Reinforcement Learning in Large or Unknown MDPs

Ambuj Tewari

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-126
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-126.html

October 25, 2007

Reinforcement Learning in Large or Unknown MDPs

by

Ambuj Tewari

B.Tech. (Indian Institute of Technology, Kanpur) 2002
M.A. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science
and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Peter L. Bartlett, Chair
Professor Stuart J. Russell
Professor Peter J. Bickel

Fall 2007

The dissertation of Ambuj Tewari is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2007

Reinforcement Learning in Large or Unknown MDPs

Ambuj Tewari

Abstract

Reinforcement Learning in Large or Unknown MDPs

by

Ambuj Tewari

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Peter L. Bartlett, Chair

Reinforcement learning is a central problem in artificial intelligence. Unlike supervised learning, in reinforcement learning the learning agent does not receive any input from a supervisor about what to do in different situations. The agent has to learn from its own experience taking into account any uncertainty in the outcomes of its actions.

Markov Decision Processes (MDPs) have been the dominant formalism used to mathematically state and investigate the problem of reinforcement learning. Classical algorithms like value iteration and policy iteration can compute optimal policies for MDPs in time polynomial in the description of the MDP. This is fine for small problems but makes it impractical to apply these algorithms to real world MDPs where the number of states is enormous, even infinite. Another drawback is that these algorithms assume that the MDP parameters are precisely known. To quantify learning in an unknown MDP, the notion of regret has been defined and studied in the literature.

This dissertation consists of two parts. In the first part, we study two methods that have been proposed to handle large MDPs. PEGASUS is a policy search method that uses simulators and approximate linear programming is a general methodology that tries to obtain a good policy by solving linear programs of reasonable size. We give performance bounds for policies produced by these methods. In the second part, we study the problem of learning an unknown MDP. We begin by considering bounded parameter MDPs. These arise when we have confidence intervals associated with each MDP parameter. Finally, we give a new algorithm that achieves logarithmic regret in an irreducible but otherwise unknown MDP. This is a provably optimal rate up to a constant.

Professor Peter L. Bartlett
Dissertation Committee Chair

To the memory of my grandmother,

Shrimati Vimla Tewari

मेरी पूज्य दादीजी

श्रीमती विमला तिवारी

की स्मृति में

Contents

Acknowledgments

It is a pleasure to acknowledge the help and support I have received during the time the research described in this dissertation was conducted.

My advisor, Professor Peter Bartlett, has been a great mentor and collaborator for the past four years. His knowledge, patience and the willingness to work with his students down to the minutest details of proofs have always amazed me. I could not have done this research without his constant support and encouragement. I also thank my dissertation committee, Professor Peter Bickel and Professor Stuart Russell, for their comments on earlier drafts of this dissertation.

Members of Peter's research group: Jacob Abernethy, Alekh Agarwal, Wei-Chun Kao, Alexander Rakhlin, David Rosenberg, Benjamin Rubinstein and Mikhail Traskin, provided a stimulating research environment through their ideas and discussions. I am also indebted to Krishnendu Chatterjee, Ankit Jain, Vishnu Narayanan and XuanLong Nguyen for all those numerous occasions when I would describe my own research and also listen to what they were doing.

I have also benefited from meetings with Peter Auer, Nicolò Cesa-Bianchi, Sham Kakade, Csaba Szepesvári and Bin Yu.

Life in Berkeley would not have been fun without these wonderful friends: Ankit, Gautam, Jake, Jaydeep, Koushik, Kranthi, Krishnendu, Long, Mani, Marghoob, Puneet, Sasha, Saurabh, Shankar, Shariq, Sonesh, Sourav, Sumit and Vishnu. Special thanks to Anurag, my roommate for the past five years. I have more things to thank him for than this short space will allow. If I were to pick just one of them, it would be our countless

discussions on history, philosophy, literature, science and life in general.

My father, mother and brother, though living far away in India, have always been close to me in my heart. Mere words cannot express my gratitude to them.

Chapter 1

Introduction

Reinforcement learning is learning how to behave without having a teacher or supervisor tell which actions to take in given situations. The learning agent interacts with the environment collecting rewards as it tries out different actions in different situations. This experience should help it decide which actions are better long term consequences in terms of maximizing the rewards obtained.

Reinforcement learning problems are often modeled by Markov Decision Processes (MDPs). We give a precise definition in the next chapter but for now it is sufficient to know that there are four components to an MDP:

1. A state space S: This could be finite or infinite. Even if finite, the number of elements in the state space is huge for any practical problem.

2. An action space A: This is the set of actions available to the agent. It can also depend on the state the agent is in. Then A becomes a function of state. Again, A could be finite or infinite.

3. A reward function R: This is the numerical reward that the agent receives in the course of its interaction with the environment. We assume that the agent receives a reward $R(i, a)$ when it takes action a in state i. Sometimes, the reward is also a function of the state that is reached after taking the action. But we will follow the above convention.

4. The transition function P: This describes how the state of the agent changes when it takes various actions. Since there is uncertainty in outcomes of actions, the precise outcome of taking action a in state i is not determined. Instead, there is a certain probability that the agent will end up in state j as a result of this action. We denote this probability by $P_{i,a}(j)$.

A crucial concept in the theory of MDPs is that of a policy. A policy provides the agent with a way to map situations onto actions. To assess the quality of a policy, various performance criteria have been considered in the literature. The next chapter describes two of the most popular ones: the discounted sum of rewards and the average reward criteria. When the parameters R and P are given explicitly, there are standard algorithms to compute optimal policies. The running time of these algorithms typically scales polynomially in the sizes $|S|$ and $|A|$ of the state and action spaces. For MDPs with large state spaces, these algorithms are impractical. Another issue is that we often do not have direct access to these parameters. Instead, we might be faced with an unknown MDP and the only way to obtain information about the parameters is through interaction with the environment. This brings the famous 'exploration versus exploitation' issue to the fore. On the one hand, the agent has to take seemingly inferior actions to explore previously unvisited parts of the

state space. On the other hand, it is necessary to exploit past experience by executing those actions that have performed well so far.

This thesis is divided into two parts. The first part, consisting of Chapters 3 and 4, deals with the problem of large state spaces. Chapters 5 and 6 form the second part and deal with the problem of acting in an unknown MDP. We give a brief summary of the contributions of these chapters below.

Chapter 3 considers PEGASUS, a policy search method due to Ng and Jordan [23]. In policy search, we fix a class of policies and try to find the best policy in that class. The hope is that by giving up the goal of producing the truly optimal policy, we have made the problem tractable. PEGASUS involves the use of a simulator for the underlying MDP. Such simulators are often available for many tasks and PEGASUS has been successfully applied to autonomous helicopter flight. We provide bounds on sample complexity, i.e. the number of sample trajectories that need to be generated during the simulation in order to achieve a given level of performance. This chapter also illustrates that the issue of sample complexity is not straightforward since bounds cannot be obtained by just assuming that the policy class and the MDP dynamics are simple according to the usual notions of complexity like Pollard's pseudodimension or the fat-shattering dimension.

Chapter 4 considers the approximate linear programming approach pioneered by Van Roy and his colleagues. The idea behind this approach is to approximate the value function (a measure of performance for a given policy) by a linear combination of a set of basis functions. This approximation is then plugged into the standard linear program for discounted MDPs. The hope is that the computation will then scale with the number of ba-

sis functions and not with the size of the state space. Obviously, making the approximation means the policy thus obtained will be sub-optimal. We obtain a bound that quantifies the difference in the performance of the policy derived from approximate linear programming with constraint sampling and that of an optimal policy. This chapter also considers approximation in the dual linear program which has begun to be explored in the recent literature. We provide some preliminary results about the solution of the approximate dual. We hope that in future this will allow us to obtain more understanding of the power of the dual approximation approach.

Chapter 5 considers bounded parameter MDPs (BMDPs). Among other situations, they arise when we are exploring an unknown MDP. Due to limited experience, we do not have precise estimates of the parameters R and P. Instead, each parameter has a confidence interval associated with it. A BMDP is simply a set of MDPs whose parameters lie in given intervals. There are at least two ways to define the value function of a given policy in a BMDP. The optimistic value function of a policy is its value in the *best* MDP for that policy. The pessimistic value function is its value in the *worst* MDP for that policy. We give algorithms to compute the optimal value function in the optimistic as well the pessimistic sense.

The recent online reinforcement learning algorithm of Auer and Ortner [5] needs to compute an optimal policy (in the optimistic sense) for a BMDP obtained from the current confidence intervals. The optimal policy is then executed for some time, the confidence intervals are then updated and the process repeats. Thus, our algorithm for BMDPs can be used as a subroutine of their algorithm. In the robust control literature, Nilim and El

Ghaoui [24] considered pessimistic value functions and gave algorithms to compute optimal value functions and policies for the finite horizon and discounted criteria. Our work can thus also be seen as extending their work to the case of the average reward criterion.

Chapter 6 considers the problem of minimizing regret in an unknown MDP. The definition of regret can be found in Section 6.2 of that chapter. It measures the difference between the accumulated reward of the agent and that of an optimal policy. Intuitively, low regret algorithms can be thought of as making a successful exploration-exploitation trade-off. It is remarkable that Burnetas and Katehakis [11] have given an algorithm following which the agent can ensure that its regret will only increase logarithmically with time. Their algorithm used KL-divergence in its main step and as a result needed to know the support of the distributions $P_{i,a}$. We simplify their algorithm by using L_1 distance instead of KL-divergence. Our use of the L_1 distance not only makes the algorithm and its analysis simpler but also gets rid of the requirement to know the support of the distributions $P_{i,a}$. Recently, Auer and Ortner [5] have also given a logarithmic regret algorithm that works without this knowledge. We provide a comparison of the constant appearing in our bound with the one appearing in theirs.

Chapter 2

Markov Decision Processes

A Markov decision process is a model for sequential decision making under uncertainty. Mathematically, a Markov decision process (MDP) is a 4-tuple $M = \langle S, A, P, R \rangle$ where S is the *state space*, A is the *action space*, P is the *transition function* and R is the *reward function*. For every $i \in S$, $a \in A$, $P_{i,a}$ is probability distribution over S and $P_{i,a}(j)$ represents the probability of moving to state j after taking action a in state i. Also, a reward of $R(i, a)$ is received when action a is taken in state i. A good reference for the theory of MDPs is the Puterman's book [26].

The fundamental issue in making sequential decisions is that the current decision can have long-term consequences. The reward $R(i, a)$ is a short term measure of the value of taking action a in state i. In order to discuss long-term values, we need to introduce the notion of a policy. Intuitively, a policy is a rule for making a decision based on what has happened in the past. A *history* is a sequence $\sigma_t = (i_0, k_0, \ldots, i_{t-1}, k_{t-1}, i_t)$ such that $i_l \in S, 0 \leq l \leq t$ and $k_l \in A, 0 \leq l < t$. A history dependent policy π is a sequence

$\{\pi_t\}$ of probability distributions on A given σ_t. We denote the set of all policies by Π. A randomized *stationary* policy only depends on the current state and can thus be thought of as a map $\mu : S \times A \mapsto [0,1]$ such that $\sum_a \mu(i,a) = 1$ for all $i \in S$. When in states i, the policy μ takes action i with probability $\mu(i,a)$. We denote the set of all randomized stationary policies by Π_R. A *deterministic stationary* policy is simply a mapping $\mu : S \mapsto A$. We denote the set of all deterministic stationary policies by Π_D.

Given an MDP M, a start state i_0 and a policy π, we can define a natural stochastic process $s_0, a_0, s_1, a_1, \ldots$ where $s_0 = i_0$, a_t is drawn from the distribution π_t given σ_t and s_{t+1} is drawn from P_{s_t,a_t}. Denote the expectations and probabilities with respect this process by $\mathbb{E}_{i_0}^{\pi,M}$ and $\mathbb{P}_{i_0}^{\pi,M}$ respectively. If the start state s_0 is not fixed but is instead chosen from a distribution α over states, we denote the corresponding expectations and probabilities by $\mathbb{E}_{\alpha}^{\pi,M}$ and $\mathbb{P}_{\alpha}^{\pi,M}$ respectively.

2.1 Discounted Reward Criterion

Given an MDP M, a discount factor $\gamma \in [0,1)$ and policy π define the γ-discounted *value* of a policy π starting from state i to be

$$V_{\gamma,\pi,M}(i) := \mathbb{E}_i^{\pi,M}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] .$$

If rewards are in some bounded interval, say $[0, R_{\max}]$, then the above limit exists for each policy $\pi \in \Pi$. Define the *optimal value function* by

$$V_{\gamma,M}^*(i) := \sup_{\pi \in \Pi} V_{\gamma,\pi,M}(i) .$$

It is a fact from MDP theory that there is a policy $\mu^* \in \Pi_D$ such that

$$V^*_{\gamma,M} = V_{\gamma,\mu^*,M} \ .$$

We call such a policy *optimal* for the γ-discounted reward criterion. When the MDP M and discount factor γ are fixed, we will denote $V_{\gamma,\mu,M}$ and $V^*_{\gamma,M}$ simply by V_μ and V^* respectively.

Given a policy $\mu \in \Pi_D$, define the matrix P_μ and the vector R_μ by

$$P_\mu(i,j) := P_{i,\mu(i)}(j) \ , \qquad\qquad R_\mu(i) := R(i,\mu(i)) \ .$$

Define the *occupation measure* of μ with initial distribution α as

$$\psi_{\mu,\alpha}(i) := \mathbb{E}^{\mu,M}_\alpha \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{1}\left[s_t = i\right] \right] \ .$$

Note that we also have

$$\psi_{\mu,\alpha} = \alpha^\top (I - \gamma P_\mu)^{-1} \ .$$

The occupation measure can be extended to state-action pairs as follows.

$$\psi_{\mu,\alpha}(i,a) := \psi_\mu(i)\mu(i,a) \ .$$

It will usually be clear from context as to which occupation measure is meant.

Define the following operators that act on $V : S \to \mathbb{R}$ producing a $V' : S \to \mathbb{R}$.

$$\left(T_{\gamma,\mu,M}V\right)(i) := R(i,a) + \gamma \sum_j P_{i,\mu(i)}(j)V(j) \ ,$$

$$\left(T_{\gamma,M}V\right)(i) := \max_{a \in A} \left[R(i,a) + \gamma \sum_j P_{i,a}(j)V(j) \right] \ .$$

MDP theory tells us that the fixed point of these operators are $V_{\gamma,\mu,M}$ and $V^*_{\gamma,M}$ respectively. Again, we omit the subscripts M and γ if those quantities are fixed in a given context.

The *greedy policy* with respect to a function $V : S \to \mathbb{R}$ is defined by

$$(\text{Greedy}(V))(i) := \arg\max_a \left(R(i,a) + \gamma P_{i,a}^\top V \right) .$$

If V^* is the optimal value function then Greedy(V) is an optimal policy.

2.2 Average Reward Criterion

Define the average reward value function of a policy π as follows.

$$U_{\mu,M}(i) := \liminf_{T\to\infty} \frac{\mathbb{E}_i^{\mu,M}\left[\sum_{t=0}^{T-1} R(s_t,a_t)\right]}{T} .$$

This limit exists for all $\mu \in \Pi_D$. Moreover, for *irreducible* MDPs considered in Chapter 6 the limit is independent of the start state i and is denoted there by λ_μ. For $\mu \in \Pi_D$, the 'lim inf' in the above definition can be replaced by a 'lim' and there exists $h_{\mu,M} : S \mapsto \mathbb{R}$ such that the following important relation holds (see, for example, Section 8.2.2 in [26]) between $V_{\gamma,\mu,M}$ and $U_{\mu,M}$,

$$\forall i,\ U_{\mu,M}(i) = (1-\gamma)V_{\gamma,\mu,M}(i) + (1-\gamma)h_{\mu,M}(i) + O\left(|1-\gamma|^2\right) . \tag{2.1}$$

Also note that for an event A and a natural number m, we use the following notation in the rest of the thesis.

$$\mathbf{1}[A] := 1 \text{ if } A \text{ is true, } 0 \text{ otherwise ,}$$

$$[m] := \{1, \ldots, m\} .$$

Chapter 3

Sample Complexity of Policy Search

In this chapter, we consider an example of a *policy search* method. These methods try to find a good policy for an MDP by choosing one from a given class. In the method we investigate, the policy is chosen based on its empirical performance in simulations. We are interested in conditions on the complexity of the policy class that ensure the success of such simulation based policy search methods. We show that under bounds on the amount of computation involved in computing policies, transition dynamics and rewards, uniform convergence of empirical estimates to true value functions occurs. Previously, such results were derived by assuming boundedness of pseudodimension and Lipschitz continuity. These assumptions and ours are both stronger than the usual combinatorial complexity measures. We show, via minimax inequalities, that this is essential: boundedness of pseudodimension or fat-shattering dimension alone is not sufficient.

3.1 Policy Search Using Simulators

Except for toy problems with a few states, computing an optimal policy for an MDP is usually out of the question. Some relaxations need to be done if our aim is to develop tractable methods for achieving near optimal performance. One possibility is to avoid considering all possible policies by restricting oneself to a smaller class Π of policies. Given a simulator for the environment, we try to pick the best policy from Π. The hope is that if the policy class is appropriately chosen, the best policy in Π would not be too much worse than the true optimal policy.

Use of simulators introduces an important issue: how is one to be sure that performance of policies in the class Π on a few simulations is indicative of their true performance? This is reminiscent of the situation in statistical learning. There the aim is to learn a concept and one restricts attention to a hypotheses class which may or may not contain the "true" concept. The sample complexity question then is: how many labeled examples are needed in order to be confident that error rates on the training set are close to the true error rates of the hypotheses in our class? The answer turns out to depend on "complexity" of the hypothesis class as measured by combinatorial quantities associated with the class such as the VC dimension, the pseudodimension and the fat-shattering dimension.

Some progress [21, 23] has already been made to obtain uniform bounds on the difference between value functions and their empirical estimates, where the value function of a policy is the expected long term reward starting from a certain state and following the policy thereafter. We continue this line of work by further investigating what properties of the policy class determine the rate of uniform convergence of value function estimates. The

key difference between the usual statistical learning setting and ours is that we not only have to consider the complexity of the class Π but also of the classes derived from Π by composing the functions in Π with themselves and with the state evolution process implied by the simulator.

Ng and Jordan [23] used a finite pseudodimension condition along with Lipschitz continuity to derive uniform bounds. The Lipschitz condition was used to control the covering numbers of the iterated function classes. We provide a uniform convergence result (Theorem 1) under the assumption that policies are parameterized by a finite number of parameters and that the computations involved in computing the policy, the single-step simulation function and the reward function all require a bounded number of arithmetic operations on real numbers. The number of samples required grows linearly with the dimension of the parameter space but is independent of the dimension of the state space. Ng and Jordan's and our assumptions are both stronger than just assuming finiteness of some combinatorial dimension. We show that this is unavoidable by constructing two examples where the fat-shattering dimension and the pseudodimension respectively are bounded, yet no simulation based method succeeds in estimating the true values of policies well. This happens because iteratively composing a function class with itself can quickly destroy finiteness of combinatorial dimensions. Additional assumptions are therefore needed to ensure that these iterates continue to have bounded combinatorial dimensions.

Although we restrict ourselves to MDPs for ease of exposition, the analysis presented in this chapter also carries over easily to the case of partially observable MDPs (POMDPs), provided the simulator also simulates the conditional distribution of observa-

tions given state using a bounded amount of computation.

The plan of the rest of the chapter is as follows. We set up notation and terminology in Section 3.2. In the same section, we describe the model of computation over reals that we use. Section 3.3 proves Theorem 1, which gives a sample complexity bound for achieving a desired level of performance within the policy class. In Section 3.4, we give two examples of policy classes whose combinatorial dimensions are bounded. Nevertheless, we can prove strong minimax lower bounds implying that no method of choosing a policy based on empirical estimates can do well for these examples.

3.2 Bounded Computation Assumption

Suppose we have an MDP $M = \langle S, A, P, R \rangle$. In this chapter, we will assume that S and A are Euclidean spaces of dimensionality d_S and d_A respectively. Further, we do not restrict the reward function to be a deterministic function of the state. Instead, R maps states to distributions over a bounded interval $[0, R_{\max}]$. Moreover, we also assume that there is some initial distribution D over states such that the initial state s_0 is drawn from D.

Let π be a randomized policy. For a discount factor $\gamma \in [0, 1)$ consider the value of the policy,

$$V_{\gamma,\pi,M}(D) = \mathbb{E}_D^{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t \rho_t\right] ,$$

where ρ_t is a random reward drawn from the distribution $R(s_t, a_t)$ where a_t in turn is drawn from $\pi(s_t)$. Since M, γ and D will mostly be fixed throughout this chapter, we will denote the value function of π simply by V_π. Given a horizon length $H > 0$, also define the

truncated value function,

$$V^H_{\gamma,\pi,M}(D) = \mathbb{E}^{\pi,M}_D \left[\sum_{t=0}^{H} \gamma^t \rho_t \right] .$$

We need a model of real computation to state the results in this chapter. We use a degree bounded version of the Blum-Shub-Smale [9] model of computation over reals. At each time step, we can perform one of the four arithmetic operations $+, -, \times, /$ or can branch based on a comparison (say $<$). While Blum et al. allow an arbitrary fixed rational map to be computed in one time step, we further require that the degree of any of the polynomials appearing at computation nodes be at most 1.

Definition 1. Let k, l, m, τ be positive integers, f a function from $\mathbb{R}^k$ to probability distributions over $\mathbb{R}^l$ and Ξ a probability distribution over $\mathbb{R}^m$. The function f is **(Ξ, τ)-computable** if there exists a degree bounded finite dimensional machine $\mathbb{M}$ over $\mathbb{R}$ with input space $\mathbb{R}^{k+m}$ and output space $\mathbb{R}^l$ such that the following hold.

1. For every $x \in \mathbb{R}^k$ and $\xi \in \mathbb{R}^m$, the machine halts with halting time $T_{\mathbb{M}}(x, \xi) \leq \tau$.

2. For every $x \in \mathbb{R}^k$, if $\xi \in \mathbb{R}^m$ is distributed according to Ξ the input-output map $\Phi_{\mathbb{M}}(x, \xi)$ is distributed as $f(x)$.

Informally, the definition states that given access to an oracle which generates samples from Ξ, we can generate samples from $f(x)$ by doing a bounded amount of computation. For precise definitions of the input-output map and halting time, we refer the reader to Chapter 2 of [9].

In Section 3.3, we assume that the policy class Π is parameterized by a finite dimensional parameter $\theta \in \mathbb{R}^d$. In this setting $\pi(s; \theta)$, $P_{s,a}$ and $R(s, a)$ are distributions

over $\mathbb{R}^{d_A}$, $\mathbb{R}^{d_S}$ and $[0, R]$ respectively. The following assumption states that all these maps are computable within τ time steps in our model of computation.

Assumption 1. There exists a probability distribution Ξ over $\mathbb{R}^m$ and a positive integer τ such that the functions

- $s \mapsto \pi(s; \theta)$,
- $(s, a) \mapsto P_{s,a}$, and
- $s \mapsto R(s, a)$

are (Ξ, τ)-computable. Let $\mathbb{M}_\pi$, $\mathbb{M}_P$ and $\mathbb{M}_r$ respectively be the machines that compute them.

This assumption will be satisfied if we have three "programs" that make a call to a random number generator for distribution Ξ, do a fixed number of floating-point operations and simulate the policies in our class, the state-transition dynamics and the rewards respectively. The following two examples illustrate this for the state-transition dynamics.

Example 1. Linear Dynamical System with Additive Noise

Suppose A and B are $d_S \times d_S$ and $d_S \times d_A$ matrices and the system dynamics is given by

$$s_{t+1} = As_t + Ba_t + \xi_t \ , \tag{3.1}$$

where ξ_t are i.i.d. from some distribution Ξ. Since computing (3.1) takes $2(d_S^2 + d_S d_A + d_S)$ operations, $P_{s,a}$ is (Ξ, τ)-computable for $\tau = O(d_S(d_S + d_A))$. In this case, the *realizable dynamics*, i.e. the mapping from state to next state for a given policy class, is not uniformly

Lipschitz if policies allow unbounded actions. So previously known bounds [23] are not applicable even in this simple setting.

Example 2. Discrete States and Actions

Suppose $S = \{1, 2, \ldots, |S|\}$ and $A = \{1, 2, \ldots, |A|\}$. For some fixed s, a, $P_{s,a}$ is described by n numbers $(p_1, \ldots, p_{n_S})$, $\sum_i p_i = 1$. Let $P_k = \sum_{i=1}^{k} p_i$. For $\xi \in (0, 1]$, set $f(\xi) = \min\{k : P_k \geq \xi\}$. Thus, if ξ has uniform distribution on $(0, 1]$, then $f(\xi) = k$ with probability p_k. Since the P_k's are non-decreasing, $f(\xi)$ can be computed in $\log |S|$ steps using binary search. But this was for a fixed s, a pair. Finding which $P_{s,a}$ to use, further takes $\log(|S||A|)$ steps using binary search. So if Ξ denotes the uniform distribution on $(0, 1]$ then $P_{s,a}$ is (Ξ, τ)-computable for $\tau = O(\log |S| + \log |A|)$.

For a small ϵ, let H be the ϵ horizon time, i.e. ignoring rewards beyond time H does not affect the value of any policy by more than ϵ. To obtain sample rewards, given initial state s_0 and policy $\pi_\theta = \pi(\cdot; \theta)$, we first compute the trajectory $s_0, \ldots, s_H$ sampled from the Markov chain induced by π_θ. This requires H "calls" each to $\mathbb{M}_\pi$ and $\mathbb{M}_P$. A further $H + 1$ calls to $\mathbb{M}_r$ are then required to generate the rewards ρ_0 through ρ_H. These calls require a total of $3H + 1$ samples from Ξ. The empirical estimates are computed as follows. Suppose, for $1 \leq i \leq n$, $(s_0^{(i)}, \vec{\xi}_i)$ are i.i.d. samples generated from the joint distribution $D \times \Xi^{3H+1}$. Define the empirical estimate of the value of the policy π by

$$\hat{V}_{\pi_\theta}^{H} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{H} \gamma^t \rho_t(s_0^{(i)}, \theta, \vec{\xi}_i) \ .$$

Define an ϵ-approximate maximizer of $\hat{V}$ to be a policy π' such that

$$\hat{V}_{\pi'}^{H} \geq \sup_{\pi \in \Pi} \hat{V}_{\pi}^{H} - \epsilon \ .$$

Finally, we mention the definitions of three standard combinatorial dimensions. Let $\mathcal{X}$ be some space and consider classes $\mathcal{G}$ and $\mathcal{F}$ of $\{-1, +1\}$ and real valued functions on $\mathcal{X}$, respectively. Fix a finite set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$. We say that $\mathcal{G}$ *shatters* X if for all bit vectors $\vec{b} \in \{0,1\}^n$ there exists $g \in \mathcal{G}$ such that for all i, $b_i = 0 \Rightarrow g(x_i) = -1$, $b_i = 1 \Rightarrow g(x_i) = +1$. We say that $\mathcal{F}$ *shatters* X if there exists $\vec{r} \in \mathbb{R}^n$ such that, for all bit vectors $\vec{b} \in \{0,1\}^n$, there exists $f \in \mathcal{F}$ such that for all i, $b_i = 0 \Rightarrow f(x_i) < r_i$, $b_i = 1 \Rightarrow f(x_i) \geq r_i$. We say that $\mathcal{F}$ *ϵ-shatters* X if these exists $\vec{r} \in \mathbb{R}^n$ such that, for all bit vectors $\vec{b} \in \{0,1\}^n$, there exists $f \in \mathcal{F}$ such that for all i, $b_i = 0 \Rightarrow f(x_i) \leq r_i - \epsilon$, $b_i = 1 \Rightarrow f(x_i) \geq r_i + \epsilon$. We then have the following definitions,

$$\mathrm{VCdim}(\mathcal{G}) = \max\{|X| : \mathcal{G} \text{ shatters } X\} ,$$

$$\mathrm{Pdim}(\mathcal{F}) = \max\{|X| : \mathcal{F} \text{ shatters } X\} ,$$

$$\mathrm{fat}_{\mathcal{F}}(\epsilon) = \max\{|X| : \mathcal{F} \ \epsilon\text{-shatters } X\} .$$

3.3 Sample Complexity Bound

We now state the main result of this chapter. It gives the number of trajectories that need to be generated to guarantee that an approximate maximizer of $\hat{V}$ is also an approximate maximizer of V, the true value function.

Theorem 1. *Fix an MDP M, a discount factor $\gamma \in [0,1)$, a policy class $\Pi = \{s \mapsto \pi(s;\theta) : \theta \in \mathbb{R}^d\}$, and an $\epsilon > 0$. Suppose Assumption 1 holds. Then*

$$n > O\left(\frac{R_{\max}^2 H}{(1-\gamma)^2\epsilon^2} \cdot d \cdot \tau \cdot \log \frac{R_{\max}}{\epsilon(1-\gamma)}\right)$$

ensures that

$$\mathbb{E}\left[\sup_{\pi\in\Pi} V_\pi - V_{\pi_n}\right] \leq 3\epsilon + \epsilon' \ ,$$

where π_n *is an* ϵ'*-approximate maximizer of* $\hat{V}$ *and* $H = \log_{1/\gamma}(2R_{\max}/(\epsilon(1-\gamma)))$ *is the* $\epsilon/2$ *horizon time.*

Proof. The proof consists of three steps: (1) Assumption 1 is used to get bounds on pseudodimension; (2) The pseudodimension bound is used to prove uniform convergence of empirical estimates to true value functions; (3) Uniform convergence and the definition of ϵ'-approximate maximizer gives the bound on expected regret.

STEP 1. Given initial state s_0, parameter θ and random numbers ξ_1 through ξ_{3H+1}, we first compute the trajectory as follows. Recall that $\Phi_{\mathbb{M}}$ refers to the input-output map of a machine $\mathbb{M}$.

$$s_t = \Phi_{\mathbb{M}_P}(s_{t-1}, \Phi_{\mathbb{M}_\pi}(\theta, s, \xi_{2t-1}), \xi_{2t}),\ 1 \leq t \leq H \ . \tag{3.2}$$

The rewards are then computed by

$$\rho_t = \Phi_{\mathbb{M}_r}(s_t, \Phi_{\mathbb{M}_\pi}(\theta, s, \xi_{2t-1}), \xi_{2H+t+1}),\ 0 \leq t \leq H \ . \tag{3.3}$$

The H-step discounted reward sum is computed as

$$\sum_{t=0}^{H} \gamma^t \rho_t = \rho_0 + \gamma(\rho_1 + \gamma(\rho_2 + \ldots(\rho_{H-1} + \gamma\rho_H)\ldots)) \ . \tag{3.4}$$

Define the function class $\mathcal{R} = \{(s_0, \vec{\xi}) \mapsto \sum_{t=0}^{H} \gamma^t \rho_t(s_0, \theta, \vec{\xi}) : \theta \in \mathbb{R}^d\}$, where we have explicitly shown the dependence of ρ_t on s_0, θ and $\vec{\xi}$. Let us count the number of arithmetic operations needed to compute a function in this class. Using Assumption 1, we see that steps (3.2) and (3.3) require no more than $2\tau H$ and $\tau(H+1)$ operations respectively.

Step (3.4) requires H multiplications and H additions. This gives a total of $2\tau H + \tau(H + 1) + 2H \leq 6\tau H$ operations. Goldberg and Jerrum [17] showed that the VC dimension of a function class can be bounded in terms of an upper bound on the number of arithmetic operations it takes to compute the functions in the class. Since the pseudodimension of $\mathcal{R}$ can be written as

$$\text{Pdim}(\mathcal{R}) = \text{VCdim}\{(s_0, \vec{\xi}, c) \mapsto \text{sign}(f(s_0, \vec{\xi}) - c) : f \in \mathcal{R}, c \in \mathbb{R}\} \ ,$$

we get the following bound by Theorem 8.4 in [2],

$$\text{Pdim}(\mathcal{R}) \leq 4d(6\tau H + 3) \ . \tag{3.5}$$

Step 2. Recall that V_π^H denotes the truncated value function for π. For the choice of H stated in the theorem, we have for all π, $|V_\pi^H - V_\pi| \leq \epsilon/2$. Therefore,

$$\mathbb{P}(\exists \pi \in \Pi : |\hat{V}_\pi^H - V_\pi| > \epsilon) \leq \mathbb{P}(\exists \pi \in \Pi : |\hat{V}_\pi^H - V_\pi^H| > \epsilon/2) \ . \tag{3.6}$$

Functions in $\mathcal{R}$ are positive and bounded above by $R' = R_{\max}/(1-\gamma)$. There are well-known bounds for deviations of empirical estimates from true expectations for bounded function classes in terms of the pseudodimension of the class (see, for example, Theorems 3 and 5 in [20]; also see Pollard's book [25]). Using a weak form of these results, we get

$$\mathbb{P}(\exists \pi \in \Pi : |\hat{V}_\pi^H - V_\pi^H| > \epsilon) \leq 8 \left(\frac{32eR'}{\epsilon}\right)^{2\,\text{Pdim}(\mathcal{R})} e^{-\epsilon^2 n/64R'^2} \ .$$

In order to ensure that $\mathbb{P}(\exists \pi \in \Pi : |\hat{V}_\pi^H - V_\pi^H| > \epsilon/2) < \delta$, we need

$$8 \left(\frac{64eR'}{\epsilon}\right)^{2\,\text{Pdim}(\mathcal{R})} e^{-\epsilon^2 n/256R'^2} < \delta \ ,$$

Using the bound (3.5) on $\text{Pdim}(\mathcal{R})$, we get that

$$\mathbb{P}\left(\sup_{\pi \in \Pi} \left|\hat{V}_\pi^H - V_\pi\right| > \epsilon\right) < \delta \ , \tag{3.7}$$

provided

$$n > \frac{256R^2}{(1-\gamma)^2\epsilon^2}\left(\log\left(\frac{8}{\delta}\right) + 8d(6\tau H+3)\log\left(\frac{64eR}{(1-\gamma)\epsilon}\right)\right) .$$

STEP 3. We now show that (3.7) implies

$$\mathbb{E}\left[\sup_{\pi\in\Pi} V_\pi - V_{\pi_n}\right] \leq \frac{R_{\max}\delta}{(1-\gamma)} + (2\epsilon+\epsilon') .$$

The theorem them immediately follows by setting $\delta = (1-\gamma)\epsilon/R$.

Suppose that for all $\pi \in \Pi$, $|\hat{V}_\pi^H - V_\pi| \leq \epsilon$. This implies that for all $\pi \in \Pi$, $V_\pi \leq \hat{V}_\pi^H + \epsilon$. Since π_n is an ϵ'-approximate maximizer of $\hat{V}$, we have for all $\pi \in \Pi$, $\hat{V}_\pi^H \leq \hat{V}_{\pi_n}^H + \epsilon'$. Thus, for all $\pi \in \Pi$, $V_\pi \leq \hat{V}_{\pi_n}^H + \epsilon + \epsilon'$. Taking the supremum over $\pi \in \Pi$ and using the fact that $\hat{V}^H(\pi_n) \leq V(\pi_n) + \epsilon$, we get $\sup_{\pi\in\Pi} V_\pi \leq V_{\pi_n} + 2\epsilon + \epsilon'$. Thus, if (3.7) holds then we have

$$\mathbb{P}\left(\sup_{\pi\in\Pi} V_\pi - V_{\pi_n} > 2\epsilon + \epsilon'\right) < \delta .$$

Denoting the event $\{\sup_{\pi\in\Pi} V_\pi - V_{\pi_n} > 2\epsilon + \epsilon'\}$ by E, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{\pi\in\Pi} V_\pi - V_{\pi_n}\right] &= \mathbb{E}\,\mathbf{1}\left[E\right]\left[\sup_{\pi\in\Pi} V_\pi - V_{\pi_n}\right] + \mathbb{E}\,\mathbf{1}\left[\neg E\right]\left[\sup_{\pi\in\Pi} V_\pi - V_{\pi_n}\right] \\ &\leq R_{\max}\delta/(1-\gamma) + (2\epsilon+\epsilon') .\end{aligned}$$

where we used the fact that the random variable whose expectation we are bounding is upper bounded by $R_{\max}/(1-\gamma)$. □

3.4 Two Policy Classes Having Bounded Combinatorial Dimensions

We will describe two policy classes for which we can prove that there are strong limitations on the performance of any method (of choosing a policy out of a policy class) that

has access only to empirically observed rewards. Somewhat surprisingly, one can show this for policy classes which are "simple" in the sense that standard combinatorial dimensions of these classes are bounded. This shows that sufficient conditions for the success of simulation based policy search (such as the assumptions in [23] and in our Theorem 1) have to be necessarily stronger than boundedness of standard combinatorial dimensions.

The first example is a policy class $\mathcal{F}_1$ for which $\mathrm{fat}_{\mathcal{F}_1}(\epsilon) < \infty$ for all $\epsilon > 0$. The second example is a class $\mathcal{F}_2$ for which $\mathrm{Pdim}(\mathcal{F}_2) = 1$. Since finiteness of pseudodimension is a stronger condition, the second example makes our point more forcefully than the first one. However, the first example is considerably less contrived than the second one.

Policy Class 1

Let $M_D = \langle S, A, P, R \rangle$ be an MDP with initial state distribution D and

$$S = [-1, +1] \ ,$$

$$A = [-2, +2] \ ,$$

$$P_{s,a}(s') = \begin{cases} 1 & \text{if } s' = \max(-1, \min(s + a, 1))) \\ 0 & \text{otherwise} \end{cases} \ ,$$

$$R = \text{deterministic reward that maps } s \text{ to } s$$

Let γ be some fixed discount factor in $[0, 1)$.

For a function $f : [-1, +1] \mapsto [-1, +1]$, let π_f denote the (deterministic) policy which takes action $f(s) - s$ in state s. Given a class $\mathcal{F}$ of functions, we define an associated policy class $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$.

We now describe a specific function class $\mathcal{F}_1$. Fix $\epsilon_1 > 0$. Let T be an arbitrary finite subset of $(0,1)$. Let $\delta(x) = (1 - |x|)_+$ be the "triangular spike" function. Let

$$f_T(x) = \begin{cases} -1 & -1 \leq x < 0 \\ 0 & x = 0 \\ \max\limits_{y \in T} \left(\frac{\epsilon_1}{|T|} \delta \left(\frac{x-y}{\epsilon_1/|T|} \right) - \frac{\epsilon_1}{|T|} \right) & 0 < x \leq 1 \end{cases} .$$

There is a spike at each point in T and the tips of the spikes just touch the X-axis (see Figure 3.1). Let

$$f_T^n := \underbrace{f_T \circ f_T \circ \ldots \circ f_T}_{n \text{ times}} .$$

Since -1 and 0 are fixed points of $F_T(x)$, it is straightforward to verify that

$$f_T^2(x) = \begin{cases} -1 & -1 \leq x < 0 \\ 0 & x = 0 \\ \mathbf{1}\left[x \in T\right] - 1 & 0 < x \leq 1 \end{cases} . \tag{3.8}$$

Also, $f_T^n = f_T^2$ for all $n > 2$. Define $\mathcal{F}_1 = \{f_T : T \subset (\epsilon_1, 1), |T| < \infty\}$. By construction, functions in $\mathcal{F}_1$ have bounded total variation and so, $\text{fat}_{\mathcal{F}_1}(\epsilon)$ is $O(1/\epsilon)$ (see, for example, Chapter 11 in [2]). Moreover, $f_T(x)$ satisfies the Lipschitz condition everywhere (with constant $L = 1$) except at 0. This is striking in the sense that the loss of the Lipschitz property at a single point allows us to prove the following lower bound.

Proposition 2. *Let g_n range over functions from S^n to $\mathcal{F}_1$. Let D range over probability distributions on S. Then,*

$$\inf_{g_n} \sup_{D} \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} \left[\sup_{\pi \in \Pi_{\mathcal{F}_1}} V_{M_D,\pi} - V_{M_D,\pi_{g_n(s^1,\ldots,s^n)}} \right] \geq \frac{\gamma^2}{1-\gamma} - 2\epsilon_1 .$$

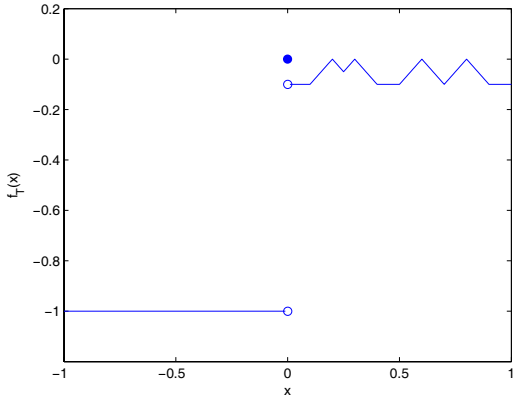

Figure 3.1: Plot of the function f_T with $T = \{0.2, 0.3, 0.6, 0.8\}$. Note that, for $x > 0$, $f_T(x)$ is 0 iff $x \in T$. Also, $f_T(x)$ satisfies the Lipschitz condition (with constant 1) everywhere except at 0.

This says that for any method that maps random initial states $s^1, \ldots, s^n$ to a policy in $\Pi_{\mathcal{F}_1}$, there is an initial state distribution such that the expected regret of the selected policy is at least $\gamma^2/(1-\gamma) - 2\epsilon_1$. This is in sharp contrast to Theorem 1 where we could reduce, by using sufficiently many samples, the expected regret down to any positive number given the ability to maximize the empirical estimates $\hat{V}$.

Let us see how maximization of empirical estimates behaves in this case. Since $\text{fat}_{\mathcal{F}_1}(\epsilon) < \infty$ for all $\epsilon > 0$, the law of large numbers holds uniformly (see Theorem 2.5 in [1]) over the class $\mathcal{F}_1$. The transitions, policies and rewards here are all deterministic. The reward function is just the identity. This means that the 1-step reward function family is just $\mathcal{F}_1$. So the estimates of 1-step rewards are still uniformly concentrated around their expected values. Since the contribution of rewards from time step 2 onwards can be no more than $\gamma^2 + \gamma^3 + \ldots = \gamma^2/(1-\gamma)$, we can claim that, for any initial distribution D,

$$\mathbb{E}\left[\sup_{\pi \in \Pi_{\mathcal{F}_1}} V_{M_D,\pi} - V_{M_D,\pi_n}\right] \leq \frac{\gamma^2}{1-\gamma} + e_n \ ,$$

where $e_n \to 0$. Thus the bound in Proposition 2 above is essentially tight.

Before we prove Proposition 2, we need the following lemma whose proof is given in Appendix A, Section A.1.

Lemma 3. *Fix an interval (a, b) and let $\mathcal{T}$ be the set of all its finite subsets. Let g_n range over functions from $(a, b)^n$ to $\mathcal{T}$. Let D range over probability distributions on (a, b). Then,*

$$\inf_{g_n} \sup_{D} \left(\sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim D} \mathbf{1}\left[X \in T\right] - \mathbb{E}_{(X_1,\ldots,X_n)\sim D^n} \mathbb{E}_{(X \sim D)} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n)\right] \right) \geq 1 \ .$$

Proof of Proposition 2. We will prove the inequality when D ranges over distributions on $(0, 1)$ which, obviously, implies the proposition.

Since, for all $f \in \mathcal{F}_1$ and $n > 2$, $f^n = f^2$, we have

$$\begin{aligned}
&\sup_{\pi \in \Pi_{\mathcal{F}_1}} V_{M_D,\pi} - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} V_{M_D}(\pi_{g_n(s^1,\ldots,s^n)}) \\
&= \sup_{f \in \mathcal{F}_1} \mathbb{E}_{s \sim D} \left[s + \gamma f(s) + \frac{\gamma^2}{1-\gamma} f^2(s) \right] \\
&\quad - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} \left[\mathbb{E}_{s \sim D}[s + \gamma g_n(s^1, \ldots, s^n)(s) + \frac{\gamma^2}{1-\gamma} g_n(s^1, \ldots, s^n)^2(s)] \right] \\
&= \sup_{f \in \mathcal{F}_1} \mathbb{E}_{s \sim D} \left[\gamma f(s) + \frac{\gamma^2}{1-\gamma} f^2(s) \right] \\
&\quad - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} \left[\mathbb{E}_{s \sim D}[\gamma g_n(s^1, \ldots, s^n)(s) + \frac{\gamma^2}{1-\gamma} g_n(s^1, \ldots, s^n)^2(s)] \right] \ .
\end{aligned}$$

For all f_1, f_2, $|\mathbb{E}f_1 - \mathbb{E}f_2| \leq \mathbb{E}|f_1 - f_2| \leq \epsilon_1$. Therefore, we can get rid of the first terms in both sub-expressions above without changing the value by more than $2\gamma\epsilon_1$. Thus, continuing

to bound the expression above, we get

$$\geq \sup_{f\in\mathcal{F}_1} \mathbb{E}_{s\sim D}\left[\frac{\gamma^2}{1-\gamma}f^2(s)\right] - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n}\left[\mathbb{E}_{s\sim D}[\frac{\gamma^2}{1-\gamma}g_n(s^1,\ldots,s^n)^2(s)]\right]$$

$$- 2\gamma\epsilon_1$$

$$=\frac{\gamma^2}{1-\gamma}\left(\sup_{f\in\mathcal{F}_1} \mathbb{E}_{s\sim D}\left[f^2(s)+1\right] - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n}\mathbb{E}_{s\sim D}[g_n(s^1,\ldots,s^n)^2(s)+1]\right)$$

$$- 2\gamma\epsilon_1 \ .$$

From (3.8), we know that $f_T^2(x)+1$ restricted to $x\in(0,1)$ is the same as $\mathbf{1}[x\in T]$. Therefore, restricting D to probability measures on $(0,1)$ and applying Lemma 3, we get

$$\inf_{g_n}\sup_{D}\left(\sup_{\pi\in\Pi_{\mathcal{F}_1}} V_{M_D,\pi} - \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} V_{M_D,\pi_{g_n(s^1,\ldots,s^n)}}\right) \geq \frac{\gamma^2}{1-\gamma} - 2\gamma\epsilon_1 \ .$$

Noting that $\gamma<1$ finishes the proof. □

Policy Class 2

We use the MDP of the previous section with a different policy class which we now describe. For a real number $x,y\in(0,1)$ with binary expansions (choose the terminating representation for rationals) $0.b_1b_2b_3\ldots$ and $0.c_1c_2c_3\ldots$, define

$$\mathrm{mix}(x,y) = 0.b_1c_1b_2c_2\ldots \qquad \mathrm{stretch}(x) = 0.b_10b_20b_3\ldots$$

$$\mathrm{even}(x) = 0.b_2b_4b_6\ldots \qquad \mathrm{odd}(x) = 0.b_1b_3b_5\ldots$$

Some obvious identities are $\mathrm{mix}(x,y) = \mathrm{stretch}(x) + \mathrm{stretch}(y)/2$, $\mathrm{odd}(\mathrm{mix}(x,y)) = x$ and $\mathrm{even}(\mathrm{mix}(x,y)) = y$. Now fix $\epsilon_2>0$. Since, finite subsets of $(0,1)$ and irrationals in $(0,\epsilon_2)$ have the same cardinality, there exists a bijection h which maps every finite subset T of

$(0,1)$ to some irrational $h(T) \in (0, \epsilon_2)$. For a finite subset T of $(0,1)$, define

$$f_T(x) = \begin{cases} 0 & x = -1 \\ \mathbf{1}\left[\text{odd}(-x) \in h^{-1}(\text{even}(-x)\right] & -1 < x < 0 \\ 0 & x = 0 \\ -\min(x, h(T)) & 0 < x < 1 \\ 1 & x = 1 \end{cases} .$$

It is easy to check that with this definition, $f_T^2(x) = \mathbf{1}[x \in T]$ for $x \in (0,1)$. Finally, let $\mathcal{F}_2 = \{f_T : T \subset (0,1), |T| < \infty\}$. To calculate the pseudodimension of this class, note that using the identity $\min(x,y) = \text{stretch}(x) + \text{stretch}(y)/2$, every function f_T in the class can be written as $f_T = f_0 + \tilde{f}_T$ where f_0 is a fixed function (does not depend on T) and $\tilde{f}_T$ is given by

$$\tilde{f}_T(x) = \begin{cases} 0 & -1 \le x \le 0 \\ -\text{stretch}(h(T))/2 & 0 < x < 1 \\ 0 & x = 1 \end{cases} .$$

Let $\mathcal{H} = \{\tilde{f}_T : T \subset (0,1), |T| < \infty\}$. Since $\text{Pdim}(\mathcal{H} + f_0) = \text{Pdim}(\mathcal{H})$ for any class $\mathcal{H}$ and a fixed function f_0, we have $\text{Pdim}(\mathcal{F}_2) = \text{Pdim}(\mathcal{H})$. As each function $\tilde{f}_T(x)$ is constant on $(0,1)$ and zero elsewhere, we cannot shatter even two points using $\mathcal{H}$. Thus, $\text{Pdim}(\mathcal{H}) = 1$.

Proposition 4. *Let g_n range over functions from S^n to $\mathcal{F}_2$. Let D range over probability distributions on S. Then,*

$$\inf_{g_n} \sup_{D} \mathbb{E}_{(s^1,\ldots,s^n)\sim D^n} \left[\sup_{\pi \in \Pi_{\mathcal{F}_2}} V_{M_D,\pi} - V_{M_D,\pi_{g_n(s^1,\ldots,s^n)}} \right] \ge \frac{\gamma^2}{1-\gamma} - \epsilon_2 .$$

Sketch. Let us only check that the properties of $\mathcal{F}_1$ that allowed us to proceed with the proof of Proposition 2 are also satisfied by $\mathcal{F}_2$. First, for all $f \in \mathcal{F}_2$ and $n > 2$, $f^n = f^2$. Second, for all $f_1, f_2 \in \mathcal{F}_2$ and $x \in [-1, +1]$, $|f_1(x) - f_2(x)| \leq \epsilon_2/2$. This is because f_{T_1} and f_{T_2} can differ only for $x \in (0, 1)$. For such an x, $|f_{T_1}(x) - f_{T_2}(x)| = |\operatorname{mix}(x, h(T_1) - \operatorname{mix}(x, h(T_2))| = |\operatorname{stretch}(h(T_1)) - \operatorname{stretch}(h(T_2))|/2 \leq \epsilon_2/2$. $D =$ some distribution on $[-1, +1]$, $D =$ some distribution on $[-1, +1]$, Third, the restriction of f_T^2 to $(0, 1)$ is $\mathbf{1}[x \in T]$. □

Chapter 4

Approximate Linear Programming

In this chapter, we consider the approximate linear programming approach to solving large MDPs that has been pioneered by Van Roy and his colleagues. We first consider the exact linear program that can be used to solve for the optimal value function of an explicitly given MDP. We then make a linear approximation for the value function resulting in the approximate linear program. We give a new proof of one of Van Roy's bounds by directly appealing to LP duality. Then we derive a bound on the performance of the greedy policy obtained from the solution of the approximate linear program with constraint sampling. The bound that previously appeared in the literature could only be applied to the solution of the approximate linear program without constraint sampling. We also look at the idea of approximating the exact dual linear program and prove bounds on the amount by which the optimum shifts as a result of making the approximation.

4.1 The Exact LP and its Dual

Given an MDP $M = \langle S, A, P, R \rangle$ and a discount factor $\gamma \in [0, 1)$, there are several ways to compute an optimal policy μ^* for the γ-discounted sum of rewards criterion. One way is to solve the following linear program.

PrimalLP

$$\text{minimize } \alpha^\top V$$

$$\text{subject to } V(i) \geq R(i,a) + \gamma P_{i,a}^\top V \qquad \forall (i,a) \in S \times A$$

Here α is probability distribution over states such that $\alpha(i) > 0$ for all $i \in S$. For any such α, the optimal value function $V^* = V_{\mu^*}$ is the unique optimal solution of PrimalLP. The variables of the above linear program are $V(i)$, $i \in S$. It therefore has $|S|$ variables and $|S||A|$ constraints.

Introducing the dual variable $\phi(i, a)$ for the constraint corresponding to the state-action pair (i, a), we get the following linear program which is the *dual* of PrimalLP.

DualLP

$$\text{maximize } R^\top \phi$$

$$\text{subject to } \sum_a \phi(i,a) = \alpha(i) + \gamma \sum_{j,a} P_{j,a}(i)\phi(j,a) \qquad \forall i \in S$$

$$\phi(i,a) \geq 0 \qquad \forall (i,a) \in S \times A$$

The following theorem relates the feasible solutions of the dual linear program to state-action visitation frequencies of stationary randomized policies. It also relates the

optimal solutions to the visitation frequencies of optimal policies. This can be found in standard textbooks, for example, in Section 6.9 of [26].

Theorem 5. *Consider the linear program* DUALLP*.*

1. *For any $\mu \in \Pi_R$, $\psi_{\mu,\alpha}(i,a)$ is a feasible solution.*

2. *Suppose $\phi(i,a)$ is a feasible solution. Define the policy $\mu \in \Pi_R$ by*

$$\mu(i,a) = \frac{\phi(i,a)}{\sum_a \phi(i,a)} \ .$$

   Then, $\phi(i,a) = \psi_{\mu,\alpha}(i,a)$.

3. *If $\mu \in \Pi_R$ is an optimal policy then $\psi_{\mu,\alpha}(i,a)$ is an optimal solution.*

4. *Suppose $\phi^*(i,a)$ is an optimal solution. Define the policy $\mu \in \Pi_R$ by*

$$\mu(i,a) = \frac{\phi^*(i,a)}{\sum_a \phi^*(i,a)} \ .$$

   Then, μ is an optimal policy.

4.2 Basis Functions and the Approximate LP

For large MDPs, solving the linear program PRIMALLP or its dual DUALLP is impractical. To address this issue, Van Roy [12] and his colleagues have pioneered the *approximate linear programming* approach where we choose a set of K *basis functions* $\{h_1, \ldots, h_K\}$. Each basis function $h_i : S \mapsto \mathbb{R}$ is a function on the state space. We try to approximate the true value function V^* by a linear combination of the basis functions,

$$V^* \approx \sum_{l=1}^{K} w_l h_l = H\mathbf{w} \ ,$$

where $\mathbf{w} = (w_1, \ldots, w_K)^\top$ and $H = (h_1, \ldots, h_K)$ is a $|S| \times K$ matrix. Plugging $V = \sum_l w_l h_l$ into PrimalLP, we get the following linear program.

ApproxPrimal

$$\text{minimize } \alpha^\top H\mathbf{w}$$

$$\text{subject to } (H\mathbf{w})(i) \geq R(i,a) + \gamma P_{i,a}^\top H\mathbf{w} \qquad \forall (i,a) \in S \times A$$

The variables in the above LP are $w_1, \ldots, w_K$. Thus, we have reduced the number of variables from $|S|$ to K. Unfortunately, the number of constraints is still $|S||A|$. *Constraint sampling* has been suggested in the literature [13] to handle this issue. However, let us first focus on the quality of the value function $H\mathbf{w}^*$ obtained from a solution $\mathbf{w}^*$ of ApproxPrimal. The issue of feasibility arises here even before we can talk about the solution $\mathbf{w}^*$. An assumption that guarantees feasibility of ApproxPrimal is that our basis includes a constant function. We make this assumption for the rest of the chapter.

Assumption 2. There exists $h_l \in \{h_1, \ldots, h_K\}$ such that $h_l(i) = 1$ for all $i \in S$.

For later reference we also state the dual of ApproxPrimal.

DualApproxPrimal

$$\text{minimize } R^\top \phi$$

$$\text{subject to}$$

$$\sum_{i,a} \phi(i,a) h_l(i) = \sum_i \alpha(i) h_l(i) + \gamma \sum_{j,a,i} P_{j,a}(i) \phi(j,a) h_l(i) \qquad \forall l \in [K]$$

$$\phi(i,a) \geq 0 \qquad \forall (i,a) \in S \times A$$

Define the weighted 1-norm,

$$\|V\|_{1,\alpha} := \sum_{i \in S} \alpha(i)|V(i)| \ .$$

De Farias and Van Roy proved the following simple bound.

Theorem 6. *If* $\mathbf{w}^*$ *is a solution to* ApproxPrimal*, then*

$$\|V^* - H\mathbf{w}^*\|_{1,\alpha} \leq \frac{2}{1-\gamma} \min_{\mathbf{w}} \|V^* - H\mathbf{w}\|_\infty \ .$$

The above result gives a guarantee on the quality of the solution obtained from the approximate LP ApproxPrimal provided that the span of the basis functions can approximate the optimal value function well in an infinity-norm sense. The latter assumption is a very strong one and de Farias and Van Roy went on to derive further results which relied on weaker assumptions. Nevertheless, Theorem 6 remains a key result offering support for the approximate linear programming approach. It is derived in [12] using the properties of the dynamic programming operator. Below we provide a general result for linear programs from which Theorem 6 may be deduced as a simple corollary. Our derivation also clearly indicates that the only property of the dynamic programming operator T that we use is that $V \geq TV$ implies $V \geq V^*$. Consider a general linear program,

$$\begin{aligned} &\text{minimize } c^\top x \\ &\text{subject to } Ax \geq b \end{aligned} \tag{4.1}$$

and its dual,

$$\begin{aligned} &\text{maximize } b^\top y \\ &\text{subject to } A^\top y = c \\ &y \geq \mathbf{0} \end{aligned} \tag{4.2}$$

Here the dimensions of c, x, A, b and y are $n \times 1$, $n \times 1$, $m \times n$, $m \times 1$ and $m \times 1$ respectively. Suppose, we now try to approximate x as a linear combination Hw of the columns of a $n \times K$ matrix H leading to the linear program,

$$\begin{aligned} &\text{minimize } c^\top Hw \\ &\text{subject to } AHw \geq b \end{aligned} \tag{4.3}$$

and its dual,

$$\begin{aligned} &\text{maximize } b^\top y \\ &\text{subject to } H^\top A^\top y = H^\top c \\ &y \geq \mathbf{0} \end{aligned} \tag{4.4}$$

An interesting thing to notice here is that making the approximation $x \approx Hw$ in the primal (4.1) amounts to approximating the feasible region of the dual (4.2). The K equality constraints in (4.4) are linear combinations of the n equality constraints in (4.2). Suppose all four linear programs above are feasible and bounded. Denote the solutions to them by x^*,y^*, w^* and $\tilde{y}$ respectively. From our discussion above, it is clear that the value $c^\top x^* = b^\top y^*$ of the original linear program is less than the value $c^\top Hw^* = b^\top \tilde{y}$ obtained after approximation. This is because approximating the feasible region of a maximization

problem by a superset can only result in a larger value for the optimum. The result below gives a bound on the amount by which the optimum shifts.

Theorem 7. *Suppose the linear programs* (4.1)–(4.4) *are feasible and bounded. Denote their solution by* x^*,y^*, w^* *and* $\tilde{y}$ *respectively. Define the* dual violation *vector* $\tilde{\Delta} := A^\top \tilde{y} - c$. *Then, we have*

$$c^\top H w^* - c^\top x^* \leq \|\tilde{\Delta}\|_p \min_w \|x^* - Hw\|_q \ ,$$

for any $p, q \geq 1$ *satisfying* $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. By strong duality, non-negativity of $\tilde{y}$ and feasibility of x^* in (4.1), we have

$$c^\top H w^* = \tilde{y}^\top b \leq \tilde{y}^\top A x^* \ .$$

Thus,

$$c^\top H w^* - c^\top x^* \leq (A^\top \tilde{y} - c)^\top x^* = \tilde{\Delta}^\top x^* \ . \tag{4.5}$$

By feasibility of $\tilde{y}$ in (4.4), we have $H^\top (A^\top \tilde{y} - c) = \mathbf{0}$ and hence, for any w,

$$w^\top H^\top (A^\top \tilde{y} - c) = 0 \ .$$

Combining this with (4.5), we get

$$c^\top H w^* - c^\top x^* \leq \tilde{\Delta}^\top (x^* - Hw) \leq \|\tilde{\Delta}\|_p \|x^* - Hw\|_q \ .$$

Minimizing over w finishes the proof. □

Using the above result we can now derive Theorem 6.

Proof of Theorem 6. We apply Theorem 7 to PrimalLP, DualLP, ApproxPrimal and DualApproxPrimal. Denote their solutions by $V^*, \phi^*, \mathbf{w}^*$ and $\tilde{\phi}$. Assumption 2 guarantees that they are all feasible and bounded and that $\|\tilde{\phi}\|_1 = 1/(1-\gamma)$. The dual violations

are, by definition,

$$\tilde{\Delta}(i) := \sum_a \tilde{\phi}(i,a) - \gamma \sum_{j,a} P_{j,a}(i)\tilde{\phi}(j,a) - \alpha(i) \ .$$

We can bound the 1-norm of $\tilde{\Delta}$ as follows.

$$\begin{aligned}
\sum_i |\tilde{\Delta}(i)| &= \sum_i \left| \sum_a \tilde{\phi}(i,a) - \gamma \sum_{j,a} P_{j,a}(i)\tilde{\phi}(j,a) - \alpha(i) \right| \\
&\leq \|\tilde{\phi}\|_1 + \gamma \|\tilde{\phi}\|_1 + \|\alpha\|_1 \\
&= \frac{1}{1-\gamma} + \frac{\gamma}{1-\gamma} + 1 \\
&= \frac{2}{1-\gamma}
\end{aligned}$$

Now we can use Theorem 7 with $p = 1$ and $q = \infty$ to get

$$\alpha^\top H\mathbf{w}^* - \alpha^\top V^* \leq \frac{2}{1-\gamma} \min_{\mathbf{w}} \|V^* - H\mathbf{w}\|_\infty \ .$$

We finish the proof by noting that the constraints in ApproxPrimal ensure $H\mathbf{w}^* \geq T(H\mathbf{w}^*)$ which implies that $H\mathbf{w}^* \geq V^*$. Therefore, $\alpha^\top H\mathbf{w}^* - \alpha^\top V^* = \|V^* - H\mathbf{w}^*\|_{1,\alpha}$. □

4.3 Performance Bound for Greedy Policies

The previous section only gave a bound on the norm of the difference $V^* - H\mathbf{w}^*$. We are ultimately interested in the performance of the greedy policy

$$\mu(\mathbf{w}^*) := \text{Greedy}(H\mathbf{w}^*)$$

derived from $\mathbf{w}^*$. De Farias and Van Roy [12] provided such a result that we quote below.

Theorem 8. *Let V be such that $V \geq TV$. Let $\mu = \text{Greedy}(V)$. Then, for any probability distribution ν over states,*

$$\|V^* - V_\mu\|_{1,\nu} \leq \frac{1}{1-\gamma}\|V^* - V\|_{1,\psi_{\mu,\nu}} \ .$$

Since we know that the solution $\mathbf{w}^*$ to the approximate linear program APPROX-PRIMAL satisfies $H\mathbf{w}^* \geq T(H\mathbf{w}^*)$, the above theorem can be applied to it. However, as we noted above, the approximate linear program has $|S||A|$ constraints. Therefore, we cannot hope to find $\mathbf{w}^*$ efficiently for large MDPs. To tackle this problem, de Farias and Van Roy suggested that we sample the constraints by putting a probability distribution over $S \times A$. Let $\hat{\mathbf{w}}$ denote that solution of the linear program with just the sampled constraints. When the sampling distribution is

$$\psi^*(i,a) := \frac{\psi_{\mu^*,\alpha}(i)}{K} \ , \tag{4.6}$$

they showed that $\|V^* - H\hat{\mathbf{w}}\|_{1,\alpha}$ is not much larger than $\|V^* - H\mathbf{w}^*\|_{1,\alpha}$ with high probability provided we sample enough constraints. It turns out that the number of sampled constraints only needs to be larger than some polynomial function of K, $|A|$ and other parameters of the MDP but it does not depend on the state space size $|S|$. However, no bound is currently known to bound the performance of the greedy policy obtained from $H\hat{\mathbf{w}}$. Theorem 8 is inapplicable since it is no longer true that $H\hat{\mathbf{w}} \geq T(H\hat{\mathbf{w}})$. However, we do know that

$$\psi_{\mu^*,\alpha}\left(\{i \in S \ : \ (H\hat{\mathbf{w}})(i) < (T(H\hat{\mathbf{w}}))(i)\}\right)$$

is small with high probability (see Theorem 2.1 in [12]). This motivates our next result.

Suppose $\mu = \text{Greedy}(V)$. The following theorem bounds the norm of the difference

$V^* - V_\mu$ in terms of the norm of $V^* - V$ and the set of violated constraints

$$C_V := \{i \in S \,:\, V(i) < (TV)(i)\} \ .$$

Note that if $C_V = \emptyset$, we recover the bound of Theorem 8.

Theorem 9. *Let* $\mu = \text{Greedy}(V)$, $\mu_k = \text{Greedy}(T^k V)$ *and* $N = \|TV - V\|_\infty$. *For* $X \subseteq S$, *define*

$$\sigma(X) := \sum_{k=1}^{\infty} \gamma^k P_{\mu_k} P_{\mu_{k-1}} \dots P_{\mu_1} \mathbf{1}_X \ ,$$

where $\mathbf{1}_X$ *is a vector defined by* $\mathbf{1}_X(i) = \mathbf{1}[i \in X]$. *Then, we have*

$$\|V^* - V_\mu\|_{1,\nu} \leq \frac{1}{1-\gamma} \|V^* - V\|_{1,\psi_{\mu,\nu}} + N \left(\frac{\psi_{\mu,\nu}^\top \sigma(C_V)}{1-\gamma} + \nu(C_V) + \nu^\top \sigma(C_V) \right) \ .$$

Proof. We first prove by induction that, for all $k \geq 1$,

$$T^{k+1} V \leq T^k J + N \gamma^k P_{\mu_k} P_{\mu_{k-1}} \dots P_{\mu_1} \mathbf{1}_{C_V} \ . \tag{4.7}$$

For the base case, we need to prove

$$T^2 V \leq TV + N \gamma P_{\mu_1} \mathbf{1}_{C_V} \ . \tag{4.8}$$

Towards this end, write

$$
\begin{aligned}
(T(TV))(i) &= (T_{\mu_1}(TJ))(i) \\
&\quad [\because \mu_1 = \text{Greedy}(TV)] \\
&= R(i, \mu_1(i)) + \gamma \sum_j P_{i,\mu_1(i)}(j)(TJ)(j) \\
&= R(i, \mu_1(i)) + \gamma \sum_{j \in \bar{C_V}} P_{i,\mu_1(i)}(j)(TV)(j) + \gamma \sum_{j \in C_V} P_{i,\mu_1(i)}(j)(TV)(j) \\
&\leq R(i, \mu_1(i)) + \gamma \sum_{j \in \bar{C_V}} P_{i,\mu_1(i)}(j)V(j) + \gamma \sum_{j \in C_V} P_{i,\mu_1(i)}(j)(TV)(j) \\
&\quad [\because TV \leq V \text{ outside } C_V] \\
&= R(i, \mu_1(i)) + \gamma \sum_j P_{i,\mu_1(i)}(j)V(j) + \gamma \sum_{j \in C_V} P_{i,\mu_1(i)}(j)((TV)(j) - J(j)) \\
&\leq R(i, \mu_1(i)) + \gamma \sum_j P_{i,\mu_1(i)}(j)V(j) + N\gamma \sum_j P_{i,\mu_1(j)}(j)\mathbf{1}_{C_V}(j) \\
&\quad [\text{definitions of } N \text{ and } \mathbf{1}_{C_V}] \\
&= R(i, \mu_1(i)) + \gamma \sum_j P_{i,\mu_1(i)}(j)V(j) - N\gamma(P_{\mu_1}\mathbf{1}_{C_V})(i) \\
&\quad [\text{definition of } P_{\mu_1}] \\
&\leq (TV)(i) + N\gamma(P_{\mu_1}\mathbf{1}_{C_V})(i) \ . \\
&\quad [\text{definition of } T]
\end{aligned}
$$

We have thus shown (4.8). Now assume that (4.7) holds for k. Then we have,

$$
\begin{aligned}
(T^{k+2}J)(i) &= (T_{\mu_{k+1}}(T^{k+1}J))(i) \\
&\quad [\because \mu_{k+1} = \text{Greedy}(T^{k+1}V)] \\
&= R(i,\mu_{k+1}(i)) + \gamma \sum_j P_{i,\mu_{k+1}(i)}(j)(T^{k+1}V)(j) \\
&\leq R(i,\mu_{k+1}(i)) + \gamma \sum_j P_{i,\mu_{k+1}(i)}(j)((T^kV)(j) + N\gamma^k(P_{\mu_k}P_{\mu_{k-1}}\dots P_{\mu_1}\mathbf{1}_{C_V}))(j) \\
&\quad [\text{Induction hypothesis}] \\
&= R(i,\mu_{k+1}(i)) + \gamma \sum_j P_{i,\mu_{k+1}(i)}(j)(T^kV)(j) + N\gamma^{k+1}(P_{\mu_{k+1}}P_{\mu_k}\dots P_{\mu_1}\mathbf{1}_{C_V})(i) \\
&\quad [\text{definition of } P_{\mu_{k+1}}] \\
&\leq (T^{k+1}V)(i) + N\gamma^{k+1}(P_{\mu_{k+1}}P_{\mu_k}\dots P_{\mu_1}\mathbf{1}_{C_V})(i) \\
&\quad [\text{definition of } T]
\end{aligned}
$$

Thus, we have shown that (4.7) holds for $k+1$ too. Since $T^kV \to V^*$ for any V, applying (4.7) for different values of k, adding them and taking limit as $k \to \infty$, we get

$$V^* \leq TV + N\sigma(C_V)\ . \tag{4.9}$$

Now, we have

$$\begin{aligned}
\|V^* - V_\mu\|_{1,\nu} &= \nu^\top (V^* - V_\mu) \\
&\quad [\because V_\mu \leq V^*] \\
&\leq \nu^\top (TV + N\sigma(C_V) - V_\mu) \\
&\quad [\text{by (4.9)}] \\
&= \nu^\top (TV - V_\mu) + N\nu^\top \sigma(C_V) \\
&\leq \nu^\top (V - V_\mu) + N\nu(C_V) + N\nu^\top \sigma(C_V) \\
&\quad [\text{definitions of } N \text{ and } C_V] \\
&= \frac{1}{1-\gamma} \psi_{\mu,\nu}^\top (V - TV) + N\nu(C_V) + N\nu^\top \sigma(C_V) \\
&\quad [\text{ see proof of Theorem 3.1 in [12] }] \\
&\leq \frac{1}{1-\gamma} \psi_{\mu,\nu}^\top (V - V^* + N\sigma(C_V)) + N\nu(C_V) + N\nu^\top \sigma(C_V) \\
&\quad [\text{by (4.9)}] \\
&\leq \frac{1}{1-\gamma} \psi_{\mu,\nu}^\top (V - V^*) + N \left(\frac{\psi_{\mu,\nu}^\top \sigma(C_V)}{1-\gamma} + \nu(C_V) + \nu^\top \sigma(C_V) \right) .
\end{aligned}$$

The theorem follows by noting that $\psi_{\mu,\nu}^\top (V - V^*) \leq \|V^* - V\|_{1,\psi_{\mu,\nu}}$. □

As opposed to de Farias and Van Roy's result, the above bound holds for any V. However, a number of issues need to be worked out before it can be successfully applied to $V = H\hat{\mathbf{w}}$. First of all, we only know that $\|V^* - H\hat{\mathbf{w}}\|_{1,\alpha}$ is small with high probability in case we can sample from $\psi^*(i,a)$ (recall the definition given in (4.6)) which requires access to the occupation measure of an optimal policy. But an optimal policy is unavailable in the first place! We can assume oracle access to samples from the trajectory of an optimal

policy. Such an assumption might be reasonable when we do not know an optimal policy but have blackbox access to an implementation of an optimal or near-optimal policy. For example, in the autonomous vehicle navigation setting, we do have expert human drivers available to us.

Even with this assumption, the above bound is useful only if

$$\alpha \approx \psi_{\mu(\hat{\mathbf{w}}),\nu} \ .$$

Note the recursive nature of the above relation. The vector $\hat{\mathbf{w}}$ itself is obtained as a solution to a random linear program whose objective function involves α. An approach worth investigating is to view the above as a fixed point equation and try to find a solution by repeated iterations of the mapping whose fixed point is required. This is not very straightforward as the mapping is random (note that obtaining $\hat{\mathbf{w}}$ involves randomly sampling constraints) and so the iterates might not converge.

Another issue is that of controlling the additional terms appearing in the bound. They all measure the "size" of the set $C_{H\hat{\mathbf{w}}}$ in various ways. We do know that $\psi_{\mu^*,\alpha}(C_{H\hat{\mathbf{w}}})$ is small with high probability. This suggests that we choose $\nu \approx \psi_{\mu^*,\alpha}$.

4.4 Approximating the Dual

So far in this chapter, we have approximated the variables in the *primal* linear program PrimalLP. Recently, the idea of approximating the variables in the *dual* linear program DualLP has been receiving some attention [14, 28]. We follow the strategy of the previous section by first stating a result for general linear programs and then deriving a result for MDPs as a simple corollary. Consider the linear programs (4.1) and (4.2). Instead

of x, suppose we try to approximate y as a linear combination Gu of the columns of a $m \times K$ matrix G leading to the linear program,

$$\begin{aligned} &\text{maximize } (Gu)^\top b \\ &\text{subject to } A^\top Gu = c \\ &Gu \geq 0 \end{aligned} \tag{4.10}$$

and its dual,

$$\begin{aligned} &\text{minimize } c^\top x \\ &\text{subject to } G^\top (b - Ax + \lambda) = \mathbf{0} \\ &\lambda \geq \mathbf{0} \end{aligned} \tag{4.11}$$

Note that making the approximation $y \approx Gu$ in the dual (4.2) amounts to approximating the feasible region of the primal (4.1). Assuming feasibility and boundedness of the concerned linear programs, denote the solutions to (4.1), (4.2), (4.10), (4.11) by x^*, y^*, u^* and $\tilde{x}, \tilde{\lambda}$ respectively. It is clear that the value $c^\top x^* = b^\top y^*$ of the original linear program is more than the value $c^\top \tilde{x} = b^\top G u^*$ obtained after approximation. This is because we have relaxed the constraints by approximating the feasible region of a minimization problem. This can only result in a smaller value for the optimum. The result below gives two bounds on the amount by which the optimum shifts.

Theorem 10. *Suppose the linear programs* (4.1), (4.2), (4.10) *and* (4.11) *are feasible and bounded. Denote their solutions by* x^*, y^*, u^* *and* $\tilde{x}, \tilde{\lambda}$ *respectively. Then, we have*

$$(y^*)^\top b - (Gu^*)^\top b \leq \|b - A\tilde{x} + \tilde{\lambda}\|_p \min_u \|y^* - Gu\|_q \ , \tag{4.12}$$

and

$$(y^*)^\top b - (Gu^*)^\top b \leq \|b - A\tilde{x}\|_p \min_{Gu \geq \mathbf{0}} \|y^* - Gu\|_q \ , \tag{4.13}$$

for any $p, q \geq 1$ *satisfying* $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. By strong duality and feasibility of y^* in (4.2), we have

$$(Gu^*)^\top b = c^T \tilde{x} = (y^*)^\top A\tilde{x} \ .$$

Thus,

$$(y^*)^\top b - (Gu^*)^\top b = (y^*)^\top (b - A\tilde{x}) \ . \tag{4.14}$$

By feasibility of $\tilde{x}, \tilde{\lambda}$ in (4.11), we have $G^\top(b - A\tilde{x} + \tilde{\lambda}) = \mathbf{0}$ and hence for all u,

$$(Gu)^\top (b - A\tilde{x} + \tilde{\lambda}) = 0 \ .$$

Combining this with (4.14), we get

$$(y^*)^\top b - (Gu^*)^\top b = (y^*)^\top (b - A\tilde{x}) - (Gu)^T (b - A\tilde{x} + \tilde{\lambda}) \ .$$

Since $y^*, \tilde{\lambda} \geq 0$, we have $(y^*)^\top \tilde{\lambda} \geq 0$ and thus adding it to the right hand side we get

$$(y^*)^\top b - (Gu^*)^\top b \leq (y^* - Gu)^T (b - A\tilde{x} + \tilde{\lambda}) \leq \|y^* - Gu\|_q \|b - A\tilde{x} + \tilde{\lambda}\|_p \ .$$

Minimizing over u gives (4.12). Alternatively, if we know that $Gu \geq \mathbf{0}$ then we can add the non-negative quantity $(Gu)^\top \tilde{\lambda}$ to the right hand side to get

$$(y^*)^\top b - (Gu^*)^\top b \leq (y^* - Gu)^T (b - A\tilde{x}) \leq \|y^* - Gu\|_q \|b - A\tilde{x}\|_p \ .$$

Minimizing over u such that $Gu \geq \mathbf{0}$ gives (4.13). □

We now show how to use the above result to say something about the linear program when we plug the approximation $\phi \approx G\mathbf{u}$ in DualLP.

ApproxDual

$$\text{maximize } (G\mathbf{u})^\top R$$

$$\text{subject to } \sum_a (G\mathbf{u})(i,a) = \alpha(i) + \gamma \sum_{j,a} P_{j,a}(i)(G\mathbf{u})(j,a) \qquad \forall i \in S$$

$$(G\mathbf{u})(i,a) \geq 0 \qquad \forall (i,a) \in S \times A$$

Here, G is a $|S||A| \times K$ matrix and u is a $K \times 1$ vector. The idea is to approximate the occupation measure $\psi_{\mu^*,\alpha} = \phi^*$ of an optimal policy by a linear combination of the columns of G. The dual of the above linear program is the following.

DualApproxDual

$$\text{minimize } \alpha^T V$$

$$\text{subject to } \sum_{i,a} G_l(i,a) \Big(R(i,a) + \gamma P_{i,a}^\top V - V(i) + \lambda(i,a) \Big) = 0 \quad \forall l \in [K]$$

$$\lambda(i,a) \geq 0 \qquad \forall (i,a) \in S \times A$$

In order to ensure that the above linear programs are feasible and bounded, we make the following assumption.

Assumption 3. One of the columns G_l of G is the occupation measure $\psi_{\mu,\alpha}$ of some policy $\mu \in \Pi_R$.

We now use Theorem 10 to relate the value of ApproxDual to that of DualLP.

Theorem 11. *Let* $V^*, \phi^*, \mathbf{u}^*$ *and* $\tilde{V}, \tilde{\lambda}$ *be solutions to the linear programs* PrimalLP, DualLP, ApproxDual *and* DualApproxDual *respectively. Define*

$$\tilde{\Delta}_1(i,a) := R(i,a) + \gamma P_{i,a}^\top \tilde{V} - \tilde{V}(i) + \tilde{\lambda}(i,a) \ ,$$

$$\tilde{\Delta}_2(i,a) := R(i,a) + \gamma P_{i,a}^\top \tilde{V} - \tilde{V}(i) \ .$$

Then, we have

$$(\phi^*)^\top R - (G\mathbf{u}^*)^\top R \leq \|\tilde{\Delta}_1\|_\infty \min_{\mathbf{u}} \|\phi^* - G\mathbf{u}\|_1 \ ,$$

and

$$(\phi^*)^\top R - (G\mathbf{u}^*)^\top R \leq \|\tilde{\Delta}_2\|_\infty \min_{G\mathbf{u} \geq \mathbf{0}} \|\phi^* - G\mathbf{u}\|_1 \ .$$

Proof. Assumption (3) guarantees that the four linear programs under consideration are feasible and bounded. So, the result follows immediately from Theorem 10. □

No performance guarantees are currently available in the literature for methods that approximate the dual variables rather than the primal variables. The above result is a small step in this direction but it falls short of being satisfactory due to a number of reasons as described below.

- The bound is *a posteriori*. Unlike the previous section, we cannot find an easy way to derive an *a priori* bound on either $\|\tilde{\Delta}\|_\infty$ or $\|\tilde{\Delta}\|_\infty$. So, the bounds can only be applied after having solved the linear programs.

- The result only bounds the amount by which the optimum shifts. Unfortunately, this does not immediately imply a bound on $\|\phi^* - G\mathbf{u}^*\|$ for any choice of the norm $\|\cdot\|$. This is because we do not have $\phi^* \geq G\mathbf{u}^*$.

- The span of columns of G is required to approximate the occupation measure ϕ^* of the optimal policy well in L_1-norm. This is a very stringent requirement.

Chapter 5

Bounded Parameter Markov Decision Processes

Bounded parameter Markov Decision Processes (BMDPs) were defined by Givan et al. [16] to address the issue of uncertainty in the parameters of an MDP. Unlike the case of an MDP, the notion of an optimal policy for a BMDP is not entirely straightforward. We consider two notions of optimality based on optimistic and pessimistic criteria. These have been analyzed for the discounted rewards criterion in [24]. In this chapter, we will consider the average reward criterion. We will establish a fundamental relationship between the discounted and the average reward problems, prove the existence of Blackwell optimal policies and, for both notions of optimality, derive algorithms that converge to the optimal value function.

5.1 Uncertainty in MDP Parameters

In an MDP, the uncertainty involved in the outcome of making a decision in a certain state is represented using various probabilities. However, these probabilities themselves may not be known precisely. This can happen for a variety of reasons. The probabilities might have been obtained via an estimation process. In such a case, it is natural that confidence intervals will be associated with them. State aggregation, where groups of similar states of a large MDP are merged to form a smaller MDP, can also lead to a situation where probabilities are no longer known precisely but are only known to lie in an interval.

In this chapter, we are concerned with such higher level uncertainty, namely uncertainty about the parameters of an MDP. Bounded parameter MDPs (BMDPs) have been introduced in the literature [16] to address this problem. They use intervals (or equivalently, lower and upper bounds) to represent the set in which the parameters of an MDP can lie. We obtain an entire family, say $\mathcal{M}$, of MDPs by taking all possible choices of parameters consistent with these intervals. For an exact MDP M and a deterministic stationary policy μ (which is a mapping specifying the actions to take in various states), the γ-discounted value from state i, $V_{\gamma,\mu,M}(i)$ and the long term average value $U_{\mu,M}(i)$ are two standard ways of measuring the quality of μ with respect to M. When we have a family $\mathcal{M}$ of MDPs, we are immediately faced with the problem of finding a way to measure the quality of a policy. An optimal policy will then be the one that maximizes the particular performance measure chosen.

We might choose to put a distribution over $\mathcal{M}$ and define the value of a policy as its average value under this distribution. In this chapter, however, we will avoid taking

this approach. Instead, we will consider the worst and the best MDP for each policy and accordingly define two performance measures,

$$U_\mu^{\text{opt}}(i) := \sup_{M \in \mathcal{M}} U_{\mu,M}(i)$$

$$U_\mu^{\text{pes}}(i) := \inf_{M \in \mathcal{M}} U_{\mu,M}(i)$$

where the superscripts denote that these are optimistic and pessimistic criteria respectively. Analogous quantities for the discounted case were defined in [16] and algorithms were given to compute them. However, as the discounted and average reward criteria are quite different, these results do not immediately yield algorithms for the average reward case. In this chapter, we give such algorithms.

The optimistic criterion is motivated by the *optimism in the face of uncertainty* principle. Several learning algorithms for MDPs [5, 10, 15, 27] proceed in the following manner. Faced with an unknown MDP, they start collecting data which yields confidence intervals for the parameters of the MDP. Then they choose a policy which is optimal in the sense of the optimistic criterion. This policy is followed for the next phase of data collection and the process repeats. In fact, the algorithm of Auer and Ortner [5] requires, as a blackbox, an algorithm to compute the optimal (with respect to the optimistic criterion) value function for a BMDP.

The pessimistic criterion is related to research on robust control of MDPs [24]. If nature is adversarial, then once we pick a policy μ it will pick the worst possible MDP M from $\mathcal{M}$. In such a scenario, it is reasonable to choose a policy which is best in the worst case. The results of this chapter can also be seen as extending this line of research to the case of the average reward criterion.

The rest of the chapter is organized as follows. In Section 5.2 we define two kinds of optimal value functions based on the optimistic and pessimistic criteria mentioned above. The main result in Section 5.3 is a relation between the optimal value functions for the average reward and discounted criteria. This result depends on a finite subset property (Theorem 14) satisfied by BMDPs. Section 5.4 proves the existence of Blackwell optimal policies. In the exact MDP case, a Blackwell optimal policy is a policy that is optimal for an entire range of discount factors in the neighborhood of 1. Existence of Blackwell optimal policies is an important result in the theory of MDPs. We extend this result to BMDPs. Then, in Section 5.5, we build up on the previous sections and exploit the relationship between the discounted and average returns together with the existence of a Blackwell optimal policy to derive algorithms that converge to optimal value functions for both optimistic as well as pessimistic criteria. In Section 5.6, we consider an extension of BMDPs by allowing semi-algebraic constraints on MDP parameters. That section also proves the existence of Blackwell optimal policies for semi-algebraically constrained MDPs.

5.2 Optimistic and Pessimistic Value Functions

Recall that an MDP is a tuple $\langle S, A, P, R\rangle$. In this chapter, both the state space S and the action space A will be finite.

Let $\mu : S \mapsto A$ be a deterministic stationary policy. Recall that, for $\gamma \in [0, 1)$, we denote the γ-discounted value function of μ by $V_{\gamma,\mu,M}$ and the optimal value function by $V^*_{\gamma,M}$. Also recall that there is a stationary deterministic policy μ^* such that

$$V_{\gamma,\mu^*,M} = V^*_{\gamma,M} \ .$$

Also recall the definition of the average reward value function.

$$U_{\mu,M}(i) := \liminf_{T\to\infty} \frac{\mathbb{E}_i^{\mu,M}\left[\sum_{t=0}^{T-1} R(s_t, a_t)\right]}{T} \ .$$

A bounded parameter MDP (BMDP) is a collection of MDPs specified by bounds on the parameters of the MDPs. For simplicity, we will assume that the reward function is fixed, so that the only parameters that vary are the transition probabilities. Suppose, for each state-action pair i, a, we are given lower and upper bounds, $l(i, j, a)$ and $u(i, j, a)$ respectively, on the transition probability $P_{i,a}(j)$. We assume that the bounds are legitimate, that is

$$\forall i, a, j, \ 0 \le l(i, j, a) \le u(i, j, a) \ ,$$

$$\forall i, a, \ \sum_j l(i, j, a) \le 1 \ \wedge \ \sum_j u(i, j, a) \ge 1 \ .$$

This means that the set defined by

$$\mathcal{C}_{i,a} := \{q \in \mathbb{R}_+^{|S|} \ : \ q^\top \mathbf{1} = 1 \ \wedge \ \forall j, \ l(i, j, a) \le q_j \le u(i, j, a)\}$$

is non-empty for each state-action pair i, a. Fix S, A and R and define the collection of MDPs

$$\mathcal{M} := \{ \ \langle S, A, P, R\rangle \ : \ \forall i, a, \ P_{i,a} \in \mathcal{C}_{i,a} \ \} \ .$$

We could also choose lower and upper bounds for $R(i, a)$ and also let the rewards vary. But for simplicity, we keep a fixed reward function for all MDPs.

Given a BMDP $\mathcal{M}$ and a stationary deterministic policy μ, there are two natural choices for the value function: an optimistic and a pessimistic one,

$$V_{\gamma,\mu}^{\text{opt}}(i) := \sup_{M\in\mathcal{M}} V_{\gamma,\mu,M}(i) \qquad\qquad V_{\gamma,\mu}^{\text{pes}}(i) := \inf_{M\in\mathcal{M}} V_{\gamma,\mu,M}(i) \ . \tag{5.1}$$

We also define the undiscounted value functions,

$$U_{\mu}^{\text{opt}}(i) := \sup_{M \in \mathcal{M}} U_{\mu,M}(i) \qquad\qquad U_{\mu}^{\text{pes}}(i) := \inf_{M \in \mathcal{M}} U_{\mu,M}(i) \ . \tag{5.2}$$

Optimal value functions are defined by maximizing over policies.

$$\mathbf{V}_{\gamma}^{\text{opt}}(i) := \max_{\mu} V_{\gamma,\mu}^{\text{opt}}(i) \qquad\qquad \mathbf{V}_{\gamma}^{\text{pes}}(i) := \max_{\mu} V_{\gamma,\mu}^{\text{pes}}(i) \tag{5.3}$$

$$\mathbf{U}^{\text{opt}}(i) := \max_{\mu} U_{\mu}^{\text{opt}}(i) \qquad\qquad \mathbf{U}^{\text{pes}}(i) := \max_{\mu} U_{\mu}^{\text{pes}}(i) \tag{5.4}$$

5.3 Relation between Discounted and Average Reward Criteria

In this chapter, we are interested in computing $\mathbf{U}^{\text{opt}}$ and $\mathbf{U}^{\text{pes}}$. Algorithms to compute $\mathbf{V}_{\gamma}^{\text{opt}}$ and $\mathbf{V}_{\gamma}^{\text{pes}}$ have already been proposed in the literature [24]. Let us review some of the results pertaining to the discounted case. We note that the results in this section, with the exception of Theorem 15, either appear or can easily be deduced from results appearing in [16]. However, we provide self-contained proofs of these in Appendix B. Before we state the results, we need to introduce a few important operators. Note that, since $\mathcal{C}_{i,a}$ is a closed, convex set, the maximum (or minimum) of $q^{\top}V$ (a linear function of q) appearing in the definitions below is achieved.

$$\left(T_{\gamma,\mu}^{\text{opt}}V\right)(i) := R(i,a) + \gamma \max_{q \in \mathcal{C}_{i,\mu(i)}} q^{\top}V$$

$$\left(T_{\gamma}^{\text{opt}}V\right)(i) := \max_{a \in A}\left[R(i,a) + \gamma \max_{q \in \mathcal{C}_{i,a}} q^{\top}V\right]$$

$$\left(T_{\gamma,\mu}^{\text{pes}}V\right)(i) := R(i,a) + \gamma \min_{q \in \mathcal{C}_{i,\mu(i)}} q^{\top}V$$

$$\left(T_{\gamma}^{\text{pes}}V\right)(i) := \max_{a \in A}\left[R(i,a) + \gamma \min_{q \in \mathcal{C}_{i,a}} q^{\top}V\right]$$

Recall that an operator T is a contraction mapping with respect to a norm $\|\cdot\|$ if there is a $\gamma \in [0,1)$ such that

$$\forall V_1, V_2,\ \|TV_1 - TV_2\| \leq \gamma \|V_1 - V_2\| \ .$$

A contraction mapping has a unique solution to the fixed point equation $TV = V$ and the sequence $\{T^k V_0\}$ converges to that solution for any choice of V_0. It is straightforward to verify that the four operators defined above are contraction mappings (with factor γ) with respect to the norm

$$\|V\|_\infty := \max_i |V(i)| \ .$$

We saw in Chapter 2 that the fixed points of $T_{\gamma,\mu,M}$ and $T_{\gamma,M}$ are $V_{\gamma,\mu,M}$ and $V^*_{\gamma,M}$ respectively. The following theorem tells us what the fixed points of the above four operators are.

Theorem 12. *The fixed points of* $T^{\rm opt}_{\gamma,\mu}, T^{\rm opt}_{\gamma}, T^{\rm pes}_{\gamma,\mu}$ *and* $T^{\rm pes}_{\gamma}$ *are* $V^{\rm opt}_{\gamma,\mu}, \mathbf{V}^{\rm opt}_{\gamma}, V^{\rm pes}_{\gamma,\mu}$ *and* $\mathbf{V}^{\rm pes}_{\gamma}$ *respectively.*

Existence of optimal policies for BMDPs is established by the following theorem.

Theorem 13. *For any* $\gamma \in [0,1)$*, there exist optimal policies* μ_1 *and* μ_2 *such that, for all* $i \in S$*,*

$$V^{\rm opt}_{\gamma,\mu_1}(i) = \mathbf{V}^{\rm opt}_{\gamma}(i) \ ,$$

$$V^{\rm pes}_{\gamma,\mu_2}(i) = \mathbf{V}^{\rm pes}_{\gamma}(i) \ .$$

A very important fact is that out of the uncountably infinite set $\mathcal{M}$, only a finite set is of real interest.

Theorem 14. *There exist finite subsets $\mathcal{M}_{\rm opt}, \mathcal{M}_{\rm pes} \subset \mathcal{M}$ with the following property. For all $\gamma \in [0,1)$ and for every policy μ there exist $M_1 \in \mathcal{M}_{\rm opt}$, $M_2 \in \mathcal{M}_{\rm pes}$ such that*

$$V_{\gamma,\mu}^{\rm opt} = V_{\gamma,\mu,M_1} \ ,$$

$$V_{\gamma,\mu}^{\rm pes} = V_{\gamma,\mu,M_2} \ .$$

Theorem 15. *The optimal undiscounted value functions are limits of the optimal discounted value functions. That is, for all $i \in S$, we have*

$$\lim_{\gamma\to 1} \ (1-\gamma)\mathbf{V}_{\gamma}^{\rm opt}(i) = \mathbf{U}^{\rm opt}(i) \ , \tag{5.5}$$

$$\lim_{\gamma\to 1} \ (1-\gamma)\mathbf{V}_{\gamma}^{\rm pes}(i) = \mathbf{U}^{\rm pes}(i) \ . \tag{5.6}$$

Proof. Fix $i \in S$. We first prove (5.5). We have

$$\begin{aligned}
\liminf_{\gamma\to 1} \ (1-\gamma)\mathbf{V}_{\gamma}^{\rm opt}(i) &= \liminf_{\gamma\to 1} \max_{\mu} \sup_{M\in\mathcal{M}} (1-\gamma)V_{\gamma,\mu,M}(i) \\
&= \max_{\mu} \liminf_{\gamma\to 1} \sup_{M\in\mathcal{M}} (1-\gamma)V_{\gamma,\mu,M}(i) \\
&\geq \max_{\mu} \sup_{M\in\mathcal{M}} \lim_{\gamma\to 1} (1-\gamma)V_{\gamma,\mu,M}(i) \\
&= \max_{\mu} \sup_{M\in\mathcal{M}} U_{\mu,M}(i) \\
&= \mathbf{U}^{\rm opt}(i)
\end{aligned}$$

This first equality follows by definition. The second equality holds because there are a finite number of policies. The inequality is elementary. The third equality follows from (2.1). Theorem 14 gives us

$$\mathbf{V}_{\gamma}^{\rm opt}(i) = \max_{\mu} \max_{M\in\mathcal{M}_{\rm opt}} V_{\gamma,\mu,M}(i) \ .$$

Using this, we have

$$\begin{aligned}
\limsup_{\gamma\to 1}\ (1-\gamma)\mathbf{V}_{\gamma}^{\mathrm{opt}}(i) &= \limsup_{\gamma\to 1}\max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}}(1-\gamma)V_{\gamma,\mu,M}(i)\\
&= \max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}}\lim_{\gamma\to 1}(1-\gamma)V_{\gamma,\mu,M}(i)\\
&= \max_{\mu}\max_{M\in\mathcal{M}_{\mathrm{opt}}}U_{\mu,M}(i)\\
&\leq \mathbf{U}^{\mathrm{opt}}(i)\ .
\end{aligned}$$

The second equality holds because lim sup and max over a finite set commute. Note that finiteness is crucial here since lim sup and sup do not commute. The third equality follows from (2.1). We have thus established (5.5).

To prove (5.6), one repeats the steps above with appropriate changes. □

5.4 Existence of Blackwell Optimal Policies

For an MDP, Blackwell [8] showed that there exist policies that are optimal for the discounted criteria for all discount factors γ in some interval $(\gamma_0, 1)$. Such policies are called *Blackwell optimal* and their existence is a classical result in the theory of MDPs. These policies have the nice property that they are optimal for the average reward criterion too. Given the importance of Blackwell optimal policies in the theory of MDPs, it is natural to speculate whether they exist for BMDPs. The following theorem answers the question in the affirmative.

Theorem 16. *There exist $\gamma_{\mathrm{opt}} \in (0,1)$, a policy μ_{opt} and an MDP $M_{\mathrm{opt}} \in \mathcal{M}_{\mathrm{opt}}$ such that*

$$\forall\gamma \in (\gamma_{\mathrm{opt}}, 1),\ V_{\gamma,\mu_{\mathrm{opt}},M_{\mathrm{opt}}} = \mathbf{V}_{\gamma}^{\mathrm{opt}}\ .$$

Similarly, there exist $\gamma_{\text{pes}} \in (0,1)$*, a policy* μ_{pes} *and an MDP* $M_{\text{pes}} \in \mathcal{M}_{\text{pes}}$ *such that*

$$\forall \gamma \in (\gamma_{\text{pes}}, 1),\ V_{\gamma,\mu_{\text{pes}},M_{\text{pes}}} = \mathbf{V}^{\text{pes}}_{\gamma}\ .$$

Proof. Given an MDP $M = \langle S, A, P, R\rangle$ and a policy μ, recall that the associated matrix P_μ and the vector R_μ are defined as

$$P_\mu(i,j) := P_{i,\mu(i)}(j)\ ,$$

$$R_\mu(i) := R(i,\mu(i))\ .$$

The value function $V_{\gamma,\mu,M}$ has a closed form expression.

$$V_{\gamma,\mu,M} = (I - \gamma P_\mu)^{-1} R_\mu$$

Therefore, for all i, the map $\gamma \mapsto V_{\gamma,\mu,M}(i)$ is a rational function of γ. Two rational functions are either identical or intersect each other at a finite number of points. Further, the number of policies and the number of MDPs in $\mathcal{M}_{\text{opt}}$ is finite. Therefore, for each i, there exists $\gamma_i \in [0,1)$ such that no two functions in the set

$$\{\gamma \mapsto V_{\gamma,\mu,M}(i)\ :\ \mu : S \mapsto A,\ M \in \mathcal{M}_{\text{opt}}\}$$

intersect each other in the interval $(\gamma_i, 1)$. Let $\gamma_{\text{opt}} = \max_i \gamma_i$. By Theorem 13, there is an optimal policy, say μ_{opt}, such that

$$V^{\text{opt}}_{\gamma_{\text{opt}},\mu_{\text{opt}}} = \mathbf{V}^{\text{opt}}_{\gamma_{\text{opt}}}\ .$$

By Theorem 14, there is an MDP, say M_{opt}, in $\mathcal{M}_{\text{opt}}$ such that

$$V_{\gamma_{\text{opt}},\mu_{\text{opt}},M_{\text{opt}}} = V^{\text{opt}}_{\gamma_{\text{opt}},\mu_{\text{opt}}} = \mathbf{V}^{\text{opt}}_{\gamma_{\text{opt}}}\ . \tag{5.7}$$

We now claim that

$$V_{\gamma,\mu_{\text{opt}},M_{\text{opt}}} = \mathbf{V}^{\text{opt}}_{\gamma_{\text{opt}}}$$

for all $\gamma \in (\gamma_{\text{opt}}, 1)$. If not, there is an $\gamma' \in (\gamma_{\text{opt}}, 1)$, a policy μ' and an MDP $M' \in \mathcal{M}_{\text{opt}}$ such that

$$V_{\gamma',\mu_{\text{opt}},M_{\text{opt}}}(i) < V_{\gamma',\mu',M'}(i)$$

for some i. But this yields a contradiction, since (5.7) holds and by definition of γ_{opt}, the functions

$$\gamma \mapsto V_{\gamma,\mu_{\text{opt}},M_{\text{opt}}}(i)$$

and

$$\gamma \mapsto V_{\gamma,\mu',M'}(i)$$

cannot intersect in $(\gamma_{\text{opt}}, 1)$.

The proof of the existence of $\gamma_{\text{pes}}, \mu_{\text{pes}}$ and M_{pes} is based on similar arguments. □

5.5 Algorithms to Compute the Optimal Value Functions

5.5.1 Optimistic Value Function

The idea behind our algorithm (Algorithm 1) is to start with some initial vector and perform a sequence of updates while increasing the discount factor at a certain rate. The following theorem guarantees that the sequence of value functions thus generated converges to the optimal value function. Note that if we held the discount factor constant at some value, say γ, the sequence would converge to $(1-\gamma)\mathbf{V}^{\text{opt}}_{\gamma}$.

Algorithm 1 Algorithm to Compute $\mathbf{U}^{\text{opt}}$

$U^{(0)} \leftarrow \mathbf{0}$

for $k = 0, 1, \ldots$ **do**

$\gamma_k \leftarrow \frac{k+1}{k+2}$

for all $i \in S$ **do**

$U^{(k+1)}(i) \leftarrow \max_{a \in A} \left[(1-\gamma_k)R(i,a) + \gamma_k \max_{q \in \mathcal{C}_{i,a}} q^\top U^{(k)}\right]$

end for

end for

Theorem 17. *Let $\{U^{(k)}\}$ be the sequence of functions generated by Algorithm 1. Then we have, for all $i \in S$,*

$$\lim_{k \to \infty} U^{(k)}(i) = \mathbf{U}^{\text{opt}}(i) \ .$$

We need a few intermediate results before proving this theorem. Let γ_{opt}, μ_{opt} and M_{opt} be as given by Theorem 16. To avoid too many subscripts, let μ and M denote μ_{opt} and M_{opt} respectively for the remainder of this subsection. From (2.1), we have that for k large enough, say $k \geq k_1$, we have,

$$\left|(1-\gamma_k)V_{\gamma_k,\mu,M}(i) - (1-\gamma_{k+1})V_{\gamma_{k+1},\mu,M}(i)\right| \leq K(\gamma_{k+1} - \gamma_k) \ , \tag{5.8}$$

where K can be taken to be $\|h_{\mu,M}\|_\infty + 1$. Since $\gamma_k \uparrow 1$, we have $\gamma_k > \gamma_{\text{opt}}$ for all $k > k_2$ for some k_2. Let $k_0 = \max\{k_1, k_2\}$. Define

$$\delta_{k_0} := \|V^{(k_0)} - (1-\gamma_{k_0})V_{\gamma_{k_0},\mu,M}\|_\infty \ . \tag{5.9}$$

Since rewards are in $[0, R_{\max}]$, we have $\delta_{k_0} \leq R_{\max}$. For $k \geq k_0$, define δ_{k+1} recursively as

$$\delta_{k+1} := K(\gamma_{k+1} - \gamma_k) + \gamma_k \delta_k \ . \tag{5.10}$$

The following lemma shows that this sequence bounds the norm of the difference between $V^{(k)}$ and $V_{\gamma_k,\mu,M}$.

Lemma 18. *Let $\{U^{(k)}\}$ be the sequence of functions generated by Algorithm 1. Further, let μ, M denote $\mu_{\text{opt}}, M_{\text{opt}}$ mentioned in Theorem 16. Then, for $k \geq k_0$, we have*

$$\|U^{(k)} - (1-\gamma_k)V_{\gamma_k,\mu,M}\|_\infty \leq \delta_k \ .$$

Proof. Base case of $k = k_0$ is true by definition of δ_{k_0}. Now assume we have proved the claim till $k \geq k_0$. So we know that,

$$\max_i \left| U^{(k)}(i) - (1-\gamma_k)V_{\gamma_k,\mu,M}(i) \right| \leq \delta_k \ . \tag{5.11}$$

We wish to show

$$\max_i \left| U^{(k+1)}(i) - (1-\gamma_{k+1})V_{\gamma_{k+1},\mu,M}(i) \right| \leq \delta_{k+1} \ . \tag{5.12}$$

Recall that $\mathbf{V}_\gamma^{\text{opt}}$ is the fixed point of T_α^{opt} by Theorem 12. We therefore have, for all i,

$$\begin{aligned}
(1-\gamma_k)V_{\gamma_k,\mu,M}(i) &= (1-\gamma_k)\left(T_{\gamma_k}^{\text{opt}} V_{\gamma_k,\mu,M}\right)(i) \\
&\quad [\ \gamma_k > \gamma_{\text{opt}} \text{ and } V_{\gamma,\mu,M} = \mathbf{V}_\gamma^{\text{opt}} \text{ for } \gamma > \gamma_{\text{opt}} \ ] \\
&= \max_{a\in A}[\ (1-\gamma_k)R(i,a) + \gamma_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)(1-\gamma_k)V_{\gamma_k,\mu,M}(j) \ ] \\
&\quad [\ \text{defn. of } T_{\gamma_k}^{\text{opt}}] \\
&\leq \max_{a\in A}[\ (1-\gamma_k)R(i,a) + \gamma_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)U^{(k)}(j) \ ] + \gamma_k\delta_k \\
&\quad [\ (5.11) \text{ and } \sum_j q(j)\delta_k = \delta_k] \\
&= U^{(k+1)}(i) + \gamma_k\delta_k \ . \\
&\quad [\ \text{defn. of } U^{(k+1)}(i) \ ]
\end{aligned}$$

Similarly, for all i,

$$
\begin{aligned}
U^{(k+1)}(i) &= \max_{a\in A}[\ (1-\gamma_k)R(i,a) + \gamma_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)U^{(k)}(j)\ ] \\
&\quad [\text{ defn. of } U^{(k+1)}(i)] \\
&\le \max_{a\in A}[\ (1-\gamma_k)R(i,a) + \gamma_k \max_{q\in\mathcal{C}_{i,a}} \sum_j q(j)(1-\gamma_k)V_{\gamma_k,\mu,M}(j)\ ] + \gamma_k\delta_k \\
&\quad [\text{ (5.11) and } \sum_j q(j)\delta_k = \delta_k] \\
&= (1-\gamma_k)\left(T^{\text{opt}}_{\gamma_k} V_{\gamma_k,\mu,M}\right)(i) + \gamma_k\delta_k \\
&\quad [\text{ defn. of } T^{\text{opt}}_{\gamma_k}] \\
&= (1-\gamma_k)V_{\gamma_k,\mu,M}(i) + \gamma_k\delta_k\ . \\
&\quad [\ \gamma_k > \gamma_{\text{opt}} \text{ and } V_{\gamma,\mu,M} = \mathbf{V}^{\text{opt}}_\gamma \text{ for } \gamma > \gamma_{\text{opt}}\ ]
\end{aligned}
$$

Thus, for all i,

$$\left|U^{(k+1)}(i) - (1-\gamma_k)V_{\gamma_k,\mu,M}(i)\right| \le \gamma_k\delta_k\ .$$

Combining this with (5.8) (as $k \ge k_0 \ge k_1$), we get

$$\left|U^{(k+1)}(i) - (1-\gamma_{k+1})V_{\gamma_{k+1},\mu,M}(i)\right| \le \gamma_k\delta_k + K(\gamma_{k+1}-\gamma_k)\ .$$

Thus we have shown (5.12). □

The sequence $\{\delta_k\}$ can be shown to converge to zero using elementary arguments.

Lemma 19. *The sequence* $\{\delta_k\}$ *defined for* $k \ge k_0$ *by equations* (5.9) *and* (5.10) *converges to* 0.

Proof. Plugging $\gamma_k = \frac{k+1}{k+2}$ into the definition of δ_{k+1} we get,

$$\begin{aligned}\delta_{k+1} &= K\left(\frac{k+2}{k+3} - \frac{k+1}{k+2}\right) + \frac{k+1}{k+2}\delta_k \\ &= \frac{K}{(k+3)(k+2)} + \frac{k+1}{k+2}\delta_k \ .\end{aligned}$$

Applying the recursion again for δ_k, we get

$$\begin{aligned}\delta_{k+1} &= \frac{K}{(k+3)(k+2)} + \frac{k+1}{k+2}\left(\frac{K}{(k+2)(k+1)} + \frac{k}{k+1}\delta_{k-1}\right) \\ &= \frac{K}{k+2}\left(\frac{1}{k+3} + \frac{1}{k+2}\right) + \frac{k}{k+2}\delta_{k-1} \ .\end{aligned}$$

Continuing in this fashion, we get for any $j \geq 0$,

$$\delta_{k+1} = \frac{K}{k+2}\left(\frac{1}{k+3} + \frac{1}{k+2} + \ldots + \frac{1}{k-j+3}\right) + \frac{k-j+1}{k+2}\delta_{k-j} \ .$$

Setting $j = k - k_0$ above, we get

$$\delta_{k+1} = \frac{K}{k+2}\left(H_{k+3} - H_{k_0+2}\right) + \frac{k_0+1}{k+2}\delta_{k_0} \ ,$$

where $H_n = 1 + \frac{1}{2} + \ldots + \frac{1}{n}$. This clearly tends to 0 as $k \to \infty$ since $H_n = O(\log n)$ and $\delta_{k_0} \leq R_{\max}$. □

We can now prove Theorem 17.

Proof. (*of Theorem 17*) Fix $i \in S$. We have,

$$\begin{aligned}|U^{(k)}(i) - \mathbf{U}^{\text{opt}}(i)| \leq &\underbrace{|U^{(k)}(i) - V_{\gamma_k,\mu,M}(i)|}_{\leq \delta_k} + \underbrace{|V_{\gamma_k,\mu,M}(i) - \mathbf{V}^{\text{opt}}_{\gamma_k}(i)|}_{\epsilon_k} \\ &+ \underbrace{|\mathbf{V}^{\text{opt}}_{\gamma_k}(i) - \mathbf{U}^{\text{opt}}(i)|}_{\zeta_k} \ .\end{aligned}$$

We use Lemma 18 to bound the first summand on the right hand side by δ_k. By Lemma 19, $\delta_k \to 0$. Also, $\epsilon_k = 0$ for sufficiently large k because $\gamma_k \uparrow 1$ and $V_{\gamma,\mu,M}(i) = \mathbf{V}^{\text{opt}}_{\gamma}(i)$ for γ sufficiently close to 1 (by Theorem 16). Finally, $\zeta_k \to 0$ by Theorem 15. □

5.5.2 Pessimistic Value Function

Algorithm 2 Algorithm to Compute $\mathbf{U}^{\text{pes}}$

$V^{(0)} \leftarrow \mathbf{0}$

for $k = 0, 1, \ldots$ **do**

$\gamma_k \leftarrow \frac{k+1}{k+2}$

for all $i \in S$ **do**

$U^{(k+1)}(i) \leftarrow \max_{a \in A} \left[(1-\gamma_k) R(i,a) + \gamma_k \min_{q \in \mathcal{C}_{i,a}} q^\top U^{(k)} \right]$

end for

end for

Algorithm 2 is the same as Algorithm 1 except that the max over $\mathcal{C}_{i,a}$ appearing inside the innermost loop gets replaced by a min. The following analogue of Theorem 17 holds.

Theorem 20. *Let $\{U^{(k)}\}$ be the sequence of functions generated by Algorithm 2. Then we have, for all $i \in S$,*

$$\lim_{k \to \infty} U^{(k)}(i) = \mathbf{U}^{\text{pes}}(i) \ .$$

To prove this theorem, we repeat the argument given in the previous subsection with appropriate changes. Let γ_{pes}, μ_{pes} and M_{pes} be as given by Theorem 16. For the remainder of this subsection, let μ and M denote μ_{pes} and M_{pes} respectively. Let k_1, k_2 be large enough so that, for all $k \geq k_1$,

$$\left| V_{\gamma_k, \mu, M}(i) - V_{\gamma_{k+1}, \mu, M}(i) \right| \leq K(\gamma_{k+1} - \gamma_k) \ ,$$

for some constant K (which depends on μ, M), and $\gamma_k > \gamma_{\text{pes}}$ for $k > k_2$. Set $k_0 = \max\{k_1, k_2\}$ and define the sequence $\{\delta_k\}_{k \geq k_0}$ as before (equations (5.9) and (5.10)).

The proof of the following lemma can be obtained from that of Lemma 18 by fairly straightforward changes and is therefore omitted.

Lemma 21. *Let $\{U^{(k)}\}$ be the sequence of functions generated by Algorithm 2. Further, let μ, M denote $\mu_{\text{pes}}, M_{\text{pes}}$ mentioned in Theorem 16. Then, for $k \geq k_0$, we have*

$$\|U^{(k)} - (1-\gamma_k)V_{\gamma_k,\mu,M}\|_\infty \leq \delta_k \ .$$

Theorem 20 is now proved in exactly the same fashion as Theorem 17 and we therefore omit the proof.

5.6 Semi-algebraic Constraints

In this section, we consider a generalization of BMDPs. A BMDP consists of all MDPs satisfying certain constraints on their parameters. The constraints had the form

$$l(i,j,a) \leq P_{i,a}(j) \leq u(i,j,a) \ .$$

These are simple inequality constraints with two hyperparameters $l(i,j,a)$ and $u(i,j,a)$. The set of possible values of $P_{i,a}(j)$ is an interval, a very simple semi-algebraic set. A semi-algebraic set is one that can be expressed as a finite boolean combination of polynomial equalities and inequalities. For a formal definition, we refer the reader to the book by Benedetti and Risler [6]. Some examples of semi-algebraic sets are as follows.

$$\{(x,y) \ : \ 2 \leq x^2 + y^2 \leq 5\} \subseteq \mathbb{R}^2$$

$$\{(x,y,z) \ : \ x^2 > 6yz\} \subseteq \mathbb{R}^3$$

$$\{(x,y) \ : \ xy > 0 \ \wedge \ (x^2 + y^3 < 1 \ \vee \ x^5 + y > 3)\} \subseteq \mathbb{R}^2$$

Note that any semi-algebraic set has a finite description and thus can be given explicitly. For each $i \in S, a \in A$, let $C(i,a)$ be a given closed semi-algebraic set. Define,

$$\mathcal{C}_{i,a} = \{q \in \mathbb{R}_+^{|S|} \ : \ q^\top \mathbf{1} = 1 \ \wedge \ q \in C(i,a)\}$$

Assume that all the sets $\mathcal{C}_{i,a}$ are non-empty. Note that $\mathcal{C}_{i,a}$ is a compact semi-algebraic set. A semi-algebraically constrained MDP (SAMDP) $\mathcal{M}$ is a collection of MDPs with the above constraints,

$$\mathcal{M} = \{\langle S, A, P, R\rangle \ : \ \forall i, a, P_{i,a} \in \mathcal{C}_{i,a}\} \ .$$

Note that any BMDP is an SAMDP. We can extend the definitions (5.1) through (5.4) to SAMDPs. We do not know of algorithms to compute these value functions for SAMDPs. However, Blackwell optimal policies can be shown to exist even in the case of SAMDPs thanks to the Tarski-Seidenberg theorem. This theorem can be stated in several different ways. We choose a form most useful for us in the present context. For more details, see Section 2.3 in [6].

Theorem (Tarski-Seidenberg). *Suppose $n > 1$ and $Q \in \mathbb{R}^n$ is a semi-algebraic set. Then the following two sets in $\mathbb{R}^{n-1}$ are also semi-algebraic,*

$$\{(x_2, \ldots, x_n) \ : \ \exists x_1 \ s.t. \ (x_1, x_2, \ldots, x_n) \in Q\} \ ,$$

$$\{(x_2, \ldots, x_n) \ : \ \forall x_1, \ (x_1, x_2, \ldots, x_n) \in Q\} \ .$$

We can now state the main result of this section.

Theorem 22. *Consider an SAMDP $\mathcal{M}$ and extend the definitions* (5.1) *through* (5.4) *to $\mathcal{M}$. Then, there exist $\gamma_{\text{opt}} \in (0,1)$ and a policy μ_{opt} such that*

$$\forall \gamma \in (\gamma_{\text{opt}}, 1), \ V_{\gamma,\mu_{\text{opt}}}^{\text{opt}} = \mathbf{V}_\gamma^{\text{opt}} \ .$$

Similarly, there exist $\gamma_{\text{pes}} \in (0,1)$ *and a policy* μ_{pes} *such that*

$$\forall \gamma \in (\gamma_{\text{pes}}, 1),\ V^{\text{pes}}_{\gamma,\mu_{\text{pes}}} = \mathbf{V}^{\text{pes}}_{\gamma}\ .$$

Proof. Let μ_1, μ_2 be two deterministic stationary policies. Let $i \in S$. We show that there exists $\gamma' = \gamma'(i, \mu_1, \mu_2) < 1$ such that the proposition

$$V^{\text{opt}}_{\gamma,\mu_1}(i) \geq V^{\text{opt}}_{\gamma,\mu_2}(i)$$

has the same truth value for all $\gamma \in (\gamma', 1)$. Since

$$V^{\text{opt}}_{\gamma,\mu_j}(i) = \sup_{M \in \mathcal{M}} (1-\gamma)(I - \gamma P_{\mu_j})^{-1} R_{\mu_j}$$

for $j \in \{1,2\}$, we see that $\gamma \mapsto V^{\text{opt}}_{\gamma,\mu_j}(i)$ can be written as

$$V^{\text{opt}}_{\gamma,\mu_j} = \sup_{\mathbf{p} \in Q} \frac{g_j(\gamma, \mathbf{p})}{h_j(\gamma, \mathbf{p})}$$

where Q is a semi-algebraic set, g_j, h_j are polynomials and $\mathbf{p}$ is a long vector denoting all the transition probabilities. Now, $V^{\text{opt}}_{\gamma,\mu_1}(i) \geq V^{\text{opt}}_{\gamma,\mu_2}(i)$ precisely when γ is in the set

$$\{\gamma\ :\ \exists \mathbf{p} \in Q \text{ s.t. } \forall \mathbf{q} \in Q,\ g_1(\gamma, \mathbf{p}) h_2(\gamma, \mathbf{q}) \geq g_2(\gamma, \mathbf{q}) h_1(\gamma, \mathbf{p})\}$$

$$= \{\gamma\ :\ \exists \mathbf{p} \forall \mathbf{q},\ \mathbf{p} \in Q\ \wedge\ (\mathbf{q} \in Q \Rightarrow g_1(\gamma, \mathbf{p}) h_2(\gamma, \mathbf{q}) \geq g_2(\gamma, \mathbf{q}) h_1(\gamma, \mathbf{p})\}$$

Let us call the above set T. Since Q is semi-algebraic and $g_1 h_2 \geq g_2 h_1$ is a polynomial inequality, we can use the Tarski-Seidenberg theorem to claim that that $T \subseteq \mathbb{R}$ is semi-algebraic. A semi-algebraic set in $\mathbb{R}$ is either empty or is a finite union of intervals. In either case, there exists $\gamma' < 1$ such that the proposition $\gamma \in T$ has the same truth value for all $\gamma \in (\gamma', 1)$.

To finish the proof, set $\gamma_{\text{opt}} = \max_{i,\mu_1,\mu_2} \gamma'(i, \mu_1, \mu_2)$. Let μ_{opt} be an optimal policy for discount factor γ_{opt}. By the choice of γ_{opt}, it remains optimal for all discount factors in $(\gamma_{\text{opt}}, 1)$.

The proof for the existence of γ_{pes} and μ_{pes} uses similar arguments. □

Before we close this chapter, a couple of remarks are in order. We chose to represent the uncertainty in the parameters of an MDP by intervals. One can ask whether similar results can be derived for other representations. If the intervals for $P_{i,a}(j)$ are equal for all j then our representation corresponds to an L_∞ ball around a probability vector. It will be interesting to investigate cases where we consider balls defined by other metrics. We can also consider sets defined by inequalities involving non-metrics like relative entropy (for an example of an algorithm using sets defined by relative entropy, see [11]).

Our last remark is regarding the convergence rate of the algorithms given in Section 5.5. Examining the proofs, one can verify that the number of iterations required to get to within ϵ accuracy is $O(\frac{1}{\epsilon})$. This is a pseudo-polynomial convergence rate. It might be possible to obtain algorithms where the number of iterations required to achieve ϵ-accuracy is $\text{poly}(\log \frac{1}{\epsilon})$.

Chapter 6

Logarithmic Regret Bound for Irreducible MDPs

In this chapter, we present an algorithm called Optimistic Linear Programming (OLP) for learning to optimize average reward in an irreducible but otherwise unknown MDP. OLP uses its experience so far to estimate the MDP. It chooses actions by optimistically maximizing estimated future rewards over a set of next-state transition probabilities that are close to the estimates, a computation that corresponds to solving linear programs. We show that the total expected reward obtained by OLP up to time T is within $C(P)\log T$ of the reward obtained by the optimal policy, where $C(P)$ is an explicit, MDP-dependent constant. OLP is closely related to an algorithm proposed by Burnetas and Katehakis with four key differences: OLP is simpler, it *does not* require knowledge of the supports of transition probabilities, the proof of the regret bound is simpler, but our regret bound is a constant factor larger than the regret of their algorithm. OLP is also similar in flavor to

an algorithm recently proposed by Auer and Ortner [5]. But OLP is simpler and its regret bound has a better dependence on the size of the MDP.

6.1 The Exploration-Exploitation Trade-off and Regret

Given complete knowledge of the parameters of an MDP, there are standard algorithms to compute optimal policies. A frequent criticism of these algorithms is that they assume an explicit description of the MDP which is seldom available. The parameters constituting the description are themselves estimated by simulation or experiment and are thus not known with complete reliability. Taking this into account brings us to the well known *exploration vs. exploitation* trade-off. On one hand, we would like to *explore* the system as well as we can to obtain reliable knowledge about the system parameters. On the other hand, if we keep exploring and never *exploit* the knowledge accumulated, we will not behave optimally.

Given a policy π, how do we measure its ability to handle this trade-off? Suppose the agent gets a numerical reward at each time step and we measure performance by the accumulated reward over time. Then, a meaningful quantity to evaluate the policy π is its *regret* over time. To understand what regret means, consider an omniscient agent who knows all parameters of the MDP accurately and behaves optimally. Let V_T be the expected reward obtained by this agent up to time T. Let V_T^π denote the corresponding quantity for π. Then the regret $R_T^\pi = V_T - V_T^\pi$ measures how much π is hurt due to its incomplete knowledge of the MDP up to time T. If we can show that the regret R_T^π grows slowly with time T, for all MDPs in a sufficiently big class, then we can safely conclude that π is making

a judicious trade-off between exploration and exploitation. It is rather remarkable that for this notion of regret, logarithmic bounds have been proved in the literature [5, 11]. This means that there are policies π with $R_T^\pi = O(\log T)$. Thus the per-step regret R_T^π/T goes to zero very quickly.

Call a policy *uniformly good* if for all MDPs and all $\epsilon > 0$, $R_T^\pi = o(T^\epsilon)$ as $T \to \infty$. Burnetas and Katehakis [11] proved that for any uniformly good policy π, $R_T^\pi \geq C_B(P) \log T$ where they identified the constant $C_B(P)$. This constant depends on the transition function P of the MDP. They also gave an algorithm (we call it BKA) that achieves this rate and is therefore optimal in a very strong sense. However, besides assuming that the MDP is irreducible (see Assumption 4 below) they assumed that the support sets of the transition distributions $P_{i,a}$ are known for all state-action pairs. We not only get rid of this assumption but our optimistic linear programming (OLP) algorithm is also computationally simpler. At each step, OLP considers certain parameters in the vicinity of the estimates. Like BKA, OLP makes *optimistic* choices among these. But now, making these choices only involves solving *linear programs* (LPs) to maximize linear functions over L_1 balls. BKA instead required solving non-linear (though convex) programs due to the use of KL-divergence. Another benefit of using the L_1 distance is that it greatly simplifies a significant part of the proof. The price we pay for these advantages is that the regret of OLP is $C(P) \log T$ asymptotically, for a constant $C(P) \geq C_B(P)$.

A number of algorithms in the literature have been inspired by the *optimism in the face of uncertainty* principle [3, 4, 10, 22, 27]. One such algorithm is that of Auer and Ortner (we refer to it as AOA) and it achieves logarithmic regret for irreducible MDPs. AOA does

not solve an optimization problem at every time step but only when a confidence interval is halved. But then the optimization problem they solve is more complicated because they find a policy to use in the next few time steps by optimizing over a *set of MDPs*. The regret of AOA is $C_A(P) \log T$ where

$$C_A(P) = c\frac{|S|^5|A|T_w(P)\kappa(P)^2}{\Delta^*(P)^2} \ ,$$

for some universal constant c. Here $|S|, |A|$ denote the state and action space size, $T_w(P)$ is the worst case hitting time over deterministic policies (see Eqn. (6.7)) and $\Delta^*(P)$ is the difference between the long term average return of the best policy and that of the next best policy. The constant $\kappa(P)$ is also defined in terms of hitting times. Under Auer and Ortner's assumption of bounded rewards, we can show that the constant for OLP satisfies

$$C(P) \leq \frac{2|S||A|T(P)^2}{\Phi^*(P)} \ .$$

Here $T(P)$ is the hitting time of an optimal policy is therefore necessarily smaller than $T_w(P)$. We get rid of the dependence on $\kappa(P)$ while replacing $T_w(P)$ with $T(P)^2$. Most importantly, we significantly improve the dependence on the state space size. The constant $\Phi^*(P)$ can roughly be thought of as the minimum (over states) difference between the quality of the best and the second best action (see Eqn. (6.5)). The constants $\Delta^*(P)$ and $\Phi^*(P)$ are similar though not directly comparable. Nevertheless, note that $C(P)$ depends inversely on $\Phi^*(P)$ not $\Phi^*(P)^2$.

6.2 The Irreducibility Assumption

Consider an MDP $\langle S, \mathcal{A}, P, R\rangle$ where S is the set of states, $\mathcal{A} = \cup_{i\in S} A(i)$ is the set of actions. Note that in this chapter, we allow for the possibility of having different actions available at different states: $A(i)$ being the actions available in state i. For simplicity of analysis, we assume that the rewards are known to us beforehand. We *do not* assume that we know the support sets of the distributions $P_{i,a}$.

The definitions given in Chapter 2 can easily be extended to the case of state-dependent action spaces in the following way. The *history* σ_t up to time t is a sequence $(i_0, k_0, \ldots, i_{t-1}, k_{t-1}, i_t)$ such that $k_s \in A(i_s)$ for all $s < t$. A *policy* π is a sequence $\{\pi_t\}$ of probability distributions on A given σ_t such that $\pi_t(A(s_t)|\sigma_t) = 1$ where s_t denotes the random variable representing the state at time t. The set of all policies is denoted by Π. A *deterministic stationary* policy is simply a function $\mu : S \to \mathcal{A}$ such that $\mu(i) \in A(i)$. Denote the set of deterministic stationary policies by Π_D. If $\mathcal{D}$ is a subset of $\mathcal{A}$, let $\Pi(\mathcal{D})$ denote the set of policies that take actions in $\mathcal{D}$. Given history σ_t, define the counts,

$$
\begin{aligned}
N_t(i) &:= \sum_{t'=0}^{t-1} \mathbf{1}\left[i_t = i\right] \ , \\
N_t(i,a) &:= \sum_{t'=0}^{t-1} \mathbf{1}\left[(i_t, k_t) = (i,a)\right] \ , \\
N_t(i,a,j) &:= \sum_{t'=0}^{t-1} \mathbf{1}\left[(i_i, k_t, i_{t+1}) = (i,a,j)\right] \ .
\end{aligned}
$$

If $M = \langle S, \mathcal{A}, P, R\rangle$ and $S, \mathcal{A}$ and R are clear from context, we will denote the expectation operator $\mathbb{E}_i^{\pi,M}[\cdot]$ by $\mathbb{E}_i^{\pi,P}[\cdot]$.

We make the following irreducibility assumption regarding the MDP.

Assumption 4. For all $\mu \in \Pi_D$, the transition matrix $P^\mu = \left(P_{i,\mu(i)}(j)\right)_{i,j\in S}$ is irreducible (i.e. it is possible to reach any state from any other state).

Fix an MDP M and consider the rewards accumulated by the policy π before time T,

$$V_T^\pi(i_0, P) := \mathbb{E}_{i_0}^{\pi,P}\left[\sum_{t=0}^{T-1} R(s_t, a_t)\right] ,$$

where a_t is the random variable representing the action taken by π at time t. Let $V_T(i_0, P)$ be the maximum possible sum of expected rewards before time T,

$$V_T(i_0, P) := \sup_{\pi\in\Pi} V_T^\pi(i_0, P) .$$

The *regret* of a policy π at time T is a measure of how well the expected rewards of π compare with the above quantity,

$$R_T^\pi(i_0, P) := V_T(i_0, P) - V_T^\pi(i_0, P) .$$

Define the long term average reward of a policy π as

$$\lambda_\pi(P) := \liminf_{T\to\infty} \frac{V_T^\pi(i_0, P)}{T} .$$

Under Assumption 4, the above limit exists and is independent of the starting state i_0. Given a restricted set $\mathcal{D} \subseteq \mathcal{A}$ of actions, the *gain* or the best long term average performance is

$$\lambda(P, \mathcal{D}) := \sup_{\pi\in\Pi(\mathcal{D})} \lambda_\pi(P) .$$

As a shorthand, define $\lambda^*(P) := \lambda(P, \mathcal{A})$.

6.3 Optimality Equations and Critical Pairs

A restricted problem $(P, \mathcal{D})$ is obtained from the original MDP by choosing subsets $D(i) \subseteq A(i)$ and setting $\mathcal{D} = \cup_{i \in S} D(i)$. The transition and reward functions of the restricted problems are simply the restrictions of P and r to $\mathcal{D}$. Assumption 4 implies that there is a *bias* vector $h(P, \mathcal{D}) = \{h(i; P, \mathcal{D})\}_{i \in S}$ such that the gain $\lambda(P, \mathcal{D})$ and bias $h(P, \mathcal{D})$ are the unique solutions to the *average reward optimality equations*:

$$\forall i \in S,\ \lambda(P, \mathcal{D}) + h(i; P, D) = \max_{a \in D(i)} [R(i, a) + P_{i,a}^{\top} h(P, \mathcal{D})] \ . \tag{6.1}$$

We will use $h^*(P)$ to denote $h(P, \mathcal{A})$. Also, denote the infinity norm $\|h^*(P)\|_\infty$ by $H^*(P)$. Note that if $h^*(P)$ is a solution to the optimality equations and $\mathbf{e}$ is the vector of ones, then $h^*(P) + c\mathbf{e}$ is also a solution for any scalar c. We can therefore assume $\exists i^* \in S, h^*(i^*; P) = 0$ without any loss of generality.

It will be convenient to have a way to denote the quantity inside the 'max' that appears in the optimality equations. Accordingly, define

$$\mathcal{L}(i, a, p, h) := R(i, a) + p^{\top} h \ ,$$

$$\mathcal{L}^*(i; P, \mathcal{D}) := \max_{a \in D(i)} \mathcal{L}(i, a, P_{i,a}, h(P, D)) \ .$$

To measure the degree of suboptimality of actions available at a state, define

$$\phi^*(i, a; P) = \mathcal{L}^*(i; P, \mathcal{A}) - \mathcal{L}(i, a, P_{i,a}, h^*(P)) \ .$$

Note that the optimal actions are precisely those for which the above quantity is zero.

$$O(i; P, \mathcal{D}) := \{a \in D(i) \ : \ \phi^*(i, a; P) = 0\} \ ,$$

$$\mathcal{O}(P, \mathcal{D}) := \Pi_{i \in S} O(i; P, \mathcal{D}) \ .$$

Any policy in $\mathcal{O}(P, \mathcal{D})$ is an optimal policy, i.e.,

$$\forall \mu \in \mathcal{O}(P, \mathcal{D}),\ \lambda_\mu(P) = \lambda(P, \mathcal{D}) \ .$$

From now on, Δ^+ will denote the probability simplex of dimension determined by context. For a suboptimal action $a \notin O(i; P, \mathcal{A})$, the following set contains probability distributions q such that if $P_{i,a}$ is changed to q, the quality of action a comes within ϵ of an optimal action. Thus, q *makes* a look almost *optimal*:

$$\text{MakeOpt}(i, a; P, \epsilon) := \{q \in \Delta^+ \ :\ \mathcal{L}(i, a, q, h^*(P)) \geq \mathcal{L}^*(i; P, \mathcal{A}) - \epsilon\} \ .$$

Those suboptimal state-action pairs for which MakeOpt is never empty, no matter how small ϵ is, play a crucial role in determining the regret. We call these *critical* state-action pairs,

$$\text{Crit}(P) := \{(i, a) \ :\ a \notin O(i; P, \mathcal{A}) \wedge (\forall \epsilon > 0,\ \text{MakeOpt}(i, a; P, \epsilon) \neq \emptyset)\} \ .$$

Define the function,

$$J_{i,a}(p; P, \epsilon) := \inf\{\|p - q\|_1^2 \ :\ q \in \text{MakeOpt}(i, a; P, \epsilon)\} \ . \tag{6.2}$$

To make sense of this definition, consider $p = P_{i,a}$. The above infimum is then the least distance (in the L_1 sense) one has to move away from $P_{i,a}$ to make the suboptimal action a look ϵ-optimal. Taking the limit of this as ϵ decreases gives us a quantity that also plays a crucial role in determining the regret,

$$K(i, a; P) := \lim_{\epsilon \to 0} J_{i,a}(P_{i,a}; P, \epsilon) \ . \tag{6.3}$$

Intuitively, if $K(i, a; P)$ is small, it is easy to confuse a suboptimal action with an optimal one and so it should be difficult to achieve small regret. The constant that multiplies $\log T$

in the regret bound of our algorithm OLP (see Algorithm 3 and Theorem 25 below) is the following:

$$C(P) := \sum_{(i,a)\in \mathrm{Crit}(P)} \frac{2\phi^*(i,a;P)}{K(i,a;P)} \ . \tag{6.4}$$

This definition might look a bit hard to interpret, so we give an upper bound on $C(P)$ just in terms of the infinity norm $H^*(P)$ of the bias and $\Phi^*(P)$. This latter quantity is defined below to be the minimum degree of suboptimality of a critical action.

Proposition 23. *Suppose $A(i) = A$ for all $i \in S$. Define*

$$\Phi^*(P) := \min_{(i,a)\in \mathrm{Crit}(P)} \phi^*(i,a;P) \ . \tag{6.5}$$

Then, for any P,

$$C(P) \leq \frac{2|S||A|H^*(P)^2}{\Phi^*(P)} \ .$$

Proof. Fix $(i,a) \in \mathrm{Crit}(P)$. Drop dependence of ϕ^*, K, MakeOpt, H^* on i, a, P for readability. Also, let $J(\epsilon) = J_{i,a}(P_{i,a}; P, \epsilon)$. Since $|\mathrm{Crit}(P)| \leq |S||A|$, it suffices to show that $\phi^*/K \leq (H^*)^2/\phi^*$. Let $\epsilon < \phi^*$ be arbitrary. By definition of MakeOpt(ϵ) and ϕ^*, we have, for all $q \in \mathrm{MakeOpt}(\epsilon)$, $q^\top h^* - P_{i,a}^\top h^* \geq \phi^* - \epsilon$. This implies $\|P_{i,a} - q\|_1 \geq (\phi^* - \epsilon)/H^*$ since $H^* = \|h^*\|_\infty$. Thus, by definition of $J(\epsilon)$, we have

$$J(\epsilon) \geq \frac{(\phi^* - \epsilon)^2}{(H^*)^2} \qquad \Rightarrow \qquad \lim_{\epsilon \to 0} J(\epsilon) \geq \frac{(\phi^*)^2}{(H^*)^2} \ .$$

By definition, the left hand side is K. Thus, $\phi^*/K \leq (H^*)^2/\phi^*$. □

6.4 Hitting Times

It turns out that we can bound the infinity norm of the bias in terms of the hitting time of an optimal policy. For any policy μ define its *hitting time* to be the worst case

expected time to reach one state from another:

$$T_\mu(P) := \max_{i \neq j} \mathbb{E}_j^{\mu,P}[\min\{t > 0 \ : \ s_t = i\}] \ .$$

The following constant is the minimum hitting time among optimal policies:

$$T(P) := \min_{\mu \in \mathcal{O}(P,\mathcal{D})} T_\mu(P) \ . \tag{6.6}$$

The following constant is defined just for comparison with results in [5]. It is the worst case hitting time over all policies:

$$T_w(P) := \max_{\mu \in \Pi_D} T_\mu(P) \ . \tag{6.7}$$

We can now bound $C(P)$ just in terms of the hitting time $T(P)$ and $\phi^*(P)$.

Proposition 24. *Suppose $A(i) = A$ for all $i \in S$. Then for any P,*

$$C(P) \leq \frac{2|S||A|R_{\max}^2 T(P)^2}{\Phi^*(P)} \ .$$

Proof. Fix $\mu \in \mathcal{O}(P, \mathcal{A})$. Drop the dependence of $h^*, H^*, \lambda^*, T_\mu$ on P. It suffices to prove that $H^* \leq T_\mu$ for the result then follows from Proposition 2 and definition of $T(P)$. Since rewards and hence the gain λ^* are in $[0, R_{\max}]$, and λ^*, h^* satisfy, for all $i \in S$,

$$\lambda^* + h^*(i) = R(i, \mu(i)) + P_{i,\mu(i)}^\top h^* \ ,$$

we have, for all $i \in S$,

$$P_{i,\mu(i)}^\top h^* \geq h^*(i) - R_{\max} \ . \tag{6.8}$$

Start the policy μ in state j and define the random variables $Y_t := h(s_t)$. Clearly $Y_0 = h(j)$. Fix a state $i \neq j$. Define the stopping time

$$\tau := \min\{t > 0 \ : \ s_t = i\} \ .$$

Because of (6.8), we have

$$\mathbb{E}_j^{\mu,P}[Y_{t+1} \mid Y_t] \geq Y_t - R_{\max} \ .$$

Adding $R_{\max}(t+1)$ to both sides we see that $Y_t + R_{\max}t$ is a submartingale and hence using the optional stopping theorem (see, for example, page 489 in [19]), we have

$$\mathbb{E}_j^{\mu,P}[Y_\tau + R_{\max}\tau] \geq Y_0 \ .$$

By definition of T_μ, we have

$$\mathbb{E}_j^{\mu,P}[\tau] \leq T_\mu \ .$$

Thus, noting that $Y_\tau = h^*(i)$ and $Y_0 = h^*(j)$, we have

$$h^*(i) + R_{\max}T_\mu \geq h^*(j) \ .$$

But this is true for all $i \neq j$. Also, there is some $i^* \in S$ such that $h^*(i^*) = 0$. Therefore, $H^* = \|h^*\|_\infty \leq R_{\max}T_\mu$. □

6.5 The Optimistic LP Algorithm and its Regret Bound

Algorithm 3 is the Optimistic Linear Programming algorithm. It is inspired by the algorithm of Burnetas and Katehakis [11] but uses L_1 distance instead of KL-divergence. At each time step t, the algorithm computes the empirical estimates for transition probabilities. It then forms a restricted problem ignoring relatively undersampled actions. An action $a \in A(i)$ is considered "undersampled" if $N_t(i,a) < \log^2 N_t(i)$. The solutions $h(\hat{P}^t, \mathcal{D}_t), \lambda(\hat{P}^t, \mathcal{D}_t)$ might be misleading due to estimation errors. To avoid being misled by empirical samples we compute optimistic "indices" $U_t(s_t, a)$ for all legal actions $a \in A(s_t)$

Algorithm 3 Optimistic Linear Programming

for $t = 0, 1, 2, \ldots$ **do**

  $s_t \leftarrow$ current state

  ▷ Compute solution for "empirical MDP" excluding "undersampled" actions

  $\forall i, j \in S, a \in A(i),\ \hat{p}^t_{i,j}(a) \leftarrow \frac{1+N_t(i,a,j)}{|A(i)|+N_t(i,a)}$

  $\forall i \in S, D_t(i) \leftarrow \{a \in A(i)\ :\ N_t(i,a) \geq \log^2 N_t(i)\}$

  Solve the equations (6.1) with $P = \hat{P}^t, \mathcal{D} = \mathcal{D}_t$

  ▷ Compute indices of *all* actions for the current state

  $\forall a \in A(s_t),\ U_t(s_t, a) \leftarrow \sup_{q \in \Delta^+}\{R(s_t,a) + q^\top h(\hat{P}^t, \mathcal{D}_t)\ :\ \|\hat{p}^t_{s_t}(a) - q\|_1 \leq \sqrt{\frac{2\log t}{N_t(s_t,a)}}\}$

  ▷ Optimal actions (for the current problem) that are about to become "undersampled"

  $\Gamma^1_t \leftarrow \{a \in O(s_t; \hat{P}^t, \mathcal{D}_t)\ :\ N_t(s_t,a) < \log^2(N_t(s_t)+1)\}$

  ▷ The index maximizing actions

  $\Gamma^2_t \leftarrow \arg\max_{a \in A(s_t)} U_t(s_t, a)$

  if $\Gamma^1_t = O(s_t; \hat{P}^t, \mathcal{D}_t)$ **then**

    $a_t \leftarrow$ any action in Γ^1_t

  else

    $a_t \leftarrow$ any action in Γ^2_t

  end if

end for

where s_t is the current state. The index for action a is computed by looking at an L_1-ball around the empirical estimate $\hat{p}^t_{s_t}(a)$ and choosing a probability distribution q that maximizes $\mathcal{L}(i, a, q, h(\hat{P}^t, \mathcal{D}_t))$. Note that if the estimates were perfect, we would take an action maximizing $\mathcal{L}(i, a, \hat{p}^t_{s_t}(a), h(\hat{P}^t, \mathcal{D}_t))$. Instead, we take an action that maximizes the index. There is one case where we are forced not to take an index-maximizing action. It is when all the optimal actions of the current problem are about to become undersampled at the next time step. In that case, we are forced to take one of these actions (steps 16–20). Note that both steps 6 and 9 can be done by solving LPs. The LP for solving optimality equations can be found in several textbooks (see, for example, page 391 in [26]). The LP in step 9 is even simpler: the L_1 ball has only $2|S|$ vertices and so we can maximize over them efficiently.

Like the original Burnetas-Katehakis algorithm, the modified one also satisfies a logarithmic regret bound as stated in the following theorem. Unlike the original algorithm, OLP does not need to know the support sets of the transition distributions.

Theorem 25. *Let π_0 denote the policy implemented by Algorithm 3. Then we have, for all $i_0 \in S$ and for all P satisfying Assumption 4,*

$$\limsup_{T\to\infty} \frac{R^{\pi_0}_T(i_0, P)}{\log T} \leq C(P) \ ,$$

where $C(P)$ is the MDP-dependent constant defined in (6.4).

Proof. From Proposition 1 in [11], it follows that

$$R^{\pi_0}_T(i_0, P) = \sum_{i\in S} \sum_{a\notin O(i;P,\mathcal{A})} \mathbb{E}^{\pi_0,P}_{i_0}[N_T(i, a)]\phi^*(i, a; P) + O(1) \ . \tag{6.9}$$

Define the event

$$A_t := \{\|h(\hat{P}^t, \mathcal{D}_t) - h^*(P)\|_\infty \leq \epsilon \wedge \mathcal{O}(\hat{P}^t, \mathcal{D}_t) \subseteq \mathcal{O}(P)\} \ . \tag{6.10}$$

Define,

$$N_T^1(i,a;\epsilon) := \sum_{t=0}^{T-1} \mathbf{1}\left[(s_t,a_t) = (i,a) \wedge A_t \wedge U_t(i,a) \geq \mathcal{L}^*(i;P,\mathcal{A}) - 2\epsilon\right] \ ,$$

$$N_T^2(i,a;\epsilon) := \sum_{t=0}^{T-1} \mathbf{1}\left[(s_t,a_t) = (i,a) \wedge A_t \wedge U_t(i,a) < \mathcal{L}^*(i;P,\mathcal{A}) - 2\epsilon\right] \ ,$$

$$N_T^3(\epsilon) := \sum_{t=0}^{T-1} \mathbf{1}\left[\bar{A}_t\right] \ ,$$

where $\bar{A}_t$ denotes the complement of A_t. For all $\epsilon > 0$,

$$N_T(i,a) \leq N_T^1(i,a;\epsilon) + N_T^2(i,a;\epsilon) + N_T^3(\epsilon) \ . \tag{6.11}$$

The result then follows by combining (6.9) and (6.11) with the following three propositions and then letting $\epsilon \to 0$ sufficiently slowly. □

Proposition 26. *For all P and $i_0 \in S$, we have*

$$\lim_{\epsilon \to 0} \limsup_{T \to \infty} \sum_{i \in S} \sum_{a \notin O(i;P,\mathcal{A})} \frac{\mathbb{E}_{i_0}^{\pi_0,P}[N_T^1(i,a;\epsilon)]}{\log T} \phi^*(i,a;P) \leq C(P) \ .$$

Proposition 27. *For all P, $i_0, i \in S$, $a \notin O(i;P,\mathcal{A})$ and ϵ sufficiently small, we have*

$$\mathbb{E}_{i_0}^{\pi_0,P}[N_T^2(i,a;\epsilon)] = o(\log T) \ .$$

Proposition 28. *For all P satisfying Assumption 4, $i_0 \in S$ and $\epsilon > 0$, we have*

$$\mathbb{E}_{i_0}^{\pi_0,P}[N_T^3(\epsilon)] = o(\log T) \ .$$

6.6 Proofs of Auxiliary Propositions

In this section, we prove Propositions 26, 27 and 28. The proof of Proposition 27 is considerably simpler (because of the use of L_1 distance rather than KL-divergence) than the analogous Proposition 4 in [11].

Proof of Proposition 26. There are two cases depending on whether $(i,a) \in \mathrm{Crit}(P)$ or not. If $(i,a) \notin \mathrm{Crit}(P)$, there is an $\epsilon_0 > 0$ such that $\mathrm{MakeOpt}(i,a;P,\epsilon_0) = \emptyset$. On the event A_t (recall the definition given in (6.10)), we have $|q^\top h(\hat{P}^t, \mathcal{D}_t) - q^\top h^*(P)| \le \epsilon$ for any $q \in \Delta^+$. Therefore,

$$\begin{aligned}
U_t(i,a) &\le \sup_{q \in \Delta^+} \{R(i,a) + q^\top h(\hat{P}^t, \mathcal{D}_t)\} \\
&\le \sup_{q \in \Delta^+} \{R(i,a) + q^\top h^*(P)\} + \epsilon \\
&< \mathcal{L}^*(i;P,\mathcal{A}) - \epsilon_0 + \epsilon \qquad [\because \mathrm{MakeOpt}(i,a;P,\epsilon_0) = \emptyset] \\
&< \mathcal{L}^*(i;P,\mathcal{A}) - 2\epsilon \text{ provided that } 3\epsilon < \epsilon_0
\end{aligned}$$

Therefore for $\epsilon < \epsilon_0/3$, $N_T^1(i,a;\epsilon) = 0$.

Now suppose $(i,a) \in \mathrm{Crit}(P)$. The event $U_t(i,a) \ge \mathcal{L}^*(i;P,\mathcal{A}) - 2\epsilon$ is equivalent to

$$\exists q \in \Delta^+ \text{ s.t. } \left(\|\hat{p}_i^t(a) - q\|_1^2 \le \frac{2\log t}{N_t(i,a)}\right) \wedge \left(R(i,a) + q^\top h(\hat{P}^t, \mathcal{D}_t) \ge \mathcal{L}^*(i;P,\mathcal{A}) - 2\epsilon\right) .$$

On the event A_t, we have $|q^\top h(\hat{P}^t, \mathcal{D}_t) - q^\top h^*(P)| \le \epsilon$ and thus the above implies

$$\exists q \in \Delta^+ \text{ s.t. } \left(\|\hat{p}_i^t(a) - q\|_1^2 \le \frac{2\log t}{N_t(i,a)}\right) \wedge \left(R(i,a) + q^\top h^*(P) \ge \mathcal{L}^*(i;P,\mathcal{A}) - 3\epsilon\right) .$$

Recalling the definition (6.2) of $J_{i,a}(p;P,\epsilon)$, we see that this implies

$$J_{i,a}(\hat{p}_i^t(a);P,3\epsilon) \le \frac{2\log t}{N_t(i,a)} .$$

We therefore have,

$$\begin{aligned} N_T^1(i,a;\epsilon) &\leq \sum_{t=0}^{T-1} \mathbf{1}\left[(s_t,a_t) = (i,a) \wedge J_{i,a}(\hat{p}_i^t(a);P,3\epsilon) \leq \frac{2\log t}{N_t(i,a)}\right] \\ &\leq \sum_{t=0}^{T-1} \mathbf{1}\left[(s_t,a_t) = (i,a) \wedge J_{i,a}(P_{i,a};P,3\epsilon) \leq \frac{2\log t}{N_t(i,a)} + \delta\right] \\ &\quad + \sum_{t=0}^{T-1} \mathbf{1}\left[(s_t,a_t) = (i,a) \wedge J_{i,a}(P_{i,a};P,3\epsilon) > J_{i,a}(\hat{p}_i^t(a);P,3\epsilon) + \delta\right] \end{aligned} \tag{6.12}$$

where $\delta > 0$ is arbitrary. Each time the pair (i,a) occurs $N_t(i,a)$ increases by 1, so the first count is no more than

$$\frac{2\log T}{J_{i,a}(P_{i,a};P,3\epsilon) - \delta}\ . \tag{6.13}$$

To control the expectation of the second sum, note that continuity of $J_{i,a}$ in its first argument implies that there is a function f such that $f(\delta) > 0$ for $\delta > 0$, $f(\delta) \to 0$ as $\delta \to 0$ and $J_{i,a}(P_{i,a};P,3\epsilon) > J_{i,a}(\hat{p}_i^t(a);P,3\epsilon)+\delta$ implies that $\|P_{i,a}-\hat{p}_i^t(a)\|_1 > f(\delta)$. By a Chernoff-type bound [29], we have, for some constant C_1,

$$\mathbb{P}_{i_0}^{\pi_0,P}[\|P_{i,a} - \hat{p}_i^t(a)\|_1 > f(\delta) \,|\, N_t(i,a) = m] \leq C_1 \exp(-mf(\delta)^2)\ .$$

and so the expectation of the second sum is no more than

$$\mathbb{E}_{i_0}^{\pi_0,P}[\sum_{t=0}^{T-1} C_1 \exp(-N_t(i,a)f(\delta)^2)] \leq \sum_{m=1}^{\infty} C_1 \exp(-mf(\delta)^2) = \frac{C_1}{1-\exp(-f(\delta)^2)}\ . \tag{6.14}$$

Combining the bounds (6.13) and (6.14) and plugging them into (6.12), we get

$$\mathbb{E}_{i_0}^{\pi_0,P}[N_T^1(i,a;\epsilon)] \leq \frac{2\log T}{J_{i,a}(P_{i,a};P,3\epsilon) - \delta} + \frac{C_1}{1-\exp(-f(\delta)^2)}\ .$$

Letting $\delta \to 0$ sufficiently slowly, we get that for all $\epsilon > 0$,

$$\mathbb{E}_{i_0}^{\pi_0,P}[N_T^1(i,a;\epsilon)] \leq \frac{2\log T}{J_{i,a}(P_{i,a};P,3\epsilon)} + o(\log T)\ .$$

Therefore,

$$\lim_{\epsilon\to 0}\limsup_{T\to\infty}\frac{\mathbb{E}_{i_0}^{\pi_0,P}[N_T^1(i,a;\epsilon)]}{\log T}\leq\lim_{\epsilon\to 0}\frac{2}{J_{i,a}(P_{i,a};P,3\epsilon)}=\frac{2}{K(i,a;P)}\ ,$$

where the last equality follows from the definition (6.3) of $K(i,a;P)$. The result now follows by summing over (i,a) pairs in $\mathrm{Crit}(P)$. □

Proof of Proposition 27. Define the event

$$A_t'(i,a;\epsilon):=\{(s_t,a_t)=(i,a)\wedge A_t\wedge U_t(i,a)<\mathcal{L}^*(i;P,\mathcal{A})-2\epsilon\}\ ,$$

so that we can write

$$N_T^2(i,a;\epsilon)=\sum_{t=0}^{T-1}\mathbf{1}\left[A_t'(i,a;\epsilon)\right]\ . \tag{6.15}$$

Note that on $A_t'(i,a;\epsilon)$, we have $\Gamma_t^1\subseteq O(i;\hat{P}^t,\mathcal{D}_t)\subseteq O(i;P,\mathcal{A})$. So, $a\notin O(i;P,\mathcal{A})$. But a was taken at time t, so it must have been in Γ_t^2 which means it maximized the index. Therefore, for all optimal actions $a^*\in O(i;P,\mathcal{A})$, we have, on the event $A_t'(i,a;\epsilon)$,

$$U_t(i,a^*)\leq U_t(i,a)<\mathcal{L}^*(i;P,\mathcal{A})-2\epsilon\ .$$

Since $\mathcal{L}^*(i;P,\mathcal{A})=R(i,a^*)+P_{i,a^*}^\top h^*(P)$, this implies

$$\forall q\in\Delta^+,\ \|q-\hat{p}_i^t(a^*)\|_1\leq\sqrt{\frac{2\log t}{N_t(i,a^*)}}\Rightarrow q^\top h(\hat{P}^t,\mathcal{D}_t)<P_{i,a^*}^\top h^*(P)-2\epsilon\ .$$

Moreover, on the event A_t, $|q^\top h(\hat{P}^t,\mathcal{D}_t)-q^\top h^*(P)|\leq\epsilon$. We therefore have, for any

$a^* \in O(i; P, \mathcal{A})$,

$$\begin{aligned}
A'_t(i,a;\epsilon) &\subseteq \left\{ \forall q \in \Delta^+,\ \|q - \hat{p}_i^t(a)\|_1 \le \sqrt{\frac{2\log t}{N_t(i,a)}} \Rightarrow q^\top h^*(P) < P_{i,a}^\top h^*(P) - \epsilon \right\} \\
&\subseteq \left\{ \forall q \in \Delta^+,\ \|q - \hat{p}_i^t(a)\|_1 \le \sqrt{\frac{2\log t}{N_t(i,a)}} \Rightarrow \|q - P_{i,a}\|_1 > \frac{\epsilon}{\|h^*(P)\|_\infty} \right\} \\
&\subseteq \left\{ \|\hat{p}_i^t(a) - P_{i,a}\|_1 > \frac{\epsilon}{h^*(P)} + \sqrt{\frac{2\log t}{N_t(i,a)}} \right\} \\
&\subseteq \bigcup_{m=1}^{t} \left\{ N_t(i,a) = m \wedge \|\hat{p}_i^t(a) - P_{i,a}\|_1 > \frac{\epsilon}{\|h^*(P)\|_\infty} + \sqrt{\frac{2\log t}{N_t(i,a)}} \right\}
\end{aligned}$$

Using a Chernoff-type bound, we have, for some constant C_1,

$$\mathbb{P}_{i_0}^{\pi_0,P}[\|\hat{p}_i^t(a) - P_{i,a}\|_1 > \delta \,|\, N_t(i,a) = m] \le C_1 \exp(-m\delta^2/2) \ .$$

Using a union bound, we therefore have,

$$\begin{aligned}
\mathbb{P}_{i_0}^{\pi_0,P}[A'_t(i,a;\epsilon)] &\le \sum_{m=1}^{t} C_1 \exp\left(-\frac{m}{2} \left(\frac{\epsilon}{\|h^*(P)\|_\infty} + \sqrt{\frac{2\log t}{m}} \right)^2 \right) \\
&\le \frac{C_1}{t} \sum_{m=1}^{\infty} \exp\left(-\frac{m\epsilon^2}{2\|h^*(P)\|_\infty^2} - \frac{\epsilon\sqrt{2m\log t}}{\|h^*(P)\|_\infty} \right) = o\left(\frac{1}{t}\right) \ .
\end{aligned}$$

Combining this with (6.15) proves the result. □

Before we prove Proposition 28, we need a few lemmas. Let m denote the number of deterministic stationary policies. For large enough t, we divide the first time interval up to time t into $m+1$ subintervals. Interval number ν is denoted by I_ν^t. The event B_t defined below says that after the first interval, all states have been visited often enough and the transition probability estimates are sufficiently accurate. Lemmas 30–33 work out different consequences of the event B_t. Once these lemmas are in place, we can prove the proposition.

The precise definitions of the intervals are as follows. Fix $\beta < 1/(m+1)$. For sufficiently large t, define the subintervals

$$I_\nu^t := \{t' \,:\, b_\nu^t \le t' < b_{\nu+1}^t\} \ ,$$

where $b_0^t = 0$ and $b_\nu^t = t - (m+1-\nu)\lfloor k/(m+1)\rfloor$. Note that, by our choice of β, the length of each subinterval I_ν^t is at least βt and thus $b_\nu^t > \nu\beta t$. The number of times the state action pair (i,a) and the state j is visited during the interval I_ν^t is denoted by $\Delta_\nu(i,a)$ and $\Delta_\nu(j)$ respectively.

Define the events,

$$\begin{aligned} B_t(\zeta) := \Big\{&\forall j \in S, \forall \nu \in [m], \forall t' \ge b_1^t, \forall a \in D_{t'}(j), \\ &\Delta_\nu(j) > \rho\beta t \wedge \|\hat{p}_j^{t'}(a) - P_{j,a}\|_\infty \le \zeta \Big\} \ , \end{aligned} \tag{6.16}$$

$$\begin{aligned} C_\nu^t(\delta) := \Big\{&\forall t' \in I_\nu^t,\ h(P, \mathcal{D}_{t'}) = h(P, \mathcal{D}_{t'+1}) \ \Rightarrow \\ &\forall j, a, t' \in I_\nu^t,\ U_{t'}(j,a) > \mathcal{L}(j, a; P_{j,a}, h(P; \mathcal{D}_{t'})) - \delta \Big\} \ , \end{aligned} \tag{6.17}$$

$$\begin{aligned} G_t := \Big\{ &\Big(\forall \nu \ge 1,\ \lambda(P, \mathcal{D}_{b_{\nu+1}^t}) > \lambda(P, \mathcal{D}_{b_\nu^t})\Big) \\ &\vee \Big(\exists \nu \text{ s.t. } \lambda(P, \mathcal{D}_{b_{\nu+1}^t}) = \lambda(P, \mathcal{D}_{b_\nu^t}) = \lambda^*(P)\Big)\Big\} \ . \end{aligned} \tag{6.18}$$

We quote the following result from the Burnetas and Katehakis paper [11].

Proposition 29. *For all P satisfying Assumption 4 there exist $A > 0$ and $\rho_0 > 0$ such that for all $i_0 \in S, t \ge |S|, \rho > 0$ and any policy π,*

$$\mathbb{P}_{i_0}^{\pi,P}[N_t(i) \le \rho t] \le A \exp((\rho - \rho_0)t) \ .$$

As a consequence, for all $\rho < \rho_0$, $i_0 \in S$ and any policy π,

$$\mathbb{P}_{i_0}^{\pi,P}[N_t(i) \le \rho t] = o(1/t)$$

as $k \to \infty$.

The following two lemmas ensure that certain events happen under the event B_t. Their proofs can be found in Appendix A, Section A.2.

Lemma 30. *For all* $\delta, \epsilon > 0$, $\exists \zeta_1, \zeta_2, \zeta_3 > 0, t_0$ *such that the following hold true. For all* $\zeta < \zeta_1$, $B_t(\zeta) \subseteq$

$$\{\forall t' \geq b_1^t, \forall \mu \in \Pi(\mathcal{D}_{t'}),\ |\lambda_\mu(\hat{P}^{t'}) - \lambda_\mu(P)| \leq \epsilon,\ \|h_\mu(\hat{P}^{t'}) - h_\mu(P)\|_\infty \leq \epsilon\}$$

For all $\zeta < \zeta_2$, $B_t(\zeta) \subseteq$

$$\{\forall t' \geq b_1^t, \mathcal{O}(\hat{P}^{t'}, \mathcal{D}_{t'}) \subseteq \mathcal{O}(P, \mathcal{D}_{t'}) \ \wedge\ \|h(\hat{P}^{t'}, \mathcal{D}_{t'}) - h(P, \mathcal{D}_{t'})\|_\infty \leq \epsilon\}$$

For all $\zeta < \zeta_3, t > t_0$, $B_t(\zeta) \subseteq$

$$\{\forall t' \geq b_1^t, j \in S, a \in D_{t'}(j),\ U_{t'}(j,a) \leq \mathcal{L}(j, a; P_{j,a}, h(P, \mathcal{D}_{t'})) + \delta\} \tag{6.19}$$

Lemma 31. *On the event* $B_t(\zeta)$, *the following hold true. Let*

$$\tau = \rho\beta t/(2\log(\rho\beta t)) - 1\ .$$

For $t' > b_1^t, j \in S, a \in A(j)$, *if* $(s_{t'}, a_{t'}) = (j, a)$ *with* $a \in \Gamma^1_{t'}$, *then for all* ζ,

$$\sum_{l=0}^{\tau} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge\ a \in \Gamma^1_{t'+l}\right] \leq 1 \tag{6.20}$$

For all $t' \geq b_1^t$ *and sufficiently small* ζ,

$$\lambda(P, \mathcal{D}_{t'+1}) \geq \lambda(P, \mathcal{D}_{t'})\ . \tag{6.21}$$

Lemma 32. *For all* $\delta > 0$, $\exists \zeta'$ *such that for all* $\zeta < \zeta'$, $\nu \in [m]$,

$$\mathbb{P}_{i_0}^{\pi_0, P}[B_t(\zeta)\overline{C_\nu^t(\delta)}] = o(1/t)\ .$$

Proof. By Lemma 30, we know that for sufficiently small ζ, on the event $B_t(\zeta)$, we have, for $t' \geq b_1^t$, $\|h(\hat{P}^{t'}, \mathcal{D}_{t'}) - h(P, \mathcal{D}_{t'}\|_\infty \leq \delta/2$. Therefore, on the event $B_t(\zeta)\overline{C_\nu^t(\delta)}$, we have, for some $j, a, t' \in I_\nu^t$,

$$\forall q \in \Delta^+, \|q - \hat{p}_j^{t'}(a)\|_1 \leq \sqrt{\frac{2\log t'}{N_{t'}(j,a)}} \Rightarrow q^\top h(P, \mathcal{D}_{t'}) \leq P_{j,a}^\top h(P, \mathcal{D}_{t'}) - \delta/2$$

On the event $B_t(\zeta)\overline{C_\nu^t(\delta)}$, we also have $h(P, \mathcal{D}_{t'}) = h(P, \mathcal{D}_{b_\nu^t})$ for all $t' \in I_\nu^t$. Moreover, $t' \geq b_\nu^t$ which allows us to write the above event as

$$\forall q \in \Delta^+, \|q - \hat{p}_j^{t'}(a)\|_1 \leq \sqrt{\frac{2\log b_\nu^t}{N_{t'}(j,a)}} \Rightarrow q^\top h(P, \mathcal{D}_{b_\nu^t}) \leq P_{j,a}^\top h(P, \mathcal{D}_{b_\nu^t}) - \delta/2$$

Since $h(P, \mathcal{D}_{b_\nu^t}) = h_\mu(P)$ for some $\mu \in \Pi_D$ and the sets $\Pi_D, S, A(j)$ are finite, it suffices to bound the probability of the following event for fixed $\mu, j, a \in A(j)$: $\exists t' \in I_\nu^t$ such that

$$\forall q \in \Delta^+, \|q - \hat{p}_j^{t'}(a)\|_1 \leq \sqrt{\frac{2\log b_\nu^t}{N_{t'}(j,a)}} \Rightarrow q^\top h_\mu(P) \leq P_{j,a}^\top h_\mu(P) - \delta/2 \ .$$

Since $q^\top h_\mu(P) \leq P_{j,a}^\top h_\mu(P) - \delta/2$ implies $\|q - P_{j,a}\|_1 \geq \delta/(2\|h_\mu(P)\|_\infty)$, it suffices to consider the event: $\exists t' \in I_\nu^t$ such that

$$\|\hat{p}_j^{t'}(a) - P_{j,a}\|_1 \geq \sqrt{\frac{2\log b_\nu^t}{N_{t'}(j,a)}} + \frac{\delta}{2\|h_\mu(P)\|_\infty} \ .$$

Conditioning on the possible values $1, \ldots, b_{\nu+1}^t$ of $N_{t'}(j,a)$ and using a bound [29] for the L_1 deviation of $\hat{p}_j^{t'}(a)$ from $P_{j,a}$, we can bound probability of the above event by

$$\begin{aligned}
&\sum_{m=1}^{b_{\nu+1}^t} C_1 \exp\left(-\frac{m}{2}\left(\sqrt{\frac{2\log b_\nu^t}{m}} + \frac{\delta}{2\|h_\mu(P)\|_\infty}\right)^2\right) \\
&\leq \frac{C_1}{b_\nu^t} \sum_{m=1}^{\infty} \exp\left(-\frac{\delta\sqrt{2m\log b_\nu^t}}{2\|h_\mu(P)\|_\infty} - \frac{\delta^2}{8\|h_\mu(P)\|_\infty^2}\right) \\
&= o\left(\frac{1}{b_\nu^t}\right)
\end{aligned}$$

for some constant C_1. Noting that $b_\nu^t \geq \nu\beta t$ for all $\nu \in [m]$ finishes the proof. $\square$

Lemma 33. *For all sufficiently small ζ,*

$$\mathbb{P}_{i_0}^{\pi_0,P}[B_t(\zeta)\overline{G_t}] = o(1/t)$$

Proof. We assume that there exist two policies $\mu, \mu' \in \Pi_D$ such that $\lambda_\mu(P) \neq \lambda_{\mu'}(P)$ because otherwise the lemma holds trivially. This ensures that the constant δ_{gap} defined below is strictly positive. Define the event,

$$F_\nu^t := \left\{\lambda(P;\mathcal{D}_{b_{\nu+1}^t}) = \lambda(P;\mathcal{D}_{b_\nu^t}) < \lambda^*(P)\right\} \ . \tag{6.22}$$

Under $B_t(\zeta)$ we have,

$$B_t(\zeta)\overline{G_t} \subseteq B_t(\zeta)\bigcup_{\nu=1}^{m} F_\nu^t \tag{6.23}$$

because of (6.21). Having obtained the solution of a restricted problem $(P,\mathcal{D})$, we define the set of improving actions as

$$\text{Impr}(j;P,\mathcal{D}) := \{a \in A(j) \ : \ \lambda(P,\mathcal{D}) + h(j;P,\mathcal{D}) < \mathcal{L}(j,a;P_{j,a},h(j;P,\mathcal{D}))\} \ .$$

When $\mathcal{D}$ is a specific policy μ, the set $\text{Impr}(j;P,\mu)$ is simply

$$\{a \in A(j) \ : \ \lambda_\mu(P) + h_\mu(j;P) < \mathcal{L}(j,a;P_{j,a},h_\mu(j;P))\} \ .$$

Define the constant,

$$\delta_{\text{gap}} := \min_{\mu,j,a}\{\mathcal{L}(j,a;P_{j,a},h_\mu(j;P)) - \lambda_\mu(P) - h_\mu(j;P)\} \ , \tag{6.24}$$

where μ ranges over Π_D, j ranges over states such that $\text{Impr}(j;P,\mu) \neq \emptyset$ and a ranges over $\text{Impr}(j;P,\mu)$. We now claim that it suffices to show that if ζ is sufficiently small, then for large t and all $\nu \in [m]$,

$$B_t(\zeta)C_\nu^t(\delta_{\text{gap}}/2)F_\nu^t = \emptyset \ . \tag{6.25}$$

Indeed if this holds then we have,

$$\begin{aligned}
\mathbb{P}_{i_0}^{\pi_0,P}[B_t(\zeta)\overline{G_t}] &\le \sum_{\nu=1}^{m} \mathbb{P}_{i_0}^{\pi_0,P}[B_t(\zeta)F_\nu^t] && \text{[by (6.23)]} \\
&\le \sum_{\nu=1}^{m} \mathbb{P}_{i_0}^{\pi_0,P}[B_t(\zeta)\overline{C_\nu^t(\delta_{\text{gap}}/2)}] && \text{[by (6.25)]} \\
&= o(1/t) && \text{[by Lemma 32]}
\end{aligned}$$

To prove (6.25), it suffices to show that

$$B_t(\zeta)C_\nu^t(\delta_{\text{gap}}/2)F_\nu^t \subseteq \{\exists j \text{ s.t. } \Delta_\nu(j) \ge \rho\beta t \ \wedge \ \Delta_\nu(j) = O(\log^2 t)\} \ . \tag{6.26}$$

This is because the last event cannot happen for large t.

The rest of the proof is devoted to proving (6.26). Under F_ν^t, there is a state j such that an improving action exists for j, i.e. $\text{Impr}(j; P, \mathcal{D}_{b_\nu^t}) \neq \emptyset$. Fix such a j and for any $a \in A(j)$, split the occurrences of (j,a) into two groups,

$$\Delta_\nu^{(1)}(j,a) := \sum_{t' \in I_\nu^t} \mathbf{1}\left[(s_{t'}, a_{t'}) = (j,a) \ \wedge \ a \in \Gamma_{t'}^1\right]$$

$$\Delta_\nu^{(2)}(j,a) := \sum_{t' \in I_\nu^t} \mathbf{1}\left[(s_{t'}, a_{t'}) = (j,a) \ \wedge \ a \in \Gamma_{t'}^2\right]$$

and write

$$\Delta_\nu(j) = \sum_{a \in A(j)} \Delta^{(1)}(j,a) + \sum_{a \in \text{Impr}(j;P,\mathcal{D}_{b_\nu^t})} \Delta^{(2)}(j,a) + \sum_{a \notin \text{Impr}(j;P,\mathcal{D}_{b_\nu^t})} \Delta^{(2)}(j,a) \ .$$

We bound each term on the right hand side separately.

First term. It is easy to bound $\Delta^{(1)}(j,a)$ using (6.20). The event being counted occurs at most once in a time period of length $\tau + 1$ and so

$$\Delta_\nu^{(1)}(j,a) \le (b_{\nu+1}^t - b_\nu^t)/(\tau+1) = O(\log t) \ . \tag{6.27}$$

SECOND TERM. To bound the rest of the terms, we first need to show that under $B_t(\zeta)F_\nu^t$, for all $t' \in I_\nu^t$,

$$\lambda(P; \mathcal{D}_{t'}) = \lambda(P; \mathcal{D}_{t'+1}) \tag{6.28}$$

$$h(P; \mathcal{D}_{t'}) = h(P; \mathcal{D}_{t'+1}) \tag{6.29}$$

$$\mathrm{Impr}(j; P, \mathcal{D}_{t'}) = \mathrm{Impr}(j; P, \mathcal{D}_{t'+1}) \tag{6.30}$$

The definition of F_ν^t and (6.21) immediately imply (6.28). Consider a time t' with $s_t = i$. Then, from the definition of the policy π_0, we have

$$\forall i' \neq i,\ D_{t'}(i') = D_{t'+1}(i') \ . \tag{6.31}$$

Thus, $\lambda(P, \mathcal{D}_{t'})$, $h(P, \mathcal{D}_{t'})$ satisfy the optimality equations of $(P, \mathcal{D}_{t'+1})$ for all states $i' \neq i$. Furthermore, we established before that, under $B_t(\zeta)$,

$$D_{t'+1}(i) \cap O(i; P, \mathcal{D}_t) \neq \emptyset \ .$$

Therefore, if $\lambda(P, \mathcal{D}_{t'})$ and $h(P, \mathcal{D}_{t'})$ did not satisfy the optimality equation of $(P, \mathcal{D}_{t'+1})$ for state i, we would have

$$\lambda(P, \mathcal{D}_{t'}) + h(i; P, \mathcal{D}_{t'}) < \max_{a \in D_{t'+1}(i)} \mathcal{L}(i, a; P, h(P, \mathcal{D}_{t'})) \ .$$

Together with (6.31) this implies that there exists $\mu \in \mathcal{D}_{t'+1}$ such that

$$\lambda(P; \mathcal{D}_{t'+1}) \geq \lambda_\mu(P; \mathcal{D}_{t'+1}) > \lambda(P; \mathcal{D}_{t'}) \ ,$$

which contradicts (6.28). It therefore follows that $\lambda(P, \mathcal{D}_{t'})$ and $h(P, \mathcal{D}_{t'})$ do satisfy the optimality equations of $(P, \mathcal{D}_{t'+1})$. Hence $h(P, \mathcal{D}_{t'}) = h(P, \mathcal{D}_{t'+1})$ and (6.29) holds. Equation (6.30) is an immediate consequence of (6.28) and (6.29).

For $a \in \mathrm{Impr}(j; P, \mathcal{D}_{b^t_\nu})$, (6.30) implies that $a \in \mathrm{Impr}(j; P, \mathcal{D}_{b^t_{\nu+1}})$ and hence $a \notin \mathcal{D}_{b^t_{\nu+1}}$ by definition of improving actions. Therefore, for such an a, $\Delta^{(2)}(j,a)$ can bounded as follows.

$$\begin{aligned}
\Delta^{(2)}_\nu(j,a) &\le \Delta_\nu(j,a) \le N_{b^t_{\nu+1}}(j,a) \\
&< \log^2 b^t_{\nu+1} && [\because a \notin \mathcal{D}_{b^t_{\nu+1}}] \\
&\le \log^2 t\ . && (6.32)
\end{aligned}$$

Third term. Next we prove that for sufficiently small ζ and large enough t, on the events $B_t(\zeta) C^t_\nu(\delta_{\mathrm{gap}}/2) F^t_\nu$ and $(s_t, a_t) = (j,a)$, we have

$$D_{t'}(j) \cap \Gamma^2_{t'} = \emptyset \tag{6.33}$$

for all $t' \in I^t_\nu$. To see this, note that on the event $B_t(\zeta) F^t_\nu$, we have for all $t' \in I^t_\nu, a \in D_{t'}(j)$ and large t,

$$\begin{aligned}
U_{t'}(j,a) &\le \mathcal{L}(j,a; P_{j,a}, h(P; \mathcal{D}_{t'})) + \delta_{\mathrm{gap}}/2 && [\text{by (6.19)}] \\
&\le \lambda(P; \mathcal{D}_{t'}) + h(j, P; \mathcal{D}_{t'}) + \delta_{\mathrm{gap}}/2 && [\text{by optimality eqns.}] \\
&= \lambda(P; \mathcal{D}_{b^t_\nu}) + h(j, P; \mathcal{D}_{b^t_\nu}) + \delta_{\mathrm{gap}}/2 && [\text{by (6.28), (6.29)}] \qquad (6.34)
\end{aligned}$$

Moreover, for any $a' \in \mathrm{Impr}(j; P, \mathcal{D}_{b^t_\nu})$, on the event $B_t(\zeta) C^t_\nu(\delta_{\mathrm{gap}}/2) F^t_\nu$, we have, for any $t' \in I^t_\nu$,

$$\begin{aligned}
U_{t'}(j,a') &> \mathcal{L}(j,a'; P_{j,a'}, h(P; \mathcal{D}_{t'}) - \delta_{\mathrm{gap}}/2 && [\because C^t_\nu, (6.29) \text{ are true}] \\
&= \mathcal{L}(j,a'; P_{j,a'}, h(P; \mathcal{D}_{b^t_\nu}) - \delta_{\mathrm{gap}}/2 && [\text{by (6.29)}] \\
&\ge \lambda(P; \mathcal{D}_{b^t_\nu}) + h(j, P; \mathcal{D}_{b^t_\nu}) + \delta_{\mathrm{gap}} - \delta_{\mathrm{gap}}/2 && [\text{defn. of } \delta_{\mathrm{gap}}] \qquad (6.35)
\end{aligned}$$

Combining (6.34) and (6.35), we see that no action in $D_{t'}(j)$ can maximize the index and hence be in Γ_t^2. Thus, (6.33) is proved.

We can finally bound $\Delta_\nu^{(2)}(j,a)$ for $a \notin \text{Impr}(j; P, \mathcal{D}_{b_\nu^t})$.

$$\begin{aligned}
\Delta_\nu^{(2)}(j,a) &\leq \sum_{t' \in I_\nu^t} \mathbf{1}\left[(s_{t'}, a_{t'}) = (j,a) \ \wedge \ a \notin D_{t'}(j)\right] && \text{[by (6.33)]} \\
&\leq \sum_{t' \in I_\nu^t} \mathbf{1}\left[(s_{t'}, a_{t'}) = (j,a) \ \wedge \ N_{t'}(j,a) < \log^2 t'\right] \\
&\leq \sum_{t' \in I_\nu^t} \mathbf{1}\left[(s_{t'}, a_{t'}) = (j,a) \ \wedge \ N_{t'}(j,a) < \log^2 b_{\nu+1}^t\right] && [\because t' \leq b_{\nu+1}^t] \\
&\leq \log^2 b_{\nu+1}^t < \log^2 t \ . && (6.36)
\end{aligned}$$

Combining (6.27),(6.32) and (6.36), we see that, under the event

$$B_t(\zeta) C_\nu^t(\delta_{\text{gap}}/2) F_\nu^t \ ,$$

we have $\Delta_\nu(j) = O(\log^2 t)$. Moreover, under $B_t(\zeta)$, $\Delta_\nu(j) \geq \rho\beta t$. Thus (6.26) is proved. □

We can now finally prove Proposition 28.

Proof. Fix $\epsilon > 0$. It suffices to establish the following two claims.

$$\forall \epsilon > 0, \ \exists \zeta_1 \text{ s.t. } B_t(\zeta_1) G_t \subseteq A_t \ , \tag{6.37}$$

and

$$\mathbb{P}_{i_0}^{\pi_0, P}[\overline{B_t(\zeta_1) G_t}] = o(1/t) \ . \tag{6.38}$$

We first prove (6.37). Recall that m is the total number of deterministic stationary policies. Since monotonic increase of $\lambda(P, \mathcal{D}_{b_\nu^t})$ is not possible for more than m intervals I_ν^t, it must be the case that at the end of the last interval, we have $\mathcal{D}_t \cap \mathcal{O}(P) \neq \emptyset$ and thus,

$$G_t \subseteq \{\mathcal{O}(P, \mathcal{D}_t) \subseteq \mathcal{O}(P), \ \lambda(P, \mathcal{D}_t) = \lambda^*(P), \ h(P, \mathcal{D}_t) = h^*(P)\} \ . \tag{6.39}$$

Moreover, by Lemma (30), part 2, we have, for sufficiently small ζ,

$$B_t(\zeta) \subseteq \{\mathcal{O}(\hat{P}^t, \mathcal{D}_t) \subseteq \mathcal{O}(P, \mathcal{D}_t),\ \|h(\hat{P}^t, \mathcal{D}_t) - h(P, \mathcal{D}_t)\| < \epsilon\} \tag{6.40}$$

Combining (6.39) and (6.40) and recalling the definition of the event A_t given in (6.10) proves (6.37).

To prove (6.38), write

$$\mathbb{P}_{i_0}^{\pi_0,P}[\overline{B_t(\zeta_1)G_t}] \leq \mathbb{P}_{i_0}^{\pi_0,P}[B_t(\zeta_1)\overline{G_t}] + \mathbb{P}_{i_0}^{\pi_0,P}[\overline{B_t(\zeta_1)}]$$

Lemma 33 shows that the first term is $o(1/t)$. We now show that the second term is also $o(1/t)$.

Using the definition of $B_t(\zeta_1)$ and the union bound we have,

$$\begin{aligned}\mathbb{P}_{i_0}^{\pi_0,P}[\overline{B_t(\zeta_1)}] \leq \sum_{j\in S}\sum_{\nu=0}^{m} \mathbb{P}_{i_0}^{\pi_0,P}[\Delta_\nu(j) \leq \rho\beta t] \\ + \sum_{j\in S} \mathbb{P}_{i_0}^{\pi_0,P}[\forall \nu \in [m], \Delta_\nu(j) > \rho\beta t \ \wedge\ \exists t' \geq b_1^t, a \in D_{t'}(j) \text{ s.t.} \\ \|\hat{p}_j^{t'}(a) - P_{j,a}\|_\infty > \zeta_1] \ .\end{aligned}$$

Since the length of the interval I_ν^t is at least βt, it follows from Proposition 29 that $\mathbb{P}_{i_0}^{\pi_0,P}[\Delta_\nu(j) \leq \rho\beta t]$ is $o(1/t)$. To bound the second term, note that on the event $\{\forall \nu \in [m],\ \Delta_\nu(j) > \rho\beta t\}$, we have $N_{t'}(j) \geq \rho\beta t \geq \rho\beta t'$ for all $t' \geq b_1^t$. Therefore, the possible values of $N_{t'}(j)$ are $\rho\beta t, \ldots, t$. Given that $N_{t'}(j) = q$, the possible values of $N_{t'}(j,a)$ for $a \in D_{t'}(j)$ are $\log^2 q, \ldots, q$. Using a union bound, we can now bound the second term as

follows (where $A = \max_i |A(i)|$).

$$\sum_{j\in S}\sum_{t'=b_1^t}^{t}\sum_{a\in D_t(j)}\sum_{q=\rho\beta t'}^{t'}\sum_{r=\log^2 q}^{q}\mathbb{P}_{i_0}^{\pi_0,P}[\|\hat{p}_j^{t'}(a)-P_{j,a}\|_\infty>\zeta_1 \mid N_{t'}(j,a)=r]$$

$$\leq \sum_{j\in S}\sum_{t'=b_1^t}^{t}\sum_{a\in D_t(j)}\sum_{q=\rho\beta t'}^{t'}\sum_{r=\log^2 q}^{q} C\exp(-r\zeta_1^2)$$

$$\leq C|S|A\sum_{t'=b_1^t}^{t}\sum_{q=\rho\beta t'}^{t'}\sum_{r=\log^2 q}^{q}\exp(-r\zeta_1^2)$$

$$\leq C|S|A\sum_{t'=b_1^t}^{t}\sum_{q=\rho\beta t'}^{t'}\sum_{r=\log^2 q}^{\infty}\exp(-r\zeta_1^2)$$

$$= \frac{C|S|A}{1-\exp(-\zeta_1^2)}\sum_{t'=b_1^t}^{t}\sum_{q=\rho\beta t'}^{t'}\exp(-\log^2 q\cdot\zeta_1^2)$$

$$\leq \frac{C|S|A}{1-\exp(-\zeta_1^2)}\cdot t^2\cdot\exp(-\log^2(\rho\beta b_1^t)\cdot\zeta_1^2)$$

Now $b_1^t \geq \beta t$ and thus the last expression above, ignoring leading constants, is less than

$$t^2\cdot\exp(-(\log\rho\beta^2+\log t)^2\zeta_1^2)$$

$$= \exp(-\zeta_1^2\log^2\rho\beta^2)\cdot t^{2-2\zeta_1^2\log(\rho\beta^2)-\zeta_1^2\log t}$$

which is $o(1/t)$ as $t\to\infty$. □

Chapter 7

Conclusion

The last four chapters have considered various aspects of reinforcement learning in large or unknown MDPs. We close this dissertation by pointing out the major lessons learned and describing some problems for future research.

Chapter 3 considered the sample complexity of policy search using simulators. In that chapter, we showed, using Propositions 2 and 4, that it is not possible to prove convergence bounds by just assuming that the pseudodimension or the fat-shattering dimension of the policy class is finite. Some extra assumption is needed to ensure that the interaction of the policy class with the system transition dynamics does not lead to a blow-up in statistical complexity. Ng and Jordan [23] had earlier used a Lipschitz assumption for this purpose. We essentially replaced their assumption by another one concerning the amount of computation involved in computing policies, the system dynamics and the reward mapping. However, neither of these assumptions implies the other. So, a necessary and sufficient condition for simulation based policy search to work remains to be found.

Chapter 4 considered the approximate linear programming approach for solving large MDPs. Theorem 9, the main result in that chapter, gives a performance bound for the greedy policy derived from a function $V : S \mapsto \mathbb{R}$. However, our goal is to get a performance bound for the greedy policy derived from $H\hat{\mathbf{w}}$, where $\hat{\mathbf{w}}$ is the solution to the approximate linear program with constraint sampling. An important direction for future research is to find conditions on the MDP and the basis functions under which the remainder terms in that bound are small. In that chapter, we also looked at the approach of approximating the dual linear program. Theorem 11 provides a preliminary bound for this approach. While being a first step towards understanding the power of the dual approximation approach, the bound is unsatisfactory for a number of reasons as mentioned towards the end of that chapter. More work is therefore needed to derive better performance bounds.

In Chapter 5, we considered bounded parameter MDPs (BMDPs) that arise when MDP parameters are only known to lie in some intervals. We presented algorithms for computing the optimal value functions for average reward BMDPs. We also related the discounted and average reward criteria and proved the existence of Blackwell optimal policies thus extending two fundamental results from the theory of MDPs to the BMDP setting. A fundamental question left open by our work is that of *perfect duality* in the case of the pessimistic optimality criterion for average reward BMDPs. From Nilim and El Ghaoui's work [24], we know that perfect duality holds for the pessimistic optimality criterion for discounted BMDPs. That is, for any $\gamma \in [0, 1)$, any BMDP $\mathcal{M}$ and any state $i \in S$, we have

$$\max_{\mu} \inf_{M \in \mathcal{M}} V_{\gamma,\mu,M}(i) = \inf_{M \in \mathcal{M}} \max_{\mu} V_{\gamma,\mu,M}(i) \ ,$$

where the max is taken over deterministic stationary policies. We do not know whether a

similar relation holds true for average reward MDPs. That is, we do not know whether

$$\max_{\mu} \inf_{M \in \mathcal{M}} U_{\mu,M}(i) = \inf_{M \in \mathcal{M}} \max_{\mu} U_{\mu,M}(i) \ .$$

In Chapter 6, we gave an algorithm that achieves logarithmic regret in any irreducible MDP. Auer and Ortner's algorithm [5] also achieves logarithmic regret but with a different constant. Neither algorithm requires the knowledge of the supports of the transition functions $P_{i,a}$. The original algorithm of Burnetas and Katehakis [11] required this knowledge. An important advantage of Auer and Ortner's analysis is that their bounds hold uniformly over time. Our bounds, like those of Burnetas and Katehakis, only hold for sufficiently large times and the initial period when the bound does not hold can be quite large, even exponential in the size of the state space. We do not know whether this is a limitation of our analysis or of the algorithm.

It is also interesting to compare our algorithm with Auer and Ortner's. At any time step t, let $\mathcal{M}_t$ denote the BMDP obtained by taking high confidence intervals around the empirical estimates of MDP parameters. Auer and Ortner do not update $\mathcal{M}_t$ at every time step, but when they compute a new policy, it is the optimal policy for $\mathcal{M}_t$ in the optimistic sense. In contrast to this, in our algorithm $\mathcal{M}_t$ gets updated at every time step. But then, we choose our current action by effectively executing just one iteration of the algorithm to compute the optimistically optimal policy for $\mathcal{M}_t$. So, these two algorithms seem to be at two extreme ends of a spectrum of algorithms. Viewing these two algorithms in this way might lead to a better understanding of logarithmic regret algorithms for MDPs.

Bibliography

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.

[5] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems 19*, pages 49–56. MIT Press, 2007.

[6] Riccardo Benedetti and Jean-Jacques Risler. *Real algebraic and semi-algebraic sets*. Actualités Mathématiques, Paris, 1990.

[**7**] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 1995.

[**8**] David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2):719–726, 1962.

[**9**] Lenore Blum, Felipe Cucker, Michael Shub, and Stephen Smale. *Complexity and Real Computation*. Springer-Verlag, 1998.

[**10**] Ronen I. Brafman and Moshe Tennenholtz. R-MAX – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

[**11**] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

[**12**] Daniela P. de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[**13**] Daniela P. de Farias and Benjamin Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.

[**14**] Dmitri A. Dolgov and Edmund H. Durfee. Symmetric approximate linear programming for factored MDPs with application to constrained problems. *Annals of Mathematics and Artificial Intelligence*, 47(3–4):273–293, 2006.

[**15**] Eyal Even-Dar and Yishay Mansour. Convergence of optimistic and incremental

Q-learning. In *Advances in Neural Information Processing Systems 14*, pages 1499–1506. MIT Press, 2001.

[16] Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.

[17] Paul W. Golberg and Mark R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148, 1995.

[18] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. John Hopkins University Press, 1996.

[19] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001.

[20] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[21] Rahul Jain and Pravin P. Varaiya. Simulation-based uniform value function estimates of Markov decision processes. *SIAM Journal on Control and Optimization*, 45(5):1633–1656, 2006.

[22] T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[23] Andrew Y. Ng and Michael I. Jordan. PEGASUS: A policy search method for

MDPs and POMDPs. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, pages 405–415. Morgan Kauffman Publishers, 2000.

[24] Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[25] David Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1990.

[26] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[27] Alexander L. Strehl and Michael Littman. A theoretical analysis of model-based interval estimation. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, pages 857–864. ACM Press, 2005.

[28] Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *Proceedings of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51. IEEE Press, 2007.

[29] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J. Weinberger. Inequalities for the L1 deviation of the empirical distribution. Technical Report HPL-2003-97, Information Theory Research Group, HP Laboratories, Palo Alto, 2003.

Appendix A

Proofs of Auxiliary Lemmas

A.1 Lemma from Chapter 3

Lemma 3. *Fix an interval (a, b) and let $\mathcal{T}$ be the set of all its finite subsets. Let g_n range over functions from $(a, b)^n$ to $\mathcal{T}$. Let D range over probability distributions on (a, b). Then,*

$$\inf_{g_n} \sup_{D} \left(\sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim D} \mathbf{1}\left[X \in T\right] - \mathbb{E}_{(X_1,\ldots,X_n)\sim D^n} \mathbb{E}_{(X \sim D)} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n)\right] \right) \geq 1 \ .$$

Proof. Fix n, g_n. Our proof uses the probabilistic method. We have,

$$\begin{aligned}
&\sup_{D} \left(\sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim D} \mathbf{1}\left[X \in T\right] - \mathbb{E}_{(X_1,\ldots,X_n)\sim D^n} \mathbb{E}_{(X \sim D)} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n)\right] \right) \\
&\geq \sup_{D_m} \left(\sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim D_m} \mathbf{1}\left[X \in T\right] - \mathbb{E}_{(X_1,\ldots,X_n)\sim D_m^n} \mathbb{E}_{(X \sim D_m)} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n)\right] \right) \\
&= 1 - \inf_{D_m} \left(\mathbb{E}_{(X,X_1,\ldots,X_n)\sim D_m^{(n+1)}} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n)\right] \right) \ ,
\end{aligned}$$

where D_m ranges over discrete distribution supported on m points. The last equality holds because for such distributions, the first term inside the sup over D_m is 1. We now use the probabilistic method to upper bound the infimum over D_m by an expectation over D_m

when D_m is a random probability distribution supported on $Y_1, \ldots, Y_m$ where

$$Y_1, \ldots, Y_m \sim \text{i.i.d. Uniform}(a, b) \ .$$

Further, let $I_0, I_1, \ldots, I_n$ be independent of the Y's with

$$I_0, I_1, \ldots, I_n \sim \text{i.i.d. Discrete Uniform}\{1, \ldots, m\} \ .$$

Finally, set $X = Y_{I_0}$ and $X_j = Y_{I_j}, 1 \leq j \leq n$. Thus, we have

$$\begin{aligned}
&\inf_{D_m} \left(\mathbb{E}_{(X, X_1, \ldots, X_n) \sim D_m^{(n+1)}} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n) \right] \right) \\
&\leq \mathbb{P}(X \in g_n(X_1, \ldots, X_n)) \\
&\leq \mathbb{P}(X \in g_n(X_1, \ldots, X_n) | \neg E) + \mathbb{P}(E) \ ,
\end{aligned}$$

where E is the event "there exist duplicates amongst the I_j's". Conditioned on $\neg E$, the X_j's are i.i.d. Uniform(a, b) and hence the first term above is 0. To bound the second term note that

$$\begin{aligned}
\mathbb{P}(E) &= 1 - \frac{m(m-1)\ldots(m-n+1)}{m^n} \\
&\leq 1 - \frac{(m-n)^n}{m^n} \\
&\leq 1 - \left(1 - \frac{n}{m}\right)^n
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\sup_D \left(\sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim D} \mathbf{1}\left[X \in T \right] - \mathbb{E}_{(X_1, \ldots, X_n) \sim D^n} \mathbb{E}_{(X \sim D)} \mathbf{1}\left[X \in g_n(X_1, \ldots, X_n) \right] \right) \\
&\geq \left(1 - \frac{n}{m}\right)^n
\end{aligned}$$

Since this is true for all $m > n$ and the right hand side tends to 1 as $m \to \infty$ (with n fixed), the lemma is proved. $\square$

A.2 Lemmas from Chapter 6

Lemma 30. *For all $\delta, \epsilon > 0$, $\exists \zeta_1, \zeta_2, \zeta_3 > 0, t_0$ such that the following hold true. For all $\zeta < \zeta_1$, $B_t(\zeta) \subseteq$*

$$\{\forall t' \geq b_1^t, \forall \mu \in \Pi(\mathcal{D}_{t'}),\ |\lambda_\mu(\hat{P}^{t'}) - \lambda_\mu(P)| \leq \epsilon,\ \|h_\mu(\hat{P}^{t'}) - h_\mu(P)\|_\infty \leq \epsilon\}$$

For all $\zeta < \zeta_2$, $B_t(\zeta) \subseteq$

$$\{\forall t' \geq b_1^t, \mathcal{O}(\hat{P}^{t'}, \mathcal{D}_{t'}) \subseteq \mathcal{O}(P, \mathcal{D}_{t'}) \ \wedge\ \|h(\hat{P}^{t'}, \mathcal{D}_{t'}) - h(P, \mathcal{D}_{t'})\|_\infty \leq \epsilon\}$$

For all $\zeta < \zeta_3, t > t_0$, $B_t(\zeta) \subseteq$

$$\{\forall t' \geq b_1^t, j \in S, a \in D_{t'}(j),\ U_{t'}(j,a) \leq \mathcal{L}(j, a; P_{j,a}, h(P, \mathcal{D}_{t'})) + \delta\}$$

Proof. (First part) Fix $t' \geq b_1^t$. Given a policy $\mu \in \Pi(\mathcal{D}_{t'})$, let

$$H_\mu(P) = (\lambda_\mu(P), h_\mu(2; P), \ldots, h_\mu(|S|; P))^\top\ .$$

Since, $h_\mu(P)$ is unique only up to an additive constant, we assume that $h_\mu(1; P) = 0$. With this convention, the average reward equations become

$$A_\mu(P) H_\mu(P) = R_\mu$$

where $A_\mu(P)$ is an invertible matrix. For an $n \times n$ matrix A, let $\|A\|_\infty$ be the row sum norm,

$$\|A\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}|\ .$$

Theorem 2.7.2 in [18] says that if

$$Ax = b \qquad\qquad (A + \Delta)y = b$$

with $\|\Delta\|_\infty \cdot \|A^{-1}\|_\infty \leq 1$, then

$$\|y - x\|_\infty \leq \frac{2\|\Delta\|_\infty \cdot \|A^{-1}\|_\infty}{1 - \|\Delta\|_\infty \cdot \|A^{-1}\|_\infty} \|x\|_\infty \ . \tag{A.1}$$

Let $\hat{P} = \hat{P}^{t'}$ and

$$H_\mu(\hat{P}) = (\lambda_\mu(\hat{P}), h_\mu(2; \hat{P}), \ldots, h_\mu(|S|; \hat{P}))^\top \ .$$

The average reward equations for the problem $(\hat{P}, \mathcal{D}_{t'})$ can be written as

$$A_\mu(\hat{P}) H_\mu(\hat{P}) = R_\mu$$

where $A_\mu(\hat{P})$ is an invertible matrix with the estimated probabilities in its entries. Under the event $B_t(\zeta)$, we have

$$\|A_\mu(\hat{P}) - A_\mu(P)\|_\infty \leq |S|\zeta \ .$$

Now using (A.1) with $y = H_\mu(\hat{P})$, $x = H_\mu(P)$, $A = A_\mu(P)$ and $\Delta = A_\mu(\hat{P}) - A_\mu(P)$, we get

$$\|H_\mu(\hat{P}) - H_\mu(P)\|_\infty \leq \epsilon$$

provided that

$$\zeta \leq \zeta_1(\epsilon; \mu) := \frac{\epsilon}{|S| \cdot \|A_\mu(P)^{-1}\|_\infty \cdot (2\|H_\mu(P)\|_\infty + \epsilon)} \ .$$

We finish the proof of the first part by setting $\zeta_1 = \min_{\mu \in \Pi_D} \zeta_1(\epsilon; \mu)$.

(Second part) Define the constant,

$$\delta_{\text{diff}} := \min\{|\lambda_\mu(P) - \lambda_{\mu'}(P)| \ : \ \mu, \mu' \in \Pi_D, \mu \neq \mu'\} \ .$$

Assume that $\delta_{\text{diff}} > 0$, otherwise the result is trivial. Given $\epsilon > 0$, choose ζ_2 small enough such that the previous part of the lemma holds both for ϵ and for $\delta_{\text{diff}}/2$. Then, for all $\zeta < \zeta_2$, under $B_t(\zeta)$, any μ_0 that is optimal for $(\hat{P}^{t'}, \mathcal{D}_t)$ is $\delta_{\text{diff}}/2$-optimal for $(P, \mathcal{D}_{t'})$.

By the definition of δ_{diff}, any $\delta_{\text{diff}}/2$-optimal policy is actually optimal. Therefore, under $B_t(\zeta)$, $\mathcal{O}(\hat{P}^{t'}, \mathcal{D}_{t'}) \subseteq \mathcal{O}(P, \mathcal{D}_{t'})$. Moreover, for such a μ_0 we have $h(\hat{P}^{t'}, \mathcal{D}_{t'}) = h_{\mu_0}(\hat{P}^{t'})$ and $h(P, \mathcal{D}_{t'}) = h_{\mu_0}(P)$ and so we also have $\|h(\hat{P}^{t'}, \mathcal{D}_{t'}) - h(P, \mathcal{D}_{t'})\|_\infty \leq \epsilon$ by our choice of ζ_2.

(THIRD PART) Fix $t' \geq b_1^t, j \in S, a \in D_{t'}(j)$. It suffices to prove that, under $B_t(\zeta)$, for any $q \in \Delta^+$, such that $\|\hat{p}_j^{t'}(a) - q\| \leq \sqrt{2 \log t' / N_{t'}(j,a)}$ we have

$$\mathcal{L}(j, a, q, h(\hat{P}^{t'}, \mathcal{D}_{t'})) \leq \mathcal{L}(j, a, P_{j,a}, h(P, \mathcal{D}_{t'})) + \delta \ . \tag{A.2}$$

Fix such a q. Using the previous part, if ζ is sufficiently small then, under $B_t(\zeta)$, we have $\|h(\hat{P}^{t'}, \mathcal{D}_{t'}) - h(P, \mathcal{D}_{t'})\| \leq \delta/3$ and hence,

$$\mathcal{L}(j, a, q, h(\hat{P}^{t'}, \mathcal{D}_{t'})) \leq \mathcal{L}(j, a, q, h(P, \mathcal{D}_{t'})) + \delta/3 \ . \tag{A.3}$$

Under $B_t(\zeta)$, $N_{t'}(j) \geq \rho\beta t$ and since $a \in D_{t'}(j)$, $N_{t'}(j,a) \geq \log^2(\rho\beta t)$. As $t' \leq t$, we have $\sqrt{2 \log t' / N_{t'}(j,a)} \leq \sqrt{2 \log t / \log^2(\rho\beta t)}$ which goes to 0 as $t \to \infty$. As $\mathcal{L}(j, a, p, h)$ is continuous in p, there exists t_0 such that for all $t > t_0$, we have

$$\mathcal{L}(j, a, q, h(P, \mathcal{D}_{t'})) \leq \mathcal{L}(j, a, \hat{p}_j^{t'}(a), h(P, \mathcal{D}_{t'})) + \delta/3 \ . \tag{A.4}$$

Under $B_t(\zeta)$, $\|\hat{p}_j^{t'}(a) - P_{j,a}\|_\infty \leq \zeta$ and thus for small enough ζ we have,

$$\mathcal{L}(j, a, \hat{p}_j^{t'}(a), h(P, \mathcal{D}_{t'})) \leq \mathcal{L}(j, a, P_{j,a}, h(P, \mathcal{D}_{t'})) + \delta/3 \ . \tag{A.5}$$

Here, we again used the continuity of $L(j, a, p, h)$ in p. Combining (A.3),(A.4) and (A.5) gives (A.2). □

Lemma 31. *On the event $B_t(\zeta)$, the following hold true. Let*

$$\tau = \rho\beta t / (2 \log(\rho\beta t)) - 1 \ .$$

For $t' > b_1^t, j \in S, a \in A(j)$, if $(s_{t'}, a_{t'}) = (j, a)$ with $a \in \Gamma_{t'}^1$, then for all ζ,

$$\sum_{l=0}^{\tau} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ a \in \Gamma_{t'+l}^1\right] \leq 1$$

For all $t' \geq b_1^t$ and sufficiently small ζ,

$$\lambda(P, \mathcal{D}_{t'+1}) \geq \lambda(P, \mathcal{D}_{t'}) \ .$$

Proof. (First part) Fix $t' > b_1^t$ such that $(s_{t'}, a_{t'}) = (j, a)$ with $a \in \Gamma_{t'}^1$. Let $n = N_{t'}(j)$ and $\sigma = n/(2 \log n) - 1$. With this choice of σ, we have

$$\log^2(n + \sigma + 1) < \log^2 n + 1 \tag{A.6}$$

for sufficiently large n. Note that, for any l, if $(s_{t'+l}, a_{t'+l}) = (j, a)$ then

$$\Gamma_{t'+l}^1 \subseteq O(j; \hat{P}^{t'+l}, \mathcal{D}_{t'+l}) \subseteq D_{t'+l}(j)$$

and therefore $N_{t'+l}(j, a) \geq \log^2 N_{t'+l}(j)$. Moreover, by the definition of $\Gamma_{t'+l}^1$, $N_{t'+l}(j, a) < \log^2(N_{t'+l}(j) + 1)$. Hence,

$$\begin{aligned}
&\sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ a \in \Gamma_{t'+l}^1\right] \\
&\leq \sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ \log^2 N_{t'+l}(j) \leq N_{t'+l}(j, a) < \log^2(N_{t'+l}(j) + 1)\right] \\
&\leq \sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ \log^2 n \leq N_{t'+l}(j, a) < \log^2(N_{t'+l}(j) + 1)\right] \\
&\leq \sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ \log^2 n \leq N_{t'+l}(j, a) < \log^2(n + \sigma + 1)\right] \\
&\leq \sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ \log^2 n \leq N_{t'+l}(j, a) < \log^2 n + 1\right] \\
&\leq \sum_{l=0}^{\sigma} \mathbf{1}\left[(s_{t'+l}, a_{t'+l}) = (j, a) \ \wedge \ N_{t'+l}(j, a) = \lfloor \log^2 n \rfloor\right] \\
&\leq 1
\end{aligned}$$

The second inequality follows because $N_{t'+l}(j) \geq N_{t'}(j) = n$. The third inequality follows because, for $l \leq \sigma$, $N_{t'+l}(j) \leq N_{t'+\sigma}(j) \leq N_{t'}(j) + \sigma = n + \sigma$. The fourth inequality follows from (A.6). Now, for all ζ, on the event $B_t(\zeta)$ we have $n = N_{t'}(j) \geq N_{b_1^t}(j) \geq \rho\beta t$. Thus, $\sigma \geq \tau$.

(Second part) It suffices to show that for all $t' \geq b_1^t$,

$$\mathcal{D}_{t'+1} \cap \mathcal{O}(\hat{P}^{t'}, \mathcal{D}_{t'}) \neq \emptyset \ , \tag{A.7}$$

because, by the previous part, for sufficiently small ζ, on the event $B_t(\zeta)$ we have

$$\mathcal{O}(\hat{P}^{t'}, \mathcal{D}_{t'}) \subseteq \mathcal{O}(P, \mathcal{D}_{t'}) \ .$$

Hence, (A.7) implies $\mathcal{D}_{t'+1} \cap \mathcal{O}(P, \mathcal{D}_{t'}) \neq \emptyset$. This, in turn, implies that

$$\lambda(P, \mathcal{D}_{t'+1}) \geq \lambda(P, \mathcal{D}_{t'}) \ .$$

Suppose $s_{t'} = j$. Note that, for all $j' \neq j$, $D_{t'+1}(j') = D_{t'}(j')$ and therefore $D_{t'+1}(j') \cap \mathcal{O}(j'; \hat{P}^{t'}, \mathcal{D}_{t'}) \neq \emptyset$ holds trivially. To finish the proof we thus need to prove $D_{t'+1}(j) \cap \mathcal{O}(j; \hat{P}^{t'}, \mathcal{D}_{t'}) \neq \emptyset$. For that, consider two cases. First, if $\Gamma_t^1 = \mathcal{O}(j; \hat{P}^t, \mathcal{D}_t)$ then consider the action $a = a_t \in \Gamma_t^1$ that π_0 takes at time t. For this action,

$$\begin{aligned}
N_{t'+1}(j,a) &= N_{t'}(j,a) + 1 && [\because a_{t'} = a] \\
&\geq \log^2(N_{t'}(j)) + 1 && [\because a \in \mathcal{D}_{t'}] \\
&> \log^2(N_{t'}(j) + 1) \ . && [\because \log^2 n + 1 > \log^2(n+1)] \\
&= \log^2(N_{t'+1}(j)) && [\because s_{t'} = j] \ .
\end{aligned}$$

This means that $a \in D_{t'+1}(j)$.

The second case is when $\Gamma_t^1 \subset \mathcal{O}(j; \hat{P}^t, \mathcal{D}_t)$. Consider any $a \in \mathcal{O}(j; \hat{P}^t, \mathcal{D}_t) - \Gamma_t^1$. For this action,

$$\begin{aligned} N_{t'+1}(j,a) &\geq N_{t'}(j,a) \\ &\geq \log^2(N_{t'}(j)+1) && [\because a \notin \Gamma_t^1] \\ &= \log^2(N_{t'+1}(j)) && [\because s_{t'} = j] \ . \end{aligned}$$

Again, this means that $a \in D_{t'+1}(j)$. □

Appendix B

Some Results for BMDPs

In this appendix, we provide self-contained proofs of the three theorems mentioned without proof in Section 5.3 of Chapter 5. Throughout this appendix, vector inequalities of the form $V_1 \leq V_2$ are to be interpreted to mean $V_1(i) \leq V_2(i)$ for all i.

Proofs of Theorems 12 and 13

Lemma 34. *If $V_1 \leq V_2$ then, for all $M \in \mathcal{M}$,*

$$T_{\gamma,\mu,M} V_1 \leq T^{\text{opt}}_{\gamma,\mu} V_2 \ ,$$

$$T^{\text{pes}}_{\gamma,\mu} V_1 \leq T_{\gamma,\mu,M} V_2 \ .$$

Proof. We prove the first inequality. Fix an MDP $M \in \mathcal{M}$. Let $P_{i,a}(j)$ denote transition

probabilities of M. We then have,

$$\begin{aligned}
\left(T_{\gamma,\mu,M}V_1\right)(i) &= (1-\gamma)R(i,\mu(i)) + \gamma\sum_j P_{i,\mu(i)}(j)V_1(j) \\
&\leq (1-\gamma)R(i) + \gamma\sum_j P_{i,\mu(i)}(j)V_2(j) && [\ \because V_1 \leq V_2\ ] \\
&\leq (1-\gamma)R(i,\mu(i)) + \gamma \max_{q\in\mathcal{C}_{i,\mu(i)}} q^\top V_2 && [\ \because M \in \mathcal{M}\ ] \\
&= \left(T^{\text{opt}}_{\gamma,\mu}V_2\right)(i)\ .
\end{aligned}$$

The proof of the second inequality is similar. □

Lemma 35. *If $V_1 \leq V_2$ then, for any policy μ,*

$$T^{\text{opt}}_{\gamma,\mu}V_1 \leq T^{\text{opt}}_{\gamma}V_2\ ,$$

$$T^{\text{pes}}_{\gamma,\mu}V_1 \leq T^{\text{pes}}_{\gamma}V_2\ .$$

Proof. Again, we prove only the first inequality. Fix a policy μ. We then have,

$$\begin{aligned}
\left(T^{\text{opt}}_{\gamma,\mu}V_1\right)(i) &= (1-\gamma)R(i,\mu(i)) + \gamma \max_{q\in\mathcal{C}_{i,\mu(i)}} q^\top V_1 \\
&\leq (1-\gamma)R(i) + \gamma \max_{q\in\mathcal{C}_{i,\mu(i)}} q^\top V_2 \\
&\leq \max_{a\in A}\left[(1-\gamma)R(i,a)) + \gamma \max_{q\in\mathcal{C}_{i,a}} q^\top V_2\right] \\
&= \left(T^{\text{opt}}_{\gamma}V_2\right)(i)
\end{aligned}$$

□

Proof. (*of Theorems 12 and 13*) Let $\tilde{V}$ be the fixed point of $T^{\text{opt}}_{\gamma,\mu}$. This means that for all $i \in S$,

$$\tilde{V}(i) = (1-\gamma)R(i,\mu(i)) + \gamma \max_{q\in\mathcal{C}_{i,\mu(i)}} q^\top \tilde{V}\ .$$

We wish to show that $\tilde{V} = V^{\text{opt}}_{\gamma,\mu}$. Let q_i be the probability vector that achieves the maximum above. Construct an MDP $M_1 \in \mathcal{M}$ as follows. Set the transition probability vector $P_{i,\mu(i)}$ to be q_i. For $a \neq \mu(i)$, choose $P_{i,a}$ to be any element of $\mathcal{C}_{i,a}$. It is clear that $\tilde{V}$ satisfies, for all $i \in S$,

$$\tilde{V}(i) = (1-\gamma)R(i,\mu(i)) + \gamma \sum_j P_{i,\mu(i)}(j)\tilde{V}(j) \ ,$$

and therefore $\tilde{V} = V_{\gamma,\mu,M_1} \leq V^{\text{opt}}_{\gamma,\mu}$. It remains to show that $\tilde{V} \geq V^{\text{opt}}_{\gamma,\mu}$. For that, fix an arbitrary MDP $M \in \mathcal{M}$. Let V_0 be any initial vector. Using Lemma 34 and straightforward induction, we get

$$\forall k \geq 0, \ (T_{\gamma,\mu,M})^k V_0 \leq (T^{\text{opt}}_{\gamma,\mu})^k V_0 \ .$$

Taking limits as $k \to \infty$, we get $V_{\gamma,\mu,M} \leq \tilde{V}$. Since $M \in \mathcal{M}$ was arbitrary, for any $i \in S$,

$$V^{\text{opt}}_{\gamma,\mu}(i) = \sup_{M \in \mathcal{M}} V_{\gamma,\mu,M}(i) \leq \tilde{V}(i) \ .$$

Therefore, $\tilde{V} = V^{\text{opt}}_{\gamma,\mu}$.

Now let $\tilde{V}$ be the fixed point of T^{opt}_{γ}. This means that for all $i \in S$,

$$\tilde{V}(i) = \max_{a \in A} \left[(1-\gamma)R(i,a) + \gamma \max_{q \in \mathcal{C}_{i,a}} q^\top \tilde{V} \right] \ .$$

We wish to show that $\tilde{V} = \mathbf{V}^{\text{opt}}_{\gamma}$. Let $\mu_1(i)$ be any action that achieves the maximum above. Since $\tilde{V}$ satisfies, for all $i \in S$,

$$\tilde{V}(i) = (1-\gamma)R(i,\mu_1(i)) + \gamma \max_{q \in \mathcal{C}_{i,\mu_1(i)}} q^\top \tilde{V} \ ,$$

we have $\tilde{V} = V^{\text{opt}}_{\gamma,\mu_1} \leq \mathbf{V}^{\text{opt}}_{\gamma}$. It remains to show that $\tilde{V} \geq \mathbf{V}^{\text{opt}}_{\gamma}$. For that, fix an arbitrary policy μ. Let V_0 be any initial vector. Using Lemma 35 and straightforward induction, we get

$$\forall k \geq 0, \ (T^{\text{opt}}_{\gamma,\mu})^k V_0 \leq (T^{\text{opt}}_{\gamma})^k V_0 \ .$$

Taking limits as $k \to \infty$, we get $V^{\text{opt}}_{\gamma,\mu} \leq \tilde{V}$. Since μ was arbitrary, for any $i \in S$,

$$\mathbf{V}^{\text{opt}}_{\gamma}(i) = \max_{\mu} V^{\text{opt}}_{\gamma,\mu}(i) \leq \tilde{V}(i) \ .$$

Therefore, $\tilde{V} = \mathbf{V}^{\text{opt}}_{\gamma}$. Moreover, this also proves the first part of Theorem 13 since

$$V^{\text{opt}}_{\gamma,\mu_1} = \tilde{V} = \mathbf{V}^{\text{opt}}_{\gamma} \ .$$

The claim that the fixed points of $T^{\text{pes}}_{\gamma,\mu}$ and T^{pes}_{γ} are $V^{\text{pes}}_{\gamma,\mu}$ and $\mathbf{V}^{\text{pes}}_{\gamma}$ respectively, is proved by making a few obvious changes to the argument above. Further, as it turned out above, the argument additionally yields the proof of the second part of Theorem 13.

□

Proof of Theorem 14

We prove the existence of $\mathcal{M}_{\text{opt}}$ only. The existence of $\mathcal{M}_{\text{pes}}$ is proved in the same way. Note that in the proof presented in the previous subsection, given a policy μ, we explicitly constructed an MDP M_1 such that $V^{\text{opt}}_{\gamma,\mu} = V_{\gamma,\mu,M_1}$. Further, the transition probability vector $P_{i,\mu(i)}$ of M_1 was a vector that achieved the maximum in

$$\max_{\mathcal{C}_{i,\mu(i)}} q^{\top} V^{\text{opt}}_{\gamma,\mu} \ .$$

Recall that the set $\mathcal{C}_{i,\mu(i)}$ has the form

$$\{q \ : \ q^{\top}\mathbf{1} = 1, \ \forall j \in S, \ l_j \leq q_j \leq u_j\} \ , \tag{B.1}$$

where $l_j = l(i,j,\mu(i)), \ u_j = u(i,j,\mu(i))$. Therefore, all that we require is the following lemma.

Lemma 36. *Given a set $\mathcal{C}$ of the form* (B.1)*, there exists a finite set $Q = Q(\mathcal{C})$ of cardinality no more than $|S|!$ with the following property. For any vector V, there exists $\tilde{q} \in Q$ such that*

$$\tilde{q}^\top V = \max_{q \in \mathcal{C}} q^\top V \ .$$

We can then set

$$\mathcal{M}_{\text{opt}} = \{ \ \langle S, A, P, R \rangle \ : \ \forall i, a, \ P_{i,a} \in Q(\mathcal{C}_{i,a}) \ \} \ .$$

The cardinality of $\mathcal{M}_{\text{opt}}$ is at most $(|S||A|)|S|!$

Proof. (*of Lemma 36*) A simple greedy algorithm (Algorithm 4) can be used to find a maximizing $\tilde{q}$. The set $\mathcal{C}$ is specified using upper and lower bounds, denoted by u_i and l_i respectively. The algorithm uses the following idea recursively. Suppose i^* is the index of a largest component of V. It is clear that we should set $\tilde{q}(i^*)$ as large as possible. The value of $\tilde{q}(i^*)$ has to be less than u_i. Moreover, it has to be less than $1 - \sum_{i \neq i^*} l_i$. Otherwise, the remaining lower bound constraints cannot be met. So, we set $\tilde{q}(i^*)$ to be the minimum of these two quantities.

Note that the output depends only on the sorted order of the components of V. Hence, there are only $|S|!$ choices for $\tilde{q}$. □

Algorithm 4 A greedy algorithm to maximize $q^\top V$ over $\mathcal{C}$.

INPUTS The vector V and the set $\mathcal{C}$. The latter is specified by bounds $\{l_i\}_{i \in S}$ and $\{u_i\}_{i \in S}$ that satisfy $\forall i, 0 \leq l_i \leq u_i$ and $\sum_i l_i \leq 1 \leq \sum_i u_i$.

OUTPUT A maximizing vector $\tilde{q} \in \mathcal{C}$.

▷ **order**(V) gives the indices of the largest to smallest elements of V

indices $\leftarrow$ **order**(V)

massLeft $\leftarrow 1$

indicesLeft $\leftarrow S$

for all $i \in$ indices **do**

  elem $\leftarrow V(i)$

  lowerBoundSum $\leftarrow \sum_{j \in \text{indicesLeft}, j \neq i} l_j$

  $\tilde{q}(i) \leftarrow \min(u_i, \text{massLeft} - \text{lowerBoundSum})$

  massLeft $\leftarrow$ massLeft $- \tilde{q}(i)$

  indicesLeft $\leftarrow$ indicesLeft $- \{i\}$

end for

return $\tilde{q}$