A linear programming approach to nonstationary infinite-horizon Markov decision processes

Archis Ghate* Robert L Smith†

July 24, 2012

Abstract

Nonstationary infinite-horizon Markov decision processes (MDPs) generalize the most well-studied class of sequential decision models in operations research, namely, that of stationary MDPs, by relaxing the restrictive assumption that problem data do not change over time. Linear programming (LP) has been very successful in obtaining structural insights and devising solution methods for stationary MDPs. However, an LP approach for nonstationary MDPs is currently missing. This is because the LP formulation of a nonstationary infinite-horizon MDP includes countably infinite variables and constraints, and research on such infinite-dimensional LPs has traditionally faced several hurdles. For instance, duality results may not hold; an extreme point may not be a basic feasible solution; and in the context of a Simplex algorithm, a pivot operation may require infinite data and computations, and a sequence of improving extreme points need not converge in value to optimal. In this paper, we tackle these challenges and establish (1) weak and strong duality, (2) complementary slackness, (3) a basic feasible solution characterization of extreme points, (4) a one-to-one correspondence between extreme points and deterministic Markovian policies, and (5) devise a Simplex algorithm for an infinite-dimensional LP formulation of nonstationary infinite-horizon MDPs. Pivots in this Simplex algorithm use finite data, perform finite computations, and generate a sequence of improving extreme points that converges in value to optimal. Moreover, this sequence of extreme points gets arbitrarily close to the set of optimal extreme points. We also prove that decisions prescribed by these extreme points are eventually exactly optimal in all states of the nonstationary infinite-horizon MDP in early periods.

1 Introduction

Nonstationary infinite-horizon Markov decision processes (MDPs) [13] (henceforth called nonstationary MDPs) are one of the most general sequential decision models studied in operations research. Nonstationary MDPs extend the more well-studied stationary MDPs [38, 42] by relaxing the restrictive assumption that problem data do not change over time. From a practical viewpoint, nonstationary MDPs incorporate temporal changes in underlying economic and technological conditions into the decision-making process, and have been used to model problems such as asset selling [13] and stochastic inventory control [14]. They can be described as follows. A dynamic system is observed at the beginning of periods $n = 1, 2, \ldots$ by a decision maker to be in state $s \in \mathcal{S}$, where $\mathcal{S} \triangleq \{1, 2, \ldots, S\}$ is a finite set. The decision maker then chooses an action $a \in \mathcal{A}$, where $\mathcal{A} \triangleq \{1, 2, \ldots, A\}$ is also a finite set. Given that action a was chosen in state s

*Industrial and Systems Engineering, University of Washington, Seattle, USA; archis@uw.edu.

†Industrial and Operations Engineering, University of Michigan, Ann Arbor, USA; rlsmith@umich.edu.

in period n, the system makes a transition to state s' with probability $p_n(s'|s,a)$, incurring cost $0 \leq c_n(s,a;s') \leq c < \infty$. This procedure continues ad infinitum. Let $c_n(s,a)$ denote the expected cost incurred on choosing action a in state s in period n. That is, $c_n(s,a) = \sum_{s' \in \mathcal{S}} p_n(s'|s,a) c_n(s,a;s')$, and note that $0 \leq c_n(s,a) \leq c$. The decision maker's goal is to find a decision rule that minimizes total infinite-horizon discounted expected cost when the discount factor[1] is $0 < \alpha < 1$. This is an infinite-dimensional optimization problem. In fact, owing to infinite data requirements, it is not in general possible even to completely specify an instance of a truly nonstationary MDP. Thus the question is whether optimal solutions to a nonstationary MDP can be well-approximated by forecasting only a finite amount of probability and cost data (see Section 2 in [21] for a rigorous discussion of this issue). The only existing approach involves approximation by a sequence of longer and longer finite-horizon MDPs. This "planning horizon" approach is somewhat similar to the value iteration method [38, 42] for stationary MDPs, and it has been applied to various deterministic and stochastic sequential decision problems in [7, 8, 9, 11, 13, 14, 19, 20, 24, 25, 26, 27, 30, 29, 43, 45] and references therein. Reviews of this approach are available in [12, 21].

Success of the linear programming approach to stationary MDPs: Linear programming (LP) formulations [16, 17, 33] of stationary MDPs have recently been very successful in approximate solution of large-scale problems that were previously considered intractable [1, 2, 3, 4, 18, 32, 35, 36, 46, 47, 48, 52]. This is partly because the LP approach draws heavily from the power of LP duality, basic feasible solution characterization of extreme points, and efficient algorithms like the Simplex method combined with column generation. For instance, deterministic Markovian policies are in one-to-one correspondence with basic feasible solutions, and hence extreme points, of the dual of the LP formulation of a stationary MDP [38]. In addition, a violated constraint in this LP formulation provides an opportunity for policy improvement by pivoting in the corresponding variable in a Simplex algorithm for the dual problem. This leads to a close connection between Howard's classic policy iteration method [31], which updates actions in multiple states simultaneously, and the Simplex algorithm with so-called block pivots. In fact, the Simplex method has been called simple policy iteration, which updates an action in only one state at a time [51]. A new result by Ye [51] shows that Dantzig's original Simplex method with the most negative reduced cost pivoting rule [15] is strongly polynomial for solving stationary MDPs. This complexity bound is better than the polynomial performance of value iteration [49, 51], and in fact, is superior to the only known strongly polynomial time interior point algorithm [50] for solving stationary MDPs. Also see Chapter 6 of [38] for several insightful structural results from LP formulations of stationary MDPs. LP duality results have also been extended to stationary MDPs with uncountable state- and action-spaces [28]. Unfortunately, such LP-based theoretical and algorithmic advances have proven elusive for nonstationary MDPs.

Challenges in developing a linear programming approach to nonstationary MDPs: The major hurdle in developing an LP approach to nonstationary MDPs is that the LP formulation of a nonstationary MDP includes a countably infinite number of variables and constraints, and hence belongs to the class of countably infinite linear programs (CILPs) [6, 22]. Research on CILPs has traditionally faced several mathematical hurdles. For instance, the nonnegative orthant in $\Re^\infty$ has an empty interior in the product topology and thus standard interior point sufficient conditions (e.g., Theorem 3.13 in [6]) for strong duality do not hold. It is possible to construct examples

[1] A nonstationary MDP with time-dependent discount factors $0 < \alpha_n < 1$ can be converted into a nonstationary MDP with a constant discount factor $0 < \alpha < 1$ if α_n are uniformly bounded above by α. To see this, note that $\alpha_n = q_n \alpha$ for some $0 < q_n \leq 1$ for all n. Then define new cost functions $\gamma_n(\cdot,\cdot)$ by $\gamma_n(s,a) = q_n^{n-1} c_n(s,a)$ and note that $0 \leq \gamma_n(s,a) \leq c$ for all $n \in \mathbb{N}$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$. Thus we only consider nonstationary MDPs under the standard assumption of a time-invariant discount factor as in [13].

where a CILP and its dual possess a duality gap, and in fact, even weak duality can fail [40, 41] (also see Section 3 for an example). A CILP in the nonnegative orthant in $\Re^\infty$ that has an optimal solution need not have an extreme point optimal solution (see Section 3.7 in [6]; also recall that this cannot happen in a finite-dimensional LP by Theorem 2.7 in [10]). An extreme point of a CILP need not be a basic feasible solution [23, 39]. A pivot operation may require infinite computations and hence may not be implementable [6, 44]. Finally, a sequence of CILP extreme points with strictly improving objective function values may not converge in value to optimal [22]. Noticing such pathologies, Anderson and Nash commented on page 73 in their seminal book [6], "*... any algorithm will be difficult to implement; it is hard even to check feasibility*." Since then, to the best of our knowledge, only two Simplex-type algorithms have been published on CILPs. Sharkey and Romeijn [44] presented a Simplex method for minimum cost flow problems in a class of infinite networks. However, since the CILP formulation of nonstationary MDPs does not belong to this class, their approach is not applicable here. Ghate et al. [22] presented a Simplex-type method for a larger class of CILPs that subsumes nonstationary MDPs. That algorithm however was akin to a planning horizon approach and even though it produced a sequence of adjacent extreme points, it did not utilize duality and basic feasible solution characterization of extreme points, and in particular did not guarantee that the sequence was improving in objective values.

Contributions of this paper: Our goal in this paper is to overcome the above hurdles and develop a comprehensive LP approach to nonstationary MDPs. We first present a CILP formulation of the above nonstationary MDP and its dual CILP. We note that the dual CILP can be visualized as a minimum cost flow problem in a staged hypernetwork with infinite stages. We establish that this dual has an extreme point optimal solution. Then we prove that weak duality, complementary slackness, and strong duality hold owing to our choice of the variable-space for the primal and the structure of constraints in the dual. We then provide a definition of basic feasible solutions of the dual CILP and show that they are equivalent to its extreme points even though a "strictly positive support" condition that has recently been shown in [23] to be sufficient for such an equivalence cannot be established directly in our case. A one-to-one correspondence between deterministic Markovian policies and basic feasible solutions (or equivalently, extreme points) is also established. Finally, we present a Simplex method to solve the dual CILP. Each iteration of this Simplex method uses a finite amount of data, can be implemented finitely, and achieves the necessary magnitude of improvemnt in objective function value so that the sequence of extreme points visited converges in value to optimal. Similar to Dantzig's strongly polynomial time Simplex method with the most negative reduced cost pivot rule for stationary MDPs, our infinite-dimensional Simplex method uses a most negative approximate reduced cost rule (in the infinite-dimensional case, the most negative reduced cost cannot be found in finite time). As in stationary MDPs, our Simplex method can be viewed as simple policy iteration for nonstationary MDPs. The resulting sequence of extreme points gets arbitrarily close to the set of optimal extreme points. This fact is then used to prove that decisions prescribed by the Simplex method in early periods in all states of the nonstationary MDP are eventually exactly optimal.

2 A CILP formulation of nonstationary MDPs

To develop an LP formulation of the above nonstationary MDP, we first observe that it is equivalent to a stationary MDP with a countable state-space that is constructed by appending states $s \in \mathcal{S}$ with time-indices n. The states in this stationary MDP are given by $(n, s) \in \mathbb{N} \times \mathcal{S}$, where $\mathbb{N} \triangleq \{1, 2, \ldots\}$. Consequently, we adapt the CILP formulation for countable-state, finite-action, stationary MDPs given in [42] to our finite-state, finite-action, nonstationary MDP. In particular,

let $v_n(s)$ be the minimum infinite-horizon expected cost incurred starting time-period n in state s; $v_n(\cdot) : \mathcal{S} \to \Re$, for $n \in \mathbb{N}$, are called the optimal cost-to-go functions. Suppose $\beta \triangleq \{\beta_n\}$ is any sequence of *positive* vectors in $\Re^S$ such that $\sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \beta_n(s) < \infty$. Let $z \triangleq \{z_n\}$, with $z_n \in \Re^S$ for each $n \in \mathbb{N}$, denote sequences in $\prod_{n=1}^{\infty} \Re^S$. Also let $Z \subset \prod_{n=1}^{\infty} \Re^S$ be the subspace of such sequences with $\sup_{(n,s)} |z_n(s)| < \infty$. It follows from arguments in [42] (page 41) that values $v_n(s)$ equal the (unique) optimal values of variables $z_n(s)$ in the CILP

$$\max \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \beta_n(s) z_n(s) \tag{1}$$

$$z_n(s) - \alpha \sum_{s'\in\mathcal{S}} p_n(s'|s,a) z_{n+1}(s') \leq c_n(s,a), \text{ for } s \in \mathcal{S},\ a \in \mathcal{A},\ n \in \mathbb{N}, \tag{2}$$

$$z \in Z. \tag{3}$$

Owing to this interpretation of optimal values of variables $z_n(s)$, $0 \leq z_n(s) \leq \frac{c}{1-\alpha}$ for all (n,s) without loss of optimality in the above CILP.

It will be more convenient to work with an equivalent variant of the above CILP. We rewrite (1)-(3) by using a variable transformation and by making a specific choice for the sequence $\{\beta_n(s)\}$. In particular, we multiply the inequality constraint (2) by α^{n-1} and employ the variable transformation $y_n(s) = \alpha^{n-1} z_n(s)$ for all (n,s). Let $Y \subset \prod_{n=1}^{\infty} \Re^S$ be the subspace of all sequences $y \triangleq \{y_n\}$, with $y_n \in \Re^S$ for each $n \in \mathbb{N}$, such that $|y_n(s)| \leq \alpha^{n-1}\tau_y$ for all (n,s). Here, τ_y is some finite constant that may depend on y. Notice that if we did not allow τ_y to depend on y, then Y would not be a linear subspace. We also set $\beta_n(s) = \alpha^{n-1}$ for all (n,s), and note that for this choice, $\sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \beta_n(s) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \alpha^{n-1} = \frac{S}{1-\alpha} < \infty$ as required. This transforms the above CILP into the equivalent problem

$$(P)\ \max\ g(y) \triangleq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} y_n(s) \tag{4}$$

$$y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \leq \alpha^{n-1} c_n(s,a), \text{ for } s \in \mathcal{S},\ a \in \mathcal{A},\ n \in \mathbb{N}, \tag{5}$$

$$y \in Y. \tag{6}$$

The infinite series in the objective function of (P) converges absolutely for each $y \in Y$. To see this, note that

$$\sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} |y_n(s)| \leq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \alpha^{n-1}\tau_y = \tau_y S \sum_{n\in\mathbb{N}} \alpha^{n-1} = \frac{\tau_y S}{1-\alpha}.$$

We remark that whereas optimal value of variable $z_n(s)$ equals the optimal cost-to-go $v_n(s)$, the optimal value of variable $y_n(s)$ equals $v_n(s)$ discounted back to the first decision epoch. Therefore, without loss of optimality in (P), we have that $0 \leq y_n(s) \leq \alpha^{n-1}\frac{c}{1-\alpha}$ for all (n,s).

Let $x \triangleq \{x_n\}$, with $x_n \in \Re^{SA}$ for each $n \in \mathbb{N}$, denote sequences in $\prod_{n=1}^{\infty} \Re^{SA}$. We define the dual of (P) as

$$(D)\ \min\ f(x) \triangleq \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) \tag{7}$$

$$\sum_{a\in\mathcal{A}} x_1(s,a) = 1, \text{ for } s \in \mathcal{S}, \tag{8}$$

$$\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{n-1}(s|s',a)x_{n-1}(s',a) = 1, \text{ for } s \in \mathcal{S},\ n \in \mathbb{N}\setminus\{1\}, \tag{9}$$

$$x_n(s,a) \geq 0, \text{ for } s \in \mathcal{S},\ a \in \mathcal{A},\ n \in \mathbb{N}. \tag{10}$$

Lemma 2.1. *Suppose x is feasible to (D). Then, for each $n \in \mathbb{N}$, $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a) = nS$; since $x_n(s,a)$ are nonnegative, this also implies that $x_n(s,a) \leq nS$.*

Proof. In Appendix A. □

The infinite series in the objective function in (D) converges for each x that is feasible to (D). To see this, note that

$$\sum_{n\in\mathbb{N}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) \leq c\sum_{n\in\mathbb{N}}\alpha^{n-1}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a) \leq cS\sum_{n\in\mathbb{N}}\alpha^{n-1}n = \frac{cS}{(1-\alpha)^2},$$

because $\alpha < 1$.

Throughout this paper, we use the product topology. Thus a sequence $\{y^k\}$ converges to y if and only if $\{y_n^k\}$ converges in the usual Euclidean metric[2] in $\Re^S$ to y_n for every $n \in \mathbb{N}$. Similarly, a sequence $\{x^k\}$ converges to x if and only if $\{x_n^k\}$ converges in the usual Euclidean metric[3] in $\Re^{SA}$ to x_n for every $n \in \mathbb{N}$. Note that this product topology is a countable product of metrizable (see, for example, page Theorem 3.36 in [5]). For instance, the product topology on $\prod_{n=1}^{\infty}\Re^S$ is induced by the metric

$$\rho_1(y,y') = \sum_{n=1}^{\infty}\frac{1}{2^n}\frac{d_1(y_n,y_n')}{1+d_1(y_n,y_n')}, \tag{11}$$

where $y = \{y_n\}$ and $y' = \{y_n'\}$ with $y_n, y_n' \in \Re^S$ for each n, and $d_1(\cdot,\cdot)$ is the usual Euclidean metric on $\Re^S$. Similarly, the product topology on $\prod_{n=1}^{\infty}\Re^{SA}$ is induced by the metric

$$\rho_2(x,x') = \sum_{n=1}^{\infty}\frac{1}{2^n}\frac{d_2(x_n,x_n')}{1+d_2(x_n,x_n')}, \tag{12}$$

where $x = \{x_n\}$ and $x' = \{x_n'\}$ with $x_n, x_n' \in \Re^{SA}$ for each n, and $d_2(\cdot,\cdot)$ is the usual Euclidean metric on $\Re^{SA}$.

It is easy to show (see for example Proposition 2.7 in [22]) that the objective function in (D) is continuous, and the feasible region is nonempty and compact. Hence it has an optimal solution, justifying our use of min instead of inf.

It is helpful to visualize (D) as a staged, minimum cost flow problem in a hypernetwork with infinite stages. Stage n corresponds to the nth period in the nonstationary MDP. Each stage includes S nodes, each representing one state in $\mathcal{S}$. Each node has a supply of one unit as evident from the right hand sides of constraints (8)-(9). There are A hyperarcs emanating from each such node. Hyperarc (n,s,a) corresponds to action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ in stage n. Then $x_n(s,a)$ is the flow in this hyperarc, and $\alpha^{n-1}c_n(s,a)$ is the cost of sending unit flow through this hyperarc. For

[2]This in turn happens if and only if $y_n^k(s)$ converge as sequences of real numbers to $y_n(s)$ for every $s \in \mathcal{S}$.

[3]Again, this happens if and only if $x_n^k(s,a)$ converge as sequences of real numbers to $x_n(s,a)$ for every $(s,a) \in \mathcal{S}\times\mathcal{A}$.

each (n, s, a), let $\mathcal{J}_n(s, a)$ be the set of nodes in stage $n + 1$ that are reachable on choosing action a in node s in stage n. That is,

$$\mathcal{J}_n(s, a) = \{s' \in \mathcal{S} : p_n(s'|s, a) > 0\}. \tag{13}$$

Then, the hyperarc corresponding to action $a \in \mathcal{A}$ that emanates from the node representing state $s \in \mathcal{S}$ in stage $n \in \mathbb{N}$ has $|\mathcal{J}_n(s, a)|$ "heads". Furthermore, the flow reaching from node s to node $s' \in \mathcal{J}_n(s, a)$ equals $p_n(s'|s, a)x_n(s, a)$. Constraints (8) and (9) imply that flow is conserved at nodes (n, s) for all $n \in \mathbb{N}$ and $s \in \mathcal{S}$. For a feasible flow x, flow conservation implies that $\sum_{a\in\mathcal{A}} x_n(s, a) > 0$ for all $n \in \mathbb{N}$ and $s \in \mathcal{S}$. Figure 1 illustrates the structure of this hypernetwork.

3 Duality results

Challenges in proving duality results for CILPs have been well-documented [6, 40, 41, 44]. Here we present a motivating example adapted from [44] to illustrate that weak and strong duality can fail in CILPs. Consider a minimum cost flow problem in an infinite network with nodes numbered $i = 1, 2, \ldots$. There is a supply of 1 at node 1 and no supply or demand at other nodes. All arcs in the network are of the form $(i, i+1)$ with unit flow cost $1/2^i$ for $i = 1, 2, \ldots$. Using $x_{i,i+1}$ to denote the flow in arc $(i, i + 1)$, this problem is modeled by the CILP

$$\begin{aligned} \min \sum_{i=1}^{\infty} \frac{1}{2^i} x_{i,i+1} \\ x_{1,2} = 1 \\ x_{i,i+1} - x_{i-1,i} = 0,\ i = 2, 3, \ldots \\ x_{i,i+1} \geq 0,\ i = 1, 2, \ldots. \end{aligned}$$

Its dual is given by

$$\begin{aligned} \max\ y_1 \\ y_i - y_{i+1} \leq \frac{1}{2^i},\ i = 1, 2, \ldots. \end{aligned}$$

There is only one feasible solution to the primal, namely, the one wherein $x_{i,i+1} = 1$ for $i = 1, 2, \ldots$. Its cost equals 1. Hence this is the optimal primal cost. Moreover, for any number θ, solutions of the form $y_i = \theta$ for $i = 1, 2, \ldots$ are feasible to the dual. Thus, for any $\theta > 1$, we have a dual feasible solution with objective value larger than the optimal primal cost. Thus weak duality fails. In fact, the dual is unbounded and hence strong duality does not hold.

Romeijn et al. [41] and Romeijn and Smith [40] established a condition under which duality results hold for CILPs where every constraint includes a finite number of variables (as in (P)). This condition was presented in the context of primal and dual problems that only included inequality constraints with a lower staircase structure and nonnegative variables. Thus, to use their condition as is, we would need to convert (D) into that format. Although such a conversion is in principle possible, it is unnecessary. We instead establish duality results for (P) and (D) directly, using the method of proof in [40].

Theorem 3.1. (Weak Duality). *Suppose y and x are feasible to (P) and (D), respectively. Then*

$$f(x) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s, a)x_n(s, a) \geq g(y) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} y_n(s). \tag{14}$$

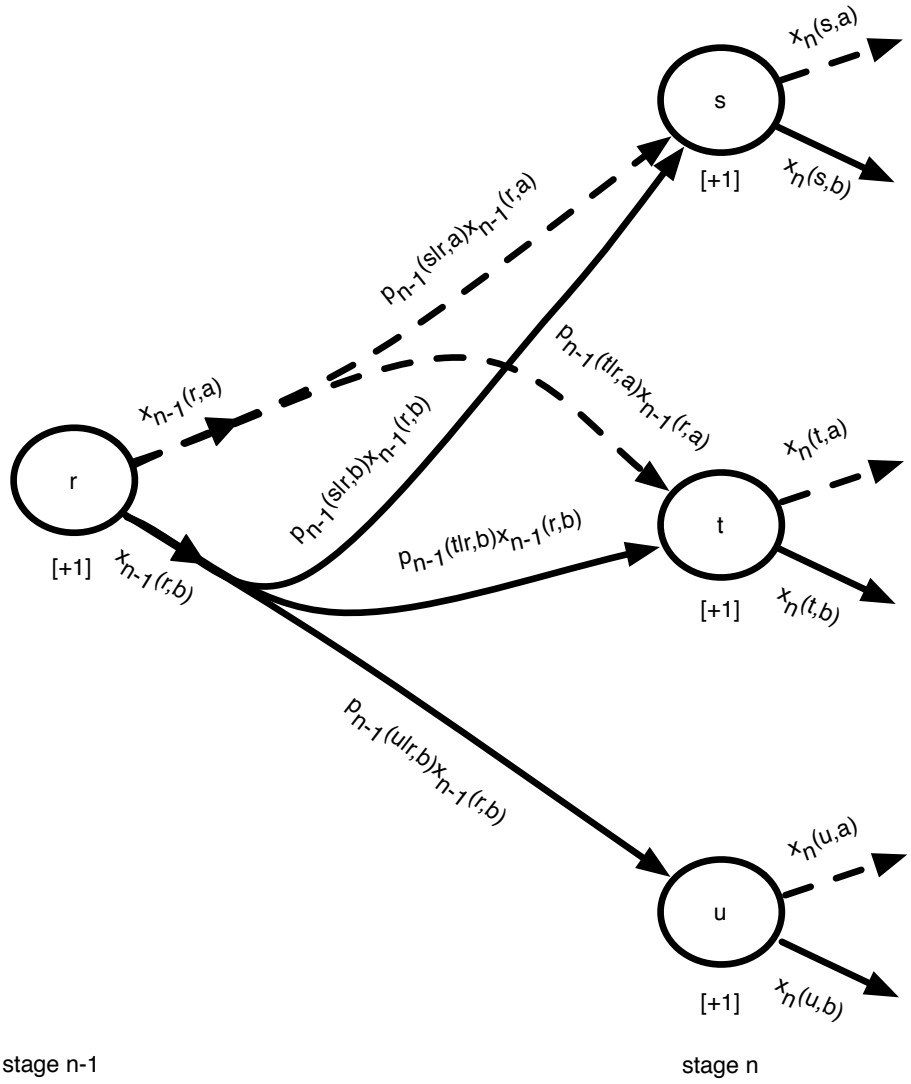

Figure 1: The picture shows a small piece of two stages in the hypernetwork associated with problem (D) wherein $\mathcal{S} = \{r, s, t, u\}$ and $\mathcal{A} = \{a, b\}$. Only state r in stage $n-1$ and states s, t, u in stage n are shown to avoid crowding. A flow supply of $[+1]$ is available at each state. Two hyperarcs emanating from state r in stage $n-1$ are shown. The dotted hyperarc corresponds to choosing action a in state r in stage $n-1$ and has two heads, namely, states s and t. The solid hyperarc corresponds to choosing action b in state r in stage $n-1$ and has three heads, namely, states s, t, and u. The dotted hyperarc carries flow $x_{n-1}(r, a)$, which is split into two portions: flow $p_{n-1}(s|r, a)x_{n-1}(r, a)$ reaches state s whereas flow $p_{n-1}(t|r, a)x_{n-1}(r, a)$ reaches state t. The solid hyperarc carries flow $x_{n-1}(r, b)$, which is split into three portions: flow $p_{n-1}(s|r, b)x_{n-1}(r, b)$ reaches state s, flow $p_{n-1}(t|r, b)x_{n-1}(r, b)$ reaches state t, and flow $p_{n-1}(u|r, b)x_{n-1}(r, b)$ reaches state u. Flows in hyperarcs corresponding to actions a and b in states s, t, u in stage n are also shown so that the reader can visualize flow conservation constraints (9) in (D).

Proof. In Appendix B. □

Corollary 3.2. *Suppose y and x are feasible to (P) and (D), respectively. If equality holds in (14), then y is optimal to (P) and x is optimal to (D).*

Definition 3.3. *Suppose x is feasible to (D) and $y \in Y$. Then we say that x and y satisfy* complementary slackness *if*

$$x_n(s,a)\Big[\alpha^{n-1}c_n(s,a) - \Big(y_n(s) - \sum_{s' \in \mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\Big)\Big] = 0,\ \forall s \in \mathcal{S},\ a \in \mathcal{A},\ n \in \mathbb{N}. \tag{15}$$

Theorem 3.4. (Complementary Slackness Sufficiency). *Suppose x is feasible to (D) and satisfies complementary slackness with some $y \in Y$. Then*

$$f(x) = g(y). \tag{16}$$

If y is feasible to (P), then y and x are optimal to (P) and (D), respectively.

Proof. In Appendix C. □

Recall that while stating problem (P), we had argued, based on [42] that it has an optimal solution. Also recall from Section 2 that (D) has an optimal solution. The next result establishes that there is no duality gap between (P) and (D).

Theorem 3.5. (Strong Duality). *Problems (P) and (D) have optimal solutions and their optimal objective function values are equal.*

Proof. In Appendix D. □

Theorem 3.6. (Complementary Slackness Necessity). *Suppose y and x are optimal to (P) and (D), respectively. Then (15) holds.*

Proof. In Appendix E. □

4 Characterization of extreme points

In finite-dimensional LPs with equality constraints and nonnegative variables, a solution is called basic if it can be obtained as the unique solution to the system of equations formed by setting a subset of the variables to zero. If this solution is nonnegative, then it is called a basic feasible solution. The variables that are selected to set to zero are called nonbasic whereas the remaining ones are called basic. A feasible solution is called an extreme point if it cannot be expressed as a strict convex combination of two other distinct, feasible solutions. It is well-known that a feasible solution is an extreme point if and only if it is a basic solution [10]. This equivalence does not hold in CILPs — a basic feasible solution is an extreme point but an extreme point need not be a basic feasible solution [23, 39]. We show that this pathological scenario does not arise in (D).

Definition 4.1. *A feasible solution x to (D) is called an* extreme point *if it cannot be written as $x = \lambda w + (1-\lambda)z$, where $\lambda \in (0,1)$ and $w \neq z$ are distinct from x and are feasible to (D).*

Recall from Section 2 that (D) has an optimal solution. In fact, since the feasible region of (D) is convex and the objective function is linear, Bauer's Maximum Principle [5] implies that (D) has an extreme point optimal solution (also see Proposition 2.7 in [22]). Also recall from Section 2 that, in any feasible solution x to (D), $x_n(s,a) > 0$ for *at least one* $a \in \mathcal{A}$ for each $n \in \mathbb{N}$ and $s \in \mathcal{S}$. We then have

Definition 4.2. *Suppose x is feasible to (D). We call it a* basic feasible solution *of (D) if, for every $n \in \mathbb{N}$ and $s \in \mathcal{S}$, there is* exactly one *action $a_n(s) \in \mathcal{A}$ for which $x_n(s, a_n(s)) > 0$.*

The actions (or equivalently, the hyperarcs) $a_n(s)$ for which $x_n(s, a_n(s)) > 0$ will be called *basic actions*. Other actions will be called *nonbasic*. Note that any selection of basic hyperarcs $a_n(s)$ for all $n \in \mathbb{N}$ and $s \in \mathcal{S}$ uniquely determines flows $x_n(s, a_n(s))$. Specifically, they are given recursively in the order $n = 1, 2, \ldots$ by

$$x_n(s, a_n(s)) = 1 + \sum_{s' \in \mathcal{I}_{n-1}(s)} p_{n-1}(s|s', a_{n-1}(s')) x_{n-1}(s', a_{n-1}(s')), \tag{17}$$

where

$$\mathcal{I}_{n-1}(s) = \{s' \in \mathcal{S} : p_{n-1}(s|s', a_{n-1}(s')) > 0\},$$

with the convention that $\mathcal{I}_0(s) = \emptyset$ for all $s \in \mathcal{S}$. Since $x_{n-1}(s', a_{n-1}(s')) \geq 0$, the above recursion implies that $x_n(s, a_n(s)) \geq 1 > 0$. That is, every basic feasible solution is "nondegenerate."

We now characterize extreme points of (D) as basic feasible solutions of (D). It is known that a basic feasible solution of (D) is also an extreme point of (D) [23]. We nevertheless provide a short proof of this fact here for completeness. But it is not evident at first glance whether every extreme point x of (D) is also a basic feasible solution. For this to hold, it is sufficient for x to have a "strictly positive support" [23]; that is, if $\Omega(x)$ is the set of hyperarcs (n, s, a) such that $x_n(s, a) > 0$, then $\left[\inf_{(n,s,a) \in \Omega(x)} x_n(s, a)\right] > 0$. Unfortunately, there is no obvious way to establish directly that this condition holds at every extreme point of (D). Nevertheless, we show below that every extreme point of (D) is indeed a basic feasible solution using a more concrete argument that uses the structure of the hypernetwork underlying (D).

Theorem 4.3. *A feasible solution x of (D) is an extreme point if and only if it is a basic feasible solution.*

Proof. In Appendix F. □

The "only if" part of the above theorem yields

Corollary 4.4. *If x is an extreme point of (D), then $x_n(s, a) \geq 1$ for all hyperarcs (n, s, a) in the support set $\Omega(x)$. That is, x has strictly positive support.*

Since (D) has an extreme point optimal solution, Theorem 4.3 yields

Corollary 4.5. *(D) has an optimal solution that is a basic feasible solution.*

The term deterministic Markovian policy refers to a decision rule that assigns one action to each possible state, irrespective of the earlier states visited and of the previous actions taken, over the infinite-horizon [38, 42]. Under the discounted cost optimality criterion, deterministic Markovian policies are optimal to countable state stationary MDPs with bounded costs and finite action sets in each state (see Theorem 2.2 on page 32 in [42]). Consequently, we can also limit attention to such policies without loss of optimality in our nonstationary MDP. Definition 4.2 establishes a one-to-one correspondence between basic feasible solutions of (D) and deterministic Markovian policies for the nonstationary MDP. In particular, if x is a basic feasible solution of (D), then the basic actions $a_n(s)$ define a unique deterministic Markovian poclicy. Similarly, if π is a deterministic Markovian policy, then we can construct a unique basic feasible solution to (D) using the actions $\pi_n(s)$ prescribed by pocliy π in states (n, s) as the basic actions. Our interest in basic feasible solutions stems from the following result, which implies that an optimal basic feasible solution to (D) defines an optimal deterministic Markovian policy for the nonstationary MDP.

Theorem 4.6. *Suppose x^* is an optimal basic feasible solution to (D). Then for each $n \in \mathbb{N}$ and $s \in \mathcal{S}$, the action $a_n(s)$ with $x^*_n(s, a_n(s)) > 0$ is optimal for the nonstationary MDP in state s in period n.*

Proof. In Appendix G. □

Lemma 4.7. *Suppose x is a basic feasible solution of (D). For all $n \in \mathbb{N}$ and $s \in \mathcal{S}$, let $y_n(s)$ be the expected cost-to-go, discounted back to the first period, incurred on implementing the deterministic Markovian policy defined by x, starting in state s in period n. Then this y is the unique element of Y that satisfies complementary slackness with x. Note that this y need not be feasible to (P).*

Proof. In Appendix H. □

The y defined above will be called *the* solution complementary to the basic feasible solution x. Since complementary variables $y_n(s)$ equal discounted expected costs-to-go, they satisfy $0 \leq y_n(s) \leq \alpha^{n-1} \frac{c}{1-\alpha}$ because $0 \leq c_n(s, a) \leq c$ for all $n \in \mathbb{N}$, $s \in \mathcal{S}$, and $a \in \mathcal{A}$.

5 Simplex algorithm

In finite-dimensional minimization LPs with equality constraints and nonnegative variables, the Simplex method works as follows. It starts at an initial extreme point, and at each iteration, moves along an edge of the feasible polytope to a new adjacent extreme point. This geometric notion can be implemented algebraically by swapping one nonbasic variable in a basic feasible solution with a basic variable [10]. This is called a pivot operation. The concept of a reduced cost is used to ensure that the objective function value is improved in each pivot operation. The algorithm reaches an optimal extreme point after a finite number of iterations and then stops. The difficulties in replicating this in the context of CILPs, even when duality results and basic feasible solution characterization of extreme points are available, have been outlined in [6, 22, 44]. Even checking feasibility of a given solution may in general require infinite data and computations. It is not possible in general to "store" a solution on a computer. Moving from one extreme point to an adjacent, improving extreme point may require infinite computations. To make matters worse, unlike in finite-dimensional LPs, a strictly improving sequence of extreme points may not converge in value to optimal as demonstrated by example in [22]. Our Simplex method successfully overcomes these hurdles.

Suppose x is a basic feasible solution of (D) and let $y \in Y$ be its complementary solution defined in Lemma 4.7. For every hyperarc (n, s, a),

$$\gamma_n(s, a) \triangleq \alpha^{n-1} c_n(s, a) + \sum_{s' \in \mathcal{S}} p_n(s'|s, a) y_{n+1}(s') - y_n(s) \tag{18}$$

is the slack in the corresponding constraint (5) in (P). In view of Definition 4.2 and complementary slackness equations (15), $\gamma_n(s, a) = 0$ if hyperarc (n, s, a) is basic. Moreover, if $\gamma_n(s, a) \geq 0$ for all nonbasic hyperarcs (n, s, a), then x is optimal to (D) because y is feasible to (P). A pivot operation involves finding a nonbasic hyperarc and adding it to the set of basic hyperarcs. If nonbasic hyperarc (n, s, a) is chosen for this purpose, then basic hyperarc $(n, s, a_n(s))$ must leave the set of basic arcs according to Definition 4.2. The new values of basic variables are then uniquely defined by the equality constraints in (D). Let z denote this new basic feasible solution.

Proposition 5.1. *The difference in objective function values at basic feasible solution x and the new basic feasible solution z in the aforementioned pivot operation is given by*

$$f(z) - f(x) = (1+\theta)\gamma_n(s,a), \tag{19}$$

where $\theta > 0$ is a constant that depends on x, n and s.

Proof. In Appendix I. □

This shows that, similar to finite-dimensional LPs, the slack $\gamma_n(s,a)$ can be interpreted as the reduced cost of hyperarc (n,s,a). In particular, selecting a nonbasic hyperarc with negative reduced cost guarantees that the pivot operation will improve the objective function value. We thus have

Corollary 5.2. *If basic feasible solution x is optimal to (D), then its complementary solution y is feasible and hence optimal to (P).*

Proof. Since x is optimal, all reduced costs, that is, slacks in constraints (5) in (P) for the complementary solution y must be nonnegative. That is, y must be feasible to (P). Consequently, by Theorem 3.4, y must also be optimal to (P). □

However, as noted above, it is not adequate to simply construct a sequence of improving extreme points. Intuitively, we need to ensure that "sufficient" improvement is made in each pivot operation. Although we cannot find the direction of greatest improvement finitely, our goal is to devise a Simplex algorithm that uses only finite amount of data and finite computations to find a nonbasic arc with a sufficiently negative reduced cost in each iteration and to move to a new extreme point so that the resulting sequence of solutions converges in value to optimal. We will show that the Simplex algorithm below accomplishes this objective.

A Simplex algorithm

1. Initialize: Set iteration counter $k = 1$. Fix basic actions $a_n^1(s)$ for $s \in \mathcal{S}$ and $n \in \mathbb{N}$[4]. We denote the corresponding basic feasible solution of (D) by x^1.

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

   (a) Set $m = 1$ and define $m(k) \triangleq \infty$ and $\gamma^{k,\infty} \triangleq 0$.

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

   $$y_n^{k,m}(s) = \alpha^{n-1} c_n(s, a_n^k(s)) + \sum_{s' \in \mathcal{S}} p_n(s'|s, a_n^k(s)) y_{n+1}^{k,m}(s'), \text{ for } s \in \mathcal{S},\ n \le m, \tag{20}$$

   $$y_{m+1}^{k,m}(s) = 0. \tag{21}$$

   (c) Compute approximate nonbasic reduced costs

   $$\gamma_n^{k,m}(s,a) = \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y_{n+1}^{k,m}(s') - y_n^{k,m}(s) \tag{22}$$

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

[4]This set of basic actions can be described finitely. For example, since set $\mathcal{A}$ is finite, the "first" action from $\mathcal{A}$ can be the basic action for every (n,s).

(d) Compute the smallest approximate nonbasic reduced cost

$$\gamma^{k,m} = \min_{\substack{n \le m \\ s \in \mathcal{S}, a \in \mathcal{A} \\ a \neq a_n^k(s)}} \gamma_n^{k,m}(s,a). \tag{23}$$

(e) If $\gamma^{k,m} < -\alpha^m \frac{c}{1-\alpha}$, set $m(k) = m$, let (n^k, s^k, a^k) be the argmin in (23), set $a_{n^k}^{k+1}(s^k) = a^k$ as the new basic action in state s^k in period n^k and go to Step 3 below; else set $m = m+1$ and go to Step 2(b) above.

3. Obtain basic hyperarc flows in the first $m(k)$ periods of x^{k+1}: Compute $x_n^{k+1}(s, a_n^{k+1}(s))$ for all $s \in \mathcal{S}$ using formula (17) in the order $n = 1, 2, \ldots, m(k)$.

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

Let y^k be the complementary primal solution of the dual extreme point x^k. The $y^{k,m}$ calculated in Step 2(b) is an approximation of this y^k. Although the reduced costs computed are approximate, we show below that the sequence of objective function values is strictly improving to optimality. In particular, we emphasize that this Simplex algorithm does *not* solve to optimality a sequence of longer and longer finite-horizon LPs. It instead works directly with extreme points x^k of (D). To ensure a finite implementation of pivots using finite data and strictly improving objective function values, the algorithm uses a good enough approximation of y^k for the reduced cost calculation in Step 2(c). In particular, this is *not* a planning horizon approach but rather what we call a *strategy horizon* approach. As in stationary MDPs, the Simplex algorithm is a simple policy iteration method because Step 2 of the algorithm finds a period n^k, a state s^k, and updates the decision in (n^k, s^k) from $a_{n^k}^{k-1}(s^k)$ to a^k. Consistent with this view, it is not necessary to compute the basic hyperarc flows in x^{k+1} in Step 3 because these flow values are not used by the algorithm — it suffices to simply know the set of basic actions. We nevertheless include Step 3 in the algorithm to emphasize the hypernetwork flow interpretation of our nonstationary MDP.

Let $f^k \triangleq f(x^k)$ be the sequence of objective function values of basic feasible solutions x^k of (D) visited by the above Simplex algorithm. The rest of this section is devoted to proving the following key theorem.

Theorem 5.3. *Let f^* be the optimal value of (D). Then $\lim_{k\to\infty} f^k = f^*$. Moreover, for any $\epsilon > 0$, there exists an iteration k_ϵ such that $\rho_2(x^k, x^{*k}) < \epsilon$ for some optimal basic feasible solution x^{*k} of (D) for all $k \ge k_\epsilon$.*

The second claim above means that the sequence x^k of basic feasible solutions eventually stays arbitrarily close to some optimal basic feasible solution to (D). The proof of Theorem 5.3 is quite long so we break it into multiple parts. The first five parts are established in five separate lemmas below. The last part of the proof uses these five lemmas. The first lemma provides quality-of-approximation bounds for $y^{k,m}$.

Lemma 5.4. *The approximation $y^{k,m}$ of y^k in Step 2(c) of the Simplex algorithm satisfies*

$$y_n^{k,m}(s) \le y_n^k(s) \le y_n^{k,m}(s) + \alpha^m \frac{c}{1-\alpha} \text{ for } s \in \mathcal{S},\ n = 1, 2, \ldots, m+1. \tag{24}$$

Proof. By complementary slackness, for each iteration k, y^k is the solution of the infinite system

$$y_n^k(s) = \alpha^{n-1} c_n(s, a_n^k(s)) + \sum_{s' \in \mathcal{S}} p_n(s'|s, a_n^k(s)) y_{n+1}^k(s'), \text{ for } s \in \mathcal{S},\ n \in \mathbb{N}. \tag{25}$$

Since $y^{k,m}$ is the solution of the m-horizon truncation (20)-(21) of this infinite system, and since $y^k_n(s) \geq 0$ for all n by the discussion following Lemma 4.7, we have that

$$y^k_n(s) \geq y^{k,m}_n(s) \text{ for } n = 1, 2, \ldots, m+1. \tag{26}$$

Moreover, since $y^k_{m+1}(s) \leq \alpha^m \frac{c}{1-\alpha}$ and $y^{k,m}_{m+1}(s) = 0$, equations (20) and (25) imply that

$$y^k_m(s) \leq y^{k,m}_m(s) + \alpha^m \frac{c}{1-\alpha} \text{ for } s \in \mathcal{S}. \tag{27}$$

Using this recursively in (20) and (25), we get

$$y^k_n(s) \leq y^{k,m}_n(s) + \alpha^m \frac{c}{1-\alpha} \text{ for } s \in \mathcal{S},\ n = 1, 2, \ldots, m+1. \tag{28}$$

Combining this with (26), we get (24). □

Lemma 5.5. *Step 2 of the Simplex algorithm terminates at a finite value of m if and only if x^k is not optimal to (D).*

Proof. Suppose x^k is not optimal to (D). Since x^k is not optimal, y^k must not be feasible to (P) by Theorem 3.4. Thus there exist a period n, a state $s \in \mathcal{S}$, an action $a \in \mathcal{A}$, and an $\epsilon > 0$ such that

$$-\epsilon = \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y^k_{n+1}(s') - y^k_n(s).$$

Then using (24) from Lemma 5.4 we get

$$-\epsilon \geq \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y^{k,m}_{n+1}(s') - y^{k,m}_n(s) - \alpha^m \frac{c}{1-\alpha}, \text{ for all } m \geq n.$$

That is,

$$-\epsilon + \alpha^m \frac{c}{1-\alpha} \geq \gamma^{k,m}_n(s,a), \text{ for all } m \geq n.$$

But since $\alpha^m \frac{c}{1-\alpha} < \epsilon/2$ for all sufficiently large m, we have that $-\epsilon/2 > \gamma^{k,m}_n(s,a) \geq \gamma^{k,m}$ for all m large enough. Here, the last inequality follows from the definition of $\gamma^{k,m}$ in (23). Now notice that $-\alpha^m \frac{c}{1-\alpha} > -\epsilon/2$ for all such m. Consequently, $-\alpha^m \frac{c}{1-\alpha} > \gamma^{k,m}$ for all such m. Thus the condition in Step 2(e) is eventually met and Step 2 terminates.

Now suppose that x^k is optimal to (D). Suppose Step 2 terminates at some $m(k)$. Then $\gamma^{k,m(k)} + \alpha^{m(k)} \frac{c}{1-\alpha} < 0$. That is, $\gamma^{k,m(k)}_{n^k}(s^k, a^k) + \alpha^{m(k)} \frac{c}{1-\alpha} < 0$, where (n^k, s^k, a^k) is the argmin in (23). Then using (24) from Lemma 5.4 we get

$$\alpha^{n^k-1} c_{n^k}(s^k, a^k) + \sum_{s' \in \mathcal{S}} p_{n^k}(s'|s^k, a^k) y^k_{n^k+1}(s') - y^k_{n^k}(s) < 0.$$

Thus y^k is not feasible to (P). But this contradicts Corollary 5.2. □

The "only if" part of the above lemma implies that Step 2 does not terminate finitely when x^k is an optimal solution. In this sense, the algorithm cannot tell that an optimal solution has been found. This may, at first sight, appear to be a weakness of our algorithm. However, it is in fact due to a feature of problem (D) itself, and more generally, of nonstationary infinite-horizon optimization problems — optimality of a given solution cannot be affirmed with finite computations (see [13]). Fortunately, this does not undermine the validity of Theorem 5.3 as its conclusions are trivially true if x^k is optimal for some k and we simply repeat this solution for all subsequent k. The next lemma establishes that the Simplex algorithm produces an improving sequence of basic feasible solutions.

Lemma 5.6. *If x^k is not optimal to (D), then $f^{k+1} < f^k$. Moreover, $\left[\alpha^{m(k)}\frac{c}{1-\alpha} + \gamma^{k,m(k)}\right] \to 0$ as $k \to \infty$.*

Proof. Since x^k is not optimal, Step 2 of the algorithm terminates finitely by Lemma 5.5. By Proposition 5.1, the difference between objective function values of y^{k+1} and y^k, and hence between f^{k+1} and f^k is given by

$$\delta^k \triangleq f^{k+1} - f^k = (1+\theta_k)\Big[\alpha^{n^k-1}c_{n^k}(s^k,a^k) + \sum_{s'\in\mathcal{S}} p_{n^k}(s'|s^k,a^k)y^k_{n^k+1}(s') - y^k_{n^k}(s^k)\Big] \tag{29}$$

for some $\theta_k > 0$. Since $n^k \le m(k)$ by construction, we can use (24) in Lemma 5.4 to bound δ^k. In particular, we have

$$\begin{aligned}\delta^k &\le (1+\theta_k)\Big[\alpha^{n^k-1}c_{n^k}(s^k,a^k) + \sum_{s'\in\mathcal{S}} p_{n^k}(s'|s^k,a^k)y^{k,m(k)}_{n^k+1}(s') + \alpha^{m(k)}\frac{c}{1-\alpha} - y^{k,m(k)}_{n^k}(s)\Big]\\ &= (1+\theta_k)\Big[\alpha^{m(k)}\frac{c}{1-\alpha} + \gamma^{k,m(k)}\Big] < 0,\end{aligned}$$

because $\gamma^{k,m(k)} < -\alpha^{m(k)}\frac{c}{1-\alpha}$ by Step 2(e) of the Simplex algorithm. This shows that $f^{k+1} < f^k$. Now, for the second claim, note that if x^k is optimal for any k, then Step 2 does not terminate, and hence $\left[\alpha^{m(k)}\frac{c}{1-\alpha} + \gamma^{k,m(k)}\right] = 0$. If x^k is not optimal for any k, then the algorithm produces a sequence of basic feasible solutions with $\theta_k > 0$ and hence $f^{k+1} < f^k + \left[\alpha^{m(k)}\frac{c}{1-\alpha} + \gamma^{k,m(k)}\right]$. Since sequence f^k is bounded below by zero and $f^1 < \infty$, this implies that $\left[\alpha^{m(k)}\frac{c}{1-\alpha} + \gamma^{k,m(k)}\right] \to 0$ as $k \to \infty$. □

Lemma 5.7. *The sequence $m(k) \to \infty$ as $k \to \infty$. Also, $\gamma^{k,m(k)} \to 0$ as $k \to \infty$.*

Proof. The Lemma holds trivially if x^k is optimal for any k. So we focus on the situation where this is not the case.

For the first claim, we need to show that for every period n, there exists an integer M_n such that $m(k) \ge n$ for all $k \ge M_n$. Suppose not. Then there exists some period n such that $m(k) < n$ for infinitely many k. As a result, there is an integer $M < n$ such that $m(k) = M$ for infinitely many k. Let k_i, for $i = 1, 2, \ldots$, define the infinite subsequence of iterations in which this occurs. Let $\pi^{k_i,M}$ be the M-horizon deterministic Markovian policy defined by basic actions $a^{k_i}_n(s)$, for $n = 1, 2, \ldots, M$ and $s \in \mathcal{S}$, in the basic feasible solution x^{k_i} to (D) in iteration k_i. A close look at the finite system of equations solved in Step 2(b) of the algorithm affirms that values $y^{k_i,M}$ depend only on basic actions in the first M periods, that is, only on $\pi^{k_i,M}$. As a result, the reduced costs $\gamma^{k_i,M}$ also depend only on $\pi^{k_i,M}$, and hence we denote them by $\gamma(\pi^{k_i,M})$. Note that there are only a finite number of deterministic Markovian policies for the M-horizon truncation of the nonstationary MDP. As a result, there must exist an M-horizon deterministic Markovian policy $\pi^{*,M}$ and a corresponding infinite subsequence k_{i_j} of iterations k_i such that $\pi^{k_{i_j},M} = \pi^{*,M}$. As in the proof of Lemma 5.6 we have $f^{k_{i_j}+1} < f^{k_{i_j}} + \left[\alpha^{m(k_{i_j})}\frac{c}{1-\alpha} + \gamma^{k_{i_j},m(k_{i_j})}\right]$. But since $m(k_{i_j}) = M$, we get

$$\begin{aligned}f^{k_{i_j}+1} &< f^{k_{i_j}} + \Big[\alpha^M\frac{c}{1-\alpha} + \gamma^{k_{i_j},M}\Big] = f^{k_{i_j}} + \Big[\alpha^M\frac{c}{1-\alpha} + \gamma(\pi^{k_{i_j},M})\Big]\\ &= f^{k_{i_j}} + \Big[\alpha^M\frac{c}{1-\alpha} + \gamma(\pi^{*,M})\Big] = f^{k_{i_j}} - \epsilon,\end{aligned}$$

where $\epsilon = -\left[\alpha^M \frac{c}{1-\alpha} + \gamma(\pi^{*,M})\right] > 0$ is a constant that depends only on M and $\pi^{*,M}$. This implies that the objective value in (D) is reduced by at least $\epsilon > 0$ in each iteration that belongs to the infinite sequence of iterations k_{i_j}. But this is impossible since sequence f^k is bounded below by zero and $f^1 < \infty$. This proves the first claim by contradiction.

To prove the second claim, we recall from Lemma 5.6 that $\alpha^{m(k)} \frac{c}{1-\alpha} + \gamma^{k,m(k)} \to 0$ as $k \to \infty$. Moreover, $\alpha^{m(k)} \frac{c}{1-\alpha} \to 0$ as $k \to \infty$ because $m(k) \to \infty$. Hence we must have that $\gamma^{k,m(k)} \to 0$ as $k \to \infty$. □

Lemma 5.8. *Let x^k be any convergent sequence of basic feasible solutions of (D) and $\bar{x}$ be its limit. Then $\bar{x}$ is also a basic feasible solution of (D).*

Proof. Since x^k are feasible, it is easy to see, as in the proof of Theorem 5.3 below, that $\bar{x}$ is feasible. Suppose it is not basic. That is, there is an $n \in \mathbb{N}$ and $s \in \mathcal{S}$ and two distinct actions $a, b \in \mathcal{A}$ such that $\bar{x}_n(s,a) > 0$ and $\bar{x}_n(s,b) > 0$. Let $\delta = \min\{\bar{x}_n(s,a), \bar{x}_n(s,b)\}$. Since x^k converges to $\bar{x}$ in the product topology, there exists a K such that $0 < \bar{x}_n(s,a) - \delta/2 < x^k_n(s,a) < \bar{x}_n(s,a) + \delta/2$ and $0 < \bar{x}_n(s,b) - \delta/2 < x^k_n(s,b) < \bar{x}_n(s,b) + \delta/2$. This contradicts the fact that x^k is a basic feasible solution. □

Now we are ready to complete the proof of Theorem 5.3. As noted earlier, conclusions of the theorem are trivially true if x^k is optimal to (D) for any k. We therefore assume that x^k is not optimal for any k. Let x^{k_i} be a convergent subsequence of x^k with $\lim_{i\to\infty} x^{k_i} = \bar{x}$. Such a sequence exists because the feasible region of (D) is compact in the metrizable product topology by Lemma 2.1 and Tychonoff's product theorem (Theorem 37.3 in [34]). Let y^{k_i} be the corresponding subsequence of y^k. Subsequence y^{k_i} has a further convergent subsequence because y^{k_i} belongs to set $\mathcal{C} = \left\{y : 0 \le y_n(s) \le \alpha^{n-1}\frac{c}{1-\alpha},\ \forall s \in \mathcal{S},\ n \in \mathbb{N}\right\}$ that is compact in the metrizable product topology by Tychonoff's product theorem. We denote this by $y^{k_{i_j}}$ and let $\lim_{j\to\infty} y^{k_{i_j}} = \bar{y}$. We also let $x^{k_{i_j}}$ be the corresponding subsequence of x^{k_i} and note that $x^{k_{i_j}}$ also must converge to $\bar{x}$. Similarly, $\gamma^{k_{i_j},m(k_{i_j})}$ is the corresponding subsequence of $\gamma^{k,m(k)}$ and $\lim_{j\to\infty} \gamma^{k_{i_j},m(k_{i_j})} = 0$. We show that $\bar{x}$ is feasible to (D), $\bar{y}$ is feasible to (P) and $\bar{x}$ and $\bar{y}$ satisfy complementary slackness conditions. This will imply that $\bar{x}$ is optimal to (D) by Theorem 3.4.

Since $x^{k_{i_j}}$ are feasible to (D) for all j, they satisfy (8)-(10). That is,

$$\sum_{a\in\mathcal{A}} x_1^{k_{i_j}}(s,a) = 1, \text{ for } s \in \mathcal{S},$$
$$\sum_{a\in\mathcal{A}} x_n^{k_{i_j}}(s,a) - \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{n-1}(s|s',a) x_{n-1}^{k_{i_j}}(s',a) = 1, \text{ for } s \in \mathcal{S},\ n \in \mathbb{N}\setminus\{1\},$$
$$x_n^{k_{i_j}}(s,a) \ge 0, \text{ for } s \in \mathcal{S},\ a \in \mathcal{A},\ n \in \mathbb{N}.$$

Taking limits as $j \to \infty$ in the above three, it is clear that $\bar{x}$ also satisfies (8)-(10) and hence is feasible to (D).

Now suppose that $\bar{y}$ is not feasible to (P). This implies that there exist some $n \in \mathbb{N}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $\epsilon > 0$ such that

$$\bar{y}_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)\bar{y}_{n+1}(s') - \alpha^{n-1}c_n(s,a) = \epsilon.$$

That is,

$$\lim_{k\to\infty}\Big(y_n^{k_{i_j}}(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}^{k_{i_j}}(s')-\alpha^{n-1}c_n(s,a)\Big)=\epsilon.$$

Thus there exists a positive integer J such that

$$\frac{-\epsilon}{2}\geq\Big[\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}^{k_{i_j}}(s')+\alpha^{n-1}c_n(s,a)-y_n^{k_{i_j}}(s)\Big]\geq\frac{-3\epsilon}{2}$$

for all $j\geq J$. But for all j that are large enough, $n\leq m(k_{i_j})$, since $m(k_{i_j})\to\infty$ as $j\to\infty$ by Lemma 5.7. Therefore, for all such j, we have from (24) that

$$\begin{aligned}&\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}^{k_{i_j}}(s')+\alpha^{n-1}c_n(s,a)-y_n^{k_{i_j}}(s)\\&\geq\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}^{k_{i_j},m(k_{i_j})}(s')+\alpha^{n-1}c_n(s,a)-\frac{c\alpha^{m(k_{i_j})}}{1-\alpha}-y_n^{k_{i_j},m(k_{i_j})}(s)\\&\geq\gamma^{k_{i_j},m(k_{i_j})}-\frac{c\alpha^{m(k_{i_j})}}{1-\alpha}.\end{aligned}$$

Consequently, $-\epsilon/2\geq\gamma^{k_{i_j},m_{k_{i_j}}}-\frac{c\alpha^{m(k_{i_j})}}{1-\alpha}$ for all these j. But this contradicts the fact that both $\gamma^{k_{i_j},m_{k_{i_j}}}$ and $\frac{c\alpha^{m(k_{i_j})}}{1-\alpha}$ converge to zero as $j\to\infty$ by Lemma 5.7.

Since $x^{k_{i_j}}$ and $y^{k_{i_j}}$ satisfy complementary slackness conditions, we have

$$x_n^{k_{i_j}}(s,a)\Big[\alpha^{n-1}c_n(s,a)-\Big(y_n^{k_{i_j}}(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}^{k_{i_j}}(s')\Big)\Big]=0,\ \text{for all } s\in\mathcal{S},\ a\in\mathcal{A},\ n\in\mathbb{N}.$$

Taking limits as $j\to\infty$, this implies that

$$\bar{x}_n(s,a)\Big[\alpha^{n-1}c_n(s,a)-\Big(\bar{y}_n(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)\bar{y}_{n+1}(s')\Big)\Big]=0,\ \text{for all } s\in\mathcal{S},\ a\in\mathcal{A},\ n\in\mathbb{N}.$$

That is, $\bar{x}$ and $\bar{y}$ satisfy complementary slackness conditions. Thus we have shown that $\bar{x}$ is optimal to (D) and $\bar{y}$ is optimal to (P). By continuity of the objective function in (D), this implies that

$$\lim_{i\to\infty}f^{k_i}=f^*. \tag{30}$$

But since f^k is a monotone decreasing sequence that is bounded below, it converges, and in fact, must converge to f^* because f^* is the limit of one of its subsequences as stated in (30).

Now suppose that the second claim is not true. Then there exists some $\epsilon>0$ and a subsequence k_i such that $\rho_2(x^{k_i},x^*)>\epsilon$ for all optimal basic feasible solutions x^* to (D) and all i. But k_i must have a further subsequence k_{i_j} that converges to some $\bar{x}$ because x^{k_i} belongs to a compact set in the metrizable product topology by Lemma 2.1 and Tychonoff product theorem. Consequently, there exists a J such that $\rho_2(x^{k_{i_j}},\bar{x})<\epsilon$ for all $j\geq J$. But as shown above, $\bar{x}$ must be optimal to (D), and by Lemma 5.8, $\bar{x}$ must be a basic feasible solution of (D). Contradiction.

This completes the proof of Theorem 5.3.

Corollary 5.9. *If (D) has a unique optimal solution x^* (this solution must be a basic feasible solution by Proposition 4.6), then $\lim_{k\to\infty}x^k=x^*$.*

Proof. Fix any $\epsilon > 0$. Since (D) has a unique optimal solution, the second claim in Theorem 5.3 implies that there exists an iteration k_ϵ such that $\rho_2(x^k, x^*) < \epsilon$ for all $k \geq k_\epsilon$. That is, $\lim_{k\to\infty} x^k = x^*$. □

Our final result establishes eventual optimality of the decisions prescribed by our Simplex algorithm in all states in early periods.

Theorem 5.10. *For any period n, there exists an iteration counter K_n such that for all $k \geq K_n$, actions $a_m^k(s)$ are optimal for the nonstationary MDP in all states $s \in \mathcal{S}$ and all periods $m \leq n$.*

Proof. The conclusion trivially holds if x^k is optimal for any k. When this is not the case, we claim that given any $\epsilon > 0$ and any period n, there exists a K_n such that for all $k \geq K_n$, $|x_m^k(s,a) - x_m^{*k}(s,a)| < \epsilon$ for all $m \leq n$, all $s \in \mathcal{S}$ and all $a \in \mathcal{A}$, where x^{*k} are optimal basic feasible solutions to (D). Suppose not. Then there exist an $m \leq n$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and a subsequence k_i such that $|x_m^{k_i}(s,a) - x_m^*(s,a)| > \epsilon$ for all k for all optimal basic feasible solutions x^* to (D). But k_i has a further subsequence k_{i_j} that converges to an optimal basic feasible solution $\bar{x}$ as in the proof of Theorem 5.3. This yields a contradiction. Now fix $0 < \epsilon < 1$ and a period n and consider any iteration $k \geq K_n$. Then for any $m \leq n$ and any $s \in \mathcal{S}$, $|x_m^k(s, a_m(s)) - x_m^{*k}(s, a_m(s))| < \epsilon$ for some optimal basic feasible solution x^{*k} of (D), where $a_m(s)$ is the basic action in state s in period m in x^{*k}. But we know from Definition 4.2 that $x_m^{*k}(s, a_m(s)) \geq 1$. Thus $x_m^k(s, a_m(s)) > 1 - \epsilon > 0$. As a result, the basic action in state s in period m in x^k is also $a_m(s)$. The conclusion of the theorem then follows. □

This theorem only establishes the existence of iterations K_n with the stated property — we cannot tell whether we have reached K_n since it is not possible in general to finitely establish optimality of early decisions in nonstationary MDPs [13].

6 Numerical example

In this section we apply our Simplex algorithm to a nonstationary MDP example and compare it with an efficient implementation of the Shadow Simplex method [22]. This example has two states and two actions. That is, $\mathcal{S} = \{1, 2\}$ and $\mathcal{A} = \{1, 2\}$. Thus the data in each period n is characterized by four costs: $c_n(1,1)$, $c_n(1,2)$, $c_n(2,1)$, and $c_n(2,2)$, and four transition probabilities: $p_n(1|1,1)$, $p_n(1|1,2)$, $p_n(1|2,1)$, $p_n(1|2,2)$ (note that transition probabilities $p_n(2|s,a)$ equal $1 - p_n(1|s,a)$ for $s \in \{1,2\}$ and $a \in \{1,2\}$).

The Shadow Simplex method for an infinite-horizon time-staged CILP solves to optimality N-horizon truncations of the CILP, for $N = 1, 2, 3, \ldots$. These truncations are themselves finite-dimensional LPs, and are solved using the finite-dimensional Simplex method. An optimal basis for the N-horizon LP is used to construct an initial basic feasible solution for the $N+1$-horizon LP. As $N \to \infty$, the Shadow Simplex method converges in value to the optimal value of the infinite-horizon CILP (see [22]). When this method is applied to the CILP formulation (D) of a nonstationary MDP, there is a one-to-one correspondence between basic feasible solutions of the N-horizon LPs and deterministic policies for the N-horizon stochastic dynamic programs obtained by truncating the nonstationary MDP. As a result, pivots in the finite-dimensional Simplex method for the N-horizon LP are equivalent to policy improvement steps in the N-horizon stochastic dynamic program. This leads to an efficient implementation of the Shadow Simplex method that solves a sequence of N-horizon stochastic dynamic programs to optimality using backward induction, for $N = 1, 2, \ldots$. In particular, we start with some initial infinite-horizon policy, and while solving the N-horizon stochastic dynamic program, if backward induction finds a state s and a period n

in which the action prescribed by the current policy is not optimal for the N-horizon stochastic dynamic program, then the current policy is modified by exchanging the inferior action for the optimal one. Our goal here is to highlight the difference between pivots, that is, action swaps, in this implementation of the Shadow Simplex method and pivots in our Simplex method. We achieve this by starting both methods with the same initial policy and tracking, over pivots, the costs of the sequences of policies they produce.

We used discount factor $\alpha = 0.95$. Problem instances were created by drawing cost and transition probability values from a uniform(0,1) random number generator. This yields $0 \leq c_n(s,a) \leq 1$ for all n, s, a, and hence the cost bound $c = 1$. Both methods were initialized with the same randomly generated policy. We illustrate results for one representative problem instance in Figure 2 since all problem instances produced a similar qualitative pattern. To plot this figure, the cost[5] of each infinite-horizon policy was approximated with the cost incurred by that policy in the first five thousand periods. Two hundred iterations of our Simplex method were run. Since one iteration of our Simplex method performs one pivot, two hundred pivots of both methods are plotted. The figure illustrates how Simplex pivots uniformly dominated planning horizon pivots in terms of cost. Simplex pivots also generated a sequence of infinite-horizon policies with monotonically decreasing costs unlike the planning horizon approach in this instance. This cost us some computational overhead — running the Shadow Simplex method required about 0.6 seconds on average over 100 instances compared with about 6 seconds for the Simplex method. Future research would be to investigate how to accelerate our Simplex method by a less costly computation of which actions to swap in and still retain cost improving pivots and convergence to optimal.

7 Conclusions and future work

The contribution of this paper is two-fold. First, it provides a strategy horizon alternative to the planning horizon approach for solving nonstationary MDPs. Second, it identifies a class of CILPs to which an implementable Simplex algorithm can be successfully applied.

In this paper, we have laid an LP foundation for nonstationary discounted MDPs. It would be interesting to investigate whether it naturally leads to an LP-based approximate dynamic programming approach as in stationary MDPs [18] when S and A are themselves very large. One approach would be to substitute a sequence of lower-dimensional value function approximations in place of $y_n(\cdot)$ for all $n \in \mathbb{N}$ in (P) and then design an efficient variant of our algorithm that is rooted in constraint sampling/column generation.

Analysis of average reward MDPs in the stationary case requires assumptions relating to recurrence and communication structures of the Markov chains induced by stationary policies (see Section 8.3 in [38]). Analogous recurrence and communication properties for nonhomogenous Markov chains would likely present significant challenges to formulate and establish. Moreover in the LP formulation of unichain MDPs (wherein the Markov chain corresponding to every stationary policy consists of a single recurrent class plus a potentially empty set of transient states), there need not be a one-to-one correspondence between deterministic policies and basic feasible solutions (Example 8.8.2 in [38]). We believe however that LP formulations of special classes of nonstationary average reward MDPs would be an interesting and fruitful direction for future research.

Three other extensions of our Simplex algorithm could potentially be fruitfully investigated in the future. The first one would be for stationary discounted MDPs with a countably infinite state-space. The challenge there is that the corresponding hypernetwork is not staged and in

[5]Here we mean the total cost-to-go of all states in all periods discounted back to the first decision epoch; that is, $\sum_{n \in \mathbb{N}} \sum_{s \in \mathcal{S}} y_n(s)$, or in other words, f^k in the case of our Simplex algorithm by Equation (16).

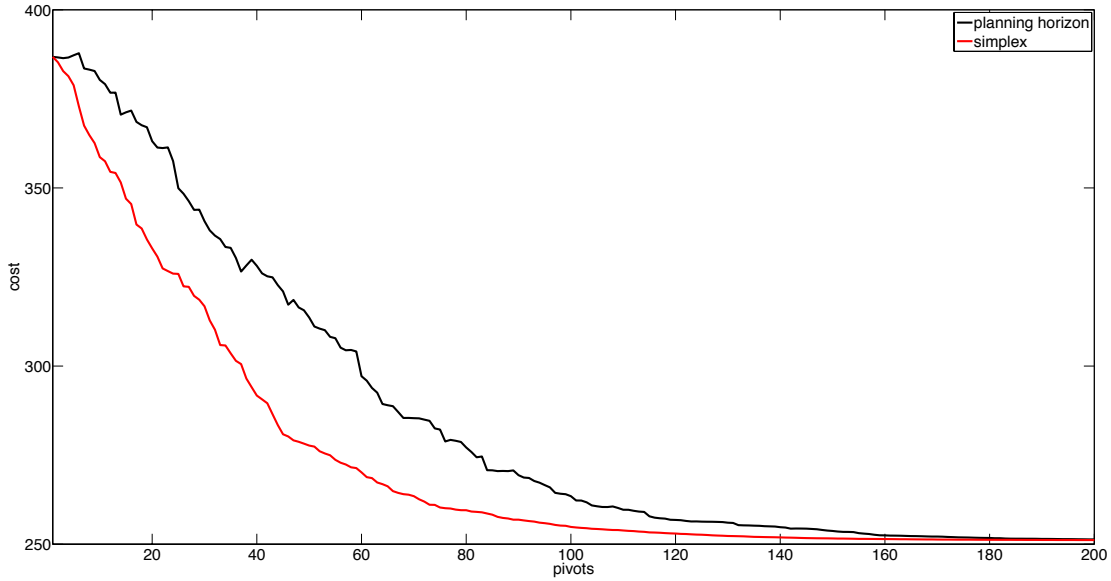

Figure 2: An illustration of monotone improvement in cost achieved by Simplex pivots for one test instance of our example. The graph also shows that Simplex pivots uniformly dominated pivots in the planning horizon approach in [22] in terms of costs attained for this instance. Graphs for all test instances were qualitatively similar.

particular includes cycles. Thus a basic feasible solution characterization of extreme points might be difficult. The second generalization would be to minimum cost flow problems in infinite-dimensional hypernetworks. Finally, perhaps the most difficult extension involves a subset of CILPs in [22] for which there exist sequences of adjacent, improving extreme points with strictly positive support that converge in value to optimal.

More generally, several questions about countably infinite mathematical programs in $\Re^\infty$ remain unanswered. For instance, the feasible region of (D) is not polyhedral in the traditional sense in that it cannot be represented using a finite number of inequalities [5]. Note that this standard definition is too restrictive. According to this definition, even the infinite-dimensional cube $[0,1]^\infty$ is not a polyhedron. However, the feasible region of (D) and the cube $[0,1]^\infty$ do share a key structural property typical of finite-dimensional polyhedra — it is possible to move along an edge of the feasible region from one extreme point to an adjacent one. Can this provide an alternative definition for polyhedra in $\Re^\infty$? We hope that this paper will attract others to study such questions.

References

[1] D Adelman. Price-directed replenishment of subsets: methodology and its application to inventory routing. *Manufacturing and Service Operations Management*, 5(4):348–371, 2003.

[2] D Adelman. A price-directed approach to stochastic inventory routing. *Operations Research*, 52(4):499–514, 2004.

[3] D Adelman. Dynamic bid-prices in revenue management. *Operations Research*, 55(4):647–661, 2007.

[4] D Adelman and A J Mersereau. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research*, 56(3):712–727, 2008.

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

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

[7] J C Bean, J Lohmann, and R L Smith. A dynamic infinite horizon replacement economy decision model. *Engineering Economist*, 30:99–120, 1985.

[8] J C Bean and R L Smith. Conditions for the existence of planning horizons. *Mathematics of Operations Research*, 9:391–401, 1984.

[9] J C Bean and R L Smith. Conditions for the discovery of solution horizons. *Mathematical Programming*, 59:215–229, 1993.

[10] D Bertsimas and J N Tsitsiklis. *Introduction to linear optimization*. Athena Scientific, Belmon, MA, USA, 1997.

[11] C Bes and S Sethi. Concepts of forecast and decision horizons: Application to dynamic stochastic optimization problems. *Mathematics of Operations Research*, 13:295–310, 1988.

[12] S Chand, V Hsu, and S Sethi. Forecast, solution and rolling horizons in operations management problems: A classified bibliography. *Manufacturing and Service Operations Management*, 4:25–43, 2002.

[13] T Cheevaprawatdomrong, I E Schochetman, R L Smith, and A Garcia. Solution and forecast horizons for infinite-horizon non-homogeneous Markov decision processes. *Mathematics of Operations Research*, 32(1):51–72, 2007.

[14] T Cheevaprawatdomrong and R L Smith. Infinite horizon production scheduling in time-varying systems under stochastic demand. *Operations Research*, 52(1), 2004.

[15] G B Dantzig. *Linear programming and extensions*. Princeton University Press, Princeton, New Jersey, USA, 1963.

[16] G de Ghellinck. Les probléms de décisions séquentielles. *Cahiers du Centre d'Etudes de Recherche Opérationnelle*, 2:161–179, 1960.

[17] F D'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10:98–108, 1963.

[18] D P De Farias and B Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[19] A Federgruen and M Tzur. Fast solution and detection of minimal forecast horizons in dynamic programs with a single indicator of future: Application to dynamic lot-sizing models. *Management Science*, 41:874–893, 1995.

[20] A Garcia and R L Smith. Solving nonstationary infinite horizon dynamic optimization problems. *Journal of Mathematical Analysis and Applications*, 244:304–317, 2000.

[21] A Ghate. Infinite horizon problems. In J Cochran, editor, *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, 2011.

[22] A Ghate, D Sharma, and R L Smith. A shadow simplex method for infinite linear programs. *Operations Research*, 58(4):865–877, 2010.

[23] A Ghate and R L Smith. Characterizing extreme points as basic feasible solutions in infinite linear programs. *Operations Research Letters*, 37(1):7–10, 2009.

[24] A Ghate and R L Smith. Optimal backlogging over an infinite horizon under time varying convex production and inventory costs. *Manufacturing and Service Operations Management*, 11(2):362–368, 2009.

[25] R C Grinold. Finite horizon approximations of infinite horizon linear programs. *Mathematical Programming*, 12:1–17, 1977.

[26] O Hernandez-Lerma. *Adaptive Markov control processes*. Springer, New York, NY, USA, 1989.

[27] O Hernandez-Lerma and J B Lasserre. A forecast horizon and a stopping rule for general Markov decision processes. *Journal of Mathematical Analysis and Applications*, 132:388–400, 1988.

[28] O Hernandez-Lerma and J B Lasserre. The linear programming approach. In E Feinberg and A Shwartz, editors, *Handbook of Markov decision processes: methods and algorithms*, chapter 12, pages 377–408. Kluwer, Boston, MA, USA, 2002.

[29] W Hopp. Identifying forecast horizons in nonhomogeneous Markov decision processes. *Operations Research*, 37:339–343, 1989.

[30] W J Hopp, J C Bean, and R L Smith. A new optimality criterion for non-homogeneous markov decision processes. *Operations Research*, 35:875–883, 1987.

[31] R A Howard. *Dynamic programming and Markov processes*. PhD thesis, MIT, Cambridge, MA, USA, 1960.

[32] S. Kunnumkal and H. Topaloglu. A duality-based relaxation and decomposition approach for inventory distribution systems. *Naval Research Logistics*, 55(7):612–631, 2008.

[33] A S Manne. Linear programming and sequential decisions. *Management Science*, 6:259–267, 1960.

[34] J R Munkres. *Topology*. Prentice-Hall, 2000.

[35] J Patrick, M L Puterman, and M Queyranne. Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

[36] W B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality of dimensionality*. John Wiley and Sons, Hoboken, New Jersey, USA, 2007.

[37] M Puterman. *Markov Decision Processes*. John Wiley and Sons, New Jersey, 1994.

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

[39] H E Romeijn, D Sharma, and R L Smith. Extreme point solutions for infinite network flow problems. *Networks*, 48(4):209–222, 2006.

[40] H E Romeijn and R L Smith. Shadow prices in infinite dimensional linear programming. *Mathematics of Operations Research*, 23(1):239–256, 1998.

[41] H E Romeijn, R L Smith, and J C Bean. Duality in infinite dimensional linear programming. *Mathematical Programming*, 53:79–97, 1992.

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

[43] I E Schochetman and R L Smith. Finite dimensional approximation in infinite dimensional mathematical programming. *Mathematical Programming*, 54(3):307–333, 1992.

[44] T C Sharkey and H E Romeijn. A simplex algorithm for minimum cost network flow problems in infinite networks. *Networks*, 52(1):14–31, 2008.

[45] R L Smith and R Zhang. Infinite horizon production planning in time varying systems with convex production and inventory costs. *Management Science*, 44(9):1313–1320, 1998.

[46] H Topaloglu. Using lagrangian relaxation to compute capacity-dependent bid prices in network revenue management. *Operations Research*, 57(3):637–649, 2009.

[47] H Topaloglu and W B Powell. A distributed decision-making structure for dynamic resource allocation using nonlinear functional approximations. *Operations Research*, 53(2):281–297, 2005.

[48] H Topaloglu and W B Powell. Sensitivity analysis of a dynamic fleet management model using approximate dynamic programming. *Operations Research*, 55:319–331, 2007.

[49] P Tseng. Solving h-horizon, stationary Markov decision problems in time proportional to log(h). *Operations Research Letters*, 9(5):287–297, 1990.

[50] Y Ye. A new complexity result on solving the Markov decision problem. *Mathematics of Operations Research*, 30(3):733–749, 2005.

[51] Y Ye. The simplex and policy iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. Technical report, Stanford University, 2011.

[52] D Zhang and D Adelman. An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science*, 43(3):381–394, 2009.

A Proof of Lemma 2.1

The proof uses induction on n and relies on the equality constraints in (D). For $n = 1$, equality constraint (8) implies

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_1(s,a) = S$$

as required. Now suppose, as the inductive hypothesis, that $\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_{n-1}(s,a) = (n-1)S$ for some $n \in \mathbb{N}\setminus\{1\}$. Then using constraint (9) we get

$$\begin{aligned}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_n(s,a) &= \sum_{s\in\mathcal{S}} 1 + \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_{n-1}(s|s',a)x_{n-1}(s',a)\\ &= S + \sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_{n-1}(s',a)\sum_{s\in\mathcal{S}} p_{n-1}(s|s',a) = S + (n-1)S = nS.\end{aligned}$$

This restores the inductive hypothesis.

B Proof of Theorem 3.1

Inequality constraints (5) in (P) imply that, for any integer $N \geq 1$,

$$\begin{aligned}&\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) \geq \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \Big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\Big)x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} y_n(s)x_n(s,a) - \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_n(s'|s,a)y_{n+1}(s')x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_n(s'|s,a)x_n(s,a)\\
&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\Big(\sum_{a\in\mathcal{A}} x_{n+1}(s',a) - 1\Big),\end{aligned}$$

where the last equality follows from equality constraint (9) in (D). The above right hand side in turn simplifies as

$$\begin{aligned}&= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\sum_{a\in\mathcal{A}} x_{n+1}(s',a) + \sum_{n=1}^{N}\sum_{s'\in\mathcal{S}} y_{n+1}(s')\\
&= \sum_{s\in\mathcal{S}} y_1(s)\sum_{a\in\mathcal{A}} x_1(s,a) + \sum_{n=2}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=2}^{N+1}\sum_{s'\in\mathcal{S}} y_n(s')\sum_{a\in\mathcal{A}} x_n(s',a) + \sum_{n=2}^{N+1}\sum_{s'\in\mathcal{S}} y_n(s')\\
&= \sum_{s\in S} y_1(s) + \sum_{n=2}^{N}\sum_{s\in\mathcal{S}} y_n(s)\sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{n=2}^{N+1}\sum_{s'\in\mathcal{S}} y_n(s')\sum_{a\in\mathcal{A}} x_n(s',a) + \sum_{n=2}^{N+1}\sum_{s'\in\mathcal{S}} y_n(s'),\end{aligned}$$

where the last equality follows from equality constraint (8) in (D). The above right hand side equals

$$\sum_{n=1}^{N+1}\sum_{s\in\mathcal{S}} y_n(s) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{a\in\mathcal{A}} x_{N+1}(s,a) = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s) + \sum_{s\in\mathcal{S}} y_{N+1}(s)\Big(1-\sum_{a\in\mathcal{A}} x_{N+1}(s,a)\Big)$$
$$= \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s) - \sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_N(s|s',a)x_N(s',a),$$

where the last equality follows from constraint (9) in (D). Thus, we have shown that

$$\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) \geq \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s) - \underbrace{\sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_N(s|s',a)x_N(s',a)}_{\text{error}(N)}. \quad (31)$$

Now we wish to take limits as $N\to\infty$ on both sides to prove (14). Toward that end, we first show that the limit of the error term as $N\to\infty$ is zero[6]. Since $x_{N+1}(s,a)\geq 0$ for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$, and $|y_{N+1}(s)|\leq \alpha^N\tau_y$ for all $s\in\mathcal{S}$ because $y\in Y$, the error term is bounded below and above as

$$-\alpha^N\tau_y\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_N(s',a) \leq \text{error}(N) \leq \alpha^N\tau_y\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} x_N(s',a).$$

Now by Lemma 2.1, the above bounds simplify as

$$-\alpha^N\tau_y NS^2 \leq \text{error}(N) \leq \alpha^N\tau_y NS^2.$$

Since $\lim_{N\to\infty} N\alpha^N = 0$, the above bounds imply that the limit of the error term is zero. Then taking limits as $N\to\infty$ on both sides of (31) yields (14).

C Proof of Theorem 3.4

From the complementary slackness condition (15), we have

$$\alpha^{n-1}c_n(s,a)x_n(s,a) = \Big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\Big)x_n(s,a),\ \forall s\in\mathcal{S},\ a\in\mathcal{A},\ n\in\mathbb{N}.$$

By adding the above equations over all $s\in\mathcal{S}$, $a\in\mathcal{A}$, and $n=1,2,\ldots,N$ for any integer $N\geq 1$, we get

$$\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \Big(y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a)y_{n+1}(s')\Big)x_n(s,a).$$

Then since x is feasible to (D), using algebraic simplifications identical to the proof of Theorem 3.1, we obtain

$$\sum_{n=1}^{N}\sum_{s\in S}\sum_{a\in\mathcal{A}} \alpha^{n-1}c_n(s,a)x_n(s,a) = \sum_{n=1}^{N}\sum_{s\in\mathcal{S}} y_n(s) - \underbrace{\sum_{s\in\mathcal{S}} y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_N(s|s',a)x_N(s',a)}_{\text{error}(N)}.$$

[6]This property is called transversality in [40].

Then taking limits as $N \to \infty$, and noting, from the proof of Theorem 3.1, that $\lim_{N\to\infty} \text{error}(N) = 0$ since x is feasible to (D), we get

$$\sum_{n\in\mathbb{N}} \sum_{s\in S} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) = \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} y_n(s).$$

If y is also feasible to (P), then Corollary 3.2 implies that y is optimal to (P) and x is optimal to (D), respectively.

D Proof of Theorem 3.5

We consider the following N-horizon truncation of (P)

$$\begin{aligned}
(P(N))\ \max & \sum_{n\in\mathbb{N}} \sum_{s\in\mathcal{S}} y_n(s) \\
& y_n(s) - \sum_{s'\in\mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \le \alpha^{n-1} c_n(s,a),\ \text{for } s\in\mathcal{S},\ a\in\mathcal{A},\ n=1,\dots,N-1, \quad (32) \\
& y_N(s) \le \alpha^{N-1} c_N(s,a),\ \text{for } s\in\mathcal{S},\ a\in\mathcal{A}, \quad (33)
\end{aligned}$$

and its dual

$$\begin{aligned}
(D(N))\ \min & \sum_{n\in\mathbb{N}} \sum_{s\in S} \sum_{a\in\mathcal{A}} \alpha^{n-1} c_n(s,a) x_n(s,a) \\
& \sum_{a\in\mathcal{A}} x_1(s,a) = 1,\ \text{for } s\in\mathcal{S}, \quad (34) \\
& \sum_{a\in\mathcal{A}} x_n(s,a) - \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_{n-1}(s|s',a) x_{n-1}(s',a) = 1,\ \text{for } s\in\mathcal{S},\ n=2,\dots,N, \quad (35) \\
& x_n(s,a) \ge 0,\ \text{for } s\in\mathcal{S},\ a\in\mathcal{A},\ n=1,2,\dots,N. \quad (36)
\end{aligned}$$

Both these are finite-dimensional linear programs. By arguments identical to Lemma 2.1, $(D(N))$ has a bounded feasible region and hence does have an optimal solution. By strong duality, $(P(N))$ also has an optimal solution. Let y^N and x^N denote optimal solutions to $(P(N))$ and $(D(N))$, respectively. It is easy to show that y^N satisfies $0 \le y_n^N(s) \le \alpha^{n-1}\frac{c}{1-\alpha}$ for all $s\in\mathcal{S}$ and for $n=1,2,\dots,N$. Similarly, x^N satisfies $x_n^N(s,a) \le nS$ for all $s\in\mathcal{S}$ and for $n=1,2,\dots,N$ as in Lemma 2.1. By appending y^N and x^N with infinite strings of zeros, we view y^N as an element of the set

$$\mathcal{C} = \Big\{ y : 0 \le y_n(s) \le \alpha^{n-1}\frac{c}{1-\alpha},\ \forall s\in\mathcal{S},\ n\in\mathbb{N} \Big\}, \quad (37)$$

and x^N as an element of the set

$$\mathcal{K} = \Big\{ x : 0 \le x_n(s,a) \le nS,\ \forall s\in\mathcal{S},\ a\in\mathcal{A},\ n\in\mathbb{N} \Big\}. \quad (38)$$

Now consider the sequence of pairs $(y^N, x^N) \in \mathcal{C}\times\mathcal{K}$. Both $\mathcal{C}$ and $\mathcal{K}$ are compact in the metrizable product topology by Tychonoff's product theorem, and so is $\mathcal{C}\times\mathcal{K}$, again by Tychonoff. Therefore, (y^N, x^N) has a convergent subsequence, say (y^{N_k}, x^{N_k}), with limit $(\bar{y},\bar{x}) \in \mathcal{C}\times\mathcal{K}$ as $k\to\infty$. Then, it is easy to show, by taking limits in the constraints of $(P(N))$ and $(D(N))$, that $\bar{y}$ is feasible to (P) and $\bar{x}$ is feasible to (D). Similarly, it is easy to show, by taking a limit of the finite-dimensional complementary slackness condition, that $\bar{y}$ and $\bar{x}$ satisfy the complementary slackness condition (15). Thus, Theorem 3.4 implies that $\bar{y}$ and $\bar{x}$ are optimal to (P) and (D), respectively, and their objective function values are equal.

E Proof of Theorem 3.6

Since y and x are optimal to (P) and (D), respectively, their objective function values are equal by Theorem 3.5. That is,

$$\lim_{N\to\infty}\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}c_n(s,a)x_n(s,a)=\lim_{N\to\infty}\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}y_n(s).$$

Recall from the proof of Theorem 3.1 that the limit of $\sum_{s\in\mathcal{S}}y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}p_N(s|s',a)x_N(s',a)$ as $N\to\infty$ is zero. Therefore, subtracting this limit from the right hand side of the above equation does not alter the equation. Thus we have

$$\begin{aligned}
&\lim_{N\to\infty}\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}c_n(s,a)x_n(s,a)\\
&=\lim_{N\to\infty}\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}y_n(s)-\lim_{N\to\infty}\sum_{s\in\mathcal{S}}y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}p_N(s|s',a)x_N(s',a)\\
&=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}y_n(s)-\sum_{s\in\mathcal{S}}y_{N+1}(s)\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}p_N(s|s',a)x_N(s',a)\right).
\end{aligned}$$

Then using the algebraic simplification in the proof of Theorem 3.1 we obtain

$$\begin{aligned}
&\lim_{N\to\infty}\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\alpha^{n-1}c_n(s,a)x_n(s,a)\\
&=\lim_{N\to\infty}\left(\sum_{n=1}^{N}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\Big(y_n(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}(s')\Big)x_n(s,a)\right).
\end{aligned}$$

That is,

$$\lim_{N\to\infty}\sum_{n=1}^{N}\underbrace{\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}x_n(s,a)\Big[\alpha^{n-1}c_n(s,a)-\Big(y_n(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}(s')\Big)\Big]}_{\psi(n)}=0.$$

Since y and x are feasible to (P) and (D), respectively, we have (i) $x_n(s,a)\geq 0$ for all $s\in\mathcal{S}$ and all $a\in\mathcal{A}$, and (ii) $\alpha^{n-1}c_n(s,a)\geq y_n(s)-\sum_{s'\in\mathcal{S}}p_n(s'|s,a)y_{n+1}(s')$ for all $s\in\mathcal{S}$ and all $a\in\mathcal{A}$. Consequently, $\psi(n)\geq 0$ for all n. In fact, since $\sum_{n\in\mathbb{N}}\psi(n)=0$, we have that $\psi(n)=0$ for all n. This implies (15) in light of (i) and (ii) above.

F Proof of Theorem 4.3

Suppose x is a basic feasible solution but not an extreme point. Then there exists a $\lambda\in(0,1)$ and $w,z\in X$ that are distinct from x and are feasible to (D) such that $x=\lambda w+(1-\lambda)z$. Since $w\geq 0$ and $z\geq 0$ by constraint (10) in (D), $x_n(s,a)=0$ for any $s\in\mathcal{S}$, $a\in\mathcal{A}$ and $n\in\mathbb{N}$ implies $w_n(s,a)=z_n(s,a)=0$. That is, the sets of basic actions in x, w, and z are identical. Uniqueness

of flows in basic actions then implies that $x = w = z$. This contradicts the hypothesis that w and z are distinct from x.

Suppose x is an extreme point but not a basic feasible solution. Then there exists some $n \in \mathbb{N}$ and some $s \in \mathcal{S}$ and two distinct actions $a, b \in \mathcal{A}$ such that $x_n(s,a) = \delta > 0$ and $x_n(s,b) = \epsilon > 0$. In fact, let n be the smallest such period. Without loss of generality, we assume that $\delta > \epsilon$. We show that x is a midpoint of two distinct solutions w and z that are feasible to (D), thus contradicting that x is an extreme point. For $k = n+1, n+2, \ldots$ let $\mathcal{S}_k(x) \subseteq \mathcal{S}$ be the subset of states that receive any portion of flow δ originating in hyperarc (n,s,a) in solution x. Moreover, for any $s_k \in \mathcal{S}_k(x)$, let $\mathcal{A}_k(s_k)$ be the subset of $a_k \in \mathcal{A}$ such that $x_k(s_k, a_k) > 0$. Let $\mathcal{F}_n(x)$ be the sub-hypernetwork formed by node s, hyperarc (n,s,a), nodes in $\bigcup_{k=n+1}^{\infty} \mathcal{S}_k(x)$ and hyperarcs in $\bigcup_{k=n+1}^{\infty} \bigcup_{s_k \in \mathcal{S}_k} \mathcal{A}_k(s_k)$. For any $s_k \in \mathcal{S}_k(x)$ and $a_k \in \mathcal{A}_k(s_k)$, let $\phi_k(s_k, a_k) = x_k(s_k, a_k) / \sum_{a \in \mathcal{A}_k(s_k)} x_k(s_k, a)$. Put a supply of ϵ on node s and a supply of 0 on all other nodes in sub-hypernetwork $\mathcal{F}_n(x)$, and then construct a flow u recursively in periods $n, n+1, \ldots$ in this sub-hypernetwork as follows. First set $u_n(s,a) = \epsilon$. Then for each $s_{n+1} \in \mathcal{S}_{n+1}$ and each $a_{n+1} \in \mathcal{A}_{n+1}(s_{n+1})$, set $u_{n+1}(s_{n+1}, a_{n+1}) = \epsilon p_n(s_{n+1}|s,a)\phi_{n+1}(s_{n+1}, a_{n+1})$. More generally, for each $s_k \in \mathcal{S}_k$ and $a_k \in \mathcal{A}_k(s_k)$ for $k = n+2, n+3, \ldots$, set

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

We claim that $x_k(s_k, a_k) \geq u_k(s_k, a_k)$ for all hyperarcs (s_k, a_k) in sub-hypernetwork $\mathcal{F}_n(x)$. To see this, we note that

$$\begin{aligned}
u_k(s_k, a_k) &= \phi_k(s_k, a_k) \sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s_{k-1}, a_{k-1}) u_{k-1}(s_{k-1}, a_{k-1}) \\
&= \frac{x_k(s_k, a_k)}{\sum_{a \in \mathcal{A}_k(s_k)} x_k(s_k, a)} \sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s_{k-1}, a_{k-1}) u_{k-1}(s_{k-1}, a_{k-1}) \\
&= \frac{x_k(s_k, a_k) \sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s_{k-1}, a_{k-1}) u_{k-1}(s_{k-1}, a_{k-1})}{1 + \sum_{s_{k-1} \in \mathcal{S}} \sum_{a_{k-1} \in \mathcal{A}} p_{k-1}(s_k|s', a) x_{k-1}(s', a)}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{x_k(s_k, a_k) \sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s_{k-1}, a_{k-1}) u_{k-1}(s_{k-1}, a_{k-1})}{\sum_{s_{k-1} \in \mathcal{S}} \sum_{a_{k-1} \in \mathcal{A}} p_{k-1}(s_k|s_{k-1}, a_{k-1}) x_{k-1}(s_{k-1}, a_{k-1})} \\
&\leq \frac{x_k(s_k, a_k) \sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s_{k-1}, a_{k-1}) u_{k-1}(s_{k-1}, a_{k-1})}{\sum_{s_{k-1} \in \mathcal{S}_{k-1}} \sum_{a_{k-1} \in \mathcal{A}_{k-1}(s_{k-1})} p_{k-1}(s_k|s', a) x_{k-1}(s_{k-1}, a_{k-1})} \leq x_k(s_k, a_k).
\end{aligned}$$

Similarly, let $\mathcal{T}_k(x) \subseteq \mathcal{S}$ be the subset of states that receive any portion of flow ϵ originating in hyperarc (n,s,b) in solution x. Moreover, for any $t_k \in \mathcal{T}_k(x)$, let $\mathcal{B}_k(t_k)$ be the subset of $b_k \in \mathcal{A}$ such that $x_k(t_k, b_k) > 0$. Let $\mathcal{G}_n(x)$ be the sub-hypernetwork formed by node s, hyperarc (n,s,b), nodes in $\bigcup_{k=n+1}^{\infty} \mathcal{T}_k(x)$ and hyperarcs in $\bigcup_{k=n+1}^{\infty} \bigcup_{t_k \in \mathcal{T}_k} \mathcal{B}_k(t_k)$. For any $t_k \in \mathcal{T}_k(x)$ and $b_k \in \mathcal{B}_k(s_k)$, let $\lambda_k(t_k, b_k) = x_k(t_k, b_k) / \sum_{b \in \mathcal{B}_k(t_k)} x_k(t_k, b)$. Put a supply of ϵ on node s and a supply of 0 on all other

nodes in sub-hypernetwork $\mathcal{G}_n(x)$, and then construct a flow v recursively in periods $n, n+1, \ldots$ in this sub-hypernetwork as follows. First set $v_n(s,b) = \epsilon$. Then for each $t_{n+1} \in \mathcal{T}_{n+1}$ and each $b_{n+1} \in \mathcal{B}_{n+1}(t_{n+1})$, set

$$v_{n+1}(t_{n+1}, b_{n+1}) = \epsilon p_n(t_{n+1}|s,b)\lambda_{n+1}(t_{n+1}, b_{n+1}).$$

More generally, for each $t_k \in \mathcal{T}_k$ and $b_k \in \mathcal{B}_k(t_k)$ for $k = n+2, n+3, \ldots$, set

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

Again, we note that $v_k(t_k, b_k) \leq x_k(t_k, b_k)$ for all hyperarcs (t_k, b_k) in sub-hypernetwork $\mathcal{G}_n(x)$. We construct a new solution w to (D) from x as follows. Set $w_k(s_k, a_k) = x_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are not in $\mathcal{F}_n(x)$ or $\mathcal{G}_n(x)$. Then set $w_k(s_k, a_k) = x_k(s_k, a_k) - u_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are in $\mathcal{F}_n(x)$, and $w_k(s_k, a_k) = x_k(s_k, a_k) + v_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are in $\mathcal{G}_n(x)$. That is, w is constructed from x by rerouting a total flow of ϵ from sub-hypernetwork $\mathcal{F}_n(x)$ through sub-hypernetwork $\mathcal{G}_n(x)$, and hence it is feasible to the flow balance constraints in (D). Similarly, we construct a new feasible solution z to (D) from x as follows. Set $z_k(s_k, a_k) = x_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are not in $\mathcal{F}_n(x)$ or $\mathcal{G}_n(x)$. Then set $z_k(s_k, a_k) = x_k(s_k, a_k) + u_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are in $\mathcal{F}_n(x)$, and $z_k(s_k, a_k) = x_k(s_k, a_k) - v_k(s_k, a_k)$ for hyperarcs (s_k, a_k) that are in $\mathcal{G}_n(x)$. That is, z is constructed by rerouting a total flow of ϵ from sub-hypernetwork $\mathcal{G}_n(x)$ through sub-hypernetwork $\mathcal{F}_n(x)$, and hence it is feasible to the flow balance constraints in (D) Then observe that $x = (z+w)/2$ contradicting the assumption that x is an extreme point.

G Proof of Proposition 4.6

Suppose that $y^* \in Y$ is optimal to (P). Then by Theorem 3.6, x^* and y^* satisfy complementary slackness conditions. This implies that, since $x^*_n(s, a_n(s)) > 0$,

$$y^*_n(s) = \alpha^{n-1} c_n(s, a_n(s)) + \sum_{s' \in \mathcal{S}} p_n(s'|s, a_n(s)) y^*_{n+1}(s'),$$

which in turn equals

$$\min_{a \in \mathcal{A}} \left\{ \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y^*_{n+1}(s') \right\}$$

because $y^*_n(s) \leq \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y^*_{n+1}(s')$ for all $a \in \mathcal{A}$ by constraints (5) in (P). Thus $a_n(s)$ achieves the minimum in Bellman's equations of optimality in dynamic programming and hence is optimal in state s in period n (Theorem 2.2 on page 32 of [42]).

H Proof of Lemma 4.7

Let $a_n(s)$, for $n \in \mathbb{N}$ and $s \in \mathcal{S}$, be the basic actions in x. Since $x_n(s, a_n(s)) > 0$ for all $n \in \mathbb{N}$ and $s \in \mathcal{S}$, Equation (15) in Definition 3.3 implies that any $y \in Y$ that satisfies complementary slackness with x must be a solution of the infinite system of equations

$$y_n(s) = \alpha^{n-1} c_n(s, a_n(s)) + \sum_{s' \in \mathcal{S}} p_n(s'|s, a_n(s)) y_{n+1}(s'), \text{ for all } n \in \mathbb{N}, \ s \in \mathcal{S}.$$

Theorem 6.1.1 in [37] shows[7] that this system has a unique bounded solution wherein values $y_n(s)$, for all $n \in \mathbb{N}$ and $s \in \mathcal{S}$, equal the expected cost-to-go, discounted back to the first period, incurred on implementing the deterministic Markovian policy defined by actions $a_n(s)$ starting in state s in period n.

I Proof of Proposition 5.1

Let y be the solution complementary to x and w be the solution complementary to z. Then we know by Equation (15) in Theorem 3.4 that $f(x) = g(y)$ and $f(z) = g(w)$. Thus, to prove that $f(z) - f(x) = (1+\theta)\gamma_n(s,a)$, we show that $g(w) - g(y) = (1+\theta)\gamma_n(s,a)$.

Since basic hyperarcs in periods $n+1, n+2, \ldots$ do not change in the pivot operation, $w_k(s,a) = y_k(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ for $k = n+1, n+2, \ldots$. Since basic hyperarcs in other states $t \neq s$ in period n do not change in the pivot operation, this also implies that $w_n(t) = y_n(t)$ for all $t \neq s$. Moreover, since (n, s, a) is a basic hyperarc in the new basic feasible solution z, $z_n(s,a) > 0$. Thus

$$w_n(s) = \alpha^{n-1} c_n(s,a) + \sum_{s' \in \mathcal{S}} p_n(s'|s,a) y_{n+1}(s') \tag{39}$$

by complementary slackness. Now let $\mathcal{S}_{n-1} \subseteq \mathcal{S}$ be the set of states t in period $n-1$ such that $s \in \mathcal{J}_{n-1}(t, a_{n-1}(t))$. That is, state s in period n is reachable by choosing basic actions in states in $\mathcal{S}_{n-1}$. Thus, for $t \in \mathcal{S}_{n-1}$, we have by complementary slackness that

$$w_{n-1}(t) = \alpha^{n-2} c_{n-1}(t, a_{n-1}(t)) + p_{n-1}(s|t, a_{n-1}(t)) w_n(s) + \sum_{s' \in \mathcal{S} \setminus \{s\}} p_{n-1}(s'|t, a_{n-1}(t)) y_n(s').$$

Also, by complementary slackness, we have

$$y_{n-1}(t) = \alpha^{n-2} c_{n-1}(t, a_{n-1}(t)) + p_{n-1}(s|t, a_{n-1}(t)) y_n(s) + \sum_{s' \in \mathcal{S} \setminus \{s\}} p_{n-1}(s'|t, a_{n-1}(t)) y_n(s').$$

Hence, for $t \in \mathcal{S}_{n-1}$, we get

$$w_{n-1}(t) - y_{n-1}(t) = p_{n-1}(s|t, a_{n-1}(t))(w_n(s) - y_n(s)) \triangleq \theta_{n-1}(s,t)(w_n(s) - y_n(s)).$$

On the other hand, for $t \notin \mathcal{S}_{n-1}$, $w_{n-1}(t) = y_{n-1}(t)$. Now for $k = 1, 2, \ldots, n-2$, we recursively define $\mathcal{S}_k \subseteq \mathcal{S}$ as the set of states t in period k such that $\mathcal{I}_k(t, a_k(t)) \triangleq \mathcal{J}_k(t, a_k(t)) \cap \mathcal{S}_{k+1} \neq \emptyset$. Thus, for $t \in \mathcal{S}_k$, we have that

$$w_k(t) = \alpha^{k-1} c_k(t, a_k(t)) + \sum_{s' \in \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) w_{k+1}(s') + \sum_{s' \notin \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) y_{k+1}(s').$$

Again, by complementary slackness, we have

$$y_k(t) = \alpha^{k-1} c_k(t, a_k(t)) + \sum_{s' \in \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) y_{k+1}(s') + \sum_{s' \notin \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) y_{k+1}(s').$$

Thus, for $t \in \mathcal{S}_k$, we get

$$w_k(t) - y_k(t) = \sum_{s' \in \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t))(w_{k+1}(s') - y_{k+1}(s'))$$

[7] The result in [37] is for stationary, infinite-horizon, countable state, finite action MDPs but it is valid here, since as noted earlier, our MDP can be viewed that way.

$$
\begin{aligned}
&= \sum_{s' \in \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) \theta_{k+1}(s, s')(w_n(s) - y_n(s)) \\
&= (w_n(s) - y_n(s)) \sum_{s' \in \mathcal{I}_k(t, a_k(t))} p_k(s'|t, a_k(t)) \theta_{k+1}(s, s') \\
&\triangleq \theta_k(s, t)(w_n(s) - y_n(s)).
\end{aligned}
$$

On the other hand, for $t \notin \mathcal{S}_k$, $w_k(t) = y_k(t)$. Then the difference between objective values of w and y is given by

$$
\begin{aligned}
g(w) - g(y) &= \sum_{k \in \mathbb{N}} \sum_{t \in \mathcal{S}} (w_k(t) - y_k(t)) = \sum_{k=1}^{n} \sum_{t \in \mathcal{S}} (w_k(t) - y_k(t)) \\
&= (w_n(s) - y_n(s)) + \sum_{k=1}^{n-1} \sum_{t \in \mathcal{S}_k} (w_k(t) - y_k(t)) \\
&= (w_n(s) - y_n(s)) + \sum_{k=1}^{n-1} \sum_{t \in \mathcal{S}_k} \theta_k(s, t)(w_n(s) - y_n(s)) \\
&\triangleq (1 + \theta)(w_n(s) - y_n(s)),
\end{aligned}
$$

where $\theta \triangleq \sum_{k=1}^{n-1} \sum_{t \in \mathcal{S}_k} \theta_k(s, t)$. Then Equation (39) implies that $g(w) - g(y) = (1 + \theta)\gamma_n(s, a)$.