MATHEMATICS OF OPERATIONS RESEARCH
Vol. 32, No. 2, May 2007, pp. 365–373
ISSN 0364-765X | EISSN 1526-5471 | 07 | 3202 | 0365

DOI 10.1287/moor.1060.0238

On the Speed of Convergence of Value Iteration on Stochastic Shortest-Path Problems

Blai Bonet
Departamento de Computación, Universidad Simón Bolívar, Caracas 89000, Venezuela,
bonet@ldc.usb.ve, http://www.ldc.usb.ve/~bonet

We establish a bound on the convergence time of the value iteration algorithm on stochastic shortest-path problems. The bound, which applies for *admissible* initial vectors as, for example, $J \equiv 0$, implies a polynomial-time convergence of value iteration for all problems with polynomially bounded $\|J^*\|/\underline{g}$. This result gives a partial answer to the open problem of bounding the convergence time of value iteration on arbitrary initial vectors. The proof is obtained by analyzing a stochastic process associated with the shortest-path problem.

Key words: Markov decision processes; stochastic shortest-path problems; value iteration
MSC2000 subject classification: Primary: 90C40, 68Q25; secondary: 68T20, 60J20
OR/MS subject classification: Primary: dynamic programming; secondary: Markov finite state
History: Received October 3, 2005; revised August 28, 2006.

1. Introduction. Markov decision processes (MDPs) have become the standard formalism for modeling sequential decision tasks with uncertainty and complete information, Bellman [2], Puterman [13], Bertsekas [3]. In this setting, the environment is modeled as a stochastic system in discrete time whose evolution depends on the decisions (controls) taken by an abstract agent. Each time the agent executes a control, the system makes a transition from the current state to a next state, and a cost (respectively, reward) is generated. The goal is to obtain a decision function or policy that minimizes (respectively, maximizes) the expected cost (respectively, reward).

MDPs can be divided into two groups: discounted and undiscounted models. In the former group, future costs are diminished at a geometric rate using a *discount* factor. These discounts, which have a natural interpretation within the context of an inflationary economy, often appear as unpleasant artifacts used to facilitate the mathematical analysis of the model. Undiscounted models, on the other hand, provide a closer representation of the world, yet their complexity have limited rigorous analyses of the model and algorithms.

In this paper, we focus on formal issues about the class of infinite-horizon undiscounted models with positive costs, which are also known as stochastic shortest-path problems (SSPs); see Eaton and Zadeh [8], Bertsekas and Tsitsiklis [5], Bertsekas [3], Puterman [13]. In particular, we establish a bound on the convergence time of the value iteration algorithm on stochastic shortest-path problems when the initial cost vector is *admissible*, i.e., $0 \leq J \leq J^*$ where J^* is the optimal cost vector.

The bound is a function of the ratio $\|J^*\|/\underline{g}$ where $\underline{g}$ refers to the minimum positive cost of the problem. An important consequence of this result is that the value iteration algorithm converges in polynomial time over the class of models in which $\|J^*\|/\underline{g}$ is polynomially bounded.

This is a fairly general result about the convergence time of value iteration on stochastic shortest-path problems. Weaker related results have a narrow scope, as, for example, when all stationary policies of the problem are proper, or when there exist consistently improving policies. In the former case, it can be shown that the dynamic programming operator (see below) is a contraction with respect to a weighted supremum norm, and so is straightforward to bound the speed of convergence of value iteration; see Bertsekas and Tsitsiklis [4], Tseng [14], and Bertsekas [3, Vol. 2, p. 128]. In the latter case, labeling methods like variants of Dijkstra's algorithm for deterministic shortest-path problems can be used to solve stochastic problems; such algorithms perform at most a number of iterations equal to the number of states; see Bertsekas [3, Vol. 2, p. 90] and references therein.

The paper is self-contained, yet preferred readers are those familiar with the theory of MDPs, SSPs, and the value iteration algorithm. The main result of this paper is obtained by analyzing a stochastic process associated with the SSP. Excellent references for the theory of optimal control in discrete time are the books of Bertsekas [3] and Puterman [13]; we use a notation similar to the one found in the former.

The paper is organized as follows. Section 2 contains standard definitions, known facts, and notation. Section 3 contains the assumptions for the analysis and a discussion on them. Section 4 contains the main results and the proofs. Because the bound on the convergence time is a function of the quantity $\|J^*\|/\underline{g}$, §5 provides a simple bound on such quantity when a proper policy is known. The paper concludes with a discussion that includes a summary and some open problems.

2. Definitions, known facts, and notation. Stochastic shortest-path problems provide a model for sequential decision tasks with uncertainty and complete information. The setting characterizes a physical system that evolves in discrete time and is controlled by an agent. The system dynamics are governed by probabilistic transition functions that map states and controls into states. Additionally, the agent incurs a cost at every time step that depends on the current state of the system and the chosen control. Formally, a stochastic shortest-path problem is characterized by:

S1. a finite state space S,

S2. a finite set of controls $U(s)$ for each state $s \in S$,

S3. transition probabilities $p(s, u, s')$ for all $u \in U(s)$ that are equal to the probability of the next state being s' after applying control u in state s,

S4. positive (and finite) costs $g(s, u)$ for each $u \in U(s)$ and $s \in S$, and

S5. a *unique target* (or goal) state $t \in S$ such that $p(t, u, t) = 1$ and $g(t, u) = 0$ for all $u \in U(t)$.

The two conditions on the target state are often referred to as that t is an "absorbing" state. Observe that the target state is the only state in which costs equal zero; other costs are assumed to be positive. The assumption that there is a unique absorbing state facilitates the analysis, yet is not restrictive because multiple target states can be collapsed into a single one.

Given such a model, the main computational task is to find a control strategy (also known as policy) that minimizes the total expected cost over the *infinite-horizon time* setting.

In a general form, a policy π is an infinite sequence $(\mu_0, \mu_1, \dots)$ of decision functions where μ_k maps states to controls such that the agent applies $\mu_k(s)$ in state $x_k = s$ at time k, the only restriction being $\mu_k(s) \in U(s)$ for all $s \in S$. In this case, π is called a *nonstationary policy*, i.e., one whose behavior depends on time. A stationary policy, on the other hand, is one of the form $\pi = (\mu, \mu, \dots)$, which is simply denoted by μ.

The cost associated with policy π when the system "starts" from state s is defined as the expected cumulative cost incurred by π over the infinite horizon; i.e.,

$$J_\pi(s) \doteq E_\pi\left[\sum_{k=0}^{\infty} g(X_k, \mu_k(X_k)) \,\middle|\, X_0 = s\right]$$

where the X_ks are random variables that represent the state of the system at time k, distributed according to the distribution P_π induced by the transition probabilities and π, and E_π is the expectation with respect to P_π. Such a function that maps states into reals is called a value function or value vector.

The SSP task is formally defined as *finding an optimal policy* π^* satisfying

$$J^*(s) \doteq J_{\pi^*}(s) \le J_\pi(s), \quad \forall s \in S,$$

and for every other policy π. Although there could be no, or more than one, optimal policy, the optimal cost vector J^* (if defined) is unique. The existence of an optimal policy π^* and how to compute it are nontrivial mathematical problems.

Under very mild assumptions (Bertsekas [3]), the optimal cost vector J^* exists and is the unique solution to *Bellman's optimality* equations:

$$J^*(s) = \begin{cases} 0 & s = t, \\ \min\limits_{u \in U(s)} g(s, u) + \sum\limits_{s' \in S} p(s, u, s') J^*(s') & s \ne t. \end{cases} \tag{1}$$

Once J^* is known, an optimal policy can be recovered by choosing controls *greedily* with respect to J^*. That is, the *stationary* policy μ^* defined as

$$\mu^*(s) \doteq \arg\min_{u \in U(s)} \left\{ g(s, u) + \sum_{s' \in S} p(s, u, s') J^*(s') \right\}, \quad s \in S,$$

where ties are broken arbitrarily, is an optimal policy for the model S1–S5. The most popular (and often effective) method for solving (1) consists of treating such an equation as an assignment. Thus, from an initial iterate J_0 satisfying $J_0(t) = 0$ (typically $J_0 \equiv 0$), one computes a sequence of iterates $\{J_k\}_{k \ge 0}$ as

$$J_{k+1}(s) \doteq \min_{u \in U(s)} \left\{ g(s, u) + \sum_{s' \in S} p(s, u, s') J_k(s') \right\}, \quad s \in S.$$

This method is known as the algorithm of value iteration. The same conditions that guarantee the existence of J^* also guarantee that value iteration converges to J^*; i.e., $J_k(s) \to J^*(s)$ for all $s \in S$. An important open question, for many years now, is to bound the speed of convergence of value iteration on stochastic shortest-path problems; that is, to bound the number of iterates J_k.

Because in general value iteration takes an infinite number of iterations to converge, more precision is required when talking about speed of convergence. In particular, we are interested in upper bounding the number of iterations k needed to bring the *residual*, defined as $\max_{s\in S} |J^*(s) - J_k(s)|$, to a quantity no greater than a fixed parameter $\epsilon > 0$.

A class of models $\mathcal{M}$ is said to have the quantity $\|J^*\|/\underline{g}$ polynomially bounded from above, or just that $\mathcal{M}$ has polynomially bounded $\|J^*\|/\underline{g}$, if there is a polynomial p such that $\|J^*_M\|/\underline{g}_M \le p(|M|)$ for each $M \in \mathcal{M}$, where J^*_M and $\underline{g}_M$ are the optimal value function and minimum positive cost for model M, respectively, and $|M|$ is the size in bits of a suitable encoding of M.

We are now in a position to precisely describe one of our main results. Let $\mathcal{M}$ be a class of models with polynomially bounded $\|J^*\|/\underline{g}$. Then, the value iteration algorithm converges in polynomial time for any $M \in \mathcal{M}$ when started from an initial iterate J_0 satisfying $0 \le J_0 \le J^*_M$, e.g., for $J_0 \equiv 0$. A value function that satisfies such condition is called an admissible value function or lower bound, a term borrowed from the field of heuristic search in artificial intelligence (Hart et al. [9], Pearl [12]). In control theory, such functions have proven to be important in the context of asynchronous DP algorithms, e.g., real-time dynamic programming of Barto et al. [1].

Observe that the convergence time of value iteration must depend on the ratio $\|J^*\|/\underline{g}$ and not only on $\|J^*\|$, because if all costs in a problem are uniformly scaled, the value $\|J^*\|$ changes while the behavior of value iteration does not.

A more convenient description of value iteration is obtained by considering the dynamic programming operator T, which maps value functions into value functions:

$$(TJ)(s) \doteq \min_{u\in U(s)} \left\{ g(s,u) + \sum_{s'\in S} p(s,u,s')J(s') \right\}.$$

The value iteration algorithm is then described by $J_k = TJ_{k-1} = T^k J_0$, where T^k denotes the kth-fold composition of T with itself. The statement about convergence of value iteration then becomes $T^k J_0 \to J^*$ as $k \to \infty$.

Besides admissible value functions, we will also deal with value functions that satisfy $0 \le J \le TJ$. These value functions have been called uniformly improving and/or consistent, where the latter term is also borrowed from heuristic search. In heuristic search, the value function is interpreted as an estimate of the distances from states to the goal state. Thus, an admissible function refers to estimates that are lower bounds on the costs of the minimum-cost paths. Consistency, on the other hand, amounts to a type of triangular inequality that such estimates must satisfy. Indeed, if s' is the result of applying control u in s (in a deterministic model), then $J \le TJ$ implies $J(s) \le g(s,u) + J(s')$, which intuitively says that the estimate of the distance between s and the goal is less than or equal to the distance between s and s' plus the estimate of the distance between s' and the goal. In stochastic models, consistency also refers to a type of triangular inequality, but in expectation because it implies $J(s) \le g(s,u) + E[J(X_{k+1}) \mid X_k = s, u]$. The class of all consistent value functions for model M is denoted by $\mathcal{J}_M$ or just $\mathcal{J}$ if M is clear from context.

Finally, we make use of two norms: the sup (or L_∞) norm defined as $\|J\| \doteq \sup_{s\in S} |J(s)|$ and the L_1 norm defined as $\|J\|_{L_1} \doteq \sum_{s\in S} |J(s)|$. Some known properties are the following:

(a) The zero-constant value function is admissible and consistent,
(b) Every consistent value function is admissible,
(c) If J is admissible, $J(t) = 0$ (because $J^*(t) = 0$),
(d) (Monotonicity) $TJ \le TJ'$ if $J \le J'$, and
(e) (Nonexpansion) $\|TJ - TJ'\| \le \|J - J'\|$ for all value functions $J, J' \in \mathbb{R}^S$.

3. Assumptions. As is standard in the proofs of the existence and uniqueness of solutions for SSPs as well as in the proof of convergence of value iteration (Bertsekas and Tsitsiklis [5], Bertsekas [3]), we assume the existence of a proper policy. A proper policy is a stationary policy that *eventually* achieves the goal from any state; that is, every trajectory[1] generated by the policy reaches the target state (and remains in it) after some time. In symbols, a stationary policy μ is proper if and only if

$$\min_{s\in S} P_\mu(\exists_{n\ge 0} \forall_{m\ge n} \{X_m = t\} \mid X_0 = s) = 1.$$

[1] That is, those that do not satisfy the property make up an event of probability zero.

To see that this equation captures the intuition, suppose the equality is indeed a strict inequality and consider the complementary event that would have positive probability:

$$\min_{s\in S} P_\mu(\forall_{n\geq 0}\exists_{m\geq n}\{X_m \neq t\} \mid X_0 = s) > 0,$$

that is, there are trajectories starting at s such that for any time index n, there exists $m \geq n$ such that the state at time m is not target. Observe that being a proper policy does not mean that there exists a *uniform* bound n such that every trajectory hits the target before time n. Indeed, for most of the interesting problems such a bound does not exist.

Let M be a model and μ an improper stationary policy for M. Because μ is improper, there exists a state s and a set of trajectories starting at s, with positive probability $p > 0$, that do not reach the target state. Thus, μ incurs an infinite cost over those trajectories, because all costs are positive, and so $J_\mu(s) \geq p \cdot \infty = \infty$. That is, every improper policy incurs in an infinite cost for at least one state. Therefore, if M does not have a proper policy, and because $\|TJ - J\|$ is finite for any J, then $T^k J \leq \infty$ for all k and value iteration cannot converge over M.[2] Our first assumption is then

Assumption 3.1. *The model* S1–S5 *has a proper policy.*

A necessary and sufficient condition for μ being a proper policy is that $\max_{s\in S} P_\mu(X_{|S|} \neq t \mid X_0 = s) < 1$ (Bertsekas [3]); the quantity in the left-hand side of the inequality is denoted by ρ_μ. It is not hard to show that a proper policy μ incurs in finite expected costs for every state; that is, $J_\mu(s) < \infty$ for all $s \in S$.

The second assumption needed refers to the transition dynamics of the model. It is a technicality that can be assumed to hold without loss of generality, as shown below.

Assumption 3.2. *There are no self-loops in the state transition diagram except for t; i.e., $p(s,u,s) = 0$ for all $s \neq t$ and $u \in U(s)$. Note that, t being absorbing, $p(t,u,t) = 1$ for all $u \in U(t)$.*

As shown by Bertsekas [3, Vol. 2, p. 89], this assumption is not restrictive because a model M can be "converted" into an equivalent model $\widetilde{M}$ that satisfies Assumption 3.2. Indeed, $\widetilde{M}$ is identical to M except that transition probabilities and costs are modified as follows:

$$\tilde{p}(s,u,s') \doteq \begin{cases} 1 & \text{if } s = s' = t, \\ 0 & \text{if } s = s' \quad \text{and} \quad s \neq t, \\ p(s,u,s')/(1-p(s,u,s)) & \text{otherwise,} \end{cases}$$

$$\tilde{g}(s,u) \doteq g(s,u)/(1-p(s,u,s)) \quad \text{for } s \neq t.$$

By equivalent models, it is meant that the set of valid policies for each model coincides as well as the cost functions associated with each policy. Hence, the optimal value functions for both models are identical as well as the sets of optimal policies.

4. Convergence time of value iteration. In this section we present the results and their proofs. Throughout the rest of the section, consider a model M according to S1–S5 that satisfies Assumptions 3.1 and 3.2, and let $\underline{g} \doteq \min\{g(s,u)\colon s \neq t,\ u \in U(s)\}$ be the minimum positive cost.

The first result shows a geometric rate of convergence of value iteration when the initial iterate J is consistent and $\|TJ - J\|$ is sufficiently small.

Theorem 4.1. *Let J be a consistent function. If there are constants $K \geq 0$ and $0 \leq \alpha < 1$ such that*

$$(1+|S|)\|J^*\| + (1+K)^2\|TJ - J\| \leq \alpha(1+K)\underline{g}, \tag{2}$$

then

$$\|J^* - T^{nK}J\| \leq \alpha^n \|J^* - J\|.$$

Remarks. Typically $(1+|S|)\|J^*\|$ is fairly large with respect to $\underline{g}$. Thus, for (2) to hold, the constant K must be relatively large, and so $\|TJ - J\|$ must be small enough to take care of the $(1+K)^2$ term.

[2] Under weaker conditions on the model, such as having nonnegative costs instead of strictly positive costs, stronger assumptions would be required; e.g., that every improper policy incurs an infinite cost for at least one state (Bertsekas [3]).

Once the conditions of Theorem 4.1 are achieved for some J, $T^{nK}J$ converges to J^* at a geometric rate. This result is used as follows. From an initial consistent function J, find integer m such that T^mJ satisfies the condition (2) of Theorem 4.1. Then, using the geometric convergence, find n such that $\|J^* - T^{m+nK}J\| \leq \epsilon$. By upper bounding the number of iterations m and n (for appropriately chosen K and α), we obtain the main results of the paper:

THEOREM 4.2. *Let $J \in \mathcal{J}$ be a consistent value function and $\epsilon > 0$. Then, $\|J^* - T^nJ\| \leq \epsilon$ for all $n \geq O(\|J^*\|^2|S|^2/\underline{g}^2 + (\log\|J^*\| + |\log\epsilon|)\|J^*\||S|/\underline{g})$.*

COROLLARY 4.1. *Let J be an admissible value function. Then, the value iteration algorithm achieves a residual of $\|J^* - T^nJ\| \leq \epsilon$ in at most $O(\|J^*\|^2|S|^2/\underline{g}^2 + (\log\|J^*\| + |\log\epsilon|)\|J^*\||S|/\underline{g})$ iterations.*

PROOF. Let $J_0 \equiv 0$. Because $J_0 \leq J$, then, by the monotonicity of T, $0 \leq T^kJ_0 \leq T^kJ \leq J^*$. Finish with $\|J^* - T^nJ\| \leq \|J^* - T^nJ_0\|$ and the result of the theorem applied to J_0. □

COROLLARY 4.2. *Let $\mathcal{M}$ be a class of models with polynomially bounded $\|J^*\|/\underline{g}$, and $M \in \mathcal{M}$. Value iteration converges in polynomial time over M when started from an admissible value function.*

PROOF. Because each iteration takes polynomial time, we need to show that the number of iterations required to bring the residual to at most ϵ is bounded by a polynomial function of $|M| + |\log\epsilon|$; i.e., we need to show that the three quantities $\|J^*\|^2|S|^2/\underline{g}^2$, $\|J^*\||S|\log\|J^*\|/\underline{g}$, and $\|J^*\||S||\log\epsilon|/\underline{g}$ are polynomially bounded. Because $|S| = O(|M|)$, it is direct that the first and last quantities are such bounded. For the second, the assumption and $|\log\underline{g}| = O(|M|)$ implies $\log\|J^*\|$ is polynomially bounded. □

4.1. Proofs. Consider the Markov process $(X_k, U_k, J_k)_{0\leq k\leq N}$ generated by a pair (s, J) of initial state and value function, where N is a positive integer whose value shall be determined later. The process evolves according to the following rules:

(a) At time 0: $X_0 = s$ and $J_0 = T^{N+1}J$,
(b) At time $k \geq 0$: choose a control $U_k \doteq \arg\min_{u\in U(X_k)}\{g(X_k, u) + \sum_{s\in S} p(X_k, u, s)T^{N-k-1}J(s)\}$,
(c) Jump to state $X_{k+1} = s'$ with probability $p(X_k, U_k, s')$, and
(d) Compute vector J_{k+1} defined as $J_{k+1}(s) \doteq [\![s = X_k]\!](T^{N-k}J)(s) + [\![s \neq X_k]\!]J_k(s)$.

The notation $[\![\varphi]\!]$ refers to one or zero whether the formula φ is true or false. It is easy to check that the process satisfies the Markov property and that its dynamics are determined by the pair (s, J). Let $\hat{P}_{s,J}$ be the probability measure associated with the process and $\hat{E}_{s,J}$ the expectation with respect to it.

The motivation for this process comes from studying the equation of the residual $\|J^* - T^nJ\|$. Indeed, for $s = s_0 \in S$,

$$\begin{aligned}
|J^*(s) - (T^nJ)(s)| &= J^*(s_0) - (T^nJ)(s_0) \\
&= \min_{u\in U(s_0)}\left\{g(s_0, u) + \sum_{s_1\in S} p(s_0, u, s_1)J^*(s_1)\right\} - \min_{u\in U(s_0)}\left\{g(s_0, u) + \sum_{s_1\in S} p(s_0, u, s_1)(T^{n-1}J)(s_1)\right\} \\
&\leq \sum_{s_1\in S} p(s_0, u_0, s_1)|J^*(s_1) - (T^{n-1}J)(s_1)| \\
&= \sum_{s_1\neq t} p(s_0, u_0, s_1)|J^*(s_1) - (T^{n-1}J)(s_1)|,
\end{aligned}$$

where $u_0 \in U(s_0)$ is the greedy control with respect to $T^{n-1}J$. Repeat this calculation n times to obtain

$$|J^*(s) - (T^nJ)(s)| \leq \sum_{s_1\ldots s_n\neq t}\prod_{k=0}^{n-1} p(s_k, u_k, s_{k+1})(J^*(s_n) - J(s_n)) \leq \|J^* - J\| \sum_{s_1\ldots s_n\neq t}\prod_{k=0}^{n-1} p(s_k, u_k, s_{k+1}). \quad (3)$$

The last quantity in the right-hand side is the probability of $\{X_n \neq t\}$ in the stochastic process generated by the pair (s, J). Thus, for $n = N$, $|J^*(s) - (T^n)(s)| \leq \|J^* - J\|\hat{P}_{s,J}(X_n \neq t)$. Some properties of the process are given in the following lemmas.

LEMMA 4.1. *If $J \in \mathcal{J}$, then $J_k \geq J_{k+1}$ and $J_k \geq T^{N-k+1}J$ for $k \geq 0$.*

PROOF. Both results are shown simultaneously with induction on k. The base cases are direct from the definition of J_0. Assume the validity for k, then

$$\begin{aligned} J_{k+1}(s) &= [\![s = X_k]\!](T^{N-k}J)(s) + [\![s \neq X_k]\!]J_k(s) \\ &\geq [\![s = X_k]\!](T^{N-k}J)(s) + [\![s \neq X_k]\!](T^{N-k+1}J)(s) \\ &\geq [\![s = X_k]\!](T^{N-k}J)(s) + [\![s \neq X_k]\!](T^{N-k}J)(s) \\ &= (T^{N-k}J)(s), \end{aligned}$$

the first inequality by inductive hypothesis and the second since $J \in \mathcal{J}$. For the second result,

$$\begin{aligned} J_{k+1}(s) &= [\![s = X_k]\!](T^{N-k}J)(s) + [\![s \neq X_k]\!]J_k(s) \\ &\leq [\![s = X_k]\!](T^{N-k+1}J)(s) + [\![s \neq X_k]\!]J_k(s) \\ &= J_k(s) - [\![s = X_k]\!][J_k(s) - (T^{N-k+1}J)(s)] \\ &\leq J_k(s), \end{aligned}$$

the first inequality because $J \in \mathcal{J}$ and the second by inductive hypothesis. □

LEMMA 4.2. *If $J \in \mathcal{J}$, then for all $0 \leq k \leq N$ and $n \geq 0$*

$$\max_{s \in S} \hat{P}_{s, T^n J}(X_k \neq t) \leq \frac{(1 + |S|)\|J^*\| + (1 + N)(1 + k)\|TJ - J\|}{(1 + k)\,\underline{g}}. \tag{4}$$

PROOF. Consider the sequence $\{W_k\}_{k \geq 0}$ defined as

$$\begin{aligned} W_0 &\doteq \|J_0 - J\|_{L_1} - J_0(X_0), \\ W_{k+1} &\doteq \|J_{k+1} - J\|_{L_1} - J_k(X_{k+1}). \end{aligned}$$

Let $\mathcal{F}_k \doteq \{X_0, U_0, J_0, \ldots, X_k, U_k, J_k\}$ be the random history of the process until time k. Then,

$$\begin{aligned} W_{k+1} &= \sum_{s \in S} \big[J_{k+1}(s) - J(s)\big] - J_k(X_{k+1}) \\ &= \sum_{s \neq X_k} \big[J_{k+1}(s) - J(s)\big] + J_{k+1}(X_k) - J(X_k) - J_k(X_{k+1}) \\ &= \sum_{s \neq X_k} \big[J_k(s) - J(s)\big] + J_{k+1}(X_k) - J(X_k) - J_k(X_{k+1}) \\ &= \|J_k - J\|_{L_1} - J_k(X_k) + J_{k+1}(X_k) - J_k(X_{k+1}) \\ &= \|J_k - J\|_{L_1} - J_{k-1}(X_k) + J_{k+1}(X_k) - J_k(X_{k+1}) \\ &= W_k + (T^{N-k}J)(X_k) - J_k(X_{k+1}) \\ &= W_k + g(X_k, U_k) + \sum_{s \in S} p(X_k, U_k, s)(T^{N-k-1}J)(s) - J_k(X_{k+1}) \\ &= W_k + g(X_k, U_k) + \hat{E}_{s, J}[(T^{N-k-1}J)(X_{k+1}) \mid \mathcal{F}_k] - J_k(X_{k+1}) \end{aligned}$$

using the facts $J_{k+1}(s) = J_k(s)$ for $s \neq X_k$ (3rd equality), $J_k(X_k) = J_{k-1}(X_k)$ because $X_k \neq X_{k-1}$ by Assumption 3.2 (5th equality), and the definition of U_k (7th equality). Taking conditional expectations with respect to $\mathcal{F}_k$,

$$\begin{aligned} \hat{E}_{s, J}[W_{k+1} \mid \mathcal{F}_k] &= W_k + g(X_k, U_k) - \hat{E}_{s, J}[J_k(X_{k+1}) - (T^{N-k-1}J)(X_{k+1}) \mid \mathcal{F}_k] \\ &\geq W_k + g(X_k, U_k) - \hat{E}_{s, J}[J_0(X_{k+1}) - J(X_{k+1}) \mid \mathcal{F}_k] \\ &\geq W_k + \underline{g}[\![X_k \neq t]\!] - \|J_0 - J\|, \end{aligned}$$

where the first inequality is by Lemma 4.1 and $J \in \mathcal{J}$. Unfold the first term on the right-hand side and integrate with respect to previous $\mathcal{F}_k$s to obtain

$$\hat{E}_{s,J}[W_{k+1} \mid \mathcal{F}_{k-i}] \geq W_{k-i} + \underline{g} \sum_{j=0}^{i} \hat{P}_{s,J}(X_{k-j} \neq t \mid \mathcal{F}_{k-i}) - (i+1)\|J_0 - J\|$$
$$\geq W_{k-i} + \underline{g}(i+1)\hat{P}_{s,J}(X_k \neq t) - (i+1)\|J_0 - J\|$$

because the event $\{X_k \neq t\} \supseteq \{X_{k+1} \neq t\}$. Therefore,

$$\hat{E}_{s,J}[W_{k+1}] = \hat{E}_{s,J}[W_{k+1} \mid \mathcal{F}_0] \geq W_0 + \underline{g}(k+1)\hat{P}_{s,J}(X_k \neq t) - (k+1)\|J_0 - J\|.$$

To obtain the result for $n = 0$, use the following bounds and the fact that the right-hand side of (4) is independent of s:

$$W_0 = \|J_0 - J\|_{L_1} - J_0(s_0) \geq \|J_0 - J\|_{L_1} - \|J^*\|,$$
$$W_{k+1} = \|J_{k+1} - J\|_{L_1} - J_k(X_{k+1}) \leq \|J^* - J\|_{L_1},$$
$$\|J^* - J\|_{L_1} - \|J_0 - J\|_{L_1} \leq |S| \cdot \|J^*\|,$$
$$\|J_0 - J\| \leq (N+1)\|TJ - J\|.$$

The result for general n follows directly because $\|T^{n+1}J - T^nJ\| \leq \|TJ - J\|$. □

Lemma 4.3. *If $J \in \mathcal{J}$ and $\epsilon > 0$, then $\|T^{k+1}J - T^kJ\| \leq \epsilon$ for all $k \geq \|J^* - J\|_{L_1}/\epsilon$.*

Proof. Because $J \in \mathcal{J}$, the updates decrease the value of no state. For s such that $(TJ)(s) - J(s) > \epsilon$, an update increases its value by at least ϵ. Thus, there cannot be more than $\|J^* - J\|_{L_1}/\epsilon$ updates. □

Proof of Theorem 4.1. By (3), we know

$$|J^*(s) - (T^nJ)(s)| \leq \|J^* - J\|\hat{P}_{s,J}(X_n \neq t).$$

Take $n = N = K$ and use the assumptions on the constants K and α and Lemma 4.2 to get $\hat{P}_{s,J}(X_K \neq t) \leq \alpha$. Therefore, $\|J^* - T^{nK}J\| \leq \|J^* - T^{(n-1)K}J\| \max_{s \in S} \hat{P}_{s,T^KJ}(X_K \neq t) \leq \alpha^n \|J^* - J\|$. □

Proof of Theorem 4.2. We divide the convergence in two phases. An initial phase of slow convergence with m iterations finishing when $\|T^{m+1}J - T^mJ\|$ is sufficiently small, and a second phase of geometric convergence.

Fix $\epsilon > 0$ and let $\alpha = 1/2$. We will choose values of m, K, and δ such that

$$\|T^{m+1}J - T^mJ\| \leq \delta, \tag{5}$$
$$(1 + |S|)\|J^*\| + (1+K)^2\|T^{m+1}J - T^mJ\| \leq \alpha(1+K)\underline{g}. \tag{6}$$

Thus, the condition (5) determines the number of iterations for the first phase, and (6) the number of iterations for the final phase. By Lemma 4.3, it is sufficient to set $m = \|J^* - J\|_{L_1}/\delta$, so we need to choose K and δ such that

$$K^2\delta + K(2\delta - \omega) + (\beta + \delta - \omega) \leq 0 \tag{7}$$

where $\beta = (1 + |S|)\|J^*\|$ and $\omega = \underline{g}/2$. This condition is only possible when Equation (7) has two real roots because the coefficients δ and $\beta + \delta - \omega$ are both positive. In such a case, it is sufficient to set $K = (\omega - 2\delta)/2\delta$ (the average of the roots). Equation (7) has real roots only if $\Delta(\delta) = \omega^2 - 4\delta\beta \geq 0$. Therefore, it is enough to take

$$\delta = \frac{\omega^2}{4\beta} = \Theta(\underline{g}^2/\|J^*\||S|),$$
$$m = \frac{4\beta\|J^* - J\|_{L_1}}{\omega^2} = O(\|J^*\|^2|S|^2/\underline{g}^2),$$
$$K = \frac{2\beta - \omega}{\omega} = \Theta(\|J^*\||S|/\underline{g}).$$

Let $J' \doteq T^mJ$ be the value function that results after the first m iterations of value iteration. We now bound the number of iterations in the second phase of convergence that starts at J'. In order to bring the residual $\|J^* - T^{nK}J'\| \leq 2^{-n}\|J^* - J'\| \leq \epsilon$, n must be at least $(\log\|J^*\| + |\log\epsilon|)$. Thus, the total number of iterations that value iteration requires to bring the residual no greater than ϵ is

$$m + nK = O(\|J^*\|^2|S|^2/\underline{g}^2 + (\log\|J^*\| + |\log\epsilon|)\|J^*\||S|/\underline{g}).$$ □

5. A simple bound on $\|J^*\|/\underline{g}$. Denote with $\tilde{P}$ the problem that results from replacing all positive costs with their maximum $\bar{g}$, and with $\tilde{J}^*$ its optimal cost function. Similarly, denote with $\hat{P}$ the problem that results from replacing all positive costs with one, and with $\hat{J}^*$ its optimal cost function. Clearly, $J^* \leq \tilde{J}^*$. Furthermore, $\tilde{J}^* = \bar{g}\hat{J}^*$ because $\tilde{T}^k J_0 = \bar{g}\hat{T}^k J_0$ for $k \geq 0$ where $\tilde{T}$ and $\hat{T}$ are the DP operators for $\tilde{P}$ and $\hat{P}$, respectively. Therefore, $\|J^*\| \leq \bar{g}\|\hat{J}^*\|$.

If μ is a proper policy for problem P, then it is also a proper policy for $\hat{P}$. Denote with $\hat{J}_\mu$ the cost function for μ with respect to problem $\hat{P}$. By definition, $\|\hat{J}^*\| \leq \|\hat{J}_\mu\|$ so a bound on the latter quantity implies a bound on $\|J^*\|$:

$$\begin{aligned}
\hat{J}_\mu(s) &= 1 + \sum_{s' \in S} p(s, \mu(s), s')\hat{J}_\mu(s') \\
&= 1 + \sum_{s' \in S} p(s, \mu(s), s')\left[1 + \sum_{s'' \in S} p(s', \mu(s'), s'')\hat{J}_\mu(s'')\right] \\
&= 2 + \sum_{s', s''} p(s, \mu(s), s')p(s', \mu(s'), s'')\hat{J}_\mu(s'') \\
&\leq |S| + \hat{P}_\mu(X_{|S|} \neq t \mid X_0 = s)\|\hat{J}_\mu\| \leq |S| + \rho_\mu\|\hat{J}_\mu\|.
\end{aligned}$$

Therefore, $\|\hat{J}_\mu\| \leq |S|/(1-\rho_\mu)$ and $\|J^*\|/\underline{g} \leq \bar{g}|S|/(1-\rho_\mu)\underline{g}$. The following is direct.

THEOREM 5.1. *Value iteration converges in a polynomial number of iterations, and hence in polynomial time, for those classes of stochastic shortest-path problems in which $\bar{g}/\underline{g}$ is polynomially bounded, and that also have a proper policy μ such that $1-\rho_\mu$ is polynomially bounded away from zero.*

6. Summary and conclusions. The main result of the paper is a bound on the number of iterations required by the value iteration algorithm to achieve a given residual when the initial vector is admissible. This result implies polynomial time convergence of value iteration on classes of models that have polynomially bounded $\|J^*\|/\underline{g}$.

This is a partial answer, restricted to an important class of initial vectors, towards a general resolution for the speed of convergence of value iteration on stochastic shortest-path problems. Our proof is based on the analysis of a stochastic process associated with the stochastic shortest-path problem. A similar proof method was used in Koenig [10] for the analysis of Minmax-LRTA*. Bonet [6] contains an analysis of the Real-Time Dynamic Programming algorithm (Barto et al. [1]), an asynchronous version of dynamic programming, using a similar approach.

It is important to remark that consistent and admissible estimates are fundamental ideas in heuristic search, where they have been used to establish the optimality and resources needed by algorithms like A* and IDA*; Pearl [12], Dechter and Pearl [7], Korf [11].

Another important open problem is to bound the suboptimality of greedy policies. If μ is the greedy policy with respect to J, the suboptimality of μ is defined as $\|J^* - J_\mu\|$. In general, J and J_μ can differ substantially, and there is no known obvious relation between $\|J^* - J\|$ and $\|J^* - J_\mu\|$. By a bound on the suboptimality of μ, we mean an expression in terms of $\|TJ - J\|$ or $\|J^* - J\|$.

Acknowledgments. The author thanks Héctor Geffner and Henryk Gzyl for fruitful discussions related to this research. He also thanks the anonymous reviewers for comments and suggestions that helped to improve this paper.

References

[1] Barto, A., S. Bradtke, S. Singh. 1995. Learning to act using real-time dynamic programming. *Artificial Intelligence* **72** 81–138.
[2] Bellman, R. 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.
[3] Bertsekas, D. 1995. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA.
[4] Bertsekas, D., J. Tsitsiklis. 1989. *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, Englewood Cliffs, NJ.
[5] Bertsekas, D., J. Tsitsiklis. 1991. An analysis of stochastic shortest-path problems. *Math. Oper. Res.* **16** 580–595.
[6] Bonet, B. 2003. *Modeling and Solving Sequential Decision Tasks with Uncertainty and Partial Information*. Ph.D. thesis, University of California, Los Angeles, CA.
[7] Dechter, R., J. Pearl. 1985. Generalized best-first search strategies and the optimality of A*. *J. Assoc. Comput. Mach.* **32**(3) 505–536.
[8] Eaton, J. H., L. A. Zadeh. 1962. Optimal pursuit strategies in discrete state probabilistic systems. *Trans. ASME, Series D, J. Basic Engrg.* **84** 23–29.

[9] Hart, P., N. Nilsson, B. Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Systems Sci. Cybernetics* **4** 100–107.
[10] Koenig, S. 2001. Minimax real-time heuristic search. *Artificial Intelligence* **129** 165–197.
[11] Korf, R. 1985. Depth-first iterative-deepening: An optimal admissible tree search. *Artificial Intelligence* **27**(1) 97–109.
[12] Pearl, J. 1983. *Heuristics*. Morgan Kaufmann, San Francisco, CA.
[13] Puterman, M. 1994. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.
[14] Tseng, P. 1990. Solving H-horizon, stationary Markov decision problems in time proportional to $\log(H)$. *Oper. Res. Lett.* **9** 289–297.