Analysis and Simplex-type Algorithms for Countably Infinite Linear Programming Models of Markov Decision Processes

by

Ilbin Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2015

Doctoral Committee:

Associate Professor Marina A. Epelman, Co-chair
Professor Edwin Romeijn, Co-chair
Assistant Professor Laura K. Balzano
Professor Emeritus Robert L. Smith

For Junghee

ACKNOWLEDGEMENTS

At the end of May in 2010, I got a PhD admission from IOE. I was the last PhD student admitted to IOE in that year while the other students got their admissions in March. Moreover, it was the only admission I received. I have been heartily grateful that I was given this wonderful opportunity of learning and training. Now while finishing up my PhD study, I would like to express my gratitudes to everyone who has helped and supported me.

This dissertation would have been impossible without ceaseless intellectual supports from my advisors, Prof. Marina Epelman, Prof. Edwin Romeijn, and Prof. Robert Smith. Their constructive criticism has prepared me to be an independent researcher. They have been always patient in teaching me how to conduct research and present research findings. I sincerely thank them for providing the GSRA support throughout my PhD years and supporting my travels to conferences. Also, I cannot thank enough for their effort and time to write recommendation letters.

I am grateful to Prof. Laura Balzano for serving as a dissertation committee member. She also gave me opportunities to do make-up lectures of IOE 600, through which I found myself enjoying teaching. I thank Prof. Mark Daskin for allowing me to teach IOE 202 as a primary instructor and Prof. Romesh Saigal for a fruitful discussion about norms of matrices. I would also like to thank Prof. Eunshin Byon, Prof. Jon Lee, and Prof. Viswanath Nagarajan for giving me valuable career advices.

My graduate school experience would have been less enjoyable without numerous helps and warm regards from IOE staffs. I thank Tina for helping me with administrative processes and office supplies. I am grateful to Wanda and Chris for helping me when I taught IOE

202. I also thank Candy and Matt for processing my health insurance while I was away.

My friends in IOE rescued me from solitary dissertation research time to time. Troy taught me a lot about English, American culture, and useful softwares. I really enjoyed talking about NBA with Victor. Zohar told me many helpful tips about research and career. I especially thank all my friends for delivering foods when Grace and Rachel were born.

I thank my family members and relatives for supporting my decision to pursue a PhD. I am especially grateful to my grandmother for taking care of my mother while I am away. I am greatly indebted to my wife, Junghee Hong, to whom I am dedicating this thesis. She is always supportive with my career decisions and works so hard to raise our daughters. Lastly, I thank God for giving me this great opportunity and guiding my career.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Analysis and Simplex-type Algorithms for Countably Infinite Linear Programming Models of Markov Decision Processes

by

Ilbin Lee

Chair: Marina A. Epelman and H. Edwin Romeijn

The class of Markov decision processes (MDPs) provides a popular framework which covers a wide variety of sequential decision-making problems. We consider infinite-horizon discounted MDPs with countably infinite state space and finite action space. Our goal is to establish theoretical properties and develop new solution methods for such MDPs by studying their linear programming (LP) formulations. The LP formulations have countably infinite numbers of variables and constraints and therefore are called countably infinite linear programs (CILPs). General CILPs are challenging to analyze or solve, mainly because useful theoretical properties and techniques of finite LPs fail to extend to general CILPs. Another goal of this thesis is to deepen the limited current understanding of CILPs, resulting in new algorithmic approaches to find their solutions.

Recently, Ghate and Smith (2013) developed an implementable simplex-type algorithm for solving a CILP formulation of a non-stationary MDP with finite state space. We establish rate of convergence results for their simplex algorithm with a particular pivoting rule and another existing solution method for such MDPs, and compare empirical performance of the algorithms. We also present ways to accelerate their simplex algorithm. The class

of non-stationary MDPs with finite state space can be considered to be a subclass of stationary MDPs with countably infinite state space. We present a simplex-type algorithm for solving a CILP formulation of a stationary MDP with countably infinite state space that is implementable (using only finite data and computation in each iteration). We show that the algorithm finds a sequence of policies that improves monotonically and converges to optimality in value, and present a numerical illustration. An important extension of MDPs considered so far are constrained MDPs, which optimize an objective function while satisfying constraints, typically on budget, quality, and so on. For constrained non-stationary MDPs with finite state space, we provide a necessary and sufficient condition for a feasible solution of its CILP formulation to be an extreme point. Since simplex-type algorithms are expected to navigate between extreme points, this result sets a foundation for developing a simplex-type algorithm for constrained non-stationary MDPs.

CHAPTER I

Introduction

Sequential decision-making problems arise in various application areas that require long-term planning, such as production planning [7, 54, 52], queueing control problems [5, 30, 34, 44], facility maintenance [12, 25, 61], medical treatment planning [1, 48], and stochastic unit commitment [33]. In such problems, there are multiple stages of decisions to be made on a dynamic system. When a decision maker makes a decision, she should take into account not only the decision's immediate cost (or reward), but also how the decision will affect the system and her future decision-making. For example, a decision that incurs the minimum immediate cost may put the system in a position that leads to high costs in future. The class of Markov decision processes (MDPs) provides a popular framework that covers a wide variety of sequential decision-making problems (see [10, 18, 39, 60] and the references cited above).

Consider a dynamic system that evolves over discrete time periods. In period $n \in \mathbb{N} = \{1, 2, \ldots\}$, the system is observed in a state $s \in \mathcal{S}$ and a decision maker chooses an action $a \in \mathcal{A}$. Given that action a is taken in state s, the system makes a transition to a next state $t \in \mathcal{S}$ with probability $p(t|s, a)$ incurring a nonnegative cost $c(s, a; t)$. This procedure continues indefinitely, and future costs are discounted by a fixed discount factor $\alpha \in (0, 1)$, thus, we consider infinite-horizon discounted problems throughout this thesis. Let $c(s, a)$ denote the expected cost incurred by taking action a at state s, i.e., $c(s, a) = \sum_{t \in \mathcal{S}} p(t|s, a)c(s, a; t)$.

The goal of the decision maker is to minimize expected total discounted cost over infinite horizon.

There is a vast literature about solving the case where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$ are both finite (e.g., [10, 39]). We call such MDPs the base case. In this thesis, we will consider three extensions of the base case in each of Chapters II, III, and IV, that arise in different applications. The first extension we consider in Chapter II is obtained from the base case by relaxing the assumption that transition probabilities and cost function are stationary, which is often violated in practice. We call this class of problems *non-stationary MDPs*. That is, non-stationary MDPs have finite state space, finite action space, and non-stationary problem data. To note the dependence on period index, transition probabilities $p_n(t|s,a)$ and cost function $c_n(s,a)$ will have a subscript n. The class of non-stationary MDPs covers a variety of applications, such as production planning under non-stationary cost and demand data [22, 52], capacity expansion under nonlinear demand [9], and equipment replacement under technological change [8]. The second extension considered in Chapter III is obtained from the base case by allowing the state set $\mathcal{S}$ to be countably-infinite. We call this class *countable-state MDPs*. Thus, countable-state MDPs have countably-infinite state space, finite action space, and stationary problem data. Applications that require a countably-infinite state space include inventory management and queueing control where there is no specific limit on the size of inventory or queue [4, 39]. The third class of MDPs considered in Chapter IV is obtained from a non-stationary MDP by adding side constraints, typically on budget or quality. We call this *constrained non-stationary MDPs*. That is, constrained non-stationary MDPs have finite state space, finite action space, non-stationary problem data, and side constraints. MDPs with side constraints often arise in data communications and facility maintenance (e.g., [5, 61]).

In this dissertation, we analyze the above three classes of MDPs by studying their linear programming (LP) formulations. We will introduce new solution algorithms for the MDPs by providing algorithms for solving the LP formulations. The main goal of this thesis is

developing simplex-type algorithms for solving the LP formulations (a simplex-type algorithm navigates through adjacent extreme points of a feasible region, improves an objective function in each move, and converges to optimality) and analyzing the algorithms. Whenever necessary, we will also establish theoretical properties that are required for developing simplex-type algorithms.

Solving an equivalent LP formulation is a popular solution method for MDPs. It is well known that policy iteration, one of the popular solution methods for base case MDPs, can be viewed as the simplex method applied to an equivalent LP formulation of the MDP. A recent result in [62] showed that for base case MDPs, simplex method with Dantzig's pivoting rule (for minimization, choosing a non-basic variable with the most negative reduced cost) is strongly polynomial for a fixed discount factor, and the complexity bound is better than that of the other solution methods. Furthermore, LP models of MDPs can easily incorporate additional constraints while other popular solution methods for MDPs such as value iteration and policy iteration do not naturally extend to constrained MDPs.

For the three classes of MDPs considered in this thesis, equivalent LP formulations have a countably-infinite number of variables and a countably-infinite number of constraints. Such LPs are called *countably-infinite linear programs (CILPs)*. In general, CILPs arise in various additional applications, such as multistage stochastic programs [51], infinite network flow problems [50], and games with partial informations [14, 20]. However, general CILPs are challenging to analyze or solve mainly because one encounters many obstacles in trying to extend useful theoretical properties and techniques of finite LPs to general CILPs. First, it is possible that a CILP and its dual have a duality gap (see [42] for an example). Also, for finite LPs, a feasible solution is an extreme point if and only if it is a basic solution, but such an algebraic characterization of extreme points does not extend to CILPs in general, which is one of the difficulties in devising a simplex-type algorithm [22]. In addition, it is possible for a CILP to have an optimal solution but not an extreme point optimal solution [6]. Even for CILPs that have extreme point optimal solutions, extending the simplex

Classes of MDPs	Characterization of extreme points	Simplex-type algorithm	Convergence rate of simplex algorithm
Stationary, finite-state (base case)	[39]	[15, 31]	[62]
Non-stationary, finite-state	[4, 18]	[24]	Chapter 2
Stationary, countable-state	[4, 18]	Chapter 3	
Constrained, non-stationary, finite-state	Chapter 4		

Table 1.1: An organization of the contributions of this thesis and representative related work

method is challenging. A pivot operation may require infinite computation, and hence not be implementable [6, 50]. Moreover, [21] provided an example of a CILP in which a strictly improving sequence of extreme points may not converge in value to optimality, which indicates that proving convergence to optimality requires careful considerations (for more details, see [24, 21]). However, we note that some of LP results can be extended by considering more structured CILPs [24, 41, 42] or finding appropriate sequence spaces [20].

Due to these hurdles, there are only a few algorithmic approaches to CILPs in the literature [24, 21, 50] (all of these are simplex-type algorithms). In particular, for a CILP formulation of a non-stationary MDP, [24] recently provided duality results, characterization of extreme points, and a simplex-type algorithm that improves in every iteration and converges to optimality. Their algorithm is implementable in the sense that each of its iterations requires only a finite amount of data and computation. In Chapter II, we establish rate of convergence results for their simplex algorithm with a particular pivoting rule and an existing solution method for non-stationary MDPs, called receding horizon approach, and compare the two results. Also, we introduce a technique that accelerates the simplex algorithm and present experimental results for the two algorithms.

The class of non-stationary MDPs can be considered to be a subclass of countable-state

MDPs. In Chapter III, we introduce an implementable simplex algorithm for solving a CILP formulation of a countable-state MDP. The contributions of this development is twofolds.

First, this is the first algorithm for countable-state MDPs that computes a sequence of policies that not only converges to optimality but also improves monotonically. The existing solution methods of countable-state MDPs obtain a sequence of policies that converges to optimality in value (to be more precise, value functions of the policies converge to the optimal value function, see Section 1.1 for precise definitions). However, the policies obtained by those methods may not improve in every iteration. In other words, a policy obtained in a later iteration may be worse than a previously obtained policy. In practice, one can run those algorithms only for a finite time, obtaining a finite sequence of policies. Upon termination, it should be determined which policy to execute. Without monotone improvement of obtained policies, those policies should be evaluated in order to find the best one. However, exact evaluation of even one policy takes an infinite amount of computation for countable-state MDPs, and even if the policies are evaluated approximately, it still requires a considerable amount of computation (in addition to the running time of the algorithm). On the other hand, if the sequence of obtained policies is guaranteed to improve monotonically, then the last obtained policy is always guaranteed to be the best one so far. Furthermore, consider cases where the decision-maker has been executing a policy (e.g., one based on a rule-of-thumb) and wants to improve her decision-making by employing a solution algorithm. If an algorithm improves monotonically and also converges to optimality, then by running the algorithm starting from the policy she has been executing, she can obtain a sequence of policies that are better than the initial policy and also converge to optimality. Meanwhile, the existing solution methods do not provide any guarantee regarding how long they should be run in order to obtain a policy that is better than a given one.

Second, there are simplex-type algorithms for solving classes of CILPs in literature [24, 21, 50] but classes of CILPs considered therein have a special structure that each constraint has only a finite number of variables and each variable appears only in a finite number

of constraints. In the CILP formulation of countable-state MDPs we consider, each constraint may have an infinite number of variables and each variable may appear in an infinite number of constraints (i.e., the coefficient matrix of the CILP can be "dense"). Another key contribution is that we show that even without restrictions on positions of nonzeros in the coefficient matrix, "the MDP structure" in the coefficient matrix of the CILP formulation of a countable-state MDP still enables us to establish the standard LP results and develop a simplex-type algorithm.

Decision-making problems with multiple criteria are often approached by optimizing one criterion while satisfying constraints on the other criteria. In Chapter IV, we consider constrained MDPs, which optimize an objective function while satisfying constraints, typically on budget, quality, and so on. Specifically, we analyze constrained non-stationary MDPs (constrained MDPs with finite state space and non-stationary problem data) by studying their CILP formulations. In this chapter, we focus on characterizing extreme point solutions of such CILPs, and corresponding policies for constrained MDPs. By Bauer's Maximum Principle (e.g., see Theorem 7.69 of [2]), there exists an extreme point optimal solution for finite LPs, and the CILP formulations of constrained non-stationary MDPs as well. For finite LPs, a feasible solution is an extreme point if and only if it is a basic solution. This equivalency translates the geometric concept of an extreme point to the algebraic object of a basic solution. However, such an algebraic characterization of extreme points does not extend to CILPs in general [22]. We provide algebraic necessary conditions for a feasible solution of the CILP formulation of a constrained non-stationary MDP to be an extreme point of its feasible region. Using those necessary conditions, we also establish a necessary and sufficient condition for a feasible solution to be an extreme point, which can be checked by considering a familiar finite dimensional polyhedron. This yields a complete algebraic characterization of extreme points for CILPs representing constrained non-stationary MDPs, and thus, sets foundations towards development of a simplex-type algorithm. Also, by characterizing extreme points, we provide an alternative proof that there exists an optimal policy that is

K-randomized, where K is the number of constraints and a policy is K-randomized if it uses K "more" actions than a Markov deterministic policy (for more precise definitions, see Section 4.4).

1.1 Technical Background

In this section, we first introduce notation and definitions that will be used throughout the thesis. Then we review theoretical properties of base case MDPs (finite state space and stationary problem data) and an LP approach to these MDPs. This provides a common base for the following three chapters because each of the chapters can be viewed as an extension of the LP approach in this section to the class of MDPs studied in the chapter.

A policy π is a sequence $\pi = \{\pi_1, \pi_2, \ldots\}$ of probability distributions $\pi_n(\cdot|h_n)$ over the action set $\mathcal{A}$, where $h_n = (s_0, a_0, s_1, a_2, \ldots, a_{n-1}, s_n)$ is the whole observed history of state visited and actions taken at the beginning of period n. Given an initial state s, each policy π induces a probability measure P_π^s on sequences $\{(s_n, a_n)\}_{n=1}^\infty$, where $(s_n, a_n) \in \mathcal{S} \times \mathcal{A}$ for $n = 1, 2, \ldots$ and defines the state process $\{S_n\}_{n=1}^\infty$ and the action process $\{A_n\}_{n=1}^\infty$. The corresponding expectation operator is denoted by E_π^s. Let

$$J_\pi(s) \triangleq E_\pi^s \left[\sum_{n=1}^{\infty} \alpha^{n-1} c_n(S_n, A_n) \right] \text{ for } s \in \mathcal{S}, \tag{1.1}$$

which is the expected total discounted cost of policy π starting from state s. The function J_π on $\mathcal{S}$ is called the *value function* of the policy π. Let

$$J^\star(s) \triangleq \sup_{\pi \in \Pi} J_\pi(s) \text{ for } s \in \mathcal{S},$$

where Π denotes the set of all policies. The function $J^\star$ on $\mathcal{S}$ is called the *optimal value function*. Then, the goal of an MDP problem is finding a policy that is optimal for any starting state, i.e., a policy $\pi^\star$ such that $J_{\pi^\star,1}(s) = J^\star(s)$ for all $s \in \mathcal{S}$.

A policy π is called *Markov* if the distributions π_n depend only on the current state and time, i.e., $\pi_n(\cdot|h_n) = \pi_n(\cdot|s_n)$. A Markov policy π is called *stationary* if the distributions π_n do not depend on time n, i.e., $\pi_n(\cdot|s) = \pi_m(\cdot|s)$ for $s \in \mathcal{S}$ and time periods m and n. A policy π is said to be *deterministic* if each distribution $\pi_n(\cdot|h_n)$ is concentrated on one action. Let $\Pi, \Pi_M, \Pi_{MD}, \Pi_S$, and Π_{SD} denote the set of all policies, Markov policies, Markov deterministic policies, stationary policies, and stationary deterministic policies, respectively. For a Markov deterministic policy π, let $\pi_n(s)$ denote the action chosen by π at state s in period n, and for a stationary deterministic policy σ (in this thesis, notation σ is used to emphasize the choice of a stationary policy), $\sigma(s)$ denotes the action chosen by σ at state s.

For base case MDPs, it is well known that there exists an optimal policy that is stationary and deterministic and that we can compute such an optimal policy and the optimal value function $J^\star$ by solving Bellman equations.

Theorem 1.1 (cf. Proposition 1.2.2 and 1.2.3 of [10])**.** *MDPs with finite state spaces and stationary problem data (base case) satisfy the following.*

(1) There exists an optimal policy that is stationary and deterministic.

(2) The optimal value function $V^\star$ is the unique solution of Bellman equations:

$$y(s) = \min_{a \in \mathcal{A}} \left\{ c(s,a) + \alpha \sum_{t=1}^{S} p(t|s,a) y(t) \right\} \text{ for } s \in \mathcal{S}.$$

Moreover, the actions that achieve the above minimum form a stationary and deterministic optimal policy.

In particular, for any stationary deterministic policy σ, J_σ equals the optimal value function of a new MDP obtained by allowing only one action $\sigma(s)$ for $s \in \mathcal{S}$, and thus, J_σ is the unique solution of

$$y(s) = c(s,\sigma(s)) + \alpha \sum_{t=1}^{S} p(t|s,\sigma(s)) y(t) \text{ for } s \in \mathcal{S}.$$

In the rest of this section, we summarize the well-known LP approach to base case MDPs

(for more details, e.g., see Section 6.9 of [39]).

Consider the following LP formulation of a base case MDP:

$$
\begin{aligned}
(P)\ \min_{x}\ & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} c(s,a)x(s,a) \\
\text{s.t. } & \sum_{a\in\mathcal{A}} x(s,a) - \alpha\sum_{t=1}^{S}\sum_{a\in\mathcal{A}} p(s|t,a)x(t,a) = \beta(s) \text{ for } s\in\mathcal{S} \\
& x \geq 0,
\end{aligned}
$$

where $\beta : \mathcal{S} \to \mathbb{R}$ satisfies $\beta(s) > 0$ for $s \in \mathcal{S}$ and $\sum_{s\in\mathcal{S}} \beta(s) = 1$ (β can be interpreted as an initial state distribution).

The following LP is its dual:

$$
\begin{aligned}
(D)\ \max_{y}\ & \sum_{s\in\mathcal{S}} \beta(s)y(s) \\
\text{s.t. } & y(s) - \alpha\sum_{t=1}^{S} p(t|s,a)y(t) \leq c(s,a) \text{ for } s\in\mathcal{S}, \quad a\in\mathcal{A}.
\end{aligned}
$$

There is a unique optimal solution to (D) and from Theorem 1.1, it is easy to check that the optimal solution of (D) equals the optimal value function $J^\star$.

The above two LPs are finite LPs and dual to each other, and $y(s) = 0$ for $s \in \mathcal{S}$ is feasible to (D). Thus, their optimal objective function values are equal (strong duality) and complementary slackness is necessary and sufficient for optimality (e.g., see Section 4.3 of [11]).

It is well known that there exists a one-to-one correspondence between feasible solutions of (P) and stationary policies. For a stationary policy, the corresponding solution of (P) is the occupancy measure of the policy, which is the expected total discounted time spent in different state-action pairs under the policy (see Section 3.5.1 for a precise definition). Also, a feasible solution x of this LP is basic (i.e., an extreme point) if and only if for any $s \in \mathcal{S}$ there exists a unique $a(s) \in \mathcal{A}$ such that $x(s, a(s)) > 0$. Note that this equivalent condition

naturally defines a stationary deterministic policy formed by the actions $a(s)$ for $s \in \mathcal{S}$ and in fact, a basic feasible solution x is the occupancy measure of the stationary deterministic policy.

One can obtain a stationary deterministic optimal policy for the MDP by obtaining an optimal basic feasible solution of (P). One of the popular methods to do it is the simplex method. Note that for (P), choosing a basis is equivalent to assigning a unique action to each state. One iteration of the simplex method applied to (P) can be written as follows.

Simplex method for base case MDPs

0. A basis: Have a unique action $\sigma(s)$ for each state s

1. Compute the complementary dual solution: by solving

$$y(s) - \alpha \sum_{t=1}^{S} p(t|s, \sigma(s))y(t) = r(s, \sigma(s)) \text{ for } s = 1, \ldots, S$$

2. Compute reduced costs: for $s = 1, \ldots, S$ and $a \neq \sigma(s)$,

$$\gamma(s, a) \triangleq r(s, a) + \alpha \sum_{t=1}^{S} p(t|s, a)y(t) - y(s)$$

3. Pivot: Choose a nonbasic variable $x(s, a)$ with a negative reduced cost and make it basic

Let us discuss some challenges in extending the above LP approach to the three classes of MDPs considered in this thesis. As previously mentioned, LP formulations of the three classes (non-stationary MDPs, countable-state MDPs, and constrained non-stationary MDPs) are CILPs.[1] Since the useful theoretical properties of finite LPs such as duality results and

[1]Although the LP formulations of MDPs we consider are CILPs, they will look similar to (P) and (D), and simplex-type algorithms we develop or analyze will be also similar to the simplex method presented in this section. When coming across LP formulations of MDPs and simplex-type algorithms in later chapters, we recommend the readers to re-visit this section and check the similarity.

the characterization of extreme points as basic feasible solutions do not directly extend to general CILPs, we have to establish those properties before developing simplex-type algorithms. However, even with those properties in hand, it is challenging to extend the above simplex method to the CILP formulations of MDPs. First, a mere extension of the simplex method to the CILPs may require performing infinite computation and using infinite data in one iteration. For example, the system of equations solved in Step 1 of the above simplex method becomes an infinite system of equations with an infinite number of variables when applied to the CILPs we consider. In addition, Step 2 computes reduced cost of all nonbasic variables, but in the CILPs we consider there is an infinite number of nonbasic variables. Thus, in order to develop an *implementable* extension of the above simplex method, it is necessary to develop proper approximation schemes. Consequently, there are challenges in ensuring that despite approximations we will have to introduce, simplex-type algorithms we develop generate a sequence of extreme points that improves an objective function in every iteration and converges to optimality. In particular, we remind the reader that for CILPs, strictly improving sequences of extreme points may not converge to optimality [21], so a careful consideration is required to ensure that a simplex-type algorithm converges to optimality.

In Chapter II, we establish a complexity bound for the simplex algorithm for non-stationary MDPs introduced in [24] to achieve near-optimality and suggest ways to accelerate the algorithm. We experimentally illustrate the performance of the algorithm and the improvements. In Chapter III, we extend the major theoretical extreme point and duality results to a CILP formulation of countable-state MDPs under standard assumptions for analyzing MDPs with countably-infinite state spaces. Under an additional technical assumption which is satisfied by several applications of interest, we present a simplex-type algorithm that is implementable. We show that the algorithm finds a sequence of policies which improves monotonically and converges to optimality in value. A numerical illustration for inventory management problems is also presented. In Chapter IV, we provide duality results and a

complete characterization of extreme points of a CILP formulation of a constrained non-stationary MDP. The resulting necessary and sufficient condition to be an extreme point can be checked by considering a familiar finite dimensional polyhedron. We also illustrate the condition for special cases. As a corollary, we obtain a new proof of the existence of a K-randomized optimal policy, where K is the number of constraints. We conclude in Chapter V with a summary of contributions and future research directions.

CHAPTER II

Analysis of Algorithms for Non-stationary Markov Decision Processes

2.1 Introduction

Non-stationary MDPs are MDPs with finite state space, finite action space, and non-stationary problem data. Specifically, the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$ are finite with $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$, and given that action a is taken at state s in period n, the system makes a transition to a next state t with probability $p_n(t|s,a)$ incurring cost $0 \leq c_n(s,a;t) \leq c < \infty$, where c is a uniform upper bound on immediate costs. Let $c_n(s,a)$ denote the expected cost incurred by taking action a at state s in period n. The goal of the decision maker is to find a policy that minimizes the expected total discounted cost over infinite horizon. This problem is called a *non-stationary MDP*. Non-stationary MDPs arise in various applications, such as production planning under non-stationary cost and demand data [23, 52], capacity expansion under nonlinear demand [9], and equipment replacement under technological change [8].

Note that a non-stationary MDP can be also viewed as a stationary MDP with a countable number of states by appending states $s \in S$ with time-indices $n \in \mathbb{N}$. The states in the stationary MDP counterpart are $(n,s) \in \mathbb{N} \times \mathcal{S}$. It is well known that for stationary MDPs with a countable number of states and uniformly bounded costs, there exists an optimal

policy that is stationary and deterministic (e.g., see Theorem 6.10.4 of [39]). A stationary deterministic policy in the stationary MDP counterpart corresponds to a Markov deterministic policy in the original non-stationary MDP. Therefore, we can limit our attention to Markov deterministic policies. For a Markov deterministic policy π, let $\pi_n(s)$ denote the action chosen by π at state s in period n.

The simplex algorithm for solving non-stationary MDPs introduced in [24] is a simplex-type algorithm for solving CILP formulations of non-stationary MDPs. We will review the algorithm and prove its convergence in Section 2.2. In short, the algorithm finds a sequence of extreme points of the CILP feasible region whose objective function values improve monotonically and converge to the optimal objective function value. There is a one-to-one correspondence between extreme points of the CILP and Markov deterministic policies. This implies that the simplex algorithm obtains a sequence of improving policies that converges to optimality (see Section 2.2 for more details).

Another solution method for non-stationary MDPs is to solve successively larger but finite horizon truncations to optimality, i.e., for $N = 1, 2, \ldots$, solve the truncated problem of periods $1, \ldots, N$ obtained from the original problem by assuming no cost is incurred after period N. Each truncated problem is solved by backward induction. We call this algorithm the receding horizon approach (RHA). RHA can be viewed as a special case of the shadow simplex method [21]. Also, RHA can be considered as a version of algorithm in [57] that solves stationary MDPs with countable state space by solving successively larger but finite *state* truncation to optimality. RHA for non-stationary MDPs can be viewed as a version of the algorithm in [57] applied to the stationary MDP counterpart of non-stationary MDPs. RHA computes a sequence of policies whose value functions converge to the optimal value function but may not improve monotonically.[1]

In this chapter, we establish rate of convergence results of the simplex algorithm and RHA. For base case MDPs (finite state space and stationary problem data), the simplex

[1] In [57], the convergence of value functions of the policies generated by the algorithm was not established, but it can be proven by arguments similar to those in the paper.

method with Dantzig's pivoting rule (choosing a non-basic variable with the most negative reduced cost) was shown to be strongly polynomial for a fixed discount factor [62]. For non-stationary MDPs, we establish a number of iterations for the simplex algorithm with a particular pivoting rule to find a solution whose objective function value is within a given threshold from the optimal value. Note that one cannot expect any algorithm to solve a non-stationary MDP in finite time because of the non-stationarity of problem data. We also derive a similar result for RHA and compare the two results.

We also introduce a modification to the simplex algorithm that greatly accelerates its empirical performance. The simplex algorithm for non-stationary MDPs in [24] performs only one pivot operation (a pivot operation is replacing a basic variable by a nonbasic variable) per iteration. Policy iteration, one of the popular solution methods for base case MDPs, is the simplex method performing block pivot operations in each iteration. In this spirit, we modify the simplex algorithm so that it performs multiple pivot operations in one iteration.

In Section 2.2, we review CILP formulations of non-stationary MDPs, the simplex algorithm introduced by [24], and related results. We also provide a proof of convergence that is different from the one in [24], which is useful in later sections. In Section 2.3 and 2.4, we establish rate of convergence results of the simplex algorithm and RHA, respectively. Section 2.5 introduces multiple pivoting. In Section 2.6, we provide experimental results for the simplex algorithms introduced in this chapter and RHA for inventory management problems.

2.2 Technical Background

2.2.1 CILP Formulations

Let $x = \{x_n(s, a)\}$ denote a sequence indexed by $n \in \mathbb{N}, s \in \mathcal{S}$, and $a \in \mathcal{A}$. We consider the following CILP formulation of a non-stationary MDP introduced in [24] (see also Chapter

8 of [4] and Chapter 12 of [18] for similar LP formulations of more general classes of MDPs):

$$\text{(NP)} \ \min \ f(x) = \sum_{n=1}^{\infty} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) \tag{2.1}$$

$$\text{s.t.} \ \sum_{a \in \mathcal{A}} x_1(s,a) = 1 \text{ for } s \in \mathcal{S} \tag{2.2}$$

$$\sum_{a \in \mathcal{A}} x_n(s,a) - \sum_{t \in \mathcal{S}} \sum_{a \in \mathcal{A}} p_{n-1}(s|t,a) x_{n-1}(t,a) = 1 \text{ for } n \in \mathbb{N} \setminus \{1\}, \ s \in \mathcal{S} \tag{2.3}$$

$$x_n(s,a) \geq 0 \text{ for } n \in \mathbb{N}, \ s \in \mathcal{S}, \ a \in \mathcal{A}. \tag{2.4}$$

To gain intuition, it is convenient to interpret solutions of (NP) as flows in a directed staged *hypernetwork* with an infinite number of stages (cf. [22]). Stage n in the hypernetwork corresponds to period n of the MDP, and each stage includes S nodes, one for each state in $\mathcal{S}$. There are A directed *hyperarcs* emanating from each node, one for each action in $\mathcal{A}$; thus, a hyperarc (n,s,a) corresponds to action a at state s in stage n. A hyperarc (in a hypernetwork) can connect its "tail" node to multiple "head" nodes; here, a hyperarc (n,s,a) has (n,s) as its tail node, and all nodes $(n+1,t)$ such that $p_n(t|s,a) > 0$ as its head nodes. If all nodes (n,s) have supply of 1 units for $n \in \mathbb{N}$ and $s \in \mathcal{S}$, any x satisfying the constraints of (NP) can be visualized as a flow in this hypernetwork. See Figure 2.1. Specifically, $x_n(s,a)$ is the flow in the hyperarc (n,s,a), and the flow reaching from node (n,s) to node $(n+1,t)$ through this hyperarc equals $p_n(t|s,a)x_n(s,a)$. Moreover, constraints (2.2) and (2.3) ensure *flow balance* at each node. We will refer to any x feasible to (NP) as a *flow* in the hypernetwork. Let $\mathcal{P}$ denote the feasible region of (NP).[2]

For any Markov policy π for the non-stationary MDP, the corresponding flow x can be found as

$$x_n(s,a) = \pi_n(a|s) \sum_{t \in \mathcal{S}} P_\pi^t(S_n = s) \text{ for } n \in \mathbb{N}, \ s \in \mathcal{S}, \ a \in \mathcal{A}.$$

That is, $x_n(s,a)$ is proportional to the probability, under π, of using action a at state s in

[2]In each of the following chapters, $\mathcal{P}$ will denote the feasible region of the primal CILP considered in the chapter.

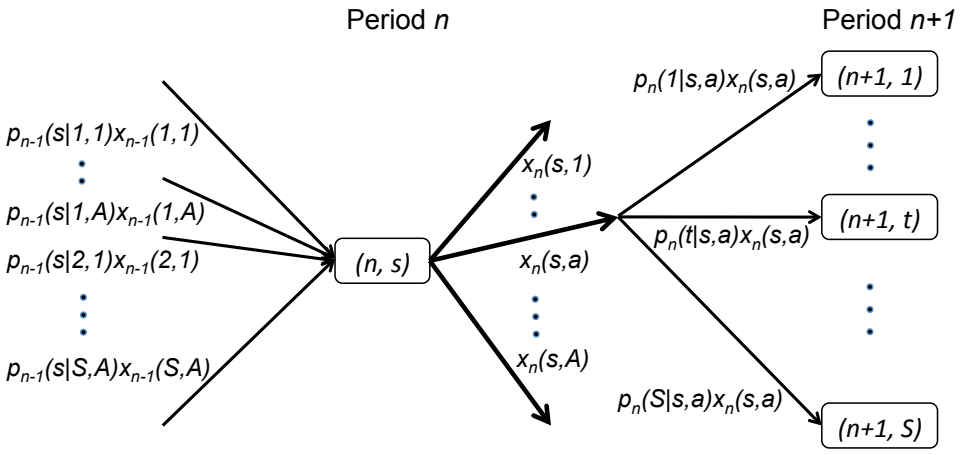


Figure 2.1: Hypernetwork of (NP)

period n, scaled by the sum of the probabilities of reaching this state in period n under the policy π for different initial states, while the total inflow into node (n, s) is precisely the sum $\sum_{t\in\mathcal{S}} P_\pi^t(S_n = s)$. In light of this interpretation, a Markov policy corresponding to any flow x is also easy to identify, and thus, there exists a one-to-one correspondence between the set of Markov policies and the set of flows.

Using the following lemma from [24], one can easily show that the infinite series in the objective function of (NP) converges absolutely for any $x \in \mathcal{P}$.

Lemma 2.1 (Lemma 2.1 of [24])**.** *For a feasible solution x of (NP), for each $n \in \mathbb{N}$,*

$$\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x_n(s, a) = nS, \tag{2.5}$$

and this also implies $x_n(s, a) \leq nS$.

We consider another CILP formulation of a non-stationary MDP which can be derived

by following arguments in Chapter 2.5 of [45] (see also [24]):

$$\begin{aligned}
\text{(ND)}\ \max\ & g(y) = \sum_{n=1}^{\infty} \sum_{s \in \mathcal{S}} y_n(s) \\
\text{s.t.}\ & y_n(s) - \sum_{t \in \mathcal{S}} p_n(t|s,a) y_{n+1}(t) \leq \alpha^{n-1} c_n(s,a) \text{ for } n \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A} \\
& y \in Y,
\end{aligned}$$

where Y is the subspace of all sequences $y = \{y_n(s)\}$ indexed by $(n,s) \in \mathbb{N} \times \mathcal{S}$ such that $|y_n(s)| \leq \alpha^{n-1}\tau_y$ for all (n,s), for which τ_y is a finite constant that depends on y. Note that the infinite series in the objective function converges absolutely for any $y \in Y$. We remark that there exists a unique optimal solution $y^\star$ of (ND) and that $y_n^\star(s)$ equals $\alpha^{n-1} J_n^\star(s)$ where $J_n^\star(s)$ denotes the minimum expected total discounted cost starting from the state s in period n, i.e., $y_n^\star(s)$ equals $J_n^\star(s)$ discounted back to the initial period [24].

Duality results between (NP) and (ND) can be found in literature. Chapter 8 of [4] and Chapter 12 of [18] presented duality results for similar LP formulations of more general classes of MDPs. Specifically for the above LP formulations of non-stationary MDPs, [24] provided duality results, which we state below without proofs.

Theorem 2.2 (Strong duality, Theorem 3.5 of [24] or see Chapter 9 of [4])**.** *(NP) and (ND) have optimal solutions and their optimal objective function values are equal.*

Definition 2.3 (Complementary slackness)**.** Suppose x is feasible to (NP) and $y \in Y$. Then we say that x and y satisfy *complementary slackness* if

$$x_n(s,a) \left(\alpha^{n-1} c_n(s,a) - y_n(s) + \sum_{t \in \mathcal{S}} p_n(t|s,a) y_{n+1}(t) \right) = 0 \text{ for } n \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A}. \quad (2.6)$$

Theorem 2.4 (Theorem 3.4 and 3.6 of [24])**.** *(1) Suppose x is feasible to (NP) and complementary with some $y \in Y$. Then $f(x) = g(y)$. If y is feasible to (ND), then x and y are optimal to (NP) and (ND), respectively.*

(2) Suppose x and y are optimal to (NP) and (ND), respectively. Then, they are complementary.

2.2.2 Basic Feasible Solution and Pivot Operation

By Theorem 2.2 in [24], (NP) has an extreme point optimal solution and the goal of the simplex algorithm we will analyze is to generate a sequence of improving extreme points whose objective function values converge to the optimal value. We define a basic feasible solution of (NP) which provides an algebraic characterization of extreme points of $\mathcal{P}$, as shown by the theorem following the definition.

Definition 2.5 (Basic feasible solution)**.** Suppose x is feasible to (NP). We call it a basic feasible solution of (NP) if, for every $n \in \mathbb{N}$ and $s \in \mathcal{S}$, there is exactly one action $a_n(s) \in \mathcal{A}$ for which $x_n(s, a_n(s)) > 0$.

Theorem 2.6 (Theorem 11.3 of [18] or Theorem 4.3 of [24])**.** *A feasible solution x of (NP) is an extreme point of $\mathcal{P}$ if and only if it is a basic feasible solution.*

Note that a basic feasible solution x naturally defines a Markov deterministic policy π. By the one-to-one correspondence between the set of Markov policies and the set of flows in the hypernetwork, it is easy to show that x is the flow in the hypernetwork corresponding to the policy π. Given a basic feasible solution x, for $n \in \mathbb{N}$ and $s \in \mathcal{S}$, the unique action $a_n(s) \in \mathcal{A}$ such that $x_n(s, a_n(s)) > 0$ is said to be the *basic action* of x at state s in period n.

Let us extend the definition (1.1) as follows: for a policy π,

$$J_{\pi,n}(s) \triangleq E_\pi^s \left[\sum_{k=n}^{\infty} \alpha^{k-1} c_k(S_k, A_k) \right] = \alpha^{n-1} E_\pi^s \left[\sum_{k=n}^{\infty} \alpha^{k-n} c_k(S_k, A_k) \right] \text{ for } n \in \mathbb{N}, s \in \mathcal{S}, \quad (2.7)$$

that is, the expected total discounted cost of policy π starting from state s in period n, discounted back to the initial period. By Theorem 1.1, for a Markov deterministic policy π,

$J_\pi \triangleq \{J_{\pi,n}(s)\}_{(n,s)\in\mathbb{N}\times\mathcal{S}}$ is the unique solution of the following system of equations:

$$y_n(s) = \alpha^{n-1}c_n(s,\pi_n(s)) + \sum_{t\in\mathcal{S}} p_n(t|s,a)y_n(t) \text{ for } n \in \mathbb{N} \text{ and } s \in \mathcal{S}. \tag{2.8}$$

That is, for a basic feasible solution x and the corresponding policy π, J_π is the unique complementary solution of x. From the definition of J_π, it is evident that

$$0 \le J_{\pi,n}(s) \le \alpha^{n-1}\frac{c}{1-\alpha} \text{ for } \pi \in \Pi_{MD}, n \in \mathbb{N}, s \in \mathcal{S}, \tag{2.9}$$

and thus, $J_\pi \in Y$.

Let $f^\star$ denote the optimal objective function value of (NP). The following theorem shows that if an algorithm generates a sequence of basic feasible solutions of (NP) whose objective function values converge to $f^\star$, then the algorithm generates a sequence of policies converging to optimality for non-stationary MDPs.

Theorem 2.7. *Let x^k be a sequence of basic feasible solutions and π^k be the policy corresponding to x^k. If $\lim_{k\to\infty} f(x^k) = f^\star$, then $\lim_{k\to\infty} J_{\pi^k,n}(s) = J_n^\star(s)$ for $n \in \mathbb{N}, s \in \mathcal{S}$.*

Proof: Because of the definition of $J^\star$, we have $J_{\pi^k,n}(s) - J_n^\star(s) \ge 0$ for $n \in \mathbb{N}, s \in \mathcal{S}$. Moreover, we have

$$\begin{aligned} f(x^k) - f^\star = g(J_{\pi^k}) - g(J^\star) &= \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} J_{\pi^k,n}(s) - \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} J_n^\star(s) \\ &= \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} (J_{\pi^k,n}(s) - J_n^\star(s)), \end{aligned}$$

where the first equality is obtained from complementary slackness and the third inequality is obtained because J_π and $J^\star$ are both feasible to (ND). Since each term in the last sum is nonnegative and the sum converges to zero, each term converges to zero, and therefore, the theorem is proven. □

In the rest of this subsection, we define a pivot operation and provide an expression of

the change in objective function value made by a pivot operation. Consider a basic feasible solution x and its complementary solution y. We first define reduced costs.

Definition 2.8. Given a basic feasible solution x and its complementary solution y, *reduced cost* $\gamma_n(s,a)$ of a hyperarc (n,s,a) is defined as the slack in the corresponding constraint in (ND):

$$\gamma_n(s,a) \triangleq \alpha^{n-1}c_n(s,a) + \sum_{t\in\mathcal{S}} p_n(t|s,a)y_{n+1}(t) - y_n(s). \tag{2.10}$$

Recall that $y = J_\pi$ where π is the Markov deterministic policy corresponding to x. From the definition of the reduced cost and (2.9), we can easily prove the next lemma.

Lemma 2.9. *The reduced cost γ satisfies*

$$-\alpha^{n-1}\frac{c}{1-\alpha} \le \gamma_n(s,a) \le \alpha^{n-1}\frac{c}{1-\alpha} \text{ for } n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}. \tag{2.11}$$

Fix a period n, a state s, and an action $a \neq \pi_n(s)$. Consider a Markov deterministic policy ρ obtained from π by changing the basic action at state s in period n to a. We call this procedure for obtaining ρ from π a *pivot operation*. Let z be the basic feasible solution corresponding to ρ. The next proposition shows the relation between the change in objective function value made by this pivot operation and the reduced cost $\gamma_n(s,a)$.

Proposition 2.10. *The difference in objective function values at basic feasible solution x and the adjacent basic feasible solution z is given by*

$$f(z) - f(x) = (1+\theta^n(s))\gamma_n(s,a) \tag{2.12}$$

where $\theta^n(s) = x_n(s,a_n(s)) - 1 \ge 0$ for $n\in\mathbb{N}, s\in\mathcal{S}$.

Proof: Proposition 5.1 of [24] showed that $f(z) - f(x) = (1+\theta^n(s))\gamma_n(s,a)$ for some $\theta^n(s) \ge 0$. Thus, we only have to prove that $1+\theta^n(s) = x_n(s,a_n(s))$, which is done in Section 2.7.1. □

In this proposition, x and z are two extreme points whose basic actions differ only at one period-state pair. We define such two extreme points to be *adjacent*. Note that the set of nonnegativity constraints (2.4) binding at x differs from those binding at z only by one constraint. Thus, the definition of adjacent extreme points of $\mathcal{P}$ naturally aligns with the general notion of adjacency in a finite-dimensional polyhedron. Notice also that for a given extreme point of $\mathcal{P}$, there are a countably infinite number of adjacent extreme points.

2.2.3 The Simplex Algorithm for Non-stationary MDPs

Proposition 2.10 provides an exact expression of the difference in objective function value made by a pivot operation. Also, the proposition implies that, if one finds a nonbasic variable that has a negative reduced cost, then the solution resulting from the pivot operation has a strictly lower objective function value. However, in order to compute a reduced cost, one should first compute the complementary solution y and computing y requires either solving the infinite system of equations (2.8) or computing (2.7) exactly, thus an infinite amount of computation would be required. Consequently, the implementable simplex algorithm in [24] (stated below) approximates y and reduced costs using an m-horizon truncation of the non-stationary MDP and increases m as necessary to enhance the approximation until it finds a nonbasic variable with a negative reduced cost. In addition, recall that a strictly improving sequence of extreme points may not converge to optimality for general CILPs, as mentioned in Chapter I. In other words, the improvement in each iteration should be "sufficiently large" for the sequence of objective function values at obtained extreme points not to get stalled prematurely before converging to optimality. It was shown in [24] that the following algorithm, which we denote as SA in the rest of this chapter, generates a sequence of improving extreme points that converges to optimality.

The simplex algorithm (SA)

1. Initialize: Set iteration counter $k := 0$. Fix basic actions $a_n^0(s) \in A$ for $s \in S$ and $n \in \mathbb{N}$. Set $m := 1$.

2. Find a nonbasic arc with the most negative *approximate* reduced cost:

   (a) Define $m(k) \triangleq \infty$ and $\gamma^{k,\infty} \triangleq 0$.

   (b) Let $y^{k,m}$ be the solution of the *finite* system of equations

$$y_n^{k,m}(s) = \alpha^{n-1} c_n(s, a_n^k(s)) + \sum_{t \in S} p_n(t|s, a_n^k(s)) y_{n+1}^{k,m}(t) \text{ for } s \in S, n \leq m, \quad (2.13)$$

$$y_{m+1}^{k,m}(s) = 0. \quad (2.14)$$

   (c) Compute *approximate* nonbasic reduced costs

$$\gamma_n^{k,m}(s,a) \triangleq \alpha^{n-1} c_n(s,a) + \sum_{t \in S} p_n(t|s,a) y_{n+1}^{k,m}(t) - y_n^{k,m}(s) \quad (2.15)$$

   for $n \leq m$, $s \in S, a \in A$ such that $a \neq a_n^k(s)$.

   (d) Compute the smallest *approximate* nonbasic reduced cost

$$\gamma_{\min}^{k,m} \triangleq \min_{(n,s,a)} \gamma_n^{k,m}(s,a) \quad (2.16)$$

   (e) If $\gamma_{\min}^{k,m} < -\alpha^m \frac{c}{1-\alpha}$, set $m(k) = m$, $(n^k, s^k, a^k) = \arg\min_{(n,s,a)} \gamma_n^{k,m}(s,a)$, $a_{n^k}^{k+1}(s^k) = a^k$ and $a_n^{k+1}(s) = a_n^k(s)$ for $(n,s) \neq (n^k, s^k)$, and go to Step 3 below; else set $m := m + 1$ and go to Step 2(b) above.

3. Set $k := k + 1$ and go to Step 2.

In this algorithm, m is the number of periods used to approximate reduced costs and is called *strategy horizon*. The above simplex algorithm is different from the one in [24] only in the way it sets the strategy horizon at the beginning of each iteration: in Step 2(a), the algorithm in [24] resets $m := 1$ at the beginning of each iteration, but this begins iteration k with strategy horizon $m = m(k-1)$, i.e., the strategy horizon at which the previous iteration found a pivot operation satisfying the conditions. We consider this modified version

because we empirically experienced that the modification greatly accelerates convergence of the algorithm. It can be easily shown that all of the results in [24] still hold after this change because a starting value of m in each iteration does not play any role in their proofs.

Let x^k denote the basic feasible solution found in iteration k of SA with basic actions $a_n^k(s)$ for $s \in \mathcal{S}$ and $n \in \mathbb{N}$. Let y^k be the complementary solution of x^k. Then, $y^{k,m}$ computed in Step 2(b) approximates y^k with the following error bound.

Lemma 2.11 (Lemma 5.4 of [24])**.** *The approximation $y^{k,m}$ of y^k satisfies*

$$y_n^{k,m}(s) \leq y_n^k(s) \leq y_n^{k,m}(s) + \alpha^m \frac{c}{1-\alpha} \text{ for } s \in S, n \leq m+1. \tag{2.17}$$

Remark 2.12. The above inequality has the following interpretation. $y_n^k(s)$ is the infinite-horizon cost-to-go of the current policy at the kth iteration starting from the period-state pair (n, s) discounted back to the initial period, whereas $y_n^{k,m}(s)$ is the cost-to-go of the current policy at the kth iteration starting from the period-state pair (n, s) *only to period* m discounted back to the initial period. From period $m+1$, an upper bound on the cost that the current policy can incur is obtained by paying c in every period from period $m+1$, and a lower bound is obtained by paying 0 from period $m+1$ onward.

In iteration k of the algorithm, the approximate reduced costs computed in Step 2(c) satisfy the following error bound.

Lemma 2.13. *The approximation $\gamma^{k,m}$ of γ^k satisfies*

$$\gamma_n^{k,m}(s,a) - \alpha^m \frac{c}{1-\alpha} \leq \gamma_n^k(s,a) \leq \gamma_n^{k,m}(s,a) + \alpha^m \frac{c}{1-\alpha} \text{ for } n \leq m, s \in \mathcal{S}, a \in \mathcal{A}. \tag{2.18}$$

Remark 2.14. The above inequality has a similar interpretation with the one of (2.17). $\gamma_n^k(s,a)$ is the difference of infinite-horizon cost incurred by changing the basic action of the period-state pair (n, s) from the action given by the current policy to a new basic action a. $\gamma_n^{k,m}(s,a)$ is the difference of m-horizon cost incurred by the same procedure. Assume

that from the period $m+1$, taking the current basic action in period n at state s causes the maximum cost c in periods $m+1, m+2, \ldots$ while the action a causes cost zero in periods $m+1, m+2, \ldots$. This case gives the lower bound (the most optimistic case for the new basic action) on the approximate reduced cost and it is the left side of the above inequality. The upper bound is derived in a similar way.

Let $\gamma^k_{\min}$ be the minimum (exact) reduced cost over all nonbasic variables in iteration k. The following lemma shows the relation between $\gamma^k_{\min}$, the approximate reduced cost $\gamma^{k,m(k)}_{n^k}(s^k, a^k)$ of the nonbasic variable chosen by the algorithm, and its (exact) reduced cost $\gamma^k_{n^k}(s^k, a^k)$.

Lemma 2.15. *The minimum reduced cost $\gamma^k_{\min}$, the approximate reduced cost $\gamma^{k,m(k)}_{n^k}(s^k, a^k)$ of the nonbasic variable chosen by the algorithm, and its (exact) reduced cost $\gamma^k_{n^k}(s^k, a^k)$ satisfy*

$$\gamma^k_{n^k}(s^k, a^k) - 2\alpha^{m(k)}\frac{c}{1-\alpha} \le \gamma^{k,m(k)}_{n^k}(s^k, a^k) - \alpha^{m(k)}\frac{c}{1-\alpha} \le \gamma^k_{\min}. \tag{2.19}$$

Proof: Let a hyperarc $(\bar{n}, \bar{s}, \bar{a})$ attain the minimum reduced cost $\gamma^k_{\min}$, i.e., $\gamma^k_{\bar{n}}(\bar{s}, \bar{a}) = \gamma^k_{\min}$. Suppose $\bar{n} > m(k)$. By the choice of $m(k)$ and (n^k, s^k, a^k),

$$\gamma^{k,m(k)}_{n^k}(s^k, a^k) < -\alpha^{m(k)}\frac{c}{1-\alpha} < -\alpha^{\bar{n}}\frac{c}{1-\alpha} \le \gamma^k_{\bar{n}}(\bar{s}, \bar{a}).$$

If $\bar{n} \le m(k)$, then

$$\gamma^{k,m(k)}_{n^k}(s^k, a^k) - \alpha^{m(k)}\frac{c}{1-\alpha} \le \gamma^{k,m(k)}_{\bar{n}}(\bar{s}, \bar{a}) - \alpha^{m(k)}\frac{c}{1-\alpha} \le \gamma^k_{\bar{n}}(\bar{s}, \bar{a}),$$

by (2.18). By combining the above two inequalities, we obtain

$$\gamma^{k,m(k)}_{n^k}(s^k, a^k) - \alpha^{m(k)}\frac{c}{1-\alpha} \le \gamma^k_{\bar{n}}(\bar{s}, \bar{a}) = \gamma^k_{\min}.$$

By applying (2.18) again, we obtain

$$\gamma^k_{n^k}(s^k, a^k) - 2\alpha^{m(k)}\frac{c}{1-\alpha} \leq \gamma^{k,m(k)}_{n^k}(s^k, a^k) - \alpha^{m(k)}\frac{c}{1-\alpha} \leq \gamma^k_{\min}.$$

□

The next lemma states that SA generates an improving sequence of extreme points.

Lemma 2.16 (Lemma 5.5 and 5.6 of [24]). *If x^k is not optimal to (D), then Step 2 terminates and $f(x^{k+1}) < f(x^k)$.*

2.2.4 Proof of Convergence

In this subsection, we prove that SA generates a sequence of extreme points whose objective function values converge to optimality. This proof is different from the one in [24]. This alternative proof will be helpful in Section 2.3 in analyzing convergence rate of the simplex algorithm.

In addition to the above preliminary results, [24] showed the following lemma about $m(k)$ and $\gamma^{k,m(k)}_{\min} = \gamma^{k,m(k)}_{n^k}(s^k, a^k)$ of SA, which is necessary to prove convergence of the algorithm.

Lemma 2.17 (Lemma 5.7 of [24]). *The sequence $m(k) \to \infty$ as $k \to \infty$. Also, $\gamma^{k,m(k)}_{n^k}(s^k, a^k)$ $\to 0$ as $k \to \infty$.*

Let $x^\star$ denote an extreme point optimal solution of (NP) and $a^\star_n(s)$ denote its (optimal) basic action at state s in period n. Then, the optimality gap of a basic feasible solution x can be written using reduced costs as follows.

Lemma 2.18. *The optimality gap of a basic feasible solution x is written as*

$$f(x) - f^\star = \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} x^\star_n(s, a^\star_n(s))(-\gamma_n(s, a^\star_n(s))). \tag{2.20}$$

Proof: In Section 2.7.2.

Remark 2.19. In Appendix A, we provide another proof of Lemma 2.18 by analyzing geometry of $\mathcal{P}$. We show that given an extreme point of $\mathcal{P}$, any point in $\mathcal{P}$ can be represented as the sum of the extreme point and a conic combination of vectors from the extreme point to its adjacent extreme points, and using the result, we prove Lemma 2.18.

Using the lemmas established so far, we provide a proof of convergence of SA that does not use the duality results, unlike the proof in [24].

Theorem 2.20. *Let x^k be the sequence of basic feasible solutions generated by SA. Then* $\lim_{k\to\infty} f(x^k) = f^\star$.

To prove this theorem, it suffices to show that for any number $\epsilon > 0$, there exists a positive integer $K(\epsilon)$ such that for $k \geq K(\epsilon)$, $f(x^k) - f^\star \leq \epsilon$. In the proof and also in Section 2.3, we will repeatedly use the following positive integer $N(\epsilon)$:

$$N(\epsilon) \triangleq \left\lceil \frac{\log(\epsilon(1-\alpha)^3/2cS)}{\log\alpha + 1 - \alpha} \right\rceil. \tag{2.21}$$

For any basic feasible solution x and its complementary solution y, we have $f(x) = g(y) = \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} y_n(s)$ and $y_n(s) \leq \alpha^{n-1}\frac{c}{1-\alpha}$ by (2.9), and thus, $f(x) \leq \frac{cS}{(1-\alpha)^2}$. Therefore, any $\epsilon \geq \frac{cS}{(1-\alpha)^2}$ is not of interest to us. For $\epsilon < \frac{cS}{(1-\alpha)^2}$, the numerator of the fraction in (2.21) is negative. Since $\log\alpha < \alpha - 1$ for $0 < \alpha < 1$, $N(\epsilon)$ is positive for all ϵ of our interest.

This number $N(\epsilon)$ will be used to bound contribution of decisions after period $N(\epsilon)$ to the objective function f. The property of $N(\epsilon)$ we will mainly use is described in the following proposition.

Proposition 2.21. *The positive integer $N(\epsilon)$ in (2.21) satisfies*

$$\frac{cS}{1-\alpha}\sum_{n=N(\epsilon)+1}^{\infty} n\alpha^{n-1} = \frac{cS\alpha^{N(\epsilon)}}{(1-\alpha)^2}\left(N(\epsilon) + \frac{1}{1-\alpha}\right) \leq \frac{\epsilon}{2}. \tag{2.22}$$

Proof: In Section 2.7.3 □

Proof of Theorem 2.20: Fix an arbitrary $\epsilon > 0$. By Lemma 2.18,

$$\begin{aligned}
f(x^k) - f^\star &= \sum_{n=1}^{\infty} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) \\
&= \sum_{n=1}^{N(\epsilon)} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) + \sum_{n=N(\epsilon)+1}^{\infty} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) \\
&\leq \sum_{n=1}^{N(\epsilon)} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) + \sum_{n=N(\epsilon)+1}^{\infty} nS\alpha^{n-1}\frac{c}{1-\alpha} \\
&\leq \sum_{n=1}^{N(\epsilon)} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) + \frac{\epsilon}{2}
\end{aligned}$$

where the first inequality is obtained from Lemma 2.1 and 2.9, and the second inequality is obtained from Proposition 2.21. By Lemma 2.15,

$$-\gamma_n^k(s, a_n^\star(s)) \leq -\gamma_{\min}^k \leq \alpha^{m(k)}\frac{c}{1-\alpha} - \gamma_{n^k}^{k,m(k)}(s^k, a^k).$$

Note that the right hand side above is strictly positive by the way (n^k, s^k, a^k) is chosen. Then, continuing the chain of inequalities

$$\begin{aligned}
f(x^k) - f^\star &\leq \sum_{n=1}^{N(\epsilon)} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^k(s, a_n^\star(s))) + \frac{\epsilon}{2} \\
&\leq \sum_{n=1}^{N(\epsilon)} nS\left(\alpha^{m(k)}\frac{c}{1-\alpha} - \gamma_{n^k}^{k,m(k)}(s^k, a^k)\right) + \frac{\epsilon}{2} \\
&\leq \frac{N(\epsilon)(N(\epsilon)+1)S}{2}\left(\alpha^{m(k)}\frac{c}{1-\alpha} - \gamma_{n^k}^{k,m(k)}(s^k, a^k)\right) + \frac{\epsilon}{2}.
\end{aligned}$$

By Lemma 2.17, we can find $K(\epsilon)$ such that for $k \geq K(\epsilon)$,

$$\alpha^{m(k)}\frac{c}{1-\alpha} - \gamma_{n^k}^{k,m(k)}(s^k, a^k) \leq \frac{\epsilon}{N(\epsilon)(N(\epsilon)+1)S}.$$

For such k, we have $f(x^k) - f^\star \leq \epsilon$ and the theorem is proven. □

2.3 Convergence Rate of Simplex Algorithm

In the previous section, we proved that for any $\epsilon > 0$, there exists $K(\epsilon)$ such that for $k \geq K(\epsilon)$, $f(x^k) - f^\star \leq \epsilon$. In this section, we derive such a $K(\epsilon)$ for the simplex algorithm with a particular pivoting rule, i.e., we provide a rate of convergence result. We first introduce the pivoting rule.

2.3.1 Simplex Algorithm with Accuracy δ

For a given $\delta > 0$, let $M(\delta)$ be the smallest positive integer satisfying $2\alpha^{M(\delta)}\frac{c}{1-\alpha} < \delta$. Then, the only modification we make to SA is to set $m = M(\delta)$ in Step 1. All the results presented in the previous section still hold after this modification. In addition, the following proposition shows that the reduced cost of the nonbasic variable chosen by the SA(δ) is within δ from the true minimum reduced cost.

Proposition 2.22. *For a given $\delta > 0$, the nonbasic variable (n^k, s^k, a^k) found in iteration k of the SA(δ) satisfies*

$$\gamma^k_{n^k}(s^k, a^k) - \delta \leq \gamma^k_{\min}. \tag{2.23}$$

Proof: By the modified Step 2, we have $m(k) \geq M(\delta)$. By Lemma 2.15,

$$\gamma^k_{n^k}(s^k, a^k) - \delta < \gamma^k_{n^k}(s^k, a^k) - 2\alpha^{M(\delta)}\frac{c}{1-\alpha} \leq \gamma^k_{n^k}(s^k, a^k) - 2\alpha^{m(k)}\frac{c}{1-\alpha} \leq \gamma^k_{\min}.$$

□

We call the modified algorithm the *simplex algorithm with accuracy δ*, or in short, SA(δ).

2.3.2 Convergence Rate of SA(δ)

For a given $\epsilon > 0$, if a solution x satisfies $f(x) - f^\star \leq \epsilon$, then x is said to be ϵ-optimal. We will establish an upper bound on the number of iterations for the simplex algorithm with a certain accuracy to obtain an ϵ-optimal solution.

Let

$$\delta(\epsilon) \triangleq \frac{\epsilon}{N(\epsilon)(N(\epsilon)+1)S}. \tag{2.24}$$

From Lemma 2.16 and Theorem 2.20, we know that SA($\delta(\epsilon)$) generates a sequence of solutions x^k such that $f(x^k)$ is decreasing and converges to $f^\star$. For a given $\epsilon > 0$, the next key lemma illustrates how fast $f(x^k)$ approaches $f^\star + \epsilon$.

Lemma 2.23. *For a given $\epsilon > 0$, let $N(\epsilon)$ be defined as (2.21). Let x^k be the sequence of solutions generated by SA($\delta(\epsilon)$) starting from x^0. If $f(x^{k-1}) - (f^\star + \epsilon) > 0$, then for $l = 0, 1, \ldots, k-1$,*

$$\frac{f(x^{l+1}) - (f^\star + \epsilon)}{f(x^l) - (f^\star + \epsilon)} \leq 1 - \frac{2\delta(\epsilon)}{\epsilon} = 1 - \frac{2}{N(\epsilon)(N(\epsilon)+1)S}, \tag{2.25}$$

and thus, we have

$$\frac{f(x^k) - (f^\star + \epsilon)}{f(x^0) - (f^\star + \epsilon)} \leq \left(1 - \frac{2\delta(\epsilon)}{\epsilon}\right)^k = \left(1 - \frac{2}{N(\epsilon)(N(\epsilon)+1)S}\right)^k. \tag{2.26}$$

Proof: Note that since $f(x^{k-1}) - (f^\star + \epsilon) > 0$ and $f(x^l)$ is decreasing in l, we have $f(x^l) - (f^\star + \epsilon) > 0$ for $l = 0, 1, \ldots, k-1$. By following the same steps as in the proof of Theorem 2.20, for any positive integer l, we obtain

$$\begin{aligned} f(x^l) - f^\star &\leq \sum_{n=1}^{N(\epsilon)} \sum_{s \in \mathcal{S}} x_n^\star(s, a_n^\star(s))(-\gamma_n^l(s, a_n^\star(s))) + \frac{\epsilon}{2} \\ &\leq -\gamma_{\min}^l \sum_{n=1}^{N(\epsilon)} nS + \frac{\epsilon}{2} \\ &= -\gamma_{\min}^l \frac{N(\epsilon)(N(\epsilon)+1)S}{2} + \frac{\epsilon}{2} = -\gamma_{\min}^l \frac{\epsilon}{2\delta(\epsilon)} + \frac{\epsilon}{2}. \end{aligned} \tag{2.27}$$

By Proposition 2.10, for some $\theta \geq 0$,

$$\begin{aligned} f(x^l) - f(x^{l+1}) &= (1+\theta)(-\gamma^l_{n^l}(s^l, a^l)) \geq -\gamma^l_{n^l}(s^l, a^l) \geq -\gamma^l_{\min} - \delta(\epsilon) \\ &\geq \frac{2(f(x^l) - (f^\star + \epsilon/2))\delta(\epsilon)}{\epsilon} - \delta(\epsilon) = \frac{2(f(x^l) - (f^\star + \epsilon))\delta(\epsilon)}{\epsilon} \end{aligned} \tag{2.28}$$

where the first inequality is obtained from $\gamma^l_{n^l}(s^l, a^l) < 0$, the second inequality from (2.23), and the third inequality from (2.27). By rearranging terms in the left and right hand sides of (2.28), we obtain, for $l = 0, 1, \ldots, k-1$,

$$\frac{f(x^{l+1}) - (f^\star + \epsilon)}{f(x^l) - (f^\star + \epsilon)} \leq 1 - \frac{2\delta(\epsilon)}{\epsilon} = 1 - \frac{2}{N(\epsilon)(N(\epsilon)+1)S}.$$

By multiplying the above inequality for $l = 0, 1, \ldots, k-1$, we obtain

$$\frac{f(x^k) - (f^\star + \epsilon)}{f(x^0) - (f^\star + \epsilon)} \leq \left(1 - \frac{2\delta(\epsilon)}{\epsilon}\right)^k = \left(1 - \frac{2}{N(\epsilon)(N(\epsilon)+1)S}\right)^k$$

and this proves the lemma. □

In the proof of the above lemma, (2.28) gives a positive lower bound on $f(x^l) - f(x^{l+1})$ only when $f(x^l) - (f^\star + \epsilon) > 0$ (and for this, we need the condition $f(x^{k-1}) - (f^\star + \epsilon) > 0$). In other words, we can only provide a lower bound on the amount of improvement achieved by an iteration of the simplex algorithm when the current iterate has a sufficiently large optimality gap.

Let $\phi \triangleq f(x^0) - f^\star$ denote the initial optimality gap and let $\bar{\phi}$ be an upper bound on ϕ. Later, we will provide a way to compute such an upper bound. The following main theorem establishes the number of iterations of SA($\delta(\epsilon/2)$) to obtain an ϵ-optimal solution.

Theorem 2.24. *Let x^0 be an initial basic feasible solution and $\bar{\phi}$ be an upper bound of the initial optimality gap $f(x^0) - f^\star$. Then the sequence of solutions x^k generated by SA($\delta(\epsilon/2)$)*

satisfies $f(x^k) - f^\star \leq \epsilon$ *for* $k \geq K_1(\bar{\phi}, \epsilon)$ *where*

$$K_1(\bar{\phi}, \epsilon) \triangleq \left\lceil \frac{\epsilon}{4\delta(\epsilon/2)} \log\left(\frac{2\bar{\phi}}{\epsilon} - 1\right) \right\rceil.$$

Proof: If we show that $f(x^{K_1(\bar{\phi},\epsilon)}) - f^\star \leq \epsilon$, then the conclusion trivially holds by monotonicity of $f(x^k)$.

Suppose $f(x^{K_1(\bar{\phi},\epsilon)}) - f^\star > \epsilon > \epsilon/2$. From Lemma 2.23 (with $\epsilon/2$ in place of ϵ), for $k \leq K_1(\bar{\phi}, \epsilon)$,

$$\begin{aligned} f(x^k) - f^\star &\leq \frac{\epsilon}{2} + \left(f(x^0) - f^\star - \frac{\epsilon}{2}\right)\left(1 - \frac{4\delta(\epsilon/2)}{\epsilon}\right)^k \\ &\leq \frac{\epsilon}{2} + \left(\bar{\phi} - \frac{\epsilon}{2}\right)\left(1 - \frac{4\delta(\epsilon/2)}{\epsilon}\right)^k. \end{aligned} \tag{2.29}$$

By the definition of $K_1(\bar{\phi}, \epsilon)$, for $k \geq K_1(\bar{\phi}, \epsilon)$,

$$k \log\left(1 - \frac{4\delta(\epsilon/2)}{\epsilon}\right) \leq -k\frac{4\delta(\epsilon/2)}{\epsilon} \leq \log \frac{1}{2\bar{\phi}/\epsilon - 1} \tag{2.30}$$

where the first inequality is obtained from $\log(1-u) \leq -u$ for $u < 1$. Applying (2.30) to (2.29), we obtain $f(x^{K_1(\bar{\phi},\epsilon)}) - f^\star \leq \epsilon$. □

If $c, \alpha, \bar{\phi}$ are considered fixed constants, then the number of iterations that guarantees ϵ-optimality is $K_1(\bar{\phi}, \epsilon) = \mathcal{O}(S(\log S)^2(\log \epsilon)^3)$.

By applying arguments nearly identical to the above proof, we can prove the following theorem as well.

Theorem 2.25. *Let* z^0 *be an initial basic feasible solution and suppose that* $f(z^0) - f^\star \leq \epsilon$. *Then the sequence of solutions* z^k *generated by SA*$(\delta(\epsilon/4))$ *satisfies* $f(z^k) - f^\star \leq \frac{\epsilon}{2}$ *for* $k \geq K_2(\epsilon)$ *where*

$$K_2(\epsilon) \triangleq \left\lceil \frac{N(\epsilon/4)(N(\epsilon/4)+1)S}{2} \log 3 \right\rceil.$$

This theorem shows that the number of iterations guaranteed to reduce optimality gap

from ϵ to its half is $K_2(\epsilon) = \mathcal{O}(S(\log S)^2(\log \epsilon)^2)$ given $c, \alpha, \bar{\phi}$ are considered fixed constants.

As promised, we provide an upper bound $\bar{\phi}$ on ϕ. Since costs $c_n(s,a)$ are nonnegative, by Lemma 2.11 and (2.9), for any positive integer m, we can obtain an upper bound $\bar{\phi}$ as follows:

$$\begin{aligned}
\phi = f(x^0) - f^\star \le f(x^0) = g(y^0) &= \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} y_n^0(s) = \sum_{n=1}^{m}\sum_{s\in\mathcal{S}} y_n^0(s) + \sum_{n=m+1}^{\infty}\sum_{s\in\mathcal{S}} y_n^0(s) \\
&\le \sum_{n=1}^{m}\sum_{s\in\mathcal{S}} \left[y_n^{0,m}(s) + \alpha^m \frac{c}{1-\alpha} \right] + \sum_{n=m+1}^{\infty}\sum_{s\in\mathcal{S}} \alpha^{n-1}\frac{c}{1-\alpha} \\
&= \sum_{n=1}^{m}\sum_{s\in\mathcal{S}} y_n^{0,m}(s) + mS\alpha^m \frac{c}{1-\alpha} + S\alpha^m \frac{c}{(1-\alpha)^2} \triangleq \bar{\phi}.
\end{aligned}$$

In fact, for any basic feasible solution x^k and any positive integer m, we have

$$\begin{aligned}
\sum_{n=1}^{m}\sum_{s\in\mathcal{S}} y_n^{k,m}(s) &\le \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} y_n^{k}(s) (= g(y^k)) \\
&\le \sum_{n=1}^{m}\sum_{s\in\mathcal{S}} y_n^{k,m}(s) + mS\alpha^m \frac{c}{1-\alpha} + S\alpha^m \frac{c}{(1-\alpha)^2}. \qquad (2.31)
\end{aligned}$$

Therefore, we can estimate the objective function value $f(x^k) = g(y^k)$ with arbitrary precision by computing the left hand side of (2.31) for a large enough m. In Section 2.6, we use this to evaluate policies generated by solution algorithms.

2.4 Convergence Rate of RHA

Since SA(δ) performs one pivot operation per iteration, Theorems 2.24 and 2.25 provide an upper bound on the number of pivot operations for SA(δ) to obtain a near-optimal solution. In this section, we provide a similar result for RHA. RHA can be formally written as follows.

Receding horizon approach (RHA)

1. Initialize: Set $N := 1$. Fix basic actions $a_n^0(s) \in \mathcal{A}$ for $s \in \mathcal{S}$ and $n \in \mathbb{N}$.

2. Let $w_{N+1}(s) = 0$ for $s \in \mathcal{S}$.

3. For $n = N, N-1, \ldots, 1$,

   (a) For $s \in \mathcal{S}$,

      (i) Compute

$$w_n(s) = \min_{a \in \mathcal{A}} \left\{ c_n(s,a) + \alpha \sum_{t \in \mathcal{S}} p_n(t|s,a) w_{n+1}(t) \right\},$$
$$a_n^N(s) = \arg\min_{a \in \mathcal{A}} \left\{ c_n(s,a) + \alpha \sum_{t \in \mathcal{S}} p_n(t|s,a) w_{n+1}(t) \right\}. \tag{2.32}$$

4. Set $N := N + 1$ and go to Step 2.

Let z^N be the basic feasible solution corresponding to the policy formed by the basic actions $a_n^N(s)$ for $n \in \mathbb{N}$ and $s \in \mathcal{S}$ computed by RHA.

For a given $\epsilon > 0$, the next theorem gives a value of N for which z^N is an ϵ-optimal solution. In Step 3(a)(i) of RHA, basic action of one period-state pair (n, s) is updated, which is a pivot operation. The following theorem also gives the number of pivot operations for RHA to obtain an ϵ-optimal solution.

Theorem 2.26. *Let z^N be the sequence of basic feasible solutions generated by RHA. Then $f(z^N) - f^\star \leq \epsilon$ for $N \geq N'(\epsilon)$ where*

$$N'(\epsilon) \triangleq \left\lceil \frac{\log(\epsilon(1-\alpha)^2/cS)}{\log \alpha + 1 - \alpha} \right\rceil. \tag{2.33}$$

The number of pivot operations for RHA to obtain $z^{N'(\epsilon)}$ is at most $N'(\epsilon)(N'(\epsilon)+1)S/2$.

Proof: In Section 2.7.4.

This theorem shows that, if c and α are considered fixed constants, the number of pivot operations for RHA to achieve ϵ-optimality is $N'(\epsilon)(N'(\epsilon)+1)S/2 = \mathcal{O}(S(\log S)^2(\log \epsilon)^2)$. The number of pivot operations for SA($\delta(\epsilon/2)$) to achieve ϵ-optimality is $\mathcal{O}(S(\log S)^2(\log \epsilon)^3)$,

given by Theorem 2.24. Thus, the number of pivot operations for RHA has a lower order in ϵ by a factor of $\log \epsilon$ than that of SA($\delta(\epsilon/2)$). However, our experimental results in Section 2.6 show that the simplex algorithm takes fewer pivot operations to find a near-optimal solution than RHA does, as opposed to the theoretical guarantees.

We should point out that the amounts of computation per pivot operation of the two algorithms can be vastly different. In order to find a pivot operation that improves the current policy, the simplex algorithm approximately evaluates the current policy and increases the strategy horizon if such a pivot operation is not detected. As we cannot predict at what strategy horizon an improving pivot operation will be detected, we cannot exactly estimate the amount of computation per pivot operation for the simplex algorithm. On the other hand, RHA's pivot operation is computing (2.32) (which is often called a backup operation in literature) whose computational complexity is $\mathcal{O}(SA)$. However, a pivot operation performed by RHA may make its policy worse as demonstrated in our experiments in Section 2.6.

2.5 Multiple Pivoting in the Simplex Algorithm

The simplex algorithm in Section 2.2.3 (SA) performs only one improving pivot operation in each iteration. In this section, we introduce a modification to SA so that it performs possibly multiple improving pivot operations in one iteration, which greatly improves its empirical performance as illustrated in Section 2.6. The resulting algorithm also generates a sequence of policies that improves monotonically and converges to optimality, as the original SA does. We call it the *simplex algorithm with multiple pivoting*, in short, SAMP.

SAMP is obtained from SA by replacing Step 2(e) by the following:

The simplex algorithm with multiple pivoting (SAMP)

(only the replaced parts)

2. (e′) If $\gamma_{\min}^{k,m} < -\alpha^m c/(1-\alpha)$, set $m(k) = m$, compute

$$C_n^k \triangleq \left\{ (n,s,a) \;\middle|\; a = \arg\min_{a' \in A} \gamma_n^{k,m}(s,a'), \gamma_n^{k,m}(s,a) < -\alpha^m \frac{c}{1-\alpha} \right\} \text{ and}$$
$$C^k \triangleq \bigcup_{n=1}^{m} C_n^k,$$

set $a_n^{k+1}(s) = a$ for all $(n,s,a) \in C^k$ and $a_n^{k+1}(s) = a_n^k(s)$ for all (n,s) such that there is no a satisfying $(n,s,a) \in C^k$, and go to Step 3; else set $m := m+1$ and go to Step 2(b).

That is, C_n^k contains all hyperarcs (n,s,a) in period n such that (n,s,a) can be guaranteed to have a negative reduced cost (based on the inequality in the definition of C_n^k) and where a has the smallest approximate reduced cost among all actions at state s in period n. C^k contains all such hyperarcs in periods 1 to $m(k)$.

Note that in SAMP, the inequality that triggers pivot operations is still the same as in SA, but when the inequality is satisfied, SAMP performs all pivot operations in C^k. The next proposition extends Proposition 2.10 and gives an expression for the change in objective function value made by pivot operations in C^k.

Proposition 2.27. *Let x be the basic feasible solution in iteration k of SAMP. Then, the difference in objective function values of x and the new basic feasible solution z obtained by applying all pivot operations in C^k to x is given by*

$$f(z) - f(x) = \sum_{n=1}^{m(k)} \sum_{i=1}^{l_n} (1 + \theta^n(s_n^i)) \gamma_n(s_n^i, a_n^i) \tag{2.34}$$

where $C_n^k = \{(n,s_n^1,a_n^1),(n,s_n^2,a_n^2),\ldots,(n,s_n^{l_n},a_n^{l_n})\}$ for $n = 1,2,\ldots,m(k)$ and $\theta^n(s_n^i) \geq 0$ for $n = 1,2,\ldots,m(k)$ and $i = 1,2,\ldots,l_n$.

Proof: In Section 2.7.5.

SAMP and SA share the same scheme to approximate reduced costs (Step 2(b) and (c) of SA). Thus, Lemma 2.11, 2.13, and 2.15 hold for SAMP. Furthermore, using the above

proposition, we can easily prove that Lemma 2.16 and 2.17 also hold for SAMP. Therefore, we have the following theorem, convergence of SAMP, and the proof is omitted as it is nearly identical to the one of Theorem 2.20.

Theorem 2.28. *SAMP produces a sequence of basic feasible solutions x^k whose objective function values $f(x^k)$ converge to f^*, i.e., $\lim_{k\to\infty} f(x^k) = f^*$.*

2.6 Experimental Results

The objectives of this section are: (1) to compare empirical performance of the simplex algorithms introduced in this section, in particular, to evaluate how the multiple pivoting affects performance and (2) to compare the simplex algorithms and RHA empirically.

2.6.1 Comparison of the Simplex Algorithms

As a testing ground for the algorithms, we consider the following inventory management problem over discrete time periods with stochastic non-stationary demands, non-stationary costs, and lost sales. At the beginning of period $n = 1, 2, \ldots$, a decision maker observes the current integer-valued inventory level i_n and determines an integer-valued order quantity, denoted as a_n. A nonnegative integer-valued demand D_n is realized from a distribution function F_n, and then the inventory level at the beginning of the next period is $(i_n+a_n-D_n)^+$ and the unmet demand $(D_n - i_n - a_n)^+$ is lost. We assume no lead time for delivery. Given i_n, a_n, and D_n, cost incurred in period n is $p_n(a_n) + h_n(i_n + a_n) + s_n((D_n - i_n - a_n)^+)$ where p_n, h_n, s_n are functions for purchase, holding, and shortage cost in period n, respectively. We assume that items ordered in period n takes up inventory space in that period, so the holding cost in period n is $h_n(i_n + a_n)$. We also assume that there is a limit I on inventory level, i.e., $i_n + a_n \leq I$ for all n, that demands are uniformly bounded by a constant $D < I$, and that the marginal costs of purchase, holding, and shortage are uniformly bounded from below and above, i.e., $0 < \underline{p} \leq p_n(a) - p_n(a-1) \leq \bar{p} < \infty$ for all integers $a > 0$, $0 < \underline{h} \leq$

	Average CPU time (secs)	Average number of pivots
SA	1453.3	2290.9
SA(δ)	3010.61	2291.4
SAMP	280.52	2283.85

Table 2.1: Average performance of the simplex algorithms for $\epsilon = 0.01$

$h_n(i) - h_n(i-1) \leq \bar{h} < \infty$ for all integers $i > 0$, and $0 < \underline{s} \leq s_n(u) - s_n(u-1) \leq \bar{s} < \infty$ for all integers $u > 0$, for all $n = 1, 2, \ldots$.

For experiments, we used the following five parameter sets

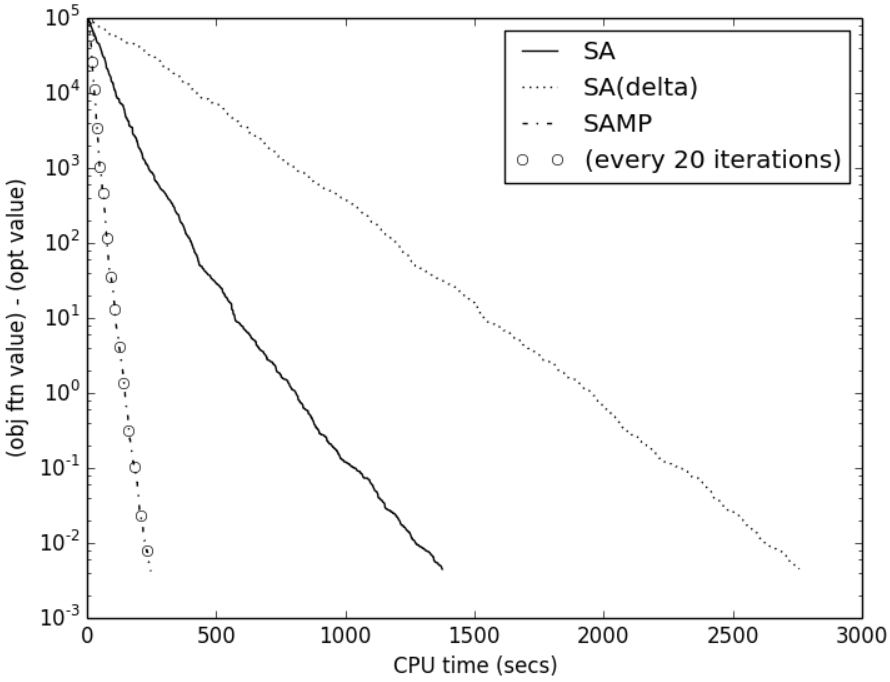
$$(D, I, \underline{p}, \bar{p}, \underline{h}, \bar{h}, \underline{s}, \bar{s}) = (10, 20, 100, 10, 3, 1, 150, 10),$$

$$(10, 20, 100, 10, 10, 5, 150, 10), (15, 20, 100, 10, 3, 1, 150, 10),$$

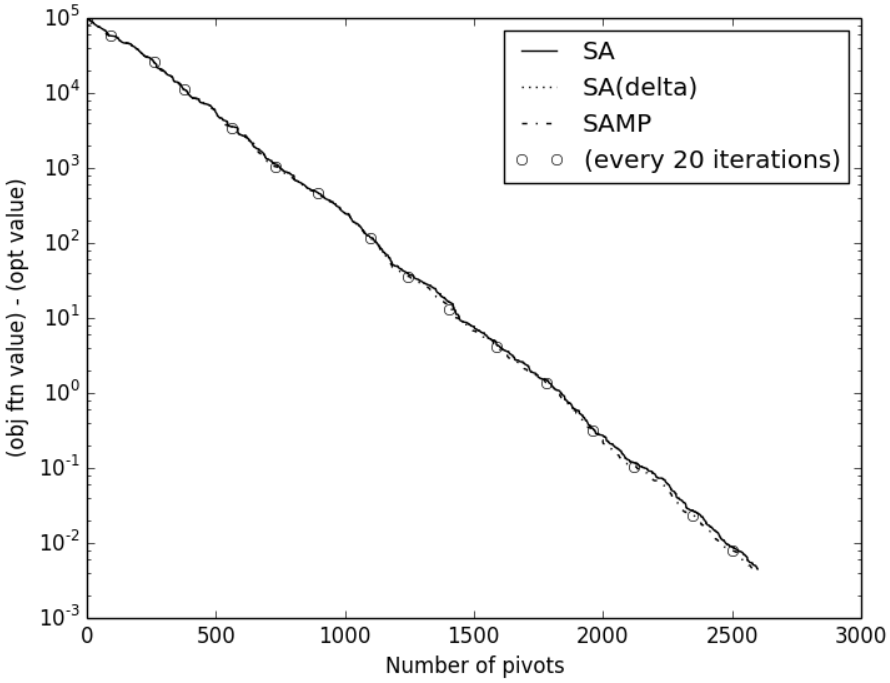
$$(10, 25, 100, 10, 3, 1, 150, 10), (10, 20, 100, 50, 3, 1, 150, 100),$$

and randomly generated four instances using each parameter set. Discount factor was 0.9. For the 20 instances, we measured CPU time and number of pivot operations for the three simplex algorithms to achieve 0.01-optimality ("a penny away" from the optimal cost). In order to compare the theoretical guarantee given by Theorem 2.24 and empirical performance, we used $\delta = \delta(0.005)$ for SA(δ). The objective function value of each policy generated by the algorithms was estimated using (2.31). The optimal objective function value was estimated by RHA. For inventory level i, the initial policy was ordering $(D - i)^+$, i.e., filling up to the maximum one-period demand. The algorithms were implemented in Python and ran on 2.93 GHz Intel Xeon CPU.

Table 2.1 shows average CPU time and average number of pivots for the three algorithms to achieve 0.01-optimality. Note that SAMP may perform multiple pivot operations in one iteration. We counted the pivot operations performed in one iteration separately, and evaluated all of the intermediate policies obtained by applying the pivot operations one by one. Figure 2.2 shows optimality gap progress of the simplex algorithms as a function of (a)



(a) For CPU time



(b) For number of pivots

Figure 2.2: Optimality gap progress of the simplex algorithms for inventory management problems

CPU time and (b) number of pivot operations, for one of the 20 instances. (The plots look very similar for the other instances.) The vertical axis of Figure 2.2 is optimality gap, the difference between the objective function value of policies obtained by the simplex algorithms and the optimal objective function value. The circles mark every 20 iterations of SAMP.

In Table 2.1, SAMP is at least five times faster than SA, thus, the multiple pivoting greatly accelerates the simplex algorithm. SAMP performed 7.59 pivot operations per iteration on average. On the other hand, SA(δ) is about two times slower than SA. SA(δ) uses larger strategy horizons than SA does, thus takes further future into account in making pivoting decisions, but it also means more computation. We can deduce that the additional computational load overwhelmed the gain from using more information for pivoting decisions.

The average numbers of pivot operations to achieve 0.01-optimality in Table 2.1 are nearly identical. Also, in Figure 2.2(b), we observe that the simplex algorithms improve at nearly identical paces as the number of pivots increases although their pivoting decisions are made quite differently. For the 20 instances we generated, the guaranteed numbers of iterations for SA($\delta(0.005)$) to achieve 0.01-optimality given by Theorem 2.24 are larger than a billion, as opposed to a few thousands in Table 2.1. Thus, for the inventory management problem, the theoretical guarantee given by Theorem 2.24 is a pessimistic bound.

For the sequence of solutions generated by SA($\delta(0.005)$), we compared the left and right hand sides of (2.25) in Lemma 2.23, which are the actual improvement rate and the guaranteed improvement rate, respectively. Figure 2.3 plots the actual improvement rate of solutions generated by SA($\delta(0.005)$) divided by the guaranteed improvement rate, for the same instance as the one of Figure 2.2. Thus, in Figure 2.3, dots close to zero indicate improvements bigger than the guarantee.

In Figure 2.3, we observe that although the actual improvement rates are mostly close to the guaranteed one (a lot of dots concentrated around one), there are occasional big improvements, which make the actual improvement a lot faster than the guarantee, as the

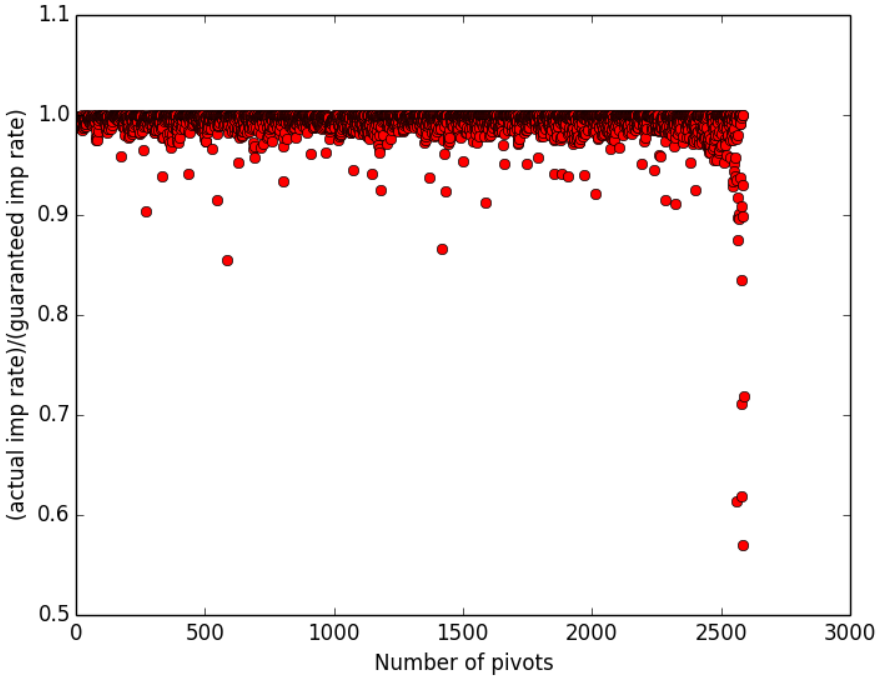


Figure 2.3: Comparison of guaranteed/actual improvement rates

	Avg. CPU time (secs)	Avg. number of pivots	Avg. number of iterations
SAMP	280.52	2283.85	306.60
RHA	67.90	5167.65	181.65

Table 2.2: Average performance of SAMP and RHA for $\epsilon = 0.01$

total improvement rate over multiple iterations is multiplicative.

2.6.2 Comparison of SAMP and RHA

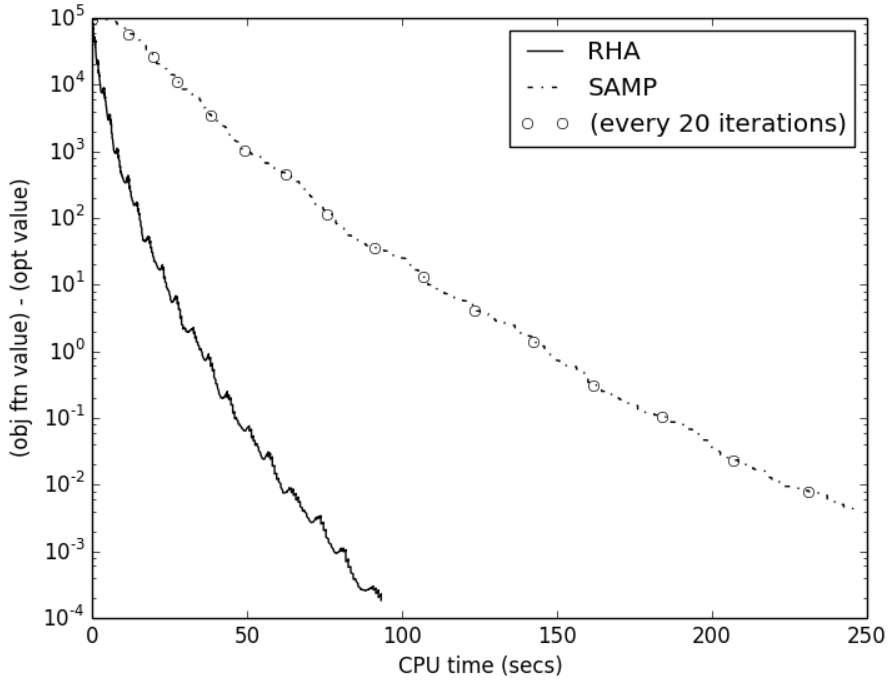
We empirically compared SAMP and RHA for the 20 instances of the inventory management problem (see Table 2.2). For RHA, we evaluated every policy obtained after each pivot operation. Recall that not all pivot operations performed by RHA improve the objective function. We define the CPU time of RHA to achieve 0.01-optimality as the time when the objective function value of a policy obtained by RHA enters the interval $[f^\star, f^\star + 0.01]$ and does not go out of the interval afterwards, and the second column of Table 2.2 shows its average. The other two metrics of RHA in the table were computed in similar manners.

Figure 2.4 shows optimality gap progress of SAMP and RHA as a function of (a) CPU time and (b) number of pivot operations, for the same instance as the one used for Figure 2.2 and Figure 2.3. For the other instances, the plots look similar.

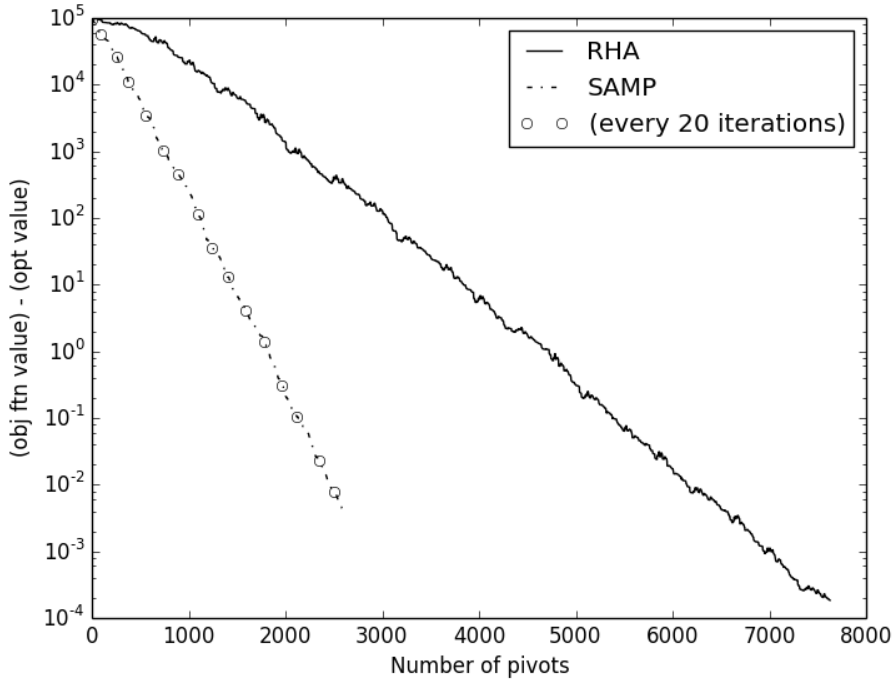
In Table 2.2 and in Figure 2.4 (a), we observe that RHA converges faster than SAMP. Note that in Figure 2.4 (a), RHA often goes up (i.e., performs a pivot operation that worsens its policy) while SAMP always goes down (i.e., always improves its policy). In Table 2.2 and Figure 2.4 (b), we observe that SAMP takes a smaller number of pivot operations than RHA to achieve the same level of near-optimality. Recall that SAMP approximately evaluates its current policy and enhances the approximation until an improving pivot operation is found. On the other hand, RHA performs only a backup operation per pivot operation, but the chosen pivot may not improve its policy. We can deduce that although SAMP performs "smarter" pivot operations than RHA does, for the inventory management problem, the additional computation of SAMP to find smarter pivot operations overwhelms its benefit.

Recall that RHA solves successively longer finite horizon truncations to optimality, and thus, it requires costs and transition probabilities of further future periods as it runs. SAMP also requires problem data of further future periods as the strategy horizon increases. Figure 2.5 shows improvement of the two algorithms as a function of number of periods for which problem data are requested. For the inventory management problem, SAMP required more data than RHA to achieve the same level of near-optimality.

In Table 2.2, RHA achieved 0.01-optimality after 181.65 iterations or 5167.65 pivot operations on average. The number of iterations and the number of pivot operations of RHA guaranteed to achieve 0.01-optimality given by Theorem 2.26 are in order of thousands and hundreds of millions, respectively. Thus, the theoretical guarantee of RHA given by Theorem 2.26 is also pessimistic.



(a) For CPU time



(b) For number of pivots

Figure 2.4: Optimality gap progress of RHA and SAMP for inventory management problems

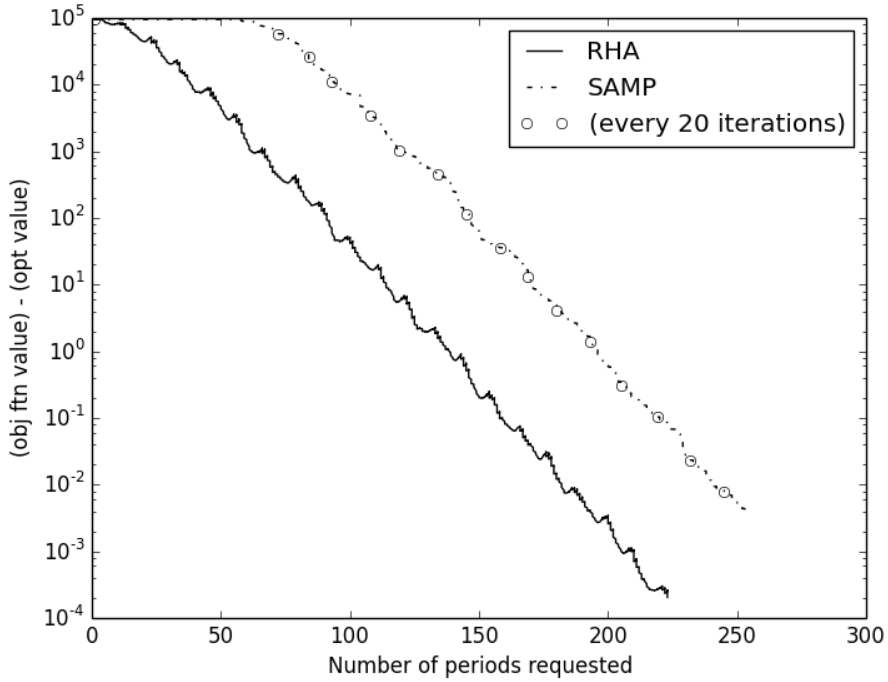


Figure 2.5: Comparison of data requirement of RHA and SAMP

2.7 Technical Proofs

2.7.1 Proof of Proposition 2.10

In the proof of Proposition 5.1 in [24], $\theta^n(s)$ is given as follows:

$$\theta^n(s) \triangleq \sum_{k=1}^{n-1} \sum_{t \in \mathcal{S}} \theta_k^n(s,t)$$

where

$$\theta_{n-1}^n(s,t) \triangleq p_{n-1}(s|t, a_{n-1}(t)) \text{ and}$$
$$\theta_k^n(s,t) \triangleq \sum_{s' \in \mathcal{S}} p_k(s'|t, a_k(t)) \theta_{k+1}^n(s,s') \text{ for } k = 1, 2, \ldots, n-2.$$

Interpretations of the above definitions are in order. $\theta_{n-1}^n(s,t)$ is the probability of transition from t to s in period $n-1$ under the policy given by x. One can easily verify that $\theta_k^n(s,t)$ is

the probability of going from period-state pair (k,t) to (n,s) under the policy x. Therefore, $\theta^n(s)$ is the sum of probabilities reaching (n,s) from (k,t) for $k = 1,\ldots,n-1$ and $t \in \mathcal{S}$.

We prove $1 + \theta^n(s) = x_n(s, a_n(s))$ for $s \in \mathcal{S}$ by induction on n. For $n = 1$ and $s \in \mathcal{S}$, $1 + \theta^1(s) = 1 = x_1(s, a_1(s))$. Suppose that for a positive integer k, $1 + \theta^k(s) = x_k(s, a_k(s))$ for $s \in \mathcal{S}$. Then for $s \in \mathcal{S}$,

$$\begin{aligned}
x_{k+1}(s, a_{k+1}(s)) &= 1 + \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s')) x_k(s', a_k(s')) \\
&= 1 + \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s'))(1 + \theta^k(s')) \\
&= 1 + \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s')) + \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s'))\theta^k(s') \\
&= 1 + \sum_{s' \in \mathcal{S}} \theta_k^{k+1}(s, s') + \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s')) \sum_{l=1}^{k-1} \sum_{t \in \mathcal{S}} \theta_l^k(s', t) \\
&= 1 + \sum_{s' \in \mathcal{S}} \theta_k^{k+1}(s, s') + \sum_{l=1}^{k-1} \sum_{t \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p_k(s|s', a_k(s')) \theta_l^k(s', t) \\
&= 1 + \sum_{s' \in \mathcal{S}} \theta_k^{k+1}(s, s') + \sum_{l=1}^{k-1} \sum_{t \in \mathcal{S}} \theta_l^{k+1}(s, t) \\
&= 1 + \sum_{l=1}^{k} \sum_{t \in \mathcal{S}} \theta_l^{k+1}(s, t) = 1 + \theta^{k+1}(s),
\end{aligned}$$

where the first equality is obtained from feasibility of x to (NP) and the second equality is obtained from the induction hypothesis. Thus, by induction, the lemma is proven. □

2.7.2 Proof of Lemma 2.18

Recall that $x^\star$ is a basic feasible solution and $a_n^\star(s)$ denotes its basic action at (n,s). For a positive integer N,

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}} x_n^\star(s,a_n^\star(s))(-\gamma_n(s,a_n^\star(s))) = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n^\star(s,a)(-\gamma_n(s,a)) \\
&\quad = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n^\star(s,a)\left[-\alpha^{n-1}c_n(s,a) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s') + y_n(s)\right] \\
&\quad = -\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n^\star(s,a) \\
&\qquad -\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left[\sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s') - y_n(s)\right] x_n^\star(s,a) \qquad (2.35)
\end{aligned}$$

where y is the complementary solution of x. We can simplify the second term of (2.35) as follows:

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left(\sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s') - y_n(s)\right) x_n^\star(s,a) \\
&\quad = \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_n(s'|s,a)x_n^\star(s,a) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n^\star(s,a) \\
&\quad = \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\left(\sum_{a\in\mathcal{A}} x_{n+1}^\star(s',a) - 1\right) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n^\star(s,a) \\
&\quad = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_{n+1}(s)\sum_{a\in\mathcal{A}} x_{n+1}^\star(s,a) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n^\star(s,a) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_{n+1}(s) \\
&\quad = \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}^\star(s,a) - \sum_{s\in\mathcal{S}} y_1(s)\sum_{a\in\mathcal{A}} x_1^\star(s,a) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_{n+1}(s) \\
&\quad = \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}^\star(s,a) - \sum_{n=1}^{N+1}\sum_{s\in\mathcal{S}} y_n(s)
\end{aligned}$$

where the second and the last equalities are obtained from feasibility of $x^\star$ to (NP). However, since $0 \le y_{N+1}(s) \le \alpha^N \frac{c}{1-\alpha}$ by (2.9) and $0 \le x^\star_{N+1}(s,a) \le (N+1)S$ by Lemma 2.1, we have

$$0 \le \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x^\star_{N+1}(s,a) \le S^2 A(N+1)\alpha^N \frac{c}{1-\alpha},$$

thus, $\sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x^\star_{N+1}(s,a) \to 0$ as $N \to \infty$. Therefore, taking $N \to \infty$ in (2.35) gives

$$\sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} x^\star_n(s, a^\star_n(s))(-\gamma_n(s, a^\star_n(s))) = -f^\star + \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} y_n(s) = f(x) - f^\star$$

where the last equality is obtained from the fact that y is complementary with x. Thus, the lemma is proven. □

2.7.3 Proof of Proposition 2.21

Consider a function $F(N) = \log(N + \frac{1}{1-\alpha})$. Since $F(N)$ is a concave function, we have $F(N(\epsilon)) \le F(0) + F'(0)N(\epsilon)$ and that is,

$$\log(N(\epsilon) + \frac{1}{1-\alpha}) \le (1-\alpha)N(\epsilon) - \log(1-\alpha).$$

Thus, we have

$$\begin{aligned} N(\epsilon)\log\alpha + \log(N(\epsilon) + \frac{1}{1-\alpha}) &\le N(\epsilon)\log\alpha + (1-\alpha)N(\epsilon) - \log(1-\alpha) \\ &= (\log\alpha + 1 - \alpha)N(\epsilon) - \log(1-\alpha). \end{aligned}$$

Since $0 < \alpha < 1$, we have $\log\alpha + 1 - \alpha < 0$. By the definition of $N(\epsilon)$ in (2.21),

$$(\log\alpha + 1 - \alpha)N(\epsilon) - \log(1-\alpha) \le \log(\epsilon(1-\alpha)^2/2cS) - \log(1-\alpha) = \log(\epsilon(1-\alpha)^2/2cS).$$

Thus,

$$N(\epsilon)\log\alpha + \log(N(\epsilon) + \frac{1}{1-\alpha}) \le \log(\epsilon(1-\alpha)^2/2cS)$$

. By taking exponential function on both sides, we obtain the result. □

2.7.4 Proof of Theorem 2.26

Let π be the policy corresponding to $z^{N'(\epsilon)}$, i.e., the policy obtained after solving the $N'(\epsilon)$-horizon truncated problem. Note that π restricted to $N'(\epsilon)$-horizon is optimal for the $N'(\epsilon)$-horizon truncated problem. We first show that π is ϵ-optimal for the original problem. Let x denote the basic feasible solution of π and y denote its complementary solution. Let w denote the optimal solution to (ND). Then

$$
\begin{aligned}
0 \le f(x) - f^\star = g(y) - g(w) &= \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} y_n(s) - \sum_{n=1}^{\infty}\sum_{s\in\mathcal{S}} w_n(s) \\
&= \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}(y_n(s) - w_n(s)) + \sum_{n=N'(\epsilon)+1}^{\infty}\sum_{s\in\mathcal{S}}(y_n(s) - w_n(s)) \\
&= \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}(y_n^{N'(\epsilon)}(s) - w_n^{N'(\epsilon)}(s)) + \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}(y_n(s) - y_n^{N'(\epsilon)}(s)) \\
&\quad + \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}(w_n^{N'(\epsilon)}(s) - w_n(s)) + \sum_{n=N'(\epsilon)+1}^{\infty}\sum_{s\in\mathcal{S}}(y_n(s) - w_n(s)) \\
&\le \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}(y_n(s) - y_n^{N'(\epsilon)}(s)) + \sum_{n=N'(\epsilon)+1}^{\infty}\sum_{s\in\mathcal{S}}(y_n(s) - w_n(s)) \\
&\le \sum_{n=1}^{N'(\epsilon)}\sum_{s\in\mathcal{S}}\alpha^{N'(\epsilon)}\frac{c}{1-\alpha} + \sum_{n=N'(\epsilon)+1}^{\infty}\sum_{s\in\mathcal{S}}\alpha^{n-1}\frac{c}{1-\alpha} \\
&= \frac{cS\alpha^{N'(\epsilon)}}{1-\alpha}N'(\epsilon) + \frac{cS\alpha^{N'(\epsilon)}}{(1-\alpha)^2} = \frac{cS\alpha^{N'(\epsilon)}}{1-\alpha}\left(N'(\epsilon) + \frac{1}{1-\alpha}\right) \le \epsilon
\end{aligned}
$$

where $y^{N'(\epsilon)}$ and $w^{N'(\epsilon)}$ denote the $N'(\epsilon)$-horizon approximations of y and w (computed in the same way as (2.13) and (2.14)), respectively; the second inequality is obtained from the fact that π is optimal for the $N'(\epsilon)$-horizon truncated problem (i.e., $y^{N'(\epsilon)} \le w^{N'(\epsilon)}$) and Lemma 2.11; the third inequality is from Lemma 2.11 and (2.9); and the last inequality is obtained by using arguments similar to the proof of Proposition 2.21. Therefore, π is ϵ-optimal.

We conclude the proof by computing the number of pivot operations to obtain π. RHA solves $1, 2, \ldots, N'(\epsilon)$-horizon problems by backward induction until π is obtained. For a positive integer m, backward induction solving an m-horizon problem performs at most mS pivot operations. Thus, the ϵ-optimal policy π is obtained after at most $N'(\epsilon)(N'(\epsilon)+1)S/2$ pivot operations. $\square$

2.7.5 Proof of Proposition 2.27

Since the iteration counter k is fixed in the proposition, we omit it in the notation C_n^k and C^k and denote them as C_n and C, respectively. We first prove the following proposition, which is a simpler version of Proposition 2.27.

Proposition 2.29. *For a fixed $n \leq m(k)$, the difference in the objective function values at a basic feasible solution x and the new basic feasible solution z obtained by applying the pivot operations in C_n to x is given by*

$$f(z) - f(x) = \sum_{i=1}^{l_n} (1 + \theta^n(s_n^i))\gamma_n(s_n^i, a_n^i)$$

where $\theta^n(s_n^i) \geq 0$ for $i = 1, 2, \ldots, l_n$.

Proof: Let y and w be the complementary solutions of x and z, respectively. Let $\mathcal{S}_n = \{s_n^1, s_n^2, \ldots, s_n^{l_n}\}$. Since basic actions of periods $n+1, n+2, \ldots$ are not changed, $y_j(s) = w_j(s)$ for all $s \in \mathcal{S}$ and for $j = n+1, n+2, \ldots$. Since basic actions of states $s \notin \mathcal{S}_n$ in period n are not changed, $y_n(s) = w_n(s)$ for $s \notin \mathcal{S}_n$. By complementary slackness, for $i = 1, 2, \ldots, l_n$

$$w_n(s_n^i) = \alpha^{n-1} c_n(s_n^i, a_n^i) + \sum_{s' \in \mathcal{S}} p_n(s'|s_n^i, a_n^i) y_{n+1}(s'). \tag{2.36}$$

Let $\mathcal{S}_{n-1,i} \subset \mathcal{S}$ be the set of states t in period $n-1$ such that $s_n^i \in \mathcal{J}_{n-1}(t, a_{n-1}(t))$ where $\mathcal{J}_{n-1}(t, a_{n-1}(t))$ is the set of states that are reachable by choosing action $a_{n-1}(t)$ (the basic action of x at state t in period $n-1$) at state t in period $n-1$. Let $\mathcal{S}_{n-1} = \cup_{i=1}^{l_n} \mathcal{S}_{n-1,i}$. For

$t \in \mathcal{S}_{n-1}$, by complementary slackness,

$$\begin{aligned} w_{n-1}(t) &= \alpha^{n-2} c_{n-1}(t, a_{n-1}(t)) + \sum_{i=1}^{l_n} p_{n-1}(s_i|t, a_{n-1}(t)) w_n(s_n^i) \\ &\quad + \sum_{s' \in \mathcal{S} \setminus \mathcal{S}_n} p_{n-1}(s'|t, a_{n-1}(t)) y_n(s') \\ y_{n-1}(t) &= \alpha^{n-2} c_{n-1}(t, a_{n-1}(t)) + \sum_{i=1}^{l_n} p_{n-1}(s_i|t, a_{n-1}(t)) y_n(s_n^i) \\ &\quad + \sum_{s' \in \mathcal{S} \setminus \mathcal{S}_n} p_{n-1}(s'|t, a_{n-1}(t)) y_n(s'). \end{aligned}$$

Thus, for $t \in \mathcal{S}_{n-1}$,

$$\begin{aligned} w_{n-1}(t) - y_{n-1}(t) &= \sum_{i=1}^{l_n} p_{n-1}(s_n^i|t, a_{n-1}(t))(w_n(s_n^i) - y_n(s_n^i)) \\ &= \sum_{i=1}^{l_n} \theta_{n-1}^n(s_n^i, t)(w_n(s_n^i) - y_n(s_n^i)) \end{aligned}$$

(see Section 2.7.1 for definition of $\theta_{n-1}^n(s_n^i, t)$). For $t \notin \mathcal{S}_{n-1}$, $w_{n-1}(t) = y_{n-1}(t)$.

Now for $j = 1, 2, \ldots, n-2$, we recursively define $\mathcal{S}_j \subset \mathcal{S}$ as the set of states t in period j such that $\mathcal{I}_j(t, a_j(t)) \triangleq \mathcal{J}_j(t, a_j(t)) \cap \mathcal{S}_{j+1} \neq \emptyset$. Then for $t \in \mathcal{S}_j$, we have

$$\begin{aligned} w_j(t) &= \alpha^{j-1} c_j(t, a_j(t)) + \sum_{s' \in \mathcal{I}_j(t, a_j(t))} p_j(s'|t, a_j(t)) w_{j+1}(s') + \sum_{s' \notin \mathcal{I}_j(t, a_j(t))} p_j(s'|t, a_j(t)) y_{j+1}(s'), \\ y_j(t) &= \alpha^{j-1} c_j(t, a_j(t)) + \sum_{s' \in \mathcal{I}_j(t, a_j(t))} p_j(s'|t, a_j(t)) y_{j+1}(s') + \sum_{s' \notin \mathcal{I}_j(t, a_j(t))} p_j(s'|t, a_j(t)) y_{j+1}(s'). \end{aligned}$$

Thus,

$$\begin{aligned}
w_j(t) - y_j(t) &= \sum_{s' \in \mathcal{I}_j(t,a_j(t))} p_j(s'|t, a_j(t))(w_{j+1}(s') - y_{j+1}(s')) \\
&= \sum_{s' \in \mathcal{I}_j(t,a_j(t))} p_j(s'|t, a_j(t)) \sum_{i=1}^{l_n} \theta_{j+1}^n(s_n^i, s')(w_n(s_n^i) - y_n(s_n^i)) \\
&= \sum_{i=1}^{l_n} (w_n(s_n^i) - y_n(s_n^i)) \sum_{s' \in \mathcal{I}_j(t,a_j(t))} p_j(s'|t, a_j(t)) \theta_{j+1}^n(s_n^i, s') \\
&= \sum_{i=1}^{l_n} \theta_j^n(s_n^i, t)(w_n(s_n^i) - y_n(s_n^i)).
\end{aligned}$$

For $t \notin \mathcal{S}_j$, $w_j(t) = y_j(t)$. Therefore, the difference between objective function values of w and y is

$$\begin{aligned}
g(w) - g(y) &= \sum_{j \in \mathbb{N}} \sum_{t \in \mathcal{S}} (w_j(t) - y_j(t)) = \sum_{j=1}^{n} \sum_{t \in \mathcal{S}} (w_j(t) - y_j(t)) \\
&= \sum_{i=1}^{l_n} (w_n(s_n^i) - y_n(s_n^i)) + \sum_{j=1}^{n-1} \sum_{t \in \mathcal{S}_j} (w_j(t) - y_j(t)) \\
&= \sum_{i=1}^{l_n} (w_n(s_n^i) - y_n(s_n^i)) + \sum_{j=1}^{n-1} \sum_{t \in \mathcal{S}_j} \sum_{i=1}^{l_n} \theta_j^n(s_n^i, t)(w_n(s_n^i) - y_n(s_n^i)) \\
&= \sum_{i=1}^{l_n} \left[1 + \sum_{j=1}^{n-1} \sum_{t \in \mathcal{S}_j} \theta_j^n(s_n^i, t) \right] (w_n(s_n^i) - y_n(s_n^i)) \\
&\triangleq \sum_{i=1}^{l_n} (1 + \theta^n(s_n^i))(w_n(s_n^i) - y_n(s_n^i)),
\end{aligned}$$

and (2.36) gives $w_n(s_n^i) - y_n(s_n^i) = \gamma_n(s_n^i, a_n^i)$, thus the proposition is proven. □

Now we prove Proposition 2.27. Let $z^0 = x$. Inductively, for $n = 1, 2, \ldots, m(k)$, let z^n be the basic feasible solution obtained by applying the pivot operations in C_n to z^{n-1}. By definition, $z^{m(k)} = z$.

Fix $n \leq m(k)$. For $n' \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$, let $\tilde{\gamma}_{n'}(s, a)$ be the reduced cost of variable

$x_{n'}(s,a)$ computed at the basic feasible solution z^{n-1}. Then, by Proposition 2.29, we have

$$f(z^n) - f(z^{n-1}) = \sum_{i=1}^{l_n} (1 + \theta^n(s_n^i))\tilde{\gamma}_n(s_n^i, a_n^i)$$

where $\theta^n(s_n^i) \geq 0$ for $i = 1, 2, \ldots, l_n$. The basic feasible solution z^{n-1} is obtained by applying the pivot operations of $C_1, C_2, \ldots, C_{n-1}$ to $z^0 = x$. The sets $C_1, C_2, \ldots, C_{n-1}$ contain nonbasic variables in periods $1, 2, \ldots, n-1$. Thus, the basic actions of z^{n-1} in periods $n, n+1, \ldots$ coincide with those of x. This implies that $\tilde{\gamma}_{n'}(s,a)$ equals $\gamma_{n'}(s,a)$, the reduced cost of the same nonbasic variable computed at x, for $n' \geq n, s \in \mathcal{S}, a \in \mathcal{A}$. Therefore, we have

$$f(z^n) - f(z^{n-1}) = \sum_{i=1}^{l_n} (1 + \theta^n(s_n^i))\gamma_n(s_n^i, a_n^i)$$

and by adding this equality for $n = 1, 2, \ldots, m(k)$, the proposition is proven. □

CHAPTER III

Simplex Algorithm for Countable-state Markov Decision Processes

3.1 Introduction

In this chapter, we introduce a simplex-type algorithm for solving countable-state MDPs. Consider MDPs that have a countable state space, a finite action space, and stationary transition probabilities and reward function, but whose objective is to *maximize* expected total discounted reward.[1] Specifically, the set of states $\mathcal{S}$ is countably-infinite, and given that action a is taken in state s, reward $r(s,a,t)$ is obtained with probability $p(t|s,a)$. Let $r(s,a)$ denote the expected reward incurred by choosing action a at state s, i.e., $r(s,a) = \sum_{t\in\mathcal{S}} p(t|s,a)r(s,a,t)$. Note that the transition probabilities and the rewards do not depend on period index, i.e., are stationary. Countable-state MDPs arise in inventory management and queueing control where there is no specific limit on the size of inventory or queue, as we illustrate in Section 3.1.3. We let $\mathcal{S} = \{1,2,\ldots\}$ and $\mathcal{A} = \{1,2,\ldots,A\}$ unless otherwise specified.

We define the value function of a policy for an arbitrary initial state distribution, extending (1.1) in Section 1.1. Given an initial state distribution β, each policy π induces

[1] This chapter studies countable-state MDPs under assumptions given in [39] (Section 6.10) which studies maximization problems. For consistency with the reference, we also study countable-state MDPs maximizing expected total discounted reward. Since we allow rewards to be negative in this chapter, all results in this chapter can be easily converted to minimization problems.

a probability distribution P_π^β and defines the state process $\{S_n\}_{n=1}^\infty$ and the action process $\{A_n\}_{n=1}^\infty$. We denote by E_π^β the corresponding expectation operator. The expected total discounted reward of a policy π with initial state distribution β is defined as

$$J_\pi(\beta) \triangleq E_\pi^\beta \left[\sum_{n=1}^{\infty} \alpha^{n-1} r(S_n, A_n) \right]. \tag{3.1}$$

We call $J_\pi(\beta)$ *the value of policy* π *with initial state distribution* β, or simply *the value of policy* π whenever it is clear which initial state distribution is used. Recall that $J_\pi(s)$ denotes (3.1) for those initial state distributions concentrated on one state s.

A policy $\pi^\star$ is said to be *optimal for initial state distribution* β if $J_{\pi^\star}(\beta) = J^\star(\beta) \triangleq \sup_{\pi \in \Pi} J_\pi(\beta)$. A policy $\pi^\star$ is said to be *optimal for initial state* s if $J_{\pi^\star}(s) = J^\star(s) = \sup_{\pi \in \Pi} J_\pi(s)$. Recall $J^\star : \mathcal{S} \to \mathbb{R}$ is called the optimal value function. A policy $\pi^\star$ is optimal if $J_{\pi^\star}(s) = J^\star(s)$ for all $s \in \mathcal{S}$. The goal of the decision maker is to find an optimal policy.

The main contribution of this chapter is that we introduce a simplex-type algorithm for solving a CILP formulation of countable-state MDPs. It is the first solution algorithm for countable-state MDPs that finds a sequence of policies whose value functions not only converge to the optimal value function but also improve in every iteration. Countable-state MDPs were studied by many researchers, including [13, 28, 39, 56, 57, 58], with predominant solution methods summarized as the three algorithms in [56, 57] and [58]. In the next section, we review the three existing solution methods for countable-state MDPs as discussed in [56, 57, 58].

3.1.1 Literature Review

The algorithm suggested in [58] is an extension of value iteration to countable-state MDPs. In general, value iteration computes a sequence of real-valued functions on $\mathcal{S}$ that converges to the optimal value function. To remind the readers, value iteration for finite-state MDPs starts with a function $J^0 : \mathcal{S} \to \mathbb{R}$ and for $k = 1, 2, \ldots$, computes a function

$J^k : \mathcal{S} \to \mathbb{R}$ in iteration k by the following recursion formula:

$$J^k(s) \triangleq \max_{a \in \mathcal{A}} \left\{ r(s,a) + \alpha \sum_{t \in \mathcal{S}} p(t|s,a) J^{k-1}(t) \right\} \text{ for } s \in \mathcal{S}. \tag{3.2}$$

Extending this to countable-state MDPs requires an adjustment in order for each iteration to finish in finite time. The value iteration for countable-state MDPs first selects a function $u : \mathcal{S} \to \mathbb{R}$ and then, in iteration k, computes $J^k(s)$ by (3.2) only for $s \leq k$ and lets $J^k(s) = u(s)$ for $s > k$. In [58], it was shown that J^k converges pointwise to the optimal value function $J^\star$ and error bounds on the approximations were provided. In iteration k, a policy $\pi^k : \mathcal{S} \to \mathcal{A}$ (i.e., a stationary and deterministic policy) is obtained by assigning the action that achieves the maximum in (3.2) to $s \leq k$ and an arbitrary action to $s > k$. It was also shown in [58] that J_{π^k} converges pointwise to $J^\star$ but the convergence may not be monotone.

The solution method in [56] is an extension of policy iteration, another popular solution method for finite-state MDPs. Recall that, given $\pi^0 : \mathcal{S} \to \mathcal{A}$, the kth iteration of policy iteration for finite-state MDPs is as follows:

1. Obtain $J^k : \mathcal{S} \to \mathbb{R}$ that satisfies

$$J^k(s) = r(s, \pi^{k-1}(s)) + \alpha \sum_{t \in \mathcal{S}} p(t|s, \pi^{k-1}(s)) J^k(t) \text{ for } s \in \mathcal{S}; \tag{3.3}$$

2. Choose $\pi^k : \mathcal{S} \to \mathcal{A}$ that satisfies

$$\pi^k(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \alpha \sum_{t \in \mathcal{S}} p(t|s,a) J^k(t) \right\} \text{ for } s \in \mathcal{S}. \tag{3.4}$$

Each iteration consists of computing the value function of the current (stationary and deterministic) policy (Step 1) and obtaining a new policy based on the evaluation (Step 2). The extension of policy iteration to countable state space is as follows: after selecting a function

$u : \mathcal{S} \to \mathbb{R}$, in iteration k, it computes J^k that satisfies (3.3) for $s \leq k$ and $J^k(s) = u(s)$ for $s > k$, and then finds π^k that satisfies (3.4) only for $s \leq k$. It was shown in [56] that J^k obtained by this method also converges pointwise to $J^\star$ and error bounds on the approximations were provided. One can extend π^k to the entire state space $\mathcal{S}$ by assigning an arbitrary action to $s > k$; then J_{π^k} converges pointwise to $J^\star$ but again, the convergence may not be monotone.

Another method, proposed in [57], is to solve successively larger but finite-state approximations of the original MDP to optimality. The real-valued functions on $\mathcal{S}$ obtained by this method were also proven to converge pointwise to $J^\star$. A sequence of policies covering $\mathcal{S}$ is also obtained by this algorithm in a similar manner but pointwise convergence of their value functions was not established in the paper.

It should be pointed out that the above three papers only considered the case where the reward function is uniformly bounded. However, in the aforementioned applications of countable-state MDPs, immediate reward typically goes to infinity as the inventory level or the number of customers in queue goes to infinity, which suggests the need to consider countable-state MDPs with unbounded immediate reward functions. For brevity, let us refer to a set of assumptions on transition probabilities and rewards as a *setting* in the following literature review. Under three different settings with unbounded rewards, [27, 35, 55] studied properties of countable-state MDPs. In [59], each of the three settings with unbounded rewards in [27, 35, 55] was equivalently transformed into a bounded one. Therefore, the algorithms and results mentioned in previous paragraphs for bounded case were extended to the three unbounded problems in [27, 35, 55]. Meanwhile, [13] extensively reviewed conditions under which the extension of the value iteration in [58] converges to optimality in value and studied its rate of convergence. The setting in [13] is more general than the settings in [27, 35, 55] but one cannot check whether it holds for given transition probabilities and rewards without solving the MDP since it includes an assumption on the optimal value function. In this chapter, we consider the setting in [39] (Section 6.10) for countable-state

MDPs with unbounded rewards (Assumptions A1, A2, and A3 in Section 3.1.2). One can easily show that this setting covers the three settings in [27, 35, 55]; it is a special case of the one in [13] but it is checkable for given parameters without solving the MDP and has enough generality to cover many applications of interest.

3.1.2 Assumptions

Let us define additional notation that will come in handy in the rest of the chapter. Given a policy π and states $s, t \in \mathcal{S}$, $P_\pi^n(t|s)$ denotes the probability of reaching state t after n transitions starting from state s when policy π is applied, with $P_\pi^0(t|s) \triangleq \mathbf{1}\{t = s\}$. P_π^n denotes the transition probability matrix of policy π for n transitions with both rows and columns indexed by states. P_π^0, defined similarly, is denoted as I. For simplicity, we denote $P_\pi^1(t|s)$ and P_π^1 as $P_\pi(t|s)$ and P_π, respectively. For a stationary policy σ (recall that notation σ is used to emphasize the choice of a stationary policy) and a state $s \in \mathcal{S}$, $r_\sigma(s)$ denotes the expected immediate reward at s when σ is applied, and r_σ denotes the reward vector indexed by states. For a stationary and deterministic policy σ and a state s, $\sigma(s)$ denotes the action chosen by π at s. Throughout this chapter, we will make the following assumptions, which enable us to analyze countable-state MDPs with unbounded rewards:

Assumption (cf. Assumptions 6.10.1 and 6.10.2 of [39]) *There exists a positive real-valued function w on $\mathcal{S}$ satisfying the following:*

A1 $|r(s,a)| \leq w(s)$ *for all* $a \in \mathcal{A}$ *and* $s \in \mathcal{S}$*;*

A2 *There exists* κ, $0 \leq \kappa < \infty$, *for which*

$$\sum_{t=1}^{\infty} p(t|s,a)w(t) \leq \kappa w(s)$$

for all $a \in \mathcal{A}$ *and* $s \in \mathcal{S}$*;*

A3 *There exists* λ, $0 \leq \lambda < 1$, *and a positive integer* J *such that*

$$\alpha^J \sum_{t=1}^{\infty} P_\pi^J(t|s) w(t) \leq \lambda w(s)$$

for all $\pi \in \Pi_{MD}$.

Using the infinite matrix and infinite vector notation, the above three assumptions can be written as: **A1** $|r_\sigma| \leq w$ for all $\sigma \in \Pi_{SD}$, **A2** $P_\sigma w \leq \kappa w$ for all $\sigma \in \Pi_{SD}$, and **A3** $\alpha^J P_\pi^J w \leq \lambda w$ for all $\pi \in \Pi_{MD}$, where the inequalities are component-wise. We can easily show that the above assumptions imply that $|r_\sigma| \leq w$ and $P_\sigma w \leq \kappa w$ for all $\sigma \in \Pi_S$, and $\alpha^J P_\pi^J w \leq \lambda w$ for all $\pi \in \Pi_M$, i.e., they also hold for the corresponding class of randomized policies.

Assumption 1 tells us that the absolute value of the reward function is bounded by the function w. In other words, the function w provides a "scale" of reward that can be obtained in each state. Assumption 2 can be interpreted as that the transition probabilities prevent the expected scale of immediate reward after one transition from being larger than the scale in the current state (multiplied by κ). Assumption 3 can be interpreted similarly, but for J transitions. However, note that λ is strictly less than one, which is important because λ will play a role similar to that of the discount factor α in our following analysis.

3.1.3 Examples

We give two examples of countable-state MDPs with unbounded costs that satisfy Assumptions A1, A2, and A3.

Example 3.1 (Example 6.10.2 in [39])**.** Consider an infinite-horizon inventory management problem with a single product and unlimited inventory capacity where the objective is to

maximize the expected total discounted profit. Let $\mathcal{S} = \{0, 1, \ldots\}$, $\mathcal{A} = \{0, 1, \ldots, M\}$, and

$$p(t|s,a) = \begin{cases} 0 & t > s + a \\ p_{s+a-t} & s + a \geq t > 0 \\ q_{s+a} & t = 0, \end{cases}$$

where p_k denotes the probability of demand of k units in any period, and $q_k = \sum_{j=k}^{\infty} p_j$ denotes the probability of demand of at least k units in any period. For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the reward $r(s,a)$ is given as

$$r(s,a) = F(s+a) - O(a) - h \cdot (s+a),$$

where

$$F(s+a) = \sum_{j=0}^{s+a-1} bjp_j + b(s+a)q_{s+a},$$

with $b > 0$ representing the per-unit price, $O(a) = K + ca$ for $a > 0$ and $O(0) = 0$ representing the ordering cost, and $h > 0$ representing the cost of storing one unit of product for one period. It is reasonable to assume $\sum_{k=0}^{\infty} kp_k < \infty$, i.e., the expected demand is finite. Then,

$$|r(s,a)| \leq b(s+M) + K + cM + h(s+M) = K + M(b+c+h) + (b+h)s \triangleq C + Ds$$

by letting $C \triangleq K + M(b+c+h)$ and $D \triangleq b+h$. Let $w(s) \triangleq C + Ds$ so that A1 holds. Since

$$\begin{aligned} \sum_{t=0}^{\infty} p(t|s,a)w(t) &= \sum_{t=1}^{s+a} p_{s+a-t} \cdot w(t) + q_{s+a} \cdot w(0) = \sum_{t=0}^{s+a-1} p_t \cdot w(s+a-t) + q_{s+a} \cdot w(0) \\ &= C + D \sum_{t=0}^{s+a-1} (s+a-t)p_t \leq C + D(s+a) \leq w(s) + DM, \end{aligned}$$

by Proposition 6.10.5(a) in [39], A2 and A3 are also satisfied.

Example 3.2 (Generalized flow and service control)**.** This example is a generalization of the

flow and service rate control problem in [4]. Consider a discrete-time single-server queue with an infinite buffer. State is defined as the number of customers in the queue at the beginning of a period, so $\mathcal{S} = \{0, 1, \ldots\}$. Let $\mathcal{A}^1$ and $\mathcal{A}^2$ be finite sets of nonnegative numbers and let $\mathcal{A} = \mathcal{A}^1 \times \mathcal{A}^2$. If the decision-maker chooses $(a^1, a^2) \in \mathcal{A}^1 \times \mathcal{A}^2$ in a period, then the number of arrivals in the period is a Poisson random variable with mean a^1 and the number of served (thus, leaving) customers in the period is the minimum of a Poisson random variable with mean a^2 and the number of customers in the system at the beginning of the period plus the number of arrivals in the period. (That is, we assume that order of the events in a period is: the decision-maker observes the current state and chooses two numbers $a^1 \in \mathcal{A}^1$ and $a^2 \in \mathcal{A}^2$, arrivals occur, and then services are provided and served customers leave.) For $s \in \mathcal{S}, a = (a^1, a^2) \in \mathcal{A}$, the immediate reward is

$$r(s, a) = -cs - d^1(a^1) - d^2(a^2),$$

where c is a positive constant, $d^1(\cdot)$ is the flow control cost function, and $d^2(\cdot)$ is the service control cost function. The reward is linear in s, which is justified by the well-known Little's Law.

In the flow and service control problem in [4], it was assumed that in a period, at most one customer arrives and at most one customer leaves the system, which no longer holds in this example.

Let $C \triangleq c$ and $D \triangleq \max_{a^1 \in \mathcal{A}^1} |d^1(a^1)| + \max_{a^2 \in \mathcal{A}^2} |d^2(a^2)|$. Then A1 is satisfied with $w(s) \triangleq Cs + D$. In addition,

$$\begin{aligned}
\sum_{t \in \mathcal{S}} p(t|s,a) w(t) &= D + C \sum_{t=0}^{\infty} p(t|s,a) t \\
&= D + C \left[\sum_{t=0}^{s-1} p(t|s,a) t + \sum_{u=0}^{\infty} p(s+u|s,a)(s+u) \right] \\
&\leq D + Cs + \sum_{u=0}^{\infty} p(s+u|s,a) u \leq w(s) + \sum_{u=0}^{\infty} \frac{e^{-a^1_{\max}} (a^1_{\max})^u}{u!} u = w(s) + C a^1_{\max},
\end{aligned}$$

where the second inequality is obtained by considering maximum arrival rate $a^1_{\max} = \max \mathcal{A}^1$ and zero service rate. Therefore, by Proposition 6.10.5(a) in [39], A2 and A3 are satisfied.

Parts (b) and (c) of Proposition 6.10.5 in [39] provide two other sufficient conditions to satisfy A2 and A3.

3.1.4 Background

In this section we review some technical preliminaries that were established in the literature and will be used in this chapter.

The following theorem is an extension of Theorem 1.1 to countable-state MDPs. By this theorem, we can limit our attention to policies that are stationary and deterministic.

Theorem 3.3 (cf. Theorem 6.10.4 of [39])**.** *Countable-state MDPs under Assumptions A1, A2, and A3 satisfy the following.*

(1) There exists an optimal policy that is stationary and deterministic.

(2) The optimal value function $J^\star$ is the unique solution of

$$y(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a) y(t) \right\} \text{ for } s \in \mathcal{S}.$$

Moreover, the actions that achieve the above maximum form a stationary and deterministic optimal policy.

In particular, for any stationary deterministic policy σ, J_σ equals the optimal value function of a new MDP obtained by allowing only one action $\sigma(s)$ for $s \in \mathcal{S}$, and thus, J_σ is the unique solution of

$$y(s) = r(s,\sigma(s)) + \alpha \sum_{t=1}^{\infty} p(t|s,\sigma(s)) y(t) \text{ for } s \in \mathcal{S},$$

or $y = r_\sigma + \alpha P_\sigma y$ in the infinite vector and matrix notation.

Define

$$L \triangleq \begin{cases} \frac{J}{1-\lambda} & \text{if } \alpha\kappa = 1 \\ \frac{1}{1-\lambda}\frac{1-(\alpha\kappa)^J}{1-(\alpha\kappa)} & \text{otherwise.} \end{cases}$$

It has been shown that the value function of any Markov policy is bounded by Lw:

Proposition 3.4 (cf. Proposition 6.10.1 of [39]). *If Assumptions A1, A2, and A3 are satisfied,*

$$|J_\pi(s)| \leq Lw(s) \text{ for any } s \in \mathcal{S} \text{ and } \pi \in \Pi_M. \tag{3.5}$$

In the rest of this subsection, we review some real analysis results that will be used in this chapter for exchanging two infinite sums, an infinite sum and an expectation, or a limit and an expectation.

Proposition 3.5 (cf. Tonelli's theorem on page 309 of [46]). *Given a double sequence $\{a_{ij}\}$ for $i = 1, 2, \ldots$, $j = 1, 2, \ldots$, if $a_{ij} \geq 0$ for all i and j, then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}.$$

Proposition 3.6 (Theorem 8.3 in [47]). *Given a double sequence $\{a_{ij}\}$ for $i = 1, 2, \ldots$, $j = 1, 2, \ldots$, if $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |a_{ij}|$ converges, then*

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij} < \infty.$$

This proposition is a special case of Fubini-Tonelli theorem, which is obtained by combining Fubini's theorem (see Theorem 19 on page 307 of [46]) and Tonelli's theorem. We will also use monotone convergence theorem (MCT) and dominated convergence theorem (DCT).

Proposition 3.7 (Series version of monotone convergence theorem, Corollary 5.3.1 of [40]). *If X_i are nonnegative random variables for $i = 1, 2, \ldots$, then*

$$E\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} E[X_i].$$

Proposition 3.8 (Dominated convergence theorem, Theorem 5.3.3 of [40]). *If a sequence of random variables $\{X_i\}_{i=1}^{\infty}$ converges to a random variable X and there exists a dominating random variable Z such that $|X_i| \leq Z$ for $i = 1, 2, \ldots$ and $E[|Z|] < \infty$, then*

$$E[X_i] \to E[X].$$

3.2 CILP Formulations

In this section, we introduce primal and dual CILP formulations of countable-state MDPs. We start with a straightforward result which was used in [39] and [45] without being explicitly stated.

Lemma 3.9. *A policy is optimal if and only if it is optimal for an initial state distribution that has a positive probability at every state.*

Proof: For a policy π and an initial state distribution β, observe that

$$J_\pi(\beta) = E_\pi^\beta\left[\sum_{n=1}^{\infty} \alpha^{n-1} r(S_n, A_n)\right] = \sum_{s=1}^{\infty} \beta(s) E_s^\pi\left[\sum_{n=1}^{\infty} \alpha^{n-1} r(S_n, A_n)\right] = \sum_{s=1}^{\infty} \beta(s) J_\pi(s).$$

Since $J_\pi(s) \leq J^\star(s)$ for any $s \in \mathcal{S}$, and $\beta(s) > 0$ for any $s \in \mathcal{S}$, a policy π maximizes $J_\pi(\beta)$ if and only if it maximizes $J_\pi(s)$ for each state s, and thus, the equivalency is proven. □

Using this lemma, we equivalently consider finding an optimal policy for a fixed initial state distribution that satisfies

$$\beta(s) > 0 \text{ for all } s \in \mathcal{S}. \tag{3.6}$$

Additionally, we require that β satisfies

$$\beta^T w = \sum_{s=1}^{\infty} \beta(s) w(s) < \infty. \tag{3.7}$$

(3.7) will help us show that a variety of infinite series we consider in this chapter converge. Note that β is not a given problem parameter and that there are many functional forms of w that allow us to choose β satisfying (3.6) and (3.7). For example, if $w \in \mathcal{O}(s^m)$ for some positive number m (in other words, w is asymptotically dominated by a polynomial in s), then we can easily find β satisfying the conditions by modifying an exponential function appropriately.

Now we introduce a CILP formulation of a countable-state MDP.

$$\text{(CP)} \ \max \ f(x) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a) x(s,a) \tag{3.8}$$

$$\text{s.t.} \ \sum_{a=1}^{A} x(s,a) - \alpha \sum_{t=1}^{\infty} \sum_{a=1}^{A} p(s|t,a) x(t,a) = \beta(s) \text{ for } s \in \mathcal{S} \tag{3.9}$$

$$x \geq 0, \ x \in l_1,$$

where l_1 is the space of absolutely summable sequences: $x \in l_1$ means

$$\sum_{s=1}^{\infty} \sum_{a=1}^{A} |x(s,a)| < \infty.$$

Derivations of (CP) can be found in the literature even for more general classes of MDPs. (Chapter 12.3 of [18] introduced a similar CILP formulation for a more general class of MDPs, but for the average reward criterion. Additionally, in Chapter 8.8 of [4], a similar CILP formulation was derived for constrained MDPs, with regular MDPs a special case. However, assumptions used in the latter are quite different from ours and it is not known either if his set of assumptions implies ours or vice versa.) In Section 3.5.1, we provide a high-level derivation of (CP). Briefly, (CP) is derived by a convex analytic approach which

considers the MDP as a convex optimization problem maximizing a linear functional over the convex set of occupancy measures. (An occupancy measure corresponding to a policy is the total expected discounted time spent in different state-action pairs under the policy; for a precise definition, see Section 3.5.1.) It is well known that $\mathcal{P}$, which denotes the set of feasible solutions to (CP), coincides with the set of occupancy measures of stationary policies. For any stationary policy σ and its occupancy measure $x \in \mathcal{P}$, $J_\sigma(\beta)$ is equal to the objective function value of (CP) at x. An optimal stationary policy (which is known to be an optimal policy) can therefore be obtained from an optimal solution to (CP) by computing the corresponding stationary policy (for more details, see Section 3.5.1).

The following visualization of constraint (3.9) will help readers understand the structure of (CP). Using infinite matrix and vector notation, constraint (3.9) can be written as

$$[M^1|M^2|\ldots|M^s|\ldots]\,x = \beta. \tag{3.10}$$

Here, for $s \in \mathcal{S}$, M^s is an $\infty \times A$ matrix whose rows are indexed by states and $M^s = E^s - \alpha P^s$, where each column of E^s is the unit vector e^s, and the ath column of P^s is the probability distribution $p(\cdot|s,a)$.[2]

Remark 3.10. Let us re-visit Example 2. For any state s, for any action $a = (a^1, a^2)$ such that $a^1 > 0$, transition to any state $t \geq s$ has a positive probability. That is, the ath column of P^s has an infinite number of positive entries. On the other hand, any state s can be reached by a transition from any state $t \geq s$ by an action $a = (a^1, a^2)$ such that $a^2 > 0$. That is, for any $t \geq s$, the entry of P^t at the sth row and the ath column is positive. Consequently, in the CILP (CP) for Example 2, there are variables that appear in an infinite number of constraints (unless $\mathcal{A}^1 = \{0\}$) and each constraint has an infinite number of variables (unless $\mathcal{A}^2 = \{0\}$).

We consider another CILP formulation of countable-state MDPs. Let $\|y\|_w \triangleq \sup_{s\in\mathcal{S}} \frac{|y(s)|}{w(s)}$

[2]Note that the rows of P^s are indexed by next states. Meanwhile, given a stationary deterministic policy σ, the rows of P_σ are indexed by current states and its columns are indexed by next states.

for $y \in \mathbb{R}^\infty$ and $Y_w \triangleq \{y \in \mathbb{R}^\infty : \|y\|_w < \infty\}$. Consider the following CILP:

$$\text{(CD)} \ \min \ g(y) = \sum_{s=1}^{\infty} \beta(s) y(s) \tag{3.11}$$

$$\text{s.t. } y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a) y(t) \geq r(s,a) \text{ for } s \in \mathcal{S} \text{ and } a \in \mathcal{A} \tag{3.12}$$

$$y \in Y_w. \tag{3.13}$$

A CILP formulation consisting of (3.11) and (3.12) was introduced in Chapter 2.5 of [45] for MDPs with uniformly bounded rewards. By adding constraint (3.13), one can apply essentially the same arguments to countable-state MDPs being considered in this chapter (i.e., ones with unbounded rewards but satisfying Assumptions A1, A2, and A3) to show that the optimal value function $J^\star$ is equal to the unique optimal solution of (CD).

A few remarks about (CD) are in order. Note that for any $y \in Y_w$, the objective function value is always finite because of (3.7). Also, the infinite sum in each constraint, $\sum_{t=1}^{\infty} p(t|s,a) y(t)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, can be shown to be finite for any $y \in Y_w$ by using Assumption A2. Also, under Assumptions A1, A2, and A3, value functions of all Markov policies belong to Y_w due to Proposition 3.4, so (3.13) does not exclude any solution of interest. Since, for any optimal policy $\pi^\star$,

$$J^\star(\beta) = J_{\pi^\star}(\beta) = \sum_{s=1}^{\infty} \beta(s) J_{\pi^\star}(s) = \sum_{s=1}^{\infty} \beta(s) J^\star(s), \tag{3.14}$$

the optimal value of (CD) equals $J^\star(\beta)$. Lastly, we note that Y_w is a Banach space, so (CD) is a problem of minimization of a linear function in a Banach space while satisfying linear inequalities.

Section 3.5.1 proves that the optimal objective function value of (CP) is $J^\star(\beta)$, and thus, we have the following strong duality theorem.

Theorem 3.11. *Strong duality holds between (CP) and (CD), i.e., $f(x^*) = g(y^*)$, where x^**

and y^* are optimal solutions of (CP) and (CD), respectively.

Note that (CP) has only equality constraints and non-negativity constraints, and thus can be said to be in standard form. The main goal of this chapter is to develop a simplex-type algorithm that solves (CP). A simplex-type algorithm is expected to move along an edge between two adjacent extreme points, improving the objective function value at every iteration, and converge to an extreme point optimal solution. The following characterization of extreme points of $\mathcal{P}$ is also well known in literature (e.g., Theorem 11.3 of [18]).

Theorem 3.12. *A feasible solution x of (CP) is an extreme point of $\mathcal{P}$ if and only if for any $s \in \mathcal{S}$, there exists $a(s) \in \mathcal{A}$ such that $x(s, a(s)) > 0$ and $x(s, b) = 0$ for all $b \neq a(s)$. That is, the extreme points of $\mathcal{P}$ correspond to stationary deterministic policies.*

Therefore, it is natural to define basic feasible solution in the following way.

Definition 3.13. A feasible solution x to (CP) is defined to be a *basic feasible solution* of (CP) if for any $s \in \mathcal{S}$, there exists $a(s) \in \mathcal{A}$ such that $x(s, a(s)) > 0$ and $x(s, b) = 0$ for all $b \neq a(s)$.

Note that a basic feasible solution is determined by choosing one column from each block matrix M^s in (3.10) for $s \in \mathcal{S}$. For a basic feasible solution x and for $s \in \mathcal{S}$, the unique action $a(s)$ that satisfies $x(s, a(s)) > 0$ is called a *basic action* of x at state s. Basic actions of x naturally define a stationary deterministic policy, say, σ. Recall that $\mathcal{P}$ is the set of occupancy measures of stationary policies; moreover, the set of extreme points of $\mathcal{P}$ coincides with the set of occupancy measures of stationary deterministic policies. Thus, conversely, the extreme point x is the occupancy measure of the stationary deterministic policy σ.

The next theorem follows immediately, based on the existence of an optimal policy that is stationary and deterministic and the correspondence between stationary deterministic policies and extreme points (Theorem 11.3 of [18]).

Theorem 3.14. *(CP) has an extreme point optimal solution.*

In Appendix B, we provide an alternative proof of the above theorem. Briefly, the proof shows that $\mathcal{P}$ is convex and compact in $\mathbb{R}^\infty$ under the product topology and that the objective function $f(x)$ of (CP) is continuous over the feasible region $\mathcal{P}$, and then, by Bauer's Maximum Principle (e.g., see Theorem 7.69 of [2]), that (CP) has an extreme point optimal solution.

Next, we define complementary slackness between solutions of (CP) and (CD), and prove its equivalence to optimality.

Definition 3.15. *(Complementary slackness)* Suppose $x \in \mathcal{P}$ and $y \in Y_w$. x and y are said to satisfy *complementary slackness* (or be *complementary*) if

$$x(s,a)\left[r(s,a) - \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] = 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{3.15}$$

Theorem 3.16. (Complementary slackness sufficiency) *Suppose $x \in \mathcal{P}$ and $y \in Y_w$ are complementary. Then $f(x) = g(y)$, and if y is feasible to (CD), then x and y are optimal to (CP) and (CD), respectively.*

Proof: In Section 3.5.2.

Theorem 3.17. (Complementary slackness necessity) *If x and y are optimal to (CP) and (CD), respectively, then they are complementary.*

Proof: In Section 3.5.3.

Given a basic feasible solution x, let σ be the corresponding stationary deterministic policy. By Theorem 3.3(2) and the definition of complementary slackness, a $y \in Y_w$ is complementary with x if and only if y is the value function of σ. Since the value function of a policy is unique, for any basic feasible solution x, there exists a unique $y \in Y_w$ that is complementary with x, and moreover, y satisfies $|y| \leq Lw$ by Proposition 3.4.

Recently in [20], it was shown that for general CILPs, weak duality and complementary slackness could be established by choosing appropriate sequence spaces for primal and dual,

and the result was applied to CILP formulations of countable-state MDPs with bounded rewards. In the paper, one of the conditions for the choice of sequence space is that the objective function should converge for any sequence in the sequence space. However, in (CP), the sequence space l_1 does not guarantee convergence of the objective function (but the objective function converges for any feasible solution of (CP) as shown in Section 3.5.1). Thus, for countable-state MDPs with unbounded rewards being considered in this chapter, applying the choice of sequence spaces in [20] would yield a different CILP formulation from (CP), in which the feasible region may not coincide the set of occupancy measures of stationary policies.

We conclude this section with the next lemma which will be useful in later sections.

Lemma 3.18. *Any $x \in \mathcal{P}$ satisfies*

$$\sum_{s=1}^{\infty} \sum_{a=1}^{A} x(s,a) = \frac{1}{1-\alpha}.$$

Proof: In Section 3.5.4.

3.3 Simplex Algorithm

To devise a simplex-type algorithm for (CP), let us recall how the simplex method for finite LPs works (in case of maximization). It starts with an initial basic feasible solution, and in each iteration, computes reduced costs of nonbasic variables, chooses a nonbasic variable with a positive reduced cost, and then replaces a basic variable with this nonbasic variable to move to an adjacent basic feasible solution (this step is called a pivot operation). The difficulties in replicating this for general CILPs are summarized in [21, 24]: 1) for a given solution, checking feasibility may require infinite data and computation, 2) it generally requires infinite memory to store a solution, 3) there are an infinite number of nonbasic variables to consider for pivot operation, 4) computing reduced cost of even one nonbasic variable may require infinite data and computation. In addition to these difficulties in implementation,

[21] provided an example of a CILP in which a strictly improving sequence of adjacent extreme points may not converge in value to optimality. Therefore, an implementable simplex algorithm for (CP) should store each iterate in finite memory, and approximate reduced costs of only a finite number of nonbasic variables using only finite computation and data in every iteration. We should also ensure that the algorithm improves in every iteration and converges to optimality despite the above restrictions.

In [24], a simplex algorithm for non-stationary MDPs with finite state space that satisfies all of the requirements was introduced. Here we introduce a simplex algorithm that satisfies all of the requirements for a larger class of MDPs, namely, *countable-state MDPs*.

3.3.1 Approximating Reduced Costs

In this section we describe how we approximate reduced costs and prove an error bound for the approximation. Let x be a basic feasible solution to (CP) and let $y \in Y$ be its complementary solution. We first define reduced costs.

Definition 3.19. Given a basic feasible solution x and the corresponding complementary solution y, *reduced cost* $\gamma(s,a)$ of state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is defined as negative of the slack in the corresponding constraint in (CD):

$$\gamma(s,a) \triangleq r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t) - y(s). \tag{3.16}$$

For a state-action pair (s,a) such that $x(s,a) > 0$, the reduced cost $\gamma(s,a)$ is zero by complementarity. If $\gamma(s,a) \leq 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, it means that y is feasible to (CD), and thus, x is optimal to (CP) by Theorem 3.16.

Let σ be the stationary deterministic policy corresponding to x. Fix a state s and an action $a \neq \sigma(s)$ and consider a stationary deterministic policy τ obtained from σ by changing the basic action at state s to a. We call this procedure for obtaining τ from σ a *pivot operation*. Let z be the basic feasible solution corresponding to τ. The next

proposition shows the relation between the change in objective function value made by this pivot operation and the reduced cost $\gamma(s,a)$.

Proposition 3.20. *In the aforementioned pivot operation, the difference in objective function values of x and z is given by*

$$f(z) - f(x) = \gamma(s,a) \sum_{t=1}^{\infty} \beta(t) \sum_{n=0}^{\infty} \alpha^n P_\tau^n(s|t).^3$$

If the reduced cost $\gamma(s,a)$ is positive, then the objective function strictly increases after the pivot operation, by at least $\beta(s)\gamma(s,a)$.

Proof: First, note that we can easily show that the infinite sum on the right hand side is finite because probabilities are less than or equal to one. Let y and v be the complementary solutions of x and z, respectively. Then, we have $y = r_\sigma + \alpha P_\sigma y$ and $v = r_\tau + \alpha P_\tau v$. Thus, $v - y = r_\tau + \alpha P_\tau v - y = r_\tau + \alpha P_\tau (v-y) + \alpha P_\tau y - y$ where the last equality follows because each entry of $P_\tau v$ and $P_\tau y$ is finite (since $|v|$ and $|y|$ are bounded by Lw and each entry of $P_\tau w$ is finite by Assumption A2). By Theorem C.2 in [39], $(I - \alpha P_\tau)^{-1}$ exists for any stationary policy τ and we have[4]

$$(I - \alpha P_\tau)^{-1} \triangleq I + \alpha P_\tau + \alpha^2 P_\tau^2 + \dots.$$

Therefore, we have $v - y = (I - \alpha P_\tau)^{-1}(r_\tau + \alpha P_\tau y - y)$. Entries of the infinite vector $r_\tau + \alpha P_\tau y - y$ are

$$(r_\tau + \alpha P_\tau y - y)(t) = r(t,\tau(t)) + \alpha \sum_{t'=1}^{\infty} p(t'|t,\tau(t)) y(t') - y(t) = \begin{cases} \gamma(s,a) & \text{if } t = s \\ 0 & \text{otherwise.} \end{cases}$$

[3]Note that in the second sum, the index n starts from 0 since it is not a period index, but denotes a number of transitions.

[4]Because αP_τ is a bounded linear operator on Y_w equipped with the norm $\|\cdot\|_w$ and the spectral radius of αP_τ is strictly less than one, the conditions of the theorem is satisfied.

Therefore,

$$f(z) - f(x) = \beta^T v - \beta^T y = \beta^T (v - y) = \beta^T (I - \alpha P_\tau)^{-1} (r_\tau + \alpha P_\tau y - y)$$
$$= \gamma(s,a) \sum_{t=1}^{\infty} \beta(t) \sum_{n=0}^{\infty} \alpha^n P_\tau^n(s|t),$$

establishing the first result. Because

$$\sum_{t=1}^{\infty} \beta(t) \sum_{n=0}^{\infty} \alpha^n P_\tau^n(s|t) \geq \beta(s) P_\tau^0(s|s) = \beta(s) > 0,$$

the second claim is also proven. □

Computing the reduced cost of even one state-action pair requires computing y. Recall that y is the value function of the policy σ. Computing y requires an infinite amount of computation and an infinite amount of data, no matter how it is computed, either by computing the infinite sum (3.1) or solving the infinite system of equations $y = r_\sigma + \alpha P_\sigma y$.

For a given policy σ, we consider approximating the complementary solution y by solving the following *N-state truncation* of the infinite system of equations $y = r_\sigma + \alpha P_\sigma y$. Let N be a positive integer. The approximate complementary solution, which we denote as y^N, is defined to be the solution of the following *finite* system of equations:

$$y^N(s) = r_\sigma(s) + \alpha \sum_{t=1}^{N} P_\sigma(t|s) y^N(t) \text{ for } s = 1, \ldots, N. \tag{3.17}$$

Note that y^N is the value function of policy σ for a new MDP obtained by replacing states greater than N by an absorbing state in which no reward is earned, and thus, y^N is an approximation of y obtained from the N-state truncation of the original MDP. The next lemma provides an error bound for the approximate complementary solution.

Lemma 3.21. *For any positive integer N, the approximate complementary solution y^N*

satisfies

$$|y^N(s) - y(s)| \leq L \sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s) w(t) \text{ for } s = 1, \ldots, N.$$

The error bound on the right hand side converges to zero as $N \to \infty$. Therefore, y^N converges pointwise to y as $N \to \infty$.

Proof: In Section 3.5.5.

Using the approximate complementary solution, we define approximate reduced costs of nonbasic variables that belong to the N-state truncation:

$$\gamma^N(s,a) \triangleq r(s,a) + \alpha \sum_{t=1}^{N} p(t|s,a) y^N(t) - y^N(s) \text{ for } s = 1, \ldots, N, a \in \mathcal{A}. \tag{3.18}$$

Note that $\gamma^N(s,a)$ is an approximation of reduced cost $\gamma(s,a)$ computed by using y^N in place of y. The next lemma provides an error bound on the approximate reduced cost.

Lemma 3.22. *For any positive integer N, the approximate reduced cost γ^N satisfies*

$$|\gamma^N(s,a) - \gamma(s,a)| \leq \delta(\sigma, s, a, N) \text{ for } s = 1, \ldots, N, a \in \mathcal{A},$$

where we define

$$\begin{aligned} \delta(\sigma, s, a, N) \triangleq & L \sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s) w(t) + \alpha L \sum_{t=1}^{N} p(t|s,a) \sum_{t'>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t'|t) w(t') \\ & + \alpha L \sum_{t>N} p(t|s,a) w(t). \end{aligned} \tag{3.19}$$

Proof: By Lemma 3.21 and (3.5), for any $s \leq N$ and $a \in \mathcal{A}$,

$$
\begin{aligned}
&|\gamma^N(s,a) - \gamma(s,a)| \\
&\leq \alpha \sum_{t=1}^{N} p(t|s,a)|y^N(t) - y(t)| + |y^N(s) - y(s)| + \alpha \sum_{t>N} p(t|s,a)|y(t)| \\
&\leq \alpha L \sum_{t=1}^{N} p(t|s,a) \sum_{t'>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t'|t) w(t') + L \sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s) w(t) + \alpha L \sum_{t>N} p(t|s,a) w(t) \\
&= \delta(\sigma, s, a, N),
\end{aligned}
$$

which proves the lemma. □

By using Assumptions A2 and A3 and arguments similar to those in Section 3.5.5, it is not hard to prove the following proposition about $\delta(\sigma, s, a, N)$.

Proposition 3.23. *For any positive integer N and for $\sigma \in \Pi_{SD}, s = 1, \ldots, N$, and $a \in \mathcal{A}$,*

$$\delta(\sigma, s, a, N) \leq L(L + \alpha\kappa L + \alpha\kappa) w(s), \tag{3.20}$$

and for any $\sigma \in \Pi_{SD}, s = 1, \ldots, N$, and $a \in \mathcal{A}$,

$$\delta(\sigma, s, a, N) \to 0 \text{ as } N \to \infty. \tag{3.21}$$

Thus, by this proposition and Lemma 3.22, we have $\gamma^N(s,a) \to \gamma(s,a)$ as $N \to \infty$ for any state-action pair (s,a).

To design a convergent simplex-like algorithm for solving (CP), we need to assume the existence of a uniform (policy independent) upper bound on $\delta(\sigma, s, a, N)$, i.e., $\bar{\delta}(s, a, N) \geq \delta(\sigma, s, a, N)$ for all $\sigma \in \Pi_{SD}$, positive integer N, $s \leq N$, and $a \in \mathcal{A}$, such that

$$\bar{\delta}(s, a, N) \to 0 \text{ as } N \to \infty \text{ for any } (s,a). \tag{3.22}$$

Additionally, for the algorithm to be implementable, we require $\bar{\delta}(s, a, N)$ to be computable

in finite time, using finite data. In Section 3.3.4, we show how such an upper bound $\bar{\delta}(s, a, N)$ can be computed for Examples 1 and 2.

3.3.2 Simplex Algorithm

Our simplex algorithm finds a sequence of stationary deterministic policies whose value functions strictly improve in every iteration and converge to the optimal value function. Let σ^k denote the stationary deterministic policy our algorithm finds in iteration k. Let x^k denote the corresponding basic feasible solution of (CP) and y^k denote the complementary solution.

The intuition behind the algorithm can be described as follows. If σ^k is not optimal, y^k is not feasible to (CD), and thus, there is at least one nonbasic variable (state-action pair) whose reduced cost is positive. To identify such a variable with finite computation, in each iteration we consider N-state truncations of the MDP, increasing N as necessary. As N increases, the variable's approximate reduced cost approaches its exact value, and for sufficiently large N becomes sufficiently large to deduce (by Lemma 4.4) that the (exact) reduced cost of the variable is positive. Moreover, in choosing a variable for the pivot operation, the algorithm selects a nonbasic variable that not only has a positive reduced cost, but also has the largest approximate reduced cost (weighted by β) among all nonbasic variables in the N-state truncation; this choice is similar to the Dantzig pivoting rule for finite LPs. Choosing a nonbasic variable with a positive reduced cost ensures strict improvement, and choosing one with the largest weighted approximate reduced cost enables us to prove convergence to optimality. (As demonstrated by a counter-example in [21], an arbitrary sequence of improving pivot operations may lead to convergence to a suboptimal value.) A unique feature of our algorithm is that in each iteration it adjusts N, the size of finite-state truncation, dynamically until a condition for performing a pivot operation is satisfied, whereas existing solution methods for countable-state MDPs increase the size by one in every iteration.

An implementable simplex algorithm for countable-state MDPs

1. Initialize: Set iteration counter $k = 1$. Fix basic actions $\sigma^1(s) \in \mathcal{A}$ for $s \in \mathcal{S}$.[5]

2. Find a nonbasic variable with the most positive *approximate* reduced cost:

   (a) Set $N := 1$ and set $N(k) := \infty$.

   (b) Compute the approximate complementary solution, $y^{k,N}(s)$ for $s = 1, \ldots, N$ by solving (3.17).

   (c) Compute the approximate reduced costs, $\gamma^{k,N}(s, a)$ for $s = 1, \ldots, N, a \in \mathcal{A}$ by (3.18).

   (d) Find the nonbasic variable achieving the largest *approximate* nonbasic reduced cost weighted by β:

$$(s^{k,N}, a^{k,N}) = \arg\max_{(s,a)} \beta(s)\gamma^{k,N}(s, a). \tag{3.23}$$

   (e) If $\gamma^{k,N}(s^{k,N}, a^{k,N}) > \bar{\delta}(s^{k,N}, a^{k,N}, N)$, set $N(k) = N$, $(s^k, a^k) = (s^{k,N}, a^{k,N})$, and $\sigma^{k+1}(s^k) = a^k$, $\sigma^{k+1}(s) = \sigma^k(s)$ for $s \neq s^k$, and go to Step 3; else set $N := N + 1$ and go to Step 2(b).

3. Set $k = k + 1$ and go to Step 2.

3.3.3 Proof of Convergence

In this section we show that the simplex algorithm of Section 3.3.2 strictly improves in every iteration and that it converges to optimality.

In Step 2(e) of the algorithm, a pivot operation is performed only if $\gamma^{k,N}(s^k, a^k) > \bar{\delta}(s^k, a^k, N)$. This inequality implies that the reduced cost of variable $x(s^k, a^k)$ is positive as

[5]Note that we can select an initial policy that can be described finitely. For example, for $\mathcal{A} = \{1, \ldots, A\}$, we can let $\sigma^1(s) = 1$ for all $s \in \mathcal{S}$. Then, the algorithm stores only deviations from the initial policy, which total at most k at the kth iteration.

shown in the following lemma. For $s \in \mathcal{S}, a \in \mathcal{A}$, and $k = 1, 2, \ldots$, we use $\gamma^k(s, a)$ to denote the reduced cost of variable $x(s, a)$ where the current policy is σ^k.

Lemma 3.24. *The reduced cost $\gamma^k(s^k, a^k)$ of the state-action pair chosen in iteration k of the simplex algorithm is strictly positive.*

Proof: We have

$$\gamma^k(s^k, a^k) \geq \gamma^{k,N}(s^k, a^k) - \delta(\sigma^k, s^k, a^k, N) \geq \gamma^{k,N}(s^k, a^k) - \bar{\delta}(s^k, a^k, N) > 0$$

where the first inequality follows by Lemma 3.22, the second by the definition of $\bar{\delta}(s^k, a^k, N)$, and the last by Step 2(e) of the algorithm. □

By this lemma and Proposition 3.20, the following corollary is immediate. We denote $f(x^k)$ as f^k for simplicity.

Corollary 3.25. *The objective function of (CP) is strictly improved by the simplex algorithm in every iteration, i.e., $f^{k+1} > f^k$ for $k = 1, 2, \ldots$.*

The next corollary shows that the value function of the policies found by the algorithm improves in every iteration.

Corollary 3.26. *The value function of the policies obtained by the simplex algorithm is nondecreasing in every state and strictly improves in at least one state in every iteration, i.e., for any k, $y^{k+1} \geq y^k$ and there exists $s \in \mathcal{S}$ for which $y^{k+1}(s) > y^k(s)$.*

Proof: As shown in the proof of Proposition 3.20,

$$y^{k+1} - y^k = (I - \alpha P_{\sigma^{k+1}})^{-1}(r_{\sigma^{k+1}} + \alpha P_{\sigma^{k+1}} y^k - y^k),$$

and for $s \in \mathcal{S}$,

$$y^{k+1}(s) - y^k(s) = \gamma^k(s^k, a^k) \sum_{n=0}^{\infty} \alpha^n P^n_{\sigma^{k+1}}(s^k | s).$$

Since $\gamma^k(s^k, a^k) > 0$, we have $y^{k+1}(s) - y^k(s) \geq 0$ for all $s \in \mathcal{S}$. Moreover,

$$y^{k+1}(s^k) - y^k(s^k) = \gamma^k(s^k, a^k) \sum_{n=0}^{\infty} \alpha^n P^n_{\sigma^{k+1}}(s^k|s^k) \geq \gamma^k(s^k, a^k) P^0_{\sigma^{k+1}}(s^k|s^k) = \gamma^k(s^k, a^k) > 0.$$

□

From the above corollaries, the next corollary is trivial.

Corollary 3.27. *The simplex algorithm does not repeat any non-optimal basic feasible solution.*

The next lemma shows that the algorithm finds a pivot operation satisfying the conditions as long as the current basic feasible solution is not optimal.

Lemma 3.28. *Step 2 of the algorithm terminates if and only if x^k is not optimal to (CP).*

Proof: In Section 3.5.6.

In the rest of this section we show that the algorithm converges in value to optimality. We begin by proving a few useful lemmas.

From Proposition 3.20, we know that $\beta(s^k)\gamma^k(s^k, a^k)$ is a lower bound on the improvement of the objective function in iteration k. The next lemma shows that f^k converges, and thus the guaranteed improvement should converge to zero.

Lemma 3.29. *The sequence f^k has a finite limit and $\beta(s^k)\gamma^k(s^k, a^k)$ tends to zero as $k \to \infty$.*

Proof: For any k,

$$f^k = f(x^k) = g(y^k) = \sum_{s=1}^{\infty} \beta(s) y^k(s) \leq L \sum_{s=1}^{\infty} \beta(s) w(s) < \infty,$$

where the second equality follows by Theorem 3.16, the first inequality by (3.5), and the last inequality by (3.7). By Corollary 3.25, the sequence f^k is an increasing sequence. Therefore, f^k has a finite limit, and thus $f^{k+1} - f^k$ converges to zero as $k \to \infty$. Since $\beta(s^k)\gamma^k(s^k, a^k)$

is nonnegative for any k, by Proposition 3.20, we can conclude that $\beta(s^k)\gamma(s^k, a^k)$ converges to zero. □

The next lemma shows that $N(k)$, the size of the finite truncation at which the simplex algorithm finds a state-action pair satisfying the conditions of Step 2(e), tends to infinity as $k \to \infty$.

Lemma 3.30. *$N(k) \to \infty$ as $k \to \infty$.*

Proof: This proof is similar to the proof of Lemma 5.7 in [24].

The lemma holds trivially if x^k is optimal for any k. Suppose that this is not the case, and that there exists an integer M such that $N(k) = M$ for infinitely many k. Let $\{k_i\}_{i=1}^{\infty}$ be the infinite subsequence of iteration counters in which this occurs. Let $\sigma^{k_i,M}$ be the stationary deterministic policy in the M-state truncation defined by $\sigma^{k_i}(s)$ for $s = 1, \ldots, M$. Note that in the M-state truncation of the original MDP, since $\mathcal{A}$ is finite, there are only a finite number of stationary deterministic policies. Thus, there exists a stationary deterministic policy of the M-state truncation that appears for infinitely many k_i. Let $\sigma^{*,M}$ denote the M-state stationary deterministic policy and, passing to a subsequence if necessary, let $\sigma^{k_i,M} = \sigma^{*,M}$.

In the simplex algorithm, the nonbasic variable chosen by the algorithm is completely characterized by the basic feasible solution of the M-state truncation. Thus, in iteration k_i for $i = 1, 2, \ldots$, the state-action pair chosen for a pivot operation is the same. Let (s^*, a^*) denote this state-action pair. For $i = 1, 2, \ldots$, in iteration k_i of the simplex algorithm, the improvement of the objective function is

$$\begin{aligned} f^{k_i+1} - f^{k_i} &\geq \beta(s^*)\gamma^{k_i}(s^*, a^*) \geq \beta(s^*)(\gamma^{k_i,M}(s^*, a^*) - \delta(\sigma^{k_i}, s^*, a^*, M)) \\ &\geq \beta(s^*)(\gamma^{k_i,M}(s^*, a^*) - \bar{\delta}(s^*, a^*, M)) > 0, \end{aligned}$$

where the first inequality follows by Proposition 3.20, the second by Lemma 3.22, the third by the definition of $\bar{\delta}(s^*, a^*, M)$, and the last by Step 2(e) of the algorithm. Note that the approximate reduced cost $\gamma^{k_i,M}(s^*, a^*)$ is also solely determined by the basic feasible solution

of the M-state truncation. Thus, the last nonzero expression in the above inequalities, $\beta(s^*)(\gamma^{k_i,M}(s^*,a^*) - \bar{\delta}(s^*,a^*,M))$, is a positive constant. This implies that the objective function is increased by at least a fixed amount in iteration k_i for $i = 1, 2, \ldots$. However, we know that f^k is an increasing convergent sequence from Corollary 3.25 and Lemma 3.29. Thus, we established the result by contradiction. □

Theorem 3.31. *Let $f^\star$ be the optimal value of (CP). The simplex algorithm converges to optimality in value, i.e.,* $\lim_{k\to\infty} f^k = f^\star$.

Proof: The main steps of this proof are similar to the steps of the proof of Theorem 5.3 in [24], but details of each step are quite different. We borrowed some of their notation.

This theorem trivially holds if x^k is optimal for any k, so suppose that this is not the case.

There exists a sequence of positive integers $\{r_k\}$ such that $s^{r_k} \to \infty$ as $k \to \infty$. Indeed, recall that s^k is the state where the algorithm performs a pivot operation in iteration k. Suppose that there exists N' such that $s^k < N'$ for all k. Then, the algorithm performs pivot operations only for states less than N', and thus, can encounter only a finite number of basic feasible solutions, since the action set $\mathcal{A}$ is finite. However, we assumed that x^k is not optimal for any k and the algorithm performs a pivot operation as long as it does not reach an optimal solution (Lemma 3.28) and never repeats any non-optimal basic feasible solutions (Corollary 3.27). Thus, we reached a contradiction.

We will next show that the sequence x^{r_k} has a converging subsequence whose limit is an optimal solution to (CP). The fact that $s^{r_k} \to \infty$ as $k \to \infty$ will play a role in showing the optimality of the limit, later in this proof.

For any k, x^{r_k} belongs to $\mathcal{P}$ which is shown to be compact in Theorem 11.3 of [18] or Corollary 10.1 of [4], and thus, there exists a convergent subsequence x^{t_k} of x^{r_k} with $\lim_{k\to\infty} x^{t_k} = \bar{x}$. Note that $\bar{x} \in \mathcal{P}$. Let y^{t_k} be the corresponding subsequence of y^k. Let $Y_L \triangleq \{y \in \mathbb{R}^n \ : \ \|y\|_w \leq L\}$, then Y_L is a compact set of $\mathbb{R}^\infty$ under the product topology by Tychonoff's theorem (e.g., see Theorem 2.61 of [2]). By (3.5), we have $y^{t_k} \in Y_L$ for all k, and

thus, the subsequence y^{t_k} also has a further convergent subsequence y^{u_k}. Let $\lim_{k\to\infty} y^{u_k} = \bar{y}$, and note that $\lim_{k\to\infty} x^{u_k} = \bar{x}$. We will show that $\bar{x}$ and $\bar{y}$ are complementary and $\bar{y}$ is feasible to (CD), and thus, that $\bar{x}$ is optimal for (CP).

Since x^{u_k} and y^{u_k} are complementary, we have

$$x^{u_k}(s,a)\left[r(s,a) - \left(y^{u_k}(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y^{u_k}(t)\right)\right] = 0 \text{ for } s \in \mathcal{S}, a \in \mathcal{A}. \tag{3.24}$$

Recall that, by (3.5), $|y^{u_k}(t)| \le Lw(t)$ for any state t and

$$\sum_{t=1}^{\infty} p(t|s,a)Lw(t) \le \kappa Lw(s) \text{ for any } s \in \mathcal{S}.$$

Thus, by Proposition 3.8, we have

$$\lim_{k\to\infty}\sum_{t=1}^{\infty} p(t|s,a)y^{u_k}(t) = \sum_{t=1}^{\infty} p(t|s,a)\bar{y}(t).$$

Consequently, by taking $k \to \infty$ in (3.24), we obtain

$$\bar{x}(s,a)\left[r(s,a) - \left(\bar{y}(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)\bar{y}(t)\right)\right] = 0 \text{ for } s \in \mathcal{S}, a \in \mathcal{A}.$$

Therefore, $\bar{x}$ and $\bar{y}$ are complementary.

Suppose that $\bar{y}$ is not feasible to (CD). That is, there exists $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\epsilon > 0$ such that

$$r(s,a) + \alpha\sum_{t=1}^{\infty} p(t|s,a)\bar{y}(t) - \bar{y}(s) = \epsilon.$$

Thus, there exists K such that for $k \ge K$,

$$r(s,a) + \alpha\sum_{t=1}^{\infty} p(t|s,a)y^{u_k}(t) - y^{u_k}(s) \ge \frac{1}{2}\epsilon. \tag{3.25}$$

Since $\lim_{k\to\infty} N(u_k) = \infty$ by Lemma 3.30, $s \leq N(u_k)$ for sufficiently large k. For all such k,

$$\begin{aligned} r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a) y^{u_k}(t) - y^{u_k}(s) &= \gamma^{u_k}(s,a) \\ \leq \gamma^{u_k,N(u_k)}(s,a) + \delta(\sigma^{u_k}, s, a, N(u_k)) &\leq \gamma^{u_k,N(u_k)}(s,a) + \bar{\delta}(s,a,N(u_k)) \end{aligned} \tag{3.26}$$

by Lemma 3.22 and the definition of $\bar{\delta}(s,a,N(u_k))$. By Lemma 3.30, we know that

$$\bar{\delta}(s,a,N(u_k)) \to 0 \text{ as } k \to \infty.$$

We will also show that $\gamma^{u_k,N(u_k)}(s,a)$ becomes nonpositive as $k \to \infty$, which will contradict (3.25), and thus, we will conclude that $\bar{y}$ is feasible to (CD).

We have:

$$\begin{aligned} \beta(s)\gamma^{u_k,N(u_k)}(s,a) &\leq \beta(s^{u_k})\gamma^{u_k,N(u_k)}(s^{u_k},a^{u_k}) \\ &\leq \beta(s^{u_k})\gamma^{u_k}(s^{u_k},a^{u_k}) + \beta(s^{u_k})\delta(\sigma^{u_k},s^{u_k},a^{u_k},N(u_k)) \end{aligned} \tag{3.27}$$

where the first inequality is due to (3.23). By Lemma 3.29, the first term of the right hand side of (3.27) tends to zero as $k \to \infty$. Also, by (3.20), the second term of the right hand side of (3.27) is bounded as follows:

$$\beta(s^{u_k})\delta(\sigma^{u_k},s^{u_k},a^{u_k},N(u_k)) \leq L(L + \alpha\kappa L + \alpha\kappa)\beta(s^{u_k})w(s^{u_k}).$$

The right hand side tends to zero as $k \to \infty$ because $\beta(s)w(s) \to 0$ as $s \to \infty$ by (3.7) and $s^{u_k} \to \infty$ as $k \to \infty$ by the choice of sequence u_k. Therefore, the right hand side of (3.27) converges to zero as $k \to \infty$. Since $\beta(s) > 0$, we obtain that $\limsup_k \gamma^{u_k,N(u_k)}(s,a) \leq 0$. Thus, (3.25) is contradicted and so $\bar{y}$ is feasible to (CD).

Thus, we have shown that $\bar{x}$ is optimal to (CP). By following arguments similar to those of Lemma 8.5 of [4], one can show that f, the objective function of (CP), is continuous on

$\mathcal{P}$ under the product topology. Thus, f^{u_k} converges to $f^\star$ as $k \to \infty$. However, f^k converges by Lemma 3.29 and its limit should be the same as the limit of its subsequence. Therefore, f^k converges to $f^\star$ as $k \to \infty$. □

3.3.4 Examples (continued)

Recall that our simplex algorithm relies on $\bar{\delta}(s, a, N)$ — a finitely computable upper bound on $\delta(\sigma, s, a, N)$ that converges to zero as N increases. Let us demonstrate how this bound can be computed for the examples of inventory management and queueing from Section 3.1.3.

Example 3.1 (continued)**.** In the inventory example, recall that the maximum inventory level that can be reached by n transitions from state s is $s + nM$. An upper bound on the first term in (3.19) can be computed as follows:

$$\begin{aligned} L \sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s) w(t) &= L \sum_{n=1}^{\infty} \alpha^n \sum_{t>N} P_\sigma^n(t|s)(C + Dt) \\ &\leq L \sum_{n=1}^{\infty} \alpha^n [C + D(s + nM)] \mathbf{1}\{N < s + nM\} \\ &= L \sum_{n=\nu}^{\infty} \alpha^n [C + D(s + nM)] \\ &= \frac{L\alpha^\nu}{1-\alpha} \left[(C + Ds) + DM \frac{\alpha + \nu - \alpha\nu}{1-\alpha} \right], \end{aligned}$$

where the exchange of infinite sums follows by Proposition 3.5 and $\nu \triangleq \lfloor \frac{N-s}{M} \rfloor + 1$. An upper bound on the rest of (3.19) can be found similarly. Thus, we obtain the following upper

bound on $\delta(\sigma, s, a, N)$:[6]

$$\begin{aligned}
\delta(\sigma, s, a, N) \leq & \frac{L\alpha^{\nu}}{1-\alpha}\left[(C+Ds)+DM\frac{\alpha+\nu-\alpha\nu}{1-\alpha}\right] \\
& + \frac{L\alpha^{2}}{1-\alpha}\left(C+DN+\frac{DM}{1-\alpha}\right)\mathbf{1}\{N<s+M\} \\
& + \frac{L\alpha^{\nu}}{1-\alpha}\left[C+Ds+DM\frac{\alpha+\nu-\alpha\nu}{1-\alpha}\right]\mathbf{1}\{N\geq s+M\} \\
& + L\alpha(C+Ds+DM)\mathbf{1}\{N<s+M\} \\
& \triangleq \bar{\delta}(s,a,N)
\end{aligned}$$

and $\bar{\delta}(s, a, N)$ decreases to zero as N increases. It can also be denoted as $\bar{\delta}(s, N)$ since it does not depend on action a.

In the above example, w is a linear function of the state. However, note that for any polynomial function w, one can easily find $\bar{\delta}(s, a, N)$ that converges to zero as $N \to \infty$ by following arguments similar to the above steps.

Example 3.2 (continued)**.** For $n = 1, 2, \ldots$, let X_n be a Poisson random variable with mean $na^1_{\max}$ and let Y be a random variable that equals X_n with probability $(1-\alpha)\alpha^{n-1}$ for $n = 1, 2, \ldots$. Let μ denote the expected value of Y. For a random variable X, let f_X and F_X denote the probability distribution function and the cumulative distribution function of X, respectively.

Then, for $s \in \mathcal{S}$ and $N = 1, 2, \ldots$, we define $\bar{\delta}(s, a, N)$ as

$$\bar{\delta}(s,a,N) \triangleq L \cdot h_1(s,N) + \alpha L \sum_{t=0}^{N} p(t|s,a) \cdot g(t,N) + \alpha \cdot L \cdot h_2(s,N) \tag{3.28}$$

[6]In (3.19), it is assumed that $\mathcal{S} = \{1, 2, \ldots\}$. However, note that in Examples 1 and 2, $\mathcal{S} = \{0, 1, \ldots\}$. Thus, we derive an upper bound on $\delta(\sigma, s, a, N)$ for which the first sum in the second term in (3.19) starts with $t = 0$ instead of $t = 1$. Then, $\delta(\sigma, s, a, N)$ is an error bound of the approximate reduced cost computed from the $(N+1)$-state truncation.

where

$$h_1(s,N) \triangleq \frac{\alpha}{1-\alpha}\left[C\left(\mu - \sum_{u=0}^{N-s} u f_Y(u)\right) + (Cs+D)(1-F_Y(N-s))\right]$$

and

$$h_2(s,N) \triangleq C\left(a^1_{\max} - \sum_{u=0}^{N-s} u f_{X_1}(u)\right) + (Cs+D)(1-F_{X_1}(N-s)).$$

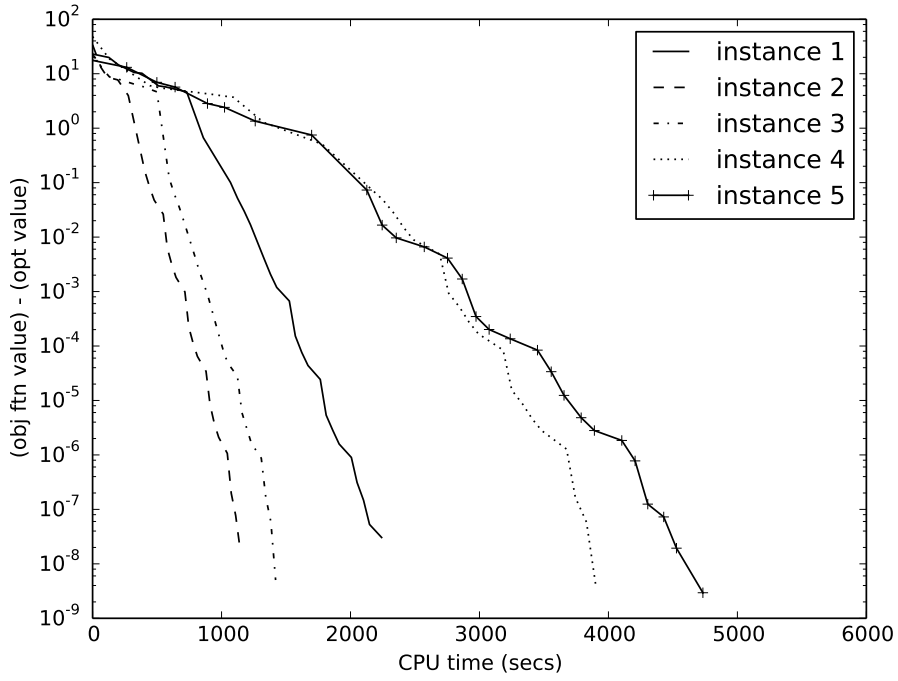
In Section 3.5.7, we prove that $\delta(\sigma,s,a,N) \leq \bar{\delta}(s,a,N)$ and that $\bar{\delta}(s,a,N) \to 0$ as $N \to \infty$, and illustrate how $\bar{\delta}(s,a,N)$ can be computed finitely.

3.4 Numerical Illustration

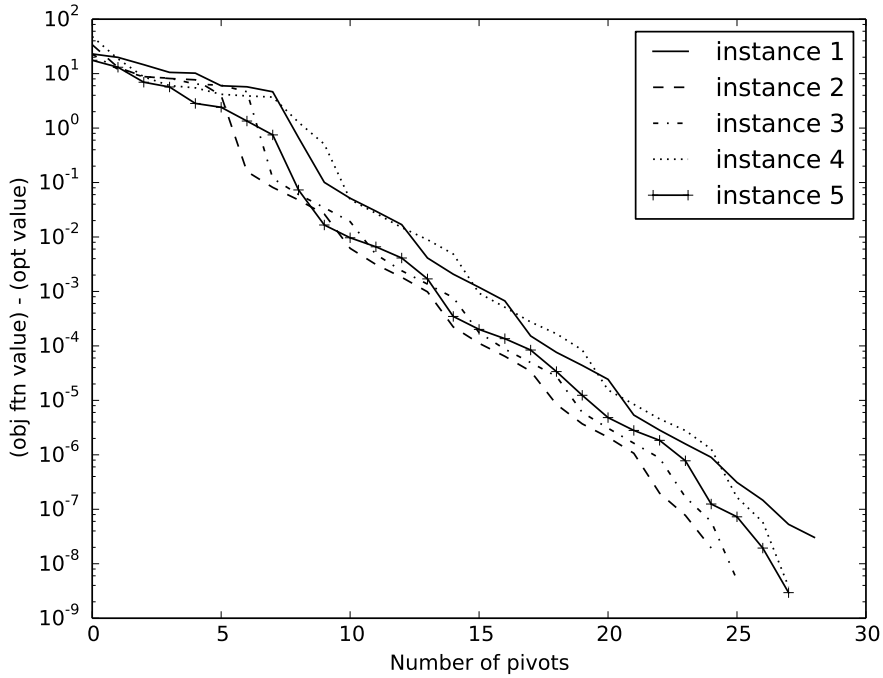
We implemented the simplex algorithm and tested it on five instances of the inventory management problem of Example 3.1. Recall that b, K, c, h, M denote the per-unit price, the fixed ordering cost, the per-unit ordering cost, the per-unit inventory cost, and the maximum ordering level, respectively, and let d denote the expected demand in one period. The parameters of the five instances were $(b,K,c,h,d,M) = (15,3,5,0.1,2,4)$, $(10,5,7,0.1,2,4)$, $(10,3,5,0.2,2,4)$, $(10,3,5,0.2,2,5)$, $(10,3,5,0.2,3,5)$, respectively. For all instances, demand in each period follows Poisson distribution with the specified expected value. We used discount factor $\alpha = 0.9$. The simplex algorithm was written in Python and ran on 2.93 GHz Intel Xeon CPU.

Figure 3.1 shows cost improvement of the simplex algorithm for the above instances of the inventory management problem as a function of (a) CPU time and (b) number of pivot operations. The vertical axis of Figure 3.1 is the difference between the objective function value of (CP) of policies obtained by the simplex algorithm and the optimal objective function value. For $s \in \mathcal{S}$, the initial basic action $\sigma^1(s)$ was the remainder of $s+3$ divided by M. The objective function values of each policy were estimated by computing $\sum_{s=1}^{N} \beta(s) y^N(s)$ for increasing N until the change of the value in consecutive iterations was less than a threshold, where y^N is obtained by solving (3.18).

Figure 3.1 illustrates that for all instances, the difference between the objective function



(a) For CPU time



(b) For number of pivots

Figure 3.1: Optimality gap progress of the simplex algorithm for inventory management problems

value of policies and the optimal value decreased monotonically and converged to zero. As shown in Figure 3.1(b), the algorithm converges at similar rates for all instances as the number of iterations increases, but CPU time of one iteration is longer on average for instances 4 and 5 as shown in Figure 3.1(a), possibly because of the higher maximum ordering level.

3.5 Technical Proofs

3.5.1 Derivation of (CP) and Proof of Strong Duality

Here we provide some intuition behind problem (CP) by illustrating its relationship to the MDP problem. We also prove strong duality between (CP) and (CD).

We first define an *occupancy measure* and will show that the feasible region of (CP) coincides with the set of occupancy measures of all policies. In order to introduce the concept of an occupancy measure, we consider *the expected total reward criterion*, instead of the discounted one. It is well known (e.g., see Chapter 10 of [4]) that we can transform an MDP with the expected total discounted reward criterion into an equivalent MDP with the expected total reward criterion, by adding an absorbing state, say, 0. Let $\tilde{\mathcal{S}} = \mathcal{S} \cup \{0\} = \{0, 1, 2, \ldots\}$ and set the transition probabilities and rewards for $s \in \tilde{\mathcal{S}}$ and $a \in \mathcal{A}$ as:

$$\tilde{p}(t|s,a) \triangleq \begin{cases} \alpha p(t|s,a) & \text{if } s \neq 0, t \neq 0 \\ 1-\alpha & \text{if } s \neq 0, t = 0 \\ 1 & \text{if } s = t = 0 \end{cases},$$

$$\tilde{r}(s,a) \triangleq \begin{cases} r(s,a) & \text{if } s \neq 0 \\ 0 & \text{if } s = 0. \end{cases}$$

Extend β by letting $\beta(0) = 0$ and π by arbitrarily choosing an action at state 0. The

expected total reward is defined as:

$$\tilde{V}_\pi(\beta) \triangleq \tilde{E}_\pi^\beta \left[\sum_{n=1}^{\infty} \tilde{r}(\tilde{S}_n, \tilde{A}_n) \right],$$

where $\tilde{P}_\pi^\beta$ and $\tilde{E}_\pi^\beta$ are defined similarly for the new MDP, and the processes $\{\tilde{S}_n\}$ and $\{\tilde{A}_n\}$ are also defined accordingly. Then it is easy to show that $J_\pi(\beta) = \tilde{V}_\pi(\beta)$ for any policy π. We call this *the absorbing MDP formulation* of the original discounted MDP. It is said to be absorbing since it has a finite expected lifetime before entering 0 under any policy, i.e., $E_\pi^\beta T = 1/(1-\alpha) < \infty$ for any policy π, where $T = \min\{n \geq 1 : s_n = 0\}$. Since the original discounted MDP and its absorbing MDP formulation can be considered equivalent, we use the same notation for both; it will be clear which one is discussed from the context.

For $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the occupancy measure of the state-action pair is denoted as $Q_\pi^\beta(s,a)$ and defined as the expectation of the number of visits to (s,a) until entering the absorbing state 0 under policy π with the initial state distribution β, that is, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\begin{aligned} Q_\pi^\beta(s,a) &\triangleq E_\pi^\beta \sum_{n=1}^{T-1} \mathbf{1}\{S_n = s, A_n = a\} = E_\pi^\beta \sum_{n=1}^{\infty} \mathbf{1}\{S_n = s, A_n = a\} \\ &= \sum_{n=1}^{\infty} P_\pi^\beta\{S_n = s, A_n = a\}, \end{aligned} \tag{3.29}$$

where the last equality is due to Proposition 3.7. (An equivalent alternative interpretation of the occupancy measure is as the total expected discounted time spent in different state-action pairs in the original discounted MDP.)

It is well known (e.g., Theorem 8.1 of [4]) that for any policy π, there exists a stationary policy σ such that $Q_\pi^\beta = Q_\sigma^\beta$, namely,

$$\sigma(a|s) = \frac{Q_\pi^\beta(s,a)}{\sum_{b \in \mathcal{A}} Q_\pi^\beta(s,b)}, \tag{3.30}$$

where $\sigma(a|s)$ denotes the probability of σ choosing a at s. This result implies that $\mathcal{Q} = \mathcal{Q}_M = \mathcal{Q}_S$ where $\mathcal{Q}, \mathcal{Q}_M$, and $\mathcal{Q}_S$ denote the sets of occupancy measures of all policies, Markov policies, and stationary policies, respectively.

It is also well known (e.g., Theorem 11.3 of [18] and Corollary 10.1 of [4]) that $\mathcal{Q}_S$ coincides with the set of nonnegative and summable solutions of the following set of equations:

$$\sum_{a=1}^{A} x(s,a) = \beta(s) + \alpha \sum_{t=1}^{\infty} \sum_{a=1}^{A} p(s|t,a)x(t,a) \text{ for } s \neq 0. \tag{3.31}$$

Therefore, the feasible region of (CP) is the set of occupancy measures of all stationary policies, and thus, it is the set of occupancy measures of all policies.

By using arguments similar to those in the proof of Theorem 8.3 of [4], one can show that for any Markov policy π,

$$J_\pi(\beta) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a) Q_\pi^\beta(s,a). \tag{3.32}$$

From Proposition 3.4 and (3.7), we know that $J_\pi(\beta)$ is finite, and thus the right hand side of the above equation is finite, for any Markov policy π. Since the feasible region of (CP) is the set of occupancy measures of all stationary policies, the objective function of (CP) is finite for any feasible solution. Moreover, by Lemma 3.9, a stationary policy whose occupancy measure is an optimal solution to (CP) is also optimal for the MDP. Given an optimal solution of (CP), a stationary optimal policy can be obtained by (3.30).

By following the arguments of Lemma 8.5 of [4], one can show that f, the objective function of (CP), is continuous on its feasible region under the usual product topology. In addition, it is also well known (e.g., Theorem 11.3 of [18] and Corollary 10.1 of [4]) that the feasible region of (CP) is a compact subset of $\mathbb{R}^\infty$ under the product topology. Therefore, the maximum of f is attained in the feasible region of (CP).

Recall that the optimal value of (CD) is $V^\star(\beta)$. As we just discussed, the optimal value of (CP) is the maximum of $J_\pi(\beta)$ over all policies π, thus (CD) and (CP) satisfy strong

duality.

3.5.2 Proof of Theorem 3.16

Since x and y are complementary, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$r(s,a)x(s,a) = \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t) \right) x(s,a).$$

By summing up both sides for $s = 1, 2, \ldots, N$ and $a = 1, 2, \ldots, A$,

$$\begin{aligned}
&\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) \\
&\quad = \sum_{s=1}^{N}\sum_{a=1}^{A} \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t) \right) x(s,a) \\
&\quad = \sum_{s=1}^{N}\sum_{a=1}^{A} y(s)x(s,a) - \alpha \sum_{s=1}^{N}\sum_{a=1}^{A}\sum_{t=1}^{\infty} p(t|s,a)y(t)x(s,a) \\
&\quad = \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) - \sum_{t=1}^{\infty} y(t)\alpha \sum_{s=1}^{N}\sum_{a=1}^{A} p(t|s,a)x(s,a) \\
&\quad = \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) - \sum_{t=1}^{\infty} y(t) \left(\sum_{a=1}^{A} x(t,a) - \beta(t) - \alpha \sum_{s=N+1}^{\infty}\sum_{a=1}^{A} p(t|s,a)x(s,a) \right),
\end{aligned} \tag{3.33}$$

where the exchange of sums in the third equality is justified by the fact that the sum $\sum_{t=1}^{\infty} p(t|s,a)y(t)$ is finite for any s and a, and the last equality is obtained from feasibility of x to (CP). We will find the limit of (3.33) as $N \to \infty$.

We use the fact that $\|y\|_w$ is finite to observe the following:

$$\sum_{s=1}^{\infty}\sum_{a=1}^{A} |y(s)x(s,a)| = \sum_{s=1}^{\infty} |y(s)| \sum_{a=1}^{A} x(s,a) \le \|y\|_w \sum_{s=1}^{\infty} w(s) \sum_{a=1}^{A} x(s,a),$$

and since x is feasible to (CP), there exists a stationary policy σ such that (here we consider

the absorbing MDP formulation introduced in Section 3.5.1)

$$\begin{aligned}\|y\|_w \sum_{s=1}^{\infty} w(s) \sum_{a=1}^{A} x(s,a) &= \|y\|_w \sum_{s=1}^{\infty} w(s) \sum_{a=1}^{A} Q_\sigma^\beta(s,a) \\ &= \|y\|_w \sum_{s=1}^{\infty} \sum_{a=1}^{A} \sum_{n=1}^{\infty} P_\sigma^\beta(S_n = s, A_n = a) w(s) \\ &= \|y\|_w \sum_{s=1}^{\infty} \sum_{n=1}^{\infty} \sum_{a=1}^{A} P_\sigma^\beta(S_n = s, A_n = a) w(s) \\ &= \|y\|_w \sum_{s=1}^{\infty} \sum_{n=1}^{\infty} P_\sigma^\beta(S_n = s) w(s), \end{aligned} \tag{3.34}$$

where the third equality is obtained from Proposition 3.5. However,

$$\begin{aligned}&\sum_{n=1}^{\infty} \sum_{s=1}^{\infty} P_\sigma^\beta(S_n = s) w(s) = \beta^T (w + \alpha P_\sigma w + \alpha^2 P_\sigma^2 w + \ldots) \\ &\leq \beta^T [(w + (\alpha\kappa)w + (\alpha\kappa)^2 w + \ldots + (\alpha\kappa)^{J-1} w) + (\lambda w + \lambda(\alpha\kappa)w + \ldots + \lambda(\alpha\kappa)^{J-1} w) + \ldots] \\ &= L\beta^T w < \infty\end{aligned}$$

by Assumptions A2 and A3, and (3.7). Thus, the sum (3.34) is finite by Proposition 3.5. Therefore, we have

$$\sum_{s=1}^{\infty} y(s) \sum_{a=1}^{A} x(s,a) < \infty. \tag{3.35}$$

We will also prove that

$$\sum_{t=1}^{\infty} y(t) \sum_{s=N+1}^{\infty} \sum_{a=1}^{A} p(t|s,a) x(s,a) < \infty \tag{3.36}$$

and that the above sum tends to zero as $N \to \infty$. We first show that the following sum is

finite:

$$\begin{aligned}\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\left|\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a)\right| &\leq \sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\sum_{a=1}^{A}|y(t)|\,p(t|s,a)x(s,a)\\ &= \sum_{s=1}^{\infty}\sum_{a=1}^{A}x(s,a)\sum_{t=1}^{\infty}p(t|s,a)\,|y(t)| \leq \sum_{s=1}^{\infty}\sum_{a=1}^{A}x(s,a)\|y\|_w\sum_{t=1}^{\infty}p(t|s,a)w(t)\\ &\leq \kappa\|y\|_w\sum_{s=1}^{\infty}\sum_{a=1}^{A}w(s)x(s,a) < \infty,\end{aligned}$$

where the interchange of sums in the equality follows by Proposition 3.5, the second inequality by $y \in Y_w$, the third inequality by Assumption A2, and the last infinite sum is finite as shown before. Then, by Proposition 3.6,

$$\sum_{s=1}^{\infty}\sum_{t=1}^{\infty}\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a) = \sum_{t=1}^{\infty}\sum_{s=1}^{\infty}\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a) < \infty.$$

Therefore,

$$\sum_{s=N+1}^{\infty}\sum_{t=1}^{\infty}\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a) \to 0 \text{ as } N \to \infty.$$

Using the same arguments, we can also show that, for any N,

$$\sum_{s=N+1}^{\infty}\sum_{t=1}^{\infty}\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a) = \sum_{t=1}^{\infty}\sum_{s=N+1}^{\infty}\sum_{a=1}^{A} y(t)p(t|s,a)x(s,a) < \infty.$$

Therefore, (3.36) is proven and its left hand side converges to zero as $N \to \infty$. Also, we know

$$\sum_{t=1}^{\infty}\beta(t)|y(t)| \leq \|y\|_w\sum_{t=1}^{\infty}\beta(t)w(t) < \infty. \tag{3.37}$$

Then, by (3.35), (3.36), and (3.37), we can write (3.33) as

$$\begin{aligned}
&\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) \\
&= \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) - \sum_{t=1}^{\infty} y(t) \sum_{a=1}^{A} x(t,a) + \sum_{t=1}^{\infty} \beta(t)y(t) \\
&\quad + \alpha \sum_{t=1}^{\infty} y(t) \sum_{s=N+1}^{\infty} \sum_{a=1}^{A} p(t|s,a)x(s,a).
\end{aligned}$$

By letting $N \to \infty$ on both sides, we obtain

$$\sum_{s=1}^{\infty}\sum_{a=1}^{A} r(s,a)x(s,a) = \sum_{t=1}^{\infty} \beta(t)y(t),$$

and thus, the theorem is proven. □

3.5.3 Proof of Theorem 3.17

Since y and x are feasible to (CD) and (CP), respectively, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$x(s,a)\left[r(s,a) - \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] \leq 0.$$

By summing up the above for $s = 1, 2, \ldots, N$ and $a = 1, 2, \ldots, A$, we obtain

$$\begin{aligned}
0 \geq &\sum_{s=1}^{N}\sum_{a=1}^{A} x(s,a)\left[r(s,a) - \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] \\
= &\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) - \sum_{s=1}^{N}\sum_{a=1}^{A} \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t)\right) x(s,a) \\
= &\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) - \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) + \sum_{t=1}^{\infty} y(t) \sum_{a=1}^{A} x(t,a) - \sum_{t=1}^{\infty} \beta(t)y(t) \\
&- \alpha \sum_{t=1}^{\infty} y(t) \sum_{s=N+1}^{\infty} \sum_{a=1}^{A} p(t|s,a)x(s,a), \qquad (3.38)
\end{aligned}$$

where the last equality is obtained similarly to the proof of Theorem 3.16. Note that by strong duality we have

$$\sum_{s=1}^{\infty}\sum_{a=1}^{A} r(s,a)x(s,a) = \sum_{t=1}^{\infty}\beta(t)y(t).$$

Therefore, by letting $N \to \infty$ in (3.38) and using arguments similar to the proof of Theorem 3.16, we obtain that

$$0 \geq \sum_{s=1}^{\infty}\sum_{a=1}^{A} x(s,a)\left[r(s,a) - \left(y(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] = 0,$$

and thus, for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$x(s,a)\left[r(s,a) - \left(y(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] = 0,$$

i.e., y and x are complementary. □

3.5.4 Proof of Lemma 3.18

For any $x \in \mathcal{F}$, there exists a stationary policy σ such that $x(s,a) = Q_\sigma^\beta(s,a)$ for $s \in \mathcal{S}, a \in \mathcal{A}$. By (3.29), it suffices to show (here we consider the absorbing MDP formulation introduced in Section 3.5.1)

$$\sum_{s=1}^{\infty}\sum_{a=1}^{A}\sum_{n=1}^{\infty} P_\sigma^\beta\{S_n = s, A_n = a\} = \frac{1}{1-\alpha}.$$

Using Proposition 3.5 to interchange the sums, we have:

$$\sum_{n=1}^{\infty}\sum_{s=1}^{\infty}\sum_{a=1}^{A} P_\sigma^\beta\{S_n = s, A_n = a\} = \sum_{n=1}^{\infty}\sum_{s=1}^{\infty} P_\sigma^\beta\{S_n = s\} = \sum_{n=1}^{\infty}\alpha^{n-1} = \frac{1}{1-\alpha}.$$

□

3.5.5 Proof of Lemma 3.21

We prove this lemma for a more general case of an arbitrary stationary policy σ (rather than just stationary deterministic policy). For $s = 1, \ldots, N$,

$$\begin{aligned} y^N(s) =& r_\sigma(s) + \alpha \sum_{t_1=1}^{N} P_\sigma(t_1|s) y^N(t_1) \\ =& r_\sigma(s) + \alpha \sum_{t_1=1}^{N} P_\sigma(t_1|s) r_\sigma(t_1) + \alpha^2 \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) \\ &+ \alpha^3 \sum_{t_1=1}^{N} \sum_{t_2=1}^{N} \sum_{t_3=1}^{N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) + \ldots . \end{aligned}$$

On the other hand, for $s = 1, \ldots, N$,

$$\begin{aligned} y(s) =& r_\sigma(s) + \alpha \sum_{t_1=1}^{\infty} P_\sigma(t_1|s) r_\sigma(t_1) + \alpha^2 \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) \\ &+ \alpha^3 \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) + \ldots . \end{aligned}$$

Then, for $s = 1, \ldots, N$,

$$
\begin{aligned}
|y(s) - y^N(s)| = \Bigg| & \alpha \sum_{t_1 > N} P_\sigma(t_1|s) r_\sigma(t_1) \\
& + \alpha^2 \left[\sum_{t_1 > N} \sum_{t_2 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) + \sum_{t_1 \le N} \sum_{t_2 > N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) \right] \\
& + \alpha^3 \Bigg[\sum_{t_1 > N} \sum_{t_2 = 1}^{\infty} \sum_{t_3 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) \\
& \quad + \sum_{t_1 \le N} \sum_{t_2 > N} \sum_{t_3 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) \\
& \quad + \sum_{t_1 \le N} \sum_{t_2 \le N} \sum_{t_3 > N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) \Bigg] + \ldots \Bigg| \\
\le \; & \alpha \sum_{t_1 > N} P_\sigma(t_1|s) w(t_1) \\
& + \alpha^2 \left[\sum_{t_1 > N} \sum_{t_2 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) + \sum_{t_1 \le N} \sum_{t_2 > N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) \right] \\
& + \alpha^3 \Bigg[\sum_{t_1 > N} \sum_{t_2 = 1}^{\infty} \sum_{t_3 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \\
& \quad + \sum_{t_1 \le N} \sum_{t_2 > N} \sum_{t_3 = 1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \\
& \quad + \sum_{t_1 \le N} \sum_{t_2 \le N} \sum_{t_3 > N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \Bigg] + \ldots
\end{aligned}
\tag{3.39}
$$

by Assumption A1. Note that the terms of the infinite sum on the right hand side of (3.39) can be reordered by Proposition 3.5. In particular, consider rearranging the terms in (3.39)

as follows:

$$\begin{aligned}
&\Bigg[\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1>N}\sum_{t_2=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) \\
&\quad + \alpha^3 \sum_{t_1>N}\sum_{t_2=1}^{\infty}\sum_{t_3=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots\Bigg] \\
&+ \Bigg[\alpha^2 \sum_{t_1\le N}\sum_{t_2>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) + \alpha^3 \sum_{t_1\le N}\sum_{t_2>N}\sum_{t_3=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots\Bigg] \\
&+ \Bigg[\alpha^3 \sum_{t_1\le N}\sum_{t_2\le N}\sum_{t_3>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) \\
&\quad +\alpha^4 \sum_{t_1\le N}\sum_{t_2\le N}\sum_{t_3>N}\sum_{t_4=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)P_\sigma(t_4|t_3)w(t_4) + \dots\Bigg] \\
&+ \dots. \qquad (3.40)
\end{aligned}$$

First, let us compute an upper bound on the first bracket of (3.40) by considering groups of J terms, and establishing bounds using A2 and A3:

$$\begin{aligned}
&\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1>N}\sum_{t_2=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) \\
&\quad + \alpha^3 \sum_{t_1>N}\sum_{t_2=1}^{\infty}\sum_{t_3=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots \\
&\le \left[\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1>N} P_\sigma(t_1|s)\kappa w(t_1) + \dots + \alpha^J \sum_{t_1>N} P_\sigma(t_1|s)\kappa^{J-1}w(t_1)\right] \\
&\quad + \left[\alpha \sum_{t_1>N} P_\sigma(t_1|s)\lambda w(t_1) + \alpha^2 \sum_{t_1>N} P_\sigma(t_1|s)\lambda\kappa w(t_1) + \dots + \alpha^J \sum_{t_1>N} P_\sigma(t_1|s)\lambda\kappa^{J-1}w(t_1)\right] + \dots \\
&= \alpha\frac{1}{1-\lambda}[1 + (\alpha\kappa) + \dots + (\alpha\kappa)^{J-1}] \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) = L\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1).
\end{aligned}$$

Applying similar arguments to the other terms of (3.40), we obtain the following upper bound:

$$
\begin{aligned}
L \Bigg[& \alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1\le N} \sum_{t_2>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) \\
& + \alpha^3 \sum_{t_1\le N} \sum_{t_2\le N} \sum_{t_3>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots \Bigg] \\
\le L \Bigg[& \alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1=1}^{\infty} \sum_{t_2>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) \\
& + \alpha^3 \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} \sum_{t_3>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots \Bigg] \\
= L \Bigg[& \alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_2>N} \sum_{t_1=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) \\
& + \alpha^3 \sum_{t_3>N} \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots \Bigg] \\
= L & \sum_{n=1}^{\infty} \sum_{t>N} \alpha^n P_\sigma^n(t|s)w(t). \qquad (3.41)
\end{aligned}
$$

by Proposition 3.5. Combining these, we obtain

$$
|y(s) - y^N(s)| \le L \sum_{n=1}^{\infty} \sum_{t>N} \alpha^n P_\sigma^n(t|s)w(t) = L \sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s)w(t), \qquad (3.42)
$$

again by Proposition 3.5. The right hand side of (3.42) converges to zero as $N \to \infty$ because

$$
\sum_{n=1}^{\infty} \sum_{t=1}^{\infty} \alpha^n P_\sigma^n(t|s)w(t) \le Lw(s),
$$

which can be shown by using Assumptions A2 and A3 following already familiar steps. Thus, the lemma is proven. □

3.5.6 Proof of Lemma 3.28

In order to prove this lemma, we introduce a new interpretation of the approximate reduced cost $\gamma^{k,N}$. As explained before, $\gamma^{k,N}$ is the reduced cost (defined in (3.16)) of policy σ^k for the N-state truncation of the original MDP, obtained by replacing states bigger than N with an absorbing state where no reward is earned. We can extend the N-state truncation into a countable-state MDP by adding artificial states that have zero initial probabilities and are never reached. It is easy to prove that the countable-state version of the N-state truncation satisfies all the assumptions in Section 3.1.2. Then, $y^{k,N}$ and $\gamma^{k,N}$ are the *exact* value function and the *exact* reduced cost of policy σ^k in the new countable-state MDP, respectively. Therefore, $y^{k,N}$ also satisfies $|y^{k,N}(s)| \leq Lw(s)$ for $s = 1, \ldots, N$.

Thus, for $s = 1, \ldots, N$ and $a \in \mathcal{A}$,

$$\begin{aligned}
|\gamma^{k,N}(s,a)| &= \left| r(s,a) + \alpha \sum_{t=1}^{N} p(t|s,a) y^{k,N}(t) - y^{k,N}(s) \right| \\
&\leq |r(s,a)| + \alpha \sum_{t=1}^{N} p(t|s,a) \, |y^{k,N}(t)| + |y^{k,N}(s)| \\
&\leq w(s) + \alpha \sum_{t=1}^{N} p(t|s,a) L w(t) + L w(s) \\
&\leq w(s) + \alpha\kappa L w(s) + L w(s) = [1 + (1+\alpha\kappa)L] w(s).
\end{aligned}$$

Suppose that x^k is not optimal to (CP). Then, y^k must not be feasible to (CD), so there exists a state-action pair $(\hat{s}, \hat{a})$ such that $\gamma^k(\hat{s}, \hat{a}) = \epsilon > 0$. Since we have $\lim_{N\to\infty} \gamma^{k,N}(\hat{s}, \hat{a}) = \gamma^k(\hat{s}, \hat{a}) = \epsilon$, there exists $N_1 \geq \hat{s}$ such that for $N \geq N_1$, $\gamma^{k,N}(\hat{s}, \hat{a}) \geq \frac{3}{4}\epsilon$.

Since $\sum_{s=1}^{\infty} \beta(s) w(s)$ is finite, we know $\lim_{s\to\infty} \beta(s) w(s) = 0$. Thus, there exists $s_1 > \hat{s}$ such that for $s \geq s_1$, $[1 + (1+\alpha\kappa)L]\beta(s)w(s) < \frac{3}{4}\beta(\hat{s})\epsilon$. Then for $N \geq \max\{N_1, s_1\}$, $s \in \mathcal{S}$ such that $s_1 \leq s \leq N$, and $a \in \mathcal{A}$,

$$\beta(\hat{s})\gamma^{k,N}(\hat{s}, \hat{a}) \geq \frac{3}{4}\beta(\hat{s})\epsilon > [1 + (1+\alpha\kappa)L]\beta(s)w(s) \geq \beta(s)\gamma^{k,N}(s,a).$$

That is, for $N \geq \max\{N_1, s_1\}$, state-action pair (s, a) such that $s_1 \leq s \leq N$ cannot achieve the maximum in Step 2(d) of the simplex algorithm. Thus, for $N \geq \max\{N_1, s_1\}$, we can limit our attention to the state-action pairs (s, a) such that $s < s_1$ to find the maximum in Step 2(d) of the simplex algorithm.

Let C be the set of state-action pairs (s, a) such that $s < s_1$ and $\gamma^k(s, a) > 0$. Note that C is a finite set and $(\hat{s}, \hat{a}) \in C$. Since $\lim_{N\to\infty} \bar{\delta}(s, a, N) = 0$ and $\lim_{N\to\infty} \gamma^{k,N}(s, a) = \gamma^k(s, a)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exists N_2 such that for $N \geq N_2$, any $(s, a) \in C$ satisfies $\gamma^{k,N}(s, a) \geq \frac{1}{2}\gamma^k(s, a) > \bar{\delta}(s, a, N)$. There also exists N_3 such that for $N \geq N_3$, any $(s', a') \notin C$ such that $s' < s_1$ satisfies $\beta(s')\gamma^{k,N}(s', a') < \min_{(s,a)\in C} \frac{1}{2}\beta(s)\gamma^k(s, a)$. Then, for $N \geq \max\{N_1, N_2, N_3, s_1\}$ and for any $(s, a) \in C$ and any $(s', a') \notin C$ such that $s' < s_1$, we have $\beta(s)\gamma^{k,N}(s, a) > \frac{1}{2}\beta(s)\gamma^k(s, a) > \beta(s')\gamma^{k,N}(s', a')$, i.e., $\beta(s)\gamma^{k,N}(s, a) > \beta(s')\gamma^{k,N}(s', a')$. Thus, for $N \geq \max\{N_1, N_2, N_3, s_1\}$, the maximum in Step 2(d) of the algorithm is achieved by an element of C and the inequality $\gamma^{k,N}(s, a) > \bar{\delta}(s, a, N)$ is satisfied for any $(s, a) \in C$. Therefore, the Step 2 terminates with some $N \geq \max\{N_1, N_2, N_3, s_1\}$.

Now suppose that x^k is optimal for (CP). Then y^k is feasible to (CD), so $\gamma^k(s, a) \leq 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Suppose that the Step 2 terminates. Then, (s^k, a^k) satisfies $\gamma^{k,N}(s^k, a^k) > \bar{\delta}(s^k, a^k, N)$. However, by Lemma 3.22, we have $\gamma^k(s^k, a^k) \geq \gamma^{k,N}(s^k, a^k) - \delta(\sigma^k, s^k, a^k, N) \geq \gamma^{k,N}(s^k, a^k) - \bar{\delta}(s^k, a^k, N) > 0$, which is a contradiction. □

3.5.7 Example 3.2 (continued)

Let us first show that $\bar{\delta}(s, a, N)$ is an upper bound of $\delta(\sigma, s, a, N)$ for any $\sigma \in \Pi_{SD}$.

The first term in (3.19) is bounded as follows:

$$
\begin{aligned}
L\sum_{t>N}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(t|s)w(t) &= L\sum_{u>N-s}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(s+u|s)w(s+u)\\
&\le L\sum_{u>N-s}\sum_{n=1}^{\infty}\alpha^n P\{X_n=u\}(Cs+Cu+D)\\
&= CL\sum_{u>N-s}\sum_{n=1}^{\infty}\alpha^n\frac{e^{na^1_{\max}}(na^1_{\max})^u}{u!}u + L(Cs+D)\sum_{u>N-s}\sum_{n=1}^{\infty}\alpha^n\frac{e^{na^1_{\max}}(na^1_{\max})^u}{u!}\\
&= L\left[\frac{\alpha C}{1-\alpha}\sum_{u>N-s}\sum_{n=1}^{\infty}(1-\alpha)\alpha^{n-1}\frac{e^{na^1_{\max}}(na^1_{\max})^u}{u!}u\right.\\
&\quad\left.+\frac{\alpha}{1-\alpha}(Cs+D)\sum_{u>N-s}\sum_{n=1}^{\infty}(1-\alpha)\alpha^{n-1}\frac{e^{na^1_{\max}}(na^1_{\max})^u}{u!}\right]\\
&= \frac{\alpha L}{1-\alpha}\left[C\left(\mu-\sum_{u=0}^{N-s}uf_Y(u)\right)+(Cs+D)(1-F_Y(N-s))\right]\\
&= Lh_1(s,N),
\end{aligned}
$$

where the first equality is a change of variable $u \triangleq t-s$, the inequality follows by assuming the maximum arrival rate ($a^1_{\max}$) and zero service rate, and the following equalities follow by the definitions of X_n, Y, f_Y, and F_Y.

Similarly, the second term in (3.19) is bounded as follows:

$$
\alpha L\sum_{t=0}^{N}p(t|s,a)\sum_{t'>N}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(t'|t)w(t') \le \alpha L\sum_{t=0}^{N}p(t|s,a)g(t,N). \tag{3.43}
$$

The last term in (3.19) is also bounded as follows:

$$\begin{aligned}
\alpha L \sum_{t>N} p(t|s,a)w(t) &= \alpha L \sum_{u>N-s} p(s+u|s,a)(Cs+Cu+D) \\
&\leq \alpha L \sum_{u>N-s} P\{X_1 = u\}(Cs+Cu+D) \\
&= \alpha L \left[C \sum_{u>N-s} P\{X_1 = u\}u + (Cs+D) \sum_{u>N-s} P\{X_1 = u\} \right] \\
&= \alpha L \left[C \left(a^1_{\max} - \sum_{u=0}^{N-s} u f_{X_1}(u) \right) + (Cs+D)(1 - F_{X_1}(N-s)) \right] \\
&= \alpha L h_2(s,N)
\end{aligned}$$

by using similar arguments. Therefore, we showed that $\bar{\delta}(s,a,N) \geq \delta(\sigma,s,a,N)$.

Now we show that $\bar{\delta}(s,a,N) \to 0$ as $N \to \infty$. Since the expectations of Y and X_1 are finite, it is clear that $h_1(s,N)$ and $h_2(s,N)$ converge to zero as $N \to \infty$. Thus, it suffices to prove that the second term of $\bar{\delta}(s,a,N)$ in (3.28) converges to zero as $N \to \infty$. This term can be written as follows:

$$\begin{aligned}
\alpha L \sum_{t=0}^{N} p(t|s,a)g(t,N) =& \frac{\alpha^2 CL}{1-\alpha} \left(\mu \sum_{t=0}^{N} p(t|s,a) - \sum_{t=0}^{N} p(t|s,a) \sum_{u=0}^{N-s} u f_Y(u) \right) \\
&+ \frac{\alpha^2 L(Cs+D)}{1-\alpha} \left(\sum_{t=0}^{N} p(t|s,a) - \sum_{t=0}^{N} p(t|s,a) F_Y(N-s) \right). \quad (3.44)
\end{aligned}$$

As $N \to \infty$, $\mu \sum_{t=0}^{N} p(t|s,a)$ converges to μ. Also, as $N \to \infty$, $\sum_{t=0}^{N} p(t|s,a) \sum_{u=0}^{N-s} u f_Y(u)$ converges to μ as well. Therefore, the first big parenthesis in (3.44) converges to zero as $N \to \infty$. We can also similarly show that the second big parenthesis in (3.44) converges to zero. Therefore, we proved that the second term of $\bar{\delta}(s,a,N)$ in (3.28) converges to zero as $N \to \infty$, and thus, $\bar{\delta}(s,a,N) \to 0$ as $N \to \infty$.

Lastly, we illustrate that we can compute $\bar{\delta}(s,a,N)$ finitely. Clearly, we can compute $h_2(s,N)$ finitely. To show that $h_1(s,N)$ can be computed finitely, we only have to show that

$F_Y(U)$ can be computed finitely for any nonnegative integer U. For any U,

$$\begin{aligned} F_Y(U) &= \sum_{u=0}^{U} P\{Y=u\} = \sum_{u=0}^{U}\sum_{n=1}^{\infty}(1-\alpha)\alpha^{n-1}\frac{e^{-na^1_{\max}}(na^1_{\max})^u}{u!} \\ &= \sum_{u=0}^{U}\frac{(1-\alpha)(a^1_{\max})^u}{\alpha(u!)}\sum_{n=1}^{\infty} n^u(\alpha e^{-a^1_{\max}})^n. \end{aligned}$$

Thus, in order to show that $F_Y(U)$ can be computed finitely, it suffices to show that $B_u \triangleq \sum_{n=1}^{\infty} n^u\zeta^n$ can be computed finitely for any $\zeta \in (0,1)$ and nonnegative integer u. Clearly, B_0 can be computed finitely. Suppose that $B_0, B_1, \ldots, B_{u-1}$ can be computed finitely. Then,

$$\begin{aligned} (1-\zeta)B_u &= \sum_{n=1}^{\infty} n^u\zeta^n - \sum_{n=1}^{\infty} n^u\zeta^{n+1} = \sum_{n=1}^{\infty} n^u\zeta^n - \sum_{n=1}^{\infty}(n-1)^u\zeta^n = \sum_{n=1}^{\infty}[n^u-(n-1)^u]\zeta^n \\ &= \sum_{n=1}^{\infty}\left(\sum_{l=0}^{u-1}\binom{u}{l}n^l(-1)^{u-l+1}\right)\zeta^n = \sum_{l=0}^{u-1}\binom{u}{l}(-1)^{u-l+1}\sum_{n=1}^{\infty} n^l\zeta^n \\ &= \sum_{l=0}^{u-1}\binom{u}{l}(-1)^{u-l+1}B_l, \end{aligned}$$

where the sum exchange is justified by the fact that B_l is finite for $l = 0, 1, \ldots, u-1$. Thus, B_u can be computed finitely. By induction, B_u can be computed finitely for $\zeta \in (0,1)$ and any nonnegative integer u, and therefore, $h_1(s,N)$ can be computed finitely. This implies that we can compute $\bar{\delta}(s,a,N)$ finitely.

CHAPTER IV

A Linear Programming Approach to Constrained Non-stationary Markov Decision Processes

4.1 Introduction

For the last couple of decades, growing attention has been given to solving constrained Markov decision processes (MDPs). Constrained MDPs are MDPs optimizing an objective function while satisfying constraints, typically on budget, quality, etc. In addition, decision making problems with multiple criteria are often approached by optimizing one criterion while satisfying constraints on the other criteria, which also turns into a constrained MDP. One setting where such problems often arise is data communications. In queueing systems with service rate control, the average throughput is maximized with constraints on the average delay [30, 34]. Priority queueing systems with a fixed service rate are another example [5, 37, 44]. Here, one optimizes the queueing time of non-interactive traffic while satisfying a constraint on the average end-to-end delay of interactive traffic. For these problems, [49] considered a case where service rate costs and penalty costs of delay are actually incurred in discrete time periods and it is desired to minimize the discounted service rate cost with constraints on the discounted delay cost. Facility maintenance is another type of problems modeled by constrained MDPs. Examples are finding an optimal maintenance policy for each mile of a network of highways [25] and a problem in building management

[61]. In the models for these problems, the total cost is minimized subject to constraints on quality of facilities.

In this chapter, we consider constrained MDPs that have a finite state space, a finite action space, and non-stationary transition probabilities and reward function, which we call constrained non-stationary MDPs. Specifically, the state set $\mathcal{S}$ and the action set $\mathcal{A}$ are both finite and given that action a is taken at state s in period n, multiple kinds of costs, denoted by $c_n(s,a)$ and $d_n^k(s,a)$ for $k = 1,\ldots,K$, are incurred, and the system makes a transition to be observed in a state t at the beginning of period $n+1$ with probability $p_n(t|s,a)$. The costs are assumed to be nonnegative and uniformly bounded, i.e., there exist c and d^k for $k = 1,\ldots,K$ such that $0 \le c_n(s,a) \le c$, $0 \le d_n^k(s,a) \le d^k$ for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$, and $k = 1,\ldots,K$. The goal is to minimize the expected total discounted "c-cost" satisfying K constraints on the expected total discounted "d^k-costs" for $k = 1,\ldots,K$, with a common discount factor $0 < \alpha < 1$. Let

$$C_\pi(\beta) \triangleq E_\beta^\pi \left[\sum_{n=1}^{\infty} \alpha^{n-1} c_n(S_n, A_n) \right],$$
$$D_\pi^k(\beta) \triangleq E_\beta^\pi \left[\sum_{n=1}^{\infty} \alpha^{n-1} d_n^k(S_n, A_n) \right] \text{ for } k = 1,\ldots,K,$$

and let $\Pi_f \triangleq \{\pi \,|\, D_\pi^k(\beta) \le V_k \text{ for } k = 1,\ldots,K\}$. The optimization problem can then be written as

$$\text{(Q)} \quad \min_{\pi \in \Pi_f} C_\pi(\beta).$$

In [19] it was shown that an optimal policy for a constrained MDP may depend on the initial state; more generally, we formulate (Q) with a fixed initial state distribution β.

In this chapter, we study CILP formulations of the problem (Q), constrained non-stationary MDPs. As mentioned in Introduction, duality results and the algebraic characterization of extreme points as basic feasible solutions do not extend to CILPs in general [22]. After introducing primal and dual CILP formulations, we introduce the duality re-

sults proven in [4], define complementary slackness, and establish its relation to optimality. We provide algebraic necessary conditions for a feasible solution of the CILP formulation of a constrained non-stationary MDP to be an extreme point of its feasible region. Using those necessary conditions, we also establish a necessary and sufficient condition for a feasible solution to be an extreme point, which can be checked by considering a familiar finite dimensional polyhedron. This yields a complete algebraic characterization of extreme points for CILPs representing constrained non-stationary MDPs. Thus, this chapter sets important foundations towards the development of a simplex-type algorithm for solving constrained non-stationary MDPs.

Under typical settings for constrained MDPs, there exists a stationary optimal policy but a deterministic stationary optimal policy may not exist [19]. Thus, an often pursued goal in literature is to prove existence of an optimal policy that is as close to deterministic as possible. In particular, this means that we are interested in the existence of a K-randomized optimal policy, where K is the number of constraints and a policy is K-randomized if it uses K "more" actions than a deterministic stationary policy (for a more precise definition, see Section 4.4). It is well-known that extreme points of LP formulations of unconstrained MDPs with a finite number of states correspond to deterministic policies. Now consider a constrained MDP obtained by adding linear constraints to an unconstrained MDP. Then an extreme point of the LP formulation of the constrained MDP is a convex combination of extreme points of the LP formulation of the unconstrained MDP, i.e., deterministic policies, and this explains how randomization is introduced. The existence of $(K+1)$-randomized optimal policy was shown for constrained MDPs with Borel state space and stationary problem data in [26] using the Carathéodory's theorem. For constrained MDPs with finite state space, there exists a K-randomized optimal policy and it can be found by obtaining an optimal basic feasible solution of the corresponding finite LP formulation [29, 32, 43]. For constrained MDPs with a countably infinite number of states, a K-randomized optimal policy is proven to exist for the single constraint case using the Lagrangian multiplier approach in [49] and for the

general case in [17] by studying the Pareto frontier of the performance set. In this chapter, we obtain the existence of a K-randomized optimal policy for constrained non-stationary MDPs as a byproduct of characterizing extreme points of the CILP formulation.

We conclude this section by discussing a set of policies to which we can limit our attention. Constrained non-stationary MDPs (which has finite state space) can be reformulated as a constrained stationary MDP with a countable number of states by appending the states $s \in \mathcal{S}$ with time-indices $n \in \mathbb{N}$. For constrained stationary MDPs, it was shown in [4] that, without loss of optimality, we can restrict our attention to Markov policies. In the stationary MDP counterpart of constrained non-stationary MDPs, a Markov policy is also stationary because each period-state pair is visited only once. Moreover, any stationary policy in the stationary MDP counterpart corresponds to a Markov policy in the original constrained non-stationary MDP, and thus, we can restrict our attention to Markov policies for constrained non-stationary MDPs.

4.2 CILP Formulations

It was proven that (Q) has an equivalent CILP formulation [3, 4], which can be written as:

$$\text{(CNP)} \ \min \ f(x) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) \tag{4.1}$$

$$\text{s.t.} \ \sum_{a\in\mathcal{A}} x_1(s,a) = \beta(s) \text{ for } s \in \mathcal{S} \tag{4.2}$$

$$\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{n-1}(s|s',a) x_{n-1}(s',a) = 0 \text{ for } n \geq 2, s \in \mathcal{S} \tag{4.3}$$

$$\sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \leq V_k \text{ for } k = 1, \ldots, K \tag{4.4}$$

$$x \geq 0. \tag{4.5}$$

Note that (CNP) is similar to (NP) in Section 2.2.1, but they are different in that (CNP)

has the side constraints (4.4) and the right hand side of (4.2) is $\beta(s)$ instead of 1. Let $\mathcal{P}$ denote the feasible region of (CNP). Constraints (4.2) and (4.3) imply that for any $x \in \mathcal{P}$,

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a) = 1 \text{ for } n \in \mathbb{N}.^1 \tag{4.6}$$

Since x is nonnegative, we have $0 \leq x_n(s,a) \leq 1$ for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$. Because all objective and constraint cost functions are uniformly bounded, the infinite sums in (4.1) and (4.4) exist.

Solutions of (CNP) can be interpreted as flows in a directed staged hypernetwork with infinite stages, in a way similar to the interpretation of (NP) in Section 2.2.1. Structure of the hypernetwork for (CNP) is exactly the same as the one for (NP) except the nodes $(1,s)$ have supply of $\beta(s)$ units for $s \in \mathcal{S}$, instead of 1 (due to the different right hand side of constraint (4.2)). Any x satisfying (4.2), (4.3), and (4.5) can be visualized as a flow in the hypernetwork, thus as we did in Section 2.2.1, we will refer to any x satisfying (4.2), (4.3), and (4.5) as a *flow* in the corresponding hypernetwork. This interpretation provides particularly helpful intuition for proofs in Section 4.4.2.

For any Markov policy π for the non-stationary MDP, the corresponding flow x can be found as

$$x_n(s,a) = \pi_n(a|s) \cdot P_\beta^\pi(S_n = s),\ n \in \mathbb{N},\ s \in \mathcal{S},\ a \in \mathcal{A},$$

i.e., $x_n(s,a)$ is proportional to the probability, under π, of using action a in state s in period n, scaled by the probability of reaching this state under the probability measure induced by π and β. Thus, $x_n(s,a)$ can also be interpreted as the probability of encountering hyperarc (n,s,a) under policy π for the given initial state distribution β, while the total inflow into node (n,s) is precisely $P_\beta^\pi(S_n = s)$. In light of this interpretation, a Markov policy corresponding to any flow x is also easy to identify, with the following caveat: for a given flow x, there may be some nodes (n,s) that receive no incoming flow, and thus have

[1] Note the difference from Lemma 2.1

$\sum_{a\in\mathcal{A}} x_n(s,a) = 0$ and no outgoing flow. For those nodes, we can define $\pi_n(s)$ arbitrarily, i.e., we do not distinguish between policies that have the same corresponding flows. Under this convention, there exists a one-to-one correspondence between the set of policies and the set of flows. There also exists an obvious one-to-one correspondence between $\mathcal{P}$ and Π_f. We refer to a (feasible) policy and the corresponding (feasible) flow interchangeably.

Finally, the quantity $\alpha^{n-1}x_n(s,a),\ n\in\mathbb{N},\ s\in\mathcal{S},\ a\in\mathcal{A}$ can be interpreted as the occupancy measure studied in [16]. The next result was shown to hold for a more general setting (e.g., Theorem 9 in [38]), but to make this paper self-contained, we provide the theorem and its proof.

Theorem 4.1. *If (CNP) is feasible, then it has an extreme point optimal solution.*

Proof: It is easy to show that $\mathcal{P}$ is a closed and convex subset of $\mathbb{R}^\infty$. By Tychonoff's product theorem (see [2]) and (4.6), $\mathcal{P}$ is a subset of a compact set and thus, it is compact. Since the objective function is continuous and convex, by Bauer's Maximum Principle (e.g., Theorem 7.69 of [2]), (CNP) has an extreme point optimal solution. □

[4] (see Section 9.5 and Theorem 9.11) defined dual of (CNP) as follows.

$$\text{(CND)}\ \max\ g(y,\mu) = \sum_{s\in\mathcal{S}} \beta(s)y_1(s) - \sum_{k=1}^{K} \mu_k V_k \tag{4.7}$$

$$\text{s.t. } y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s') - \alpha^{n-1}\sum_{k=1}^{K}\mu_k d_n^k(s,a) \le \alpha^{n-1}c_n(s,a)$$

$$\text{for } n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A} \tag{4.8}$$

$$\mu \ge 0 \tag{4.9}$$

$$y \in l_\infty, \tag{4.10}$$

where l_∞ denotes the set of sequences whose supremum norm is finite, i.e.,

$$\sup_{(n,s)\in\mathbb{N}\times\mathcal{S}} |y_n(s)| < \infty.$$

4.3 Duality Results

In this section, we present strong duality between (CNP) and (CND), define complementary slackness, and prove its relation to optimality. Throughout this section, we make the following assumption, which is known as Slater's condition.

Assumption 4.2. (Slater's condition) *(CNP) has a strict feasible solution, i.e., there exists a feasible solution x to (CNP) that satisfies all of the inequality constraints (4.4) strictly.*

Under this assumption, strong duality holds between (CNP) and (CND).

Theorem 4.3. (Theorem 9.11 of [4]) *Under Assumption 4.2, the optimal objective function values of (CNP) and (CND) coincide.*

We now define complementary slackness and show that feasible solutions of (CNP) and (CND) are optimal to their corresponding problems if and only if they satisfy the complementary slackness. Under Assumption 4.2, [4] showed necessity of complementary slackness for optimality for a more general class of MDPs by using an interpretation of constrained MDPs as an inf-sup problem with Lagrangian multipliers. For constrained non-stationary MDPs, we provide an alternative proof for necessity and establish sufficiency of complementary slackness for optimality.

Definition 4.4. *(Complementary slackness)* Suppose x is feasible to (CNP). Then we say that x and (y, μ) satisfy complementary slackness if

$$x_n(s,a)\left[\alpha^{n-1}\left(c_n(s,a)+\sum_{k=1}^{K} d_n^k(s,a)\mu_k\right) - y_n(s) + \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\right] = 0$$

$$\text{for } n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}, \tag{4.11}$$

$$\mu_k\left[d_k - \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a)x_n(s,a)\right] = 0 \text{ for } k = 1,2,\ldots,K. \tag{4.12}$$

Theorem 4.5. (Complementary slackness sufficiency) *Suppose that Assumption 4.2 holds and that x is feasible to (CNP) and complementary with some (y, μ). Then $f(x) = g(y, \mu)$. If (y, μ) is feasible to (CND), then x and (y, μ) are optimal to (CNP) and (CND), respectively.*

Proof: In Section 4.7.1.

Theorem 4.6. (Complementary slackness necessity) *Suppose that Assumption 4.2 holds, and x and (y, μ) are optimal to (CNP) and (CND), respectively. Then x and (y, μ) are complementary.*

Proof: In Section 4.7.2.

4.4 Splitting Randomized Policies

One of the main objectives in this chapter is to study extreme points of $\mathcal{P}$, the feasible region of (CNP). An extreme point of a convex set is defined as a point in the set that cannot be represented as a non-trivial convex combination of other points in the set. This section provides preliminary results in the form of two different representations of a randomized MDP policy as a convex combination of other policies (the results in this section are not limited to constrained problems). The first one, which describes a convenient characterization of representations of a randomized policy as a convex combination of deterministic policies, will be needed in Section 4.6 to derive a necessary and sufficient condition for a point in $\mathcal{P}$ to be an extreme point. The second one is needed in Section 4.5 to provide necessary conditions for an extreme point.

The following definitions will be helpful in describing the two representations. Following [17], we define a *submodel* of the MDP to be an MDP that is identical to the original one except that the action sets are limited to $B_n(s) \subseteq \mathcal{A}$ for $n \in \mathbb{N}, s \in \mathcal{S}$. For a given policy x, we define a *submodel defined by* x as a submodel such that $B_n(s) = \{a \in \mathcal{A} \,|\, x_n(s, a) > 0\}$ for $n \in \mathbb{N}, s \in \mathcal{S}$. We also say that a policy x *belongs to a submodel* if $\{a \in \mathcal{A} \,|\, x_n(s, a) > 0\} \subseteq B_n(s)$. We refer to the number $M = \sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} (|B_n(s)| - 1)$ as the *index of the submodel*.

A randomized policy that belongs to a submodel with index M can be interpreted as using at most M "more" actions than a deterministic policy. Since in each period of the original MDP there are S states, and each state has A action choices, in each period a policy can use up to $S(A-1)$ "more" actions than a deterministic policy.

Definition 4.7. A randomized policy that belongs to a submodel with index M is called an *M-randomized policy*. An M-randomized policy that does not belong to any submodel with index less than M is called an *exactly M-randomized policy*. A randomized policy that does not belong to any submodel with a finite index is called an *∞-randomized policy*.

4.4.1 Splitting into deterministic policies

The following result is well-known:

Lemma 4.8 (cf. Theorem 5.1 in [17])**.** *For any positive integer M, any exactly M-randomized policy is a convex combination of $M+1$ 0-randomized (i.e., deterministic) policies.*

In addition, it was recently shown in [16] that, for any positive integer M, it is possible to represent an M-randomized policy as a convex combination of $M+1$ deterministic policies that can be ordered so that each pair of consecutive policies differ at only one period-state pair, and an efficient algorithm to find such a convex combination of deterministic policies was introduced.

Consider an exactly M-randomized policy x for a positive integer M. Let B be the submodel defined by x. Since M is finite, the number of deterministic policies in the submodel B is also finite, say, N. Let $x^1, \ldots, x^N$ be these deterministic policies. Let

$$\Lambda(x) = \left\{ \lambda \in \mathbb{R}^N \,|\, x = \sum_{i=1}^{N} \lambda_i x^i,\ \sum_{i=1}^{N} \lambda_i = 1,\ \lambda \geq 0 \right\} \tag{4.13}$$

be the set of all weights of convex combinations of $x^1, \ldots, x^N$ that equal x (this set plays an important role in the analysis of Section 4.6.). Although in (4.13) $\Lambda(x)$ is defined as the

set of solutions of an infinite number of linear equations, the following theorem shows that a finite number is sufficient.

Theorem 4.9. *Let x be an exactly M-randomized policy and N be the number of deterministic policies in the submodel defined by x. Then there exists a matrix $A \in \mathbb{R}^{M \times N}$ and a vector $b \in \mathbb{R}^M$ such that $\Lambda(x) = \{\lambda \in \mathbb{R}^N \mid A\lambda = b, \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$, and matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ has full row rank.*

Proof: In Section 4.7.3.

Theorem 4.9 provides a way to construct $\Lambda(x)$ for a given exactly M-randomized policy x. Define $E(x)$ as the subset of $\Lambda(x)$ whose elements have at most $M+1$ nonzeros; then we can easily show $\Lambda(x) = \text{conv}E(x)$. Indeed, by Theorem 4.9, $\Lambda(x)$ is the set of feasible solutions of a standard form LP with $M+1$ constraints. Thus, $E(x)$ contains all extreme points of $\Lambda(x)$ and therefore, we have $\Lambda(x) = \text{conv}E(x)$. One can construct $E(x)$ by finding every representation of x as a convex combination of $M+1$ deterministic policies among $x^1, \ldots, x^N$, which can be done by applying the procedure described in the proof of Theorem 5.1 in [17] or Algorithm 1 in [16] in a straightforward way.

4.4.2 Splitting into "less" randomized policies

In this section, we introduce another representation of a randomized policy as a convex combination of "less randomized" policies, in two lemmas. The particular representation satisfies a set of properties which will help us establish necessary conditions for an extreme point of $\mathcal{P}$ in Section 4.5. Lemma 4.10 considers the case of an exactly M-randomized policy for a positive integer M.

Lemma 4.10. *Let M be a positive integer. Any exactly M-randomized policy x can be represented as a convex combination of $M+1$ $(M-1)$-randomized policies $x^1, x^2, \ldots, x^{M+1}$ such that the weights of the representation are uniquely determined (by the policies) and positive.*

Proof: We use induction on M.

Base case. For $M = 1$, let x be an exactly 1-randomized policy, randomizing in period-state pair (n, s) over actions a and b ($x_n(s, a) = \delta > 0$, $x_n(s, b) = \epsilon > 0$, and $x_n(s, a') = 0$ for $a' \in \mathcal{A} \setminus \{a, b\}$). We show that x is a convex combination of two 0-randomized (i.e., deterministic) policies, w and z. To construct them, we first define two *sub-flows*, u and v. Here, and in the rest of this section, the steps to define sub-flows are similar to the proof of Theorem 4.3 of [24], and we also borrowed their notation.

For $k \geq n+1$, since x does not randomize in those periods, let $a_k(s')$ for $s' \in \mathcal{S}$ denote the action chosen by x at (k, s'), i.e., $x_k(s', a_k(s')) > 0$. Let $\mathcal{S}_{n+1}(x) = \{s' \in \mathcal{S} \,|\, p_n(s'|s, a) > 0\}$. For $k \geq n + 2$, recursively define

$$\mathcal{S}_k(x) = \{s' \in \mathcal{S} \,|\, p_{k-1}(s'|\tilde{s}, a_{k-1}(\tilde{s})) > 0 \text{ for some } \tilde{s} \in \mathcal{S}_{k-1}(x)\}.$$

That is, $\mathcal{S}_k(x)$ is the set of states in period k that receive any portion of flow δ originating in hyperarc (n, s, a) under policy x. Let $\mathcal{F}(x)$ be the sub-hypernetwork formed by the node (n, s), the hyperarc (n, s, a), nodes in $\cup_{k=n+1}^{\infty} \mathcal{S}_k(x)$ and hyperarcs $\cup_{k=n+1}^{\infty} \{(k, s_k, a_k(s_k)) \,|\, s_k \in \mathcal{S}_k(x)\}$. We construct a sub-flow u in $\mathcal{F}(x)$ recursively in the following way. Node (n, s) is the only source node in the sub-hypernetwork, with supply of 1. Set $u_n(s, a) = 1$, and for each $s_{n+1} \in \mathcal{S}_{n+1}(x)$, set $u_{n+1}(s_{n+1}, a_{n+1}(s_{n+1})) = p_n(s_{n+1}|s, a)$. For $k \geq n + 2$ and for each $s_k \in \mathcal{S}_k(x)$,

$$u_k(s_k, a_k(s_k)) = \sum_{s_{k-1} \in \mathcal{S}_{k-1}(x)} p_{k-1}(s_k|s_{k-1}, a_{k-1}(s_{k-1}))u_{k-1}(s_{k-1}, a_{k-1}(s_{k-1})).$$

By construction, we can easily see that $x_n(s, a) = \delta u_n(s, a)$ and $x_k(s_k, a_k(s_k)) \geq \delta u_k(s_k, a_k(s_k))$ for any other hyperarc $(k, s_k, a_k(s_k))$ in $\mathcal{F}(x)$. To see this, note that for hyperarcs $(k, s_k, a_k(s_k))$ in $\mathcal{F}(x)$, $u_k(s_k, a_k(s_k))$ can be interpreted as the conditional probability of encountering hyperarc $(k, s_k, a_k(s_k))$ by following policy x, given that we encountered hyperarc (n, s, a). Fix a hyperarc $(k, s_k, a_k(s_k))$ in $\mathcal{F}(x)$. Let A be an event of encountering the hyperarc $(k, s_k, a_k(s_k))$

by following the policy x and let B be an event of encountering hyperarc (n, s, a) by following policy x. Then, $P(A|B) = u_k(s_k, a_k(s_k))$, $P(B) = \delta$, and $P(A) = x_k(s_k, a_k(s_k))$. Therefore, we have $x_k(s_k, a_k(s_k)) = P(A) \geq P(A \cap B) = P(A|B)P(B) = \delta u_k(s_k, a_k(s_k))$.

Similarly, for $k \geq n+1$, let $\mathcal{T}_k(x) \subset \mathcal{S}$ be the set of states in period k receiving any portion of flow ϵ in hyperarc (n, s, b) under policy x. For any $t_k \in \mathcal{T}_k(x)$, there exists a unique action $b_k(t_k)$ such that $x_k(t_k, b_k(t_k)) > 0$. Let $\mathcal{G}(x)$ be the sub-hypernetwork defined similarly to $\mathcal{F}(x)$, formed by the node (n, s), the hyperarc (n, s, b), nodes in $\cup_{k=n+1}^{\infty} \mathcal{T}_k(x)$ and hyperarcs in $\cup_{k=n+1}^{\infty} \{(k, t_k, b_k(t_k)) \mid t_k \in \mathcal{T}_k(x)\}$. We construct a sub-flow v in $\mathcal{G}(x)$ similarly to construction of u. Let the node (n, s) be a source of supply 1 and all other nodes in sub-hypernetwork $\mathcal{G}(x)$ have no supply. Set $v_n(s, b) = 1$, then for each $t_{n+1} \in \mathcal{T}_{n+1}(x)$, set $v_{n+1}(t_{n+1}, b_{n+1}(t_{n+1})) = p_n(t_{n+1}|s, b)$. For $k = n+2, n+3, \ldots$ and for each $t_k \in \mathcal{T}_k(x)$, set

$$v_k(t_k, b_k(t_k)) = \sum_{t_{k-1} \in \mathcal{T}_{k-1}(x)} p_{k-1}(t_k|t_{k-1}, b_{k-1}(t_{k-1})) v_{k-1}(t_{k-1}, b_{k-1}(t_{k-1})).$$

By construction, $x_n(s, b) = \epsilon v_n(s, b)$ and by using the same interpretation as u, we can easily check $x_k(t_k, b_k(t_k)) \geq \epsilon v_k(t_k, b_k(t_k))$ for any other hyperarc $(k, t_k, b_k(t_k))$ in $\mathcal{G}(x)$.

We construct a new flow w as follows.

$$w_k(s_k, a_k) = \begin{cases} x_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ not in } \mathcal{F}(x) \text{ or } \mathcal{G}(x) \\ x_k(s_k, a_k) - \delta u_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{F}(x) \setminus \mathcal{G}(x) \\ x_k(s_k, a_k) + \delta v_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{G}(x) \setminus \mathcal{F}(x) \\ x_k(s_k, a_k) + \delta(v_k(s_k, a_k) - u_k(s_k, a_k)) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{F}(x) \cap \mathcal{G}(x). \end{cases}$$

Since we have $x_k(s_k, a_k) \geq \delta u_k(s_k, a_k)$ for any hyperarc (k, s_k, a_k) in $\mathcal{F}(x)$, w is nonnegative. Note that w is obtained from x by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}(x)$. Thus, w satisfies the flow balance constraints and is 0-randomized.

z is constructed similarly, by redirecting flow ϵ from $\mathcal{G}(x)$ to $\mathcal{F}(x)$. More precisely,

$$z_k(s_k, a_k) = \begin{cases} x_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ not in } \mathcal{F}(x) \text{ or } \mathcal{G}(x) \\ x_k(s_k, a_k) + \epsilon u_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{F}(x) \setminus \mathcal{G}(x) \\ x_k(s_k, a_k) - \epsilon v_k(s_k, a_k) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{G}(x) \setminus \mathcal{F}(x) \\ x_k(s_k, a_k) + \epsilon(u_k(s_k, a_k) - v_k(s_k, a_k)) & \text{if } (k, s_k, a_k) \text{ in } \mathcal{F}(x) \cap \mathcal{G}(x). \end{cases}$$

By construction, z also satisfies the flow balance constraints and is 0-randomized. Moreover, $x = \frac{\delta z + \epsilon w}{\delta + \epsilon}$, i.e., x is a non-trivial convex combination of two 0-randomized flows. Note that the weights in the convex combination are both positive, and are uniquely determined (by w and z).

Inductive step. Suppose the statement holds for $M = M' - 1$. Let x be an exactly M'-randomized policy. There are finitely many period-state pairs at which x randomizes; among them, let (n, s) be the one with the largest period index (break ties arbitrarily). At (n, s), x randomizes over actions $a, b^1, \ldots, b^l$ for some $l \geq 1$; let $x_n(s, a) = \delta > 0$ and $x_n(s, b^i) = \epsilon_i > 0$ for $i = 1, \ldots, l$, with $\epsilon = \sum_{i=1}^{l} \epsilon_i$. We show that x is a convex combination of two $(M' - 1)$-randomized flows, denoted by w and z. To define w and z, we first introduce sub-flows u and v^i for $i = 1, \ldots, l$. Construction of u is the same as in the base case, and v^i, for $i = 1, \ldots, l$, is constructed in the same way as v in the base case except that the starting hyperarc is (n, s, b^i), with $\mathcal{G}^i(x)$ denoting the corresponding sub-hypernetwork. Then, we construct flow w from x by subtracting δu in $\mathcal{F}(x)$ and adding $(\delta\epsilon_i/\epsilon)v^i$ in $\mathcal{G}^i(x)$ for $i = 1, \ldots, l$ (i.e., by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s, maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s), and construct flow z from x by adding ϵu in $\mathcal{F}(x)$ and subtracting $\epsilon_i v^i$ in $\mathcal{G}^i(x)$ for $i = 1, \ldots, l$ (i.e., redirecting total flow ϵ from $\mathcal{G}^i(x)$'s to $\mathcal{F}(x)$). By construction, w and z are nonnegative and satisfy the flow balance constraints. Moreover, note that, except at (n, s), x does not have any randomization in either $\mathcal{F}(x)$ or $\mathcal{G}^i(x)$ for $i = 1, \ldots, l$. Therefore, w is exactly $(M' - 1)$-randomized and z is exactly $(M' - l)$-randomized. By construction,

$x = \frac{\delta z + \epsilon w}{\delta + \epsilon}$, i.e., x is a nontrivial convex combination of these two $(M'-1)$-randomized flows.

By the induction hypothesis, w can be represented as a convex combination of M' $(M'-2)$-randomized flows, say $w^1, \ldots, w^{M'}$, with unique positive weights $\lambda_1, \ldots, \lambda_{M'}$. Thus, x is a convex combination of z and $w^1, \ldots, w^{M'}$, i.e., $M'+1$ $(M'-1)$-randomized policies. Now we have to show that the representation of x as a convex combination of z and $w^1, \ldots, w^{M'}$ is unique and all of the weights are positive. Let

$$x = \lambda_z z + \sum_{i=1}^{M'} \lambda_i w^i, \tag{4.14}$$

where $\lambda_z \in [0,1]$, $\lambda_i \in [0,1]$ for $i = 1, \ldots, M'$, and $\lambda_z + \sum_{i=1}^{M'} \lambda_i = 1$. By construction of w and z, and since w is a convex combination of $w^1, \ldots, w^{M'}$, we have $z_n(s,a) = \delta + \epsilon > 0$ and $w_n^i(s,a) = 0$ for $i = 1, \ldots, M'$. Since $x_n(s,a) = \delta > 0$, we should have $\lambda_z = \frac{\delta}{\delta+\epsilon} > 0$. From (4.14), we obtain

$$\sum_{i=1}^{M'} \lambda_i w^i = x - \lambda_z z = \frac{\delta z + \epsilon w}{\delta + \epsilon} - \frac{\delta z}{\delta + \epsilon} = \frac{\epsilon w}{\delta + \epsilon}.$$

Since $\frac{\epsilon}{\delta+\epsilon} > 0$, by dividing the both sides by $\frac{\epsilon}{\delta+\epsilon}$ we obtain

$$w = \sum_{i=1}^{M'} \lambda_i' w^i, \tag{4.15}$$

where $\lambda_i' = \frac{\delta+\epsilon}{\epsilon}\lambda_i$ and $\sum_{i=1}^{M'} \lambda_i' = 1$. By the induction hypothesis, the representation in (4.15) is unique and has positive weights. Thus, there exist unique and positive λ_i's for $i = 1, \ldots, M'$ that satisfy (4.14) along with $\lambda_z = \frac{\delta}{\delta+\epsilon}$. Therefore, representation of x as a convex combination of z and $w^1, \ldots, w^{M'}$ is unique, and all of the weights are positive. By induction, the lemma is proven. □

By the above lemma, for any positive integer M and any exactly M-randomized policy x we can find $M+1$ $(M-1)$-randomized policies $x^1, \ldots, x^{M+1}$ that, by construction, belong to the submodel defined by x such that we can uniquely represent x as a convex combination of $x^1, \ldots, x^{M+1}$ and the weights of the convex combination are positive. Note that we can also

find $M+1$ deterministic policies to represent x via a convex combination, but there may not exist $M+1$ deterministic policies such that the convex combination representation of x is *unique*, and *all the weights are positive* [16, Example 6.1]. For ∞-randomized policies, we have a somewhat extended result with the same properties.

Lemma 4.11. *For any ∞-randomized policy x and for any positive integer L, there exist an integer $\bar{L} \geq L$ such that x can be represented as a convex combination of policies $x^1, \ldots, x^{\bar{L}}$ that belong to the submodel defined by x, such that the weights of the representation are uniquely determined and positive.*

Proof: In Section 4.7.4.

Remark 4.12. Since we have argued that nonstationary MDPs with finite state space can be seen as a special case of stationary MDPs with countably infinite state space, it is natural to consider extending the results presented here to the more general problem class. At present, we do not know whether the generalization is possible. The proofs in this section, which provides the foundation for the following results, have relied on the staged structure of the hypernetwork corresponding to nonstationary MDPs (*e.g.*, by finding the smallest or the largest period index of a state at which randomization occurs), and thus are not directly extendable to the stationary case. For example, to generalize results of subsection 4.4.2 to stationary MDPs with countably infinite state space, for a given M- or ∞-randomized policy x, we may follow steps similar to the proofs of Lemmas 4.10 and 4.11 to obtain the split into two policies, w and z. For nonstationary MDPs with finite state space, it is clear by construction that w is $(M-1)$-randomized if x is M-randomized and that we can find w that has any number of randomizations as needed if x is ∞-randomized. However, for stationary MDPs with countably infinite state space, the randomization of w is unknown and should be studied to generalize our results.

4.5 Necessary Conditions for an Extreme Point

We now return to constrained MDPs. In this section we provide necessary conditions for a feasible solution of (CNP) to be an extreme point, while the next section deals with a necessary and sufficient condition. Although many researchers have studied constrained MDPs, as far as we know, algebraic characterizations of extreme points of CILPs that represent constrained MDPs with countably infinite number of states have not been studied before. In this section, the existence of a K-randomized optimal policy, which was proven in [17] for a more general class of constrained MDPs, is given as a corollary of Theorem 4.13.

Theorem 4.13. *Any extreme point of $\mathcal{P}$ is K-randomized.*

Proof: Let x be an extreme point of $\mathcal{P}$. Suppose that x is exactly M-randomized for some $K+1 \leq M \leq \infty$. Combining Lemmas 4.10 and 4.11, we can state that there exist a (finite) integer $N > K+1$ and N policies $x^1, \ldots, x^N$ that belong to the submodel defined by x, such that x can be represented as their convex combination with uniquely determined positive weights $\lambda_1, \ldots, \lambda_N$. Note that the N policies $x^1, \ldots, x^N$ may not be in $\mathcal{P}$. Consider polyhedron $\mathcal{F}_1$ of weights of convex combination of $x^1, \ldots, x^N$ that belong to $\mathcal{P}$. That is, $\mathcal{F}_1 = \{(\nu_1, \ldots, \nu_N) \geq 0 \,|\, \sum_{i=1}^N \nu_i = 1,\ x' = \sum_{i=1}^N \nu_i x^i \in \mathcal{P}\}$. We can easily show that any convex combination of flows is a flow. Thus, in order for x' to belong to $\mathcal{P}$, it only has to satisfy inequality constraints (4.4), i.e., for $k = 1, \ldots, K$,

$$V_k \geq \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x'_n(s,a) = \sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(\alpha^{n-1} d_n^k(s,a) \sum_{i=1}^N \nu_i x_n^i(s,a) \right)$$
$$= \sum_{i=1}^N \nu_i \left(\sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n^i(s,a) \right) \triangleq \sum_{i=1}^N \nu_i D^k(x^i).$$

The exchange of sums is justified because $x^1, \ldots, x^N$ are flows, so they satisfy (4.6), and thus $D^k(x^i)$ exists for $k = 1, \ldots K$ and $i = 1, \ldots, N$. To use matrix notation, let $D = \{D_{k,i}\} \in \mathbb{R}^{K\times N}$, where $D_{k,i} \triangleq D^k(x^i)$, $\nu = (\nu_1, \ldots, \nu_N)^T \in \mathbb{R}^N$, and $v = (V_1, \ldots, V_K)^T \in \mathbb{R}^K$. Then

$\mathcal{F}_1$ can be written as

$$\mathcal{F}_1 = \{(\nu, t) \,|\, D\nu + t = v,\ \mathbf{1}^T\nu = 1,\ (\nu, t) \geq 0\}. \tag{4.16}$$

$\mathcal{F}_1$ is a polyhedron in standard form with $K+1$ equality constraints and $N+K$ variables, so any extreme point of $\mathcal{F}_1$ has at most $K+1$ nonzeros. Note that $\nu = \lambda = (\lambda_1, \ldots, \lambda_N)^T$ belongs to $\mathcal{F}_1$ with some slack t_λ since $x \in \mathcal{P}$ and $\lambda > 0$. Since $N > K+1$, (λ, t_λ) is not an extreme point of $\mathcal{F}_1$. Therefore, it is a convex combination of extreme points of $\mathcal{F}_1$, say, $(\nu^1, t^1), \ldots, (\nu^m, t^m)$ for some $m \geq 2$. Set $z^j \triangleq \sum_{i=1}^N \nu_i^j x^i$ for $j = 1, \ldots, m$. For $j = 1, \ldots, m$, $z^j \in \mathcal{P}$ because $(\nu^j, t^j) \in \mathcal{F}_1$. Since $x = \sum_{i=1}^N \lambda_i x^i$ and λ is a convex combination of $\nu^1, \ldots, \nu^m$, we can easily show that x is a convex combination of $z^1, \ldots, z^m$.

Recall that $\nu^1, \ldots, \nu^m$ have at most $K+1$ nonzeros whereas λ has $N > K+1$ nonzeros. Since $\lambda > 0$ is the unique weight vector to represent x via a convex combination of $x^1, \ldots, x^N$, $x \neq z^j$ for $j = 1, \ldots, m$. That is, x is a convex combination of points in $\mathcal{P}$ that are different from x, contradicting the assumption that x is an extreme point of $\mathcal{P}$. Therefore, the theorem is proven. □

Theorem 4.1 and Theorem 4.13 lead to the following corollary.

Corollary 4.14. *(Q) has a K-randomized optimal policy.*

The existence of a K-randomized optimal policy for constrained stationary MDPs with countably infinite number of states, which covers the stationary MDP counterpart of constrained nonstationary MDPs with finite number of states, was proven in [17]. However, the approach in [17] was based on vector optimization and geometry of the performance set, defined as the set of vectors $(C(\beta, \pi), D^1(\beta, \pi), D^2(\beta, \pi), \ldots, D^K(\beta, \pi))$ for any $\pi \in \Pi_f$. Our proof is conceptually simpler and gives insights into geometry of the feasible region of the CILP representation of a subclass of constrained MDPs with countably infinite number of states.

The next theorem shows that, at an extreme point x that uses M "more" actions than a

deterministic policy, at least M inequality constraints (4.4) are binding. To illustrate this, consider a feasible set $\{(x, s) \mid Ax + s = b, x \geq 0, s \geq 0\}$ of a finite LP. At an extreme point (x, s), the number of basic variables equals the number of equality constraints. That is, the number of non-basic (and hence 0) slack variables is greater than or equal to the number of positive components of x. Theorem 4.15 extends this condition to the CILP (CNP).

Theorem 4.15. *For any integer $M \leq K$, at an extreme point of $\mathcal{P}$ that is exactly M-randomized, at least M of inequality constraints* (4.4) *are binding.*

Proof: Let x be an exactly M-randomized extreme point of $\mathcal{P}$. Suppose that only $k < M$ inequalities of (4.4) are binding at x. Let $x^1, \ldots, x^{M+1}$ be the $M + 1$ $(M - 1)$-randomized policies and $\lambda > 0$ be the weight found by Lemma 4.10. Consider polyhedron $\mathcal{F}_2$ of weights of convex combination of $x^1, \ldots, x^{M+1}$ that belong to $\mathcal{P}$. Using similar notation, $\mathcal{F}_2$ has the form (4.16), but with $M + 1 + K$ variables and $K + 1$ equality constraints. Since $x \in \mathcal{P}$, λ belongs to $\mathcal{F}_2$ with some slack t_λ. Since only k of the constraints (4.4) are binding at x, the slack t_λ has $K - k$ nonzeros. Therefore, (λ, t_λ) has $M + 1 + K - k$ nonzeros and since $k < M$, we have $M + 1 + K - k > K + 1$, i.e., (λ, t_λ) is not an extreme point of $\mathcal{F}_2$. Then, (λ, t_λ) is a convex combination of extreme points of $\mathcal{F}_2$, say, $(\nu^1, s^1), \ldots, (\nu^m, s^m)$ for some $m \geq 2$. Note that this convex combination is not a trivial one. Also, note that slack variables are determined by the weight variables, i.e., having $\lambda = \nu^j$ for some j implies $(\lambda, t_\lambda) = (\nu^j, s^j)$. Thus, $\nu^1, \ldots, \nu^m$ are different from λ. Set $z^j \triangleq \sum_{i=1}^{N} \nu_i^j x^i$ for $j = 1, \ldots, m$. Then, by Lemma 4.10, $z^1, \ldots, z^m$ are different from x. Similarly to the proof of Theorem 4.13, $z^j \in \mathcal{P}$ for $j = 1, \ldots, m$ and x is a convex combination of $z^1, \ldots, z^m$, contradicting that x is an extreme point of $\mathcal{P}$. □

4.6 A Necessary and Sufficient Condition for an Extreme Point

Theorems 4.13 and 4.15 lead to the following necessary condition for $x \in \mathcal{P}$ to be an extreme point: it should be exactly M-randomized for some $M \leq K$ and at least M of

constraints (4.4) should be binding at x. In this section, we establish a necessary and sufficient condition for a point in $\mathcal{P}$ to be an extreme point.

Definition 4.16 ([53])**.** A convex subset E of a convex set D is called *extreme* if any representation $x = \lambda z + (1-\lambda)w$ for $0 < \lambda < 1$, with $z, w \in D$, of a point $x \in E$ implies $z, w \in E$.

For example, a face of a polyhedron in a finite dimensional space is an extreme set of the polyhedron. (A subset E of a convex set D is called *exposed* if there is a hyperplane H supporting E such that $E = H \cap D$. In general, an exposed subset of a convex set is extreme but the converse may not hold [17].)

Guided by the necessary conditions from the previous section, we consider an exactly M-randomized feasible policy x with $M \leq K$ at which M of inequality constraints (4.4) are binding. Let B be the submodel defined by x. Let N be the number of deterministic policies in submodel B and let $x^1, \ldots, x^N$ be these deterministic policies. (Notice that policies x^i for $i = 1, \ldots, N$ considered here are different from those used in the proof of Theorem 4.13; rather, they correspond to the decomposition discussed in Theorem 4.9.)

Consider polyhedron $\mathcal{G}$ of weights of convex combinations of $x^1, \ldots, x^N$ that belong to $\mathcal{P}$. Using the same notation as in (4.16), $\mathcal{G}$ is described as:

$$\mathcal{G} = \{(\nu, t) \,|\, D\nu + t = v,\ \mathbf{1}^T\nu = 1,\ (\nu, t) \geq 0\}.$$

We emphasize again that, unlike $\mathcal{F}_1$ and $\mathcal{F}_2$, $\mathcal{G}$ is formed using N *deterministic* policies in submodel B.

Let $\lambda = (\lambda_1, \ldots, \lambda_N)^T$ be an element of $\Lambda(x)$ defined in (4.13). Since $x = \sum_{i=1}^N \lambda_i x^i \in \mathcal{P}$, $\nu = \lambda$ belongs to $\mathcal{G}$ together with some slacks. Since the values of the slack variables are determined by x (and do not depend on a specific $\lambda \in \Lambda(x)$), let t_x denote the vector of

slacks corresponding to x. Let

$$\tilde{\Lambda}(x) = \left\{(\lambda, t_x) \in \mathbb{R}^N \times \mathbb{R}^K_+ \,|\, \lambda \in \Lambda(x),\ t_x = v - D\lambda\right\}.$$

Note that the last K components of elements of $\tilde{\Lambda}(x)$ (the slack part) are fixed at t_x. Since $x \in \mathcal{P}$, we have $\tilde{\Lambda}(x) \subseteq \mathcal{G}$. We state the following theorem (a proof will be provided later in this section).

Theorem 4.17. *A feasible exactly M-randomized policy x for some $M \leq K$ at which at least M of inequality constraints* (4.4) *are binding is an extreme point of $\mathcal{P}$ if and only if $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$.*

Theorems 4.13, 4.15, and 4.17 lead to the following corollary, a necessary and sufficient condition for a feasible solution of (CNP) to be an extreme point that can be checked using the finite polyhedron $\mathcal{G}$.

Corollary 4.18. *A point $x \in \mathcal{P}$ is an extreme point of $\mathcal{P}$ if and only if it is an exactly M-randomized policy for some $M \leq K$, at least M of inequality constraints* (4.4) *are binding at x, and $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$.*

We first illustrate the above corollary for the case of $K = 1$. For $K = 1$, the only candidates to be considered are deterministic policies and exactly 1-randomized policies for which constraint (4.4) is binding. Let x be a feasible deterministic policy, and let t_x be the corresponding slack. Then $\tilde{\Lambda}(x) = \{(1, t_x)\}$. It is easy to show that the corresponding polyhedron $\mathcal{G}$ consists of one point, $(1, t_x)$ and $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$. Now, let x be a feasible 1-randomized policy with $t_x = 0$. There exists $\lambda \in (0, 1)$ and deterministic policies x^1, x^2 such that $x = \lambda x^1 + (1 - \lambda)x^2$, and we have $\tilde{\Lambda}(x) = \{(\lambda, 1 - \lambda, 0)^T\}$. Since $\tilde{\Lambda}(x)$ is a singleton, it is an extreme set if and only if the point $(\lambda, 1 - \lambda, 0)^T$ is an extreme point of $\mathcal{G}$. Since $K = 1$, by dropping the constraint index k, $\mathcal{G}$ can be written as

$$\mathcal{G} = \{(\nu_1, \nu_2, t) \,|\, D(x^1)\nu_1 + D(x^2)\nu_2 + t = V,\ \nu_1 + \nu_2 = 1,\ (\nu_1, \nu_2, t) \geq 0\}.$$

Since $(\lambda, 1-\lambda, 0) \in \mathcal{G}$, we have either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$ or $D(x^1) = D(x^2) = V$. The point $(\lambda, 1-\lambda, 0)$ is an extreme point if and only if the corresponding basis matrix is nonsingular, which is equivalent to $D(x^1) \neq D(x^2)$. Consequently, $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$ if and only if either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$. Therefore, according to Corollary 4.18, for $K = 1$, a point $x \in \mathcal{P}$ is an extreme point if and only if x is either a feasible deterministic policy or a feasible exactly 1-randomized policy such that the inequality constraint is binding at x and it is a non-trivial convex combination of two deterministic policies x^1 and x^2 for which either $D(x^1) < V < D(x^2)$ or $D(x^2) < V < D(x^1)$ holds.

To gain intuition, consider the intersection of a polyhedron and a halfspace in a finite dimensional space (Figure 4.1). Extreme points of the intersection of the polyhedron P and the halfspace defined by an additional constraint $\{x : d^T x \leq v\}$ are either extreme points of P that belong to the halfspace (such as x_3 in Figure 4.1) or points where an edge of P intersects the hyperplane defined by the halfspace (such as x' in Figure 4.1, which is a convex combination of adjacent extreme points x^1 and x^2). Consider now an unconstrained MDP obtained by excluding the linear inequality constraint (4.4) from (CNP). A feasible solution to the unconstrained MDP is an extreme point if and only if it is a deterministic policy (e.g., Theorem 4.3 of [24]). Then, the necessary and sufficient condition for $K = 1$ shows that the characterization of extreme points of the intersection of a polyhedron and a halfspace in finite dimensional space naturally extends to $\mathcal{P}$, which is the intersection of the infinite dimensional feasible region of the unconstrained MDP and the set satisfying the (linear) inequality constraint.

Proof of Theorem 4.17: Suppose that $\tilde{\Lambda}(x)$ is not an extreme set of $\mathcal{G}$. Then there exist (σ, t_1) and (τ, t_2) in $\mathcal{G}$ such that $(\theta\sigma + (1-\theta)\tau, \theta t_1 + (1-\theta)t_2) \in \tilde{\Lambda}(x)$ for some $\theta \in (0,1)$, but either $(\sigma, t_1) \notin \tilde{\Lambda}(x)$ or $(\tau, t_2) \notin \tilde{\Lambda}(x)$, or both. Without loss of generality, suppose $(\sigma, t_1) \notin \tilde{\Lambda}(x)$. Let $z \triangleq \sum_{i=1}^{N} \sigma_i x^i$ and $w \triangleq \sum_{i=1}^{N} \tau_i x^i$; by construction, z and w are in $\mathcal{P}$. If $z = x$, then $(\sigma, t_1) \in \tilde{\Lambda}(x)$ since the slack of z, t_1, should equal the slack of x. Thus,

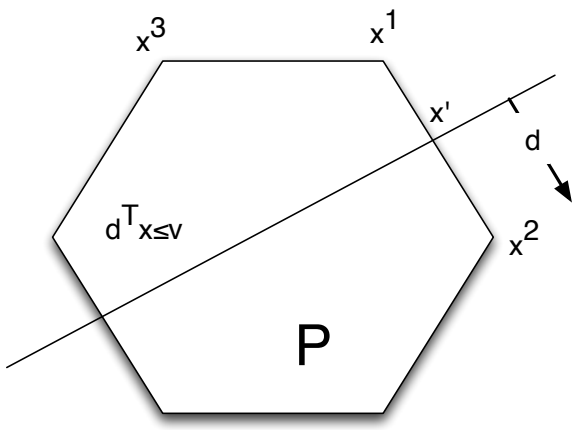


Figure 4.1: Extreme points for $K = 1$

z is not equal to x. However, $x = \sum_{i=1}^{N}[\theta\sigma_i + (1-\theta)\tau_i]x^i = \theta z + (1-\theta)w$, where $\theta \in (0,1)$ and $z \neq x$. Since both z and w are in $\mathcal{P}$, x is not an extreme point of $\mathcal{P}$. We showed that if x is an extreme point of $\mathcal{P}$, then $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$.

For the converse, suppose $x \in \mathcal{P}$ is not an extreme point. Then there exist z and w in $\mathcal{P}$ such that $x = \theta z + (1-\theta)w$ for some $\theta \in (0,1)$. It can be shown that z and w belong to the submodel defined by x and thus can be expressed as convex combinations of $x^1, \ldots, x^N$, say, $z = \sum_{i=1}^{N} \sigma_i x^i$ and $w = \sum_{i=1}^{N} \tau_i x^i$. By construction, σ and τ belong to $\mathcal{G}$ with slacks t_z and t_w, respectively. Since z and w are different from x, (σ, t_z) and (τ, t_w) are not in $\tilde{\Lambda}(x)$. However,

$$\sum_{i=1}^{N}[\theta\sigma_i + (1-\theta)\tau_i]x^i = \theta\sum_{i=1}^{N}\sigma_i x^i + (1-\theta)\sum_{i=1}^{N}\tau_i x^i = \theta z + (1-\theta)w = x,$$

and moreover,

$$\theta t_z + (1-\theta)t_w = \theta(v - D\sigma) + (1-\theta)(v - D\tau) = v - D(\theta\sigma + (1-\theta)\tau) = t_x.$$

Therefore, $(\theta\sigma + (1-\theta)\tau, \theta t_z + (1-\theta)t_w) \in \tilde{\Lambda}(x)$ and it is a convex combination of (σ, t_z) and (τ, t_w) which are not in $\tilde{\Lambda}(x)$. That is, $\tilde{\Lambda}(x)$ is not an extreme set of $\mathcal{G}$. Therefore, if $\tilde{\Lambda}(x)$ is an extreme set of $\mathcal{G}$, then x is an extreme point of $\mathcal{P}$. □

The next example illustrates Theorem 4.17 for one possible type of a 2-randomized policy and $K = 2$.

Example 4.19. Let $K = 2$. Consider an exactly 2-randomized policy x such that both inequality constraints (4.4) are binding at x, and x randomizes only at a period-state pair (n, s) over three actions, say, a^1, a^2, and a^3. Then in the submodel defined by x, there are three deterministic policies, say, x^1, x^2, x^3, where x^i chooses a^i at (n, s) for $i = 1, 2, 3$, and $x = \sum_{i=1}^{3} \lambda_i x^i$ where $\lambda > 0$ and $\mathbf{1}^T\lambda = 1$. Since both inequality constraints are binding at x, its corresponding t_x is 0. We can check that $\tilde{\Lambda}(x) = \{(\lambda_1, \lambda_2, \lambda_3, 0, 0)\}$. Then, Theorem 4.17 implies that x is an extreme point of $\mathcal{P}$ if and only if $(\lambda_1, \lambda_2, \lambda_3, 0, 0)$ is an extreme point of $\mathcal{G}$, which is equivalent to the following basis matrix being nonsingular:

$$D_B = \begin{bmatrix} D^1(x^1) & D^1(x^2) & D^1(x^3) \\ D^2(x^1) & D^2(x^2) & D^2(x^3) \\ 1 & 1 & 1 \end{bmatrix}.$$

Consider x^i for $i = 1, 2, 3$ as vectors in $\mathbb{R}^\infty$, easily shown to be linearly independent. Then the subspace S of $\mathbb{R}^\infty$ spanned by x^1, x^2, x^3 has dimension 3. Define an isomorphism linear operator $T : S \to \mathbb{R}^3$ as $T(\nu_1 x^1 + \nu_2 x^2 + \nu_3 x^3) = (\nu_1, \nu_2, \nu_3)$. Then Tx belongs to the hyperplane $H \triangleq \{\nu \in \mathbb{R}^3 \mid \mathbf{1}^T\nu = 1\}$. $D^1(\cdot)$ is a linear functional on $\mathbb{R}^\infty$ and $H^1_{\text{inf}} \triangleq \{x' : D^1(x') = v_1\}$ is a hyperplane in $\mathbb{R}^\infty$. Then $H^1 \triangleq T(H^1_{\text{inf}} \cap S) = \{\nu \in \mathbb{R}^3 \mid D^1(x^1)\nu_1 + D^1(x^2)\nu_2 + D^1(x^3)\nu_3 = v_1\}$, and H^2 (defined analogously using $D^2(\cdot)$ and v_2) are also hyperplanes in $\mathbb{R}^3$. Then, nonsingularity of D_B is equivalent to the hyperplanes H, H^1 and H^2 intersecting at a single point in $\mathbb{R}^3$. However, $Tx = (\lambda_1, \lambda_2, \lambda_3)$ is in H and x also satisfies $D^1(x) = v_1$ and $D^2(x) = v_2$. Thus, if the intersection of H, H^1 and H^2 is a single point, then Tx is that point. Therefore, the necessary and sufficient condition of Theorem 4.17 is equivalent to $H \cap H^1 \cap H^2 = \{Tx\}$.

4.7 Technical Proofs

4.7.1 Proof of Theorem 4.5

From the complementary slackness condition (4.11), we have

$$\alpha^{n-1}c_n(s,a)x_n(s,a) = \left(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\right)x_n(s,a) - \alpha^{n-1}\left(\sum_{k=1}^{K} d_n^k(s,a)\mu_k\right)x_n(s,a)$$

for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$. By summing up both sides for $n = 1, 2, \ldots, N$, $s \in \mathcal{S}, a \in \mathcal{A}$, we obtain

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) \\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\right)x_n(s,a) \\
&\quad - \sum_{k=1}^{K}\mu_k \sum_{n=1}^{N}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d_n^k(s,a)x_n(s,a). \qquad (4.17)
\end{aligned}$$

We simplify the first sum of the right hand side as follows:

$$\begin{aligned}
&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \left(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\right)x_n(s,a) \\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_n(s'|s,a)x_n(s,a) \\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{a\in\mathcal{A}} x_{n+1}(s',a) \\
&= \sum_{s\in\mathcal{S}} y_1(s)\sum_{a\in\mathcal{A}} x_1(s,a) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a) \\
&= \sum_{s\in\mathcal{S}} \beta(s)y_1(s) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a). \qquad (4.18)
\end{aligned}$$

By substituting (4.18) into (4.17), we have

$$\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}c_n(s,a)x_n(s,a)$$
$$=\sum_{s\in\mathcal{S}}\beta(s)y_1(s)-\sum_{s\in\mathcal{S}}y_{N+1}(s)\sum_{a\in\mathcal{A}}x_{N+1}(s,a)-\sum_{k=1}^{K}\mu_k\sum_{n=1}^{N}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}d_n^k(s,a)x_n(s,a). \tag{4.19}$$

The second term above goes to zero as N increases, because

$$-S\alpha^N\tau_y\le\sum_{s\in\mathcal{S}}y_{N+1}(s)\sum_{a\in\mathcal{A}}x_{N+1}(s,a)\le S\alpha^N\tau_y. \tag{4.20}$$

By the second condition (4.12) of complementary slackness, for $k=1,2,\ldots,K$,

$$\mu_k\sum_{n=1}^{N}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}d_n^k(s,a)x_n(s,a)\to V_k\mu_k \text{ as } N\to\infty.$$

Thus, taking $N\to\infty$ on both sides of (4.19) gives $f(x)=g(y,\mu)$. The second statement of the theorem follows by weak duality. □

4.7.2 Proof of Theorem 4.6

Since x and (y,μ) are feasible to (CNP) and (CND), respectively, we have $x_n(s,a)\ge 0$ and

$$\alpha^{n-1}\left(c_n(s,a)+\sum_{k=1}^{K}d_n^k(s,a)\mu_k\right)\ge y_n(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}(s') \text{ for } n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}.$$

Thus, the left hand side of the first condition (4.11) of complementary slackness is nonnegative for $n\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}$. By summing up the left hand side of (4.11) for $n=1,2,\ldots,N$,

$s \in \mathcal{S}$, and $a \in \mathcal{A}$, we obtain

$$\begin{aligned}
0 \leq & \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x_n(s,a) \left[\alpha^{n-1} \left(c_n(s,a) + \sum_{k=1}^{K} d_n^k(s,a)\mu_k \right) - y_n(s) + \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \right] \\
= & \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) + \sum_{k=1}^{K} \mu_k \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_n^k(s,a) x_n(s,a) \\
& - \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left[y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \right] x_n(s,a) \\
= & \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) + \sum_{k=1}^{K} \mu_k \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d_n^k(s,a) x_n(s,a) \\
& - \sum_{s\in\mathcal{S}} \beta(s) y_1(s) + \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x_{N+1}(s,a)
\end{aligned}$$

where the last equality is obtained from (4.18). By taking $N \to \infty$ on both sides,

$$\begin{aligned}
0 \leq & \lim_{N\to\infty} \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} x_n(s,a) \left[\alpha^{n-1} \left(c_n(s,a) + \sum_{k=1}^{K} d_n^k(s,a)\mu_k \right) - y_n(s) \right. \\
& \left. + \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \right] \\
\leq & f(x) + \sum_{k=1}^{K} V_k \mu_k - \sum_{s\in\mathcal{S}} \beta(s) y_1(s) = f(x) - g(y,\mu) = 0,
\end{aligned}$$

by strong duality. This shows that the sum of the left hand side of (4.11) equals zero. However, we know the left hand side of (4.11) for $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$ is nonnegative, and thus, each of them equals zero. Therefore, (4.11) holds.

We now prove (4.12). Since x is feasible to (CNP),

$$V_k \geq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \text{ for } k = 1, 2, \ldots, K.$$

Since $\mu_k \geq 0$, the left hand side of (4.12) is nonnegative for $k = 1, 2, \ldots, K$. By summing up the left hand side of (4.12) for $k = 1, 2, \ldots, K$, we obtain

$$
\begin{aligned}
0 &\leq \sum_{k=1}^{K} \mu_k \left[V_k - \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \right] \\
&= \sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) - \sum_{k=1}^{K} \mu_k \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} d_n^k(s,a) x_n(s,a) \\
&= \lim_{N\to\infty} \left[\sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} \left(c_n(s,a) + \sum_{k=1}^{K} d_n^k(s,a)\mu_k \right) x_n(s,a) \right] \\
&\leq \lim_{N\to\infty} \left[\sum_{s\in\mathcal{S}} \beta(s) y_1(s) - \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \right) x_n(s,a) \right] \\
&= \lim_{N\to\infty} \sum_{s\in\mathcal{S}} y_{N+1}(s) \sum_{a\in\mathcal{A}} x_{N+1}(s,a)
\end{aligned}
$$

where the first equality is obtained from strong duality, the second inequality from the constraint (4.8) of (CND), and the last equality from (4.18). We know that the last expression is zero from (4.20) and thus, (4.12) holds. □

4.7.3 Proof of Theorem 4.9

For simplicity, we will assume that x does not allow any node that has zero incoming flow. Our proof can be easily extended to the general case.

Let $(n_1, s_1), (n_2, s_2), \ldots, (n_m, s_m)$ be the period-state pairs at which x randomizes and suppose that they are ordered so that the period index is nondecreasing. For $i = 1, 2, \ldots, m$, assume that x randomizes over $a^{i,1}, a^{i,2}, \ldots, a^{i,l_i}$ at (n_i, s_i). We have $\sum_{i=1}^{m}(l_i - 1) = M$ and $\prod_{i=1}^{m} l_i = N$. Let $\Lambda^0(x) = \{\lambda \in \mathbb{R}^N \mid \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. Since policy x does not randomize in periods 1 to $n_1 - 1$, the policies $x^1, x^2, \ldots, x^N$ choose the same action with x in those periods. Consequently, for any $\lambda \in \Lambda^0(x)$, $\sum_{i=1}^{N} \lambda_i x^i$ and x have the same flows in periods 1 to $n_1 - 1$. This implies that they also have the same flow on hyperarcs from the period-state pairs in period n_1 where x does not randomize. Let $\Lambda^1(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^{N} x_{n_1}^i(s_1, a^{1,1})\lambda_i = x_{n_1}(s_1, a^{1,1}), \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. Then, for any $\lambda \in \Lambda^1(x)$, $\sum_{i=1}^{N} \lambda_i x^i$ and x have the same

flows in periods 1 to $n_1 - 1$ and on those hyperarcs from the period-state pairs in period n_1 where x does not randomize. Moreover, in period n_1, they have the same flow on hyperarc $(n_1, s_1, a^{1,1})$. Let $\Lambda^{l_1-1}(x) = \{\lambda \in \mathbb{R}^N \mid \sum_{i=1}^N x^i_{n_1}(s_1, a^{1,j})\lambda_i = x_{n_1}(s_1, a^{1,j})$ for $j = 1, 2, \ldots, l_1 - 1, \mathbf{1}^T\lambda = 1, \lambda \geq 0\}$. For any $\lambda \in \Lambda^{l_1-1}(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in periods 1 to $n_1 - 1$ and on those hyperarcs from the period-state pairs in period n_1 where x does not randomize, and additionally in period n_1, they have the same flow on hyperarc $(n_1, s_1, a^{1,j})$ for $j = 1, 2, \ldots, l_1 - 1$. Then, they also have the same flow on hyperarc (n_1, s_1, a^{1,l_1}), and thus, they coincide on all hyperarcs emanating from (n_1, s_1). Note that x randomizes over l_1 actions at (n_1, s_1) and we added $l_1 - 1$ equations to obtain $\Lambda^{l_1-1}(x)$ from $\Lambda^0(x)$.

We can apply the same procedure repeatedly to the other period-state pairs (n_2, s_2), $(n_3, s_3), \ldots, (n_m, s_m)$, in order of nondecreasing period index. Then we obtain $\Lambda^M(x)$ such that for any $\lambda \in \Lambda^M(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in periods 1 to $n_m - 1$ (which implies that they also have the same flow on all hyperarcs in period n_m where x does not randomize). Moreover, in period n_m, they have the same flow on all hyperarcs from any period-state pair where x randomizes due to the added equality constraints on λ. Thus, for any $\lambda \in \Lambda^M(x)$, $\sum_{i=1}^N \lambda_i x^i$ and x coincide in all periods, i.e., $\Lambda^M(x) \subset \Lambda(x)$. We can easily see that any $\lambda \in \Lambda(x)$ satisfies all of the equalities that define $\Lambda^M(x)$. Therefore, we showed $\Lambda^M(x) = \Lambda(x)$.

Let A be the coefficient matrix of the M equalities added to obtain $\Lambda^M(x)$ from $\Lambda^0(x)$. We will show that the rows of the matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ are linearly independent. The rows of A corresponds to the M equalities that define $\Lambda^M(x)$ from $\Lambda^0(x)$. For $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, l_i - 1$, let $A_{(i,j)}$ denote the row of A corresponding to the equality $\sum_{k=1}^N x^k_{n_i}(s_i, a^{i,j})\lambda_k = x_{n_i}(s_i, a^{i,j})$. Then, $A_{(i,j)}$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, l_i - 1$ form all rows of A. Let the rows of A be sorted in lexicographic order of their subscripts. Recall that the columns of A correspond to the deterministic policies $x^1, x^2, \ldots, x^N$ and that a row $A_{(i,j)}$ has nonzeros in those columns that correspond to the deterministic policies that choose the action $a^{i,j}$ at (n_i, s_i). Let's fix a row $A_{(i,j)}$ and let's focus on the columns corresponding to policies

choosing a^{k,l_k} at (n_k, s_k) for $k = 1, 2, \ldots, i-1$ and $a^{i,j}$ at (n_i, s_i). In these columns, the row $A_{(i,j)}$ has nonzeros. However, by construction, all rows above the row $A_{(i,j)}$ have zeros in those columns. Therefore, we showed that each row of A has a nonzero in a column at which other rows above the row have zeros. Moreover, all the rows of A have zeros at the column of policy choosing a^{k,l_k} at (n_k, s_k) for $k = 1, \ldots, m$ (i.e., always choosing the last action), and thus, the row with ones has a nonzero in the column at which all rows of A have zeros. Therefore, we proved that the matrix $\begin{bmatrix} A \\ \mathbf{1}^T \end{bmatrix}$ has a full row rank. □

4.7.4 Proof of Lemma 4.11

Let $\mathcal{H}(x, n) \triangleq \{(n, s, a) \,|\, x_n(s, a) > 0,\ s \in \mathcal{S},\ a \in \mathcal{A}\}$, and let $r_n(x) = |\mathcal{H}(x, n)| - |S|$, i.e., $r_n(x)$ is the number of "additional" actions used by x compared to a deterministic policy in period n. Also, for any $m \in \mathbb{N}$, $t \in \mathcal{S}$, $b \in \mathcal{A}$, let $\phi_m(t, b) = x_m(t, b) / \sum_{a \in \mathcal{A}} x_m(t, a)$, i.e., $\phi_m(t, b)$ is the probability that x will select action b in state t in period m.

We prove the lemma by induction on L. For $L = 1$, we can simply let $\bar{L} = L = 1$ and $x^1 = x$.

Suppose the statement holds for $L = L' - 1 \geq 1$. Since x is an ∞-randomized policy, $\sum_{n=1}^{\infty} r_n(x) = \infty$. Let $n = \min\{\bar{n} \,|\, \sum_{n'=1}^{\bar{n}} r_{n'}(x) \geq L' - 1\}$. Let $s \in \mathcal{S}$ be any state such that in period-state pair (n, s), x randomizes over multiple actions, say, $a, b^1, b^2, \ldots, b^l$. Let $x_n(s, a) = \delta > 0$, $x_n(s, b^i) = \epsilon_i > 0$, $i = 1, \ldots, l$, and let $\epsilon = \sum_{i=1}^{l} \epsilon_i$. We will represent x as a convex combination of two flows, w and z. Again, we define sub-flows u and v^i for $i = 1, \ldots, l$ to construct w and z.

For $k \geq n + 1$, let $\mathcal{T}_k^i(x) \subset \mathcal{S}$ be the set of states in period k that receive any portion of flow ϵ_i originating in hyperarc (n, s, b^i) under policy x, and for $t_k \in \mathcal{T}_k^i(x)$, let $\mathcal{B}_k^i(t_k) = \{b_k \in \mathcal{A} \,|\, x_k(t_k, b_k) > 0\}$. Let $\mathcal{G}^i(x)$ be the sub-hypernetwork formed by the node (n, s), hyperarc (n, s, b^i), nodes in $\cup_{k=n+1}^{\infty} \mathcal{T}_k^i(x)$, and hyperarcs in $\cup_{k=n+1}^{\infty} \cup_{t_k \in \mathcal{T}_k^i(x)} \mathcal{B}_k^i(t_k)$. Then, define sub-flow v^i in the following way. Let node (n, s) be the source node with supply 1.

Set $v_n^i(s, b^i) = 1$ and for each $t_{n+1} \in \mathcal{T}_{n+1}^i(x)$ and each $b_{n+1} \in \mathcal{B}_{n+1}^i(t_{n+1})$,

$$v_{n+1}^i(t_{n+1}, b_{n+1}) = \phi_{n+1}(t_{n+1}, b_{n+1}) p_n(t_{n+1}|s, b^i).$$

For $k \geq n+2$ and for $t_k \in \mathcal{T}_k^i(x)$ and $b_k \in \mathcal{B}_k^i(t_k)$, set

$$v_k^i(t_k, b_k) = \phi_k(t_k, b_k) \sum_{t_{k-1} \in \mathcal{T}_{k-1}^i(x)} \sum_{b_{k-1} \in \mathcal{B}_{k-1}^i(t_{k-1})} p_{k-1}(t_k|t_{k-1}, b_{k-1}) v_{k-1}^i(t_{k-1}, b_{k-1}).$$

Sub-flow u is defined similarly in the sub-hypernetwork consisting of the node (n, s), hyperarc (n, s, a) and the part of the hypernetwork receiving any portion of the flow δ.

As in the proof of Lemma 4.10, w is obtained from x by redirecting flow δ from $\mathcal{F}(x)$ to $\mathcal{G}^i(x)$'s, and z is obtained from x by redirecting flow ϵ from $\mathcal{G}^i(x)$'s to $\mathcal{F}(x)$, maintaining the original proportion of flows in $\mathcal{G}^i(x)$'s. By construction, w and z satisfy the flow balance constraints, and we have $x = \frac{\delta z + \epsilon w}{\delta + \epsilon}$.

In the construction of w, the hyperarc (n, s, a) is the only randomization removed from x in periods $1, 2, \ldots, n$. Since $\sum_{n'=1}^{n} r_{n'}(x) - 1 \geq L' - 2$, w is at least $(L'-2)$-randomized. We consider the following two cases regarding the randomization of w.

If w is exactly $\bar{N}$-randomized for some positive integer $\bar{N} \geq L' - 2$, then by Lemma 4.10, there exists $\bar{N}+1$ $(\bar{N}-1)$-randomized policies $w^1, \ldots, w^{\bar{N}+1}$ such that w is their convex combination with uniquely determined, positive weights. By arguments in the proof of Lemma 4.10, we can show that in representing x as a convex combination of z and $w^1, \ldots, w^{\bar{N}+1}$ the weight of z is $\frac{\delta}{\delta+\epsilon} > 0$, and therefore x can be represented as a convex combination of $\bar{N} + 2 (\geq L')$ policies z and $w^1, \ldots, w^{\bar{N}+1}$, and the representation has uniquely determined positive weights.

If w is ∞-randomized, by the induction hypothesis, there exists a positive integer $N' \geq L' - 1$ and policies $w^1, \ldots, w^{N'}$ such that w is uniquely represented as their convex combination and the weights are positive. Similarly, we can show that all of z and $w^1, \ldots, w^{N'}$ are necessary to represent x as their convex combination and the weights are uniquely de-

termined.

Therefore, by induction, the lemma is proven. □

CHAPTER V

Conclusion and Future Research

In this dissertation, we studied CILP formulations of non-stationary MDPs, countable-state MDPs, and constrained non-stationary MDPs. We established foundations for developing simplex-type algorithms for solving the CILP formulations, developed simplex-type algorithms, and analyzed performance of the algorithms.

In Chapter II, we established rate of convergence results of the simplex algorithm for non-stationary MDPs introduced in [24] with a particular pivoting rule and RHA. We introduced the multiple pivoting technique which greatly accelerates the simplex algorithm. We also compared the simplex algorithm and RHA empirically and analyzed its performance. The experiments showed that the upper bound on the number of iterations to achieve near-optimality was pessimistic for the inventory management problem. It is a future research either to derive a tighter theoretical guarantee or to find a non-stationary MDP example for which the bound is not so pessimistic or even tight. In the experimental section, we only compared how fast the two algorithms converge to optimality. However, the cost of data acquisition is another important issue. As those algorithms proceed, they require transition probabilities and rewards in more periods. In practice, obtaining the problem data can be expensive. It will be interesting to compare how much data the algorithms require to achieve ϵ-optimality for a given ϵ.

In Chapter III, we extended the major theoretical extreme point and duality results to

the CILP formulation of countable-state MDPs and introduced the implementable simplex algorithm for solving the CILP formulation. This algorithm is the first solution algorithm for countable-state MDPs that generates a sequence of policies whose value functions improve monotonically and converge to optimality. Also, unlike existing simplex-type algorithms for CILPs, our algorithm solves a class of CILPs in which each constraint may contain an infinite number of variables and each variable may appear in an infinite number of constraints.

A possible future research is comparing the simplex algorithm to the existing methods for countable-state MDPs. However, it is not straightforward how to make an empirical comparison, also because of the issue of data acquisition. As those algorithms for countable-state MDPs proceed, they require transition probabilities and rewards from more states. For example, if an algorithm converges to optimality in value faster than another algorithm but requires significantly more data, then it is not clear which one is a better solution method. Another future research direction is to study convergence rate of the simplex algorithm for countable-state MDPs, possibly in a way similar to our analysis in Chapter II on the simplex algorithm for non-stationary MDPs. Then, it would be possible to compare the convergence rates of the algorithms for countable-state MDPs by comparing the result for the simplex algorithm to the ones in [58, 56]. However, the convergence rates provide us with upper bounds on number of iterations (or computational complexity) to achieve near-optimality so this theoretical comparison would also be incomplete.

In Chapter IV, we established duality results for the CILP formulation of constrained non-stationary MDPs and provided a complete algebraic characterization of extreme points of the CILP formulation. The existence of a K-randomized optimal policy followed this characterization. It is a natural next step to develop a simplex-type algorithm for solving the CILP based on the characterization of extreme points. However, there is a major difference between the CILP (CNP) and other primal CILPs in this thesis which makes the development difficult: the left hand side of (4.4) cannot be exactly computed using finite computation, in other words, feasibility of a solution cannot be determined, while in (NP) and (CP) basic

solutions are trivially feasible. Thus, we will instead consider solving a Lagrangian dual of (CNP). Let's first consider the following Lagrangian relaxation of (CNP) obtained by relaxing the side constraints (4.4):

$$\begin{aligned}
\min_{x} \; L(x,\mu) = & \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} \left(c_n(s,a) + \sum_{k=1}^{K} \mu_k d_n^k(s,a) \right) x_n(s,a) - \mu^T V \\
\text{s.t.} \; & \sum_{a\in\mathcal{A}} x_1(s,a) = \beta(s) \text{ for } s \in \mathcal{S} \\
& \sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{n-1}(s|s',a) x_{n-1}(t,a) = 0 \text{ for } n \geq 2, s \in \mathcal{S} \\
& x \geq 0.
\end{aligned}$$

Let $L(\mu)$ be the optimal value of the Lagrangian relaxation for a multiplier μ. Then we consider the following Lagrangian dual of (CNP):

$$\text{(LCNP)} \; \max_{u\geq 0} L(u).$$

It was shown in [4] that (LCNP) is a strong dual of (CNP) under Assumption 4.2. We will develop subgradient algorithms for solving (LCNP). Note that exact evaluation of $L(\cdot)$ and its subgradient involve solving the Lagrangian relaxation exactly, that is, solving an unconstrained non-stationary MDP to optimality. Therefore, challenges are in ensuring convergence of subgradient algorithms despite the fact that $L(\cdot)$ and its subgradient are only approximately computed.

In this thesis, we extended the standard LP results and the simplex method to the CILPs representing the three classes of MDPs. In order to extend the LP approach to more general classes of CILPs, one would have to understand what aspects of the CILPs considered in this thesis enabled the success of the LP approach. In (NP), each variable appears only in a finite number of constraints and each constraint has only a finite number of variables, and this property was shown to be useful in proving duality results [41, 42]. In (CNP) of Chapter IV,

each variable appears only in a finite number of constraints but the side constraints have an infinite number of variables. Meanwhile, the coefficient matrix of (CP) in Chapter III can be dense, but it still has the MDP structure. Thus, this thesis is an encouraging first step to extend the standard LP results and the simplex method to CILPs that do not have any sparsity structure. We will analyze what characteristics of the MDP structure made the LP approach successful. We will also study how the assumptions on problem parameters (such as uniformly bounded costs for non-stationary MDPs or the assumptions in Section 3.1.2 for countable-state MDPs) helped the success.

APPENDICES

APPENDIX A

Extreme Point Cone Inclusion

Consider a finite-dimensional polyhedron that has an extreme point. Then any point in the polyhedron is represented as a sum of the vector from origin to the extreme point and a conic combination of vectors from the extreme point to its adjacent extreme points. In this section we prove this statement for any bounded polyhedron in finite dimension and also for the feasible region of (NP) which is infinite-dimensional. Using the result for (NP), we provide an alternative proof of Lemma 2.18.

A.1 In Finite-dimensional Spaces

Consider a finite-dimensional polyhedron $\mathcal{H}$ that is bounded (such polyhedron is also called polytope). Then, $\mathcal{H}$ is represented by a set of linear equations:

$$\mathcal{H} = \{x \in \mathbb{R}^N \mid \sum_{n=1}^{N} a_{m,n} x_n \leq b_m, m = 1, 2, \ldots, M\}.$$

For simplicity of illustration, suppose that $\mathcal{H}$ is a full dimensional polytope. $\mathcal{H}$ has an extreme point, say x^0. There are a finite number, say K, of adjacent extreme points of x^0. Let us denote them as $x^1, x^2, \ldots, x^K$. Consider a cone generated by conic combinations of

vectors from x^0 to the adjacent extreme points and let $\mathcal{C}_1$ be the cone transferred by x^0, i.e.,

$$\mathcal{C}_1 \triangleq \{x \in \mathbb{R}^n \mid x = x^0 + \sum_{k=1}^{K} \mu_i (x^k - x^0), \mu_k \geq 0, k = 1, 2, \ldots, K\}.$$

On the other hand, among the constraints that are binding at x^0, there are N constraints that are linearly independent (constraints are linearly independent if vectors of the coefficients appended by the right hand side are linearly independent). Let them be indexed by $m(i)$ for $i = 1, 2, \ldots, N$, i.e., $\sum_{n=1}^{N} a_{m(i),n} x_n \leq b_{m(i)}$ for $i = 1, 2, \ldots, N$. Let $\mathcal{C}_2$ be the points that satisfy those constraints, i.e.,

$$\mathcal{C}_2 = \{x \in \mathbb{R}^N \mid \sum_{n=1}^{N} a_{m(i),n} x_n \leq b_{m(i)}, i = 1, 2, \ldots, I\}.$$

It is clear that $\mathcal{H} \subset \mathcal{C}_2$. We show that $\mathcal{C}_1 = \mathcal{C}_2$, which implies that any point in $\mathcal{H}$ can be represented as the sum of x^0 and a conic combination of vectors $x^k - x^0$ for $k = 1, 2, \ldots, K$.

Theorem A.1. *The sets $\mathcal{C}_1$ and $\mathcal{C}_2$ coincide.*

Proof: By definition, $\mathcal{C}_2$ is a polyhedron. We can easily see that x^0 is the only extreme point of $\mathcal{C}_2$ as follows. For any extreme point of $\mathcal{C}_2$, there should be n linearly independent constraints of $\mathcal{C}_2$ that are binding at the extreme point. However, $\mathcal{C}_2$ has only n constraints, thus, x^0 is the only extreme point of $\mathcal{C}_2$.

We also show that $x^k - x^0$ for $k = 1, 2, \ldots, K$ form the complete set of extreme rays of $\mathcal{C}_2$. For $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, I$, and $\mu \geq 0$,

$$\begin{aligned}\sum_{n=1}^{N} a_{m(i),n}(x_n^0 + \mu(x_n^k - x_n^0)) &= (1-\mu)\sum_{n=1}^{N} a_{m(i),n} x_n^0 + \mu \sum_{n=1}^{N} a_{m(i),n} x_n^k \\ &\leq (1-\mu) b_{m(i)} + \mu b_{m(i)} = b_{m(i)}\end{aligned}$$

where the inequality is obtained from the fact that the constraint $m(i)$ is binding at x^0 and satisfied at x^k. Also, for $k = 1, 2, \ldots, K$, x^k has n binding constraints that are linearly

independent and share $n-1$ of them with x^0 because they are adjacent. Thus, $x^k - x^0$ satisfies $n-1$ linearly independent constraints of $\mathcal{C}_2$. This shows that $x^k - x^0$ for $k = 1, 2, \ldots, K$ are extreme rays of $\mathcal{C}_2$. We also can easily show that they are only extreme rays of $\mathcal{C}_2$ using the fact that $x^1, x^2, \ldots, x^K$ are the complete list of adjacent extreme points of x^0.
By the theorem of representation of polyhedra (e.g., Theorem 4.15 of [11]), $\mathcal{C}_1 = \mathcal{C}_2$. □

Corollary A.2. *The polytope $\mathcal{H}$ is contained in $\mathcal{C}_1$.*

A.2 In the Feasible Region of (NP)

In this section, we will prove a similar result to Corollary A.2 for $\mathcal{F}$, the feasible region of (NP).

For $n \in \mathbb{N}, s \in \mathcal{S}$, and $a \in \mathcal{A} \setminus \{a_n(s)\}$, let $e^{(n,s,a)}$ denote the extreme point adjacent to x whose basic action differs only at state s in period n and the basic action is a. Then we prove the following theorem which is a counterpart of Corollary A.2 for the infinite-dimensional set $\mathcal{F}$.

Theorem A.3. *Any point $z \in \mathcal{F}$ is represented as the sum of x and a conic combination of vectors $e^j - x$ for $j = 1, 2, \ldots$, i.e.,*

$$z = x + \sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \lambda_{n,s,a}(e^{(n,s,a)} - x) \tag{A.1}$$

where the convergence is pointwise (i.e., we consider the product topology of $\mathbb{R}^\infty$) and $\lambda \geq 0$.

To prove this theorem we first establish a special case of the theorem where z is an extreme point. Let $a_n(s)$ and $b_n(s)$ be basic actions of x and z at state s in period n, respectively.

Lemma A.4. *Any extreme point $z \in \mathcal{F}$ is represented as the sum of x and a conic combi-*

nation of vectors $e^{(n,s,a)} - x$ *for* $n \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$*, i.e.,*

$$z = x + \sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} \lambda_{n,s}(e^{(n,s,b_n(s))} - x) \tag{A.2}$$

where

$$\lambda_{n,s} = \frac{z_n(s, b_n(s))}{x_n(s, a_n(s))} \tag{A.3}$$

for $n \in \mathbb{N}$ *and* $s \in \mathcal{S}$*.*

Proof: Let

$$z^N \triangleq x + \sum_{n=1}^{N} \sum_{s \in \mathcal{S}} \lambda_{n,s}(e^{(n,s,b_n(s))} - x). \tag{A.4}$$

We will prove by induction that z^N and z coincide in periods up to N, i.e., $z_n^N(s,a) = z_n(s,a)$ for $n = 1, 2, \ldots, N, s \in \mathcal{S}, a \in \mathcal{A}$.

First, consider $z^1 = x + \sum_{s \in \mathcal{S}} \lambda_{1,s}(e^{(1,s,b_1(s))} - x)$. For $s, t \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$e_1^{(1,s,b_1(s))}(t,a) - x_1(t,a) = \begin{cases} 1 & \text{if } t = s \text{ and } a = b_1(s) \\ -1 & \text{if } t = s \text{ and } a = a_1(s) \\ 0 & \text{otherwise.} \end{cases}$$

Also, $\lambda_{1,s} = \frac{z_1(s,b_1(s))}{x_1(s,a_1(s))} = 1$ for $s \in \mathcal{S}$. Then, we can easily show that for $t \in \mathcal{S}$, $z_1^1(t, b_1(t)) = 1$ and $z_1^1(t,a) = 1$ for $a \neq b_1(t)$. Thus, z^1 and z coincide in period 1. Also, we can easily show that z^1 satisfies the flow balance constraints, (2.2) and (2.3). In addition, note that for a state s, $e^{(1,s,b_1(s))}$ and x have the same basic actions in periods bigger than 1, that is, $e_m^{(1,s,b_1(s))}(t,a) = x_m(t,a) = 0$ for $m \geq 2, t \in \mathcal{S}$, and any action $a \neq a_m(t)$. Thus, we also have $z_m^1(t,a) = 0$ for $m \geq 2, t \in \mathcal{S}$, and any action $a \neq a_m(t)$. From the flow balance constraints, we can easily see that $z_m^1(t, a_m(t))$ should be positive for $m \geq 2, t \in \mathcal{S}$, and thus, z^1 is nonnegative. Therefore, z^1 is a flow in the hypernetwork, whose basic actions are equal to those of z in period 1 and those of x in period bigger than 1 (but note that the flows of z^1

after period 1 may be different from x because the basic actions in period 1 are different).

Suppose that z^{k-1} is a flow whose basic actions are equal to those of z in period up to $k-1$ and those of x in periods from k. We have

$$z^k = x + \sum_{n=1}^{k} \sum_{s \in \mathcal{S}} \lambda_{n,s}(e^{(n,s,b_n(s))} - x) = z^{k-1} + \sum_{s \in \mathcal{S}} \lambda_{k,s}(e^{(k,s,b_k(s))} - x).$$

First, note that $e^{(k,s,b_k(s))}$ and x coincide in periods up to $k-1$. Thus, in periods up to $k-1$, z^{k-1} and z^k coincide, and therefore, z^k and z coincide. For $s, t \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$e_k^{(k,s,b_k(s))}(t,a) - x_k(t,a) = \begin{cases} x_k(s, a_k(s)) & \text{if } t = s \text{ and } a = b_k(s) \\ -x_k(s, a_k(s)) & \text{if } t = s \text{ and } a = a_k(s) \\ 0 & \text{otherwise} \end{cases}$$

where the first case is obtained from the fact that $e^{(k,s,b_k(s))}$ and x have the same incoming flow to the node (k, s). Thus, in period k,

$$\begin{aligned} z_k^k(t,a) &= z_k^{k-1}(t,a) + \sum_{s \in \mathcal{S}} \lambda_{k,s}(e_k^{(k,s,b_k(s))}(t,a) - x_k(t,a)) \\ &= z_k^{k-1}(t,a) + \frac{z_k(t, b_k(t))}{x_k(t, a_k(t))}(e_k^{(k,t,b_k(t))}(t,a) - x_k(t,a)) \end{aligned}$$

Since z^{k-1} is the same with z in periods up to $k-1$ and has the same basic actions with x in period k, the flow of z^{k-1} in period k is given as, for $t \in \mathcal{S}$, $z_k^{k-1}(t, a_k(t)) = z_k(t, b_k(t))$ and $z_k^{k-1}(t,a) = 0$ for $a \neq a_k(t)$. Then, we can easily check that for $t \in \mathcal{S}$, $z_k^k(t, b_k(t)) = z_k(t, b_k(t))$ and $z_k^k(t,a) = z_k(t,a) = 0$ for $a \neq b_k(t)$. That is, z^k and z coincides in period k. Using similar arguments we used for z^1, it is straightforward to show that z^k is a flow and its basic actions in period bigger than k are equal to those of x.

By induction, we showed that z^N converges to z as $N \to \infty$.

Proof of Theorem A.3: $\mathcal{F}$ is compact and convex (see [24]). Thus, Krein-Milman

theorem (e.g., see Theorem 7.68 of [2]) gives us every feasible point as a convex combination of extreme points. Hence it suffices to show Theorem A.3) for extreme point z, thus the theorem is proven.

Using Theorem A.3, we can prove the following proposition.

Proposition A.5. *For any extreme point z of (D), f satisfies*

$$f(z) = f(x) + \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \lambda_{n,s}\big(f(e^{(n,s,b_n(s))}) - f(x)\big).$$

Proof: Define z^N as in (A.4). We showed $z^N \to z$ as $N \to \infty$. We know that z and z^N coincide in periods up to N. Therefore,

$$\begin{aligned}
|f(z) - f(z^N)| &= |\sum_{n=N+1}^{\infty} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a)(z_n(s,a) - z_n^N(s,a))| \\
&\leq \sum_{n=N+1}^{\infty} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} |c_n(s,a)||z_n(s,a) - z_n^N(s,a)| \\
&\leq \sum_{n=N+1}^{\infty} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} cnS
\end{aligned}$$

where the last inequality is obtained from Lemma 2.1. Thus, $f(z^N) \to f(z)$ as $N \to \infty$. By linearity of f,

$$f(z^N) = f(x + \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} (e^{(n,s,a_n(s))} - x)) = f(x) + \sum_{n=1}^{N} \sum_{s\in\mathcal{S}} (f(e^{(n,s,a_n(s))}) - f(x))$$

and by taking $N \to \infty$ on both sides,

$$f(z) = f(x) + \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} (f(e^{(n,s,a_n(s))}) - f(x)).$$

□

Another proof of Lemma 2.18: From Proposition A.5, we have

$$f^\star = f(x^\star) = f(x) + \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \lambda_{n,s}\big(f(e^{(n,s,a_n^\star(s))}) - f(x)\big)$$

where $\lambda_{n,s} = \frac{x_n^\star(s,a_n^\star(s))}{x_n(s,a_n(s))}$, $x^\star$ is an optimal basic feasible solution, and $a_n(s)$ and $a_n^\star(s)$ are the basic actions of x and $x^\star$ at state s in period n, respectively. By Proposition 2.10, we have $f(e^{(n,s,a_n^\star(s))}) - f(x) = x_n(s,a_n(s))\gamma_n(s,a_n^\star(s))$ (recall Definition 2.8). Therefore,

$$f(x) - f^\star = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} x_n^\star(s,a_n^\star(s))(-\gamma_n(s,a_n^\star(s)))$$

and the lemma is proven. □

APPENDIX B

Proof for Theorem 3.14 using Bauer's Maximum Principle

In this section, we present an alternative proof of Theorem 3.14 using Bauer's Maximum Principle (e.g., see Theorem 7.69 of [2]). A similar proof for a more general class of MDPs than countable-state MDPs is available in Section 7 and 8 of [4] but our proof is much simpler because it is focused on discrete state space and finite action space. While proving the theorem, we will use the integral version of dominated convergence theorem (DCT) which is called Lebesgue DCT and more general than Proposition 3.8. Also, we will use a generalization of the Lebesgue DCT.

Proposition B.1 (Lebesgue DCT, Theorem 16 on page 267 of [46]). *Let $\bar{h}$ be an integrable function over E, and suppose that $\{h_n\}$ is a sequence of measurable functions such that $|h_n(x)| \leq \bar{h}(x)$ on E, and $h_n(x) \to h(x)$ a.e. on E. Then*

$$\int_E h = \lim \int_E h_n.$$

Proposition B.2 (Generalized DCT, Proposition 18 on page 270 of [46]). *Let $(X, \mathcal{B})$ be a measurable space and $\{\mu_n\}$ a sequence of measures on $\mathcal{B}$ which converge setwise to a measure*

μ. Let $\{h_n\}$ and $\{\bar{h}_n\}$ be two sequences of measurable functions which converge pointwise to h and $\bar{h}$, respectively. Suppose $|h_n| \leq \bar{h}_n$ and that

$$\lim \int \bar{h}_n \, d\mu_n = \int \bar{h} \, d\mu < \infty.$$

Then

$$\lim \int h_n \, d\mu_n = \int h \, d\mu.$$

We will show that $\mathcal{P}$, the feasible region of (CP), is a compact subset of $\mathbb{R}^\infty$ under the product topology and the objective function $f(x)$ of (CP) is continuous over the feasible region $\mathcal{P}$. It is straightforward to show that $\mathcal{P}$ is convex. Therefore, by Bauer's Maximum Principle, we will obtain that (CP) has an extreme point optimal solution. We present this proof by using only the absorbing MDP formulation defined in Section 3.5.1. Recall that 0 denotes the absorbing state. Let $\mathcal{S} = \{0, 1, 2, \ldots\}$ and extend w by letting $w(0) = 0$.

To show that $\mathcal{P}$ is compact in $\mathbb{R}^\infty$ and that $f(x)$ is continuous on $\mathcal{P}$, we first define a topology on Π_S, the set of stationary policies. Let $M(\mathcal{A})$ be the set of probability measures on the finite action set $\mathcal{A}$. Then, $M(\mathcal{A}) = [0,1]^A \cap \{p \in \mathbb{R}^A : \mathbf{1}^T p = 1\}$ (which is often called the unit A-simplex) and this is compact in $\mathbb{R}^A$. A stationary policy can be considered as assigning a probability measure on $\mathcal{A}$ to each state, and thus, Π_S is represented as $\prod_{s=1}^{\infty} M(\mathcal{A}) \subset (\mathbb{R}^A)^\infty$ and $(\mathbb{R}^A)^\infty$ is trivially isomorphic to $\mathbb{R}^\infty$. By Tychonoff's theorem (e.g., see Theorem 2.61 of [2]), Π_S is compact under the product topology of $\mathbb{R}^\infty$. Also, note that $\mathbb{R}^\infty$ equipped with the product topology is metrizable (e.g., see Theorem 20.5 of [36]), so a function f from Π_S is continuous if and only if for any sequence $\{\sigma^m\} \subset \Pi_S$ converging to $\bar{\sigma} \in \Pi_S$, $\{f(\sigma^m)\}$ converges to $f(\bar{\sigma})$ (e.g., see Theorem 21.3 of [36]).

Consider a function $Q^\beta_{(\cdot)} : \Pi_S \to \mathbb{R}^\infty$ that maps a stationary policy σ to its occupancy measure $Q^\beta_\sigma \in \mathcal{P}$. We call $Q^\beta_{(\cdot)}$ the occupancy measure function. By (3.30) and the definition of occupancy measure (3.29), we can easily show that $Q^\beta_{(\cdot)}$ is a one-to-one mapping onto $\mathcal{P} \subset \mathbb{R}^\infty$. Since we know Π_S is compact, proving continuity of $Q^\beta_{(\cdot)}$ is sufficient to show the

compactness of $\mathcal{P}$. Toward the end, we first prove the following two lemmas.

Lemma B.3. *For $s \in \mathcal{S}$, an initial distribution β, and a nonnegative integer n, $P_\sigma^\beta(S_n = s)$ is a continuous function of σ on Π_S.*

Proof: We first show that $P_\sigma^t(S_n = s)$ is continuous for $t \in \mathcal{S}$ by induction on n. For $n = 1$, $P_\sigma^t(S_1 = s) = \mathbf{1}\{t = s\}$, which is a constant, so continuous. Suppose that $P_\sigma^t(S_m = s) = \alpha^{m-1} P_\sigma^{m-1}(s|t)$ is continuous for some positive integer m. We have

$$P_\sigma^t(S_{m+1} = s) = \alpha^m P_\sigma^m(s|t) = \sum_{t' \in \mathcal{S}} \alpha P_\sigma(t'|t) \alpha^{m-1} P_\sigma^{m-1}(s|t').$$

Consider a sequence of stationary policies $\{\sigma^k\}$ converging to $\bar{\sigma}$. Because of the induction hypothesis,

$$\alpha P_{\sigma^k}(t'|t) \alpha^{m-1} P_{\sigma^k}^{m-1}(s|t') \to \alpha P_{\bar{\sigma}}(t'|t) \alpha^{m-1} P_{\bar{\sigma}}^{m-1}(s|t') \text{ as } k \to \infty.$$

Also, for any $\sigma \in \Pi_S$, $|\alpha P_\sigma(t'|t) \alpha^m P_\sigma^m(s|t')| \le \alpha P_\sigma(t'|t)$ and $\sum_{t' \in \mathcal{S}} \alpha P_\sigma(t'|t) = \alpha$. Thus, by Proposition B.1,

$$\lim_{k \to \infty} \sum_{t' \in \mathcal{S}} \alpha P_{\sigma^k}(t'|t) \alpha^m P_{\sigma^k}^m(s|t') = \sum_{t' \in \mathcal{S}} \alpha P_{\bar{\sigma}}(t'|t) \alpha^m P_{\bar{\sigma}}^m(s|t'),$$

that is, $P_\sigma^t(S_{m+1} = s)$ is continuous in σ. Therefore, $P_\sigma^t(S_n = s)$ is continuous for $t \in \mathcal{S}$ and for any n. Then, we can easily show that $P_\sigma^\beta(S_n = s)$ is continuous for a general initial state distribution β by using Proposition B.1 or Proposition 3.7. □

Lemma B.4. *For $s \in \mathcal{S}$, $\sum_{t \in \mathcal{S}} P_\sigma(t|s) w(t)$ is a continuous function of σ on Π_S.*

Proof: Consider a sequence $\{\sigma^k\}$ in Π_S that converges to $\bar{\sigma} \in \Pi_S$. For any k, we have

$$\sum_{t \in \mathcal{S}} P_\sigma(t|s) w(t) = \sum_{t \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sigma^k(a|s) p(t|s,a) w(t) = \sum_{a \in \mathcal{A}} \sigma^k(a|s) \sum_{t \in \mathcal{S}} p(t|s,a) w(t) \tag{B.1}$$

where the exchange of sums follows by Proposition 3.5. By the assumption A2, we know that $\sum_{t\in\mathcal{S}} p(t|s,a)w(t)$ is finite for any (s,a). Since $\lim_{k\to\infty}\sigma^k(a|s) = \bar{\sigma}(a|s)$ for all (s,a) and $\mathcal{A}$ is a finite set, (B.1) converges to

$$\sum_{a\in\mathcal{A}} \bar{\sigma}(a|s) \sum_{t\in\mathcal{S}} p(t|s,a)w(t)$$

as $k\to\infty$, and thus, the lemma is proven. □

Lemma B.5. *For $s\in\mathcal{S}$, an initial distribution β, and a positive integer n, $E^\beta_\sigma[w(S_n)] = \sum_{s\in\mathcal{S}} P^\beta_\sigma(S_n = s)w(s)$ is a continuous function of σ on Π_S.*

Proof: Proof of this lemma is similar to that of the previous lemma. First, for $t\in\mathcal{S}$, we show that $E^t_\sigma[w(S_n)]$ is continuous by induction on n. For $n=1$, $E^t_\sigma[w(S_1)] = w(t)$, which is a constant. Suppose that $E^t_\sigma[w(S_m)]$ is continuous for a positive integer m. We have

$$E^t_\sigma[w(S_{m+1})] = \sum_{t'\in\mathcal{S}} \alpha P_\sigma(t'|t) E^{t'}_\sigma[w(S_m)]. \tag{B.2}$$

Because of the induction hypothesis, each term of the above sum is continuous. By the assumption A2,

$$\alpha P_\sigma(t'|t) E^{t'}_\sigma[w(S_m)] \le \alpha\kappa^{m-1} P_\sigma(t'|t)w(t').$$

By Lemma B.4,

$$\sum_{t'\in\mathcal{S}} \alpha\kappa^{m-1} P_\sigma(t'|t)w(t') = \alpha\kappa^{m-1}\sum_{t'\in\mathcal{S}} P_\sigma(t'|t)w(t')$$

is continuous in σ. Then, by Proposition B.2, (B.2) is continuous in σ. By induction, we conclude that $E^t_\sigma[w(S_n)]$ is continuous in σ for any n. Since

$$E^\beta_\sigma[w(S_n)] = \sum_{t\in\mathcal{S}} \beta(t) E^t_\sigma[w(S_n)],$$

it is also easy to show that $E^\beta_\sigma[w(S_n)]$ is continuous in σ by using Assumption A2 and (3.7), and applying Proposition B.1. □

The next lemma shows that $Q^{\beta}_{(\cdot)}$ is weakly continuous in the following sense.

Lemma B.6. *For any real-valued function r on $\mathcal{S}\times\mathcal{A}$ satisfying $|r(s,a)|\le Lw(s)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, the function $\langle r, Q^{\beta}_{(\cdot)}\rangle : \Pi_S \to \mathbb{R}$ defined as*

$$\langle r, Q^{\beta}_{\sigma}\rangle \triangleq \sum_{s=1}^{\infty}\sum_{a=1}^{A} r(s,a)Q^{\beta}_{\sigma}(s,a) \tag{B.3}$$

is a continuous function of σ on Π_S.

Proof: By (3.32), we have

$$\langle r, Q^{\beta}_{\sigma}\rangle = \sum_{s=1}^{\infty}\sum_{a=1}^{A} r(s,a)Q^{\beta}_{\sigma}(s,a) = V_{\sigma}(\beta) = E^{\beta}_{\sigma}\big[\sum_{n=1}^{\infty} r(S_n,A_n)\big]. \tag{B.4}$$

Then

$$\begin{aligned}
E^{\beta}_{\sigma}\big[\sum_{n=1}^{\infty}|r(S_n,A_n)|\big] =& \beta^T[|r_\sigma| + \alpha P_\sigma|r_\sigma| + \alpha^2 P^2_\sigma|r_\sigma| + \ldots] \\
\le& \beta^T[w + \alpha\kappa w + \alpha^2\kappa^2 w + \ldots + \alpha^{J-1}\kappa^{J-1}w \\
&+ \lambda w + \lambda\alpha\kappa w + \lambda\alpha^2\kappa^2 w + \ldots + \lambda\alpha^{J-1}\kappa^{J-1}w \\
&+ \ldots] \\
=& L\beta^T w < \infty,
\end{aligned}$$

where the inequality follows by considering groups of J terms and applying Assumptions A2 and A3. Therefore, we have

$$E^{\beta}_{\sigma}\big[\sum_{n=1}^{\infty} r(S_n,A_n)\big] = \sum_{n=1}^{\infty} E^{\beta}_{\sigma}[r(S_n,A_n)]$$

by Fubini-Tonelli theorem, which is the combination of Fubini's theorem (see Theorem 19 on page 307 of [46]) and Tonelli's theorem (see Theorem 20 on page 309 of [46]), and is a more general version of Proposition 3.6.

We will show that the partial sums $\sum_{n=1}^{N} E_\sigma^\beta[r(S_n, A_n)]$ converges uniformly to the infinite sum on Π_S as $N \to \infty$. Let

$$V_\sigma^N(\beta) = \sum_{n=N+1}^{\infty} E_\sigma^\beta[r(S_n, A_n)],$$

and let V_σ^N denote the infinite vector indexed by states whose value at s is $V_\sigma^N(s)$. Let $N = kJ + l$ for some nonnegative integers k and $l \leq J$, then

$$\begin{aligned}
|V_\sigma^N| \leq & \alpha^N P_\sigma^N |r_\sigma| + \alpha^{N+1} P_\sigma^{N+1} |r_\sigma| + \dots \\
= & \alpha^{kJ+l} P_\sigma^{kJ+l} |r_\sigma| + \alpha^{kJ+l+1} P_\sigma^{kJ+l+1} |r_\sigma| + \dots \\
\leq & \alpha^l P_\sigma^l (\alpha^J P_\sigma^J)^k w + \alpha^{l+1} P_\sigma^{l+1} (\alpha^J P_\sigma^J)^k w + \dots + \alpha^{J-1} P_\sigma^{J-1} (\alpha^J P_\sigma^J)^k w \\
& + (\alpha^J P_\sigma^J)^{k+1} w + \alpha P_\sigma (\alpha^J P_\sigma^J)^{k+1} w + \dots + \alpha^{J-1} P_\sigma^{J-1} (\alpha^J P_\sigma^J)^{k+1} w \\
& + (\alpha^J P_\sigma^J)^{k+2} w + \alpha P_\sigma (\alpha^J P_\sigma^J)^{k+2} w + \dots + \alpha^{J-1} P_\sigma^{J-1} (\alpha^J P_\sigma^J)^{k+2} w \\
& + \dots \\
\leq & \alpha^l \kappa^l \lambda^k w + \alpha^{l+1} \kappa^{l+1} \lambda^k w + \dots + \alpha^{J-1} \kappa^{J-1} \lambda^k w \\
& + \lambda^{k+1} w + \alpha\kappa \lambda^{k+1} w + \dots + \alpha^{J-1} \kappa^{J-1} \lambda^{k+1} w \\
& + \lambda^{k+2} w + \alpha\kappa \lambda^{k+2} w + \dots + \alpha^{J-1} \kappa^{J-1} \lambda^{k+2} w \\
& + \dots \\
= & \left\{ [\lambda^k (\alpha\kappa)^l [1 + \alpha\kappa + \dots + (\alpha\kappa)^{J-l-1}] + \frac{\lambda^{k+1}}{1-\lambda} [1 + \alpha\kappa + \dots + (\alpha\kappa)^{J-1}] \right\} w,
\end{aligned}$$

where the second inequality follows by applying Assumption A1 and considering a group the first $J - l$ terms and groups of J terms for the rest; and the third inequality follows by Assumptions A2 and A3. Therefore,

$$\begin{aligned}
|V_\sigma^N(\beta)| &= |\beta^T V_\sigma^N| \\
&\leq \left\{ [\lambda^k (\alpha\kappa)^l [1 + \alpha\kappa + \dots + (\alpha\kappa)^{J-l-1}] + \frac{\lambda^{k+1}}{1-\lambda} [1 + \alpha\kappa + \dots + (\alpha\kappa)^{J-1}] \right\} \beta^T w
\end{aligned}$$

In the last expression, $\beta^T w$ is finite by (3.7) and its preceding multiplier tends to zero as $k \to \infty$. That is, given $\epsilon > 0$, for each $0 \leq l \leq J-1$, there exists k_l such that for $k \geq k_l$, we have $|V_\sigma^{kJ+l}(\beta)| < \epsilon$ for any $\sigma \in \Pi_S$. Then, for $N \geq \max_{0 \leq l \leq J-1}\{k_l J + l\}$, we have $|V_\sigma^N(\beta)| < \epsilon$ for any $\sigma \in \Pi_S$. Therefore, as $N \to \infty$, the partial sums $\sum_{n=0}^N E_\sigma^\beta[r(S_n, A_n)]$ converges to the infinite sum uniformly on Π_S.

Secondly, we will show that $E_\sigma^\beta[r(S_n, A_n)]$ is continuous in σ for any n, then it will imply that (B.3) is continuous on Π_S. By conditioning on state, we have

$$E_\sigma^\beta[r(S_n, A_n)] = \sum_{s \in \mathcal{S}} P_\sigma^\beta(S_n = s) r(s, \sigma) \tag{B.5}$$

where $r(s, \sigma) = \sum_{a=1}^A \sigma(a|s) r(s, a)$. It is easy to show that $r(s, \sigma)$ is continuous in σ. By Lemma B.3, $P_\sigma^\beta(S_n = s) r(s, \sigma)$ is also continuous in σ. By the assumption A1, $|P_\sigma^\beta(S_n = s) r(s, \sigma)| \leq P_\sigma^\beta(S_n = s) w(s)$, and by Lemma B.5, $\sum_{s \in \mathcal{S}} P_\sigma^\beta(S_n = s) w(s)$ is continuous in σ. Therefore, by Proposition B.2, (B.5) is continuous in σ. Thus, the partial sum $\sum_{n=1}^N E_\sigma^\beta[r(S_n, A_n)]$ is continuous in σ. Since the partial sum converges to the infinite sum uniformly over σ,

$$\langle r, Q_\sigma^\beta \rangle = \sum_{n=0}^\infty E_\sigma^\beta[r(S_n, A_n)]$$

is continuous in σ. □

Using the weak continuity of $Q_{(\cdot)}^\beta$, it is easy to show that it is actually continuous in the product topology.

Lemma B.7. *The occupancy measure function $Q_{(\cdot)}^\beta : \Pi_S \to \mathbb{R}^\infty$ is continuous.*

Proof: By the previous lemma, we know that $\langle r, Q_{(\cdot)}^\beta \rangle$ is continuous over Π_S for any r such that $|r(s, a)| \leq L w(s)$ for $s \in \mathcal{S}$. Consider a sequence $\{\sigma^m\} \subset \Pi_S$ converging to $\bar{\sigma} \in \Pi_S$. Fix an arbitrary state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Let $r(s, a) = w(s)$ and $r(t, b) = 0$ for $(t, b) \neq (s, a)$. Then, by the weak continuity, $\langle r, Q_{\sigma^m}^\beta \rangle = w(s) Q_{\sigma^m}^\beta(s, a)$ converges to $\langle r, Q_{\bar{\sigma}}^\beta \rangle = w(s) Q_{\bar{\sigma}}^\beta(s, a)$, i.e., $Q_{\sigma^m}^\beta(s, a)$ converges to $Q_{\bar{\sigma}}^\beta(s, a)$ for an arbitrary state-action

pair (s, a), and therefore, $Q^{\beta}_{\sigma^m}$ converges to Q^{β}_{σ} in the product topology of $\mathbb{R}^{\infty}$. Since the product topology of $\mathbb{R}^{\infty}$ is metrizable, the function $Q^{\beta}_{(\cdot)}$ is continuous. □

Since $\mathcal{P} = \mathcal{Q}_S$ is the image of Π_S by $Q^{\beta}_{(\cdot)}$, the compactness of Π_S and the above lemma implies the following corollary.

Corollary B.8. *$\mathcal{P}$ is a compact subset of $\mathbb{R}^{\infty}$.*

Using Lemma B.6, we can easily prove the next lemma by which we conclude the proof of Theorem 3.14.

Lemma B.9. *The objective function $f(x)$ of (CP) is continuous in its feasible region $\mathcal{P}$.*

Proof: For any $x \in \mathcal{P}$, there exists a unique stationary policy σ defined by (3.30) such that $Q^{\beta}_{\sigma} = x$. Let $\xi : \mathcal{P} \to \Pi_S$ denote the function that maps an occupancy measure $x \in \mathcal{P}$ to its corresponding stationary policy, i.e., $\xi(x) \triangleq \sigma$. Then,

$$f(x) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a)x(s,a) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a) Q^{\beta}_{\xi(x)}(s,a) = \langle r, Q^{\beta}_{\xi(x)} \rangle.$$

By Lemma B.6, $\langle r, Q^{\beta}_{(\cdot)} \rangle$ is continuous on Π_S. If ξ is continuous, then f is also continuous on x. To show that ξ is continuous, consider a sequence $\{x^m\}$ in $\mathcal{P}$ converging to $\bar{x} \in \mathcal{P}$, that is, $x^m(s,a)$ converges to $\bar{x}(s,a)$ for each state-action pair (s,a). Then the corresponding stationary policies $\sigma^m = \xi(x^m)$ for $m = 1, 2, \ldots$ and $\bar{\sigma} = \xi(\bar{x})$ are computed by, respectively,

$$\sigma^m(a|s) = \frac{x^m(s,a)}{\sum_{b=1}^{A} x^m(s,b)} \text{ and } \bar{\sigma}(a|s) = \frac{\bar{x}(s,a)}{\sum_{b=1}^{A} \bar{x}(s,b)}.$$

Therefore, $\sigma^m(a|s)$ also converges to $\bar{\sigma}(a|s)$ for each state-action pair (s,a), and thus, $\sigma^m = \xi(x^m)$ converges to $\bar{\sigma} = \xi(\bar{x})$ in the product topology. We showed that ξ is continuous, and thus, f is continuous in x. □

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] O. Alagoz, H. Hsu, A. J. Schaefer, and M. S. Roberts. Markov decision processes: A tool for sequential decision making under uncertainty. *Medical Decision Making*, 30(4):474–483, 2010.

[2] C. Aliprantis and K. Border. *Infinite-dimensional analysis: a hitchhiker's guide*. Springer-Verlag, Berlin, Germany, 1994.

[3] E. Altman. Denumerable constrained Markov decision processes and finite approximations. *Mathematics of Operations Research*, 19:169–191, 1994.

[4] E. Altman. *Constrained Markov decision processes*. Chapman and Hall, CRC, 1998.

[5] E. Altman and A. Shwartz. Optimal priority assignment: a time sharing approach. *IEEE Trans. on Auto. Control*, AC-34:1089–1102, 1989.

[6] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. John Wiley and Sons, Chichester, UK, 1987.

[7] K. Baker. An experimental study of the effectiveness of rolling schedules in production planning. *Decision Science*, 8:19–27, 1977.

[8] J. Bean, J. Lohmann, and R. L. Smith. A dynamic infinite horizon replacement economy decision model. *The Engineering Economist*, 30(2):99–120, 1984.

[9] J. Bean and R. L. Smith. Optimal capacity expansion over an infnite horizon. *Management Science*, 31:1523–1532, 1985.

[10] D. P. Bertsekas. *Dynamic programming and Optimal Control*. Athena Scientific, 1995.

[11] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, Belmon, MA, USA, 1997.

[12] E. Byon and Y. Ding. Season-dependent condition-based maintenance for a wind turbine using a partially observed markov decision process. *Power Systems, IEEE Transactions on*, 25(4):1823–1834, 2010.

[13] R. Cavazos-Cadena. Finite-state approximations for denumerable state discounted Markov decision processes. *Applied Mathematics and Optimization*, 14:1–26, 1986.

[14] W. Cook, C. Field, and M. Kirby. Infinite linear programming in games with partial information. *Operations Research*, 23(5):996–1010, 1975.

[15] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1963.

[16] E. A. Feinberg and U. G. Rothblum. Splitting randomized stationary policies in total-reward Markov decision processes. *Mathematics of Operations Research*, 37:129–153, 2012.

[17] E. A. Feinberg and A. Shwartz. Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21:922–945, 1996.

[18] E. A. Feinberg and A. Shwartz. *Handbook of Markov decision processes*. Kluwer International Series, 2002.

[19] E. B. Frid. On optimal strategies in control problems with constraints. *Theory of Probability and Its Applications*, 17:188–192, 1972.

[20] A. Ghate. Duality in countably infinite linear programs. 2014. Working paper.

[21] A. Ghate, D. Sharma, and R. L. Smith. A shadow simplex method for infinite linear programs. *Operations Research*, 58:865–877, 2010.

[22] A. Ghate and R. L. Smith. Characterizing extreme points as basic feasible solutions in infinite linear programs. *Operations Research Letters*, 33:7–10, 2009.

[23] A. Ghate and R. L. Smith. Optimal backlogging over an infinite horizon under time-varying convex production and inventory costs. *Manufacturing & Service Operations Management*, 11(2):362–368, 2009.

[24] A. Ghate and R. L. Smith. A linear programming approach to nonstationary infinite-horizon Markov decision processes. *Operations Research*, 61:413–425, 2013.

[25] K. Golabi, R. B. Kulkarni, and G. B. Way. A statewide pavement management system. *Interfaces*, 12:5–21, 1982.

[26] J. González-Hernández and O. Hernández-Lerma. Extreme points of sets of randomized strategies in constrained optimization and control problems. *SIAM Journal on Optimization*, 15:1085–1104, 2005.

[27] J. Harrison. Discrete dynamic programming with unbounded rewards. *Annals of Mathematical Statistics*, 43:636–644, 1972.

[28] O. Hernández-Lerma. Finite-state approximations for denumerable multidimensional state discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 113:382–389, 1986.

[29] D. P. Heyman and M. J. Sobel. *Stochastic models in operations research. Vol 2: Stochastic optimization*. McGraw-Hill, N.Y., 1984.

[30] A. Hordijk and F. Spieksma. Constrained admission control to a queueing system. *Adv. Appl. Probab.*, 21:409–431, 1989.

[31] R. A. Howard. *Dynamic programming and Markov processes*. MIT, Cambridge, MA., 1960.

[32] L. C. M. Kallenberg. Linear programming and finite Markovian control problems. *Mathematical Centre Tracts*, 148:1–245, 1983.

[33] J. Kim and W. Powell. Optimal energy commitments with storage and intermittent supply. *Operations Research*, 59(6):1347 – 1360, 2011.

[34] A. Lazar. Optimal flow control of a class of queueing networks in equilibrium. *IEEE Trans. on Auto. Control*, 28:1001–1007, 1983.

[35] S. Lippman. On dynamic programming with unbounded rewards. *Management Science*, 21:1225–1233, 1975.

[36] J. Munkres. *Topology*. Prentice Hall, Inc., 2000.

[37] P. Nain and K. W. Ross. Optimal priority assignment with hard constraint. *IEEE Trans. on Auto. Control*, 31:883–888, 1986.

[38] A. Piunovskiy. Controlled random sequences: methods of convex analysis and problems with functional constraints. *Russian Mathematical Surveys*, 53:1233–1293, 1998.

[39] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley and Sons, New York, NY, USA, 1994.

[40] S. Resnick. *A Probability Path*. Birkhäuser, 1999.

[41] H. E. Romeijn and R. L. Smith. Shadow prices in infinite dimensional linear programming. *Mathematics of Operations Research*, 23:239–256, 1998.

[42] H. E. Romeijn, R. L. Smith, and J. Bean. Duality in infinite dimensional linear programming. *Mathematical Programming*, 53:79–97, 1992.

[43] K. W. Ross. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research*, 37:474–477, 1989.

[44] K. W. Ross and B. Chen. Optimal scheduling of interactive and noninteractive traffic in telecommunications systems. *IEEE Trans. on Auto. Control*, 33:261–267, 1988.

[45] S. Ross. *Introduction to stochastic dynamic programming*. Academic Press, New York, NY, USA, 1983.

[46] H. Royden. *Real Analysis*. Macmillan Publishing Company, New York, 1988.

[47] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., 1976.

[48] A. J. Schaefer, M. D. Bailey, S. M. Shechter, and M. S. Roberts. Modeling medical treatment using Markov decision processes. In *Operations Research and Health Care: A Handbook of Methods and Applications*. Springer, 2004.

[49] L. I. Sennott. Constrained discounted Markov decision chains. *Probability in the Engineering and Informational Sciences*, 5:463–475, 1991.

[50] T. C. Sharkey and H. E. Romeijn. A simplex algorithm for minimum cost network flow problems in infinite networks. *Networks*, 52:14–31, 2008.

[51] S. Siegrist. *A Complementary Approach to Multistage Stochastic Linear Programs*. PhD thesis, University of Zürich, 2006.

[52] R. L. Smith and R. Zhang. Infinite horizon production planning in time varying systems with convex production and inventory costs. *Management Science*, 44:1313–1320, 1998.

[53] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions 1*. Springer-Verlag, New York, 1970.

[54] C. T. and R. L. Smith. Infinite horizon production scheduling in time-varying systems under stochastic demand. *Operations Research*, 52:105–115, 2004.

[55] J. Wessels. Markov programming by successive approximations with respect to weighted supremum norms. *Journal of Mathematical Analysis and Applications*, 58:326–335, 1977.

[56] D. White. Finite state approximations for denumerable-state infinite horizon contracted Markov decision processes: The policy space method. *Journal of Mathematical Analysis and Applications*, 72:512–523, 1979.

[57] D. White. Finite state approximations for denumerable state infinite horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 74:292–295, 1980.

[58] D. White. Finite state approximations for denumerable state infinite horizon discounted markov decision processes: The method of successive approximations. In R. Hartley, L. Thomas, and D. White, editors, *Recent Developments in Markov Decision Processes*. Academic Press, New York, 1980.

[59] D. White. Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards. *Journal of Mathematical Analysis and Applications*, 86:292–306, 1982.

[60] D. White. A survey of applications of markov decision processes. *The Journal of the Operational Research Society*, 44(11):1073–1096, 1993.

[61] C. V. Winden and R. Dekker. Markov decision models for building maintenance: A feasibility study. *Journal of the Operations Research Society*, 49:928–935, 1998.

[62] Y. Ye. The Simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36:593–603, 2011.