Identifying the Optimal Value Function of a Negative Markov Decision Process: An Integer Programming Approach

Amin Dehghanian
Department of Industrial Engineering, Amirkabir University of Technology, Tehran, Iran, amin.dehghanian@aut.ac.ir

Mathematical programming formulation to identify the optimal value function of a negative Markov decision process (MDP) is non-convex, non-smooth, and computationally intractable. Also note that other well-known solution methods of MDP do not work properly for a negative MDP. More specifically, the policy iteration diverges, and the value iteration converges but does not provide an error bound in a finite number of iterations. To overcome this challenge, we develop a mixed integer linear programming (MILP) formulation using the theory of disjunctive programming. Then, we discuss its strength and how to extract the optimal value function and an optimal policy from this MILP formulation. Moreover, we exemplify our formulation for a class of stopping problems studied by Ross (1983). At the end, we briefly present another formulation developed by the traditional big-M method.

Key words: Negative Markov Decision Process, Disjunctive Programming, Integer Programming, Optimal Stopping

1. Introduction

An infinite horizon MDP is a decision-making tool to control a system in a stochastic dynamic environment. Primitives of an MDP consist of a single decision maker, periods, states, actions, costs, and transition probabilities. The system evolves as follows. In each period, the system occupies a state $s \in S$, where S denotes the set of all possible states. After observing the state, the decision maker chooses an action a from the set of permissible actions A_s in state s, and incurs a cost of $c(s,a)$. The system then moves into a new state j in the next period according to the transition probability $p(j|s,a)$, and so on. The decision maker seeks to minimize a cost criterion, e.g., total expected cost.

Researchers have adopted several performance criteria in MDPs, including 1. expected total discounted cost, 2. average cost, and 3. expected total cost (Puterman 2005). Based on the sign of $c(s,a)$, MDPs with the third criterion are subdivided into two classes: 3.1. positive MDPs in which $c(s,a) \leq 0$ for all $s \in S$ and $a \in A_s$, 3.2. negative MDPs in which $c(s,a) \geq 0$ for all $s \in S$ and $a \in A_s$ (Puterman 2005, Ross 1983, Strauch 1966). (Definition of positive MDP models is more general in

Puterman (2005), but we do not involve ourselves in those details since those models are not the focus of this paper.)

MDPs are applied to many practical settings such as healthcare, maintenance, inventory control, revenue management, transportation (Puterman 2005, Ross 1983, Talluri and van Ryzin 2004, Powell 2011). From methodological and practical standpoints, it is crucial to be able to solve MDPs in the sense of finding an optimal policy and the optimal value function. There are three general solution approaches for MDPs including value iteration, policy iteration, and linear programming (LP) (Puterman 2005, Ross 1983). Functionality of these approaches depends on the considered performance criterion, and it is summarized in Table 1. All the solution approaches work properly for discounted, average cost, and positive MDP settings, but this is not the case for negative MDPs as described at what follows.

Negative MDPs were originally studied by Strauch (1966), and ever since they have been used to model problems such as optimal stopping, optimal search, and Bayesian sequential analysis (Ross 1983). For negative MDPs, the value iteration is the only functional approach in the sense that it converges to the optimal value function. Although this convergence is methodologically important, but it is of little computational value since it is unable to provide error bounds. More specifically, when we terminate the value iteration algorithm after a finite number of iterations, it is unknown how far our current solution is from the optimal value function. To sum up, there is no general solution approach to compute the optimal value function of a negative MDP, which in turn has limited its applicability.

Table 1 Solution approaches for different MDP settings.

	MDP model type			
Solution Method	Discounted	Average	Positive	Negative
Value iteration*	✓	✓	✗	✗
Policy iteration	✓	✓	✓	✗
LP	✓	✓	✓	✗

*Note that here we investigate the existence of an error bound in finite termination of the value iteration rather than its convergence.

Our contribution in this paper is to present the first systematic approach to solve negative MDPs. More specifically, we develop two MILP formulations, which are able to identify the optimal value function and an optimal policy of a negative MDP. The first formulation is built on the theory of disjunctive programming that provides an ideal formulation for the union of polytopes. We demonstrate the validity and tightness of this formulation, and describe how to derive the optimal value function and an optimal policy from the optimal solution of this MILP formulation. We

adopt our formulation for a class of stopping problems studied by Ross (1983). Finally, we develop our second formulation by the traditional big-M method. This formulation enjoys fewer variables at the cost of a weaker LP relaxation.

The remainder of the paper is organized as follows. In Section 2, we present our disjunctive programming formulation, analyze its strength, and illustrate it for the stopping example. Section 3 presents our big-M formulation. Some concluding remarks are provided in Section 4.

2. Disjunctive Formulation

In this section, we present our first MILP formulation. For this purpose, we introduce our notation at what follows. We adopt the convention that a term in bold refers to a real-valued vector; e.g., $\mathbf{v}$ refers to the vector $\langle v(s)\rangle_{s\in S}$. Moreover, $\mathbf{0}$ denotes a vector of appropriate dimension whose components are all equal to zero. Let $\mathbf{v}^* := \langle v^*(s)\rangle_{s\in S}$ denote the optimal value function of our negative MDP model. It is clear that $v^*(s) \geq 0$ for all $s \in S$. For convenience, we use the term *policy* to refer to *stationary deterministic policy*. Throughout the paper, we assume that both state and action spaces are finite, and this implies that an optimal policy exists (Puterman 2005). The optimality equation is as follows:

$$v(s) = \min_{a\in A_s}\{c(s,a) + \sum_{j\in S} p(j|s,a)v(j)\} \qquad \forall s \in S. \tag{1}$$

THEOREM 1. *(Puterman 2005, Proposition 7.3.2.) The vector* $\mathbf{v}^*$ *is the minimal nonnegative solution of the following set of inequalities:*

$$v(s) \geq \min_{a\in A_s}\{c(s,a) + \sum_{j\in S} p(j|s,a)v(j)\} \qquad \forall s \in S. \tag{2}$$

Let $\overline{\mathbf{v}}$ be an upper-bound vector of $\mathbf{v}^*$. In Subsection 2.2, we describe how to obtain such a bound by evaluating the value function of an arbitrary stationary policy. For each $s \in S$, let Γ_s denote the set of solutions for inequality associated with s in (2), which are (componentwise) bounded from above by $\overline{\mathbf{v}}$. More formally,

$$\Gamma_s := \{\mathbf{v} \in \mathbb{R}^{|S|} : v(s) \geq \min_{a\in A_s}\{c(s,a) + \sum_{j\in S} p(j|s,a)v(j)\}, \mathbf{0} \leq \mathbf{v} \leq \overline{\mathbf{v}}\}. \tag{3}$$

It follows that:

$$\Gamma_s = \bigcup_{a\in A_s} \{\mathbf{v} \in \mathbb{R}^{|S|} : v(s) - \sum_{j\in S} p(j|s,a)v(j) \geq c(s,a), \mathbf{0} \leq \mathbf{v} \leq \overline{\mathbf{v}}\}. \tag{4}$$

Note that Γ_s is the union of a finite number of polytopes, and such a set is called *disjunctive constraint* or *disjunctive set* in mathematical programming literature (Vielma 2015). There is an extensive literature on *disjunctive programming* which addresses how to formulate the convex hull

of the union of polyhedra as an LP (Balas 1998). We adopt this literature to provide an MILP formulation for Γ_s, and we discuss its strength later on in Subsection 2.1.

Let $\Gamma := \cap_{s\in S}\Gamma_s$, which consists of solutions to the set of inequalities (2). To characterize Γ, we present the feasible region of our MILP formulation as follows:

$$v(j) = \sum_{a\in A_s} w(s,j,a) \qquad \forall s,j \in S, \tag{5a}$$

$$w(s,s,a) - \sum_{j\in S} p(j|s,a)w(s,j,a) \geq c(s,a)x(s,a) \qquad \forall s\in S, a\in A_s, \tag{5b}$$

$$0 \leq w(s,j,a) \leq \overline{v}(j)x(s,a) \qquad \forall s,j\in S, a\in A_s, \tag{5c}$$

$$\sum_{a\in A_s} x(s,a) = 1 \qquad \forall s\in S, \tag{5d}$$

$$x(s,a) \in \{0,1\} \qquad \forall s\in S, a\in A_s. \tag{5e}$$

Let $\Delta := \{\langle \mathbf{v},\mathbf{w},\mathbf{x}\rangle \,|(5a)-(5e)\}$, where $\mathbf{w} := \langle w(s,j,a)\rangle_{s,j\in S,a\in A_s}$ and $\mathbf{x} := \langle x(s,a)\rangle_{s\in S,a\in A_s}$. Indeed, the vector $\mathbf{w}$ consists of $\sum_{s\in S}|A_s|$ copies of $\mathbf{v}$ as required in disjunctive programming. More specifically, for each $s\in S$, we create $|A_s|$ copies of the vector $\mathbf{v}$ needed to describe the set Γ_s. So, to describe the set Γ this procedure creates in total $\sum_{s\in S}|A_s|$ copies of $\mathbf{v}$, denoted by the vector $\mathbf{w}$.

THEOREM 2. *The vector* $\mathbf{v}$ *belongs to* Γ *if and only if there exist* $\mathbf{w}$ *and* $\mathbf{x}$ *such that* $\langle \mathbf{v},\mathbf{w},\mathbf{x}\rangle \in \Delta$.

PROOF. $\Rightarrow$: Suppose $\mathbf{v}\in\Gamma$. For each $s\in S$, there exists an action $a_s\in A_s$ such that $v(s) - \sum_{j\in S} p(j|s,a_s)v(j) \geq c(s,a_s)$. If there are multiple actions with such a property, pick one arbitrarily. Let $x(s,a_s) := 1$ and $x(s,a) := 0$ for all $a\in A_s\backslash a_s$. Moreover, let $w(s,j,a_s) := v(j)$ for all $j\in S$, and let $w(s,j,a) := 0$ for all $a\in A_s\backslash\{a_s\}$ and $j\in S$. It is clear that $\langle \mathbf{v},\mathbf{w},\mathbf{x}\rangle\in\Delta$.

$\Leftarrow$: Suppose $\langle \mathbf{v},\mathbf{w},\mathbf{x}\rangle\in\Delta$. For each $s\in S$, (5d) implies that there is an action $a_s\in A_s$ such that $x(s,a_s)=1$ and $x(s,a)=0$ for all $a\in A_s\backslash\{a_s\}$. Therefore, (5c) implies that $w(s,j,a)=0$ for all $a\in A_s\backslash\{a_s\}$ and $j\in S$. Then, (5a) implies that $w(s,j,a_s)=v(j)$ for all $j\in S$. Given (5b) for a_s, it follows that $v(s) - \sum_{j\in S} p(j|s,a_s)v(j) \geq c(s,a_s)$ for all $s\in S$, so $\mathbf{v}\in\Gamma$. $\square$

Choose $\alpha(s), s\in S$ to be strictly positive numbers. The optimal solution of the following MILP provides the optimal value function and an optimal policy of our negative MDP.

$$\min \sum_{s\in S} \alpha(s)v(s) \tag{6a}$$

$$s.t.$$

$$\langle \mathbf{v},\mathbf{w},\mathbf{x}\rangle \in \Delta. \tag{6b}$$

THEOREM 3. *(i) There exist* $\mathbf{w}^*$ *and* $\mathbf{x}^*$ *such that* $\langle \mathbf{v}^*,\mathbf{w}^*,\mathbf{x}^*\rangle$ *is an optimal solution to the MILP* (6a)-(6b).

(ii) If $\langle \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{x}} \rangle$ *is an optimal solution to the MILP* (6a)-(6b)*, then* $\hat{\mathbf{v}} = \mathbf{v}^*$.
(iii) Given an optimal solution $\langle \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{x}} \rangle$ *to the MILP* (6a)-(6b)*, let* $d^{\infty}_{\hat{\mathbf{x}}}$ *be a policy under which in each state* $s \in S$*, we choose action* $a_s \in A_s$ *such that* $\hat{x}(s, a_s) = 1$*. Then,* $d^{\infty}_{\hat{\mathbf{x}}}$ *is an optimal policy.*

PROOF. *(i)* Let us choose $\mathbf{w}^*$ and $\mathbf{x}^*$ by the procedure described in the proof of Theorem 2. Then, $\langle \mathbf{v}^*, \mathbf{w}^*, \mathbf{x}^* \rangle \in \Delta$ by Theorem 2, and it is sufficient to show that this point is optimal. Let $\langle \mathbf{v}, \mathbf{w}, \mathbf{x} \rangle$ denote an arbitrary feasible solution to Δ, so $\mathbf{v} \in \Gamma$ by Theorem 2. Then, Theorem 1 implies that $v^*(s) \leq v(s)$ for all $s \in S$. As $\alpha(s) > 0$ for all $s \in S$, it follows that $\sum_{s \in S} \alpha(s) v^*(s) \leq \sum_{s \in S} \alpha(s) v(s)$, implying that $\langle \mathbf{v}^*, \mathbf{w}^*, \mathbf{x}^* \rangle$ is an optimal solution.

(ii) Theorems 1 and 2 imply that $\hat{v}(s) \geq v^*(s)$ for all $s \in S$. As $\langle \hat{\mathbf{v}}, \hat{\mathbf{w}}, \hat{\mathbf{x}} \rangle$ is an optimal solution to the MILP (6a)-(6b), part *(i)* implies that $\sum_{s \in S} \alpha(s) \big(\hat{v}(s) - v^*(s) \big) = 0$. Since $\alpha(s) > 0$ for all $s \in S$, it follows that $\hat{v}(s) = v^*(s)$ for all $s \in S$.

(iii) Part *(ii)* implies that $\hat{\mathbf{v}} = \mathbf{v}^*$. By a proof similar to that of Theorem 2, it can be seen that $\hat{x}(s, a_s) = 1$ and $\hat{x}(s, a) = 0$ for all $a \in A_s \backslash \{a_s\}$, $\hat{w}(s, j, a) = 0$ for all $a \in A_s \backslash \{a_s\}$ and $j \in S$, and $\hat{w}(s, j, a_s) = \hat{v}(j)$ for all $j \in S$. Given (5b) for a_s, it follows that $\hat{v}(s) \geq c(s, a_s) + \sum_{j \in S} p(j|s, a_s) \hat{v}(j)$ for all $s \in S$. By MDP optimality equations, it follows that:

$$\hat{v}(s) = \min_{a \in A_s} \{ c(s, a) + \sum_{j \in S} p(j|s, a) \hat{v}(j) \} \qquad \forall s \in S.$$

As a_s is the argument minimizing the right-hand side of the above equation, it follows from Theorem 7.3.5 of Puterman (2005) that $d^{\infty}_{\hat{\mathbf{x}}}$ is an optimal policy. □

2.1. Strength of the MILP Formulation

In this subsection, we essentially provide Theorem 4 that enhances our understanding of the tightness of the MILP formulation (6a)-(6b). For this purpose, we present the following definition.

Definition 1. (Vielma 2015) An MILP formulation is called *ideal* when extreme points of its LP relaxation automatically satisfy integrality requirements of integer variables.

The tightness of the LP relaxation for an MILP formulation significantly affects its solution time by state-of-the-art MILP solvers. Hence, the tightness is adopted to measure strength of an MILP formulation (Vielma 2015). In principle, a formulation with a tighter LP relaxation is favorable since state-of-the-art MILP solvers adopt the branch-and-cut algorithm which repeatedly solves the LP relaxation, and a tighter LP relaxation results in a faster convergence (Vielma 2015). For an ideal formulation, we may simply solve its LP relaxation instead of the MILP formulation. Note that the number of variables in the formulation (6a)-(6b) is $\sum_{s \in S} \big(|A_s|(|S| + 1) + 1 \big)$, which is used as another indicator to measure strength of the formulation (Vielma 2015).

In the following, we present a regularity condition under which the operations of projection and intersection commute, and it is used in the proof of Theorem 4. Given a set $Q \subset \mathbb{R}^{n+p}$, define its projection onto the linear subspace $\mathbb{R}^n \times \{0\}^p$ as follows:

$$proj_{\mathbf{x}}(Q) := \{\mathbf{x} \in \mathbb{R}^n : \exists \mathbf{y} \in \mathbb{R}^p \text{ s.t. } (\mathbf{x}, \mathbf{y}) \in Q\}.$$

Let $\mathscr{I} := \{1, \ldots, I\}$ be an index set. For each $i \in \mathscr{I}$, let $P_i := \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+p_i} : \mathbf{A}_i\mathbf{x} + \mathbf{B}_i\mathbf{y}_i \leq \mathbf{b}_i\}$, where $\mathbf{A}_i$ and $\mathbf{B}_i$ are, respectively, $m_i \times n$ and $m_i \times p_i$ matrices, and $\mathbf{b}_i$ is an m_i-dimensional column vector. Therefore, $\{P_i\}_{i\in\mathscr{I}}$ is a set of polyhedra that share the same set of the first n variables, and the rest of their variables are different. To investigate $\cap_{i\in\mathscr{I}} P_i$, we clearly need to consider the space of $\mathbb{R}^{n+\sum_{i=1}^{I} p_i}$ including all variables $\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_I$. In such a space, we may represent P_i as follows:

$$P_i := \{(\mathbf{x}, \mathbf{y}_1, \ldots, \mathbf{y}_I) \in \mathbb{R}^{n+\sum_{i=1}^{I} p_i} : \mathbf{A}_i\mathbf{x} + \mathbf{B}_i\mathbf{y}_i \leq \mathbf{b}_i\} \qquad \forall i \in \mathscr{I}.$$

LEMMA 1. *The following holds:*

$$proj_{\mathbf{x}}(\cap_{i\in\mathscr{I}} P_i) = \cap_{i\in\mathscr{I}} proj_{\mathbf{x}}(P_i).$$

PROOF. $\Rightarrow$: Suppose $\hat{\mathbf{x}} \in proj_{\mathbf{x}}(\cap_{i\in\mathscr{I}} P_i)$. Then, there exist $\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_I$ such that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_I) \in \cap_{i\in\mathscr{I}} P_i$. As a result, $\hat{\mathbf{x}} \in proj_{\mathbf{x}}(P_i)$ for all $i \in \mathscr{I}$, which implies that $\hat{\mathbf{x}} \in \cap_{i\in\mathscr{I}} proj_{\mathbf{x}}(P_i)$.

$\Leftarrow$: Suppose $\hat{\mathbf{x}} \in \cap_{i\in\mathscr{I}} proj_{\mathbf{x}}(P_i)$. Then, there exists $\hat{\mathbf{y}}_i$ such that $(\hat{\mathbf{x}}, \hat{\mathbf{y}}_i) \in P_i$ for all $i \in I$. We may restate this in the space of $\mathbb{R}^{n+\sum_{i=1}^{I} p_i}$ as $(\hat{\mathbf{x}}, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_I) \in P_i$ for all $i \in I$. Then, $(\hat{\mathbf{x}}, \hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_I) \in \cap_{i\in\mathscr{I}} P_i$, which implies that $\hat{\mathbf{x}} \in proj_{\mathbf{x}}(\cap_{i\in\mathscr{I}} P_i)$. $\square$

Let us fix $s \in S$ in the formulation of Δ, which results in the following formulation:

$$v(j) = \sum_{a\in A_s} w(s,j,a) \qquad \forall j \in S, \tag{7a}$$

$$w(s,s,a) - \sum_{j\in S} p(j|s,a)w(s,j,a) \geq c(s,a)x(s,a) \qquad \forall a \in A_s, \tag{7b}$$

$$0 \leq w(s,j,a) \leq \overline{v}(j)x(s,a) \qquad \forall a \in A_s, \tag{7c}$$

$$\sum_{a\in A_s} x(s,a) = 1 \tag{7d}$$

$$x(s,a) \in \{0,1\} \qquad \forall a \in A_s, \tag{7e}$$

$$v(s) \geq 0. \tag{7f}$$

Let $\Delta_s := \{\langle \mathbf{v}, \mathbf{w}, \mathbf{x}\rangle \,|(7a) - (7f)\}$. So that $\Delta = \cap_{s\in S}\Delta_s$. Moreover, let $LP_R(\Delta_s)$ and $LP_R(\Delta)$ denote the LP relaxations of Δ_s and Δ, respectively.

THEOREM 4. *(i) For each $s \in S$, the convex hull of Γ_s is equal to the projection of $LP_R(\Delta_s)$ onto* $\mathbf{v}$*-space.*
(ii) The intersection of convex hull of Γ_s over all $s \in S$ is equal to the projection of $LP_R(\Delta)$ onto $\mathbf{v}$*-space.*

PROOF. *(i)* For each $s \in S$, (4) represents the union of $|A_s|$ polytopes. Theorems 3.3 & 3.4 of Balas (1985) or Proposition 3.2. of Jeroslow and Lowe (1984) implies that the convex hull of solutions to (4) is the projection of $LP_R(\Delta_s)$ onto $\mathbf{v}$-space.

(ii) Given a set $F \subseteq \mathbb{R}^n$, let $conv(F)$ denote its convex hull.

$$\begin{aligned} proj_{\mathbf{v}}\big(LP_R(\Delta)\big) &= proj_{\mathbf{v}}\Big(\bigcap_{s\in S} LP_R(\Delta_s)\Big) \\ &= \bigcap_{s\in S} proj_{\mathbf{v}}\big(LP_R(\Delta_s)\big) \\ &= \bigcap_{s\in S} conv(\Gamma_s), \end{aligned}$$

where the second and third equalities follow, respectively, from Lemma 1 and part (i). □

In summary, it is desirable to provide an ideal formulation for Γ, which is the intersection of the union of polyhedra. To the best of our knowledge, providing such a formulation seems unlikely based on the available results in the integer programming literature (Vielma 2015). However, there are well-known results in disjunctive programming which provides an ideal formulation for the union of polyhedra. We adopt them to provide an ideal formulation for Γ_s for all $s \in S$, and combine them to provide a tight formulation for Γ. Theorem 4 illuminates our understanding regarding the tightness of the formulation.

REMARK 1. Compared to LPs, MILPs may more effectively incorporate any information available regarding a special structure of an optimal policy. More specifically, many real-life applications of MDPs exhibit a monotone structure of an optimal policy (Puterman 2005, Ross 1983). For instance, suppose that $A_s = A$ for all $s \in S$, both state and action spaces are ordered sets, and an optimal policy has a monotone structure in the sense that if $s_1 \geq s_2$, then the optimal action in s_1 is at least as big as that of s_2. This structure may be captured in the formulation of (6a)-(6b) by adding $x(s_1, a_1) \leq x(s_2, a_2)$ for all $s_1 \geq s_2$ and $a_1 \leq a_2$. For the case where S is a subset of integer numbers, this can be more compactly represented as $x(s+1, a_1) \leq x(s, a_2)$ for all $s \in S$ and $a_1 \leq a_2$. Adding these constraints to (6a)-(6b) tightens its LP relaxation, and improves its solution time.

2.2. Computing the Value Function of a Policy

In this subsection, we present an LP formulation to compute the upper-bound $\overline{\mathbf{v}}$ used in the formulation of Δ. As noted earlier, $\overline{\mathbf{v}}$ could be the value function of an arbitrary policy, chosen

in an ad hoc manner, with a finite value function for all states. Let d^∞ and $\mathbf{v}^{d^\infty}$ denote such a policy and its value function, respectively. Theorem 1 implies that $\mathbf{v}^{d^\infty}$ is the minimal nonnegative solution of the following set of inequalities:

$$v(s) \geq c(s, d(s)) + \sum_{j \in S} p(j|s, d(s))v(j) \qquad \forall s \in S. \tag{8}$$

Therefore, $\mathbf{v}^{d^\infty}$ is the optimal solution of the following LP:

$$\min \sum_{s \in S} \alpha(s)v(s) \tag{9a}$$

$$s.t.$$

$$v(s) \geq c(s, d(s)) + \sum_{j \in S} p(j|s, d(s))v(j) \qquad \forall s \in S. \tag{9b}$$

Note that although d^∞ may be arbitrarily chosen, we should attempt to choose it as close as possible to an optimal policy. This is because a near-optimal policy possesses a smaller value function, which leads to a tighter MILP formulation for Δ and a smaller solution time.

2.3. An Example: Our Formulation for a Class of Stopping Problems

In this subsection, we illustrate our formulation for a class of stopping problems studied by Ross (1983) as follows: A decision maker observes the current state of the system $i \in S$, and then he may either choose to *stop* and receive the reward $r(i) \geq 0$, or *continue* at the cost of $c(i) \geq 0$. If the decision maker decides to continue, the next state will be j with probability $p(j|i)$. If the decision maker decides to stop, the system moves into an absorbing state Ω, where the system remains forever with no cost nor a reward.

As the decision maker may receive a reward $r(i) > 0$, the problem is not within the framework of a negative MDP. To circumvent this issue, Ross (1983) assumed a stability condition under which $\inf_{i \in S} c(i) > 0$ and $\sup_{i \in S} r(i) < \infty$, and then transformed the problem into an "equivalent" negative MDP as follows. Consider a related optimal stopping problem for which the stopping *reward* $r(i)$ of the original problem is changed to a stopping *cost* $r - r(i)$ where $r := \sup_{i \in S} r(i)$. Ross (1983) observed that for each policy π that stops almost surely, $\underline{v}_\pi(i) = v_\pi(i) + r$ where $\mathbf{v}_\pi$ and $\underline{\mathbf{v}}_\pi$ denote the value function of π for the original MDP and the modified one, respectively. He noted that investigation of such policies suffices for finding an optimal policy as the other policies have an infinite value function at least at one state. Therefore, a policy is optimal for the original problem if and only if it is optimal for the transformed problem. To characterize the optimal policy, Ross (1983) defined the set of states for which stopping is not worse than continuing for exactly one more period and then stopping as $B := \{i \in S : -r(i) \leq c(i) - \sum_{j \in S} p(j|i)r(j)\}$. Then, he showed that if the set B is closed in the sense that if the process enters B, it will never leave it, then the

optimal policy is to stop as soon as the process enters B. Finally, Ross (1983) adopted this class of stopping problems for three specific applications: selling an asset, a burglar problem, and searching for distinct types.

In many practical and theoretical settings, the set B is not necessarily closed. Hence, it is crucial to develop a method which can identify the optimal value function in such settings. For this purpose, we present our MILP formulation for the transformed stopping problem. Letting 0 and 1, respectively, denote the actions of continuing and stopping, our MILP formulation is as follows:

$$\min \sum_{i\in S} \alpha(i)v(i) \tag{10a}$$

$$s.t.$$

$$v(j) = w(i,j,0) + w(i,j,1) \qquad \forall i,j \in S, \tag{10b}$$

$$w(i,i,0) - \sum_{j\in S} p(j|i)w(i,j,0) \geq c(i)x(i,0) \qquad \forall i \in S, \tag{10c}$$

$$w(i,i,1) \geq \big(r - r(i)\big)x(i,1) \qquad \forall i \in S, \tag{10d}$$

$$0 \leq w(i,j,0) \leq \big(r - r(j)\big)x(i,0) \qquad \forall i,j \in S, \tag{10e}$$

$$0 \leq w(i,j,1) \leq \big(r - r(j)\big)x(i,1) \qquad \forall i,j \in S, \tag{10f}$$

$$x(i,0) + x(i,1) = 1 \qquad \forall i \in S, \tag{10g}$$

$$x(i,0), x(i,1) \in \{0,1\} \qquad \forall i \in S, \tag{10h}$$

where we used the policy that stops in all states to evaluate the upper bound vector of $\overline{\mathbf{v}}$ in the formulation of (6a) - (6b), which resulted in $\overline{v}(j) = r - r(j)$ for all $j \in S$.

3. Big-M Formulation

The formulation (5a)-(5e) for Γ possesses $\sum_{s\in S}\big(|A_s|(|S|+1)+1\big)$ variables. A formulation with fewer variables may be provided by the *big-M* method as follows:

$$\min \sum_{s\in S} \alpha(s)v(s) \tag{11a}$$

$$s.t.$$

$$v(s) - \sum_{j\in S} p(j|s,a)v(j) \geq c(s,a) - Mx(s,a) \qquad \forall s \in S, a \in A_s, \tag{11b}$$

$$\sum_{a\in A_s} \big(1 - x(s,a)\big) \geq 1 \qquad \forall s \in S, \tag{11c}$$

$$x(s,a) \in \{0,1\} \qquad \forall s \in S, a \in A_s, \tag{11d}$$

$$v(s) \geq 0 \qquad \forall s \in S, \tag{11e}$$

where M is a sufficiently large constant. Let $\Theta := \{\langle \mathbf{v}, \mathbf{x}\rangle \,|(11b) - (11e)\}$. The next theorem shows that the optimal value function and an optimal policy may be derived from the optimal solution of (11a)-(11e), and vice versa.

THEOREM 5. *(i) The vector* $\mathbf{v}$ *belongs to* Γ *if and only if there exists* $\mathbf{x}$ *such that* $\langle \mathbf{v}, \mathbf{x} \rangle \in \Theta$.
(ii) There exists $\mathbf{x}^*$ *such that* $\langle \mathbf{v}^*, \mathbf{x}^* \rangle$ *is an optimal solution to the MILP* (11a)-(11e).
(iii) If $\langle \hat{\mathbf{v}}, \hat{\mathbf{x}} \rangle$ *is an optimal solution to the MILP* (11a)-(11e), *then* $\hat{\mathbf{v}} = \mathbf{v}^*$.
(iv) Given an optimal solution $\langle \hat{\mathbf{v}}, \hat{\mathbf{x}} \rangle$ *to the MILP* (11a)-(11e), *let* $d^\infty_{\hat{\mathbf{x}}}$ *be a policy under which in each state* $s \in S$, *we choose arbitrarily only one action* $a_s \in A_s$ *for which* $\hat{x}(s, a_s) = 0$. *Then,* $d^\infty_{\hat{\mathbf{x}}}$ *is an optimal policy.*

PROOF. The proof is similar to those of Theorems 2 & 3. □

The number of variables in this formulation is $\sum_{s \in S}(1 + |A_s|)$, which is relatively smaller than that of Δ. However, its LP relaxation is weaker than that of Δ, and assigning a value to the big-M coefficient is a computational challenge on its own. Hence, the formulation of Δ is superior to that of Θ.

4. Conclusion

LP formulations are available to solve discounted, average cost, and positive MDPs. An LP formulation is unavailable for negative MDPs, and identifying their optimal value functions and their optimal policies had remained a challenge. To overcome such a challenge, we developed two MILP formulations. Although MILPs are computationally more challenging than LPs, they can effectively be solved by state-of-the-art MILP solvers (Vielma 2015). As an advantage over LPs, MILPs may more efficiently incorporate structural properties of an optimal policy through adding constraints reflecting those properties, and this decreases their solution times.

References

Egon Balas. Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM Journal on Algebraic Discrete Methods*, 6(3):466–486, 1985.

Egon Balas. Disjunctive programming: Properties of the convex hull of feasible points. *Discrete Applied Mathematics*, 89(1-3):3–44, 1998.

Robert G Jeroslow and James K Lowe. Modelling with integer variables. In *Mathematical Programming at Oberwolfach II*, pages 167–184. Springer, 1984.

Warren B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality.* John Wiley & Sons, Inc., New York, NY, USA, 2 edition, 2011. ISBN 9780470604458.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, USA, 2 edition, 2005.

Sheldon M Ross. *Introduction to Stochastic Dynamic Programming.* Academic press, 1983.

Ralph E Strauch. Negative dynamic programming. *The Annals of Mathematical Statistics*, 37(4):871–890, 1966.

Kalyan Talluri and Garrett van Ryzin. *The Theory and Practice of Revenue Management*. Springer US, 1 edition, 2004. ISBN 978-1-4020-7701-2.

Juan Pablo Vielma. Mixed integer linear programming formulation techniques. *SIAM Review*, 57(1):3–57, 2015.