Complexity Estimates and Reductions to Discounting for Total and Average-Reward Markov Decision Processes and Stochastic Games

A Dissertation presented

by

Jefferson Huang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Operations Research)

Stony Brook University

August 2016

Stony Brook University

The Graduate School

Jefferson Huang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation

Eugene A. Feinberg
Distinguished Professor, Applied Mathematics and Statistics

Joseph S. B. Mitchell
Distinguished Professor and Chair, Applied Mathematics and Statistics

Jiaqiao Hu
Associate Professor, Applied Mathematics and Statistics

Jie Gao
Associate Professor, Computer Science

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

Complexity Estimates and Reductions to Discounting for Total and Average-Reward Markov Decision Processes and Stochastic Games

by

Jefferson Huang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Operations Research)

Stony Brook University

2016

Recently there has been a resurgence of interest in the complexity of algorithms for Markov decision processes (MDPs) and stochastic games. Much of this work was inspired by recent groundbreaking results on the complexity of policy iteration algorithms for MDPs under the Blum-Shub-Smale (BSS) model of computation. In particular, for discounted MDPs with a fixed discount factor, Yinyu Ye showed that the number of arithmetic operations needed by two classic variants of policy iteration can be bounded above by a polynomial in the number of state-action pairs only.

A natural question is whether a similar complexity estimate exists for the value iteration algorithm, which is another classic approach to computing optimal policies for MDPs. Our first main contribution is a negative answer to this question. Using a deterministic MDP with four state-action pairs, we show that under the BSS model there is no upper bound on the number of iterations needed by value iteration to return an optimal policy. We also show that the same example implies the same result for a broad class of so-called optimistic policy iteration algorithms, which includes algorithms of interest to the reinforcement learning community such as λ-policy iteration and modified policy iteration.

Another natural question is whether Ye's approach can yield results for MDPs under other optimality criteria. Our second main contribution is a formulation of conditions, which to our knowledge are the most general ones known, under which MDPs and two-player zero-sum stochastic games with Borel state and action spaces can be reduced to discounted ones. For undiscounted total-reward MDPs and stochastic games, the transformations we formulate are based on an idea due to Alan Hoffman. For average-rewards, the transformations are extensions to Borel state and action spaces of one proposed recently for finite stochastic games. In addition to implying the existence of ϵ-optimal policies for total and average rewards, these reductions lead to estimates of the number of arithmetic operations needed to compute optimal policies for such models with finite state and actions sets, as well as complexity estimates for computing ϵ-optimal policies for MDPs with Euclidean state and action spaces.

To Mom, Dad, and Aric.

Contents

Preface **1**

Acknowledgments **3**

1 Model Description & Literature Survey **5**

1.1 Description of the Decision Models 5

1.1.1 Markov Decision Processes (MDPs) 6

1.1.2 Two-Player Zero-Sum Stochastic Games 8

1.2 Complexity of MDPs & Games 10

1.2.1 Models of Computation 10

1.2.2 Upper Bounds 12

1.2.3 Lower Bounds 17

1.3 Reductions to Discounted Models 20

1.3.1 of Total-Reward Models 20

1.3.2 of Average-Reward Models 20

I Markov Decision Processes **22**

2 Discounted MDPs **23**

2.1 Optimality Criterion 23

2.2 Existence of Optimal Policies 24

2.3 Complexity Estimates 26

2.3.1 Policy Iteration (PI) 26

2.3.2 Optimistic PI 29

2.3.3 λ-PI, Modified PI, & Value Iteration 31

2.3.4 Computing ϵ-Optimal Policies 33

3 Reduction of Transient MDPs **38**
3.1 Transience Assumption 38
3.1.1 Applications 43
3.2 Optimality Criterion 44
3.3 Hoffman-Veinott (HV) Transformation 45
3.4 Existence of Optimal Policies 46
3.5 Complexity Estimates 49
3.5.1 Constructing the Transformation 50
3.5.2 Computing Optimal Policies 50
3.5.3 Computing ϵ-Optimal Policies 51

4 Reduction of Average-Reward MDPs **59**
4.1 Hitting Time Assumption 59
4.1.1 Applications 61
4.2 Optimality Criterion 62
4.3 Akian-Gaubert (AG) Transformation 62
4.4 Existence of Optimal Policies 63
4.5 Complexity Estimates 67
4.5.1 Constructing the Transformation 67
4.5.2 Computing Optimal Policies 68
4.5.3 Computing ϵ-Optimal Policies 69

II Two-Player Zero-Sum Stochastic Games **73**

5 Discounted Stochastic Games **74**
5.1 Optimality Criterion 74
5.2 Existence of ϵ-Optimal Strategies 75
5.2.1 Existence of Optimal Strategies 77

6 Reduction of Transient Stochastic Games **78**
6.1 Transience Assumption 78
6.2 Optimality Criterion 81
6.3 Hoffman-Veinott (HV) Transformation 82
6.4 Existence of ϵ-Optimal Strategies 88
6.4.1 Existence of Optimal Strategies 88
6.5 Complexity Estimates 90
6.5.1 Constructing the Transformation 90

6.5.2 Computing Optimal Strategies 91

7 Reduction of Average-Payoff Stochastic Games 93

7.1 Hitting Time Assumption 93

7.2 Optimality Criterion 94

7.3 Akian-Gaubert (AG) Transformation 95

7.4 Existence of ϵ-Optimal Strategies 97

7.4.1 Existence of Optimal Strategies 103

7.5 Complexity Estimates 104

7.5.1 Constructing the Transformation 104

7.5.2 Computing Optimal Strategies 105

Bibliography 106

Preface

Recently there has been a resurgence of interest in the complexity of algorithms for Markov decision processes (MDPs) and stochastic games; the latter can be viewed as a generalization of the former to allow for multiple decision-makers. Much of this work was inspired by Yinyu Ye's [121] groundbreaking results on the complexity of policy iteration algorithms for MDPs under the Blum-Shub-Smale (BSS) model of computation. In particular, for discounted MDPs with a fixed discount factor, Ye showed that two classic variants of policy iteration run in polynomial time under the BSS model. This answers an important special case of the more general question of whether there is an algorithm for linear programming that runs in polynomial time under the BSS model. The latter is a long-standing open problem in optimization and complexity theory; see e.g. Tardos [109] and Smale [108]. Hansen et al. [51] subsequently showed via a refinement of Ye's analysis that two-player zero-sum stochastic games of perfect information can also be solved in polynomial time under the BSS model, using a game-theoretic generalization of policy iteration.

A natural question is whether a similar complexity estimate exists for the value iteration algorithm, which is another classic approach to computing optimal policies for MDPs and stochastic games. The first main contribution of this dissertation is a negative answer to this question. In fact, using a deterministic MDP with four state-action pairs, we show that under the BSS model there is no upper bound (polynomial or otherwise) on the number of iterations needed by the value iteration algorithm to return an optimal policy. We also show that the same example implies an analogous result for a broad class of so-called optimistic policy iteration algorithms, which are designed to combine advantageous aspects of both value and policy iteration. This class includes algorithms of interest to the reinforcement learning community, such as λ-policy iteration and modified policy iteration.

Another natural question is whether Ye's approach can lead to results for MDPs under other optimality criteria. The second main contribution of this dissertation is motivated by the idea, which has been known since the 1960s, that certain undiscounted total-reward and average-reward MDPs can be reduced to discounted ones. We provide conditions, which to our knowledge are the most general ones known, under which MDPs and two-player zero-sum stochastic games with Borel state and action spaces can be reduced to discounted ones. For undiscounted total-reward MDPs and stochastic games, the transformation that we formulate is based on an idea that Pete Veinott [115] attributed to Alan Hoffman. For the average-reward criterion, the transformation is an extension to Borel state and action spaces of a transformation proposed by Akian & Gaubert [1]. These reductions lead, via recent work on discounted MDPs, to estimates of the number of arithmetic operations needed to compute optimal policies for MDPs and games with finite state and actions sets, as well as ϵ-optimal policies for MDPs with Euclidean state and action spaces.

For transient MDPs and games with transition kernels that are not necessarily substochastic, we also provide model formulations that allow for randomized history-dependent policies. To our knowledge, this has not been done rigorously before. Application areas of such models include pursuit-evasion games [88] and controlled branching processes [90].

Acknowledgments

First and foremost, I'd like to thank my dissertation advisor, Eugene Feinberg; it has been an honor and a privilege to be your student. Thank you for giving me the confidence to pursue a career in research. I hope we have many productive years of collaboration ahead of us! In addition, I'm grateful to Joe Mitchell, Estie Arkin, and Loretta Au for supporting me during my first forays into teaching, and to Jiaqiao Hu for introducing me to stochastic modeling. And, I'm glad to have had Andrea, Cathy, Christine, Janice, Laurie, and Victor around when I needed help.

I'd also like to thank Manasa Mandava; I wonder if I'll ever meet anyone else I can talk to seriously about both descriptive set theory and Ashtanga yoga! I hope I'll see you in Mysore someday. I'm also glad to have shared an office with Muqi Li; thanks for teaching me about power systems and parallel computing! In addition, I'm grateful for conversations with Shingo Omori about reinforcement learning, with Yan Liang about inventory control, with Francisco Hawas about machine learning, and with Chelsea Kennedy about being a grad student. I also want to thank Chen Cai for organizing the AMS Graduate Reading Group; I wish I could've done something like that when I was a first-year student!

I feel lucky to have spent the summer of 2015 as an intern at IBM's T. J. Watson Research Center. I'd like to thank Rui Zhang and the Smarter Energy group for being so welcoming and supportive, Bo Zhang for inviting me to give an AP for Lunch talk, and Marek Petrik for the stimulating discussions. I also had fun times with Adham, Arzoo, Avner, Bowen, Carlos, Josh, Leke, Mahanth, Maria, Muhammad, Parnika, Qiancheng, Saad, Wubai, Zhaowei, and Zi.

I'm also grateful to the RLAI group at the University of Alberta for a lovely three days in Edmonton. I had wonderful conversations with Tor Lattimore, Jaeyoung Lee, Viliam Lisý, Rupam Mahmood, Martin Müller, Patrick Pilarski, Roshan Shariff, Rich Sutton, Csaba Szepesvári, and Janey Yu. You made me feel that academia is where I belong.

Finally, I'd like to thank Danielle Goldstein for being there for me both as a teacher and as a friend, Beryl Bender Birch for making me think more deeply about how I spend my time and energy, and Christine Hoar for showing me that I'm capable of much more than I think I am. Along with Adrienn, Allan, Alyssa, Chris & Olivia, Cuiyu, Eva, Elizabeth, Gab, Dhruva & Jackie, Janet, Jessica, Jill, Karen, Kat, Katie, Kristen, Laura B., Laura T. P., Lisa, Lizzy, Manasa, Megan, Tom, Steve & Fran, and everyone else I had the privilege of practicing with, you made waking up with the sun a joy.

Chapter 1

Model Description & Literature Survey

In this introductory chapter, we set the notation to be used in the sequel for the decision models that we will consider, and provide a review of relevant literature. In Section 1.1, we define the Markov decision process (MDP) (Section 1.1.1) and stochastic game (Section 1.1.2) models. In Section 1.2, we review the literature on the computational complexity of obtaining optimal policies for MDPs and stochastic games, including work on both upper bounds (Section 1.2.2) and lower bounds (Section 1.2.3). Finally, in Section 1.3 we review prior work on reducing undiscounted MDPs and stochastic games to discounted ones.

1.1 Description of the Decision Models

We begin by recalling some definitions that will be used to define the models considered in the sequel. It is assumed that the reader is familiar with the basics of general topology and measure-theoretic probability; see e.g. [83] and [63], respectively.

A *Polish space* is a separable topological space that is completely metrizable. Given a Polish space S, let $\mathcal{B}(S)$ and $\mathcal{P}(S)$ respectively denote the Borel σ-algebra on S and the set of all probability measures on $(S, \mathcal{B}(S))$. A function $f : S \to \mathbb{R}$ on a Polish space S is *universally measurable* if for every $p \in \mathcal{P}(S)$ there is a Borel function $f_p : S \to \mathbb{R}$ such that $f(s) = f_p(s)$ for p-almost every s; see e.g. [10, Lemma 7.27]. For Polish spaces S and

T, a *universally* (resp. *Borel-*) *measurable stochastic kernel* on T given S is a function $\kappa : \mathsf{S} \times \mathcal{P}(\mathsf{T}) \to [0, 1]$ such that $\kappa(\cdot|s)$ is a probability measure on $(\mathsf{S}, \mathcal{B}(\mathsf{S}))$ for $s \in \mathsf{S}$, and $\kappa(B|\cdot)$ is a universally (resp. Borel-) measurable function for $B \in \mathcal{B}(\mathsf{T})$.

For any sets S, T we use the convention $(\mathsf{S})^0 \times \mathsf{T} := \mathsf{T}$. Further, unless stated otherwise, all products of families $\{\mathsf{S}_\alpha : \alpha \in \Lambda\}$ of topological spaces are endowed with their corresponding product topologies and product σ-algebras defined by the Borel σ-algebras on the spaces S_α, $\alpha \in \Lambda$.

1.1.1 Markov Decision Processes (MDPs)

A *Markov decision process (MDP)* is defined by a tuple $(\mathbb{X}, \mathbb{A}, \{A(x) : x \in \mathbb{X}\}, r, p)$. The elements of this tuple are associated with a controlled stochastic system as follows. On each time step $t = 0, 1, \dots$, the decision-maker observes the current state $x_t \in \mathbb{X}$ of the system, selects an available action $a_t \in A(x_t) \subseteq \mathbb{A}$, and earns a reward $r(x_t, a_t)$. The actions performed by the decision-maker affect the system she is controlling in the following way: if the state of the system is $x_t \in \mathbb{X}$ at time t and action $a_t \in A(x_t)$ is performed, then the state she will observe at time $t + 1$ belongs to the (measurable) subset B of $\mathbb{X}$ with probability $p(B|x_t, a_t)$.

Remark 1. Throughout, we assume that the one-step reward function $r : \mathrm{Gr}(A) \to \mathbb{R}$ is bounded.

More precisely, we consider the usual framework of MDPs with possibly uncountable state and action spaces; see e.g. [10], [25], [53]. Namely, the *state space* $\mathbb{X}$ and *action space* $\mathbb{A}$ are Borel subsets of Polish spaces. Furthermore, the *set of available actions* $A(x)$ when the current state is $x \in \mathbb{X}$ is a nonempty Borel-measurable subset of $\mathbb{A}$, and the set

$$\mathrm{Gr}(A) := \{(x, a) : x \in \mathbb{X}, a \in A(x)\}$$

is a Borel subset of $\mathbb{X} \times \mathbb{A}$. Finally, the *one-step reward function* $r : \mathrm{Gr}(A) \to \mathbb{R}$ is Borel-measurable, and p is a Borel-measurable stochastic kernel on $\mathbb{X}$ given $\mathrm{Gr}(A)$ which defines the *transition probabilities*.

Policies. The decision-maker controls the system via a *policy*, according to which actions can be selected randomly based on the entire observed history of the system up to the current time step. To define the notion of a policy, for $t = 0, 1, \dots$ let $\mathbb{H}_t := (\mathbb{X} \times \mathbb{A})^t \times \mathbb{X}$ denote the *space of histories* up to time step t. A *decision rule* for time step t is defined by a universally measurable stochastic kernel π_t on $\mathbb{A}$ given $\mathbb{H}_t$, where for each $h_t := x_0 a_0 \cdots a_{t-1} x_t \in \mathbb{H}_t$ the support of the probability measure $\pi_t(\cdot|h_t)$ is $A(x_t)$. A *policy* is a sequence $\pi = \{\pi_t\}_{t=0}^{\infty}$ of decision rules; when the decision-maker follows a given policy π, and the observed history up to time step t is h_t, she selects an action belonging to the set $B \in \mathcal{B}(\mathbb{A})$ with probability $\pi_t(B|h_t)$. Let Π denote the set of all policies.

Deterministic Stationary Policies. An important class of policies is the class of so-called *deterministic stationary* policies, which is identified with the class of universally measurable selectors of the correspondence $x \mapsto A(x)$. More precisely, a policy $\pi \in \Pi$ is a deterministic stationary policy if there is a universally measurable function ϕ from $\mathbb{X}$ to $\mathbb{A}$ where $\phi(x) \in A(x)$ for all $x \in \mathbb{X}$ and $\pi_t(\cdot|h_t) = \delta_{\phi(x_t)}(\cdot)$ for $t = 0, 1, \dots$ and all $h_t = x_0 a_0 \cdots a_{t-1} x_t \in \mathbb{H}_t$, in which case we identify π with ϕ. Hence, when the decision-maker follows the deterministic stationary policy ϕ, she selects the action $\phi(x)$ whenever the system is in state x. Let $\mathbb{F}$ denote the set of all deterministic stationary policies.

Probability Spaces Associated with Policies. When the initial state is $x \in \mathbb{X}$ and the decision-maker follows the policy $\pi \in \Pi$, the observed sequence of state-action pairs can be viewed as a realization of a discrete-time stochastic process. To make this statement precise, let $\mathbb{H}_\infty := (\mathbb{X} \times \mathbb{A})^\infty$ denote the *space of trajectories* of the MDP, and for $t = 0, 1, \dots$ define the random variables ξ_t and υ_t for $\omega = x_0 a_0 x_1 a_1 \cdots \in \mathbb{H}_\infty$ by $\xi_t(\omega) := x_t$ and $\upsilon_t(\omega) := a_t$. According to [10, Proposition 7.45], for every initial state $x \in \mathbb{X}$ and policy $\pi \in \Pi$ there is a unique probability measure $\mathbb{P}_x^\pi$ on $\mathbb{H}_\infty$, called a *strategic measure*, that satisfies the following conditions:

1. $\mathbb{P}_x^\pi(\mathbb{H}_\infty) = 1$;

2. $\mathbb{P}_x^\pi(\xi_0 = x) = 1$;

3. $\mathbb{P}_x^\pi(\upsilon_t \in B|h_t) = \pi_t(B|h_t)$ for all $B \in \mathcal{B}(\mathbb{A})$, $h_t \in \mathbb{H}_t$, and $t = 0, 1, \dots$;

4. $\mathbb{P}_x^\pi(\xi_{t+1} \in B|h_t, a_t) = p(B|x_t, a_t)$ for all $B \in \mathcal{B}(\mathbb{X})$, $h_t = x_0 a_0 \cdots x_t \in \mathbb{H}_t$, and $t = 0, 1, \dots$;

For $x \in \mathbb{X}$ and $\pi \in \Pi$, the expectation operator corresponding to $\mathbb{P}_x^\pi$ is denoted by $\mathbb{E}_x^\pi$.

Optimality Criteria. The relative attractiveness of policies is usually assessed via a chosen optimality criterion that is defined in terms of strategic measures; see e.g. [34]. Two commonly used criteria are expected total rewards and long-run expected average rewards per unit time; see e.g. [43, Chapter 6] and [9], respectively. For definitions of the former see Sections 2.1 and 3.2, and for a definition of the latter see Section 4.2.

1.1.2 Two-Player Zero-Sum Stochastic Games

A two-player zero-sum stochastic game can be viewed as a generalization of MDPs that allows for two decision-makers. In particular, instead of a single decision-maker selecting actions that affect the rewards earned and the trajectory of the system, the actions selected by two decision-makers, referred to as *players*, jointly affect the reward earned by one of them (i.e. the cost incurred by the other) and the distribution of the subsequent state.

More precisely, a *two-player zero-sum stochastic game* is defined by a tuple $(\mathbb{X}, \mathbb{A}^1, \mathbb{A}^2, \{A^1(x) : x \in \mathbb{X}\}, \{A^2(x) : x \in \mathbb{X}\}, r, p)$ where, analogously to the case of an MDP, $\mathbb{X}$ is the *state space*, $\mathbb{A}$ is the *action space*, and $A^i(x)$ is the *set of available actions* for player $i = 1, 2$ when the system is in state $x \in \mathbb{X}$. If the system is in state $x \in \mathbb{X}$ and players 1 and 2 select action $a^1 \in A^1(x)$ and $a^2 \in A^2(x)$, respectively, then player 2 pays player 1 $r(x, a^1, a^2)$ and the distribution of the next state is given by the probability measure $p(\cdot|x, a^1, a^2)$ on $\mathbb{X}$. Here we also assume that $\mathbb{X}$ and $\mathbb{A}$ are Borel subsets of Polish spaces, that for $i = 1, 2$ the sets $A^i(x)$ for $x \in \mathbb{X}$ are nonempty Borel subsets of $\mathbb{A}^i$, and that the set

$$\mathrm{Gr}(A^1 \times A^2) := \{(x, a^1, a^2) : x \in \mathbb{X}, a^i \in A^i(x), i = 1, 2\}$$

is a Borel subset of $\mathbb{X} \times \mathbb{A}^1 \times \mathbb{A}^2$. In addition, player 1's *payoff function* $r : \mathrm{Gr}(A^1 \times A^2) \to \mathbb{R}$ is Borel-measurable, and the transition probabilities are defined by the Borel-measurable stochastic kernel p on $\mathbb{X}$ given $\mathrm{Gr}(A^1 \times A^2)$.

Remark 2. Throughout, we also assume that player 1's one-step payoff function $r : \mathrm{Gr}(A^1 \times A^2) \to \mathbb{R}$ is bounded.

Strategies. Both players control the system via *strategies*, according to which actions can be selected randomly based on the entire observed history of the system up to the current time step. More precisely, here we let $H_t := (\mathbb{X} \times \mathbb{A}^1 \times \mathbb{A}^2)^t \times \mathbb{X}$ for $t = 0, 1, \ldots$ denote the *space of histories* up to time step t. For player $i = 1, 2$, a *decision rule* for time step t is defined by a universally measurable stochastic kernel π^i_t on $\mathbb{A}^i$ given H_t, where for each $h_t := x_0 a^1_0 a^2_0 \cdots a^1_{t-1} a^2_{t-1} x_t \in H_t$ the support of the probability measure $\pi^i_t(\cdot|h_t)$ is $A^i(x_t)$. For $i = 1, 2$ a *strategy* for player i is a sequence $\pi^i = \{\pi^i_t\}^{\infty}_{t=0}$ of decision rules; when player i follows a given strategy π^i, and the observed history up to time step t is h_t, she selects an action belonging to the set $B \in \mathcal{B}(\mathbb{A}^i)$ with probability $\pi^i_t(B|h_t)$. Let Π^i denote the set of all strategies for player i, for $i = 1, 2$.

Stationary Strategies. For stochastic games we will consider conditions under which *stationary* strategies are at least nearly optimal in some sense. A strategy π^i for player $i = 1, 2$ is stationary if there is a universally measurable stochastic kernel φ^i on $\mathbb{A}^i$ given $\mathbb{X}$ satisfying $\pi^i_t(\cdot|h_t) = \varphi^i(\cdot|x_t)$ for $t = 0, 1, \ldots$ and all $h_t = x_0 a^1_0 a^2_0 \cdots a^1_{t-1} a^2_{t-1} x_t \in H_t$, in which case we identify π^i with φ^i. When player i follows the stationary strategy φ^i, she selects an action belonging to the set $B \in \mathbb{A}^i$ with probability $\varphi^i(B|x)$ whenever the system is in state x. For $i = 1, 2$, let Φ^i denote the set of all stationary strategies for player i. In addition, a stationary strategy φ^i for player $i = 1, 2$ is *deterministic* if for every $x \in \mathbb{X}$ the measure $\pi(\cdot|x)$ is a Dirac measure; let $\mathbb{F}^i$ denote the set of all deterministic stationary strategies for player i.

Probability Spaces Associated with Strategy Pairs. Let $H_\infty := (\mathbb{X} \times \mathbb{A}^1 \times \mathbb{A}^2)^\infty$ denote the *space of trajectories* of the stochastic game, and for $t = 0, 1, \ldots$ define the random variables ξ_t, υ^1_t, and υ^2_t for $\omega = x_0 a^1_0 a^2_0 \cdots \in H_\infty$ by $\xi_t(\omega) := x_t$, $\upsilon^1_t(\omega) := a^1_t$, and $\upsilon^2_t(\omega) := a^2_t$. By [10, Proposition 7.45], for every initial state $x \in \mathbb{X}$ and pair of strategies $\pi^1 \in \Pi^1$ and $\pi^2 \in \Pi^2$ there is a unique probability measure $P^{\pi^1\pi^2}_x$ on H_∞ that satisfies the following conditions:

1. $P^{\pi^1\pi^2}_x(H_\infty) = 1$;

2. $P^{\pi^1\pi^2}_x(\xi_0 = x) = 1$;

3. for $i = 1, 2$, $P_x^{\pi^1\pi^2}(\upsilon_t^i \in B|h_t) = \pi_t^i(B|h_t)$ for all $B \in \mathcal{B}(\mathbb{A}^i)$, $h_t \in H_t$, and $t = 0, 1, \ldots$;
4. $P_x^{\pi^1\pi^2}(\xi_{t+1} \in B|h_t, a_t^1, a_t^2) = p(B|x_t, a_t^1, a_t^2)$ for all $B \in \mathcal{B}(\mathbb{X})$, $h_t = x_0 a_0^1 a_0^2 \cdots x_t \in H_t$, and $t = 0, 1, \ldots$;

For $x \in \mathbb{X}$, $\pi^1 \in \Pi^1$, and $\pi^2 \in \Pi^2$, let $E_x^{\pi^1\pi^2}$ denote the expectation operator corresponding to $P_x^{\pi^1\pi^2}$.

Optimality Criteria. Two commonly used optimality criteria for stochastic games involve total and long-run expected average payoffs. For definitions of the former see Sections 5.1 and 6.2, and for a definition of the latter see Section 7.2.

1.2 Complexity of MDPs & Games

In this section, we provide a review of the literature on the complexity of computing optimal policies for MDPs and stochastic games. In Section 1.2.1, we describe two models of computation that have been used in deriving complexity estimates for MDPs and stochastic games, and the notion of a polynomial-time algorithm. In Section 1.2.2, we review the literature on upper bounds for the complexity of computing optimal policies, including upper bounds for models with special structure such as ergodic MDPs and stochastic games and deterministic MDPs, and for MDPs with sensitive discount optimality criteria. Finally, in Section 1.2.3 we consider lower bounds on the complexity of computing optimal policies.

1.2.1 Models of Computation

The computational complexity of algorithms is typically analyzed in terms of the amount of resources (e.g. time, space) that the algorithm needs, as a function of the size of the input. Classically, inputs to the algorithm of interest are taken to be numbers represented by finite strings of symbols, and the algorithm itself is formalized as a Turing machine; see e.g. [107, Part Three] and [14, Section 1.3]. This is often referred to as the *Turing model of computation*.

More recently, Blum et al. [15] (see also [14]) proposed a model of computation that accommodates algorithms operating on real numbers (or more generally, on elements of a ring). This model, referred to as the *Blum-Shub-Smale (BSS) model of computation*, was designed to provide a rigorous foundation for the analysis of algorithms used in the field of numerical analysis (e.g. Newton's method), and can be viewed as a generalization of the classical theory of computation. For algorithms operating on real numbers, the BSS model is also referred to as the *arithmetic model of computation*; see e.g. [48, p. 32].

Polynomial-Time Algorithms

The resource that we are primarily interested in is time, which is typically measured via the number of elementary operations (e.g. the number of arithmetic operations) required by the algorithm of interest. A central notion in the analysis of the time complexity of algorithms is that of a *polynomial-time algorithm*, which is a formalization of the intuitive idea of an "efficient algorithm" that originated from the work of Cobham [16] and Edmonds [27]. In particular, a *polynomial-time algorithm* is an algorithm where the required number of elementary operations for an input of size $\mathcal{S}$ is bounded above by a polynomial in $\mathcal{S}$. A polynomial-time algorithm is also referred to as a *polynomial* algorithm, or as an algorithm that *runs in polynomial time*. We say that a problem can be *solved in polynomial time* if there exists a polynomial algorithm that, given any instance of that problem, computes a solution to the problem.

Turing Model. Under the Turing model of computation, an algorithm is *polynomial* if the total number of steps taken by its corresponding Turing machine can be bounded above by a polynomial function of the total bit-size of the input data. In other words, under this model of computation the elementary operations are the steps taken by the corresponding Turing machine, and the size $\mathcal{S}$ of an input is taken to be the total number of bits needed to represent that input. Hence the Turing model can be used to study algorithms operating on rational inputs, for example, but may not be appropriate when the inputs are taken to be any real numbers. One common way to derive upper bounds in this setting is to upper bound the number of arithmetic operations multiplied by the maximum number of bits needed to encode any input number; see e.g. [14, p. 19]. Polynomial algorithms under the Turing model are also called *weakly polynomial*; see e.g. [104, p. 48].

BSS Model. Under the BSS model of computation, an algorithm is *polynomial* if the total number of arithmetic operations that the algorithm requires in order to terminate can be bounded above by a polynomial function of the total number of input elements. For example, if an algorithm operating on real numbers takes any $n \times n$ real matrix as input, then the number of input elements for any given input is n^2. If a polynomial algorithm under the BSS model is such that, for rational inputs, the amount of space needed[1] can be bounded above by a polynomial in the total bit-size of the inputs, then it is also called *strongly polynomial*; see e.g. [48, p. 32] and [104, p. 47].

1.2.2 Upper Bounds

In this section, we provide a review of the literature on upper bounds for the complexity of computing an optimal policy. In Chapter 2, we describe the work of Feinberg & Huang [38] in detail.

Linear Programming

It has been known since almost the beginning of the development of the theory of MDPs that many important classes of these models can be solved via linear programming; see e.g. Kallenberg [64] and the references contained therein. According to the work of Khachiyan [68], it follows that these classes of models, which include discounted and average-reward MDPs, can be solved in polynomial time.

The study of interior-point methods for linear programming, which was stimulated by the work on Karmarkar [65], has led to improved complexity estimates; for example, by considering a combinatorial interior-point method, Ye [120] was the first to show that discounted MDPs with a fixed discount factor can be solved in polynomial time under the BSS

[1] Here, the amount of space needed is taken to be the maximum length of any string that appears on the tape of the algorithm's corresponding Turing machine while it computes the output; see e.g. [48, p. 24].

model of computation. In addition, Alagoz et al. [2] have obtained empirical evidence, using randomly generated problems and a real-life problem arising in healthcare, indicating that interior-point methods perform favorably compared to policy iteration.

Value Iteration

To our knowledge, the only published upper bounds for value iteration are for discounted MDPs. In this case, Tseng [112] showed that when the discount factor is fixed, the classic value iteration algorithm is weakly polynomial. In particular, each iteration can be performed in strongly polynomial time, and the total number of iterations can be bounded above by

$$\frac{nL + n\log_2 n}{1-\beta},$$

where n is the total number of states, L is the total bit-size of the input data, and β is the discount factor. This can be used to show that Howard's [62] policy iteration algorithm is also weakly polynomial; see e.g. [74].

Policy Iteration

While the weak polynomiality of Howard's policy iteration (PI) follows from Tseng's [112] result on value iteration, this conclusion was actually reached by Meister & Holzbaur [78] in a paper published four years earlier. In particular, they showed that the number of iterations required by Howard's PI is bounded above by a constant times $nL/[-\log(\beta)]$, where n is the number of states, L is the total bit-size of the input data, and β is the discount factor. Since each iteration of Howard's PI can be performed in strongly polynomial time, it follows that this algorithm is weakly polynomial.

A breakthrough result on the complexity of policy iteration and the question of whether MDPs or, more generally, linear programming problems can be solved in strongly polynomial time was published in 2011 [121]. In that paper, Yinyu Ye improved on his earlier result [120] on the solvability of discounted MDPs with a fixed discount factor in strongly polynomial time. In particular, this was done by showing that both the version of policy iteration corresponding to using the simplex method with Dantzig's pivoting rule, as well as Howard's [62] classic variant of policy iteration, actually have a superior iteration bound compared to the combinatorial interior-point method proposed in [120].

Ye's work subsequently led to a number of improvements and generalizations. Hansen et al. [51] both improved Ye's [121] bound for Howard's policy iteration algorithm, and showed that the same bound applies to the strategy iteration algorithm, proposed by Rao et al. [93], when it is used to solve discounted two-player zero-sum stochastic games of perfect information. This solved a long-standing open problem regarding the complexity of computing Nash equilibria for such discounted games. According to Andersson & Miltersen [7] the complexity of solving such games is closely related to the complexity of solving a number of other classes of games, including simple stochastic games [17], mean-payoff games of perfect information on graphs [28], and parity games [29]. More recently, Akian & Gaubert [1] improved on Hansen et al.'s [51] iteration bound, and used this result along with techniques from nonlinear Perron-Frobenius theory to derive a new iteration bound for the Hoffman-Karp algorithm [57] for solving a certain class of two-player zero-sum average-payoff stochastic games of perfect information.

Scherrer [103] improved on Ye's [121] and Hansen et al.'s [51] complexity estimates even further. Namely, it turns out that for discounted MDPs with a fixed discount factor, the number of iterations required by the simplex method with Dantzig's pivoting rule is linear in the number of state-action pairs m times the number of state n, and the number required by Howard's [62] policy iteration algorithm is linear in m. The latter agrees with what has been observed empirically about policy iteration; see [121].

Another research direction regarding upper bounds on the complexity of computing optimal policies for MDPs was initiated by Mansour & Singh [76] who, for discounted MDPs, derived the first nontrivial (but still exponential) iteration bound for policy iteration that does not depend on the discount factor. In particular, for any discounted MDP with n states and at most k actions per state, Howard's policy iteration terminates with an optimal policy after at most $13k^n/n$ iterations. Recently, Hollanders et al. [59] made the first progress in this direction since the work of Mansour & Singh. Using an abstract formulation of policy iteration, it was shown that for both discounted and average-reward MDPs, Howard's policy iteration requires at most

$$\frac{k}{k-1} \cdot \frac{k^n}{n} + \frac{k^n}{n^2}$$

iterations to return an optimal policy, where again n denotes the number of states and k denotes the maximum number of actions available at any state.

Finally, we remark that Ye's [121] method of analyzing the complexity of policy iteration has been generalized to a wider class of linear programming problems. Kitahara & Mizuno [70] showed that if every basic feasible solution of a linear program with m variables and n constraints is bounded above and below by γ and δ, respectively, then the simplex method with Dantzig's pivoting rule terminates after at most

$$m \left\lceil n \cdot \frac{\gamma}{\delta} \left(m \cdot \frac{\gamma}{\delta} \right) \right\rceil$$

iterations. Ye's [121] result on the simplex method with Dantzig's rule follows as a special case, because the linear programming formulation of discounted MDPs with discount factor β that is analyzed in [121] has m variables and n constraints, and every basic feasible solution of this linear program is bounded above by $\gamma = (1 - \beta)^{-1}$ and bounded below by 1. Another corollary that is observed in [70] is that, for linear programs with totally unimodular constraint matrices[2] and integral right-hand side vector b, the total number of iterations required by the simplex method with Dantzig's rule is at most

$$m \lceil n \|b\|_1 \ln n \|b\|_1 \rceil, \tag{1.1}$$

where $\| \cdot \|_1$ denotes the L^1 norm. This bound (1.1) applies to the linear programming formulations of the shortest path problem, the max-flow problem, and weighted bipartite matching; see [87, Section 13.2].

Models with Special Structure

The work of Ye [121] has also inspired a number of results on the complexity of computing optimal policies for MDPs and stochastic games with transition probabilities having certain special properties.

Ergodic MDPs & Games. An assumption on the transition probabilities that is often considered in the context of the average-reward criterion is the condition that the Markov chains associated with the deterministic

[2] i.e. the determinant of every square nonsingular submatrix is 1 or -1.

stationary policies are ergodic in some sense. For surveys of the kinds of assumptions that have been considered in the literature, see e.g.[33], [110], and [55]. One of the first complexity results in this direction was actually obtained a few years prior to Ye's [121] groundbreaking paper. Namely, Zadorojniy et al. [123] showed for MDPs that if all of the Markov chains induced by deterministic stationary policies are both irreducible and satisfy a so-called coupling property, then optimal policies can be computed in strongly polynomial time under both the discounted and average-reward criteria[3]. More recently, Feinberg & Huang [36] used Ye's [121] results and a reduction of certain average-reward MDPs to discounted ones due to Ross [95, 94] to show that MDPs modeling maintenance and replacement problems with a fixed failure probability can be solved in strongly polynomial time using policy iteration for average-reward MDPs. Akian & Gaubert [1] subsequently generalized the complexity estimates in Feinberg & Huang [36] to stochastic games and a more general ergodicity condition, using methods from nonlinear Perron-Frobenius theory.

Deterministic MDPs. A number of strongly polynomial algorithms exist for MDPs where the state transitions occur deterministically. For deterministic average-reward MDPs, the problem of computing an optimal policy is reducible to the problem of finding a minimum mean-weight cycle in a directed graph, and hence is solvable in strongly polynomial time using the algorithm proposed by Karp [66]. For the discounted-reward criterion, strongly polynomial algorithms have been proposed by Andersson & Vorobyov [8] and Madani et al. [75]. On the other hand, Post & Ye [91] showed that the classic simplex method with Dantzig's pivoting rule is strongly polynomial for discounted deterministic MDPs, regardless of the discount factor. Post & Ye's result was further improved by Hansen et al. [50], who also showed that the minimum cost to time ratio cycle problem is also solvable in strongly polynomial time.

[3] Here the discount factor is not necessarily fixed. In addition, we remark that the algorithm proposed by Zadorojniy et al. [123] was later shown by Even & Zadorojniy [31] to be equivalent to using the simplex method with the Gass-Saaty shadow vertex pivoting rule.

Sensitive Discount Optimality Criteria

There has also been some recent progress on the complexity of computing policies that are optimal under so-called sensitive discount optimality criteria, which generalize the average-reward criterion; see e.g. Veinott [116]. In his PhD thesis under Veinott, O'Sullivan [86] showed for the first time that Blackwell-optimal policies (see e.g. [61]), which have a number of attractive properties including being average-reward optimal and optimal under the discounted reward criterion for all discount factors sufficiently close to 1, can be computed in polynomial time. This was done by introducing a new algorithm that involves a reduction of the original problem to a sequence of linear programs.

On the other hand, it is unknown whether the classic policy iteration algorithm for computing optimal policies under sensitive discount optimality criteria (as well as Blackwell-optimal policies) proposed by Miller & Veinott [80] (see also [116]) is a polynomial-time algorithm. In particular, the best known upper bound for this algorithm, which follows from the work of Hollanders et al. [59], is exponential in the number of states.

1.2.3 Lower Bounds

In this section, we provide a review of the literature on lower bounds for the complexity of computing an optimal policy. In Chapter 2, we describe the work of Feinberg & Huang [37] and Feinberg et al. [39] in detail.

Value Iteration

To our knowledge, the only published lower bounds on the time complexity of value iteration are for the discounted-reward criterion, which is defined in Section 2.1 for MDPs and in Section 5.1 for two-player zero-sum stochastic games.

Consider a discount factor $\beta \in [0, 1)$. In an early survey of complexity estimates for value and policy iteration algorithms for MDPs, Littman et al. [74] provide an example of an MDP where value iteration requires at least

$$\frac{1}{2(1-\beta)} \cdot \log \frac{1}{1-\beta} \tag{1.2}$$

iterations to compute the optimal policy. This indicates that in general, it is not possible to derive an upper bound on the time complexity of value iteration that does not depend on the discount factor. In particular, it is not possible to remove the dependence on $(1-\beta)^{-1}$ of the upper bound for value iteration derived by Tseng [112]; see Section 1.2.2.

The discounted MDP that Littman et al. [74] used to prove their lower bound (1.2) on the time complexity of value iteration consists of four state-action pairs. Recently, by modifying the one-step reward associated with one of the actions, Feinberg et al. [39] showed that the number of iterations required by a broad class of so-called optimistic policy iteration algorithms to compute the optimal policy can grow arbitrarily quickly. Their result holds when the discount factor is fixed, and applies to value iteration, modified policy iteration [92], and λ-policy iteration [12]. In particular, it follows that unlike certain classic versions of policy iteration [121], the value iteration algorithm is not strongly polynomial; see also Feinberg & Huang [37]. In fact, the computational complexity of any of the aforementioned optimistic policy iteration algorithms is unbounded in the BSS model [15] of computation.

We remark that, since MDPs can be viewed as special cases of stochastic games, the preceding lower bounds also apply to value iteration for such games. In particular, they apply to the classical total-payoff two-player zero-sum stochastic game considered by Shapley [105]. For an example of a so-called simple stochastic game for which value iteration requires an exponential number of iterations to return a vector within a constant factor of the optimal value vector, see Condon [18].

Linear Programming & Policy Iteration

It has been known since at least the 1970s that there is a close relationship between policy iteration algorithms and the simplex method for solving linear programming problems. In particular, for discounted and certain average-reward MDPs there is a one-to-one correspondence between rules for updating actions during policy iteration and pivoting rules for the simplex method applied to a certain linear program; see e.g. Mine & Osaki [81, Sections 2.4, 3.5] and Kallenberg [64, pp. 67-68, 122].

Beginning with the seminal work of Klee & Minty [71], who showed that the simplex method with Dantzig's [21, p. 98] pivoting rule[4] can require an exponential number of iterations, many deterministic pivoting rules have been shown to take exponential time in the worst case; see e.g. Amenta & Ziegler [4] and Todd [111]. Similarly, many versions of policy iteration have been shown to have exponential iteration lower bounds. To our knowledge, the first such lower bound is due to Melekopoglou & Condon [79], who showed that the version of policy iteration corresponding to applying the simplex method with Bland's [13] pivoting rule can require an exponential number of iterations to compute an optimal policy under both the discounted and average-reward criteria. For the versions of policy iteration for undiscounted total and average-reward MDPs that use Howard's rule for updating actions[5], Fearnley [32] showed that an exponential number of iterations may be required. By modifying Fearnley's example, Hollanders et al. [58] showed that for a suitably large discount factor, policy iteration with Howard's rule may also require an exponential number of iterations. In addition, Friedmann [46, 47] obtained superpolynomial (namely, subexponential) lower bounds on the required number of iterations taken by versions of policy iteration for discounted MDPs corresponding to the simplex method with Zadeh's [122] pivoting rule and with Cunningham's [20] pivoting rule, respectively. We remark that Friedmann et al. [45] derived analogous superpolynomial lower bounds on the number of iterations taken by Dantzig's [21] random-edge pivoting rule Matoušek et al.'s [77] random-facet pivoting rule for discounted MDPs with a certain discount factor.

On the other hand, for discounted MDPs with a fixed discount factor, the best known lower bound on the number of iterations required by Howard's [62] policy iteration in the worst case are linear in the number of states; see the example due to John Tsitsiklis in [102] and Andersson et al. [6]. Hence the (strongly) polynomial upper bounds for both Howard's policy iteration and the simplex method with Dantzig's pivoting rule, which we describe in Section 1.2.2, are not known to be tight. It is interesting to note that, while Dantzig's classic pivoting rule served as the first example of an exponential pivoting rule [71], it also served as one of the first action switching rules that was shown to turn policy iteration into a polynomial-time algorithm under the BSS model of computation [121].

[4] Pivot the variable with the most negative reduced cost into the basis.

[5] For each state, switch to the action with the most negative reduced cost.

1.3 Reductions to Discounted Models

In this section, we provide a review of the literature on reducing the problem of solving undiscounted-reward MDPs and stochastic games to the problem of solving discounted ones. In Section 1.3.1 we consider the total reward criterion, while in Section 1.3.2 we consider average rewards.

1.3.1 of Total-Reward Models

A reduction of so-called transient MDPs, under the total reward criterion, to discounted MDPs is implicit in the work of Veinott [115]. There, a so-called positive similarity transformation based on ideas due to Alan Hoffman was used to reduce a transient MDP, where the (not necessarily substochastic) spectral radii of the transition matrices corresponding to the deterministic stationary policies are all less than one, to a transient MDP with substochastic transition matrices. The associated reduction to a discounted problem is made explicit in Kallenberg [64, pp. 75-77].

The transformation given in Veinott [115] has also been used to study MDPs with unbounded reward functions. Wessels [118] considered the convergence of value iteration for total-reward MDPs with discrete state sets where, for every deterministic stationary policy, the associated one-step reward function and transition matrix are bounded with respect to some weighted supremum norm. The model studied by Wessels [118] is also referred to as *contracting*; see e.g. [114], [113, p. 100]. The weight function used in Veinott [115] is also a special case of a kind of bounding function that van Hee and Wessels [114] call *strongly excessive*. In [114], characterizations of the existence of strongly excessive functions for MDPs are given in terms of the random drift through a partition of the state space, the lifetime distribution of the process under the Markov policies, and the spectral radii of the transition matrices induced by the Markov policies. van der Wal [113, Section 5.2] shows that for infinite-horizon total rewards, three sets of assumptions related to contracting MDPs are equivalent to the assumption that the MDP is discounted.

1.3.2 of Average-Reward Models

To our knowledge, the first reduction of average-reward MDPs to discounted ones is due to Sheldon Ross. In [95] a transformation is provided for MDPs with discrete state sets that, when there is a state to which the process transitions with probability at least $\alpha > 0$ under any action, allows one to compute an optimal policy under the average-reward criterion by solving a discounted MDP. Ross also provided an analogous transformation and result in [94] for MDPs with standard Borel state space and finite action sets. For a textbook treatment, see Ross [96, pp. 98-99]. Gubenko & Štatland [49] subsequently showed for MDPs with standard Borel state and action spaces that under a minorant condition on the transition probabilities, which generalizes the condition considered by Ross [94], a reduction of the original average-cost MDP to a discounted one is possible; see also Dynkin & Yushkevich [25, pp. 186-188]. More recently, Akian & Gaubert [1] showed that a reduction of average-payoff two-player zero-sum stochastic games with perfect information is possible under a condition that is more general than the one considered by Ross [95].

Gubenko & Štatland [49] also showed that a reduction to a kind of discounted MDP is possible when the transition probabilities satisfy a so-called majorant condition. Namely, in this case the resulting discounted MDP has a negative discount factor that belongs to the interval $(-1, 0)$. Hence, while the associated optimality operator still defines a contraction mapping, this mapping no longer has the monotonicity property that has played a central role in the study of the structure of sequential decision problems that was initiated by Denardo [22]; for a modern treatment with some extensions, see Bertsekas [11]. In particular, the techniques used by Ye [121], Hansen et al. [51], and Scherrer [103] to analyze discounted MDPs with a discount factor that lies on the interval $(0, 1)$ do not apply. To our knowledge, besides the work of Gubenko & Štatland [49] there are no other published results on discounted MDPs with a negative discount factor. We remark, however, that the work of Ames and Ginsburg [5] on iterative algorithms for nonlinear partial differential equations, where the same kind of oscillatory contraction mapping arises, may be relevant.

Part I

Markov Decision Processes

Chapter 2

Discounted MDPs

In this chapter, we review some results on discounted MDPs. In Section 2.1, discounted-reward optimality criterion is defined, and in Section 2.2 we state some results on the existence and characterization of optimal policies that will be used in Chapters 3 and 4. Finally, Section 2.3 provides results on the complexity of computing optimal and ϵ-optimal policies for discounted MDPs. In particular, Section 2.3.1 contains statements of some known results on the complexity of computing optimal policies, Section 2.3.2 contains the statements and proofs of our results on optimistic policy iteration, Section 2.3.3 contains the resulting corollaries for modified policy iterations and value iteration, and Section 2.3.4 contains the statement and proof of our result on the complexity of computing ϵ-optimal policies.

2.1 Optimality Criterion

Let β denote the discount factor. When the initial state is $x \in \mathbb{X}$ and the decision-maker follows the policy $\pi \in \Pi$, the total expected β-*discounted reward* earned is

$$v_\beta^\pi(x) := \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \beta^n r(\xi_t, \upsilon_t).$$

A policy $\pi_* \in \Pi$ is β-*optimal* if $v_\beta^{\pi_*}(x) = \sup_{\pi \in \Pi} v_\beta^\pi(x) =: v_\beta(x)$ for all $x \in \mathbb{X}$.

2.2 Existence of Optimal Policies

For $x \in \mathbb{X}$, consider the sets

$$A_\beta(x) := \left\{ a \in A(x) : v_\beta(x) = r(x,a) + \beta \int_{\mathbb{X}} v_\beta(y) p(dy|x,a) \right\}. \tag{2.1}$$

The statement of Theorem 2, which is due to Feinberg et al. [41] and provides general sufficient conditions for the existence of β-optimal policies that are deterministic stationary, makes use of a few additional definitions.

Definition 1 ($\mathbb{K}$-sup-compactness [40])**.** The one-step reward function r is $\mathbb{K}$-*sup-compact* if for every nonempty compact $K \subseteq \mathbb{X}$, the sets

$$\{(x,a) \in (K \times \mathbb{A}) \cap \mathrm{Gr}(A) : r(x,a) \geqslant \lambda\}, \qquad \lambda \in \mathbb{R},$$

are compact.

Proposition 1. *The one-step reward function $r : Gr(A) \to \mathbb{R}$ is $\mathbb{K}$-sup-compact if and only if*

(i) *r is upper semicontinuous, and*

(ii) *for any sequence $\{x_k\}$ in $\mathbb{X}$ that converges to some $x \in \mathbb{X}$, any sequence $\{a_k\}$ in $\mathbb{A}$ where $a_k \in A(x_k)$ for all k and $\{r(x_k, a_k)\}$ is bounded below has an accumulation point belonging to $A(x)$.*

Proof. Recall that $\mathbb{X}$ is a Borel subset of a Polish space, and hence is metrizable. Since any metrizable space is compactly generated (see e.g. [83, Lemma 46.3]), the proposition follows from [40, Corollary 2.2]. □

Definition 2 (Weak continuity)**.** The transition probabilities p are *weakly continuous* if for every bounded continuous function $f : \mathbb{X} \to \mathbb{R}$ the mapping

$$(x,a) \mapsto \int_{\mathbb{X}} f(y) p(dy|x,a)$$

is continuous on $\mathrm{Gr}(A)$.

Theorem 2 ([41, Theorem 2]). *Suppose* $r : Gr(A) \to \mathbb{R}$ *is bounded above and* $\mathbb{K}$*-sup-compact, and* p *is weakly continuous. Then*

(i) the value function v_β *is upper semicontinuous and satisfies*

$$v_\beta(x) = \max_{a \in A(x)} \left[r(x,a) + \beta \int_{\mathbb{X}} v_\beta(y) p(dy|x,a) \right], \quad x \in \mathbb{X}; \qquad (2.2)$$

(ii) there is a β*-optimal deterministic stationary policy;*

(iii) a deterministic stationary policy $\phi \in \mathbb{F}$ *is* β*-optimal if and only if* $\phi(x) \in A_\beta(x)$ *for all* $x \in \mathbb{X}$*, where* A_β *is defined by (2.1).*

Remark 3. In order for the conclusions of Theorem 2 to hold, the one-step rewards r need not be bounded below. If r is bounded, however, then the $\mathbb{K}$-sup-compactness of r implies that the all action sets $A(x)$, $x \in \mathbb{X}$, are compact; see [40, Theorem 2.1(ii)].

Theorem 3 ([54, Section 8.5]). *Suppose* $r : Gr(A) \to \mathbb{R}$ *is bounded and upper semicontinuous,* $A(x)$ *is compact for all* $x \in \mathbb{X}$*, and* p *is weakly continuous. Then*

(i) the value function v_β *is the unique bounded upper semicontinuous function that satisfies (2.2);*

(ii) statements (ii) and (iii) of Theorem 2 hold.

To state Theorem 4 below, we say that the set-valued mapping $x \mapsto A(x)$ is *compact-valued* if $A(x)$ is compact for all $x \in \mathbb{X}$, and is *continuous* if for every open subset V of $\mathbb{A}$ the sets $\{x \in \mathbb{X} : A(x) \subseteq V\}$ and $\{x \in \mathbb{X} : A(x) \cap V \neq \emptyset\}$ are open subsets of $\mathbb{X}$.

Theorem 4 ([52, Theorem 2.8]). *Suppose* $r : Gr(A) \to \mathbb{R}$ *is bounded and continuous, the set-valued mapping* $x \mapsto A(x)$ *is compact-valued and continuous, and* p *is weakly continuous. Then*

(i) the value function v_β *is the unique bounded continuous function that satisfies (2.2);*

(ii) statements (ii) and (iii) of Theorem 2 hold.

2.3 Complexity Estimates

In this section, we assume that the state and action sets are *finite*, with the exception of Section 2.3.4 where the state and action sets are possibly uncountable Euclidean spaces. When the state set $\mathbb{X}$ and the action set $\mathbb{A}$ are finite, let $m := \sum_{x\in\mathbb{X}} |A(x)|$ denote the total number of state-action pairs and let $n := |\mathbb{X}|$ denote the total number of states.

To describe the algorithms that we will consider, it is convenient to define some operators on the set of all functions $f : \mathbb{X} \to \mathbb{R}$, i.e. the set of all elements of $\mathbb{R}^{|\mathbb{X}|}$. For $f : \mathbb{X} \to \mathbb{R}$, let

$$T^a_\beta f(x) := r(x,a) + \beta \sum_{y\in\mathbb{X}} p(y|x,a)f(y), \qquad a \in A(x),\ x \in \mathbb{X}$$

and define the *optimality operator* by

$$T_\beta f(x) := \max_{a\in A(x)} T^a_\beta f(x), \qquad x \in \mathbb{X}.$$

In addition, for $f : \mathbb{X} \to \mathbb{R}$ and $\phi \in \mathbb{F}$ we define

$$T^\phi_\beta f(x) := T^{\phi(x)}_\beta f(x) \qquad x \in \mathbb{X}.$$

Asymptotic Notation. A common way to state complexity estimates for algorithms is via asymptotic notation. Given two real-valued functions f and g on the natural numbers, the statement $f(n) \in O(g(n))$ means that there is a constant $C \in \mathbb{R}$ satisfying $f(n) \leqslant Cg(n)$ for all sufficiently large natural numbers n; when $f(n) \in O(g(n))$, we say that $f(n)$ is $O(g(n))$. Here, f usually denotes the number of elementary operations needed for the algorithm to terminate for an input of size n.

2.3.1 Policy Iteration (PI)

In this section, we state the best known complexity estimates for policy iteration, which are due to Scherrer [103]. These estimates will be used to provide complexity estimates for total-reward MDPs in Chapter 3 and for average-reward MDPs in Chapter 4, via reductions to discounted MDPs.

Scherrer [103] obtained improved complexity estimates for the two versions of policy iteration that were considered by Ye [121]. One version, which was stated in the first monograph on MDPs, is due to Howard [62]. To state it, define the set-valued mapping $\mathcal{G}$ from the set of all functions on the state set $\mathbb{X}$ to the set of all deterministic stationary policies $\mathbb{F}$ by

$$\mathcal{G}(f) := \{\phi \in \mathbb{F} : T_\beta^\phi f = T_\beta f\} \qquad \text{for } f : \mathbb{X} \to \mathbb{R}. \tag{2.3}$$

Given $f : \mathbb{X} \to \mathbb{R}$, a deterministic stationary policy $\phi \in \mathcal{G}(f)$ is called *greedy* with respect to f.

Howard's PI

Input: $V_0 : \mathbb{X} \to \mathbb{R}$.

Output: Optimal policy $\phi_* \in \mathbb{F}$.

```
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{G}(V_{j-1})$
3: if $v_\beta^{\phi_j} = T_\beta v_\beta^{\phi_j}$ then
4:     return $\phi_* = \phi_j$
5: else
6:     Set $V_{j+1} = v_\beta^{\phi_j}$ and $j = j + 1$.
7:     go to line 2.
```

The other version of policy iteration that Ye [121] considered corresponds to applying the simplex method with Dantzig's pivoting rule to a certain linear program. To state it, for $f : \mathbb{X} \to \mathbb{R}$ let

$$x_f := \arg\max_{x \in \mathbb{X}} [T_\beta f(x) - f(x)],$$

where ties are broken arbitrarily, and define the set-valued mapping $\mathcal{D}$ from the set of all functions on the state set $\mathbb{X}$ and the set of all deterministic stationary policies $\mathbb{F}$, to the set $\mathbb{F}$, by

$$\mathcal{D}(f, \psi) := \{\phi \in \mathbb{F} : \phi(x_f) \in \arg\max_{a \in A(x_f)} [T_\beta^a f(x_f) - f(x_f)], \phi(x) = \psi(x) \forall x \neq x_f\}$$

for $f : \mathbb{X} \to \mathbb{R}$ and $\psi \in \mathbb{F}$. Observe that x_f is the state x that has the action a for which the quantity

$$r(x, a) + \beta \sum_{y \in \mathbb{X}} p(y|x, a) f(y) - f(x)$$

is the most positive, and any $\phi \in \mathcal{D}(f, \psi)$ is obtained by switching the action $\psi(x_f)$ to some action $a \in A(x_f)$ maximizing $T^a_\beta f(x_f) - f(x_f)$, and letting ϕ be identical to ψ for all of the remaining states.

Simplex-PI

```
Input: $V_0 : \mathbb{X} \to \mathbb{R}$.
Output: Optimal policy $\phi_* \in \mathbb{F}$.
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{D}(V_{j-1})$
3: if $v^{\phi_j}_\beta = T_\beta v^{\phi_j}_\beta$ then
4:     return $\phi_* = \phi_j$
5: else
6:     Set $V_{j+1} = v^{\phi_j}_\beta$ and $j = j + 1$.
7:     go to line 2.
```

Theorem 5. *[[103, Theorems 3,4]]*

(i) The number of iterations required by Howard's PI is

$$O\left(\frac{m}{1-\beta} \log \frac{1}{1-\beta}\right) \tag{2.4}$$

(ii) The number of iterations required by Simplex-PI is

$$O\left(\frac{nm}{1-\beta} \log \frac{1}{1-\beta}\right) \tag{2.5}$$

Proposition 6. *Each iteration of Howard's PI and Simplex-PI requires at most* $O(mn + n^3)$ *arithmetic operations.*

Proof. Both Howard's PI and Simplex-PI require the computation of $v^{\phi_j}_\beta$ for every iteration j, which may require $O(n^3)$ arithmetic operations if Gaussian elimination is used. Both the computation of a greedy policy $\phi_j \in \mathcal{G}(V_{j-1})$ and a policy $\phi_j \in \mathcal{D}(V_{j-1})$ can be accomplished by computing, for every state-action pair $(x, a) \in \mathrm{Gr}(A)$, the quantity

$$r(x, a) + \beta \sum_{y \in \mathbb{X}} p(y|x, a) V_{j-1}(y);$$

hence a total of $O(mn)$ arithmetic operations are needed in both cases. □

Corollary 7. *For discounted MDPs with a fixed discount factor, both Howard's PI and Simplex-PI are polynomial under the BSS model of computation.*

Proof. This follows from Theorem 5 and Proposition 6. □

2.3.2 Optimistic PI

While policy iteration has attractive complexity estimates, the fact that computing $v_\beta^{\phi_j}$ for every iteration j involves the solution of the linear system of equations

$$(I - \beta P^{\phi_j})f = r^{\phi_j}, \qquad f \in \mathbb{R}^n,$$

where $P^{\phi_j}(x, y) := p(y|x, \phi_j(x))$ and $r^{\phi_j}(x) := r(x, \phi_j(x))$ for $x, y \in \mathbb{X}$, can make the algorithm infeasible in practice when the number of states is large. Namely, a total of $O(n^3 + n^2 m)$ arithmetic operations may be needed if Gaussian elimination is used. On the other hand, every iteration of the classic value iteration algorithm can be executed using $O(n^2 m)$ arithmetic operations.

These two properties of policy and value iteration have led to the development of so-called *optimistic* policy iteration algorithms. These algorithms are intended to combine the attractive iteration complexity of policy iteration with the attractive per-iteration complexity of value iteration, by replacing the computation of $v_\beta^{\phi_j}$ on every iteration j with an approximation of it. A classic optimistic policy iteration algorithm is *modified* policy iteration, which is due to Puterman & Shin [92], involves iteratively applying the operator $T_\beta^{\phi_j}$ a certain number n_j of times to the approximation of $v_\beta^{\phi_{j-1}}$ obtained in the previous iteration. Another optimistic policy iteration algorithm, which is of interest to the reinforcement learning community (see Bertsekas & Tsitsiklis [12, Section 2.3.1]) is called λ-policy iteration or *temporal difference-based* policy iteration. In addition, value iteration can also be considered as a certain kind of optimistic policy iteration algorithm; namely, it corresponds to the version of modified policy iteration where $n_j = 1$ for all j.

To define optimistic policy iteration, recall the definition of a greedy policy with respect to an $f : \mathbb{X} \to \mathbb{R}$ given by (2.3), and let $\bar{\mathbb{N}} := \{1, 2, \dots\} \cup \{\infty\}$.

Optimistic PI

Input: $V_0 : \mathbb{X} \to \mathbb{R}$ and a $\bar{\mathbb{N}}$-valued stochastic sequence $\{N_j\}_{j=1}^{\infty}$ with associated probability measure $\mathbf{Q}$ and expectation operator $\mathbf{E}$.
Output: Optimal policy $\phi_* \in \mathbb{F}$.

```
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{G}(V_{j-1})$
3: if $v_\beta^{\phi_j} = T_\beta v_\beta^{\phi_j}$ then
4:     return $\phi_* = \phi_j$
5: else
6:     Set $V_{j+1} = \mathbf{E}[(T_\beta^{\phi_j})^{N_j} V_j]$ and $j = j + 1$.
7:     go to line 2.
```

Theorem 8. *Consider any discount factor $\beta \in (0, 1)$. If $V_0 \equiv 0$ and $Q(N_j < \infty) > 0$ for $j = 1, 2, \ldots$, then for any positive integer N there exists a deterministic MDP with four state-action pairs for which optimistic PI requires at least N iterations to return an optimal policy.*

Proof. To prove the theorem, we consider a family deterministic MDPs parameterized by $R \in \mathbb{R}$. For each of these MDPs, the state set is $\mathbb{X} = \{1, 2, 3\}$ and the action sets are $A(1) = \{\lambda, \rho\}$ and $A(2) = A(3) = \{\sigma\}$. For $R \in (0, \infty)$, the one-step reward for action λ in state 1 is $r(1, \lambda) = R \in \mathbb{R}$, and the remaining one-step rewards are $r(1, \rho) = r(2, \sigma) = 0$ and $r(3, \sigma) = 1$. Finally the transition probabilities are defined by $p(2|1, \lambda) = p(3|1, \rho) = p(2|2, \sigma) = p(3|3, \sigma) = 1$. See Figure 2.1 below.

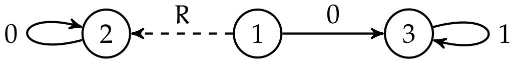

Figure 2.1: A deterministic MDP; each arrow corresponds to an action.

Note that $\mathbf{E}[\beta^{N_j}] > 0$ for every $j \in \mathbb{N}$; since

$$\mathbf{Q}\{N_j < \infty\} = \sum_{n=0}^{\infty} \mathbf{Q}\{N_j = n\} > 0$$

implies that there is an $n_0 \in \mathbb{N}$ such that $P\{N_j = n_0\} > 0$, we have

$$\mathbf{E}[\beta^{N_j}] = \sum_{n=1}^{\infty} \beta^n \mathbf{Q}\{N_j = n\} \geqslant \beta^{n_0} \mathbf{Q}\{N_j = n_0\} > 0.$$

Given N, let R satisfy

$$\frac{\beta}{1-\beta} > R > \frac{\beta(1-\prod_{i=1}^{N-1}\mathbf{E}[\beta^{N_i}])}{1-\beta}.$$

Then the unique optimal policy is characterized by taking action ρ (i.e. "going right") in state 1. Moreover, this policy is obtained on iteration j only if $r(1,\lambda) + \beta V_{j-1}(2) \leqslant r(1,\rho) + \beta V_{j-1}(3)$, i.e. only if

$$R \leqslant \beta V_{j-1}(3) = \beta \cdot \left(\frac{1-\prod_{i=1}^{j-1}\mathbf{E}[\beta^{N_i}]}{1-\beta} \right).$$

But for $j = 1, 2, \ldots N$,

$$R > \frac{\beta(1-\prod_{i=1}^{N-1}\mathbf{E}[\beta^{N_i}])}{1-\beta} \geqslant \frac{\beta(1-\prod_{i=1}^{j-1}\mathbf{E}[\beta^{N_i}])}{1-\beta}.$$

Hence for $j = 1, 2, \ldots, N$, the policy $\phi^j \in \mathcal{G}(V_{j-1})$ is not optimal. □

Corollary 9. *Under the BSS model of computation, there is no upper bound on the number of iterations needed by optimistic policy iteration to return an optimal policy.*

Proof. If such a bound exists, then it would follow that the number of iterations needed by optimistic PI to compute the optimal policy for the MDP defined in the proof of Theorem 8 is some constant N for any value of R, contradicting Theorem 8. □

2.3.3 λ-PI, Modified PI, & Value Iteration

The results in Section 2.3.2 also apply to the variants of optimistic policy iteration mentioned in Section 2.3.2. We first state the λ-policy iteration algorithm, which is due to Bertsekas & Tsitsiklis [12, Section 2.3.1].

λ-PI

```
Input: $V_0 : \mathbb{X} \to \mathbb{R}$ and $\lambda \in (0, 1)$.
Output: Optimal policy $\phi_* \in \mathbb{F}$.
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{G}(V_{j-1})$.
3: if $v_\beta^{\phi_j} = T_\beta v_\beta^{\phi_j}$ then
4:     return $\phi_* = \phi_j$.
5: else
6:     Set $V_{j+1} = (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n (T_\beta^{\phi_j})^n V_j$ and $j = j + 1$.
7:     go to line 2.
```

Next, we recall the modified policy iteration algorithm, which was proposed by Puterman & Shin [92].

Modified PI

```
Input: $V_0 : \mathbb{X} \to \mathbb{R}$ and a sequence $\{n_j\}_{j=1}^{\infty}$ of nonnegative integers.
Output: Optimal policy $\phi_* \in \mathbb{F}$.
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{G}(V_{j-1})$.
3: if $v_\beta^{\phi_j} = T_\beta v_\beta^{\phi_j}$ then
4:     return $\phi_* = \phi_j$.
5: else
6:     Set $V_{j+1} = (T_\beta^{\phi_j})^{n_j} V_j$ and $j = j + 1$.
7:     go to line 2.
```

To our knowledge, the idea of iteratively applying the optimality operator T_β to an initial function V_0 was first considered by Shapley [105] in the context of stochastic games. Below we state a version of value iteration that is known to return an optimal policy in a finite number of iterations; see e.g. Bertsekas [11, Proposition 2.3.1].

Value Iteration

```
Input: $V_0 : \mathbb{X} \to \mathbb{R}$.
Output: Optimal policy $\phi_* \in \mathbb{F}$.
1: Set $j = 1$.
2: Select a policy $\phi_j \in \mathcal{G}(V_{j-1})$.
3: if $v_\beta^{\phi_j} = T_\beta v_\beta^{\phi_j}$ then
4:     return $\phi_* = \phi_j$.
5: else
6:     Set $V_{j+1} = T_\beta V_j$ and $j = j + 1$.
7:     go to line 2.
```

Corollary 10. *For λ-policy iteration, modified policy iteration, and value iteration, under the BSS model of computation there is no upper bound on the number of iterations needed to return an optimal policy.*

Proof. First, λ-policy iteration corresponds to the special case of optimistic policy iteration where each N_j is a geometrically distributed random variable with parameter λ. Next, modified policy iteration corresponds to the version of optimistic policy iteration where each N_j is a degenerate random variable. Finally, value iteration corresponds to the special case of modified policy iteration where $N_j \equiv 1$ for all j. Hence λ-PI, modified PI, and value iteration are each special cases of optimistic PI where $Q(N_j < \infty) = 1$ for $j = 1, 2, \ldots$; hence the corollary follows from Theorem 8. □

2.3.4 Computing ϵ-Optimal Policies

In this section, we consider the complexity of computiong ϵ-optimal policies for total and average-reward MDPs, with Euclidean state and action spaces, that satisfy certain Lipschitz-type assumptions. Throughout this section, we assume that $A(x) \equiv \mathbb{A}$ for all $x \in \mathbb{X}$.

Preliminaries

We first recall some relevant definitions from metric space topology and the convergence of probability measures.

Product Metric. Recall that if (X, ρ_X) and (Y, ρ_Y) are metric spaces, then the product topology on $X \times Y$ is induced by the metric $\rho_{X\times Y}$ where

$$\rho_{X\times Y}((x_1, y_1), (x_2, y_2)) := \max\{\rho_X(x_1, x_2), \rho_Y(y_1, y_2)\}$$

for $(x_1, y_1), (x_2, y_2) \in X \times Y$. A function $u : X \times Y \to \mathbb{R}$ is *Lipschitz continuous on* $X \times Y$ *with modulus* L_u if there is a constant $L_u < \infty$ that satisfies

$$|u(x_1, y_1) - u(x_2, y_2)| \leqslant L_u \rho_{X\times Y}((x_1, y_1), (x_2, y_2))$$

for all $(x_1, y_1), (x_2, y_2) \in X \times Y$.

Total Variation Distance. Given a set S and a σ-algebra Σ on S, the *total variation distance* between two probability measures ν_1 and ν_2 on the measurable space (S, Σ) is

$$\rho_{TV}(\nu_1, \nu_2) := \sup\left\{\left|\int_S f(x)\nu_1(dx) - \int_S f(x)\nu_2(dx)\right| : f : S \to [-1, 1] \text{ Borel}\right\}.$$

Lemma 11. *Let* ν_1, ν_2 *be probability measures on the measurable space* (S, Σ). *If* $f : S \to \mathbb{R}$ *is bounded and Borel-measurable, then*

$$\left|\int_{\mathbb{X}} f(x)\nu_1(dx) - \int_{\mathbb{X}} f(x)\nu_2(dx)\right| \leqslant \left(\sup_{x\in\mathbb{X}} |f(x)|\right) \cdot \rho_{TV}(\nu_1, \nu_2).$$

Proof. See [106, p. 432, Lemma 1] and [53, 172, 185]. □

Complexity Estimate

The complexity estimate that we derive for discounted MDPs is based on the recent work of Saldi et al. [99], [100], [101]. In particular, in [99] rates of convergence of optimal policies for approximating MDPs with finite action sets are derived, and in [100] analogous results are obtained for MDPs with finite state sets. Combining the results in [99] and [100] leads to estimates of the size of a finite state and action MDP that are sufficient to define to an ϵ-optimal policy for the original MDP. The rate of convergence results in [99] and [100] that we will use rely on the following assumption.

Assumption L. *The MDP is such that*

(a) there exist positive integers $d_{\mathbb{X}}$ and $d_{\mathbb{A}}$ such that the state space $\mathbb{X}$ is a compact subspace of $\mathbb{R}^{d_{\mathbb{X}}}$ and the action space $\mathbb{A}$ is a compact subspace of $\mathbb{R}^{d_{\mathbb{A}}}$;

(b) the one-step reward function $r : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ is bounded and Lipschitz-continuous with modulus L_r;

(c) there is a constant $L_p < \infty$ satisfying

$$\rho_{TV}(p(\cdot|x_1, a_1), p(\cdot|x_2, a_2)) \leqslant L_p \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for all $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$.

Remark 4. Assumption L(c) implies that the transition probabilities p are *continuous in total variation*, i.e. any sequence $\{(x_k, a_k)\}_{k=0}^{\infty}$ in $\mathbb{X} \times \mathbb{A}$ that converges to $(x, a) \in \mathbb{X} \times \mathbb{A}$ satisfies

$$\lim_{n\to\infty} \rho_{TV}(p(\cdot|x_k, a_k), p(\cdot|x, a)) = 0.$$

This implies that p is also weakly continuous; see e.g. [42, Theorem 2.5].

Remark 5. Recall that a subset S of a metric space is *totally bounded* if for every $\epsilon > 0$ there exists a finite collection of open balls of radius ϵ that cover S, and that if S is compact then S is totally bounded. In particular, for any compact $S \subseteq \mathbb{R}^d$ there exists a sequence $\{S_k\}_{k=1}^{\infty}$ of finite $S_k \subseteq S$ satisfying $|S_k| = k$ for all k such that, for some $\alpha \in [0, \infty)$,

$$\max_{s_1\in S} \min_{s_2\in S_k} \|s_1 - s_2\|_2 \leqslant \alpha(1/k)^{1/d}, \qquad k = 1, 2, \ldots$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

When Assumption L(i) holds, we define the finite sets $\mathbb{X}_k$ for $k = 1, 2, \ldots$ by substituting $\mathbb{X}$ for S in Remark 5. The finite sets $\mathbb{A}_k$ for $k = 1, 2, \ldots$ are defined analogously. In addition, let

$$\alpha_{\mathbb{X}} := \inf\left\{\alpha \geqslant 0 : \max_{x_1\in\mathbb{X}} \min_{x_2\in\mathbb{X}_k} \|x_1 - x_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{X}}},\ |\mathbb{X}_k| = k\ \forall k\right\} \tag{2.6}$$

and

$$\alpha_{\mathbb{A}} := \inf\left\{\alpha \geqslant 0 : \max_{a_1\in\mathbb{A}} \min_{a_2\in\mathbb{A}_k} \|a_1 - a_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{A}}},\ |\mathbb{A}_k| = k\ \forall k\right\}. \tag{2.7}$$

Theorem 12. *Suppose Assumptions L and HT hold. Then the number of arithmetic operations needed to compute an ϵ-β-optimal policy is at most a constant times*

$$\left\lceil \left(\frac{2M_{\mathbb{X},\beta}}{\epsilon}\right)^{d_{\mathbb{X}}} \right\rceil^4 \cdot \left\lceil \left(\frac{2M_{\mathbb{A},\beta}}{\epsilon}\right)^{d_{\mathbb{A}}} \right\rceil^2 \cdot \frac{1}{1-\beta} \log \frac{1}{1-\beta}, \tag{2.8}$$

where

$$M_{\mathbb{X},\beta} := \frac{2\alpha_{\mathbb{X}}}{1-\beta} \left[\left((2+\beta)\beta L_p + \frac{\beta^2+4\beta+2}{(1-\beta)^2} \right) \cdot \frac{L_r}{1-\beta L_p} + \frac{2L_r}{1-\beta} \right]$$

and

$$M_{\mathbb{A},\beta} := \frac{\alpha_{\mathbb{A}}}{1-\beta} \left[L_r - \beta L_p \|r\|_\infty + \left(\frac{2\beta \|r\|_\infty L_p}{1-\beta} \right) \right],$$

where $\|r\|_\infty := \sup_{(x,a)\in\mathbb{X}\times\mathbb{A}} |r(x,a)|$.

Proof. Consider the the MDP where the original action set is replaced with the finite action set $\mathbb{A}_{k^*}$ where

$$|\mathbb{A}_{k^*}| = k^* := \left\lceil \left(\frac{2M_{\mathbb{A},\beta}}{\epsilon}\right)^{d_{\mathbb{A}}} \right\rceil.$$

Let v_{β,k^*} denote the value function for this MDP. By Theorem 3, this MDP has a deterministic stationary optimal policy ϕ_*. According to Saldi et al. [99, Theorem 4.1], it follows that

$$v_\beta(x) - v_{\beta,k^*}(x) \leqslant M_{\mathbb{A},\beta}(1/k^*)^{1/d_{\mathbb{A}}} \leqslant \epsilon/2, \quad x \in \mathbb{X}. \tag{2.9}$$

Next, following [100], consider the MDP with finite state set $\mathbb{X}_{n^*}$ where

$$|\mathbb{X}_{n^*}| = n^* := \left\lceil \left(\frac{2M_{\mathbb{X},\beta}}{\epsilon}\right)^{d_{\mathbb{X}}} \right\rceil.$$

Let the finite action set be $\mathbb{A}_{n^*}$, and define the one-step rewards r_{n^*} and transition probabilities p_{n^*} as follows. Take any probability measure ν_{n^*} on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ that assigns positive measure to the sets

$$S_{n^*}(x_{n^*}) := \left\{ x \in \mathbb{X} : \arg\min_{y\in\mathbb{X}_{n^*}} \|x-y\| = x_{n^*} \right\}, \quad x_{n^*} \in \mathbb{X}_{n^*},$$

and let

$$r_{n^*}(x_{n^*}, a_{k^*}) := \int_{S_{n^*}(x_{n^*})} r(x, a_{k^*}) \nu_{n^*}(dx), \quad (x_{n^*}, a_{k^*}) \in \mathbb{X}_{n^*} \times \mathbb{A}_{k^*},$$

and

$$p_{n^*}(y_{n^*}|x_{n^*}, a_{k^*}) := \int_{S_{n^*}(x_{n^*})} p(S_{n^*}(y_{n^*})|x, a_{k^*}) \nu_{n^*}(dx)$$

for $x_{n^*}, y_{n^*} \in \mathbb{X}_{n^*}$ and $a_{k^*} \in \mathbb{A}_{k^*}$.

Let v_{β,n^*,k^*} denote the value function for the MDP defined n the preceding paragraph, let ϕ_{n^*,k^*} denote a deterministic optimal policy for that MDP, and consider the deterministic stationary policy ϕ_ϵ for the original MDP defined by

$$\phi_\epsilon(x) := \phi_{n^*,k^*}(\arg\min_{x_{n^*} \in \mathbb{X}_{n^*}} \|x - x_{n^*}\|_2).$$

According to [100, Theorem 5.2], we have

$$v_{\beta,k^*}(x) - v_\beta^{\phi_\epsilon}(x) \leqslant M_{\mathbb{X},\beta}(1/n^*)^{1/d_{\mathbb{X}}} \leqslant \epsilon/2, \quad x \in \mathbb{X}. \tag{2.10}$$

Hence it follows from (2.9) and (2.10) that

$$v_\beta(x) - v_\beta^{\phi_\epsilon}(x) \leqslant \epsilon/2 + \epsilon/2 = \epsilon, \quad x \in \mathbb{X}.$$

□

Chapter 3

Reduction of Transient MDPs

In this chapter, we provide a formulation of transient MDPs that allows for the consideration of randomized history-dependent policies. To our knowledge, previous work on transient MDPs whose transition kernels are not necessarily substochastic only considers optimality over Markov policies; see [90].

In Section 3.1, we provide an alternative formulation of the usual notion of transience, and describe some application areas in Section 3.1.1. In Section 3.2 we define the total-reward optimality criterion for transient MDPs. Next, in Section 3.3 we define a transformation of the original transient MDP to a discounted one, and show how it can lead to the existence and characterization of deterministic stationary optimal policies in Section 3.4. Finally, in Section 3.5 we provide complexity estimates for constructing the HV transformation (Section 3.5.1), computing a deterministic stationary optimal policy (Section 3.5.2, and computing an ϵ-optimal policy (Section 3.5.3).

3.1 Transience Assumption

Consider a nonnegative real-valued Borel-measurable *discount function* α on $\mathrm{Gr}(A)$. In the sequel, we assume that α satisfies the following assumption, which generalizes the case of constant discounting considered in Chapter 2 where $\alpha \equiv \beta \in [0,1)$.

Assumption T (Transience). *There is a constant* $K \geqslant 1$ *satisfying*

$$\mathbb{E}_x^\phi \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} \alpha(\xi_k, \upsilon_k) \leqslant K < \infty \qquad \textit{for all } \; x \in \mathbb{X}, \; \phi \in \mathbb{F}. \tag{3.1}$$

To state the following proposition, recall that a subset E of a Polish space S is *analytic* if there is a Polish space T and a set $B \in \mathcal{B}(S \times T)$ such that E is the projection $\text{proj}_S B$ of B into S. A function $f : S \to \mathbb{R}$ on a Polish space S is *upper* (resp. *lower*) *semianalytic* if for every $\lambda \in \mathbb{R}$ the set $\{s \in S : f(s) \geqslant \lambda\}$ (resp. $\{s \in S : f(s) \leqslant \lambda\}$ is an analytic subset of S.

Proposition 13. *If Assumption T holds, then there is an upper semianalytic function* $\mu : \mathbb{X} \to [1, \infty)$ *that is bounded above by* K *and satisfies*

$$\mu(x) \geqslant 1 + \alpha(x, a) \int_{\mathbb{X}} \mu(y) p(dy|x, a) \quad \textit{for all } \; (x, a) \in Gr(A). \tag{3.2}$$

Proof. Consider the operator $\mathcal{U}$ defined for bounded upper semianalytic functions $u : \mathbb{X} \to [0, \infty)$ by

$$\mathcal{U}u(x) := \sup_{A(x)} \left[1 + \alpha(x, a) \int_{\mathbb{X}} u(y) p(dy|x, a) \right], \quad x \in \mathbb{X}. \tag{3.3}$$

Let $u_0 \equiv 0$, and for $n = 1, 2, \ldots$ let $u_n = \mathcal{U}u_{n-1}$. Note that for each $n \geqslant 0$, u_n is upper semianalytic (see e.g. [10, Propositions 7.47] and $1 \equiv u_1 \leqslant u_n \leqslant u_{n+1}$. Letting $\mu(x) := \lim_{n\to\infty} u_n(x) \geqslant 1$ for $x \in \mathbb{X}$, it follows from [10, Lemma 7.30] that μ is upper semianalytic. We will show that $\mu \leqslant K$ and $\mu = \mathcal{U}\mu$.

We first show that $u_n \leqslant K$ for all $n \geqslant 0$. Note that $u_0 \equiv 0 \leqslant K$. Next, suppose $u_n \leqslant K$ for some $n \geqslant 0$ and consider an arbitrary $\epsilon > 0$. For $\phi \in \mathbb{F}$, define the operator Q_ϕ for bounded upper semianalytic functions $u : \mathbb{X} \to [0, \infty)$ by

$$Q_\phi u(x) := \alpha(x, \phi(x)) \int_{\mathbb{X}} u(y) p(dy|x, \phi(x)), \qquad x \in \mathbb{X},$$

let $Q_\phi^0 u := u$, and for $n = 1, 2, \ldots$ let $Q_\phi^n u := Q_\phi(Q_\phi^{n-1} u)$. Since $K > 0$, according to the definition of $\mathcal{U}$ there is a $\phi^\epsilon \in \mathbb{F}$ satisfying

$$1 + Q_{\phi^\epsilon} u_n(x) \geqslant \mathcal{U}u_n(x) - \frac{\epsilon}{K} \quad \text{for each } x \in \mathbb{X}.$$

Let $\tilde{u}_0 := u_n$, and for $N = 1, 2, \ldots$ let $\tilde{u}_N := 1 + Q_{\phi^\epsilon} \tilde{u}_{N-1}$. Then, letting $e(x) := 1$ for $x \in \mathbb{X}$,

$$\tilde{u}_N(x) = \sum_{i=0}^{N-1} Q^i_{\phi^\epsilon} e(x) + Q^N_{\phi^\epsilon} u_n(x) \quad \text{for each } N \geqslant 1,\ x \in \mathbb{X}. \tag{3.4}$$

By Assumption T, $\sum_{i=0}^{\infty} Q^i_{\phi^\epsilon} e \leqslant K$. Since u_n is bounded, it follows that $Q^N_{\phi^\epsilon} u_n(x) \to 0$ for each $x \in \mathbb{X}$. Letting $N \to \infty$ on both sides of (3.4) gives

$$\lim_{N\to\infty} \tilde{u}_N(x) = \sum_{i=0}^{\infty} Q^i_{\phi^\epsilon} e(x) \leqslant K \quad \text{for each } x \in \mathbb{X}. \tag{3.5}$$

Next, we claim that

$$\tilde{u}_N(x) \geqslant u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^i_{\phi^\epsilon} e(x) \quad \text{for each } N \geqslant 1,\ x \in \mathbb{X}. \tag{3.6}$$

To prove (3.6), first note that for $x \in \mathbb{X}$

$$\begin{aligned} \tilde{u}_1(x) &= 1 + Q_{\phi^\epsilon} \tilde{u}_0(x) \\ &= 1 + Q_{\phi^\epsilon} u_n(x) \geqslant \mathcal{U} u_n(x) - \frac{\epsilon}{K} = u_{n+1}(x) - \frac{\epsilon}{K} Q^0_{\phi^\epsilon} e(x). \end{aligned}$$

Now suppose (3.6) holds for some $N \geqslant 1$. Then for $x \in \mathbb{X}$

$$\begin{aligned} \tilde{u}_{N+1}(x) = 1 + Q_{\phi^\epsilon} \tilde{u}_N(x) &\geqslant 1 + Q_{\phi^\epsilon} u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^{i+1}_{\phi^\epsilon} e(x) \\ &\geqslant 1 + Q_{\phi^\epsilon} u_n(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^{i+1}_{\phi^\epsilon} e(x) \qquad (3.7) \\ &\geqslant \mathcal{U} u_n(x) - \frac{\epsilon}{K} - \frac{\epsilon}{K} \sum_{i=1}^{N} Q^i_{\phi^\epsilon} e(x) \\ &= u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{(N+1)-1} Q^i_{\phi^\epsilon} e(x), \end{aligned}$$

where (3.7) holds since $u_n \leqslant u_{n+1}$. Hence (3.6) holds by induction. Letting $N \to \infty$ on both sides of (3.6), it follows from (3.5) that

$$K \geqslant u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=1}^{\infty} Q^i_{\phi^\epsilon} e(x) \geqslant u_{n+1}(x) - \epsilon \quad \text{for each } x \in \mathbb{X}, \tag{3.8}$$

where the rightmost inequality holds because of Assumption T. Since $\epsilon > 0$ was arbitrary, this means $u_{n+1} \leqslant K$. By induction, $u_n \leqslant K$ for all $n = 0, 1, \dots$. Therefore, $\mu \leqslant K$.

To complete the proof, note that since $u_n \uparrow \mu$, Lebesgue's monotone convergence theorem implies that for $x \in \mathbb{X}$ and $a \in A(x)$

$$\int_{\mathbb{X}} u_n(y) p(dy|x,a) \uparrow \int_{\mathbb{X}} \mu(y) p(dy|x,a) \quad \text{as } n \to \infty.$$

Since $u_n \uparrow \mu$ implies that $\mathcal{U}u_n = u_{n+1} \uparrow \mu$, for $x \in \mathbb{X}$

$$\begin{aligned}\mu(x) = \lim_{n\to\infty} \mathcal{U}u_n(x) &= 1 + \lim_{n\to\infty} \sup_{A(x)} \int_{\mathbb{X}} u_n(y) p(dy|x,a) \\ &= 1 + \sup_{n\geqslant 0} \sup_{A(x)} \int_{\mathbb{X}} u_n(y) p(dy|x,a) \\ &= 1 + \sup_{A(x)} \lim_{n\to\infty} \int_{\mathbb{X}} u_n(y) p(dy|x,a) = \mathcal{U}\mu(x).\end{aligned}$$

□

The following proposition, which states that the sequence $\{u_n\}_{n\geqslant 0}$ defined in the proof of Proposition 13 converges uniformly, will be used to prove the continuity of μ under certain assumptions. This in turn will be used to ensure that the transition probabilities defined by the HV transformation, which is defined in Section 3.3, are weakly continuous if the original transition probabilities are weakly continuous.

Proposition 14. *Suppose Assumption T holds, and consider the operator $\mathcal{U}$ defined by (3.3). Then the sequence $\{u_n\}_{n\geqslant 0}$ where $u_0 \equiv 0$ and $u_n := \mathcal{U}u_{n-1}$ for $n = 1, 2, \dots$ converges uniformly.*

Proof. According to Proposition 13, Assumption T implies that there exists an upper semianalytic function $\mu : \mathbb{X} \to [1, \infty)$ that is bounded above by the constant $K < \infty$ that satisfies (3.2). We claim that for any bounded measurable functions f and g on $\mathbb{X}$,

$$\sup_{x\in\mathbb{X}} \frac{|\mathcal{U}f(x) - \mathcal{U}g(x)|}{\mu(x)} \leqslant \left(\frac{K-1}{K}\right) \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)}. \tag{3.9}$$

To prove (3.9), note that the positivity of μ implies

$$f(x) \leqslant g(x) + \mu(x) \cdot \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)} \quad \text{for all } \ x \in \mathbb{X}.$$

Hence for $x \in \mathbb{X}$, it follows from the nonnegativity of $\alpha(x, a)$ and (3.2) that

$$\begin{aligned} \mathcal{U}f(x) &\leqslant \mathcal{U}g(x) + \left[\sup_{A(x)} \alpha(x, a) \int_{\mathbb{X}} \mu(y) p(dy|x, a)\right] \cdot \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)} \\ &\leqslant \mathcal{U}g(x) + [\mu(x) - 1] \cdot \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)} \\ &\leqslant \mathcal{U}g(x) + \mu(x) \cdot \frac{K-1}{K} \cdot \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)}, \end{aligned}$$

so

$$\frac{\mathcal{U}f(x) - \mathcal{U}g(x)}{\mu(x)} \leqslant \left(\frac{K-1}{K}\right) \sup_{x\in\mathbb{X}} \frac{|f(x) - g(x)|}{\mu(x)} \quad \text{for all } \ x \in \mathbb{X}.$$

By reversing the roles of f and g, (3.9) follows.

According to (3.9), for $n = 0, 1, \ldots$

$$\sup_{x\in\mathbb{X}} \frac{|u_{n+1}(x) - u_n(x)|}{\mu(x)} \leqslant \left(\frac{K-1}{K}\right)^n \sup_{x\in\mathbb{X}} \frac{|u_1(x) - u_0(x)|}{\mu(x)} \leqslant \left(\frac{K-1}{K}\right)^n. \tag{3.10}$$

Hence, letting $\|f\|_\infty := \sup_{x\in\mathbb{X}} |f(x)|$ denote the supremum norm of the function f, for any nonnegative integers m, n where $m > n$,

$$\begin{aligned}\|u_m - u_n\|_\infty &\leqslant K \cdot \sup_{x \in \mathbb{X}} \frac{|u_m(x) - u_n(x)|}{\mu(x)} \\ &\leqslant K \cdot \sum_{k=0}^{m-n-1} \sup_{x \in \mathbb{X}} \frac{|u_{n+k+1}(x) - u_{n+k}(x)|}{\mu(x)} \\ &\leqslant K \cdot \sum_{k=0}^{m-n-1} \left(\frac{K-1}{K}\right)^{n+k} \\ &\leqslant K \cdot \left(\frac{K-1}{K}\right)^{n} \sum_{k=0}^{\infty} \left(\frac{K-1}{K}\right)^{k} = K^2 \cdot \left(\frac{K-1}{K}\right)^{n}.\end{aligned}$$

This means $\{u_n\}_{n \geqslant 0}$ is a Cauchy sequence with respect to the supremum norm in the space of bounded functions $B(\mathbb{X})$ on $\mathbb{X}$. Since $B(\mathbb{X})$ is a Banach space, it follows that $\{u_n\}_{n \geqslant 0}$ converges uniformly. □

3.1.1 Applications

The case where the discount function α may be greater than one under some states and actions is relevant to the study of the control of certain population processes, which have variously been referred to as *branching Markov decision chains* (e.g. [98], [97]), *Markov population decision chains* (e.g. [117], [26]) and *controlled multitype branching processes* (e.g. [89], [90]). For such models, the usual approach has been to consider MDPs where the transition probabilities are replaced with so-called *transition rates* $q(\cdot|x, a)$ for $x \in \mathbb{X}$ and $a \in A(x)$, which are nonnegative and may take on values greater than one.

Note that, equivalently to considering transition rates $q(\cdot|x, a)$, one can consider an MDP with transition probabilities $p(\cdot|x, a)$ and discount function $\alpha : \mathbb{X} \to [0, \infty)$. In particular, given an MDP in the latter form, we can let $q(\cdot|x, a) := \alpha(x, a)p(\cdot|x, a)$ for $x \in \mathbb{X}$ and $a \in A(x)$; conversely, given a controlled multitype branching process with transition rates $q(\cdot|x, a)$, we can let $\alpha(x, a) := q(\mathbb{X}|x, a)$ and $p(\cdot|x, a) := q(\cdot|x, a)/q(\mathbb{X}|x, a)$ for $x \in \mathbb{X}$ and $a \in A(x)$. Assumption T can then be interpreted as the condition that, regardless of the initial population and the actions selected, the total number of individuals that are subsequently born is finite.

The decision process associated with a controlled multitype branching process proceeds as follows. At each decision epoch, the decision-maker controls a finite population of individuals, each of which is in some state $x \in \mathbb{X}$. For each individual that is in state $x \in \mathbb{X}$, the decision-maker selects an action $a \in A(x)$ to perform, whereupon a reward $r(x, a)$ is earned. Further, if the action a is selected for an individual in state x, at the next decision epoch that individual will have given birth to a random number of individuals of different types, independently of the other individuals. In particular, for every Borel subset B of $\mathbb{X}$, the expected number of individuals whose states belong to B is $q(B|x, a)$; when $q(B|x, a) \leqslant 1$ for all $x \in \mathbb{X}$ and $a \in A(x)$, one obtains the standard MDP model as a special case. The goal is to control the population in a way that maximizes the expected total reward earned over either a finite or infinite horizon.

Such models are applicable in a diverse array of contexts. For example, Pliska [89] describes their relevance to the control of infinite particle systems, marketing, and population genetics. A number of other references to applications are given in Eaves & Veinott [26] and Etessami & Yannakakis [30]; in particular, the latter provides an indication of the relevance of controlled multitype branching processes to problems in computer science. We remark that branching processes, which were first referred to as such by Kolmogorov and Dmitriev [73], have been used to study populations in the contexts of biology and demography, as well as various physical phenomena such as cosmic-ray cascades; see e.g. [82, Section 1.1].

3.2 Optimality Criterion

When the initial state is $x \in \mathbb{X}$ and the decision-maker follows the policy $\pi \in \Pi$, the total expected *α-discounted reward* earned is

$$\nu_\alpha^\pi(x) := \mathbb{E}_x^\pi \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} \alpha(\xi_k, \upsilon_k) r(\xi_t, \upsilon_t).$$

For $\epsilon \geqslant 0$, a policy $\pi_* \in \Pi$ is ϵ-α-optimal if $\nu_\alpha^{\pi_*}(x) \geqslant \sup_{\pi \in \Pi} \nu_\beta^\pi(x) - \epsilon$ for all $x \in \mathbb{X}$. A 0-α-optimal policy is called *α-optimal*, and we refer to the function on $\mathbb{X}$ defined by $\sup_{\pi \in \Pi} \nu_\beta^\pi(x) =: \nu_\alpha(x)$, for $x \in \mathbb{X}$, as the *α-value function*.

3.3 Hoffman-Veinott (HV) Transformation

We now define a transformation based on an idea that Veinott [115] attributes to Alan Hoffman; we therefore refer to the transformation as the Hoffman-Veinott (HV) transformation.

By Proposition 13, there is an upper semianalytic function $\mu : \mathbb{X} \to [1, \infty)$ that is bounded above by $K < \infty$ and satisfies (3.2). Objects associated with the transformed MDP will be indicated by a tilde. The state space is $\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$, where $\tilde{x} \notin \mathbb{X}$ is a reward-free absorbing state that is isolated from $\mathbb{X}$. Letting $\tilde{a}$ denote the only action available at state $\tilde{x}$, the action space is $\tilde{\mathbb{A}} := \mathbb{A} \cup \{\tilde{a}\}$ and for $x \in \tilde{\mathbb{X}}$ the set of available actions is unchanged if $x \in \mathbb{X}$, namely

$$\tilde{A}(x) := \begin{cases} A(x), & \text{if } x \in \mathbb{X}, \\ \{\tilde{a}\}, & \text{if } x = \tilde{x}. \end{cases}$$

Define the one-step rewards $\tilde{r}$ by

$$\tilde{r}(x, a) := \begin{cases} \mu(x)^{-1} r(x, a), & \text{if } x \in \mathbb{X},\ a \in A(x), \\ 0, & \text{if } (x, a) = (\tilde{x}, \tilde{a}). \end{cases}$$

To complete the definition of the discounted MDP, choose a discount factor

$$\tilde{\beta} \in \left[\frac{K-1}{K}, 1\right),$$

and let

$$\tilde{p}(B|x, a) := \begin{cases} \frac{\alpha(x,a)}{\tilde{\beta}\mu(x)} \int_B \mu(y) p(dy|x, a), & \text{if } B \in \mathcal{B}(\mathbb{X}),\ x \in \mathbb{X},\ a \in A(x), \\ 1 - \frac{\alpha(x,a)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y) p(dy|x, a), & \text{if } B = \{\tilde{x}\},\ x \in \mathbb{X},\ a \in A(x), \\ 1 & \text{if } B = \{\tilde{x}\},\ x = \tilde{x},\ a = \tilde{a}. \end{cases} \tag{3.11}$$

Note that Lebesgue's monotone convergence theorem implies that $\tilde{p}(\cdot|x, a)$ is a probability measure on $(\tilde{\mathbb{X}}, \mathcal{B}(\tilde{\mathbb{X}}))$ for each $x \in \tilde{\mathbb{X}}$ and $a \in \tilde{A}(x)$. Also, $\tilde{p}(B|\cdot)$ is a lower semianalytic function for each $B \in \mathcal{B}(\mathbb{X})$; see [10, Proposition 7.48].

Since $\tilde{A}(\tilde{x})$ is a singleton, the sets of policies for these two models coincide. Given $x \in \tilde{\mathbb{X}}$ and $\pi \in \Pi$, let $\tilde{\mathbb{E}}_x^\pi$ denote the expectation operator for the $\tilde{\beta}$-discounted MDP with state space $\tilde{\mathbb{X}}$, action space $\tilde{\mathbb{A}}$, sets of available actions $\tilde{\mathbb{A}}$, one-step rewards $\tilde{r}$, and transition probabilities $\tilde{p}$. Let $\tilde{v}^{\pi}_{\tilde{\beta}}(x)$ denote the $\tilde{\beta}$-discounted reward incurred under the policy π when the initial state of this MDP is $x \in \tilde{\mathbb{X}}$, and let $\tilde{v}_{\tilde{\beta}}(x) := \sup_{\pi \in \Pi} \tilde{v}^{\pi}_{\tilde{\beta}}(x)$ for $x \in \tilde{\mathbb{X}}$.

3.4 Existence of Optimal Policies

We now consider conditions under which the HV transformation leads to the existence of (deterministic stationary) optimal policies for the transient MDP under the optimality criterion defined in Section 3.2.

Given $\pi \in \Pi$, the following proposition relates the α-discounted rewards incurred in the original MDP with those incurred in the MDP with a constant discount factor defined by the HV transformation. In particular, for every $x \in \mathbb{X}$ the total reward earned in the original and transformed model when the initial state is x differ by the constant $\mu(x)$.

Proposition 15. *Suppose Assumption T holds. Then* $v^{\pi}(x) = \mu(x)\tilde{v}^{\pi}_{\tilde{\beta}}(x)$ *for each* $\pi \in \Pi$ *and* $x \in \mathbb{X}$.

Proof. For $x \in \mathbb{X}$,

$$\tilde{\mathbb{E}}^{\pi}_{x}|\tilde{r}(x_0, a_0)| = \int_{\tilde{\mathbb{A}}} |\tilde{r}(x, a_0)|\pi_0(da_0|x) = \int_{\mathbb{A}} \frac{|r(x,a_0)|}{\mu(x)}\pi_0(a_0|x) = \frac{\mathbb{E}^{\pi}_{x}|r(x_0,a_0)|}{\mu(x)}.$$

In addition, for $x \in \mathbb{X}$ and $t = 1, 2, \dots$, since $\tilde{r}(\tilde{x}, \tilde{a}) = 0$

$$\begin{aligned}
\tilde{\mathbb{E}}^{\pi}_{x}|\tilde{\beta}^{t}\tilde{r}(x_t, a_t)| &= \int_{\tilde{\mathbb{A}}}\int_{\tilde{\mathbb{X}}} \cdots \int_{\tilde{\mathbb{A}}}\int_{\tilde{\mathbb{X}}}\int_{\tilde{\mathbb{A}}} |\tilde{\beta}^{t}\tilde{r}(x_t, a_t)|\pi_n(da_t|xa_0 \cdots x_t)\tilde{p}(dx_t|x_{t-1}, a_{t-1}) \cdots \\
&\qquad \cdots \pi_1(da_1|xa_0x_1)\tilde{p}(dx_1|x, a_0)\pi_0(da_0|x) \\
&= \tilde{\beta}^{t} \int_{\mathbb{A}}\int_{\mathbb{X}} \cdots \int_{\mathbb{A}}\int_{\mathbb{X}}\int_{\mathbb{A}} \frac{|r(x_t, a_t)|}{\mu(x_t)}\pi_n(da_t|x_t)\frac{\alpha(x_{t-1}, a_{t-1})}{\tilde{\beta}\mu(x_{t-1})}\mu(x_t)p(dx_t|x_{t-1}, a_{t-1}) \cdots \\
&\qquad \cdots \pi_1(da_1|x_1)\frac{\alpha(x, a_0)}{\tilde{\beta}\mu(x)}\mu(x_1)p(dx_1|x, a_0)\pi_0(da_0|x) \\
&= \frac{1}{\mu(x)} \int_{\mathbb{A}}\int_{\mathbb{X}} \cdots \int_{\mathbb{A}}\int_{\mathbb{X}}\int_{\mathbb{A}} |r(x_t, a_t)|\alpha(x_{t-1}, a_{t-1}) \cdots \\
&\qquad \cdots \alpha(x, a_0)\pi_n(da_t|xa_0 \cdots x_t)p(dx_t|x_{t-1}, a_{t-1}) \cdots \\
&\qquad \cdots \pi_1(da_1|xa_0x_1)p(dx_1|x, a_0)\pi_0(da_0|x) \\
&= \frac{1}{\mu(x)}\mathbb{E}^{\pi}_{x}\left|\prod_{k=0}^{t-1} \alpha(x_k, a_k)r(x_t, a_t)\right|.
\end{aligned}$$

Since r is bounded, the boundedness of μ by Proposition 13 implies that $\tilde{r}$ is also bounded. Hence

$$\sum_{t=0}^{\infty} \frac{1}{\mu(x)} \mathbb{E}_x^{\pi} \left| \prod_{k=0}^{t-1} \alpha(x_k, a_k) r(x_t, a_t) \right| = \sum_{t=0}^{\infty} \tilde{\mathbb{E}}_x^{\pi} |\tilde{\beta}^t \tilde{r}(x_t, a_t)| < \infty$$

which (see e.g. [63, Theorem 9.2]) implies that $v^{\pi}(x)/\mu(x) = \tilde{v}_{\tilde{\beta}}^{\pi}(x)$. □

To state the main assumption in this section, recall that the set-valued mapping $x \mapsto A(x)$, $x \in \mathbb{X}$, is *continuous* if for every open subset V of $\mathbb{A}$ the sets $\{x \in \mathbb{X} : A(x) \subseteq V\}$ and $\{x \in \mathbb{X} : A(x) \cap V \neq \emptyset\}$ are open.

Assumption W.

(i) the bounded one-step reward function $r : Gr(A) \to \mathbb{R}$ is continuous;

(ii) $A(x)$ is compact for each $x \in \mathbb{X}$, and the multifunction $x \mapsto A(x)$ is continuous;

(iii) the transition probabilities p are weakly continuous;

(iv) the discount function $\alpha : Gr(A) \to \mathbb{R}$ is continuous.

Lemma 16. *Suppose Assumption T and statements (ii)-(iv) of Assumption W hold. Then there exists a continuous function $\mu : \mathbb{X} \to [1, \infty)$ that is bounded above by K and satisfies (3.2)*

Proof. Recall the operator $\mathcal{U}$ defined by (3.2) in the proof of Proposition 13. Letting $u_0 \equiv 0$, and $u_n := \mathcal{U}u_{n-1}$ for $n = 1, 2, \ldots$, it was shown that $\{u_n\}_{n \geqslant 0}$ increases to an upper semianalytic function μ that is bounded above by K and satisfies (3.2). Note that the continuity of $\alpha(x, a)$ in $(x, a) \in Gr(A)$ and the weak continuity of $p(\cdot|x, a)$ in $(x, a) \in Gr(A)$ imply that $(x, a) \mapsto \alpha(x, a) \int_{\mathbb{X}} f(y) p(dy|x, a)$ is continuous in $(x, a) \in Gr(A)$ for any bounded continuous $f : \mathbb{X} \to \mathbb{R}$. Hence the Berge Maximum Theorem (see e.g. [3, Theorem 17.31]) implies that for $n = 1, 2, \ldots$ the bounded function u_n is continuous. Since $\{u_n\}$ in fact converges uniformly to μ according to Proposition 14, it follows that μ is also continuous. □

Lemma 17. *Suppose Assumptions T and W hold. Then the discounted MDP defined by the HV transformation satisfies the hypotheses of Theorem 4.*

Proof. Lemma 16 and statement (i) of Assumption W imply that $\tilde{r}(x, a)$ is bounded and continuous in $(x, a) \in \{(x, a) : x \in \tilde{\mathbb{X}}, a \in \tilde{A}(x)\}$. In addition, statement (ii) of Assumption W implies that $\tilde{A}(x)$ is compact for each $x \in \tilde{\mathbb{X}}$, and the isolatedness of $\tilde{x}$ implies that $x \mapsto \tilde{A}(x)$ is continuous; see e.g. [3, Theorems 17.20, 17.21]. Next, note that the continuity of μ implies that $\tilde{p}(B|\cdot)$ is a measurable function on $\{(x, a) : x \in \tilde{\mathbb{X}}, a \in \tilde{A}(x)\}$ for each $B \in \mathcal{B}(\tilde{\mathbb{X}})$; see e.g. [10, Proposition 7.29]. In addition, for any bounded continuous function $f : \tilde{\mathbb{X}} \to \mathbb{R}$, since $\tilde{x}$ is isolated from $\mathbb{X}$ Lemma 16 and statements (iii)-(iv) of Assumption W imply that

$$\int_{\tilde{\mathbb{X}}} f(y)\tilde{p}(dy|x, a) = \frac{\alpha(x, a)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} f(y)\mu(y)p(dy|x, a) + \left[1 - \frac{\alpha(x, a)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y)p(dy|x, a)\right] f(\tilde{x})$$

is continuous in $(x, a) \in \{(x, a) : x \in \tilde{\mathbb{X}}, a \in \tilde{A}(x)\}$. □

To state Theorem 18 below, for $x \in \mathbb{X}$ consider the sets of actions

$$A^*_\alpha(x) := \left\{a \in A(x) \;\middle|\; v_\alpha(x) = c(x, a) + \alpha(x, a) \textstyle\int_{\mathbb{X}} v_\alpha(y)p(y|x, a)\right\}$$

which, as will be shown, characterizes the set of optimal actions for that state.

Theorem 18. *Suppose the original MDP with discount function $\alpha(x, a)$ satisfies Assumptions T and W. Then:*

(i) the value function $v_\alpha = \mu\tilde{v}_{\tilde{\beta}}$ is the unique bounded continuous function that satisfies the optimality equation

$$v_\alpha(x) = \max_{A(x)} \left[r(x, a) + \alpha(x, a) \int_{\mathbb{X}} v_\alpha(y)p(dy|x, a)\right], \quad x \in \mathbb{X}; \quad (3.12)$$

(ii) there is a stationary α-discounted cost optimal policy;

(iii) a policy $\phi \in \mathbb{F}$ is α-discounted cost optimal if and only if $\phi(x) \in A^*_\alpha(x)$ for all $x \in \mathbb{X}$, and

$$A^*_\alpha(x) = \left\{a \in A(x) \;\middle|\; \tilde{v}_{\tilde{\beta}}(x) = \tilde{r}(x, a) + \tilde{\beta} \int_{\tilde{\mathbb{X}}} \tilde{v}_{\tilde{\beta}}(y)\tilde{p}(y|x, a)\right\} \quad (3.13)$$

for $x \in \mathbb{X}$; in other words, the sets of optimal actions for the original MDP and for the transformed MDP with a constant discount factor coincide.

Proof. By Lemma 17, the transformed discounted MDP satisfies the hypotheses of Theorem 4. Hence the conclusions of Theorem 4 hold for the transformed MDP.

Proposition 15 implies that $v_\alpha = \mu \tilde{v}_{\tilde{\beta}}$. According to Lemma 16 and Theorem 2, $\mu \geqslant 0$ is continuous and $\tilde{v}_{\tilde{\beta}}$ is continuous, which means v_α is also continuous. Further, since $\tilde{v}_{\tilde{\beta}}(\tilde{x}) = 0$, (7.10) follows from Theorem 4. To show that v_α is the unique bounded continuous function satisfying (3.12), note that if the bounded continuous function $u : \mathbb{X} \to \mathbb{R}$ satisfies (3.12) then the bounded continuous function u/μ satisfies the optimality equation (2.2) for the $\tilde{\beta}$-discounted MDP defined by the HV transformation. According to Theorem 4, this implies that $u = \mu \tilde{v}_{\tilde{\beta}} = v_\alpha$. Hence (i) holds.

Next, according to Theorem 4 there is a $\phi_* \in \mathbb{F}$ that is $\tilde{\beta}$-optimal for the transformed MDP. By Proposition 15, $v_\alpha^{\phi_*} = \mu \tilde{v}_{\tilde{\beta}}^{\phi_*} = \mu \tilde{v}_{\tilde{\beta}} = v_\alpha$, so ϕ_* is α-discounted cost optimal for the original MDP. Therefore (ii) holds.

It follows from the definitions of $\tilde{\mathbb{X}}$, $\tilde{\mathbb{A}}$, $\tilde{r}$, $\tilde{\beta}$, and $\tilde{p}$ that (3.13) holds. Suppose $\phi \in \mathbb{F}$ is α-discounted cost optimal for the original MDP. Then $v_\alpha^\phi = v_\alpha$, so since $v_\alpha^\phi(x) = c(x, \phi(x)) + \alpha(x, \phi(x)) \int_{\mathbb{X}} v_\alpha^\phi(y) p(dy|x, a)$ for every $x \in \mathbb{X}$, it follows that $\phi(x) \in A_\alpha^*(x)$ for all $x \in \mathbb{X}$. Conversely, if $\phi(x) \in A_\alpha^*(x)$ for all $x \in \mathbb{X}$, then according to Theorem 4 and (3.13) the policy ϕ is $\tilde{\beta}$-optimal for the transformed MDP. By Proposition 15, this means ϕ is α-discounted cost optimal for the original MDP. Hence (iii) holds. □

Corollary 19. *Suppose Assumption T and Assumption W hold. Then any algorithm that computes an optimal policy for the discounted MDP defined by the HV transformation is an algorithm for the original α-discounted cost MDP.*

3.5 Complexity Estimates

In this section, we provide complexity estimates related to applying the HV transformation. In Section 3.5.1, we provide an upper bound on the number of arithmetic operations needed to compute a function μ for the HV transformation. Then, in Section 3.5.2 we consider the computation of optimal policies for finite transient MDPs. Finally, in Section 3.5.3 we consider the computation of ϵ-optimal policies for transient MDPs with Euclidean state and action spaces satisfying certain Lipschitz-type conditions. We remark that, according to Veinott [116], Assumption T can be checked in strongly polynomial time.

3.5.1 Constructing the Transformation

Note that, given a suitable function μ, the MDP defined by the HV transformation can be constructed with a number of arithmetic operations that is polynomial in the number of state-action pairs m of the original MDP. The following theorem provides an estimate of the complexity of computing a function μ that can be used for the HV transformation.

Theorem 20. *Suppose the state set $\mathbb{X}$ and action set $\mathbb{A}$ are finite, and that Assumption T holds. Then the number of arithmetic opterations needed to compute a function μ satisfying the hypotheses of Proposition 13 is*

$$O((n^2m + n^3)mK\log K).$$

Proof. To compute a function satisfying the hypotheses of Proposition 13, it suffices to compute a bounded nonnegative function μ that satisfies

$$\mu(x) = \max_{a \in A(x)} \left[1 + \alpha(x,a) \sum_{y \in \mathbb{X}} p(y|x,a)\mu(y) \right] \tag{3.14}$$

for all $x \in \mathbb{X}$. Let $q(y|x,a) := \alpha(x,a)p(y|x,a)$ for $x, y \in \mathbb{X}$ and $a \in A(x)$, and consider the Markov decision process with state set $\mathbb{X}$, action sets $A(x)$ for $x \in \mathbb{X}$, transition rates $q(y|x,a)$ for $x, y \in \mathbb{X}$ and $a \in A(x)$, and one-step rewards identically equal to one. According to Assumption T, this MDP is transient; see [23, Hypothesis 1]. Hence it follows from [23, Theorem 2] that the number of arithmetic operations needed, to compute a nonnegative function that is bounded above by K and satisfies (3.14), is $O((n^2m + n^3)mK\log K)$. □

3.5.2 Computing Optimal Policies

The results in both [121] and [23] were obtained without reducing the original problem to a discounted one. On the other hand, Corollary 19 makes the complexity results on discounted MDPs obtained by Ye [121], Hansen et al. [51], and Scherrer [103] immediately applicable to the study of algorithms for transient MDPs. In particular, Corollary 19 implies that an optimal policy for the original transient MDP can be computed by solving the linear program (LP)

$$\begin{aligned}
&\text{minimize} && \sum_{x\in\tilde{\mathbb{X}}}\sum_{a\in\tilde{A}(x)} \tilde{r}(x,a)z_{x,a} && \\
&\text{such that} && \sum_{a\in\tilde{A}(x)} z_{x,a} - \tilde{\beta}\sum_{y\in\tilde{\mathbb{X}}}\sum_{a\in\tilde{A}(y)} \tilde{p}(x|y,a)z_{y,a} = 1 && \text{for all } x\in\tilde{\mathbb{X}}, \\
& && z_{x,a}\geqslant 0 && \text{for all } x\in\tilde{\mathbb{X}},\ a\in\tilde{A}(x).
\end{aligned} \tag{3.15}$$

Let $m := \sum_{x\in\mathbb{X}} |A(x)|$ denote the total number of state-action pairs, and let $n = |\mathbb{X}|$ denote the total number of states. If $\tilde{\beta} = (K-1)/K$ and $K > 1$, then Scherrer's [103] results imply that the LP (3.15) can be solved using $O(mK\log K)$ iterations of the block-pivoting simplex method corresponding to Howard's policy iteration algorithm, or in $O(mnK\log K)$ iterations using the simplex method with Dantzig's rule. If $K = 1$, then $\tilde{\beta} = 0$ and the problem can be solved by simply selecting, for each $x \in \mathbb{X}$, an action maximizing $r(x,a)$ over $a \in A(x)$.

Remark 6. If Assumption T holds, it holds with the same upper bound K if the values $q(y|x,a) := \alpha(x,a)p(y|x,a)$ are replaced with $\beta q(y|x,a)$, where $\beta \in (0,1]$. Hence the number of arithmetic operations needed to compute an optimal policy for a discounted MDP with transition rates $q(y|x,a)$ satisfying Assumption T can be bounded by a polynomial in m that does not depend on the discount factor $\beta \in (0,1]$. The bounds provided in the previous paragraph are applicable to all discount factors $\beta \in (0,1]$. If $\beta = 0$, the discounted problem becomes a one-step problem, which is equivalent to a problem with $K = 1$; this case was discussed at the end of the previous paragraph.

3.5.3 Computing ϵ-Optimal Policies

In this section, we assume that the original transient MDP satisfies several Lipschitz-type assumptions.

Assumption LT. *The MDP is such that*

(a) there exist positive integers $d_{\mathbb{X}}$ *and* $d_{\mathbb{A}}$ *such that the state space* $\mathbb{X}$ *is a compact subspace of* $\mathbb{R}^{d_{\mathbb{X}}}$ *and the action space* $\mathbb{A}$ *is a compact subspace of* $\mathbb{R}^{d_{\mathbb{A}}}$*;*

(b) the one-step reward function $r : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ *is bounded and Lipschitz-continuous with modulus* L_r*;*

(c) there is a constant $L_p < \infty$ *satisfying*

$$\rho_{TV}(p(\cdot|x_1, a_1), p(\cdot|x_2, a_2)) \leqslant L_p \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for all $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$*;*

(d) the discount function $\alpha : \mathbb{X} \times \mathbb{A} \to \mathbb{R}$ *is bounded and Lipschitz-continuous with modulus* L_α*.*

Proposition 21. *Suppose Assumptions LT and HT hold. Then there exists a Borel-measurable function* $\mu : \mathbb{X} \to [1, \infty)$ *that is bounded above by* K *and satisfies*

$$\mu(x) = \max_{a\in\mathbb{A}} \left[1 + \alpha(x, a) \int_{\mathbb{X}} \mu(y)p(dy|x, a)\right], \qquad \textit{for all } \ x \in \mathbb{X}. \tag{3.16}$$

Proof. Consider the operator $\mathcal{U}$ defined for nonnegative bounded Borel-measurable functions u on $\mathbb{X}$ by

$$\mathcal{U}u(x) := \sup_{a\in\mathbb{A}} \left[1 + \alpha(x, a) \int_{\mathbb{X}} u(y)p(dy|x, a)\right], \qquad x \in \mathbb{X}.$$

Let $u_0 :\equiv 0$, and for $t = 1, 2, \dots$ let $u_t := \mathcal{U}u_{t-1}$. To prove the proposition, we first verify that for $t = 1, 2, \dots$ the function $u_t \geqslant 1$ is Borel-measurable. Then, we show that the Borel-measurable function

$$\mu := \lim_{t\to\infty} u_t$$

satisfies $1 \leqslant \mu \leqslant K$ and (3.16).

Clearly $u_1 \equiv 1$ is Borel-measurable. Next, suppose u_t is Borel for some $t \geqslant 1$. Since the continuity in total variation of p implies that p is setwise continuous, it follows that the mapping

$$(x, a) \mapsto \alpha(x, a) \int_{\mathbb{X}} u_t(y)p(dy|x, a)$$

is continuous on $\mathbb{X} \times \mathbb{A}$. By [56, Theorem 2], it follows that u_{t+1} is Borel-measurable.

Since $1 \equiv u_1 \leqslant u_t \leqslant u_{t+1}$ for each $t \geqslant 1$, the sequence $\{u_t\}_{n=0}^{\infty}$ increases to a Borel-measurable function

$$\mu := \lim_{n \to \infty} u_t \geqslant 1.$$

Moreover, since $\mathcal{U}u_t = u_{t+1} \uparrow \mu$, Lebesgue's monotone convergence theorem implies that

$$\left.\begin{aligned} \mu(x) = \lim_{t \to \infty} \mathcal{U}u_t(x) = 1 + \lim_{t \to \infty} \sup_{a \in \mathbb{A}} \alpha(x,a) \int_{\mathbb{X}} u_t(y) p(dy|x,a) \\ = 1 + \sup_{a \in \mathbb{A}} \lim_{t \to \infty} \alpha(x,a) \int_{\mathbb{X}} u_t(y) p(dy|x,a) = \mathcal{U}\mu(x) \end{aligned}\right. \quad \text{for } x \in \mathbb{X}; \tag{3.17}$$

this will be used to verify (3.16).

Next, we show that $u_t \leqslant K$ for each $t \geqslant 0$; since $u_t \uparrow \mu$, it follows that $\mu \leqslant K$. First, note that $u_0 \equiv 0 \leqslant K$. Next, suppose $u_t \leqslant K$ for some $t \geqslant 0$. For $\phi \in \mathbb{F}$, define the operator Q_ϕ for bounded Borel-measurable functions $u : \mathbb{X} \to [0, \infty)$ by

$$Q_\phi u(x) := \alpha(x,a) \int_{\mathbb{X}} u(y) p(dy|x, \phi(x)), \qquad x \in \mathbb{X}.$$

Letting $e(x) := 1$ for $x \in \mathbb{X}$, by [56, Theorem 2], there exists a $\phi^t \in \mathbb{F}$ satisfying

$$e + Q_{\phi^t} u_t = \mathcal{U}u_t. \tag{3.18}$$

Consider $\tilde{u}_0 := u_t$ and, for $T = 1, 2, \dots$ let $\tilde{u}_T := e + Q_{\phi^t} \tilde{u}_{T-1}$. Then, letting $Q^0_{\phi^t} u := u$ and $Q^i_{\phi^t} u := Q_{\phi^t}(Q^{i-1}_{\phi^t} u)$ for $i = 1, 2, \dots$ and any bounded Borel-measurable $u : \mathbb{X} \to [0, \infty)$,

$$\tilde{u}_T = \sum_{i=0}^{T-1} Q^i_{\phi^t} e + Q^t_{\phi^t} u_t \qquad \text{for } T = 1, 2, \dots . \tag{3.19}$$

Since Assumption HT holds and $u_t \leqslant K$, it follows from (3.19) that

$$\lim_{T \to \infty} \tilde{u}_T = \sum_{i=0}^{\infty} Q^i_{\phi^t} e \leqslant K. \tag{3.20}$$

We claim that $u_{t+1} = \tilde{u}_1$ which, by (3.20), implies that $u_{n+1} \leqslant K$. This claim holds because, by (3.18),

$$\tilde{u}_1 = e + Q_{\phi^t}\tilde{u}_0 = e + Q_{\phi^t}u_t = \mathcal{U}u_t = u_{t+1}.$$

Hence it follows by induction that $u_t \leqslant K$ for each $n \geqslant 0$, which implies that $\mu \leqslant K$.

Finally, (3.16) holds as a consequence of (3.17) and [56, Theorem 2], because the transition probabilities $p(\cdot|x,a)$ are setwise continuous, and μ is a bounded Borel-measurable function. □

In what follows, we denote the *supremum norm* of a real-valued function f on a Polish space S by $\|f\|_\infty$.

Proposition 22. *If Assumptions LT and T hold, then there exists a Lipschitz-continuous function* $\mu : \mathbb{X} \to [1,\infty)$ *with modulus* $L_\mu := K(\|\alpha\|_\infty L_p + L_\alpha)$ *that is bounded above by* K *and satisfies*

$$\mu(x) \geqslant 1 + \alpha(x,a)\int_{\mathbb{X}} \mu(y)p(dy|x,a) \qquad \textit{for all} \;\; (x,a) \in \mathbb{X}\times\mathbb{A}.$$

Proof. According to Proposition 21, Assumption T implies that there exists a Borel-measurable function $\mu : \mathbb{X} \to [1,\infty)$ that is bounded above by K and satisfies

$$\mu(x) = \max_{a\in\mathbb{A}}\left[1 + \alpha(x,a)\int_{\mathbb{X}} \mu(y)p(dy|x,a)\right], \quad (x,a) \in \mathbb{X}\times\mathbb{A}; \tag{3.21}$$

For $(x,a) \in \mathbb{X}\times\mathbb{A}$, let $u(x,a) := \int_{\mathbb{X}} \mu(y)p(dy|x,a)$. Then it follows from Lemma 11 and Assumption LT that for (x_1,a_1) and (x_2,a_2) belonging to $\mathbb{X}\times\mathbb{A}$,

$$\begin{aligned} |u(x_1,a_1) - u(x_2,a_2)| &= \left|\int_{\mathbb{X}} \mu(y)p(dy|x_1,a_1) - \int_{\mathbb{X}} \mu(y)p(dy|x_2,a_2)\right| \\ &\leqslant \left(\sup_{x\in\mathbb{X}} \mu(x)\right)\cdot \rho_{TV}(p(\cdot|x_1,a_1), p(\cdot|x_2,a_2)) \\ &\leqslant K L_p \rho_{\mathbb{X}\times\mathbb{A}}((x_1,a_1),(x_2,a_2)). \end{aligned} \tag{3.22}$$

Next, let $u_\alpha(x,a) := 1 + \alpha(x,a)u(x,a)$ for $(x,a) \in \mathbb{X}\times\mathbb{A}$. Since $\alpha : \mathbb{X}\times\mathbb{A} \to \mathbb{R}$ is Lipschitz-continuous with modulus L_α, it follows from (3.22) that for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}|u_\alpha(x_1, a_1) - u_\alpha(x_2, a_2)| &= |\alpha(x_1, a_1)u(x_1, a_1) - \alpha(x_2, a_2)u(x_2, a_2)| \\ &\leqslant |\alpha(x_1, a_1)u(x_1, a_1) - \alpha(x_1, a_1)u(x_2, a_2)| \\ &\qquad + |\alpha(x_1, a_1)u(x_2, a_2) - \alpha(x_2, a_2)u(x_2, a_2)| \\ &\leqslant \|\alpha\|_\infty |u(x_1, a_1) - u(x_2, a_2)| \\ &\qquad\quad + K|\alpha(x_1, a_1) - \alpha(x_2, a_2)| \\ &\leqslant (\|\alpha\|_\infty K L_p + K L_\alpha)\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2)).\end{aligned} \tag{3.23}$$

Next, consider any $x_1, x_2 \in \mathbb{X}$. By (3.21), there is an action $a_1^* \in \mathbb{A}$ satisfying $u_\alpha(x_1, a_1^*) = \mu(x_1)$. Further, from (3.23) it follows that

$$\begin{aligned}\mu(x_1) - \mu(x_2) &= u_\alpha(x_1, a_1^*) - \mu(x_2) \\ &\leqslant u_\alpha(x_1, a_1^*) - u_\alpha(x_2, a_1^*) \\ &\leqslant (\|\alpha\|_\infty K L_p + K L_\alpha)\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1^*), (x_2, a_1^*)) \\ &= (\|\alpha\|_\infty K L_p + K L_\alpha)\rho_{\mathbb{X}}(x_1, x_2).\end{aligned}$$

Reversing the roles of x_1 and x_2 gives

$$|\mu(x_1) - \mu(x_2)| \leqslant (\|\alpha\|_\infty K L_p + K L_\alpha)\rho_{\mathbb{X}}(x_1, x_2).$$

□

Lemma 23. *If the function $\mu : \mathbb{X} \to [1, \infty)$ is Lipschitz-continuous with modulus L_μ, then the mapping $x \mapsto 1/\mu(x)$ is Lipschitz-continuous with the same modulus.*

Proof. Since $\mu \geqslant 1$, for $x_1, x_2 \in \mathbb{X}$

$$\left|\frac{1}{\mu(x_1)} - \frac{1}{\mu(x_2)}\right| = \frac{|\mu(x_1) - \mu(x_2)|}{\mu(x_1)\mu(x_2)} \leqslant |\mu(x_1) - \mu(x_2)| \leqslant L_\mu \rho_{\mathbb{X}}(x_1, x_2).$$

□

Consider the MDP with state space $\mathbb{X}$, action space $\mathbb{A}$, one-step rewards $\tilde{r}(x, a) := r(x, a)/\mu(x)$ for $(x, a) \in \mathbb{X} \times \mathbb{A}$, and transition dynamics defined by the Borel-measurable substochastic kernel $\tilde{p}$ on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{A}$ where

$$\tilde{p}(B|x, a) := \frac{\alpha(x, a)}{\tilde{\beta}\mu(x)} \int_B \mu(y)p(dy|x, a), \quad (x, a) \in \mathbb{X} \times \mathbb{A},\ B \in \mathcal{B}(\mathbb{X}).$$

Proposition 24. *If Assumptions LT and T hold, then $\tilde{r}$ is bounded and Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus*

$$L_{\tilde{r}} := \max\{1, \|r\|_\infty\}(L_r + L_\mu), \tag{3.24}$$

where $L_\mu := K(\|\alpha\|_\infty L_p + L_\alpha)$.

Proof. The boundedness of $\tilde{r}$ follows from the boundedness of c and μ. For $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned} |\tilde{r}(x_1, a_1) - \tilde{r}(x_2, a_2)| &= \left| \frac{r(x_1, a_1)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_2)} \right| \\ &\leqslant \left| \frac{r(x_1, a_1)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_1)} \right| + \left| \frac{r(x_2, a_2)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_2)} \right| \\ &= \frac{1}{\mu(x_1)} \cdot |r(x_1, a_1) - r(x_2, a_2)| + |r(x_2, a_2)| \cdot \left| \frac{1}{\mu(x_1)} - \frac{1}{\mu(x_2)} \right| \\ &\leqslant [\max\{1, \|r\|_\infty (L_c + L_\mu)\} \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2)), \end{aligned}$$

where the second inequality follows from Lemma 23. □

Proposition 25. *If Assumptions LT and HT hold, then $\tilde{p}$ satisfies*

$$\rho_{TV}(\tilde{p}(\cdot|x_1, a_1), \tilde{p}(\cdot|x_2, a_2)) \leqslant L_{\tilde{p}} \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for all $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$, *where*

$$L_{\tilde{p}} := \max\left\{K, \frac{\|\alpha\|_\infty K}{K-1}\right\} \left(\frac{K(\|\alpha\|_\infty L_\mu + K L_\alpha)}{K-1} + K L_p \right). \tag{3.25}$$

Proof. Consider any Borel-measurable $g : \mathbb{X} \to [-1, 1]$. Note that for $(x, a) \in \mathbb{X} \times \mathbb{A}$,

$$\int_{\mathbb{X}} g(y) \tilde{p}(dy|x, a) = \frac{\alpha(x, a)}{\tilde{\beta} \tilde{\mu}(x)} \int_{\mathbb{X}} g(y) \mu(y) p(dy|x, a).$$

Hence for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned} &\left| \int_{\mathbb{X}} g(y) \tilde{p}(dy|x_1, a_1) - \int_{\mathbb{X}} g(y) \tilde{p}(dy|x_2, a_2) \right| \\ &\qquad = \left| \frac{\alpha(x_1, a_1)}{\tilde{\beta} \tilde{\mu}(x_1)} \int_{\mathbb{X}} g(y) \mu(y) p(dy|x_1, a_1) - \frac{\alpha(x_2, a_2)}{\tilde{\beta} \tilde{\mu}(x_2)} \int_{\mathbb{X}} g(y) \mu(y) p(dy|x_2, a_2) \right|. \end{aligned}$$

Consider the functions u_1, u_2 on $\mathbb{X} \times \mathbb{A}$ defined for $(x, a) \in \mathbb{X} \times \mathbb{A}$ by

$$u_1(x, a) := \alpha(x, a)/(\tilde{\beta}\mu(x))$$

and

$$u_2(x, a) := \int_{\mathbb{X}} g(y)\mu(y)p(dy|x, a).$$

Since $\tilde{\beta} = (K-1)/K$, the function u_1 is bounded by $\|\alpha\|_\infty K/(K-1)$. Further, by Lemma 23 and the Lipschitz-continuity of α, the function u_1 is Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus

$$L_{u_1} := K(\|\alpha\|_\infty L_\mu + KL_\alpha)/(K-1).$$

Further, u_2 is bounded by K and is Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus $L_{u_2} := KL_p$; the latter holds because for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}|u_2(x_1, a_1) - u_2(x_2, a_2)| &= \left|\int_{\mathbb{X}} g(y)\mu(y)[p(dy|x_1, a_1) - p(dy|x_2, a_2)]\right| \\ &\leqslant K\rho_{TV}(p(\cdot|x_1, a_1), p(\cdot|x_2, a_2)) \\ &\leqslant KL_p\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2)).\end{aligned}$$

It follows that for any $g : \mathbb{X} \to [-1, 1]$ and $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}&\left|\int_{\mathbb{X}} g(y)\tilde{p}(dy|x_1, a_1) - \int_{\mathbb{X}} g(y)\tilde{p}(dy|x_2, a_2)\right| \\ &\qquad = |u_1(x_1, a_1)u_2(x_1, a_1) - u_1(x_2, a_2)u_2(x_2, a_2)| \\ &\qquad \leqslant K|u_1(x_1, a_1) - u_1(x_2, a_2)| \\ &\qquad\qquad + \frac{\|\alpha\|_\infty K}{K-1}|u_2(x_1, a_1) - u_2(x_2, a_2)| \\ &\qquad \leqslant \max\left\{K, \frac{\|\alpha\|_\infty K}{K-1}\right\}(L_{u_1} + L_{u_2})\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2)).\end{aligned}$$

Hence

$$\rho_{TV}(\tilde{p}(\cdot|x_1, a_1), \tilde{p}(\cdot|x_2, a_2)) \leqslant L_{\tilde{p}}\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$. □

When Assumption LT(i) holds, define the finite sets $\mathbb{X}_k$ for $k = 1, 2, \ldots$ by substituting $\mathbb{X}$ for S in Remark 5. The finite sets $\mathbb{A}_k$ for $k = 1, 2, \ldots$ are defined analogously. In addition, let

$$\alpha_{\mathbb{X}} := \inf\left\{\alpha \geqslant 0 : \max_{x_1 \in \mathbb{X}} \min_{x_2 \in \mathbb{X}_k} \|x_1 - x_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{X}}},\ |\mathbb{X}_k| = k\ \forall k\right\} \quad (3.26)$$

and

$$\alpha_{\mathbb{A}} := \inf\left\{\alpha \geqslant 0 : \max_{a_1 \in \mathbb{A}} \min_{a_2 \in \mathbb{A}_k} \|a_1 - a_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{A}}},\ |\mathbb{A}_k| = k\ \forall k\right\}. \tag{3.27}$$

Theorem 26. *Suppose Assumptions LT and HT hold. Then the number of arithmetic operations needed to compute an ϵ-optimal policy is at most*

$$\left\lceil \left(\frac{2M_{\mathbb{X}}}{\epsilon}\right)^{d_{\mathbb{X}}} \right\rceil^4 \cdot \left\lceil \left(\frac{2M_{\mathbb{A}}}{\epsilon}\right)^{d_{\mathbb{A}}} \right\rceil^2 \cdot K \log K, \tag{3.28}$$

where

$$M_{\mathbb{X}} := 2\alpha_{\mathbb{X}} K L_{\tilde{r}} \left[2K + \frac{7K^2 - 6K + 1 + L_{\tilde{p}}(3 - \frac{4}{K} + \frac{1}{K^2})}{1 - L_{\tilde{p}} + \frac{L_{\tilde{p}}}{K}}\right]$$

and

$$M_{\mathbb{A}} := \alpha_{\mathbb{A}} K \left[L_{\tilde{r}} - \|r\|_\infty L_{\tilde{p}} \left(2K - 3 + \frac{1}{K}\right)\right];$$

here $\alpha_{\mathbb{X}}$ and $\alpha_{\mathbb{A}}$ are respectively defined by (3.26) and (3.27), $L_{\tilde{r}}$ and $L_{\tilde{p}}$ are respectively defined by (3.24) and (3.25).

Proof. According to Propositions 24 and 25, the MDP defined by the HV transformation satisfies the hypotheses of Assumption L. Hence the theorem follows by applying Theorem 12 to the $\tilde{\beta}$-discounted MDP defined by the HV transformation. □

Chapter 4

Reduction of Average-Reward MDPs

In this chapter, we consider the reduction of average-reward MDPs to discounted ones. In Section 4.1, we describe the hitting time assumption that will be used in the reduction, and in Section 4.1.1 we describe some application areas where the assumption is relevant. In Section 4.2 we define the average-reward optimality criterion, and in Section 4.3 we define a transformation of the original MDP to a discounted one. This leads to complexity estimates for computing optimal policies for finite average-reward MDPs (Section 4.5) and for computing ϵ-optimal policies for average-reward MDPs with Euclidean state and action spaces that satisfy certain Lipschitz-type conditions.

4.1 Hitting Time Assumption

For $x \in \mathbb{X}$, let τ_x denote the *hitting time* to state x after time 0, which is also referred to as the *first return time* to state x. In particular, $\tau_x(\omega) := \inf\{t \geqslant 1 : \xi_t(\omega) = x\}$ for $\omega = x_0 a_0 \cdots \in \mathbb{H}_\infty$.

Assumption HT. *There is a special state $\ell \in \mathbb{X}$ and a constant $L \geqslant 1$ satisfying*

$$\mathbb{E}_x^\phi \tau_\ell \leqslant L < \infty \qquad \text{for all } \; x \in \mathbb{X}, \; \phi \in \mathbb{F}. \tag{4.1}$$

In words, Assumption HT means that under every deterministic stationary policy, the expected time until the process transitions to state ℓ from any initial state is bounded above by L.

Proposition 27. *Suppose Assumption HT holds. Then there exists an upper semianalytic function $\bar{\mu} : \mathbb{X} \to [1, \infty)$ that is bounded above by $\bar{K} := L + 1$ and satisfies*

$$\bar{\mu}(x) \geqslant 1 + \int_{\mathbb{X} \setminus \{\ell\}} \bar{\mu}(y) p(dy|x, a) \quad \textit{for all} \ \ (x, a) \in Gr(A). \tag{4.2}$$

Proof. For $(x, a) \in \mathrm{Gr}(A)$, let

$$q(B|x, a) := \frac{p(B \setminus \{\ell\}|x, a)}{p(\mathbb{X} \setminus \{\ell\}|x, a)}, \qquad B \in \mathcal{B}(\mathbb{X}), \tag{4.3}$$

and $\alpha(x, a) := p(\mathbb{X} \setminus \{\ell\}|x, a)$. Note that $0 \leqslant q(B|x, a)$ and $q(\mathbb{X}|x, a) = 1$ for $(x, a) \in \mathrm{Gr}(A)$. Moreover, Lebesgue's monotone convergence theorem implies that for each $(x, a) \in \mathrm{Gr}(A)$ the set function $B \mapsto q(B|x, a)$ is countably additive. Also, since p is a stochastic kernel on $\mathbb{X}$ given $\mathrm{Gr}(A)$, for each $B \in \mathcal{B}(\mathbb{X})$ the mapping $(x, a) \mapsto q(B|x, a)$ is Borel-measurable. Hence q is a stochastic kernel on $\mathbb{X}$ given $\mathrm{Gr}(A)$. In addition, the mapping $(x, a) \mapsto p(\mathbb{X} \setminus \{\ell\}|x, a) = \alpha(x, a)$ is Borel-measurable by [10, Proposition 7.29].

Replace the transition probabilities p with q, and for $x \in \mathbb{X}$ and $\pi \in \Pi$ let Q_x^π and E_x^π respectively denote the corresponding strategic measure and expectation operator for the resulting MDP. By Assumption HT and the definition of q, for $x \in \mathbb{X}$ and $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$

$$\infty > \bar{K} := L + 1 \geqslant \sum_{n=0}^{\infty} Q_x^{\phi}\{\tau_\ell > n\} = \sum_{n=0}^{\infty} E_x^{\phi} \prod_{k=1}^{n-1} \alpha(x_k, a_k).$$

According to Proposition 13, it follows that there is an upper semianalytic function $\bar{\mu} : \mathbb{X} \to [1, \infty)$ that is bounded above by $\bar{K}$ and satisfies

$$\bar{\mu}(x) \geqslant 1 + \alpha(x, a) \int_{\mathbb{X}} \bar{\mu}(y) q(dy|x, a) = 1 + \int_{\mathbb{X} \setminus \{\ell\}} \bar{\mu}(y) p(dy|x, a)$$

for all $(x, a) \in \mathrm{Gr}(A)$. □

4.1.1 Applications

An important type of problem where MDPs satisfying Assumption HT are relevant is the problem of finding good maintenance and/or replacement policies for systems, such as manufacturing equipment, where the time of failure may be unknown. For such systems, it is often reasonable to assume that, regardless of its initial condition and what the decision-maker does, the time until the system fails is bounded above by a constant. This holds, for example, if it is assumed that at every decision epoch the probability that the system fails is at least a constant $\gamma > 0$, regardless of what the decision-maker does. Such a problem is studied in Dynkin & Yushkevich [25, pp. 188-193], where the time until the system transitions to the state of having just been replaced is bounded above by $L := 1/\gamma$. The models of a stochastically deteriorating systems that are considered in Klein [72] and Derman [24] also satisfy this assumption. Such models are a special case of discrete-state MDPs with a so-called *minorant*; see [25, Section 10] As was shown by Ross [95], the problem of solving such average-reward MDPs can be reduced to solving a related discounted MDP; see also [25, Section 10].

Alternatively, it may be more appropriate to assume that regardless of what the initial state is and what the decision-maker does, the probability that the system fails in N steps, for some positive integer N, is bounded below by a constant $\gamma > 0$. In this case, for any initial state, the time until the system fails is bounded above by $L := N/\gamma$, regardless of what actions are selected. Kim & Thomas [69] consider such an equipment failure model, which is noted to be relevant to the maintenance of standby equipment such as emergency power supplies for hospitals, emergency response vehicles, and military defense equipment. MDPs that model such systems are special cases of MDPs satisfying what Hordijk [60, Chapter 11] refers to as the *simultaneous Doeblin condition*.

4.2 Optimality Criterion

When the initial state is $x \in \mathbb{X}$ and the decision-maker follows the policy $\pi \in \Pi$, the expected long-run *average reward* earned is

$$w^{\pi}(x) := \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}_x^{\pi} \sum_{t=0}^{T-1} r(\xi_t, \upsilon_t).$$

For $\epsilon \geqslant 0$, a policy $\pi_* \in \Pi$ is ϵ*-optimal* if $w^{\pi_*}(x) \geqslant \sup_{\pi \in \Pi} w^{\pi}(x)$ for all $x \in \mathbb{X}$. A 0-optimal policy is called *optimal*, and we refer to the function on $\mathbb{X}$ defined by $\sup_{\pi \in \Pi} w^{\pi}(x) =: w(x)$, for $x \in \mathbb{X}$, as the *value function*.

4.3 Akian-Gaubert (AG) Transformation

Objects associated with the discounted MDP will be indicated by a horizontal bar. The state space is $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, where $\bar{x} \notin \mathbb{X}$ is a reward-free absorbing state that is isolated from $\mathbb{X}$. Letting $\bar{a}$ denote the only action available at state $\bar{x}$, the action space is $\bar{\mathbb{A}} := \mathbb{A} \cup \{\bar{a}\}$ and for $x \in \bar{\mathbb{X}}$ the set of available actions is unchanged if $x \in \mathbb{X}$, namely

$$\bar{A}(x) := \begin{cases} A(x), & \text{if } x \in \mathbb{X}, \\ \{\bar{a}\}, & \text{if } x = \bar{x}. \end{cases}$$

Define the one-step rewards $\bar{r}$ by

$$\bar{r}(x,a) := \begin{cases} \bar{\mu}(x)^{-1} rz(x,a), & \text{if } x \in \mathbb{X},\ a \in A(x), \\ 0, & \text{if } (x,a) = (\bar{x},\bar{a}). \end{cases}$$

To complete the definition of the discounted MDP, choose a discount factor

$$\bar{\beta} \in \left[\frac{\bar{K}-1}{\bar{K}}, 1\right),$$

and let

$$\bar{p}(B|x,a) := \begin{cases} \frac{1}{\bar{\beta}\bar{\mu}(x)} \int_B \bar{\mu}(y) p(dy|x,a), & B \in \mathcal{B}(\mathbb{X} \setminus \{\ell\}),\ x \in \mathbb{X},\ a \in A(x), \\ \frac{1}{\bar{\beta}\bar{\mu}(x)} [\bar{\mu}(x) - 1 - \int_{\mathbb{X} \setminus \{\ell\}} \bar{\mu}(y) p(dy|x,a)], & B = \{\ell\},\ x \in \mathbb{X},\ a \in A(x) \\ 1 - \frac{1}{\bar{\beta}\bar{\mu}(x)} [\bar{\mu}(x) - 1], & B = \{\bar{x}\},\ x \in \mathbb{X},\ a \in A(x) \\ 1, & B = \{\bar{x}\},\ (x,a) = (\bar{x},\bar{a}). \end{cases}$$

Lebesgue's monotone convergence theorem implies that $\bar{p}(\cdot|x, a)$ is a probability measure on $(\bar{\mathbb{X}}, \mathcal{B}(\bar{\mathbb{X}}))$ for each $x \in \bar{\mathbb{X}}$ and $a \in \bar{A}(x)$. Also, $\bar{p}(B|\cdot)$ is a lower semianalytic function on $\{(x, a) : x \in \bar{\mathbb{X}}, a \in \bar{A}(x)\}$ for each $B \in \mathcal{B}(\bar{\mathbb{X}})$; see [10, Proposition 7.48].

Since $\bar{A}(\bar{x})$ is a singleton, the sets of policies for these two models coincide. Given $x \in \bar{\mathbb{X}}$ and $\pi \in \Pi$, let $\bar{\mathbb{E}}_x^\pi$ denote the expectation operator for the $\bar{\beta}$-discounted MDP with state space $\bar{\mathbb{X}}$, action space $\bar{\mathbb{A}}$, sets of available actions $\bar{A}$, one-step rewards $\bar{r}$, and transition probabilities $\bar{p}$. Let $\bar{v}_{\bar{\beta}}^\pi(x)$ denote the $\bar{\beta}$-discounted reward incurred when the initial state of the transformed MDP is $x \in \bar{\mathbb{X}}$ and the policy π is used.

Remark 7. Ross [95] [94] considered MDPs satisfying the special case of Assumption HT where there is a constant α such that

$$p(\{\ell\}|x, a) \geqslant \alpha > 0 \quad \text{for all } \; x \in \mathbb{X}, \; a \in A(x),$$

and introduced a transformation of the transition probabilities that can be used to reduce the average-reward MDP to a discounted one. In fact, Ross's [95] [94] transformation can be viewed as a special case of the AG transformation. Namely, taking $\bar{\mu} \equiv \bar{K} = 1/\alpha$, the resulting transition probabilities are the same in both cases and the one-step rewards differ by a factor of α.

4.4 Existence of Optimal Policies

The proofs of Proposition 29 and Theorem 32 below rely on the following lemma.

Lemma 28. *If a bounded measurable function* $f : \bar{\mathbb{X}} \to \mathbb{R}$ *satisfies* $f(\bar{x}) = 0$, *then for any* $x \in \mathbb{X}$ *and* $a \in A(x)$

$$\begin{aligned}\bar{r}(x, a) + \bar{\beta} \int_{\bar{\mathbb{X}}} f(y)\bar{p}(dy|x, a) \\ = \frac{1}{\mu(x)} \left[r(x, a) + \int_{\mathbb{X}} \mu(y)[f(y) - f(\ell)]p(dy|x, a) + [\mu(x) - 1]f(\ell) \right].\end{aligned}$$

Proof. According to the definition of $\bar{r}$, $\bar{\beta}$, for $x \in \mathbb{X}$ and $a \in A(x)$

$$\begin{aligned}\bar{r}(x,a)+\bar{\beta}\int_{\bar{\mathbb{X}}}\bar{p}(dy|x,a)f(y) &= \frac{r(x,a)}{\mu(x)}+\frac{1}{\mu(x)}\int_{\mathbb{X}\setminus\{\ell\}}\mu(y)f(y)p(dy|x,a)\\ &\quad+\frac{1}{\mu(x)}\left[\mu(x)-1-\int_{\mathbb{X}\setminus\{\ell\}}\mu(y)p(dy|x,a)\right]f(\ell)\\ &=\frac{1}{\mu(x)}\left[r(x,a)+\int_{\mathbb{X}}\mu(y)[f(y)-f(\ell)]p(dy|x,a)+[\mu(x)-1]f(\ell)\right].\end{aligned}$$

□

For $\phi\in\mathbb{F}$, the following proposition relates the average rewards incurred in the original MDP with the discounted rewards incurred in the MDP constructed using the AG transformation.

Proposition 29. *Let $\phi\in\mathbb{F}$ be a stationary policy and $h^{\phi}(x):=\bar{\mu}(x)[\bar{v}^{\phi}_{\bar{\beta}}(x)-\bar{v}^{\phi}_{\bar{\beta}}(\ell)]$ for $x\in\mathbb{X}$. Then*

$$\bar{v}^{\phi}_{\bar{\beta}}(\ell)+h^{\phi}(x)=r(x,\phi(x))+\int_{\mathbb{X}}h^{\phi}(y)p(dy|x,\phi(x)),\quad x\in\mathbb{X}. \tag{4.4}$$

Further, if the one-step rewards c are bounded, then $w^{\phi}\equiv\bar{v}^{\phi}_{\bar{\beta}}(\ell)$.

Proof. Since the state $\bar{x}$ in the discounted MDP defined by the AG transformation is reward-free and absorbing, (4.4) follows from the fact that

$$\bar{v}^{\phi}_{\bar{\beta}}(x)=\bar{r}(x,\phi(x))+\bar{\beta}\int_{\bar{\mathbb{X}}}\bar{v}^{\phi}_{\bar{\beta}}(y)\bar{p}(dy|x,\phi(x)),\quad x\in\mathbb{X},$$

and Lemma 28. Next, suppose c is bounded. Iterating (4.4) gives

$$N\bar{v}^{\phi}_{\bar{\beta}}(\ell)+h^{\phi}(x)=\mathbb{E}^{\phi}_{x}\sum_{n=0}^{N-1}r(x_n,a_n)+\mathbb{E}^{\phi}_{x}h^{\phi}(x_N) \tag{4.5}$$

for $x\in\mathbb{X},\ N=1,2,\dots$. Since r and $\bar{\mu}$ are bounded, the function $h^{\phi}(x)=\bar{\mu}(x)[\bar{v}^{\phi}_{\bar{\beta}}(x)-\bar{v}^{\phi}_{\bar{\beta}}(\ell)]$ is bounded as well. The equality $w^{\phi}\equiv\bar{v}^{\phi}_{\bar{\beta}}(\ell)$ then follows by dividing both sides of (4.5) by N and letting $N\to\infty$. □

Lemma 30. *Suppose Assumption HT holds with an isolated state ℓ, and Assumption W holds. Then there exists a continuous function $\bar{\mu}:\mathbb{X}\to[1,\infty)$ that is bounded above by $\bar{K}:=L+1$ and satisfies (4.2).*

Proof. This follows by applying Lemma 16 to the transient MDP with state space $\mathbb{X}$, action space $\mathbb{A}$, sets of available actions $A(x)$ for $x \in \mathbb{X}$, one-step rewards identically equal to one, transition probabilities q defined by

$$q(B|x,a) := \frac{p(\mathbb{X} \setminus \{\ell\}|x,a)}{p(\mathbb{X} \setminus \{\ell\}|x,a)}, \quad B \in \mathcal{B}(\mathbb{X}),\ (x,a) \in \mathrm{Gr}(A),$$

and discount function α defined by

$$\alpha(x,a) := p(\mathbb{X} \setminus \{\ell\}|x,a), \quad (x,a) \in \mathrm{Gr}(A).$$

□

Remark 8. According to the theory of positive dynamic programming (see e.g. [35]), the function $\bar{\mu}$ constructed in the proof of Lemma 30 satisfies $\bar{\mu}(x) = \sup_{\pi \in \Pi} \mathbb{E}_x^{\pi} \inf\{n \geqslant 1 : x_n = \ell\}$ for all $x \in \mathbb{X}$. This function may not be continuous if the state ℓ satisfying Assumption HT is not isolated. For example, let $\ell \in (0, (\sqrt{5}-1)/2)$ and consider the Markov chain with state space $\mathbb{X} := [0,\ell]$ and transition kernel $P(\cdot|x)$ defined by $P(\{\ell\}|0) := 1$, $P(\{\ell\}|\ell) := 1-\ell$, $P(\{0\}|\ell) := \ell$, and for $x \in (0,\ell)$, $P(\{0\}|x) := x$, $P(\{x\}|x) := x^2$, and $P(\{\ell\}|x) := 1 - x - x^2$. It is straightforward to verify that $P(\cdot|x)$ is weakly continuous in $x \in \mathbb{X}$. Hence statements (ii)-(iii) of Assumption W hold for the corresponding MDP with a single available action for every state (where the rewards, since they do not play a role here, may be defined arbitrarily). In addition, since $\bar{\mu}(0) = 1$, $\bar{\mu}(x) = (1-x)^{-1}$ for $x \in (0,1)$, and $\bar{\mu}(\ell) = 1 + \ell$, Assumption HT holds for state ℓ and the constant $L := (1-\ell)^{-1}$. Therefore the hypotheses of Lemma 30 hold except for the isolatedness of ℓ. However, for any sequence $\{x_n\}_{n \geqslant 0}$ in $[0,\ell)$ that converges to ℓ, $\lim_{n\to\infty} \bar{\mu}(x_n) = (1-\ell)^{-1} > 1 + \ell = \bar{\mu}(\ell)$.

Lemma 31. *Suppose Assumptions HT holds with an isolated state ℓ, and Assumption W holds. Then the discounted MDP defined by the AG transformation satisfies Assumption W.*

Proof. Lemma 30 and Assumption W imply that $\bar{r}(x,a)$ is bounded and continuous in $(x,a) \in \{(x,a) : x \in \bar{\mathbb{X}}, a \in \bar{A}(x)\}$. In addition, statement (ii) of Assumption W and the isolatedness of $\bar{x}$ imply that $x \mapsto \bar{A}(x)$ is compact-valued and continuous; see e.g. [3, Theorems 17.20, 17.21]. Next, note that the measurability of $\bar{\mu}$ implies that $\bar{p}(B|\cdot)$ is a measurable function on $\{(x,a) : x \in \bar{\mathbb{X}}, a \in \bar{A}(x)\}$ for each $B \in \mathcal{B}(\bar{\mathbb{X}})$; see e.g. [10, Proposition 7.29]. In addition, for any bounded continuous function $f : \bar{\mathbb{X}} \to \mathbb{R}$, since the states $\bar{x}, \ell$ are isolated and $\bar{\mu}$ is continuous on $\mathbb{X}$, Lemma 30 and the weak continuity of the transition probabilities p imply that

$$\int_{\bar{\mathbb{X}}} f(y)\bar{p}(dy|x,a) = \frac{1}{\bar{\beta}\bar{\mu}(x)} \int_{\mathbb{X}\setminus\{\ell\}} f(y)\bar{\mu}(y) + \frac{1}{\bar{\beta}\bar{\mu}(x)} \left[\bar{\mu}(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y)p(dy|x,a)\right] f(\ell) \\ + \left[1 - \frac{1}{\bar{\beta}\bar{\mu}(x)}[\bar{\mu}(x) - 1]\right] f(\bar{x})$$

is continuous in $(x,a) \in \{(x,a) : x \in \bar{\mathbb{X}}, a \in \bar{A}(x)\}$. □

For $x \in \mathbb{X}$, and a constant w and function $h : \mathbb{X} \to \mathbb{R}$ satisfying the average-reward optimality equation (4.6) given in the statement of Theorem 32 below, consider the sets of actions

$$A^*_{\mathrm{av}}(x) := \left\{ a \in A(x) \;\middle|\; w + h(x) = r(x,a) + \int_{\mathbb{X}} h(y)p(dy|x,a) \right\}, \quad x \in \mathbb{X}.$$

Theorem 32. *Suppose the original MDP satisfies Assumption HT with an isolated state ℓ, and Assumption W holds. Then:*

(i) the constant $w = \bar{v}_{\bar{\beta}}(\ell)$ and the bounded function $h(x) = \mu(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfy the optimality equation

$$w + h(x) = \max_{A(x)} \left[r(x,a) + \int_{\mathbb{X}} h(y)p(dy|x,a) \right], \quad x \in \mathbb{X}, \tag{4.6}$$

and $\bar{v}_{\bar{\beta}}(\ell)$ is the optimal average reward for each initial state.

(ii) there is a stationary average-reward optimal policy;

(iii) any $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{av}(x)$ for all $x \in \mathbb{X}$ is average-reward optimal, and

$$A^*_{av}(x) = \left\{ a \in A(x) \;\middle|\; \bar{v}_{\bar{\beta}}(x) = \bar{r}(x,a) + \bar{\beta} \int_{\bar{\mathbb{X}}} \bar{v}_{\bar{\beta}}(y)\bar{p}(dy|x,a) \right\} \tag{4.7}$$

for $x \in \mathbb{X}$.

Proof. Lemma 31 implies that the conclusions of Theorem 4 hold for the transformed MDP. In particular, there is a stationary $\bar{\beta}$-optimal policy for the transformed MDP.

By applying Lemma 28 to the optimality equation for the $\bar{\beta}$-discounted MDP defined by the AG transformation, it follows that $w = \bar{v}_{\bar{\beta}}(\ell)$ and

$h(x) = \bar{\mu}(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$, $x \in \mathbb{X}$, satisfy (4.6). Note that, since the MDP satisfies Assumption HT and Assumption W, it also satisfies Assumptions ($\underline{B}$) and (W^*) in [41]. Hence, according to [41, Theorem 3], Proposition 29 implies that the optimal average reward for each state is $\inf_{\phi \in \mathbb{F}} w^{\phi} \equiv \inf_{\phi \in \mathbb{F}} \bar{v}^{\phi}_{\bar{\beta}}(\ell) = \bar{v}_{\bar{\beta}}(\ell)$, so (i) holds.

Let $\phi_* \in \mathbb{F}$ be a $\bar{\beta}$-optimal policy for the transformed MDP. According to Proposition 29 and the previous paragraph, $w^{\phi_*} \equiv \bar{v}^{\phi_*}_{\bar{\beta}}(\ell) = \bar{v}_{\bar{\beta}}(\ell) \equiv \inf_{\phi \in \mathbb{F}} w^{\phi} = \inf_{\pi \in \Pi} w^{\pi}$, which means ϕ_* is average-reward optimal for the original MDP. Hence (ii) holds.

Lemma 28 implies that (4.7) holds. Moreover, since the function h is bounded, it follows that any $\phi \in \mathbb{F}$ satisfying $\phi(x) \in A^*_{av}(x)$ for all $x \in \mathbb{X}$ is average-reward optimal; see e.g., [53, Theorem 5.2.4]. Therefore (iii) holds. □

Corollary 33. *Suppose Assumption HT holds with an isolated state, and Assumption W hold. Then any algorithm that computes a stationary optimal policy for the discounted MDP defined by the AG transformation is an algorithm for the original average-reward MDP.*

4.5 Complexity Estimates

In this section, we provide complexity estimates related to applying the AG transformation. In Section 4.5.1 we provide an upper bound on the number of arithmetic operations needed to compute a function $\bar{\mu}$ for the AG transformation. Then, in Section 4.5.2 we consider the computation of optimal policies for finite average-reward MDPs. Finally, in Section 4.5.3 we consider the computation of ϵ-optimal policies for transient MDPs with Euclidean state and action spaces satisfying certain Lipschitz-type conditions. We remark that, according to Feinberg & Yang [44], checking whether Assumption HT holds can be done in strongly polynomial time.

4.5.1 Constructing the Transformation

Note that, given a suitable function $\bar{\mu}$, the MDP defined by the AG transformation can be constructed using a number of arithmetic operations that is polynomial in the number of state-action pairs m of the original MDP. The following theorem provides an estimate of the complexity of computing a funtion $\bar{\mu}$ that can be used for the AG transformation.

Theorem 34. *Suppose the state set $\mathbb{X}$ and action set A are finite, and that Assumption HT holds. Then the number of arithmetic operations needed to compute a function μ satisfying the hypotheses of Proposition 27 is*

$$O((n^2m + n^3)mL \log L).$$

Proof. To compute a function satisfying the hypotheses of Proposition 27, it suffices to compute a bounded nonnegative function $\bar{\mu}$ that satisfies

$$\bar{\mu}(x) = \max_{a \in A(x)} \left[1 + \sum_{y \in \mathbb{X} \setminus \{\ell\}} p(y|x,a)\bar{\mu}(y) \right], \qquad \text{for all } x \in \mathbb{X}. \tag{4.8}$$

Let

$$q(y|x,a) := 1_{\mathbb{X}\setminus\{\ell\}}(y)p(y|x,a)$$

for $x, y \in \mathbb{X}$ and $a \in A(x)$, and consider the Markov decision process with state set $\mathbb{X}$, action sets $A(x)$ for $x \in \mathbb{X}$, transition rates $q(y|x,a)$ for $x, y \in \mathbb{X}$ and $a \in A(x)$, and one-step rewards identically equal to one. According to Assumption HT, this MDP is transient; see [23, Hypothesis 1]. Hence it follows from [23, Theorem 2] that the number of arithmetic operations needed, to compute a nonnegative function that is bounded above by $\bar{K} := L + 1$ and satisfies (4.8), is $O((n^2m + n^3)m\bar{K}\log\bar{K}) = O((n^2m + n^3)mL\log L)$. □

4.5.2 Computing Optimal Policies

For a finite state and action MDP that satisfies Assumption HT, Corollary 33 implies that a stationary average-reward optimal policy can be computed by solving the LP

$$\begin{aligned}
&\text{minimize} && \sum_{x \in \bar{\mathbb{X}}} \sum_{a \in \bar{A}(x)} \bar{r}(x,a) z_{x,a} && \\
&\text{such that} && \sum_{a \in \bar{A}(x)} z_{x,a} - \bar{\beta} \sum_{y \in \bar{\mathbb{X}}} \sum_{a \in \bar{A}(y)} \bar{p}(x|y,a) z_{y,a} = 1 && \text{for all } x \in \bar{\mathbb{X}}, \\
& && z_{x,a} \geqslant 0 && \text{for all } x \in \bar{\mathbb{X}},\ a \in \bar{A}(x).
\end{aligned} \tag{4.9}$$

Recall that $m = \sum_{x \in \mathbb{X}} |A(x)|$ and $n = |\mathbb{X}|$. If $\bar{K} = L + 1 > 1$, Scherrer's [103] bounds imply that the LP (4.9) can be solved using $O(m\bar{K}\log(\bar{K}))$ iterations of the block-pivoting simplex method corresponding to Howard's policy iteration algorithm, or in $O(mn\bar{K}\log(\bar{K}))$ iterations using the simplex method with Dantzig's rule. Observe that $\bar{K} = L + 1 = 1$ means that $L = 0$, i.e. the only state ℓ is absorbing under each stationary policy, and a stationary policy ϕ is average-reward optimal if and only if $r(\ell, \phi(\ell)) = \min\{r(\ell, a) : a \in A(\ell)\}$.

4.5.3 Computing ϵ-Optimal Policies

The following is a generalization, to infinite state and action spaces, of a transformation introduced in [1]. Let

$$\bar{\beta} := (K-1)/K,$$

and consider the MDP with state space $\mathbb{X}$, action space $\mathbb{A}$, one-step rewards $\bar{r}(x, a) := r(x, a)/\mu(x)$ for $(x, a) \in \mathbb{X} \times \mathbb{A}$, and transition dynamics defined by the Borel-measurable substochastic kernel $\bar{p}$ on $\mathbb{X}$ given $\mathbb{X} \times \mathbb{A}$ where

$$\bar{p}(B|x, a) := \begin{cases} \frac{1}{\bar{\beta}\mu(x)} \int_B \mu(y) p(dy|x, a) & \text{if } B \in \mathcal{B}(\mathbb{X} \setminus \{\ell\}), (x, a) \in \mathbb{X} \times \mathbb{A}, \\ \frac{1}{\bar{\beta}\mu(x)} [\mu(x) - 1 - \int_{\mathbb{X} \setminus \{\ell\}} \mu(y) p(dy|x, a)] & \text{if } B = \{\ell\}, (x, a) \in \mathbb{X} \times \mathbb{A}, \end{cases}$$

Proposition 35. *If Assumptions L and HT hold, then the one-step reward function $\bar{r}$ is bounded and Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus*

$$L_{\bar{r}} := \max\{1, \|r\|_\infty\}(L_r + KL_p). \tag{4.10}$$

Proof. The boundedness of $\bar{r}$ follows from the boundedness of c and μ. For $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned} |\bar{c}(x_1, a_1) - \bar{c}(x_2, a_2)| &= \left| \frac{r(x_1, a_1)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_2)} \right| \\ &\leqslant \left| \frac{r(x_1, a_1)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_1)} \right| + \left| \frac{r(x_2, a_2)}{\mu(x_1)} - \frac{r(x_2, a_2)}{\mu(x_2)} \right| \\ &= \frac{1}{\mu(x_1)} \cdot |r(x_1, a_1) - r(x_2, a_2)| + |r(x_2, a_2)| \cdot \left| \frac{1}{\mu(x_1)} - \frac{1}{\mu(x_2)} \right| \\ &\leqslant [\max\{1, \|r\|_\infty\}(L_c + L_\mu)]\rho_{\mathbb{X} \times \mathbb{A}}((x_1, a_1), (x_2, a_2)), \end{aligned}$$

where the second inequality follows from Lemma 23. □

Proposition 36. *If Assumptions L and HT hold, then $\bar{p}$ satisfies*

$$\rho_{TV}(\bar{p}(\cdot|x_1, a_1), \bar{p}(\cdot|x_2, a_2)) \leqslant L_{\bar{p}} \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for all $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$, *where*

$$L_{\bar{p}} := \max\left\{2(K+1), \frac{K}{K-1}\right\}\left(2K + \frac{K^2}{K-1}\right) L_p. \tag{4.11}$$

Proof. Consider any Borel-measurable $g : \mathbb{X} \to [-1, 1]$. Note that for $(x, a) \in \mathbb{X} \times \mathbb{A}$,

$$\int_{\mathbb{X}} g(y)\bar{p}(dy|x, a) = \frac{1}{\bar{\beta}\bar{\mu}(x)} \int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)p(dy|x, a) + \left[1 - \frac{\mu(x) - 1}{\bar{\beta}\mu(x)}\right] [g(\bar{x}) - g(\ell)].$$

Hence for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}
&\left|\int_{\mathbb{X}} g(y)\bar{p}(dy|x_1, a_1) - \int_{\mathbb{X}} g(y)\bar{p}(dy|x_2, a_2)\right| \\
&\qquad = \Bigg|\frac{1}{\bar{\beta}\bar{\mu}(x_1)} \int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)p(dy|x_1, a_1) + \left[1 - \frac{\mu(x_1) - 1}{\bar{\beta}\mu(x_1)}\right] [g(\bar{x}) - g(\ell)] \\
&\qquad\qquad - \frac{1}{\bar{\beta}\bar{\mu}(x_2)} \int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)p(dy|x_1, a_1) - \left[1 - \frac{\mu(x_2) - 1}{\bar{\beta}\mu(x_2)}\right] [g(\bar{x}) - g(\ell)]\Bigg| \\
&\qquad = \Bigg|\frac{1}{\bar{\beta}\bar{\mu}(x_1)} \left(\int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)p(dy|x_1, a_1) + g(\bar{x}) - g(\ell)\right) \\
&\qquad\qquad - \frac{1}{\bar{\beta}\bar{\mu}(x_2)} \left(\int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)p(dy|x_2, a_2) + g(\bar{x}) - g(\ell)\right)\Bigg|.
\end{aligned}$$

Consider the functions u_1, u_2 on $\mathbb{X} \times \mathbb{A}$ where $u_1(x, a) := 1/[\bar{\beta}\mu(x)]$ and $u_2(x, a) := \int_{\mathbb{X}\setminus\{\ell\}}[g(y) - g(\ell)]\mu(y)p(dy|x, a) + g(\bar{x}) - g(\ell)$. Since $\bar{\beta} = (K-1)/K$, u_1 is bounded by $K/(K-1)$ and, by Lemma 23, is Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus $K^2 L_p/(K-1)$. Further, u_2 is bounded by $2(K+1)$ and is Lipschitz-continuous on $\mathbb{X} \times \mathbb{A}$ with modulus $2KL_p$; the latter holds because for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}
|u_2(x_1, a_1) - u(x_2, a_2)| &= \left|\int_{\mathbb{X}\setminus\{\ell\}} [g(y) - g(\ell)]\mu(y)[p(dy|x_1, a_1) - p(dy|x_2, a_2)]\right| \\
&\leqslant \left|\int_{\mathbb{X}} 2\mu(y)[p(dy|x_1, a_1) - p(dy|x_2, a_2)]\right| \\
&\leqslant 2K\rho_{TV}(p(\cdot|x_1, a_1) - p(\cdot|x_2, a_2)) \\
&\leqslant 2KL_p\rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2)).
\end{aligned}$$

It follows that for any $g : \mathbb{X} \to [-1, 1]$ and $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$,

$$\begin{aligned}
&\left|\int_{\mathbb{X}} g(y)\bar{p}(dy|x_1, a_1) - \int_{\mathbb{X}} g(y)\bar{p}(dy|x_2, a_2)\right| \\
&\qquad = |u_1(x_1, a_1)u_2(x_1, a_1) - u_1(x_2, a_2)u_2(x_2, a_2)| \\
&\qquad \leqslant |u_1(x_1, a_1)u_2(x_1, a_1) - u_1(x_2, a_2)u_2(x_1, a_1)| \\
&\qquad\qquad + |u_1(x_1, a_2)u_2(x_1, a_1) - u_1(x_2, a_2)u_2(x_2, a_2)| \\
&\qquad \leqslant 2(K+1)|u_1(x_1, a_1) - u_1(x_2, a_2)| \\
&\qquad\qquad + [K/(K-1)]|u_2(x_1, a_1) - u_2(x_2, a_2)| \\
&\qquad \leqslant \max\left\{2(K+1), \frac{K}{K-1}\right\}\left(2K + \frac{K^2}{K-1}\right) L_p \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2));
\end{aligned}$$

Hence

$$\rho_{TV}(\bar{p}(\cdot|x_1, a_1), \bar{p}(\cdot|x_2, a_2)) \leqslant L_{\bar{p}} \rho_{\mathbb{X}\times\mathbb{A}}((x_1, a_1), (x_2, a_2))$$

for $(x_1, a_1), (x_2, a_2) \in \mathbb{X} \times \mathbb{A}$. □

When Assumption L(i) holds, we define the finite sets $\mathbb{X}_k$ for $k = 1, 2, \ldots$ by substituting $\mathbb{X}$ for S in Remark 5. The finite sets $\mathbb{A}_k$ for $k = 1, 2, \ldots$ are defined analogously. In addition, let

$$\alpha_{\mathbb{X}} := \inf\left\{\alpha \geqslant 0 : \max_{x_1\in\mathbb{X}} \min_{x_2\in\mathbb{X}_k} \|x_1 - x_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{X}}},\ |\mathbb{X}_k| = k\ \forall k\right\} \tag{4.12}$$

and

$$\alpha_{\mathbb{A}} := \inf\left\{\alpha \geqslant 0 : \max_{a_1\in\mathbb{A}} \min_{a_2\in\mathbb{A}_k} \|a_1 - a_2\|_2 \leqslant \alpha(1/k)^{1/d_{\mathbb{A}}},\ |\mathbb{A}_k| = k\ \forall k\right\}. \tag{4.13}$$

Theorem 37. *Suppose Assumptions L and HT hold. Then the number of arithmetic operations needed to compute an ϵ-optimal policy is at most*

$$\left\lceil\left(\frac{M_{\mathbb{X}}}{\epsilon}\right)^{d_{\mathbb{X}}}\right\rceil^3 \cdot \left\lceil\left(\frac{2M_{\mathbb{A}}}{\epsilon}\right)^{d_{\mathbb{A}}}\right\rceil^2 \cdot K \log K, \tag{4.14}$$

where

$$M_{\mathbb{X}} := 2\alpha_{\mathbb{X}} K L_{\bar{r}}\left[2K + \frac{7K^2 - 6K + 1 + L_{\bar{p}}(3 - \frac{4}{K} + \frac{1}{K^2})}{1 - L_{\bar{p}} + \frac{L_{\bar{p}}}{K}}\right]$$

and

$$M_{\mathbb{A}} := \alpha_{\mathbb{A}} K\left[L_{\bar{r}} - \|c\|_\infty L_{\bar{p}}\left(2K - 3 + \frac{1}{K}\right)\right];$$

here $\alpha_{\mathbb{X}}$ *and* $\alpha_{\mathbb{A}}$ *are respectively defined by (4.12) and (4.13),* $L_{\bar{c}}$ *and* $L_{\bar{p}}$ *are respectively defined by (4.10) and (4.11).*

Proof. According to Propositions 35 and 36, the MDP defined by the AG transformation satisfies the hypotheses of Assumption L. Hence the theorem follows by applying Theorem 12 to the $\bar{\beta}$-discounted MDP defined by the AG transformation. □

Part II

Two-Player Zero-Sum Stochastic Games

Chapter 5

Discounted Stochastic Games

In this chapter, we review some definitions and results for two-player zero-sum discounted stochastic games with Borel state and action spaces. In Section 5.1, the optimality criterion for discounted stochastic games is defined. Then, in Section 5.2 we state conditions under which both players have ϵ-optimal stationary strategies, and in Section 5.2.1 conditions are given under which both players have optimal stationary strategies.

5.1 Optimality Criterion

Consider a discount factor $\beta \in [0,1)$. When the initial state is $x \in \mathbb{X}$ and players 1 and 2 follow the strategies $\pi^1 \in \Pi^1$ and $\pi^2 \in \Pi^2$, respectively, the expected total *discounted payoff* that player 1 receives from player 2 is

$$v_\beta^{\pi^1\pi^2}(x) := \mathbb{E}_x^{\pi^1\pi^2} \sum_{t=0}^{\infty} \beta^n r(\xi_t, \upsilon_t^1, \upsilon_t^2).$$

Define the *lower value* function ${}_*v_\beta : \mathbb{X} \to \mathbb{R}$ by

$${}_*v_\beta(x) := \sup_{\pi^1 \in \Pi^1} \inf_{\pi^2 \in \Pi^2} v_\beta^{\pi^1\pi^2}$$

and the *upper value* function ${}^*v_\beta : \mathbb{X} \to \mathbb{R}$ by

$${}^*v_\beta(x) := \inf_{\pi^2 \in \Pi^2} \sup_{\pi^1 \in \Pi^2} v_\beta^{\pi^1\pi^2}.$$

According to the minimax inequality stated in Proposition 38 below,

$$ {}_*v_\beta(x) \leqslant {}^*v_\beta(x) \qquad \text{for all } \; x \in \mathbb{X}. $$

If ${}_*v_\beta = {}^*v_\beta =: v_\beta$, then v_β is called the β-*discounted value* of the discounted stochastic game.

Proposition 38 (Minimax inequality)**.** *Let* Y *and* Z *be any sets. Then for any* $f : Y \times Z \to \mathbb{R}$,

$$ \sup_{y \in Y} \inf_{z \in Z} f(y,z) \leqslant \inf_{z \in Z} \sup_{y \in Y} f(y,z). $$

Proof. Observe that

$$ f(y_0, z_0) \leqslant \sup_{y \in Y} f(y, z_0) \qquad \text{for all } \; (y_0, z_0) \in Y \times Z. $$

Hence every $y_0 \in Y$ satisfies $\inf_{z \in Z} f(y_0, z) \leqslant \inf_{z \in Z} \sup_{y \in Y} f(y,z)$, so

$$ \sup_{y \in Y} \inf_{z \in Z} f(y,z) \leqslant \inf_{z \in Z} \sup_{y \in Y} f(y,z). $$

□

For $\epsilon \geqslant 0$, a strategy $\pi^1_* \in \Pi^1$ is ϵ-β-*optimal for player 1* if

$$ \inf_{\pi^2 \in \Pi^2} v_\beta^{\pi^1_*, \pi^2}(x) \geqslant {}^*v_\beta(x) - \epsilon \qquad \text{for all } \; x \in \mathbb{X}, $$

and $\pi^2_* \in \Pi^2$ is ϵ-β-*optimal for player 2* if

$$ \sup_{\pi^1 \in \Pi^1} v_\beta^{\pi^1, \pi^2_*}(x) \leqslant {}_*v_\beta(x) + \epsilon \qquad \text{for all } \; x \in \mathbb{X}. $$

For both players, a 0-β-optimal strategy is called β-*optimal*.

5.2 Existence of ϵ-Optimal Strategies

The following assumption is a sufficient condition for the maximizing player to have an ϵ-optimal stationary strategy, and for the minimizing player to have an optimal stationary strategy, under the discounted-payoff criterion.

Assumption P2.

(i) $A^2(x)$ *is compact for all* $x \in \mathbb{X}$*;*

(ii) $r(x, a^1, \cdot)$ *is lower semicontinuous on* $A^2(x)$ *for all* $x \in \mathbb{X}$ *and* $a^1 \in A^1(x)$*;*

(iii) for each $B \in \mathcal{B}(\mathbb{X})$, $x \in \mathbb{X}$*, and* $a^1 \in A^1(x)$*, the function* $p(B|x, a^1, \cdot)$ *is continuous on* $A^2(x)$*;*

(iv) for each $x \in \mathbb{X}$ *and* $a^1 \in A^1(x)$*, the function* $\alpha(x, a^1, \cdot)$ *is continuous on* $A^2(x)$*.*

For probability measures ν^1, ν^2 on $\mathbb{A}^1$ and $\mathbb{A}^2$, respectively, define the operator $\mathcal{L}_\beta(\nu^1, \nu^2)$ for bounded upper semianalytic functions $f : \mathbb{X} \to \mathbb{R}$ by

$$\mathcal{L}_\beta(\nu^1, \nu^2) f(x) := \int_{\mathbb{A}^2} \int_{\mathbb{A}^1} \left(r(x, a^1, a^2) + \beta \int_{\mathbb{X}} f(y) p(dy|x, a^1, a^2) \right) \nu^1(da^1) \nu^2(da^2)$$

for $x \in \mathbb{X}$, and for $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$ let

$$\mathcal{L}_\beta^{\varphi^1 \varphi^2} f(x) := \mathcal{L}_\beta(\varphi^1(x), \varphi^2(x)) f(x), \quad x \in \mathbb{X}.$$

Further, define the optimality operators for bounded upper semianalytic functions $f : \mathbb{X} \to \mathbb{R}$ by

$$\mathcal{L}_\beta^{\varphi^2} f(x) := \sup_{\varphi^1 \in \Phi^1} \mathcal{L}_\beta^{\varphi^1 \varphi^2} f(x), \quad \varphi^2 \in \Phi^2,\ x \in \mathbb{X}$$

and

$$\mathcal{L}_\beta f(x) := \inf_{\varphi^2 \in \Phi^2} \mathcal{L}_\beta^{\varphi^2} f(x), \quad x \in \mathbb{X}.$$

Theorem 39 ([85, Theorem 5.3])**.** *Suppose Assumption P2 holds with* $\alpha \equiv \beta \in [0, 1)$*. Then player 1 has an* ϵ*-*β*-optimal stationary strategy for any* $\epsilon > 0$*, and player 2 has a* β*-optimal stationary strategy. Further, the game has a* β*-value* v_β *which is bounded, upper semianalytic, and uniquely satisfies*

$$v_\beta(x) = \mathcal{L}_\beta v_\beta(x) \quad \text{for all } x \in \mathbb{X}. \tag{5.1}$$

5.2.1 Existence of Optimal Strategies

Under the following assumption, it can be shown that both players have stationary optimal strategies under the discounted-payoff criterion.

Assumption P1.

(i) $A^1(x)$ *is compact for all* $x \in \mathbb{X}$*;*

(ii) $r(x, \cdot, a^2)$ *is upper semicontinuous on* $A^1(x)$ *for all* $x \in \mathbb{X}$ *and* $a^1 \in A^1(x)$*;*

(iii) for each $B \in \mathcal{B}(\mathbb{X})$ *and* $x \in \mathbb{X}$*, the function* $p(B|x, \cdot, a^2)$ *is continuous on* $A^1(x)$*;*

(iv) for each $x \in \mathbb{X}$ *and* $a^2 \in A^2(x)$*, the function* $\alpha(x, \cdot, a^2)$ *is continuous on* $A^1(x)$*.*

Theorem 40 ([84, Theorems 5.1, 5.3, 5.4])**.** *Suppose both Assumptions P1 and P2 hold. Then both players have* β*-optimal stationary strategies that are Borel-measurable. Further, the discounted game has a* β*-value* v_β*, which is Borel-measurable and satisfies (5.1).*

Chapter 6

Reduction of Transient Stochastic Games

In this chapter, we consider the reduction of so-called transient stochastic games, under the total-payoff criterion, to discounted ones. In Section 6.1, the transience assumption is stated, and in Section 6.2 the optimality criterion that we consider for such games is defined. Next, a transformation to a discounted stochastic game is given in Section 6.3, which extends an idea of Alan Hoffman (Veinott [115]) to a two-player setting. This transformation is shown to lead to the existence of ϵ-optimal strategies for both players in Section 6.4, and to the existence of optimal strategies for both players in Section 6.4.1.

6.1 Transience Assumption

Let $\alpha : \mathrm{Gr}(A^1 \times A^2) \to [0, \infty)$ denote a Borel-measurable *discount function* that satisfies the following assumption.

Assumption T (Transience)**.** *There is a constant* $K \geqslant 1$ *satisfying*

$$\mathbb{E}_x^{\phi^1\phi^2} \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} \alpha(\xi_k, \upsilon_k^1, \upsilon_k^2) \leqslant K < \infty \tag{6.1}$$

for all $x \in \mathbb{X}$ *and* $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$.

Observe that Assumption T generalizes the case of constant discounting considered in Chapter 5, where $\alpha \equiv \beta \in [0,1)$.

Proposition 41. *If Assumption T holds, then there is an upper semianalytic function* $\mu : \mathbb{X} \to [1,\infty)$ *that is bounded above by* $K < \infty$ *and satisfies*

$$\mu(x) \geqslant 1 + \alpha(x,a^1,a^2)\int_{\mathbb{X}} \mu(y)p(dy|x,a^1,a^2) \quad \textit{for all} \ (x,a^1,a^2) \in \mathbb{K}. \tag{6.2}$$

Proof. Consider the operator $\mathcal{U}$ defined for bounded upper semianalytic functions $u : \mathbb{X} \to [0,\infty)$ by

$$\mathcal{U}u(x) := \sup_{a^1\in A^1(x)} \sup_{a^2\in A^2(x)} \left[1 + \alpha(x,a^1,a^2)\int_{\mathbb{X}} u(y)p(dy|x,a^1,a^2)\right], \quad x \in \mathbb{X}. \tag{6.3}$$

Let $u_0 :\equiv 0$, and for $n = 1,2,\dots$ let $u_n := \mathcal{U}u_{n-1}$. Note that for each $n \geqslant 0$, u_n is upper semianalytic (see e.g. [10, Proposition 7.47, 7.48]) and $1 \equiv u_1 \leqslant u_n \leqslant u_{n+1}$. Letting $\mu(x) := \lim_{n\to\infty} u_n(x) \geqslant 1$ for $x \in \mathbb{X}$, it follows from [10, Lemma 7.30] that μ is upper semianalytic. We will show that $\mu \leqslant K$ and $\mu = \mathcal{U}\mu$.

We first show that $u_n \leqslant K$ for all $n \geqslant 0$. Note that $u_0 \equiv 0 < K$. Next, suppose $u_n \leqslant K$ for some $n \geqslant 0$ and consider an arbitrary $\epsilon > 0$. For $(\phi^1,\phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$, define the operator $Q_{\phi^1\phi^2}$ for bounded upper semianalytic functions $u : \mathbb{X} \to [0,\infty)$ by

$$Q_{\phi^1\phi^2}u(x) := \alpha(x,\phi^1(x),\phi^2(x))\int_{\mathbb{X}} u(y)p(dy|x,\phi^1(x),\phi^2(x)), \qquad x \in \mathbb{X},$$

let $Q^0_{\phi^1\phi^2}u := u$, and for $n = 1,2,\dots$ let $Q^n_{\phi^1\phi^2}u := Q_{\phi^1\phi^2}(Q^{n-1}_{\phi^1\phi^2}u)$. Since $K > 0$, according to [10, Propositions 7.47, 7.48, 7.50] there exist $\phi^1_\epsilon \in \mathbb{F}^1$ and $\phi^2_\epsilon \in \mathbb{F}^2$ satisfying

$$1 + Q_{\phi^1_\epsilon\phi^2_\epsilon}u_n(x) > \mathcal{U}u_n(x) - \frac{\epsilon}{K} \quad \text{for each } x \in \mathbb{X}.$$

Let $\tilde{u}_0 := u_n$, and for $N = 1,2,\dots$ let $\tilde{u}_N := 1 + Q_{\phi^1_\epsilon\phi^2_\epsilon}\tilde{u}_{N-1}$. Then, letting $e(x) := 1$ for $x \in \mathbb{X}$,

$$\tilde{u}_N(x) = \sum_{i=0}^{N-1} Q^i_{\phi^1_\epsilon\phi^2_\epsilon}e(x) + Q^N_{\phi^1_\epsilon\phi^2_\epsilon}u_n(x) \quad \text{for each } N \geqslant 1,\ x \in \mathbb{X}. \tag{6.4}$$

By Assumption T, $\sum_{i=0}^{\infty} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e \leqslant K$. Since u_n is bounded, it follows that $Q^N_{\phi^1_\epsilon \phi^2_\epsilon} u_n(x) \to 0$ for each $x \in \mathbb{X}$. Letting $N \to \infty$ on both sides of (6.4) gives

$$\lim_{N \to \infty} \tilde{u}_N(x) = \sum_{i=0}^{\infty} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e(x) \leqslant K \quad \text{for each } x \in \mathbb{X}. \tag{6.5}$$

Next, we claim that

$$\tilde{u}_N(x) > u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e(x) \quad \text{for each } N \geqslant 1,\ x \in \mathbb{X}. \tag{6.6}$$

To prove (6.6), first note that for $x \in \mathbb{X}$

$$\begin{aligned} \tilde{u}_1(x) = 1 + Q_{\phi^1_\epsilon \phi^2_\epsilon} \tilde{u}_0(x) &= 1 + Q_{\phi^1_\epsilon \phi^2_\epsilon} u_n(x) \\ &> \mathcal{U} u_n(x) - \frac{\epsilon}{K} = u_{n+1}(x) - \frac{\epsilon}{K} Q^0_{\phi^1_\epsilon \phi^2_\epsilon} e(x). \end{aligned}$$

Now suppose (6.6) holds for some $N \geqslant 1$. Then for $x \in \mathbb{X}$

$$\begin{aligned} \tilde{u}_{N+1}(x) = 1 + Q_{\phi^1_\epsilon \phi^2_\epsilon} \tilde{u}_N(x) &\geqslant 1 + Q_{\phi^1_\epsilon \phi^2_\epsilon} u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^{i+1}_{\phi^1_\epsilon \phi^2_\epsilon} e(x) \\ &\geqslant 1 + Q_{\phi^1_\epsilon \phi^2_\epsilon} u_n(x) - \frac{\epsilon}{K} \sum_{i=0}^{N-1} Q^{i+1}_{\phi^1_\epsilon \phi^2_\epsilon} e(x) \qquad (6.7) \\ &> \mathcal{U} u_n(x) - \frac{\epsilon}{K} - \frac{\epsilon}{K} \sum_{i=1}^{N} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e(x) \qquad (6.8) \\ &= u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=0}^{(N+1)-1} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e(x), \end{aligned}$$

where (6.7) holds since $u_n \leqslant u_{n+1}$. Hence (6.6) holds by induction. Letting $N \to \infty$ on both sides of (6.6), it follows from (6.5) that

$$K \geqslant u_{n+1}(x) - \frac{\epsilon}{K} \sum_{i=1}^{\infty} Q^i_{\phi^1_\epsilon \phi^2_\epsilon} e(x) > u_{n+1}(x) - \epsilon \quad \text{for each } x \in \mathbb{X}, \tag{6.9}$$

where the rightmost inequality holds because of Assumption T. Since $\epsilon > 0$ was arbitrary, this means $u_{n+1} \leqslant K$. By induction, $u_n \leqslant K$ for all $n = 0, 1, \ldots$. Therefore, $\mu \leqslant K$.

To complete the proof, note that since $u_n \uparrow \mu$, Lebesgue's monotone convergence theorem implies that for $x \in \mathbb{X}$, $a^1 \in \mathbb{A}^1$, and $a^2 \in \mathbb{A}^2$,

$$\int_{\mathbb{X}} u_n(y)p(dy|x, a^1, a^2) \uparrow \int_{\mathbb{X}} \mu(y)p(dy|x, a^1, a^2) \quad \text{as } n \to \infty.$$

Since $u_n \uparrow \mu$ implies that $\mathcal{U}u_n = u_{n+1} \uparrow \mu$, for $x \in \mathbb{X}$

$$\begin{aligned}
\mu(x) = \lim_{n\to\infty} \mathcal{U}u_n(x) &= 1 + \lim_{n\to\infty} \sup_{a^1\in A^1(x)} \sup_{a^2\in A^2(x)} \int_{\mathbb{X}} u_n(y)p(dy|x, a^1, a^2) \\
&= 1 + \sup_{n\geqslant 0} \sup_{a^1\in A^1(x)} \sup_{a^2\in A^2(x)} \int_{\mathbb{X}} u_n(y)p(dy|x, a^1, a^2) \\
&= 1 + \sup_{a^1\in A^1(x)} \sup_{a^2\in A^2(x)} \lim_{n\to\infty} \int_{\mathbb{X}} u_n(y)p(dy|x, a^1, a^2) \\
&= \mathcal{U}\mu(x).
\end{aligned}$$

□

6.2 Optimality Criterion

When the initial state is $x \in \mathbb{X}$ and players 1 and 2 follow the strategies $\pi^1 \in \Pi^1$ and $\pi^2 \in \Pi^2$, respectively, the total expected *α-discounted payoff* that player 1 receives from player 2 is

$$v_\alpha^{\pi^1\pi^2}(x) := \mathbb{E}_x^{\pi^1\pi^2} \sum_{t=0}^{\infty} \prod_{k=0}^{t-1} \alpha(\xi_k, v_k^1, v_k^2) r(\xi_t, v_k^1, v_k^2).$$

Define the *lower value* function ${}_*v_\alpha : \mathbb{X} \to \mathbb{R}$ by

$${}_*v_\alpha(x) := \sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} v_\alpha^{\pi^1\pi^2}(x)$$

and the *upper value* function ${}^*v_\alpha : \mathbb{X} \to \mathbb{R}$ by

$${}^*v_\alpha(x) := \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} v_\alpha^{\pi^1\pi^2}(x).$$

According to Proposition 38,

$$ {}_*v_\alpha(x) \leqslant {}^*v_\alpha(x) \qquad \text{for all } \ x \in \mathbb{X}. $$

If ${}_*v_\alpha = {}^*v_\alpha =: v_\alpha$, then v_α is called the *α-discounted value* of the stochastic game.

For $\epsilon \geqslant 0$, a strategy $\pi^1_* \in \Pi^1$ is *ϵ-α-optimal for player 1* if

$$ \inf_{\pi^2 \in \Pi^2} v_\alpha^{\pi^1_*, \pi^2}(x) \geqslant {}^*v_\alpha(x) - \epsilon \qquad \text{for all } \ x \in \mathbb{X}, $$

and $\pi^2_* \in \Pi^2$ is *ϵ-α-optimal for player 2* if

$$ \sup_{\pi^1 \in \Pi^1} v_\alpha^{\pi^1, \pi^2_*}(x) \leqslant {}_*v_\alpha(x) + \epsilon \qquad \text{for all } \ x \in \mathbb{X}. $$

For both players, a 0-α-optimal strategy is called *α-optimal*.

6.3 Hoffman-Veinott (HV) Transformation

We first describe a transformation of the original game to a discounted-payoff game with a constant discount factor less than one, which we call the *HV (Hoffman–Veinott) transformation*. Objects associated with the transformed game will be denoted by a tilde. The state space is $\tilde{\mathbb{X}} := \mathbb{X} \cup \{\tilde{x}\}$, where $\tilde{x} \notin \mathbb{X}$ is a payoff-free absorbing state. Letting $\tilde{a}$ denote the only action available to players 1 and 2 at state $\tilde{x}$, the action space for player $i = 1, 2$ is $\tilde{\mathbb{A}}^i := \mathbb{A}^i \cup \{\tilde{a}\}$. For $x \in \tilde{\mathbb{X}}$ the set of available actions for player $i = 1, 2$ is unchanged if $x \in \mathbb{X}$, namely

$$ \tilde{A}^i(x) := \begin{cases} A^i(x), & \text{if } x \in \mathbb{X}, \\ \{\tilde{a}\}, & \text{if } x = \tilde{x}, \end{cases} \tag{6.10} $$

and let $\tilde{\mathbb{K}} := \{(x, a^1, a^2) : x \in \tilde{\mathbb{X}}, a^i \in \tilde{A}^i(x), i = 1, 2\}$

Define the payoff function for player 1 (i.e. the payment function for player 2) by

$$ \tilde{r}(x, a^1, a^2) := \begin{cases} \mu(x)^{-1} r(x, a^1, a^2), & \text{if } (x, a^1, a^2) \in \mathbb{K}, \\ 0, & \text{if } (x, a^1, a^2) = (\tilde{x}, \tilde{a}, \tilde{a}). \end{cases} $$

To complete the definition of the discounted-payoff game, choose a discount factor

$$\tilde{\beta} \in \left[\frac{K-1}{K}, 1\right),$$

and let

$$\tilde{p}(B|x,a^1,a^2) := \begin{cases} \frac{\alpha(x,a^1,a^2)}{\tilde{\beta}\mu(x)} \int_B \mu(y) p(dy|x,a^1,a^2), & \text{if } B \in \mathcal{B}(\mathbb{X}),\ (x,a^1,a^2) \in \mathbb{K}, \\ 1 - \frac{\alpha(x,a^1,a^2)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y) p(dy|x,a^1,a^2), & \text{if } B = \{\tilde{x}\},\ (x,a^1,a^2) \in \mathbb{K}, \\ 1, & \text{if } B = \{\tilde{x}\},\ (x,a^1,a^2) = (\tilde{x},\tilde{a},\tilde{a}). \end{cases}$$

Lebesgue's monotone convergence theorem implies that $\tilde{p}(\cdot|x,a^1,a^2)$ is a probability measure on $(\tilde{\mathbb{X}}, \mathcal{B}(\tilde{\mathbb{X}}))$ for each $(x,a^1,a^2) \in \tilde{\mathbb{K}}$. In addition, according to Proposition 41 and [10, Proposition 7.46], for each $B \in \mathcal{B}(\tilde{\mathbb{X}})$ the function $\tilde{p}(B|\cdot)$ is universally measurable.

Since $\tilde{a}$ is the only action available to players 1 and 2 in state $\tilde{x}$, the sets of all strategies of each player for the $\tilde{\beta}$-discounted game and the original α-discounted game coincide. Given $x \in \tilde{\mathbb{X}}$ and $(\pi^1,\pi^2) \in \Pi^1 \times \Pi^2$, let $\tilde{\mathbb{E}}_x^{\pi^1\pi^2}$ denote the expectation operator (defined via [10, Proposition 7.45]) associated with the discounted-payoff game, and

$$\tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi^2) := \tilde{\mathbb{E}}_x^{\pi^1\pi^2} \sum_{n=0}^{\infty} \tilde{\beta}^n r(x_n, a_n^1, a_n^2).$$

Given an initial state $x \in \tilde{\mathbb{X}}$, let

$${}_*\tilde{v}_{\tilde{\beta}}(x) := \sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} \tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi^2), \quad {}^*\tilde{v}_{\tilde{\beta}}(x) := \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} \tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi^2).$$

For $\epsilon \geqslant 0$, a strategy $\pi_*^1 \in \Pi^1$ is ϵ-$\tilde{\beta}$-*optimal for player 1* if

$$\inf_{\pi^2\in\Pi^2} \tilde{v}_{\tilde{\beta}}(x,\pi_*^1,\pi^2) \geqslant {}^*\tilde{v}_{\tilde{\beta}}(x) - \epsilon$$

for all $x \in \tilde{\mathbb{X}}$; $\pi_*^2 \in \Pi^2$ is ϵ-$\tilde{\beta}$-*optimal for player 2* if

$$\sup_{\pi^1\in\Pi^1} \tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi_*^2) \leqslant {}_*\tilde{v}_{\tilde{\beta}}(x) + \epsilon$$

for all $x \in \tilde{\mathbb{X}}$. A 0-$\tilde{\beta}$-optimal policy for either player is called $\tilde{\beta}$-*optimal*. If ${}_*\tilde{v}_{\tilde{\beta}} = {}^*\tilde{v}_{\tilde{\beta}} =: \tilde{v}_{\tilde{\beta}}$, then $\tilde{v}_{\tilde{\beta}}$ is the $\tilde{\beta}$-*discounted value* of the game.

Proposition 42. *If Assumption T holds, then $v_\alpha(x,\pi^1,\pi^2) = \mu(x)\tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi^2)$ for each $x \in \mathbb{X}$ and $(\pi^1,\pi^2) \in \Pi^1 \times \Pi^2$.*

Proof. For $x \in \mathbb{X}$,

$$\begin{aligned}\tilde{\mathbb{E}}_x^{\pi^1\pi^2}|\tilde{r}(x_0,a_0^1,a_0^2)| &= \int_{\tilde{\mathbb{A}}^2}\int_{\tilde{\mathbb{A}}^1}|\tilde{r}(x,a_0^1,a_0^2)|\pi_0^1(da_0^1|x)\pi_0^2(da_0^2|x)\\ &= \int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\frac{|r(x,a_0^1,a_0^2)|}{\mu(x)}\pi_0^1(da_0^1|x)\pi_0^2(da_0^2|x) = \frac{\mathbb{E}_x^{\pi^1\pi^2}|r(x_0,a_0^1,a_0^2)|}{\mu(x)}.\end{aligned}$$

For $x \in \mathbb{X}$ and $t = 1,2,\ldots$, let $h_n := xa_0^1a_0^2\ldots x_n$. Since $\tilde{r}(\tilde{x},\tilde{a},\tilde{a}) = 0$,

$$\begin{aligned}\tilde{\mathbb{E}}_x^{\pi^1\pi^2}|\tilde{\beta}^t\tilde{r}(x_t,a_t^1,a_t^2)| &= \int_{\tilde{\mathbb{A}}^2}\int_{\tilde{\mathbb{A}}^1}\cdots\int_{\tilde{\mathbb{X}}}\int_{\tilde{\mathbb{A}}^2}\int_{\tilde{\mathbb{A}}^1}|\tilde{\beta}^t\tilde{r}(x_t,a_t^1,a_t^2)|\pi_t^1(da_t^1|h_t)\pi_t^2(da_t^2|h_t)\tilde{p}(dx_t|x_{t-1},a_{t-1}^1,a_{t-1}^2)\cdots\\ &\qquad\cdots\pi_0^1(da_0^1|x)\pi_0^2(da_0^2|x)\\ &= \tilde{\beta}^t\int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\cdots\int_{\mathbb{X}}\int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\frac{|r(x_t,a_t^1,a_t^2)|}{\mu(x_t)}\pi_t^1(da_t^1|h_t)\pi_t^2(da_t^2|h_t)\frac{\alpha(x_{t-1},a_{t-1}^1,a_{t-1}^2)}{\tilde{\beta}\mu(x_{t-1})}\mu(x_t)p(dx_t|x_{t-1},a_{t-1}^1,a_{t-1}^2)\cdots\\ &\qquad\cdots\pi_0^1(da_0^1|x)\pi_0^2(da_0^2|x)\\ &= \frac{1}{\mu(x)}\int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\cdots\int_{\mathbb{X}}\int_{\mathbb{A}^2}\int_{\mathbb{A}^1}|r(x_t,a_t^1,a_t^2)|\pi_t^1(da_t^1|h_t)\pi_t^2(da_t^2|h_t)\alpha(x_{t-1},a_{t-1}^1,a_{t-1}^2)p(dx_t|x_{t-1},a_{t-1}^1,a_{t-1}^2)\cdots\\ &\qquad\cdots\pi_0^1(da_0^1|x)\pi_0^2(da_0^2|x)\\ &= \frac{1}{\mu(x)}\mathbb{E}_x^{\pi^1\pi^2}\left|\prod_{k=0}^{n-1}\alpha(x_k,a_k^1,a_k^2)r(x_t,a_t^1,a_t^2)\right|.\end{aligned}$$

Since r is bounded, the boundedness of μ by Proposition 13 implies that $\tilde{r}$ is also bounded. Hence

$$\begin{aligned}\sum_{t=0}^{\infty}\frac{1}{\mu(x)}\mathbb{E}_x^{\pi^1\pi^2}&\left|\prod_{k=0}^{t-1}\alpha(x_k,a_k^1,a_k^2)r(x_t,a_t^1,a_t^2)\right|\\ &= \sum_{t=0}^{\infty}\tilde{\mathbb{E}}_x^{\pi^1,\pi^2}|\tilde{\beta}^t\tilde{r}(x_t,a_t^1,a_t^2)| < \infty,\end{aligned}$$

which (see e.g. [63, Theorem 9.2]) implies that

$$v_\alpha(x,\pi^1,\pi^2)/\mu(x) = \tilde{v}_{\tilde{\beta}}(x,\pi^1,\pi^2).$$

□

Corollary 43. *Suppose Assumption T holds. Then for $\epsilon \geqslant 0$, a strategy for player 1 (resp. player 2) is ϵ-α-optimal for the original α-discounted game if and only if that strategy for player 1 (resp. player 2) is ϵ-$\tilde{\beta}$-optimal for the $\tilde{\beta}$-discounted game defined by the HV transformation. Further, the former game has a value if and only if the latter game has one.*

Proof. This follows from Proposition 42 and the fact that $\mu(x) \geqslant 1$ for all $x \in \mathbb{X}$. □

Remark 9. Corollary 43 implies that in order to compute a pair of optimal strategies for the α-discounted game, it suffices to compute a pair of optimal strategies for the $\tilde{\beta}$-discounted game defined by the HV transformation.

Lemma 44. *Suppose the conclusions of Proposition 41 hold with a Borel function μ. Then*

(i) $\tilde{r} : \tilde{\mathbb{K}} \to \mathbb{R}$ *is bounded and Borel-measurable,*

(ii) $\tilde{p}$ *is a Borel-measurable stochastic kernel on* $\tilde{\mathbb{X}}$ *given* $\tilde{\mathbb{K}}$*, and*

(iii) if Assumption P2 holds, then the stochastic game defined by the HV transformation also satisfies Assumption P2 with $\alpha \equiv \tilde{\beta}$.

Proof.

(i) This follows from the boundedness and Borel-measurability of both r and μ.

(ii) Fix $(x, a^1, a^2) \in \tilde{\mathbb{K}}$. By Proposition 41, $0 \leqslant \tilde{p}(B|x, a^1, a^2) \leqslant 1$ for all $B \in \mathcal{B}(\tilde{\mathbb{X}})$. Since Lebesgue's monotone convergence theorem implies that the set function

$$B \mapsto \int_B \mu(y) p(dy|x, a^1, a^2), \qquad B \in \mathcal{B}(\mathbb{X}),$$

is countably additive, it follows that $\tilde{p}(\cdot|x, a^1, a^2)$ is a probability measure on $(\tilde{\mathbb{X}}, \mathcal{B}(\tilde{\mathbb{X}}))$.

Next, fix $B \in \mathcal{B}(\tilde{\mathbb{X}})$. Let δ_x denote the Dirac measure on $(\tilde{\mathbb{X}}, \mathcal{B}(\tilde{\mathbb{X}}))$ sitting at $x \in \tilde{\mathbb{X}}$, and let 1_B denote the indicator function on $\tilde{\mathbb{X}}$ for $B \in \mathcal{B}(\tilde{\mathbb{X}})$. Note that for $(x, a^1, a^2) \in \tilde{\mathbb{K}}$,

$$\tilde{p}(B|x, a^1, a^2) = 1_{\mathbb{X}}(x) \left(\frac{\alpha(x, a^1, a^2)}{\tilde{\beta}\mu(x)} \int_{B \setminus \{\tilde{x}\}} \mu(y) p(dy|x, a^1, a^2) + \left[1 - \frac{\alpha(x, a^1, a^2)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y) p(dy|x, a^1, a^2) \right] \delta_{\tilde{x}}(B) \right) + 1_{\{\tilde{x}\}}(x) \delta_{\tilde{x}}(B).$$

Hence, according to the Borel-measurability of μ and [10, Proposition 7.29], the mapping $(x, a^1, a^2) \mapsto \tilde{p}(B|x, a^1, a^2)$ on $\tilde{\mathbb{K}}$ is Borel-measurable.

(iii) According to the definition of the HV transformation, $\tilde{A}^2(x)$ is compact for all $x \in \tilde{\mathbb{X}}$. Further, Assumption P2 and Proposition 41 imply that $\tilde{r}$ is bounded on $\tilde{\mathbb{K}}$ and $\tilde{r}(x, a^1, \cdot)$ is lower semicontinuous on $\tilde{A}^2(x)$ for all $x \in \tilde{\mathbb{X}}$ and $a^1 \in \tilde{A}^1(x)$. Finally, note that for any bounded Borel-measurable $f : \tilde{\mathbb{X}} \to \mathbb{R}$,

$$\int_{\tilde{\mathbb{X}}} f(y)\tilde{p}(dy|x, a^1, a^2) = 1_{\mathbb{X}}(x) \left(\frac{\alpha(x, a^1, a^2)}{\tilde{\beta}\mu(x)} \int_{B\setminus\{\tilde{x}\}} f(y)\mu(y)p(dy|x, a^1, a^2) + \right.$$
$$\left. \left[1 - \frac{\alpha(x, a^1, a^2)}{\tilde{\beta}\mu(x)} \int_{\mathbb{X}} \mu(y)p(dy|x, a^1, a^2)\right] f(\tilde{x}) \right) + 1_{\{\tilde{x}\}}(x)f(\tilde{x});$$

hence Assumption P2(iii) and [53, Proposition C.4] imply that for each $B \in \mathcal{B}(\tilde{\mathbb{X}})$, $x \in \tilde{\mathbb{X}}$, and $a^1 \in \tilde{A}^1(x)$, the function $\tilde{p}(B|x, a^1, \cdot)$ is continuous on $\tilde{A}^2(x)$.

□

Recall that a measure ν_1 is *absolutely continuous* with respect to another measure ν_2 on the same measurable space, written $\nu_1 \ll \nu_2$, if every ν_2-null set is also ν_1-null.

Assumption AC. *There is a* $\nu \in \mathcal{P}(\mathbb{X})$ *such that*

$$p(\cdot|x, a^1, a^2) \ll \nu \qquad \textit{for all } (x, a^1, a^2) \in \mathbb{K}. \tag{6.11}$$

Proposition 45. *Suppose Assumptions T and AC hold. Then there is a Borel-measurable function* $\mu_\nu : \mathbb{X} \to [1, \infty)$ *that is bounded above by* $2K$ *and satisfies (6.2).*

Proof. According to Proposition 41, there is an upper semianalytic $\mu : \mathbb{X} \to [1, \infty)$ that is bounded above by K and satisfies (6.2). Let $\mathbb{Q}$ denote the set of all rational numbers, and let $\mathbb{Q}_K := \{q \in \mathbb{Q} : 1 \leqslant q \leqslant K\}$. Since every analytic set is universally measurable (see e.g. [10, p. 171]), the sets

$$U(q) := \{x \in \mathbb{X} : \mu(x) \geqslant q\} \qquad q \in \mathbb{Q},$$

are universally measurable. By [10, Lemma 7.26], this implies that for each $q \in \mathbb{Q}$ there is a Borel subset $B(q)$ of $\mathbb{X}$ satisfying $\nu(U(q) \triangle B(q)) = 0$.

Noting that $\mu(x) = \sup\{q \in \mathbb{Q}_K : x \in U(q)\}$ for $x \in \mathbb{X}$, let

$$g_\nu(x) := \begin{cases} \sup\{q \in \mathbb{Q}_K : x \in B(q)\} & \text{if } x \in B(q), \\ 1 & \text{if } x \in \mathbb{X} \setminus B(q). \end{cases}$$

Observe that $g_\nu \geqslant 1$ is bounded above by K. Also, letting

$$f_q(x) := \begin{cases} q & \text{if } x \in B(q), \\ 1 & \text{otherwise,} \end{cases}$$

we have $g_\nu(x) = \sup_{q \in \mathbb{Q}_K} f_q(x)$ for $x \in \mathbb{X}$; hence g_ν is Borel-measurable.

Further, since

$$N := \{x \in \mathbb{X} : g_\nu(x) \neq \mu(x)\} \subseteq \bigcup_{q \in \mathbb{Q}_K} [U(q) \triangle B(q)],$$

the function g_ν is ν-almost everywhere equal to μ. Letting 1_N denote the indicator function for the set N, define μ_ν by

$$\mu_\nu(x) := g_\nu(x) + K1_N(x), \qquad x \in \mathbb{X}.$$

Consider $(x, a^1, a^2) \in \mathbb{K}$. If $x \notin N$, then (6.2) and Assumption AC imply that

$$\begin{aligned} \mu_\nu(x) = g_\nu(x) &\geqslant 1 + \alpha(x, a^1, a^2) \int_{\mathbb{X}} g_\nu(y) p(dy|x, a^1, a^2) \\ &= 1 + \alpha(x, a^1, a^2) \int_{\mathbb{X}} \mu_\nu(y) p(dy|x, a^1, a^2). \end{aligned}$$

On the other hand, if $x \in N$ then

$$\begin{aligned} 1+&\alpha(x, a^1, a^2) \int_{\mathbb{X}} \mu_\nu(y) p(dy|x, a^1, a^2) \\ &= 1 + \alpha(x, a^1, a^2) \int_{\mathbb{X}} \mu(y) p(dy|x, a^1, a^2) \\ &\leqslant \mu(x) \leqslant g_\nu(x) + K = \mu_\nu(x). \end{aligned}$$

□

6.4 Existence of ϵ-Optimal Strategies

For $(\varphi^1, \varphi^2) \in \Phi^1 \times \Phi^2$, define the operators $T_\alpha^{\varphi^1\varphi^2}$, $T_\alpha^{\varphi^2}$, and T_α for bounded upper semianalytic functions u on $\mathbb{X}$ by

$$T_\alpha^{\varphi^1\varphi^2}u(x) := \int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\left(r(x,a^1,a^2) + \alpha(x,a^1,a^2)\int_{\mathbb{X}} u(y)p(dy|x,a^1,a^2)\right)\varphi^1(da^1|x)\varphi^2(da^2|x),$$

$T_\alpha^{\varphi^2}u := \sup_{\varphi^1\in\Phi^1} T_\alpha^{\varphi^1\varphi^2}u$, and $T_\alpha u := \inf_{\varphi^2\in\Phi^2} T_\alpha^{\varphi^2}u$.

Theorem 46. *Suppose Assumptions T, AC, and P2 hold. Then player 1 has an ϵ-α-optimal stationary strategy for any $\epsilon > 0$, and player 2 has an α-optimal stationary strategy. Further, the game has an α-value v_α which is bounded, upper semianalytic, and uniquely satisfies*

$$v_\alpha(x) = T_\alpha v_\alpha(x), \quad x \in \mathbb{X}. \tag{6.12}$$

Proof. By Lemma 44, the discounted stochastic game with $\alpha \equiv \tilde{\beta} \in [0,1)$ defined by the HV transformation satisfies Assumption P2. Hence the conclusions of Proposition 39 hold for this game. According to Corollary 43, it follows that player 1 has an ϵ-α-optimal stationary strategy, and player 2 has an α-optimal stationary strategy, for the original α-discounted game.

In addition, Proposition 39 and Corollary 43 also imply that $v_\alpha = \mu\tilde{v}_{\tilde{\beta}}$ is the value of the original α-discounted game, which by (5.1) satisfies (6.12). Moreover, v_α is bounded by the boundedness of μ and $\tilde{v}_{\tilde{\beta}}$, and is upper semianalytic by Proposition 45 and [10, Lemma 7.30(4)]. To show that v_α is the unique bounded function satisfying (6.12), suppose u is a bounded function that satisfies (6.12). Then u/μ is a bounded function that satisfies (5.1) with $\beta = \tilde{\beta}$; by Proposition 39, this implies that $u = \mu\tilde{v}_{\tilde{\beta}} = v_\alpha$. □

6.4.1 Existence of Optimal Strategies

Lemma 47. *Suppose the conclusions of Proposition 41 hold with a Borel function μ. If Assumption P1 holds, then the stochastic game defined by the HV transformation also satisfies Assumption P1 with $\alpha \equiv \tilde{\beta}$.*

Proof. This follows *mutatis mutandis* from the proof of statement (iii) of Lemma 44. □

Proposition 48. *Suppose Assumptions T, P1 and P2 hold. Then there is a Borel-measurable function $\mu : \mathbb{X} \to [1, \infty)$ that is bounded above by K and satisfies (6.2).*

Proof. Recall the operator $\mathcal{U}$ defined by (6.3). Letting $u_0 \equiv 0$ and $u_n := \mathcal{U}u_{n-1}$ for $n = 1, 2, \ldots$, it was shown that $\{u_n\}_{n \geqslant 0}$ increases to an upper semianalytic function μ that is bounded above by K and satisfies (6.2).

We now show by induction that for $n = 1, 2, \ldots$ the function u_n is Borel-measurable, from which the Borel-measurability of $\mu = \lim_{n\to\infty} u_n$ follows. First, note that $u_1 \equiv 1$ is Borel-measurable. Next, suppose that for some $n \geqslant 1$ the function u_n is Borel-measurable. For $(x, a^1, a^2) \in \mathbb{K}$, let

$$\eta_n(x, a^1, a^2) := 1 + \alpha(x, a^1, a^2) \int_{\mathbb{X}} u_n(y) p(dy|x, a^1, a^2).$$

Consider any $\lambda \in \mathbb{R}$. According to [10, Proposition 7.29], the function $\eta_n : \mathbb{K} \to \mathbb{R}$ is Borel-measurable, which means the set $\mathcal{D}_{\eta_n}(\lambda) := \{(x, a^1, a^2) \in \mathbb{K} : \eta_n(x, a^1, a^2) \geqslant \lambda\}$ is Borel. Further, the continuity of $\alpha(x, \cdot, \cdot)$ and $p(B|x, \cdot, \cdot)$ for $B \in \mathcal{B}(\mathbb{X})$ imply that for $x \in \mathbb{X}$ the function $\eta_n(x, \cdot, \cdot)$ on $A^1(x) \times A^2(x)$ is continuous. This means that $\mathcal{D}_{\eta_n}(\lambda)$ has closed x-sections for $x \in \mathbb{X}$, and that

$$\begin{aligned}\{x \in \mathbb{X} : u_{n+1}(x) \geqslant \lambda\} &= \{x \in \mathbb{X} : \eta_n(x, a^1, a^2) \\ &\geqslant \lambda \text{ for some } (a^1, a^2) \in A^1(x) \times A^2(x)\} \\ &= \operatorname{proj}_{\mathbb{X}} \mathcal{D}_{\eta_n}(\lambda).\end{aligned}$$

As the compactness of $A^1(x)$ and $A^2(x)$ for $x \in \mathbb{X}$ imply that $\mathbb{K} \supseteq \mathcal{D}_{\eta_n}(\lambda)$ has compact x-sections for $x \in \mathbb{X}$, it follows from the Arsenin-Kunugui Theorem (see e.g. [67, Theorem 35.46]) that $\{x \in \mathbb{X} : u_{n+1}(x) \geqslant \lambda\}$ is Borel. Hence u_{n+1} is also Borel-measurable. □

Theorem 49. *Suppose Assumptions T, P1, and P2 hold. Then both players have α-optimal stationary strategies that are Borel-measurable. Further, the game has an α-value v_α, which is Borel-measurable and satisfies (6.12).*

Proof. According to Proposition 48, it follows from Lemma 47 and statement (iii) of Lemma 44 that the discounted stochastic game with $\alpha \equiv \tilde{\beta} \in [0, 1)$ defined by the HV transformation satisfies Assumptions P1 and P2. Hence the conclusions of Proposition 40 hold for this game. According to Corollary 43, it follows that both players have α-optimal stationary strategies. The existence and Borel-measurability of the value v_α follow *mutatis mutandis* from the proof of Theorem 46. □

6.5 Complexity Estimates

In this section, we provide complexity estimates related to applying the HV transformation for stochastic games. In Section 6.5.1, we provide an upper bound on the number of arithmetic operations needed to compute a function μ for the HV transformation. Then, in Section 6.5.2 we provides estimates for the number of arithmetic operations needed to compute a pair of optimal strategies for two-player zero-sum transient stochastic games with perfect information.

6.5.1 Constructing the Transformation

Note that, given a suitable function μ, the two-player zero-sum stochastic game defined by the HV transformation can be constructed with a number of arithmetic operations that is polynomial in the total number of state-action triples m of the original game. The following theorem provides an estimate of the complexity of computing a function μ that can be used for the HV transformation.

Theorem 50. *Suppose the state set $\mathbb{X}$ and action sets $\mathbb{A}^i$, $i = 1, 2$, are finite, and that Assumption T holds. Then the number of arithmetic opterations needed to compute a function μ satisfying the hypotheses of Proposition 41 is*

$$O(mK \log K),$$

where $m := \sum_{x\in\mathbb{X}} |A^1(x)| \cdot |A^2(x)|$.

Proof. To compute a function satisfying the hypotheses of Proposition 41, it suffices to compute a bounded nonnegative function μ that satisfies

$$\mu(x) = \max_{(a^1,a^2)\in A^1(x)\times A^2(x)} \left[1 + \alpha(x, a^1, a^2) \sum_{y\in\mathbb{X}} p(y|x, a^1, a^2)\mu(y) \right] \tag{6.13}$$

for all $x \in \mathbb{X}$. Let $q(y|x, a^1, a^2) := \alpha(x, a^1, a^2)p(y|x, a^1, a^2)$ for $x, y \in \mathbb{X}$ and $(a^1, a^2) \in A^1(x) \times A^2(x)$, and consider the Markov decision process with state set $\mathbb{X}$, action sets $A(x) := A^1(x) \times A^2(x)$ for $x \in \mathbb{X}$, transition rates $q(y|x, a)$ for $x, y \in \mathbb{X}$ and $a \in A(x)$, and one-step rewards identically equal to one. According to Assumption T, this MDP is transient; see [23, Hypothesis 1]. Hence it follows from [23, Theorem 2] that the number of arithmetic operations needed, to compute a nonnegative function that is bounded above by K and satisfies (6.13), is at most a constant times $mK \log K$. □

6.5.2 Computing Optimal Strategies

Assumption PI. *The stochastic game is one of perfect information, i.e. there exist disjoint sets $\mathbb{X}^1, \mathbb{X}^2 \subseteq \mathbb{X}$ such that $|A^2(x)| = 1$ for all $x \in \mathbb{X}^1$ and $|A^1(x)| = 1$ for all $x \in \mathbb{X}^2$.*

Note that each state of a perfect-information stochastic game can be viewed as being controlled by only one of the players. For $i = 1, 2$, let $n_i := |\mathbb{X}^i|$ and $m_i := \sum_{x \in \mathbb{X}^i} |A^i(x)|$ respectively denote the total number of states that player i controls and the total number of actions in those states. Further, let $n := n_1 + n_2$ and $m := m_1 + m_2$.

Theorem 51. *Suppose the state set $\mathbb{X}$ and action sets $\mathbb{A}^i$, $i = 1, 2$, are finite, and that Assumptions T and PI hold. Then both players have α-optimal deterministic stationary strategies, and the number of arithmetic operations needed to compute a pair of such strategies is*

$$O\left(\left((n_1^3 + n_1^2 m_1) m_1 K \log K + n^3 + m_2 n_2^2\right) \cdot mK \log nK\right).$$

Proof. Consider the strategy iteration algorithm for discounted stochastic games described by Rao et al. [93]; see also Hansen et al. [51, Algorithm 2]. Beginning with an arbitrary deterministic stationary strategy $\phi_0^2 \in \mathbb{F}^2$ for the minimizing player, each iteration $k = 0, 1, \ldots$ of the algorithm generates a deterministic stationary strategy $\phi_k^1 \in \mathbb{F}^1$ for the maximizing player and a new strategy $\phi_{k+1}^2 \in \mathbb{F}^2$ for the minimizing player as follows.

First, ϕ_k^1 is computed by solving the discounted total-reward Markov decision process obtained by fixing player 2's strategy to ϕ_k^0; when the discount factor is $\beta \in (0, 1)$, according to Scherrer [103, Theorem 3] the number of arithmetic operations needed to accomplish this as at most a constant times $(n_1^3 + n_1^2 m_1) m_1 (1 - \beta)^{-1} \log(1 - \beta)^{-1}$. Next, the total discounted payoff function $v_\beta^{\phi_k^1 \phi_k^2}$ for the maximizing player is computed by solving an $n \times n$ system of linear equations; the number of arithmetic operations needed for this is at most a constant times n^3 via e.g. Gaussian elimination[1]. The last step on iteration k consists of selecting a new deterministic stationary strategy $\phi^2_{k+1} \in \mathbb{F}^2$ for the minimizing player satisfying

$$\phi^2_{k+1}(x) \in \operatorname*{arg\,min}_{a^2 \in A^2(x)} \left[r(x, \phi^1_k, a^2) + \beta \sum_{y \in \mathbb{X}} p(y|x, \phi^1_k, a^2) v_\beta^{\phi^1_k \phi^2_k}(y) \right]$$

for all $x \in \mathbb{X}$; here the required number of arithmetic operations is at most a constant times $m_2 n_2^2$. Moreover, according to Hansen et al. [51, Theorem 7.5], the total number of iterations of the strategy iteration algorithm needed to return a pair of optimal deterministic stationary strategies is at most a constant times $m(1-\beta)^{-1} \log(1-\beta)^{-1}$. Hence the total number of arithmetic operations needed to compute a pair of optimal deterministic stationary strategies is at most a constant times

$$\left((n_1^3 + n_1^2 m_1) \frac{m_1}{1-\beta} \log \frac{1}{1-\beta} + n^3 + m_2 n_2^2 \right) \cdot \frac{m}{1-\beta} \log \frac{n}{1-\beta}. \tag{6.14}$$

According to Theorem 50, the number of arithmetic operations needed to compute a function μ that is bounded above by $K < \infty$ for the Hoffman-Veinott transformation is at most a constant times $mK \log K$. In addition, with $\beta := (K-1)K^{-1}$, the number of arithmetic operations needed to compute a pair of optimal deterministic stationary strategies for the resulting discounted stochastic game is at most a constant times (6.14). Hence the total number of arithmetic operations needed is at most a constant times

$$mK \log K + \left((n_1^3 + n_1^2 m_1) m_1 K \log K + n^3 + m_2 n_2^2 \right) \cdot mK \log nK;$$

the theorem then follows from Proposition 42. □

[1] Alternatively, using Williams' [119] improvements on Coppersmith & Winograd's [19] algorithm, on the order of $n^{2.3727}$ arithmetic operations are needed.

Chapter 7

Reduction of Average-Payoff Stochastic Games

In this chapter, we consider the reduction of certain two-player zero-sum average-payoff stochastic games to discounted ones. In Section 7.1 we state the hitting time assumption that will be used in the reduction, and in Section 7.2 we define the average-payoff optimality criterion. In Section 7.3 we use the hitting time assumption to construct a discounted-payoff stochastic game. The existence of ϵ-optimal strategy pairs is considered in Section 7.4, and the existence of optimal strategy pairs is considered in Section 7.4.1. Finally, complexity estimates for computing optimal strategy pairs are given for two-player zero-sum average-payoff stochastic games of perfect information in Section 7.5.

7.1 Hitting Time Assumption

For $x \in \mathbb{X}$, let τ_x denote the *hitting time* to state x after time 0, which is sometimes called the *first return time* to state x. In particular, $\tau_x(\omega) := \inf\{t \geqslant 1 : \xi_t(\omega) = x\}$ for $\omega = x_0 a_0^1 a_0^2 \cdots \in H_\infty$.

Assumption HT. *There is a special state* $\ell \in \mathbb{X}$ *and a constant* $L \geqslant 1$ *satisfying*

$$\mathbb{E}_x^{\phi^1\phi^2}\tau_\ell \leqslant L < \infty \tag{7.1}$$

for all $x \in \mathbb{X}$ *and* $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$.

Proposition 52. *Suppose Assumption HT holds. Then there is an upper semianalytic function* $\bar{\mu} : \mathbb{X} \to [1, \infty)$ *that is bounded above by* $\bar{K} := L + 1$ *and satisfies*

$$\bar{\mu}(x) \geqslant 1 + \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y) p(dy|x, a^1, a^2) \quad \textit{for all} \ \ (x, a^1, a^2) \in \mathbb{K}. \tag{7.2}$$

Proof. For $(x, a^1, a^2) \in \mathbb{K}$, let

$$q(B|x, a^1, a^2) := \frac{p(B \setminus \{\ell\}|x, a^1, a^2)}{p(\mathbb{X} \setminus \{\ell\}|x, a^1, a^2)}, \qquad B \in \mathcal{B}(\mathbb{X}), \tag{7.3}$$

and $\alpha(x, a^1, a^2) := p(\mathbb{X} \setminus \{\ell\}|x, a^1, a^2)$. Note that $0 \leqslant q(B|x, a^1, a^2)$ and $q(\mathbb{X}|x, a^1, a^2) = 1$ for $(x, a^1, a^2) \in \mathbb{K}$. Moreover, Lebesgue's monotone convergence theorem implies that for each $(x, a^1, a^2) \in \mathbb{K}$ the set function $B \mapsto q(B|x, a^1, a^2)$ is countably additive. Also, since p is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$, for each $B \in \mathcal{B}(\mathbb{X})$ the mapping $(x, a^1, a^2) \mapsto q(B|x, a^1, a^2)$ is Borel-measurable. Hence q is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$. In addition, the mapping $(x, a^1, a^2) \mapsto p(\mathbb{X} \setminus \{\ell\}|x, a^1, a^2) = \alpha(x, a^1, a^2)$ is Borel-measurable by [10, Proposition 7.29].

Replace the transition probabilities p with q, and for $x \in \mathbb{X}$ and any pair $(\pi^1, \pi^2) \in \Pi^1 \times \Pi^2$ of strategies for players 1 and 2, let $Q_x^{\pi^1\pi^2}$ and $E_x^{\pi^1\pi^2}$ respectively denote the corresponding strategic measure and expectation operator for the resulting stochastic game. By Assumption HT and the definition of q, for $x \in \mathbb{X}$ and $(\phi^1, \phi^2) \in \mathbb{F}^1 \times \mathbb{F}^2$

$$\infty > \bar{K} := L + 1 \geqslant \sum_{n=0}^{\infty} Q_x^{\phi^1\phi^2}\{\tau_\ell > n\} = \sum_{n=0}^{\infty} E_x^{\phi^1\phi^2} \prod_{k=1}^{n-1} \alpha(x_k, a_k^1, a_k^2).$$

According to Proposition 41, it follows that there is an upper semianalytic function $\bar{\mu} : \mathbb{X} \to [1, \infty)$ that is bounded above by $\bar{K}$ and satisfies

$$\bar{\mu}(x) \geqslant 1 + \alpha(x, a^1, a^2) \textstyle\int_{\mathbb{X}} \bar{\mu}(y) q(dy|x, a^1, a^2) = 1 + \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y) p(dy|x, a^1, a^2)$$

for all $(x, a^1, a^2) \in \mathbb{K}$. □

7.2 Optimality Criterion

When the initial state is $x \in \mathbb{X}$ and players 1 and 2 follow the strategies $\pi^1 \in \Pi^1$ and $\pi^2 \in \Pi^2$, respectively, the expected long-run *average payoff* that player 1 receives from player 2 is

$$w^{\pi^1\pi^2}(x) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_x^{\pi^1\pi^2}\sum_{t=0}^{T-1} r(\xi_t, \upsilon_t^1, \upsilon_t^2).$$

Define the *lower value* function ${}_*w : \mathbb{X} \to \mathbb{R}$ by

$${}_*w(x) := \sup_{\pi^1\in\Pi^1}\inf_{\pi^2\in\Pi^2} w^{\pi^1\pi^2}(x), \qquad x \in \mathbb{X}.$$

and the *upper value* function ${}^*w : \mathbb{X} \to \mathbb{R}$ by

$${}^*w(x) := \inf_{\pi^2\in\Pi^2}\sup_{\pi^1\in\Pi^1} w^{\pi^1\pi^2}(x), \qquad x \in \mathbb{X}.$$

7.3 Akian-Gaubert (AG) Transformation

We now describe a transformation to a discounted game, which we call the *Akian-Gaubert (AG) transformation*. Objects associated with the discounted game will be indicated by a horizontal bar. The state space is $\bar{\mathbb{X}} := \mathbb{X} \cup \{\bar{x}\}$, where $\bar{x} \notin \mathbb{X}$ is a cost-free absorbing state. Letting $\bar{a}$ denote the only action available to players 1 and 2 at state $\bar{x}$, the action space for player i is $\bar{\mathbb{A}}^i := \mathbb{A}^i \cup \{\bar{a}\}$, $i = 1, 2$, and for $x \in \bar{\mathbb{X}}$ the set of available actions is unchanged if $x \in \mathbb{X}$, namely for $i = 1, 2$

$$\bar{A}^i(x) := \begin{cases} A^i(x), & \text{if } x \in \mathbb{X}, \\ \{\bar{a}\}, & \text{if } x = \bar{x}. \end{cases}$$

Define the payoff function $\bar{r}$ for player 1 (i.e. the payment function for player 2) by

$$\bar{r}(x, a^1, a^2) := \begin{cases} \bar{\mu}(x)^{-1} r(x, a^1, a^2), & \text{if } x \in \mathbb{X},\ a^1 \in \mathbb{A}^1,\ a^2 \in \mathbb{A}^2, \\ 0, & \text{if } (x, a) = (\bar{x}, \bar{a}). \end{cases}$$

To complete the definition of the discounted game, choose a discount factor

$$\bar{\beta} \in \left[\frac{L-1}{L}, 1\right),$$

and let

$$\bar{p}(B|x,a^1,a^2) := \begin{cases} \frac{1}{\bar{\beta}\bar{\mu}(x)}\int_B \bar{\mu}(y)p(dy|x,a^1,a^2), & B \in \mathcal{B}(\mathbb{X}\setminus\{\ell\}),\ x \in \mathbb{X},\ a^1 \in \mathbb{A}^1,\ a^2 \in \mathbb{A}^2, \\ \frac{1}{\bar{\beta}\bar{\mu}(x)}[\bar{\mu}(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y)p(dy|x,a^1,a^2)], & B = \{\ell\},\ x \in \mathbb{X},\ a^1 \in \mathbb{A}^1,\ a^2 \in \mathbb{A}^2, \\ 1 - \frac{1}{\bar{\beta}\bar{\mu}(x)}[\bar{\mu}(x) - 1], & B = \{\bar{x}\},\ x \in \mathbb{X},\ a^1 \in \mathbb{A}^1,\ a^2 \in \mathbb{A}^2, \\ 1, & B = \{\bar{x}\},\ (x,a^1,a^2) = (\bar{x},\bar{a},\bar{a}). \end{cases}$$

Since $\bar{A}^i(\bar{x})$ is a singleton for $i = 1,2$, the sets of all strategies available to each player for the discounted-payoff game and the original game coincide. Given $x \in \bar{\mathbb{X}}$ and $(\pi^1,\pi^2) \in \Pi^1 \times \Pi^2$, let $\bar{\mathbb{E}}_x^{\pi^1\pi^2}$ denote the expectation operator associated with the discounted-payoff game and

$$\bar{v}_{\bar{\beta}}(x,\pi^1,\pi^2) := \bar{\mathbb{E}}_x^{\pi^1\pi^2}\sum_{n=0}^{\infty}\bar{\beta}^n\bar{r}(x_n,a_n^1,a_n^2).$$

Given an initial state $x \in \bar{\mathbb{X}}$, let

$${}_*\bar{v}_{\bar{\beta}}(x) := \sup_{\pi^1\in\Pi^1}\inf_{\pi^2\in\Pi^2}\bar{v}_{\bar{\beta}}(x,\pi^1,\pi^2), \quad {}^*\bar{v}_{\bar{\beta}}(x) := \inf_{\pi^2\in\Pi^2}\sup_{\pi^1\in\Pi^1}\bar{v}_{\bar{\beta}}(x,\pi^1,\pi^2).$$

For $\epsilon \geqslant 0$, a strategy $\pi_*^1 \in \Pi^1$ is ϵ-$\bar{\beta}$-*optimal for player 1* if $\bar{v}_{\bar{\beta}}(x,\pi_*^1,\sigma^2) \geqslant {}^*\bar{v}_{\bar{\beta}}(x) - \epsilon$ for all $x \in \bar{\mathbb{X}}$ and $\sigma^2 \in \Pi^2$, and $\pi_*^2 \in \Pi^2$ is ϵ-$\bar{\beta}$-*optimal for player 2* if $\bar{v}_{\bar{\beta}}(x,\sigma^1,\pi_*^2) \leqslant {}_*\bar{v}_{\bar{\beta}}(x) + \epsilon$ for all $x \in \bar{\mathbb{X}}$ and $\sigma^1 \in \Pi^1$. A 0-optimal strategy is called *optimal*. If ${}_*\bar{v}_{\bar{\beta}} = {}^*\bar{v}_{\bar{\beta}} =: \bar{v}_{\bar{\beta}}$, then $\bar{v}_{\bar{\beta}}$ is the $\bar{\beta}$-*discounted value* of the game.

For $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$ and $\beta \in [0,1]$, define the operators $T_\beta^{\varphi^1\varphi^2}$, $T_\beta^{\varphi^2}$, and T_β for bounded upper semianalytic functions u on $\mathbb{X}$ by

$$T_\beta^{\varphi^1\varphi^2}u(x) := \int_{\mathbb{A}^2}\int_{\mathbb{A}^1}\left(r(x,a^1,a^2) + \beta\int_{\mathbb{X}}u(y)p(dy|x,a^1,a^2)\right)\varphi^1(da^1|x)\varphi^2(da^2|x),$$

$T_\beta^{\varphi^2}u := \sup_{\varphi^1\in\Phi^1}T_\beta^{\varphi^1\varphi^2}u$, and $T_\beta u := \inf_{\varphi^2\in\Phi^2}T_\beta^{\varphi^2}u$. For the stochastic game defined by the AG transformation, the operators $\bar{T}_\beta^{\varphi^1\varphi^2}$, $\bar{T}_\beta^{\varphi^2}$, and $\bar{T}_\beta$ are defined analogously.

Proposition 53. *Consider* $(\varphi^1,\varphi^2) \in \Phi^1 \times \Phi^2$*, and let*

$$h(x,\varphi^1,\varphi^2) := \bar{\mu}(x)[\bar{v}_{\bar{\beta}}(x,\varphi^1,\varphi^2) - \bar{v}_{\bar{\beta}}(\ell,\varphi^1,\varphi^2)]$$

for $x \in \mathbb{X}$*. If* r *is bounded, then* $w(\cdot,\varphi^1,\varphi^2) \equiv \bar{v}_{\bar{\beta}}(\ell,\varphi^1,\varphi^2)$*.*

Proof. Since the state $\overline{x}$ in the stochastic game defined by the AG transformation is payoff-free and absorbing for both players, and

$$\bar{v}_{\bar{\beta}}(x, \varphi^1, \varphi^2) = \bar{T}_1 \bar{v}_{\bar{\beta}}(x)$$

for $x \in \bar{\mathbb{X}}$, it follows from the definition of $h(\cdot, \varphi^1, \varphi^2)$ that for $x \in \mathbb{X}$,

$$\bar{v}_{\bar{\beta}}(\ell, \varphi^1, \varphi^2) + h(x, \varphi^1, \varphi^2) = T_1 h(x). \tag{7.4}$$

Iterating (7.4) gives

$$N\bar{v}_{\bar{\beta}}(\ell, \varphi^1, \varphi^2) + h(x, \varphi^1, \varphi^2) = \mathbb{E}_x^{\varphi^1\varphi^2} \textstyle\sum_{n=0}^{N-1} r(x_n, a_n^1, a_n^2) + \mathbb{E}_x^{\varphi^1\varphi^2} h(x_N, \varphi^1, \varphi^2) \tag{7.5}$$

for $N = 1, 2, \ldots$ and $x \in \mathbb{X}$. Since r and $\bar{\mu}$ are bounded, $h(\cdot, \varphi^1, \varphi^2)$ is bounded as well; hence the equality $w(\cdot, \varphi^1, \varphi^2) \equiv \bar{v}_{\bar{\beta}}(\ell, \varphi^1, \varphi^2)$ follows by dividing both sides of (7.5) by N and letting $N \to \infty$. □

7.4 Existence of ϵ-Optimal Strategies

Proposition 54. *Suppose Assumption P2 holds, consider any $\beta \in [0, 1)$, and let u be any bounded upper semianalytic function on $\mathbb{X}$. Then for any $\epsilon > 0$ there exist stationary strategies $\varphi^1_\epsilon \in \Phi^1$ and $\varphi^2_* \in \Phi^2$ satisfying*

$$\sup_{\varphi^1 \in \Phi^1} T_\beta^{\varphi^1 \varphi^2_*} u \leqslant T_\beta u \leqslant \inf_{\varphi^2 \in \Phi^2} T_\beta^{\varphi^1_\epsilon \varphi^2} u + \epsilon. \tag{7.6}$$

Proof. Suppose u is bounded below by $M > -\infty$. Letting $\underline{u} := u - M$, it follows from [85, Theorem 5.1] xthat there exist $\varphi^1_\epsilon \in \Phi^1$ and $\varphi^2_* \in \Phi^2$ satisfying

$$\sup_{\varphi^1 \in \Phi^1} T_\beta^{\varphi^1 \varphi^2_*} \underline{u} \leqslant T_\beta \underline{u} \leqslant \inf_{\varphi^2 \in \Phi^2} T_\beta^{\varphi^1_\epsilon \varphi^2} \underline{u} + \epsilon,$$

from which (7.6) follows by the definition of $\underline{u}$. □

Lemma 55. *Let u be any bounded upper semianalytic function on $\mathbb{X}$, and consider any $\epsilon \geqslant 0$.*

(i) If player 1 has a strategy $\varphi^1_\epsilon \in \Phi^1$ satisfying $\inf_{\varphi^2 \in \Phi^2} T_1^{\varphi^1_\epsilon \varphi^2} u \geqslant T_1 u - \epsilon$, then

$$\inf_{\pi^2 \in \Pi^2} w(\cdot, \varphi^1_\epsilon, \pi^2) \geqslant \inf_{x \in \mathbb{X}} [T_1 u(x) - u(x)] - \epsilon. \tag{7.7}$$

(ii) *If player 2 has a strategy* $\varphi^2_\epsilon \in \Phi^2$ *satisfying* $\sup_{\varphi^1\in\Phi^1} T_1^{\varphi^1\varphi^2_\epsilon}u \leqslant T_1u + \epsilon$, *then*

$$\sup_{\pi^1\in\Pi^1} w(\cdot,\pi^1,\varphi^2_\epsilon) \leqslant \sup_{x\in\mathbb{X}}[T_1u(x) - u(x)] + \epsilon. \tag{7.8}$$

Proof. For $i = 1,2$, a strategy π^i for player i is a *Markov strategy* if for $n = 0,1,\dots$ there is a universally measurable stochastic kernel φ^i_t on $\mathbb{A}^i$ given $\mathbb{X}$ satisfying $\pi^i_n(\cdot|h_n) = \varphi^i_n(\cdot|x_n)$ for all $h_n = x_0a^1_0a^2_0\cdots x_n \in \mathbb{H}_n$; each such φ^i_n is called a *decision rule* for player i. Let Π^i_M denote the set of all Markov strategies for player $i = 1,2$.

According to the sufficiency of Markov policies for discrete-time MDPs (see e.g. [35, Corollary 6.1]), it suffices to prove that the infimum in (7.7) and the supremum in (7.8) respectively hold when Π^2 and Π^1 are replaced with Π^2_M and Π^1_M, respectively.

(i) Consider any sequence $\{\varphi^2_n\}_{n=1}^\infty$ of decision rules for player 2. Since $\inf_{\varphi^2\in\Phi^2} T_1^{\varphi^1_\epsilon\varphi^2}u \geqslant T_1u - \epsilon$, it follows that

$$T_1^{\varphi^1_\epsilon\varphi^2_1}u \geqslant u + T_1u - u - \epsilon \geqslant u + \inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - \epsilon.$$

Further, if $T_1^{\varphi^1_\epsilon\varphi^2_n}\cdots T_1^{\varphi^1_\epsilon\varphi^2_1}u \geqslant u + n\inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - n\epsilon$ for some positive integer n, then

$$\begin{aligned} T_1^{\varphi^1_\epsilon\varphi^2_{n+1}}T_1^{\varphi^1_\epsilon\varphi^2_n}\cdots T_1^{\varphi^1_\epsilon\varphi^2_1}u &\geqslant T_1^{\varphi^1_\epsilon\varphi^2_{n+1}}u + n\inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - n\epsilon \\ &\geqslant T_1u - \epsilon + n\inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - n\epsilon \\ &\geqslant u + (n+1)\inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - (n+1)\epsilon. \end{aligned}$$

It follows that for any initial state $x \in \mathbb{X}$, any Markov strategy $\sigma^2 \in \Pi^2_M$ for player 2, and any positive integer N,

$$\begin{aligned} \mathbb{E}_x^{\varphi^1_\epsilon\sigma^2}\sum_{n=0}^{N-1} r(x_n,a^1_n,a^2_n) &\geqslant \mathbb{E}_x^{\varphi^1_\epsilon\sigma^2}\left[\sum_{n=0}^{N-1} r(x_n,a^1_n,a^2_n) + u(x_N)\right] - \sup_{x\in\mathbb{X}} u(x) \\ &\geqslant u(x) + N\inf_{x\in\mathbb{X}}[T_1u(x) - u(x)] - N\epsilon - \sup_{x\in\mathbb{X}} u(x). \end{aligned}$$

Therefore the boundedness of u implies that for any initial state $x \in \mathbb{X}$ and any Markov strategy σ^2 for player 2,

$$\begin{aligned} w(x, \varphi^1_\epsilon, \sigma^2) &\geqslant \liminf_{N\to\infty} \frac{u(x) - \sup_{x\in\mathbb{X}} u(x)}{N} + \inf_{x\in\mathbb{X}} [T_1 u(x) - u(x)] - \epsilon \\ &= \inf_{x\in\mathbb{X}} [T_1 u(x) - u(x)] - \epsilon. \end{aligned}$$

(ii) Consider any sequence $\{\varphi^1_n\}_{n=1}^\infty$ of decision rules for player 1. Since $\sup_{\varphi^1\in\Phi^2} T_1^{\varphi^1\varphi^2_\epsilon} u \leqslant T_1 u + \epsilon$, it follows that

$$T_1^{\varphi^1_1\varphi^2_\epsilon} u \leqslant u + T_1 u - u + \epsilon \leqslant u + \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + \epsilon.$$

Further, if $T_1^{\varphi^1_n\varphi^2_\epsilon} \cdots T_1^{\varphi^1_1\varphi^2_\epsilon} u \leqslant u + n \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + n\epsilon$ for some positive integer n, then

$$\begin{aligned} T_1^{\varphi^1_{n+1}\varphi^2_\epsilon} T_1^{\varphi^1_n\varphi^2_\epsilon} \cdots T_1^{\varphi^1_1\varphi^2_\epsilon} u &\leqslant T_1^{\varphi^1_{n+1}\varphi^2_\epsilon} u + n \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + n\epsilon \\ &\leqslant T_1 u + \epsilon + n \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + n\epsilon \\ &\leqslant u + (n+1) \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + (n+1)\epsilon \end{aligned}$$

It follows that for any initial state $x \in \mathbb{X}$, any Markov strategy $\sigma^1 \in \Pi^1_M$ for player 1, and any positive integer N,

$$\begin{aligned} \mathbb{E}_x^{\sigma^1\varphi^2_\epsilon} \sum_{n=0}^{N-1} r(x_n, a^1_n, a^2_n) &\leqslant \mathbb{E}_x^{\sigma^1\varphi^2_\epsilon} \left[\sum_{n=0}^{N-1} r(x_n, a^1_n, a^2_n) + u(x_N) \right] - \inf_{x\in\mathbb{X}} u(x) \\ &\leqslant u(x) + N \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + N\epsilon - \inf_{x\in\mathbb{X}} u(x). \end{aligned}$$

Therefore the boundedness of u implies that for any initial state $x \in \mathbb{X}$ and any Markov strategy σ^1 for player 1,

$$\begin{aligned} w(x, \sigma^1, \varphi^2_\epsilon) &\leqslant \liminf_{N\to\infty} \frac{u(x) - \inf_{x\in\mathbb{X}} u(x)}{N} + \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + \epsilon \\ &= \sup_{x\in\mathbb{X}} [T_1 u(x) - u(x)] + \epsilon. \end{aligned}$$

□

Proposition 56. *Suppose the bounded function* u *on* $\mathbb{X}$ *satisfies*

$$\sup_{x\in\mathbb{X}}[T_1 u(x) - u(x)] = \inf_{x\in\mathbb{X}}[T_1 u(x) - u(x)] =: \rho,$$

and that for any $\epsilon > 0$ *there exist* $\varphi^1_\epsilon \in \Phi^1$ *and* $\varphi^2_* \in \Phi^2$ *satisfying*

$$\sup_{\varphi^1\in\Phi^1} T_1^{\varphi^1\varphi^2_*} u \leqslant T_1 u \leqslant \inf_{\varphi^2\in\Phi^2} T_1^{\varphi^1_\epsilon\varphi^2} u + \epsilon. \tag{7.9}$$

Then player 1 has an ϵ*-optimal strategy* φ^1_ϵ *for any* $\epsilon > 0$, φ^2_* *is optimal for player 2, and the average-payoff stochastic game has a value* $w \equiv \rho$.

Proof. Since $\inf_{\varphi^2\in\Phi^2} T_1^{\varphi^1_\epsilon\varphi^2} u \geqslant T_1 u - \epsilon$, Lemma 55(i) implies that

$$\inf_{\pi^2\in\Pi^2} w(\cdot, \varphi^1_\epsilon, \pi^2) \geqslant \rho - \epsilon. \tag{7.10}$$

Since a $\varphi^1_\epsilon \in \Phi^1$ satisfying (7.9) exists for any $\epsilon > 0$, it follows that

$$\sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} w(\cdot, \pi^1, \pi^2) \geqslant \rho. \tag{7.11}$$

Also, since $\sup_{\varphi^1\in\Phi^1} T_1^{\varphi^1\varphi^2_*} u \leqslant T_1 u$, Lemma 55(ii) implies that

$$\sup_{\pi^1\in\Pi^1} w(\cdot, \pi^1, \varphi^2_*) \leqslant \rho. \tag{7.12}$$

Hence

$$\inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} w(\cdot, \pi^1, \pi^2) \leqslant \rho. \tag{7.13}$$

Combining (7.11) and (7.13) with the fact that

$$\sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} w(\cdot, \pi^1, \pi^2) \leqslant \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} w(\cdot, \pi^1, \pi^2),$$

it follows that

$$\rho = \sup_{\pi^1\in\Pi^1} \inf_{\pi^2\in\Pi^2} w(\cdot, \pi^1, \pi^2) = \inf_{\pi^2\in\Pi^2} \sup_{\pi^1\in\Pi^1} w(\cdot, \pi^1, \pi^2) =: w$$

is the value of the average-payoff game. Further, (7.10) implies that player 1 has an ϵ-optimal strategy $\varphi^1_\epsilon \in \Phi^1$ for any $\epsilon > 0$, while (7.12) implies that $\varphi^2_* \in \Phi^2$ is optimal for player 2. □

Lemma 57. *Suppose the conclusions of Proposition 52 hold with a Borel function* $\bar{\mu}$. *Then*

(i) $\bar{r} : \bar{\mathbb{K}} \to \mathbb{R}$ *is bounded and Borel-measurable,*

(ii) $\bar{p}$ *is a Borel-measurable stochastic kernel on* $\bar{\mathbb{X}}$ *given* $\bar{\mathbb{K}}$, *and*

(iii) if Assumption P2 holds, then the stochastic game defined by the AG transformation also satisfies Assumption P2 with $\alpha \equiv \bar{\beta}$.

Proof.

(i) This follows from the boundedness and Borel-measurability of both r and $\bar{\mu}$.

(ii) Fix $(x, a^1, a^2) \in \bar{\mathbb{K}}$. By Proposition 52, $0, \leqslant \bar{p}(B|x, a^1, a^2) \leqslant 1$ for all $B \in \mathcal{B}(\bar{\mathbb{X}})$. Since Lebesgue's monotone convergence theorem implies that the set function

$$B \mapsto \int_{B\setminus\{\ell\}} \mu(y)p(dy|x, a^1, a^2), \qquad B \in \mathcal{B}(\mathbb{X}),$$

is countably additive, it follows that $\bar{p}(\cdot|x, a^1, a^2)$ is a probability measure on $(\bar{\mathbb{X}}, \mathcal{B}(\bar{\mathbb{X}}))$.

Next, fix $B \in \mathcal{B}(\bar{\mathbb{X}})$. Let δ_x denote the Dirac measure on $(\bar{\mathbb{X}}, \mathcal{B}(\bar{\mathbb{X}}))$ sitting at $x \in \bar{\mathbb{X}}$, and let 1_B denote the indicator function on $\bar{\mathbb{X}}$ for $B \in \mathcal{B}(\bar{\mathbb{X}})$. Note that for $(x, a^1, a^2) \in \bar{\mathbb{K}}$,

$$\begin{aligned}\bar{p}(B|x, a^1, a^2) = 1_{\mathbb{X}}(x) \Bigg(\frac{1}{\bar{\beta}\bar{\mu}(x)} \Bigg[\int_{B\setminus\{\ell\}} \bar{\mu}(y)p(dy|x, a^1, a^2) \\ + \left(\bar{\mu}(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y)p(dy|x, a^1, a^2) \right) \delta_\ell(B) \Bigg] + \left(1 - \frac{\bar{\mu}(x) - 1}{\bar{\beta}\bar{\mu}}\right) \delta_{\bar{x}}(B) \Bigg) \\ + 1_{\{\bar{x}\}}(x)\delta_{\bar{x}}(B).\end{aligned}$$

Hence, according to the Borel-measurability of $\bar{\mu}$ and [10, Proposition 7.29], the mapping $(x, a^1, a^2) \mapsto \bar{p}(B|x, a^1, a^2)$ on $\bar{\mathbb{K}}$ is Borel-measurable.

(iii) According to the definition of the AG transformation, $\bar{A}^2(x)$ is compact for all $x \in \bar{\mathbb{X}}$. Further, Assumption P2 and Proposition 52 imply that $\bar{r}$ is bounded on $\bar{\mathbb{K}}$ and $\bar{r}(x, a^1, \cdot)$ is lower semicontinuous on $\bar{A}^2(x)$ for all $x \in \bar{\mathbb{X}}$ and $a^1 \in \bar{A}^1(x)$. Finally, note that for any bounded Borel-measurable $f : \bar{f} : \bar{\mathbb{X}} \to \mathbb{R}$,

$$\begin{aligned}\bar{p}(B|x,a^1,a^2) = 1_{\mathbb{X}}(x)\Bigg(\frac{1}{\bar{\beta}\bar{\mu}(x)}\Bigg[\int_{B\setminus\{\ell\}} f(y)\bar{\mu}(y)p(dy|x,a^1,a^2) \\ + \left(\bar{\mu}(x) - 1 - \int_{\mathbb{X}\setminus\{\ell\}} \bar{\mu}(y)p(dy|x,a^1,a^2)\right) f(\ell)\Bigg] + \left(1 - \frac{\bar{\mu}(x)-1}{\bar{\beta}\bar{\mu}}\right) f(\bar{x})\Bigg) \\ + 1_{\{\bar{x}\}}(x)f(\bar{x});\end{aligned}$$

hence Assumption P2(iii) and [53, Proposition C.4] imply that for each $B \in \mathcal{B}(\bar{\mathbb{X}})$, $x \in \bar{\mathbb{X}}$, and $a^1 \in \bar{A}^1(x)$, the function $\bar{p}(B|x,a^1,\cdot)$ is continuous on $\bar{A}^2(x)$.

□

Theorem 58. *Suppose Assumptions HT, AC, and P2 hold. Then player 1 has an ϵ-optimal stationary strategy for any $\epsilon > 0$, and player 2 has an optimal stationary strategy. Further, the game has a value $w \equiv \bar{v}_{\bar{\beta}}(\ell)$ which, along with $h(x) := \bar{\mu}(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$ for $x \in \mathbb{X}$, satisfies*

$$w + h(x) = T_1 h(x) \tag{7.14}$$

for $x \in \mathbb{X}$.

Proof. By Lemma 57, the discounted stochastic game with $\alpha \equiv \bar{\beta} \in [0,1)$ defined by the AG transformation satisfies Assumption P2. Hence the conclusions of Proposition 39 hold for this game.

In particular, $\bar{v}_{\bar{\beta}} = \bar{T}_{\bar{\beta}}\bar{v}_{\bar{\beta}}$ is a bounded upper semianalytic function on $\bar{\mathbb{X}}$. According to Proposition 54, this implies that for any $\epsilon > 0$ there exist stationary strategies $\varphi^1_\epsilon \in \Phi^1$ and $\varphi^2_* \in \Phi^2$ satisfying

$$\sup_{\varphi^1 \in \Phi^1} \bar{T}_{\bar{\beta}}^{\varphi^1\varphi^2_*}\bar{v}_{\bar{\beta}} \leqslant \bar{T}_{\bar{\beta}}\bar{v}_{\bar{\beta}} = \bar{v}_{\bar{\beta}} \leqslant \inf_{\varphi^2 \in \Phi^2} \bar{T}_{\bar{\beta}}^{\varphi^1_\epsilon\varphi^2}\bar{v}_{\bar{\beta}} + \epsilon. \tag{7.15}$$

By the definitions of the AG transformation, w, and h, it follows from (7.15) that

$$\sup_{\varphi^1 \in \Phi^1} T_1^{\varphi^1\varphi^2_*} h \leqslant w + h \leqslant \inf_{\varphi^2 \in \Phi^2} T_1^{\varphi^1_\epsilon\varphi^2} h + \epsilon. \tag{7.16}$$

But since $\bar{v}_{\bar{\beta}} = \bar{T}_{\bar{\beta}}\bar{v}_{\bar{\beta}}$ implies that $w + h = T_1 h$, and h is bounded by the boundedness of $\bar{r}$, it follows from Proposition 56 that player 1 has an ϵ-optimal strategy φ^1_ϵ for any $\epsilon > 0$, φ^2_* is optimal for player 2, and the average-payoff stochastic game has a value $w \equiv \bar{v}_{\bar{\beta}}(\ell)$. □

7.4.1 Existence of Optimal Strategies

Proposition 59. *Suppose the bounded function* u *on* $\mathbb{X}$ *satisfies*

$$\sup_{x \in \mathbb{X}} [T_1 u(x) - u(x)] = \inf_{x \in \mathbb{X}} [T_1 u(x) - u(x)] =: \rho,$$

and that there exist $\varphi^1_* \in \Phi^1$ *and* $\varphi^2_* \in \Phi^2$ *satisfying*

$$\sup_{\varphi^1 \in \Phi^1} T_1^{\varphi^1 \varphi^2 y} u \leqslant T_1 u \leqslant \inf_{\varphi^2 \in \Phi^2} T_1^{\varphi^1_* \varphi^2} u. \tag{7.17}$$

Then φ^1_* *and* φ^2_* *are optimal for players 1 and 2, respectively, and the average-payoff stochastic game has a value* $w \equiv \rho$.

Proof. This follows from Lemma 55. □

Proposition 60. *Suppose Assumptions HT, P1, and P2 hold. Then there is a Borel-measurable function* $\bar{\mu} : \mathbb{X} \to [1, \infty)$ *that is bounded above by* $\bar{K} := L + 1$ *and satisfies (7.2).*

Proof. This follows from Proposition 48 by considering the stochastic game with transition probabilities q defined by (7.3) and discount function $\alpha := p(\mathbb{X} \setminus \{\ell\} | \cdot)$. □

Lemma 61. *Suppose the conclusions of Proposition 52 hold with a Borel function* $\bar{\mu}$. *If Assumption P1 holds, then the stochastic game defined by the AG transformation also satisfies Assumption P1 with* $\alpha \equiv \bar{\beta}$.

Proof. This follows *mutatis mutandis* from the proof of statement (iii) of Lemma 57. □

Theorem 62. *Suppose Assumptions HT, P1, and P2 hold. Then both players have optimal stationary strategies that are Borel-measurable. Further, the game has a value* $w \equiv \bar{v}_{\bar{\beta}}(\ell)$ *which, along with* $h(x) := \bar{\mu}(x)[\bar{v}_{\bar{\beta}}(x) - \bar{v}_{\bar{\beta}}(\ell)]$ *for* $x \in \mathbb{X}$, *satisfies (7.14).*

Proof. By Lemmas 57 and 61, the discounted stochastic game with $\alpha \equiv \bar{\beta}$ defined by the AG transformation satisfies Assumptions P1 and P2. Hence the conclusions of Proposition 40 hold for this game. In addition, according to [84, Lemmas 4.3, 5.6] there exist Borel-measurable $\varphi^1_* \in \Phi^1$ and $\varphi^2_* \in \Phi^2$ satisfying

$$\sup_{\varphi^1 \in \Phi^1} \bar{T}_{\bar{\beta}}^{\varphi^1 \varphi^2_*} \bar{v}_{\bar{\beta}} \leqslant \bar{T}_{\bar{\beta}} \bar{v}_{\bar{\beta}} \leqslant \inf_{\varphi^2 \in \Phi^2} \bar{T}^{\varphi^1_* \varphi^2} \bar{v}_{\bar{\beta}}.$$

Since $\bar{v}_{\bar{\beta}}$ is bounded, it follows from the definition of the AG transformation and Proposition 59 and that φ_*^1 and φ_*^2 are optimal for players 1 and 2, respectively, and that average-payoff game has a value $w \equiv \bar{v}_{\bar{\beta}}(\ell)$. Further, the definition of the AG transformation and the fact that $\bar{v}_{\bar{\beta}} = \bar{T}_{\bar{\beta}}\bar{v}_{\bar{\beta}}$ imply that (7.14) holds with $w \equiv \bar{v}_{\bar{\beta}}(\ell)$ and h. □

7.5 Complexity Estimates

In this section, we provide complexity estimates related to applying the AG transformation for stochastic games. In Section 7.5.1, we provide an upper bound on the number of arithmetic operations needed to compute a function $\bar{\mu}$ for the AG transformation. Then, in Section 7.5.2 we provides estimates for the number of arithmetic operations needed to compute a pair of optimal strategies for two-player zero-sum average-payoff stochastic games with perfect information.

7.5.1 Constructing the Transformation

Note that, given a suitable function $\bar{\mu}$, the two-player zero-sum stochastic game defined by the AG transformation can be constructed with a number of arithmetic operations that is polynomial in the total number of state-action triples m of the original game. The following theorem provides an estimate of the complexity of computing a function $\bar{\mu}$ that can be used for the AG transformation.

Theorem 63. *Suppose the state set $\mathbb{X}$ and action sets $\mathbb{A}^i$, $i = 1, 2$, are finite, and that Assumption HT holds. Then the number of arithmetic operations needed to compute a function μ satisfying the hypotheses of Proposition 52 is at most a constant times $m\bar{K}\log\bar{K}$, where $\bar{K} := L + 1$.*

Proof. To compute a function satisfying the hypotheses of Proposition 52, it suffices to compute a bounded nonnegative function $\bar{\mu}$ that satisfies

$$\bar{\mu}(x) = \max_{(a^1,a^2)\in A^1(x)\times A^2(x)} \left[1 + \sum_{y\in\mathbb{X}\setminus\{\ell\}} p(y|x,a^1,a^2)\bar{\mu}(y)\right], \quad x \in \mathbb{X}. \qquad (7.18)$$

Let

$$q(y|x,a^1,a^2) := 1_{\mathbb{X}\setminus\{\ell\}}(y)p(y|x,a^1,a^2)$$

for $x, y \in \mathbb{X}$ and $(a^1, a^2) \in A^1(x) \times A^2(x)$, and consider the Markov decision process with state set $\mathbb{X}$, action sets $A(x) := A^1(x) \times A^2(x)$ for $x \in \mathbb{X}$, transition rates $q(y|x, a)$ for $x, y \in \mathbb{X}$ and $a \in A(x)$, and one-step rewards identically equal to one. According to Assumption HT, this MDP is transient; see [23, Hypothesis 1]. Hence it follows from [23, Theorem 2] that the number of arithmetic operations needed, to compute a nonnegative function that is bounded above by $\bar{K} := L + 1$ and satisfies (7.18), is at most a constant times $m\bar{K} \log \bar{K}$. □

7.5.2 Computing Optimal Strategies

Theorem 64. *Suppose the state set $\mathbb{X}$ and action sets $\mathbb{A}^i$, $i = 1, 2$, are finite, and that Assumptions HT and PI hold. Then both players have optimal deterministic stationary strategies, and the number of arithmetic operations needed to compute a pair of such strategies is at most a constant times*

$$\left((n_1^3 + n_1^2 m_1) m_1 \bar{K} \log \bar{K} + n^3 + m_2 n_2^2\right) \cdot m\bar{K} \log n\bar{K},$$

where $\bar{K} := L + 1$.

Proof. Recall from the proof of Theorem 51 that for any discount factor $\beta \in (0, 1)$, the total number of arithmetic operations needed to compute a pair of β-optimal deterministic strategies is at most a constant times

$$\left((n_1^3 + n_1^2 m_1)\frac{m_1}{1-\beta} \log \frac{1}{1-\beta} + n^3 + m_2 n_2^2\right) \cdot \frac{m}{1-\beta} \log \frac{n}{1-\beta}. \tag{7.19}$$

According to Theorem 34, the number of arithmetic operations needed to compute a function μ that is bounded above by $\bar{K} := L + 1 < \infty$ for the AG transformation is at most a constant times $m\bar{K} \log \bar{K}$. In addition, with $\beta := (\bar{K} - 1)\bar{K}^{-1}$, the number of arithmetic operations needed to compute a pair of optimal deterministic stationary strategies for the resulting discounted stochastic game is at most a constant times (7.19). Hence the total number of arithmetic operations needed is at most a constant times

$$m\bar{K} \log \bar{K} + \left((n_1^3 + n_1^2 m_1) m_1 \bar{K} \log \bar{K} + n^3 + m_2 n_2^2\right) \cdot m\bar{K} \log n\bar{K},$$

and the theorem follows from Theorem 62 and Proposition 53. □

Bibliography

[1] M. Akian and S. Gaubert. Policy iteration for perfect information stochastic mean payoff games with bounded first return times is strongly polynomial. *arXiv:1310.4953 [math]*, October 2013. arXiv: 1310.4953.

[2] O. Alagoz, M. U. S. Ayvaci, and J. T. Linderoth. Optimally solving Markov decision processes with total expected discounted reward function: Linear programming revisited. *Computers & Industrial Engineering*, 87:311–316, September 2015.

[3] C. D. Aliprantis and K. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, August 2006.

[4] N. Amenta and G. M. Ziegler. Deformed products and maximal shadows of polytopes. *Contemporary Mathematics*, 223:57–90, 1999.

[5] W. F. Ames and M. Ginsberg. Bilateral algorithms and their applications. In *Computational Mechanics*, pages 1–31. Springer, 1975.

[6] D. Andersson, T. D. Hansen, and P. B. Miltersen. Toward better bounds on policy iteration. *Preprint (June 2009)*, 2009.

[7] D. Andersson and P. B. Miltersen. The Complexity of Solving Stochastic Games on Graphs. In Yingfei Dong, Ding-Zhu Du, and Oscar Ibarra, editors, *Algorithms and Computation*, number 5878 in Lecture Notes in Computer Science, pages 112–121. Springer Berlin Heidelberg, December 2009. DOI: 10.1007/978-3-642-10631-6_13.

[8] D. Andersson and S. Vorobyov. Fast algorithms for monotonic discounted linear programs with two variables per inequality. *Preprint*

NI06019-LAA, Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, 2006.

[9] A. Arapostathis, V. Borkar, E. Fernndez-Gaucherand, M. Ghosh, and S. Marcus. Discrete-Time Controlled Markov Processes with Average Cost Criterion: A Survey. *SIAM Journal on Control and Optimization*, 31(2):282–344, March 1993.

[10] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, 1978.

[11] Dimitri P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, 2013.

[12] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-dynamic Programming*. Athena Scientific, 1996.

[13] R. G. Bland. New finite pivoting rules for the simplex method. *Mathematics of operations Research*, 2(2):103–107, 1977.

[14] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer Science & Business Media, 1998.

[15] L. Blum, M. Shub, and S. Smale. On a theory of computation and complexity over the real numbers: - completeness, recursive functions and universal machines. *Bulletin of the American Mathematical Society*, 21(1):1–46, 1989.

[16] A. Cobham. The Intrinsic Computational Difficulty of Functions. In *Proceedings of the Third International Congress for Logic, Methodology and Philosophy of Science*, Amsterdam, 1965. North-Holland.

[17] A. Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, February 1992.

[18] A. Condon. On Algorithms for Simple Stochastic Games. In *Advances in Computational Complexity Theory, volume 13 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 51–73. American Mathematical Society, 1993.

[19] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6. ACM, 1987.

[20] W. H. Cunningham. Theoretical Properties of the Network Simplex Method. *Mathematics of Operations Research*, 4(2):196–208, May 1979.

[21] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

[22] E. V. Denardo. Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Review*, 9(2):165–177, April 1967.

[23] E. V. Denardo. Nearly strongly polynomial algorithms for transient dynamic programs. *Preprint*, February 2016.

[24] C. Derman. Optimal replacement and maintenance under Markovian deterioration with probability bounds on failure. *Management Science*, 9(3):478–481, 1963.

[25] E. B. Dynkin and A. A. Yushkevich. *Controlled Markov Processes*. Springer New York, 1979.

[26] B. C. Eaves and A. F. Veinott. Maximum-Stopping-Value Policies in Finite Markov Population Decision Chains. *Mathematics of Operations Research*, 39(3):597–606, January 2014.

[27] J. Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17(3):449–467, 1965.

[28] A. Ehrenfeucht and J. Mycielski. Positional strategies for mean payoff games. *International Journal of Game Theory*, 8(2):109–113, June 1979.

[29] E. A. Emerson and C. S. Jutla. Tree automata, μ-calculus and determinacy. In *, 32nd Annual Symposium on Foundations of Computer Science, 1991. Proceedings*, pages 368–377, October 1991.

[30] K. Etessami and M. Yannakakis. Recursive Markov Decision Processes and Recursive Stochastic Games. *J. ACM*, 62(2):11:1–11:69, May 2015.

[31] G. Even and A. Zadorojniy. Strong polynomiality of the Gass-Saaty shadow-vertex pivoting rule for controlled random walks. *Annals of Operations Research*, 201(1):159–167, August 2012.

[32] J. Fearnley. Exponential Lower Bounds for Policy Iteration. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *Automata, Languages and Programming*, number 6199 in Lecture Notes in Computer Science, pages 551–562. Springer Berlin Heidelberg, July 2010. DOI: 10.1007/978-3-642-14162-1_46.

[33] A. Federgruen, A. Hordijk, and H. C. Tijms. *Recurrence conditions in denumerable state Markov decision processes*. Stichting Math. Centrum, 1977.

[34] E. A. Feinberg. Controlled Markov Processes with Arbitrary Numerical Criteria. *Theory of Probability & Its Applications*, 27(3):486–503, January 1983.

[35] E. A. Feinberg. Total reward criteria. In *Handbook of Markov decision processes*, pages 173–207. Springer, 2002.

[36] E. A. Feinberg and J. Huang. Strong polynomiality of policy iterations for average-cost MDPs modeling replacement and maintenance problems. *Operations Research Letters*, 41(3):249–251, May 2013.

[37] E. A. Feinberg and J. Huang. The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, 42(2):130–131, March 2014.

[38] E. A. Feinberg and J. Huang. On the Reduction of Total-Cost and Average-Cost MDPs to Discounted MDPs. *arXiv:1507.00664 [math]*, July 2015. arXiv: 1507.00664.

[39] E. A. Feinberg, J. Huang, and B. Scherrer. Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, 42(67):429–431, September 2014.

[40] E. A. Feinberg, P. O. Kasyanov, and M. Voorneveld. Berges maximum theorem for noncompact image sets. *Journal of Mathematical Analysis and Applications*, 413(2):1040–1046, May 2014.

[41] E. A. Feinberg, P. O. Kasyanov, and N. V. Zadoianchuk. Average Cost Markov Decision Processes with Weakly Continuous Transition Probabilities. *Mathematics of Operations Research*, 37(4):591–607, September 2012.

[42] E. A. Feinberg, P. O. Kasyanov, and M. Zgurovsky. Convergence of probability measures and Markov decision models with incomplete information. *Proceedings of the Steklov Institute of Mathematics*, 287(1):96–117, January 2015.

[43] E. A. Feinberg and A. Shwartz. *Handbook of Markov Decision Processes: Methods and Applications*. Springer Science & Business Media, December 2012.

[44] E. A. Feinberg and F. Yang. On polynomial cases of the unichain classification problem for Markov Decision Processes. *Operations Research Letters*, 36(5):527–530, September 2008.

[45] O. Friedmann, T. D. Hansen, and U. Zwick. A Subexponential Lower Bound for the Random Facet Algorithm for Parity Games. In *Proceedings of the Twenty-second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, pages 202–216, Philadelphia, PA, USA, 2011. Society for Industrial and Applied Mathematics.

[46] O. Friedmann, T. D. Hansen, and U. Zwick. Subexponential Lower Bounds for Randomized Pivoting Rules for the Simplex Algorithm. In *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, STOC '11, pages 283–292, New York, NY, USA, 2011. ACM.

[47] O. Friedmann. A subexponential lower bound for the least recently considered rule for solving linear programs and games. In *The Annual Workshop of the ESF Networking Programme on Games for Design and Verification, GAMES*, 2012.

[48] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, December 1988.

[49] L. G. Gubenko and E. S. Štatland. On controlled, discrete-time Markov decision processes. *Theory Probab. Math. Statist*, 7:47–61, 1975.

[50] T. D. Hansen, H. Kaplan, and U. Zwick. Dantzig's Pivoting Rule for Shortest Paths, Deterministic MDPs, and Minimum Cost to Time Ratio Cycles. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 847–860, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics.

[51] T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy Iteration Is Strongly Polynomial for 2-Player Turn-Based Stochastic Games with a Constant Discount Factor. *J. ACM*, 60(1):1:1–1:16, February 2013.

[52] O. Hernández-Lerma. *Adaptive Markov Control Processes*. Springer Singapore Pte. Limited, 1989.

[53] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer New York, December 1995.

[54] O. Hernández-Lerma and J. B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer Science & Business Media, 1998.

[55] O. Hernández-Lerma, R. Montes-De-Oca, and R. Cavazos-Cadena. Recurrence conditions for Markov decision processes with Borel state space: A survey. *Annals of Operations Research*, 28(1):29–46, December 1991.

[56] C. J. Himmelberg, T. Parthasarathy, and F. S. Van Vleck. Optimal Plans for Dynamic Programming Problems. *Mathematics of Operations Research*, 1(4):390–394, November 1976.

[57] A. J. Hoffman and R. M. Karp. On Nonterminating Stochastic Games. *Management Science*, 12(5):359–370, January 1966.

[58] R. Hollanders, J. Delvenne, and R.M. Jungers. The complexity of Policy Iteration is exponential for discounted Markov Decision Processes. In *2012 IEEE 51st Annual Conference on Decision and Control (CDC)*, pages 5997–6002, December 2012.

[59] R. Hollanders, B. Gerencsér, J. C. Delvenne, and R. M. Jungers. Improved bound on the worst case complexity of Policy Iteration. *Operations Research Letters*, 44(2):267–272, March 2016. bibtex: hollanders_improved_2016-1.

[60] A. Hordijk. *Dynamic Programming and Markov Potential Theory*. Mathematisch Centrum, 1974.

[61] A. Hordijk and A. A. Yushkevich. Blackwell optimality. In *Handbook of Markov decision processes*, pages 231–267. Springer, 2002.

[62] R. A. Howard. *Dynamic Programming and Markov Processes*. Published jointly by the Technology Press of the Massachusetts Institute of Technology and, 1960.

[63] J. Jacod and P. Protter. *Probability Essentials*. Springer Science & Business Media, December 2012.

[64] L. C. M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Mathematisch Centrum, 1980.

[65] N. Karmarkar. A New Polynomial-time Algorithm for Linear Programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing*, STOC '84, pages 302–311, New York, NY, USA, 1984. ACM.

[66] R. M. Karp. A characterization of the minimum cycle mean in a digraph. *Discrete Mathematics*, 23(3):309–311, January 1978.

[67] A. Kechris. *Classical Descriptive Set Theory*. Springer New York, January 1995.

[68] L. G. Khachiyan. A polynomial algorithm in linear programming. *Doklady Akademii Nauk SSSR*, 244:1093–1096, 1979. MSC2010: 90C31 = Sensitivity, stability, parametric optimization MSC2010: 90C05 = Linear programming MSC2010: 68Q25 = Analysis of algorithms and problem complexity MSC2010: 65K05 = Mathematical programming (numerical methods).

[69] Y. H. Kim and L. C. Thomas. Repair Strategies in an Uncertain Environment: Stochastic Game Approach. In Tadashi Dohi and Toshio

Nakagawa, editors, *Stochastic Reliability and Maintenance Modeling*, number 9 in Springer Series in Reliability Engineering, pages 123–140. Springer London, 2013. DOI: 10.1007/978-1-4471-4971-2_6.

[70] T. Kitahara and S. Mizuno. A bound for the number of different basic solutions generated by the simplex method. *Mathematical Programming*, 137(1-2):579–586, August 2011.

[71] V. Klee and G. J. Minty. How good is the simplex algorithm? In *Inequalities, III (Proc. Third Sympos., Univ. California, Los Angeles, Calif., 1969; dedicated to the memory of Theodore S. Motzkin)*, pages 159–175. Academic Press, New York, 1972.

[72] M. Klein. Inspection-maintenance-replacement schedules under Markovian deterioration. *Management Science*, 9(1):25–32, 1962.

[73] A. N. Kolmogorov and N. A. Dmitriev. Branching stochastic processes. *Doklady Akademii Nauk SSSR*, 56:5–8, 1947.

[74] M. L. Littman, T. L. Dean, and L. P. Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 394–402, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[75] O. Madani, M. Thorup, and U. Zwick. Discounted deterministic Markov decision processes and discounted all-pairs shortest paths. *ACM Transactions on Algorithms (TALG)*, 6(2):33, 2010.

[76] Y. Mansour and S. Singh. On the complexity of policy iteration. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 401–408. Morgan Kaufmann Publishers Inc., 1999.

[77] J. Matoušek, M. Sharir, and E. Welzl. A subexponential bound for linear programming. *Algorithmica*, 16(4-5):498–516, 1996.

[78] U. Meister and U. Holzbaur. A polynomial time bound for Howard's policy improvement algorithm. *Operations-Research-Spektrum*, 8(1):37–40, March 1986.

[79] M. Melekopoglou and A. Condon. On the complexity of the policy improvement algorithm for Markov decision processes. *ORSA Journal on Computing*, 6(2):188–192, May 1994.

[80] B. L. Miller and A. F. Veinott. Discrete dynamic programming with a small interest rate. *The Annals of Mathematical Statistics*, 40(2):366–370, 1969.

[81] H. Mine and S. Osaki. *Markovian Decision Processes*. American Elsevier Pub. Co., 1970.

[82] C. J. Mode. *Multitype Branching Processes: Theory and Applications*. American Elsevier Pub. Co., 1971.

[83] J. R. Munkres. *Topology*. Prentice Hall, Incorporated, 2000.

[84] A. S. Nowak. On zero sum stochastic games with general state space. i. *Probability and Mathematical Statistics*, 4(1):13–32, 1984.

[85] A. S. Nowak. Universally measurable strategies in zero-zum stochastic games. *The Annals of Probability*, 13(1):269–287, 1985.

[86] M. O'Sullivan. *New Methods for Dynamic Programming over an Infinite Time Horizon*. PhD thesis, Stanford University, 2002.

[87] C. H. Papadimitriou and K. Steiglitz. *Combinatorial Optimization: Algorithms and Complexity*. Courier Corporation, 1982.

[88] S. Patek and D. P. Bertsekas. Stochastic shortest path games. *SIAM Journal on Control and Optimization*, 37(3):804–824, January 1999.

[89] S. R. Pliska. Optimization of multitype branching processes. *Management Science*, 23(2):117–124, 1976.

[90] S. R. Pliska. On the transient case for Markov decision chains with general state spaces. *Dynamic Programming and Its Applications,(ML Puterman, Ed.). New York: Springer*, 1978.

[91] I. Post and Y. Ye. The simplex method is strongly polynomial for deterministic Markov decision processes. *Mathematics of Operations Research*, 40(4):859–868, February 2015.

[92] M. L. Puterman and M. Shin. Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, 24(11):1127–1137, July 1978.

[93] S. S. Rao, R. Chandrasekaran, and K. P. K. Nair. Algorithms for discounted stochastic games. *Journal of Optimization Theory and Applications*, 11(6):627–637, 1973.

[94] S. M. Ross. Arbitrary state Markovian decision processes. *The Annals of Mathematical Statistics*, 39(6):2118–2122, 1968.

[95] S. M. Ross. Non-discounted denumerable Markovian decision models. *The Annals of Mathematical Statistics*, 39(2):412–423, 1968.

[96] S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[97] U. G. Rothblum and A. F. Veinott. Markov Branching Decision Chains: Immigration-Induced Optimality. Technical Report, 1992.

[98] U. G. Rothblum and P. Whittle. Growth optimality for branching Markov decision chains. *Mathematics of Operations Research*, 7(4):582–601, 1982.

[99] N. Saldi, T. Linder, and S. Yüksel. Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control. *IEEE Transactions on Automatic Control*, 60(2):553–558, February 2015.

[100] N. Saldi, S. Yüksel, and T. Linder. Asymptotic optimality of finite approximations to Markov decision processes with Borel spaces. *arXiv:1503.02244 [cs, math]*, March 2015. arXiv: 1503.02244.

[101] N. Saldi, S. Yüksel, and T. Linder. Near optimality of quantized policies in stochastic control under weak continuity conditions. *Journal of Mathematical Analysis and Applications*, 435(1):321–337, March 2016.

[102] M. Santos and J. Rust. Convergence properties of policy iteration. *SIAM Journal on Control and Optimization*, 42(6):2094–2115, January 2004.

[103] B. Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41(3):758–774, February 2016.

[104] Alexander Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer Science & Business Media, December 2002.

[105] L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, October 1953.

[106] A. N. Shiryaev. *Probability-1, Graduate Texts in Mathematics*. Springer New York, New York, NY, 2016.

[107] M. Sipser. *Introduction to the Theory of Computation*. Cengage Learning, June 2012.

[108] S. Smale. Mathematical problems for the next century. *The Mathematical Intelligencer*, 20(2):7–15, January 2009.

[109] E. Tardos. Strongly polynomial and combinatorial algorithms in optimization. In *Proceedings of the International Congress of Mathematicians: August 21-29, 1990, Kyoto, Japan : ICM-90 Kyoto 2. 2.*, pages 1467–1478. Springer, 1991.

[110] L. C. Thomas. *Connectedness conditions for denumerable state Markov decision processes*. University of Manchester. Department of Decision Theory, 1978.

[111] M. J. Todd. The many facets of linear programming. *Mathematical Programming*, 91(3):417–436, 2002.

[112] P. Tseng. Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$. *Operations Research Letters*, 9(5):287–297, September 1990.

[113] J. van der Wal. *Stochastic Dynamic Programming: Successive Approximations and Nearly Optimal Strategies for Markov Decision Processes and Markov Games*. Mathematisch Centrum, 1984.

[114] K. M. van Hee and J. Wessels. Markov decision processes and strongly excessive functions. *Stochastic Processes and their Applications*, 8(1):59–76, 1978.

[115] A. F. Veinott. Discrete dynamic programming with sensitive discount optimality criteria. *The Annals of Mathematical Statistics*, 40(5):1635–1660, 1969.

[116] A. F. Veinott. Markov Decision Chains. Technical Report, DTIC Document, 1973.

[117] A. F. Veinott. Lectures in Dynamic Programming and Stochastic Control. Course Notes, Stanford University, 2008.

[118] J. Wessels. Markov programming by successive approximations with respect to weighted supremum norms. *Journal of mathematical analysis and applications*, 58(2):326–335, 1977.

[119] V. V. Williams. Multiplying matrices faster than Coppersmith-Winograd. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 887–898, New York, NY, USA, 2012. ACM.

[120] Y. Ye. A new complexity result on solving the Markov decision problem. *Mathematics of Operations Research*, 30(3):733–749, August 2005.

[121] Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, November 2011.

[122] N. Zadeh. What is the worst case behavior of the simplex algorithm? Technical Report 27, Stanford University, Department of Operations Research, 1980.

[123] A. Zadorojniy, G. Even, and A. Shwartz. A strongly polynomial algorithm for controlled queues. *Mathematics of Operations Research*, 34(4):992–1007, October 2009.