Radboud University Nijmegen

Faculty of Science

Average reward and Blackwell optimality in Robust POMDPs

Master Thesis Computing Science

Daily supervisor:
Eline Bovy, MSc
eline.bovy@ru.nl

Author:
Niels van Duijl
s1153708

First supervisor/assessor:
prof. dr. Nils Jansen
n.jansen@cs.ru.nl

Second assessor:
dr. Sebastian Jungens
sebastian.junges@ru.nl

June 2026

Abstract

Robust partially observable Markov decision processes (RPOMDPs) seek a policy that is optimal, for some predefined goal, in the worst case over an uncertain model, under partial observability of the state. Together, partial observability and model uncertainty capture the imperfect information and limited model knowledge that arise when modelling real-world sequential decision-making problems. For instance, in medical treatment planning, decisions are based on incomplete patient observations and an uncertain model of the patient's response to treatment. Existing work on RPOMDPs has focused on the finite horizon and discounted infinite horizon settings, leaving average reward and Blackwell optimality (a notion of optimality across all discount factors sufficiently close to 1) unexplored. Both objectives are already non-trivial in the special cases of partially observable MDPs (POMDPs) and robust MDPs (RMDPs): average optimal policies are not guaranteed to exist in general POMDPs, and there exist RMDP instances without a Blackwell optimal policy. For POMDPs, exact Blackwell optimality even remains an open question. However, sufficient conditions are known for average optimality in both POMDPs and RMDPs, and for Blackwell optimality in RMDPs.

In this thesis, we investigate whether such results can be combined to obtain average and Blackwell optimality results for RPOMDPs. We show that an average optimal policy exists within the class of agent policies given by mixed finite-state controllers (FSCs) with a bounded memory size, requiring no further assumptions on the uncertainty set. We further address Blackwell optimality in POMDPs, a previously open question, and give a sufficient condition under which Blackwell optimal policies are guaranteed to exist. Finally, we take first steps towards Blackwell optimality in RPOMDPs, showing that whenever a deterministic FSC is discount optimal, a deterministic Blackwell optimal FSC exists within the class of mixed FSCs, given the same bound on memory size.

Contents

1 Introduction

Markov decision processes (MDPs) are a standard mathematical model for sequential decision-making in probabilistic environments [1]. An agent chooses actions that influence its reward and the future state of the environment. The environment transitions to a new state according to a probabilistic transition function that depends on the current state and chosen action. The agent chooses its actions based on a so-called *policy*, a function that selects an action at every time step. The goal is to find an optimal policy for some predefined goal, for example maximising the expected reward. MDPs are useful for modelling real-world problems of many types, e.g. a spoken dialogue system, where the system gets voice input from the user and must quickly determine how to react to satisfy the user [2]. They are also applicable to many problems in healthcare, such as epidemic control, transplantations, and cancer treatment [3].

The standard MDP framework relies on the assumption that the agent can observe the full state of the environment. Defining an MDP also requires precise knowledge of the transition probabilities, as small changes can have a big impact on the optimal policy [4]. When we try to model real-world problems as MDPs, we often only know the transition probabilities approximately [4]. Furthermore, we may not always be able to fully observe the state of the system [5]. To deal with these problems, various generalisations of MDPs have been studied.

One of these generalisations is the partially observable MDP (POMDP), which relaxes the assumption of full observability, allowing the agent to receive observations that do not reveal the entire state [5]. Instead, by keeping track of the history of actions and observations, the agent maintains a *belief*: a probability distribution over the possible states. For example, in healthcare, where the state of the MDP corresponds to the state of the patient, it is not realistic to observe the entire state. Instead, medical tests give some observations about this state, but don't reveal it fully. We can turn a POMDP into a fully observable MDP with continuous state space, called the belief MDP, where the beliefs of the agent serve as the states of the MDP.

Another generalisation is the Robust MDP (RMDP), which relaxes the assumption that the transition dynamics are known exactly, by allowing them to range over some uncertainty set [6]. The typically considered objective is to compute the policy that maximises the reward in the worst case over this set. We can then simplify the problem by making assumptions about *rectangularity*, which relax possible constraints in the uncertainty set by assuming independence of probabilities of, for example, different state-action pairs. Again, take the healthcare example, where actions now correspond to different treatments of a disease. We do not know the exact probability that the treatment will work, as this depends on the patient. It is desirable to look at the worst-case scenario, aiming for a treatment with a guarantee for all patients. The chance of success of the same treatment in different stages of the disease (states of the MDP) can depend on common factors of the patient, but rectangularity assumptions neglect these constraints.

Robust partially observable MDPs (RPOMDPs) combine both extensions, addressing settings in which the agent must act under uncertainty about both the current state of the environment, and its transition and observation dynamics.

For any of these types of MDP problems, there are different goals that we can choose to optimise for. We focus on maximisation of expected reward, for which there are a few different measures. If we look at a finite time horizon, i.e. the system stops after a certain time, it makes sense to simply maximise the total expected reward. However, when we want to consider an infinite time horizon, the total reward will often be infinite, making

Type of MDP	**Average optimality**	**Blackwell optimality**
Regular MDPs	By [1], always exists.	By [1], always exists.
RMDPs	By [9], exists if the uncertainty set is sa-rectangular and compact [9].	By [9], exists if the uncertainty is sa-rectangular, compact and definable (Def 6.14).
POMDPs	By [10], exists under the (UB) condition (Def 5.1).	By [11], a continuous MDP has an optimal policy under a number of assumptions. We apply this result to POMDPs in Section 6.1.

Table 1: Literature overview regarding existence guarantees of optimal policies for average and Blackwell optimality

it unsuitable for comparing policies. There are two main ways of solving this, namely discounted reward and average reward. Discounted reward works with a discount factor $\gamma < 1$ that applies to all future rewards, therefore prioritising immediate reward. It is the most commonly studied reward measure due to its favourable theoretical properties. However, for many applications, it does not make much sense to prioritise immediate reward over the long-run average [7]. This is where the average reward objective comes in, although this comes with its own problems. Since it only considers asymptotic behaviour, there can be a pair of policies with the same average reward for which one outperforms the other in every finite time horizon. Blackwell optimality offers a good middle ground, as it requires a policy to be discount optimal for all discount factors sufficiently close to 1. A Blackwell optimal policy is always both discount and average optimal. [8]

Having introduced these different objectives, we now turn to the current state of art. For RPOMDPs, results currently exist for the finite horizon and discounted infinite horizon settings [12] [13], but, to the best of our knowledge, average and Blackwell optimality remain unexplored. However, more is known about regular POMDPs and RMDPs. We give an overview of these relevant results in Table 1.
For RMDPs, both average and Blackwell optimality are well understood, having clear conditions under which optimal policies exist [9]. For POMDPs, there are results for average [10] and ε-Blackwell [14] optimality, although there is a gap in exact Blackwell optimality. This thesis tries to bridge this gap for POMDPs, building on results from [11] for continuous MDPs, and investigates whether the existing results for POMDPs and RMDPs can be combined to obtain similar results about average and Blackwell optimality in RPOMDPs.

Contributions

We first investigate the existence of average optimal policies in RPOMDPs. While we attempt to obtain results for general agent policies, we find that restricting to mixed finite-state controllers (FSCs) with a given bound on the memory is sufficient to guarantee existence. We prove that an average optimal policy exists within this class, notably requiring no rectangularity assumptions on the uncertainty set, which are standard in RMDP results, and only that nature's policies are finitely randomised.

We then turn to Blackwell optimality in POMDPs. Building on existing results for continuous state space MDPs, we give a sufficient condition under which Blackwell optimal policies are guaranteed to exist. This is a fairly natural result, and it is somewhat surprising that it had not been established before.

Finally, we consider both exact and approximate Blackwell optimality for RPOMDPs. While a full existence theory remains open, we show that whenever a deterministic FSC is discount optimal, a deterministic Blackwell optimal FSC exists within the class of mixed FSCs, given the same bound on memory size.

Outline

We begin by discussing the relevant literature on RMDPs, POMDPs and RPOMDPs in Section 2. Then, we define the basic theory needed for understanding this thesis in Section 3, and give the formal definitions for different types of MDPs in Section 4. In Section 5, we discuss our different approaches for proving the existence of average optimal policies, and give conditions under which one exists. We formulate our Blackwell optimality results for POMDPs and RPOMDPs in Section 6. Finally, we discuss leftover open questions from our methods in Section 7, and conclude our results in Section 8.

2 Related work

As noted before, our focus is on finding a policy that maximises the expected reward. For RMDPs, this is done for the worst-case transition probabilities. Here, [9] gives a complete picture of both average and Blackwell optimality. They consider static uncertainty, where the transition probabilities stay the same over time. However, their results also extend to dynamic uncertainty, as there is no difference for fully observable RMDPs. They show that an average optimal policy exists for every sa-rectangular RMDP instance with a compact uncertainty set, and that there exists a single policy that is ε-Blackwell optimal for every $\varepsilon > 0$ and simultaneously average optimal. If the uncertainty set is polyhedral, exact Blackwell optimality is also achieved. Algorithmically, they give methods that converge to the optimal average reward for definable sa-rectangular uncertainty sets, though without theoretical convergence rate guarantees, and they note that efficiently computing an average optimal policy remains a long-standing open problem.

For POMDPs, [10] gives good theoretical results for the existence of average optimal policies. They provide a uniform boundedness condition on the belief space for the existence of an average optimal policy, though without an algorithm to compute one. Roughly, this condition states that the difference in discounted value between any two beliefs is bounded uniformly across discount factors. For Blackwell optimality, there are few results available. [14] only considers ε-Blackwell optimality, proving existence in POMDPs without requiring further assumptions. For exact Blackwell optimality, we have to turn to literature about the more general case of continuous MDPs, into which POMDPs can be reformulated via the belief MDP, such as [15] and [11]. This twin paper gives a number of conditions that, together, are sufficient for the existence of Blackwell optimal policies, in the setting of continuous state and action spaces and possibly unbounded rewards. However, this paper has not been applied to POMDPs yet. We bridge this gap in Section 6.1.

For RPOMDPs, there are a number of results for finite horizon and discounted infinite horizon, but none for average or Blackwell optimality. [12] and [16] focus on approximating the optimal discounted reward for dynamic, sa-rectangular uncertainty.

[17] and [18] investigate the usage of finite-state controllers as agent policies. [17] gives a robust value function for evaluating such an agent policy, with dynamic uncertainty. They also give algorithmic methods for minimising the cost acquired before reaching one of the goal states. [18] instead works with static uncertainty, and considers maximisation of reward before reaching a goal state. [19] also works with static uncertainty, focusing on robustly satisfying temporal logic and expected cost specifications.

[20] focuses on uncertainty sets defined by the total variation distance to a nominal distribution. Finally, [13] focuses on the connection to partially observable stochastic games. Using this connection, they show that the value functions differ between static and dynamic uncertainty, and that the order of play between nature and the agent matters as well.

All in all, to the best of our knowledge, there is a gap in RPOMDP literature regarding average and Blackwell optimality. In this thesis, we attempt to bridge this gap.

3 Preliminaries

3.1 Notation

For a set S, we write $\Delta(S)$ as the set of probability distributions (the simplex) over S. We write $\mathbf{e}_s \in \Delta(S)$ for the degenerate distribution concentrated on $s \in S$. That is, $\mathbf{e}_s(s) = 1$ and $\mathbf{e}_s(s') = 0$ for $s' \neq s$.
We write $\mathbf{1}_A : X \to \{0, 1\}$ for the indicator function of $A \subseteq X$, defined by $\mathbf{1}_A(x) = \begin{cases} 0, & x \notin A; \\ 1, & x \in A. \end{cases}$

3.2 Measure Theory

We start with some basics on topology and measure theory. This will mainly be applied in Section 6. For a more detailed explanation of this material, see the Analysis books [21] and [22].

A *topological space* is a set X, together with a collection τ of subsets of X which satisfy:

- $\emptyset$ and X are in τ;
- if $\mathcal{O} \subseteq \tau$, then $\cup_{O \in \mathcal{O}} O \in \tau$;
- if O_1 and O_2 are in τ, then $O_1 \cap O_2 \in \tau$.

Note that the different notations for the union and intersection rules make a subtle difference: the union may be infinite, but the intersection has to be finite. We call τ the *topology* on X, and the sets in τ *open*. An important example is the *discrete topology*, which we use in the proof of Theorem 6.1, where for a set X we let $\tau = P(X)$, the power set of X.

Continuity on functions is then usually defined using neighbourhoods $N \subseteq X$ of elements $x \in X$, which contain an open set $O \subseteq N$ with $x \in O$. For our use, it is easier to work with the following equivalent:

Proposition 3.1 (Theorem 13.1.7 of [21]). *A function $f : A \to B$ is continuous if and only if for every open $U \subseteq B$, $f^{-1}(U)$ is open in A.*

$$f^{-1}(U) = \{a \in A : f(a) \in U\}.$$

We will now work towards defining measures, for which we first need some other concepts. At the end, we will give the probability space as a illustrative example.

A set S of subsets of X is called a *σ-algebra* if the following hold:

1. $\emptyset \in S$ and $X \in S$.
2. if $A, B \in S$ then $A \cup B \in S$ and $A \setminus B \in S$.
3. if $(A_j)_{j=1}^{\infty}$ is a sequence of sets in S then $\cap_{j=1}^{\infty} A_j \in S$.

For a set $\mathcal{F}$ of subsets of X, $\sigma(\mathcal{F})$ denotes the smallest σ-algebra of subsets of X that contains $\mathcal{F}$.

A *measurable space* is a pair (X, Σ) where X is a set and Σ is a σ-algebra of subsets of X. The σ-algebra $\mathcal{B}$ generated by the open sets is the *Borel σ-algebra*, and its elements are called *Borel sets*.

A function $f : (X_1, \Sigma_1) \to (X_2, \Sigma_2)$ between two measurable spaces is called *measurable* if $f^{-1}(A) \in \Sigma_1$ for each $A \in \Sigma_2$.

For an interval $X \subseteq \mathbb{R}$, a real-valued function $f : X \to \mathbb{R}$ is *Borel measurable* if it is a measurable function of $(X, \mathcal{B})$ into $(\mathbb{R}, \mathcal{B})$. If a function $f : A \to \mathbb{R}$ is continuous, it is also Borel measurable. [22]

$\mu : \Sigma \to \mathbb{R}^+$ is a *measure* if for any sequence A_n of disjoint elements of Σ, we have $\mu(\cup_n A_n) = \sum_n \mu(A_n)$. A *finite measure space* is a triple (X, Σ, μ) where (X, Σ) is a measurable space and $\mu : \Sigma \to \mathbb{R}^+$ is a measure. It is called finite since $\mu(X)$ is a real number, and thus a finite measure. A *σ-finite measure space* instead has a sequence $(I_k)_{k=1}^{\infty}$ of disjoint elements of Σ whose union is exactly X, with $\mu(I_k) < \infty$ for all k, but now $\mu(X)$ may be infinite.

The probability space (Ω, Σ, P), where $P(\Omega) = 1$, is a finite measure space. This gives a formalisation of a random process. The event space Σ is a σ-algebra over the sample space. That is, if A and B are events, $A \cup B$ is as well. And, if A and B are disjoint, the probability of the event $A \cup B$ occurring is $P(A \cup B) = P(A) + P(B)$. A random variable is a measurable function $X : \Omega \to \mathbb{R}$. Measurability ensures that probabilities of events like $\{X \in A\}$ are well-defined.

An important example of a measure is the Dirac measure [23], defined by

$$\delta_x(A) = \mathbf{1}_A(x) = \begin{cases} 0, & x \notin A; \\ 1, & x \in A. \end{cases}$$

Another example is the counting measure, where $\mu(A)$ is simply the number of elements in A. If X is a countable set, this is a σ-finite measure. We can see the counting measure as a sum of Dirac measures, as $|A| = \sum_{x \in X} \delta_x(A)$ for a subset $A \subseteq X$.

We can take integrals over measure, formally defined as:

$$\int_X f \, d\mu = \int_0^{\infty} \lambda_f(t) \, dt,$$

where $\lambda_f(t) = \mu(\{x \in X : f(x) > t\})$ for $t \geq 0$. It is easy to verify that for a Dirac measure, we have

$$\int_X f \, d\delta_x = f(x), \tag{1}$$

since

$$\delta_y(\{x \in X : f(x) > t\}) = \begin{cases} 0, & t \geq f(y); \\ 1, & t < f(y). \end{cases}$$

For the counting measure $\mu = \sum_{x \in X} \delta_x$, we have that

$$\int_X f(y)\mu(dy) = \int_X f(y) \sum_{x \in X} \delta_x(dy) = \sum_{x \in X} \int_X f(y)\delta_x(dy) = \sum_{x \in X} f(x). \tag{2}$$

4 The Model

In this section, we formally state the basics about Markov Decision Processes and three extensions of it, which add uncertainty about the transitions, the states, or both.

4.1 Markov Decision Processes

Markov decision processes (MDPs) are a basis for the more complicated cases that we want to investigate. An MDP is a model for a sequential decision-making problem, working in a probabilistic environment. An agent observes the state of the system, and chooses actions to influence the future state. The agent also receives a reward from this, and tries to choose the actions that give a large expected reward.

Definition 4.1 (Markov Decision Process)**.** An MDP is a tuple (S, A, P, r, s_0), where:

- S is a finite set of states.
- A is a finite set of actions.
- $P : S \times A \to \Delta(S)$ is the state transition function, giving a distribution over states given the chosen action and previous state.
- $r : S \times A \to \mathbb{R}$ is the reward function, giving the expected immediate reward for the agent.
- s_0 is the initial state.

When the agent chooses action a in state s, we write $r(s, a)$ for the expected immediate reward, and $P(s, a, s')$ for the probability of ending up in state s'. In this model, both of these functions only depend on the previous state and the action chosen. That is, the conditional probability of ending up in state s', given the entire history of states and actions in the system, is simply $P(s, a, s')$.

4.1.1 Policies

The agent chooses actions based on its *policy* π, sometimes also called a *scheduler* or *strategy*. We write Π for the set of policies. Generally, a policy can depend on the entire history of the system, and the current time t. However, we mostly work with history-based policies that do not depend on t directly, defined as such:

$$\pi : (SA)^* S \to \Delta(A).$$

A policy is stationary when it only depends on the current state, that is $\pi : S \to \Delta(A)$. A policy is deterministic when it maps to a single action, that is $\pi : (SA)^* S \to A$.

We write Π_S for the set of stationary policies, Π_D for the set of deterministic policies, and $\Pi_{SD} = \Pi_S \cap \Pi_D$ for the set of stationary and deterministic policies.

4.1.2 Reward measures

We want to compute the *optimal* policy for the agent, such that some measure of the long-term expected reward is maximised. There are several options for these reward measures that we can maximise for.

The first one is finite-horizon optimality, for which there is some T given where the process stops. The agent then needs to maximise $\mathbb{E}[\sum_{t=0}^{T} r_t]$, where we use r_t as a shorthand for $r(s_t, a_t)$. That is, r_t is the reward gained by the agent at time step t.

However, we are mainly interested in optimising for infinite horizon. Usually, the total expected reward over an infinite time $\mathbb{E}[\sum_{t=0}^{\infty} r_t]$ diverges, so we need some different reward measure to be able to compare different policies. There are two main ways to do this. One is discounted optimality, where we introduce a discount factor $0 < \gamma < 1$, to prioritise earlier rewards gained. Then, the agent maximises

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right].$$

The other option is to give every r_t the same weight, regardless of t. Then, we look at average optimality. An intuitive definition would be the limit of $\mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T} r_t\right]$ as $T \to \infty$. However, this limit may not exist [9]. Therefore, we use the following definition for the average reward:

$$\liminf_{T\to\infty} \mathbb{E}\left[\frac{1}{T+1}\sum_{t=0}^{T} r_t\right].$$

Of these two, discounted reward is by far the most studied, as, due to its simplicity, it is both easier to compute and obtain theoretical results about. However, average optimality is a more useful metric for many applications where discounting future reward makes little sense. That being said, optimising only for average reward is also not desirable, as a pair of average optimal policies may exist for which one obtains a better total reward than the other for any finite time horizon. A good middle ground that solves these issues is Blackwell optimality [24].

Definition 4.2 (Blackwell optimality)**.** A policy π^* is Blackwell optimal if there exists some discount factor $\gamma_0 \in (0, 1)$ such that π^* is discount optimal for any $\gamma \in (\gamma_0, 1)$. That is:

$$\mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \geq \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \forall \pi \in \Pi, \gamma \in (\gamma_0, 1).$$

Any Blackwell optimal policy is also average optimal. This seems like the most desirable type of optimality, yet it is the least studied due to its complexity. This thesis will focus on average and Blackwell optimality, as, to the best of our knowledge, these are unexplored in RPOMDP literature.

4.1.3 Policy evaluation

If we are given a policy $\pi \in \Pi_{SD}$, we can compute the expected reward of π using the value function. The expected infinite horizon discounted reward of policy π when starting in state s is given by:

$$V_\pi(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} P(s, \pi(s), s') V_\pi(s').$$

For average reward, we cannot simply remove the discount factor from this equation, as this would cause it to diverge. Instead, we need to subtract the average reward g from each value. That is:

$$V_\pi(s) = r(s, \pi(s)) + \sum_{s' \in S} P(s, \pi(s), s') V_\pi(s') - g.$$

For MDPs, these values of states can be used to compute an optimal policy, using Dynamic Programming techniques.

4.2 Robustness

When we want to model a real-world problem as an MDP, we need to know the transition probabilities exactly. This is often a problem, as we can only approximate the exact probabilities, and a small error can lead to a completely different optimal policy [4]. A Robust Markov decision process (RMDP) tries to solve this issue, as it adds uncertainty to the transition function P. Now, P can be anywhere in a defined uncertainty set $\mathcal{U}$.

Definition 4.3 (Robust Markov Decision Process)**.** An RMDP is a tuple $(S, A, \mathcal{U}, r, s_0)$, where

- S is a finite set of states.
- A is a finite set of actions.
- $\mathcal{U} \subseteq (\Delta(S))^{S \times A}$, is the uncertainty set, which is the set of possible transition functions $P : S \times A \to \Delta(S)$ in the model.
- $r : S \times A \to \mathbb{R}$ is the reward function, giving the expected immediate reward for the agent.
- s_0 is the initial state.

For RMDPs, it is common to optimise for the worst-case probabilities, to get a robust policy that guarantees the best lower bound possible. Generally, it is NP-hard to compute the optimal policy for an RMDP [25]. So, we need some assumptions on the structure of the uncertainty set. These assumptions, about so-called *rectangularity*, may neglect constraints of the actual uncertainty set. Since we optimise for the worst case, this gives us a conservative lower bound of the actual value. That is, we get a guarantee of the expected reward that the optimal policy will at least get, even if our assumptions on the structure of the uncertainty set do not hold.

Two common types are s-rectangularity and sa-rectangularity. An sa-rectangular uncertainty set is a product $\mathcal{U} = \bigtimes_{(s,a) \in S \times A} \mathcal{U}_{sa}$, $\mathcal{U}_{sa} \subseteq \Delta(S)$. That is, the probabilities of any state-action pair (s, a) can be picked independently of any other pair.

An s-rectangular uncertainty set is a product $\mathcal{U} = \bigtimes_{s \in S} \mathcal{U}_s$, $\mathcal{U}_s \subseteq (\Delta(S))^A$. That is, each $\mathcal{U}_s$ is a set of functions from the action a to a distribution over states. Then, the probabilities of any state can be picked independently, but there can be constraints between the different actions in the same starting state.

We can look at different types of uncertainty, namely static and dynamic. For static uncertainty, the transition function $P \in \mathcal{U}$ does not change over time. For dynamic uncertainty, it can.

Since we look at the worst-case, an RMDP can be seen as a game between the agent (controlling the actions) and nature (controlling the transition probabilities). We write Θ for the set of all nature policies. In the case of static uncertainty, we restrict to nature policies that simply pick a transition function, so we write $\mathbf{P} \in \mathcal{U}$ for such a policy.

The reward now depends on both the agent policy π and the nature policy θ. For a policy pair (π, θ), we write $R_{\text{avg}}(\pi, \theta)$ for the average reward, and $R_\gamma(\pi, \theta)$ for the discounted reward with discount factor γ. They are defined as:

$$R_{\text{avg}}(\pi, \theta) = \liminf_{T \to \infty} \mathbb{E}_{\pi,\theta} \left[\frac{1}{T+1} \sum_{t=0}^{T} r(s_t, a_t) \right]$$

$$R_\gamma(\pi, \theta) = \mathbb{E}_{\pi,\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right].$$

For the existence of average optimal policies, we now need to check if a solution π exists to

$$\sup_{\pi \in \Pi} \inf_{\theta \in \Theta} R_{\text{avg}}(\pi, \theta).$$

[9] shows this existence under the condition that the uncertainty set is compact and sa-rectangular.

If we are given a stationary policy $\pi \in \Pi_S$, and the RMDP is s-rectangular, the problem of computing $\inf_{\theta \in \Theta} R_{\text{avg}}(\pi, \mathbf{P})$ is also an MDP problem. This is called the adversarial MDP. In this MDP, the set of states is just S, and the set of actions at $s \in S$ is the set $\mathcal{U}_s$. From the adversarial MDP we can compute the optimal nature policy. Since this is just a normal MDP, an optimal stationary and deterministic policy exists. Therefore, restricting to static uncertainty makes no difference for s-rectangular RMDPs [9].

4.3 Partial Observability

A different issue with MDPs is that it assumes that the agent can observe the full state of the system. This is also something that might not be feasible in practice. We can instead work with Partially Observable Markov decision processes (POMDPs). In this setting, the agent does not know what state it is in, but instead receives some observation based on the state.

Definition 4.4. A POMDP is a tuple $(S, A, P, r, Z, O, \mathbf{b}_0)$, where:

- S is a finite set of states.
- A is a finite set of actions.
- $P : S \times A \to \Delta(S)$ is the state transition function, giving a distribution over states given the chosen action and previous state.
- $r : S \times A \to \mathbb{R}$ is the reward function.
- Z is a finite set of observations.
- $O : S \to \Delta(Z)$ is the observation function, giving a distribution over observations given the resulting state.
- $\mathbf{b}_0 \in \Delta(S)$ is the initial belief state.

An important distinction is that now, the immediate reward gained can not be observed anymore, as this would give away information about the true state of the agent. We write $O(s, o)$ for the probability of observing o when in state s.

The agent cannot observe the state, but it can keep track of an internal belief state $\mathbf{b}$ instead. This is a distribution over states that represents the probability that the agent is currently in each state. The agent starts with the initial belief b_0. In every time step, the belief state is updated based on the observation received. Given an action a and observation o, we write $\mathbf{b}^{a,o}$ for the updated belief state (from $\mathbf{b}$), which is computed as

follows:

$$\begin{aligned}
\mathbf{b}^{a,o}(s') &= \Pr(s' \mid a, o, \mathbf{b}) \\
&= \frac{\Pr(o \mid s', a, \mathbf{b}) \cdot \Pr(s' \mid a, \mathbf{b})}{\Pr(o \mid a, \mathbf{b})} \\
&= \frac{\Pr(o \mid s', a) \cdot \sum_{s \in \mathcal{S}} \Pr(s' \mid a, \mathbf{b}, s) \cdot \mathbf{b}(s)}{\Pr(o \mid a, \mathbf{b})} \\
&= \frac{O(s', o) \cdot \sum_{s \in \mathcal{S}} P(s, a, s') \cdot \mathbf{b}(s)}{\sum_{s' \in \mathcal{S}} O(s', o) \cdot \sum_{s \in \mathcal{S}} P(s, a, s') \cdot \mathbf{b}(s)}
\end{aligned}$$

Here $\Pr(o \mid a, \mathbf{b}) = \sum_{s' \in \mathcal{S}} O(s', o) \cdot \sum_{s \in \mathcal{S}} P(s, a, s') \cdot \mathbf{b}(s)$ is the normalisation constant, independent of s', that makes $\mathbf{b}^{a,o}$ a valid probability distribution.

It is important to note that these beliefs form a *sufficient statistic* for the history and initial belief state. That is, the current belief captures all known information about the current state, meaning that no additional data about past actions or observations would refine the agent's uncertainty about it [26]. This allows us to construct an MDP equivalent to the POMDP with the beliefs as states. This results in an MDP with a continuous state space. It is defined as follows:

Definition 4.5 (Belief Markov Decision Process)**.** Given a POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$, the equivalent belief MDP is a tuple $(\mathcal{B}, A, \mathcal{P}, \bar{r}, \mathbf{b}_0)$

- $\mathcal{B} \subseteq \Delta(S)$ is the set of reachable belief states.
- A is the same finite set of actions.
- $\mathcal{P} : S \times A \to \Delta(S)$ is the belief state transition function, defined as:

$$\mathcal{P}(\mathbf{b}, a, \mathbf{b}') = \Pr(\mathbf{b}' \mid \mathbf{b}, a) = \sum_{o \in Z} \Pr(\mathbf{b}' \mid a, \mathbf{b}, o) \Pr(o \mid a, \mathbf{b}).$$

- $\bar{r} : \mathcal{B} \times A \to \mathbb{R}$ is the reward function, giving the expected immediate reward for belief states, defined as:

$$\bar{r}(\mathbf{b}, a) = \sum_{s} \mathbf{b}(s) r(s, a).$$

- $\mathbf{b}_0 \in \mathcal{B}$ is the initial belief state.

Note that $\Pr(\mathbf{b}' \mid a, \mathbf{b}, o)$ is always either 1 or 0, depending on whether $\mathbf{b}^{a,o} = \mathbf{b}'$ or not.

From the belief update, we can see that, even though the belief space is infinite, each state still has a finite number of possible transitions, since both A and Z are finite. This makes $\mathcal{B}$ a countable set, since it can be described as a union of finite sets $\bigcup_n D_n$ where D_i is the set of belief states that can be reached from $\mathbf{b}_0$ in exactly i steps. [10]

Since the agent cannot observe the state, our usual definitions for policies do not work anymore. Instead, we can talk about policies $\pi : (ZA)^* Z \to \Delta(A)$ that are based on the entire history of observations and actions. Another type is belief-based policies $\pi : \mathcal{B} \to \Delta(A)$, that simply pick an action based on the belief state. These policies are stationary in the belief MDP. However, since the belief states themselves are based on the full history of observations, these belief-based policies are still history-based. We write Π_B for the set of belief-based policies, and Π_{DB} for the set of deterministic belief-based policies.

4.4 Combining the two

For this thesis, we consider Robust Partially Observable Markov Decision Processes (RPOMDP), which combine the ideas of RMDPs and POMDPs. That is, they have uncertain transitions *and* uncertain states.

Definition 4.6 (Robust Partially Observable Markov Decision Process)**.** An RPOMDP is a tuple $(S, A, \mathcal{U}, r, Z, O, s_0)$, where:

- S is a finite set of states.
- A is a finite set of actions.
- $\mathcal{U} \subseteq (\Delta(S))^{S \times A}$, is the uncertainty set, which is the set of all possible transition functions $P : S \times A \to \Delta(S)$.
- $r : S \times A \to \mathbb{R}$ is the reward function, giving the expected immediate reward for the agent.
- Z is a finite set of observations.
- $O : S \to \Delta(Z)$ is the observation function, giving a distribution over observations given the resulting state.
- s_0 is the initial state

We assume that the process is only partially observable for the agent. That is, nature can observe the state of the system. The definitions of average and discounted reward are just the same as for RMDPs. Some papers also consider uncertainty in the observation function, alongside the transition probabilities, but this makes no difference [13].

Just as in RMDPs, we can again talk about static and dynamic uncertainty, and s-rectangular and sa-rectangular uncertainty sets. These terms are defined exactly the same. Other than for RMDPs, for RPOMDPs there can be a difference in static and dynamic uncertainty, even under sa-rectangularity [13].

We cannot directly work with beliefs as in POMDPs, as the belief update function is now uncertain. So, instead of belief-based policies, we have to work with general history-dependent policies $\pi : (ZA)^* Z \to \Delta(A)$.

5 Average reward optimality in RPOMDPs

We now turn to investigating the maximisation of average reward in RPOMDPs. Previous work on RPOMDPs has focused on discounted reward, leaving average optimality unexplored. Our main objective will be to prove that there exists an average optimal policy. However, we will not be able to prove this for any general RPOMDP, as there are already examples of both POMDPs [10] and RMDPs [9] without an average optimal policy. So, we want to find sufficient conditions on the RPOMDP for this existence.

We would like these conditions to restrict the nature policy. For example, we could restrict to sa-rectangular and static uncertainty, which is the same setting that [9] shows existence of average optimal policies for, in the case of RMDPs. In this section, we first show our attempt at extending their proof for the case of RPOMDPs. However, this attempt was not successful.

We then show our attempt to prove existence of average optimal policies under restrictions of the agent policy, without assumptions on the structure of the uncertainty set. This is based on [27], a paper about so-called convex semi-finite games. This will lead to our main result of this section.

5.1 A natural first approach

We base this approach on Theorem 3.4 of [9], to see if we can still follow these steps if we add partial observability. This proof does use static uncertainty, so we assume the nature policy is simply a transition function $\mathbf{P} \in \mathcal{U}$. Note that, under s-rectangularity, this is no different than dynamic for RMDPs [9], but for RPOMDPs this does matter [13].

The proof in [9] shows the following equation:

$$\sup_{\pi\in\Pi}\inf_{\mathbf{P}\in\mathcal{U}} R_{\text{avg}}(\pi,\mathbf{P}) \leq \inf_{\mathbf{P}\in\mathcal{U}}\sup_{\pi\in\Pi} R_{\text{avg}}(\pi,\mathbf{P}) = \inf_{\mathbf{P}\in\mathcal{U}}\max_{\pi\in\Pi} R_{\text{avg}}(\pi,\mathbf{P}) = \max_{\pi\in\Pi}\inf_{\mathbf{P}\in\mathcal{U}} R_{\text{avg}}(\pi,\mathbf{P}).$$

If all of these hold for RPOMDPs, we have proven existence of average optimal policies. The inequality $\sup_{\pi\in\Pi}\inf_{\mathbf{P}\in\mathcal{U}} R_{\text{avg}}(\pi,\mathbf{P}) \leq \inf_{\mathbf{P}\in\mathcal{U}}\sup_{\pi\in\Pi} R_{\text{avg}}(\pi,\mathbf{P})$ holds from weak duality.

The first equality $\inf_{\mathbf{P}\in\mathcal{U}}\sup_{\pi\in\Pi} R_{\text{avg}}(\pi,\mathbf{P}) = \inf_{\mathbf{P}\in\mathcal{U}}\max_{\pi\in\Pi} R_{\text{avg}}(\pi,\mathbf{P})$, we can show under certain conditions. For this, we need to show that, given any $\mathbf{P} \in \mathcal{U}$, the resulting POMDP, with $\mathbf{P}$ as transition function, has an average optimal policy. For this, we can use [10], which gives conditions on the POMDP for this existence. This is where the static uncertainty assumption is important, as then this induces a finite-state POMDP.

One of these conditions is that the belief space is countable. Since we only care about the reachable belief space starting from $\mathbf{b}_0$, this is always satisfied. The other condition is more involved. It is a uniform boundedness condition, defined as follows:

Definition 5.1 (UB)**.** The condition (UB) is fulfilled if there is a constant $M > 0$ such that

$$|V_\gamma(\mathbf{b}) - V_\gamma(\mathbf{b}')| \leq M,\ \forall 0 < \gamma < 1, \forall \mathbf{b}, \mathbf{b}' \in \mathcal{B}, \tag{3}$$

where $V_\gamma(\mathbf{b})$ are the values of the belief states, derived from the Bellman equation:

$$V_\gamma(\mathbf{b}) = \max_{a\in A} \bar{r}(\mathbf{b}, a) + \gamma \sum_{\mathbf{b}'\in\mathcal{B}} \mathcal{P}(\mathbf{b}, a, \mathbf{b}')V_\gamma(\mathbf{b}').$$

Condition (UB) does not bound $V_\gamma(\mathbf{b})$ itself, which may, in general, diverge as $\gamma \to 1$. Instead, it bounds the spread between values of different belief states, uniformly in γ. Intuitively, even though $V_\gamma(\mathbf{b})$ grows unboundedly as $\gamma \to 1$ (at a rate determined by the optimal average reward), this growth is the same for every $\mathbf{b} \in \mathcal{B}$, so that pairwise differences remain bounded. (UB) states this boundedness directly, for every γ. This is the basis for the vanishing discount approach that [10] use to solve the average reward optimality equation.

So, if (UB) holds given any $\mathbf{P} \in \mathcal{U}$, then we can use the results of [10] to show existence of average optimal policies. That is, the equality $\inf_{\mathbf{P}\in\mathcal{U}} \sup_{\pi\in\Pi} R_{\text{avg}}(\pi, \mathbf{P}) = \inf_{\mathbf{P}\in\mathcal{U}} \max_{\pi\in\Pi} R_{\text{avg}}(\pi, \mathbf{P})$ holds.

However, the last equality is a problem. In [9], this follows from results on stochastic games with perfect information. The RMDP can be viewed as a two-player game between the agent and nature, for which, under the assumption that $\mathcal{U}$ is sa-rectangular and polyhedral, there exist saddle-point policies. Then, the value of the game is independent of who acts first, giving

$$\inf_{\mathbf{P}\in\mathcal{U}} \max_{\pi\in\Pi} R_{\text{avg}}(\pi, \mathbf{P}) = \max_{\pi\in\Pi} \inf_{\mathbf{P}\in\mathcal{U}} R_{\text{avg}}(\pi, \mathbf{P}).$$

For RPOMDPs, we need to look at partially observable stochastic games instead. There, we do not have this type of saddle-point result. If we do assume this equality, then an average optimal policy exists for any RPOMDP with static uncertainty for which the induced POMDP given any $\mathbf{P} \in \mathcal{U}$ satisfies (UB). However, this equality is a big step of this proof, that we do not have. For future work, [28] is an interesting paper to look at regarding partially observable stochastic games literature. It might be possible to find sufficient conditions for existence for RPOMDPs based on their results. However, it is not clear what their assumptions mean, as they use notation that is not defined in the paper.

5.2 Convex semi-infinite games

However, we can prove existence of an average optimal policy within a specific class of agent policies. For this, we use [27], a paper about so-called convex semi-infinite games. They are two-person zero-sum games (as RPOMDPs are), but with the following sets of strategies:
Player I: $\Gamma = \{\lambda = (\lambda_t)_{t\in T} |$ only finitely many $\lambda_t \neq 0, \lambda_t \geq 0, \sum \lambda_t = 1\}$.
Player II: $C =$ nonempty closed convex set in $\mathbb{R}^n$.
Let $\mathscr{F}_T = \{f_t, t \in T\}$ be a family of convex functions defined in $\mathbb{R}^n$ with finite values.
The reward for player I is $P(\lambda, x) \sum_{t\in T} \lambda_t f_t(x)$ for $\lambda \in \Gamma, x \in C$.

Existence is shown under one additional condition, which is trivially satisfied when C is bounded. Then, the following equality holds:

$$\inf_{x\in C} \sup_{\lambda\in\Gamma} P(\lambda, x) = \sup_{\lambda\in\Gamma} \inf_{x\in C} P(\lambda, x).$$

There also exists an optimal strategy for player II. We can translate RPOMDPs to such a game by viewing the agent as one of the players and nature as the other. We have two options to choose from. The first is that we interpret player I as nature, and player II as the agent. Then, existence is immediate, since we know that there exists an optimal policy for player II. Then T is the set of deterministic nature policies, so we restrict to finitely randomised policies for nature.

The problem arises when we try to convert the policy space of the agent into a set in $\mathbb{R}^n$. It requires a finite-dimensional representation of the policy, which is, in general, not

possible when we work with policies that use unbounded histories. We also need the value function $f_t(x) = R_{avg}(x,t)$ to be convex in x for our representation of the policy in $\mathbb{R}^n$. So, we need a restriction on agent policies to satisfy this.

For the convexity, we restrict to mixed policies $x \in \Delta(\Pi_D)$, that are a probability distribution over deterministic policies. Then, the reward functions are immediately convex since we have $f_t(x) = \sum_{\pi \in \Pi_D} x(\pi) R_{avg}(\pi, t)$.
$R_{avg}(\pi, t)$ is then simply the long-run average reward of the induced (possibly infinite) Markov Chain. Since the RPOMDP has bounded rewards, the average value function cannot exceed that, so it always exists. It is important that we restrict to mixed policies, and not allow stochastic ones. With stochastic policies, this sum over deterministic policies does not work and the reward function is generally not convex. We will show a simple example of an MDP where the reward function is not convex in Section 5.2.1.

We can represent a mixed policy $x \in \Delta(\Pi_D)$ very easily as an element of $\mathbb{R}^n$ if we have that $\Delta(\Pi_D) \subset \mathbb{R}^n$. For this to be the case, the set of deterministic policies Π_D needs to be finite. This is generally not the case, but we can make it finite by restricting to deterministic finite-state controllers (FSCs) with a cap on the amount of memory states. Using FSCs as policies for RPOMDPs is also explored in [17] and [18].

Definition 5.2 (Finite-state Controller)**.** An FSC is a tuple (N, α, η), where:

- N is the set of memory states $n \in N$.
- $\alpha : N \times Z \to A$ is the action mapping, giving the action selection given the current memory state and observation.
- $\eta : N \times Z \times A \to N$ is the memory update, giving the next memory state given the observation, chosen action, and current memory state.

Generally, the action mapping and memory update functions return a distribution over actions and memory states, respectively. But here, as noted above, we restrict to deterministic FSCs. We write Π_{FSC_k} for the set of deterministic FSCs with $|N| = k$ memory states.

A deterministic FSC induces a history-based deterministic policy. Since we put a cap on $|N|$, there are only finitely many FSCs, since Z and A are finite sets. So, the policy space is a convex compact subset of $\mathbb{R}^{|\Pi_{FSC_k}|}$. We can therefore prove existence of an optimal agent policy, given any memory bound k. We will now formally state and prove this result. We first state the condition (R) from [27], which is needed for their results.

Condition (R). The functions of $\mathscr{F}_T$ have no common direction of recession which is also a direction of recession of C.

Lemma 5.3. *Condition (R) is satisfied when C is bounded.*

Proof. Since C is bounded, it has no direction of recession. Therefore, the functions of $\mathscr{F}_T$ have no common direction of recession that is also a direction of recession of C. □

Theorem 5.4. *Suppose we are given an RPOMDP $(S, A, \mathcal{U}, r, Z, O, s_0)$ with finitely randomised nature policies, and a memory bound k. Then, there exists a policy that is average optimal in the set $\Delta(\Pi_{FSC_k})$ of mixed policies over deterministic FSCs (N, α, η) with $|N| = k$ memory nodes.*

Proof. Given an RPOMDP $(S, A, \mathcal{U}, r, Z, O, s_0)$, and a memory bound k, we convert it into a convex semi-infinite game as follows:

Player I (Nature): $\Gamma = \{\lambda = (\lambda_t)_{t\in T} |$ only finitely many $\lambda_t \neq 0, \lambda_t \geq 0, \sum \lambda_t = 1\}$, where T is the set of deterministic (possibly dynamic) nature policies.
Player II (Agent): $C = \Delta(\Pi_{FSC_k})$, where Π_{FSC_k} is the set of deterministic FSCs (N, α, η) with $|N| = k$ memory nodes.

The reward for player I is given by $P(\lambda, x) = \sum_{t\in T} \lambda_t f_t(x)$ for $\lambda \in \Gamma, x \in C$, where $f_t(x) = \sum_{\pi\in\Pi_{FSC_k}} x(\pi) R_{avg}(\pi, t)$.

Finally, define $\mathscr{F}_T := \{f_t, t \in T\}$. We now show that $(\mathscr{F}_T, C)$ is an (R)-game, as defined in [27].

From the definition of $f_t(x)$, it is convex in x for any t. Since N, Z and A are all finite sets, Π_{FSC_k} is finite. Therefore, C is a convex compact set in $\mathbb{R}^{|\Pi_{FSC_k}|}$. So, $(\mathscr{F}_T, C)$ is an (R)-game, and there exists an optimal finite-state controller with k memory states by Theorem 4.1 of [27]. □

Note that, perhaps surprisingly, our only condition on *nature* is the restriction to finitely randomised policies. We don't need any assumptions on the structure of the uncertainty set, like rectangularity or compactness.

5.2.1 Seeing the agent as player I

There is also the option of seeing the agent as player I, and nature as player II. Then, we let T be the set of deterministic agent policies Π_D, so the only restriction is that we have a finite randomisation over agent policies. If we now restrict to static nature, with an sa-rectangular convex uncertainty set, C is immediately a compact convex set in $\mathbb{R}^n$, as it is simply the uncertainty set itself. Existence is not immediate, as there are only guarantees for optimal policies for player II, but there are additional conditions listed under which player I also has optimal policies.

The main problem is that the reward functions $f_t(x) = R_{avg}(t, x)$ need to be convex in x, that is, in the nature policy. We can easily construct an example where this is not the case.

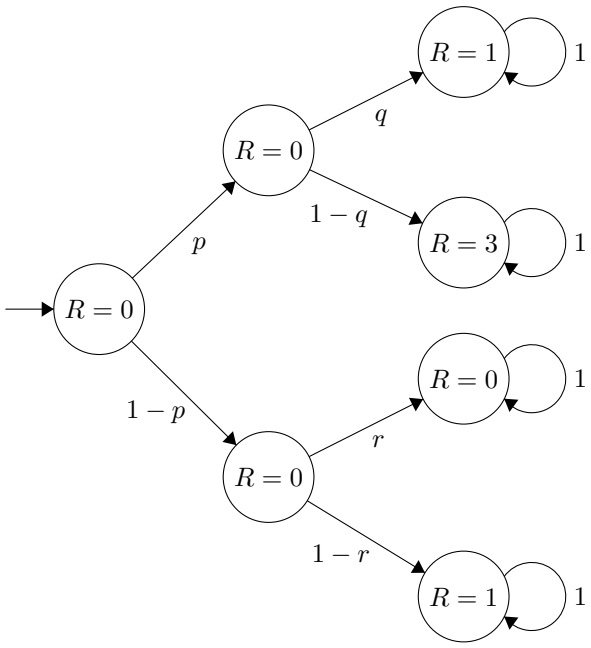

Figure 1: RMDP with no convexity in the nature policy

Figure 1 is a simple RMDP without agent actions where the reward function is not convex in x. Here, the arrows show the transition probabilities, and the numbers in the states correspond to the reward. Clearly, for (p, q, r) equal to $(1, 1, 0), (0, 0, 0), (1, 1, 1)$, or $(0, 1, 0)$, the average reward is 1. However, $(0.5, 0.5, 0.5)$ has an average reward of 1.25, and $(0.5, 1, 0.5)$ has an average reward of 0.75. That is, there are convex combinations that have lower and higher average reward. Thus, the value function is not convex in x.

We can use the same example to show why we cannot use randomised FSCs, as shown in Figure 2. Now, instead of transition probabilities, the arrows show the actions a and b. That is, it is now an MDP with deterministic transitions. The idea is the exact same, as the previous parameters can now be seen as the probability that action a is chosen in the corresponding state. Thus, the value function is also not convex in this case. This is why we have to use mixed policies that are a distribution over deterministic FSCs.

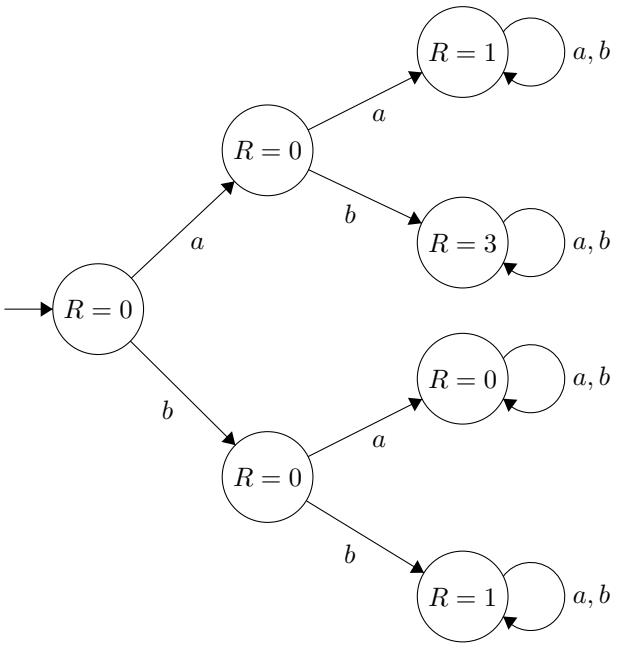

Figure 2: MDP with no convexity in the FSC action mapping

One could ask if we might be able to work with FSCs with deterministic action selection but randomised memory update. It is a bit more complicated as we need to encode it in the memory instead of the MDP itself, but it is possible to construct a similar example like this to show that the value function is also not convex in this case.

6 Blackwell optimality

We now focus on results about Blackwell optimality. As discussed in Section 4.1.2, Blackwell optimality offers a good middle ground between long-term payoff and immediate reward. Our goal is to study Blackwell optimality in RPOMDPs. However, there are very few results available about Blackwell optimality in POMDPs. Therefore, we start by giving sufficient conditions for the existence of Blackwell optimality policies in POMDPs. This is our main result of this section.

We continue with investigating Blackwell optimality in RPOMDPs, proving that there exists a Blackwell optimal policy in the class of mixed FSC policies with memory cap m. Note that we need this same restriction as in the previous section, since a Blackwell optimal policy is always average optimal as well.

6.1 Blackwell optimality in POMDPs

Even though there are few results about POMDPs, there are more results about Blackwell optimality for MDPs with a Borel (continuous) state space. Remember that for any POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$, the equivalent belief MDP $(\mathcal{B}, A, \mathcal{P}, \bar{r})$ has such a state space. So, we use the results of [15] and [11], which is a twin paper about Blackwell optimality in MDPs with Borel state space.

For this, we first define some notation for MDPs with continuous state space, taken from [11]. For a set of belief states $B \subseteq \mathcal{B}$, we denote by $\mathbf{P}(\mathbf{b}, a, B)$ the transition probability of moving to any belief state in B. Note that, since $\mathcal{B}$ is countable, this transition function is simply a discrete sum, so we can define it as $\mathbf{P}(\mathbf{b}, a, B) := \sum_{\mathbf{b}' \in B} \mathcal{P}(\mathbf{b}, a, \mathbf{b}')$. For any real-valued function f on $\mathcal{B} \times A$, we define $\hat{f}(\mathbf{b}) := \sup_{a \in A} |f(\mathbf{b}, a)|$.

By using [15] and [11], we prove that any POMDP satisfying the following assumption has a deterministic belief-based Blackwell optimal policy:

Assumption 1. *There exists a belief $\mathbf{b}' \in \mathcal{B}$ and a number $\delta > 0$ where for any $\mathbf{b} \in \mathcal{B}$, $a \in A$, we have $\mathcal{P}(\mathbf{b}, a, \mathbf{b}') \geq \delta > 0$.*

6.1.1 Revealing POMDPs

Assumption 1 is somewhat related to revealing POMDPs, a concept that is studied in [29] and [30]. We use the following definitions: an observation $o \in Z$ is revealing for a state s if $O(s, o) > 0$ and for any $a \in A$, $\mathbf{b} \in \mathcal{B}$, we have $\mathbf{b}^{a,o} = \mathbf{e}_s$; that is, the new belief is that the state is s with probability 1. A POMDP is strongly revealing if every state $s \in S$ has a revealing observation. A POMDP is weakly revealing if for any policy, the agent almost surely receives infinitely many revealing observations.

Now, a condition implying Assumption 1 would be the existence of a state $s' \in S$ that:

1. Can be reached from any other state, given any action (for any $s \in S$, $a \in A$, $P(s, a, s') > 0$).
2. Has a revealing observation.

Note that this condition implies the POMDP is weakly revealing. Now, under this condition, Assumption 1 holds with $\mathbf{b}' = \mathbf{e}_{s'}$ and $\delta = o_{\min} \cdot p_{\min}$, where $p_{\min}$ and $o_{\min}$ are the smallest strictly positive transition and observation probabilities respectively. It holds since the probability of reaching state s' is at least $p_{\min}$, and the probability of then revealing the state is at least $o_{\min}$.

An example of a type of weakly revealing POMDPs given in [30] are "POMDPs that 'reset' infinitely often, and whose reset can be observed with a dedicated signal". If this reset can also happen from any state, it would imply Assumption 1.

6.1.2 Proving Blackwell optimality

We prove existence of Blackwell optimal policies based on Theorem 2.1 of [11], which gives this result under a couple of assumptions, namely assumptions 2.1-2.5. We will see that most of these assumptions always hold for a POMDP, as they are simple assumptions about measurability, continuity and boundedness of different functions of the model.

The only assumption that we need our assumption 1 for is assumption 2.4. To give the essence of the assumption, we define it here informally as follows:

Assumption 2.4. *For every policy $\pi \in \Pi_B$, the t-step transition distribution $\mathbf{P}_\pi^t(\mathbf{b}, \cdot)$ converges as $t \to \infty$ to a stationary distribution $\mathbf{P}_\pi^\infty(\mathbf{b}, \cdot)$. Moreover, the rate of convergence is at least geometric. That is, there exist constants $C > 0$ and $\beta < 1$, independent of π and $\mathbf{b}$, such that the distance between $\mathbf{P}_\pi^t$ and $\mathbf{P}_\pi^\infty$ is at most $C\beta^t$ for all $t \geq 0$.*

This assumption, along with a bound on the rewards, allows a Laurent series expansion of the discounted reward given a belief-based policy for discount factors close to 1. Laurent series are a complex analogue of Taylor series that also allow negative powers, making them suitable for expanding functions around a pole. Even though we are not working with complex numbers, the Laurent series expansion is still used since the value function has a pole at $\gamma = 1$.

These Laurent series have well-defined coefficients h_σ^n. Blackwell optimality is then equivalent to maximising these coefficients lexicographically, which reduces the problem to a single optimality equation. The other assumptions about measurability and continuity ensure the continuity of the Laurent coefficients, which in turn guarantees the existence of a solution.

We will not show Assumption 2.4 directly. Instead, we use Theorem 5.1 of [11], which gives a few other assumptions, namely 5.1, 5.3 and 5.4, that imply 2.4 when they hold for the same set D. We will use the set $D = \mathcal{B}$, which, as we will see, makes our assumption 1 directly imply assumption 5.1, and makes 5.3 and 5.4 hold trivially. Note that our choice of $D = \mathcal{B}$ is mostly used to get a simple representation of a class of POMDPs for which Blackwell optimal policies exist, since then only Assumption 1 is needed.

We are now ready to prove our main result of this section.

Theorem 6.1. *A POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$ satisfying Assumption 1 has a deterministic belief-based policy $\pi^* \in \Pi_{DB}$ that is Blackwell optimal in the class Π_B of belief-based policies.*

We start by giving some helper lemmas, and continue by showing that all the necessary assumptions hold. We then conclude by proving the theorem.

6.1.3 Helper lemmas

From (2.3) in [15], for a function f on $\mathcal{B}$ for which it is well-defined, $\mathbf{P}f$ is given by:

$$\mathbf{P}f(x, a) := \int_{\mathcal{B}} f(y)\, \mathbf{P}(x, a, dy).$$

This operator $\mathbf{P}$ is used a number of times in the assumptions, mainly with the bounding function that we will need to give on the rewards in Assumption 2.2. Because of the countability of $\mathcal{B}$, we can simply turn this integral into a sum.

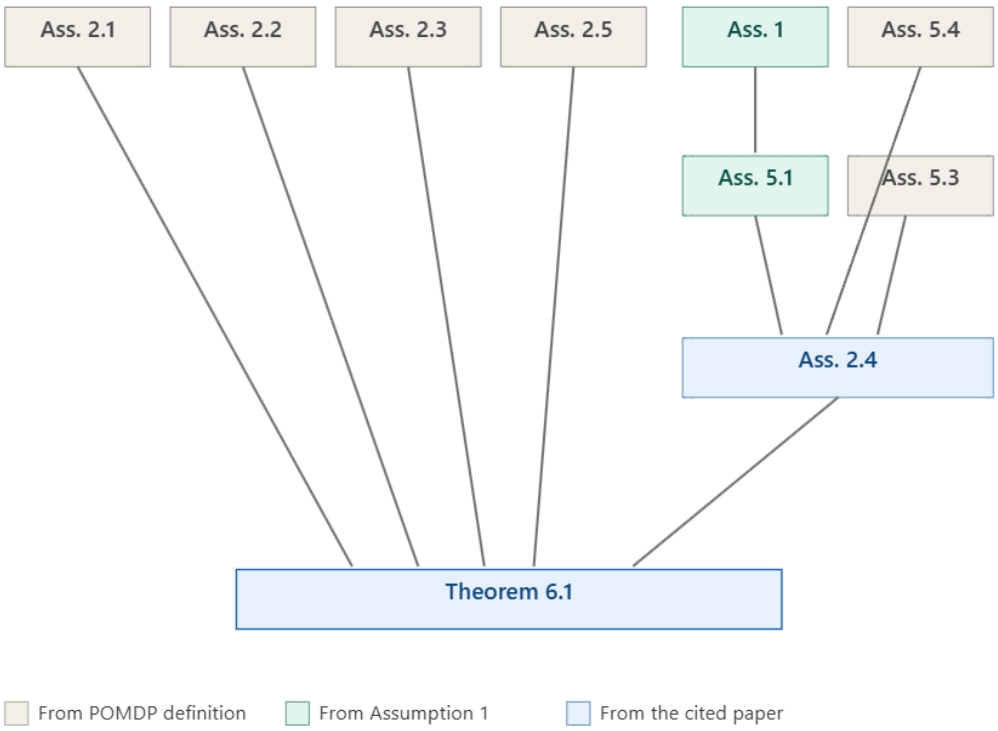

Figure 3: Proof structure for Blackwell optimality

Lemma 6.2. $\mathbf{P}f(x,a) = \sum_{y' \in \mathcal{B}} f(y')\mathcal{P}(x,a,y')$.

Proof. We can write the transition function $\mathbf{P}(x,a,dy)$ for $dy \subset \mathcal{B}$ as:

$$\mathbf{P}(x,a,dy) = \sum_{y' \in \mathcal{B}} \mathcal{P}(x,a,y')\delta_{y'}(dy),$$

where $\delta_{y'}$ is the Dirac measure. Now, from (1) we have:

$$\begin{aligned}
\mathbf{P}f(x,a) &= \int_{\mathcal{B}} f(y)\,\mathbf{P}(x,a,dy) \\
&= \int_{\mathcal{B}} f(y) \left(\sum_{y' \in \mathcal{B}} \mathcal{P}(x,a,y')\delta_{y'}(dy) \right) \\
&= \sum_{y' \in \mathcal{B}} \mathcal{P}(x,a,y') \int_{\mathcal{B}} f(y)\,\delta_{y'}(dy) \\
&= \sum_{y' \in \mathcal{B}} \mathcal{P}(x,a,y') f(y') \\
&= \sum_{y' \in \mathcal{B}} f(y')\mathcal{P}(x,a,y').
\end{aligned}$$

□

For a few functions, we will need to show continuity in a. This next lemma will help.

Lemma 6.3. *Any function $f : A \to \mathbb{R}$ is continuous (under the discrete topology).*

Proof. Let $U \subseteq \mathbb{R}$ be open. Consider the preimage

$$f^{-1}(U) = \{a \in A : f(a) \in U\}.$$

Since $f^{-1}(U)$ is a subset of A, and every subset of a discrete topological space is open, it follows that $f^{-1}(U)$ is open in A.

Therefore, f is continuous by Proposition 3.1.

□

6.1.4 Proving the assumptions

We now show that all the required assumptions hold. For clarity, the assumption definitions are copied here, with changed notation to be consistent with the rest of this paper.

Assumption 2.1. *The state space $\mathcal{B}$ is a Borel space, the action space A is a topological Borel space, the action sets A_b are nonempty compact b-sections of a Borel measurable set K in $\mathcal{B} \times A$. The reward function $\bar{r}(\mathbf{b}, a)$ and the transition probability $\boldsymbol{P}(\mathbf{b}, a, B)$ for any Borel set $B \subseteq \mathcal{B}$ are measurable functions of $(\mathbf{b}, a)$ on K.*

Lemma 6.4. *The Belief MDP of any POMDP satisfies Assumption 2.1.*

Proof. The belief state space $\mathcal{B}$ is a Borel space [10].
A *topological Borel space*, as defined in [15], corresponds to the definition of *Borel space* in [31]. The action set A satisfies this definition under the discrete topology, since it is a finite set [31].
We assume $A_b = A$ as usual in a POMDP [5] , so they are compact sets. $K = \mathcal{B} \times A$ is measurable since it is a product of two Borel sets.
$\bar{r}$ is clearly continuous in $\mathbf{b}$ by definition, and therefore measurable. Since A is finite $\bar{r}$ itself is also measurable in $\mathcal{B} \times A$.

Checking that the transition function is measurable is more complicated.

For each $a \in A$ and $o \in Z$, the observation likelihood

$$\Pr(o \mid a, \mathbf{b}) = \sum_{s' \in S} O(s', o) \sum_{s \in S} P(s, a, s')\, \mathbf{b}(s)$$

is linear in $\mathbf{b}$, and therefore continuous. The belief update

$$\mathbf{b}^{a,o}(s') = \frac{O(s', o) \sum_{s \in S} P(s, a, s')\, \mathbf{b}(s)}{\Pr(o \mid a, \mathbf{b})}$$

is a ratio of continuous functions whose denominator is strictly positive whenever $\Pr(o \mid a, \mathbf{b}) > 0$. Therefore, it is continuous in $\mathbf{b}$ on its domain. Consequently, for every Borel set $B \subset \mathcal{B}$,

$$\mathbf{P}(\mathbf{b}, a, B) = \sum_{o \in Z} \mathbf{1}_B\big(\mathbf{b}^{a,o}\big)\ \Pr(o \mid a, \mathbf{b})$$

is a finite sum of products of continuous functions of $\mathbf{b}$. Hence, this is continuous and therefore measurable in $\mathbf{b}$. Since A is finite, joint measurability in $(\mathbf{b}, a)$ on K follows immediately. So $\mathbf{P}$ is measurable. Therefore, Assumption 2.1 holds. □

Assumption 2.2. *A measurable bounding function $\mu \geq 1$ is given, and the reward function $\bar{r}$ and transition operator $\boldsymbol{P}$ satisfy the conditions:*

(a) $\sup_{\mathbf{b}\in\mathcal{B}} \frac{|\hat{\bar{r}}(\mathbf{b})|}{\mu(\mathbf{b})} \leq 1$.

(b) for some constant $C > 0$, $\boldsymbol{P}\mu(\mathbf{b}, a) \leq C\mu(\mathbf{b})$ for all $(x, a) \in K$.

Lemma 6.5. *The Belief MDP of any POMDP satisfies Assumption 2.2.*

Proof. Recall that $\hat{\bar{r}}(\mathbf{b}) := \sup_{a\in A} |\bar{r}(\mathbf{b}, a)|$.

We define $\mu(\mathbf{b}) = M$, where $M \geq 1$ is an upper bound of the absolute value of the reward function, that is: $M \geq \max_{s,a} |r(s, a)|$. Then, we have

$$\sup_{\mathbf{b}\in\mathcal{B}} \frac{|\hat{\bar{r}}(\mathbf{b})|}{\mu(\mathbf{b})} = \sup_{(\mathbf{b},a)\in\mathcal{B}\times A} \frac{|\bar{r}(\mathbf{b}, a)|}{M} \leq 1,$$

and, by Lemma 2:

$$\mathbf{P}\mu(x, \mathbf{b}) = \sum_{\mathbf{b}'\in\mathcal{B}} \mu(\mathbf{b}')\mathcal{P}(\mathbf{b}, a, \mathbf{b}') = M = \mu(\mathbf{b}).$$

μ is measurable since it is a constant. So this definition of μ satisfies Assumption 2.2 with $C = 1$. □

Assumption 2.3. *(a) The reward function $\bar{r}(\mathbf{b}, a)$, $(\mathbf{b}, a) \in K$, is continuous in a.*

(b) $\boldsymbol{P}f(\mathbf{b}, a)$ is continuous in a for every bounded measurable function $f : \mathcal{B} \to \mathbb{R}$.

Lemma 6.6. *The Belief MDP of any POMDP satisfies Assumption 2.3.*

Proof. Both functions are continuous by Lemma 6.3, so Assumption 2.3 is satisfied. □

Assumption 2.5. *A σ-finite reference measure m on the space $\mathcal{B}$ is given, and*

(a) the transition probabilities are defined by means of a given transition density function $p(\mathbf{b}, a, \mathbf{b}')$, $(\mathbf{b}, a) \in K$, $\mathbf{b}' \in \mathcal{B}$, so that

$$\boldsymbol{P}(\mathbf{b}, a, B) = \int_B p(\mathbf{b}, a, \mathbf{b}')\, m(d\mathbf{b}'), \quad (\mathbf{b}, a) \in K,\ B \subseteq \mathcal{B},$$

where p is measurable, nonnegative, and its integral over $\mathcal{B}$ is identically 1;

(b) $p(\mathbf{b}, a, \mathbf{b}')$ is continuous in a;

(c) the transition density p is related to the bounding function μ by the condition

$$\int_{\mathcal{B}} \hat{p}(\mathbf{b}, \mathbf{b}')\, \mu(\mathbf{b}')\, m(d\mathbf{b}') \leq C\mu(\mathbf{b}), \quad \mathbf{b} \in \mathcal{B},$$

where $\hat{p}(\mathbf{b}, \mathbf{b}') \coloneqq \max_{a\in A_{\mathbf{b}}} p(\mathbf{b}, a, \mathbf{b}')$.

Lemma 6.7. *The Belief MDP of any POMDP satisfies Assumption 2.5.*

Proof. We take m to be the counting measure. That is $m := |X|$ for any set $X \subseteq \mathcal{B}$. Because $\mathcal{B}$ is countable, $\mathcal{B} = \bigcup_{\mathbf{b}\in\mathcal{B}}\{\mathbf{b}\}$ is a countable union of singletons each with $m(\{\mathbf{b}\}) = 1 < \infty$. Therefore m is a σ-finite measure.

The transition density function is then simply the transition function of the belief MDP, since (from (2))

$$\mathbf{P}(\mathbf{b}, a, B) = \sum_{\mathbf{b}' \in B} \mathcal{P}(\mathbf{b}, a, \mathbf{b}') = \sum_{\mathbf{b}' \in B} p(\mathbf{b}, a, \mathbf{b}') = \int_B p(\mathbf{b}, a, \mathbf{b}')\, m(d\mathbf{b}'),$$

and its integral over $\mathcal{B}$ is indeed $\mathbf{P}(\mathbf{b}, a, \mathcal{B}) = 1$. By Lemma 3, this function is continuous in a. We also have

$$\int_{\mathcal{B}} \hat{p}(\mathbf{b}, \mathbf{b}')\, \mu(\mathbf{b}')\, m(d\mathbf{b}') = \sum_{\mathbf{b}' \in \mathcal{B}} \hat{p}(\mathbf{b}, \mathbf{b}')\, M \leq M \sum_{a \in A} \sum_{\mathbf{b}' \in \mathcal{B}} p(\mathbf{b}, a, \mathbf{b}') = |A| M = |A| \mu(\mathbf{b}).$$

So Assumption 2.5 holds with this measure m. □

Assumption 5.1. *There exist two sets $D, X' \subseteq \mathcal{B}$ with $m(D) > 0$, $m(X') > 0$, and a number δ such that the transition density $p(\mathbf{b}, a, \mathbf{b}') \geq \delta > 0$ for $\mathbf{b} \in D$, $a \in A_{\mathbf{b}}$, $\mathbf{b}' \in X'$.*

Lemma 6.8. *For any POMDP satisfying Assumption 1, Assumption 5.1 holds with $D = \mathcal{B}$, the entire reachable belief space.*

Proof. From Assumption 1 there is a belief $\mathbf{b}'$ where for any $\mathbf{b} \in \mathcal{B}$, $a \in A$, we have $\mathcal{P}(\mathbf{b}, a, \mathbf{b}') \geq \delta > 0$. Then, picking $D = \mathcal{B}$ and $X' = \{\mathbf{b}'\}$, Assumption 5.1 is satisfied since $m(\mathcal{B}) > 0, m(\{\mathbf{b}'\}) > 0$, and $p(\mathbf{b}, a, \mathbf{b}') = \mathcal{P}(\mathbf{b}, a, \mathbf{b}') \geq \delta > 0$ for any $\mathbf{b} \in \mathcal{B}, a \in A$. □

Assumption 5.3. *There exist: 1) a set $D \subseteq \mathcal{B}$ with $m(D) > 0$ and $\sup_{\mathbf{b} \in D} \mu(\mathbf{b}) < \infty$, and 2) numbers $0 < \beta < 1$, $c > 0$, such that*

$$\boldsymbol{P}\mu(\mathbf{b}, a) \leq \beta \mu(\mathbf{b}) + c \cdot \mathbf{1}_D(\mathbf{b}) \quad \textit{for all } (\mathbf{b}, a) \in K.$$

Lemma 6.9. *The Belief MDP of any POMDP satisfies Assumption 5.3 with $D = \mathcal{B}$.*

Proof. Since $\mu(\mathbf{b}) = M$ for any $\mathbf{b} \in \mathcal{B}$, Assumption 5.3 is satisfied with $D = \mathcal{B}$, $\beta = 0.5$, $c = M$, since then we have for any $x \in \mathcal{B}$, $a \in A$:

$$P\mu(\mathbf{b}, a) = M \leq 0.5M + M = \beta \mu(\mathbf{b}) + c = \beta \mu(\mathbf{b}) + c \cdot \mathbf{1}_D(x)$$

□

Assumption 5.4. *There exists a set $D \subseteq \mathcal{B}$ such that for every sublevel set $M_c := \{\mathbf{b} \in \mathcal{B} : \mu(\mathbf{b}) \leq c\}$ there are an integer $N \geq 1$ and a number $\eta > 0$ such that uniformly in $\mathbf{b} \in M_c$ and $\pi \in \Pi_B$:*

$$P^N(\pi, \mathbf{b}, D) := \Pr_{\mathbf{b}}^{\pi}\{\mathbf{b}_N \in D\} \geq \eta > 0,$$

where $\mathbf{b}_N$ denotes the belief state after N steps starting from $\mathbf{b}$ under policy π.

Lemma 6.10. *The Belief MDP of any POMDP satisfies Assumption 5.4 with $D = \mathcal{B}$.*

Proof. The system will never leave $\mathcal{B}$, thus Assumption 5.4 is satisfied with $D = \mathcal{B}$, since for any $\mathbf{b} \in \mathcal{B}$, $\pi \in \Pi_B$ we have

$$\Pr_{\mathbf{b}}^{\pi}\{\mathbf{b}_1 \in D\} = 1.$$

So we can simply take $N = \eta = 1$ for any c. □

We now have all the necessary lemmas to prove the theorem.

Theorem 6.1. *A POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$ satisfying Assumption 1 has a deterministic belief-based policy $\pi^* \in \Pi_{DB}$ that is Blackwell optimal in the class Π_B of belief-based policies.*

Proof. By Theorem 5.1 of [11], the belief MDP satisfies assumption 2.4. Thus, by Theorem 2.1 of [11], there exists a deterministic belief-based policy that is Blackwell optimal in the class Π_B of belief-based policies. □

6.1.5 Blackwell optimality in the class of all policies

We have proven that a policy exists that is Blackwell optimal in the class of belief-based policies, but we would like this policy to be Blackwell optimal in the class of all policies. [11] does give this result under an additional assumption, namely Assumption 2.6. We do not use this assumption directly, but instead use the following restriction on p and m, which are given in Assumption 2.5.

Lemma 6.11. *If $m(\mathcal{B}) < \infty$ and the transition density $p(x, a, y)$ is bounded, then Assumption 2.6 holds.*

Proof. If $m(\mathcal{B}) < \infty$, Assumption 5.5 of [11] follows directly. Thus, by Lemma 5.2 of [11], Assumption 2.6 holds. □

So, we can strengthen our result if we redefine the measure m in a way so that $m(\mathcal{B}) < \infty$. Then we can use Theorem 2.2 of [11] to prove that we also have Blackwell optimality in the class of all policies.

For this, we need an enumeration of the belief state space, which is possible from $\mathcal{B}$ being countable. We can get an enumeration in a way similar to breadth-first search. We take $\mathbf{b}_0$ to be the initial belief, and assume an ordering on A and Z, which exists since both sets are finite. We can then iterate over the possible transitions from $\mathbf{b}_0$, starting from $\mathbf{b}_1 := \mathbf{b}_0^{a_1,o_1}$, $\mathbf{b}_2 := \mathbf{b}_0^{a_1,o_2}$, etc. After having gone through all of these possible transitions, we continue with the transition from $\mathbf{b}_1$, and so on. We do need to be careful that we are not redefining belief states that are already in the enumeration. In this way, we can always compute the enumeration of the belief state space.

Then, we can define $m(\mathbf{b}_i) = 2^{-i}$ so that $m(\mathcal{B}) = \sum_{i=0}^{\infty} m(\mathbf{b}_i) \leq 2$. This inequality becomes an equality if we do indeed have an infinite set of reachable beliefs, but this may not be the case.

We do now need to verify that Assumption 2.5 still holds for our new m: Since $m = \sum_i 2^{-i}\delta_{\mathbf{b}_i}$, integration with respect to m still reduces to a sum similar to (2), but now with weights 2^{-i}:

$$\int_B f(\mathbf{b}')\, m(d\mathbf{b}') = \sum_{\mathbf{b}_i \in B} f(\mathbf{b}_i) \cdot 2^{-i}.$$

We define the transition density $p(\mathbf{b}, a, \mathbf{b}_i) := 2^i \cdot \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i)$. Then for any $B \subseteq \mathcal{B}$:

$$\int_B p(\mathbf{b}, a, \mathbf{b}')\, m(d\mathbf{b}') = \sum_{\mathbf{b}_i \in B} 2^i \cdot \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i) \cdot 2^{-i} = \sum_{\mathbf{b}_i \in B} \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i) = \mathbf{P}(\mathbf{b}, a, B),$$

and p integrates to 1 since

$$\int_{\mathcal{B}} p(\mathbf{b}, a, \mathbf{b}')\, m(d\mathbf{b}') = \sum_i 2^i \cdot \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i) \cdot 2^{-i} = \sum_i \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i) = 1.$$

By Lemma 6.3, this function is continuous in a. For the bounding condition, the weights cancel:

$$\int_{\mathcal{B}} \hat{p}(\mathbf{b}, \mathbf{b}')\, \mu(\mathbf{b}')\, m(d\mathbf{b}') = \sum_i 2^i \cdot \hat{\mathcal{P}}(\mathbf{b}, \mathbf{b}_i) \cdot M \cdot 2^{-i}$$
$$\leq M \sum_{a \in A} \sum_i \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i) = |A| M = |A| \mu(\mathbf{b}).$$

So Assumption 2.5 still holds with this change of measure m. However, our new definition of the transition density $p(\mathbf{b}, a, \mathbf{b}_i) := 2^i \cdot \mathcal{P}(\mathbf{b}, a, \mathbf{b}_i)$ is unbounded for general POMDPs with an infinite reachable belief space. Without additional structure on the POMDP guaranteeing a decay of $\mathcal{P}(\mathbf{b}, a, \mathbf{b}_i)$ for an enumeration of $\mathcal{B}$ such that the transition density is bounded, we were unable to resolve this.

A different approach Even though we cannot easily satisfy Assumption 2.6, we can still directly prove Blackwell optimality in the class of all policies, without using further results of [11]. Namely, we know that for POMDPs, beliefs are a sufficient statistic, and there exists a discount optimal belief-based policy. From this we can derive that the policy π^* from Theorem 6.1 is also Blackwell optimal in the class of all policies Π. We can show that, for any discount factor $\gamma > 1$, π^* is discount optimal in Π since no belief-based policy outperforms it. Therefore, π^* is Blackwell optimal in Π. We now show this formally.

Lemma 6.12. *Given a POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$, suppose a policy π^* is Blackwell optimal in the class of belief-based policies Π_B. Then, it is also Blackwell optimal in the class of all policies Π.*

Proof. In the associated belief MDP, for any discount factor $\gamma \in (0, 1)$, there exists a stationary deterministic policy that is discount optimal (Theorem 6.2.9 of [1]). Stationary policies in the belief MDP correspond to belief-based policies in the POMDP. Therefore, for any γ, no policy outside Π_B can outperform the optimal policy within Π_B.

Since π^* is Blackwell optimal in Π_B, there cannot be another policy in Π_B that outperforms π^* for any γ sufficiently close to 1. Combined with the above, we have that π^* is discount optimal in Π for any γ sufficiently close to 1. That is, π^* is Blackwell optimal in the class of all policies. □

Theorem 6.13. *A POMDP $(S, A, P, r, Z, O, \mathbf{b}_0)$ satisfying Assumption 1 has a deterministic belief-based policy $\pi^* \in \Pi_{DB}$ that is Blackwell optimal in the class of all policies.*

Proof. Follows immediately from Theorem 6.1 and Lemma 6.12. □

6.2 Blackwell optimality in RPOMDPs

In Section 5, we proved the existence of average optimal policies in the class of mixed FSC policies. It is an interesting problem to see if we can prove Blackwell optimality for RPOMDPs within this same class of policies. We can prove this in a similar way that [9] proves it for RMDPs. Here, we will need dynamic sa-rectangular uncertainty.

It does need a result about discount optimality: we need that in this class of mixed FSC policies, for any discount factor γ, there exists a deterministic FSC that is discount optimal. There is some literature on discount optimality in RPOMDPs (see Section 2), but they mainly focus on approximation algorithms and do not give any results that

would show this property. We leave this as an open question and continue the proof by assuming this property.

Assumption 2. *For any memory bound k and discount factor γ, there exists a deterministic FSC $\pi \in \Pi_{FSC_k}$ that is discount optimal in the class of mixed FSCs with $|N| = k$ memory states.*

Other than for our average optimality results, this time we do need restrictions on the nature policies as well. Like for the RMDP case in [9], we need that the uncertainty set $\mathcal{U}$ is sa-rectangular and *definable*. Informally, the definition of a definable set is that it can be described using polynomials, the exponential function, and projections that eliminate variables. For completeness, we state the formal definition here [9].

Definition 6.14 (Definable set and definable function)**.** A subset of $\mathbb{R}^n$ is *definable* if it is the image, under a canonical projection $\mathbb{R}^{n+k} \to \mathbb{R}^n$ that eliminates any set of k variables, of a set of the form

$$\left\{x \in \mathbb{R}^{n+k} \mid \text{Poly}(x_1, \ldots, x_{n+k}, e^{x_1}, \ldots, e^{x_{n+k}}) = 0\right\},$$

where Poly$(\cdot)$ is a real polynomial in $2(n+k)$ variables. For definable subsets $\Omega \subseteq \mathbb{R}^n$ of $\mathbb{R}^n$, a function $f : \Omega \to \mathbb{R}^m$ is *definable* if its graph

$$\{(x, y) \in \Omega \times \mathbb{R}^m \mid y = f(x)\}$$

is a definable subset of $\mathbb{R}^{n+m}$.

However, we will not use this definition directly, since we can prove our results using lemmas from [9]. See [32] for a detailed coverage of definability.

We also require dynamic uncertainty. Then, for policy evaluation of FSCs, we can construct a Robust Markov chain that is equivalent to the RPOMDP, given a deterministic FSC $\pi_f \in \Pi_{FSC_k}$ [17] [33]. Here, the state space is the product of RPOMDP states S and FSC memory states N.

The uncertain transition and reward functions of the induced Markov chain are:

$$P^{\pi_f}(\langle s', n'\rangle \mid \langle s, n\rangle) = \sum_{o \in Z} O(s, o) \cdot P(s, \alpha(n, o))(s') \cdot \mathbf{1}[\eta(n, o, \alpha(n, o)) = n'],$$

$$r^{\pi_f}(\langle s, n\rangle) = \sum_{o \in Z} O(s, o) \cdot r(s, \alpha(n, o)).$$

That is, the new uncertainty set is $\mathcal{U}^{\pi_f} = \{P^{\pi_f} \mid P \in \mathcal{U}\}$, where each P^{π_f} is constructed from $P \in \mathcal{U}$ via the equation above.

We can now evaluate states with the robust value function. Here, sa-rectangularity and dynamic uncertainty are important since it allows us to take the infimum in each state of the induced Robust Markov chain. The robust value V^{π_f} is the fixed point of the following robust Bellman equation:

$$V_\gamma^{\pi_f}(\langle s, n\rangle) = r^{\pi_f}(\langle s, n\rangle) + \gamma \inf_{P^{\pi_f} \in \mathcal{U}^{\pi_f}} \sum_{s' \in S} \sum_{n' \in N} P^{\pi_f}(\langle s', n'\rangle \mid \langle s, n\rangle)\, V_\gamma^{\pi_f}(\langle s', n'\rangle). \quad (4)$$

For the existence of Blackwell optimal FSCs, we need to show that the function $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$ is definable for any mixed FSC policy π_f (where $\mathbf{V}$ is used as vector notation). We first show that this function is definable for any deterministic FSC, from which the result follows immediately. This will then lead to the existence of Blackwell optimal policies.

We now prove the definability of the robust value function. For this, we can follow the proof of Proposition 4.24 of [9]. The value function here is different, but still has the same form. The same steps still work, even though we need dynamic uncertainty, and [9] assumes static.

Lemma 6.15. *For an RPOMDP $(S, A, \mathcal{U}, r, Z, O, s_0)$ with dynamic uncertainty, suppose $\mathcal{U}$ is sa-rectangular and definable. Then, for any memory bound k, a deterministic FSC $\pi_f \in \Pi_{FSC_k}$ has a definable value function $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$.*

Proof. $\mathbf{V}_\gamma^{\pi_f}$ is the unique fixed point of the operator $V \mapsto T_\gamma^{\pi_f,\mathcal{U}}(V)$, where $T_\gamma^{\pi_f,\mathcal{U}} : \mathbb{R}^{S\times N} \to \mathbb{R}^{S\times N}$ is defined as

$$T_{\gamma,\langle s,n\rangle}^{\pi_f,\mathcal{U}}(V) = r^{\pi_f}(\langle s,n\rangle) + \gamma \inf_{P^{\pi_f}\in\mathcal{U}^{\pi_f}} \sum_{s'\in S}\sum_{n'\in N} P^{\pi_f}(\langle s',n'\rangle \mid \langle s,n\rangle)\, V(\langle s',n'\rangle),$$

for all $\langle s,n\rangle \in S \times N$ and $V \in \mathbb{R}^{S\times N}$. The proof proceeds in two steps: we first show that $(V,\gamma) \mapsto T_\gamma^{\pi_f,\mathcal{U}}(V)$ is definable, and then use this to conclude that $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$ is definable.

First step. We show that $(V,\gamma) \mapsto T_\gamma^{\pi_f,\mathcal{U}}(V)$ is definable. Since a function is definable if and only if each of its components is definable [32], it suffices to show that

$$(V,\gamma) \mapsto r^{\pi_f}(\langle s,n\rangle) + \gamma \inf_{P^{\pi_f}\in\mathcal{U}^{\pi_f}} \sum_{s'\in S}\sum_{n'\in N} P^{\pi_f}(\langle s',n'\rangle \mid \langle s,n\rangle)\, V(\langle s',n'\rangle)$$

is definable for each fixed $\langle s,n\rangle \in S \times N$.

$\mathcal{U}^{\pi_f}$ is the image of the definable set $\mathcal{U}$ under the map $P \mapsto P^{\pi_f}$. This map is linear, hence definable (See Example 4.14 of [9]). Thus, by Lemma 4.15 of [9], $\mathcal{U}^{\pi_f}$ is a definable set.

The map

$$(V, P^{\pi_f}) \in \mathbb{R}^{S\times N} \times \mathcal{U}^{\pi_f} \mapsto \sum_{s'\in S}\sum_{n'\in N} P^{\pi_f}(\langle s',n'\rangle \mid \langle s,n\rangle)\, V(\langle s',n'\rangle)$$

is polynomial in its arguments and defined on a definable set, hence definable. By Lemma 4.15 of [9], this implies that

$$V \in \mathbb{R}^{S\times N} \mapsto \inf_{P^{\pi_f}\in\mathcal{U}^{\pi_f}} \sum_{s'\in S}\sum_{n'\in N} P^{\pi_f}(\langle s',n'\rangle \mid \langle s,n\rangle)\, V(\langle s',n'\rangle)$$

is also definable. Since multiplying by γ and adding the constant $r^{\pi_f}(\langle s,n\rangle)$ preserves definability (Lemma 4.15 of [9]), the full map $(\gamma, V) \mapsto T_{\gamma,\langle s,n\rangle}^{\pi_f,\mathcal{U}}(V)$ is definable.

Second step. We show that $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$ is definable by showing its graph is a definable set. Since $\mathbf{V}_\gamma^{\pi_f}$ is the unique fixed point of $T_\gamma^{\pi_f,\mathcal{U}}$, the graph

$$\left\{(V,\gamma) \in \mathbb{R}^{S\times N} \times \mathbb{R} \mid V = \mathbf{V}_\gamma^{\pi_f}\right\}$$

equals

$$\left\{(V,\gamma) \in \mathbb{R}^{S\times N} \times (0,1) \mid T_\gamma^{\pi_f,\mathcal{U}}(V) = V\right\},$$

which is the projection onto $\mathbb{R}^{S\times N} \times (0,1)$ of the intersection

$$\left\{(V,\gamma,Z) \in \mathbb{R}^{S\times N} \times \mathbb{R} \times \mathbb{R}^{S\times N} \;\middle|\; Z = T_\gamma^{\pi_f,\mathcal{U}}(V)\right\} \cap \left\{(V,\gamma,Z) \in \mathbb{R}^{S\times N} \times \mathbb{R} \times \mathbb{R}^{S\times N} \;\middle|\; Z = V\right\}.$$

The first set is the graph of $(V, \gamma) \mapsto T_\gamma^{\pi_f, \mathcal{U}}(V)$, which is definable by the first step. The second set is set of zeroes of the affine map $(V, Z) \mapsto V - Z$, which is definable. Their intersection is therefore definable, and its projection onto $\mathbb{R}^{S \times N} \times (0, 1)$ is definable by the definition of definable sets. Hence $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$ is definable. □

Lemma 6.16. *For any mixed FSC policy $x \in \Delta(\Pi_{FSC_k})$ with memory bound k, the value function $\gamma \mapsto V_\gamma^{\pi_f}$ is definable.*

Proof. The value function of a mixed FSC policy is a convex combination of the value functions of deterministic FSC policies:

$$\mathbf{V}_\gamma^x = \sum_{\pi_f \in \Pi_{FSC_k}} x(\pi_f)\, \mathbf{V}_\gamma^{\pi_f}.$$

By Lemma 6.15, each $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f}$ is definable. Since Π_{FSC_k} is a finite set, $\mathbf{V}_\gamma^x$ is a finite linear combination of definable functions with constant coefficients $x(\pi_f) \geq 0$. By Lemma 4.15 of [9], scalar multiplication and addition preserve definability, hence $\gamma \mapsto \mathbf{V}_\gamma^x$ is definable. □

We are now ready to show Blackwell optimality. As noted before, we do need deterministic FSCs to be discount optimal. This gives us a finite set of possible Blackwell optimal policies. From the definability of the value function, it follows that a Blackwell optimal policy must exist.

Theorem 6.17. *For an RPOMDP $(S, A, \mathcal{U}, r, Z, O, s_0)$ with dynamic uncertainty, suppose $\mathcal{U}$ is sa-rectangular and definable, and suppose Assumption 2 holds. Then, for any memory bound k, there exists a deterministic policy $\pi \in \Pi_{FSC_k}$ that is Blackwell optimal in the class of mixed FSC policies with $|N| = k$ memory states.*

Proof. From Lemma 4.15 of [9] and Lemma 6.16, the function $\gamma \mapsto \mathbf{V}_\gamma^{\pi_f} - \mathbf{V}_\gamma^{\pi_f'}$ is definable for each pair (π_f, π_f') of deterministic FSC policies in Π_{FSC_k}. By Theorem 4.22 of [9], it follows that $\gamma \mapsto V_\gamma^{\pi_f}(\langle s, n \rangle) - V_\gamma^{\pi_f'}(\langle s, n \rangle)$ does not change signs in a neighbourhood $(1 - \delta(\pi_f, \pi_f', \langle s, n \rangle), 1)$ for some $\delta(\pi_f, \pi_f', \langle s, n \rangle) > 0$, for each $\langle s, n \rangle \in S \times N$.

We define

$$\bar{\gamma} := \max_{\pi_f, \pi_f' \in \Pi_{FSC_k}, \langle s, n \rangle \in S \times N} 1 - \delta(\pi_f, \pi_f', \langle s, n \rangle).$$

Since both Π_{FSC_k} and $S \times N$ are finite, the maximum is taken over a finite set, thus exists, so $\bar{\gamma} < 1$.

Take a $\gamma_0 \in (\bar{\gamma}, 1)$. By Assumption 2, there exists a deterministic FSC $\pi^* \in \Pi_{FSC_k}$ that is discount optimal for γ_0 in the class of mixed FSCs.

This policy is Blackwell optimal since, from the definition of $\bar{\gamma}$, the sign of $\gamma \mapsto V_\gamma^{\pi^*}(\langle s, n \rangle) - V_\gamma^{\pi_f}(\langle s, n \rangle)$ does not change on $(\bar{\gamma}, 1)$ for any $\pi_f \in \Pi_{FSC_k}$ and $\langle s, n \rangle \in S \times N$. Since it is non-negative at γ_0, it remains non-negative on all of $(\bar{\gamma}, 1)$, so no deterministic FSC outperforms π^* for any $\gamma \in (\bar{\gamma}, 1)$.

Therefore, there exists a deterministic policy $\pi \in \Pi_{FSC_k}$ that is Blackwell optimal in the class of mixed FSC policies with $|N| = k$ memory states. □

6.3 ε-Blackwell optimality in RPOMDPs

Instead of looking at exact Blackwell optimality, we can also look at ε-Blackwell optimality. Then a policy just needs to be approximately discount optimal for all large discount factors.

Definition 6.18 (ε-Blackwell optimality)**.** For a given $\varepsilon > 0$, a policy π^* is ε-Blackwell optimal if there exists some discount factor $\gamma_\varepsilon \in (0, 1)$ such that:

$$\inf_{\theta \in \Theta} (1-\gamma)\mathbb{E}_{\pi^*,\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \geq \sup_{\pi \in \Pi} \inf_{\theta \in \Theta} (1-\gamma)\mathbb{E}_{\pi,\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] - \varepsilon, \quad \forall \gamma \in (\gamma_\varepsilon, 1).$$

This definition includes normalisation of the total discounted reward, since in general this may diverge as $\gamma \to 1$.

We could expect to be able to prove ε-Blackwell optimality for a larger class of agent policies. [14] guarantees ε-Blackwell optimality in POMDPs without any assumptions. For RMDPs, [9] gives results whenever the uncertainty set is sa-rectangular and compact.

It is natural to look at both these proofs and see if they can be replicated for RPOMDPs. However, in both cases, there are problems. The results of [9] are much stronger, as they prove that there is a single policy that is ε-Blackwell optimal for any $\varepsilon > 0$. The proof mainly focuses on that, and not on simple existence of ε-Blackwell optimal policies. [14] shows that this stronger result already does not hold for POMDPs, so we certainly don't have it for RPOMDPs.

The issue with [14] is more that the proof is very long and technical, and uses a lot of lemmas for POMDPs. To replicate all these steps for RPOMDPs could be possible in theory, but it does not seem feasible.

It does not rule this approach out completely, but due to time constraints we have not looked at this any further. It could be interesting for future work to look further into the proof of [14], and see if it is possible to replicate for RPOMDPs.

7 Discussion

We have left a number of open questions throughout this thesis. Here we will discuss them briefly.

Regarding average optimality in RPOMDPs, our result is based on a restriction to mixed FSCs as agent policies. We would like to have a result that doesn't restrict agent policies, and uses assumptions about the uncertainty set instead. We have explored this direction, but it remains an open question. One avenue for future work is to look more into the missing equality in Section 5.1. As noted there, looking at [28] is a promising direction.

In the proof for Blackwell optimality in POMDPs (Section 6.1), our choice of $D = \mathcal{B}$ is mostly used to get a simple representation of sufficient conditions for the existence of Blackwell optimal policies in a POMDP, as we only need Assumption 1. For future work, it could be interesting to characterise which POMDPs satisfy the assumptions of [11] for a different choice of D, broadening the class of POMDPs for which Blackwell optimality is guaranteed.

Due to time constraints, we leave ε-Blackwell optimality for RPOMDPs as an open question. A possible direction is to look further into the proof of [14], which gives this result for POMDPs, and investigate whether the proof can be extended to the robust setting.

8 Conclusion

In this thesis, we have made a first attempt at understanding average and Blackwell optimality in robust partially observable MDPs. We have also looked at Blackwell optimality in POMDPs, a previous gap in the literature.

We first looked at the existence of average optimal policies in RPOMDPs, for which we have found sufficient conditions on the agent policy. We did this using a distribution over deterministic FSCs as agent policies, proving that, given a bound on the memory, there exists an average optimal one within this set. With this method, the only restriction on nature policies is that they need to be finitely randomised. Surprisingly, no rectangularity assumptions are needed.

We then investigated Blackwell optimality in POMDPs, based on results about continuous state space MDPs. We were able to give Assumption 1 as a sufficient condition for the existence of Blackwell optimal policies.

Finally, we shortly looked at both ε-Blackwell and exact Blackwell optimality for RPOMDPs, and were able to prove a result about Blackwell optimal FSCs. Under the assumption that a deterministic FSC is discount optimal, we showed that there exists a deterministic Blackwell optimal FSC in the class of mixed FSCs, given a bound on the memory.

For future work, it would be of interest to show existence of average optimal policies under assumptions about the uncertainty set instead of restrictive conditions for the agent policies. There are also interesting options for broadening our Blackwell optimality results, both for POMDPs and RPOMDPs. This is discussed in more detail in Section 7.

References

[1] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[2] Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174, April 2010.

[3] Andrew J. Schaefer, Matthew D. Bailey, Steven M. Shechter, and Mark S. Roberts. Modeling Medical Treatment Using Markov Decision Processes. In Margaret L. Brandeau, François Sainfort, and William P. Pierskalla, editors, *Operations Research and Health Care: A Handbook of Methods and Applications*, pages 593–612. Springer US, Boston, MA, 2004.

[4] Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, October 2005.

[5] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, May 1998.

[6] Garud Iyengar. Robust Dynamic Programming. *Math. Oper. Res.*, 30:257–280, May 2005.

[7] Anton Schwartz. A Reinforcement Learning Method for Maximizing Undiscounted Rewards. pages 298–305, December 1993.

[8] Rommert Dekker and Arie Hordijk. Average, Sensitive and Blackwell Optimal Policies in Denumerable Markov Decision Chains with Unbounded Rewards. *Mathematics of Operations Research*, 13(3):395–420, August 1988.

[9] Julien Grand-Clément, Marek Petrik, and Nicolas Vieille. Beyond discounted returns: Robust Markov decision processes with average and Blackwell optimality, January 2025. arXiv:2312.03618 [math].

[10] Emmanuel Fernández-Gaucherand, Aristotle Arapostathis, and Steven I. Marcus. On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Annals of Operations Research*, 29(1):439–469, December 1991.

[11] Arie Hordijk and Alexander A. Yushkevich. Blackwell optimality in the class of all policies in Markov decision chains with a Borel state space and unbounded rewards. *Mathematical Methods of Operations Research*, 50(3):421–448, December 1999.

[12] Takayuki Osogami. Robust partially observable Markov decision process. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 106–115. PMLR, June 2015.

[13] Eline M. Bovy, Marnix Suilen, Sebastian Junges, and Nils Jansen. Imprecise Probabilities Meet Partial Observability: Game Semantics for Robust POMDPs, July 2024. arXiv:2405.04941 [cs].

[14] Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Blackwell Optimality in Markov Decision Processes with Partial Observation. *The Annals of Statistics*, 30(4):1178–1193, 2002.

[15] Arie Hordijk and Alexander A. Yushkevich. Blackwell optimality in the class of stationary policies in Markov decision chains with a Borel state space and unbounded rewards. *Mathematical Methods of Operations Research*, 49(1):1–39, March 1999.

[16] Hideaki Nakao, Ruiwei Jiang, and Siqian Shen. Distributionally Robust Partially Observable Markov Decision Process with Moment-based Ambiguity, December 2020. arXiv:1906.05988 [math].

[17] Maris F. L. Galesloot, Marnix Suilen, Thiago D. Simão, Steven Carr, Matthijs T. J. Spaan, Ufuk Topcu, and Nils Jansen. Pessimistic Iterative Planning with RNNs for Robust POMDPs, August 2025. arXiv:2408.08770 [cs.AI].

[18] Murat Cubuktepe, Nils Jansen, Sebastian Junges, Ahmadreza Marandi, Marnix Suilen, and Ufuk Topcu. Robust Finite-State Controllers for Uncertain POMDPs, March 2021. arXiv:2009.11459 [cs].

[19] Marnix Suilen, Nils Jansen, Murat Cubuktepe, and Ufuk Topcu. Robust Policy Synthesis for Uncertain POMDPs via Convex Optimization, January 2020. arXiv:2001.08174 [math.OC].

[20] Ioannis Tzortzis and Charalambos D. Charalambous. Partially Observable Markov Decision Subject to Total Variation Distance Ambiguity. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4794–4799, December 2022. ISSN: 2576-2370.

[21] D. J. H. Garling. *A Course in Mathematical Analysis: Volume 2: Metric and Topological Spaces, Functions of a Vector Variable*, volume 2. Cambridge University Press, Cambridge, 2014.

[22] D. J. H. Garling. *A Course in Mathematical Analysis: Volume 3: Complex Analysis, Measure and Integration*, volume 3. Cambridge University Press, Cambridge, 2014.

[23] Jean Dieudonné. *Treatise on Analysis*. Academic Press, 1969. Google-Books-ID: hxXvAAAAMAAJ.

[24] David Blackwell. Discrete Dynamic Programming. *The Annals of Mathematical Statistics*, 33(2):719–726, June 1962.

[25] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, February 2013.

[26] Richard D. Smallwood and Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon. *Operations Research*, 21(5):1071–1088, 1973.

[27] M. A. Lopez and D. E. Vercher. Convex semi-infinite games. *Journal of Optimization Theory and Applications*, 50(2):289–312, August 1986.

[28] Subhamay Saha. Zero-Sum Stochastic Games with Partial Information and Average Payoff, September 2014. arXiv:1409.4030 [math].

[29] Ali Asadi, Krishnendu Chatterjee, David Lurie, and Raimundo Saona. Revealing POMDPs: Qualitative and Quantitative Analysis for Parity Objectives, December 2025. arXiv:2511.13134 [cs.CC].

[30] Marius Belly, Nathanaël Fijalkow, Hugo Gimbert, Florian Horn, Guillermo A. Pérez, and Pierre Vandenhove. Revelations: A Decidable Class of POMDPs with Omega-Regular Objectives, December 2024. arXiv:2412.12063 [cs.AI].

[31] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, December 1996.

[32] M. Coste. AN INTRODUCTION TO O-MINIMAL GEOMETRY. 2002.

[33] Nicolas Meuleau, Kee-Eung Kim, Leslie Pack Kaelbling, and Anthony R. Cassandra. Solving POMDPs by searching the space of finite policies. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 417–426, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.