Simplex Algorithm for Countable-state Discounted Markov Decision Processes*

Ilbin Lee Marina A. Epelman H. Edwin Romeijn Robert L. Smith

November 16, 2014

Abstract

We consider discounted Markov Decision Processes (MDPs) with countably-infinite state spaces, finite action spaces, and unbounded rewards. Typical examples of such MDPs are inventory management and queueing control problems in which there is no specific limit on the size of inventory or queue. Existing solution methods obtain a sequence of policies that converges to optimality in value but may not improve monotonically, i.e., a policy in the sequence may be worse than preceding policies. Our proposed approach considers countably-infinite linear programming (CILP) formulations of the MDPs (a CILP is defined as a linear program (LP) with countably-infinite numbers of variables and constraints). Under standard assumptions for analyzing MDPs with countably-infinite state spaces and unbounded rewards, we extend the major theoretical extreme point and duality results to the resulting CILPs. Under an additional technical assumption which is satisfied by several applications of interest, we present a simplex-type algorithm that is implementable in the sense that each of its iterations requires only a finite amount of data and computation. We show that the algorithm finds a sequence of policies which improves monotonically and converges to optimality in value. Unlike existing simplex-type algorithms for CILPs, our proposed 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 numerical illustration for inventory management problems is also presented.

1 Introduction

The class of Markov decision processes (MDPs) provides a popular framework which covers a wide variety of sequential decision-making problems. An MDP is classified by its criterion being optimized, its state and action space, whether the problem data is stationary, etc., and those different settings are mostly motivated by applications. For example, in inventory management and queueing control, often times there is no specific limit on the size of inventory or queue, and such problems can be modeled by MDPs with countable state space.

We consider MDPs whose objective is to maximize expected total discounted reward, and that have a countable state space, a finite action space, and unbounded immediate rewards. Specifically, consider a dynamic system that evolves over discrete time periods. In time periods $n = 0, 1, \ldots$, a decision-maker observes the current state of the system $s \in \mathcal{S}$, where the set of states $\mathcal{S}$ is countably-infinite. The decision-maker then chooses an action $a \in \mathcal{A}$, where the action set $\mathcal{A}$ is a finite set. Given that action a is taken in state s, the system makes a transition to a next state

*Submitted to *Operations Research*; preliminary version.

$t \in \mathcal{S}$ with probability $p(t|s,a)$ and reward $r(s,a,t)$ is obtained. Let $r(s,a)$ denote the expected reward incurred by choosing action a at state s, i.e., $r(s,a) = \sum_{t\in\mathcal{S}} p(t|s,a)r(s,a,t)$. A policy is defined as a decision rule that dictates which action to execute and the rule may in general be conditioned on current state, time, and the whole history of visited states and actions taken. The goal is to find a policy that maximizes the expected total discounted reward, with discount factor $\alpha \in (0,1)$, over the infinite-horizon for any starting state. In this paper, we refer to this problem as a *countable-state MDP* for short.

For a policy π, $V_\pi(s)$ denotes the expected total discounted reward obtained by executing π starting from state s, and we call V_π the value function of policy π. For $s \in \mathcal{S}$, let $V^\star(s)$ denote the supremum of $V_\pi(s)$ over all policies π; then $V^\star$ is the optimal value function. Thus, the goal of the MDP problem can be rephrased as finding a policy whose value function coincides with the optimal value function. For more precise definitions, see Section 2.1.

1.1 Motivation and Contribution

Countable-state MDPs were studied by many researchers, including [4, 10, 12, 21, 22, 23], with predominant solution methods summarized as the three algorithms in [21, 22] and [23]. We will review these in Section 1.2 in detail. Each of the three algorithms computes a sequence of real-valued functions on $\mathcal{S}$ that converges to the optimal value function $V^\star$. Also, for the algorithms in [21] and [23], methods to obtain a sequence of policies whose value functions converge pointwise to the optimal value function were provided as well. However, the value function of policies obtained by these two methods may not improve in every iteration. In other words, a policy obtained in a later iteration may be worse (for some starting states) 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 monotonicity of value functions of obtained policies, the value functions of those policies should be exactly computed in order to find the best one. However, computing the value function 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 value functions of obtained policies is guaranteed to be monotone, then the last obtained policy is always guaranteed to be the best one so far. The key motivation of this paper is to develop an algorithm that finds a sequence of policies whose value functions improve in every iteration and converge to the optimal value function. We propose an algorithm that has both of the characteristics under the assumptions considered in [12] to analyze countable-state MDPs with unbounded rewards (introduced in Section 2.1) and an additional technical assumption (introduced at the end of Section 4.1), which are satisfied by application examples of interest as we show in this paper.

Our algorithm is a simplex-type algorithm for solving a linear programming (LP) formulation of countable-state MDPs. 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 MDPs with finite state space, can be viewed as the simplex method applied to an equivalent LP formulation of the MDP. A recent result in [25] showed that for finite-state MDPs, simplex method with Dantzig's pivoting rule (for maximization, choosing a non-basic variable with the most positive reduced cost) is strongly polynomial for a fixed discount factor, and the complexity bound is better than that of the other solution methods.

For countable-state MDPs, the 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)*. General CILPs are challenging to analyze or solve mainly because useful theoretical properties of finite LPs (such as duality) fail to extend to general CILPs. We summarize the challenges for general CILPs in Section 1.2. Due to these hurdles, there are only a few algorithmic approaches to CILPs in the literature ([7, 8, 19]) (all of these are simplex-type algorithms, i.e., they navigate through extreme points of the feasible region). In particular, for a CILP formulation of non-stationary MDPs with finite state space, which can be considered to be a subclass of countable-state MDPs, [8] recently provided duality results, characterization of extreme points, and an implementable simplex algorithm that improves in every iteration and converges to optimality.

However, classes of CILPs considered so far ([7, 8, 19]) 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 present in Section 3, 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") as we illustrate by Example 2. Another key contribution of this paper is that we show that even without restrictions on positions of nonzeros in the coefficient matrix, the dynamic programming structure in the coefficient matrix of the CILP formulation of countable-state MDPs still enables us to establish the standard LP results and develop a simplex-type algorithm.

1.2 Literature Review

In this section, we first review the existing solution methods for countable-state MDPs as discussed in [21, 22, 23].

The algorithm suggested in [23] 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 $V^0 : \mathcal{S} \to \mathbb{R}$ and for $k = 1, 2, \ldots$, computes a function $V^k : \mathcal{S} \to \mathbb{R}$ in iteration k by the following recursion formula:

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

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 $V^k(s)$ by (1) only for $s \leq k$ and lets $V^k(s) = u(s)$ for $s > k$. In [23], it was shown that V^k converges pointwise to the optimal value function $V^\star$ 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 (1) to $s \leq k$ and an arbitrary action to $s > k$. It was also shown in [23] that V_{π^k} converges pointwise to $V^\star$ but the convergence may not be monotone.

The solution method in [21] 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 $V^k : \mathcal{S} \to \mathbb{R}$ that satisfies

$$V^k(s) = r(s, \pi^{k-1}(s)) + \alpha \sum_{t \in \mathcal{S}} p(t|s, \pi^{k-1}(s)) V^k(t) \text{ for } s \in \mathcal{S}; \tag{2}$$

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

$$\pi^k(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \alpha \sum_{t \in \mathcal{S}} p(t|s,a) V^k(t) \right\} \text{ for } s \in \mathcal{S}. \tag{3}$$

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 V^k that satisfies (2) for $s \leq k$ and $V^k(s) = u(s)$ for $s > k$, and then finds π^k that satisfies (3) only for $s \leq k$. It was shown in [21] that V^k obtained by this method also converges pointwise to $V^\star$ 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 V_{π^k} converges pointwise to $V^\star$ but again, the convergence may not be monotone.

Another method, proposed in [22], 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 $V^\star$. 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, [9, 11, 20] studied properties of countable-state MDPs. In [24], each of the three settings with unbounded rewards in [9, 11, 20] 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 [9, 11, 20]. Meanwhile, [4] extensively reviewed conditions under which the extension of the value iteration in [23] converges to optimality in value and studied its rate of convergence. The setting in [4] is more general than the settings in [9, 11, 20] 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 paper, we consider the setting in [12] (Section 6.10) for countable-state MDPs with unbounded rewards (Assumptions A1, A2, and A3 in Section 2.1 of this paper). One can easily show that this setting covers the three settings in [9, 11, 20]; it is a special case of the one in [4] but it is checkable for given parameters without solving the MDP and has enough generality to cover many applications of interest.

As mentioned in the previous section, an important contribution of this paper is that we developed a simplex-type algorithm for 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. To put our findings in perspective, let us review some of the difficulties in analyzing and solving general CILPs

(for more details, see [7, 8]). First, there is an example in which a CILP and its dual have a duality gap ([15]). Also, there is a CILP that has an optimal solution but does not have an extreme point optimal solution ([3]). Even for CILPs that have an extreme point optimal solution, extending the simplex method is challenging. A pivot operation may require infinite computation, and hence not be implementable ([3, 19]). Moreover, [7] 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. However, some of LP results can be extended by considering more structured CILPs ([8, 14, 15]) or finding appropriate sequence spaces ([6]).

1.3 Organization

The rest of this paper is organized as follows. In Section 2, we formally define countable-state MDPs, introduce assumptions on problem parameters, give two application examples, and review some results established in literature. In Section 3, we introduce primal and dual CILP formulations for countable-state MDPs, prove strong duality, define complementary slackness, and prove the equivalence of complementary slackness and optimality. Then, in Section 4, we introduce an implementable simplex-type algorithm solving the dual CILP. We show that the algorithm obtains a sequence of policies whose value functions strictly improve in every iteration and converge to the optimal value function. Section 5 illustrates computational behavior of the algorithm for inventory management problem instances and Section 6 concludes the paper with discussion and some future research directions.

2 Countable-State MDP

2.1 Problem Formulation

We continue to use the notation introduced in the previous section: countably-infinite state set $\mathcal{S}$, finite action set $\mathcal{A}$, transition probabilities $p(t|s,a)$ and immediate reward $r(s,a,t)$ (along with the expected immediate reward $r(s,a)$) for $s,t \in \mathcal{S}$ and $a \in \mathcal{A}$, and discount factor $0 < \alpha < 1$. We let $\mathcal{S} = \{1,2,\ldots\}$ and $\mathcal{A} = \{1,2,\ldots,A\}$ unless otherwise specified.

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 at the beginning of period n. 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 all $s \in \mathcal{S}$ 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. For a stationary and deterministic policy π and a state s, $\pi(s)$ denotes the action chosen by π at s. 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.

Given an initial state distribution β, each policy π induces a probability distribution P_π^β on sequences $\{(s_n, a_n)\}_{n=0}^\infty$, where $(s_n, a_n) \in \mathcal{S} \times \mathcal{A}$ for $n = 0, 1, \ldots$, and defines a state process $\{S_n\}_{n=0}^\infty$ and an action process $\{A_n\}_{n=0}^\infty$. 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

$$V_\pi(\beta) \triangleq E_\pi^\beta \left[\sum_{n=0}^{\infty} \alpha^n r(S_n, A_n) \right]. \tag{4}$$

We call $V_\pi(\beta)$ *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. For those initial state distributions concentrated on one state s, we use a slight abuse of notation in (4): $V_\pi(s)$ where $\beta(s) = 1$ for a state s. $V_\pi : \mathcal{S} \to \mathbb{R}$ is called *the value function of policy* π.

A policy $\pi^\star$ is said to be *optimal for initial state distribution* β if $V_{\pi^\star}(\beta) = V^\star(\beta) \triangleq \sup_{\pi \in \Pi} V_\pi(\beta)$. A policy $\pi^\star$ is said to be *optimal for initial state* s if $V_{\pi^\star}(s) = V^\star(s) \triangleq \sup_{\pi \in \Pi} V_\pi(s)$. We call $V^\star : \mathcal{S} \to \mathbb{R}$ *the optimal value function of the MDP* or simply *the optimal value function.* A policy $\pi^\star$ is defined to be *optimal* if $V_{\pi^\star}(s) = V^\star(s)$ for all $s \in \mathcal{S}$, i.e., if it is optimal for any initial state. The goal of the decision maker is to find an optimal policy.

Let us define additional notation that will come in handy in the rest of the paper. 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 (in this paper, notation σ is used to emphasize the choice of a stationary policy) and a state $s \in \mathcal{S}$, $r_\sigma(s)$ denotes $r(s, \sigma(s))$, the expected immediate reward at s when σ is applied, and r_σ denotes the reward vector indexed by states.

Throughout the paper, 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 [12]) *There exists a positive real-valued function* w *on* $\mathcal{S}$ *satisfying the following:*

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

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

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

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

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

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

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

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

2.2 Examples

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

Example 1 (Example 6.10.2 in [12]) 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

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

by Proposition 6.10.5(a) in [12], A2 and A3 are also satisfied.

Example 2 (Generalized flow and service control) This example is a generalization of the flow and service rate control problem in [2]. 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 [2], 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,

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

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 [12], A2 and A3 are satisfied.

Parts (b) and (c) of Proposition 6.10.5 in [12] provide two other sufficient conditions to satisfy A2 and A3.

2.3 Background

We conclude this section by reviewing some technical preliminaries that were established in the literature and will be used in this paper.

By the following theorem, we can limit our attention to policies that are stationary and deterministic.

Theorem 2.1 (cf. Theorem 6.10.4 of [12]) *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 $V^\star$ 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, V_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, V_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 2.2 (cf. Proposition 6.10.1 of [12]) *If Assumptions A1, A2, and A3 are satisfied,*

$$|V_\pi(s)| \leq Lw(s) \text{ for any } s \in \mathcal{S} \text{ and } \pi \in \Pi_M. \tag{5}$$

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

Proposition 2.3 (cf. Tonelli's theorem on page 309 of [17]) *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 2.4 (Theorem 8.3 in [18]) *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 [17]) and Tonelli's theorem. We will also use the following special case of monotone convergence theorem (MCT).

Proposition 2.5 (Series version of monotone convergence theorem, Corollary 5.3.1 of [13]) *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 2.6 (Dominated convergence theorem, Theorem 5.3.3 of [13]) *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 CILP Formulations

In this section, we introduce primal and dual CILP formulations of countable-state discounted MDPs. We start with a straightforward result which was used in [12] and [16] without being explicitly stated.

Lemma 3.1 *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

$$V_\pi(\beta) = E_\pi^\beta \left[\sum_{n=0}^{\infty} \alpha^n r(S_n, A_n) \right] = \sum_{s=1}^{\infty} \beta(s) E_s^\pi \left[\sum_{n=0}^{\infty} \alpha^n r(S_n, A_n) \right] = \sum_{s=1}^{\infty} \beta(s) V_\pi(s).$$

Since $V_\pi(s) \leq V^\star(s)$ for any $s \in \mathcal{S}$, and $\beta(s) > 0$ for any $s \in \mathcal{S}$, a policy π maximizes $V_\pi(\beta)$ if and only if it maximizes $V_\pi(s)$ 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

$$\beta(s) > 0 \text{ for all } s \in \mathcal{S}. \tag{6}$$

Additionally, we require that β satisfies

$$\beta^T w = \sum_{s=1}^{\infty} \beta(s) w(s) < \infty. \tag{7}$$

(7) will help us show that a variety of infinite series we consider in this paper 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 (6) and (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. Let $\|y\|_w \triangleq \sup_{s \in \mathcal{S}} \frac{|y(s)|}{w(s)}$ for $y \in \mathbb{R}^\infty$ and $Y_w \triangleq \{y \in \mathbb{R}^\infty : \|y\|_w < \infty\}$. Consider the following CILP:

$$\text{(P)} \quad \min \; g(y) = \sum_{s=1}^{\infty} \beta(s) y(s) \tag{8}$$

$$\text{s.t.} \; y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a) y(t) \geq r(s,a) \text{ for } s \in \mathcal{S} \text{ and } a \in \mathcal{A} \tag{9}$$

$$y \in Y_w. \tag{10}$$

A CILP formulation consisting of (8) and (9) was introduced in Chapter 2.5 of [16] for MDPs with uniformly bounded rewards. By adding constraint (10), one can apply essentially the same arguments to countable-state MDPs being considered in this paper (i.e., ones with unbounded rewards but satisfying Assumptions A1, A2, and A3) to show that the optimal value function $V^\star$ is equal to the unique optimal solution of (P). (Chapter 12.3 of [5] introduced a similar CILP formulation for a more general class of MDPs, but for the average reward criterion. Additionally, in Chapter 8.8 of [2], Altman derived a similar CILP formulation 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.)

A few remarks about (P) are in order. Note that for any $y \in Y_w$, the objective function value is always finite because of (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 2.2, so (10) does not exclude any solution of interest. Since, for any optimal policy $\pi^\star$,

$$V^\star(\beta) = V_{\pi^\star}(\beta) = \sum_{s=1}^{\infty} \beta(s) V_{\pi^\star}(s) = \sum_{s=1}^{\infty} \beta(s) V^\star(s), \tag{11}$$

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

We also consider the following CILP formulation of a countable-state MDP:

$$\text{(D)} \ \max \ f(x) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a) x(s,a) \tag{12}$$

$$\text{s.t.} \ \sum_{a=1}^{A} x(s,a) - \alpha \sum_{t=1}^{\infty} \sum_{a=1}^{A} p(s|t,a) x(t,a) = \beta(s) \text{ for } s \in \mathcal{S} \tag{13}$$

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

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 (D) can be found in the literature even for more general classes of MDPs (e.g., see Chapter 8 of [2] or Chapter 12 of [5]). However, we provide a high-level derivation of (D) in Appendix A, in part because it also gives a proof of strong duality between (P) and (D) (Theorem 3.3); due to this relationship, we will refer to (P) as the primal problem, and (D) — as the dual problem. Briefly, (D) 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 Appendix A.) It is well known that $\mathcal{F}$, which denotes the set of feasible solutions to (D), coincides with the set of occupancy measures of stationary policies. For any stationary policy σ and its occupancy measure $x \in \mathcal{F}$, $V_\sigma(\beta)$ is equal to the objective function value of (D) at x. An optimal stationary policy (which is known to be an optimal policy) can therefore be obtained from an optimal solution to (D) by computing the corresponding stationary policy (for more details, see Appendix A).

The following visualization of constraint (13) will help readers understand the structure of (D). Using infinite matrix and vector notation, constraint (13) can be written as

$$[M^1|M^2|\ldots|M^s|\ldots]\, x = \beta. \tag{14}$$

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 $p(\cdot|s,a)$.[1]

[1]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.

Remark 3.2 *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 (D) 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 $\mathcal{A}^2 = \{0\}$).*

The following strong duality theorem is proven in Appendix A.

Theorem 3.3 *Strong duality holds between (P) and (D), i.e., $g(y^*) = f(x^*)$, where y^* and x^* are optimal solutions of (P) and (D), respectively.*

Note that (D) has only equality constraints and non-negativity constraints, and thus can be said to be in standard form. The main goal of this paper is to develop a simplex-type algorithm that solves (D). 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 $\mathcal{F}$ is also well known in literature (e.g., Theorem 11.3 of [5]).

Theorem 3.4 *A feasible solution x of (D) is an extreme point of $\mathcal{F}$ 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 $\mathcal{F}$ correspond to stationary deterministic policies.*

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

Definition 3.5 *A feasible solution x to (D) is defined to be a* basic feasible solution *of (D) 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 (14) 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 $\mathcal{F}$ is the set of occupancy measures of stationary policies; moreover, the set of extreme points of $\mathcal{F}$ 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 [5]).

Theorem 3.6 *(D) has an extreme point optimal solution.*

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

Definition 3.7 (Complementary slackness) *Suppose $x \in \mathcal{F}$ and $y \in Y_w$. x and y are said to satisfy* complementary slackness *(or be* complementary*) if*

$$x(s,a)\left[r(s,a) - \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] = 0 \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}. \tag{15}$$

Theorem 3.8 (Complementary slackness sufficiency) *Suppose $x \in \mathcal{F}$ and $y \in Y_w$ are complementary. Then $f(x) = g(y)$, and if y is feasible to (P), then y and x are optimal to (P) and (D),*

respectively.

Proof: In Appendix B.

Theorem 3.9 (Complementary slackness necessity) *If y and x are optimal to (P) and (D), respectively, then they are complementary.*

Proof: In Appendix C.

Given a basic feasible solution x, let σ be the corresponding stationary deterministic policy. By Theorem 2.1(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 2.2.

Recently in [6], 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 (D), the sequence space l_1 does not guarantee convergence of the objective function (but the objective function converges for any feasible solution of (D) as shown in Appendix A). Thus, for countable-state MDPs with unbounded rewards being considered in this paper, applying the choice of sequence spaces in [6] would yield a different CILP formulation from (D), 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 3.10 *Any $x \in \mathcal{F}$ satisfies*

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

Proof: In Appendix D.

4 Simplex Algorithm

To devise a simplex-type algorithm for (D), 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 [7, 8]: 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, [7] 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 (D) 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 [8], 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*.

4.1 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 (D) and let $y \in Y$ be its complementary solution. We first define reduced costs.

Definition 4.1 *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 (P):*

$$\gamma(s,a) \triangleq r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a)y(t) - y(s). \tag{16}$$

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 (P), and thus, x is optimal to (D) by Theorem 3.8.

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

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

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 [12], $(I - \alpha P_\tau)^{-1}$ exists for any stationary policy τ and we have[2]

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

[2]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, 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

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

Therefore,

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

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 (4) 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:

$$y^N(s) = r_\sigma(s) + \alpha \sum_{t=1}^{N} P_\sigma(t|s) y^N(t) \text{ for } s = 1, \ldots, N. \tag{17}$$

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 4.3 *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 Appendix E.

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

$$\gamma^N(s,a) \triangleq r(s,a) + \alpha \sum_{t=1}^{N} p(t|s,a) y^N(t) - y^N(s) \text{ for } s = 1, \ldots, N, a \in \mathcal{A}. \tag{18}$$

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 4.4 *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

$$\delta(\sigma,s,a,N) \triangleq L\sum_{t>N}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(t|s)w(t) + \alpha L\sum_{t=1}^{N} p(t|s,a)\sum_{t'>N}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(t'|t)w(t') + \alpha L\sum_{t>N} p(t|s,a)w(t). \tag{19}$$

Proof: By Lemma 4.3 and (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 Appendix E, it is not hard to prove the following proposition about $\delta(\sigma,s,a,N)$.

Proposition 4.5 *For any positive integer N and for $\sigma \in \Pi_{SD}, s = 1,\ldots,N$, and $a \in \mathcal{A}$,*

$$\delta(\sigma,s,a,N) \leq L(L + \alpha\kappa L + \alpha\kappa)w(s), \tag{20}$$

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

$$\delta(\sigma,s,a,N) \to 0 \text{ as } N \to \infty. \tag{21}$$

Thus, by this proposition and Lemma 4.4, 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 (D), 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

$$\bar{\delta}(s,a,N) \to 0 \text{ as } N \to \infty \text{ for any } (s,a). \tag{22}$$

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 4.4, we show how such an upper bound $\bar{\delta}(s,a,N)$ can be computed for Examples 1 and 2.

4.2 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 (D) 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 (P), 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 [7], 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 $s \in \mathcal{S}$.[3]
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 (17).
   (c) Compute the approximate reduced costs, $\gamma^{k,N}(s,a)$ for $s = 1, \ldots, N, a \in \mathcal{A}$ by (18).
   (d) Find the nonbasic variable achieving the largest *approximate* nonbasic reduced cost weighted by β:
   $$(s^{k,N}, a^{k,N}) = \arg\max_{(s,a)} \beta(s)\gamma^{k,N}(s,a). \tag{23}$$
   (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]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.

4.3 Proof of Convergence

In this section we show that the simplex algorithm of Section 4.2 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 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 4.6 *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 4.4, 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 4.2, the following corollary is immediate. We denote $f(x^k)$ as f^k for simplicity.

Corollary 4.7 *The objective function of (D) 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 4.8 *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 4.2,

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

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 4.9 *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 4.10 *Step 2 of the algorithm terminates if and only if x^k is not optimal to (D).*

Proof: In Appendix F.

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 4.2, 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 4.11 *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 3.8, the first inequality by (5), and the last inequality by (7). By Corollary 4.7, 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 4.2, 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 4.12 *$N(k) \to \infty$ as $k \to \infty$.*

Proof: This proof is similar to the proof of Lemma 5.7 in [8].

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 4.2, the second by Lemma 4.4, 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 4.7 and Lemma 4.11. Thus, we established the result by contradiction. □

Theorem 4.13 *Let $f^\star$ be the optimal value of (D). 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 [8], 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 4.10) and never repeats any non-optimal basic feasible solutions (Corollary 4.9). 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 (D). 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 $\mathcal{F}$ which is shown to be compact in Theorem 11.3 of [5] or Corollary 10.1 of [2], 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 $\bar{x} \in \mathcal{F}$. 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 [1]). By (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 (P), and thus, that $\bar{x}$ is optimal for (D).

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

$$x^{u_k}(s,a)\left[r(s,a) - \left(y^{u_k}(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y^{u_k}(t)\right)\right] = 0 \text{ for } s \in \mathcal{S}, a \in \mathcal{A}. \tag{24}$$

Recall that, by (5), $|y^{u_k}(t)| \leq Lw(t)$ for any state t and

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

Thus, by Proposition 2.6, 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 (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 (P). 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 $k \geq K$,

$$r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a) y^{u_k}(t) - y^{u_k}(s) \geq \frac{1}{2}\epsilon. \tag{25}$$

Since $\lim_{k\to\infty} N(u_k) = \infty$ by Lemma 4.12, $s \leq N(u_k)$ for sufficiently large k. For all such k,

$$\begin{aligned} r(s,a) + \alpha \sum_{t=1}^{\infty} p(t|s,a) y^{u_k}(t) - y^{u_k}(s) = \gamma^{u_k}(s,a) &\leq \gamma^{u_k,N(u_k)}(s,a) + \delta(\sigma^{u_k}, s, a, N(u_k)) \\ &\leq \gamma^{u_k,N(u_k)}(s,a) + \bar{\delta}(s,a,N(u_k)) \end{aligned} \tag{26}$$

by Lemma 4.4 and the definition of $\bar{\delta}(s,a,N(u_k))$. By Lemma 4.12, we know that $\bar{\delta}(s,a,N(u_k)) \to 0$ 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 (25), and thus, we will conclude that $\bar{y}$ is feasible to (P).

We have:

$$\begin{aligned} \beta(s)\gamma^{u_k,N(u_k)}(s,a) &\leq \beta(s^{u_k})\gamma^{u_k,N(u_k)}(s^{u_k},a^{u_k}) \\ &\leq \beta(s^{u_k})\gamma^{u_k}(s^{u_k},a^{u_k}) + \beta(s^{u_k})\delta(\sigma^{u_k},s^{u_k},a^{u_k},N(u_k)) \end{aligned} \tag{27}$$

where the first inequality is due to (23). By Lemma 4.11, the first term of the right hand side of (27) tends to zero as $k \to \infty$. Also, by (20), the second term of the right hand side of (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 (7) and $s^{u_k} \to \infty$ as $k \to \infty$ by the choice of sequence u_k. Therefore, the right hand side of (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, (25) is contradicted and so $\bar{y}$ is feasible to (P).

Thus, we have shown that $\bar{x}$ is optimal to (D). By following arguments similar to those of Lemma 8.5 of [2], one can show that f, the objective function of (D), is continuous on $\mathcal{F}$ under the product topology. Thus, f^{u_k} converges to $f^\star$ as $k \to \infty$. However, f^k converges by Lemma 4.11 and its limit should be the same as the limit of its subsequence. Therefore, f^k converges to $f^\star$ as $k \to \infty$. □

4.4 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 2.2.

Example 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 (19) can be computed as follows:

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

where the exchange of infinite sums follows by Proposition 2.3 and $\nu \triangleq \lfloor\frac{N-s}{M}\rfloor + 1$. An upper bound on the rest of (19) can be found similarly. Thus, we obtain the following upper bound on $\delta(\sigma, s, a, N)$:[4]

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

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 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

$$\bar{\delta}(s,a,N) \triangleq L\cdot g(s,N) + \alpha L\sum_{t=0}^{N} p(t|s,a)\cdot g(t,N) + \alpha\cdot L\cdot h(s,N) \tag{28}$$

where

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

and

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

[4] In (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 (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.

In Appendix G, 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.

5 Numerical Illustration

We implemented the simplex algorithm and tested it on five instances of the inventory management problem of Example 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.

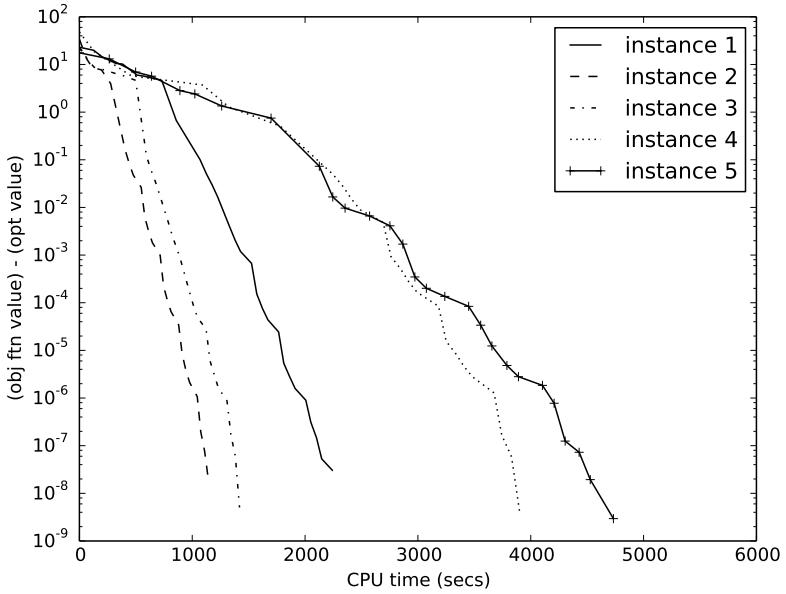
Figure 1 shows cost improvement of the simplex algorithm for the above instances of the inventory management problem as a function of (a) CPU time and (b) number of pivot operations. The vertical axis of Figure 1 is the difference between the objective function value of (D) of policies obtained by the 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 (18).

Figure 1 illustrates that for all instances, the difference between the objective function value of policies and the optimal value decreased monotonically and converged to zero. As shown in Figure 1b, 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 1a, possibly because of the higher maximum ordering level.

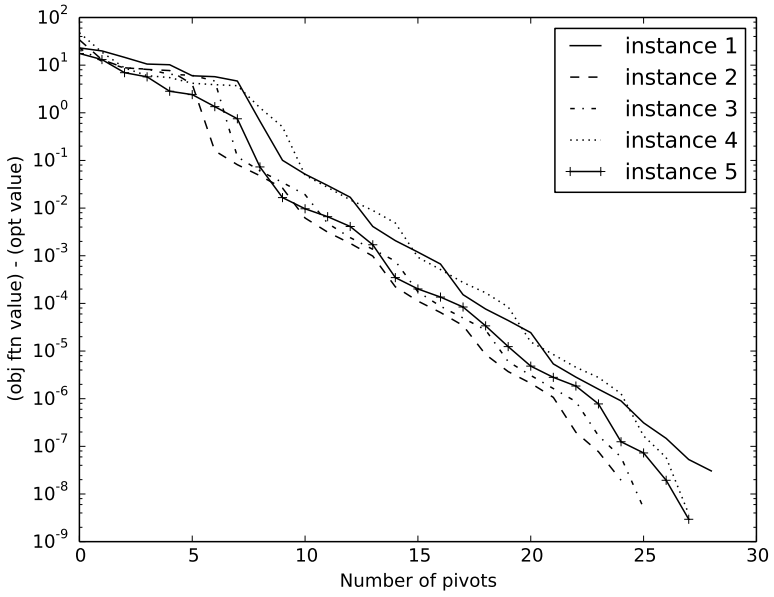
6 Discussion and Future Research

It is a natural next step to compare the simplex algorithm to the existing methods for countable-state MDPs. However, it is not straightforward how to make an empirical comparison. Aside from deciding how to implement those algorithms in a fair manner and which problem instances to try, another important issue is the cost of data acquisition. As those algorithms proceed, they require transition probabilities and rewards from more states. In practice, obtaining the problem data can be expensive. 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.

Recently, complexity of the simplex method with the Dantzig's pivoting rule for finite-state MDPs was studied in [25]. If one can derive a number of iterations (or computational complexity) for the simplex algorithm for countable-state MDPs to find a policy whose value function is within a given threshold from the optimal value function, 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 [23, 21]. However, the convergence rates provide us with upper bounds on number of iterations (or computational complexity) to achieve near-optimality so this theoretical comparison



(a) For CPU time



(b) For number of pivots

Figure 1: Cost improvement of the simplex algorithm for inventory management problems

would also be incomplete.

This paper generalized the LP approach in [8] to CILP formulations of a more general class of MDPs and introduced a simplex-type algorithm to a class of CILPs with less structure than those in the literature. To extend the LP approach of this paper to a more general class of CILPs, one would have to understand what aspects of the CILPs considered in this paper enabled the success of the LP approach. To be more precise, one needs to analyze what characteristics of the assumptions in Section 2.1 and the dynamic programming structure in the coefficient matrix (14) of (D) made it possible to establish the standard LP results and devise the simplex algorithm.

References

[1] C. Aliprantis and K. Border. *Infinite-dimensional analysis: a hitchhiker's guide*. Springer-Verlag, Berlin, Germany, 1994.

[2] E. Altman. *Constrained Markov decision processes*. Chapman and Hall, CRC, 1998.

[3] E. J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces: theory and applications*. John Wiley and Sons, Chichester, UK, 1987.

[4] R. Cavazos-Cadena. Finite-state approximations for denumerable state discounted Markov decision processes. *Applied Mathematics and Optimization*, 14:1–26, 1986.

[5] E. Feinberg and A. Shwartz. *Handbook of Markov decision processes*. Kluwer International Series, 2002.

[6] A. Ghate. Duality in countably infinite linear programs. 2014. Working paper.

[7] A. Ghate, D. Sharma, and R. L. Smith. A shadow simplex method for infinite linear programs. *Operations Research*, 58:865–877, 2010.

[8] A. Ghate and R. L. Smith. A linear programming approach to nonstationary infinite-horizon Markov decision processes. *Operations Research*, 61:413–425, 2013.

[9] J. Harrison. Discrete dynamic programming with unbounded rewards. *Annals of Mathematical Statistics*, 43:636–644, 1972.

[10] O. Hernández-Lerma. Finite-state approximations for denumerable multidimensional state discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 113:382–389, 1986.

[11] S. Lippman. On dynamic programming with unbounded rewards. *Management Science*, 21:1225–1233, 1975.

[12] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley and Sons, New York, NY, USA, 1994.

[13] S. Resnick. *A Probability Path*. Birkhäuser, 1999.

[14] H. E. Romeijn and R. L. Smith. Shadow prices in infinite dimensional linear programming. *Mathematics of Operations Research*, 23:239–256, 1998.

[15] H. E. Romeijn, R. L. Smith, and J. Bean. Duality in infinite dimensional linear programming.

Mathematical Programming, 53:79–97, 1992.

[16] S. Ross. *Introduction to stochastic dynamic programming*. Academic Press, New York, NY, USA, 1983.

[17] H. Royden. *Real Analysis*. Macmillan Publishing Company, New York, 1988.

[18] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, Inc., 1976.

[19] T. C. Sharkey and H. E. Romeijn. A simplex algorithm for minimum cost network flow problems in infinite networks. *Networks*, 52:14–31, 2008.

[20] J. Wessels. Markov programming by successive approximations with respect to weighted supremum norms. *Journal of Mathematical Analysis and Applications*, 58:326–335, 1977.

[21] 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.

[22] D. White. Finite state approximations for denumerable state infinite horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 74:292–295, 1980.

[23] 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.

[24] 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.

[25] 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.

A Derivation of (D) and Proof of Strong Duality

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

We first define an *occupancy measure* and will show that the feasible region of (D) 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 [2]) 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:

$$\tilde{V}_\pi(\beta) \triangleq \tilde{E}_\pi^\beta \left[\sum_{n=0}^{\infty} \tilde{r}(\tilde{S}_n, \tilde{A}_n) \right],$$

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 $V_\pi(\beta) = \tilde{V}_\pi(\beta)$ 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 $T = \min\{n \geq 0 : s_n = 0\}$. 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 $a \in \mathcal{A}$,

$$Q_\pi^\beta(s,a) \triangleq E_\pi^\beta \sum_{n=0}^{T-1} \mathbf{1}\{S_n = s, A_n = a\} = E_\pi^\beta \sum_{n=0}^{\infty} \mathbf{1}\{S_n = s, A_n = a\} = \sum_{n=0}^{\infty} P_\pi^\beta\{S_n = s, A_n = a\}, \tag{29}$$

where the last equality is due to Proposition 2.5. (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 [2]) that for any policy π, there exists a stationary policy σ such that $Q_\pi^\beta = Q_\sigma^\beta$, namely,

$$\sigma(a|s) = \frac{Q_\pi^\beta(s,a)}{\sum_{b \in \mathcal{A}} Q_\pi^\beta(s,b)}, \tag{30}$$

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 [5] and Corollary 10.1 of [2]) that $\mathcal{Q}_S$ coincides with the set of nonnegative and summable solutions of the following set of equations:

$$\sum_{a=1}^{A} x(s,a) = \beta(s) + \alpha \sum_{t=1}^{\infty} \sum_{a=1}^{A} p(s|t,a) x(t,a) \text{ for } s \neq 0. \tag{31}$$

Therefore, the feasible region of (D) 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 [2], one can show that for any Markov policy π,

$$V_\pi(\beta) = \sum_{s=1}^{\infty} \sum_{a=1}^{A} r(s,a) Q_\pi^\beta(s,a). \tag{32}$$

From Proposition 2.2 and (7), we know that $V_\pi(\beta)$ is finite, and thus the right hand side of the above equation is finite, for any Markov policy π. Since the feasible region of (D) is the set of occupancy measures of all stationary policies, the objective function of (D) is finite for any feasible solution. Moreover, by Lemma 3.1, a stationary policy whose occupancy measure is an optimal solution to (D) is also optimal for the MDP. Given an optimal solution of (D), a stationary optimal policy can be obtained by (30).

By following the arguments of Lemma 8.5 of [2], one can show that f, the objective function of (D), is continuous on its feasible region under the usual product topology. In addition, it is also well known (e.g., Theorem 11.3 of [5] and Corollary 10.1 of [2]) that the feasible region of (D) is a compact subset of $\mathbb{R}^\infty$ under the product topology. Therefore, the maximum of f is attained in the feasible region of (D).

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

B Proof of Theorem 3.8

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 $a = 1, 2, \ldots, A$,

$$\begin{aligned}
&\sum_{s=1}^{N} \sum_{a=1}^{A} r(s,a)x(s,a) \\
&\quad = \sum_{s=1}^{N} \sum_{a=1}^{A} \left(y(s) - \alpha \sum_{t=1}^{\infty} p(t|s,a) y(t) \right) x(s,a) \\
&\quad = \sum_{s=1}^{N} \sum_{a=1}^{A} y(s)x(s,a) - \alpha \sum_{s=1}^{N} \sum_{a=1}^{A} \sum_{t=1}^{\infty} p(t|s,a) y(t) x(s,a) \\
&\quad = \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) - \sum_{t=1}^{\infty} y(t) \alpha \sum_{s=1}^{N} \sum_{a=1}^{A} p(t|s,a) x(s,a) \\
&\quad = \sum_{s=1}^{N} y(s) \sum_{a=1}^{A} x(s,a) - \sum_{t=1}^{\infty} y(t) \left(\sum_{a=1}^{A} x(t,a) - \beta(t) - \alpha \sum_{s=N+1}^{\infty} \sum_{a=1}^{A} p(t|s,a) x(s,a) \right),
\end{aligned} \tag{33}$$

where the exchange of sums in the third equality is justified by the fact that $\sum_{t=1}^{\infty} p(t|s,a)y(t)$ is finite for any s and a and the last equality is obtained by the feasibility of x to (D). We will find the limit of (33) as $N \to \infty$.

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

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

and since x is feasible to (D), there exists a stationary policy σ such that (here we consider the absorbing MDP formulation introduced in Appendix A)

$$\begin{aligned}\|y\|_w \sum_{s=1}^{\infty} w(s)\sum_{a=1}^{A}x(s,a) &= \|y\|_w \sum_{s=1}^{\infty} w(s)\sum_{a=1}^{A}Q_\sigma^\beta(s,a) = \|y\|_w \sum_{s=1}^{\infty}\sum_{a=1}^{A}\sum_{n=0}^{\infty}P_\sigma^\beta(S_n=s,A_n=a)w(s)\\ &= \|y\|_w \sum_{s=1}^{\infty}\sum_{n=0}^{\infty}\sum_{a=1}^{A}P_\sigma^\beta(S_n=s,A_n=a)w(s) = \|y\|_w \sum_{s=1}^{\infty}\sum_{n=0}^{\infty}P_\sigma^\beta(S_n=s)w(s), \qquad (34)\end{aligned}$$

where the third equality follows by Proposition 2.3. However,

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{s=1}^{\infty}P_\sigma^\beta(S_n=s)w(s) = \beta^T(w+\alpha P_\sigma w+\alpha^2P_\sigma^2 w+\ldots)\\ &\quad\leq \beta^T[(w+(\alpha\kappa)w+(\alpha\kappa)^2w+\ldots+(\alpha\kappa)^{J-1}w)+(\lambda w+\lambda(\alpha\kappa)w+\ldots+\lambda(\alpha\kappa)^{J-1}w)+\ldots]\\ &\quad= L\beta^T w<\infty\end{aligned}$$

by Assumptions A2 and A3, and (7). Thus, the sum (34) is finite by Proposition 2.3. Therefore, we have

$$\sum_{s=1}^{\infty} y(s)\sum_{a=1}^{A}x(s,a)<\infty. \qquad (35)$$

We will also prove that

$$\sum_{t=1}^{\infty} y(t)\sum_{s=N+1}^{\infty}\sum_{a=1}^{A}p(t|s,a)x(s,a)<\infty \qquad (36)$$

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 2.3, 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 2.4,

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

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

$$\sum_{t=1}^{\infty} \beta(t)|y(t)| \leq \|y\|_w \sum_{t=1}^{\infty} \beta(t)w(t) < \infty. \tag{37}$$

Then, by (35), (36), and (37), we can write (33) as

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

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. □

C Proof of Theorem 3.9

Since y and x are feasible to (P) and (D), 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

$$\begin{aligned}
0 \geq &\sum_{s=1}^{N}\sum_{a=1}^{A} x(s,a)\left[r(s,a) - \left(y(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y(t)\right)\right] \\
= &\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) - \sum_{s=1}^{N}\sum_{a=1}^{A}\left(y(s) - \alpha\sum_{t=1}^{\infty} p(t|s,a)y(t)\right)x(s,a) \\
= &\sum_{s=1}^{N}\sum_{a=1}^{A} r(s,a)x(s,a) - \sum_{s=1}^{N} y(s)\sum_{a=1}^{A} x(s,a) + \sum_{t=1}^{\infty} y(t)\sum_{a=1}^{A} x(t,a) - \sum_{t=1}^{\infty}\beta(t)y(t) \\
&- \alpha\sum_{t=1}^{\infty} y(t)\sum_{s=N+1}^{\infty}\sum_{a=1}^{A} p(t|s,a)x(s,a),
\end{aligned} \tag{38}$$

where the last equality is obtained similarly to the proof of Theorem 3.8. 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 (38) and using arguments similar to the proof of Theorem 3.8, 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. □

D Proof of Lemma 3.10

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 (29), it suffices to show (here we consider the absorbing MDP formulation introduced in Appendix A)

$$\sum_{s=1}^{\infty}\sum_{a=1}^{A}\sum_{n=0}^{\infty} P_\sigma^\beta\{S_n = s, A_n = a\} = \frac{1}{1-\alpha}.$$

Using Proposition 2.3 to interchange the sums, we have:

$$\sum_{n=0}^{\infty}\sum_{s=1}^{\infty}\sum_{a=1}^{A} P_\sigma^\beta\{S_n = s, A_n = a\} = \sum_{n=0}^{\infty}\sum_{s=1}^{\infty} P_\sigma^\beta\{S_n = s\} = \sum_{n=0}^{\infty}\alpha^n = \frac{1}{1-\alpha}.$$

□

E Proof of Lemma 4.3

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 $s = 1, \ldots, N$,

$$\begin{aligned} & |y(s) - y^N(s)| \\ =& \left| \alpha \sum_{t_1>N} P_\sigma(t_1|s) r_\sigma(t_1) + \alpha^2 \left[\sum_{t_1>N} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) + \sum_{t_1\le N} \sum_{t_2>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) r_\sigma(t_2) \right] \right. \\ & + \alpha^3 \left[\sum_{t_1>N} \sum_{t_2=1}^{\infty} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) + \sum_{t_1\le N} \sum_{t_2>N} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) \right. \\ & \left. \left. + \sum_{t_1\le N} \sum_{t_2\le N} \sum_{t_3>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) r_\sigma(t_3) \right] + \ldots \right| \\ \le& \alpha \sum_{t_1>N} P_\sigma(t_1|s) w(t_1) + \alpha^2 \left[\sum_{t_1>N} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) + \sum_{t_1\le N} \sum_{t_2>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) \right] \\ & + \alpha^3 \left[\sum_{t_1>N} \sum_{t_2=1}^{\infty} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) + \sum_{t_1\le N} \sum_{t_2>N} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \right. \\ & \left. + \sum_{t_1\le N} \sum_{t_2\le N} \sum_{t_3>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \right] + \ldots \end{aligned} \tag{39}$$

by Assumption A1. Note that the terms of the infinite sum on the right hand side of (39) can be reordered by Proposition 2.3. In particular, consider rearranging the terms in (39) as follows:

$$\begin{aligned} & \left[\alpha \sum_{t_1>N} P_\sigma(t_1|s) w(t_1) + \alpha^2 \sum_{t_1>N} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) + \alpha^3 \sum_{t_1>N} \sum_{t_2=1}^{\infty} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) + \ldots \right] \\ & + \left[\alpha^2 \sum_{t_1\le N} \sum_{t_2>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) w(t_2) + \alpha^3 \sum_{t_1\le N} \sum_{t_2>N} \sum_{t_3=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) + \ldots \right] \\ & + \left[\alpha^3 \sum_{t_1\le N} \sum_{t_2\le N} \sum_{t_3>N} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) w(t_3) \right. \\ & \quad \left. + \alpha^4 \sum_{t_1\le N} \sum_{t_2\le N} \sum_{t_3>N} \sum_{t_4=1}^{\infty} P_\sigma(t_1|s) P_\sigma(t_2|t_1) P_\sigma(t_3|t_2) P_\sigma(t_4|t_3) w(t_4) + \ldots \right] \\ & + \ldots. \end{aligned} \tag{40}$$

First, let us compute an upper bound on the first bracket of (40) by considering groups of J terms, and establishing bounds using A2 and A3:

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

Applying similar arguments to the other terms of (40), we obtain the following upper bound:

$$L\left[\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1\leq N} \sum_{t_2>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) + \alpha^3 \sum_{t_1\leq N} \sum_{t_2\leq N} \sum_{t_3>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots\right]$$
$$\leq L\left[\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_1=1}^{\infty} \sum_{t_2>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) + \alpha^3 \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} \sum_{t_3>N} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots\right]$$
$$= L\left[\alpha \sum_{t_1>N} P_\sigma(t_1|s)w(t_1) + \alpha^2 \sum_{t_2>N} \sum_{t_1=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)w(t_2) + \alpha^3 \sum_{t_3>N} \sum_{t_1=1}^{\infty} \sum_{t_2=1}^{\infty} P_\sigma(t_1|s)P_\sigma(t_2|t_1)P_\sigma(t_3|t_2)w(t_3) + \dots\right]$$
$$= L\sum_{n=1}^{\infty} \sum_{t>N} \alpha^n P_\sigma^n(t|s)w(t). \tag{41}$$

by Proposition 2.3. Combining these, we obtain

$$|y(s) - y^N(s)| \leq L\sum_{n=1}^{\infty} \sum_{t>N} \alpha^n P_\sigma^n(t|s)w(t) = L\sum_{t>N} \sum_{n=1}^{\infty} \alpha^n P_\sigma^n(t|s)w(t), \tag{42}$$

again by Proposition 2.3. The right hand side of (42) converges to zero as $N \to \infty$ because

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

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

F Proof of Lemma 4.10

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 (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 2.1. 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 $a \in \mathcal{A}$,

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

Suppose that x^k is not optimal to (D). Then, y^k must not be feasible to (P), 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 T be the set of state-action pairs (s,a) such that $s < s_1$ and $\gamma^k(s,a) > 0$. Note that T is a finite set and $(\hat{s}, \hat{a}) \in T$. 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 $(s,a) \in T$ 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 $(s',a') \notin T$ such that $s' < s_1$ satisfies $\beta(s')\gamma^{k,N}(s',a') < \min_{(s,a)\in T} \frac{1}{2}\beta(s)\gamma^k(s,a)$. Then, for $N \geq \max\{N_1, N_2, N_3, s_1\}$ and for any $(s,a) \in T$ and any $(s',a') \notin T$ 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 T and the inequality $\gamma^{k,N}(s,a) > \bar{\delta}(s,a,N)$ is satisfied for any $(s,a) \in T$. 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 (D). Then y^k is feasible to (P), 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 4.4, 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. □

G Example 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 (19) is bounded as follows:

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

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 (19) is bounded as follows:

$$
\alpha L\sum_{t=0}^{N}p(t|s,a)\sum_{t'>N}\sum_{n=1}^{\infty}\alpha^n P_\sigma^n(t'|t)w(t') \le \alpha L\sum_{t=0}^{N}p(t|s,a)g(t,N). \tag{43}
$$

The last term in (19) is also bounded as follows:

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

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

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 $g(s,N)$ and $h(s,N)$ converge to zero as $N\to\infty$. Thus, it suffices to prove that the second term of $\bar{\delta}(s,a,N)$ in (28) converges to zero as $N\to\infty$. This term can be written as

follows:

$$\alpha L \sum_{t=0}^{N} p(t|s,a) g(t,N) = \frac{\alpha^2 CL}{1-\alpha} \left(\mu \sum_{t=0}^{N} p(t|s,a) - \sum_{t=0}^{N} p(t|s,a) \sum_{u=0}^{N-s} u f_Y(u) \right) + \frac{\alpha^2 L(Cs+D)}{1-\alpha} \left(\sum_{t=0}^{N} p(t|s,a) - \sum_{t=0}^{N} p(t|s,a) F_Y(N-s) \right). \quad (44)$$

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 (44) converges to zero as $N \to \infty$. We can also similarly show that the second big parenthesis in (44) converges to zero. Therefore, we proved that the second term of $\bar{\delta}(s,a,N)$ in (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 $h(s,N)$ finitely. To show that $g(s,N)$ 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,

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

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, $g(s,N)$ can be computed finitely. This implies that we can compute $\bar{\delta}(s,a,N)$ finitely.