February 2014 (Revised July 2014) **Report LIDS - 2916**

Infinite-Space Shortest Path Problems and Semicontractive Dynamic Programming†

Dimitri P. Bertsekas‡

Abstract

In this paper we consider deterministic and stochastic shortest path problems with an infinite, possibly uncountable, number of states. The objective is to reach or approach a special destination state through a minimum cost path. We use an optimal control problem formulation, under assumptions that parallel those for finite-node shortest path problems, i.e., there exists a path to the destination starting from every node, and all cycles have positive or at least nonnegative length. Our analysis makes use of the recently developed theory of abstract semicontractive dynamic programming models. We investigate questions of existence and uniqueness of solution of the optimality equation, existence of optimal paths, and the validity of various algorithms patterned after the classical methods of value and policy iteration. Our analysis applies to classical shortest path problems and their countable node extensions, as well as optimal control problems where the objective is to control the state of a continuous-space system towards a target state.

† Supported in part by the Air Force Grant FA9550-10-1-0412. Helpful comments by Huizhen (Janey) Yu are gratefully acknowledged.

‡ Dimitri Bertsekas is with the Dept. of Electr. Engineering and Comp. Science, and the Laboratory for Information and Decision Systems, M.I.T., Cambridge, Mass., 02139.

1. INTRODUCTION

The classical deterministic shortest path problem is to reach a special destination node with a minimum length path from every one of the finite number of nodes in a given directed graph. This is a fundamental problem that has an enormous range of applications and has been studied extensively (see e.g., the surveys [Dre69], [GaP88], and many textbooks, including [Roc84], [AMO89], [Ber98], [Ber05]). At any node x, we may deterministically select a successor node y from a given set of possible successors, defined by the arcs (x, y) of the graph that are outgoing from x. While the problem is traditionally viewed as a combinatorial or network flow problem, it is also possible to consider it an optimal control problem. This viewpoint is particularly useful in stochastic extensions of the problem, where a formulation as a finite-state Markovian decision or optimal control problem is often most appropriate, with the role of states played by the nodes of the graph.

In this paper we consider shortest path problems with a possibly infinite number of states, using an optimal control point of view. We consider both a deterministic shortest path problem (DSP for short), and its stochastic shortest path counterpart (SSP for short). For SSP we assume that given any state and any control that can be applied at that state, the number of possible successor states is countable, so measurability issues do not arise within our framework. Based on the approach of this paper, it will be seen that DSP and SSP have similar structure, and for this reason, with appropriate adjustments in the definitions and assumptions, it will turn out that the extension of our results from DSP to SSP is straightforward. We will thus first formulate and analyze DSP, which is the simpler of the two problems.

Our starting point for DSP is a graph with a possibly (uncountable) infinite set of nodes $X \cup \{t\}$ and a set of directed arcs $\mathcal{A} \subset \big\{(x, u) \mid x, u \in X \cup \{t\}\big\}$. For each arc (x, u) we are given a real-valued length $g(x, u)$. Node t is a destination, which is absorbing and cost-free, in the sense that the only outgoing arc from t is (t, t) and $g(t, t) = 0$. A *path starting at node* $x_1 \in X$ is an arc sequence of the form

$$p = \big\{(x_1, x_2), (x_2, x_3), \dots\big\}.$$

The *length of* p is defined as

$$L_p = \sum_{k=1}^{\infty} g(x_k, x_{k+1})$$

if the series above is convergent (possibly to $+\infty$ or $-\infty$), and as

$$L_p = \limsup_{m \to \infty} \sum_{k=1}^{m} g(x_k, x_{k+1})$$

if this series is not convergent. A *terminating path starting at node* $x_1 \in X$ is a path of the form

$$p = \big\{(x_1, x_2), (x_2, x_3), \dots, (x_m, t), (t, t), \dots\big\},$$

which reaches the destination t after some finite number $m \geq 1$ of arcs. The length of such a path is

$$L_p = g(x_1, x_2) + g(x_2, x_3) + \cdots + g(x_m, t).$$

In DSP we want to find a path of minimum length, for each starting node. Note, however, that an optimal path need not be terminating: reaching the destination is not a requirement, but only a "retirement option," to be exercised only if desirable from the point of view of minimizing path length.

The preceding problem formulation is patterned after the classical shortest path problem, which is usually considered for the case of a graph with a finite number of nodes. It is also possible to consider DSP as a special case of an infinite horizon discrete-time deterministic optimal control problem, with nodes corresponding to states and arcs corresponding to controls. Here we have a discrete-time system, which is usually given by an equation of the form

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots,$$

where $x_k \in \Re^n$ is the state, $u_k \in \Re^m$ is the control, and $f : \Re^{n+m} \mapsto \Re^n$ is the (possibly nonlinear) system function (which defines the "graph" of the earlier shortest path formulation). State $x = 0$ is the destination, so $f(0, u_k) = 0$ for all u_k. The control u_k may be subject to some state-dependent constraint, $u_k \in U(x_k)$, and its choice defines the next "node" x_{k+1} in terms of the earlier graph formalism. We want to minimize the cost

$$\sum_{k=0}^{\infty} g(x_k, u_k)$$

where $g : \Re^{n+m} \mapsto [0, \infty)$ is a function, which takes value 0 for $x_k = 0$, i.e., $g(0, u_k) = 0$ for all u_k. To provide an incentive to reach the destination, $g(x, u)$ is often assumed positive for $x \neq 0$. In some problems, which are popular in the methodology of adaptive dynamic programming, the main objective is to construct a policy that stabilizes the system. Results within this context have focused among others on the convergence of value and policy iteration, and associated questions of adaptive control under some specialized assumptions (see e.g., [ALA08], [LeL12], [LiW13]). However, to the author's knowledge the DSP problem has not been investigated earlier at the level of generality of this paper, and most of the results given here have not been presented elsewhere.

We may also view DSP as a special case of SSP, which is a total cost infinite horizon Markovian decision problem with a substantial analytical and algorithmic methodology (see [Pal67], [Der70], [Pli78], [Whi82], [BeT89], [BeT91], [Put94], [HCP99], [HiW05], [JaC06], [Ber12], [BeY13], [YuB13a]). The difference of DSP from SSP is that at a given node x, in DSP the next node u is deterministically chosen among the set of possible next nodes $\big\{u \mid (x, u) \in \mathcal{A}\big\}$, but in SSP it is determined stochastically, according to a distribution over the set $\big\{u \mid (x, u) \in \mathcal{A}\big\}$. However, the strongest forms of the SSP methodology apply only to the finite-state version of the problem [for a treatment of the infinite state space case, including the associated measurability questions, and parallels the analysis of [BeT91] and has some common elements with the analysis of the present paper (mainly Prop. 2.1 and Section 5), see [JaC06] which assumes among others that $g(x, y)$ is bounded, and does not consider the cases covered by our Props. 2.2-2.4]. It turns out that thanks to the nature of analysis, the extension of our results for DSP to the SSP case is straightforward, as will be discussed in Section 5.

The analytical methodology of this paper is based on an embedding of DSP within a framework of dynamic programming (DP for short), which is abstract in the sense that it places heavy reliance on abstract properties of the associated DP mapping, such as monotonicity and contraction, and focuses the analysis on these properties. The benefit of abstraction is a unified treatment that applies to a broad class of problems. More specifically DSP (and SSP later) will be analyzed as special cases of the *semicontractive abstract DP model*, which has been formulated and analyzed recently in [Ber13], under a variety of assumptions. To this end we view as states the set of nondestination nodes X, and we view the set of successor nodes

$$U(x) = \big\{u \in X \cup \{t\} \mid (x, u) \in \mathcal{A}\big\}, \qquad x \in X,$$

as the controls that are admissible at state x. We consider policies μ, which are functions that assign at each state x a control $\mu(x) \in U(x)$, and we denote the set of all policies by $\mathcal{M}$. A policy may be identified with

a subgraph of the original graph, which has a unique outgoing arc for every node. For each nondestination node $x \in X$, this subgraph defines a unique path that starts at x, which is denoted by $p_\mu(x)$. The length (or cost) of this path is denoted by $J_\mu(x)$:

$$J_\mu(x) \equiv L_{p_\mu(x)}, \qquad x \in X.$$

The optimal path length starting from x is denoted by $J^*(x)$:

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

We will generally refer to J_μ and J^* as the *cost function of* μ and the *optimal cost function*, respectively.

Let us denote by $E(X)$ the set of functions $J : X \mapsto [-\infty, \infty]$, and let $\bar{J}$ be the identically 0 function in $E(X)$, i.e., $\bar{J}(x) \equiv 0$. In our analysis, functions in $E(X)$ will represent cost or length of some type of path that starts at x, for example $J_\mu(x)$. For each policy μ, we introduce the mapping $T_\mu : E(X) \mapsto E(X)$, defined by

$$(T_\mu J)(x) = \begin{cases} g\big(x, \mu(x)\big) + J\big(\mu(x)\big) & \text{if } \mu(x) \neq t, \\ g(x,t) & \text{if } \mu(x) = t, \end{cases} \qquad x \in X.$$

We denote by T_μ^m the m-fold composition of the mapping T_μ with itself. It can be seen from the definitions that $T_\mu^m(\bar{J})(x)$ is the sum of the lengths of the first m arcs in the path $p_\mu(x)$. Thus J_μ can equivalently be defined as

$$J_\mu(x) = \limsup_{m \to \infty} T_\mu^m(\bar{J})(x), \qquad x \in X.$$

By taking upper limit of both sides of the relation

$$T_\mu^m(\bar{J})(x) = \begin{cases} g\big(x, \mu(x)\big) + (T_\mu^{m-1}\bar{J})\big(\mu(x)\big) & \text{if } \mu(x) \neq t, \\ g(x,t) & \text{if } \mu(x) = t, \end{cases} \qquad x \in X,$$

as $m \to \infty$, we see that J_μ is a fixed point of the mapping T_μ:

$$J_\mu = T_\mu J_\mu, \qquad \forall\ \mu \in \mathcal{M}.$$

A policy μ is called *proper* if the corresponding paths $p_\mu(x)$ are terminating for all x. Similar terminology is standard in SSP (cf. [Pal67], [BeT91]). If μ is not proper, it is called *improper*. It can be seen that for a proper policy μ, the function J_μ is real-valued and is also the unique fixed point of the mapping T_μ.† For an improper policy μ, the function J_μ may or may not be real-valued, and may or may not be the unique fixed point of T_μ.

Let us now define the mapping $T : E(X) \mapsto E(X)$, by

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

† To show that J_μ is the unique fixed point of T_μ when μ is proper, let $\tilde{J}_\mu$ be any fixed point of T_μ. Then $\tilde{J}_\mu = T_\mu^m \tilde{J}_\mu$ for all $m \geq 1$, so for every $x \in X$ and its corresponding terminating path $p_\mu(x)$, we have

$$\tilde{J}_\mu(x) = L_{p_\mu(x)} = J_\mu(x).$$

In the theory of shortest path problems and DP, J^* is a fixed point of T, and the equation $J^* = TJ^*$ is commonly referred to as *Bellman's equation* or *optimality equation*. Consistently with this theory, the major issues that we will discuss, under a variety of assumptions, revolve around the properties and the computation of fixed points of T. More specifically, we will address the following questions:

(a) Is J^* a fixed point of T, and if so is it the unique fixed point within some given subset of $E(X)$?

(b) If a policy μ^* satisfies $T_{\mu^*}J^* = TJ^*$, is μ^* optimal, and reversely?

(c) Under what conditions on the starting function J do we have $T^kJ \to J^*$ for the value iteration method (or Bellman-Ford iteration in shortest path terminology)?

(d) What are valid versions of the policy iteration method of infinite horizon DP within our context?

In the next section we will introduce the assumptions under which we will investigate the preceding questions, we will give a summary of our results, and we will illustrate their validity and limitations through examples and counterexamples. In Section 3 we will make the connection with the semicontractive model methodology and prove our results. In Section 4 we will discuss algorithms patterned after the value and policy iteration methods. Finally in Section 5 we will consider SSP, and we will extend the results of the preceding sections from deterministic to stochastic problems.

Our assumptions involve various combinations of properties of the problem, principal among which are:

(1) Whether improper policies cannot be optimal because they yield infinite cost starting from some initial states.

(2) Whether $g(x, u) \geq 0$ for all $(x, u) \in \mathcal{A}$.

(3) Whether there exists an optimal proper policy.

(4) Whether J^* is bounded above (in all of our assumptions J^* is bounded below).

We consider a variety of combinations of the preceding conditions, because it turns out that the structure of our problem is quite delicate, and seemingly small variations in the assumptions may have major effect on the analysis and associated results. This can be illustrated with very simple examples, such as a two-node shortest path problem where the analysis is radically affected by the presence or absence of zero-length cycles and by the optimality of paths involving such cycles (see our subsequent Example 2.1 and Fig. 2.1).

While we have introduced our problem with terminology mostly used in the finite-node DSP literature, we will be making connections with the DP literature, where the terms "costs," "states," and "controls" are used in place of "lengths," "nodes," and "successor nodes," respectively, so we will resort to some of these terms to enhance clarity of presentation. This will also be convenient when we discuss SSP, which has traditionally been viewed as a special case of a Markovian decision problem. Regarding notation, we will operate within the set of extended-real numbers $[-\infty, \infty]$, with the usual rules for arithmetic, including the rule $\infty - \infty = \infty$. The infimum over the empty set is by definition ∞. All limits of sequences, infima, equalities, and inequalities involving functions in $E(X)$ are to be interpreted in a pointwise sense.

2. ASSUMPTIONS AND A SUMMARY OF RESULTS

One of our aims is to show that DSP has a delicate structure whereby seemingly minor variations in the assumptions may result in significant changes in the corresponding results. To this end, we will introduce five alternative sets of assumptions regarding our DSP problem, and in Section 2.1 we will summarize our results and highlight their differences. In Section 2.2 we will illustrate these results with examples.

We denote by $R(X)$ the set of real-valued functions $J : X \mapsto (-\infty, \infty)$, by $R^+(X)$ the set of nonnegative functions in $R(X)$,

$$R^+(X) = \big\{J \in R(X) \mid J \geq 0\big\},$$

and by $B(X)$ the subspace of all functions $J \in R(X)$ that are bounded,

$$B(X) = \left\{J \in R(X) \;\middle|\; -\infty < \inf_{x\in X} J(x) \leq \sup_{x\in X} J(x) < \infty\right\}.$$

We also denote by $B_b(X)$ the subset of all $J \in R(X)$ that are bounded below,

$$B_b(X) = \left\{J \in R(X) \;\middle|\; -\infty < \inf_{x\in X} J(x)\right\}.$$

If X is finite and has n elements, $R(X)$, $B(X)$, and $B_b(X)$ can all be identified with the Euclidean space $\Re^n$. As a result, when X is finite, some of the distinctions between results that relate to DSP and SSP and the spaces $R(X)$, $B(X)$, and $B_b(X)$ disappear. We denote by $\hat{J}$ the function given by

$$\hat{J}(x) = \inf_{\mu:\text{proper}} J_\mu(x), \qquad x \in X.$$

Note that if no proper policy exists, the infimum above is taken over the empty set and $\hat{J}(x) = \infty$.

Assumption 2.1: (Improper Policies Have Infinite Cost)

(a) There exists at least one proper policy and we have $\hat{J} \in B(X)$.

(b) For each improper policy μ, there is at least one $x \in X$ such that $J_\mu(x) = \infty$.

(c) The set X is a metric space, and for each $x \in X$, function $J \in B_b(X)$, and scalar λ, the set

$$\big\{\mu(x) \mid (T_\mu J)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact.

Part (a) of the preceding assumption is consistent with a standard condition in finite-node shortest path methodology, whereby the destination is assumed to be reachable with a path from every other node. In the control literature, the existence of a proper policy is alternately called a *controllability assumption*. Part (b) implies that improper policies cannot be optimal. It is satisfied if the arc lengths are strictly positive and bounded away from 0, and is patterned after the standard assumption whereby all cycles have positive

length (a similar assumption is also common in SSP; see [BeT91]). The compactness condition of part (c) is satisfied in particular if $U(x)$ is a finite set for all x.

The condition $\hat{J} \in B(X)$ in Assumption 2.1(a) can be restrictive in problems with an infinite state space. A variant of the preceding assumption, which assumes nonnegativity of $g(x,u)$ in place of $\hat{J} \in B(X)$ is the following.

Assumption 2.2: (Improper Policies Have Infinite Cost - Nonnegative Arc Lengths)

(a) There exists at least one proper policy and we have $g(x,u) \geq 0$ for all $(x,u) \in \mathcal{A}$.

(b) For each improper policy μ, there is at least one $x \in X$ such that $J_\mu(x) = \infty$.

(c) The set X is a metric space, and for each $x \in X$, function $J \in R^+(X)$, and scalar λ, the set

$$\big\{\mu(x) \mid (T_\mu J)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact.

Part (b) of Assumptions 2.1 and 2.2 fails in problems where a zero length cycle may be present. It also fails in other situations, where it is optimal to approach the destination asymptotically rather than reach it in a finite number of transitions. As an example, consider the linear-quadratic optimal control problem, with no control constraints, involving the linear system

$$x_{k+1} = Ax_k + Bu_k, \qquad k = 0, 1, \ldots,$$

and the quadratic cost function

$$\sum_{k=0}^{\infty} (x_k' Q x_k + u_k' R u_k).$$

Here x_k and u_k are the state and the control at time k, taking values in finite-dimensional Euclidean spaces, and 0 plays the role of the destination t. The matrices A and B have appropriate dimensions, and Q and R are symmetric, positive semidefinite and positive definite, respectively. For this problem, under a standard controllability assumption (which guarantees the existence of a proper policy and the finiteness of the optimal cost function), there is a unique optimal policy that is improper because it turns out that it is optimal to asymptotically approach the origin, but not to reach it in a finite number of time steps (see e.g., [Ber05], Ch. 4). In this problem J^* and $\hat{J}$ are actually equal to each other and equal to a positive definite quadratic function, which shows that the infimum of J_μ over proper policies μ is not attained and that Assumption 2.1(a) is violated.

The next assumption is intended to deal with situations such as the linear-quadratic problem above, where Assumption 2.1(a),(b) or Assumption 2.2(b) cannot be verified, as well as with problems where the compactness condition of Assumption 2.2(c) does not hold.

Assumption 2.3: (Nonnegative Arc Lengths) There holds $g(x,u) \geq 0$ for all $(x,u) \in \mathcal{A}$.

The preceding assumption does not require that improper policies have infinite cost starting from some state. There may exist an optimal improper policy, which may be unique (as in the linear-quadratic example above), or may coexist with an optimal proper policy. Some of the results under this assumption are obtained by specializing generic results for infinite horizon DP problems with nonnegative costs per stage, i.e., negative (reward) DP (see e.g., [Str66], [Put94], [Ber12]). We will show that these results are strengthened significantly when in addition to the nonnegativity condition on the arc lengths, an optimal proper policy exists, which is guaranteed in particular under some form of controllability assumption.

The next assumption bears similarity with the preceding ones. In contrast with Assumptions 2.1 and 2.2, it assumes the existence of an optimal proper policy, and in contrast with Assumptions 2.2 and 2.3, there is no requirement that the arc lengths are nonnegative.

Assumption 2.4: (An Optimal Proper Policy Exists and the Optimal Cost Function is Bounded Below)

(a) There exists an optimal proper policy and we have $J^* \in B(X)$.

(b) The set X is a metric space, and for each $x \in X$ and function $J \in B_b(X)$, the set

$$\big\{\mu(x) \mid (T_\mu J)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact.

The following assumption is similar but weaker than the preceding one. It allows for an improper policy to be optimal while all proper policies are strictly suboptimal. This occurs for example in the aforementioned linear-quadratic problem; see also the subsequent two-node shortest path Example 2.1 in Section 2.2.

Assumption 2.5: (A Proper Policy Exists and the Optimal Cost Function is Bounded Below)

(a) There exists at least one proper policy, and we have $\hat{J} \in B(X)$ and $J^* \in B(X)$.

(b) The set X is a metric space, and for each $x \in X$ and function $J \in B_b(X)$, the set

$$\big\{\mu(x) \mid (T_\mu J)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact.

Problems where J^* is unbounded below are not covered by the preceding assumptions, and may be an interesting subject for investigation. This issue has also been raised in the paper [JaC06]. Problems where $J^*(x) = -\infty$ for some $x \in X$ are also not covered. Such problems typically involve improper policies generating negative costs that accumulate to $-\infty$, possibly (but not necessarily) by reaching a cycle of

negative length. However, there are problems where $J^*(x) = -\infty$ for some x even if all policies are proper; consider for example the case $X = \{s, 1, 2, \ldots\}$, where s is an "origin" state, and for each $x = 1, 2, \ldots$, there is a single arc (s, x) with $g(s, x) = 0$, and a single arc (x, t) with $g(x, t) = -x$. A well-known somewhat pathological finite-state SSP problem where all policies are proper (as defined later in Section 5), and $J^*(x) = -\infty$ for all x is the blackmailer problem (see e.g., [Ber12], Section 3.2).

2.1. Analytical Results

We will now state our results relating to the properties of the optimal cost function J^* and the existence of optimal proper policies.

Proposition 2.1: Let Assumption 2.1 hold. Then:

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

(b) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal proper policy.

(c) We have $T^k J \to J^*$ for all $J \in B_b(X)$.

(d) For any $J \in B_b(X)$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

Proposition 2.2: Let Assumption 2.2 hold. Then:

(a) The function J^* is the unique fixed point of T within $R^+(X)$.

(b) A policy μ is optimal if and only if $T_\mu J^* = TJ^*$. Moreover, there exists an optimal proper policy.

(c) We have $T^k J \to J^*$ for all $J \in R^+(X)$.

(d) For any $J \in R^+(X)$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

Proposition 2.3: Let Assumption 2.3 hold. Then:

(a) The function J^* is the unique fixed point of T within the set $\big\{J \in E(X) \mid J^* \geq J \geq 0\big\}$.

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

(c) If there exists an optimal proper policy, then J^* is the unique fixed point of T within $R^+(X)$.

(d) If there exists an optimal proper policy, X is a metric space, and for each $x \in X$, scalar λ, and integer k greater than some index $\bar{k}$, the set

$$\big\{\mu(x) \mid \big(T_\mu(T^k\bar{J})\big)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact, then we have $T^kJ \to J^*$ for every $J \in R^+(X)$.

Note the main difference between Props. 2.2 and 2.3(d). The latter requires existence of an optimal proper policy, while the former requires that there exists a proper policy, and that for each improper policy μ, there is at least one $x \in X$ such that $J_\mu(x) = \infty$ (which are often easier conditions to verify).

Proposition 2.4: Let Assumption 2.4 hold. Then:

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

(b) A proper policy μ that satisfies $T_\mu J^* = TJ^*$ is optimal. Conversely if μ is a proper optimal policy, it satisfies $T_\mu J^* = TJ^*$.

(c) We have $T^kJ \to J^*$ for every $J \in R(X)$ with $J \geq J^*$.

Note several differences between Prop. 2.3 and Prop. 2.4: T cannot have fixed points in the range $\{J \mid J^* \geq J \geq 0\}$ other than J^* according to the former, but it can according to the latter (if $J^* \geq 0$). Another subtle difference between Prop. 2.3(b) and Prop. 2.4(b) is that the optimality condition $T_\mu J^* = TJ^*$ cannot be satisfied by a nonoptimal improper policy μ according to the former, but it can according to the latter, thus leading to a breakdown in the policy iteration algorithm. There is also a further difference in the range of starting functions for which $T^kJ \to J^*$ [cf. Prop. 2.3(d) and Prop. 2.4(c)].

The proofs of the preceding proposition and the proposition that follows are based on the perturbation approach of Section 3.2.2 of [Ber13]. In this approach, we introduce a scalar $\delta \geq 0$ and a δ-perturbed version of DSP, whereby each arc length $g(x,u)$ with $x \in X$ is replaced by $g(x,u) + \delta$. As a result, for an improper policy and a state starting from which the policy accumulates finite cost without ever terminating, the finite cost is turned to ∞ when a positive perturbation δ is added. The shortest path problem obtained when $g(x,u)$ is replaced by $g(x,u) + \delta$, for all $x \in X$ and $(x,u) \in \mathcal{A}$, is referred to as the δ-DSP.

For any $\mu \in \mathcal{M}$ and $\delta > 0$, let us introduce the mappings $T_{\mu,\delta}$ and T_δ corresponding to the δ-DSP:

$$(T_{\mu,\delta}J)(x) = \begin{cases} g\big(x,\mu(x)\big) + \delta + J\big(\mu(x)\big) & \text{if } \mu(x) \neq t, \\ g(x,t) + \delta & \text{if } \mu(x) = t, \end{cases} \qquad x \in X,$$

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

Let us also denote by J^*_δ the optimal cost function of the δ-DSP. A key fact under Assumptions 2.4 and 2.5, is that J^*_δ is a fixed point of T_δ. Moreover, there exists a proper policy that is optimal for the δ-DSP. We postpone further discussion of these points to Section 3, where proofs of the propositions will be given.

Proposition 2.5: Let Assumption 2.5 hold. Then

$$\lim_{\delta \downarrow 0} J_\delta^* = \hat{J},$$

where J_δ^* is the optimal cost function of the δ-DSP.

The preceding proposition does not assert that $J^* = \hat{J}$. Furthermore, it does not guarantee existence of a proper policy μ^* that is optimal within the class of proper policies, i.e., a proper μ^* with $J_{\mu^*} = \hat{J}$, despite the fact that there exists an optimal proper policy for the δ-DSP problem for all $\delta > 0$ (this will be shown as part of the proof of Prop. 2.5). Among others, the proposition provides the basis for value and policy iteration-like algorithms for finding $\hat{J}$, which use a sequence $\delta_k \downarrow 0$ and the corresponding sequence of δ_k-DSP problems (see the discussion of Section 4).

An important difference between Props. 2.4 and 2.5 is that the latter does not assert existence of an optimal policy. For an example, let $X = \{1, 2, \ldots\}$ and for each $x \in X$, let the outgoing arcs be (x, t) and $\big(x, (x+1)\big)$ with lengths

$$g(x, t) = \frac{1}{x} - 1, \qquad g(x, x+1) = 0.$$

Then Prop. 2.5 applies, and we have $J^*(x) \equiv \hat{J}(x) \equiv \lim_{\delta \downarrow 0} J_\delta^*(x) \equiv -1$, but there does not exist an optimal policy, despite the fact that there is an optimal policy for the δ-DSP problem for all $\delta > 0$ (this can be verified using Prop. 2.1).

2.2. Applications, Examples, and Counterexamples

We will now discuss a few example problems that illustrate the nature of the preceding assumptions and the range of applicability of the corresponding propositions. The first example, which is notable for its simplicity, illustrates that all of the assumptions of Section 2.1 are relevant to common practical settings.

Example 2.1 (Finite-Node Shortest Path Problems)

Let us consider the classical finite-node shortest path problem where X is a finite set. To illustrate the nature of our analysis, it is sufficient to consider the single node case, i.e., $X = \{1\}$. Thus in this example there are only two arc lengths, which we denote by a and b (see Fig. 2.1):

$$g(1, 1) = a, \qquad g(1, t) = b,$$

and two policies: the policy $1 \to t$ which is proper and is denoted by μ, and the policy $1 \to 1$ which is improper and is denoted by $\overline{\mu}$.

The corresponding mappings T_μ, $T_{\overline{\mu}}$, and T are given by

$$(T_\mu J)(1) = b, \qquad (T_{\overline{\mu}} J)(1) = a + J(1), \qquad (TJ)(1) = \min\big\{b,\, a + J(1)\big\}.$$

We consider various cases, which correspond to the five Assumptions 2.1-2.5, and the five Props. 2.1-2.5. Note that because of the finiteness of X, the compactness conditions that appear in these assumptions are satisfied:

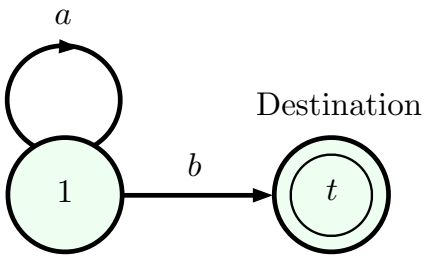

Figure 2.1. A deterministic shortest path problem with a single node 1 and a destination node t. At 1 there are two choices; a self-transition, which costs a, and a transition to t, which costs b.

(a) $a > 0$: Here Assumption 2.1 holds, and Prop. 2.1 applies (as well as Assumption 2.2 and Prop. 2.2, if $b \geq 0$). The optimal cost, $J^* = b$, is the unique fixed point of T, and the remaining results of Prop. 2.1 are seen to hold. This case corresponds to the general finite-node shortest path problem under the classical favorable assumptions, whereby all nodes are connected to the destination and all cycles have strictly positive length.

(b) $a = 0$ *and* $b = 0$: Here Assumption 2.3 holds, and since there exists an optimal proper policy, all parts of Props. 2.3 apply. The set of fixed points of T is $\{J \mid J \leq 0\}$, so $J^* = 0$ is the unique fixed point within $\{J \mid J \geq 0\}$ [cf. Prop. 2.3(c)]. Moreover, both policies are optimal and satisfy the optimality condition of Prop. 2.3(b), while value iteration converges to J^* in a single step, provided we start from $J \geq 0$.

(c) $a = 0$ *and* $b < 0$: Here Assumption 2.4 holds and Prop. 2.4 applies. The set of fixed points of T is $\{J \mid J \leq b\}$, so $J^* = b$ is the unique fixed point within $\{J \mid J \geq J^*\}$. The proper policy is optimal, yet the optimality condition $T_\mu J^* = TJ^*$ is satisfied by the suboptimal improper policy as well [cf. Prop. 2.4(b)].

(d) $a = 0$ *and* $b > 0$: Here Assumption 2.5 holds and Prop. 2.5 applies. Indeed, we have $J^* = 0 < b = \hat{J}$, while $J^*_\delta = b + \delta$, so that $\lim_{\delta \downarrow 0} J^*_\delta = \hat{J}$.

When $a < 0$, we have $J^*(1) = -\infty$ and the improper policy $\overline{\mu}$ is optimal, but this case is not covered by Props. 2.1-2.5.

Example 2.2 (Linear-Quadratic Problems)

Consider the linear-quadratic optimal control problem, with no control constraints, involving the linear system

$$x_{k+1} = Ax_k + Bu_k, \qquad k = 0, 1, \ldots,$$

and the quadratic cost function

$$\sum_{k=0}^{\infty} (x_k' Q x_k + u_k' R u_k).$$

Here, as in our earlier discussion, x_k and u_k are the state and the control at time k, taking values in $\Re^n$ and $\Re^m$, respectively, and A and B are $n \times n$ and $n \times m$ matrices, respectively. The destination is the origin $x = 0$. The matrices Q and R are symmetric, positive semidefinite and positive definite, respectively. We assume the standard controllability condition, under which the origin can be reached from every initial state in n steps or less. This is equivalent to the matrix $[B \ AB \ \cdots A^{n-1}B]$ having full rank, and guarantees the existence of a proper policy. However, it is well-known that the unique optimal policy is improper, although it can be

verified that $\hat{J} = J^*$. Propositions 2.3(a),(b) and 2.5 apply to this problem, but Props. 2.1, 2.2, and 2.4 do not apply (there is a related analysis based on semicontractive DP models, which also applies; see Section 4.4.3 of [Ber13]).

Example 2.3 (Search Problems)

Consider a classical search problem where the objective is to move within a (possibly uncountably infinite) set of states X, searching for a state to stop while minimizing the total cost up to stopping. We introduce an additional stopping/destination state t, and we formulate a DSP problem with two controls at each $x \in X$: *stop*, in which case we move to the destination t at cost $s(x)$, and *continue*, in which case we move to a state $f(x) \in X$ at cost $g(x)$, where s, f, and g are given functions.

Here there exists a proper policy (e.g., the one that stops at every state). Depending on the nature of s, f, and g, some of the propositions of the preceding section may apply. For example in the case where $s(x) \geq 0$ and $g(x) \geq 0$ for all $x \in X$, Prop. 2.3 applies. In the case where $s(x) \leq 0$ for all $x \in X$, and $g(x) \equiv 0$, we have an instance of positive (reward) DP (see e.g., [Bla65], [BeS78], [Put94], [Ber12]). Then assuming that s is bounded below, J^* and $\hat{J}$ are also real-valued and bounded below. In this case, Prop. 2.5 applies, and if in addition we assume that there exists an optimal proper policy, Prop. 2.4 also applies. However, Props. 2.1 and 2.2 do not apply because the cost function of each improper policy is real-valued. Furthermore, an optimal policy need not exist, as shown by the example given at the end of Section 2.1: $X = \{1, 2, \ldots\}$, $g(x) \equiv 0$, $f(x) = x + 1$, and $s(x) = 1/x - 1$, in which case $J^*(x) \equiv -1$ but there is no optimal policy.

Example 2.4 (A Counterexample)

We will show by example that Prop. 2.1 does not hold when $B_b(X)$ is replaced by $R(X)$, and in fact in this example T has a unique fixed point within $B_b(X)$ but an infinite number of fixed points within $R(X)$ (necessarily, each of these fixed points is a function that is unbounded below).

Assume that $X = \{1, 2, \ldots\}$, and that at each $x \in X$ there are two choices: *stop*, which leads to the destination t at cost

$$g(x, t) = \frac{1}{x} - 1,$$

and *continue*, which leads to $x + 1$ at cost

$$g(x, x + 1) = \frac{1}{x + 1}.$$

Thus a policy that does not stop after a certain state $\bar{x}$ accumulates infinite cost starting at every $x > \bar{x}$.

It can be seen that a policy μ is proper if and only if it chooses to stop at an infinite number of states $x \in X$. Moreover, since the stopping cost lies in $[-1, 0]$, we have $J_\mu \in B_b(X)$, while $\hat{J}(x) \in [-1, 0]$ for all $x \in X$, so $\hat{J} \in B(X)$. Furthermore, as noted earlier, for every improper μ, we have $J_\mu(x) = \infty$ for all x greater than the largest state at which μ stops. Thus all the conditions of Assumption 2.1 hold, implying the results of Prop. 2.1.

On the other hand the mapping T, which is given by

$$(TJ)(x) = \min\left\{\frac{1}{x} - 1,\ \frac{1}{x + 1} + J(x + 1)\right\},$$

has additional fixed points within $R(X)$. More specifically, the function $\tilde{J} \in R(X)$ given by

$$\tilde{J}(x) = -\sum_{\ell=1}^{x} \frac{1}{\ell}, \qquad x \in X,$$

is a fixed point of T, as can be seen from the identity

$$-\sum_{\ell=1}^{x}\frac{1}{\ell}=\min\left\{\frac{1}{x}-1,\ \frac{1}{x+1}-\sum_{\ell=1}^{x+1}\frac{1}{\ell}\right\},\qquad x\in X.$$

In fact T has an infinite number of fixed points, since by adding a negative constant to $\tilde{J}(x)$ we obtain another fixed point. Note that all of these fixed points are unbounded below.

3. SEMICONTRACTIVE MODELS AND SHORTEST PATH PROBLEMS

The proofs of the propositions of the preceding section are obtained by viewing the DSP problem within the framework of semicontractive problems introduced in [Ber13], and by applying the corresponding theory. We first provide a brief review of this theory, using a notation that corresponds to the one used already for DSP.

3.1. Semicontractive Models

We have a set of states X, a set of controls U, and a control constraint set $U(x)\subset U$ for each $x\in X$. A policy is a mapping $\mu: X\mapsto U$ with $\mu(x)\in U(x)$ for all $x\in X$, and the set of all policies is denoted by $\mathcal{M}$. For each policy μ, we are given a mapping $T_\mu: E(X)\mapsto E(X)$ that is monotone in the sense that for any two $J,J'\in E(X)$,

$$J\leq J'\qquad\Rightarrow\qquad T_\mu J\leq T_\mu J'.$$

The cost function of μ is defined as

$$J_\mu=\limsup_{m\to\infty}T_\mu^m\bar{J},$$

where $\bar{J}$ is some given function in $E(X)$. The objective is to find $J^*=\inf_{\mu\in\mathcal{M}}J_\mu$ and a policy attaining the infimum if one exists.

In contractive models, the mappings T_μ are assumed to be contractions, with respect to a common weighted sup-norm and with a common contraction modulus, in the subspace of functions in $E(X)$ that are bounded with respect to the weighted sup-norm. These models have a strong analytical and algorithmic theory, which dates to [Den67]; a recent extensive treatment is given in Ch. 2 of [Ber13]. Semicontractive models are a class of models where only some policies have a contraction-like property. This property is captured by the notion of S-regularity of a policy. More specifically, given a set of functions $S\subset E(X)$, we say that a policy μ is *S-regular* if:

(a) $J_\mu\in S$ and $J_\mu=T_\mu J_\mu$.

(b) $\lim_{k\to\infty}T_\mu^k J=J_\mu$ for all $J\in S$.

A policy that is not S-regular is called *S-irregular*. Roughly, μ is S-regular if J_μ is an asymptotically stable equilibrium point of T_μ within S.

There are several different choices of S, which may be useful depending on the context, such as for example $R(X)$, $R^+(X)$, $B(X)$, $B_b(X)$, $\big\{J\in E(X)\mid J\geq\bar{J}\big\}$, and others. There are also several sets of assumptions and corresponding results, which are given in [Ber13] and will be used to prove the propositions

of this paper. Generally, the analysis of semicontractive models revolves around the fixed point properties of the mapping T, and optimality conditions for policies. These are also some of the main issues in infinite horizon DP problems, with cost per stage that is either undiscounted, or is discounted but is also unbounded. Other questions of interest relate to computational methods that are motivated by the classical methods of value iteration, policy iteration, and linear programming. Note that while semicontractive models were motivated by shortest path problems of various kinds, they do not include an explicit "termination state," and their applicability ranges considerably beyond the shortest path context; for example they apply to stochastic linear quadratic optimal control problems (see [Ber13], Section 4.3.3).

We give below an assumption relating to semicontractive models, which is Assumption 3.2.1 of [Ber13].

Assumption 3.1: Consider the preceding semicontractive model with a set $S \subset R(X)$ such that the following hold:

(a) S contains $\bar{J}$, and has the property that if J_1, J_2 are two functions in S, then S contains all functions J with $J_1 \leq J \leq J_2$.

(b) The function $\hat{J}$ given by

$$\hat{J}(x) = \inf_{\mu:\, S\text{-regular}} J_\mu(x), \qquad x \in X,$$

belongs to S.

(c) For each S-irregular policy μ and each $J \in S$, there is at least one state $x \in X$ such that

$$\limsup_{k\to\infty} (T_\mu^k J)(x) = \infty.$$

(d) The control set U is a metric space, and the set

$$\big\{\mu(x) \mid (T_\mu J)(x) \leq \lambda\big\}$$

is compact for every $J \in S$, $x \in X$, and $\lambda \in \Re$.

(e) For each sequence $\{J_m\} \subset S$ with $J_m \uparrow J$ for some $J \in S$ we have

$$\lim_{m\to\infty} (T_\mu J_m)(x) = (T_\mu J)(x), \qquad \forall\ x \in X,\ \mu \in \mathcal{M}.$$

(f) For each function $J \in S$, there exists a function $J' \in S$ such that $J' \leq J$ and $J' \leq TJ'$.

The following two propositions are given in [Ber13] as Prop. 3.2.1 and Lemma 3.2.4, respectively. The proof of Prop. 2.1 will be based on these two propositions. The second proposition is useful in cases where only some of the conditions of Assumption 3.1 are satisfied. This happens for example when $\hat{J}$, the infimum of J_μ over all S-regular policies μ, is different than J^*.

Proposition 3.1: Let Assumption 3.1 hold. Then:

(a) The optimal cost function J^* is the unique fixed point of T within the set S.

(b) We have $T^k J \to J^*$ for all $J \in S$. Moreover, there exists an optimal S-regular policy.

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

(d) For any $J \in S$, if $J \leq TJ$ we have $J \leq J^*$, and if $J \geq TJ$ we have $J \geq J^*$.

Proposition 3.2: Let Assumption 3.1(b),(c),(d) hold. Then:

(a) The function $\hat{J}$ of Assumption 3.1(b) is the unique fixed point of T within S.

(b) Every policy μ satisfying $T_\mu \hat{J} = T\hat{J}$ is optimal within the set of S-regular policies, i.e., μ is S-regular and $J_\mu = \hat{J}$. Moreover, there exists at least one such policy.

There are also some other results from [Ber13], which will be used to prove Props. 2.3-2.5. More specifically, the following proposition is adapted from Prop. 4.4.1 in [Ber13], and will be used in conjunction with Assumption 2.3 and Prop. 2.3.

Proposition 3.3: Consider the preceding semicontractive model with a set $S \subset E(X)$ such that the following hold:

(1) We have

$$-\infty < \bar{J}(x) \leq (T\bar{J})(x), \qquad \forall\, x \in X.$$

(2) For each sequence $\{J_m\} \subset E(X)$ with $J_m \uparrow J$ and $\bar{J} \leq J_m$ for all $m \geq 0$, we have

$$\lim_{m\to\infty} (T_\mu J_m)(x) = (T_\mu J)(x), \qquad \forall\, x \in X,\ \mu \in \mathcal{M}.$$

(3) There exists a scalar $\alpha \in (0, \infty)$ such that for all scalars $r \in (0, \infty)$ and functions $J \in E(X)$ with $\bar{J} \leq J$, we have

$$\big(T_\mu(J + r\,e)\big)(x) \leq (T_\mu J)(x) + \alpha\, r, \qquad \forall\, x \in X,\ \mu \in \mathcal{M},$$

where e is the unit function [$e(x) \equiv 1$].

(4) There exists an optimal S-regular policy, where S is a set satisfying

$$S \subset \big\{J \in E(X) \mid J \geq \bar{J}\big\}, \qquad \bar{J} \in S. \tag{3.1}$$

Then:

(a) The optimal cost function J^* is the unique fixed point of T within S.

(b) We have $T^k J \to J^*$ for every $J \in S$ with $J \geq J^*$.

(c) If for each $x \in X$, scalar λ, and integer k greater than some index $\bar{k}$, the set

$$\big\{\mu(x) \mid \big(T_\mu(T^k \bar{J})\big)(x) \leq \lambda,\ \mu \in \mathcal{M}\big\}$$

is compact, then we have $T^k J \to J^*$ for every $J \in S$.

The following two propositions will be used in conjunction with Props. 2.4 and 2.5. To state these propositions, we introduce a δ-perturbed version of the preceding abstract DP model, generalizing the δ-DSP problem introduced in Section 2.1. More specifically, for each $\delta \geq 0$ and policy μ, we consider the mappings $T_{\mu,\delta}$ and T_δ given by

$$(T_{\mu,\delta} J)(x) = (T_\mu J)(x) + \delta, \quad x \in X, \qquad T_\delta J = \inf_{\mu \in \mathcal{M}} T_{\mu,\delta} J.$$

We define the corresponding cost functions of policies $\mu \in \mathcal{M}$, and optimal cost function J^*_δ by

$$J_{\mu,\delta}(x) = \limsup_{k \to \infty} T^k_{\mu,\delta} \bar{J}, \qquad J^*_\delta = \inf_{\pi \in \Pi} J_{\pi,\delta}.$$

We refer to the problem associated with the mappings $T_{\mu,\delta}$ as the *δ-perturbed problem*. The following proposition is given as Prop. 3.2.2 in [Ber13].

Proposition 3.4: Given a set $S \subset E(X)$, assume that:

(1) For every $\delta > 0$, there exists an optimal S-regular policy for the δ-perturbed problem.

(2) If μ is an S-regular policy, we have

$$J_{\mu,\delta} \leq J_\mu + w_\mu(\delta), \qquad \forall\ \delta > 0,$$

where w_μ is a function such that $\lim_{\delta \downarrow 0} w_\mu(\delta) = 0$.

Then

$$\lim_{\delta \downarrow 0} J^*_\delta = \inf_{\mu:\, S\text{-regular}} J_\mu,$$

where J^*_δ is the optimal cost function of the δ-perturbed problem.

A simple way to guarantee that $\lim_{\delta \downarrow 0} J_\delta^* = J^*$ is to assume that there exists an optimal S-regular policy for the unperturbed problem. This also guarantees Bellman's equation $J^* = TJ^*$, under some additional conditions that are collected in the following assumption, given as Assumption 3.2.2 in [Ber13].

Assumption 3.2: We are given a set $S \subset E(X)$ such that the following hold:

(a) There exists an S-regular policy μ^* that is optimal, i.e., $J_{\mu^*} = J^*$, and satisfies

$$J_{\mu^*,\delta} \leq J_{\mu^*} + w(\delta), \qquad \forall\ \delta > 0,$$

where w is a function such that $\lim_{\delta \downarrow 0} w(\delta) = 0$.

(b) The optimal cost function J_δ^* of the δ-perturbed problem belongs to S and satisfies the Bellman equation $J_\delta^* = T_\delta J_\delta^*$ for each $\delta > 0$.

(c) For each sequence $\{J_m\} \subset S$ with $J_m \downarrow J$ for some $J \geq J^*$, we have

$$T_\mu J_m \downarrow T_\mu J, \qquad \forall\ \mu \in \mathcal{M}.$$

Under the preceding assumption we can show the following proposition, given as Prop. 3.2.3 in [Ber13].

Proposition 3.5: Let Assumption 3.2 hold. Then:

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

(b) We have $T^k J \to J^*$ for every $J \in S$ with $J \geq J^*$.

(c) An S-regular policy μ that satisfies $T_\mu J^* = TJ^*$ is optimal. Conversely if μ is an S-regular optimal policy, it satisfies $T_\mu J^* = TJ^*$.

3.2. Proofs of Propositions

In the context of this paper, it turns out that the Assumptions 2.1-2.5, and the corresponding Props. 2.1-2.5, can be viewed as special cases of the assumptions and propositions of the preceding section for the semicontractive model, with T_μ being the monotone mapping

$$(T_\mu J)(x) = \begin{cases} g\big(x, \mu(x)\big) + J\big(\mu(x)\big) & \text{if } \mu(x) \neq t, \\ g(x,t) & \text{if } \mu(x) = t, \end{cases} \qquad x \in X, \tag{3.2}$$

with $\bar{J}$ being the zero function $[\bar{J}(x) \equiv 0]$, and with S being a suitable set of functions. A key fact in this regard is the following characterization of the connection between the notions of S-regularity and properness.

Proposition 3.6: Consider the DSP problem, viewed as a special case of the abstract semicontractive model of Section 3.1 with T_μ given by Eq. (3.2), and $\bar{J}$ being the zero function. For any set $S \subset R(X)$ that contains a nonzero constant function, a policy μ is S-regular if and only if it is proper and $J_\mu \in S$.

Proof: Let μ be proper with $J_\mu \in S$. Then, independently of the choice of J within S, for each $x \in X$ and proper policy μ, the limit $\lim_{m\to\infty}(T_\mu^m J)(x)$ is the (finite) length of the terminating path that is generated by μ starting from x, i.e., it is equal to J_μ, implying that μ is S-regular.

Conversely let μ be improper. Taking J to be a nonzero constant function, $J(x) \equiv r$, we see that

$$\limsup_{m\to\infty}(T_\mu^m J)(x) = \limsup_{m\to\infty}(T_\mu^m \bar{J})(x) + r = J_\mu(x) + r$$

for all x starting from which the path generated by μ is not terminating. This shows that μ is not S-regular. **Q.E.D.**

We now show how the proofs of the five propositions of Section 2.2 follow from the five propositions of Section 3.1.

Proof of Prop. 2.1: We verify the conditions of Assumption 3.1 with $S = B_b(X)$. The result then will follow from Prop. 3.1. To this end we first note that by Prop. 3.6, μ is $B_b(X)$-regular if and only if μ is proper. We next observe that only parts (c) and (f) of Assumption 3.1 are not immediately implied by the special structure of the DSP problem, the fact $S = B_b(X)$, or corresponding parts of Assumption 2.1. To verify Assumption 3.1(c), we note that for every policy μ and function $J \in B_b(X)$, we have

$$(T_\mu^m \bar{J})(x) + \inf_{x'\in X} J(x') \le (T_\mu^m \bar{J})(x) + J(x_m) = (T_\mu^m J)(x), \tag{3.3}$$

where x_m is the state obtained after m transitions on the path generated by μ starting from x. By taking upper limit in this relation, it follows that

$$J_\mu(x) + \inf_{x'\in X} J(x') \le \limsup_{m\to\infty} (T_\mu^m J)(x). \tag{3.4}$$

Thus, for an improper μ, the assumption that $J_\mu(x) = \infty$ for some $x \in X$ [cf. Assumption 2.1(c)] is equivalent to the assumption that for every $J \in B_b(X)$, there exists $x \in X$ such that

$$\limsup_{m\to\infty} (T_\mu^m J)(x) = \infty, \tag{3.5}$$

[cf., Assumption 3.1(c)].

Also to verify that Assumption 3.1(f) is implied by Assumption 2.1, we note that by applying Prop. 3.2 with $S = B_b(X)$, we have that $\hat{J}$ is the unique fixed point of T within $B_b(X)$. Since $\hat{J} \in B(X)$ by Assumption 2.1(a), it follows that for each function $J \in B_b(X)$, there exists a sufficiently large scalar $r > 0$ such that the function J' given by

$$J' = \hat{J} - re, \qquad \forall\, x \in X, \tag{3.6}$$

where e is the unit function, $e(x) \equiv 1$, satisfies

$$J' = \hat{J} - re = T\hat{J} - re \leq T(\hat{J} - re) = TJ', \tag{3.7}$$

as well as $J' \leq J$. Thus Assumption 3.1(f) holds and Prop. 3.1 applies, with $S = B_b(X)$. **Q.E.D.**

The preceding proof has a few fine points that are worth pointing out. One such point is that the verification of Assumption 3.1(f) via Eqs. (3.6) and (3.7) would not be possible if we assumed that $\hat{J} \in B_b(X)$ instead of $\hat{J} \in B(X)$ as in Assumption 2.1(a).

Another fine point in the preceding proof is that the verification of Assumption 3.1(c) would not be possible if we used $S = R(X)$ instead of $S = B_b(X)$, because in the former case we could have $\inf_{x' \in X} J(x') = -\infty$ and Eq. (3.4) would not follow from Eq. (3.3). For an illustration, consider the problem of Example 2.4. There part (c) of Assumption 2.1, on which the proof of Prop. 2.1 rests, would be violated if $B_b(X)$ were replaced by $R(X)$, because there exist $J \in R(X)$ and improper μ for which we have

$$\limsup_{m \to \infty} (T_\mu^m J)(x) < \infty, \qquad \forall\ x \in X. \tag{3.8}$$

For an example, consider the improper policy μ that never stops, so that

$$(T_\mu^m J)(x) = \sum_{\ell=x+1}^{x+m} \frac{1}{\ell} + J(x+m), \qquad \forall\ x \in X,\ m \geq 1,$$

[cf. Eq. (3.3)]. Then by choosing J to be the function $\tilde{J} \in R(X)$ given by

$$\tilde{J}(x) = -\sum_{\ell=1}^{x} \frac{1}{\ell}, \qquad \forall\ x \in X,$$

we obtain

$$(T_\mu^m \tilde{J})(x) = -\sum_{\ell=1}^{x} \frac{1}{\ell}, \qquad \forall\ x \in X,\ m \geq 1,$$

thereby verifying Eq. (3.8).

Proof of Prop. 2.2: The proof is essentially identical to the one of Prop. 2.1. We verify the conditions of Assumption 3.1 with $S = R^+(X)$. The result then follows from Prop. 3.1. Since the cost function of a proper policy is real-valued and there exists a proper policy, it follows that $\hat{J} \in R^+(X)$. To verify parts (c) and (f) of Assumption 3.1, the argument based on Eqs. (3.3)-(3.5) goes through, and the requirement that for each function $J \in R^+(X)$, there exists a function $J' \in R^+(X)$ such that $J' \leq J$ and $J' \leq TJ'$ [cf. Eqs. (3.6)-(3.7)] is automatically satisfied by taking $J' = \bar{J}$. **Q.E.D.**

Proof of Prop. 2.3: Parts (a) and (b) follow from generic results on negative DP [Str66] (see also [Ber12], Props. 4.1.2, 4.1.3, 4.1.5, or [Ber13], Props. 4.3.3 and 4.3.9). Parts (c) and (d) follow as a special case of Prop. 3.3 with $S = R^+(X)$; the assumptions and conclusions of Prop. 2.3 match those of Prop. 3.3. **Q.E.D.**

We will obtain the proofs of Props. 2.4 and 2.5 through an intermediate result relating to the δ-DSP problem. The essence of these proofs is the following proposition.

Proposition 3.7: Let Assumption 2.5 hold. Then for every $\delta > 0$ there exists an optimal proper policy for the δ-DSP problem, and the corresponding optimal cost function J_δ^* is a fixed point of T_δ.

Proof: We note that we cannot apply Prop. 2.1 to the δ-DSP problem because, even though we assume that $\hat{J} \in B(X)$, we do not know whether $\hat{J}_\delta$, the optimal cost function over proper policies of the δ-DSP problem, belongs to $B(X)$. We thus use a different argument, claiming that under Assumption 2.5, Prop. 3.2 applies to the δ-DSP with $S = B_b(X)$. To this end, we note that all conditions of Prop. 3.2 are fulfilled, except that we need to repeat the argument used for the proof of Prop. 2.1, involving Eqs. (3.3)-(3.5), and using the assumption $J^* \in B(X)$. Proposition 3.2 shows that $\hat{J}_\delta$ is a fixed point of T_δ, and that there exists a proper policy with cost equal to $\hat{J}_\delta$.

We conclude the proof by showing that $\hat{J}_\delta = J_\delta^*$. Indeed, assume the contrary, i.e., that there exists $x \in X$ and an improper μ such that $\hat{J}_\delta(x) > J_{\mu,\delta}(x)$. Then the path that is generated using μ starting from x must be terminating (otherwise the extra cost of δ per transition would accumulate to ∞). Consider any proper policy μ' that agrees with μ on the terminating path starting from x. Then we have $\hat{J}_\delta(x) \leq J_{\mu',\delta}(x) = J_{\mu,\delta}(x)$, a contradiction. **Q.E.D.**

Proof of Prop. 2.4: Assumption 2.4 (which implies Assumption 2.5) together with Prop. 3.7 imply Assumption 3.2 with $S = \{J \in R(X) \mid J \geq J^*\}$, under which Prop. 3.5 applies. The latter proposition proves the desired results. **Q.E.D.**

Proof of Prop. 2.5: Assumption 2.5 and Prop. 3.7 imply the conditions of Prop. 3.4 with $S = \{J \in R(X) \mid J \geq J^*\}$. The latter proposition proves the desired results. **Q.E.D.**

3.3. Local Extensions of the Analysis

The statements and proofs of Props. 2.1-2.5 require that the set S of regularity is either $S = B_b(X)$ (in Props. 2.1, 2.4, and 2.5), or $S = R^+(X)$ (in Props. 2.2 and 2.3). It is also possible to use other sets S as long as the corresponding results from semicontractive abstract DP can be applied (cf. Props. 3.1-3.5).

The incentive for using alternative sets S stems from the essential character of S: it is the "domain of local attraction" of the mapping T_μ to J_μ, in order for μ to qualify as an S-regular policy. Thus if for an interesting class of policies μ, the mapping T_μ has a domain of local attraction, which is a strict subset S of $B_b(X)$ or $R^+(X)$ rather that $S = B_b(X)$ or $S = R^+(X)$, respectively, then it may be possible to prove a local version of the corresponding Props. 2.1-2.5. For this it is necessary of course to verify the corresponding assumptions of Props. 3.1-3.5.

The preceding observation is useful when we know a specially structured class of "interesting" policies that are S-regular for some special set S, but not for $S = B_b(X)$ or $S = R^+(X)$. An example of application of this idea is the linear-quadratic problem given in [Ber13], Section 4.4.3, which is not covered well by the results of the present paper because there is no optimal proper policy, as has been noted earlier. For a more

general view of that example, assume that $g(x,u) \geq 0$ for all (x,u), so that $\bar{J} \leq T\bar{J}$ and Prop. 3.3 applies. Suppose that we know that an optimal policy μ has cost function that belongs to a special subset S of $\{J \in E(X) \mid J \geq \bar{J}\}$ with $\bar{J} \in S$, and that μ is S-regular. Then by Prop. 3.3, the optimal cost function J^* is the unique fixed point of T within S. Moreover value iteration converges to J^* starting with any $J \in S$ under the compactness condition of Prop. 3.3(c). In the linear-quadratic problem of Section 4.4.3 of [Ber13], S consists of positive definite quadratic functions [a very "small" subset of $R^+(X)$], and the S-regular policies contain the class of linear feedback controllers that stabilize the underlying dynamic system. The unique optimal policy belongs to this class.

4. COMPUTATIONAL METHODS

We have already shown as part of Props. 2.1-2.4 that the value iteration algorithm (VI for short), which generates $T^k J$ for $k \geq 0$, converges to the optimal cost function J^* under various conditions on the starting function J. We can extend this convergence property to asynchronous versions of VI based on the monotonicity and fixed point properties of the mapping T. We refer to the discussions in Sections 2.6.1, 3.3.1, and 4.3.2 of [Ber13], which apply in their entirety when specialized to the DSP problem of this paper.

The development of policy iteration (PI for short) algorithms for the DSP problem is also straightforward given the connection with semicontractive models established in the preceding section. Briefly, under Assumption 2.1, based on the analysis of Section 3.3.2 of [Ber13], there are two types of PI algorithms. The first is a natural form of PI that generates proper policies exclusively. Let μ^0 be an initial proper policy (there exists one by assumption). At the typical iteration k, we have a proper policy μ^k, and we compute a policy μ^{k+1} such that $T_{\mu^{k+1}} J_{\mu^k} = T J_{\mu^k}$. Then

$$J_{\mu^k} = T_{\mu^k} J_{\mu^k} \geq T J_{\mu^k} = T_{\mu^{k+1}} J_{\mu^k} \geq \lim_{m \to \infty} T^k_{\mu^{k+1}} J_{\mu^k} = J_{\mu^{k+1}},$$

so μ^{k+1} is proper [in view of Assumption 2.1(b)], and has improved cost over μ^k. We can thus construct a sequence of proper policies $\{\mu^k\}$ and a corresponding nonincreasing sequence $\{J_{\mu^k}\}$. Under some additional mild conditions it is then possible to show that $J_{\mu^k} \downarrow J^*$.

Unfortunately, when there are improper policies, the preceding PI algorithm is somewhat limited, because an initial proper policy may not be known, and also because when asynchronous versions of the algorithm are implemented, it is difficult to guarantee that all the generated policies are proper. There is another PI algorithm, which has been developed in [BeY10], [BeY12], [YuB13a], [YuB13b], and is described in Section 3.3.2 of [Ber13]. This algorithm works in the presence of improper policies, and can operate in a distributed asynchronous environment. The specialization of this algorithm to DSP under Assumption 2.1 or Assumption 2.2 is straightforward.

Under Assumption 2.3, the optimistic PI ideas of Section 4.4.1 of [Ber13] apply, and their specialization is straightforward. Under Assumption 2.4, the classical version of PI may fail. This can be shown by an example where the optimality condition $T_\mu J^* = T J^*$ is satisfied by a nonoptimal improper policy. Thus, starting with an optimal proper policy, the next policy generated by the classical version of PI is the nonoptimal improper policy, with an oscillation between policies resulting [see case (a) of the example in Section 3.1.2 of [Ber13]]. On the other hand, alternative PI ideas may be used. For example, the perturbation-based PI ideas of Section 3.3.3 apply, while under special assumptions, other PI algorithms are possible (see the discussion in Section 4.3.3, Example 4.3.4, and Section 4.4.1 of [Ber13]).

5. STOCHASTIC SHORTEST PATH PROBLEMS

We will now formulate SSP as the stochastic version of DSP, where the transition to the next state is determined by a probability distribution that depends on the current state and the control applied at that state. More specifically, we have a set of states $X \cup \{t\}$, where t is an absorbing and cost-free destination, a set of controls U, a control constraint set $U(x) \subset U$, $x \in X$, and a cost $g(x,u)$ for each transition. The only difference from DSP is that when a control u is used at a state $x \in X$, a transition to a state that belongs to a countable set $Y(x,u) \subset X \cup \{t\}$ occurs according to a given probability distribution $p_{xy}(u)$. This change is reflected in the mapping $T_\mu : E(X) \mapsto E(X)$, which now takes the form

$$(T_\mu J)(x) = g\big(x,\mu(x)\big) + \sum_{y \in Y\left(x,\mu(x)\right),\, y \neq t} p_{xy}\big(\mu(x)\big) J(y), \qquad x \in X. \tag{5.1}$$

Here to resolve technical ambiguities that result when J can take both the values ∞ and $-\infty$, we follow the convention $\infty - \infty = \infty$ in defining the summation in the right-hand side (this is standard in DP; see e.g., [BeS78]). Similar to DSP, the cost function of a policy μ is given by

$$J_\mu(x) = \limsup_{m \to \infty} (T_\mu^m J)(x),$$

and can be interpreted as the expected length of the paths that are generated under μ starting from x. The optimal cost function is defined as

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

We will now adapt the definition of a proper policy to the stochastic context of SSP. For each $\mu \in \mathcal{M}$, we denote by $r_m(x,\mu)$ the probability that starting from x and using μ, the destination t will not have been reached after the first m transitions in the corresponding sequence $(x, x_1, x_2, \ldots)$:

$$r_m(x,\mu) = P(x_m \neq t \mid x_0 = x,\, \mu).$$

We say that μ is proper if $J_\mu \in B(X)$ and

$$\lim_{m \to \infty} r_m(x,\mu) E\big\{J(x_m) \mid x_0 = x,\, x_m \neq t,\, \mu\big\} = 0, \qquad \forall\, x \in X, J \in B(X). \tag{5.2}$$

This definition differs slightly from the definition of properness in [JaC06] (a proper policy is called "transient" in [JaC06]).

In comparing this definition with the corresponding definition of a proper policy for DSP, we see a difference: in DSP there is no requirement that $J_\mu \in B(X)$, which can be a severe restriction when the state space is infinite. So for instance, in Example 2.4 there are (nonoptimal) policies μ that are proper in the context of DSP, for which J_μ is real-valued but unbounded above, and would therefore not qualify as proper in the context of SSP. The reason for the requirement $J_\mu \in B(X)$ in SSP is in part technical, to make the analysis go through: it can be traced to the fact that in the context of DSP, for a proper policy μ we have $r_m(x,\mu) = 0$ after a finite number of transitions, but in the context of SSP finite convergence of $r_m(x,\mu)$ to 0 is not a reasonable assumption. This difference in the definition of a proper policy will in turn be reflected in the extensions of various results from DSP to SSP. For example, the SSP counterpart of Prop. 2.1(a) [uniqueness of fixed point of T within $B_b(X)$] will take the form of uniqueness of fixed point

of T within $B(X)$. Of course, in the case where X is finite we have $B(X) = B_b(X)$, so the fine distinctions just discussed disappear.

Note that a policy μ is proper for SSP if g is bounded and $\sum_{m=1}^{\infty} r_m(x,\mu) < \infty$ for all $x \in X$, since in this case we have

$$\sup_{x\in X} \big|J_\mu(x)\big| \leq \sup_{x\in X} \big|g\big(x,\mu(x)\big)\big| \sum_{m=1}^{\infty} r_m(x,\mu) < \infty,$$

so that $J_\mu \in B(X)$, as well as

$$\lim_{m\to\infty} r_m(x,\mu) E\big\{J(x_m) \mid x_0 = x,\, x_m \neq t,\, \mu\big\} \leq \sup_{x\in X} \big|J(x)\big| \lim_{m\to\infty} r_m(x,\mu) = 0, \qquad \forall\, x \in X, J \in B(X).$$

The following proposition establishes the connection between properness and S-regularity in the context of SSP, and parallels Prop. 3.6.

Proposition 5.1: Consider the SSP problem, viewed as a special case of the abstract semicontractive model of Section 3.1 with T_μ given by Eq. (5.1), and $\bar{J}$ being the zero function. Then a policy μ is $B(X)$-regular if and only if it is proper in the sense that $J_\mu \in B(X)$ and Eq. (5.2) holds.

Proof: For every $\mu \in \mathcal{M}$ we have

$$(T_\mu^m J)(x) = (T_\mu^m \bar{J})(x) + r_m(x,\mu) E\big\{J(x_m) \mid x_0 = x,\, x_m \neq t,\, \mu\big\}, \qquad \forall\, x \in X,\ J \in B(X),\ m \geq 1. \quad (5.3)$$

Let μ be proper, so that $J_\mu \in B(X)$. By taking the upper limit in Eq. (5.3) as $m \to \infty$, and by using Eq. (5.2), we obtain $\lim_{m\to\infty} (T_\mu^m J)(x) = J_\mu(x)$ for all $x \in X$ and $J \in S$, so μ is $B(X)$-regular.

Conversely if μ is improper, then either $J_\mu \notin B(X)$ in which case μ is not $B(X)$-regular [by the definition of $B(X)$-regularity in Section 3.1], or else $J_\mu \in B(X)$ and $r_m(x,\mu) E\big\{J(x_m) \mid x_0 = x,\, x_m \neq t,\, \mu\big\}$ does not converge to 0 for some $x \in X$, and $J \in B(X)$. Thus μ is not $B(X)$-regular. **Q.E.D.**

Using this proposition, the assumptions and results of Section 2 and the line of analysis of Section 3 generalize from DSP to SSP. Note, however, that there is a subtle difference between Props. 3.6 and 5.1. In the latter proposition S is a subset of $B(X)$ while in the former S may be a subset of $R(X)$. For this reason the set $B_b(X)$ must be replaced by $B(X)$ in the statements of various parts of Assumptions 2.1-2.3 and Props. 2.1-2.3. More specifically:

(a) Assumption 2.1(c) should be modified so that $B_b(X)$ is replaced by $B(X)$. Under this modified assumption, Prop. 2.1 holds for SSP with $B_b(X)$ is replaced by $B(X)$. The proof, given in Section 3.2 carries through verbatim, except that the set S is taken to be $B(X)$ rather than $B_b(X)$, and Prop. 5.1 is invoked in place of Prop. 3.6.

(b) To extend Prop. 2.2 for SSP, $R^+(X)$ should be replaced by $\big\{J \in B(X) \mid J \geq 0\big\}$ in Assumption 2.2. However, in this case, the result obtained is a weaker version of Prop. 2.1, as modified in (a) above, and is not worth considering.

(c) Under Assumption 2.3, Prop. 2.3(a),(b) holds for SSP as stated, because its proof relies only on the assumption $g(x,u) \geq 0$ for all (x,u) and does not depend on the definition of a proper policy. However,

the statement and the proof of Prop. 2.3(c),(d), when adapted to SSP requires that $R^+(X)$ be replaced by $\big\{J \in B(X) \mid J \geq 0\big\}$.

To extend Props. 2.4 and 2.5 to SSP, it is first necessary to modify Assumptions 2.4 and 2.5 so that $B_b(X)$ is replaced by $B(X)$. However, an additional condition is needed, namely that for each improper policy μ, there exists at least one $x \in X$ such that

$$\sum_{m=1}^{\infty} r_m(x, \mu) = \infty. \tag{5.4}$$

The idea is that Eq. (5.4) guarantees that for a policy μ that is improper for SSP, there is an $x \in X$ such that $J_{\mu,\delta}(x) = \infty$ for all $\delta > 0$, while Eq. (5.2) guarantees that a proper policy for SSP is also proper for δ-SSP and $J_{\mu,\delta} \in B(X)$. With the condition (5.4) added to Assumptions 2.4 and 2.5, Prop. 3.7 holds, namely for every $\delta > 0$ there exists an optimal proper policy for the δ-SSP problem, and the corresponding optimal cost function J^*_δ is a fixed point of T_δ. This is proved by applying the analog of Prop. 2.1 to the δ-SSP problem, based on the fact that for a proper policy μ, we have $J_{\mu,\delta} \in B(X)$, while for an improper policy μ, we have $J_{\mu,\delta}(x) = \infty$, cf. the preceding discussion. The proofs of Props. 2.4 and 2.5 then follow as in the case of DSP.

An alternative to the preceding line of analysis is to change the definition of a proper policy, so that a policy is proper if and only if it is $B_b(X)$-regular. This definition would be consistent with the one for DSP, since by Prop. 3.6, properness and $B_b(X)$-regularity are equivalent in the DSP context. However, then part of the definition of properness of μ would require that $\limsup_{m\to\infty} (T^m_\mu J)(x) \to J_\mu(x)$ for all $J \in B_b(X)$ and $x \in X$, which may be difficult to verify in the context of SSP.

6. CONCLUDING REMARKS

We have shown that deterministic and stochastic shortest path problems with arbitrary state space can be analyzed in a unified way within the context of abstract semicontractive models. We have needed several different assumptions for a detailed analysis, because small changes in some of the problem's characteristics may greatly affect the nature of the results that can be obtained, as we have illustrated with the simple two-node deterministic shortest path Example 2.1. We have highlighted three characteristics whose presence or absence may have a significant effect: the nonnegativity of arc lengths, the existence of an optimal policy that is proper, and "zero-cycle effects" (more precisely, whether improper policies have or do not have infinite cost for some initial state).

In this paper we have not discussed the cases where the optimal cost function is unbounded below, or takes the value $-\infty$ for some initial states. Problems of this type remain a subject for further investigation. Extensions to stochastic shortest path problems where the number of possible transitions corresponding to a given state-control pair is uncountable, are also worth considering. A general treatment of such problems must be carried out within an appropriate measurability framework, based for example on Borel spaces and universally measurable policies, cf. [BeS78], [JaC06], and the recent paper [YuB13b].

7. REFERENCES

[ALA08] Al-Tamimi, A., Lewis, F. L., and Abu-Khalaf, M., 2008. "Discrete-Time Nonlinear HJB Solution Using Approximate Dynamic Programming: Convergence Proof," IEEE Trans. on Cybernetics, Vol. 38, pp. 943-949.

[AMO89] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B., 1989. "Network Flows," in Handbooks in Operations Research and Management Science, Vol. 1, Optimization, Nemhauser, G. L., Rinnooy-Kan, A. H. G., and Todd M. J. (eds.), North-Holland, Amsterdam, pp. 211-369.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case, Academic Press, N. Y.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., 1989. Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, Englewood Cliffs, N. J.

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

[BeY10] Bertsekas, D. P., and Yu, H., 2010. "Asynchronous Distributed Policy Iteration in Dynamic Programming," Proc. of Allerton Conf. on Communication, Control and Computing, Allerton Park, Ill, pp. 1368-1374.

[BeY12] Bertsekas, D. P., and Yu, H., 2012. "Q-Learning and Enhanced Policy Iteration in Discounted Dynamic Programming," Math. of OR, Vol. 37, pp. 66-94.

[BeY13] Bertsekas, D. P., and Yu, H., 2013. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-P-2909, MIT.

[Ber98] Bertsekas, D. P., 1998. Network Optimization: Continuous and Discrete Models, Athena Scientific, Belmont, MA.

[Ber05] Bertsekas, D. P., 2005. Dynamic Programming and Optimal Control, Vol. I, 3rd Edition, Athena Scientific, Belmont, MA.

[Ber12] Bertsekas, D. P., 2012. Dynamic Programming and Optimal Control, Vol. II, 4th Edition: Approximate Dynamic Programming, Athena Scientific, Belmont, MA.

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA.

[Bla65] Blackwell, D., 1965. "Positive Dynamic Programming," Proc. Fifth Berkeley Symposium Math. Statistics and Probability, pp. 415-418.

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

[Der70] Derman, C., 1970. Finite State Markovian Decision Processes, Academic Press, N. Y.

[Dre69] Dreyfus, S. E., 1969. "An Appraisal of Some Shortest-Path Algorithms," Vol. 17, pp. 395-412.

[GaP88] Gallo, G., and Pallottino, S., 1988. "Shortest Path Algorithms," Annals of Operations Research, Vol. 7, pp. 3-79.

[HCP99] Hernandez-Lerma, O., Carrasco, O., and Perez-Hernandez. 1999. "Markov Control Processes with the Expected Total Cost Criterion: Optimality, Stability, and Transient Models," Acta Appl. Math., Vol. 59, pp. 229-269.

[HiW05] Hinderer, K., and Waldmann, K.-H., 2005. "Algorithms for Countable State Markov Decision Models with an Absorbing Set," SIAM J. of Control and Optimization, Vol. 43, pp. 2109-2131.

[JaC06] James, H. W., and Collins, E. J., 2006. "An Analysis of Transient Markov Decision Processes," J. Appl. Prob., Vol. 43, pp. 603-621.

[LeL12] Lewis, F. L., and Liu, D., 2012. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, IEEE Press Computational Intelligence Series, N. Y.

[LiW13] Liu, D., and Wei, Q., 2013. "Finite-Approximation-Error-Based Optimal Control Approach for Discrete-Time Nonlinear Systems," IEEE Trans. on Cybernetics, Vol. 43, 2013, pp. 779-789.

[Pal67] Pallu de la Barriere, R., 1967. Optimal Control Theory, Saunders, Phila; republished by Dover, N. Y., 1980.

[Pli78] Pliska, S. R., 1978. "On the Transient Case for Markov Decision Chains with General State Spaces," in Dynamic Programming and Its Applications, M. L. Puterman (ed.), Academic Press, N. Y.

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

[Roc84] Rockafellar, R. T., 1984. Network Flows and Monotropic Programming, Wiley-Interscience, N. Y.

[Str66] Strauch, R., 1966. "Negative Dynamic Programming," Ann. Math. Statist., Vol. 37, pp. 871-890.

[Whi79] Whittle, P., 1979. "A Simple Condition for Regularity in Negative Programming," J. Appl. Prob., Vol. 16, pp. 305318.

[Whi80] Whittle, P., 1980. "Stability and Characterisation Conditions in Negative Programming," J. Appl. Prob., Vol. 17, pp. 635645.

[Whi82] Whittle, P., 1982. Optimization Over Time, Wiley, N. Y.

[YuB13a] Yu, H., and Bertsekas, D. P., 2013. "Q-Learning and Policy Iteration Algorithms for Stochastic Shortest Path Problems," Annals of Operations Research, Vol. 208, pp. 95-132.

[YuB13b] Yu, H., and Bertsekas, D. P., 2013. "A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies," Lab. for Information and Decision Systems Report LIDS-P-2905, MIT, July 2013.