Mixed Strategy Improvement Algorithm for Markov Decision Processes

Ashutosh Trivediand Yuen-Lam Voronin

Department of Computer Science, University of Colorado Boulder, USA.
{ashutosh.trivedi,yuen.voronin}@colorado.edu

Abstract

One standard way of finding an optimal policy in a given Markov decision process (MDP) is to set up a linear program whose (unique) optimal solution satisfies the Bellman equation for the value vector of any optimal policy of that MDP. We show that the dual of that standard linear program is closely related to the probabilistic policies (or mixed strategies) of the MDP, and based on that connection we propose a polynomial time policy-iteration algorithm.

1 Background

Markov decision processes (MDPs) [7] are natural and expressive formalism to capture optimal decision making situations under stochastic environment. Bellman equations provide an elegant framework to study several optimization objectives—including including expected total-reward, discounted-reward, and average-reward—that characterize optimal value via fixed points of so-called Bellman operator. Value iteration and strategy improvement are two classical algorithms to compute fixed point of Bellman operators. Value iteration algorithm works by computing an improved value at each iteration, guided by Bellman operator, until a specific termination condition is met. On the other hand, strategy improvement works by computing a sequence of pure and positional strategies till no further improvement can be made. The termination is guaranteed due to strict improvement at each step and finiteness of the set of pure and positional strategies.

There is a polynomial-time reduction [6] from MDPs to linear programming (LPs), and since the general instances of LP can be solved in polynomial-time [3], it implies that MDPs can be solve in polynomial time. However no *strongly polynomial algorithm* is known to solve LP, i.e. an algorithm that is polynomial in the number of variables and constraints with no dependence on actual numbers appearing in the description of the constraints or the objective function. In contrast, there are strongly polynomial algorithms to solve linear equation system. It is a long open equation whether there exists a strongly polynomial algorithm to solve LP. Similarly, no strongly polynomial algorithm is known to solve MDPs, i.e., an algorithm that is polynomial only in the number of states and edges, with no dependence on the actual probabilities appearing in the description of the MDP.

Dantzig's Simplex algorithm is a popular algorithm to solve LP and works well in practice. Since it is well known that optimal value of the vector lies on a vertex of the polytope characterized by the linear constraint system, Simplex algorithm begins with some vertex, and repeatedly produces vertices (according to some so-called *pivot rule*) of the polytope while ensuring that the value of the objective function for the current vertex is strictly better than the previous vertex. Since the number of vertices of constraint polytope is finite, and there is strict improvement in every iteration, the Simplex algorithm terminates in finitely many iterations. Each iteration involves solving a linear equation system, and hence is strongly polynomial. So the complexity of simplex algorithm is directly related to the number of iterations. Unfortunately, an example [4] from Klee and Minty shows that simplex algorithm can take exponentially many iterations for some instances. The original


Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

proof was for a fixed pivot rule, but later it was shown that the argument carries over for any deterministic pivot rule that does not depend upon the history of previous updates.

Interior-point algorithms [2, 5] to solve LP take a completely different route from the Simplex algorithm: instead of following vertices of constraint polytope to the optimal one, interior point algorithms start from some interior point of the polytope and at each iteration make a progress towards the optimal vertex while keeping each new point strictly in the interior of the polytope. Interior-point algorithms, e.g. [2, 5], to solve LPs are known to work in polynomial time, and the goal of this work is to find a direct algorithm, inspired by interior point algorithms, on MDPs that work in polynomial time. Similar to the analogy of simplex algorithm for LP and strategy improvement algorithm in positional strategy space for MDP, we feel that there is a direct connection between interior point algorithms for LP and strategy improvement algorithm in *randomized* (mixed and stationary) strategy space for MDP. We aim to design such strategy improvement algorithm and study its complexity for various MDPs and stochastic games.

There is a recent interest in the complexity of MDP/Stochastic games problems due to a recent result of Ye [9] that shows that simplex method is strongly polynomial for discounted Markov decision problems with fix discount factor. Building upon the same result, Hansen, Miltersen, and Zwick [1] showed that strategy improvement algorithm is strongly polynomial for discounted cost stochastic games for a fixed discount factor.

Contributions. The contributions of this article on finding near optimal strategies for fixed discount rate Markov decision processes is two-fold. We first highlight the deep connection between the dual of the classical LP formulation for solving the Bellman equation for the optimal value of fixed discount rate Markov decision processes. The solutions in the dual LP are in one-to-one correspondence with the set of mixed strategies, and the dual LP can be seen as a "surrogate" optimization problem for finding optimal strategies. Then, using this connection, we propose a polynomial time mixed strategy improvement scheme (Algorithm 4.1), which is a variant of the classical long step path following method for solving linear programs [8]. We provide and prove the convergence and complexity properties of our scheme. The complexity result depends only on the discount rate, the problem size and the user-defined tolerance. We mention in passing that while it is generally understood that standard interior point algorithms enjoy polynomial time complexity, this depends not just on the problem size, but also on the "size" of the neighborhood of the central path that the iterates must remain within. Our complexity result eliminates such a dependence, because we can show that for *all* instances, the iterates always stay in a certain "large" enough neighborhood, and the size depends only on the discount ratem and the number of states and actions.

Organization. The article is organized as follows. We first formally describe the core problem of Markov decision process in Section 2. In Section 3, we state the LP formulation for solving the Bellman equation for the optimal value vector. Then we state the mathematical connection between the dual of the LP formulation and the mixed strategies in a Markov decision process. In Section 4, we provide a mixed strategy improvement scheme (in Algorithm 4.1) that is built upon the connection explained in Section 3 (see also Section 4.1), and explain the main steps in the scheme (in Section 4.2). We then state the convergence and complexity properties of the mixed strategy improvement scheme in Section 4.3. All the proofs for various results are collected in the Appendix.

2 Preliminaries

In this article, we are interested in Markov decision processes involving finitely many states and available actions, evolving over infinite horizon. Formally, a *Markov decision process* (MDP) is a tuple $\mathcal{M} = (S, A, p(\cdot|\cdot,\cdot), r, \gamma)$, where S is a finite set of states, A is a finite set of available actions (uniform across all the states), $p : S \times S \times A \to \mathbb{R}$ is a function with $p(s'|s,a)$ being the probability of reaching state $s' \in S$ from state $s \in S$ when action $a \in A$ is taken, $r = (r_{s,a})_{s\in S, a\in A} \in \mathbb{R}^{S\times A}$ with $r_{s,a}$ being the immediate reward of taking action $a \in A$ at state $s \in S$, and $\gamma \in (0,1)$ being discount factor for future rewards.

A (stationary) *policy* or *pure strategy* in a MDP is an *a priori* choice of an action at each state, formalized via a function $\pi : S \to A$. More generally, we consider *mixed strategies*, which are vectors of the form $\tilde{x} \in \mathbb{R}^{S\times A}$ with $\tilde{x}_{s,a}$ being the probability of taking action a at state s; in the following we use $\mathcal{S}$ to denote the set of all mixed strategies, i.e., $\mathcal{S} := \{\tilde{x} \in \mathbb{R}^{S\times A} : \sum_{a\in A} \tilde{x}_{s,a} = 1,\ \forall s \in S\}$. The *value* of a mixed strategy $\tilde{x}$ is a vector $v^{\tilde{x}} \in \mathbb{R}^S$, where for each $s \in S$, $v^{\tilde{x}}_s$ is the total reward received over infinite horizon (subject to discount factor γ) if s is the initial state of the MDP. An *optimal strategy* is a strategy $\tilde{x}^*$ whose value v^* satisfies the *Bellman equation* $v^* = \mathcal{T}(v^*)$, where the *Bellman operator* $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$ is defined on any $v \in \mathbb{R}^S$ by

$$(\mathcal{T}(v))_s = \max_{a\in A} \left\{ r_{s,a} + \gamma \sum_{s'\in S} p(s'|s,a) v_{s'} \right\}, \qquad \text{for all states } s \in S.$$

There is a unique vector $v \in \mathbb{R}^S$ that satisfies the fixed point equation $v = \mathcal{T}(v)$ (i.e., the optimal value vector v^* is unique). While there may be more than one optimal policy, the optimal value v^* can always be attained by a pure strategy. A policy $\tilde{x}$ is said to be *ε-optimal* if $\|v^{\tilde{x}} - v^*\| \le \varepsilon$.

In the following, we make the following assumption on the rewards in the MDP.

▶ Assumption 1. The absolute values of all the rewards are bounded by 1, i.e., $|r_{s,a}| \le 1$, for all $s \in S$, $a \in A$.

This assumption amounts to normalizing $r_{s,a}$'s, and can be made without loss of generality.[1] We make this normalizing assumption to ensure the polynomial time complexity of the proposed Algorithm 4.1.

The main problem in MDPs is to find an optimal (pure) strategy for any given MDP. In this article, we propose a polynomial time algorithm for finding an ε-optimal mixed strategy.

3 Relation between a classical LP formulation and mixed strategies

A common approach in finding the optimal value v^* satisfying the nonlinear equation $v^* = \mathcal{T}(v^*)$ is to relax the equality constraint and at the same time minimize the components of

[1] For any general MDP with $\max_{s',a'}\{|r_{s',a'}|\} > 1$, the Bellman equation is equivalent to

$$\frac{v_s}{\max_{s',a'}\{|r_{s',a'}|\}} = \max_{a\in A} \left\{ \frac{r_{s,a}}{\max_{s',a'}\{|r_{s',a'}|\}} + \gamma \sum_{s'\in S} p(s'|s,a) \frac{v_{s'}}{\max_{s',a'}\{|r_{s',a'}|\}} \right\}, \quad \forall s \in S,$$

meaning that if we replace each $r_{s,a}$ by $r_{s,a}/\max_{s',a'}\{|r_{s',a'}|\}$ and solve the new Bellman equation for v, rescaling the solution v gives the solution to the original Bellman equation.

the value vector, resulting in the optimization problem

$$\begin{aligned} \nu = \min_{v\in\mathbb{R}^S} \quad & \textstyle\sum_{s\in S} v_s \\ \text{s.t.} \quad & v \geq \mathcal{T}(v), \end{aligned}$$

which is equivalent to the linear program (LP):

$$\begin{aligned} \nu = \min_{v\in\mathbb{R}^S} \quad & \textstyle\sum_{s\in S} v_s \\ \text{s.t.} \quad & v_s \geq r_{s,\,a} + \gamma \textstyle\sum_{s'\in S} p(s'|s,a) v_{s'}, \quad \forall\, s\in S,\ a\in A. \end{aligned} \tag{1}$$

One significance of the LP formulation (1) is that its (unique) optimal solution is always *complementary* to any optimal strategy vector. While most attention is given to the primal LP (1), in the following, we shed light on the oft-overlooked importance of its dual, which carries very rich information about the mixed strategies of the MDP. The dual allows us to indirectly search over the space of mixed strategies and the optimal dual solutions corresponds precisely to the optimal strategies of the MDP.

$$\begin{aligned} \max_{x} \quad & \textstyle\sum_{s\in S, a\in A} r_{s,a} x_{s,a} \\ \text{s.t.} \quad & \textstyle\sum_{a\in A} x_{s',a} = 1 + \gamma \sum_{s\in S, a\in A} p(s'|s,a) x_{s,a}, \quad \forall\, s'\in S, \\ & x \geq 0. \end{aligned} \tag{2}$$

(The fact that the two LPs (1) and (2) have the same optimal value ν follows from the fundamental theorem of linear programming.) While the primal feasible solutions of the LP (1) mimic value vectors in the MDP, the dual feasible solutions of (2) have a natural one-to-one correspondence with the mixed strategies in the MDP. Specifically, each dual feasible solution x corresponds uniquely to a mixed strategy $\tilde{x}$ via the normalization $\tilde{x}_{s,a} = \frac{x_{s,a}}{w_s}$, where $w_s = \sum_{a'} x_{s,a'}$. Moreover, the component sum of the value vector $v^{\tilde{x}}$ corresponding to the mixed strategy $\tilde{x}$ computed from x is precisely the dual objective value $\sum_{s,a} r_{s,a} x_{s,a}$, so the performance of the mixed strategy $\tilde{x}$ is not worse than the difference between the dual objective value and the dual optimal value.

Formally, recall that $\mathcal{S} = \{\tilde{x}\in\mathbb{R}^{S\times A} : \sum_{a\in A} \tilde{x}_{s,a} = 1\}$ is defined as the set of all mixed strategies and let $\mathcal{D} = \{x\in\mathbb{R}^{S\times A} : \sum_{a\in A} x_{s',a} = 1 + \gamma \sum_{s\in S, a\in A} p(s'|s,a) x_{s,a}, \quad \forall\, s'\in S\}$ be the set of all feasible solution to the dual LP (2). For any mixed strategy $\tilde{x}\in\mathcal{S}$, define $\mathsf{P}(\tilde{x}), \mathsf{B}(\tilde{x}) \in \mathbb{R}^{S\times S}$ by

$$\mathsf{P}(\tilde{x})_{s',s} = \sum_{a\in A} p(s'|s,a)\tilde{x}_{s,a}, \quad \forall\, s', s\in S, \quad \text{and} \quad \mathsf{B}(\tilde{x}) \triangleq I - \gamma \mathsf{P}(\tilde{x}).$$

Then the matrix $\mathsf{B}(\tilde{x})$ is an M-matrix, i.e., $\mathsf{B}(\tilde{x})$ is invertible and $\mathsf{B}(\tilde{x})^{-1}y$ is a nonnegative vector whenever y is a nonnegative vector. (See Appendix A.1.) Moreover, $\mathsf{B}(\tilde{x})^{-1}\mathbb{1}$ is a positive vector, where $\mathbb{1}$ denotes the vector of all ones of appropriate length. The matrices $\mathsf{P}(\tilde{x})$ and $\mathsf{B}(\tilde{x})$ play an important role in enabling a one-to-one correspondence between the feasible solutionsof the dual LP (2) and the mixed strategies. In addition, we can easily calculate the values of any mixed strategies using the matrices $\mathsf{P}(\tilde{x})$ and $\mathsf{B}(\tilde{x})$; see Theorem 4 below.

Now we focus on establishing a one-to-one correspondence between the feasible solutions of the dual LP (2) and the mixed strategies. Define the following functions:

$$\begin{aligned} W_{\mathcal{D}}: \quad & \mathcal{D}\to\mathbb{R}^S: \quad x\mapsto w = (w_s)_{s\in S}, && \text{where } w_s = \left(\textstyle\sum_{a\in A} x_{s,a}\right)_{s\in S}, \\ W_{\mathcal{S}}: \quad & \mathcal{S}\to\mathbb{R}^S: \quad \tilde{x}\mapsto \mathsf{B}(\tilde{x})^{-1}\mathbb{1}, && \\ f: \quad & \mathcal{D}\to\mathbb{R}^{S\times A}: \quad x\mapsto \tilde{x} = (\tilde{x}_{s,a})_{s\in S, a\in A}, && \text{where } \tilde{x}_{s,a} = \tfrac{x_{s,a}}{w_s},\ w = W_{\mathcal{D}}(x), \\ g: \quad & \mathcal{S}\to\mathbb{R}^{S\times A}: \quad \tilde{x}\mapsto x = (x_{s,a})_{s\in S, a\in A}, && \text{where } x_{s,a} = \tilde{w}_s\tilde{x}_{s,a},\ \tilde{w} = W_{\mathcal{S}}(\tilde{x}). \end{aligned} \tag{3}$$

We claim that f defines a one-to-one function mapping dual feasible solutions of (2) to mixed strategies, with the inverse of f being defined by g.

▶ **Theorem 2.** *The following holds for the functions defined in (3).*

1. *For any $x \in \mathcal{D}$, $W_{\mathcal{D}}(x) \geq \mathbb{1}$, and for any $\tilde{x} \in \mathcal{S}$, $W_{\mathcal{S}}(\tilde{x}) > 0$.*
2. *The range of the function f lies in the set of mixed strategy, i.e., $\mathcal{R}(f) \subseteq \mathcal{S}$, and the range of the function g lies in the set of feasible solutions of the dual LP (2), i.e., $\mathcal{R}(g) \subseteq \mathcal{D}$.*
3. *For any $x \in \mathcal{D}$, $W_{\mathcal{D}}(x) = W_{\mathcal{S}}(f(x))$, and for any $\tilde{x} \in \mathcal{S}$, $W_{\mathcal{S}}(\tilde{x}) = W_{\mathcal{D}}(g(\tilde{x}))$.*
4. *$f \circ g = \mathrm{id}_{\mathcal{S}}$ and $g \circ f = \mathrm{id}_{\mathcal{D}}$*

In particular, if x is an optimal solution to the dual LP (2), then $\tilde{x} = f(x)$ is an optimal strategy of the MDP. Conversely, if $\tilde{x}$ is an optimal strategy of the MDP, then $x = f^{-1}(\tilde{x}) = g(\tilde{x})$ is an optimal solution of the dual LP (2).

Theorem 2 hints at the possibility of handling the mixed strategies using the feasible solutions of the dual LP (2) as "surrogates". It turns out that the dual LP can elegantly *rank* the underlying mixed strategies in the following sense: if $x, x' \in \mathcal{D}$ are two feasible solutions of (2) and x has a better objective value than x', then the mixed strategy $\tilde{x}$ associated with x is better than the mixed strategy $\tilde{x}'$ associated with x' (see (4) in Theorem 4). To formalize this observation, we first point out some technical and useful results that allow us to evaluate the values of strategies based solely on the dual LP (2).

▶ **Lemma 3.** *The value vector of any mixed strategy $\tilde{x} \in \mathcal{S}$ is given by*

$$v^{\tilde{x}} = \mathsf{B}(\tilde{x})^{-\intercal} \tilde{r},$$

where $\tilde{r} \in \mathbb{R}^{S \times S}$ is defined by

$$\tilde{r}_s = \sum_{a \in A} r_{s,a} \tilde{x}_{s,a}.$$

The value vector is majorized by the optimal value vector, i.e., $v^{\tilde{x}} \leq v^*$. Moreover, for any $x \in \mathcal{D}$, if $\tilde{x}$ is the mixed strategy associated with x, i.e., if $\tilde{x} = f(x)$, then the total value of the mixed strategy $\tilde{x}$ coincides with the dual objective value attained by x, i.e.,

$$\sum_{s \in S} v_s^{\tilde{x}} = \sum_{s \in S, a \in A} r_{s,a} x_{s,a}.$$

Lemma 3 allows us to evaluate and compare the quality of underlying mixed strategies associated with given dual feasible solutions.

▶ **Theorem 4.** *For any feasible solution $x \in \mathcal{D}$ of the dual LP (2), let $\tilde{x} = f(x) \in \mathcal{S}$ be the corresponding mixed strategy. Then the value $v^{\tilde{x}}$ of the mixed strategy $\tilde{x}$ satisfies*

$$\|v^{\tilde{x}} - v^*\| \leq \nu - \sum_{s \in S, a \in A} r_{s,a} x_{s,a},$$

where v^* is the unique optimal value vector for the MDP and ν is the optimal value of the LP (2). Moreover, if $x' \in \mathcal{D}$ is another dual solution with a worse objective value and $\tilde{x}' \triangleq f(x')$, then $v^{\tilde{x}}$ is closer to the optimal value vector v^* than $v^{\tilde{x}'}$ is in the ℓ_1-norm, i.e.,

$$\sum_{s \in S, a \in A} r_{s,a} x_{s,a} \geq \sum_{s \in S, a \in A} r_{s,a} x'_{s,a} \quad \textit{implies that} \quad \sum_{s \in S} |v_s^* - v_s^{\tilde{x}}| \leq \sum_{s \in S} |v_s^* - v_s^{\tilde{x}'}|. \tag{4}$$

4 A polynomial time algorithm for strategy improvement

The results in Section 3 highlights the effectiveness of using the dual LP (2) to find an optimal strategy of the MDP. These observations allow us to construct an mixed-stratgey improvement algorithm based on the classical primal-dual interior point method.

4.1 Complementary slackness and duality gap

Thanks to the close connection between the dual LP (2) and the strategy optimization, the following characterization via the notion of "complementary slackness" is immediate and serves as the driving force behind our mixed strategy improvement scheme (in Algorithm 4.1, below).

▶ **Observation 1.** For any feasible solution v of (1) and any mixed strategy $\tilde{x} \in \mathcal{S}$,

$$\tilde{x}_{s,a} \cdot \left(v_s - \gamma \sum_{s' \in S} p(s'|s,a) v_{s'} - r_{s,a} \right) \geq 0, \quad \forall\, s \in S, a \in A. \tag{5}$$

Moreover, equality holds for all $s \in S$ and $a \in A$ if and only if v is the optimal value vector and $\tilde{x}$ is an optimal strategy.

In general, for any feasible solution v of (1) and any mixed strategy $\tilde{x} \in \mathcal{S}$, we can define the *duality gap* measure

$$\mu(v, \tilde{x}) \triangleq \frac{1}{|S||A|} \sum_{s \in S, a \in A} w_s \tilde{x}_{s,a} \left(v_s - \gamma \sum_{s' \in S} p(s'|s,a) v_{s'} - r_{s,a} \right), \quad \text{where } w = \mathsf{B}(\tilde{x})^{-1} \mathbb{1},$$

which is an *weighted* average of the left-hand side quantities in (1). In particular, v and $\tilde{x}$ are optimal if and only if $\mu(v, \tilde{x}) = 0$. On top of characterizing optimality, the quantity $\mu(v, \tilde{x})$ also measures the quality of the solution $(v, \tilde{x})$. As stated in (6) below, $|S||A|\mu(v, \tilde{x})$ provides an upper bound on both the distance of the value vector $v^{\tilde{x}}$ corresponding to the strategy $\tilde{x}$ from the optimal value vector, and also the difference between the objective value of (1) attained by v and the optimal value of (1).

▶ Proposition 5. For any feasible solution v of (1) and any mixed strategy $\tilde{x} \in \mathcal{S}$,

$$\mu(v, \tilde{x}) = \frac{1}{|S||A|} \left(\sum_{s \in S} v_s - \sum_{s \in S} v_s^{\tilde{x}} \right) \geq 0,$$

and in particular,

$$\|v^{\tilde{x}} - v^*\| \leq |S||A|\mu(v, \tilde{x}) \quad \text{and} \quad 0 \leq \sum_{s \in S} v_s - \nu = \sum_{s \in S} v_s - \sum_{s \in S} v_s^* \leq |S||A|\mu(v, \tilde{x}). \tag{6}$$

4.2 A mixed strategy improvement algorithm

Based on Observation 1, it makes sense to search search *simultaneously* for a feasible solution v for the primal LP (1) and a strategy vector $\tilde{x}$, such that the *complementary slackness* condition

$$\tilde{x}_{s,a} \cdot \left(v_s - \gamma \sum_{s' \in S} p(s'|s,a) v_{s'} - r_{s,a} \right) = 0, \quad \forall\, s \in S, a \in A \tag{7}$$

is satisfied. We propose Algorithm 4.1, below, that aims at finding such a solution pair $(v, \tilde{x})$. Algorithm 4.1 is a variant of the classical long-step path following method for solving LPs [8, P. 96], and enjoys polynomial time complexity. (See Theorem 8.) While Algorithm 4.1 is a mixed strategy improvement scheme, it can also immediately provide a sequence of improving solutions v for the primal LP (1). The iterates v are important in their own rights (since they converge to the solution of the Bellman equation, see Theorem 6), they also provide important insight into the quality of the mixed strategy iterates $\tilde{x}$, via the duality gap measure $\mu(v, \tilde{x})$. In particular, at termination, the inequality $\mu(v, \tilde{x}) \leq \varepsilon$ guarantees that we would get an ε-optimal strategy. (See Proposition 5).

Before we state the main results concerning the convergence of Algorithm 4.1, we briefly explain the main steps of Algorithm 4.1.

Algorithm 4.1: Mixed strategy improvement algorithm

Initialize:

Pick tolerance $\varepsilon > 0$, and pick $0 < \sigma_{\min} < \sigma_{\max} < 1$;

Pick mixed strategy $\tilde{x} = \frac{1}{|A|}\mathbb{1}$ and $v = \frac{\max_{s' \in S, a' \in A}\{|r_{s',a'}|\}+2}{1-\gamma}\mathbb{1} \in \mathbb{R}^{S \times A}$;

Compute $z \in \mathbb{R}^{S \times A}$ by $z_{s,a} = \max_{s',a'}\{|r_{s',a'}|\} - r_{s,a} + 2$ for all $s \in S$, $a \in A$;

Compute $w \leftarrow \mathsf{B}(\tilde{x})^{-1}\mathbb{1} \in \mathbb{R}^S$;

Compute $\xi \leftarrow \min_{s \in S, a \in A}\left\{\frac{w_s}{\mu(v,\tilde{x})}\tilde{x}_{s,a}z_{s,a}\right\} > 0$;

repeat

Pick $\sigma \in [\sigma_{\min}, \sigma_{\max}]$;

Solve the following linear system for the search step $(\Delta v, \Delta z, \Delta \tilde{x})$:

$$\begin{cases} \Delta z_{s,a} = \Delta v_s - \gamma \sum_{s' \in S} p(s'|s,a)\Delta v_{s'}, & \forall\, s \in S,\ a \in A, \\ w_{s'} \sum_{a \in A} \Delta \tilde{x}_{s',a} - \gamma \sum_{s,a} w_s p(s'|s,a)\Delta \tilde{x}_{s,a} = 0, & \forall\, s' \in S, \\ z_{s,a}\Delta \tilde{x}_{s,a} + \tilde{x}_{s,a}\Delta z_{s,a} = \frac{\sigma\mu(v,\tilde{x})}{w_s} - \tilde{x}_{s,a}z_{s,a}, & \forall\, s \in S,\ a \in A; \end{cases} \tag{8}$$

For $\alpha \in [0, 1]$, define

$$\begin{cases} \tilde{x}(\alpha)_{s,a} \triangleq \frac{\tilde{x}_{s,a} + \alpha\Delta\tilde{x}_{s,a}}{\sum_{a' \in A} \tilde{x}_{s,a'} + \alpha\Delta\tilde{x}_{s,a'}}, & \forall\, s \in S,\ a \in A, \\ w(\alpha) \triangleq \mathsf{B}(\tilde{x}(\alpha))^{-1}\mathbb{1}, \\ z(\alpha) \triangleq z + \alpha\Delta z, \\ v(\alpha) \triangleq v + \alpha\Delta v; \end{cases}$$

Compute step length $\alpha \in (0, 1]$ such that

$$\begin{cases} \tilde{x}(\alpha) > 0 \quad \text{and} \quad z(\alpha) > 0, \\ \tilde{x}(\alpha)_{s,a} \cdot z(\alpha)_{s,a} \geq \frac{\xi}{w(\alpha)_s}\mu(v(\alpha), \tilde{x}(\alpha)), \quad \forall\, s \in S,\ a \in A; \end{cases}$$

Update:

$\tilde{x} \leftarrow \tilde{x}(\alpha)$, $w \leftarrow w(\alpha)$, $v \leftarrow v(\alpha)$ and $z \leftarrow z(\alpha)$;

until $\mu(v, \tilde{x}) \leq \varepsilon$;

1. *Initialization.* We pick an initial mixed strategy $\tilde{x}$ and an initial v that is *strictly* feasible for the LP (1). (See Theorem 6.) For notational convenience, we also define the *slack*

variable z via $z_{s,a} = v_s - \gamma \sum_{s' \in S} p(s'|s,a)v_{s'} - r_{s,a} = \max_{s',a'}\{|r_{s',a'}|\} - r_{s,a} + 1 > 0$ for all $s \in S$, $a \in A$.

2. *The parameter* ξ. The purpose of the parameter ξ is to ensure that all the iterates $(v, z, \tilde{x})$ in the main loop satisfy the condition

$$\tilde{x}_{s,a} z_{s,a} \geq \frac{\xi}{w_s} \mu(v, \tilde{x}), \quad \forall\, s \in S,\ a \in A, \qquad \text{where } w = \mathsf{B}(\tilde{x})^{-1}\mathbb{1}. \tag{9}$$

The idea of this inequality is that, while we would like to drive the products $\tilde{x}_{s,a} z_{s,a}$ to zero, it should happen in a uniform manner, i.e., we want to avoid some product $\tilde{x}_{s,a} z_{s,a}$ going to zero much faster than the others. To ensure that the convergence is "uniform", we require that the products are not too far from their weighted average $\mu(v, \tilde{x})$, and we enforce the requirement using inequality (9).

3. *The main loop.* Inside the main loop, two key steps iteratively occur: we solve the linear equation (8) for a *search step* $(\Delta v, \Delta z, \Delta \tilde{x})$, then we update our solution using a fraction of the search step, so that the new v remains feasible for the primal LP (1) and the new $\tilde{x}$ is indeed a strategy (i.e., $\tilde{x} \in \mathcal{S}$). We give a general idea of the significance of each of the three equations in (8).

   a. The first equation involving Δz is for notational convenience (and can be eliminated by substitution into the third equation).
   b. The second equation focuses on the variable $\Delta \tilde{x}$. Its purpose is to ensure that the updated $\tilde{x}$ will remain a strategy (i.e., $\tilde{x} \in \mathcal{S}$).
   c. The third equation is the linearization of the equation

$$\tilde{x}_{s,a} z_{s,a} = \frac{\sigma \mu}{w_s}, \quad \forall\, s \in S,\ a \in A,$$

   which results from a right-hand side perturbation of the complementary slackness condition (7). The goal of the perturbation is to control the speed at which the products $\tilde{x}_{s,a} z_{s,a}$ converge to zero, and this is pivotal for guaranteeing the polynomial time complexity of the algorithm.

   After solving the equation 8, we compute the step length α that ensures the feasibility of the new iterate $(v, \tilde{x})$. In addition, we require that $\tilde{x}$ and z do not have any components that are "closer" to zero than the others. This ensures that the solution $(v, \tilde{x})$ is not too close to the boundary of the solution set—being too close to the boundary leads to simplex-method-like behavior, and this approach of lower-bounding the products $\tilde{x}_{s,a} z_{s,a}$ helps to ensure the polynomial time complexity of the algorithm.

4. *Quality of the solution.* The value $\mu(v, \tilde{x})$ functions as a quality measure of the solution tuple $(v, z, \tilde{x})$. At the end of each iteration of the main loop, we have that (abusing the notation)

$$\mu^{\text{new}} = \mu^{\text{old}}(1 - \alpha(1 - \sigma)).$$

   In other words, the choice of α and σ can control how much the value μ decreases. Adopting a long step path following variant, Algorithm 4.1 converges in polynomial time because of this sufficient decrease property.

4.3 Validity of the mixed strategy improvement algorithm

After walking through Algorithm 4.1, we lay out the formal results concerning some key features of Algorithm 4.1.

1. Algorithm 4.1 produces an *improving* sequence of solutions v for the LP formulation (1) of the Bellman equation and an *improving* sequence of mixed strategies $\tilde{x} \in \mathcal{S}$. (See Theorem 6.)
2. Algorithm 4.1 terminates in polynomial time, depending on the problem size (i.e., the number of states $|S|$ and the number of actions $|A|$) and the parameters $\sigma_{\min}, \sigma_{\max}$ and ε. (See Theorem 8.)
3. At termination, we arrive at an ε-optimal strategy.

In the following, we let $(v^{(0)}, z^{(0)}, \tilde{x}^{(0)})$ be the initial value of $(v, z, \tilde{x})$ in Algorithm 4.1, and let $(v^{(k)}, z^{(k)}, \tilde{x}^{(k)})$ be the value of $(v, z, \tilde{x})$ at the end of the k-th iteration of the main loop (for $k = 1, 2, \ldots$). Let $\sigma^{(k)}$ and $\alpha^{(k)}$ denote respectively the value of σ and the step length α in the k-th iteration.

A key feature of the mixed strategy improvement algorithm is that it does provide a sequence of improving mixed strategy, in the sense that the ℓ_1 distance of $v^{\tilde{x}^{(k)}}$ from the optimal value vector v^* decreases as the iteration count k goes up (see the second inequality in (10)). In addition, we also have a sequence of $v^{(k)}$'s that are not only feasible for (1) but also improve on the objective value of (1).

▶ **Theorem 6.** *For each $k = 0, 1, 2, \ldots$, $\tilde{x}^{(k)}$ is a valid mixed strategy (i.e., $\tilde{x}^{(k)} \in \mathcal{S}$) and $v^{(k)}$ is feasible for the LP formulation (1). Moreover,*

$$\mu(v^{(k+1)}, \tilde{x}^{(k+1)}) = \mu(v^{(k)}, \tilde{x}^{(k)})(1 - \alpha^{(k+1)}(1 - \sigma^{(k)})).$$

and

$$\sum_{s \in S} v_s^{(k+1)} \leq \sum_{s \in S} v_s^{(k)}, \quad \text{and} \quad \sum_{s \in S} |v_s^* - v_s^{\tilde{x}^{(k+1)}}| \leq \sum_{s \in S} |v_s^* - v_s^{\tilde{x}^{(k)}}|. \tag{10}$$

Another key feature of the mixed strategy improvement scheme is that it is a polynomial time algorithm. The mixed strategy improvement algorithm results from a modification of the classical long-step path following method for solving general LPs [8, P. 96], where the modification is made possible due to the connection between mixed strategies and the dual of the LP formulation (1) that we described in Section 3. The proof of the following result, which is an indirect consequence of [8, Thm. 5.12], relies much on the one-to-one correspondence between the mixed strategies and the dual feasible solutions (Theorem 2).

To establish the polynomial time complexity for Algorithm 4.1, we first need a technical result that states that the initial duality gap measure $\mu(v^{(0)}, \tilde{x}^{(0)})$ is bounded above in terms of the problem size and the discount rate, and the parameter ξ (which determines the size of a neighborhood around the so-called central path where the iterates of the algorithm are allowed to go) is bounded below (meaning that the neighborhood is always "large enough") in terms of the problem size, for *all* the MDP instances.

▶ **Lemma 7.** *For the initial point $v^{(0)}$ and $\tilde{x}^{(0)}$ defined in Algorithm 4.1, $\mu(v^{(0)}, \tilde{x}^{(0)}) \leq \frac{4|S|}{(1-\gamma)|A|}$ and $\xi \geq \frac{1-\gamma}{4|S|}$.*

Using Lemma 7, we can prove the polynomial time complexity of Algorithm 4.1, which depends only on the discount rate, the user defined parameters $\varepsilon, \sigma_{\min}, \sigma_{\max}$ and the problem size (i.e., the number of states and the number of actions.)

▶ **Theorem 8.** *For any given $\varepsilon \in (0, 1)$ and any given $0 < \sigma_{\min} < \sigma_{\max} < 1$, there exists an index $K = O(n|\log \varepsilon|)$ dependent on $\varepsilon, \sigma_{\min}, \sigma_{\max}$ such that*

$$\mu(v^{(k)}, \tilde{x}^{(k)}) \leq \varepsilon, \quad \forall\, k \geq K.$$

Finally, as an immediate consequence of Proposition 5 and the termination requirement $\mu(v, \tilde{x}) \leq \varepsilon$, the mixed strategy $\tilde{x}$ found at termination must satisfy $\|v^{\tilde{x}} - v^*\| \leq \varepsilon$, so $\tilde{x}$ is an ε-optimal strategy.

5 Conclusion

The core problem for Markov decision processes, in finding an optimal strategy, is classically tackled by a number of methods, such as value iteration, policy iteration and linear programming techniques. This article focuses on the third approach, focusing on the LP formulation for solving the Bellman equation for optimal values (and strategies) of Markov decision processes. Thanks to the discovery of polynomial-time interior point methods for solving linear programs, it is possible to solve, for instance, fixed discount rate Markov decision processes (which is the focus of this article), in polynomial time, simply by feeding the LP formulation into a standard interior point method solver. Nonetheless, this approach largely ignores the underlying properties common to all Markov decision process; one main purpose of this article is to shed light on the deep connection between the LP formulation and mixed strategy improvement, and to modify a classical interior point algorithm into a mixed strategy scheme.

We stated the connection between the dual of the classical LP formulation and mixed strategy. The solutions in the dual LP are in one-to-one correspondence with the set of mixed strategies, and the dual LP can be seen as a "surrogate" optimization problem for finding optimal strategies. Using this connection, we proposed a mixed strategy improvement scheme that is a modification of a classical polynomial time interior point algorithm. The mixed strategy improvement scheme produces a sequence of improving strategies, and at termination the output strategy is near-optimal. Moreover, the mixed strategy improvement scheme runs in polynomial time (depending on the user-defined tolerance, the problem size, and the discount rate).

References

1 T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *CoRR*, abs/1008.0530, 2010.

2 N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4, 1984.

3 L. G. Khachiyan. A polynomial time algorithm in linear programming. *Soviet Math. Dokl.*, 20, 1979.

4 F. Klee and G.J. Minty. How good is the simplex algorithm? *Inequalities III*, pages 159–175, 1972.

5 R. D. C. Montiero and I. Adler. Interior path following primal-dual algorithms. part i:linear programming. *Mathematical Programming*, 44:27–41, 1989.

6 C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12, 1987.

7 M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, 1994.

8 S. Wright. *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics, 1997.

9 Y. Ye. The simplex method is strongly polynomial for the Markov decision problem with a fixed discount rate. Available at http://www.stanford.edu/∼yyye/simplexmdp1.pdf, 2010.

A Proofs

A.1 Properties of the matrix $\mathsf{B}(\tilde{x})$

We start by explaining why, for every $\tilde{x} \in \mathcal{S}$, the matrix $\mathsf{B}(\tilde{x})$ is an M-matrix. It suffices to show that $\mathsf{B}(\tilde{x})$ has positive diagonal elements and $\mathsf{B}(\tilde{x})$ is strictly (column) diagonally dominant. First, for every $s \in S$, $\mathsf{B}(\tilde{x})_{s,s} = 1 - \gamma \sum_{a\in A} p(s|s,a)\tilde{x}_{s,a} \geq 1 - \gamma \sum_{a\in A} \tilde{x}_{s,a} = 1 - \gamma > 0$, so the diagonal of $\mathsf{B}(\tilde{x})$ is positive. Second, by definition of $\mathsf{B}(\tilde{x})$, all the off-diagonally entries of $\mathsf{B}(\tilde{x})$ are nonpositive, so for every $s \in S$,

$$\sum_{s'\in S} \mathsf{P}(\tilde{x})_{s',s} = \sum_{s'\in S, a\in A} p(s'|s,a)\tilde{x}_{s,a} = \sum_{a\in A} \tilde{x}_{s,a} = 1,$$

i.e.,

$$\mathsf{B}(\tilde{x})^\intercal \mathbb{1} = (I - \gamma \mathsf{P}(\tilde{x}))^\intercal \mathbb{1} = (1-\gamma)\mathbb{1}, \forall\, \tilde{x} \in \mathcal{S}, \tag{11}$$

which implies that $\mathsf{B}(\tilde{x})$ is strictly (column) diagonally dominant. This implies that $\mathsf{B}(\tilde{x})$ is an M-matrix for every $\tilde{x} \in \mathcal{S}$.

Note that we have also shown in passing that the matrix $\mathsf{P}(\tilde{x})^\intercal$ is a (right) stochastic matrix, so the operator norm of $\mathsf{P}(\tilde{x})$ is at most 1. It implies that the operator norm of $\mathsf{B}(\tilde{x})$ is also at most 1.

A.2 Proofs of results in Section 3

Proof of Theorem 2. **1.** To see that $w = W_{\mathcal{D}}(x) \geq \mathbb{1}$ for any $x \in \mathcal{D}$, note that, by definition, for each $s' \in S$,

$$w_{s'} = \sum_{a\in A} x_{s',a} = 1 + \gamma \sum_{s,a} p(s'|s,a)x_{s,a} \geq 0,$$

by the nonnegativity of x.

To see that $\tilde{w} = W_{\mathcal{S}}(\tilde{x}) = \mathsf{B}(\tilde{x})^{-1}\mathbb{1} > 0$, simply note that $\mathsf{B}(\tilde{x})$ being an M-matrix implies that $\mathsf{B}(\tilde{x})^{-1}$ is a nonnegative matrix. This implies that $\tilde{w} = \mathsf{B}(\tilde{x})^{-1}\mathbb{1}$ is nonnegative. If $\tilde{w}_s = 0$ for some $s \in S$, then

$$1 = (\mathsf{B}(\tilde{x})\tilde{w})_s = \sum_{s'\in S} \mathsf{B}(\tilde{x})_{s,s'}\tilde{w}_{s'} = \sum_{s'\neq s} \mathsf{B}(\tilde{x})_{s,s'}\tilde{w}_{s'} = -\gamma \sum_{s'\neq s} \mathsf{P}(\tilde{x})_{s,s'}\tilde{w}_{s'} \leq 0,$$

which is contradictory. Therefore $\tilde{w} > 0$.

2. Fix any $x \in \mathcal{D}$ and let $\tilde{x} = f(x)$. Then for any $s \in S$, $\sum_{a\in A} \tilde{x}_{s,a} = \frac{1}{w_s} \sum_{a\in A} x_{s,a} = 1$, by definition of w. Moreover, $\tilde{x} \geq 0$ because $x \geq 0$ and $w \geq \mathbb{1}$. Hence $\tilde{x} \in \mathcal{S}$, and this shows that $\mathcal{R}(f) \subseteq \mathcal{S}$.

Fix any $\tilde{x} \in \mathcal{S}$ and let $\tilde{w} = W_{\mathcal{S}}(\tilde{x})$ and $x = g(\tilde{x})$. Then $\tilde{w} > 0$ implies that $x \geq 0$. Also, for any $s' \in S$,

$$\begin{aligned}\sum_{a\in A} x_{s',a} - \gamma \sum_{s\in S, a\in A} p(s'|s,a)x_{s,a} &= \sum_{a\in A} \tilde{w}_{s'}\tilde{x}_{s',a} - \gamma \sum_{s\in S, a\in A} (p(s'|s,a)\tilde{x}_{s,a})\tilde{w}_s \\ &= \tilde{w}_{s'} - \gamma(\mathsf{P}(\tilde{x})\tilde{w})_{s'} = (\mathsf{B}(\tilde{x})\tilde{w})_{s'} = 1.\end{aligned}$$

Hence x is feasible for (2). This shows that $\mathcal{R}(g) \subseteq \mathcal{D}$.

3. Fix any $x \in \mathcal{D}$; let $w = W_{\mathcal{D}}(x)$ and $\tilde{x} = f(x)$, i.e., $\tilde{x}_{s,a} = x_{s,a}/w_s$ for all $s \in S$ and $a \in A$. Then for all $s' \in S$,

$$(\mathsf{B}(\tilde{x})w)_{s'} = w_{s'} - \gamma \sum_{s\in S, a\in A} p(s'|s,a)\tilde{x}_{s,a}w_s = \sum_{a\in A} x_{s',a} - \gamma \sum_{s\in S, a\in A} p(s'|s,a)x_{s,a} = 1,$$

by the feasibility of x. Hence $\mathsf{B}(\tilde{x})w = \mathbb{1} = \mathsf{B}(\tilde{x})\tilde{w}$, by definition of $\tilde{w}$, and the nonsingularity of $\mathsf{B}(\tilde{x})$ implies that $W_{\mathcal{D}}(\tilde{x}) = w = \tilde{w} = W_{\mathcal{S}}(f(x))$. This shows that $W_{\mathcal{D}} = W_{\mathcal{S}} \circ f$ over $\mathcal{D}$.
Fix any $\tilde{x} \in \mathcal{S}$; let $\tilde{w} = W_{\mathcal{S}}(\tilde{x}) = \mathsf{B}(\tilde{x})^{-1}\mathbb{1}$ and $x = g(\tilde{x})$, i.e., $x_{s,a} = \tilde{w}_s\tilde{x}_{s,a}$ for all $s \in S$ and $a \in A$. Let $w = W_{\mathcal{D}}(x) = W_{\mathcal{D}}(g(\tilde{x}))$. Then for all $s \in S$,

$$w_s = \sum_{a\in A} x_{s,a} = \tilde{w}_s \sum_{a\in A} \tilde{x}_{s,a} = \tilde{w}_s,$$

since $\tilde{x} \in \mathcal{S}$. This shows that $w = \tilde{w}$, i.e., $W_{\mathcal{S}}(\tilde{x}) = W_{\mathcal{D}}(g(\tilde{x}))$.

4. We first show that $f \circ g = \mathrm{id}_{\mathcal{S}}$. Fix any $\tilde{x} \in \mathcal{S}$; let $x = g(\tilde{x})$ and $\tilde{y} = f(x)$. Then by definitions of f and g, for all $s \in S$ and $a \in A$,

$$\tilde{y}_{s,a} = \frac{x_{s,a}}{W_{\mathcal{D}} \circ g(\tilde{x})_s} = \frac{W_{\mathcal{S}}(\tilde{x})_s\tilde{x}_{s,a}}{W_{\mathcal{S}}(\tilde{x})_s} = \tilde{x}_{s,a}.$$

This shows that $f \circ g(\tilde{x}) = \tilde{y} = \tilde{x}$.
We now show that $g \circ f = \mathrm{id}_{\mathcal{D}}$. Fix any $x \in \mathcal{D}$. Let $\tilde{x} = f(x)$ and $y = g(\tilde{x}) = g \circ f(x)$. Then for all $s \in S$ and $a \in A$,

$$y_{s,a} = W_{\mathcal{S}}(\tilde{x})_s\tilde{x}_{s,a} = W_{\mathcal{D}}(x)_s \cdot \frac{x_{s,a}}{W_{\mathcal{D}}(x)_s} = x_{s,a},$$

i.e., $g \circ f(x) = y = x$.

◀

Proof of Lemma 3. Fix a strategy $\tilde{x} \in \mathcal{S}$. For any $k \geq 1$ and any $s \in S$, let $v_k^{\tilde{x}}(s)$ denote the total expected discounted reward up to the k-th time step under the strategy $\tilde{x}$, if the initial state is s. Let $v_k^{\tilde{x}} \in \mathbb{R}^S$ denote the vector of the values $v_k^{\tilde{x}}(s)$ (for all $s \in S$). Then we have the following recursive relation for all $s \in S$ and $k \geq 1$:

$$\begin{aligned} v_{k+1}^{\tilde{x}}(s) &= \sum_{a\in A} \tilde{x}_{s,a}\left(r_{s,a} + \gamma \sum_{s'\in S} p(s'|s,a)v_k^{\tilde{x}}(s')\right) \\ &= \sum_{a\in A} r_{s,a}\tilde{x}_{s,a} + \sum_{a\in A, s'\in S} (\gamma p(s'|s,a)\tilde{x}_{s,a})v_k^{\tilde{x}}(s') \\ &= \tilde{r}_s + \gamma \sum_{s'\in S} \mathsf{P}(\tilde{x})_{s',s}v_k^{\tilde{x}}(s') \\ &= \tilde{r}_s + \gamma(\mathsf{P}(\tilde{x})^{\mathsf{T}}v_k^{\tilde{x}})_s. \end{aligned}$$

In other words, for all $k \geq 1$,

$$v_{k+1}^{\tilde{x}} = \tilde{r} + \gamma\mathsf{P}(\tilde{x})^{\mathsf{T}}v_k^{\tilde{x}}.$$

Taking $k \to \infty$, we have

$$\mathsf{B}(\tilde{x})^{\mathsf{T}}v^{\tilde{x}} = (I - \gamma\mathsf{P}(\tilde{x})^{\mathsf{T}})v^{\tilde{x}} = \tilde{r}.$$

Again, since $\mathsf{B}(\tilde{x})$ is invertible (see Appendix A.1), $v^{\tilde{x}} = \mathsf{B}(\tilde{x})^{-\intercal}\tilde{r}$ is well defined. This proves the first part of Lemma 3.

For the second part, note that for *any* feasible solution v of the LP (1), the feasibility means that

$$v_s \geq r_{s,a} + \gamma \sum_{s' \in S} p(s'|s,a) v_{s'}, \quad \forall\, s \in S,\ a \in A.$$

Then taking weighted sum using coefficients $\tilde{x}_{s,a}$, we have that for all $s \in S$,

$$v_s \geq \sum_{a \in A} r_{s,a}\tilde{x}_{s,a} + \gamma \sum_{s' \in S, a \in A} p(s'|s,a)\tilde{x}_{s,a} v_{s'} = \tilde{r}_s + \gamma(\mathsf{P}(\tilde{x})^{\intercal} v)_s,$$

implying that

$$\mathsf{B}(\tilde{x})^{\intercal} v \geq \tilde{r} = \mathsf{B}(\tilde{x})^{\intercal} v^{\tilde{x}}, \quad \text{i.e., } \mathsf{B}(\tilde{x})^{\intercal}(v - v^{\tilde{x}}) \geq 0.$$

Since $\mathsf{B}(\tilde{x})^{-1}$ is a nonnegative matrix, $\mathsf{B}(\tilde{x})^{-\intercal} = (\mathsf{B}(\tilde{x})^{-1})^{\intercal}$ is also a nonnegative matrix. Hence

$$v - v^{\tilde{x}} = \mathsf{B}(\tilde{x})^{-\intercal}\mathsf{B}(\tilde{x})^{\intercal}(v - v^{\tilde{x}}) \geq 0.$$

In particular, since v^* is also a feasible solution of (1), we have $v^{\tilde{x}} \leq v^*$. This proves the second claim of Lemma 3.

To prove the final claim for any $x \in \mathcal{D}$, simply note that

$$\mathbb{1}^{\intercal} v^{\tilde{x}} = \mathbb{1}^{\intercal}\mathsf{B}(\tilde{x})^{-\intercal}\tilde{r} = (\mathsf{B}(\tilde{x})^{-1}\mathbb{1})^{\intercal}\tilde{r} = \sum_{s \in S} \tilde{r}_s W_{\mathcal{S}}(\tilde{x})_s = \sum_{s \in S, a \in A} r_{s,a} x_{s,a}.$$

◀

Proof of Theorem 4. By Lemma 3, $v^{\tilde{x}} \leq v^*$, so

$$\|v^{\tilde{x}} - v^*\| \leq \sum_{s \in S} |v^{\tilde{x}}_s - v^*_s| = \sum_{s \in S} v^*_s - v^{\tilde{x}}_s = \sum_{s \in S} v^*_s - \sum_{s \in S, a \in A} r_{s,a} x_{s,a} = \nu - \sum_{s \in S, a \in A} r_{s,a} x_{s,a}.$$

This proves the first claim.

For the second claim, note that by Lemma 3, $\sum_{s \in S, a \in A} r_{s,a} x_{s,a} \geq \sum_{s \in S, a \in A} r_{s,a} x'_{s,a}$ implies that

$$\sum_{s \in S} v^{\tilde{x}}_s \geq \sum_{s \in S} v^{\tilde{x}'}_s.$$

Therefore, using $v^* \geq v^{\tilde{x}}$ and $v^* \geq v^{\tilde{x}'}$,

$$\sum_{s \in S} |v^*_s - v^{\tilde{x}}_s| = \sum_{s \in S} v^*_s - \sum_{s \in S} v^{\tilde{x}}_s \leq \sum_{s \in S} v^*_s - \sum_{s \in S} v^{\tilde{x}'}_s. = \sum_{s \in S} |v^*_s - v^{\tilde{x}'}_s|.$$

This proves the second cliam, equation (4). ◀

A.3 Proofs of results in Section 4

Proof of Proposition 5. Fix any feasible solution v of (1) and any $\tilde{x} \in \mathcal{S}$. Then by Lemma 3,

$$\sum_{s \in S} v^{\tilde{x}}_s = \sum_{s \in S, a \in A} r_{s,a} x_{s,a}, \quad \text{where } x = g(\tilde{x}).$$

Then note that, by definition of x,

$$\begin{aligned}\mu(v,\tilde{x}) &= \frac{1}{|S||A|}\sum_{s\in S,a\in A} x_{s,a}\left(v_s - r_{s,\,a} - \gamma\sum_{s'\in S} p(s'|s,a)v_{s'}\right)\\ &= \frac{1}{|S||A|}\left(\sum_{s\in S,a\in A} v_s x_{s,a} - \sum_{s'\in S}\left(\gamma\sum_{s\in S,a\in A} p(s'|s,a)x_{s,a}\right)v_{s'} - \sum_{s\in S} v_s^{\tilde{x}}\right)\\ &= \frac{1}{|S||A|}\left(\sum_{s\in S,a\in A} v_s x_{s,a} - \sum_{s'\in S}\left(\sum_{a\in A} x_{s',a} - 1\right)v_{s'} - \sum_{s\in S} v_s^{\tilde{x}}\right)\\ &= \frac{1}{|S||A|}\left(\sum_{s\in S} v_s - \sum_{s\in S} v_s^{\tilde{x}}\right).\end{aligned}$$

This proves the first claim.

To prove the first inequality in (6), note that

$$\|v^{\tilde{x}} - v^*\| \le \sum_{s\in S} v_s^* - \sum_{s\in S} v_s^{\tilde{x}} \le \sum_{s\in S} v_s - \sum_{s\in S} v_s^{\tilde{x}} = |S||A|\mu(v,\tilde{x}).$$

To prove the second inequality in (6), note that

$$\nu = \sum_{s\in S} v_s^* \ge \sum_{s\in S} v_s^{\tilde{x}},$$

so

$$\sum_{s\in S} v_s - \nu \le \sum_{s\in S} v_s - \sum_{s\in S} v_s^{\tilde{x}} = |S||A|\mu(v,\tilde{x}).$$

Hence (6) is true. ◀

Before we proceed to prove Theorem 6, we state a few useful facts, which are standard from the theory of interior point methods [8].

▶ **Observation 2.** $(\Delta v, \Delta z, \Delta\tilde{x})$ solves (8) if and only if $(\Delta v, \Delta z, \Delta x)$, where $\Delta x_{s,a} = w_s\Delta\tilde{x}_{s,a}$ for all $s \in S$ and $a \in A$, satisfies the equation

$$\begin{cases}\Delta z_{s,a} \;=\; \Delta v_s - \gamma\sum_{s'\in S} p(s'|s,a)\Delta v_{s'}, & \forall\, s\in S,\ a\in A,\\ w_{s'}\sum_{a\in A}\Delta x_{s',a} - \gamma\sum_{s,a} w_s p(s'|s,a)\Delta x_{s,a} \;=\; 0, & \forall\, s'\in S,\\ z_{s,a}\Delta x_{s,a} + x_{s,a}\Delta z_{s,a} \;=\; \sigma\mu(v,\tilde{x}) - x_{s,a}z_{s,a}, & \forall\, s\in S,\ a\in A.\end{cases} \tag{12}$$

Moreover, the first and second equations in (12) implies that $\Delta x^\intercal\Delta z = 0$.

Proof of Theorem 6. We first prove the feasibility of the initial point. It is easy to see that $\tilde{x}^{(0)} \in \mathcal{S}$: for any $s \in S$, $\sum_{a\in A}\tilde{x}^{(0)}_{s,a} = \sum_{a\in A}\frac{1}{|A|} = 1$. Since $\tilde{x}^{(0)} > 0$, we have that $W_{\mathcal{S}}(\tilde{x}^{(0)}) > 0$ (recalling that the matrix $\mathsf{B}(\tilde{x}^{(0)})$ is an M-matrix), so $g(\tilde{x}^{(0)}) > 0$ too. Next, note that

$$\frac{\max_{s,a}\{|r_{s,a}|\}+2}{1-\gamma}\cdot\left(1-\gamma\sum_{s'\in S} p(s'|s,a)\right) = \frac{\max_{s,a}\{|r_{s,a}|\}+2}{1-\gamma}\cdot(1-\gamma) \ge r_{s,a}+2$$

for all $s \in S$ and $a \in A$. Hence $v^{(0)}$ is feasible for (1).

In the following, we let $v = v^{(k)}$ and $\tilde{x} = \tilde{x}^{(k)}$ for a fixed k (and we drop the superscript (k) when necessary).

For $k \geq 1$, note that by Observation 2, $x + \alpha\Delta x$ is feasible for (1) for all $\alpha \in (0, 1]$, and $\tilde{x}(\alpha)_{s,a} = \frac{x_{s,a}+\alpha\Delta x_{s,a}}{\sum_{a' \in A} \tilde{x}_{s,a'}+\alpha\Delta\tilde{x}_{s,a}} = f(x + \alpha\Delta x)$. Hence $\tilde{x}(\alpha) \in \mathcal{S}$. On the other hand, the requirement that $z(\alpha) > 0$ implies that $v(\alpha)$ must be feasible for (1). Hence we have that $\tilde{x}^{(k)} \in \mathcal{S}$ and $v^{(k)}$ is feasible for (1) for all k.

Now we prove the second claim. Note that $\mu(v, \tilde{x}) = \frac{1}{|S||A|} x^\intercal z$, where $x = g(\tilde{x})$; also,

$$\begin{aligned}
\mu(v(\alpha), \tilde{x}(\alpha)) &= \frac{1}{|S||A|}(x^\intercal z + \alpha x^\intercal \Delta z + \alpha z^\intercal \Delta x) \\
&= \mu(v, \tilde{x}) + \alpha \left(\sigma\mu(v, \tilde{x}) - \frac{\alpha}{|S||A|} \sum_{s \in S, a \in A} x_{s,a} z_{s,a} \right) \\
&= \mu(v, \tilde{x})(1 - \alpha(1 - \sigma)).
\end{aligned}$$

where $\Delta x_{s,a} = w_s \Delta\tilde{x}_{s,a}$ for all $s \in S$ and $a \in A$. ◀

Next, to prove Theorem 8, we first note that the parameter ξ in Algorithm 4.1 is bounded away from zero for all the instances.

Proof of Lemma 7. We drop the superscript (0) for notational convenience and recall $w = \mathsf{B}(\tilde{x})^{-1}\mathbb{1}$. Let $x = g(\tilde{x})$. For all $s \in S$ and $a \in A$,

$$z_{s,a}x_{s,a} = \frac{1}{|A|} w_s \left(\max_{s' \in S, a' \in A} \{|r_{s',a'}|\} - r_{s,a} + 2 \right) \geq \frac{1}{|A|} \left(2 - \max_{s' \in S, a' \in A} \{|r_{s',a'}|\} \right) \geq \frac{1}{|A|},$$

where we have used the fact that $w \geq \mathbb{1}$ by items 1 and 4 of Theorem 2, and also Assumption 1 to bound $|r_{s,a}|$'s. In particular,

$$\begin{aligned}
\mu(v, \tilde{x}) &= \frac{1}{|S||A|} \sum_{s,a} z_{s,a}x_{s,a} \\
&= \frac{1}{|S||A|^2} \sum_{s,a} w_s \left(\max_{s' \in S, a' \in A} \{|r_{s',a'}|\} - r_{s,a} + 2 \right) \\
&\leq \frac{1}{|S||A|^2} \left(\sum_s w_s \right) \left(\sum_{s,a} \max_{s' \in S, a' \in A} \{|r_{s',a'}|\} - r_{s,a} + 2 \right) \\
&\leq \frac{1}{|S||A|^2} \left(\sum_s w_s \right) \left(2 \sum_{s,a} \max_{s' \in S, a' \in A} \{|r_{s',a'}|\} + 2 \right) \\
&\leq \frac{1}{|S||A|^2} \left(\sum_s w_s \right) \cdot (4|S||A|),
\end{aligned}$$

where the last inequality follows from Assumption 1. Now note that by Theorem 2,

$$\mathbb{1}^\intercal w = \mathbb{1}^\intercal \mathsf{B}(\tilde{x})^{-1}\mathbb{1} = (\mathsf{B}(\tilde{x})^{-\intercal}\mathbb{1})^\intercal \mathbb{1} = \frac{|S|}{1 - \gamma},$$

where the last equality follows from (11). Hence $\mu(v, \tilde{x}) \leq \frac{4|S|}{(1-\gamma)|A|}$, and

$$\xi = \min_{s \in S, a \in A} \frac{z_{s,a}x_{s,a}}{\mu} \geq \frac{(1 - \gamma)}{4|S|} \min_{s \in S, a \in A} \left\{ w_s \left(\max_{s' \in S, a' \in A} \{|r_{s',a'}|\} - r_{s,a} + 1 \right) \right\} \geq \frac{1 - \gamma}{4|S|}.$$

◀

Now we are ready to explain the validity of Theorem 8. The reasoning relies on the fact that Algorithm 4.1 results from modifying a classical interior point method, known as long-step path following method, by using the one-to-one correspondence between the feasible region of the dual (2) and the mixed strategies. More importantly, the complexity result follows because we can ensure that the iterates of Algorithm 8 *always* lies in a so-called $\mathcal{N}_{-\infty}\left(\frac{1-\gamma}{4|S|}\right)$ neighborhood of the central path *for all the instances of discounted MDPs*. This important feature means that the complexity of Algorithm 4.1 depends only on problem size, the user-defined parameters $\sigma_{\max}, \sigma_{\min}$ and the discount rate γ.

Proof of Theorem 8. By Observation 2 and the second part of the proof of Theorem 6, Algorithm 4.1 is equivalent to the long step path following method [8, P. 96], whose complexity result [8, Thm. 5.12] together with Lemma 7 implies the desired result. ◀