Affine Monotonic and Risk-Sensitive Models in Dynamic Programming

Dimitri P. Bertsekas

***Abstract*—In this paper we consider a broad class of infinite horizon discrete-time optimal control models that involve a nonnegative cost function and an affine mapping in their dynamic programming equation. They include as special cases several classical models such as stochastic undiscounted nonnegative cost problems, stochastic multiplicative cost problems, and risk-sensitive problems with exponential cost. We focus on the case where the state space is finite and the control space has some compactness properties. We assume that the affine mapping has a semicontractive character, whereby for some policies it is a contraction, while for others it is not. In one line of analysis, we impose assumptions guaranteeing that the noncontractive policies cannot be optimal. Under these assumptions, we prove strong results that resemble those for discounted Markovian decision problems, such as the uniqueness of solution of Bellman's equation, and the validity of forms of value and policy iteration. In the absence of these assumptions, the results are weaker and unusual in character: the optimal cost function need not be a solution of Bellman's equation, and an optimal policy may not be found by value or policy iteration. Instead the optimal cost function over just the contractive policies is the largest solution of Bellman's equation, and can be computed by a variety of algorithms.**

***Index Terms*—Dynamic programming, Markov decision processes, stochastic shortest paths, risk sensitive control.**

I. Introduction

We consider an infinite horizon optimal control model, characterized by an affine and monotone abstract mapping that underlies the associated Bellman equation of dynamic programming (DP for short). This model was formulated with some analysis in the author's monograph [1] as a special case of abstract DP. In the present paper we will provide a deeper analysis and more effective algorithms for the finite-state version of the model, under considerably weaker assumptions.

To relate our analysis with the existing literature, we note that DP models specified by an abstract mapping defining the corresponding Bellman equation have a long history. Models where this mapping is a sup-norm contraction over the space of bounded cost functions were introduced by Denardo [2]; see also Denardo and Mitten [3]. Their main area of application is discounted DP models of various types. Noncontractive models, where the abstract mapping is not a contraction of any kind, but is instead monotone, were considered by Bertsekas [4], [5] (see also Bertsekas and Shreve [6], Ch. 5). Among others, these models cover the important cases of positive and negative (reward) DP problems of Blackwell [7] and Strauch [8], respectively. Extensions of the analysis of [5] were given by Verdu and Poor [9], which considered additional structure that allows the development of backward and forward value iterations, and in the thesis by Szepesvari [10], [11], which introduced non-Markovian policies into the abstract DP framework. The model of [5] was also used to develop asynchronous value iteration methods for abstract contractive and noncontractive DP models; see [12], 13]. Moreover, there have been extensions of the theory to asynchronous policy iteration algorithms and approximate DP by Bertsekas and Yu ([13], [14], [15]).

A type of abstract DP model, called *semicontractive*, was introduced in the monograph [1]. In this model, the abstract DP mapping corresponding to some policies has a regularity/contraction-like property, but the mapping of others does not. A prominent example is the stochastic shortest path problem (SSP for short), a Markovian decision problem where we aim to drive the state of a finite-state Markov chain to a cost-free and absorbing termination state at minimum expected cost. In SSP problems, the contractive policies are the so-called *proper* policies, which are the ones that lead to the termination state with probability 1. The SSP problem, originally introduced by Eaton and Zadeh [16], has been discussed under a variety of assumptions, in many sources, including the books [17], [18], [19], [20], [21], [22], [23], [24], and [12], where it is sometimes referred to by other names such as "first passage problem" and "transient programming problem." It has found a wide range of applications in regulation, path planning, robotics, and other contexts.

In this paper we focus on a different subclass of semicontractive models, called *affine monotonic*, where the abstract mapping associated with a stationary policy is affine and maps nonnegative functions to nonnegative functions. These models include as special cases stochastic undiscounted nonnegative cost problems (including SSP problems with nonnegative cost per stage), and multiplicative cost problems, such as problems with exponentiated cost. A key idea in our analysis is to use the notion of a *contractive policy* (one whose affine mapping involves a matrix with eigenvalues lying strictly within the unit circle). This notion is analogous to the one of a proper policy in SSP problems and is used in similar ways.

Our analytical focus is on the validity and the uniqueness of solution of Bellman's equation, and the convergence of (possibly asynchronous) forms of value and policy iteration. Our results are analogous to those obtained for SSP problems by Bertsekas and Tsitsiklis [25], and Bertsekas and Yu [26]. As in the case of [25], under favorable assumptions where noncontractive policies cannot be optimal, we show that the optimal cost function is the unique solution of Bellman's equation, and we derive strong algorithmic results. As in the

D. P. Bertsekas is with the Laboratory for Information and Decision Systems (LIDS), M.I.T. Email: dimitrib@mit.edu. Address: 77 Massachusetts Ave., Rm 32-D660, Cambridge, MA 02139, USA.

case of [26], we consider more general situations where the optimal cost function need not be a solution of Bellman's equation, and an optimal policy may not be found by value or policy iteration. To address such anomalies, we focus attention on the optimal cost function over just the contractive policies, and we show that it is the largest solution of Bellman's equation and that it is the natural limit of value iteration.

However, there are some substantial differences from the analyses of [25] and [26]. The framework of the present paper is broader than SSP and includes in particular multiplicative cost problems. Moreover, some of the assumptions are different and necessitate a different line of analysis; for example there is no counterpart of the assumption that the optimal cost function is real-valued, which is fundamental in the analysis of [26]. As an indication, we note that deterministic shortest path problems with negative cost cycles can be readily treated within the framework of the present paper, but cannot be analyzed as SSP problems within the standard framework of [25] and the weaker framework of [26] because their optimal shortest path length is equal to $-\infty$ for some initial states.

In this paper, we also pay special attention to exponential cost problems, extending significantly some of the classical results of Denardo and Rothblum [27], and the more recent results of Patek [28]. Both of these papers impose assumptions that guarantee that the optimal cost function is the unique solution of Bellman's equation, whereas our assumptions are much weaker. The paper by Denardo and Rothblum [27] also assumes a finite control space in order to bring to bear a line of analysis based on linear programming (see also the discussion of Section II).

The paper is organized as follows. In Section II we introduce the affine monotonic model, and we show that it contains as a special case multiplicative and exponential cost models. We also introduce contractive policies and related assumptions. In Section III we address the core analytical questions relating to Bellman's equation and its solution, and we obtain favorable results under the assumption that all noncontractive policies have infinite cost starting from some initial state. In Section IV we remove this latter assumption, and we show favorable results relating to a restricted problem whereby we optimize over the contractive policies only. Algorithms such as value iteration, policy iteration, and linear programming are discussed somewhat briefly in this paper, since their analysis follows to a great extent established paths for semicontractive abstract DP models [1].

Regarding notation, we denote by $\Re^n$ the standard Euclidean space of vectors $J = \big(J(1), \ldots, J(n)\big)$ with real-valued components, and we denote by $\Re$ the real line. We denote by $\Re^n_+$ the set of vectors with nonnegative real-valued components,

$$\Re^n_+ = \big\{J \mid 0 \le J(i) < \infty,\ i = 1, \ldots, n\big\},$$

and by $\mathcal{E}^n_+$ the set of vectors with nonnegative extended real-valued components,

$$\mathcal{E}^n_+ = \big\{J \mid 0 \le J(i) \le \infty,\ i = 1, \ldots, n\big\}.$$

Inequalities with vectors are meant to be componentwise, i.e., $J \le J'$ means that $J(i) \le J'(i)$ for all i. Similarly, in the absence of an explicit statement to the contrary, operations on vectors, such as $\lim$, $\limsup$, and $\inf$, are meant to be componentwise.

II. Problem Formulation

We consider a finite state space $X = \{1, \ldots, n\}$ and a (possibly infinite) control constraint set $U(i)$ for each state i. Let $\mathcal{M}$ denote the set of all functions $\mu = \big(\mu(1), \ldots, \mu(n)\big)$ such that $\mu(i) \in U(i)$ for each $i = 1, \ldots, n$. By a *policy* we mean a sequence of the form $\pi = \{\mu_0, \mu_1, \ldots\}$, where $\mu_k \in \mathcal{M}$ for all $k = 0, 1, \ldots$. By a *stationary policy* we mean a policy of the form $\{\mu, \mu, \ldots\}$. For convenience we also refer to any $\mu \in \mathcal{M}$ as a "policy" and use it in place of the stationary policy $\{\mu, \mu, \ldots\}$, when confusion cannot arise.

We introduce for each $\mu \in \mathcal{M}$ a mapping $T_\mu : \mathcal{E}^n_+ \mapsto \mathcal{E}^n_+$ given by

$$T_\mu J = b_\mu + A_\mu J, \tag{1}$$

where b_μ is a vector of $\Re^n$ with components $b\big(i, \mu(i)\big)$, $i = 1, \ldots, n$, and A_μ is an $n \times n$ matrix with components $A_{ij}\big(\mu(i)\big)$, $i, j = 1, \ldots, n$. We assume that $b(i, u)$ and $A_{ij}(u)$ are nonnegative,

$$b(i, u) \ge 0, \quad A_{ij}(u) \ge 0, \quad \forall\ i, j = 1, \ldots, n,\ u \in U(i). \tag{2}$$

We define the mapping $T : \mathcal{E}^n_+ \mapsto \mathcal{E}^n_+$, where for each $J \in \mathcal{E}^n_+$, TJ is the vector of $\mathcal{E}^n_+$ with components

$$(TJ)(i) = \inf_{u \in U(i)} \left[b(i, u) + \sum_{j=1}^n A_{ij}(u) J(j) \right], \qquad i = 1, \ldots, n. \tag{3}$$

Note that since the value of the expression in braces on the right depends on μ only through $\mu(i)$, which is just restricted to be in $U(i)$, we have

$$(TJ)(i) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(i), \qquad i = 1, \ldots, n,$$

so that $(TJ)(i) \le (T_\mu J)(i)$ for all i and $\mu \in \mathcal{M}$.

We now define a DP-like optimization problem that involves the mappings T_μ. We introduce a special vector $\bar{J} \in \Re^n_+$, and we define the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ in terms of the composition of the mappings T_{μ_k}, $k = 0, 1, \ldots$, by

$$J_\pi(i) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(i), \qquad i = 1, \ldots, n.$$

The cost function of a stationary policy μ, is written as

$$J_\mu(i) = \limsup_{N \to \infty} (T^N_\mu \bar{J})(i), \qquad i = 1, \ldots, n.$$

(We use $\limsup$ because we are not assured that the limit exists; our analysis and results remain essentially unchanged if $\limsup$ is replaced by $\liminf$.) In contractive abstract DP models, T_μ is assumed to be a contraction for all $\mu \in \mathcal{M}$, in which case J_μ is the unique fixed point of T_μ and does not depend on the choice of $\bar{J}$. Here we will not be making such an assumption, and the choice of $\bar{J}$ may affect profoundly the character of the problem. For example, in SSP and other additive cost Markovian decision problems $\bar{J}$ is the zero function, $\bar{J}(i) \equiv 0$, while in multiplicative cost models $\bar{J}$ is the unit function, $\bar{J}(i) \equiv 1$, as we will discuss shortly. Also in SSP problems A_μ is a substochastic matrix for all $\mu \in \mathcal{M}$, while in other problems A_μ can have components or row sums that are larger and as well as smaller than 1.

We define the optimal cost function J^* by

$$J^*(i) = \inf_{\pi \in \Pi} J_\pi(i), \qquad i = 1, \ldots, n,$$

where Π denotes the set of all policies. We wish to find J^* and a policy $\pi^* \in \Pi$ that is optimal, i.e., $J_{\pi^*} = J^*$. The analysis of affine monotonic problems revolves around the equation $J = TJ$, or equivalently

$$J(i) = \inf_{u \in U(i)} \left[b(i,u) + \sum_{j=1}^{n} A_{ij}(u) J(j) \right], \quad j = 1, \ldots, n. \tag{4}$$

This is the analog of the classical infinite horizon DP equation and it is referred to as *Bellman's equation*. We are interested in solutions of this equation (i.e., fixed points of T) within $\mathcal{E}_+^n$ and within $\Re_+^n$. Usually in DP models one expects that J^* solves Bellman's equation, while optimal stationary policies can be obtained by minimization over $U(i)$ in its right-hand side. However, this is not true in general, as we will show in Section IV.

Affine monotonic models appear in several contexts. In particular, finite-state sequential stochastic control problems (including SSP problems) with nonnegative cost per stage (see, e.g., [12], Chapter 3, and Section IV) are special cases where $\bar{J}$ is the identically zero function [$\bar{J}(i) \equiv 0$]. Also, discounted problems involving state and control-dependent discount factors (for example semi-Markov problems, cf. Section 1.4 of [12], or Chapter 11 of [29]) are special cases, with the discount factors being absorbed within the scalars $A_{ij}(u)$. In all of these cases, A_μ is a substochastic matrix. There are also other special cases, where A_μ is not substochastic. They correspond to interesting classes of practical problems, including SSP-type problems with a multiplicative or an exponential (rather than additive) cost function, which we proceed to discuss.

A. Multiplicative and Exponential Cost SSP Problems

We will describe a type of SSP problem, where the cost function of a policy accumulates over time multiplicatively, rather than additively, up to the termination state. The special case where the cost from a given state is the expected value of the exponential of the length of the path from the state up to termination was studied by Denardo and Rothblum [27], and Patek [28]. We are not aware of a study of the multiplicative cost version for problems where a cost-free and absorbing termination state plays a major role (the paper by Rothblum [30] deals with multiplicative cost problems but focuses on the average cost case).

Let us introduce in addition to the states $i = 1, \ldots, n$, a cost-free and absorbing state t. There are probabilistic state transitions among the states $i = 1, \ldots, n$, up to the first time a transition to state t occurs, in which case the state transitions terminate. We denote by $p_{it}(u)$ and $p_{ij}(u)$ the probabilities of transition from i to t and to j under u, respectively, so that

$$p_{it}(u) + \sum_{j=1}^{n} p_{ij}(u) = 1, \qquad i = 1, \ldots, n, \ u \in U(i).$$

Next we introduce nonnegative scalars $h(i,u,t)$ and $h(i,u,j)$,

$$h(i,u,t) \geq 0, \quad h(i,u,j) \geq 0, \quad \forall\, i,j = 1, \ldots, n, \ u \in U(i),$$

and we consider the affine monotonic problem where the scalars $A_{ij}(u)$ and $b(i,u)$ are defined by

$$A_{ij}(u) = p_{ij}(u) h(i,u,j), \qquad i,j = 1, \ldots, n, \ u \in U(i), \tag{5}$$

and

$$b(i,u) = p_{it}(u) h(i,u,t), \quad i = 1, \ldots, n, \ u \in U(i), \tag{6}$$

and the vector $\bar{J}$ is the unit vector,

$$\bar{J}(i) = 1, \qquad i = 1, \ldots, n.$$

The cost function of this problem has a multiplicative character as we show next.

Indeed, with the preceding definitions of $A_{ij}(u)$, $b(i,u)$, and $\bar{J}$, we will prove that the expression for the cost function of a policy $\pi = \{\mu_0, \mu_1, \ldots\}$,

$$J_\pi(x_0) = \limsup_{N \to \infty} (T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J})(x_0), \qquad x_0 = 1, \ldots, n,$$

can be written in the multiplicative form

$$J_\pi(x_0) = \limsup_{N \to \infty} E\left\{ \prod_{k=0}^{N-1} h\big(x_k, \mu_k(x_k), x_{k+1}\big) \right\}, \tag{7}$$

where:

(a) $\{x_0, x_1, \ldots\}$ is the random state trajectory generated starting from x_0 and using π.

(b) The expected value is with respect to the probability distribution of that trajectory.

(c) We use the notation

$$h\big(x_k, \mu_k(x_k), x_{k+1}\big) = 1, \qquad \text{if } x_k = x_{k+1} = t,$$

(so that the multiplicative cost accumulation stops once the state reaches t).

Thus, we claim that $J_\pi(x_0)$ *can be viewed as the expected value of cost accumulated multiplicatively, starting from x_0 up to reaching the termination state t (or indefinitely accumulated multiplicatively, if t is never reached).*

To verify the formula (7) for J_π, we use the definition $T_\mu J = b_\mu + A_\mu J$, to show by induction that for every $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J} = A_{\mu_0} \cdots A_{\mu_{N-1}} \bar{J} + b_{\mu_0} + \sum_{k=1}^{N-1} A_{\mu_0} \cdots A_{\mu_{k-1}} b_{\mu_k}. \tag{8}$$

We then interpret the n components of each vector on the right as conditional expected values of the expression

$$\prod_{k=0}^{N-1} h\big(x_k, \mu_k(x_k), x_{k+1}\big) \tag{9}$$

multiplied with the appropriate conditional probability. In particular:

(a) The ith component of the vector $A_{\mu_0}\cdots A_{\mu_{N-1}}\bar{J}$ in Eq. (8) is the conditional expected value of the expression (9), given that $x_0 = i$ and $x_N \neq t$, multiplied with the conditional probability that $x_N \neq t$, given that $x_0 = i$.

(b) The ith component of the vector b_{μ_0} in Eq. (8) is the conditional expected value of the expression (9), given that $x_0 = i$ and $x_1 = t$, multiplied with the conditional probability that $x_1 = t$, given that $x_0 = i$.

(c) The ith component of the vector $A_{\mu_0}\cdots A_{\mu_{k-1}} b_{\mu_k}$ in Eq. (8) is the conditional expected value of the expression (9), given that $x_0 = i$, $x_1,\ldots,x_{k-1} \neq t$, and $x_k = t$, multiplied with the conditional probability that $x_1,\ldots,x_{k-1} \neq t$, and $x_k = t$, given that $x_0 = i$.

By adding these conditional probability expressions, we obtain the ith component of the unconditional expected value

$$E\left\{\prod_{k=0}^{N-1} h\big(x_k,\mu_k(x_k),x_{k+1}\big)\right\},$$

thus verifying the formula (7).

A special case of multiplicative cost problem is the *risk-sensitive SSP problem with exponential cost function*, where for all $i = 1,\ldots,n$, and $u \in U(i)$,

$$h(i,u,j) = \exp\big(g(i,u,j)\big), \quad j = 1,\ldots,n,t, \qquad (10)$$

and the function g can take both positive and negative values. The Bellman equation for this problem is

$$J(i) = \inf_{u\in U(i)}\Bigg[p_{it}(u)\exp\big(g(i,u,t)\big) + \sum_{j=1}^{n} p_{ij}(u)\exp\big(g(i,u,j)\big)J(j)\Bigg], \quad i = 1,\ldots,n. \qquad (11)$$

Based on Eq. (7), we have that $J_\pi(x_0)$ is the limit superior of the expected value of the exponential of the N-step additive finite horizon cost up to termination, i.e.,

$$\sum_{k=0}^{\bar{k}} g\big(x_k,\mu_k(x_k),x_{k+1}\big),$$

where $\bar{k}$ is equal to the first index prior to $N-1$ such that $x_{\bar{k}+1} = t$, or is equal to $N-1$ if there is no such index. The use of the exponential introduces risk aversion, by assigning a strictly convex increasing penalty for large rather than small cost of a trajectory up to termination (and hence a preference for small variance of the additive cost up to termination).

In the cases where $0 \leq g$ or $g \leq 0$, we also have $\bar{J} \leq T\bar{J}$ and $T\bar{J} \leq \bar{J}$, respectively, corresponding to a monotone increasing and a monotone decreasing problem, in the terminology of [1]. Both of these problems admit a favorable analysis, highlighted by the fact that J^* is a fixed point of T (see [1], Chapter 4).

The case where g can take both positive and negative values is more challenging, and is the focus of this paper. We will consider two cases, discussed in Sections III and IV of this paper, respectively. Under the assumptions of Section III, J^* is shown to be the unique fixed point of T within $\Re^n_+$. Under the assumptions of Section IV, it may happen that J^* is not a fixed point of T (see Example 2 that follows). Denardo and Rothblum [27] and Patek [28] consider only the more benign Section III case, for which J^* is a fixed point of T. Also, the approach of [27] is very different from ours: it relies on linear programming ideas, and for this reason it requires a finite control constraint set and cannot be readily adapted to an infinite control space. The approach of [28] is closer to ours in that it also descends from the paper [25]. It allows for an infinite control space under a compactness assumption that is similar to our Assumption 2.2 of the next section, but it also requires that $g(i,u,j) > 0$ for all i,u,j, so it deals only with a monotone increasing case where $T_\mu\bar{J} \geq \bar{J}$ for all $\mu \in \mathcal{M}$.

The deterministic version of the exponential cost problem where for each $u \in U(i)$, only one of the transition probabilities $p_{it}(u), p_{i1}(u),\ldots,p_{in}(u)$ is equal to 1 and all others are equal to 0, is mathematically equivalent to the classical deterministic shortest path problem (since minimizing the exponential of a deterministic expression is equivalent to minimizing that expression). For this problem a standard assumption is that there are no cycles that have negative total length to ensure that the shortest path length is finite. However, it is interesting that this assumption is not required in the present paper: when there are paths that travel perpetually around a negative length cycle we simply have $J^*(i) = 0$ for all states i on the cycle, which is permissible within our context.

B. Assumptions on Policies - Contractive Policies

We now introduce a special property of policies which is central for the purposes of this paper. We say that a given stationary policy μ is *contractive if $A_\mu^N \to 0$ as $N \to \infty$*. Equivalently, μ is contractive if all the eigenvalues of A_μ lie strictly within the unit circle. Otherwise, μ is called *noncontractive*. It follows from a classical result that a policy μ is contractive if and only if T_μ is a contraction with respect to some norm. Because $A_\mu \geq 0$, a stronger assertion can be made: μ is contractive if and only if A_μ is a contraction with respect to some weighted sup-norm (see e.g., the discussion in [22], Ch. 2, Cor. 6.2, or [12], Section 1.5.1). In the special case of SSP problems with additive cost function (all matrices A_μ are substochastic), the contractive policies coincide with the *proper* policies, i.e., the ones that lead to the termination state with probability 1, starting from every state.

A particularly favorable situation for an SSP problem arises when all policies are proper, in which case all the mappings T and T_μ are contractions with respect to some common weighted sup norm. This result was shown in the paper by Veinott [31], where it was attributed to A. J. Hoffman. Alternative proofs of this contraction property are given in Bertsekas and Tsitsiklis, [22], p. 325 and [32], Prop. 2.2, Tseng [33], and Littman [34]. The proofs of [32] and [34] are essentially identical, and easily generalize to the context of the present paper. In particular, it can be shown that if all policies are contractive, all the mappings T and T_μ are contractions with respect to some common weighted sup norm. However, we will not prove or use this fact in this paper.

Let us derive an expression for the cost function of contractive and noncontractive policies. By repeatedly applying the mapping T to both sides of the equation $T_\mu J = b_\mu + A_\mu J$, we have

$$T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu, \qquad \forall\, J \in \Re^n,\ N = 1,2,\ldots,$$

and hence

$$J_\mu = \limsup_{N\to\infty} T_\mu^N \bar J = \limsup_{N\to\infty} A_\mu^N \bar J + \sum_{k=0}^{\infty} A_\mu^k b_\mu. \tag{12}$$

From these expressions, it follows that if μ is contractive, the initial function $\bar J$ in the definition of J_μ does not matter, and we have

$$\begin{aligned} J_\mu &= \limsup_{N\to\infty} T_\mu^N J \\ &= \limsup_{N\to\infty} \sum_{k=0}^{N-1} A_\mu^k b_\mu, \quad \forall\ \mu\text{: contractive},\ J \in \Re^n. \end{aligned}$$

Moreover, since for a contractive μ, T_μ is a contraction with respect to a weighted sup-norm, the $\limsup$ above can be replaced by $\lim$, so that

$$J_\mu = \sum_{k=0}^{\infty} A_\mu^k b_\mu = (I - A_\mu)^{-1} b_\mu, \qquad \forall\ \mu\text{: contractive}. \tag{13}$$

Thus if μ is contractive, J_μ is real-valued as well as nonnegative, i.e., $J_\mu \in \Re_+^n$. If μ is noncontractive, we have $J_\mu \in \mathcal{E}_+^n$ and it is possible that $J_\mu(i) = \infty$ for some states i. We will assume throughout the paper the following.

Assumption 1. *There exists at least one contractive policy.*

The analysis of finite-state SSP problems typically assumes that the control space is either finite, or satisfies a compactness and continuity condition introduced in [25]. The following is a similar condition, and will be in effect throughout the paper.

Assumption 2 (Compactness and Continuity)**.** *The control space U is a metric space, and $p_{ij}(\cdot)$ and $b(i,\cdot)$ are continuous functions of u over $U(i)$, for all i and j. Moreover, for each state i, the sets*

$$\left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j) \le \lambda \right\}$$

are compact subsets of U for all scalars $\lambda \in \Re$ and $J \in \Re_+^n$.

The preceding assumption is satisfied if the control space U is finite. One way to see this is to simply identify each $u \in U$ with a distinct integer from the real line. Another interesting case where the assumption is satisfied is when for all i, $U(i)$ is a compact subset of the metric space U, and the functions $b(i,\cdot)$ and $A_{ij}(\cdot)$ are continuous functions of u over $U(i)$.

An advantage of allowing $U(i)$ to be infinite and compact is that it makes possible the use of randomized policies for problems where there is a *finite* set of feasible actions at each state i, call it $C(i)$. We may then specify $U(i)$ to be the set of all probability distributions over $C(i)$, which is a compact subset of a Euclidean space. In this way, our results apply to finite-state and finite-action problems where randomization is allowed, and J^* is the optimal cost function over all randomized nonstationary policies. Note, however, that the optimal cost function may change when randomized policies are introduced in this way. Basically, for our purposes, optimization over nonrandomized and over randomized policies over finite action sets $C(i)$ are two different problems, both of which are interesting and can be addressed with the methodology of this paper. However, when the sets $C(i)$ are infinite, a different and mathematically more sophisticated framework is required in order to allow randomized policies. The reason is that randomized policies over the infinite action sets $C(i)$ must obey measurability restrictions, such as universal measurability; see Bertsekas and Shreve [6], and Yu and Bertsekas [35].

The compactness and continuity part of the preceding assumption guarantees some important properties of the mapping T. These are summarized in the following proposition.

Proposition 1. *Let Assumptions 1 and 2 hold.*

(a) *The set of $u \in U(i)$ that minimize the expression*

$$b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j) \tag{14}$$

is nonempty and compact.

(b) *Let J_0 be the zero vector in $\Re^n$ [$J_0(i) \equiv 0$]. The sequence $\{T^k J_0\}$ is monotonically nondecreasing and converges to a limit $\tilde J \in \Re_+^n$ that satisfies $\tilde J \le J^*$ and $\tilde J = T\tilde J$.*

Proof. (a) The set of $u \in U(i)$ that minimize the expression in Eq. (14) is the intersection $\cap_{m=1}^\infty U_m$ of the nested sequence of sets

$$U_m = \left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j) \le \lambda_m \right\}, \quad m = 1,2,\ldots,$$

where $\{\lambda_m\}$ is a monotonically decreasing sequence such that

$$\lambda_m \downarrow \inf_{u \in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j) \right].$$

Each set U_m is nonempty, and by Assumption 2, it is compact, so the intersection is nonempty and compact.

(b) By the nonnegativity of $b(i,u)$ and $A_{ij}(u)$, we have $J_0 \le TJ_0$, which by the monotonicity of T implies that $\{T^k J_0\}$ is monotonically nondecreasing to a limit $\tilde J \in \mathcal{E}_+^n$, and we have

$$J_0 \le TJ_0 \le \cdots \le T^k J_0 \le \cdots \le \tilde J. \tag{15}$$

For all policies $\pi = \{\mu_0, \mu_1, \ldots\}$, we have $T^k J_0 \le T^k \bar J \le T_{\mu_0} \cdots T_{\mu_{k-1}} \bar J$, so by taking the limit as $k \to \infty$, we obtain $\tilde J \le J_\pi$, and by taking the infimum over π, it follows that $\tilde J \le J^*$. By Assumption 1, there exists at least one contractive policy μ, for which J_μ is real-valued [cf. Eq. (13)], so $J^* \in \Re_+^n$. It follows that the sequence $\{T^k J_0\}$ consists of vectors in $\Re_+^n$, while $\tilde J \in \Re_+^n$.

By applying T to both sides of Eq. (15), we obtain

$$(T^{k+1}J_0)(i) = \inf_{u\in U(i)} \left[b(i,u) + \sum_{j=1}^n A_{ij}(u)(T^kJ_0)(j) \right] \\ \le (T\tilde{J})(i),$$

and by taking the limit as $k \to \infty$, it follows that $\tilde{J} \le T\tilde{J}$. Assume to arrive at a contradiction that there exists a state $\tilde{i}$ such that

$$\tilde{J}(\tilde{i}) < (T\tilde{J})(\tilde{i}). \tag{16}$$

Consider the sets

$$U_k(\tilde{i}) = \left\{ u \in U(\tilde{i}) \;\middle|\; b(\tilde{i},u) + \sum_{j=1}^n A_{\tilde{i}j}(u)(T^kJ_0)(j) \le \tilde{J}(\tilde{i}) \right\}$$

for $k \ge 0$. It follows from Assumption 2 and Eq. (15) that $\{U_k(\tilde{i})\}$ is a nested sequence of compact sets. Let also u_k be a control attaining the minimum in

$$\min_{u\in U(\tilde{i})} \left[b(\tilde{i},u) + \sum_{j=1}^n A_{\tilde{i}j}(u)(T^kJ_0)(j) \right];$$

[such a control exists by part (a)]. From Eq. (15), it follows that for all $m \ge k$,

$$\begin{aligned} b(\tilde{i},u_m) &+ \sum_{j=1}^n A_{\tilde{i}j}(u_m)(T^kJ_0)(j) \\ &\le b(\tilde{i},u_m) + \sum_{j=1}^n A_{\tilde{i}j}(u_m)(T^mJ_0)(j) \\ &\le \tilde{J}(\tilde{i}). \end{aligned}$$

Therefore $\{u_m\}_{m=k}^\infty \subset U_k(\tilde{i})$, and since $U_k(\tilde{i})$ is compact, all the limit points of $\{u_m\}_{m=k}^\infty$ belong to $U_k(\tilde{i})$ and at least one such limit point exists. Hence the same is true of the limit points of the entire sequence $\{u_m\}_{m=0}^\infty$. It follows that if $\tilde{u}$ is a limit point of $\{u_m\}_{m=0}^\infty$ then

$$\tilde{u} \in \cap_{k=0}^\infty U_k(\tilde{i}).$$

This implies that for all $k \ge 0$

$$(T^{k+1}J_0)(\tilde{i}) \le b(\tilde{i},\tilde{u}) + \sum_{j=1}^n A_{\tilde{i}j}(\tilde{u})(T^kJ_0)(j) \le \tilde{J}(\tilde{i}).$$

By taking the limit in this relation as $k \to \infty$, we obtain

$$\tilde{J}(\tilde{i}) = b(\tilde{i},\tilde{u}) + \sum_{j=1}^n A_{\tilde{i}j}(\tilde{u})\tilde{J}(j).$$

Since the right-hand side is greater than or equal to $(T\tilde{J})(\tilde{i})$, Eq. (16) is contradicted, implying that $\tilde{J} = T\tilde{J}$. $\square$

III. Case of Infinite Cost Noncontractive Policies

We now turn to questions relating to Bellman's equation, the convergence of value iteration (VI for short) and policy iteration (PI for short), as well as conditions for optimality of a stationary policy. In this section we will use the following assumption, which parallels the central assumption of [25] for SSP problems. We will not need this assumption in Section IV.

Assumption 3 (Infinite Cost Condition)**.** *For every noncontractive policy μ, there is at least one state such that the corresponding component of the vector $\sum_{k=0}^\infty A_\mu^k b_\mu$ is equal to ∞.*

Note that the preceding assumption guarantees that for every noncontractive policy μ, we have $J_\mu(i) = \infty$ for at least one state i [cf. Eq. (12)]. The reverse is not true, however: $J_\mu(i) = \infty$ does not imply that the ith component of $\sum_{k=0}^\infty A_\mu^k b_\mu$ is equal to ∞, since there is the possibility that $A_\mu^N \bar{J}$ may become unbounded as $N \to \infty$ [cf. Eq. (12)].

Under Assumptions 1, 2, and 3, we will now derive results that closely parallel the standard results of [25] for additive cost SSP problems. We have the following characterization of contractive policies.

Proposition 2 (Properties of Contractive Policies)**.** *Let Assumption 3 hold.*

(a) *For a contractive policy μ, the associated cost vector J_μ satisfies*

$$\lim_{k\to\infty} (T_\mu^k J)(i) = J_\mu(i), \qquad i = 1,\dots,n,$$

for every vector $J \in \Re^n$. Furthermore, we have $J_\mu = T_\mu J_\mu$, and J_μ is the unique solution of this equation within $\Re^n$.

(b) *A stationary policy μ is contractive if and only if it satisfies $J \ge T_\mu J$ for some vector $J \in \Re^n_+$.*

Proof. (a) Follows from Eqs. (8) and (13), and by writing the equation $J_\mu = T_\mu J_\mu$ as $(I - A_\mu)J_\mu = b_\mu$.

(b) If μ is contractive, by part (a) we have $J \ge T_\mu J$ for $J = J_\mu \in \Re^n_+$. Conversely, let J be a vector in $\Re^n_+$ with $J \ge T_\mu J$. Then the monotonicity of T_μ and Eq. (8) imply that for all N we have

$$J \ge T_\mu^N J = A_\mu^N J + \sum_{k=0}^{N-1} A_\mu^k b_\mu \ge \sum_{k=0}^{N-1} A_\mu^k b_\mu \ge 0.$$

It follows that the vector $\sum_{k=0}^\infty A_\mu^k b_\mu$ is real-valued so that, by Assumption 3, μ cannot be noncontractive. $\square$

The following proposition is our main result under Assumption 3. It parallels Prop. 3 of [25] (see also Section 3.2 of [12]). In addition to the fixed point property of J^* and the convergence of the VI sequence $\{T^kJ\}$ to J^* starting from any $J \in \Re^n_+$, it shows the validity of the PI algorithm. The latter algorithm generates a sequence $\{\mu^k\}$ starting from any contractive policy μ^0. Its typical iteration consists of a computation of J_{μ^k} using the policy evaluation equation $J_{\mu^k} = T_{\mu^k}J_{\mu^k}$, followed by the policy improvement operation $T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k}$.

Proposition 3 (Bellman's Equation, Policy Iteration, Value Iteration, and Optimality Conditions)**.** *Let Assumptions 1, 2, and 3 hold.*

(a) *The optimal cost vector J^* satisfies the Bellman equation $J = TJ$. Moreover, J^* is the unique solution of this equation within $\Re^n_+$.*

(b) *Starting with any contractive policy μ^0, the sequence $\{\mu^k\}$ generated by the PI algorithm consists of contractive policies, and any limit point of this sequence is a contractive optimal policy.*

(c) *We have*

$$\lim_{k\to\infty} (T^k J)(i) = J^*(i), \qquad i = 1,\ldots,n,$$

for every vector $J \in \Re^n_+$.

(d) *A stationary policy μ is optimal if and only if $T_\mu J^* = TJ^*$.*

(e) *For a vector $J \in \Re^n_+$, if $J \le TJ$ then $J \le J^*$, and if $J \ge TJ$ then $J \ge J^*$.*

Proof. (a), (b) From Prop. 1(b), T has as fixed point the vector $\tilde J \in \Re^n_+$, the limit of the sequence $\{T^k J_0\}$, where J_0 is the identically zero vector $[J_0(i) \equiv 0]$. We will show parts (a) and (b) simultaneously and in stages. First we will show that $\tilde J$ is the unique fixed point of T within $\Re^n_+$. Then we will show that the PI algorithm, starting from any contractive policy, generates in the limit a contractive policy $\overline{\mu}$ such that $J_{\overline{\mu}} = \tilde J$. Finally we will show that $J_{\overline{\mu}} = J^*$.

Indeed, if J and J' are two fixed points, then we select μ and μ' such that $J = TJ = T_\mu J$ and $J' = TJ' = T_{\mu'}J'$; this is possible because of Prop. 1(a). By Prop. 2(b), we have that μ and μ' are contractive, and by Prop. 2(a) we obtain $J = J_\mu$ and $J' = J_{\mu'}$. We also have $J = T^k J \le T^k_{\mu'} J$ for all $k \ge 1$, and by Prop. 2(a), it follows that $J \le \lim_{k\to\infty} T^k_{\mu'} J = J_{\mu'} = J'$. Similarly, $J' \le J$, showing that $J = J'$. Thus T has $\tilde J$ as its unique fixed point within $\Re^n_+$.

We next turn to the PI algorithm. Let μ be a contractive policy (there exists one by Assumption 1). Choose μ' such that

$$T_{\mu'} J_\mu = TJ_\mu.$$

Then we have $J_\mu = T_\mu J_\mu \ge T_{\mu'} J_\mu$. By Prop. 2(b), μ' is contractive, and using the monotonicity of $T_{\mu'}$ and Prop. 2(a), we obtain

$$J_\mu \ge \lim_{k\to\infty} T^k_{\mu'} J_\mu = J_{\mu'}. \tag{17}$$

Continuing in the same manner, we construct a sequence $\{\mu^k\}$ such that each μ^k is contractive and

$$J_{\mu^k} \ge T_{\mu^{k+1}} J_{\mu^k} = TJ_{\mu^k} \ge J_{\mu^{k+1}}, \qquad k = 0,1,\ldots \tag{18}$$

The sequence $\{J_{\mu^k}\}$ is real-valued, nonincreasing, and nonnegative so it converges to some $J_\infty \in \Re^n_+$.

We claim that the sequence of vectors $\mu^k = \big(\mu^k(1),\ldots,\mu^k(n)\big)$ has a limit point $\big(\overline{\mu}(1),\ldots,\overline{\mu}(n)\big)$, with $\overline{\mu}$ being a feasible policy. Indeed, using Eq. (18) and the fact $J_\infty \le J_{\mu^{k-1}}$, we have for all $k = 1,2,\ldots,$

$$T_{\mu^k} J_\infty \le T_{\mu^k} J_{\mu^{k-1}} = TJ_{\mu^{k-1}} \le T_{\mu^{k-1}} J_{\mu^{k-1}} = J_{\mu^{k-1}} \le J_{\mu^0},$$

so $\mu^k(i)$ belongs to the set

$$\hat U(i) = \left\{ u \in U(i) \;\middle|\; b(i,u) + \sum_{j=1}^n A_{ij}(u) J_\infty(j) \le J_{\mu^0}(i) \right\},$$

which is compact by Assumption 2. Hence the sequence $\{\mu^k\}$ belongs to the compact set $\hat U(1) \times \cdots \times \hat U(n)$, and has a limit point $\overline{\mu}$, which is a feasible policy. In what follows, without loss of generality, we assume that the entire sequence $\{\mu^k\}$ converges to $\overline{\mu}$.

Since $J_{\mu^k} \downarrow J_\infty \in \Re^n_+$ and $\mu^k \to \overline{\mu}$, by taking limit as $k \to \infty$ in Eq. (18), and using the continuity part of Assumption 2, we obtain $J_\infty = T_{\overline{\mu}} J_\infty$. It follows from Prop. 2(b) that $\overline{\mu}$ is contractive, and that $J_{\overline{\mu}}$ is equal to J_∞. To show that $J_{\overline{\mu}}$ is a fixed point of T, we note that from the right side of Eq. (18), we have for all policies μ, $T_\mu J_{\mu^k} \ge J_{\mu^{k+1}}$, which by taking limit as $k \to \infty$ yields $T_\mu J_{\overline{\mu}} \ge J_{\overline{\mu}}$. By taking minimum over μ, we obtain $TJ_{\overline{\mu}} \ge J_{\overline{\mu}}$. Combining this with the relation $J_{\overline{\mu}} = T_{\overline{\mu}} J_{\overline{\mu}} \ge TJ_{\overline{\mu}}$, it follows that $J_{\overline{\mu}} = TJ_{\overline{\mu}}$. Thus $J_{\overline{\mu}}$ is equal to the unique fixed point $\tilde J$ of T within $\Re^n_+$.

We will now conclude the proof by showing that $J_{\overline{\mu}}$ is equal to the optimal cost vector J^* (which also implies the optimality of the policy $\overline{\mu}$, obtained from the PI algorithm starting from a contractive policy). By Prop. 1(b), the sequence $T^k J_0$ converges monotonically to $\tilde J$, which is equal to $J_{\overline{\mu}}$. Also, for every policy $\pi = \{\mu_0, \mu_1, \ldots\}$, we have

$$T^k J_0 \le T^k \bar J \le T_{\mu_0} \cdots T_{\mu_{k-1}} \bar J, \qquad k = 0,1,\ldots,$$

and by taking the limit as $k \to \infty$, we obtain $J_{\overline{\mu}} = \tilde J = \lim_{k\to\infty} T^k J_0 \le J_\pi$ for all π, showing that $J_{\overline{\mu}} = J^*$. Thus J^* is the unique fixed point of T within $\Re^n_+$, and $\overline{\mu}$ is an optimal policy.

(c) From the preceding proof, we have that $T^k J_0 \to J^*$, which implies that

$$\lim_{k\to\infty} T^k J = J^*, \qquad \forall\ J \in \Re^n_+ \text{ with } J \le J^*. \tag{19}$$

Also, for any $J \in \Re^n_+$ with $J \ge J^*$, we have

$$T^k_{\overline{\mu}} J \ge T^k J \ge T^k J^* = J^* = J_{\overline{\mu}},$$

where $\overline{\mu}$ is the contractive optimal policy obtained by PI in the proof of part (b). By taking the limit as $k \to \infty$ and using the fact $T^k_{\overline{\mu}} J \to J_{\overline{\mu}}$ (which follows from the contractiveness of $\overline{\mu}$), we obtain

$$\lim_{k\to\infty} T^k J = J^*, \qquad \forall\ J \in \Re^n_+ \text{ with } J \ge J^*. \tag{20}$$

Finally, given any $J \in \Re^n_+$, we have from Eqs. (19) and (20),

$$\lim_{k\to\infty} T^k\big(\min\{J, J^*\}\big) = J^*, \qquad \lim_{k\to\infty} T^k\big(\max\{J, J^*\}\big) = J^*,$$

and since J lies between $\min\{J, J^*\}$ and $\max\{J, J^*\}$, it follows that $T^k J \to J^*$.

(d) If μ is optimal, then $J_\mu = J^*$ and since by part (a) J^* is real-valued, μ is contractive. Therefore, by Prop. 2(a),

$$T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*.$$

Conversely, if $J^* = TJ^* = T_\mu J^*$, it follows from Prop. 2(b) that μ is contractive, and by using Prop. 2(a), we obtain $J^* =$

J_μ. Therefore μ is optimal. The existence of an optimal policy follows from part (b).

(e) If $J \in \Re^n_+$ and $J \leq TJ$, by repeatedly applying T to both sides and using the monotonicity of T, we obtain $J \leq T^k J$ for all k. Taking the limit as $k \to \infty$ and using the fact $T^k J \to J^*$ [cf. part (c)], we obtain $J \leq J^*$. The proof that $J \geq J^*$ if $J \geq TJ$ is similar. □

A. Computational Methods

Proposition 3(b) shows the validity of PI when starting from a contractive policy. This is similar to the case of additive cost SSP, where PI is known to converge starting from a proper policy (cf. the proof of Prop. 3 of [25]).

There is also an asynchronous version of the PI algorithm proposed for discounted and SSP models by Bertsekas and Yu [14], [15], which does not require an initial contractive policy and admits an asynchronous implementation. This algorithm extends straightforwardly to the affine monotonic model of this paper under Assumptions 2.1, 2.2, and 3.1 (see [1], Section 3.3.2, for a description of this extension to abstract DP models).

Proposition 3(c) establishes the validity of the VI algorithm that generates the sequence $\{T^k J\}$, starting from any initial $J \in \Re^n_+$. An asynchronous version of this algorithm is also valid; see the discussion of Section 3.3.1 of [1].

Finally, Prop. 3(e) shows it is possible to compute J^* as the unique solution of the problem of maximizing $\sum_{i=1}^n \beta_i J(i)$ over all $J = \big(J(1), \ldots, J(n)\big)$ such that $J \leq TJ$, where $\beta_1, \ldots, \beta_n$ are any positive scalars. This problem can be written as

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^n \beta_i J(i) \\ &\text{subject to} \quad J(i) \leq b(i,u) + \sum_{j=1}^n A_{ij}(u) J(j), \\ &\qquad\qquad i = 1, \ldots, n, \quad u \in U(i), \end{aligned} \tag{21}$$

and it is a linear program if each $U(i)$ is a finite set.

IV. Case of Finite Cost noncontractive Policies

We will now eliminate Assumption 3, thus allowing noncontractive policies with real-valued cost functions. We will prove results that are weaker yet useful and substantial. An important notion in this regard is the optimal cost that can be achieved with contractive policies only, i.e., the vector $\hat{J}$ with components given by

$$\hat{J}(i) = \inf_{\mu:\,\text{contractive}} J_\mu(i), \qquad i = 1, \ldots, n. \tag{22}$$

We will show that $\hat{J}$ is a solution of Bellman's equation, while J^* need not be. To this end, we give an important property of noncontractive policies in the following proposition.

Proposition 4. *If μ is a noncontractive policy and all the components of b_μ are strictly positive, then there exists at least one state i such that the corresponding component of the vector $\sum_{k=0}^\infty A_\mu^k b_\mu$ is ∞.*

Proof. According to the Perron-Frobenius Theorem, the nonnegative matrix A_μ has a real eigenvalue λ, which is equal to its spectral radius, and an associated nonnegative eigenvector $\xi \neq 0$ (see e.g., [22], Chapter 2, Prop. 6.6). Choose $\gamma > 0$ to be such that $b_\mu \geq \gamma \xi$, so that

$$\sum_{k=0}^\infty A_\mu^k b_\mu \geq \gamma \sum_{k=0}^\infty A_\mu^k \xi = \gamma \left(\sum_{k=0}^\infty \lambda^k \right) \xi.$$

Since some component of ξ is positive while $\lambda \geq 1$ (since μ is noncontractive), the corresponding component of the infinite sum on the right is infinite, and the same is true for the corresponding component of the vector $\sum_{k=0}^\infty A_\mu^k b_\mu$ on the left. □

A. The δ-Perturbed Problem

We now introduce a perturbation line of analysis, also used in [1] and [26], whereby we add a constant $\delta > 0$ to all components of b_μ, thus obtaining what we call the *δ-perturbed affine monotonic model*. We denote by $J_{\mu,\delta}$ and J^*_δ the cost function of μ and the optimal cost function of the δ-perturbed model, respectively. We have the following proposition.

Proposition 5. *Let Assumptions 1 and 2 hold. Then for each $\delta > 0$:*

(a) *J^*_δ is the unique solution within $\Re^n_+$ of the equation*

$$J(i) = (TJ)(i) + \delta, \qquad i = 1, \ldots, n.$$

(b) *A policy μ is optimal for the δ-perturbed problem (i.e., $J_{\mu,\delta} = J^*_\delta$) if and only if $T_\mu J^*_\delta = T J^*_\delta$. Moreover, for the δ-perturbed problem, all optimal policies are contractive and there exists at least one contractive policy that is optimal.*

(c) *The optimal cost function over contractive policies $\hat{J}$ [cf. Eq. (22)] satisfies*

$$\hat{J}(i) = \lim_{\delta \downarrow 0} J^*_\delta(i), \qquad i = 1, \ldots, n.$$

(d) *If the control constraint set $U(i)$ is finite for all states $i = 1, \ldots, n$, there exists a contractive policy $\hat{\mu}$ that attains the minimum over all contractive policies, i.e., $J_{\hat{\mu}} = \hat{J}$.*

Proof. (a), (b) By Prop. 4, we have that Assumption 3 holds for the δ-perturbed problem. The results follow by applying Prop. 3 [the equation of part (a) is Bellman's equation for the δ-perturbed problem].

(c) For an optimal contractive policy μ^*_δ of the δ-perturbed problem [cf. part (b)], we have for all μ' that are contractive

$$\hat{J} = \inf_{\mu:\,\text{contractive}} J_\mu \leq J_{\mu^*_\delta} \leq J_{\mu^*_\delta,\delta} = J^*_\delta \leq J_{\mu',\delta}.$$

Since for every contractive policy μ', we have $\lim_{\delta \downarrow 0} J_{\mu',\delta} = J_{\mu'}$, it follows that

$$\hat{J} \leq \lim_{\delta \downarrow 0} J_{\mu^*_\delta} \leq J_{\mu'}, \qquad \forall\ \mu' : \text{contractive}.$$

By taking the infimum over all μ' that are contractive, the result follows.

(d) Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and consider a corresponding sequence $\{\mu_k\}$ of optimal contractive policies for the δ_k-perturbed problems. Since the set of contractive policies is finite, some policy $\hat{\mu}$ will be repeated infinitely often within the sequence $\{\mu_k\}$, and since $\{J^*_{\delta_k}\}$ is monotonically nonincreasing, we will have

$$\hat{J} \leq J_{\hat{\mu}} \leq J^*_{\delta_k},$$

for all k sufficiently large. Since by part (c), $J^*_{\delta_k} \downarrow \hat{J}$, it follows that $J_{\hat{\mu}} = \hat{J}$. □

B. Main Results

We now show that $\hat{J}$ is the largest fixed point of T within $\Re^n_+$. This is the subject of the next proposition, which also provides a convergence result for VI as well as an optimality condition; see Fig. 1.

Proposition 6 (Bellman's Equation, Value Iteration, and Optimality Conditions)**.** *Let Assumptions 1 and 2 hold. Then:*

(a) *The optimal cost function over contractive policies, $\hat{J}$, is the largest fixed point of T within $\Re^n_+$, i.e., $\hat{J}$ is a fixed point that belongs to $\Re^n_+$, and if $J' \in \Re^n_+$ is another fixed point, we have $J' \leq \hat{J}$.*

(b) *We have $T^k J \to \hat{J}$ for every $J \in \Re^n_+$ with $J \geq \hat{J}$.*

(c) *Let μ be a contractive policy. Then μ is optimal within the class of contractive policies (i.e., $J_\mu = \hat{J}$) if and only if $T_\mu \hat{J} = T\hat{J}$.*

Proof. (a), (b) For all contractive μ, we have $J_\mu = T_\mu J_\mu \geq T_\mu \hat{J} \geq T\hat{J}$. Taking the infimum over contractive μ, we obtain $\hat{J} \geq T\hat{J}$. Conversely, for all $\delta > 0$ and $\mu \in \mathcal{M}$, we have

$$J^*_\delta = TJ^*_\delta + \delta e \leq T_\mu J^*_\delta + \delta e.$$

Taking limit as $\delta \downarrow 0$, and using Prop. 5(c), we obtain $\hat{J} \leq T_\mu \hat{J}$ for all $\mu \in \mathcal{M}$. Taking infimum over $\mu \in \mathcal{M}$, it follows that $\hat{J} \leq T\hat{J}$. Thus $\hat{J}$ is a fixed point of T.

For all $J \in \Re^n$ with $J \geq \hat{J}$ and contractive μ, we have by using the relation $\hat{J} = T\hat{J}$ just shown,

$$\hat{J} = \lim_{k\to\infty} T^k \hat{J} \leq \lim_{k\to\infty} T^k J \leq \lim_{k\to\infty} T^k_\mu J = J_\mu.$$

Taking the infimum over all contractive μ, we obtain

$$\hat{J} \leq \lim_{k\to\infty} T^k J \leq \hat{J}, \qquad \forall\ J \geq \hat{J}.$$

This proves that $T^k J \to \hat{J}$ starting from any $J \in \Re^n_+$ with $J \geq \hat{J}$. Finally, let $J' \in \Re^n_+$ be another fixed point of T, and let $J \in \Re^n_+$ be such that $J \geq \hat{J}$ and $J \geq J'$. Then $T^k J \to \hat{J}$, while $T^k J \geq T^k J' = J'$. It follows that $\hat{J} \geq J'$.

(c) If μ is a contractive policy with $J_\mu = \hat{J}$, we have $\hat{J} = J_\mu = T_\mu J_\mu = T_\mu \hat{J}$, so, using also the relation $\hat{J} = T\hat{J}$ [cf. part (a)], we obtain $T_\mu \hat{J} = T\hat{J}$. Conversely, if μ satisfies $T_\mu \hat{J} = T\hat{J}$, then from part (a), we have $T_\mu \hat{J} = \hat{J}$ and hence $\lim_{k\to\infty} T^k_\mu \hat{J} = \hat{J}$. Since μ is contractive, we obtain $J_\mu = \lim_{k\to\infty} T^k_\mu \hat{J}$, so $J_\mu = \hat{J}$. □

Note that it is possible that there exists a noncontractive policy μ that is strictly suboptimal and yet satisfies the optimality condition $T_\mu J^* = TJ^*$ (there are simple deterministic shortest path examples with a zero length cycle that can be used to show this; see [1], Section 3.1.2). Thus contractiveness of μ is an essential assumption in Prop. 6(c).

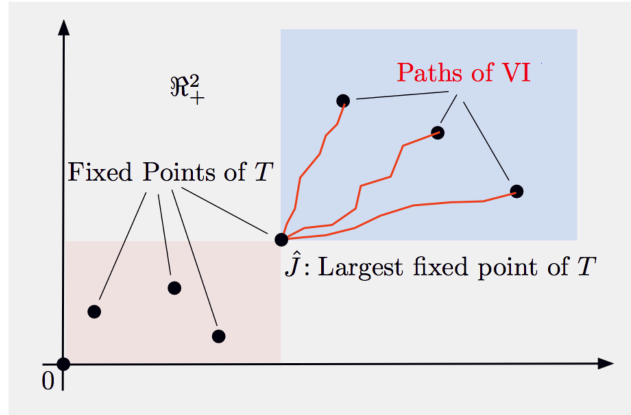

Fig. 1. Schematic illustration of Prop. 6 for a problem with two states. The optimal cost function over contractive policies, $\hat{J}$, is the largest solution of Bellman's equation, while VI converges to $\hat{J}$ starting from $J \geq \hat{J}$.

The following proposition shows that starting from any $J \geq \hat{J}$, the convergence rate of VI to $\hat{J}$ is linear. The proposition also provides a corresponding error bound. The proof is very similar to a corresponding result of [26] and will not be given.

Proposition 7 (Convergence Rate of VI)**.** *Let Assumptions 1 and 2 hold, and assume that there exists a contractive policy $\hat{\mu}$ that is optimal within the class of contractive policies, i.e., $J_{\hat{\mu}} = \hat{J}$. Then*

$$\|TJ - \hat{J}\|_v \leq \beta \|J - \hat{J}\|_v, \qquad \forall\ J \geq \hat{J},$$

where $\|\cdot\|_v$ is a weighted sup-norm for which T_{μ^*} is a contraction and β is the corresponding modulus of contraction. Moreover, we have

$$\|J - \hat{J}\|_v \leq \frac{1}{1-\beta} \max_{i=1,\ldots,n} \frac{J(i) - (TJ)(i)}{v(i)}, \qquad \forall\ J \geq \hat{J}.$$

We note that if $U(i)$ is infinite it is possible that $\hat{J} = J^*$, but the only optimal policy is noncontractive, even if the compactness Assumption 2 holds. This is shown in the following example, which is adapted from the paper [26] (Example 2.1).

Example 1 (A Counterexample on the Existence of an Optimal Contractive Policy)**.** *Consider an exponential cost SSP problem with a single state 1 in addition to the termination state t; cf. Fig. 2. At state 1 we must choose $u \in [0, 1]$. Then, we terminate at no cost [$g(1, u, t) = 0$ in Eq. (10)] with probability u, and we stay at state 1 at cost $-u$ [i.e., $g(1, u, 1) = -u$ in Eq. (10)] with probability $1 - u$. We have $b(i, u) = u \exp(0)$ and $A_{11}(u) = (1 - u) \exp(-u)$, so that*

$$H(1, u, J) = u + (1 - u) \exp(-u) J.$$

Here there is a unique noncontractive policy μ': it chooses $u = 0$ at state 1, and has cost $J_{\mu'}(1) = 1$. Every policy μ with $\mu(1) \in (0, 1]$ is contractive, and J_μ can be obtained by solving the equation $J_\mu = T_\mu J_\mu$, i.e.,

$$J_\mu(1) = \mu(1) + \big(1 - \mu(1)\big) \exp\big(-\mu(1)\big) J_\mu(1).$$

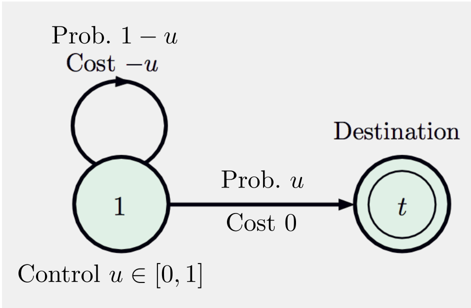

Fig. 2. The exponential cost SSP problem with a single state of Example 1.

We thus obtain

$$J_\mu(1) = \frac{\mu(1)}{\big(1-\mu(1)\big)\exp\big(-\mu(1)\big)}.$$

It can be seen that $\hat{J}(1) = J^*(1) = 0$, but there exists no optimal policy, and no optimal policy within the class of contractive policies.

Let us also show that generally, under Assumptions 1 and 2, J^* need not be a fixed point of T. The following is a straightforward adaptation of Example 2.2 of [26].

Example 2 (An Exponential Cost SSP Problem Where J^* is not a Fixed Point of T)**.** *Consider the exponential cost SSP problem of Fig. 3, involving a noncontractive policy μ whose transitions are marked by solid lines in the figure and form the two zero length cycles shown. All the transitions under μ are deterministic, except at state 1 where the successor state is 2 or 5 with equal probability $1/2$. We assume that the cost of the policy for a given state is the expected value of the exponential of the finite horizon path length. We first calculate $J_\mu(1)$. Let g_k denote the cost incurred at time k, starting at state 1, and let $s_N(1) = \sum_{k=0}^{N-1} g_k$ denote the N-step accumulation of g_k starting from state 1. We have*

$$s_N(1) = 0 \qquad \textit{if } N = 1 \textit{ or } N = 4+3t,\ t = 0,1,\ldots,$$

and

$$s_N(1) = 1 \textit{ or } s_N(1) = -1 \textit{ with probability 1/2 each}$$
$$\textit{if } N = 2+3t \textit{ or } N = 3+3t,\ t = 0,1,\ldots.$$

Thus

$$J_\mu(1) = \limsup_{N\to\infty} E\left\{e^{s_N(1)}\right\} = \frac{1}{2}(e^1 + e^{-1}).$$

On the other hand, a similar (but simpler) calculation shows that

$$J_\mu(2) = J_\mu(5) = e^1,$$

(the N-step accumulation of g_k undergoes a cycle $\{1,-1,0,1,-1,0,\ldots\}$ as N increases starting from state 2, and undergoes a cycle $\{-1,1,0,-1,1,0,\ldots\}$ as N increases starting from state 5). Thus the Bellman equation at state 1,

$$J_\mu(1) = \frac{1}{2}\big(J_\mu(2) + J_\mu(5)\big),$$

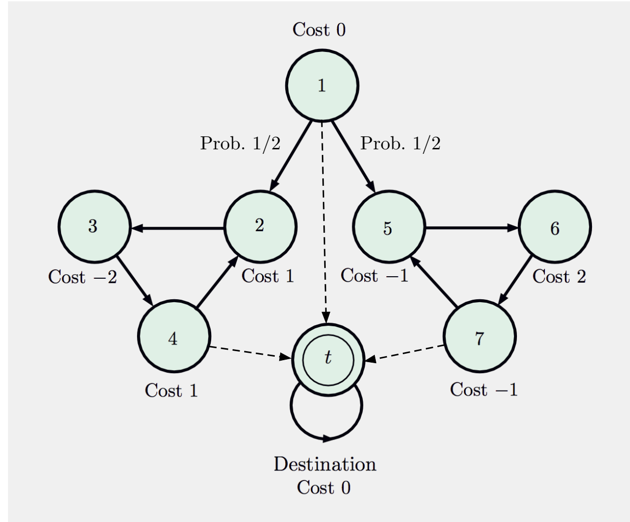

Fig. 3. An example of a noncontractive policy μ, where J_μ is not a fixed point of T_μ. All transitions under μ are shown with solid lines. These transitions are deterministic, except at state 1 where the next state is 2 or 5 with equal probability $1/2$. There are additional transitions from nodes 1, 4, and 7 to the destination (shown with broken lines) with cost $c > 2$, which create a suboptimal contractive policy. We have $J^* = J_\mu$ and J^* is not a fixed point of T.

is not satisfied, and J_μ is not a fixed point of T_μ. If for $i = 1,4,7$, we have transitions (shown with broken lines) that lead from i to t with a cost $c > 2$, the corresponding contractive policy is strictly suboptimal, so that μ is optimal, but $J_\mu = J^*$ is not a fixed point of T.

C. Computational Methods

Regarding computational methods, Prop. 6(b) establishes the validity of the VI algorithm that generates the sequence $\{T^k J\}$ and converges to $\hat{J}$, starting from any initial $J \in \Re^n_+$ with $J \geq \hat{J}$. Moreover, Prop. 7 yields a linear rate of convergence result for this VI algorithm, assuming that there exists a contractive policy $\hat{\mu}$ that is optimal within the class of contractive policies. Convergence to $\hat{J}$ starting from within the region $\{J \mid 0 \leq J \leq \hat{J}\}$ cannot be guaranteed, since there may be fixed points other than $\hat{J}$ within that region. There are also PI algorithms that converge to $\hat{J}$. As an example, we note a PI algorithm with perturbations for abstract DP problems developed in Section 3.3.3 of [1], which can be readily adapted to affine monotonic problems. Finally, it is possible to compute $\hat{J}$ by solving a linear programming problem, in the case where the control space U is finite, by using the following proposition.

Proposition 8. *Let Assumptions 1 and 2 hold. Then if a vector $J \in \Re^n$ satisfies $J \leq TJ$, it also satisfies $J \leq \hat{J}$.*

Proof. Let $J \leq TJ$ and $\delta > 0$. We have $J \leq TJ + \delta e = T_\delta J$, and hence $J \leq T_\delta^k J$ for all k. Since the infinite cost conditions hold for the δ-perturbed problem, it follows that $T_\delta^k J \to J_\delta^*$, so $J \leq J_\delta^*$. By taking $\delta \downarrow 0$ and using Prop. 5(c), it follows that $J \leq \hat{J}$. □

The preceding proposition shows that $\hat{J}$ is the unique solution of the problem of maximizing $\sum_{i=1}^{n} \beta_i J(i)$ over all $J = \big(J(1), \ldots, J(n)\big)$ such that $J \leq TJ$, where $\beta_1, \ldots, \beta_n$ are any positive scalars, i.e., the problem of Eq. (21). This problem is a linear program if each $U(i)$ is a finite set.

V. Concluding Remarks

In this paper we have expanded the SSP methodology to affine monotonic models that are characterized by an affine mapping from the set of nonnegative functions to itself. These models include among others, multiplicative and risk-averse exponentiated cost models. We have used the conceptual framework of semicontractive DP, based on the notion of a contractive policy, which generalizes the notion of a proper policy in SSP. We have provided extensions of the basic analytical and algorithmic results of SSP problems, and we have illustrated their exceptional behavior within our broader context.

Another case of affine monotonic model that we have not considered, is the one obtained when $\bar{J} \leq 0$ and

$$b(i, u) \leq 0, \qquad A_{ij}(u) \geq 0, \qquad \forall\, i, j = 1, \ldots, n,\ u \in U(i),$$

so that T_μ maps the space of nonpositive functions into itself. This case has different character from the case $\bar{J} \geq 0$ and $b(i, u) \geq 0$ of this paper, in analogy with the well-known differences in structure between stochastic optimal control problems with nonpositive and nonnegative cost per stage.

VI. References

[1] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[2] Denardo, E. V., 1967. “Contraction Mappings in the Theory Underlying Dynamic Programming,” SIAM Review, Vol. 9, pp. 165-177.

[3] Denardo, E. V., and Mitten, L. G., 1967. “Elements of Sequential Decision Processes,” J. Indust. Engrg., Vol. 18, pp. 106-112.

[4] Bertsekas, D. P., 1975. “Monotone Mappings in Dynamic Programming,” 1975 IEEE Conference on Decision and Control, pp. 20-25.

[5] Bertsekas, D. P., 1977. “Monotone Mappings with Application in Dynamic Programming,” SIAM J. on Control and Opt., Vol. 15, pp. 438-464.

[6] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y.; may be downloaded from http://web.mit.edu/dimitrib/www/home.html

[7] Blackwell, D., 1965. “Positive Dynamic Programming,” Proc. Fifth Berkeley Symposium Math. Statistics and Probability, pp. 415-418.

[8] Strauch, R., 1966. “Negative Dynamic Programming,” Ann. Math. Statist., Vol. 37, pp. 871-890.

[9] Verdu, S., and Poor, H. V., 1984. “Backward, Forward, and Backward-Forward Dynamic Programming Models under Commutativity Conditions,” Proc. 1984 IEEE Decision and Control Conference, Las Vegas, NE, pp. 1081-1086.

[10] Szepesvari, C., 1998. Static and Dynamic Aspects of Optimal Sequential Decision Making, Ph.D. Thesis, Bolyai Institute of Mathematics, Hungary.

[11] Szepesvari, C., 1998. “Non-Markovian Policies in Sequential Decision Problems,” Acta Cybernetica, Vol. 13, pp. 305-318.

[12] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control, Vol. II, 4th Edition: Approximate Dynamic Programming, Athena Scientific, Belmont, MA.

[13] Bertsekas, D. P., and Yu, H., 2010. “Asynchronous Distributed Policy Iteration in Dynamic Programming,” Proc. of Allerton Conf. on Com., Control and Comp., Allerton Park, Ill, pp. 1368-1374.

[14] Bertsekas, D. P., and Yu, H., 2012. “Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming,” Math. of OR, Vol. 37, pp. 66-94.

[15] Yu, H., and Bertsekas, D. P., 2013. “Q-Learning and Policy Iteration Algorithms for Stochastic Shortest Path Problems,” Annals of Operations Research, Vol. 208, pp. 95-132.

[16] Eaton, J. H., and Zadeh, L. A., 1962. “Optimal Pursuit Strategies in Discrete State Probabilistic Systems,” Trans. ASME Ser. D. J. Basic Eng., Vol. 84, pp.23-29.

[17] Pallu de la Barriere, R., 1967. Optimal Control Theory, Saunders, Phila; republished by Dover, N. Y., 1980.

[18] Derman, C., 1970. Finite State Markovian Decision Processes, Academic Press, N. Y.

[19] Whittle, P., 1982. Optimization Over Time, Wiley, N. Y., Vol. 1, 1982, Vol. 2, 1983.

[20] Kallenberg, L. C. M. 1983. Linear Programming and Finite Markovian Control Problems, MC Tracts 148, Amterdam.

[21] Bertsekas, D. P., 1987. Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, N. J.

[22] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.

[23] Altman, E., 1999. Constrained Markov Decision Processes, CRC Press, Boca Raton, FL.

[24] Hernandez-Lerma, O., and Lasserre, J. B., 1999. Further Topics on Discrete-Time Markov Control Processes, Springer, N. Y.

[25] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. “An Analysis of Stochastic Shortest Path Problems,” Math. of OR, Vol. 16, pp. 580-595.

[26] Bertsekas, D. P., and Yu, H., 2016. “Stochastic Shortest Path Problems Under Weak Conditions,” Lab. for Information and Decision Systems Report LIDS-2909, August 2013, revised March 2015 and January 2016.

[27] Denardo, E. V., and Rothblum, U. G., 1979. “Optimal Stopping, Exponential Utility, and Linear Programming,” Math. Programming, Vol. 16, pp. 228-244.

[28] Patek, S. D., 2001. “On Terminating Markov Decision Processes with a Risk Averse Objective Function,” Automatica, Vol. 37, pp. 1379-1386.

[29] Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming, J. Wiley, N. Y.

[30] Rothblum, U. G., 1984. "Multiplicative Markov Decision Chains," Math. of OR, Vol. 9, pp. 6-24.

[31] Veinott, A. F., Jr., 1969. "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," Ann. Math. Statist., Vol. 40, pp. 1635-1660.

[32] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[33] Tseng, P., 1990. "Solving H-Horizon, Stationary Markov Decision Problems in Time Proportional to $\log(H)$," Operations Research Letters, Vol. 9, pp. 287-297.

[34] Littman, M. L., 1996. Algorithms for Sequential Decision Making, Ph.D. thesis, Brown University, Providence, R. I.

[35] Yu, H., and Bertsekas, D. P., 2013. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," arXiv preprint arXiv:1308.3814, to appear in Math. of OR.

PLACE
PHOTO
HERE

Dimitri P. Bertsekas studied engineering at the National Technical University of Athens, Greece, obtained his MS in electrical engineering at the George Washington University, Wash. DC in 1969, and his Ph.D. in system science in 1971 at the Massachusetts Institute of Technology.

Dr. Bertsekas has held faculty positions with the Engineering-Economic Systems Dept., Stanford University (1971-1974) and the Electrical Engineering Dept. of the University of Illinois, Urbana (1974-1979). Since 1979 he has been teaching at the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (M.I.T.), where he is currently McAfee Professor of Engineering. He consults regularly with private industry and has held editorial positions in several journals. His research at M.I.T. spans several fields, including optimization, control, large-scale computation, and data communication networks, and is closely tied to his teaching and book authoring activities. He has written numerous research papers, and fourteen books, several of which are used as textbooks in MIT classes.

Professor Bertsekas was awarded the INFORMS 1997 Prize for Research Excellence in the Interface Between Operations Research and Computer Science for his book "Neuro-Dynamic Programming" (co-authored with John Tsitsiklis), the 2001 ACC John R. Ragazzini Education Award, the 2009 INFORMS Expository Writing Award, the 2014 ACC Richard E. Bellman Control Heritage Award, the 2014 Khachiyan Prize, and the SIAM/MOS 2015 George B. Dantzig Prize. In 2001, he was elected to the United States National Academy of Engineering.

Dr. Bertsekas' recent books are "Dynamic Programming and Optimal Control: 4th Edition" (2017), "Noninear Programming: 3rd Edition" (2016), "Convex Optimization Algorithms" (2015), and "Abstract Dynamic Programming" (2013) all published by Athena Scientific.