta \in \mathbb{N}$ such that $\zeta > 2L_P L_M + 2$. We then have

$$\begin{aligned}\delta_\infty(\Psi^{\text{aug}}(\pi), \Psi^{\text{aug}}(\hat{\pi})) &= \sum_{t=0}^{\infty} \zeta^{-t} \delta_{TV}(\Psi^{\text{aug}}(\pi)_t, \Psi^{\text{aug}}(\hat{\pi})_t) \\ &\leq \sum_{t=1}^{\infty} \frac{S_t}{\zeta^t}\ \delta_\Pi(\pi, \hat{\pi}) \\ &= \frac{2L_P}{2L_P L_M + 1} \left(\frac{\zeta}{\zeta - 2L_P L_M - 2} + \frac{1}{\zeta - 1} \right) \delta_\Pi(\pi, \hat{\pi})\end{aligned}$$

□

Our statement of contraction for the regularised fixed point operator of the MFG-MCDM is then essentially a corollary of the previous propositions.

Theorem 3.28. *Recall the Lipschitz constants L_p, L_P, L_R and L_M for p, p_y, r_y and the map $\nu_t \mapsto \boldsymbol{\mu}_t^{\nu}$ respectively. Let ζ and η^* be constants such that $\zeta > 2L_P L_M + 2$, and $\eta^* > \frac{2M_R}{-(1-\gamma)\log\gamma}$. Define*

$$q^* = \frac{M_R}{1-\gamma}, \quad l_{\eta^*} = \frac{L_M(L_R + 2\gamma q^* L_p)}{1 - \gamma \exp\left(\frac{2q^*}{\eta^*}\right)},$$

$$L_\Psi = \frac{2L_P}{2L_P L_M + 1} \left(\frac{\zeta}{\zeta - 2L_P L_M - 2} + \frac{1}{\zeta - 1} \right).$$

Then for any η such that

$$\eta > \frac{1}{2}|U|\,(|U| - 1)\, l_{\eta^*} L_\Psi,$$

the fixed point operator $\Psi^{\text{aug}} \circ \Phi_\eta^{\text{reg}}$ is a contraction mapping on the space $(\Delta_{\tilde{\mathcal{Y}}}^\infty, \delta_\infty)$, where the constant ζ in the metric δ_∞ (3.1) is as chosen above. In particular by Banach's fixed point theorem, there exists a unique fixed point for $\Psi^{\text{aug}} \circ \Phi_\eta^{\text{reg}}$, which is a regularised MFNE for the MFG-MCDM problem.

Proof. Let $\boldsymbol{\nu}, \hat{\boldsymbol{\nu}} \in \Delta_{\tilde{\mathcal{Y}}}^\infty$, and let $\pi = \Phi_\eta^{\text{reg}}(\boldsymbol{\nu})$ and $\hat{\pi} = \Phi_\eta^{\text{reg}}(\hat{\boldsymbol{\nu}})$. Then by Corollary 3.26 and Proposition 3.27,

$$\begin{aligned}\delta_\infty(\Psi^{\text{aug}}(\pi), \Psi^{\text{aug}}(\hat{\pi})) &\leq L_\Psi \delta_\Pi(\Phi_\eta^{\text{reg}}(\boldsymbol{\nu}), \Phi_\eta^{\text{reg}}(\hat{\boldsymbol{\nu}})) \\ &\leq L_\Psi K_{\text{soft}}^\eta \delta_\infty(\boldsymbol{\nu}, \hat{\boldsymbol{\nu}}),\end{aligned}$$

where we recall from Corollary 3.26 that $K_{\text{soft}}^\eta = \frac{|U|(|U|-1)l_{\eta^*}}{2\eta}$. Then for any η such that $\eta > \frac{1}{2}|U|\,(|U| - 1)\, l_{\eta^*} L_\Psi$, we have that $L_\Psi K_{\text{soft}}^\eta < 1$, and hence the map $\Psi^{\text{aug}} \circ \Phi_\eta^{\text{reg}}$ is a contraction. As $(\Delta_{\tilde{\mathcal{Y}}}^\infty, \delta_\infty)$ is a complete metric space (see Section 3.1.1), by the Banach fixed-point theorem, there exists a unique fixed point, which furthermore serves as a MFNE for the regularised MFG-MCDM by definition. □

We now state the analogous result for Theorem 3.28, when the KL divergence with respect to an arbitrary policy $q = (q_t)_t$ is used.

Theorem 3.29. *Let $q \in \Pi_{DM}$ be an arbitrary admissible policy, such that q is bounded above and below by $\bar{q} < \infty$ and $\underline{q} > 0$ respectively. Consider the regulariser as the KL divergence with respect to q:*

$$\Omega(\pi_t) = \sum_{u \in U} \pi_t(u) \log \frac{\pi_t(u)}{q_t(u)}.$$

Define $l_{\eta,q}$ by

$$l_{\eta,q} = \frac{L_M(L_R + 2\gamma q^* L_p)}{1 - \gamma \frac{\bar{q}}{\underline{q}} \exp\left(\frac{2q^*}{\eta}\right)},$$

where $q^* = M_R/(1-\gamma)$ as before. Let ζ and η^* be constants such that $\zeta > 2L_P L_M + 2$, and $\eta^* > \frac{2M_R}{(1-\gamma)(\log \underline{q} - \log \gamma \bar{q})}$. Then for any η such that

$$\eta > \frac{\bar{q}^2}{2\underline{q}^2} |U| \left(|U| - 1\right) L_\Psi l_{\eta^*,q},$$

the fixed point operator $\Psi^{\mathrm{aug}} \circ \Phi_\eta^{\mathrm{reg}}$ is a contraction mapping on the space $(\Delta_{\tilde{\mathcal{Y}}}^\infty, \delta_\infty)$, where the constant ζ in the metric δ_∞ (3.1) is as chosen above. In particular by Banach's fixed point theorem, there exists a unique fixed point for $\Psi^{\mathrm{aug}} \circ \Phi_\eta^{\mathrm{reg}}$, which is a regularised MFNE for the MFG-MCDM problem.

Proof. This is essentially a corollary of Theorem 3.28, by noting that the optimal regularised Q-function satisfies the dynamic programming

$$\begin{aligned} &Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t,y,u) \\ &= r_y\big(y,u,\boldsymbol{\mu}_t^\nu\big) + \gamma \sum_{y' \in \tilde{\mathcal{Y}}} p_y(y' \mid y,u,\boldsymbol{\mu}_t^\nu)\ \eta \log\left(\sum_{u' \in U} q_{t+1}(u' \mid y') \exp \frac{Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t+1,y',u')}{\eta}\right), \end{aligned}$$

and that the optimal policy π^{soft} now has the form

$$\pi_{\boldsymbol{\nu},t}^{\mathrm{soft}}(u \mid y) \coloneqq \frac{q_t(u \mid y) \exp(Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t,y,u)/\eta)}{\sum_{u' \in U} q_t(u' \mid y) \exp(Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t,y,u')/\eta)}.$$

Then the proof of Theorem 3.28 can be followed, inserting the bounds $\bar{q}$ and $\underline{q}$ into the relevant constants where appropriate. The precise proof in the classical fully observable case is shown in [23, Theorem 3]. □

3.4.1 Approximate Nash equilibria to the N-player game with controllable information delay

Recall that in the finite player case, player j's objective function is given by

$$J_j^N(\pi^{(N)}) = \mathbb{E}^{\pi^{(N)}}\left[\sum_{n=0}^{\infty}\gamma^n r(x_n^j, a_n^j, e_n^N)\right].$$

This subsection shows that the MFNE obtained from the regularised MFG with speed of information control forms an approximate Nash equilibria for sufficiently small η. Note that the theorem only states that the regularised Nash equilibria, defined via Definition 3.21, acts as an approximate Nash equilibria for the underlying finite game for sufficiently small values of η. However, one cannot infer from this the computability of the equilibria. Indeed for small η, there is no guarantee of a contractive fixed point operator, and the MFNE need not be unique. For the rest of this subsection, we only consider starting times of $t = 0$, so we shall consider the Q-functions as functions over $\tilde{\mathcal{Y}} \times U$, and write, for example, $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(y,u)$ for $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}(0,y,u)$ without loss of generality. Now for any policy π, define the associated Q-function by

$$Q^{\pi}_{\boldsymbol{\nu}}(y,u) = \mathbb{E}^{\pi}\left[\sum_{n=0}^{\infty}\gamma^n r_y\big(y_n, u_n, \boldsymbol{\mu}_n^{\nu}\big)\middle| y_0 = y, u_0 = u\right].$$

Define also the optimal (unregularised) Q-function $Q^*_{\boldsymbol{\nu}} \coloneqq \sup_{\pi\in\Pi_{DM}} Q^{\pi}_{\boldsymbol{\nu}}$. First, we shall prove the following convergence statements for the MFG with control of information speed.

Lemma 3.30. *The function $\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}$ converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}$ as $n \to \infty$, and converges uniformly over all $y \in \tilde{\mathcal{Y}}$ and $u \in U$.*

Proof. We shall prove in sequence the following statements:

1. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$ pointwise.
2. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ uniformly converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$.
3. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}(y,u)$ converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$ pointwise.
4. $\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu}^*_n)}_{\boldsymbol{\nu}}(y,u)$ uniformly converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$, uniformly over all $y \in \tilde{\mathcal{Y}}$ and $u \in U$.

We note that the corresponding convergence statements have been shown in [23] for the finite horizon case for fully observable MFGs. By Lemma 3.18, the map $\boldsymbol{\nu} \mapsto \boldsymbol{\mu}^{\nu}$ is continuous, therefore we obtain the following statements for the finite horizon problems for MFG-MCDM:

1a. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{N,*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ converges to $\boldsymbol{\nu} \mapsto Q^{N,*}_{\boldsymbol{\nu}}(y,u)$ pointwise.

2a. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{N,*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ converges uniformly to $\boldsymbol{\nu} \mapsto Q^{N,*}_{\boldsymbol{\nu}}(y,u)$.

3a. For each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, $\boldsymbol{\nu} \mapsto Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu}^*_n)}_{\boldsymbol{\nu}}(y,u)$ converges to $\boldsymbol{\nu} \mapsto Q^{N,*}_{\boldsymbol{\nu}}(y,u)$ pointwise.

4a. $\boldsymbol{\nu} \mapsto Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu}^*_n)}_{\boldsymbol{\nu}}(y,u)$ uniformly converges to $\boldsymbol{\nu} \mapsto Q^{N,*}_{\boldsymbol{\nu}}(y,u)$, uniformly over all $y \in \tilde{\mathcal{Y}}$ and $u \in U$.

We now show (1), the pointwise convergence of $\boldsymbol{\nu} \mapsto Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(t,y,u)$ to $\boldsymbol{\nu} \mapsto Q^{N,*}_{\boldsymbol{\nu}}(t,y,u)$. Fix $\boldsymbol{\nu}$, y and u. For any $n, N \in \mathbb{N}$, we have

$$\begin{aligned}
&|Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u) - Q^{*}_{\boldsymbol{\nu}}(y,u)| \\
\leq\ &|Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u) - Q^{N,*}_{\eta_n,\boldsymbol{\nu}}(y,u)| + |Q^{N,*}_{\eta_n,\boldsymbol{\nu}}(y,u) - Q^{N,*}_{\boldsymbol{\nu}}(y,u)| + |Q^{N,*}_{\boldsymbol{\nu}}(y,u) - Q^{*}_{\boldsymbol{\nu}}(y,u)|
\end{aligned} \tag{3.12}$$

By the successive approximations property, the finite horizon Q-functions converge pointwise to the infinite horizon Q-function counterparts. Hence, for any $\varepsilon > 0$, choose $N' \in \mathbb{N}$ such that the third term of (3.12) is smaller than $\frac{\varepsilon}{3}$ and $\sum_{m=N}^{\infty} \gamma^m M_r < \frac{\varepsilon}{6}$ for all $N \geq N'$. Now for the first term we have

$$\begin{aligned}
|Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u) - Q^{N,*}_{\eta_n,\boldsymbol{\nu}}(y,u)| &\leq \left| \sup_{\pi \in \Pi_{DM}} \mathbb{E}^{\pi} \left[\sum_{m=N}^{\infty} \gamma^m R^{\eta_n}\big(y_m, u_m, \boldsymbol{\mu}^{\nu}_m\big) \right] \right| \\
&= \left| \mathbb{E}^{\pi_{\mathrm{soft}}} \left[\sum_{m=N}^{\infty} \gamma^m R^{\eta_n}\big(y_m, u_m, \boldsymbol{\mu}^{\nu}_m\big) \right] \right| \\
&\leq \sum_{m=N}^{\infty} \gamma^m \big\{ M_r + \eta_n \big(\log|U| + \mathbb{E}[\log \pi_{\mathrm{soft}}(u_n \mid y_n)] \big) \big\} \\
&\leq \sum_{m=N}^{\infty} \gamma^m \{ M_r + \eta_n \log|U| \}
\end{aligned}$$

using the bound on the function $y = x \log x$. Then, given an $N \geq N'$, by (1a) we can choose $n' \in \mathbb{N}$ such that the second term of (3.12) is smaller than $\frac{\varepsilon}{3}$ and

$\sum_{m=N}^{\infty} \gamma^m \eta_n \log |U| < \frac{\varepsilon}{6}$. So that for all $n \geq n'$ we have

$$|Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(t,y,u) - Q^*_{\boldsymbol{\nu}}(t,y,u)| < \varepsilon$$

as required.

Next, to show (2), note that $Q^{N,*}_{\eta,\boldsymbol{\nu}}$ is monotonically decreasing in η, that is for any sequence $(\eta_n)_n$ such that $\eta_n \downarrow 0$, we have for each $n \in N$,

$$Q^{N,*}_{\eta_n,\boldsymbol{\nu}} \leq Q^{N,*}_{\eta_{n+1},\boldsymbol{\nu}}.$$

Hence sending $N \to \infty$ we also have

$$Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}} \leq Q^{\mathrm{reg},*}_{\eta_{n+1},\boldsymbol{\nu}}.$$

so that $Q^{\mathrm{reg},*}_{\eta,\boldsymbol{\nu}}$ is also monotonically decreasing in η. Moreover by (1), $\boldsymbol{\nu} \mapsto Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ converges to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$, which is continuous in $\boldsymbol{\nu}$ [6, Lemma 2]. Hence, by using Dini's theorem, we conclude the uniform convergence of $\boldsymbol{\nu} \mapsto Q^{\mathrm{reg},*}_{\eta_n,\boldsymbol{\nu}}(y,u)$ to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$ for each $y \in \tilde{\mathcal{Y}}$, $u \in U$.

To prove (3), the pointwise convergence of $\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}(y,u)$ to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$ for each $y \in \tilde{\mathcal{Y}}$ and $u \in U$, we proceed analogously as the proof as (1), utilising the successive approximations property and the finite horizon convergence (3a).

Finally, we show (4), the uniform convergence of $\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}(y,u)$ to $\boldsymbol{\nu} \mapsto Q^*_{\boldsymbol{\nu}}(y,u)$, uniformly over all $y \in \tilde{\mathcal{Y}}$ and $u \in U$. We demonstrate the equicontinuity of the family of functions

$$\left(\boldsymbol{\nu} \mapsto Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}(y,u)\right)_{n\in\mathbb{N}}.$$

As the reward function r_y is bounded, given $\varepsilon > 0$, we can find for some large N such that

$$\begin{aligned} &|Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu}^*_n)}_{\boldsymbol{\nu}}(y,u) - Q^{\Phi^{\mathrm{reg}}_{\eta_n}(\hat{\boldsymbol{\nu}}^*_n)}_{\hat{\boldsymbol{\nu}}}(y,u)| \\ \leq\ &|Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu}^*_n)}_{\boldsymbol{\nu}}(y,u) - Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\hat{\boldsymbol{\nu}}^*_n)}_{\hat{\boldsymbol{\nu}}}(y,u)| + \frac{\varepsilon}{2}. \end{aligned}$$

Then, by (4a) for sufficiently large n and $\delta_\infty(\boldsymbol{\nu},\hat{\boldsymbol{\nu}}) < \delta$,

$$|Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\boldsymbol{\nu})}_{\boldsymbol{\nu}}(y,u) - Q^{N,\Phi^{\mathrm{reg}}_{\eta_n}(\hat{\boldsymbol{\nu}})}_{\hat{\boldsymbol{\nu}}}(y,u)| < \frac{\varepsilon}{2}.$$

Finally we conclude uniform convergence (4) by appealing to the Arzelà-Ascoli theorem, noting that the space of measure flows is compact by Tychnoff's theorem. □

Recall that the softmax policy reads

$$\pi^{\text{soft}}_{\eta,\boldsymbol{\nu}}(u \mid y) \coloneqq \frac{\exp(Q^{\text{reg},*}_{\eta,\boldsymbol{\nu}}(y,u))}{\sum_{u' \in U} \exp(Q^{\text{reg},*}_{\eta,\boldsymbol{\nu}}(y,u'))}.$$

The uniform convergence of the Q-functions implies that the softmax policy converges to the argmax as the regulariser vanishes:

Lemma 3.31. *The softmax policy π^{soft} converges to the argmax π^*, where*

$$\pi^*(u \mid y) = \arg\max_{u \in U} Q^*_{\boldsymbol{\nu}}(y,u).$$

Proof. This follows from the uniform convergence of the Q-functions from Lemma 3.30 and the fact that the softmax function $f_c : \mathbb{R}^K \to [0,1]^K$

$$f_c(x)_i = \frac{\exp(c \cdot x_i)}{\sum_{i=1}^K \exp(c \cdot x_i)}, \quad x = (x_1, \ldots, x_K), \tag{3.13}$$

converges to the argmax f as $c \to \infty$, where $f(x)_k = 1$ if $x_k = \arg\max_i x_i$, and 0 otherwise. □

This leads to the following result showing approximate Nash for a sequence of regularised MFNE.

Theorem 3.32. *Let $(\eta_n)_n$ be a sequence with $\eta_n \downarrow 0$. For each n, let $(\pi^*_n, \boldsymbol{\nu}^*_n) \in \Pi_{DM} \times \Delta^\infty_{\mathcal{Y}}$ be the associated regularised MFNE, defined via Definition 3.21, for the MCDM-MFG. Then for any $\varepsilon > 0$, there exists $n', M' \in \mathbb{N}$ such that for all $n \geq n'$ and $M \geq M'$,*

$$J^M_j(\pi^*_n, \ldots, \pi^*_n) \geq \sup_{\pi^j \in \Pi_{DM}} J^M_j(\pi^*_n, \ldots, \pi^j, \ldots, \pi^*_n) - \varepsilon \quad \textit{for all } j \in \{1, \ldots, M\}.$$

Proof. We adapt of the proof of [23, Theorem 4], which shows the approximate Nash property for fully observable regularised MFGs in finite horizon. By the uniform convergence of the Q-functions in Lemma 3.30, we have that π^*_n converges to π^* where

$$\pi^*(u \mid y) = \arg\max_{u \in U} Q^*_{\boldsymbol{\nu}}(y,u)$$

by Lemma 3.31. By [23, Lemma B.8.11], the regularised policy is approximately optimal for the MFG: for any $\varepsilon > 0$, there exists $n' \in \mathbb{N}$ such that for all $n \geq n'$,

$$J_{\boldsymbol{\nu}^*_n}(\pi^*_n) \geq \max_{\pi \in \Pi_{DM}} J_{\boldsymbol{\nu}^*_n}(\pi) - \varepsilon.$$

By [23, Lemma B.5.6], if $\pi \in \Pi$ is an arbitrary policy and $\boldsymbol{\nu} = \Psi(\pi)$ the induced mean field, then for any sequence of policies $\{\pi^N\}_{N\in\mathbb{N}}$ we have

$$|J_1^N(\pi^N, \pi, \ldots, \pi) - J_{\boldsymbol{\nu}}(\pi^N)| \to 0.$$

Hence, we can choose a sequence of policies $\{\pi^M\}_{M\in\mathbb{N}}$ such that

$$\pi^M \in \operatorname*{arg\,max}_{\pi\in\Pi_{DM}} J_1^M(\pi, \pi_n^*, \ldots, \pi_n^*).$$

This allows us to conclude that for any $\varepsilon > 0$, there exists $n', M' \in \mathbb{N}$ such that for all $n \geq n'$, $M \geq M'$,

$$J_1^M(\pi_n^*, \ldots, \pi_n^*) \geq \max_{\pi\in\Pi_{DM}} J_1^M(\pi, \pi_n^*, \ldots, \pi_n^*) - \varepsilon \quad \text{for all } j \in \{1, \ldots, M\}$$

as desired. □

3.5 Numerical experiment in epidemiology

We demonstrate the MFG with speed of information with an epidemiology example, in the form of the SIS (susceptible-infected-susceptible) model. This is chosen as representative for a wide class of epidemiological models, including also the SIR (susceptible-infected-recovered) model and its many variants. The extension to many other variants is mathematically and computationally straightforward.

We adapt in particular the discrete-time version of the SIS model used as test case in [23]. In this model, a virus is assumed to be circulating amongst the population. Each agent can take on two states: susceptible (S), or infected (I). At each moment, the agent can decide to go out (U) or socially distance (D). Thus we have $\mathcal{X} = \{S, I\}$ and $A = \{U, D\}$. The probability of an agent being infected whilst going out is proportional to the fraction of infected people.[1] Once infected, they have a constant probability of recovering at each unit in time. We use the following parameters for the transition kernel:

$$\begin{aligned} p(S \mid I, U) &= p(S \mid I, D) = 0.2, \\ p(I \mid S, U) &= 0.9^2 \cdot \mu_t(I), \\ p(I \mid S, D) &= 0. \end{aligned}$$

[1] We will discuss later the feasibility of a more realistic extension where the probability of infection is proportional to the fraction of infected people *who also go out*.

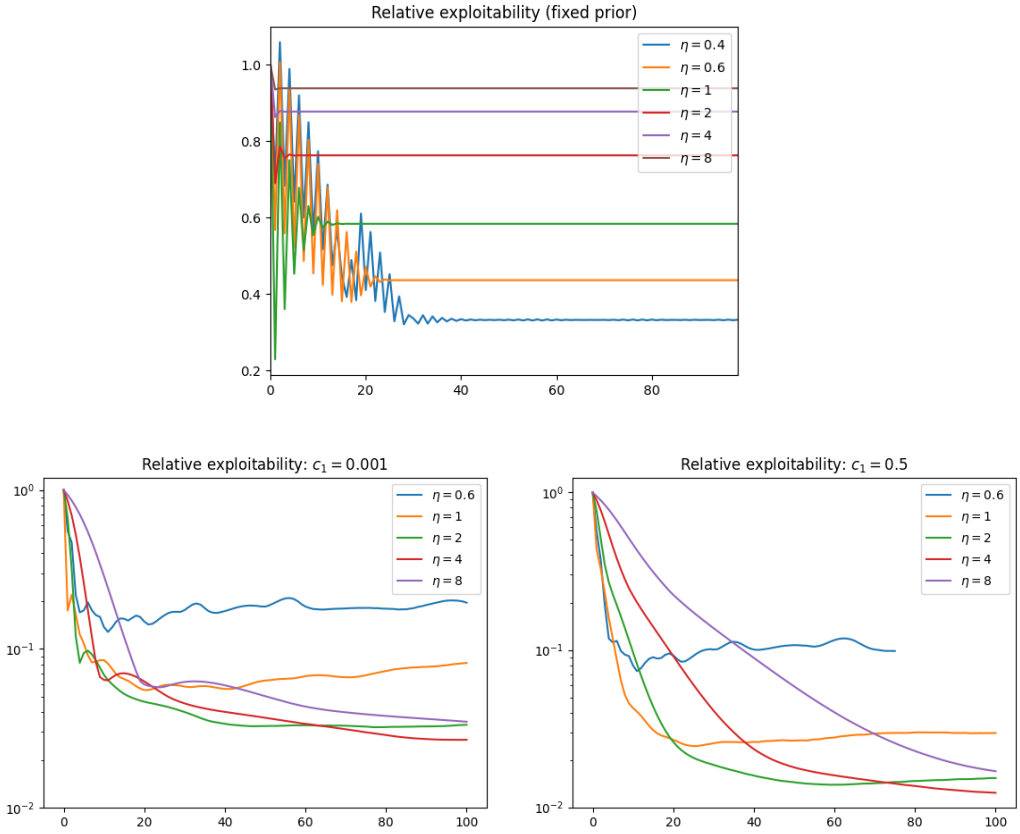

Figure 3.2: *Relative exploitability scores. Top: relative exploitability for* $c_1 = 0.5$ *when applied to a uniform policy as a reference measure, fixed across all iterations. Bottom row: relative exploitability for the prior descent algorithm for two different cost values* c_1*.*

Whilst healthy, there is a cost for socially distancing, and there are larger costs for being infected. As the infection rate in the SIS model does not depend on the proportion of people infected *and* going out, we adjust the reward to penalise this situation to reflect a desired behaviour of socially distancing whilst infected. Thus, the reward for our numerical example is given by

$$r(S,U) = 0, \quad r(S,D) = -0.3,$$
$$r(I,U) = -1.25, \quad r(I,D) = -1.0.$$

In addition to the standard model, we introduce the notion of test result times. Assume that during a pandemic, the population undergoes daily testing in order to determine whether they are infected or not. Here we assume the availability of two testing options, the free option which requires a 3-day turnaround, and a paid option

Algorithm 1: Prior descent for MFG with control of information speed

Input : Initial distribution $\boldsymbol{\nu}_0 \in \Delta_{\tilde{\mathcal{Y}}}$, prior policy $q \in \Pi_{DM}$, *tol*

Input : Number of iterations per loop I, regularisation parameter $\eta > 0$, truncation time T.

while $RelExpl > tol$ **do**

  for $i = 0, 1, \ldots, I$ **do**

    for $t = 0, 1, \ldots, T-1$ **do**

      Compute $\boldsymbol{\mu}_t^{\nu} = (\mu_{t,d})_{d=0}^{d_0} \in \Delta_{\mathcal{X}}^{d_0=1}$.

      Compute the regularised Q-function $Q_{\eta,\boldsymbol{\nu}}^{\mathrm{reg},*}(t, y, u)$ for fixed measures $\boldsymbol{\mu}_t^{\nu}$.

      Compute the softmax policy $\pi_{\eta,\boldsymbol{\nu},t}^{\mathrm{soft}} = \Phi_{\eta}^{\mathrm{reg}}(\boldsymbol{\nu})_t$.

      Compute the induced mean-field $\boldsymbol{\nu}_{t+1} = \Psi^{\mathrm{aug}}(\pi_{\eta,\boldsymbol{\nu},t}^{\mathrm{soft}})_{t+1}$.

    end

  end

  $q \leftarrow \pi_{\eta,\boldsymbol{\nu}}^{\mathrm{soft}}$.

end

which offers a next-day result. Thus we have for our model

$$\mathcal{D} = \{d_0 = 3, d_1 = 1\}, \quad \mathcal{C} = \{c_0 = 0, c_1 > 0\},$$

where we shall consider different values of c_1 in our numerical experiments.

For the computation of MFNEs for our model, we utilise the `mfglib` Python package [30]. We incorporate our own script for the mapping $\boldsymbol{\nu} \mapsto \boldsymbol{\mu}^{\nu}$ so that the existing library can be adapted for our MFG with control of information speed, and in particular to be computed on the augmented space. We first initialise with a uniform policy as the reference measure q, and repeatedly apply the mapping $\Psi^{\mathrm{aug}} \circ \Phi_{\eta}^{\mathrm{reg}}$ for a range of values of the regularisation parameter η. As a benchmark to test for the convergence towards a regularised MFNE, we utilise the exploitability score, which, for a policy π, is defined by

$$\mathrm{Expl}(\pi) := \max_{\tilde{\pi}} J_{\Psi^{\mathrm{aug}}(\tilde{\pi})}(\tilde{\pi}) - J_{\Psi^{\mathrm{aug}}(\pi)}(\pi).$$

The exploitability score measures the suboptimality gap for a policy π when computed with the measure flow induced by the map Ψ^{aug}. An exploitability score is 0 if and only if π is an MFNE for the MFG, and a score of ε indicates that π is an ε-MFNE. We refer to the literature such as [32, 46, 52] for a more detailed discussion.

As the exploitability score in general is dependent on the rewards and the initial policy, we consider instead the relative exploitability, which scales the exploitability score by

the initial value. We plot the convergence of the relative exploitability, with the uniform policy as reference measure, fixed across all iterations, in the bottom graph of Figure 3.2. We see that for lower values of η, the algorithm converges to a lower relative exploitability value. This corresponds to the fact that the regularised MFG with lower values of η approximates the unregularised MFG more closely. However, lower values of η require a larger number of iterations for convergence. For values of η smaller than 0.2, the algorithm does not even converge in our tests and explodes numerically (not plotted in the graph). This demonstrates an inherent limitation of the use of regularisation: one desires a sufficiently high value of η in order to guarantee convergence, but high values leads to a convergence to a regularised MFNE that approximates the original problem poorly. Moreover, searching for a suitable value of η is computationally expensive.

To mitigate the above issues, we utilise the prior descent algorithm [23]. Here, the reference measure is dynamically updated, by using the policy obtained from the previous iteration as the reference measure for the next iteration. The reference measure can also be updated after a number of iterations instead, creating a double loop for the algorithm. We summarise the prior descent algorithm for the MFG with control of information speed in Algorithm 1. The relative exploitability score is plotted in the top row of Figure 3.2. The score vastly outperforms the case of the fixed prior, and in general, we find that larger values of η require more iterations for convergence, but converge to a lower exploitability score. In [23], the prior descent algorithm is further improved by using the heuristic $\eta_{n+1} = \eta_n \cdot c$ for some constant $c > 1$, gradually increasing the regularisation to aid convergence. We also applied this heuristic for our problem, but for our case we do not see significant differences compared to initialising with large fixed values of η.

We now examine the MFNE for the SIS model with testing options for different values of c_1. Figure 3.3 corresponds to a low cost of $c_1 = 0.001$, whilst Figure 3.4 corresponds to a high cost of $c_1 = 0.05$. We compute the problem up to a terminal time of $T = 100$, but the truncate the graphs at $T = 75$, as the plots near terminal time are skewed by the artificially imposed terminal condition. The graphs on the right column depicts the behaviour in the choice of testing at equilibrium. The top-right shows the proportion of the population opting for the free test, whilst the bottom-right shows the optimal choice for testing for a healthy individual, given their current testing choice. We see a clear disparity across the two different costs. When the cost of the premium test is low, it is optimal to opt for this regime with probability 0.6,

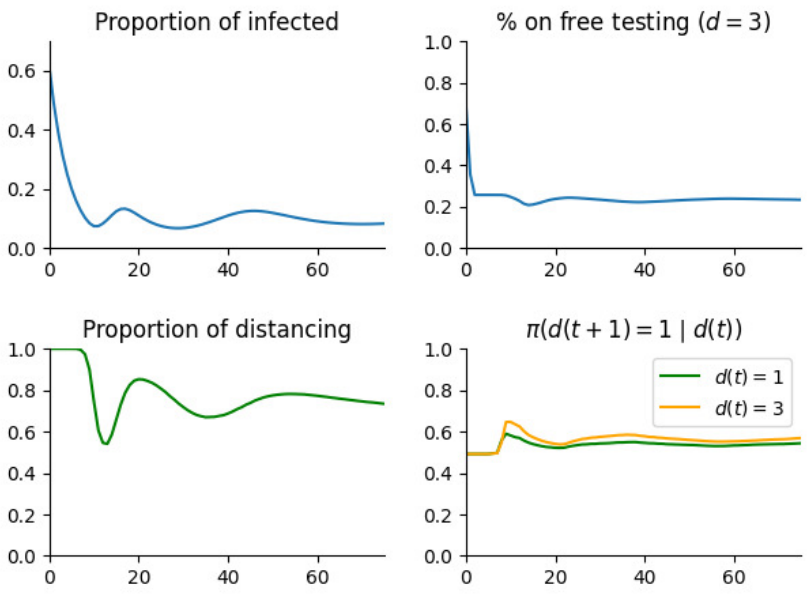

Figure 3.3: *Behaviour at MFNE for MCDM-MFG for cost* $c_1 = 0.001$.

and at equilibrium nearly 80% of the population chooses this option. In comparison, when the cost is high, it is optimal to choose the premium option only about 25% of the time, and less than half of the population use the premium option at equilibrium.

The left columns of Figure 3.3 and Figure 3.4 depict the population behaviour with social distancing, and the proportion of infected at equilibrium. At the beginning, there is a large number of infected people, so it is optimal to socially distance; once the proportion of infected is sufficiently low, a portion of the population starts to go out. This leads to a rebound in infection numbers, so that gradually the population socially distances again, and the cycle repeats. This is depicted by the periodic behaviours in the graph on the left.

For the case $c_1 = 0.001$, the low cost for premium testing leads to a higher percentage of the population with a more accurate estimation of their status. This leads to larger proportion of people going out, so that the infection occurs at a quicker rate. This in turn then leads to a quick rate of socially distancing, and so forth. This can be seen in the higher frequency of cycles in the infected and distancing graphs of Figure 3.3. Interestingly, as the proportion of infected stabilises as time passes to a similar value, regardless of the cost for the premium testing. The difference lies in the initial periods of peaks and troughs in infected numbers.

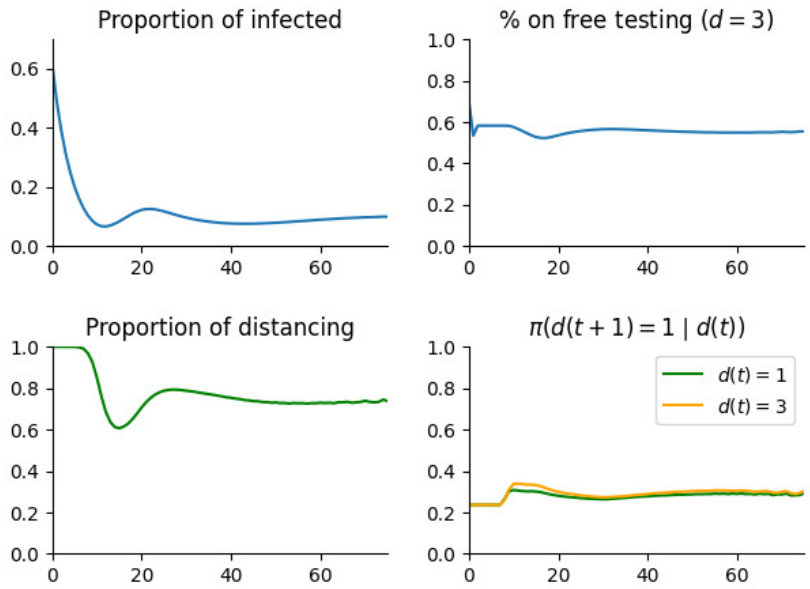

Figure 3.4: *Behaviour at MFNE for MCDM-MFG for cost* $c_1 = 0.05$.

Chapter 4

Conclusion and Outlook

This thesis contributes to the literature in partial information control, in particular of actively controlled observations. In the preceding chapters, we studied MDP frameworks with observation costs and dynamically controlled observation delays. We exploit the information structure to reduce both partially observable problems to a finite MDP, which enables straightforward numerical schemes to compute for approximate solutions. For the observation cost model, we showed that dynamic programming leads to a novel class of QVIs, which are structurally different to typical Bellman-type equations. For this class of QVIs, we utilise a penalty method as an approximation scheme to efficiently solve for the system of equations. For dynamically controlled observation delay, we extended the concept of actively controlled observations to a mean-field game setup, and formulated sufficient conditions to achieve a contraction with entropy regularisation.

We see the models explored in this thesis as a stepping stone towards the relatively unexplored literature in observation controls, particularly in continuous time. The simplicity of the models in discrete time allows more feasible analysis and computation without obfuscating intuition. There are several strands of further research directions that form a natural continuation of the material in this thesis. We shall discuss these aspects in the remainder of this chapter.

4.1 Continuous-time framework

For observation controls in continuous-time diffusions, we can proceed as follows. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_t, \mathbb{P})$ be a filtered probability space. Assume the state space $\mathcal{X} = \mathbb{R}$ and the

action space A is finite. Let $W = (W_t)_t$ be an $(\mathcal{F}_t)$- Brownian motion on $\mathbb{R}$. Consider the SDE

$$dX_t = \mu(X_t, \alpha_t)\ dt + \sigma(X_t, \alpha_t)\ dW_t, \tag{4.1}$$

where the control process $\alpha = (\alpha_t)_t$ is $(\mathcal{F}_t)_t$-adapted, $\mu : \mathbb{R} \times A \to \mathbb{R}$ and $\sigma : \mathbb{R} \times A \to \mathbb{R}$ satisfying the standard Lipschitz and growth conditions: there exists constants $K, M > 0$ such that for all $x, y \in \mathbb{R}$ and for all $a \in A$,

$$|\mu(x, a) - \mu(y, a)| + |\sigma(x, a) - \sigma(y, a)| \leq K|x - y|,$$

and

$$|\mu(x, a)| + |\sigma(x, a)| \leq M(1 + |x|).$$

For the SDE (4.1), we denote the corresponding collection of transition kernels by $\{P_t\}_t$.

The equivalent notion to the inspection variables in discrete time is given by stopping times with respect to a smaller filtration. This follows a similar definition to that in [25].

Definition 4.1. Given the SDE (4.1), let $\hat{\tau} = \{\tau_k\}_{k=0}^{\infty}$ be a sequence of strictly increasing random times, with the convention that $\tau_0 = 0$. Define for all $t \geq 0$,

$$\mathcal{F}_t^{(X,\hat{\tau})} := \sigma\big\{(\tau_j, X_{\tau_j}\mathbb{1}_{\{\tau_j \leq t\}}) : j \geq 0\big\}$$

If for each $\tau_k \in \hat{\tau}$, τ_k is a *predictable* $\mathcal{F}^{(X,\hat{\tau})}$-stopping time, then $\mathcal{F}^{(X,\hat{\tau})}$ is called an **X-observation filtration**.

Then the admissible controls is given by the following.

Definition 4.2. A control $\alpha = (\alpha_t)_t$ is admissible for the OCM if it is of the form

$$\alpha_t = \sum_{n=0}^{\infty} a_n \mathbb{1}_{[\tau_n, \tau_{n+1})}(t),$$

where $\hat{\tau} = \{\tau_n\}_{n=0}^{\infty}$ are *predictable* $\mathcal{F}^{(X,\hat{\tau})}$-stopping times and each a_n is an A-valued, $\mathcal{F}_{\tau_n}^{(X,\hat{\tau})}$-measurable random variable. The set of admissible controls is denoted by $\mathcal{A}$.

Then the goal is to seek a control $\alpha \in \mathcal{A}$ to maximise

$$\mathbb{E}\left[\int_0^{\infty} \gamma^t f(X_t, \alpha_t)\ dt - \sum_{\tau_n \in \hat{\tau}} e^{-\gamma\tau_n} \cdot c_{\text{obs}}\right],$$

where $f : \mathbb{R} \times A \to \mathbb{R}$ is the reward, such that $\mathbb{E}[\int_0^\infty \gamma^t f(X_t, \alpha_t)\, dt] < \infty$.

Proving a dynamic programming principle for the value function, together with the well-posedness and comparison principle of the associated equations, will likely require analysis on the filtered process, taking values on the space of measures $\mathcal{P}(\mathcal{X})$. A possible approach is to consider the filtered problem as a non-standard impulse control problem. At time t, the filter process Y_t represents the conditional expectation of X_t given the observation history. As no new observations occur between observation times, $Y = (Y_t)_t$ should formally satisfy

$$dY_t = LY_t dt, \quad \tau_j \leq t < \tau_{j+1},$$

where L is the generator for the SDE (4.1). Then, at the observation time τ_j, Y_{τ_j} would receive an impulse, given by

$$Y_{\tau_j} = \Gamma(\tau_j, Y_{\tau_j}^-),$$

where $\Gamma(\tau_j, Y_{\tau_j}^-)$ is a random variable with values on $\Delta_{\mathcal{X}}$, with distribution proportional to the transition kernel $P_{\tau_j - \tau_{j-1}}(X_{\tau_{j-1}}, X_{\tau_j})$. The random 'impulse' represents the change in the filter process from the new observation X_{τ_j}. This is a non-standard problem, as the impulse is random and the dynamics between observations is deterministic. To start, one can reference techniques from existing theory on impulse control analysis can be utilised to formulate the dynamic programming principle (DPP) [15, 71]. However, it is likely that conditions on either the dynamics of the SDE or conditions on the state space will have to be imposed, as the finiteness of the state space $\mathcal{X}$ in the discrete-time case was a crucial assumption for proving the comparison principle.

Assuming the DPP, one should be able to establish its equivalence to a finite integro-differential QVI, formally written in the form

$$\begin{aligned} \min \Big\{ & -\partial_s v_i(s,x) + \gamma v_i(s,x) - \mathbb{E}_i\left[f_i(X_s^x)\right], \\ & v_i(s,x) - \Big(\mathbb{E}_i[\max_{j \in A} v_j(0, X_s^x)] - (g_{ij} + c_{\text{obs}})\Big)\Big\} = 0, \end{aligned} \tag{4.2}$$

where $s \geq 0$, $x \in \mathcal{X}$, and $i \in A$ are the time passed since the last observation, the state and the action applied at the last observation respectively.

The discrete QVI (2.25) can then be seen as a discretisation of the QVI (4.2), by the use of finite differences on the differential operators, as well as on the Kolomogorov

Backwards equations for the integral terms: by writing $e_i^s(0,x) = \mathbb{E}_i[f_i(X_s^x)]$, we have that e_i^s satisfies

$$\partial_t e_i^s(t,x) + \frac{1}{2}\sigma^2(x,i)\partial_x^2 e_i^s(t,x) + \mu(x,i)\partial_x e_i^s(t,x) = 0,$$
$$e_i^s(s,x) = f_i(x), \quad 0 < t < s,\ x \in \mathbb{R},\ i \in A.$$

Then, consider a truncated state space, restricted to some interval $[x_{\min}, x_{\max}]$. Let $x = (x_l)_{0\le l\le L}$, $t = (t_n)_{0\le n\le N}$, and let $x_{l+1} - x_l = h$ and $t_{n+1} - t_n = k$. Using an implicit discretisation, we obtain a system of equations:

$$\frac{e_{i,l}^{s,n+1} - e_{i,l}^{s,n}}{k} + \frac{1}{2}\sigma^2(x_l,i)\frac{e_{i,l+1}^{s,n} - 2e_{i,l}^{s,n} + e_{i,l-1}^{s,n}}{h^2} + \mu(x_l,i)\frac{e_{i,l+1}^{s,n} - e_{i,l-1}^{s,n}}{2h} = 0, \qquad (4.3)$$

with terminal condition

$$e_{i,l}^{s,N} = f_i(x_l) =: f_{i,l}^h.$$

where $N = \frac{s}{k}$. Our desired quantity is then the vector $e_i^{s,0} = \left(e_{i,l}^{s,0}\right)_l$, which can be computed by solving the system of equations

$$\left(A_{h,k}^i\right)^N e_i^{s,0} = f_i^h,$$

where the matrix

$$A_{h,k}^i = \begin{pmatrix} b & c & & 0 \\ a & \ddots & \ddots & \\ & \ddots & \ddots & c \\ 0 & & a & b \end{pmatrix}$$

is tridiagonal with

$$a = \frac{k(h\mu(x_l,i) - \sigma^2(x_l,i))}{2h^2}, \quad b = 1 + \frac{k\sigma^2(x_l,i)}{h^2}, \quad c = -\frac{kh(\mu(x_l,i) + \sigma^2(x_l,i))}{2h^2}.$$

Omitting the indices for A for ease of notation, one obtains the discretisation of (4.2):

$$\min\left\{v_{i,l}^n - \frac{1}{1+k\gamma}\left(v_{i,l}^{n+1} + k\left(A^{-n}f_i^h\right)_l\right),\ v_{i,l}^n - \left(\left(A^{-n}\max_{j\in A} v_j^{0,h}\right)_l - c_{\text{obs}}\right)\right\} = 0.$$

A possible route to demonstrate the convergence of QVIs (2.25) to (4.2) can involve an extension of a classical result by Barles and Souganidis [12], which provides the necessary requirements for the convergence of numerical schemes to the viscosity solutions of PDEs. For the QVIs in the observation cost model, this would involve first defining the relevant notions of viscosity solutions to (4.2), then considering the Barles–Souganidis framework for integro-differential equations of the form

$$F(s,x,u,Du,\mathcal{I}[s,x,u]) = 0,$$

where $\mathcal{I}$ is an integral operator. In order to apply the framework, monotonicity, stability properties and the comparison principle for (4.2) will need to be proven.

An alternative method to show convergence is to approximate the SDE (4.1) with a suitably scaled Markov chain. In classical stochastic control, the value function can be approximated by a Markov chain under a suitable discretisation of space and time [12, 44]. The discretised value functions converge towards the value function as the step size decreases towards zero. Suppose $\{\xi_n^h\}_n$ is an MDP (under the observation cost framework) with control $\{\pi_n^h\}_n$, parametrised by h. Let $\{t_n^h\}_n$ be the interpolated time points with the intervals $\Delta t_n^h = \Delta(\xi_n^h, \pi_n^h)$. Let $t_0 = s > 0$, then the approximated reward functional becomes

$$J^h(s;(x,i);\pi) = \mathbb{E}\left[\sum_{n=0}^{\infty} f(\xi_n^h, \pi_n^h) e^{-\gamma(t_n^h - s)} \Delta t_n^h - \sum_{\tau_n \geq s} e^{-\gamma(\tau_n - s)} \cdot c_{obs}\right],$$

with value function

$$v^h(s,x,i) = \sup_{\mathcal{A}^h} J^h(s;(x,i);\alpha(\cdot)),$$

where $\mathcal{A}^h$ is the set of admissible controls for the discretisation parameter h. Under the natural scaling obtained from discretising the Kolmogorov backwards equation (4.3), we expect that $v_i^h \to v_i$ as $h \to 0$ for each $i \in A$.

Further open questions:

- Regularity of the value function – Figure 2.12 suggests that the continuous-time value function belongs to $C^1(\mathcal{X})$, and moreover is C^2 away from the free boundary. We may able to take advantage of existing regularity results for standard impulse control problems [51] to demonstrate this.
- Asymptotic behaviour as the observation cost $c_{\text{obs}} \downarrow 0$ – in impulse control problems, the value function is continuous but not differentiable with respect to the intervention cost at 0 [51]. It remains to be seen if the same holds for the continuous time observation cost model.

4.2 Costly switching between different observation streams

The observation control models in Chapters 2 and 3 have the benefit of a finite characterisation of the belief state. This allows for a tractable computation of the value functions. A more general framework would be an observation cost MDP where the

agent can switch between two general streams of information for different observation cost. Specifically, let $\mathcal{X}$ be the underlying space and $\mathcal{Y}$ be the observation space. Suppose that the agent observes $y_0, \dots, y_{t-1} \in \mathcal{Y}$ at time t, but by paying c_{obs}, they can choose to observe exactly x_t instead. Similar to general POMDP models, the relation between x_t and y_t can be given by

$$y_t = h(x_t) + \varepsilon_t,$$

for some $h : \mathcal{X} \to \mathcal{Y}$ and i.i.d. random variables $\{\varepsilon_t\}_t$, representing the noise within the observations. Thus, the belief MDP is given by

$$z_t \coloneqq \mathbb{P}^\pi(x_t \mid y_0, \dots, y_t, x_{\tau_1}, \dots, x_{\tau_k}, u_0, \dots, u_{t-1}),$$

where $\{\tau_n\}_{n=1}^k$ are the observation times between times 0 and t, and $u_t = (a_t, i_t) \in A \times \mathcal{I}$. Unfortunately, in this case the belief MDP cannot be reduced to a finite MDP, unless more information is given about the function h.

In this case, suitable approximations will be required such that the approximating problem is tractable. First, one can discretise the belief state and consider its convergence to the original problem as the mesh size tends to 0 [39]. Another difficulty arises from the computation of the belief state. The belief state depends on the entire observation sequence, which grows linearly with time, and is therefore tractable only for small time horizons. Introducing a finite memory property to the problem is a possible approach from a practical standpoint. That is, assuming that agents are unable to remember and process arbitrarily long observation sequences to compute the belief state, we impose a memory cap of N units. Thus we consider an approximation

$$\tilde{z}_t \coloneqq \mathbb{P}^\pi(x_t \mid y_{t-N}, \dots, y_t, x_{\tau_r}, \dots, x_{\tau_k}, u_{t-N}, \dots, u_{t-1}),$$

where $\{\tau_r, \dots, \tau_k\}$ are the observation times between times $t-N$ and t. Finite memory for classical POMDPs were studied in [65], and extended to the Q-learning problem in [40], where a bound on the expected suboptimal gap is given. This can be a starting point for the analysis of the observation switching model above to give bounds on its suboptimal gaps.

Further open questions:

– Suppose that the observation process satisfies instead

$$y_t = h(x_t) + \lambda\varepsilon_t,$$

for some $\lambda > 0$. Sending $\lambda \to \infty$ would correspond to the case that no information about x_t is inferred through observing y_t. It will be beneficial to show that the observation cost model studied in Chapter 2 can be considered as an asymptotic limit of $\lambda \to \infty$ of this proposed general framework.

– Due to the Markov property, the dependence of the belief state on the observation sequence would start from the time of the most recent observation of the underlying state. A lower value of c_{obs} would likely lead to more frequent observations, which acts as an inherent bound on the length of memory required. It might be of practical interest to investigate if the finite window N can be foregone for sufficiently low values of c_{obs}.

Bibliography

[1] A. B. Abel, J. C. Eberly, and S. Panageas. Optimal inattention to the stock market with information costs and transactions costs. *Econometrica*, 81(4):1455–1481, 2013.

[2] S. Adlakha, S. Lall, and A. Goldsmith. Information state for Markov decision processes with network delays. In *2008 47th IEEE Conference on Decision and Control*, pages 3840–3847. IEEE, 2008.

[3] S. Adlakha, S. Lall, and A. Goldsmith. Networked Markov decision processes with delays. *IEEE Trans. Automat. Control*, 57(4):1013–1018, 2011.

[4] E. Altman and P. Nain. Closed-loop control with delayed information. SIGMETRICS '92/PERFORMANCE '92, page 193–204, New York, NY, USA, 1992. Association for Computing Machinery.

[5] B. Anahtarcı, C. Kariksiz, and N. Saldi. Value iteration algorithm for mean-field games. *Systems Control Lett.*, 143:104744, 2020.

[6] B. Anahtarcı, C. Kariksiz, and N. Saldi. Q-learning in regularized mean-field games. *Dyn. Games Appl.*, pages 1–29, 05 2022.

[7] R. F. Anderson and A. Friedman. Optimal inspections in a stochastic control problem with costly observations. *Math. Oper. Res.*, 2(2):155–190, 1977.

[8] R. F. Anderson and A. Friedman. Optimal inspections in a stochastic control problem with costly observations, II. *Math. Oper. Res.*, 3(1):67–81, 1978.

[9] P. Azimzadeh and P. A. Forsyth. Weakly chained matrices, policy iteration, and impulse control. *SIAM J. Numer. Anal.*, 54(3):1341–1364, 2016.

[10] A. Bain and D. Crisan. *Fundamentals of Stochastic Filtering*. Stochastic Modelling and Applied Probability. Springer New York, 2009.

[11] J. L. Bander and C. White. Markov decision processes with noise-corrupted and delayed state observations. *J. Oper. Res. Soc.*, 50:660–668, 1999.

[12] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptot. Anal.*, 4(3):271–283, 1991.

[13] E. Bayraktar, E. Ekström, and J. Guo. Disorder detection with costly observations. *J. Appl. Probab.*, 59(2):338–349, 2022.

[14] E. Bayraktar and R. Kravitz. Quickest detection with discretely controlled observations. *Sequential Anal.*, 34(1):77–133, 2015.

[15] C. Belak, S. Christensen, and F. T. Seifried. A general verification result for stochastic impulse control problems. *SIAM Journal on Control and Optimization*, 55(2):627–649, 2017.

[16] C. Bellinger, R. Coles, M. Crowley, and I. Tamblyn. Active measure reinforcement learning for observation cost minimization. *arXiv:2005.12697*, 2020.

[17] C. Bellinger, A. Drozdyuk, M. Crowley, and I. Tamblyn. Balancing information with observation costs in deep reinforcement learning. In *Proceedings of the 35th Canadian Conference on Artificial Intelligence*. CAIAC, 2022.

[18] B. Bruder and H. Pham. Impulse control problem on finite horizon with execution delay. *Stochastic Process. Appl.*, 119(5):1436–1469, 2009.

[19] P. E. Caines, M. Huang, and R. P. Malhamé. Large population stochastic dynamic games: closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3):221–252, 2006.

[20] Á. Cartea and L. Sánchez-Betancourt. Optimal execution with stochastic delay. *Finance and Stoch.*, 27(1):1–47, 2023.

[21] B. Chen, M. Xu, L. Li, and D. Zhao. Delay-aware model-based reinforcement learning for continuous control. *Neurocomputing*, 450:119–128, 2021.

[22] C. Cooper and N. Hahi. An optimal stochastic control problem with observation cost. *IEEE Trans. Automat. Control*, 16(2):185–189, 1971.

[23] K. Cui and H. Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

[24] R. C. Dalang and A. N. Shiryaev. A quickest detection problem with an observation cost. *Ann. Appl. Probab.*, 25(3):1475–1512, 2015.

[25] H. Dyrssen and E. Ekström. Sequential testing of a Wiener process with costly observations. *Sequential Anal.*, 37(1):47–58, 2018.

[26] P. A. Forsyth and K. R. Vetzal. Quadratic convergence for valuing American options using a penalty method. *SIAM J. Sci. Comput.*, 23(6):2095–2122, 2002.

[27] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

[28] H.-O. Georgii. Gibbs measures and phase transitions. In *Gibbs Measures and Phase Transitions*. de Gruyter, 2011.

[29] N. Guo and V. Kostina. Optimal causal rate-constrained sampling for a class of continuous Markov processes. *IEEE Trans. Inform. Theory*, 67(12):7876–7890, 2021.

[30] X. Guo, A. Hu, M. Santamaria, M. Tajrobehkar, and J. Zhang. MFGLib: A library for mean field games. *arXiv:2304.08630*, 2023.

[31] X. Guo, A. Hu, R. Xu, and J. Zhang. Learning mean-field games. *Adv. Neural. Inf. Process. Syst.*, 32, 2019.

[32] X. Guo, A. Hu, and J. Zhang. MF-OMO: An optimization formulation of mean-field games. *arXiv preprint arXiv:2206.09608*, 2022.

[33] B. Hajek, K. Mitzel, and S. Yang. Paging and registration in cellular networks: Jointly optimal policies and an iterative algorithm. *IEEE Trans. Inform. Theory*, 54:608 – 622, 2008.

[34] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Applied mathematical sciences. Springer-Verlag, 1989.

[35] O. Hernández-Lerma and J.-B. Lasserre. *Discrete-Time Markov Control Processes*. Springer New York, 1996.

[36] C. Huang and S. Wang. A power penalty approach to a nonlinear complementarity problem. *Oper. Res. Lett.*, 38(1):72–76, 2010.

[37] Y. Huang and Q. Zhu. Self-triggered Markov decision processes. In *Proc. 60th IEEE CDC*, pages 4507–4514, 2021.

[38] K. Ito and K. Kunisch. Parabolic variational inequalities: The lagrange multiplier approach. *J. Math. Pures Appl.*, 85(3):415–449, 2006.

[39] A. D. Kara, E. Bayraktar, and S. Yüksel. Near optimality of finite memory policies for POMDPs with continuous spaces. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2301–2306. IEEE, 2022.

[40] A. D. Kara and S. Yüksel. Convergence of finite memory Q learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*, 2022.

[41] K. Katsikopoulos and S. Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE Trans. Automat. Control*, 48(4):568–574, 2003.

[42] D. Krueger, J. Leike, O. Evans, and J. Salvatier. Active reinforcement learning: Observing rewards at a cost. *arXiv:2011.06709*, 2020.

[43] H. Kushner. On the optimum timing of observations for linear control systems with unknown initial state. *IEEE T. Automat. Contr.*, 9(2):144–150, 1964.

[44] H. J. Kushner and P. G. Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, Berlin, Heidelberg, 1992.

[45] J.-M. Lasry and P.-L. Lions. Mean field games. *Jpn. J. Math.*, 2(1):229–260, 2007.

[46] M. Laurière, S. Perrin, M. Geist, and O. Pietquin. Learning mean field games: A survey. *arXiv:2205.12944*, 2022.

[47] L. Meier, J. Peschon, and R. Dressler. Optimal control of measurement subsystems. *IEEE T. Automat. Contr.*, 12(5):528–536, 1967.

[48] S. Nath, M. Baranwal, and H. Khadilkar. *Revisiting State Augmentation Methods for Reinforcement Learning with Stochastic Delays*, page 1346–1355. Association for Computing Machinery, New York, NY, USA, 2021.

[49] A. Nayyar, T. Başar, D. Teneketzis, and V. V. Veeravalli. Optimal strategies for communication and remote estimation with an energy harvesting sensor. *IEEE Trans. Automat. Control*, 58(9):2246–2260, 2013.

[50] B. Øksendal and A. Sulem. Optimal stochastic impulse control with delayed reaction. *Applied Mathematics and Optimization*, 58:243–255, 2008.

[51] B. Øksendal and A. Sulem. Stochastic control of jump diffusions stochastic control. In *Applied Stochastic Control of Jump Diffusions*, pages 93–155. Springer, 2019.

[52] J. Pérolat, S. Perrin, R. Elie, M. Laurière, G. Piliouras, M. Geist, K. Tuyls, and O. Pietquin. Scaling mean field games by online mirror descent. page 1028–1037. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[53] S. Perrin, J. Pérolat, M. Laurière, M. Geist, R. Elie, and O. Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

[54] H. Pham. *Continuous-Time Stochastic Control and Optimization with Financial Applications*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[55] C. Reisinger and J. Tam. Markov decision processes with observation costs: framework and computation with a penalty scheme. *arXiv preprint arXiv:2201.07908*, 2022.

[56] C. Reisinger and Y. Zhang. A penalty scheme for monotone systems with interconnected obstacles: Convergence and error estimates. *SIAM J. Numer. Anal.*, 57(4):1625–1648, 2019.

[57] C. Reisinger and Y. Zhang. A penalty scheme and policy iteration for nonlocal hjb variational inequalities with monotone nonlinearities. *Comput. Math. Appl.*, 93:199–213, 2021.

[58] J. Rust. Numerical dynamic programming in economics. *Handbook of computational economics*, 1:619–729, 1996.

[59] N. Saldi, T. Başar, and M. Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM J. Control Optim.*, 56(6):4256–4287, 2018.

[60] N. Saldi, T. Başar, and M. Raginsky. Approximate Nash equilibria in partially observed stochastic games with mean-field interactions. *Math. Oper. Res.*, 44(3):1006–1033, 2019.

[61] N. Saldi, T. Başar, and M. Raginsky. Partially observed discrete-time risk-sensitive mean field games. *Dyn. Games Appl.*, pages 1–32, 2022.

[62] Y. F. Saporito and J. Zhang. Stochastic control with delayed information and related nonlinear master equation. *SIAM J. Control Optim.*, 57(1):693–717, 2019.

[63] E. Schuitema, L. Buşoniu, R. Babuška, and P. Jonker. Control delay in reinforcement learning for real-time dynamic systems: A memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3226–3231. IEEE, 2010.

[64] V. Tzoumas, L. Carlone, G. J. Pappas, and A. Jadbabaie. LQG control and sensing co-design. *IEEE Trans. Automat. Control*, 66(4):1468–1483, 2020.

[65] C. C. White III and W. T. Scherer. Finite-memory suboptimal design for partially observed markov decision processes. *Operations Research*, 42(3):439–455, 1994.

[66] S. Winkelmann. *Markov Decision Processes with Information Costs*. PhD thesis, Freie Universität Berlin, Berlin, 2013.

[67] S. Winkelmann, C. Schütte, and M. v. Kleist. Markov control processes with rare state observation: Theory and application to treatment scheduling in HIV–1. *Commun. Math. Sci.*, 12(5):859–877, 2014.

[68] J. H. Witte and C. Reisinger. A penalty method for the numerical solution of Hamilton–Jacobi–Bellman (HJB) equations in finance. *SIAM J. Numer. Anal.*, 49(1):213–231, 2011.

[69] J. H. Witte and C. Reisinger. Penalty methods for the solution of discrete HJB equations—continuous control and obstacle problems. *SIAM J. Numer. Anal.*, 50(2):595–625, 2012.

[70] W. Wu and A. Arapostathis. Optimal sensor querying: General Markovian and LQG models with controlled observations. *IEEE Trans. Automat. Control*, 53(6):1392–1405, 2008.

[71] J. Yong. Systems governed by ordinary differential equations with continuous, switching and impulse controls. *Applied Mathematics and Optimization*, 20:223–235, 1989.

[72] H. Yoshioka and M. Tsujimura. Analysis and computation of an optimality equation arising in an impulse control problem with discrete and costly observations. *J. Comput. Appl. Math.*, 366:112399, 2020.

[73] H. Yoshioka, M. Tsujimura, K. Hamagami, and Y. Yoshioka. A hybrid stochastic river environmental restoration modeling with discrete and costly observations. *Optimal Control Appl. Methods*, 41(6):1964–1994, 2020.

[74] H. Yoshioka, Y. Yaegashi, M. Tsujimura, and Y. Yoshioka. Cost-efficient monitoring of continuous-time stochastic processes based on discrete observations. *Appl. Stoch. Models Bus. Ind.*, 37(1):113–138, 2021.

[75] H. Yoshioka, Y. Yoshioka, Y. Yaegashi, T. Tanaka, M. Horinouchi, and F. Aranishi. Analysis and computation of a discrete costly observation model for growth estimation and management of biological resources. *Comput. Math. Appl.*, 79(4):1072–1093, 2020.