May 2012 **Report LIDS - 2884**

Weighted Sup-Norm Contractions in Dynamic Programming: A Review and Some New Applications

Dimitri P. Bertsekas†

Abstract

We consider a class of generalized dynamic programming models based on weighted sup-norm contractions. We provide an analysis that parallels the one available for discounted MDP and for generalized models based on unweighted sup-norm contractions. In particular, we discuss the main properties and associated algorithms of these models, including value iteration, policy iteration, and their optimistic and approximate variants. The analysis relies on several earlier works that use more specialized assumptions. In particular, we review and extend the classical results of Denardo [Den67] for unweighted sup-norm contraction models, as well as more recent results relating to approximation methods for discounted MDP. We also apply the analysis to stochastic shortest path problems where all policies are assumed proper. For these problems we extend three results that are known for discounted MDP. The first relates to the convergence of optimistic policy iteration and extends a result of Rothblum [Rot79], the second relates to error bounds for approximate policy iteration and extends a result of Bertsekas and Tsitsiklis [BeT96], and the third relates to error bounds for approximate optimistic policy iteration and extends a result of Thiery and Scherrer [ThS10b].

† Dimitri Bertsekas is with the Dept. of Electr. Engineering and Comp. Science, M.I.T., Cambridge, Mass., 02139. His research was supported by NSF Grant ECCS-0801549, and by the Air Force Grant FA9550-10-1-0412.

1. INTRODUCTION

Two key structural properties of total cost dynamic programming (DP) models are responsible for most of the mathematical results one can prove about them. The first is the *monotonicity property* of the mappings associated with Bellman's equation. In many models, however, these mappings have another property that strengthens the effects of monotonicity: they are *contraction mappings* with respect to a sup-norm, unweighted in many models such as discounted finite spaces Markovian decision problems (MDP), but also weighted in some other models, discounted or undiscounted. An important case of the latter are stochastic shortest path (SSP) problems under certain conditions to be discussed in Section 7.

The role of contraction mappings in discounted DP was first recognized and exploited by Shapley [Sha53], who considered two-player dynamic games. Since that time the underlying contraction properties of discounted DP problems have been explicitly or implicitly used by most authors that have dealt with the subject. An abstract DP model, based on unweighted sup-norm contraction assumptions, was introduced in an important paper by Denardo [Den67]. This model provided generality and insight into the principal analytical and algorithmic ideas underlying the discounted DP research up to that time. Denardo's model motivated a related model by the author [Ber77], which relies only on monotonicity properties, and not on contraction assumptions. These two models were used extensively in the book by Bertsekas and Shreve [BeS78] for the analysis of both discounted and undiscounted DP problems, ranging over MDP, minimax, risk sensitive, Borel space models, and models based on outer integration. Related analysis, motivated by problems in communications, was given by Verd'u and Poor [VeP84], [VeP87]. See also Bertsekas and Yu [BeY10b], which considers policy iteration methods using the abstract DP model of [Ber77].

In this paper, we extend Denardo's model to weighted sup-norm contractions, and we provide a full set of analytical and algorithmic results that parallel the classical ones for finite-spaces discounted MDP, as well as some of the corresponding results for unweighted sup-norm contractions. These results include extensions of relatively recent research on approximation methods, which have been shown for discounted MDP with bounded cost per stage. Our motivation stems from the fact that there are important discounted DP models with unbounded cost per stage, as well as undiscounted DP models of the SSP type, where there is contraction structure that requires, however, a weighted sup-norm. We obtain among others, three new algorithmic results for SSP problems, which are given in Section 7. The first relates to the convergence of optimistic (also commonly referred to as "modified" [Put94]) policy iteration, and extends the one originally proved by Rothblum [Rot79] within Denardo's unweighted sup-norm contraction framework. The second relates to error bounds for approximate policy iteration, and extends a result of Bertsekas and Tsitsiklis [BeT96] (Prop. 6.2), given for discounted MDP, and improves on another result of [BeT96] (Prop. 6.3) for SSP. The third relates to error bounds for approximate optimistic policy iteration, and extends a result of Thiery and Scherrer [ThS10a], [ThS10b], given for discounted MDP. A recently derived error bound for a Q-learning framework for optimistic policy iteration in SSP problems, due to Yu and Bertsekas [YuB11], can also be proved using our framework.

2. A WEIGHTED SUP-NORM CONTRACTION FRAMEWORK FOR DP

Let X and U be two sets, which in view of connections to DP that will become apparent shortly, we will loosely refer to as a set of "states" and a set of "controls." For each $x \in X$, let $U(x) \subset U$ be a nonempty subset of controls that are feasible at state x. Consistent with the DP context, we refer to a function $\mu : X \mapsto U$ with $\mu(x) \in U(x)$, for all $x \in X$, as a "policy." We denote by $\mathcal{M}$ the set of all policies.

Let $R(X)$ be the set of real-valued functions $J : X \mapsto \Re$, and let $H : X \times U \times R(X) \mapsto \Re$ be a given mapping. We consider the mapping T defined by

$$(TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \qquad \forall\ x \in X.$$

We assume that $(TJ)(x) > -\infty$ for all $x \in X$, so that T maps $R(X)$ into $R(X)$. For each policy $\mu \in \mathcal{M}$, we consider the mapping $T_\mu : R(X) \mapsto R(X)$ defined by

$$(T_\mu J)(x) = H\big(x, \mu(x), J\big), \qquad \forall\ x \in X.$$

We want to find a function $J^* \in R(X)$ such that

$$J^*(x) = \inf_{u \in U(x)} H(x, u, J^*), \qquad \forall\ x \in X,$$

i.e., find a fixed point of T. We also want to obtain a policy μ^* such that $T_{\mu^*} J^* = TJ^*$.

Note that in view of the preceding definitions, H may be alternatively defined by first specifying T_μ for all $\mu \in \mathcal{M}$ [for any (x, u, J), $H(x, u, J)$ is equal to $(T_\mu J)(x)$ for any μ such $\mu(x) = u$]. Moreover T may be defined by

$$(TJ)(x) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x), \qquad \forall\ x \in X,\ J \in R(X).$$

We give a few examples.

Example 2.1 (Discounted DP Problems)

Consider an α-discounted total cost DP problem. Here

$$H(x, u, J) = E\big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\},$$

where $\alpha \in (0, 1)$, g is a uniformly bounded function representing cost per stage, w is random with distribution that may depend on (x, u), and is taken with respect to that distribution. The equation $J = TJ$, i.e.,

$$J(x) = \inf_{u \in U(x)} H(x, u, J) = \inf_{u \in U(x)} E\big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\}, \qquad \forall\ x \in X,$$

is Bellman's equation, and it is known to have unique solution J^*. Variants of the above mapping H are

$$H(x, u, J) = \min\big[V(x),\ E\big\{g(x, u, w) + \alpha J\big(f(x, u, w)\big)\big\}\big],$$

and

$$H(x, u, J) = E\big\{g(x, u, w) + \alpha \min\big[V\big(f(x, u, w)\big),\ J\big(f(x, u, w)\big)\big]\big\},$$

where V is a known function that satisfies $V(x) \geq J^*(x)$ for all $x \in X$. While the use of V in these variants of H does not affect the solution J^*, it may affect favorably the value and policy iteration algorithms to be discussed in subsequent sections.

Example 2.2 (Discounted Semi-Markov Problems)

With x, y, u as in Example 2.1, consider the mapping

$$H(x, u, J) = G(x, u) + \sum_{y=1}^{n} m_{xy}(u) J(y),$$

where G is some function representing cost per stage, and $m_{xy}(u)$ are nonnegative numbers with $\sum_{y=1}^{n} m_{xy}(u) < 1$ for all $x \in X$ and $u \in U(x)$. The equation $J = TJ$ is Bellman's equation for a continuous-time semi-Markov decision problem, after it is converted into an equivalent discrete-time problem.

Example 2.3 (Minimax Problems)

Consider a minimax version of Example 2.1, where an antagonistic player chooses v from a set $V(x,u)$, and let

$$H(x,u,J)=\sup_{v\in V(x,u)}\Big[g(x,u,v)+\alpha J\big(f(x,u,v)\big)\Big].$$

Then the equation $J=TJ$ is Bellman's equation for an infinite horizon minimax DP problem. A generalization is a mapping of the form

$$H(x,u,J)=\sup_{v\in V(x,u)}E\Big\{g(x,u,v,w)+\alpha J\big(f(x,u,v,w)\big)\Big\},$$

where w is random with given distribution, and the expected value is with respect to that distribution. This form appears in zero-sum sequential games [Sha53].

Example 2.4 (Deterministic and Stochastic Shortest Path Problems)

Consider a classical deterministic shortest path problem involving a graph of n nodes $x=1,\ldots,n$, plus a destination 0, an arc length a_{xu} for each arc (x,u), and the mapping

$$H(x,u,J)=\begin{cases} a_{xu}+J(u) & \text{if } u\neq 0,\\ a_{0t} & \text{if } u=0,\end{cases}\qquad x=1,\ldots,n,\ u=0,1,\ldots,n.$$

Then the equation $J=TJ$ is Bellman's equation for the shortest distances $J^*(x)$ from the nodes x to node 0. A generalization is a mapping of the form

$$H(x,u,J)=p_{x0}(u)g(x,u,0)+\sum_{y=1}^{n}p_{xy}(u)\big(g(x,u,y)+J(y)\big),\qquad x=1,\ldots,n.$$

It corresponds to a SSP problem, which is described in Section 7. A special case is stochastic finite-horizon, finite-state DP problems.

Example 2.5 (Q-Learning I)

Consider the case where X is the set of state-control pairs (i,w), $i=1,\ldots,n$, $w\in W(i)$, of an MDP with controls w taking values at state i from a finite set $W(i)$. Let T_μ map a Q-factor vector

$$Q=\big\{Q(i,w)\mid i=1,\ldots,n,\ w\in W(i)\big\}$$

into the Q-factor vector

$$\bar{Q}_\mu=\big\{\bar{Q}_\mu(i,w)\mid i=1,\ldots,n,\ w\in W(i)\big\}$$

with components given by

$$\bar{Q}_\mu(i,w)=g(i,w)+\alpha\sum_{j=1}^{n}p_{ij}\big(\mu(i)\big)\min_{v\in W(j)}Q(j,v),\qquad i=1,\ldots,n,\ w\in W(i).$$

This mapping corresponds to the classical Q-learning mapping of a finite-state MDP [in relation to the standard Q-learning framework, [Tsi94], [BeT96], [SuB98], μ applies a control $\mu(i)$ from the set $U(i,w)=W(i)$ independently of the value of $w\in W(i)$]. If $\alpha\in(0,1)$, the MDP is discounted, while if $\alpha=1$, the MDP is undiscounted and when there is a cost-free and absorbing state, it has the character of the SSP problem of the preceding example.

Example 2.6 (Q-Learning II)

Consider an alternative Q-learning framework introduced in [BeY10a] for discounted MDP and in [YuB11] for SSP, where T_μ operates on pairs (Q, V), and using the notation of the preceding example, Q is a Q-factor and V is a cost vector of the forms

$$\bigl\{Q(i,w) \mid i = 1,\ldots,n,\ w \in W(i)\bigr\}, \qquad \bigl\{V(i) \mid i = 1,\ldots,n\bigr\}.$$

Let T_μ map a pair (Q, V) into the pair $(\bar{Q}_\mu, \bar{V}_\mu)$ with components given by

$$\bar{Q}_\mu(i,w) = g(i,w) + \alpha \sum_{j=1}^{n} p_{ij}\bigl(\mu(i)\bigr)\nu(v \mid j) \min\bigl[V(j),\, Q(j,v)\bigr], \qquad i = 1,\ldots,n,\, w \in W(i),$$

$$\bar{V}_\mu(i) = \min_{w \in W(i)} \bar{Q}_\mu(i,w), \qquad i = 1,\ldots,n,$$

where $\nu(\cdot \mid j)$ is a given conditional distribution over $W(j)$, and $\alpha \in (0,1)$ for a discounted MDP and $\alpha = 1$ for an SSP problem.

We also note a variety of discounted countable-state MDP models with unbounded cost per stage, whose Bellman equation mapping involves a weighted sup-norm contraction. Such models are described in several sources, starting with works of Harrison [Har72], and Lippman [Lip73], [Lip75] (see also [Ber12], Section 1.5, and [Put94], and the references quoted there).

Consider a function $v : X \mapsto \Re$ with

$$v(x) > 0, \qquad \forall\, x \in X,$$

denote by $B(X)$ the space of real-valued functions J on X such that $J(x)/v(x)$ is bounded as x ranges over X, and consider the weighted sup-norm

$$\|J\| = \sup_{x \in X} \frac{\bigl|J(x)\bigr|}{v(x)}$$

on $B(X)$. We will use the following assumption.

Assumption 2.1: (Contraction) For all $J \in B(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $B(X)$. Furthermore, for some $\alpha \in (0,1)$, we have

$$\|T_\mu J - T_\mu J'\| \le \alpha \|J - J'\|, \qquad \forall\, J, J' \in B(X),\ \mu \in \mathcal{M}. \tag{2.1}$$

An equivalent way to state the condition (2.1) is

$$\frac{\bigl|H(x,u,J) - H(x,u,J')\bigr|}{v(x)} \le \alpha \|J - J'\|, \qquad \forall\, x \in X,\, u \in U(x),\ J, J' \in B(X).$$

Note that Eq. (2.1) implies that

$$\|TJ - TJ'\| \le \alpha \|J - J'\|, \qquad \forall\, J, J' \in B(X). \tag{2.2}$$

To see this we write

$$(T_\mu J)(x) \le (T_\mu J')(x) + \alpha \|J - J'\|\, v(x), \qquad \forall\ x \in X,$$

from which, by taking infimum of both sides over $\mu \in \mathcal{M}$, we have

$$\frac{(TJ)(x) - (TJ')(x)}{v(x)} \le \alpha \|J - J'\|, \qquad \forall\ x \in X.$$

Reversing the roles of J and J', we also have

$$\frac{(TJ')(x) - (TJ)(x)}{v(x)} \le \alpha \|J - J'\|, \qquad \forall\ x \in X,$$

and combining the preceding two relations, and taking the supremum of the left side over $x \in X$, we obtain Eq. (2.2).

It can be seen that the Contraction Assumption 2.1 is satisfied for the mapping H in Examples 2.1-2.3, and the discounted cases of 2.5-2.6, with v equal to the unit function, i.e., $v(x) \equiv 1$. Generally, the assumption is not satisfied in Example 2.4, and the undiscounted cases of Examples 2.5-2.6, but it will be seen later that it is satisfied for the special case of the SSP problem under the assumption that all stationary policies are proper (lead to the destination with probability 1, starting from every state). In that case, however, we cannot take $v(x) \equiv 1$, and this is one of our motivations for considering the more general case where v is not the unit function.

The next two examples show how starting with mappings satisfying the contraction assumption, we can obtain multistep mappings with the same fixed points and a stronger contraction modulus. For any $J \in R(X)$, we denote by $T_{\mu_0} \cdots T_{\mu_k} J$ the composition of the mappings $T_{\mu_0}, \ldots, T_{\mu_k}$ applied to J, i.e,

$$T_{\mu_0} \cdots T_{\mu_k} J = \big(T_{\mu_0}\big(T_{\mu_1} \cdots (T_{\mu_{k-1}}(T_{\mu_k} J)) \cdots\big)\big).$$

Example 2.7 (Multistep Mappings)

Consider a set of mappings $T_\mu : \Re^n \mapsto \Re^n$, $\mu \in \mathcal{M}$, satisfying Assumption 2.1, let m be a positive integer, and let $\bar{\mathcal{M}}$ be the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, where $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$. For each $\nu = (\mu_0, \ldots, \mu_{m-1}) \in \bar{\mathcal{M}}$, define the mapping $\bar{T}_\nu$, by

$$\bar{T}_\nu J = T_{\mu_0} \cdots T_{\mu_{m-1}} J, \qquad \forall\ J \in B(X).$$

Then we have the contraction properties

$$\|\bar{T}_\nu J - \bar{T}_\nu J'\| \le \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

and

$$\|\bar{T} J - \bar{T} J'\| \le \alpha^m \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

where $\bar{T}$ is defined by

$$(\bar{T} J)(x) = \inf_{(\mu_0, \ldots, \mu_{m-1}) \in \bar{\mathcal{M}}} (T_{\mu_0} \cdots T_{\mu_{m-1}} J)(x), \qquad \forall\ J \in B(X),\ x \in X.$$

Thus the mappings $\bar{T}_\nu$, $\nu \in \bar{\mathcal{M}}$, satisfy Assumption 2.1, and have contraction modulus α^m.

The following example considers mappings underlying weighted Bellman equations that arise in various computational contexts in approximate DP; see Yu and Bertsekas [YuB12] for analysis, algorithms, and related applications.

Example 2.8 (Weighted Multistep Mappings)

Consider a set of mappings $L_\mu : B(X) \mapsto B(X)$, $\mu \in \mathcal{M}$, satisfying Assumption 2.1, i.e., for some $\alpha \in (0,1)$,

$$\|L_\mu J - L_\mu J'\| \leq \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X),\ \mu \in \mathcal{M}.$$

Consider also the mappings $T_\mu : B(X) \mapsto B(X)$ defined by

$$(T_\mu J)(x) = \sum_{\ell=1}^{\infty} w_\ell(x)(L_\mu^\ell J)(x), \qquad x \in X,\ J \in \Re^n,$$

where $w_\ell(x)$ are nonnegative scalars such that for all $x \in X$,

$$\sum_{\ell=1}^{\infty} w_\ell(x) = 1.$$

Then it follows that

$$\|T_\mu J - T_\mu J'\| \leq \sum_{\ell=1}^{\infty} w_\ell(x)\alpha^\ell \|J - J'\|,$$

showing that T_μ is a contraction with modulus

$$\bar{\alpha} = \max_{x \in X} \sum_{\ell=1}^{\infty} w_\ell(x)\, \alpha^\ell \leq \alpha.$$

Moreover L_μ and T_μ have a common fixed point for all $\mu \in \mathcal{M}$, and the same is true for the corresponding mappings L and T.

We will now consider some general questions, first under the Contraction Assumption 2.1, and then under an additional monotonicity assumption. Most of the results of this section are straightforward extensions of results that appear in Denardo's paper [Den67] for the case where the sup-norm is unweighted [$v(x) \equiv 1$].

2.1 Basic Results Under the Contraction Assumption

The contraction property of T_μ and T can be used to show the following proposition.

Proposition 2.1: Let Assumption 2.1 hold. Then:

(a) The mappings T_μ and T are contraction mappings with modulus α over $B(X)$, and have unique fixed points in $B(X)$, denoted J_μ and J^*, respectively.

(b) For any $J \in B(X)$ and $\mu \in \mathcal{M}$,

$$\lim_{k \to \infty} T_\mu^k J = J_\mu, \qquad \lim_{k \to \infty} T^k J = J^*.$$

(c) We have $T_\mu J^* = TJ^*$ if and only if $J_\mu = J^*$.

(d) For any $J \in B(X)$,

$$\|J^* - J\| \le \frac{1}{1-\alpha}\|TJ - J\|, \qquad \|J^* - TJ\| \le \frac{\alpha}{1-\alpha}\|TJ - J\|.$$

(e) For any $J \in B(X)$ and $\mu \in \mathcal{M}$,

$$\|J_\mu - J\| \le \frac{1}{1-\alpha}\|T_\mu J - J\|, \qquad \|J_\mu - T_\mu J\| \le \frac{\alpha}{1-\alpha}\|T_\mu J - J\|.$$

Proof: We have already shown that T_μ and T are contractions with modulus α over $B(X)$ [cf. Eqs. (2.1) and (2.2)]. Parts (a) and (b) follow from the classical contraction mapping fixed point theorem. To show part (c), note that if $T_\mu J^* = TJ^*$, then in view of $TJ^* = J^*$, we have $T_\mu J^* = J^*$, which implies that $J^* = J_\mu$, since J_μ is the unique fixed point of T_μ. Conversely, if $J_\mu = J^*$, we have $T_\mu J^* = T_\mu J_\mu = J_\mu = J^* = TJ^*$.

To show part (d), we use the triangle inequality to write for every k,

$$\|T^k J - J\| \le \sum_{\ell=1}^{k} \|T^\ell J - T^{\ell-1} J\| \le \sum_{\ell=1}^{k} \alpha^{\ell-1}\|TJ - J\|.$$

Taking the limit as $k \to \infty$ and using part (b), the left-hand side inequality follows. The right-hand side inequality follows from the left-hand side and the contraction property of T. The proof of part (e) is similar to part (d) [indeed part (e) is the special case of part (d) where T is equal to T_μ, i.e., when $U(x) = \big\{\mu(x)\big\}$ for all $x \in X$]. **Q.E.D.**

Part (c) of the preceding proposition shows that there exists a $\mu \in \mathcal{M}$ such that $J_\mu = J^*$ if and only if the minimum of $H(x, u, J^*)$ over $U(x)$ is attained for all $x \in X$. Of course the minimum is attained if $U(x)$ is finite for every x, but otherwise this is not guaranteed in the absence of additional assumptions. Part (d) provides a useful error bound: we can evaluate the proximity of any function $J \in B(X)$ to the fixed point J^* by applying T to J and computing $\|TJ - J\|$. The left-hand side inequality of part (e) (with $J = J^*$) shows that for every $\epsilon > 0$, there exists a $\mu_\epsilon \in \mathcal{M}$ such that $\|J_{\mu_\epsilon} - J^*\| \le \epsilon$, which may be obtained by letting $\mu_\epsilon(x)$ minimize $H(x, u, J^*)$ over $U(x)$ within an error of $(1-\alpha)\epsilon\, v(x)$, for all $x \in X$.

2.2 The Role of Monotonicity

Our analysis so far in this section relies only on the contraction assumption. We now introduce a monotonicity property of a type that is common in DP.

Assumption 2.2: (Monotonicity) If $J, J' \in R(X)$ and $J \leq J'$, then

$$H(x,u,J) \leq H(x,u,J'), \qquad \forall\, x \in X,\ u \in U(x). \tag{2.3}$$

Note that the assumption is equivalent to

$$J \leq J' \quad \Rightarrow \quad T_\mu J \leq T_\mu J', \qquad \forall\, \mu \in \mathcal{M},$$

and implies that

$$J \leq J' \quad \Rightarrow \quad TJ \leq TJ'.$$

An important consequence of monotonicity of H, when it holds in addition to contraction, is that it implies an optimality property of J^*.

Proposition 2.2: Let Assumptions 2.1 and 2.2 hold. Then

$$J^*(x) = \inf_{\mu \in \mathcal{M}} J_\mu(x), \qquad \forall\, x \in X. \tag{2.4}$$

Furthermore, for every $\epsilon > 0$, there exists $\mu_\epsilon \in \mathcal{M}$ such that

$$J^*(x) \leq J_{\mu_\epsilon}(x) \leq J^*(x) + \epsilon\, v(x), \qquad \forall\, x \in X. \tag{2.5}$$

Proof: We note that the right-hand side of Eq. (2.5) holds by Prop. 2.1(e) (see the remark following its proof). Thus $\inf_{\mu \in \mathcal{M}} J_\mu(x) \leq J^*(x)$ for all $x \in X$. To show the reverse inequality as well as the left-hand side of Eq. (2.5), we note that for all $\mu \in \mathcal{M}$, we have $TJ^* \leq T_\mu J^*$, and since $J^* = TJ^*$, it follows that $J^* \leq T_\mu J^*$. By applying repeatedly T_μ to both sides of this inequality and by using the Monotonicity Assumption 2.2, we obtain $J^* \leq T_\mu^k J^*$ for all $k > 0$. Taking the limit as $k \to \infty$, we see that $J^* \leq J_\mu$ for all $\mu \in \mathcal{M}$. **Q.E.D.**

Propositions 2.1 and 2.2 collectively address the problem of finding $\mu \in \mathcal{M}$ that minimizes $J_\mu(x)$ simultaneously for all $x \in X$, consistently with DP theory. The optimal value of this problem is $J^*(x)$, and μ is optimal for all x if and only if $T_\mu J^* = TJ^*$. For this we just need the contraction and monotonicity assumptions. We do not need any additional structure of H, such as for example a discrete-time dynamic system, transition probabilities, etc. While identifying the proper structure of H and verifying its contraction and monotonicity properties may require some analysis that is specific to each type of problem, once this is done significant results are obtained quickly.

Note that without monotonicity, we may have $\inf_{\mu\in\mathcal{M}} J_\mu(x) < J^*(x)$ for some x. As an example, let $X = \{x_1, x_2\}$, $U = \{u_1, u_2\}$, and let

$$H(x_1,u,J) = \begin{cases} -\alpha J(x_2) & \text{if } u = u_1, \\ -1+\alpha J(x_1) & \text{if } u = u_2, \end{cases} \qquad H(x_2,u,J) = \begin{cases} 0 & \text{if } u = u_1, \\ B & \text{if } u = u_2, \end{cases}$$

where B is a positive scalar. Then it can be seen that

$$J^*(x_1) = -\frac{1}{1-\alpha}, \qquad J^*(x_2) = 0,$$

and $J_{\mu^*} = J^*$ where $\mu^*(x_1) = u_2$ and $\mu^*(x_2) = u_1$. On the other hand, for $\mu(x_1) = u_1$ and $\mu(x_2) = u_2$, we have

$$J_\mu(x_1) = -\alpha B, \qquad J_\mu(x_2) = B,$$

so $J_\mu(x_1) < J^*(x_1)$ for B sufficiently large.

Nonstationary Policies

The connection with DP motivates us to consider the set Π of all sequences $\pi = \{\mu_0, \mu_1, \ldots\}$ with $\mu_k \in \mathcal{M}$ for all k (nonstationary policies in the DP context), and define

$$J_\pi(x) = \liminf_{k\to\infty} (T_{\mu_0} \cdots T_{\mu_k} J)(x), \qquad \forall\, x \in X,$$

with J being any function in $B(X)$, where as earlier, $T_{\mu_0} \cdots T_{\mu_k} J$ denotes the composition of the mappings $T_{\mu_0}, \ldots, T_{\mu_k}$ applied to J. Note that the choice of J in the definition of J_π does not matter since for any two $J, J' \in B(X)$, we have from the Contraction Assumption 2.1,

$$\|T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J - T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J'\| \le \alpha^{k+1} \|J - J'\|,$$

so the value of $J_\pi(x)$ is independent of J. Since by Prop. 2.1(b), $J_\mu(x) = \lim_{k\to\infty} (T_\mu^k J)(x)$ for all $\mu \in \mathcal{M}$, $J \in B(X)$, and $x \in X$, in the DP context we recognize J_μ as the cost function of the stationary policy $\{\mu, \mu, \ldots\}$.

We now claim that under our Assumptions 2.1 and 2.2, J^*, the fixed point of T, is equal to the optimal value of J_π, i.e.,

$$J^*(x) = \inf_{\pi\in\Pi} J_\pi(x), \qquad \forall\, x \in X.$$

Indeed, since $\mathcal{M}$ defines a subset of Π, we have from Prop. 2.2,

$$J^*(x) = \inf_{\mu\in\mathcal{M}} J_\mu(x) \ge \inf_{\pi\in\Pi} J_\pi(x), \qquad \forall\, x \in X,$$

while for every $\pi \in \Pi$ and $x \in X$, we have

$$J_\pi(x) = \liminf_{k\to\infty} (T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J)(x) \ge \lim_{k\to\infty} (T^{k+1} J)(x) = J^*(x)$$

[the Monotonicity Assumption 2.2 can be used to show that

$$T_{\mu_0} T_{\mu_1} \cdots T_{\mu_k} J \ge T^{k+1} J,$$

and the last equality holds by Prop. 2.1(b)]. Combining the preceding relations, we obtain $J^*(x) = \inf_{\pi\in\Pi} J_\pi(x)$.

Thus, in DP terms, we may view J^* as an optimal cost function over all nonstationary policies. At the same time, Prop. 2.2 states that stationary policies are sufficient in the sense that the optimal cost can be attained to within arbitrary accuracy with a stationary policy [uniformly for all $x \in X$, as Eq. (2.5) shows].

Periodic Policies

Consider the multistep mappings $\bar{T}_\nu = T_{\mu_0} \cdots T_{\mu_{m-1}}$, $\nu \in \bar{\mathcal{M}}$, defined in Example 2.7, where $\bar{\mathcal{M}}$ is the set of m-tuples $\nu = (\mu_0, \ldots, \mu_{m-1})$, with $\mu_k \in \mathcal{M}$, $k = 1, \ldots, m-1$. Assuming that the mappings T_μ satisfy Assumptions 2.1 and 2.2, the same is true for the mappings $\bar{T}_\nu$ (with the contraction modulus of $\bar{T}_\nu$ being α^m). Thus the unique fixed point of $\bar{T}_\nu$ is J_π, where π is the nonstationary but periodic policy

$$\pi = \{\mu_0, \ldots, \mu_{m-1}, \mu_0, \ldots, \mu_{m-1}, \ldots\}.$$

Moreover the mappings $T_{\mu_0} \cdots T_{\mu_{m-1}}$, $T_{\mu_1} \cdots T_{\mu_{m-1}} T_{\mu_0}, \ldots, T_{\mu_{m-1}} T_{\mu_0} \cdots T_{\mu_{m-2}}$, have unique corresponding fixed points $J_0, J_1, \ldots, J_{m-1}$, which satisfy

$$J_0 = T_{\mu_1} J_1, \quad J_1 = T_{\mu_2} J_2, \quad \ldots \quad J_{\mu_{m-2}} = T_{\mu_{m-1}} J_{\mu_{m-1}}, \quad J_{\mu_{m-1}} = T_{\mu_0} J_0.$$

To verify the above equations, multiply the relation $J_1 = T_{\mu_1} \cdots T_{\mu_{m-1}} T_{\mu_0} J_1$ with T_{μ_0} to show that $T_{\mu_0} J_1$ is the fixed point of $T_{\mu_0} \cdots T_{\mu_{m-1}}$, i.e., is equal to J_0, etc. Note that even though $\bar{T}_\nu$ defines the cost functions of periodic policies, $\bar{T}$ has the same fixed point as T, namely J^*. This gives rise to the computational possibility of working with $\bar{T}_\nu$ in place of T_μ in an effort to find J^*. Moreover, periodic policies obtained through approximation methods, such as the ones to be discussed in what follows, may hold an advantage over stationary policies, as first shown by Scherrer [Sch12] in the context of approximate value iteration (see also the discussion of Section 4).

Error Bounds Under Monotonicity

The assumptions of contraction and monotonicity together can be characterized in a form that is useful for the derivation of error bounds.

Proposition 2.3: The Contraction and Monotonicity Assumptions 2.1 and 2.2 hold if and only if for all $J, J' \in B(X)$, $\mu \in \mathcal{M}$, and scalar $c \geq 0$, we have

$$J' \leq J + c\,v \quad \Rightarrow \quad T_\mu J' \leq T_\mu J + \alpha c\,v, \tag{2.6}$$

where v is the weight function of the weighted sup-norm $\|\cdot\|$.

Proof: Let the contraction and monotonicity assumptions hold. If $J' \leq J + c\,v$, we have

$$H(x, u, J') \leq H(x, u, J + c\,v) \leq H(x, u, J) + \alpha c\,v(x), \qquad \forall\, x \in X,\ u \in U(x), \tag{2.7}$$

where the left-side inequality follows from the monotonicity assumption and the right-side inequality follows from the contraction assumption, which together with $\|v\| = 1$, implies that

$$\frac{H(x, u, J + c\,v) - H(x, u, J)}{v(x)} \leq \alpha \|J + c\,v - J\| = \alpha c.$$

The condition (2.7) implies the desired condition (2.6). Conversely, condition (2.6) for $c = 0$ yields the monotonicity assumption, while for $c = \|J' - J\|$ it yields the contraction assumption. **Q.E.D.**

We can use Prop. 2.3 to derive some useful variants of parts (d) and (e) of Prop. 2.1 (which assumes only the contraction assumption). These variants will be used in the derivation of error bounds for computational methods to be discussed in Sections 4-6.

Proposition 2.4: (Error Bounds Under Contraction and Monotonicity) Let Assumptions 2.1 and 2.2 hold.

(a) For any $J \in B(X)$ and $c \geq 0$, we have

$$TJ \leq J + c\,v \quad \Rightarrow \quad J^* \leq J + \frac{c}{1-\alpha} v,$$

$$J \leq TJ + c\,v \quad \Rightarrow \quad J \leq J^* + \frac{c}{1-\alpha} v.$$

(b) For any $J \in B(X)$, $\mu \in \mathcal{M}$, and $c \geq 0$, we have

$$T_\mu J \leq J + c\,v \quad \Rightarrow \quad J_\mu \leq J + \frac{c}{1-\alpha} v,$$

$$J \leq T_\mu J + c\,v \quad \Rightarrow \quad J \leq J_\mu + \frac{c}{1-\alpha} v.$$

(c) For all $J \in B(X)$, $c \geq 0$, and $k = 0, 1, \ldots$, we have

$$TJ \leq J + c\,v \qquad \Rightarrow \qquad J^* \leq T^k J + \frac{\alpha^k c}{1-\alpha} v,$$

$$J \leq TJ + c\,v \qquad \Rightarrow \qquad T^k J \leq J^* + \frac{\alpha^k c}{1-\alpha} v.$$

Proof: (a) We show the first relation. Applying Eq. (2.6) with J' replaced by TJ, and taking minimum over $u \in U(x)$ for all $x \in X$, we see that if $TJ \leq J + c\,v$, then $T^2 J \leq TJ + \alpha c\,v$. Proceeding similarly, it follows that

$$T^\ell J \leq T^{\ell-1} J + \alpha^{\ell-1} \epsilon\, v.$$

We now write for every k,

$$T^k J - J = \sum_{\ell=1}^{k} (T^\ell J - T^{\ell-1} J) \leq \sum_{\ell=1}^{k} \alpha^{\ell-1} c\, v,$$

from which, by taking the limit as $k \to \infty$, we obtain $J^* \leq J + \big(c/(1-\alpha)\big) v$. The second relation follows similarly.

(b) This part is the special case of part (a) where T is equal to T_μ.

(c) We show the first relation. From part (a), the inequality $TJ \leq J + c\,v$ implies that $J^* \leq J + \big(c/(1-\alpha)\big) v$. Applying T^k to both sides of this inequality, and using the fact that T^k is a monotone sup-norm contraction of modulus α^k, with fixed point J^*, we obtain $J^* \le T^k J + \big(\alpha^k c/(1-\alpha)\big)v$. The second relation follows similarly. **Q.E.D.**

Approximations

As part of our subsequent analysis, we will consider approximations in the implementation of various VI and PI algorithms. In particular, we will assume that given any $J \in B(X)$, we cannot compute exactly TJ, but instead may compute $\bar{J} \in B(X)$ and $\mu \in \mathcal{M}$ such that

$$\|\bar{J} - TJ\| \le \delta, \qquad \|T_\mu J - TJ\| \le \epsilon, \tag{2.8}$$

where δ and ϵ are nonnegative scalars. These scalars may be unknown, so the resulting analysis will have a mostly qualitative character.

The case $\delta > 0$ arises when the state space is either infinite or it is finite but very large. Then instead of calculating $(TJ)(x)$ for all states x, one may do so only for some states and estimate $(TJ)(x)$ for the remaining states x by some form of interpolation. Alternatively, one may use simulation data [e.g., noisy values of $(TJ)(x)$ for some or all x] and some kind of least-squares error fit of $(TJ)(x)$ with a function from a suitable parametric class. The function $\bar{J}$ thus obtained will satisfy a relation such as (2.8) with $\delta > 0$. Note that δ may not be small in this context, and the resulting performance degradation may be a primary concern.

Cases where $\epsilon > 0$ may arise when the control space is infinite or finite but large, and the minimization involved in the calculation of $(TJ)(x)$ cannot be done exactly. Note, however, that it is possible that

$$\delta > 0, \qquad \epsilon = 0,$$

and in fact this occurs in several types of practical methods. In an alternative scenario, we may first obtain the policy μ subject to a restriction that it belongs to a certain subset of structured policies, so it satisfies $\|T_\mu J - TJ\| \le \epsilon$ for some $\epsilon > 0$, and then we may set $\bar{J} = T_\mu J$. In this case we have $\epsilon = \delta$.

3. LIMITED LOOKAHEAD POLICIES

A frequently used suboptimal approach in DP is to use a policy obtained by solving a finite-horizon problem with some given terminal cost function $\tilde{J}$ that approximates J^*. The simplest possibility is a *one-step lookahead policy* $\bar{\mu}$ defined by

$$\bar{\mu}(x) \in \arg\min_{u \in U(x)} H(x, u, \tilde{J}), \qquad x \in X. \tag{3.1}$$

In a variant of the method that aims to reduce the computation to obtain $\bar{\mu}(i)$, the minimization in Eq. (3.1) is done over a subset $\bar{U}(x) \subset U(x)$. Thus, the control $\bar{\mu}(x)$ used in this variant satisfies

$$\bar{\mu}(x) \in \arg\min_{u \in \bar{U}(i)} H(x, u, \tilde{J}), \qquad x \in X,$$

rather Eq. (3.1). This is attractive for example when by using some heuristic or approximate optimization, we can identify a subset $\bar{U}(x)$ of promising controls, and to save computation, we restrict attention to this subset in the one-step lookahead minimization.

The following proposition gives some bounds for the performance of such a one-step lookahead policy. The first bound [part (a) of the following proposition] is given in terms of the vector $\hat{J}$ given by

$$\hat{J}(x) = \inf_{u \in \bar{U}(i)} H(x, u, \tilde{J}), \qquad x \in X, \tag{3.2}$$

which is computed in the course of finding the one-step lookahead control at state x.

Proposition 3.1: (One-Step Lookahead Error Bounds) Let Assumptions 2.1 and 2.2 hold, and let $\bar{\mu}$ be a one-step lookahead policy obtained by minimization in Eq. (3.2).

(a) Assume that $\hat{J} \leq \tilde{J}$. Then $J_{\bar{\mu}} \leq \hat{J}$.

(b) Assume that $\bar{U}(i) = U(i)$ for all i. Then

$$\|J_{\bar{\mu}} - \hat{J}\| \leq \frac{\alpha}{1-\alpha} \|\hat{J} - \tilde{J}\|, \tag{3.3}$$

where $\|\cdot\|$ denotes the sup-norm. Moreover

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{2\alpha}{1-\alpha} \|\tilde{J} - J^*\|, \tag{3.4}$$

and

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{2}{1-\alpha} \|\hat{J} - \tilde{J}\|. \tag{3.5}$$

Proof: (a) We have

$$\tilde{J} \geq \hat{J} = T_{\bar{\mu}} \tilde{J},$$

from which by using the monotonicity of $T_{\bar{\mu}}$, we obtain

$$\tilde{J} \geq \hat{J} \geq T_{\bar{\mu}}^k \tilde{J} \geq T_{\bar{\mu}}^{k+1} \tilde{J}, \qquad k = 1, 2, \ldots$$

By taking the limit as $k \to \infty$, we have $\hat{J} \geq J_{\bar{\mu}}$.

(b) The proof of this part may rely on Prop. 2.1(e), but we will give a direct proof. Using the triangle inequality we write for every k,

$$\|T_{\bar{\mu}}^k \hat{J} - \hat{J}\| \leq \sum_{\ell=1}^{k} \|T_{\bar{\mu}}^{\ell} \hat{J} - T_{\bar{\mu}}^{\ell-1} \hat{J}\| \leq \sum_{\ell=1}^{k} \alpha^{\ell-1} \|T_{\bar{\mu}} \hat{J} - \hat{J}\|.$$

By taking the limit as $k \to \infty$ and using the fact $T_{\bar{\mu}}^k \hat{J} \to J_{\bar{\mu}}$, we obtain

$$\|J_{\bar{\mu}} - \hat{J}\| \leq \frac{1}{1-\alpha} \|T_{\bar{\mu}} \hat{J} - \hat{J}\|. \tag{3.6}$$

Since $\hat{J} = T_{\bar{\mu}} \tilde{J}$, we have

$$\|T_{\bar{\mu}} \hat{J} - \hat{J}\| = \|T_{\bar{\mu}} \hat{J} - T_{\bar{\mu}} \tilde{J}\| \leq \alpha \|\hat{J} - \tilde{J}\|,$$

and Eq. (3.3) follows by combining the last two relations.

By repeating the proof of Eq. (3.6) with $\hat{J}$ replaced by J^*, we obtain

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{1}{1-\alpha}\|T_{\bar{\mu}}J^* - J^*\|.$$

We have $T\tilde{J} = T_{\bar{\mu}}\tilde{J}$ and $J^* = TJ^*$, so

$$\begin{aligned}\|T_{\bar{\mu}}J^* - J^*\| &\leq \|T_{\bar{\mu}}J^* - T_{\bar{\mu}}\tilde{J}\|| + \|T\tilde{J} - TJ^*\| \\ &\leq \alpha\|J^* - \tilde{J}\| + \alpha\|\tilde{J} - J^*\| \\ &= 2\alpha\,\|\tilde{J} - J^*\|,\end{aligned}$$

and Eq. (3.4) follows by combining the last two relations.

Also, by repeating the proof of Eq. (3.6) with $\hat{J}$ replaced by $\tilde{J}$ and $T_{\bar{\mu}}$ replaced by T, we have using also $\hat{J} = T\tilde{J}$,

$$\|J^* - \tilde{J}\| \leq \frac{1}{1-\alpha}\|T\tilde{J} - \tilde{J}\| = \frac{1}{1-\alpha}\|\hat{J} - \tilde{J}\|.$$

We use this relation to write

$$\begin{aligned}\|J_{\bar{\mu}} - J^*\| &\leq \|J_{\bar{\mu}} - \hat{J}\| + \|\hat{J} - \tilde{J}\| + \|\tilde{J} - J^*\| \\ &\leq \frac{\alpha}{1-\alpha}\|\hat{J} - \tilde{J}\| + \|\hat{J} - \tilde{J}\| + \frac{1}{1-\alpha}\|\hat{J} - \tilde{J}\| \\ &= \frac{2}{1-\alpha}\|\hat{J} - \tilde{J}\|,\end{aligned}$$

where the second inequality follows from Eq. (3.3). **Q.E.D.**

Part (b) of the preceding proposition gives a bound on $J_{\bar{\mu}}(x)$, the performance of the one-step lookahead policy $\bar{\mu}$; the value of $\hat{J}(x)$ is obtained while finding the one-step lookahead control at x. The bound (3.4) of part (c) says that if the one-step lookahead approximation $\tilde{J}$ is within c of the optimal (in the weighted sup-norm sense), the performance of the one-step lookahead policy is within $2\alpha c/(1-\alpha)$ of the optimal. Part (b) of the preceding proposition gives bounds on $J_{\bar{\mu}}(x)$, the performance of the one-step lookahead policy $\bar{\mu}$. In particular, the bound (3.4) says that if the one-step lookahead approximation $\tilde{J}$ is within ϵ of the optimal, the performance of the one-step lookahead policy is within $2\alpha\epsilon/(1-\alpha)$ of the optimal. Unfortunately, this is not very reassuring when α is close to 1, in which case the error bound is very large relative to ϵ. Nonetheless, the following example from [BeT96], Section 6.1.1, shows that this error bound is tight in the sense that for any $\alpha < 1$, there is a problem with just two states where the error bound is satisfied with equality. What is happening is that an $O(\epsilon)$ difference in single stage cost between two controls can generate an $O\big(\epsilon/(1-\alpha)\big)$ difference in policy costs, yet it can be "nullified" in Bellman's equation by an $O(\epsilon)$ difference between J^* and $\tilde{J}$.

Example 3.1

Consider a discounted problem with two states, 1 and 2, and deterministic transitions. State 2 is absorbing, but at state 1 there are two possible decisions: move to state 2 (policy μ^*) or stay at state 1 (policy μ). The cost of each transition is 0 except for the transition from 1 to itself under policy μ, which has cost $2\alpha\epsilon$, where ϵ is a positive scalar and $\alpha \in [0,1)$ is the discount factor. The optimal policy μ^* is to move from state 1 to state 2, and the optimal cost-to-go function is $J^*(1) = J^*(2) = 0$. Consider the vector $\tilde{J}$ with $\tilde{J}(1) = -\epsilon$ and $\tilde{J}(2) = \epsilon$, so that

$$\|\tilde{J} - J^*\| = \epsilon,$$

as assumed in Eq. (3.4) [cf. Prop. 3.1(b)]. The policy μ that decides to stay at state 1 is a one-step lookahead policy based on $\tilde{J}$, because

$$2\alpha\epsilon + \alpha\tilde{J}(1) = \alpha\epsilon = 0 + \alpha\tilde{J}(2).$$

We have

$$J_\mu(1) = \frac{2\alpha\epsilon}{1-\alpha} = \frac{2\alpha}{1-\alpha}\|\tilde{J} - J^*\|,$$

so the bound of Eq. (3.4) holds with equality.

3.1 Multistep Lookahead Policies with Approximations

Let us now consider a more general form of lookahead involving multiple stages with intermediate approximations. In particular, we assume that given any $J \in B(X)$, we cannot compute exactly TJ, but instead may compute $\bar{J} \in B(X)$ and $\mu \in \mathcal{M}$ such that

$$\|\bar{J} - TJ\| \leq \delta, \qquad \|T_\mu J - TJ\| \leq \epsilon, \tag{3.7}$$

where δ and ϵ are nonnegative scalars.

In a multistep method with approximations, we are given a positive integer m and a lookahead function J_m, and we successively compute (backwards in time) $J_{m-1}, \ldots, J_0$ and policies $\mu_{m-1}, \ldots, \mu_0$ satisfying

$$\|J_k - TJ_{k+1}\| \leq \delta, \quad \|T_{\mu_k} J_{k+1} - TJ_{k+1}\| \leq \epsilon, \qquad k = 0, \ldots, m-1. \tag{3.8}$$

Note that in the context of MDP, J_k can be viewed as an approximation to the optimal cost function of an $(m-k)$-stage problem with terminal cost function J_m. We have the following proposition, which is based on the recent work of Scherrer [Sch12].

Proposition 3.2: (Multistep Lookahead Error Bound) Let Assumption 2.1 hold. The periodic policy

$$\pi = \{\mu_0, \ldots, \mu_{m-1}, \mu_0, \ldots, \mu_{m-1}, \ldots\}$$

generated by the method of Eq. (3.8) satisfies

$$\|J_\pi - J^*\| \leq \frac{2\alpha^m}{1-\alpha^m}\|J_m - J^*\| + \frac{\epsilon}{1-\alpha^m} + \frac{\alpha(\epsilon + 2\delta)(1-\alpha^{m-1})}{(1-\alpha)(1-\alpha^m)}. \tag{3.9}$$

Proof: Using the triangle inequality, Eq. (3.8), and the contraction property of T, we have for all k

$$\begin{aligned} \|J_{m-k} - T^k J_m\| &\leq \|J_{m-k} - TJ_{m-k+1}\| + \|TJ_{m-k+1} - T^2 J_{m-k+2}\| \\ &\quad + \cdots + \|T^{k-1}J_{m-1} - T^k J_m\| \\ &\leq \delta + \alpha\delta + \cdots + \alpha^{k-1}\delta, \end{aligned} \tag{3.10}$$

showing that

$$\|J_{m-k} - T^k J_m\| \le \frac{\delta(1-\alpha^k)}{1-\alpha}, \qquad k = 1, \ldots, m. \tag{3.11}$$

From Eq. (3.8), we have $\|J_k - T_{\mu_k} J_{k+1}\| \le \delta + \epsilon$, so for all k

$$\begin{aligned} \|J_{m-k} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| &\le \|J_{m-k} - T_{\mu_{m-k}} J_{m-k+1}\| \\ &\quad + \|T_{\mu_{m-k}} J_{m-k+1} - T_{\mu_{m-k}} T_{\mu_{m-k+1}} J_{m-k+2}\| \\ &\quad + \cdots \\ &\quad + \|T_{\mu_{m-k}} \cdots T_{\mu_{m-2}} J_{m-1} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| \\ &\le (\delta+\epsilon) + \alpha(\delta+\epsilon) + \cdots + \alpha^{k-1}(\delta+\epsilon), \end{aligned}$$

showing that

$$\|J_{m-k} - T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m\| \le \frac{(\delta+\epsilon)(1-\alpha^k)}{1-\alpha}, \qquad k = 1, \ldots, m. \tag{3.12}$$

Using the fact $\|T_{\mu_0} J_1 - T J_1\| \le \epsilon$ [cf. Eq. (3.8)], we obtain

$$\begin{aligned} \|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T^m J_m\| &\le \|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T_{\mu_0} J_1\| \\ &\quad + \|T_{\mu_0} J_1 - T J_1\| + \|T J_1 - T^m J_m\| \\ &\le \alpha \|T_{\mu_1} \cdots T_{\mu_{m-1}} J_m - J_1\| + \epsilon + \alpha \|J_1 - T^{m-1} J_m\| \\ &\le \epsilon + \frac{\alpha(\epsilon + 2\delta)(1-\alpha^{m-1})}{1-\alpha}, \end{aligned}$$

where the last inequality follows from Eqs. (3.11) and (3.12) for $k = m-1$.

From this relation and the fact that $T_{\mu_0} \cdots T_{\mu_{m-1}}$ and T^m are contractions with modulus α^m, we obtain

$$\begin{aligned} \|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - J^*\| &\le \|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - T_{\mu_0} \cdots T_{\mu_{m-1}} J_m\| \\ &\quad + \|T_{\mu_0} \cdots T_{\mu_{m-1}} J_m - T^m J_m\| + \|T^m J_m - J^*\| \\ &\le 2\alpha^m \|J^* - J_m\| + \epsilon + \frac{\alpha(\epsilon + 2\delta)(1-\alpha^{m-1})}{1-\alpha}. \end{aligned}$$

We also have using Prop. 2.1(e), applied in the context of the multistep mapping of Example 1.6.5 of Section 1.6,

$$\|J_\pi - J^*\| \le \frac{1}{1-\alpha^m} \|T_{\mu_0} \cdots T_{\mu_{m-1}} J^* - J^*\|.$$

Combining the last two relations, we obtain the desired result. **Q.E.D.**

Note that for $m = 1$ and $\delta = \epsilon = 0$, i.e., the case of one-step lookahead policy $\bar{\mu}$ with lookahead function J_1 and no approximation error in the minimization involved in TJ_1, Eq. (3.9) yields the bound

$$\|J_{\bar{\mu}} - J^*\| \le \frac{2\alpha}{1-\alpha} \|J_1 - J^*\|,$$

which coincides with the bound (3.4) derived earlier.

Also, in the special case where $\epsilon = \delta$ and $J_k = T_{\mu_k} J_{k+1}$ (cf. the discussion preceding Prop. 3.2), the bound (3.9) can be strengthened somewhat. In particular, we have for all k, $J_{m-k} = T_{\mu_{m-k}} \cdots T_{\mu_{m-1}} J_m$, so the right-hand side of Eq. (3.12) becomes 0 and the preceding proof yields, with some calculation,

$$\begin{aligned} \|J_\pi - J^*\| &\le \frac{2\alpha^m}{1-\alpha^m} \|J_m - J^*\| + \frac{\delta}{1-\alpha^m} + \frac{\alpha\delta(1-\alpha^{m-1})}{(1-\alpha)(1-\alpha^m)} \\ &= \frac{2\alpha^m}{1-\alpha^m} \|J_m - J^*\| + \frac{\delta}{1-\alpha}. \end{aligned}$$

We finally note that Prop. 3.2 shows that as $m \to \infty$, the corresponding bound for $\|J_\pi - J^*\|$ tends to $\epsilon + \alpha(\epsilon + 2\delta)/(1-\alpha)$, or

$$\limsup_{m\to\infty} \|J_\pi - J^*\| \le \frac{\epsilon + 2\alpha\delta}{1-\alpha}.$$

We will see that this error bound is superior to corresponding error bounds for approximate versions of VI and PI by essentially a factor $1/(1-\alpha)$. This is an interesting fact, which was first shown by Scherrer [Sch12] in the context of discounted MDP.

4. GENERALIZED VALUE ITERATION

Generalized value iteration (VI) is the algorithm that starts from some $J \in B(X)$, and generates TJ, $T^2J, \ldots$. Since T is a weighted sup-norm contraction under Assumption 2.1, the algorithm converges to J^*, and the rate of convergence is governed by

$$\|T^kJ - J^*\| \le \alpha^k \|J - J^*\|, \qquad k = 0, 1, \ldots.$$

Similarly, for a given policy $\mu \in \mathcal{M}$, we have

$$\|T_\mu^k J - J_\mu\| \le \alpha^k \|J - J_\mu\|, \qquad k = 0, 1, \ldots.$$

From Prop. 2.1(d), we also have the error bound

$$\|T^{k+1}J - J^*\| \le \frac{\alpha}{1-\alpha}\|T^{k+1}J - T^kJ\|, \qquad k = 0, 1, \ldots.$$

This bound does not rely on the Monotonicity Assumption 2.2.

Suppose now that we use generalized VI to compute an approximation $\tilde J$ to J^*, and then we obtain a policy $\bar\mu$ by minimization of $H(x, u, \tilde J)$ over $u \in U(x)$ for each $x \in X$. In other words $\tilde J$ and $\bar\mu$ satisfy

$$\|\tilde J - J^*\| \le \gamma, \qquad T_{\bar\mu}\tilde J = T\tilde J,$$

where γ is some positive scalar. Then, with an identical proof to Prop. 3.1(c), we have

$$J_{\bar\mu} \le J^* + \frac{2\alpha\,\gamma}{1-\alpha} v, \tag{4.1}$$

which can be viewed as an error bound for the performance of a policy obtained by generic one-step lookahead.

We use this bound in the following proposition, which shows that if the set of policies is finite, then a policy μ^* with $J_{\mu^*} = J^*$ may be obtained after a finite number of VI.

Proposition 4.1: Let Assumption 2.1 hold and let $J \in B(X)$. If the set of policies $\mathcal{M}$ is finite, there exists an integer $\bar k \ge 0$ such that $J_{\mu^*} = J^*$ for all μ^* and $k \ge \bar k$ with $T_{\mu^*}T^kJ = T^{k+1}J$.

Proof: Let $\bar{\mathcal{M}}$ be the set of nonoptimal policies, i.e., all μ such that $J_\mu \ne J^*$. Since $\bar{\mathcal{M}}$ is finite, we have

$$\min_{\mu\in\bar{\mathcal{M}}} \|J_\mu - J^*\| > 0,$$

so by Eq. (4.1), there exists sufficiently small $\beta > 0$ such that

$$\|\tilde J - J^*\| \le \beta \ \text{ and } \ T_\mu\tilde J = T\tilde J \quad\Rightarrow\quad \|J_\mu - J^*\| = 0 \quad\Rightarrow\quad \mu \notin \bar{\mathcal{M}}. \tag{4.2}$$

It follows that if k is sufficiently large so that $\|T^kJ - J^*\| \le \beta$, then $T_{\mu^*}T^kJ = T^{k+1}J$ implies that $\mu^* \notin \bar{\mathcal{M}}$ so $J_{\mu^*} = J^*$. **Q.E.D.**

4.1 Approximate Value Iteration

Let us consider situations where the VI method may be implementable only through approximations. In particular, given a function J, assume that we may only be able to calculate an approximation $\tilde{J}$ to TJ such that

$$\big\|\tilde{J} - TJ\big\| \leq \delta, \tag{4.3}$$

where δ is a given positive scalar. In the corresponding approximate VI method, we start from an arbitrary bounded function J_0, and we generate a sequence $\{J_k\}$ satisfying

$$\|J_{k+1} - TJ_k\| \leq \delta, \qquad k = 0, 1, \ldots. \tag{4.4}$$

This approximation may be the result of representing J_{k+1} compactly, as a linear combination of basis functions, through a projection or aggregation process, as is common in approximate DP.

We may also simultaneously generate a sequence of policies $\{\mu^k\}$ such that

$$\|T_{\mu^k} J_k - TJ_k\| \leq \epsilon, \qquad k = 0, 1, \ldots, \tag{4.5}$$

where ϵ is some scalar [which could be equal to 0, as in case of Eq. (3.8), considered earlier]. The following proposition shows that the corresponding cost vectors J_{μ^k} "converge" to J^* to within an error of order $O\big(\delta/(1-\alpha)^2\big)$ [plus a less significant error of order $O\big(\epsilon/(1-\alpha)\big)$].

Proposition 4.2: (Error Bounds for Approximate VI) Let Assumption 2.1 hold. A sequence $\{J_k\}$ generated by the approximate VI method (4.4)-(4.5) satisfies

$$\limsup_{k\to\infty} \|J_k - J^*\| \leq \frac{\delta}{1-\alpha}, \tag{4.6}$$

while the corresponding sequence of policies $\{\mu^k\}$ satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon}{1-\alpha} + \frac{2\alpha\delta}{(1-\alpha)^2}. \tag{4.7}$$

Proof: Arguing as in the proof of Prop. 3.2, we have

$$\|J_k - T^k J_0\| \leq \frac{\delta(1-\alpha^k)}{1-\alpha}, \qquad k = 0, 1, \ldots$$

[cf. Eq. (3.10)]. By taking limit as $k \to \infty$ and by using the fact $\lim_{k\to\infty} T^k J_0 = J^*$, we obtain Eq. (4.6).

We also have using the triangle inequality and the contraction property of T_{μ^k} and T,

$$\begin{aligned}\|T_{\mu^k} J^* - J^*\| &\leq \|T_{\mu^k} J^* - T_{\mu^k} J_k\| + \|T_{\mu^k} J_k - TJ_k\| + \|TJ_k - J^*\| \\ &\leq \alpha\|J^* - J_k\| + \epsilon + \alpha\|J_k - J^*\|,\end{aligned}$$

while by using also Prop. 2.1(e), we obtain

$$\|J_{\mu^k} - J^*\| \le \frac{1}{1-\alpha}\|T_{\mu^k}J^* - J^*\| \le \frac{\epsilon}{1-\alpha} + \frac{2\alpha}{1-\alpha}\|J_k - J^*\|.$$

By combining this relation with Eq. (4.6), we obtain Eq. (4.7). **Q.E.D.**

The error bound (4.7) relates to stationary policies obtained from the functions J_k by one-step lookahead. We may also obtain an m-step periodic policy π from J_k by using m-step lookahead. Then Prop. 3.2 shows that the corresponding bound for $\|J_\pi - J^*\|$ tends to $\epsilon + 2\alpha\delta/(1-\alpha)$ as $m \to \infty$, which improves on the error bound (4.7) by a factor $1/(1-\alpha)$. This is a remarkable and surprising fact, which was first shown by Scherrer [Sch12] in the context of discounted MDP.

Finally, let us note that the error bound of Prop. 4.2 is predicated upon generating a sequence $\{J_k\}$ satisfying $\|J_{k+1} - TJ_k\| \le \delta$ for all k [cf. Eq. (4.4)]. Unfortunately, some practical approximation schemes guarantee the existence of such a δ only if $\{J_k\}$ is a bounded sequence. The following simple example from [BeT96], Section 6.5.3, shows that boundedness of the iterates is not automatically guaranteed, and is a serious issue that should be addressed in approximate VI schemes.

Example 4.1 (Error Amplification in Approximate Value Iteration)

Consider a two-state discounted MDP with states 1 and 2, and a single policy. The transitions are deterministic: from state 1 to state 2, and from state 2 to state 2. These transitions are also cost-free. Thus we have $J^*(1) = J^*(2) = 0$.

We consider a VI scheme that approximates cost functions within the one-dimensional subspace of linear functions $S = \big\{(r, 2r) \mid r \in \Re\big\}$ by using a weighted least squares minimization; i.e., we approximate a vector J by its weighted Euclidean projection onto S. In particular, given $J_k = (r_k, 2r_k)$, we find $J_{k+1} = (r_{k+1}, 2r_{k+1})$, where for weights $w_1, w_2 > 0$, r_{k+1} is obtained as

$$r_{k+1} = \arg\min_r \Big[w_1\big(r - (TJ_k)(1)\big)^2 + w_2\big(2r - (TJ_k)(2)\big)^2\Big].$$

Since for a zero cost per stage and the given deterministic transitions, we have $TJ_k = (2\alpha r_k, 2\alpha r_k)$, the preceding minimization is written as

$$r_{k+1} = \arg\min_r \big[w_1(r - 2\alpha r_k)^2 + w_2(2r - 2\alpha r_k)^2\big],$$

which by writing the corresponding optimality condition yields $r_{k+1} = \alpha\beta r_k$, where $\beta = 2(w_1{+}2w_2)(w_1{+}4w_2) > 1$. Thus if $\alpha > 1/\beta$, the sequence $\{r_k\}$ diverges and so does $\{J_k\}$. Note that in this example the optimal cost function $J^* = (0,0)$ belongs to the subspace S. The difficulty here is that the approximate VI mapping that generates J_{k+1} by a least squares-based approximation of TJ_k is not a contraction. At the same time there is no δ such that $\|J_{k+1} - TJ_k\| \le \delta$ for all k, because of error amplification in each approximate VI.

5. GENERALIZED POLICY ITERATION

In generalized policy iteration (PI), we maintain and update a policy μ^k, starting from some initial policy μ^0. The $(k+1)$st iteration has the following form.

Generalized Policy Iteration

Policy Evaluation: We compute J_{μ^k} as the unique solution of the equation $J_{\mu^k} = T_{\mu^k} J_{\mu^k}$.

Policy Improvement: We obtain an improved policy μ^{k+1} that satisfies $T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$.

The algorithm requires the Monotonicity Assumption 2.2, in addition to the Contraction Assumption 2.1, so we assume these two conditions throughout this section. Moreover we assume that the minimum of $H(x, u, J_{\mu^k})$ over $u \in U(x)$ is attained for all $x \in X$, so that the improved policy μ^{k+1} is defined. The following proposition establishes a basic cost improvement property, as well as finite convergence for the case where the set of policies is finite.

Proposition 5.1: (Convergence of Generalized PI) Let Assumptions 2.1 and 2.2 hold, and let $\{\mu^k\}$ be a sequence generated by the generalized PI algorithm. Then for all k, we have $J_{\mu^{k+1}} \le J_{\mu^k}$, with equality if and only if $J_{\mu^k} = J^*$. Moreover,

$$\lim_{k \to \infty} \| J_{\mu^k} - J^* \| = 0,$$

and if the set of policies is finite, we have $J_{\mu^k} = J^*$ for some k.

Proof: We have

$$T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k} \le T_{\mu^k} J_{\mu^k} = J_{\mu^k}.$$

Applying $T_{\mu^{k+1}}$ to this inequality while using the Monotonicity Assumption 2.2, we obtain

$$T^2_{\mu^{k+1}} J_{\mu^k} \le T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k} \le T_{\mu^k} J_{\mu^k} = J_{\mu^k}.$$

Similarly, we have for all $m > 0$,

$$T^m_{\mu^{k+1}} J_{\mu^k} \le T J_{\mu^k} \le J_{\mu^k},$$

and by taking the limit as $m \to \infty$, we obtain

$$J_{\mu^{k+1}} \le T J_{\mu^k} \le J_{\mu^k}, \qquad k = 0, 1, \ldots. \tag{5.1}$$

If $J_{\mu^{k+1}} = J_{\mu^k}$, it follows that $T J_{\mu^k} = J_{\mu^k}$, so J_{μ^k} is a fixed point of T and must be equal to J^*. Moreover by using induction, Eq. (5.1) implies that

$$J_{\mu^k} \le T^k J_{\mu^0}, \qquad k = 0, 1, \ldots,$$

Since

$$J^* \le J_{\mu^k}, \qquad \lim_{k \to \infty} \| T^k J_{\mu^0} - J^* \| = 0,$$

it follows that $\lim_{k\to\infty} \|J_{\mu^k} - J^*\| = 0$. Finally, if the number of policies is finite, Eq. (5.1) implies that there can be only a finite number of iterations for which $J_{\mu^{k+1}}(x) < J_{\mu^k}(x)$ for some x, so we must have $J_{\mu^{k+1}} = J_{\mu^k}$ for some k, at which time $J_{\mu^k} = J^*$ as shown earlier. **Q.E.D.**

In the case where the set of policies is infinite, we may assert the convergence of the sequence of generated policies under some compactness and continuity conditions. In particular, we will assume that the state space is finite, $X = \{1, \dots, n\}$, and that each control constraint set $U(x)$ is a compact subset of $\Re^m$. We will view a cost vector J as an element of $\Re^n$, and a policy μ as an element of the compact set $U(1) \times \cdots \times U(n) \subset \Re^{mn}$. Then $\{\mu^k\}$ has at least one limit point $\bar{\mu}$, which must be an admissible policy. The following proposition guarantees, under an additional continuity assumption for $H(x, \cdot, \cdot)$, that every limit point $\bar{\mu}$ is optimal.

Assumption 5.1: (Compactness and Continuity)

(a) The state space is finite, $X = \{1, \dots, n\}$.

(b) Each control constraint set $U(x)$, $x = 1, \dots, n$, is a compact subset of $\Re^m$.

(c) Each function $H(x, \cdot, \cdot)$, $x = 1, \dots, n$, is continuous over $U(x) \times \Re^n$.

Proposition 5.2: Let Assumptions 2.1, 2.2, and 5.1 hold, and let $\{\mu^k\}$ be a sequence generated by the generalized PI algorithm. Then for every limit point $\bar{\mu}$ of $\{\mu^k\}$, we have $J_{\bar{\mu}} = J^*$.

Proof: We have $J_{\mu^k} \to J^*$ by Prop. 5.1. Let $\bar{\mu}$ be the limit of a subsequence $\{\mu^k\}_{k\in\mathcal{K}}$. We will show that $T_{\bar{\mu}} J^* = TJ^*$, from which it follows that $J_{\bar{\mu}} = J^*$ [cf. Prop. 2.1(c)]. Indeed, we have $T_{\bar{\mu}} J^* \geq TJ^*$, so we focus on showing the reverse inequality. From the equation $T_{\mu^k} J_{\mu^{k-1}} = TJ_{\mu^{k-1}}$ we have

$$H\big(x, \mu^k(x), J_{\mu^{k-1}}\big) \leq H(x, u, J_{\mu^{k-1}}), \qquad x = 1, \dots, n,\ u \in U(x).$$

By taking limit in this relation as $k \to \infty$, $k \in \mathcal{K}$, and by using the continuity of $H(x, \cdot, \cdot)$ [cf. Assumption 5.1(c)], we obtain

$$H\big(x, \bar{\mu}(x), J^*\big) \leq H(x, u, J^*), \qquad x = 1, \dots, n,\ u \in U(x).$$

By taking the minimum of the right-hand side over $u \in U(x)$, we obtain $T_{\bar{\mu}} J^* \leq TJ^*$. **Q.E.D.**

5.1 Approximate Policy Iteration

We now consider the PI method where the policy evaluation step and/or the policy improvement step of the method are implemented through approximations. This method generates a sequence of policies $\{\mu^k\}$ and a corresponding sequence of approximate cost functions $\{J_k\}$ satisfying

$$\|J_k - J_{\mu^k}\| \leq \delta, \qquad \|T_{\mu^{k+1}} J_k - TJ_k\| \leq \epsilon, \qquad k = 0, 1, \dots, \tag{5.2}$$

where $\|\cdot\|$ denotes the sup-norm and v is the weight vector of the weighted sup-norm (it is important to use v rather than the unit vector in the above equation, in order for the bounds obtained to have a clean form). The following proposition provides an error bound for this algorithm, which extends a corresponding result of [BeT96], shown for discounted MDP.

Proposition 5.3: (Error Bound for Approximate PI) Let Assumptions 2.1 and 2.2 hold. The sequence $\{\mu^k\}$ generated by the approximate PI algorithm (5.2) satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \tag{5.3}$$

The essence of the proof is contained in the following proposition, which quantifies the amount of approximate policy improvement at each iteration.

Proposition 5.4: Let Assumptions 2.1 and 2.2 hold. Let J, $\bar{\mu}$, and μ satisfy

$$\|J - J_\mu\| \leq \delta, \qquad \|T_{\bar{\mu}}J - TJ\| \leq \epsilon,$$

where δ and ϵ are some scalars. Then

$$\|J_{\bar{\mu}} - J^*\| \leq \alpha\|J_\mu - J^*\| + \frac{\epsilon + 2\alpha\delta}{1-\alpha}. \tag{5.4}$$

Proof: Using Eq. (5.4) and the contraction property of T and $T_{\bar{\mu}}$, which implies that $\|T_{\bar{\mu}}J_\mu - T_{\bar{\mu}}J\| \leq \alpha\delta$ and $\|TJ - TJ_\mu\| \leq \alpha\delta$, and hence $T_{\bar{\mu}}J_\mu \leq T_{\bar{\mu}}J + \alpha\delta\, v$ and $TJ \leq TJ_\mu + \alpha\delta\, v$, we have

$$T_{\bar{\mu}}J_\mu \leq T_{\bar{\mu}}J + \alpha\delta\, v \leq TJ + (\epsilon + \alpha\delta)\, v \leq TJ_\mu + (\epsilon + 2\alpha\delta)\, v. \tag{5.5}$$

Since $TJ_\mu \leq T_\mu J_\mu = J_\mu$, this relation yields

$$T_{\bar{\mu}}J_\mu \leq J_\mu + (\epsilon + 2\alpha\delta)\, v,$$

and applying Prop. 2.4(b) with $\mu = \bar{\mu}$, $J = J_\mu$, and $\epsilon = \epsilon + 2\alpha\delta$, we obtain

$$J_{\bar{\mu}} \leq J_\mu + \frac{\epsilon + 2\alpha\delta}{1-\alpha}\, v. \tag{5.6}$$

Using this relation, we have

$$J_{\bar{\mu}} = T_{\bar{\mu}}J_{\bar{\mu}} = T_{\bar{\mu}}J_\mu + (T_{\bar{\mu}}J_{\bar{\mu}} - T_{\bar{\mu}}J_\mu) \leq T_{\bar{\mu}}J_\mu + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha}\, v,$$

where the inequality follows by using Prop. 2.3 and Eq. (5.6). Subtracting J^* from both sides, we have

$$J_{\bar{\mu}} - J^* \le T_{\bar{\mu}} J_\mu - J^* + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha}\, v, \tag{5.7}$$

Also by subtracting J^* from both sides of Eq. (5.5), and using the contraction property

$$T J_\mu - J^* = T J_\mu - T J^* \le \alpha \| J_\mu - J^* \|\, v,$$

yields

$$T_{\bar{\mu}} J_\mu - J^* \le T J_\mu - J^* + (\epsilon + 2\alpha\delta)\, v \le \alpha \| J_\mu - J^* \|\, v + (\epsilon + 2\alpha\delta)\, v.$$

Combining this relation with Eq. (5.7), we obtain

$$J_{\bar{\mu}} - J^* \le \alpha \| J_\mu - J^* \|\, v + \frac{\alpha(\epsilon + 2\alpha\delta)}{1-\alpha}\, v + (\epsilon + \alpha\delta) e = \alpha \| J_\mu - J^* \|\, v + \frac{\epsilon + 2\alpha\delta}{1-\alpha}\, v,$$

which is equivalent to the desired relation (5.4). **Q.E.D.**

Proof of Prop. 5.3: Applying Prop. 5.4, we have

$$\| J_{\mu^{k+1}} - J^* \| \le \alpha \| J_{\mu^k} - J^* \| + \frac{\epsilon + 2\alpha\delta}{1-\alpha},$$

which by taking the lim sup of both sides as $k \to \infty$ yields the desired result. **Q.E.D.**

We note that the error bound of Prop. 5.3 is tight, as can be shown with an example from [BeT96], Section 6.2.3. The error bound is comparable to the one for approximate VI, derived earlier in Prop. 4.2. In particular, the error $\| J_{\mu^k} - J^* \|$ is asymptotically proportional to $1/(1-\alpha)^2$ and to the approximation error in policy evaluation or value iteration, respectively. This is noteworthy, as it indicates that contrary to the case of exact implementation, approximate PI need not hold a convergence rate advantage over approximate VI, despite its greater overhead per iteration.

On the other hand, approximate PI does not have as much difficulty with the kind of iteration instability that was illustrated by Example 4.1 for approximate VI. In particular, if the set of policies is finite, so that the sequence $\{J_{\mu^k}\}$ is guaranteed to be bounded, the assumption of Eq. (5.2) is not hard to satisfy in practice with the cost function approximation methods to be discussed in Chapters 6 and 7.

Note that when $\delta = \epsilon = 0$, Eq. (5.4) yields

$$\| J_{\mu^{k+1}} - J^* \| \le \alpha \| J_{\mu^k} - J^* \|.$$

Thus in the case of an infinite state space and/or control space, exact PI converges at a geometric rate under the contraction and monotonicity assumptions of this section. This rate is the same as the rate of convergence of exact VI.

The Case Where Policies Converge

Generally, the policy sequence $\{\mu^k\}$ generated by approximate PI may oscillate between several policies. However, under some circumstances this sequence may be guaranteed to converge to some $\bar{\mu}$, in the sense that

$$\mu^{\bar{k}+1} = \mu^{\bar{k}} = \bar{\mu} \qquad \text{for some } \bar{k}. \tag{5.8}$$

An example arises when the policy sequence $\{\mu^k\}$ is generated by *exact PI* applied with a *different* mapping $\tilde{H}$ in place of H, but the bounds of Eq. (5.2) are satisfied. The mapping $\tilde{H}$ may for example correspond to an approximation of the original problem, as in aggregation methods. In this case we can show the following bound, which is much more favorable than the one of Prop. 5.3.

Proposition 5.5: (Error Bound for Approximate PI when Policies Converge) Let Assumptions 2.1 and 2.2 hold, and let $\bar{\mu}$ be a policy generated by the approximate PI algorithm (5.2) and satisfying condition (5.8). Then we have

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{1-\alpha}. \tag{5.9}$$

Proof: Let $\bar{J}$ be the cost vector obtained by approximate policy evaluation of $\bar{\mu}$ [i.e., $\bar{J} = J_{\bar{k}}$, where $\bar{k}$ satisfies the condition (5.8)]. Then we have

$$\|\bar{J} - J_{\bar{\mu}}\| \leq \delta, \qquad \|T_{\bar{\mu}}\bar{J} - T\bar{J}\| \leq \epsilon, \tag{5.10}$$

where the latter inequality holds since we have

$$\|T_{\bar{\mu}}\bar{J} - T\bar{J}\| = \|T_{\mu^{\bar{k}+1}}J_{\bar{k}} - TJ_{\bar{k}}\| \leq \epsilon,$$

cf. Eq. (5.2). Using Eq. (5.10) and the fact $J_{\bar{\mu}} = T_{\bar{\mu}}J_{\bar{\mu}}$, we have

$$\begin{aligned}\|TJ_{\bar{\mu}} - J_{\bar{\mu}}\| &\leq \|TJ_{\bar{\mu}} - T\bar{J}\| + \|T\bar{J} - T_{\bar{\mu}}\bar{J}\| + \|T_{\bar{\mu}}\bar{J} - J_{\bar{\mu}}\| \\ &= \|TJ_{\bar{\mu}} - T\bar{J}\| + \|T\bar{J} - T_{\bar{\mu}}\bar{J}\| + \|T_{\bar{\mu}}\bar{J} - T_{\bar{\mu}}J_{\bar{\mu}}\| \\ &\leq \alpha\|J_{\bar{\mu}} - \bar{J}\| + \epsilon + \alpha\|\bar{J} - J_{\bar{\mu}}\| \\ &\leq \epsilon + 2\alpha\delta.\end{aligned} \tag{5.11}$$

Using Prop. 2.1(d) with $J = J_{\bar{\mu}}$, we obtain the error bound (5.9). **Q.E.D.**

The preceding error bound can be generalized to the case where two successive policies generated by the approximate PI algorithm are "not too different" rather than being identical. In particular, suppose that μ and $\bar{\mu}$ are successive policies, which in addition to

$$\|\bar{J} - J_{\mu}\| \leq \delta, \qquad \|T_{\bar{\mu}}\bar{J} - T\bar{J}\| \leq \epsilon,$$

[cf. Eq. (5.2)], also satisfy

$$\|T_\mu \bar{J} - T_{\bar{\mu}} \bar{J}\| \leq \zeta,$$

where ζ is some scalar (instead of $\mu = \bar{\mu}$, which is the case where policies converge exactly). Then we also have

$$\|T\bar{J} - T_{\bar{\mu}}\bar{J}\| \leq \|T\bar{J} - T_\mu \bar{J}\| + \|T_\mu \bar{J} - T_{\bar{\mu}}\bar{J}\| \leq \epsilon + \zeta,$$

and by replacing ϵ with $\epsilon + \zeta$ in Eq. (5.11), we obtain

$$\|J_{\bar{\mu}} - J^*\| \leq \frac{\epsilon + \zeta + 2\alpha\delta}{1-\alpha}.$$

When ζ is small enough to be of the order of $\max\{\delta, \epsilon\}$, this error bound is comparable to the one for the case where policies converge.

6. OPTIMISTIC POLICY ITERATION

In optimistic PI (also called "modified" PI, see e.g., [Put94]) each policy μ^k is evaluated by solving the equation $J_{\mu^k} = T_{\mu^k} J_{\mu^k}$ approximately, using a finite number of VI. Thus, starting with a function $J_0 \in B(X)$, we generate sequences $\{J_k\}$ and $\{\mu^k\}$ with the algorithm

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{6.1}$$

where $\{m_k\}$ is a sequence of positive integers.

A more general form of optimistic policy iteration, considered by Thiery and Scherrer [ThS10b], is

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = \sum_{\ell=1}^{\infty} \lambda_\ell T_{\mu^k}^{\ell} J_k, \qquad k = 0, 1, \ldots, \tag{6.2}$$

where $\{\lambda_\ell\}$ is a sequence of nonnegative scalars such that

$$\sum_{\ell=1}^{\infty} \lambda_\ell = 1.$$

An example is the λ-policy iteration method (Bertsekas and Ioffe [BeI96], Thiery and Scherrer [ThS10a], Bertsekas [Ber11], Scherrer [Sch11]), where $\lambda_\ell = (1-\lambda)\lambda^{\ell-1}$, with λ being a scalar in $(0, 1)$. For simplicity, we will not discuss the more general type of algorithm (6.2) in this paper, but some of our results admit straightforward extensions to this case, particularly the analysis of Section 6.2, and the SSP analysis of Section 7.

6.1 Convergence of Optimistic Policy Iteration

The following two propositions provide the convergence properties of the algorithm (6.1). These propositions have been proved by Rothblum [Rot79] within the framework of Denardo's model [Der67], i.e., the case of an unweighted sup-norm where v is the unit function; see also Canbolat and Rothblum [CaR11], which considers optimistic PI methods where the minimization in the policy improvement (but not the policy evaluation) operation is approximate, within some $\epsilon > 0$. Our proof follows closely the one of Rothblum [Rot79].

Proposition 6.1: (Convergence of Optimistic Generalized PI) Let Assumptions 2.1 and 2.2 hold, and let $\big\{(J_k, \mu^k)\big\}$ be a sequence generated by the optimistic generalized PI algorithm (6.1). Then

$$\lim_{k\to\infty} \|J_k - J^*\| = 0,$$

and if the number of policies is finite, we have $J_{\mu^k} = J^*$ for all k greater than some index $\bar{k}$.

Proposition 6.2: Let Assumptions 2.1, 2.2, and 5.1 hold, and let $\big\{(J_k, \mu^k)\big\}$ be a sequence generated by the optimistic generalized PI algorithm (6.1). Then every limit point $\bar{\mu}$ of $\{\mu^k\}$, satisfies $J_{\bar{\mu}} = J^*$.

We develop the proofs of the propositions through four lemmas. The first lemma collects some generic properties of monotone weighted sup-norm contractions, variants of which we have noted earlier, and we restate for convenience.

Lemma 6.1: Let $W : B(X) \mapsto B(X)$ be a mapping that satisfies the monotonicity assumption

$$J \le J' \quad \Rightarrow \quad WJ \le WJ', \qquad \forall\ J, J' \in B(X),$$

and the contraction assumption

$$\|WJ - WJ'\| \le \alpha \|J - J'\|, \qquad \forall\ J, J' \in B(X),$$

for some $\alpha \in (0, 1)$.

(a) For all $J, J' \in B(X)$ and scalar $c \ge 0$, we have

$$J \ge J' - c\,v \qquad \Rightarrow \qquad WJ \ge WJ' - \alpha c\,v. \tag{6.3}$$

(b) For all $J \in B(X)$, $c \ge 0$, and $k = 0, 1, \ldots$, we have

$$J \ge WJ - c\,v \qquad \Rightarrow \qquad W^k J \ge J^* - \frac{\alpha^k}{1-\alpha} c\,v, \tag{6.4}$$

$$WJ \ge J - c\,v \qquad \Rightarrow \qquad J^* \ge W^k J - \frac{\alpha^k}{1-\alpha} c\,v, \tag{6.5}$$

where J^* is the fixed point of W.

Proof: Part (a) essentially follows from Prop. 2.3, while part (b) essentially follows from Prop. 2.4(c). **Q.E.D.**

Lemma 6.2: Let Assumptions 2.1 and 2.2 hold, and let $J \in B(X)$ and $c \geq 0$ satisfy

$$J \geq TJ - c\,v,$$

and let $\mu \in \mathcal{M}$ be such that $T_\mu J = TJ$. Then for all $k > 0$, we have

$$TJ \geq T_\mu^k J - \frac{\alpha}{1-\alpha} c\,v, \tag{6.6}$$

and

$$T_\mu^k J \geq T(T_\mu^k J) - \alpha^k c\,v. \tag{6.7}$$

Proof: Since $J \geq TJ - c\,v = T_\mu J - c\,v$, by using Lemma 6.1(a) with $W = T_\mu^j$ and $J' = T_\mu J$, we have for all $j \geq 1$,

$$T_\mu^j J \geq T_\mu^{j+1} J - \alpha^j c\,v. \tag{6.8}$$

By adding this relation over $j = 1, \ldots, k-1$, we have

$$TJ = T_\mu J \geq T_\mu^k J - \sum_{j=1}^{k-1} \alpha^j c\,v = T_\mu^k J - \frac{\alpha - \alpha^k}{1-\alpha} c\,v \geq T_\mu^k J - \frac{\alpha}{1-\alpha} c\,v,$$

showing Eq. (6.6). From Eq. (6.8) for $j = k$, we obtain

$$T_\mu^k J \geq T_\mu^{k+1} J - \alpha^k c\,v = T_\mu(T_\mu^k J) - \alpha^k c\,v \geq T(T_\mu^k J) - \alpha^k c\,v,$$

showing Eq. (6.7). **Q.E.D.**

The next lemma applies to the optimistic generalized PI algorithm (6.1) and proves a preliminary bound.

Lemma 6.3: Let Assumptions 2.1 and 2.2 hold, let $\big\{(J_k, \mu^k)\big\}$ be a sequence generated by the PI algorithm (6.1), and assume that for some $c \geq 0$ we have

$$J_0 \geq TJ_0 - c\,v.$$

Then for all $k \geq 0$,

$$TJ_k + \frac{\alpha}{1-\alpha} \beta_k c\,v \geq J_{k+1} \geq TJ_{k+1} - \beta_{k+1} c\,v, \tag{6.9}$$

where β_k is the scalar given by

$$\beta_k = \begin{cases} 1 & \text{if } k = 0, \\ \alpha^{m_0+\cdots+m_{k-1}} & \text{if } k > 0, \end{cases} \tag{6.10}$$

with m_j, $j = 0, 1, \ldots$, being the integers used in the algorithm (6.1).

Proof: We prove Eq. (6.9) by induction on k, using Lemma 6.2. For $k = 0$, using Eq. (6.6) with $J = J_0$, $\mu = \mu^0$, and $k = m_0$, we have

$$TJ_0 \geq J_1 - \frac{\alpha}{1-\alpha} c\, v = J_1 - \frac{\alpha}{1-\alpha}\beta_0 c\, v,$$

showing the left-hand side of Eq. (6.9) for $k = 0$. Also by Eq. (6.7) with $\mu = \mu^0$ and $k = m_0$, we have

$$J_1 \geq TJ_1 - \alpha^{m_0} c\, v = TJ_1 - \beta_1 c\, v.$$

showing the right-hand side of Eq. (6.9) for $k = 0$.

Assuming that Eq. (6.9) holds for $k - 1 \geq 0$, we will show that it holds for k. Indeed, the right-hand side of the induction hypothesis yields

$$J_k \geq TJ_k - \beta_k c\, v.$$

Using Eqs. (6.6) and (6.7) with $J = J_k$, $\mu = \mu^k$, and $k = m_k$, we obtain

$$TJ_k \geq J_{k+1} - \frac{\alpha}{1-\alpha}\beta_k c\, v,$$

and

$$J_{k+1} \geq TJ_{k+1} - \alpha^{m_k}\beta_k c\, v = TJ_{k+1} - \beta_{k+1} c\, v,$$

respectively. This completes the induction. **Q.E.D.**

The next lemma essentially proves the convergence of the generalized optimistic PI (Prop. 6.1) and provides associated error bounds.

Lemma 6.4: Let Assumptions 2.1 and 2.2 hold, let $\big\{(J_k, \mu^k)\big\}$ be a sequence generated by the PI algorithm (6.1), and let $c \geq 0$ be a scalar such that

$$\|J_0 - TJ_0\| \leq c. \tag{6.11}$$

Then for all $k \geq 0$,

$$J_k + \frac{\alpha^k}{1-\alpha} c\, v \geq J_k + \frac{\beta_k}{1-\alpha} c\, v \geq J^* \geq J_k - \frac{(k+1)\alpha^k}{1-\alpha} c\, v, \tag{6.12}$$

where β_k is defined by Eq. (6.10).

Proof: Using the relation $J_0 \geq TJ_0 - c\,v$ [cf. Eq. (6.11)] and Lemma 6.3, we have

$$J_k \geq TJ_k - \beta_k c\,v, \quad k = 0, 1, \ldots.$$

Using this relation in Lemma 6.1(b) with $W = T$ and $k = 0$, we obtain

$$J_k \geq J^* - \frac{\beta_k}{1-\alpha} c\,v,$$

which together with the fact $\alpha^k \geq \beta_k$, shows the left-hand side of Eq. (6.12).

Using the relation $TJ_0 \geq J_0 - c\,v$ [cf. Eq. (6.11)] and Lemma 6.1(b) with $W = T$, we have

$$J^* \geq T^k J_0 - \frac{\alpha^k}{1-\alpha} c\,v, \qquad k = 0, 1, \ldots. \tag{6.13}$$

Using again the relation $J_0 \geq TJ_0 - c\,v$ in conjunction with Lemma 6.3, we also have

$$TJ_j \geq J_{j+1} - \frac{\alpha}{1-\alpha} \beta_j c\,v, \qquad j = 0, \ldots, k-1.$$

Applying T^{k-j-1} to both sides of this inequality and using the monotonicity and contraction properties of T^{k-j-1}, we obtain

$$T^{k-j} J_j \geq T^{k-j-1} J_{j+1} - \frac{\alpha^{k-j}}{1-\alpha} \beta_j c\,v, \qquad j = 0, \ldots, k-1,$$

cf. Lemma 6.1(a). By adding this relation over $j = 0, \ldots, k-1$, and using the fact $\beta_j \leq \alpha^j$, it follows that

$$T^k J_0 \geq J_k - \sum_{j=0}^{k-1} \frac{\alpha^{k-j}}{1-\alpha} \alpha^j c\,v = J_k - \frac{k\alpha^k}{1-\alpha} c\,v. \tag{6.14}$$

Finally, by combining Eqs. (6.13) and (6.14), we obtain the right-hand side of Eq. (6.12). **Q.E.D.**

Proof of Props. 6.1 and 6.2: Let c be a scalar satisfying Eq. (6.11). Then the error bounds (6.12) show that $\lim_{k\to\infty} \|J_k - J^*\| = 0$, i.e., the first part of Prop. 6.1. The second part (finite termination when the number of policies is finite) follows similar to Prop. 5.1. The proof of Prop. 6.2 follows using the Compactness and Continuity Assumption 5.1, and the convergence argument of Prop. 5.2. **Q.E.D.**

Convergence Rate Issues

Let us consider the convergence rate bounds of Lemma 6.4 for generalized optimistic PI, and write them in the form

$$\|J_0 - TJ_0\| \leq c \quad \Rightarrow \quad J_k - \frac{(k+1)\alpha^k}{1-\alpha} c\,v \leq J^* \leq J_k + \frac{\alpha^{m_0+\cdots+m_k}}{1-\alpha} c\,v. \tag{6.15}$$

We may contrast these bounds with the ones for generalized VI, where

$$\|J_0 - TJ_0\| \leq c \quad \Rightarrow \quad T^k J_0 - \frac{\alpha^k}{1-\alpha} c\,v \leq J^* \leq T^k J_0 + \frac{\alpha^k}{1-\alpha} c\,v \tag{6.16}$$

[cf. Prop. 2.4(c)].

In comparing the bounds (6.15) and (6.16), we should also take into account the associated overhead for a single iteration of each method: optimistic PI requires at iteration k a single application of T and $m_k - 1$ applications of T_{μ^k} (each being less time-consuming than an application of T), while VI requires a single application of T. It can then be seen that the upper bound for optimistic PI is better than the one for VI (same bound for less overhead), while the lower bound for optimistic PI is worse than the one for VI (worse bound for more overhead). This suggests that the choice of the initial condition J_0 is important in optimistic PI, and in particular it is preferable to have $J_0 \geq TJ_0$ (implying convergence to J^* from above) rather than $J_0 \leq TJ_0$ (implying convergence to J^* from below). This is consistent with the results of other works, which indicate that the convergence properties of the method are fragile when the condition $J_0 \geq TJ_0$ does not hold (see [WiB93], [BeT96], [BeY10a], [BeY10b], [YuB11]).

6.2 Approximate Optimistic Policy Iteration

We now consider error bounds for the case where the policy evaluation and policy improvement operations are approximate, similar to the nonoptimistic PI case of Section 5.1. In particular, we consider a method that generates a sequence of policies $\{\mu^k\}$ and a corresponding sequence of approximate cost functions $\{J_k\}$ satisfying

$$\|J_k - T_{\mu^k}^{m_k} J_{k-1}\| \leq \delta, \qquad \|T_{\mu^{k+1}} J_k - TJ_k\| \leq \epsilon, \qquad k = 0, 1, \ldots, \tag{6.17}$$

[cf. Eq. (5.2)]. For example, we may compute (perhaps approximately, by simulation) the values $T_{\mu^k}^{m_k}(x)$ for a subset of states x, and use a least squares fit of these values to select J_k from some parametric class of functions.

We will prove the same error bound as for the nonoptimistic case, cf. Eq. (5.3). However, for this we will need the following condition, which is stronger than the contraction and monotonicity conditions that we have been using so far.

Assumption 6.1: (Semilinear Monotonic Contraction) For all $J \in B(X)$ and $\mu \in \mathcal{M}$, the functions $T_\mu J$ and TJ belong to $B(X)$. Furthermore, for some $\alpha \in (0, 1)$, we have

$$\frac{(T_\mu J')(x) - (T_\mu J)(x)}{v(x)} \leq \alpha \sup_{y \in X} \frac{J'(y) - J(y)}{v(y)}, \qquad \forall\ J, J' \in B(X),\ \mu \leq \mathcal{M},\ x \in X. \tag{6.18}$$

This assumption implies both the Contraction and Monotonicity Assumptions 2.1 and 2.2, as can be easily verified. Moreover the assumption is satisfied in all of the discounted DP examples of Section 2, as well as the SSP problem of the next section. It holds if T_μ is a linear mapping involving a matrix with nonnegative components that has spectral radius less than 1 (or more generally if T_μ is the minimum or the maximum of a finite number of such linear mappings).

For any function $y \in B(X)$, let us use the notation

$$M(y) = \sup_{x \in X} \frac{y(x)}{v(x)}. \tag{6.19}$$

Then the condition (6.18) can be written as

$$M(T_\mu J - T_\mu J') \le \alpha M(J - J'), \qquad \forall\, J, J' \in B(X),\ \mu \in \mathcal{M}, \tag{6.20}$$

and also implies the following multistep versions,

$$T_\mu^\ell J - T_\mu^\ell J' \le \alpha^\ell M(J - J')v, \quad M(T_\mu^\ell J - T_\mu^\ell J') \le \alpha^\ell M(J - J'), \qquad \forall\, J, J' \in B(X),\ \mu \in \mathcal{M},\ \ell \ge 1, \tag{6.21}$$

which can be proved by induction using Eq. (6.20). We have the following proposition, whose proof follows closely the original one by Thiery and Scherrer [ThS10b], given for the case of a discounted MDP.

Proposition 6.3: (Error Bound for Optimistic Approximate PI) Let Assumption 6.1 hold. Then the sequence $\{\mu^k\}$ generated by the optimistic approximate PI algorithm (6.17) satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \le \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \tag{6.22}$$

Proof: Let us fix $k \ge 1$ and for simplicity let us denote

$$\underline{J} = J_{k-1}, \qquad J = J_k,$$

$$\mu = \mu^k, \qquad \bar\mu = \mu^{k+1}, \qquad m = m_k, \qquad \bar m = m_{k+1},$$

$$s = J_\mu - T_\mu^m \underline{J}, \qquad \bar s = J_{\bar\mu} - T_{\bar\mu}^{\bar m} J, \qquad t = T_\mu^m \underline{J} - J^*, \qquad \bar t = T_{\bar\mu}^{\bar m} J - J^*.$$

We have

$$J_\mu - J^* = J_\mu - T_\mu^m \underline{J} + T_\mu^m \underline{J} - J^* = s + t. \tag{6.23}$$

We will derive recursive relations for s and t, which will also involve the residual functions

$$r = T_\mu \underline{J} - \underline{J}, \qquad \bar r = T_{\bar\mu} J - J.$$

We first obtain a relation between r and $\bar r$. We have

$$\begin{aligned}
\bar r &= T_{\bar\mu} J - J \\
&= (T_{\bar\mu} J - T_\mu J) + (T_\mu J - J) \\
&\le (T_{\bar\mu} J - TJ) + \big(T_\mu J - T_\mu(T_\mu^m \underline{J})\big) + (T_\mu^m \underline{J} - J) + \big(T_\mu^m(T_\mu \underline{J}) - T_\mu^m \underline{J}\big) \\
&\le \epsilon v + \alpha M(J - T_\mu^m \underline{J})v + \delta v + \alpha^m M(T_\mu \underline{J} - \underline{J})v \\
&\le (\epsilon + \delta)v + \alpha\delta v + \alpha^m M(r)v,
\end{aligned}$$

where the first inequality follows from $T_{\bar\mu} J \ge TJ$, and the second and third inequalities follow from Eqs. (6.17) and (6.21). From this relation we have

$$M(\bar r) \le \big(\epsilon + (1+\alpha)\delta\big) + \beta M(r),$$

where $\beta = \alpha^m$. Taking $\limsup$ as $k \to \infty$ in this relation, we obtain

$$\limsup_{k\to\infty} M(r) \le \frac{\epsilon + (1+\alpha)\delta}{1-\hat{\beta}}, \tag{6.24}$$

where $\hat{\beta} = \alpha^{\liminf_{k\to\infty} m_k}$.

Next we derive a relation between s and r. We have

$$\begin{aligned} s &= J_\mu - T_\mu^m \underline{J} \\ &= T_\mu^m J_\mu - T_\mu^m \underline{J} \\ &\le \alpha^m M(J_\mu - \underline{J})v \\ &\le \frac{\alpha^m}{1-\alpha} M(T_\mu \underline{J} - \underline{J})v \\ &= \frac{\alpha^m}{1-\alpha} M(r)v, \end{aligned}$$

where the first inequality follows from Eq. (6.21) and the second inequality follows by using Prop. 2.4(b). Thus we have $M(s) \le \frac{\alpha^m}{1-\alpha} M(r)$, from which by taking $\limsup$ of both sides and using Eq. (6.24), we obtain

$$\limsup_{k\to\infty} M(s) \le \frac{\hat{\beta}\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)(1-\hat{\beta})}. \tag{6.25}$$

Finally we derive a relation between t, $\bar{t}$, and r. We first note that

$$\begin{aligned} TJ - TJ^* &\le \alpha M(J - J^*)v \\ &= \alpha M(J - T_\mu^m \underline{J} + T_\mu^m \underline{J} - J^*)v \\ &\le \alpha M(J - T_\mu^m \underline{J})v + \alpha M(T_\mu^m \underline{J} - J^*)v \\ &\le \alpha\delta v + \alpha M(t)v. \end{aligned}$$

Using this relation, and Eqs. (6.17) and (6.21), we have

$$\begin{aligned} \bar{t} &= T_{\bar{\mu}}^{\bar{m}} J - J^* \\ &= (T_{\bar{\mu}}^{\bar{m}} J - T_{\bar{\mu}}^{\bar{m}-1} J) + \cdots + (T_{\bar{\mu}}^2 J - T_{\bar{\mu}} J) + (T_{\bar{\mu}} J - TJ) + (TJ - TJ^*) \\ &\le (\alpha^{\bar{m}-1} + \cdots + \alpha) M(T_{\bar{\mu}} J - J)v + \epsilon v + \alpha\delta v + \alpha M(t)v, \end{aligned}$$

so finally

$$M(\bar{t}) \le \frac{\alpha - \alpha^{\bar{m}}}{1-\alpha} M(\bar{r}) + (\epsilon + \alpha\delta) + \alpha M(t).$$

By taking $\limsup$ of both sides and using Eq. (6.24), it follows that

$$\limsup_{k\to\infty} M(t) \le \frac{(\alpha - \hat{\beta})\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\hat{\beta})} + \frac{\epsilon + \alpha\delta}{1-\alpha}. \tag{6.26}$$

We now combine Eqs. (6.23), (6.25), and (6.26). We obtain

$$\begin{aligned} \limsup_{k\to\infty} M(J_{\mu^k} - J^*) &\le \limsup_{k\to\infty} M(s) + \limsup_{k\to\infty} M(t) \\ &\le \frac{\hat{\beta}\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)(1-\hat{\beta})} + \frac{(\alpha - \hat{\beta})\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\hat{\beta})} + \frac{\epsilon + \alpha\delta}{1-\alpha} \\ &= \frac{\big(\hat{\beta}(1-\alpha) + (\alpha - \hat{\beta})\big)\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2(1-\hat{\beta})} + \frac{\epsilon + \alpha\delta}{1-\alpha} \\ &= \frac{\alpha\big(\epsilon + (1+\alpha)\delta\big)}{(1-\alpha)^2} + \frac{\epsilon + \alpha\delta}{1-\alpha} \\ &= \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \end{aligned}$$

This proves the result, since in view of $J_{\mu^k} \geq J^*$, we have $M(J_{\mu^k} - J^*) = \|J_{\mu^k} - J^*\|$. **Q.E.D.**

Note that generally, optimistic PI with approximations is susceptible to the instability phenomenon illustrated by Example 4.1. In particular, when $m_k = 1$ for all k in Eq. (6.17), the method becomes essentially identical to approximate VI. However, it appears that choices of m_k that are significatly larger than 1 should be helpful in connection with this difficulty. In particular, it can be verified that in Example 4.1, the method converges to the optimal cost function if m_k is sufficiently large.

A remarkable fact is that approximate VI, approximate PI, and approximate optimistic PI have very similar error bounds (cf. Props. 4.2, 5.3, and 6.3). Approximate VI has a slightly better bound, but insignificantly so in practical terms.

7. STOCHASTIC SHORTEST PATH PROBLEMS

The SSP problem is a total cost infinite horizon DP problem where:

(a) There is no discounting ($\alpha = 1$).

(b) The state space is $X = \{0, 1, \ldots, n\}$ and we are given transition probabilities, denoted by

$$p_{ij}(u) = P(x_{k+1} = j \mid x_k = i, u_k = u), \qquad i, j \in X,\ u \in U(i).$$

(c) The control constraint set $U(i)$ is a compact subset of a metric space for all $i \in X$.

(d) A cost $g(i, u)$ is incurred when control $u \in U(i)$ is selected at state i.

(e) State 0 is a special destination state, which is absorbing and cost-free, i.e.,

$$p_{00}(u) = 1,$$

and for all $u \in U(0)$, $g(0, u) = 0$.

We have assumed for convenience that the cost per stage does not depend on the successor state. This amounts to using expected cost per stage in all calculations. In particular, if the cost of applying control u at state i and moving to state j is $\tilde{g}(i, u, j)$, we use as cost per stage the expected cost

$$g(i, u) = \sum_{j=0,1,\ldots,n} p_{ij}(u)\tilde{g}(i, u, j),$$

and the subsequent analysis goes through with no change.

Since the destination 0 is cost-free and absorbing, the cost starting from 0 is zero for every policy. Accordingly, for all cost functions, we ignore the component that corresponds to 0, and define

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

The mappings T and T_μ are defined by

$$(TJ)(i) = \min_{u \in U(i)} \left[g(i, u) + \sum_{j=1}^{n} p_{ij}(u)J(j) \right], \qquad i = 1, \ldots, n,$$

$$(T_\mu J)(i) = g\big(i, \mu(i)\big) + \sum_{j=1}^{n} p_{ij}\big(\mu(i)\big)J(j), \qquad i = 1, \dots, n.$$

Note that H satisfies the Monotonicity Assumption 2.2.

We say that a policy μ is *proper* if, when using this policy, there is positive probability that the destination will be reached after at most n stages, regardless of the initial state; i.e., if

$$\rho_\mu = \max_{i=1,\dots,n} P\{x_n \neq 0 \mid x_0 = i, \mu\} < 1. \tag{7.1}$$

It can be seen that μ is proper if and only if in the Markov chain corresponding to μ, each state i is connected to the destination with a path of positive probability transitions.

Throughout this section *we assume that all policies are proper*. Without this assumption, the mapping T need not be a contraction, as is well known (see [BeT91]). On the other hand, we have the following proposition [see e.g., [BeT96], Prop. 2.2; earlier proofs were given by Veinott [Vei69] (who attributes the result to A. J. Hoffman), and Tseng [Tse90]]. Note that while [BeT96] assumes that $U(i)$ is finite for all i, the proof given there, in conjunction with the results of [BeT91], extends to the more general case considered here.

Proposition 7.1: Assume that all policies are proper. Then, there exists a vector $v = \big(v(1), \dots, v(n)\big)$ with positive components such that the Contraction Assumption 2.1 holds, and the modulus of contraction is given by

$$\alpha = \max_{i=1,\dots,n} \frac{v(i) - 1}{v(i)}.$$

There is a generalization of the preceding proposition to SSP problems with a destination 0 and a *countable* number of other states, denoted $1, 2, \dots$. Let $v(i)$ be the maximum (over all policies) expected number of stages up to termination, starting from state i. Then if $v(i)$ is finite and bounded over i, the mappings T and T_μ are contraction mappings with respect to the weighted sup-norm with weight vector $v = \big(v(1), v(2), \dots\big)$. The proof is similar to the proof of Prop. 7.1, and is given in [Ber12], Section 3.5 and Exercise 2.11.

We finally note that the weighted sup-norm contraction property of Prop. 7.1 is algorithmically significant, because it brings to bear the preceding algorithms and analysis. In particular, the error bounds of Sections 4 and 5 for approximate VI and PI are valid when specialized to SSP problems, and the optimistic PI method of Section 6.1 is convergent. We provide the corresponding analysis in the next two subsections.

7.1 Approximate Policy Iteration

Consider an approximate PI algorithm that generates a sequence of stationary policies $\{\mu^k\}$ and a corresponding sequence of approximate cost vectors $\{J_k\}$ satisfying for all k

$$\|J_k - J_{\mu^k}\| \leq \delta, \qquad \|T_{\mu^{k+1}} J_k - T J_k\| \leq \epsilon, \tag{7.2}$$

where δ and ϵ are some positive scalars, and $\|\cdot\|$ is the weighted sup-norm

$$\|J\| = \max_{i=1,\ldots,n} \frac{J(i)}{v(i)}.$$

The following proposition provides error bounds that are special cases of the ones of Props. 5.3 and 5.5.

Proposition 7.2: (Error Bound for Approximate PI) The sequence $\{\mu^k\}$ generated by the approximate PI algorithm (7.2) satisfies

$$\|J_{\mu^{k+1}} - J^*\| \leq \alpha \|J_{\mu^k} - J^*\| + \frac{\epsilon + 2\alpha\delta}{1-\alpha}, \tag{7.3}$$

and

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}, \tag{7.4}$$

where $\|\cdot\|$ is the weighted sup-norm of Prop. 7.1, and α is the associated contraction modulus. Moreover, when $\{\mu^k\}$ converges to some $\bar\mu$, in the sense that

$$\mu^{\bar k+1} = \mu^{\bar k} = \bar\mu \qquad \text{for some } \bar k,$$

we have

$$\|J_{\bar\mu} - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{1-\alpha}.$$

The standard result in the literature for the approximate PI method of this section has the form

$$\limsup_{k\to\infty} \max_{i=1,\ldots,n} \big|J_{\mu^k}(x) - J^*(x)\big| \leq \frac{n(1-\alpha+n)(\epsilon+2\delta)}{(1-\rho)^2}, \tag{7.5}$$

where ρ is defined as the maximal (over all initial states and policies) probability of the Markov chain not having terminated after n transitions (see [BeT96], Prop. 6.3). While the bounds (7.4) and (7.5) involve different norms (one is weighted and the other is unweighted) and different denominators, the error bound (7.4) seems stronger, particularly for large n.

7.2 Optimistic Policy Iteration

Consider the optimistic PI method, whereby starting with a vector $J_0 \in \Re^n$, we generate sequences $\{J_k\}$ and $\{\mu^k\}$ with the algorithm

$$T_{\mu^k} J_k = T J_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{7.6}$$

where $\{m_k\}$ is a sequence of positive integers. Then the convergence result and convergence rate estimates of Section 6 hold. In particular, we have the following.

Proposition 7.3: (Convergence of Optimistic PI) Assume that all stationary policies are proper, and let $\big\{(J_k, \mu^k)\big\}$ be a sequence generated by the optimistic PI algorithm (7.6), and assume that for some $c \geq 0$ we have

$$\|J_0 - TJ_0\| \leq c,$$

where $\|\cdot\|$ is the weighted sup-norm of Prop. 7.1. Then for all $k \geq 0$,

$$J_k + \frac{\alpha^k}{1-\alpha} c\, v \geq J_k + \frac{\beta_k}{1-\alpha} c\, v \geq J^* \geq J_k - \frac{(k+1)\alpha^k}{1-\alpha} c\, v, \tag{7.7}$$

where v and α are the weight vector and the contraction modulus of Prop. 7.1, and β_k is defined by

$$\beta_k = \begin{cases} 1 & \text{if } k = 0, \\ \alpha^{m_0+\cdots+m_{k-1}} & \text{if } k > 0, \end{cases} \tag{7.8}$$

with m_j, $j = 0, 1, \ldots$, being the integers used in the algorithm (7.6). Moreover, we have

$$\lim_{k\to\infty} \|J_k - J^*\| = 0,$$

and $J_{\mu^k} = J^*$ for all k greater than some index $\bar{k}$.

To our knowledge, this is the first convergence result for optimistic PI applied to SSP problems (earlier results by Williams and Baird [WiB93] for discounted MDP, may be easily extended to SSP, but require restrictive conditions, such as $TJ_0 \leq J_0$). A similar result can be obtained when the mappings T and T_μ are replaced in the algorithm (7.6) by any monotone mappings W and W_μ that are contractions with respect to a common weighted sup-norm, and have J^* and J_μ as their unique fixed points, respectively. This latter property is true in particular if W is the Gauss-Seidel mapping based on T [this is the mapping where $(WJ)(i)$ is computed by the same equation as $(TJ)(i)$ except that the previously calculated values $(WJ)(1), \ldots, (WJ)(i-1)$ are used in place of $J(1), \ldots, J(i-1)$].

7.3 Approximate Optimistic Policy Iteration

Consider an optimistic approximate PI algorithm that generates a sequence of stationary policies $\{\mu^k\}$ and a corresponding sequence of approximate cost vectors $\{J_k\}$ satisfying for all k

$$\|J_k - T_{\mu^k}^{m_k} J_{k-1}\| \leq \delta, \qquad \|T_{\mu^{k+1}} J_k - TJ_k\| \leq \epsilon, \qquad k = 0, 1, \ldots, \tag{7.9}$$

[cf. Eq. (6.17)]. The following proposition provides an error bound that is a special case of the one of Prop. 6.3 (the Semilinear Monotonic Contraction 6.1 is clearly satisfied under our assumptions).

Proposition 7.4: (Error Bound for Optimistic Approximate PI) The sequence $\{\mu^k\}$ generated by the approximate PI algorithm (7.9) satisfies

$$\limsup_{k\to\infty} \|J_{\mu^k} - J^*\| \le \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}. \tag{7.10}$$

8. CONCLUSIONS

We have considered an abstract and broadly applicable DP model based on weighted sup-norm contractions, and provided a review of classical results and extensions to approximation methods that are the focus of current research. By virtue of its abstract character, the analysis provides insight into fundamental convergence properties and error bounds of exact and approximate VI and PI algorithms. Moreover, it allows simple proofs of results that would be difficult and/or tedious to obtain by alternative methods. The power of our analysis was illustrated by its application to SSP problems, where substantial improvements of the existing results on PI algorithms were obtained.

9. REFERENCES

[BeS78] 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

[BeT91] Bertsekas, D. P., and Tsitsiklis, J. N., 1991. "An Analysis of Stochastic Shortest Path Problems," Math. Operations Research, Vol. 16, pp. 580-595.

[BeT96] Bertsekas, D. P., and Tsitsiklis, J. N., 1996. Neuro-Dynamic Programming, Athena Scientific, Belmont, MA.

[BeY10a] Bertsekas, D. P., and Yu, H., 2010. "Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2831, MIT; to appear in Mathematics of Operations Research.

[BeY10b] Bertsekas, D. P., and Yu, H., 2010. "Asynchronous Distributed Policy Iteration in Dynamic Programming," Proc. of Allerton Conf. on Information Sciences and Systems.

[Ber77] Bertsekas, D. P., 1977. "Monotone Mappings with Application in Dynamic Programming," SIAM J. on Control and Optimization, Vol. 15, pp. 438-464.

[Ber07] Bertsekas, D. P., 2007. Dynamic Programming and Optimal Control, 3rd Edition, Vol. II, Athena Scientific, Belmont, MA.

[Ber11] Bertsekas, D. P., 2011. "λ-Policy Iteration: A Review and a New Implementation," Lab. for Information and Decision Systems Report LIDS-P-2874, MIT; to appear in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, by F. Lewis and D. Liu (eds.), IEEE Press Computational Intelligence Series.

[CaR11] Canbolat, P. G., and Rothblum, U. G., 2011. “(Approximate) Iterated Successive Approximations Algorithm for Sequential Decision Processes,” Technical Report, The Technion - Israel Institute of Technology; Annals of Operations Research, to appear.

[Den67] Denardo, E. V., 1967. “Contraction Mappings in the Theory Underlying Dynamic Programming,” SIAM Review, Vol. 9, pp. 165-177.

[Har72] Harrison, J. M., 1972. “Discrete Dynamic Programming with Unbounded Rewards,” Ann. Math. Stat., Vol. 43, pp. 636-644.

[Lip73] Lippman, S. A., 1973. “Semi-Markov Decision Processes with Unbounded Rewards,” Management Sci., Vol. 21, pp. 717-731.

[Lip75] Lippman, S. A., 1975. “On Dynamic Programming with Unbounded Rewards,” Management Sci., Vol. 19, pp. 1225-1233.

[Put94] Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming, J. Wiley, N.Y.

[Rot79] Rothblum, U. G., 1979. “Iterated Successive Approximation for Sequential Decision Processes,” in Stochastic Control and Optimization, by J. W. B. van Overhagen and H. C. Tijms (eds), Vrije University, Amsterdam.

[Sch11] Scherrer, B., 2011. “Performance Bounds for λ-Policy Iteration and Application to the Game of Tetris,” INRIA Lorraine Report, France.

[Sch12] Scherrer, B., 2012. “On the Use of Non-Stationary Policies for Infinite-Horizon Discounted Markov Decision Processes,” INRIA Lorraine Report, France.

[ThS10a] Thiery, C., and Scherrer, B., 2010. “Least-Squares λ-Policy Iteration: Bias-Variance Trade-off in Control Problems,” in ICML’10: Proc. of the 27th Annual International Conf. on Machine Learning.

[ThS10b] Thiery, C., and Scherrer, B., 2010. “Performance Bound for Approximate Optimistic Policy Iteration,” Technical Report, INRIA.

[Tse90] Tseng, P., 1990. “Solving H-Horizon, Stationary Markov Decision Problems in Time Proportional to $\log(H)$,” Operations Research Letters, Vol. 9, pp. 287-297.

[Vei69] Veinott, A. F., Jr., 1969. “Discrete Dynamic Programming with Sensitive Discount Optimality Criteria,” Ann. Math. Statist., Vol. 40, pp. 1635-1660.

[VeP84] Verd’u, 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.

[VeP87] Verd’u, S., and Poor, H. V., 1987. “Abstract Dynamic Programming Models under Commutativity Conditions,” SIAM J. on Control and Optimization, Vol. 25, pp. 990-1006.

[WiB93] Williams, R. J., and Baird, L. C., 1993. “Analysis of Some Incremental Variants of Policy Iteration: First Steps Toward Understanding Actor-Critic Learning Systems,” Report NU-CCS-93-11, College of Computer Science, Northeastern University, Boston, MA.

[YuB11] Yu, H., and Bertsekas, D. P., 2011. “Q-Learning and Policy Iteration Algorithms for Stochastic Shortest Path Problems,” Lab. for Information and Decision Systems Report LIDS-P-2871, MIT; to appear in Annals of OR.

[YuB12] Yu, H., and Bertsekas, D. P., 2012. “Weighted Bellman Eqations and their Applications in Dynamic Programming,” Lab. for Information and Decision Systems Report LIDS-P-2876, MIT.