Stochastic Shortest Path Problems
Under Weak Conditions

Dimitri P. Bertsekas† and Huizhen Yu‡

Abstract

In this paper we consider finite-state stochastic shortest path problems, and we address some of the complications due to the presence of transition cycles with nonpositive length. In particular, we assume that the optimal cost function is real-valued, that a standard compactness and continuity condition holds, and that there exists at least one proper policy, i.e., a stationary policy that terminates with probability one. Under these conditions, value and policy iteration may not converge to the optimal cost function, which in turn may not satisfy Bellman’s equation. Moreover, an improper policy may be optimal and superior to all proper policies. We propose and analyze forms of value and policy iteration for finding a policy that is optimal within the class of proper policies. In the special case where all expected transition costs are nonnegative we provide a transformation method and algorithms for finding a policy that is optimal over all policies.

1. INTRODUCTION

Stochastic shortest path (SSP) problems are a major class of infinite horizon total cost Markov decision processes (MDP) with a termination state. In this paper, we consider SSP problems with a state space $X = \{t, 1, \ldots, n\}$, where $t$ is the termination state. The control space is denoted by $U$, and the set of feasible controls at state $x$ is denoted by $U(x)$. From state $x$ under control $u \in U(x)$, a transition to state $y$ occurs with probability $p_{xy}(u)$ and incurs an expected one-stage cost $g(x,u)$. At state $t$ we have $p_{tt}(u) = 1$, $g(t,u) = 0$, for all $u \in U(t)$, i.e., $t$ is absorbing and cost-free. The goal is to reach $t$ with minimal total expected cost.

The principal issues here revolve around the solution of Bellman’s equation and the convergence of the classical algorithms of value iteration (VI for short) and policy iteration (PI for short). An important classification of stationary policies in SSP is between proper (those that guarantee eventual termination and are of principal interest in shortest path applications) and improper (those that do not). It is well-known that the most favorable results hold under the assumption that there exists at least one proper policy and that each improper policy generates infinite cost starting from at least one initial state. Then, assuming also

‡ Huizhen Yu is with the Dept. of Computing Science, University of Alberta, Edmonton, Canada, and has been supported by a grant from Alberta Innovates – Technology Futures.
a finiteness or compactness condition on $U$, and a continuity condition on $p_{xy}(\cdot)$ and $g(x, \cdot)$, the optimal cost function $J^*$ is the unique solution of Bellman’s equation, and appropriate forms of VI and PI yield $J^*$ and an optimal policy that is proper [BeT91].

In this paper, we will consider the SSP problem under weaker assumptions, where these favorable results cannot be expected to hold. In particular, we will replace the infinite cost assumption of improper policies with the condition that $J^*$ is real-valued. Under our assumptions, $J^*$ need not be a solution of Bellman’s equation, and even when it is, it may not be obtained by VI starting from any initial condition other than $J^*$ itself, while the standard form of PI may not be convergent. We will show instead that $\hat{J}$, which is the optimal cost function over proper policies only, is the unique solution of Bellman’s equation within the class of functions $J \geq \hat{J}$. Moreover VI converges to $\hat{J}$ from any initial condition $J \geq \hat{J}$, while a form of PI yields a sequence of proper policies that asymptotically attains the optimal value $\hat{J}$. Results of this type are unusual in the MDP literature, where $J^*$ is typically shown to satisfy Bellman’s equation. Our line of analysis is also unusual, and relies on a perturbation argument, which induces a more effective discrimination between proper and improper policies in terms of finiteness of their cost functions. This argument depends critically on the assumption that $J^*$ is real-valued.

In what follows in this section we will introduce our notation, terminology, and assumptions, and we will explain the nature of our analysis and its relation to the existing literature. In Section 2, we will prove our main results. In Section 3, we will consider the important special case where $g(x,u) \geq 0$ for all $(x,u)$, and show how we can effectively overcome the pathological behaviors we described. In particular, we will show how we can transform the problem to an equivalent favorably structured problem, for which $J^*$ is the unique solution of Bellman’s equation and is equal to $\hat{J}$, while powerful VI and PI convergence results hold.

1.1. Notation and Terminology

We first introduce definitions of various types of policies and cost functions. By a nonstationary policy we mean a sequence of the form $\pi = \{\mu_0, \mu_1, \ldots\}$, where each function $\mu_k$, $k = 0, 1, \ldots$, maps each state $x$ to a control in $U(x)$ (in this paper we restrict ourselves to Markov nonrandomized policies). We define the total cost of a nonstationary policy $\pi$ for initial state $x$ to be

$$J_\pi(x) = \limsup_{N \to \infty} J_{\pi,N}(x)$$

with $J_{\pi,N}(x)$ being the expected $N$-stage cost of $\pi$ for state $x$:

$$J_{\pi,N}(x) = E \left[ \sum_{k=0}^{N-1} g(x_k, u_k) \mid x_0 = x \right],$$

where $x_k$ and $u_k$ denote the state and control, respectively, at time $k$, and the expectation is with respect to the probability law of $\{x_0, u_0, \ldots, x_{N-1}, u_{N-1}\}$ induced by $\pi$. The use of lim sup in the definition of $J_\pi$
is necessary because the limit of \( J_{\pi,N}(x) \) as \( N \to \infty \) may not exist. However, the statements of our results, our analysis, and our algorithms are also valid if \( \lim \sup \) is replaced by \( \lim \inf \). The optimal cost at state \( x \), denoted \( J^*(x) \), is the infimum of \( J_\pi(x) \) over \( \pi \). Note that in general there may exist states \( x \) such that \( J_\pi(x) = \infty \) or \( J_\pi(x) = -\infty \) for some policies \( \pi \), as well as \( J^*(x) = \infty \) or \( J^*(x) = -\infty \).

Regarding notation, we denote by \( \mathbb{R} \) the set of real numbers, and by \( \mathcal{E} \) the set of extended real numbers: 
\[ \mathcal{E} = \mathbb{R} \cup \{ \infty, -\infty \}. \]
Vector inequalities of the form \( J \leq J' \) are to be interpreted componentwise, i.e., \( J(x) \leq J'(x) \) for all \( x \). Similarly, limits of function sequences are to be interpreted pointwise. Since \( J_\pi(t) = 0 \) for all \( \pi \) and \( J^*(t) = 0 \), we will ignore in various expressions the component \( J(t) \) of a cost function. Thus the various cost functions arising in our development are of the form \( J = (J(1), \ldots, J(n)) \), and they will be viewed as elements of the space \( \mathcal{E}^n \), the set of \( n \) dimensional vectors whose entries are either real numbers or \( -\infty \) or \( +\infty \), or of the space \( \mathbb{R}^n \) if their components are real. In Section 3, where we will focus on vectors with nonnegative components, \( J \) will belong to the nonnegative orthant of \( \mathbb{R}^n \), denoted \( \mathbb{R}_+^n \), or the nonnegative orthant of \( \mathcal{E}^n \), denoted \( \mathcal{E}_+^n \).

If \( \pi = \{\mu_0, \mu_1, \ldots\} \) is a stationary policy, i.e., all \( \mu_k \) are equal to some \( \mu \), \( \pi \) and \( J_\pi \) are also denoted by \( \mu \) and \( J_\mu \), respectively. The set of all stationary policies is denoted by \( \mathcal{M} \). In this paper, we will aim to find optimal policies exclusively within \( \mathcal{M} \), so in the absence of a statement to the contrary, by “policy” we will mean a stationary policy.

We say that \( \mu \in \mathcal{M} \) is proper if when using \( \mu \), there is positive probability that the termination state \( t \) will be reached after at most \( n \) stages, regardless of the initial state; i.e., if
\[
\rho_\mu = \max_{x=1, \ldots, n} P\{x_n \neq t \mid x_0 = x, \mu\} < 1.
\]
Otherwise, we say that \( \mu \) is improper. It can be seen that \( \mu \) is proper if and only if in the Markov chain corresponding to \( \mu \), each state \( x \) is connected to the termination state with a path of positive probability transitions, i.e., the only recurrent state is \( t \) and all other states are transient.

Let us introduce some notation relating to the mappings that arise in optimality conditions and algorithms. We consider the mapping \( H : \{1, \ldots, n\} \times U \times \mathcal{E}^n \to \mathcal{E} \) defined by
\[
H(x, u, J) = g(x, u) + \sum_{y=1}^{n} p_{xy}(u)J(y), \quad x = 1, \ldots, n, \ u \in U(x), \ J \in \mathcal{E}^n.
\]
For vectors \( J \in \mathcal{E}^n \) with components that take the values \( \infty \) and \( -\infty \), we adopt the rule \( \infty - \infty = \infty \) in the above definition of \( H \). However, the sum \( \infty - \infty \) never appears in our analysis. We also consider the mappings \( T_\mu, \mu \in \mathcal{M} \), and \( T \) defined by
\[
(T_\mu J)(x) = H(x, \mu(x), J), \quad (TJ)(x) = \inf_{u \in U(x)} H(x, u, J), \quad x = 1, \ldots, n, \ J \in \mathcal{E}^n.
\]
We will frequently use the monotonicity property of \( T_\mu \) and \( T \), i.e.,
\[
J \leq J' \quad \Rightarrow \quad T_\mu J \leq T_\mu J', \quad TJ \leq TJ'.
\]
The fixed point equations \( J^* = TJ^* \) and \( J_\mu = T_\mu J_\mu \) are commonly referred to as Bellman’s equations for the optimal cost function and for the cost function of \( \mu \), respectively. They are generally expected to hold in MDP models. We will see, however, in Section 2.3 that this is not the case if \( \mu \) is improper and \( g(x,u) \) can take both positive and negative values.

### 1.2. Background and Motivation

The SSP problem has been discussed in many sources, including the books [Pal67], [Der70], [Whi82], [Ber87], [BeT89], [HeL99], and [Ber12a], where it is sometimes referred to by other names such as “first passage problem” and “transient programming problem.” We may also view the SSP problem as a special case of an undiscounted MDP where both positive and negative transition costs are allowed, but no termination state is explicitly considered. The survey [Fei02] overviews the theory of these MDP, and the recent paper [Yu14] addresses the associated intricacies of the convergence of VI; see also the theory of positive bounded MDP discussed in Section 7.2 of [Put94]. However, these works require certain cost function convergence assumptions, which are restrictive in the context of the present paper. In particular, when specialized to deterministic shortest path problems, these assumptions require that each zero length cycle consists of zero length transitions.

We will now summarize the current SSP methodology, and the approaches and contributions of this paper. We first note that for any policy \( \mu \), the matrix that has components \( p_{xy}(\mu(x)) \), \( x, y = 1, \ldots, n \), is substochastic (some of its row sums may be less than 1) because there may be positive transition probability from \( x \) to \( t \). Consequently \( T_\mu \) may be a contraction for some \( \mu \), but not necessarily for all \( \mu \in \mathcal{M} \). For a proper policy \( \mu \), \( T_\mu \) is a weighted sup-norm contraction (see e.g., [Ber12a], Section 3.3), so that \( J_\mu = \lim_{k \to \infty} T_k^\mu J \) for all \( J \in \mathbb{R}^n \) [this is because \( J_0 \) is the unique fixed point of \( T_\mu \) within \( \mathbb{R}^n \), since by definition \( J_\mu = \lim_{k \to \infty} T_k^\mu J_0 \), where \( J_0 \) is the zero vector, \( J_0(x) \equiv 0 \)]. For an improper policy \( \mu \), \( T_\mu \) is not a contraction with respect to any norm. Moreover, \( T \) also need not be a contraction with respect to any norm. However, \( T \) is a weighted sup-norm contraction in the important special case where the control space is finite and all policies are proper. This was shown in [BeT96], Prop. 2.2 (see also [Ber12a], Prop. 3.3.1) for the case where the control space \( U \) is finite. The proof is easily extended to the case where \( U \) is a metric space, the set \( U(x) \) is compact for all \( x \in X \), and \( p_{xy}(\cdot) \) is continuous over \( U(x) \) for all \( x, y \in X \).

To deal with the lack of contraction property of \( T \), various assumptions have been formulated in the literature, under which results similar to the case of discounted finite-state MDP have been shown for SSP problems. These assumptions typically include the following condition, or the stronger version whereby \( U(x) \) is a finite set for all \( x \). An important consequence of this condition is that the infimum of \( H(x, \cdot, J) \) over \( U(x) \) is attained for all \( J \in \mathbb{R}^n \), i.e., there exists \( \mu \in \mathcal{M} \) such that \( T_\mu J = TJ \). However, the assumption does not guarantee the existence of an optimal policy, as shown by Example 6.7 of the total cost MDP survey
[Fei02], which is attributed to [CFM00]. This will also be seen in a very similar example, given in Section 2.2 in the context of a pathology relating to PI.

**Compactness and Continuity Condition:** The control space $U$ is a metric space, and for each $x, y \in X$, the set $U(x)$ is compact, $p_{xy}(\cdot)$ is continuous over $U(x)$, and $g(x, \cdot)$ is lower semicontinuous over $U(x)$.

In this paper we will generally assume the compactness and continuity condition, since serious anomalies may occur without it. An example is the classical blackmailer’s problem (see [Ber12a], Example 3.2.1), where

$$X = \{t, 1\}, \quad U(1) = (0, 1], \quad g(1, u) = -u, \quad p_{11}(u) = 1 - u^2, \quad H(1, u, J) = -u + (1 - u^2)J(1).$$

Here every stationary policy $\mu$ is proper and $T_\mu$ is a contraction, but $T$ is not a contraction, there is no optimal stationary policy, while the optimal cost $J^*(1)$ (which is $-\infty$) is attained by a nonstationary policy. A popular set of assumptions that we will aim to generalize, is the following.

**Classical SSP Conditions:**

(a) The compactness and continuity condition holds.

(b) There exists at least one proper policy.

(c) For every improper policy there is an initial state that has infinite cost under this policy.

Under the classical SSP conditions, it has been shown that $J^*$ is the unique fixed point of $T$ within $\mathbb{R}^n$. Moreover, a policy $\mu^*$ is optimal if and only if $T_{\mu^*}J = TJ^*$, and an optimal proper policy exists (so in particular $J^*$, being the cost function of a proper policy, is real-valued). In addition, $J^*$ can be computed by VI, $J^* = \lim_{k \to \infty} T^k J$, starting with any $J \in \mathbb{R}^n$. These results were given in [BeT91], following related results in several other works [Pal67], [Der70], [Whi82], [Ber87] (see [BeT91] for detailed references). An alternative assumption is that $g(x, u) \geq 0$ for all $(x, u)$, and that there exists a proper policy that is optimal. Our results bear some similarity to those obtained under this assumption (see [BeT91]), but are considerably stronger; see the comment preceding Prop. 2.3.

The standard form of PI generates a sequence $\{\mu^k\}$ according to

$$T_{\mu^{k+1}}J_{\mu^k} = TJ_{\mu^k},$$
starting from some initial policy $\mu^0$. It works in its strongest form when there are no improper policies. When there are improper policies and the classical SSP conditions hold, the initial policy should be proper (in which case subsequently generated policies are guaranteed to be proper). Otherwise, modifications to PI are needed, which allow it to work with improper initial policies, and also in the presence of approximate policy evaluation and asynchronism. Several types of modifications have been proposed, including a mixed VI and PI algorithm where policy evaluation is done through the use of an optimal stopping problem [BeY10], [YuB13a] (see the summary in [Ber12a], Section 3.5.3, for a discussion of this and other methods).

An important special case, with extensive literature and applications, is the deterministic shortest path problem, where for all $(x, u), p_{xy}(u)$ is equal to 1 for a single state $y$. Part (b) of the classical SSP conditions is equivalent to the existence of a proper policy, while part (c) is equivalent to all cycles of the graph of probability 1 transitions have positive length.

Many algorithms, including value iteration (VI for short) and policy iteration (PI for short), can be used to find a shortest path from every $x$ to $t$. However, if there are cycles of zero length, difficulties arise: the optimality equation can have multiple solutions, while VI and PI may break down, even in very simple problems. These difficulties are illustrated by the following example, and motivate the analysis of this paper.

![Diagram](image)

**Figure 1.1.** A deterministic shortest path problem with a single node 1 and a termination node $t$. At 1 there are two choices: a self-transition, which costs $a$, and a transition to $t$, which costs $b$.

**Example 1.1 (Deterministic Shortest Path Problem)**

Here there is a single state 1 in addition to the termination state $t$ (cf. Fig. 1.1). At state 1 there are two choices: a self-transition which costs $a$ and a transition to $t$, which costs $b$. The mapping $H$ has the form

$$H(1, u, J) = \begin{cases} a + J & \text{if } u: \text{ self transition}, \\ b & \text{if } u: \text{ transition to } t, \end{cases} J \in \mathbb{R},$$

and the mapping $T$ has the form

$$TJ = \min\{a + J, b\}, \quad J \in \mathbb{R}.$$

There are two policies: the policy that self-transitions at state 1, which is improper, and the policy that transitions from 1 to $t$, which is proper. When $a < 0$ the improper policy is optimal and we have $J^*(1) = -\infty$. The optimal cost is finite if $a > 0$ or $a = 0$, in which case the cycle has positive or zero length, respectively. Note the following:
(a) If \(a > 0\), the classical SSP conditions are satisfied, and the optimal cost, \(J^*(1) = b\), is the unique fixed point of \(T\).

(b) If \(a = 0\) and \(b \geq 0\), the set of fixed points of \(T\) (which has the form \(TJ = \min\{J, b\}\)), is the interval \((-\infty, b]\). Here the improper policy is optimal for all \(b \geq 0\), and the proper policy is also optimal if \(b = 0\).

(c) If \(a = 0\) and \(b > 0\), the proper policy is strictly suboptimal, yet its cost at state 1 (which is \(b\)) is a fixed point of \(T\). The optimal cost, \(J^*(1) = 0\), lies in the interior of the set of fixed points of \(T\), which is \((-\infty, b]\). Thus the VI method that generates \(\{T^kJ\}\) starting with \(J \neq J^*\) cannot find \(J^*\); in particular if \(J\) is a fixed point of \(T\), VI stops at \(J\), while if \(J\) is not a fixed point of \(T\) (i.e., \(J > b\)), VI terminates in two iterations at \(b \neq J^*(1)\). Moreover, the standard PI method is unreliable in the sense that starting with the suboptimal proper policy \(\mu\), it may stop with that policy because \((T\mu J_\mu)(1) = b = \min\{J_\mu(1), b\} = (TJ_\mu)(1)\) [the other/optimal policy \(\mu^*\) also satisfies \((T\mu^* J_{\mu^*})(1) = (TJ_{\mu^*})(1)\), so a rule for breaking the tie in favor of \(\mu^*\) is needed but such a rule may not be obvious in general].

(d) If \(a = 0\) and \(b < 0\), only the proper policy is optimal, and we have \(J^*(1) = b\). Here it can be seen that the VI sequence \(\{T^kJ\}\) converges to \(J^*(1)\) for all \(J \geq b\), but stops at \(J\) for all \(J < b\), since the set of fixed points of \(T\) is \((-\infty, b]\). Moreover, starting with either the proper policy \(\mu^*\) or the improper policy \(\mu\), the standard form of PI may oscillate, since \((T\mu^* J_{\mu^*})(1) = (TJ_{\mu^*})(1)\) as can be easily verified [the optimal policy \(\mu^*\) also satisfies \((T\mu^* J_{\mu^*})(1) = (TJ_{\mu^*})(1)\) but it is not clear how to break the tie; compare also with case (c) above].

As we have seen in case (c), VI may fail starting from \(J \neq J^*\). Actually in cases (b)-(d) the one-stage costs are either all nonnegative or nonpositive, so they belong to the classes of negative and positive DP models, respectively. From known results for such models, there is an initial condition, namely \(J = 0\), starting from which VI converges to \(J^*\). However, this is not necessarily the best initial condition; for example in deterministic shortest path problems initial conditions \(J \geq J^*\) are generally preferred and result in polynomial complexity computation assuming that all cycles have positive length. By contrast VI has only pseudopolynomial complexity when started from \(J = 0\). We will also see in the next section, that if there are both positive and negative one-stage costs, it may happen that \(J^*\) is not a fixed point of \(T\), so it cannot be obtained by VI or PI.

To address the distinction between the optimal cost \(J^*(x)\) and the optimal cost that may be achieved starting from \(x\) and using a proper policy, we introduce the optimal cost function over proper policies:

\[
\hat{J}(x) = \inf_{\mu: \text{proper}} J_\mu(x), \quad x \in X.
\]

Aside from its potential practical significance, the function \(\hat{J}\) plays a key role in the analysis of this paper, because when \(\hat{J} \neq J^*\), our VI and PI algorithms of Section 2 can only obtain \(\hat{J}\), and not \(J^*\).

In Section 2, we weaken part (c) of the classical SSP conditions, by assuming that \(J^*\) is real-valued instead of requiring that each improper policy has infinite cost from some initial states. We show that \(\hat{J}\) is the unique fixed point of \(T\) within the set \(\{J \mid J \geq \hat{J}\}\), and can be computed by VI starting from any \(J\)
within that set. In the two special cases where either there exists a proper policy \( \mu^* \) that is optimal, i.e., \( J^* = J_{\mu^*} \), or the cost function \( g \) is nonpositive, we show that \( J^* \) is the unique fixed point of \( T \) within the set \( \{J \mid J \geq J^*\} \), and the VI algorithm converges to \( J^* \) when started within this set. We provide an example showing that \( J^* \) may not be a fixed point of \( T \) if \( g \) can take both positive and negative values, and the SSP problem is nondeterministic.

The idea of the analysis is to introduce an additive perturbation \( \delta > 0 \) to the cost of each transition. Since \( J^* \) is real-valued, the cost function of each improper policy becomes infinite for some states, thereby bringing to bear the classical SSP conditions for the perturbed problem, while the cost function of each proper policy changes by an \( O(\delta) \) amount. We will also propose a valid version of PI that is based on this perturbation idea and converges to \( \hat{J} \), as a replacement of the standard form of PI, which may oscillate as we have seen in Example 1.1.

Since the VI and PI algorithms aim to converge to \( \hat{J} \), they cannot be used to compute \( J^* \) and an optimal policy in general. In Section 3 we resolve this issue for the case where, in addition to the compactness and continuity condition, all transition costs are nonnegative: \( g(x, u) \geq 0 \) for all \( x \) and \( u \in U(x) \) (the negative DP model). Under these conditions the classical forms of VI and PI may again fail, as demonstrated by cases (c) and (d) of Example 1.1. However, we will provide a transformation to an “equivalent” SSP problem for which the classical SSP conditions hold. Following this transformation, we may use the corresponding VI, which converges to \( J^* \) starting from any \( J \in \mathbb{R}^n \).

2. FINITE OPTIMAL COST CASE - A PERTURBATION APPROACH

In this section we allow the one-stage costs to be both positive and negative, but assume that \( J^* \) is real-valued and that there exists at least one proper policy. As a result, by adding a positive perturbation \( \delta \) to \( g \), we are guaranteed to drive to \( \infty \) the cost \( J_{\mu}(x) \) of each improper policy \( \mu \), for at least one state \( x \), thereby differentiating proper and improper policies.

We thus consider for each scalar \( \delta > 0 \) an SSP problem, referred to as the \( \delta \)-perturbed problem, which is identical to the original problem, except that the cost per stage is

\[
g_\delta(x, u) = \begin{cases} 
g(x, u) + \delta & \text{if } x = 1, \ldots, n, \\
0 & \text{if } x = t,
\end{cases}
\]

and the corresponding mappings \( T_{\mu, \delta} \) and \( T_\delta \) are given by

\[
T_{\mu, \delta}J = T_\mu J + \delta e, \quad T_\delta J = TJ + \delta e, \quad \forall J \in \mathbb{R}^n,
\]

where \( e \) is the unit function \([e(x) = 1]\). This problem has the same proper and improper policies as the original. The corresponding cost function of a policy \( \pi = \{\mu_0, \mu_1, \ldots\} \in \Pi \) is given by

\[
J_{\pi, \delta}(x) = \lim_{N \to \infty} \sup_{N} J_{\pi, \delta, N}(x),
\]
with
\[ J_{\pi, \delta, N}(x) = E \left[ \sum_{k=0}^{N-1} g_{\delta}(x_k, u_k) \mid x_0 = x \right] . \]

We denote by \( J_{\mu, \delta} \) the cost function of a stationary policy \( \mu \) for the \( \delta \)-perturbed problem, and by \( J_{\delta}^* \) the corresponding optimal cost function,
\[ J_{\delta}^* = \inf_{\pi \in \Pi} J_{\pi, \delta}. \]

Note that for every proper policy \( \mu \), the function \( J_{\mu, \delta} \) is real-valued, and that
\[ \lim_{\delta \downarrow 0} J_{\mu, \delta} = J_{\mu}. \]

This is because for a proper policy, the extra \( \delta \) cost per stage will be incurred only a finite expected number of times prior to termination, starting from any state. This is not so for improper policies, and in fact the idea behind perturbations is that the addition of \( \delta \) to the cost per stage in conjunction with the assumption that \( J^* \) is real-valued imply that if \( \mu \) is improper,
\[ J_{\mu, \delta}(x) = \lim_{k \to \infty} (T^k_{\mu, \delta} J)(x) = \infty \quad \text{for all } x \neq t \text{ that are recurrent under } \mu \text{ and all } J \in \mathbb{R}^n. \quad (2.1) \]

Thus part (c) of the classical SSP conditions holds for the \( \delta \)-perturbed problem, and the associated strong results noted in Section 1 come into play. In particular, we have the following proposition.

**Proposition 2.1:** Assume that the compactness and continuity condition holds, that there exists at least one proper policy, and that \( J^* \) is real-valued. Then for each \( \delta > 0 \):

(a) \( J_{\delta}^* \) is the unique solution of the equation
\[ J(x) = (TJ)(x) + \delta, \quad x = 1, \ldots, n. \]

(b) A policy \( \mu \) is optimal for the \( \delta \)-perturbed problem (\( J_{\mu, \delta} = J_{\delta}^* \)) if and only if \( T_{\mu} J_{\delta}^* = TJ_{\delta}^* \).

Moreover, all optimal policies for the \( \delta \)-perturbed problem are proper and there exists at least one such policy.

(c) The optimal cost function over proper policies \( \tilde{J} \) [cf. Eq. (1.2)] satisfies
\[ \tilde{J}(x) = \lim_{\delta \downarrow 0} J_{\delta}^*(x), \quad x = 1, \ldots, n. \]

**Proof:** The proof of parts (a) and (b) follows from the discussion preceding the proposition, and the results of [BeT91] noted in the introduction, which hold under the classical SSP conditions [the equation of part (a)
is Bellman’s equation for the \( \delta \)-perturbed problem. To prove part (c), we note that for an optimal proper policy \( \mu^*_\delta \) of the \( \delta \)-perturbed problem [cf. part (b)], we have

\[
\hat{J} = \inf_{\mu \text{ proper}} J_\mu \leq J_{\mu^*_\delta} \leq J_{\mu^*_\delta} = J^*_\delta \leq J_{\mu'}^*, \quad \forall \mu' : \text{ proper}. 
\]

Since for every proper policy \( \mu' \), we have \( \lim_{\delta \downarrow 0} J_{\mu', \delta} = J_{\mu'} \), it follows that

\[
\hat{J} \leq \lim_{\delta \downarrow 0} J_{\mu^*_\delta} \leq J_{\mu'}, \quad \forall \mu' : \text{ proper}.
\]

By taking the infimum over all \( \mu' \) that are proper, the result follows. Q.E.D.

2.1. Convergence of Value Iteration

The preceding perturbation-based analysis, can be used to investigate properties of \( \hat{J} \) by using properties of \( J^*_\delta \) and taking limit as \( \delta \downarrow 0 \). In particular, we use the preceding proposition to show that \( \hat{J} \) is a fixed point of \( T \), and can be obtained by VI starting from any \( J \geq \hat{J} \).

**Proposition 2.2:** Assume that the compactness and continuity condition holds, that there exists at least one proper policy, and that \( J^* \) is real-valued. Then:

(a) The optimal cost function over proper policies \( \hat{J} \) is the unique fixed point of \( T \) within the set \( \{ J \in \mathbb{R}^n \mid J \geq \hat{J} \} \).

(b) We have \( T^k J \to \hat{J} \) for every \( J \in \mathbb{R}^n \) with \( J \geq \hat{J} \).

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

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

\[
J^*_\delta = T J^*_\delta + \delta e \leq T_\mu J^*_\delta + \delta e.
\]

Taking limit as \( \delta \downarrow 0 \), and using Prop. 2.1(c), we obtain \( \hat{J} \leq T_\mu \hat{J} \) for all \( \mu \in \mathcal{M} \). Taking infimum over \( \mu \in \mathcal{M} \), it follows that \( \hat{J} \leq T \hat{J} \). Thus \( \hat{J} \) is a fixed point of \( T \).

For all \( J \in \mathbb{R}^n \) with \( J \geq \hat{J} \) and proper policies \( \mu \), we have by using the relation \( \hat{J} = T \hat{J} \) just shown,

\[
\hat{J} = \lim_{k \to \infty} T^k \hat{J} \leq \lim_{k \to \infty} T^k J \leq \lim_{k \to \infty} T^k_\mu J = J_\mu.
\]
Taking the infimum over all proper $\mu$, we obtain

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

This proves part (b) and also the claimed uniqueness property of $\hat{J}$ in part (a).

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

Note that if there exists a proper policy but $J^*$ is not real-valued, the mapping $T$ cannot have any real-valued fixed point. To see this, let $\tilde{J}$ be such a fixed point, and let $\epsilon$ be a scalar such that $\tilde{J} \leq J_0 + \epsilon e$, where $J_0$ is the zero vector. Since $T(J_0 + \epsilon e) \leq TJ_0 + \epsilon e$, it follows that $\tilde{J} = T^N \tilde{J} \leq T^N (J_0 + \epsilon e) \leq T^N J_0 + \epsilon e \leq J_{\pi,N} + \epsilon e$ for any $N$ and policy $\pi$. Taking lim sup with respect to $N$ and then infimum over $\pi$, it follows that $\tilde{J} \leq J^* + \epsilon e$. Since $J^*$ cannot take the value $\infty$ (by the existence of a proper policy), this shows that $J^*$ must be real-valued.

Note also that there may exist an improper policy $\mu$ that is strictly suboptimal and yet satisfies the optimality condition $T_\mu J^* = TJ^*$ [cf. case (d) of Example 1.1], so properness of $\mu$ is an essential assumption in Prop. 2.2(c). It is also possible that $\hat{J} = J^*$, but the only optimal policy is improper, as we will show by example in the next section (see Example 2.1). If there exists an optimal proper policy, i.e., a proper policy $\mu^*$ such that $J_{\mu^*} = \hat{J} = J^*$, we obtain the following proposition, which is new. The closest earlier result assumes in addition that $g(x,u) \geq 0$ for all $(x,u)$ (see [BeT91], Prop. 3).

**Proposition 2.3:** Assume that the compactness and continuity condition holds, and that there exists an optimal proper policy. Then:

(a) The optimal cost function $J^*$ is the unique fixed point of $T$ in the set $\{J \in \mathbb{R}^n \mid J \geq J^*\}$.

(b) We have $T^k J \to J^*$ for every $J \in \mathbb{R}^n$ with $J \geq J^*$.

(c) Let $\mu$ be a proper policy. Then $\mu$ is optimal if and only if $T_\mu J^* = TJ^*$.

**Proof:** Existence of a proper policy that is optimal implies both that $J^*$ is real-valued and that $J^* = \hat{J}$. The result then follows from Prop. 2.2. Q.E.D.

Another important special case where favorable results hold is when $g(x,u) \leq 0$ for all $(x,u)$. Then, it is well-known that $J^*$ is the unique fixed point of $T$ within the set $\{J \mid J^* \leq J \leq 0\}$, and the VI sequence $\{T^k J\}$ converges to $J^*$ starting from any $J$ within that set (see e.g., [Ber12a], Ch. 4, or [Put94], Section 7.2). In the following proposition, we will use Prop. 2.2 to obtain related results for SSP problems where $g$ may
take both positive and negative values. An example is an optimal stopping problem, where at each state $x$ we have cost $g(x, u) \geq 0$ for all $u$ except one that leads to the termination state $t$ with nonpositive cost. Classical problems of this type include searching among several sites for a valuable object, with nonnegative search costs and nonpositive stopping costs (stopping the search at every site is a proper policy guaranteeing that $\hat{J} \leq 0$).

**Proposition 2.4:** Assume that the compactness and continuity condition holds, that $\hat{J} \leq 0$, and that $J^*$ is real-valued. Then $J^*$ is equal to $\hat{J}$ and it is the unique fixed point of $T$ within the set \{ $J \in \mathbb{R}^n \mid J \geq J^*$ \}. Moreover, we have $T^k J \to J^*$ for every $J \in \mathbb{R}^n$ with $J \geq J^*$.

**Proof:** We first observe that the hypothesis $\hat{J} \leq 0$ implies that there exists at least one proper policy, so Prop. 2.2 applies, and shows that $\hat{J}$ is the unique fixed point of $T$ within the set \{ $J \in \mathbb{R}^n \mid J \geq \hat{J}$ \} and that $T^k J \to \hat{J}$ for all $J \in \mathbb{R}^n$ with $J \geq \hat{J}$. We will prove the result by showing that $\hat{J} = J^*$. Since generically we have $\hat{J} \geq J^*$, it will suffice to show the reverse inequality.

Let $J_0$ denote the zero function. Since $\hat{J}$ is a fixed point of $T$ and $\hat{J} \leq J_0$, we have

$$\hat{J} = \lim_{k \to \infty} T^k \hat{J} \leq \limsup_{k \to \infty} T^k J_0. \tag{2.2}$$

Also, for each policy $\pi = \{ \mu_0, \mu_1, \ldots \}$, we have

$$J_\pi = \limsup_{k \to \infty} T_{\mu_0} \cdots T_{\mu_{k-1}} J_0.$$  

Since

$$T^k J_0 \leq T_{\mu_0} \cdots T_{\mu_{k-1}} J_0, \quad \forall \ k \geq 0,$$

it follows that $\limsup_{k \to \infty} T^k J_0 \leq J_\pi$, so by taking the infimum over $\pi$, we have

$$\limsup_{k \to \infty} T^k J_0 \leq J^*. \tag{2.3}$$

Combining Eqs. (2.2) and (2.3), it follows that $\hat{J} \leq J^*$, so that $\hat{J} = J^*$. **Q.E.D.**

The assumption $\hat{J} \leq 0$ of the preceding proposition will be satisfied if we can find a proper policy $\mu$ such that $J_\mu \leq 0$. With this in mind, we note that the proposition applies to MDP where the compactness and continuity condition holds, and $J^*$ is real-valued and satisfies $J^* \leq 0$, even if there is no termination state. We can simply introduce an artificial termination state $t$, and for each $x = 1, \ldots, n$, a control of cost 0 that leads from $x$ to $t$, thereby creating a proper policy $\mu$ with $J_\mu = 0$, without affecting $J^*$. What is happening here is that for the subset of states $x$ for which $J^*(x)$ is maximum we must have $J^*(x) = 0$, so
this subset plays the role of a termination state. Without the assumption \( \hat{J} \leq 0 \), we may have \( J^* \neq \hat{J} \) even if \( J^* \leq 0 \), as case (c) of Example 1.1 shows.

In all of the preceding results, there is the question of finding \( J \geq \hat{J} \) with which to start VI. One possibility that may work is to use the cost function of a proper policy or an upper bound thereof. For example in a stopping problem we may use the cost function of the policy that stops at every state, and more generally we may try to introduce an artificial high stopping cost, which is our approach in Section 3. If it can be guaranteed that \( \hat{J} = J^* \), this approach will also yield \( J^* \). For example, when \( g \leq 0 \), we may use \( \hat{J} = 0 \) to start VI, as is well known. In the case where the classical SSP conditions hold, we may of course start VI with any \( J \in \mathbb{R}^n \), and it is generally recommended to use an upper bound to \( J^* \) rather than a lower bound, as noted earlier. In the case of the general convergence model, a method provided in [Fei02], p. 189, may be used to find an upper bound to \( J^* \).

We finally note that once the optimal cost function over proper policies \( \hat{J} \) is found by VI, there may still be an issue of finding a proper policy that attains \( \hat{J} \) when the control space \( U \) is finite. When \( \hat{J} = J^* \), a polynomial complexity algorithm for this is given in [FeY08]. The PI algorithm of the next section can also be used for this purpose, even if \( \hat{J} \neq J^* \), although its complexity properties are unknown at present.

### 2.2. A Policy Iteration Algorithm with Perturbations

We will now use our perturbation framework to deal with the oscillatory behavior of PI, which is illustrated in case (d) of Example 1.1. We will develop a perturbed version of the PI algorithm that generates a sequence of proper policies \( \{\mu^k\} \) such that \( J_{\mu^k} \rightarrow \hat{J} \), under the assumptions of Prop. 2.2, which include the existence of a proper policy and that \( J^* \) is real-valued. The algorithm generates the sequence \( \{\mu^k\} \) as follows.

Let \( \{\delta_k\} \) be a positive sequence with \( \delta_k \downarrow 0 \), and let \( \mu^0 \) be any proper policy. At iteration \( k \), we have a proper policy \( \mu^k \), and we generate \( \mu^{k+1} \) according to

\[
T_{\mu^{k+1}} J_{\mu^{k}, \delta_k} = T J_{\mu^{k}, \delta_k}.
\]  

(2.4)

Note that since \( \mu^k \) is proper, \( J_{\mu^k, \delta_k} \) is the unique fixed point of the mapping \( T_{\mu^k, \delta_k} \) given by

\[
T_{\mu^k, \delta_k} J = T_{\mu^k} J + \delta_k e.
\]

The policy \( \mu^{k+1} \) of Eq. (2.4) exists by the compactness and continuity condition. We claim that \( \mu^{k+1} \) is proper. To see this, note that

\[
T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k} + \delta_k e \leq T_{\mu^k} J_{\mu^k, \delta_k} + \delta_k e = J_{\mu^k, \delta_k},
\]

so that

\[
T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} \leq T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} = T J_{\mu^k, \delta_k} + \delta_k e \leq J_{\mu^k, \delta_k}, \quad \forall \ m \geq 1.
\]

(2.5)
Since \( J_{\mu^k, \delta_k} \) forms an upper bound to \( T_{\mu^{k+1}, \delta_k} \), it follows that \( \mu^{k+1} \) is proper [if it were improper, we would have \( (T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k})(x) \to \infty \) for some \( x \); cf. Eq. (2.1)]. Thus the sequence \( \{\mu^k\} \) generated by the perturbed PI algorithm (2.4) is well-defined and consists of proper policies. We have the following proposition.

**Proposition 2.5:** Assume that the compactness and continuity condition holds, that there exists at least one proper policy, and that \( J^* \) is real-valued. Then the sequence \( \{J_{\mu^k}\} \) generated by the perturbed PI algorithm (2.4) satisfies \( J_{\mu^k} \to \hat{J} \).

**Proof:** Using Eq. (2.5), we have

\[
J_{\mu^{k+1}, \delta_{k+1}} \leq J_{\mu^{k+1}, \delta_k} = \lim_{m \to \infty} T_{\mu^{k+1}, \delta_k} J_{\mu^k, \delta_k} \leq TJ_{\mu^k, \delta_k} + \delta_k e \leq J_{\mu^k, \delta_k},
\]

where the equality holds because \( \mu^{k+1} \) is proper, as shown earlier. Taking the limit as \( k \to \infty \), and noting that \( J_{\mu^{k+1}, \delta_{k+1}} \geq \hat{J} \), we see that \( J_{\mu^k, \delta_k} \downarrow J^+ \) for some \( J^+ \geq \hat{J} \), and we obtain

\[
\hat{J} \leq J^+ = \lim_{k \to \infty} TJ_{\mu^k, \delta_k}.
\]

We also have

\[
\inf_{u \in U(x)} H(x, u, J^+) \leq \lim_{k \to \infty} \inf_{u \in U(x)} H(x, u, J_{\mu^k, \delta_k}) \leq \inf_{u \in U(x)} H(x, u, J_{\mu^k, \delta_k}) \leq \lim_{k \to \infty} \inf_{u \in U(x)} H(x, u, J^+),
\]

where the first inequality follows from the fact \( J^+ \leq J_{\mu^k, \delta_k} \), which implies that \( H(x, u, J^+) \leq H(x, u, J_{\mu^k, \delta_k}) \), and the first equality follows from the continuity of \( H(x, u, \cdot) \). Thus equality holds throughout above, so that

\[
\lim_{k \to \infty} TJ_{\mu^k, \delta_k} = TJ^+.
\]

Combining Eqs. (2.6) and (2.7), we obtain \( \hat{J} \leq J^+ = TJ^+ \). Since by Prop. 2.2, \( \hat{J} \) is the unique fixed point of \( T \) within \( \{J \in \mathbb{R}^n \mid \beta \geq \hat{J}\} \), it follows that \( J^+ = \hat{J} \). Thus \( J_{\mu^k, \delta_k} \downarrow \hat{J} \), and since \( J_{\mu^k, \delta_k} \geq J_{\mu^k} \geq \hat{J} \), we have \( J_{\mu^k} \to \hat{J} \). Q.E.D.

Proposition 2.5 guarantees the monotonic convergence of \( \{J_{\mu^k, \delta_k}\} \downarrow \hat{J} \) (see the preceding proof) and the (possibly nonmonotonic) convergence \( \{J_{\mu^k}\} \to \hat{J} \). When the control space \( U \) is finite, Prop. 2.5 also implies that the generated policies \( \mu^k \) will be optimal for all \( k \) sufficiently large. The reason is that the set of policies is finite and there exists a sufficiently small \( \epsilon > 0 \), such that for all nonoptimal \( \mu \) there is some state \( x \) such that \( J_\mu(x) \geq J^*(x) + \epsilon \). This convergence behavior should be contrasted with the behavior of PI without perturbations, which may lead to difficulties, as noted earlier [cf. case (d) of Example 1.1].
However, when the control space $U$ is infinite, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. If $\{\mu^k\}_k$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that $\lim_{k \to \infty, k \in K} \mu^k(x) = \bar{\mu}(x)$ for all $x = 1, \ldots, n$, then from the preceding proof [cf. Eq. (2.7)], we have

$$T_{\mu^k} J_{\mu^{k-1}, \delta_{k-1}} = T J_{\mu^{k-1}, \delta_{k-1}} \to T \bar{J}.$$ 

Taking the limit as $k \to \infty$, $k \in K$, we obtain

$$T_{\bar{\mu}} \bar{J} \leq \lim_{k \to \infty, k \in K} T J_{\mu^{k-1}, \delta_{k-1}} = T \bar{J},$$

where the inequality follows from the lower semicontinuity of $g(x, \cdot)$ and the continuity of $p_{x,y}(\cdot)$. Since we also have $T_{\bar{\mu}} \bar{J} \geq T \bar{J}$, we see that $T_{\bar{\mu}} \bar{J} = T \bar{J}$. Thus $\bar{\mu}$ satisfies the optimality condition of Prop. 2.2(c), and $\bar{\mu}$ is optimal if it is proper. On the other hand, properness of the limit policy $\bar{\mu}$ may not be guaranteed, even if $\bar{J} = J^*$. In fact the generated sequence of proper policies $\{\mu^k\}$ satisfies $\lim_{k \to \infty} J_{\mu^k} \to \bar{J} = J^*$, yet $\{\mu^k\}$ may converge to an improper policy that is strictly suboptimal, as shown by the following example, which is similar to Example 6.7 of [Fei02].

![Figure 2.1](image)

**Figure 2.1.** A stochastic shortest path problem with two states 1, 2, and a termination state $t$. Here we have $\bar{J} = J^*$, but there is no optimal policy. Any sequence of proper policies $\{\mu^k\}$ with $\mu^k(1) \to 0$ is asymptotically optimal in the sense that $J_{\mu^k} \to J^*$, yet $\{\mu^k\}$ converges to the strictly suboptimal improper policy for which $u = 0$ at state 1.

**Example 2.1 (A Counterexample for the Perturbation-Based PI Algorithm)**

Consider two states 1 and 2, in addition to the termination state $t$; see Fig. 2.1. At state 1 we must choose $u \in [0,1]$, with expected cost equal to $u$. Then, we transition to state 2 with probability $\sqrt{u}$, and we self-transition to state 1 with probability $1 - \sqrt{u}$. From state 2 we transition to $t$ with cost -1. Thus we have

$$H(1, u, J) = u + (1 - \sqrt{u}) J(1) + \sqrt{u} J(2), \quad \forall J \in \mathbb{R}^2, \ u \in [0,1],$$

$$H(2, u, J) = -1, \quad \forall J \in \mathbb{R}^2, \ u \in U(2).$$

Here there is a unique improper policy $\mu$: it chooses $u = 0$ at state 1, and has cost $J_\mu(1) = 1$. Every policy $\mu$ with $\mu(1) \in (0,1]$ is proper, and $J_\mu$ can be obtained by solving the equation $J_\mu = T_\mu J_\mu$. We have $J_\mu(2) = -1$, so that

$$J_\mu(1) = \mu(1) + (1 - \sqrt{\mu(1)}) J_\mu(1) - \sqrt{\mu(1)},$$
and we obtain \( J_\mu(1) = \sqrt{\mu(1)} - 1 \). Thus, \( \hat{J}(1) = J^*(1) = -1 \). The perturbation-based PI algorithm will generate a sequence of proper policies \( \{\mu^k\} \) with \( \mu^k(1) \to 0 \). Any such sequence is asymptotically optimal in the sense that \( J_{\mu^k} \to \hat{J} = J^* \), yet it converges to the strictly suboptimal improper policy.

2.3. Discussion: Fixed Points of \( T \)

While \( \hat{J} \) is a fixed point of \( T \) under our assumptions, as shown in Prop. 2.2(a), an interesting question is whether and under what conditions \( J^* \) is also a fixed point of \( T \). The following example shows that if \( g(x,u) \) can take both positive and negative values, and the problem is stochastic, \( J^* \) may not be a fixed point of \( T \). Moreover, \( J_\mu \) need not be a fixed point of \( T_\mu \), when \( \mu \) is improper.

Example 2.2 (A Problem Where \( J^* \) is not a Fixed Point of \( T \))

Consider the SSP problem of Fig. 2.2, which involves a single improper policy \( \mu \) (we will introduce a proper policy later). All transitions under \( \mu \) are deterministic as shown, except at state 1 where the successor state is 3 or 5 with equal probability 1/2. Under the definition (1.1) of \( J_\mu \) in terms of lim sup, we have

\[
J_\mu(1) = 0, \quad J_\mu(2) = J_\mu(5) = 1, \quad J_\mu(3) = J_\mu(7) = 0, \quad J_\mu(4) = J_\mu(6) = 2,
\]

so that the Bellman equation at state 1

\[
J_\mu(1) = \frac{1}{2}(J_\mu(2) + J_\mu(5))
\]

is not satisfied. Thus \( J_\mu \) is not a fixed point of \( T_\mu \). If for \( x = 1, \ldots, 7 \), we introduce another control that leads from \( x \) to \( t \) with cost \( a > 1 \), we create a proper policy that is strictly suboptimal, while not affecting \( J^* \), which again is not a fixed point of \( T \).

Of course there are known cases where \( J^* \) is a fixed point of \( T \), including the SSP problem under the classical SSP conditions, the case where \( g \geq 0 \), the case where \( g \leq 0 \), the positive bounded model discussed in Section 7.2 of [Put94], and the general convergence models discussed in [Fei02] and [Yu14]. The assumptions of all these models are violated by the preceding example.

It turns out that \( J^* \) is a fixed point of \( T \) in the special case of a deterministic shortest path problem, i.e., an SSP problem where for each \( x \) and \( u \in U(x) \), there is a unique successor state denoted \( f(x,u) \). For such a problem, the mappings \( T_\mu \) and \( T \) take the form

\[
(T_\mu J)(x) = g(x,\mu(x)) + J(f(x,\mu(x))), \quad (TJ)(x) = \inf_{\mu \in \mathcal{M}} (T_\mu J)(x).
\]

Moreover, by using the definition (1.1) of \( J_\mu \) in terms of lim sup, we have for all \( \mu \in \mathcal{M} \) (proper or improper),

\[
J_\mu(x) = g(x,\mu(x)) + J_\mu(f(x,\mu(x))) = (T_\mu J_\mu)(x), \quad x = 1, \ldots, n.
\]
For any policy \( \pi = \{ \mu_0, \mu_1, \ldots \} \), using the definition (1.1) of \( J_\pi \) in terms of \( \limsup \), we have for all \( x \),

\[
J_\pi(x) = g(x, \mu_0(x)) + J_{\pi_1}(f(x, \mu_0(x))),
\]

where \( \pi_1 = \{ \mu_1, \mu_2, \ldots \} \). By taking the infimum of the left-hand side over \( \pi \) and the infimum of the right-hand side over \( \pi_1 \) and then \( \mu_0 \), we obtain \( J^* = TJ^* \). Note that this argument does not require any assumptions other than the deterministic character of the problem, and holds even if the state space is infinite. The mathematical reason why Bellman’s equation \( J_\mu = T_\mu J_\mu \) may not hold for stochastic problems and improper \( \mu \) (cf. Example 2.2) is that \( \limsup \) may not commute with the expected value that is inherent in \( T_\mu \), and the preceding proof breaks down.

The improper policy of Example 2.2 may be viewed as a randomized policy for a deterministic shortest path problem: this is the problem for which at state 1 we must (deterministically) choose one of the two successor states 2 and 5. For this deterministic problem, \( J^* \) takes the same values as in Example 2.2 for all \( x \neq 1 \), but it takes the value \( J^*(1) = 1 \) rather than \( J^*(1) = 0 \). Thus, remarkably, once we allow randomized policies into the problem in the manner of Example 2.2, the optimal cost function ceases to be a fixed point of \( T \) and simultaneously its optimal cost at state 1 is improved.

On the other hand, if we allow randomization, the optimal cost function \( \hat{J}_r \) over randomized policies that are proper will not be improved over \( \hat{J} \) under the assumptions of Prop. 2.2 and also assuming that \( U \) is a separable metric space. To see this, note that control randomization amounts to replacing each \( U(x) \) by \( P(U(x)) \), the set of probability measures over \( U(x) \), and appropriately redefining \( g(x, \cdot) \) and \( p_{xy}(\cdot) \) over \( P(U(x)) \). Then the compactness and continuity condition can be shown to hold for the randomized controls problem, assuming it holds for the nonrandomized controls problem. In particular, from Prop. 7.22
of [BeS78], \( P(U(x)) \) is a compact metric space, while Prop. 7.21 of [BeS78] characterizes convergence in this space, and together with Lemma 7.14 of [BeS78], can be used to show that the one-stage cost remains lower semicontinuous with respect to the randomized control. It can then be seen that for the DP mapping \( T_r \) of the randomized controls problem, we have \( T_r J = TJ \) for all \( J \in \mathbb{R}^n \). Since \( \hat{J} \) is a real-valued fixed point of \( T \) and hence of \( T_r \), it follows that \( J^* \), the optimal cost function over randomized Markov policies, is real-valued (see the comment following the proof of Prop. 2.2). Thus from Prop. 2.2(a), \( \hat{J}_r \) is the unique fixed point of \( T_r \), and hence also of \( T \), over all \( J \geq \hat{J}_r \). Since \( \hat{J} \geq \hat{J}_r \) and \( \hat{J} \) is a fixed point of \( T \), we obtain that \( \hat{J} = \hat{J}_r \).

The unusual behavior just described underscores the intricate mathematical behavior exhibited by SSP problems with both positive and negative costs per stage. Moreover, the reduction in optimal cost through the use of randomized policies highlights the question of deciding on the set of policies that is appropriate for a given application. In the case of practical shortest path applications, the proper policies, the focus of this paper, are arguably the most interesting. Thus optimization over these policies only is in itself an important problem. Embedding this problem within an optimization over all Markov nonrandomized possibly improper policies, as we have done, is analytically expedient as it brings to bear the analysis and insights of MDP. We have seen that by allowing improper policies the optimal cost may be reduced [cf. case (c) of Example 1.1], and by allowing improper randomized policies the optimal cost may be reduced still further (cf. Example 2.2). However, these facts are of no concern if we are solely interested in optimization over proper policies. Still the unusual behavior that we have demonstrated suggests that the analysis of SSP problems with randomized policies is a theoretically interesting as well as challenging research direction.

3. THE NONNEGATIVE ONE-STAGE COSTS CASE

In this section we aim to eliminate the difficulties of the VI and PI algorithms in the case where \( g(x, u) \geq 0 \) for all \( x \). For this case we can appeal to the results of nonnegative-cost MDP, which we summarize in the following proposition.

**Proposition 3.1:** Assume that \( g(x, u) \geq 0 \) for all \( x \) and \( u \in U(x) \). Then:

(a) \( J^* = TJ^* \), and if \( J \in \mathcal{E}_+^n \) satisfies \( J = TJ \), then \( J \geq J^* \).

(b) A policy \( \mu^* \) is optimal if and only if \( T_{\mu^*} J^* = TJ^* \).

(c) If the compactness and continuity condition holds, then there exists at least one optimal policy, and we have \( T^k J \to J^* \) for all \( J \in \mathcal{E}_+^n \) with \( J \leq J^* \).

For proofs of the various parts of the proposition and related textbook accounts, see [Put94], Ch. 7, and [Ber12a], Ch. 4. The monograph [BeS78] contains extensions to infinite state space frameworks that address
the associated measurability issues, including the convergence of VI starting from \( J \) with \( 0 \leq J \leq J^* \). The recent paper [YuB13c] provides PI algorithms that can deal with these measurability issues, and establishes the convergence of VI for a broader range of initial conditions. The paper [Ber77], and the monographs [BeS78], Ch. 5, and [Ber13], Ch. 4, provide an abstract DP formulation that points a way to extensions of the results of this section.

It is well-known that for nonnegative-cost MDP, the standard PI and VI algorithms may encounter difficulties. In particular, case (c) of Example 1.1 shows that there may exist a strictly suboptimal policy \( \mu \) satisfying \( T_\mu J_\mu = TJ_\mu \), so PI will terminate with such a policy. Moreover, there may exist fixed points \( J \) of \( T \) satisfying \( J \geq J^* \) and \( J \neq J^* \), and VI will terminate starting with such a fixed point; this occurs in case (c) of Example 1.1, where \( \hat{J} \) is a fixed point of \( T \) and \( \hat{J}(1) > J^*(1) \). In the next subsection, we will address these difficulties by introducing a suitable transformation.

### 3.1. Reformulation to an Equivalent Problem

We will define another SSP problem, which will be shown to be “equivalent” to the given problem. In the new SSP problem, all the states \( x \) in the set

\[
X^0 = \{ x \in X \mid J^*(x) = 0 \},
\]

including the termination state \( t \), are merged into a new termination state \( \tilde{t} \). This idea appears to be novel in the context of MDP, and came from a result of our recent paper [YuB13c] on convergence of VI, which shows that for a class of Borel space nonnegative-cost DP models we have \( T^k J \to J^* \) for all \( J \) such that \( J^* \leq J \leq cJ^* \) for some \( c > 1 \) (or \( 0 \leq J \leq cJ^* \) for some \( c > 1 \) if the compactness and continuity condition holds in addition); a related result is given by [Whi79]. Similar ideas, based on eliminating cycles of zero cost transitions by merging them to a single state have been mentioned in the context of deterministic shortest path algorithms.

Note that from the Bellman equation \( J^* = TJ^* \) [cf. Prop. 3.1(a)], and assuming the compactness and continuity condition (which implies that there exists \( \mu \) such that \( T_\mu J = TJ \) for all \( J \in E_+^n \)), we obtain the following useful characterization of \( X^0 \):

\[
x \in X^0 \quad \text{if and only if there exists } u \in U(x) \text{ such that } g(x, u) = 0 \text{ and } p_{xy}(u) = 0 \text{ for all } y \notin X^0. \tag{3.1}
\]

Algorithms for constructing \( X^0 \), to be given later in this section, will rely on this characterization.

It is possible that \( J^* \) is the zero vector and \( X^0 = X \) [this is true in case (b) of Example 1.1]. The algorithm for finding \( X^0 \), to be given later in this section, can still be used to verify that this is the case thereby solving the problem, but to facilitate the exposition, \textit{we assume without essential loss of generality that the set } \( X^+ \text{ given by } \)

\[
X^+ = \{ x \in X \mid J^*(x) > 0 \},
\]
is nonempty. We introduce a new SSP problem as follows.

**Definition of Equivalent SSP Problem:**

*State space:* \( \mathcal{X} = \mathcal{X}^+ \cup \{ \mathcal{I} \} \), where \( \mathcal{I} \) is a cost-free and absorbing termination state.

*Controls and one-stage costs:* For \( x \in \mathcal{X}^+ \), \( U(x) = U(x) \) and \( \overline{g}(x,u) = g(x,u) \), for all \( u \in U(x) \).

*Transition probabilities:* For \( x \in \mathcal{X}^+ \), \( u \in U(x) \), we have

\[
\overline{p}_{xy}(u) = \begin{cases} 
  p_{xy}(u) & \text{if } y \in \mathcal{X}^+, \\
  \sum_{z \in \mathcal{X}^0} p_{xz}(u) & \text{if } y = \mathcal{I}.
\end{cases}
\]

The optimal cost vector for the equivalent SSP problem is denoted by \( \overline{J} \), and is the smallest nonnegative solution of the corresponding Bellman equation \( J = \overline{T}J \), where

\[
(\overline{T}J)(x) \overset{\text{def}}{=} \inf_{u \in U(x)} \left[ \overline{g}(x,u) + \sum_{y \in \mathcal{X}^+} \overline{p}_{xy}(u)J(y) \right] = \inf_{u \in U(x)} \left[ g(x,u) + \sum_{y \in \mathcal{X}^+} p_{xy}(u)J(y) \right], \quad x \in \mathcal{X}^+, \quad (3.2)
\]

[cf. Prop. 3.1(a)].

We will now clarify the relation of the equivalent SSP problem with the given SSP problem (also referred to as the “original” SSP problem). The key fact for our purposes is that \( \overline{J} \) coincides with \( J^* \) on the set \( \mathcal{X}^+ \), and if \( J^* \) is real-valued (which can be guaranteed by the existence of a proper policy for the original SSP problem), then the equivalent SSP problem satisfies the classical SSP conditions given in the introduction. As a result we may transfer the available analytical results from the equivalent SSP problem to the original SSP problem. We may also apply the VI and PI methods discussed in Section 1 to the equivalent SSP problem, after first obtaining the set \( \mathcal{X}^0 \), in order to compute the solution of the original problem.

**Proposition 3.2:** Assume that \( g(x,u) \geq 0 \) for all \( x \) and \( u \in U(x) \), and that the compactness and continuity condition holds. Then:

(a) \( J^*(x) = \overline{J}(x) \) for all \( x \in \mathcal{X}^+ \).

(b) A policy \( \mu^* \) is optimal for the original SSP problem if and only if

\[
\mu^*(x) = \overline{\mu}(x), \quad \forall \ x \in \mathcal{X}^+,
\]

\[
g(x,\mu^*(x)) = 0, \quad p_{xy}(\mu^*(x)) = 0, \quad \forall \ x \in \mathcal{X}^0, \ y \in \mathcal{X}^+, \quad (3.3)
\]

where \( \overline{\mu} \) is an optimal policy for the equivalent SSP problem.
(c) If $J^*$ is real-valued, then in the equivalent SSP problem every improper policy has infinite cost starting from some initial state. Moreover, there exists at least one proper policy, so the equivalent SSP problem satisfies the classical SSP conditions.

(d) If $0 < J^*(x) < \infty$ for all $x = 1, \ldots, n$, then the original SSP problem satisfies the classical SSP conditions.

Proof:  
(a) Let us extend $\bar{J}$ to a vector $\hat{J}$ that has domain $X$:

\[ \hat{J}(x) = \begin{cases} \bar{J}(x) & \text{if } x \in X^+, \\ 0 & \text{if } x \in X^0. \end{cases} \]

Then from the Bellman equation $\bar{J} = T\bar{J}$ [cf. Prop. 3.1(a)], and the definition (3.2) of $T$, we have $\bar{J}(x) = (T\bar{J})(x)$ for all $x \in X^+$, while from Eq. (3.1), we have $(T\bar{J})(x) = \bar{J}(x)$ for all $x \in X^0$. Thus $\bar{J}$ is a fixed point of $T$, so that $\bar{J} \geq J^*$ [since $J^*$ is the smallest nonnegative fixed point of $T$, cf. Prop. 3.1(a)], and hence $\bar{J}(x) \geq J^*(x)$ for all $x \in X^+$. Conversely, the restriction of $J^*$ to $X^+$ is a solution of the Bellman equation $J = TJ$, with $T$ given by Eq. (3.2), so we have $\bar{J}(x) \leq J^*(x)$ for all $x \in X^+$ [since $\bar{J}$ is the smallest nonnegative fixed point of $T$, cf. Prop. 3.1(a)].

(b) A policy $\mu^*$ is optimal for the original SSP problem if and only if $J^* = TJ^* = T\mu^*J^*$ [cf. Prop. 3.1(b)], or

\[ J^*(x) = \inf_{u \in U(x)} \left[ g(x, u) + \sum_{y=1}^{n} p_{xy}(u)J^*(y) \right] = g(x, \mu^*(x)) + \sum_{y=1}^{n} p_{xy}(\mu^*(x))J^*(y), \quad \forall x = 1, \ldots, n. \]

Equivalently, $\mu^*$ is optimal if and only if

\[ J^*(x) = \inf_{u \in U(x)} \left[ g(x, u) + \sum_{y \in X^+} p_{xy}(u)J^*(y) \right] = g(x, \mu^*(x)) + \sum_{y \in X^+} p_{xy}(\mu^*(x))J^*(y), \quad \forall x \in X^+, \quad (3.4) \]

and Eq. (3.3) holds. Using part (a), the Bellman equation $\bar{J} = T\bar{J}$, and the definition (3.2) of $T$, we see that Eq. (3.4) is the necessary and sufficient condition for optimality of the restriction of $\mu^*$ to $X^+$ in the equivalent SSP problem, and the result follows.

(c) Let $\mu$ be an improper policy of the equivalent SSP problem. Then the Markov chain induced by $\mu$ contains a recurrent class $R$ that consists of states $x$ with $J^*(x) > 0$ [since we have $J^*(x) > 0$, for all $x \in X^+$]. The average cost per stage of this recurrent class must be positive [otherwise we would have $g(x, \mu(x)) = 0$ for all $x \in R$, implying that $J^*(x) = 0$ and hence $J^*(x) = 0$ for all $x \in R$]. It follows from the inequalities of [Put94, Theorem 9.4.1(a), p. 472] applied to a single policy, that $\bar{J}(x) = J^*(x) = \infty$ for all $x \in R$. To prove the existence of a proper policy, we note that by the compactness and continuity condition, the original SSP
problem has an optimal policy [cf. Prop. 3.1(c)], and since $J^*$ is real-valued this policy cannot be improper (as we have just shown, improper policies have infinite cost starting from at least one initial state).

(d) Here the original and equivalent SSP problems coincide, so the result follows from part (c). Q.E.D.

The following proposition provides analytical and computational results for the original SSP problem, using the equivalent SSP problem.

**Proposition 3.3**: Assume that $g(x,u) \geq 0$ for all $x$ and $u \in U(x)$, that the compactness and continuity condition holds, and that $J^*$ is real-valued. Consider the set

$$J = \mathbb{R}_+^n \cup \{J \notin \mathbb{R}_+^n \mid J(x) = 0, \forall x = 1, \ldots, n, \text{ with } J^*(x) = 0\}.$$

Then:

(a) $J^*$ is the unique fixed point of $T$ within $J$.

(b) We have $T^kJ \to J^*$ for any $J \in J$.

**Proof**: (a) Since the classical SSP conditions hold for the equivalent SSP problem by Prop. 3.2(c), $J$ is the unique fixed point of $\bar{T}$. From Prop. 3.2(a) and the definition of the equivalent SSP problem, it follows that $J^*$ is the unique fixed point of $T$ within the set of $J \in \mathbb{R}_+^n$ with $J(x) = 0$ for all $x \in X^0$. From Prop. 3(a) of [BeT91], we also have that $J^*$ is the unique fixed point of $T$ within $\mathbb{R}_+^n$, thus completing the proof.

(b) Similar to the proof of part (a), the VI algorithm for the equivalent SSP problem is convergent to $\bar{J}$ from any initial condition. Together with the convergence of VI starting from any $J \in \mathbb{R}_+^n$ [cf. Prop. 3(b) of [BeT91]], this implies the result. Q.E.D.

Note that Prop. 3.3(a) narrows down the range of possible fixed points of $T$ relative to known results under the nonnegativity conditions ([BeT91], Prop. 3, asserts uniqueness of the fixed point of $T$ only within $\mathbb{R}_+^n$). In particular, if $J^*(x) > 0$ for all $x = 1, \ldots, n$, $J^*$ is the unique fixed point of $T$ within $\mathbb{R}_+^n$. However, case (b) of Example 1.1 shows that the set $J$ cannot be replaced by $\mathbb{R}_+^n$ in the statement of the proposition. To make use of the proposition we should know the sets $X^0$ and $X^+$, and also be able to deal with the case where $J^*$ is not real-valued. We will look at these questions in the next two subsections, respectively.

**Algorithm for Constructing $X^0$ and $X^+$**

In practice, the sets $X^0$ and $X^+$ can often be determined by a simple analysis that relies on the special structure of the given problem. When this is not so, we may compute these sets with a simple algorithm.
that requires at most \( n \) iterations. Let
\[
\hat{U}(x) = \{ u \in U(x) \mid g(x, u) = 0 \}, \quad x \in X.
\]
Denote \( X_1 = \{ x \in X \mid \hat{U}(x) \neq \emptyset \} \), and define for \( k \geq 1 \),
\[
X_{k+1} = \{ x \in X_k \mid \text{there exists } u \in \hat{U}(x) \text{ such that } y \in X_k \text{ for all } y \text{ with } p_{xy}(u) > 0 \}.
\]

It can be seen with a straightforward induction that
\[
X_k = \{ x \in X \mid (T^k J_0)(x) = 0 \},
\]
where \( J_0 \) is the zero vector. Clearly we have \( X_{k+1} \subset X_k \) for all \( k \), and since \( X \) is finite, the algorithm terminates at some iteration \( \bar{k} \) with \( X_{\bar{k}+1} = X_{\bar{k}} \). It can be seen that the set \( X_{\bar{k}} \) is equal to \( X^0 \), since we have \( T^k J_0 \uparrow J^* \) under the compactness and continuity condition [cf. Prop. 3.1(c)]. If the number of state-control pairs is finite, say \( m \), each iteration requires \( O(m) \) computation, so the complexity of the algorithm for finding \( X^0 \) and \( X^+ \) is \( O(mn) \). Note that this algorithm finds \( X^0 \) even when \( X^0 = X \) and \( X^+ \) is empty.

### 3.2. The Case Where \( J^* \) is not Real-Valued

In order to use effectively the equivalent SSP problem, \( J^* \) must be real-valued, so that Prop. 3.3 can apply. It turns out that this restriction can be circumvented by introducing an artificial high-cost stopping action at each state, thereby making \( J^* \) real-valued.

In particular, let us introduce for each scalar \( c > 0 \), an SSP problem that is identical to the original, except that an additional control is added to each \( U(x) \), under which the transition to the termination state \( t \) occurs with probability 1 and a cost \( c \) is incurred. We refer to this problem as the \( c \)-SSP problem, and we denote its optimal cost vector by \( \hat{J}_c \). Note that
\[
\hat{J}_c(x) \leq c, \quad \hat{J}_c(x) \leq J^*(x), \quad \forall \ x \in X, \ c > 0,
\]
and that \( \hat{J}_c \) is the unique fixed point of the corresponding mapping \( \hat{T}_c \) given by
\[
(\hat{T}_c J)(x) = \min \left[ c, \inf_{u \in U(x)} \left[ g(x, u) + \sum_{y=1}^{n} p_{xy}(u)J(y) \right] \right], \quad x = 1, \ldots, n, \tag{3.5}
\]
within the set of \( J \in \mathbb{R}^n \) with \( J(x) = 0 \) for all \( x \in X^0 \) [cf. Prop. 3.3(a)]. Let
\[
X^f = \{ x \in X \mid J^*(x) < \infty \}, \quad X^\infty = \{ x \in X \mid J^*(x) = \infty \}.
\]
We have the following proposition.
Proposition 3.4: Assume that \( g(x, u) \geq 0 \) for all \( x \) and \( u \in U(x) \), and that the compactness and continuity condition holds. Then:

(a) We have
\[
\lim_{c \to \infty} \hat{J}_c(x) = J^*(x), \quad \forall \ x \in X.
\]

(b) If the control space \( U \) is finite, there exists \( \bar{c} > 0 \) such that for all \( c \geq \bar{c} \), we have
\[
\hat{J}_c(x) = J^*(x), \quad \forall \ x \in X^f,
\]
and if \( \hat{\mu} \) is an optimal policy for the \( c \)-SSP problem, then any policy \( \mu^* \) such that \( \mu^*(x) = \hat{\mu}(x) \) for \( x \in X^f \) is optimal for the original SSP problem.

We will first state the following preliminary lemma, which can be easily proved by induction.

Lemma 3.1: Let
\[
c_k = \max_{x \in X} (T^k J_0)(x), \quad k = 0, 1, \ldots \tag{3.6}
\]
where \( J_0 \) is the zero vector. Then for all \( c \geq c_k \), the VI algorithm, starting from \( J_0 \), produces identical results for the \( c \)-SSP and the original SSP problems up to iteration \( k \):
\[
\hat{T}_c^m J_0 = T^m J_0, \quad \forall \ m \leq k.
\]

Proof of Prop. 3.4: (a) Let \( J_0 \) be the zero vector. We have
\[
J^* \geq \lim_{c \to \infty} \hat{J}_c \geq \hat{J}_{c_k} \geq \hat{T}_{c_k} J_0 = T^k J_0,
\]
where \( c_k \) is given by Eq. (3.6), and the last equality follows from Lemma 3.1. Since \( T^k J_0 \uparrow J^* \) [cf. Prop. 3.1(c)], we obtain \( \lim_{c \to \infty} \hat{J}_c = J^* \).

(b) The result is clearly true if \( J^* \) is real-valued, since then for \( c \geq \max_{x \in X} J^*(x) \), the VI algorithm starting from \( J_0 \) produces identical results for the \( c \)-SSP and original SSP problems, so for such \( c \), \( \hat{J}_c = J^* \). For the case where \( X^\infty \) is nonempty, we will formulate “reduced” versions of these two problems, where the states in \( X^f \) do not communicate with the states in \( X^\infty \), so that by restricting the reduced problems to \( X^f \), we revert to the case where \( J^* \) is real-valued.
Indeed, for both the $c$-SSP problem and the original problem, let us replace the constraint set $U(x)$ by the set
\[
\hat{U}(x) = \begin{cases} 
U(x) & \text{if } x \in X^\infty, \\
\{ u \in U(x) \mid p_{xy}(u) = 0, \forall y \in X^\infty \} & \text{if } x \in X^f,
\end{cases}
\]
so that the infinite cost states in $X^\infty$ are unreachable from the finite cost states in $X^f$. We refer to the problems thus created as the “reduced” $c$-SSP problem and the “reduced” original SSP problem.

We now apply Prop. 3.1(b) to both the original and the “reduced” original SSP problems. In the original SSP problem, for each $x \in X^f$, the infimum in the expression
\[
\min_{u \in U(x)} \left[ g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J^*(y) \right],
\]
is attained for some $u \in \hat{U}(x)$ [controls $u \notin \hat{U}(x)$ are inferior because they lead with positive probability to states $y \in X^\infty$]. Thus an optimal policy for the original SSP problem is feasible for the reduced original SSP problem, and hence also optimal since the optimal cost cannot become smaller at any state when passing from the original to the reduced original problem. Similarly, for each $x \in X^f$, the infimum in the expression
\[
\min \left[ c, \min_{u \in U(x)} \left[ g(x,u) + \sum_{y=1}^{n} p_{xy}(u)J_c(y) \right] \right],
\]
[cf. Eq. (3.5)] is attained for some $u \in \hat{U}(x)$ once $c$ becomes sufficiently large. The reason is that for $y \in X^\infty$, $\hat{J}_c(y) \uparrow J^*(y) = \infty$, so for sufficiently large $c$, each control $u \notin \hat{U}(x)$ becomes inferior to the controls $u \in \hat{U}(x)$, for which $p_{xy}(u) = 0$. Thus by taking $c$ large enough, an optimal policy for the original $c$-SSP problem, becomes feasible and hence optimal for the reduced $c$-SSP problem [here the size of the “large enough” $c$ depends on $x$ and $u$, so finiteness of $X$ and $U(x)$ is important for this argument]. We have thus shown that the optimal cost vector of the reduced original SSP problem is also $J^*$, and the optimal cost vector of the reduced $c$-SSP problem is also $\hat{J}_c$ for sufficiently large $c$.

Clearly, starting from any state in $X^f$ it is impossible to transition to a state $x \in X^\infty$ in the reduced original SSP problem and the reduced $c$-SSP problem. Thus if we restrict these problems to the set of states in $X^f$, we will not affect their optimal costs for these states. Since $J^*$ is real-valued in $X^f$, it follows that for sufficiently large $c$, these optimal cost vectors are equal (as noted in the beginning of the proof), i.e., $\hat{J}_c(x) = J^*(x)$ for all $x \in X^f$. Q.E.D.

We note that the compactness and continuity condition is needed for Prop. 3.4(a) to hold, while the finiteness of $U$ is needed for Prop. 3.4(b) to hold. We demonstrate this with examples.

**Example 3.1 (Counterexamples)**

Consider the SSP problem of Fig. 3.1, and two cases:
Convex programs are special cases of NLPs.

LPs are solved by gradient/Newton methods.

Modern view: Post 1990s

(a) \( U(2) = (0, 1] \), so the compactness and continuity condition is violated. Then we have \( J^*(1) = J^*(2) = \infty \).

Let us now calculate \( \hat{J}_c(1) \) and \( \hat{J}_c(2) \) from the Bellman equation

\[
\hat{J}_c(1) = \min \left[ c, 1 + \hat{J}_c(1) \right], \quad \hat{J}_c(2) = \min \left[ c, \inf_{u \in (0, 1]} \left( 1 - \sqrt{u} + u \hat{J}_c(1) \right) \right].
\]

The equation on the left yields \( \hat{J}_c(1) = c \), and for \( c \geq 1 \), the minimization in the equation on the right takes the form

\[
\inf_{u \in (0, 1]} \left[ 1 - \sqrt{u} + uc \right].
\]

By setting to 0 the derivative with respect to \( u \), we see that the infimum is attained at \( u = 1/(2c)^2 \), yielding

\[
\hat{J}_c(2) = 1 - \frac{1}{4c}, \quad \text{for } c \geq 1.
\]

Thus we have \( \lim_{c \to \infty} \hat{J}_c(2) = 1 \), while \( J^*(2) = \infty \), so the conclusion of Prop. 3.4(a) fails to hold.

(b) \( U(2) = [0, 1] \), so the compactness and continuity condition is satisfied, but \( U \) is not finite. Then we have \( J^*(1) = \infty \), \( J^*(2) = 1 \). Similar to case (a), we calculate \( \hat{J}_c(1) \) and \( \hat{J}_c(2) \) from the Bellman equation. An essentially identical calculation as the one of case (a) yields the same results for \( c \geq 1 \):

\[
\hat{J}_c(1) = c, \quad \hat{J}_c(2) = 1 - \frac{1}{4c}.
\]

Thus we have \( \lim_{c \to \infty} \hat{J}_c(1) = J^*(1) = \infty \), and \( \lim_{c \to \infty} \hat{J}_c(2) = J^*(2) = 1 \), consistently with Prop. 3.4(a).

However, \( \hat{J}_c(2) < J^*(2) \) for all \( c \), so the conclusion of Prop. 3.4(b) fails to hold.

Proposition 3.4(b) suggests a procedure to solve a problem for which \( J^* \) is not real-valued, but the one-stage cost is nonnegative and the control space is finite:

1. Use the algorithm of Section 3.1 to compute the sets \( X^0 \) and \( X^+ \).
2. Introduce for all \( x \in X^+ \) a stopping action with a cost \( c > 0 \).
3. Solve the equivalent SSP problem and obtain a candidate optimal policy for the original SSP using Prop. 3.2(b). This step can be done with any one of the PI and VI algorithms noted in Section 1.
(4) Check that $c$ is high enough, by testing to see if the candidate optimal policy changes as $c$ is increased, or satisfies the optimality condition of Prop. 3.1(b). If it does, the current policy is optimal; if it does not, increase $c$ by some fixed factor and repeat from Step (3).

Finally, let us note that the analysis and algorithms of this section are essentially about general (not necessarily SSP-type) nonnegative-cost finite-state MDP under the compactness and continuity condition. The reason is that such MDP can be trivially converted to SSP problems by adding an artificial termination state that is not reachable from any of the other states with a feasible transition. For these MDP, with the exception of the mixed VI and PI algorithm of our recent work [YuB13c] (Section 5.2), no valid exact or approximate PI method has been known. The transformation also brings to bear the available extensive methodology for SSP under the classical SSP conditions (VI, several versions of PI, including some that are asynchronous, and linear programming), as well as simulation-based algorithms, including Q-learning and approximate PI (see e.g., [Tsi94], [BeT96], [BeY10], [Ber12a], [YuB13a], [YuB13b]).

4. CONCLUDING REMARKS

There are a few issues that we have not addressed and remain subjects for further research. Extensions to infinite-state SSP problems are interesting, as well as the further investigation of the case where the one-stage cost can take both positive and negative values. In particular, when $\hat{J} \neq J^*$, the characterization of the set of fixed points of $T$, and algorithms for computing $J^*$ and an optimal (possibly improper) policy remain open questions. The complications arising from the use of randomized policies are worth exploring. The computational complexity properties of the PI algorithm of Section 2.2 also require investigation. A broader issue relates to extensions of the notion of proper policy to DP models, which are more general than SSP problems. One such extension is the notion of a regular policy, which was introduced within the context of semicontractive models in [Ber13]. This connection points the way to generalizations of the results of this paper, among others, to affine monotonic models, including exponential cost and risk-sensitive MDP. In such models, regular policies can be related to policies that stabilize an associated linear discrete-time system (see [Ber13], Section 4.5), and an analysis that parallels the one of the present paper is possible.

We also have not fully discussed the case of the SSP problem where $\hat{J} \leq 0$ and its special case where $g(x, u) \leq 0$ for all $(x, u)$. While we have shown in Prop. 2.4 that $\hat{J} = J^*$, and that VI converges to $J^*$ starting from any $J \in \mathbb{R}^n$ with $J \geq J^*$, the standard PI algorithm may encounter difficulties [see case (d) of Example 1.1]. However, the perturbation-based PI algorithm of Section 2.2 applies. Another valid PI algorithm for the case $g \leq 0$, which does not require that $J^*$ be real-valued or the existence and iterative generation of proper policies, is the $\lambda$-PI algorithm introduced in [BeI96] and further studied in [ThS10], [Ber12b], [Sch13], [YuB12] (see Section 4.3.3 of [Ber13]). This algorithm is not specific to the SSP problem, and does not make use of the presence of a termination state. Still another possibility is a mixed VI and
PI algorithm, given in Section 4.2 of [YuB13c], which also does not rely on proper policies. This algorithm converges from above to $J^*$ (which need not be real-valued), and applies even in the case of infinite (Borel) state and control spaces.

5. REFERENCES


