March 2017 (Revised May 2017) **Report LIDS-P-3506**

Stable Optimal Control and Semicontractive Dynamic Programming

Dimitri P. Bertsekas†

Abstract

We consider discrete-time infinite horizon deterministic optimal control problems with nonnegative cost per stage, and a destination that is cost-free and absorbing. The classical linear-quadratic regulator problem is a special case. Our assumptions are very general, and allow the possibility that the optimal policy may not be stabilizing the system, e.g., may not reach the destination either asymptotically or in a finite number of steps. We introduce a new unifying notion of stable feedback policy, based on perturbation of the cost per stage, which in addition to implying convergence of the generated states to the destination, quantifies the speed of convergence. We consider the properties of two distinct cost functions: J^*, the overall optimal, and $\hat{J}$, the restricted optimal over just the stable policies. Different classes of stable policies (with different speeds of convergence) may yield different values of $\hat{J}$. We show that for any class of stable policies, $\hat{J}$ is a solution of Bellman's equation, and we characterize the smallest and the largest solutions: they are J^*, and J^+, the restricted optimal cost function over the class of (finitely) terminating policies. We also characterize the regions of convergence of various modified versions of value and policy iteration algorithms, as substitutes for the standard algorithms, which may not work in general.

1. INTRODUCTION

In this paper we consider a deterministic discrete-time infinite horizon optimal control problem involving the system

$$x_{k+1} = f(x_k, u_k), \qquad k = 0, 1, \ldots, \tag{1.1}$$

where x_k and u_k are the state and control at stage k, which belong to sets X and U, referred to as the state and control spaces, respectively, and $f : X \times U \mapsto X$ is a given function. The control u_k must be chosen from a nonempty constraint set $U(x_k) \subset U$ that may depend on the current state x_k. The cost for the kth stage, $g(x_k, u_k)$, is assumed nonnegative and possibly extended real-valued:

$$0 \leq g(x_k, u_k) \leq \infty, \qquad \forall\ x_k \in X,\ u_k \in U(x_k),\ k = 0, 1, \ldots. \tag{1.2}$$

A cost per stage that is extended real-valued may be useful in modeling conveniently additional state and control constraints. We assume that X contains a special state, denoted t, which is referred to as the

† 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.

destination, and is cost-free and absorbing:

$$f(t,u) = t, \qquad g(t,u) = 0, \qquad \forall\ u \in U(t). \tag{1.3}$$

Our terminology aims to emphasize the connection with classical problems of control where X and U are the finite-dimensional Euclidean spaces $X = \Re^n$, $U = \Re^m$, and the destination is identified with the origin of $\Re^n$. There the essence of the problem is to reach or asymptotically approach the origin at minimum cost. A special case is the classical infinite horizon linear-quadratic regulator problem. However, our formulation also includes shortest path problems with continuous as well as discrete spaces; for example the classical shortest path problem, where X consists of the nodes of a directed graph, and the problem is to reach the destination from every other node with a minimum length path.

We are interested in feedback policies of the form $\pi = \{\mu_0, \mu_1, \ldots\}$, where each μ_k is a function mapping $x \in X$ into the control $\mu_k(x) \in U(x)$. The set of all policies is denoted by Π. Policies of the form $\pi = \{\mu, \mu, \ldots\}$ are called *stationary*, and will be denoted by μ, when confusion cannot arise.

Given an initial state x_0, a policy $\pi = \{\mu_0, \mu_1, \ldots\}$ when applied to the system (1.1), generates a unique sequence of state-control pairs $\big(x_k, \mu_k(x_k)\big)$, $k = 0, 1, \ldots$, with cost

$$J_\pi(x_0) = \sum_{k=0}^{\infty} g\big(x_k, \mu_k(x_k)\big), \qquad x_0 \in X,$$

[the series converges to some number in $[0, \infty]$ thanks to the nonnegativity assumption (1.2)]. We view J_π as a function over X, and we refer to it as the cost function of π. For a stationary policy μ, the corresponding cost function is denoted by J_μ. The optimal cost function is defined as

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

and a policy π^* is said to be optimal if $J_{\pi^*}(x) = J^*(x)$ for all $x \in X$. The optimal cost $J^*(x)$ is identical to the optimal cost attained when starting at x and using open-loop sequences $\{u_0, u_1, \ldots\}$, but in this paper we consider the broader class of feedback policies to be consistent with the formalism and analysis of infinite horizon DP.

We denote by $\mathcal{E}^+(X)$ the set of functions $J : X \mapsto [0, \infty]$. All equations, inequalities, limit and minimization operations involving functions from this set are meant to be pointwise. In our analysis, we will use the set of functions

$$\mathcal{J} = \big\{J \in \mathcal{E}^+(X) \mid J(t) = 0\big\}.$$

Since t is cost-free and absorbing, this set contains J_π of all $\pi \in \Pi$, as well as J^*.

It is well known that under the cost nonnegativity assumption (1.2), J^* satisfies Bellman's equation:

$$J^*(x) = \inf_{u \in U(x)} \big\{g(x,u) + J^*\big(f(x,u)\big)\big\}, \qquad x \in X,$$

and that an optimal stationary policy (if it exists) may be obtained through the minimization in the right side of this equation (cf. Prop. 2.1 in the next section). One also hopes to obtain J^* by means of value iteration (VI for short), which starting from some function $J_0 \in \mathcal{J}$, generates a sequence of functions $\{J_k\} \subset \mathcal{J}$ according to

$$J_{k+1}(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J_k\big(f(x,u)\big) \big\}, \qquad x \in X,\ k = 0, 1, \ldots. \tag{1.4}$$

However, $\{J_k\}$ may not always converge to J^* because, among other reasons, Bellman's equation may have multiple solutions within $\mathcal{J}$.

Another possibility to obtain J^* and an optimal policy is through policy iteration (PI for short), which starting from a stationary policy μ^0, generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\big(f\big(x, \mu^k(x)\big)\big), \qquad x \in X, \tag{1.5}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{1.6}$$

We note that J_{μ^k} satisfies Eq. (1.5) (cf. Prop. 2.1 in the next section). Moreover, when referring to PI, we assume that the minimum in Eq. (1.6) is attained for all $x \in X$, which is true under some compactness condition on $U(x)$ or the level sets of the function $g(x,\cdot) + J_k\big(f(x,\cdot)\big)$, or both (see e.g., [Ber12], Ch. 4).

The uniqueness of solution of Bellman's equation within $\mathcal{J}$, and the convergence of VI to J^* and of PI to an optimal policy have been investigated in a recent paper by the author [Ber15a] under conditions guaranteeing that J^* is the unique solution of Bellman's equation within $\mathcal{J}$. This paper also gives many references from the field of adaptive dynamic programming, where the continuous-spaces version of our problem is often used as the starting point for analysis and algorithmic development; see e.g., the book [VVL13], the papers [JiJ14], [LiW13], the survey papers in the edited volumes [SBP04] and [LeL13], and the special issue [LLL08]. Our purpose here is to consider the problem under weaker conditions and to make the connection with notions of stability. This is a more complicated case, where Bellman's equation need not have a unique solution within $\mathcal{J}$, while the VI and PI algorithms may be unreliable. However, several of the favorable results of [Ber15a] will be obtained as special cases of the results of this paper; see Section 3. The type of behavior that we are trying to quantify is described in the following example.†

† In this example and later, our standard notational convention is that all vectors in $\Re^n$ are viewed as column vectors. The real line is denoted by $\Re$. A prime denotes transposition, so inner product of two vectors x and y is defined by $x'y$, and the norm is $\|x\| = \sqrt{x'x}$.

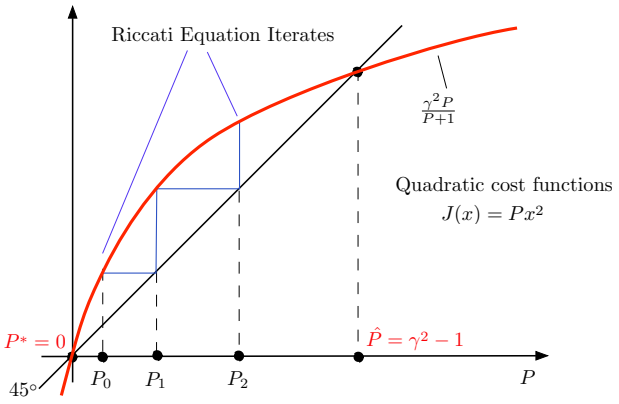

Figure 1.1 Illustration of the behavior of the Riccati equation for the linear-quadratic problem of Example 1.1, where the detectability assumption is not satisfied. The solutions of the Riccati equation are $P = 0$ (corresponds to the optimal cost) and $\hat{P} = \gamma^2 - 1$ (corresponds to the optimal cost that can be achieved with linear stable control laws).

Example 1.1 (Linear-Quadratic Problem)

Consider a linear system and a quadratic cost:

$$x_{k+1} = Ax_k + Bu_k, \qquad g(x_k, u_k) = x_k' Q x_k + u_k' R u_k,$$

where $X = \Re^n$, $U = \Re^m$, A, B, Q, and R are given matrices, with Q being positive semidefinite symmetric and R being positive definite symmetric. The classical results for this problem assume that:

(a) The pair (A, B) is stabilizable [i.e., there exists a linear policy $\mu(x) = Lx$ such that the closed-loop system $x_{k+1} = (A + BL)x_k$ is asymptotically stable].

(b) The pair (A, C), where $Q = C'C$, is detectable (i.e., if $u_k \to 0$ and $Cx_k \to 0$ then it follows that $x_k \to 0$).

Under these assumptions, it is well-known (see e.g., optimal control and estimation textbooks such as Anderson and Moore [AnM07], or the author's [Ber17], Section 3.1) that J^* has the form $J^*(x) = x'Px$, where P is the unique positive semidefinite solution of the algebraic Riccati equation

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q. \tag{1.7}$$

Furthermore the VI algorithm converges to J^* starting from any positive semidefinite quadratic function. Moreover the PI algorithm, starting from a linear stable policy, yields J^* and a linear stable optimal policy in the limit as first shown by Kleinman [Kle68].

To see what may happen when the preceding detectability condition is not satisfied, consider the scalar system

$$x_{k+1} = \gamma x_k + u_k,$$

where $\gamma > 1$ and the cost per stage is $g(x, u) = u^2$. Here we have $J^*(x) \equiv 0$, while the policy $\mu^*(x) \equiv 0$ is optimal. This policy is not stable (for any sensible definition of stability), which is not inconsistent with

optimality, since nonzero states are not penalized in this problem. The algebraic Riccati equation (1.7) for this case is

$$P = \frac{\gamma^2 P}{P+1},$$

and has *two nonnegative solutions*: $P^* = 0$ and $\hat{P} = \gamma^2 - 1$ (see Fig. 1.1). It turns out that $\hat{P}$ has an interesting interpretation: $\hat{J}(x) = \hat{P}x^2$ is the optimal cost that can be achieved using a linear stable control law, starting from x (see the analysis of [Ber17], Example 3.1.1). Moreover the VI algorithm, which generates the sequence $J_k(x) = P_k x^2$, with P_k obtained by the Riccati equation iteration

$$P_{k+1} = \frac{\gamma^2 P_k}{P_k + 1},$$

converges to $\hat{J}$ when started from any $P_0 > 0$, and stays at the zero function J^* when started from $P_0 = 0$ (see Fig. 1.1). Another interesting fact is that the PI algorithm, when started from a linear stable policy, yields in the limit $\hat{J}$ (not J^*) and the policy that is optimal within the class of linear stable policies (which turns out to be $\hat{\mu}(x) = \frac{1-\gamma^2}{\gamma}x$); see [Ber17], Section 3.1, for a verification, and Example 3.1 for an analysis of the multidimensional case.†

We note that the set of solutions of the Riccati equation has been extensively investigated starting with the papers by Willems [Wil71] and Kucera [Kuc72], [Kuc73], which were followed up by several other works (see the book by Lancaster and Rodman [LaR95] for a comprehensive treatment). In these works, the "largest" solution of the Riccati equation is referred to as the "stabilizing" solution, and the stability of the corresponding policy is shown, although the author could not find an explicit statement regarding the optimality of this policy within the class of all linear stable policies. Also the lines of analysis of these works are tied to the structure of the linear-quadratic problem and are unrelated to the analysis of the present paper.

There are also other interesting deterministic optimal control examples where Bellman's equation, and the VI and PI algorithms exhibit unusual behavior, including several types of shortest path problems (see e.g., [Ber14], [Ber15a], [BeY16], and the subsequent Example 4.1). This is typical of *semicontractive DP* theory, which is a central focal point of the author's abstract DP monograph [Ber13], and followup work [Ber15b]. The present paper is inspired by the analytical methods of this theory. In semicontractive models, roughly speaking, policies are divided into those that are "regular" in the sense that they are "well-behaved" with respect to the VI algorithm, and those that are "irregular." The optimal cost function over

† As an example of what may happen without stabilizability, consider the case when the system is instead $x_{k+1} = \gamma x_k$. Then the Riccati equation becomes $P = \gamma^2 P$ and has $P^* = 0$ as its unique solution. However, the Riccati equation iteration $P_{k+1} = \gamma^2 P_k$ diverges to ∞ starting from any $P_0 > 0$. Also, qualitatively similar behavior is obtained when there is a discount factor $\alpha \in (0, 1)$. The Riccati equation takes the form

$$P = A'\big(\alpha P - \alpha^2 PB(\alpha B'PB + R)^{-1}B'P\big)A + Q,$$

and for the given system and cost per stage, it has two solutions, $P^* = 0$ and $\hat{P} = \frac{\alpha\gamma^2 - 1}{\alpha}$. The VI algorithm converges to $\hat{P}$ starting from any $P > 0$. While the line of analysis of the present paper does not apply to discounted problems, a related analysis is given in the paper [Ber15b], using the idea of a regular policy.

the "regular" policies (under suitable conditions) is a solution of Bellman's equation, and can be found by the VI and PI algorithm, even under conditions where these algorithms may fail to find the optimal cost function J^*. Regularity in the sense of semicontractive DP corresponds to stability in the specialized context of deterministic optimal control considered here.

In this paper we address the phenomena illustrated by the linear-quadratic Example 1.1 in the more general setting where the system may be nonlinear and the cost function may be nonquadratic. Our method of analysis is to introduce a cost perturbation that involves a penalty for excursions of the state from the destination, thus resulting in a better-behaved problem. The type of perturbation used determines in turn the class of stable policies. A key aspect of our definition of a stable policy (as given in the next section) is that in addition to implying convergence of the generated states to the destination, it quantifies the speed of convergence.

A simpler approach, which involves perturbation by a constant function, has been used in the monograph [Ber13], and also in the paper by Bertsekas and Yu [BeY16]. The latter paper analyzes similarly unusual behavior in finite-state finite-control stochastic shortest path problems, where the cost per stage can take both positive and negative values (for such problems the anomalies are even more acute, including the possibility that J^* may not solve Bellman's equation).

In the analysis of the present paper, the optimal policies of the perturbed problem are stable policies, and in the limit as the perturbation diminishes to 0, the corresponding optimal cost function converges to $\hat{J}$, the optimal cost function over stable policies (not to J^*). Our central result is that $\hat{J}$ is the unique solution of Bellman's equation within a set of functions in $\mathcal{J}$ that majorize $\hat{J}$. Moreover, the VI algorithm converges to $\hat{J}$ when started from within this set. In addition, if J^+, the optimal cost function over the class of (finitely) terminating policies belongs to $\mathcal{J}$, then J^+ is the largest solution of Bellman's equation within $\mathcal{J}$. These facts are shown in Section 3, including a treatment of the multidimensional version of the linear-quadratic problem of Example 1.1. In Section 3, we also consider the favorable special case where $J^* = J^+$, and we develop the convergence properties of VI for this case. In Section 4 we consider PI algorithms, including a perturbed version (Section 4.2).

2. STABLE POLICIES

In this section, we will lay the groundwork for our analysis and introduce the notion of a stable policy. To this end, we will use some classical results for optimal control with nonnegative cost per stage, which stem from the original work of Strauch [Str66]. For textbook accounts we refer to [BeS78], [Put94], [Ber12], and for a more abstract development, we refer to the monograph [Ber13].

The following proposition gives the results that we will need (see [BeS78], Props. 5.2, 5.4, and 5.10,

[Ber12], Props. 4.1.1, 4.1.3, 4.1.5, 4.1.9, or [Ber13], Props. 4.3.3, 4.3.9, and 4.3.14). Actually, these results hold for stochastic infinite horizon DP problems with nonnegative cost per stage, and do not depend on the favorable structure of this paper (a deterministic problem with a cost-free and absorbing destination).

Proposition 2.1: The following hold:

(a) J^* is a solution of Bellman's equation and if $J \in \mathcal{E}^+(X)$ is another solution, i.e., J satisfies

$$J(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X, \tag{2.1}$$

then $J^* \leq J$.

(b) For all stationary policies μ, J_μ is a solution of the equation

$$J(x) = g\big(x,\mu(x)\big) + J\big(f\big(x,\mu(x)\big)\big), \qquad \forall\ x \in X,$$

and if $J \in \mathcal{E}^+(X)$ is another solution, then $J_\mu \leq J$.

(c) A stationary policy μ^* is optimal if and only if

$$\mu^*(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J^*\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X.$$

(d) Let $\{\bar{J}_k\}$ be the sequence generated by the VI algorithm (1.4) starting from the zero function $\bar{J}_0(x) \equiv 0$. If U is a metric space and the sets

$$U_k(x,\lambda) = \big\{ u \in U(x) \big|\ g(x,u) + \bar{J}_k\big(f(x,u)\big) \leq \lambda \big\} \tag{2.2}$$

are compact for all $x \in X$, $\lambda \in \Re$, and $k \geq 0$, then there exists at least one optimal stationary policy, and we have $J_k \to J^*$ for every sequence generated by VI starting from a function $J_0 \in \mathcal{E}^+(X)$ with $J_0 \leq J^*$.

(e) For every $\epsilon > 0$ there exists an ϵ-optimal policy, i.e., a policy π_ϵ such that

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

We introduce a *forcing function* $p : X \mapsto [0,\infty)$ such that

$$p(t) = 0, \qquad p(x) > 0, \qquad \forall\ x \neq t,$$

and the *p-δ-perturbed optimal control problem*, where $\delta > 0$ is a given scalar. This is the same problem as the original, except that the cost per stage is changed to

$$g(x,u) + \delta p(x).$$

We denote by $J_{\pi,p,\delta}$ the cost function of a policy $\pi \in \Pi$ in the p-δ-perturbed problem:

$$J_{\pi,p,\delta}(x_0) = J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k), \tag{2.3}$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π. We also denote by $\hat{J}_{p,\delta}$, the corresponding optimal cost function, $\hat{J}_{p,\delta} = \inf_{\pi\in\Pi} J_{\pi,p,\delta}$. We introduce a notion of stability involving the p-δ-perturbed problem.

Definition 2.1: Let p be a given forcing function. For a state $x \in X$, we say that a policy π is *p-stable from* x if for all $\delta > 0$ we have

$$J_{\pi,p,\delta}(x) < \infty.$$

The set of all such policies is denoted by $\Pi_{p,x}$. We define the *restricted optimal cost function* over $\Pi_{p,x}$ by

$$\hat{J}_p(x) = \inf_{\pi\in\Pi_{p,x}} J_\pi(x), \qquad x \in X. \tag{2.4}$$

We say that π is *p-stable* (without qualification) if $\pi \in \Pi_{p,x}$ for all $x \in X$ such that $\Pi_{p,x} \neq \emptyset$. The set of all p-stable policies is denoted by Π_p.

The preceding definition of a p-stable policy is novel within the general context of this paper, and is inspired from a notion of regularity, which is central in the theory of semicontractive DP; see [Ber13] and the related subsequent papers [Ber14], [Ber15b], [Ber16]. Note that the set $\Pi_{p,x}$ depends on the forcing function p. As an example, let $X = \Re^n$ and

$$p(x) = \|x\|^\rho,$$

where $\rho > 0$ is a scalar. Then roughly speaking, ρ quantifies the rate at which the destination is approached using the p-stable policies. In particular, the policies $\pi \in \Pi_{p,x_0}$ are the ones that force x_k towards 0 at a rate faster than $O(1/k^\rho)$, so slower policies would be excluded from Π_{p,x_0}.

Let us make some observations regarding p-stability:

(a) *Equivalent definition of p-stability*: Given any policy π and state $x_0 \in X$, from Eq. (2.3) it follows that

$$\pi \in \Pi_{p,x_0} \qquad \text{if and only if} \qquad J_\pi(x_0) < \infty \text{ and } \sum_{k=0}^{\infty} p(x_k) < \infty, \tag{2.5}$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π. Since the right-hand side of the preceding relation does not depend on δ, it also follows that an equivalent definition of a policy π that is p-stable from x is that $J_{\pi,p,\delta}(x) < \infty$ for some $\delta > 0$ (rather than all $\delta > 0$).

(b) *Approximation property of $J_{\pi,p,\delta}(x)$*: Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By taking the limit as $\delta \downarrow 0$ in the expression

$$J_{\pi,p,\delta}(x_0) = J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k),$$

[cf. Eq. (2.3)] and by using Eq. (2.5), it follows that

$$\lim_{\delta \downarrow 0} J_{\pi,p,\delta}(x_0) = J_\pi(x_0), \qquad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p,x_0}. \tag{2.6}$$

From this equation, we have that if $\pi \in \Pi_{p,x}$, then $J_{\pi,p,\delta}(x)$ is finite and differs from $J_\pi(x)$ by $O(\delta)$. By contrast, if $\pi \notin \Pi_{p,x}$, then $J_{\pi,p,\delta}(x) = \infty$ by the definition of p-stability.

(c) *Limiting property of $\hat{J}_p(x_k)$*: Consider a pair (π, x_0) with $\pi \in \Pi_{p,x_0}$. By breaking down $J_{\pi,p,\delta}(x_0)$ into the sum of the costs of the first k stages and the remaining stages, we have

$$J_{\pi,p,\delta}(x_0) = \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big) + \delta \sum_{m=0}^{k-1} p(x_m) + J_{\pi_k,p,\delta}(x_k), \qquad \forall\ \delta > 0,\ k > 0, \tag{2.7}$$

where $\{x_k\}$ is the sequence generated starting from x_0 and using π, and π_k is the policy $\{\mu_k, \mu_{k+1}, \ldots\}$. By taking the limit as $k \to \infty$ and using Eq. (2.3), it follows that

$$\lim_{k \to \infty} J_{\pi_k,p,\delta}(x_k) = 0, \qquad \forall \text{ pairs } (\pi, x_0) \text{ with } \pi \in \Pi_{p,x_0},\ \delta > 0. \tag{2.8}$$

Also, since $\hat{J}_p(x_k) \le \hat{J}_{p,\delta}(x_k) \le J_{\pi_k,p,\delta}(x_k)$, it follows that

$$\lim_{k \to \infty} J_{p,\delta}(x_k) = 0, \qquad \lim_{k \to \infty} \hat{J}_p(x_k) = 0, \qquad \forall\ (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p,x_0},\ \delta > 0. \tag{2.9}$$

Terminating Policies and Controllability

An important special case is when p is equal to the function

$$p^+(x) = \begin{cases} 0 & \text{if } x = t, \\ 1 & \text{if } x \neq t. \end{cases} \tag{2.10}$$

For $p = p^+$, a policy π is p^+-stable from x if and only if it is *terminating from* x, i.e., reaches t in a finite number of steps starting from x [cf. Eq. (2.5)]. The set of terminating policies from x is denoted by Π_x^+ and it is contained within every other set of p-stable policies $\Pi_{p,x}$, as can be seen from Eq. (2.5). As a result, the restricted optimal cost function over Π_x^+,

$$J^+(x) = \inf_{\pi \in \Pi_x^+} J_\pi(x), \qquad x \in X,$$

satisfies $J^*(x) \leq \hat{J}_p(x) \leq J^+(x)$ for all $x \in X$. A policy π is said to be *terminating* if it is simultaneously terminating from all $x \in X$ such that $\Pi_x^+ \neq \emptyset$. The set of all terminating policies is denoted by Π^+.

Note that if the state space X is finite, we have for every forcing function p

$$\underline{\beta}\, p^+(x) \leq p(x) \leq \bar{\beta}\, p^+(x), \qquad \forall\, x \in X,$$

for some scalars $\underline{\beta}, \bar{\beta} > 0$. As a result it can be seen that $\Pi_{p,x} = \Pi_x^+$ and $\hat{J}_p = J^+$, so in effect the case where $p = p^+$ is the only case of interest for finite-state problems.

The notion of a terminating policy is related to the notion of *controllability*. In classical control theory terms, the system $x_{k+1} = f(x_k, u_k)$ is said to be completely controllable if for every $x_0 \in X$, there exists a policy that drives the state x_k to the destination in a finite number of steps. This notion of controllability is equivalent to the existence of a terminating policy from each $x \in X$.

One of our main results, to be shown in the next section, is that J^*, $\hat{J}_p$, and J^+ are solutions of Bellman's equation, with J^* being the "smallest" solution and J^+ being the "largest" solution within $\mathcal{J}$. The most favorable situation arises when $J^* = J^+$, in which case J^* is the unique solution of Bellman's equation within $\mathcal{J}$. Moreover, in this case it will be shown that the VI algorithm converges to J^* starting with any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$ (see the subsequent Prop. 3.5), and the PI algorithm converges to J^* as well (see Section 4.1). This special case has been discussed in the paper [Ber15a].

3. RESTRICTED OPTIMIZATION OVER STABLE POLICIES

For a given forcing function p, we denote by $\widehat{X}_p$ the effective domain of $\hat{J}_p$, the set of all x where $\hat{J}_p$ is finite,

$$\widehat{X}_p = \big\{x \in X \mid \hat{J}_p(x) < \infty\big\}.$$

Since $\hat{J}_p(x) < \infty$ if and only if $\Pi_{p,x} \neq \emptyset$ [cf. Eqs. (2.4) and (2.5)], or equivalently $J_{\pi,p,\delta}(x) < \infty$ for some π and all $\delta > 0$, it follows that $\widehat{X}_p$ is also the effective domain of $\hat{J}_{p,\delta}$,

$$\widehat{X}_p = \big\{x \in X \mid \Pi_{p,x} \neq \emptyset\big\} = \big\{x \in X \mid \hat{J}_{p,\delta}(x) < \infty\big\}, \qquad \forall\, \delta > 0.$$

Note that $\widehat{X}_p$ may depend on p and may be a strict subset of the effective domain of J^*, which is denoted by

$$X^* = \big\{x \in X \mid J^*(x) < \infty\big\}.$$

The reason is that there may exist a policy π such that $J_\pi(x) < \infty$, even when there is no p-stable policy from x.

Our first objective is to show that as $\delta \downarrow 0$, the p-δ-perturbed optimal cost function $\hat{J}_{p,\delta}$ converges to the restricted optimal cost function $\hat{J}_p$.

Proposition 3.1 (Approximation Property of $\hat{J}_{p,\delta}$): Let p be a given forcing function and $\delta > 0$.

(a) We have

$$J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x), \qquad \forall\ x \in X,\ \pi \in \Pi_{p,x}, \tag{3.1}$$

where $w_{\pi,p,\delta}$ is a function such that $\lim_{\delta \downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$.

(b) We have

$$\lim_{\delta \downarrow 0} \hat{J}_{p,\delta}(x) = \hat{J}_p(x), \qquad \forall\ x \in X.$$

Proof: (a) Follows by using Eq. (2.6) for $x \in \widehat{X}_p$, and by taking $w_{p,\delta}(x) = 0$ for $x \notin \widehat{X}_p$.

(b) By Prop. 2.1(e), there exists an ϵ-optimal policy π_ϵ for the p-δ-perturbed problem, i.e., $J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. Moreover, for $x \in \widehat{X}_p$ we have $\hat{J}_{p,\delta}(x) < \infty$, so $J_{\pi_\epsilon,p,\delta}(x) < \infty$. Hence π_ϵ is p-stable from all $x \in \widehat{X}_p$, and we have $\hat{J}_p \leq J_{\pi_\epsilon}$. Using also Eq. (3.1), we have for all $\delta > 0$, $\epsilon > 0$, $x \in X$, and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) - \epsilon \leq J_{\pi_\epsilon}(x) - \epsilon \leq J_{\pi_\epsilon,p,\delta}(x) - \epsilon \leq \hat{J}_{p,\delta}(x) \leq J_{\pi,p,\delta}(x) = J_\pi(x) + w_{\pi,p,\delta}(x),$$

where $\lim_{\delta \downarrow 0} w_{\pi,p,\delta}(x) = 0$ for all $x \in X$. By taking the limit as $\epsilon \downarrow 0$, we obtain for all $\delta > 0$ and $\pi \in \Pi_{p,x}$,

$$\hat{J}_p(x) \leq \hat{J}_{p,\delta}(x) \leq J_\pi(x) + w_{\pi,p,\delta}(x), \qquad \forall\ x \in X.$$

By taking the limit as $\delta \downarrow 0$ and then the infimum over all $\pi \in \Pi_{p,x}$, we have

$$\hat{J}_p(x) \leq \lim_{\delta \downarrow 0} \hat{J}_{p,\delta}(x) \leq \inf_{\pi \in \Pi_{p,x}} J_\pi(x) = \hat{J}_p(x), \qquad \forall\ x \in X,$$

from which the result follows. **Q.E.D.**

We now consider approximately optimal policies. Given any $\epsilon > 0$, by Prop. 2.1(e), there exists an ϵ-optimal policy for the p-δ-perturbed problem, i.e., a policy π such that $J_\pi(x) \leq \hat{J}_{p,\delta}(x) + \epsilon$ for all $x \in X$. We address the question whether there exists a p-stable policy π that is ϵ-optimal for the restricted optimization over p-stable policies, i.e., a policy π that is p-stable simultaneously from all $x \in X_p$, (i.e., $\pi \in \Pi_p$) and satisfies

$$J_\pi(x) \leq \hat{J}_p(x) + \epsilon, \qquad \forall\ x \in X.$$

We refer to such a policy as a *p-ϵ-optimal policy*.

Proposition 3.2 (Existence of p-ϵ-Optimal Policy): Let p be a given forcing function and $\delta > 0$. For every $\epsilon > 0$, a policy π that is ϵ-optimal for the p-δ-perturbed problem is p-ϵ-optimal, and hence belongs to Π_p.

Proof: For any ϵ-optimal policy π_ϵ for the p-δ-perturbed problem, we have

$$J_{\pi_\epsilon,p,\delta}(x) \leq \hat{J}_{p,\delta}(x) + \epsilon < \infty, \qquad \forall\ x \in \widehat{X}_p.$$

This implies that $\pi_\epsilon \in \Pi_p$. Moreover, for all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in \widehat{X}_p$ and $\pi \in \Pi_{p,x_0}$, we have

$$J_{\pi_\epsilon}(x_0) \leq J_{\pi_\epsilon,p,\delta}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon \leq J_\pi(x_0) + \delta \sum_{k=0}^{\infty} p(x_k) + \epsilon.$$

Taking the limit as $\delta \downarrow 0$ and using the fact $\sum_{k=0}^{\infty} p(x_k) < \infty$ (since $\pi \in \Pi_{p,x_0}$), we obtain

$$J_{\pi_\epsilon}(x_0) \leq J_\pi(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p,\ \pi \in \Pi_{p,x_0}.$$

By taking infimum over $\pi \in \Pi_{p,x_0}$, it follows that

$$J_{\pi_\epsilon}(x_0) \leq \hat{J}_p(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p,$$

which in view of the fact $J_{\pi_\epsilon}(x_0) = \hat{J}_p(x_0) = \infty$ for $x_0 \notin \widehat{X}_p$, implies that π_ϵ is p-ϵ-optimal. **Q.E.D.**

Note that the preceding proposition implies that

$$\hat{J}_p(x) = \inf_{\pi \in \Pi_p} J_\pi(x), \qquad \forall\ x \in X, \tag{3.2}$$

which is a stronger statement than the definition $\hat{J}_p(x) = \inf_{\pi \in \Pi_{p,x}} J_\pi(x)$ for all $x \in X$. However, it can be shown through examples that there may not exist a restricted-optimal p-stable policy, i.e., a $\pi \in \Pi_p$ such that $J_\pi = \hat{J}_p$, even if there exists an optimal policy for the original problem. One such example is the one-dimensional linear-quadratic problem of Example 1.1 for the case where $p = p^+$. Then, there exists a unique linear stable policy that attains the restricted optimal cost $J^+(x)$ for all x (cf. Fig. 1.1), but this policy is not terminating. Note also that there may not exist a *stationary* p-ϵ-optimal policy, since generally in undiscounted nonnegative optimal control problems there may not exist a stationary ϵ-optimal policy, as is well-known (for an example, see [Ber13], following Prop. 4.3.2).

Our next proposition is preliminary for our main result. It involves the set of functions S_p given by

$$S_p = \Big\{ J \in \mathcal{J} \;\Big|\; J(x_k) \to 0 \text{ for all sequences } \{x_k\} \text{ generated from initial state-policy pairs } (\pi, x_0) \text{ with } x_0 \in X \text{ and } \pi \in \Pi_{p,x_0} \Big\}. \tag{3.3}$$

In words, S_p is the set of functions in $\mathcal{J}$ whose value is asymptotically driven to 0 by all the policies that are p-stable starting from some $x_0 \in X$ (and thus have the character of Lyapounov functions for these policies).

Note that S_p contains $\hat{J}_p$ and $\hat{J}_{p,\delta}$ for all $\delta > 0$ [cf. Eq. (2.9)]. Moreover, S_p contains all functions J such that $0 \leq J \leq h(\hat{J}_{p,\delta})$ for some $\delta > 0$ and function $h : X \mapsto X$ such that $h(J) \to 0$ as $J \to 0$. For example S_p contains all J such that $0 \leq J \leq c\,\hat{J}_{p,\delta}$ for some $c > 0$ and $\delta > 0$.

We summarize the preceding discussion in the following proposition, which also shows uniqueness of solution (within S_p) of Bellman's equation for the p-δ-perturbed problem. The significance of this is that the p-δ-perturbed problem, which can be solved more reliably than the original problem (including by VI methods), can yield a close approximation to $\hat{J}_p$ for sufficiently small δ [cf. Prop. 3.1(b)].

Proposition 3.3: Let p be a forcing function and $\delta > 0$. The function $\hat{J}_{p,\delta}$ belongs to the set S_p, and is the unique solution within S_p of Bellman's equation for the p-δ-perturbed problem,

$$\hat{J}_{p,\delta}(x) = \inf_{u \in U(x)} \Big\{ g(x,u) + \delta p(x) + \hat{J}_{p,\delta}\big(f(x,u)\big) \Big\}, \qquad x \in X. \tag{3.4}$$

Moreover, S_p contains $\hat{J}_p$ and all functions J satisfying

$$0 \leq J \leq h(\hat{J}_{p,\delta})$$

for some $h : X \mapsto X$ with $h(J) \to 0$ as $J \to 0$.

Proof: We have $\hat{J}_{p,\delta} \in S_p$ and $\hat{J}_p \in S_p$ by Eq. (2.9), as noted earlier. We also have that $\hat{J}_{p,\delta}$ is a solution of Bellman's equation (3.4) by Prop. 2.1(a). To show that $\hat{J}_{p,\delta}$ is the unique solution within S_p, let $\tilde{J} \in S_p$ be another solution, so that using also Prop. 2.1(a), we have

$$\hat{J}_{p,\delta}(x) \leq \tilde{J}(x) \leq g(x,u) + \delta p(x) + \tilde{J}\big(f(x,u)\big), \qquad \forall\, x \in X,\ u \in U(x). \tag{3.5}$$

Fix $\epsilon > 0$, and let $\pi = \{\mu_0, \mu_1, \ldots\}$ be an ϵ-optimal policy for the p-δ-perturbed problem. By repeatedly applying the preceding relation, we have for any $x_0 \in \widehat{X}_p$,

$$\hat{J}_{p,\delta}(x_0) \leq \tilde{J}(x_0) \leq \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big), \qquad \forall\, k \geq 1, \tag{3.6}$$

where $\{x_k\}$ is the state sequence generated starting from x_0 and using π. We have $\tilde{J}(x_k) \to 0$ (since $\tilde{J} \in S_p$ and $\pi \in \Pi_p$ by Prop. 3.2), so that

$$\lim_{k\to\infty} \left\{ \tilde{J}(x_k) + \delta \sum_{m=0}^{k-1} p(x_m) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big) \right\} = J_{\pi,\delta}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon. \tag{3.7}$$

By combining Eqs. (3.6) and (3.7), we obtain

$$\hat{J}_{p,\delta}(x_0) \leq \tilde{J}(x_0) \leq \hat{J}_{p,\delta}(x_0) + \epsilon, \qquad \forall\ x_0 \in \widehat{X}_p.$$

By letting $\epsilon \to 0$, it follows that $\hat{J}_{p,\delta}(x_0) = \tilde{J}(x_0)$ for all $x_0 \in \widehat{X}_p$. Also for $x_0 \notin \widehat{X}_p$, we have $\hat{J}_{p,\delta}(x_0) = \tilde{J}(x_0) = \infty$ [since $\hat{J}_{p,\delta}(x_0) = \infty$ for $x_0 \notin \widehat{X}_p$ and $\hat{J}_{p,\delta} \leq \tilde{J}$, cf. Eq. (3.5)]. Thus $\hat{J}_{p,\delta} = \tilde{J}$, proving that $\hat{J}_{p,\delta}$ is the unique solution of the Bellman Eq. (3.4) within S_p. **Q.E.D.**

We next show that $\hat{J}_p$ is the unique solution of Bellman's equation within the set of functions

$$\mathcal{W}_p = \{J \in S_p \mid \hat{J}_p \leq J\}, \tag{3.8}$$

and that the VI algorithm yields $\hat{J}_p$ in the limit for any initial $J_0 \in \mathcal{W}_p$.

Proposition 3.4: Let p be a given forcing function. Then:

(a) $\hat{J}_p$ is the unique solution of Bellman's equation

$$J(x) = \inf_{u\in U(x)} \Big\{ g(x,u) + J\big(f(x,u)\big) \Big\}, \qquad x \in X, \tag{3.9}$$

within the set $\mathcal{W}_p$ of Eq. (3.8).

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (1.4) starting with some $J_0 \in \mathcal{W}_p$, then $J_k \to \hat{J}_p$.

(c) (*Optimality Condition*) If $\hat{\mu}$ is a p-stable stationary policy and

$$\hat{\mu}(x) \in \arg\min_{u\in U(x)} \big\{ g(x,u) + \hat{J}_p\big(f(x,u)\big) \big\}, \qquad \forall\ x \in X, \tag{3.10}$$

then $\hat{\mu}$ is optimal over the set of p-stable policies. Conversely, if $\hat{\mu}$ is optimal within the set of p-stable policies, then it satisfies the preceding condition (3.10).

Proof: (a), (b) We first show that $\hat{J}_p$ is a solution of Bellman's equation. Since $\hat{J}_{p,\delta}$ is a solution [cf. Prop. 3.3] and $\hat{J}_{p,\delta} \geq \hat{J}_p$ [cf. Prop. 3.1(b)], we have for all $\delta > 0$,

$$\hat{J}_{p,\delta}(x) = \inf_{u\in U(x)} \Big\{ g(x,u) + \delta p(x) + \hat{J}_{p,\delta}\big(f(x,u)\big) \Big\}$$

$$\geq \inf_{u\in U(x)} \Big\{ g(x,u) + \hat{J}_{p,\delta}\big(f(x,u)\big) \Big\}$$
$$\geq \inf_{u\in U(x)} \Big\{ g(x,u) + \hat{J}_{p}\big(f(x,u)\big) \Big\}.$$

By taking the limit as $\delta \downarrow 0$ and using the fact $\lim_{\delta\downarrow 0} \hat{J}_{p,\delta} = \hat{J}_p$ [cf. Prop. 3.1(b)], we obtain

$$\hat{J}_p(x) \geq \inf_{u\in U(x)} \Big\{ g(x,u) + \hat{J}_{p}\big(f(x,u)\big) \Big\}, \qquad \forall\ x \in X. \tag{3.11}$$

For the reverse inequality, let $\{\delta_m\}$ be a sequence with $\delta_m \downarrow 0$. From Prop. 3.3, we have for all m, $x \in X$, and $u \in U(x)$,

$$g(x,u) + \delta_m p(x) + \hat{J}_{p,\delta_m}\big(f(x,u)\big) \geq \inf_{v\in U(x)} \Big\{ g(x,v) + \delta_m p(x) + \hat{J}_{p,\delta_m}\big(f(x,v)\big) \Big\} = \hat{J}_{p,\delta_m}(x).$$

Taking the limit as $m \to \infty$, and using the fact $\lim_{\delta_m\downarrow 0} \hat{J}_{p,\delta_m} = \hat{J}_p$ [cf. Prop. 3.1(b)], we have

$$g(x,u) + \hat{J}_{p}\big(f(x,u)\big) \geq \hat{J}_p(x), \qquad \forall\ x \in X,\ u \in U(x),$$

so that

$$\inf_{u\in U(x)} \Big\{ g(x,u) + \hat{J}_{p}\big(f(x,u)\big) \Big\} \geq \hat{J}_p(x), \qquad \forall\ x \in X. \tag{3.12}$$

By combining Eqs. (3.11) and (3.12), we see that $\hat{J}_p$ is a solution of Bellman's equation. We also have $\hat{J}_p \in S_p$ by Prop. 3.3, implying that $\hat{J}_p \in \mathcal{W}_p$ and proving part (a) except for the uniqueness assertion.

We will now prove part (b). Let $\pi = \{\mu_0, \mu_1, \ldots\} \in \Pi_p$ (which is nonempty by Prop. 3.2), and for $x_0 \in \widehat{X}_p$, let $\{x_k\}$ be the generated sequence starting from x_0 and using π. We have $J_0(x_k) \to 0$ since $J_0 \in S_p$. Since from the definition of the VI sequence $\{J_k\}$, we have

$$J_k(x) \leq g(x,u) + J_{k-1}\big(f(x,u)\big), \qquad \forall\ x \in X,\ u \in U(x),\ k = 1, 2, \ldots,$$

it follows that

$$J_k(x_0) \leq J_0(x_k) + \sum_{m=0}^{k-1} g\big(x_m, \mu_m(x_m)\big).$$

By taking limit as $k \to \infty$ and using the fact $J_0(x_k) \to 0$, it follows that $\limsup_{k\to\infty} J_k(x_0) \leq J_\pi(x_0)$. By taking the infimum over all $\pi \in \Pi_p$, we obtain $\limsup_{k\to\infty} J_k(x_0) \leq \hat{J}_p(x_0)$. Conversely, since $\hat{J}_p \leq J_0$ and $\hat{J}_p$ is a solution of Bellman's equation (as shown earlier), it follows by induction that $\hat{J}_p \leq J_k$ for all k. Thus $\hat{J}_p(x_0) \leq \liminf_{k\to\infty} J_k(x_0)$, implying that $J_k(x_0) \to \hat{J}_p(x_0)$ for all $x_0 \in \widehat{X}_p$. We also have $\hat{J}_p \leq J_k$ for all k, so that $\hat{J}_p(x_0) = J_k(x_0) = \infty$ for all $x_0 \notin \widehat{X}_p$. This completes the proof of part (b). Finally, since $\hat{J}_p \in \mathcal{W}_p$ and $\hat{J}_p$ is a solution of Bellman's equation, part (b) implies the uniqueness assertion of part (a).

(c) If μ is p-stable and Eq. (3.10) holds, then

$$\hat{J}_p(x) = g\big(x, \mu(x)\big) + \hat{J}_p\big(f(x,\mu(x))\big), \qquad x \in X.$$

By Prop. 2.1(b), this implies that $J_\mu \leq \hat{J}_p$, so μ is optimal over the set of p-stable policies. Conversely, assume that μ is p-stable and $J_\mu = \hat{J}_p$. Then by Prop. 2.1(b), we have

$$\hat{J}_p(x) = g\big(x,\mu(x)\big) + \hat{J}_p\big(f(x,\mu(x))\big), \qquad x \in X,$$

and since [by part (a)] $\hat{J}_p$ is a solution of Bellman's equation,

$$\hat{J}_p(x) = \inf_{u\in U(x)} \big\{g(x,u) + \hat{J}_p\big(f(x,u)\big)\big\}, \qquad x \in X.$$

Combining the last two relations, we obtain Eq. (3.10). **Q.E.D.**

We now consider the special case where p is equal to the function $p^+(x) = 1$ for all $x \neq t$ [cf. Eq. (2.10)]. The set of p^+-stable policies from x is Π_x^+, the set of terminating policies from x, and $J^+(x)$ is the corresponding restricted optimal cost,

$$J^+(x) = \hat{J}_{p^+}(x) = \inf_{\pi\in\Pi_x^+} J_\pi(x) = \inf_{\pi\in\Pi^+} J_\pi(x), \qquad x \in X,$$

[the last equality follows from Eq. (3.2)]. In this case, the set S_{p^+} of Eq. (3.3) is the entire set $\mathcal{J}$, since for all $J \in \mathcal{J}$ and all sequences $\{x_k\}$ generated from initial state-policy pairs (π, x_0) with $x_0 \in X$ and π terminating from x_0, we have $J(x_k) = 0$ for k sufficiently large. Thus, the set $\mathcal{W}_{p^+}$ of Eq. (3.8) is

$$\mathcal{W}_{p^+} = \{J \in \mathcal{J} \mid J^+ \leq J\}. \tag{3.13}$$

By specializing to the case $p = p^+$ the result of Prop. 3.4, we obtain the following proposition, which makes a stronger assertion than Prop. 3.4(a), namely that J^+ is the largest solution of Bellman's equation within $\mathcal{J}$ (rather than the smallest solution within $\mathcal{W}_{p^+}$).

Proposition 3.5:

(a) J^+ is the largest solution of the Bellman equation (3.9) within $\mathcal{J}$, i.e., if $\tilde{J} \in \mathcal{J}$ is a solution of Bellman's equation, then $\tilde{J} \leq J^+$.

(b) (*VI Convergence*) If $\{J_k\}$ is the sequence generated by the VI algorithm (1.4) starting with some $J_0 \in \mathcal{J}$ with $J_0 \geq J^+$, then $J_k \to J^+$.

(c) (*Optimality Condition*) If μ^+ is a terminating stationary policy and

$$\mu^+(x) \in \arg\min_{u\in U(x)} \big\{g(x,u) + J^+\big(f(x,u)\big)\big\}, \qquad \forall\, x \in X, \tag{3.14}$$

then μ^+ is optimal over the set of terminating policies. Conversely, if μ^+ is optimal within the set of terminating policies, then it satisfies the preceding condition (3.14).

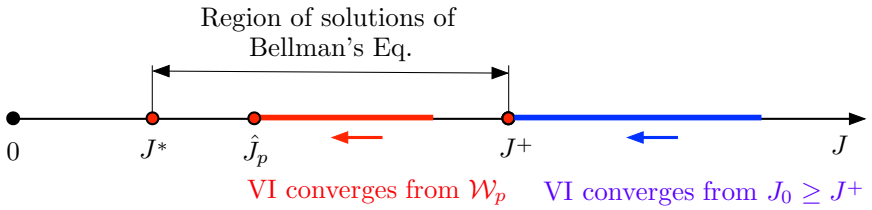

Figure 3.1 Illustration of the solutions of Bellman's equation. The smallest and the largest solutions are J^* and J^+, respectively. The VI algorithm converges to J^+ starting from any $J_0 \in \mathcal{J}$ with $J_0 \geq J^+$, and it converges to $\hat{J}_p$ starting from any $J_0 \in \mathcal{W}_p$.

Proof: In view of Prop. 3.4 and the expression (3.13) for $\mathcal{W}_{p^+}$, we only need to show that $\tilde{J} \leq J^+$ for every solution $\tilde{J} \in \mathcal{J}$ of Bellman's equation. Indeed, let $\tilde{J}$ be such a solution. We have $\tilde{J}(x_0) \leq J^+(x_0)$ for all x_0 with $J^+(x_0) = \infty$, so in order to show that $\tilde{J} \leq J^+$, it will suffice to show that for every (π, x_0) with $\pi \in \Pi^+_{x_0}$, we have $\tilde{J}(x_0) \leq J_\pi(x_0)$. Indeed, consider (π, x_0) with $\pi \in \Pi^+_{x_0}$, and let $\{x_0, \ldots, x_k, t\}$ be the terminating state sequence generated starting from x_0 and using π. Since $\tilde{J}$ solves Bellman's equation, we have

$$\tilde{J}(x_m) \leq g\big(x_m, \mu_m(x_m)\big) + \tilde{J}(x_{m+1}), \qquad m = 0, \ldots, k-1,$$

$$\tilde{J}(x_k) \leq g\big(x_k, \mu_k(x_k)\big).$$

By adding these relations, we obtain

$$\tilde{J}(x_0) \leq \sum_{m=0}^{k} g\big(x_m, \mu_m(x_m)\big) = J_\pi(x_0), \qquad \forall\ (\pi, x_0) \text{ with } \pi \in \Pi^+_{x_0},$$

and by taking the infimum of the right side over $\pi \in \Pi^+_{x_0}$, we obtain $\tilde{J}(x_0) \leq J^+(x_0)$. **Q.E.D.**

We illustrate Props. 3.4 and 3.5 in Figs. 3.1 and 3.2. In particular, each forcing function p delineates the set of initial functions $\mathcal{W}_p$ from which VI converges to $\hat{J}_p$. The function $\hat{J}_p$ is the minimal element of $\mathcal{W}_p$. Moreover, we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$, in view of the VI convergence result of Prop. 3.4(b).

Note that Prop. 3.5(b) implies that VI converges to J^+ starting from the particular initial condition

$$J_0(x) = \begin{cases} 0 & \text{if } x = t, \\ \infty & \text{if } x \neq t. \end{cases} \tag{3.15}$$

For this choice of J_0, the value $J_k(x)$ generated by VI is the optimal cost that can be achieved starting from x subject to the constraint that t is reached in k steps or less.

Suppose now that the set of terminating policies is sufficient in the sense that it can achieve the same optimal cost as the set of all policies, i.e., $J^+ = J^*$. Then, from Prop. 3.5, it follows that J^* is the unique solution of Bellman's equation within $\mathcal{J}$, and the VI algorithm converges to J^* from above, i.e., starting from any $J_0 \in \mathcal{J}$ with $J_0 \geq J^*$. Under additional conditions, such as finiteness of $U(x)$ for all $x \in X$ [cf. Prop. 2.1(d)], VI converges to J^* starting from any $J_0 \in \mathcal{J}$.

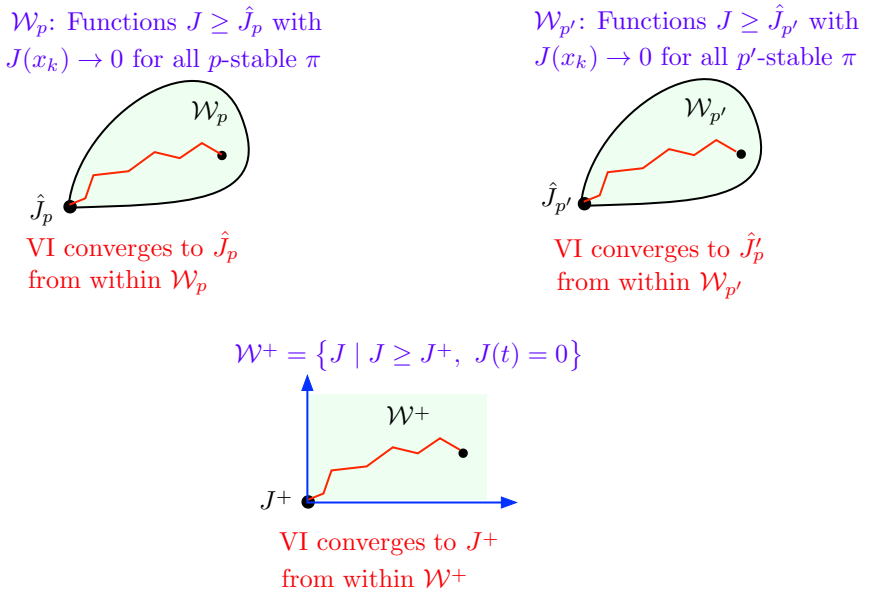

Figure 3.2 Illustration of the VI convergence results of Prop. 3.4 and 3.5. Each p defines the set of initial functions $\mathcal{W}_p$ from which VI converges to $\hat{J}_p$ from above. For two forcing functions p and p', we have $\mathcal{W}_p \cap \mathcal{W}_{p'} = \emptyset$ if $\hat{J}_p \neq \hat{J}_{p'}$.

Examples of problems where terminating policies are sufficient include linear-quadratic problems under the classical conditions of controllability and observability (cf. Example 1.1), and finite-node deterministic shortest path problems with all cycles having positive length. Note that in the former case, despite the fact $J^+ = J^*$, there is no optimal terminating policy, since the only optimal policy is a linear policy that drives the system to the origin asymptotically, but not in finite time.

Let us illustrate the results of this section with two examples.

Example 3.1 (Minimum Energy Stable Control of Linear Systems)

Consider the linear-quadratic problem of Example 1.1. We assume that the pair (A, B) is stabilizable. However, we are making no assumptions on the state weighting matrix Q other than positive semidefiniteness, so the detectability assumption may not be satisfied. This includes the case $Q = 0$, when $J^*(x) \equiv 0$. In this case an optimal policy is $\mu^*(x) \equiv 0$, which may not be stable, yet the problem of finding a stable policy that minimizes the "control energy" (a cost that is quadratic on the control with no penalty on the state) among all stable policies is meaningful.

We consider the forcing function

$$p(x) = \|x\|^2,$$

so the p-δ-perturbed problem satisfies the detectability condition and from classical results, $\hat{J}_{p,\delta}$ is a positive definite quadratic function $x'P_\delta x$, where P_δ is the unique solution of the δ-perturbed Riccati equation

$$P_\delta = A'\big(P_\delta - P_\delta B(B'P_\delta B + R)^{-1}B'P_\delta\big)A + Q + \delta I, \tag{3.16}$$

within the class of positive semidefinite matrices. By Prop. 3.1, we have $\hat{J}_p(x) = x'\hat{P}x$, where $\hat{P} = \lim_{\delta \downarrow 0} P_\delta$ is positive semidefinite, and solves the (unperturbed) Riccati equation

$$P = A'\big(P - PB(B'PB + R)^{-1}B'P\big)A + Q.$$

Moreover, by Prop. 3.4(a), $\hat{P}$ is the largest solution among positive semidefinite matrices, since all positive semidefinite quadratic functions belong to the set S_p of Eq. (3.3). By Prop. 3.4(c), any stable stationary policy $\hat{\mu}$ that is optimal among the set of stable policies must satisfy the optimality condition

$$\hat{\mu}(x) \in \arg\min_{u \in \Re^m} \big\{u'Ru + (Ax + Bu)'\hat{P}(Ax + Bu)\big\}, \qquad \forall\, x \in \Re^n,$$

[cf. Eq. (3.10)], or equivalently, by setting the gradient of the minimized expression to 0,

$$(R + B'\hat{P}B)\hat{\mu}(x) = -B'\hat{P}Ax. \tag{3.17}$$

We may solve Eq. (3.17), and check if any of its solutions $\hat{\mu}$ is p-stable; if this is so, $\hat{\mu}$ is optimal within the class of p-stable policies. Note, however, that in the absence of additional conditions, it is possible that some policies $\hat{\mu}$ that solve Eq. (3.17) are p-unstable.

In the case where the pair (A, B) is not stabilizable, the p-δ-perturbed cost function $\hat{J}_{p,\delta}$ need not be real-valued, and the δ-perturbed Riccati equation (3.16) may not have any solution (consider for example the case where $n = 1$, $A = 2$, $B = 0$, and $Q = R = 1$). Then, Prop. 3.5 still applies, but the preceding analytical approach needs to be modified.

As noted earlier, the Bellman equation may have multiple solutions corresponding to different forcing functions p, with each solution being unique within the corresponding set $\mathcal{W}_p$ of Eq. (3.8), consistently with Prop. 3.4(a). The following is an illustrative example.

Example 3.2 (An Optimal Stopping Problem)

Consider an optimal stopping problem where the state space X is $\Re^n$. We identify the destination with the origin of $\Re^n$, i.e., $t = 0$. At each $x \neq 0$, we may either stop (move to the origin) at a cost $c > 0$, or move to state γx at cost $\|x\|$, where γ is a scalar with $0 < \gamma < 1$; see Fig. 3.3.† Thus the Bellman equation has the form

$$J(x) = \begin{cases} \min\big\{c,\, \|x\| + J(\gamma x)\big\} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases} \tag{3.18}$$

† In this example, the salient feature of the policy that never stops is that it drives the system asymptotically to the destination according to an equation of the form $x_{k+1} = f(x_k)$, where f is a contraction mapping. The example admits generalization to the broader class of optimal stopping problems where the policy that never stops has this property. For simplicity in illustrating our main point, we consider here the special case where $f(x) = \gamma x$ with $\gamma \in (0, 1)$.

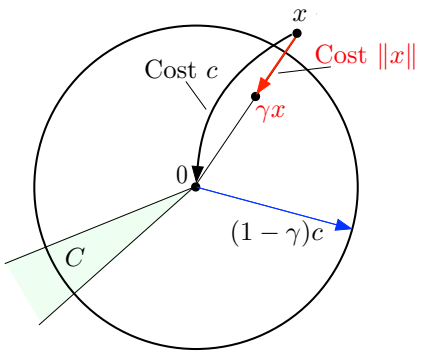

Figure 3.3 Illustration of the stopping problem of Example 3.2. The optimal policy is to stop outside the sphere of radius $(1-\gamma)c$ and to continue otherwise. Each cone C of the state space defines a different solution $\hat{J}_p$ of Bellman's equation, with $\hat{J}_p(x) = c$ for all nonzero $x \in C$, and a corresponding region of convergence of the VI algorithm.

Let us consider first the forcing function

$$p(x) = \|x\|.$$

Then it can be verified that all policies are p-stable. We have

$$J^*(x) = \hat{J}_p(x) = \min\left\{c,\ \frac{1}{1-\gamma}\|x\|\right\}, \qquad \forall\ x \in \Re^n,$$

and the optimal cost function of the corresponding p-δ-perturbed problem is

$$\hat{J}_{p,\delta}(x) = \min\left\{c + \delta\|x\|,\ \frac{1+\delta}{1-\gamma}\|x\|\right\}, \qquad \forall\ x \in \Re^n.$$

Here the set S_p of Eq. (3.3) is given by

$$S_p = \Big\{J \in \mathcal{J} \mid \lim_{x\to 0} J(x) = 0\Big\},$$

and the corresponding set $\mathcal{W}_p$ of Eq. (3.8) is given by

$$\mathcal{W}_p = \Big\{J \in \mathcal{J} \mid J^* \leq J,\ \lim_{x\to 0} J(x) = 0\Big\}.$$

Let us consider next the forcing function

$$p^+(x) = \begin{cases} 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

Then the p^+-stable policies are the terminating policies. Since stopping at some time and incurring the cost c is a requirement for a terminating policy, it follows that the optimal p^+-stable policy is to stop as soon as possible, i.e., stop at every state. The corresponding restricted optimal cost function is

$$J^+(x) = \begin{cases} c & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

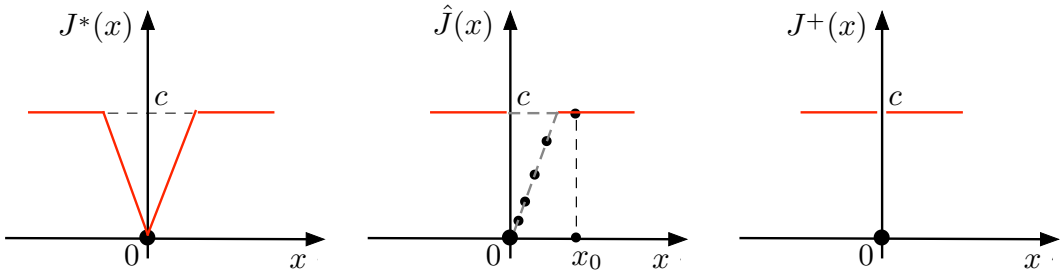

Figure 3.4 Illustration of three solutions of Bellman's equation in the one-dimensional case ($n = 1$) of the stopping problem of Example 3.2. The solution in the middle is specified by a scalar $x_0 > 0$, and has the form

$$\hat{J}(x) = \begin{cases} 0 & \text{if } x = 0, \\ \frac{1}{1-\gamma}|x| & \text{if } 0 < x < (1-\gamma)c \text{ and } x = \gamma^k x_0 \text{ for some } k \geq 0, \\ c & \text{otherwise.} \end{cases}$$

The optimal cost function of the corresponding p^+-δ-perturbed problem is

$$\hat{J}_{p^+,\delta}(x) = \begin{cases} c + \delta & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

since in the p^+-δ-perturbed problem it is again optimal to stop as soon as possible, at cost $c + \delta$. Here the set S_{p^+} is equal to $\mathcal{J}$, and the corresponding set $\mathcal{W}_{p^+}$ is equal to $\big\{J \in \mathcal{J} \mid J^+ \leq J\big\}$.

However, there are infinitely many additional solutions of Bellman's equation between the largest and smallest solutions J^* and J^+. For example, when $n > 1$, functions $J \in \mathcal{J}$ such that $J(x) = J^*(x)$ for x in some cone and $J(x) = J^+(x)$ for x in the complementary cone are solutions; see Fig. 3.3. There is also a corresponding infinite number of regions of convergence $\mathcal{W}_p$ of VI. Also VI converges to J^* starting from any J_0 with $0 \leq J_0 \leq J^*$ [cf. Prop. 2.1(d)]. Figure 3.4 illustrates additional solutions of Bellman's equation of a different character.

4. POLICY ITERATION METHODS

Generally, the standard PI algorithm (1.5), (1.6) produces unclear results under our assumptions. As an illustration, in the stopping problem of Example 3.2, if PI is started with the policy that stops at every state, it repeats that policy, and this policy is not optimal within the class of p-stable policies with respect to the forcing function $p(x) = \|x\|$. The following example provides an instance where the PI algorithm may converge to either an optimal or a strictly suboptimal policy.

Example 4.1 (Counterexample for PI)

Consider the case $X = \{0, 1\}$, $U(0) = U(1) = \{0, 1\}$, and the destination is $t = 0$. Let also

$$f(x,u) = \begin{cases} 0 & \text{if } u = 0, \\ x & \text{if } u = 1, \end{cases} \qquad g(x,u) = \begin{cases} 1 & \text{if } u = 0,\ x = 1, \\ 0 & \text{if } u = 1 \text{ or } x = 0. \end{cases}$$

This is a shortest path problem where the control $u = 0$ moves the state from $x = 1$ to $x = 0$ (the destination) at cost 1, while the control $u = 1$ keeps the state unchanged at cost 0. The policy μ^* that keeps the state unchanged is the only optimal policy, with $J_{\mu^*}(x) = J^*(x) = 0$ for both states x. However, under any forcing function p with $p(1) > 0$, the policy $\hat{\mu}$, which moves from state 1 to 0, is the only p-stable policy, and we have $J_{\hat{\mu}}(1) = \hat{J}_p(1) = 1$. The standard PI algorithm (1.5), (1.6) if started with μ^* will repeat μ^*. If this algorithm is started with $\hat{\mu}$, it may generate μ^* or it may repeat $\hat{\mu}$, depending on how the policy improvement iteration is implemented. The reason is that for both x we have

$$\hat{\mu}(x) \in \arg\min_{u \in \{0,1\}} \Big\{ g(x,u) + \hat{J}_p\big(f(x,u)\big) \Big\},$$

as can be verified with a straightforward calculation. Thus a rule for breaking a tie in the policy improvement operation is needed, but such a rule may not be obvious in general.

Motivated by the preceding example, we consider several types of PI method that bypass the difficulty above either through assumptions or through modifications. We first consider a case where the PI algorithm is reliable. This is the case where the terminating policies are sufficient, in the sense that $J^+ = J^*$.

4.1. Policy Iteration for the Case $J^* = J^+$

The PI algorithm starts with a stationary policy μ^0, and generates a sequence of stationary policies $\{\mu^k\}$ via a sequence of policy evaluations to obtain J_{μ^k} from the equation

$$J_{\mu^k}(x) = g\big(x, \mu^k(x)\big) + J_{\mu^k}\big(f\big(x, \mu^k(x)\big)\big), \qquad x \in X, \tag{4.1}$$

interleaved with policy improvements to obtain μ^{k+1} from J_{μ^k} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{4.2}$$

We implicitly assume here that the minimum in Eq. (4.2) is attained for each $x \in X$, which is true under some compactness condition on either $U(x)$ or the level sets of the function $g(x,\cdot) + J_k\big(f(x,\cdot)\big)$, or both.

Proposition 4.1: (Convergence of PI) Assume that $J^* = J^+$. Then the sequence $\{J_{\mu^k}\}$ generated by the PI algorithm (4.1), (4.2), satisfies $J_{\mu^k}(x) \downarrow J^*(x)$ for all $x \in X$.

Proof: If μ is a stationary policy and $\bar{\mu}$ satisfies the policy improvement equation

$$\bar{\mu}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\}, \qquad x \in X,$$

[cf. Eq. (4.2)], we have for all $x \in X$,

$$J_\mu(x) = g\big(x,\mu(x)\big) + J_\mu\big(f\big(x,\mu(x)\big)\big) \geq \min_{u \in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} = g\big(x,\bar{\mu}(x)\big) + J_\mu\big(f\big(x,\bar{\mu}(x)\big)\big), \quad (4.3)$$

where the first equality follows from by Prop. 2.1(b), and the second equality follows from the definition of $\bar{\mu}$. Repeatedly applying this relation, we see that the sequence $\big\{\tilde{J}_k(x_0)\big\}$ defined by

$$\tilde{J}_k(x_0) = J_\mu(x_k) + \sum_{m=0}^{k-1} g\big(x_m, \bar{\mu}(x_m)\big), \qquad k = 1, 2, \ldots,$$

is monotonically nonincreasing, where $\{x_k\}$ is the sequence generated starting from x_0 and using μ. Moreover, from Eq. (4.3) we have

$$J_\mu(x_0) \geq \min_{u \in U(x_0)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} = \tilde{J}_1(x_0) \geq \tilde{J}_k(x_0),$$

for all k. This implies that

$$J_\mu(x_0) \geq \min_{u \in U(x_0)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} \geq \lim_{k \to \infty} \tilde{J}_k(x_0) \geq \lim_{k \to \infty} \sum_{m=0}^{k-1} g\big(x_m, \bar{\mu}(x_m)\big) = J_{\bar{\mu}}(x_0),$$

where the last inequality follows since $J_\mu \geq 0$. In conclusion, we have

$$J_\mu(x) \geq \inf_{u \in U(x)} \big\{ g(x,u) + J_\mu\big(f(x,u)\big) \big\} \geq J_{\bar{\mu}}(x), \qquad x \in X. \quad (4.4)$$

Using μ^k and μ^{k+1} in place of μ and $\bar{\mu}$, we see that the sequence $\{J_{\mu^k}\}$ generated by PI converges monotonically to some function $J_\infty \in E^+(X)$, i.e., $J_{\mu^k} \downarrow J_\infty$. Moreover, from Eq. (4.4) we have

$$J_\infty(x) \geq \inf_{u \in U(x)} \big\{ g(x,u) + J_\infty\big(f(x,u)\big) \big\}, \qquad x \in X,$$

as well as

$$g(x,u) + J_{\mu_k}\big(f(x,u)\big) \geq J_\infty(x), \qquad x \in X,\ u \in U(x).$$

We now take the limit in the second relation as $k \to \infty$, then the infimum over $u \in U(x)$, and then combine with the first relation, to obtain

$$J_\infty(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J_\infty\big(f(x,u)\big) \big\}, \qquad x \in X.$$

Thus J_∞ is a solution of Bellman's equation, satisfying $J_\infty \geq J^*$ (since $J_{\mu^k} \geq J^*$ for all k) and $J_\infty \in \mathcal{J}$ (since $J_{\mu^k} \in \mathcal{J}$), so by Prop. 3.5(a), it must satisfy $J_\infty = J^*$. **Q.E.D.**

4.2. A Perturbed Version of Policy Iteration

We now consider a PI algorithm that does not require the condition $J^* = J^+$. We will provide a version of the PI algorithm that uses the forcing function p and generates a sequence $\{\mu^k\}$ of p-stable policies such that $J_{\mu^k} \to \hat{J}_p$. In this section, the forcing function p is kept fixed, and to simplify notation, we abbreviate $J_{\mu,p,\delta}$ with $J_{\mu,\delta}$. The following assumption requires that the algorithm generates p-stable policies exclusively, which can be quite restrictive. For example it is not satisfied for the problem of Example 4.1.

Assumption 4.1: For each $\delta > 0$ there exists at least one p-stable stationary policy μ such that $J_{\mu,\delta} \in S_p$. Moreover, given a p-stable stationary policy μ and a scalar $\delta > 0$, every stationary policy $\overline{\mu}$ such that

$$\overline{\mu}(x) \in \arg\min_{u \in U(x)} \Big\{ g(x,u) + J_{\mu,\delta}\big(f(x,u)\big) \Big\}, \qquad \forall\ x \in X, \tag{4.5}$$

is p-stable, and at least one such policy exists.

The perturbed version of the PI algorithm is defined as follows. Let $\{\delta_k\}$ be a positive sequence with $\delta_k \downarrow 0$, and let μ^0 be a p-stable policy that satisfies $J_{\mu^0,\delta_0} \in S_p$. One possibility is that μ^0 is an optimal policy for the δ_0-perturbed problem (cf. the discussion preceding Prop. 3.3). At iteration k, we have a p-stable policy μ^k, and we generate a p-stable policy μ^{k+1} according to

$$\mu^{k+1}(x) \in \arg\min_{u \in U(x)} \big\{ g(x,u) + J_{\mu^k,\delta_k}\big(f(x,u)\big) \big\}, \qquad x \in X. \tag{4.6}$$

Note that under Assumption 4.1, the algorithm is well-defined, and is guaranteed to generate a sequence of p-stable stationary policies.

We will use for all policies μ and scalars $\delta > 0$ the mappings $T_\mu : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ and $T_{\mu,\delta} : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ by

$$(T_\mu J)(x) = g\big(x,\mu(x)\big) + J\big(f(x,\mu(x))\big), \qquad x \in X,$$

$$(T_{\mu,\delta} J)(x) = g\big(x,\mu(x)\big) + \delta p(x) + J\big(f(x,\mu(x))\big), \qquad x \in X,$$

and the mapping $T : \mathcal{E}^+(X) \mapsto \mathcal{E}^+(X)$ given by

$$(TJ)(x) = \inf_{u \in U(x)} \big\{ g(x,u) + J\big(f(x,u)\big) \big\}, \qquad x \in X.$$

For any integer $m \geq 1$, we denote by T_μ^m and $T_{\mu,\delta}^m$ the m-fold compositions of the mappings T_μ and $T_{\mu,\delta}$, respectively. We have the following proposition.

Proposition 4.2: Let Assumption 4.1 hold. Then for a sequence of p-stable policies $\{\mu^k\}$ generated by the perturbed PI algorithm (4.6), we have $J_{\mu^k,\delta_k} \downarrow \hat{J}_p$ and $J_{\mu^k} \to \hat{J}_p$.

Proof: The algorithm definition (4.6) implies that for all integer $m \geq 1$ we have for all $x_0 \in X$,

$$J_{\mu^k,\delta_k}(x_0) \geq (TJ_{\mu^k,\delta_k})(x_0) + \delta_k p(x_0) = (T_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(x_0) \geq (T^m_{\mu^{k+1},\delta_k} J_{\mu^k,\delta_k})(x_0) \geq (T^m_{\mu^{k+1},\delta_k} \bar{J})(x_0),$$

where $\bar{J}$ is the identically zero function [$\bar{J}(x) \equiv 0$]. From this relation we obtain

$$J_{\mu^k,\delta_k}(x_0) \geq \lim_{m\to\infty} (T^m_{\mu^{k+1},\delta_k} \bar{J})(x_0) = \lim_{m\to\infty} \left\{ \sum_{\ell=0}^{m-1} \big(g(x_\ell, \mu^{k+1}(x_\ell)) + \delta_k p(x_\ell) \big) \right\} \geq J_{\mu^{k+1},\delta_{k+1}}(x_0),$$

as well as

$$J_{\mu^k,\delta_k}(x_0) \geq (TJ_{\mu^k,\delta_k})(x_0) + \delta_k p(x_0) \geq J_{\mu^{k+1},\delta_{k+1}}(x_0).$$

It follows that $\{J_{\mu^k,\delta_k}\}$ is monotonically nonincreasing, so that $J_{\mu^k,\delta_k} \downarrow J_\infty$ for some J_∞, and

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = J_\infty. \tag{4.7}$$

We also have, using the fact $J_\infty \leq J_{\mu^k,\delta_k}$,

$$\begin{aligned}
\inf_{u\in U(x)} \big\{ g(x,u) + J_\infty\big(f(x,u)\big) \big\} &\leq \lim_{k\to\infty} \inf_{u\in U(x)} \big\{ g(x,u) + J_{\mu^k,\delta_k}\big(f(x,u)\big) \big\} \\
&\leq \inf_{u\in U(x)} \lim_{k\to\infty} \big\{ g(x,u) + J_{\mu^k,\delta_k}\big(f(x,u)\big) \big\} \\
&= \inf_{u\in U(x)} \Big\{ g(x,u) + \lim_{k\to\infty} J_{\mu^k,\delta_k}\big(f(x,u)\big) \Big\} \\
&= \inf_{u\in U(x)} \big\{ g(x,u) + J_\infty\big(f(x,u)\big) \big\}.
\end{aligned}$$

Thus equality holds throughout above, so that

$$\lim_{k\to\infty} TJ_{\mu^k,\delta_k} = TJ_\infty.$$

Combining this with Eq. (4.7), we obtain $J_\infty = TJ_\infty$, i.e., J_∞ solves Bellman's equation. We also note that $J_\infty \leq J_{\mu^0,\delta_0}$ and that $J_{\mu^0,\delta_0} \in S_p$ by assumption, so that $J_\infty \in S_p$. By Prop. 3.4(a), it follows that $J_\infty = \hat{J}_p$. **Q.E.D.**

Note that despite the fact $J_{\mu^k} \to \hat{J}_p$, the generated sequence $\{\mu^k\}$ may exhibit some serious pathologies in the limit. In particular, if U is a metric space and $\{\mu^k\}_{\mathcal{K}}$ is a subsequence of policies that converges to some $\bar{\mu}$, in the sense that

$$\lim_{k\to\infty,\, k\in\mathcal{K}} \mu^k(x) = \bar{\mu}(x), \qquad \forall\ x \in X,$$

it does not follow that $\bar{\mu}$ is p-stable. In fact it is possible to construct examples where the generated sequence of p-stable policies $\{\mu^k\}$ satisfies $\lim_{k\to\infty} J_{\mu^k} = \hat{J}_p = J^*$, yet $\{\mu^k\}$ may converge to a p-unstable policy whose cost function is strictly larger than $\hat{J}_p$. Example 2.1 of the paper [BeY16] provides an instance of a stochastic shortest path problem with two states, in addition to the termination state, where this occurs.

4.3. An Optimistic Policy Iteration Method

Let us consider an optimistic variant of PI, where policies are evaluated inexactly, with a finite number of VIs. We use a fixed forcing function p. We will show that the algorithm can be used to compute $\hat{J}_p$, the restricted optimal cost function over the p-stable policies. The algorithm generates a sequence $\{J_k, \mu^k\}$ according to

$$T_{\mu^k} J_k = TJ_k, \qquad J_{k+1} = T_{\mu^k}^{m_k} J_k, \qquad k = 0, 1, \ldots, \tag{4.8}$$

where m_k is a positive integer for each k. We assume that a policy μ^k satisfying $T_{\mu^k} J_k = TJ_k$ can be found for all k, but it need not be p-stable. However, the algorithm requires that

$$J_0 \in \mathcal{J}, \qquad J_0 \geq TJ_0, \qquad J_0 \in \mathcal{W}_p. \tag{4.9}$$

This may be a restrictive assumption. We have the following proposition.

Proposition 4.3: (Convergence of Optimistic PI) Assume that there exists at least one p-stable policy $\pi \in \Pi_p$, and that J_0 satisfies Eq. (4.9). Then a sequence $\{J_k\}$ generated by the optimistic PI algorithm (4.8) belongs to $\mathcal{W}_p$ and satisfies $J_k \downarrow \hat{J}_p$.

Proof: Since $J_0 \geq \hat{J}_p$ and $\hat{J}_p = T\hat{J}_p$ [cf. Prop. 3.5(a)], all operations on any of the functions J_k with T_{μ^k} or T maintain the inequality $J_k \geq \hat{J}_p$ for all k, so that $J_k \in \mathcal{W}_p$ for all k. Also the conditions $J_0 \geq TJ_0$ and $T_{\mu^k} J_k = TJ_k$ imply that

$$J_0 = J_1 \geq T_{\mu^0}^{m_0+1} J_0 = T_{\mu^0} J_1 \geq TJ_1 = T_{\mu^1} J_1 \geq \cdots \geq J_2, \tag{4.10}$$

and continuing similarly,

$$J_k \geq TJ_k \geq J_{k+1}, \qquad k = 0, 1, \ldots. \tag{4.11}$$

Thus $J_k \downarrow J_\infty$ for some J_∞, which must satisfy $J_\infty \geq \hat{J}_p$, and hence belong to $\mathcal{W}_p$. By taking limit as $k \to \infty$ in Eq. (4.11) and using an argument similar to the one in the proof of Prop. 4.2, it follows that $J_\infty = TJ_\infty$. By Prop. 3.5(a), this implies that $J_\infty \leq \hat{J}_p$. Together with the inequality $J_\infty \geq \hat{J}_p$ shown earlier, this proves that $J_\infty = \hat{J}_p$. **Q.E.D.**

As an example, for the shortest path problem of Example 4.1, the reader may verify that for the case where $p(x) = 1$, for $x = 1$, the optimistic PI algorithm converges in a single iteration to

$$\hat{J}_p(x) = \begin{cases} 1 & \text{if } x = 1, \\ 0 & \text{if } x = 0, \end{cases}$$

provided that $J_0 \in \mathcal{W}_p = \big\{J \mid J(1) \geq 1,\ J(0) = 0\big\}$. For other starting functions J_0, the algorithm converges in a single iteration to the function

$$J_\infty(1) = \min\big\{1,\, J_0(1)\big\}, \qquad J_\infty(0) = 0.$$

All functions J_∞ of the form above are solutions of Bellman's equation, but only $\hat{J}_p$ is restricted optimal.

5. CONCLUDING REMARKS

We have considered deterministic optimal control problems with a cost-free and absorbing destination under general assumptions, which include arbitrary state and control spaces, and a Bellman's equation with multiple solutions. Within this context, we have used perturbations of the cost per stage and the ideas of semicontractive DP as a means to connect classical issues of stability and optimization. In particular, we have shown that the restricted optimal cost function over just the stable policies is a solution of Bellman's equation, and that versions of the VI and PI algorithm are attracted to it. Moreover, the restricted optimal cost J^+ over the "fastest" policies (the ones that terminate) is the largest solution of Bellman's equation. The generality of our framework makes our results a convenient starting point for analysis of related problems and methods, involving additional assumptions, and/or cost function approximation and state space discretization.

An interesting open question is how to discretize continuous-spaces problems to solve Bellman's equation numerically. As an example, consider the linear-quadratic problem of Example 3.1. Any reasonable discretization of this problem is a finite-state (deterministic or stochastic) shortest path problem, whose Bellman equation has a unique solution that approximates the solution J^* of the continuous-spaces problem, while missing entirely the solution J^+. The same is true for the optimal stopping problem of Example 3.2. In such cases, one may discretize a δ-perturbed version of the problem, which is better behaved, and use a small value of δ to obtain an approximation to J^+. However, the limiting issues as $\delta \downarrow 0$ remain to be explored.

6. REFERENCES

[AnM07] Anderson, B. D., and Moore, J. B., 2007. Optimal Control: Linear Quadratic Methods, Courier Corporation.

[BeS78] Bertsekas, D. P., and Shreve, S. E., 1978. Stochastic Optimal Control: The Discrete Time Case. New York: Academic Press; may be downloaded from http://web.mit.edu/dimitrib/www/home.html.

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

[BeY16] Bertsekas, D. P., and Yu, H., 2016. "Stochastic Shortest Path Problems Under Weak Conditions," Lab. for Information and Decision Systems Report LIDS-2909.

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

[Ber13] Bertsekas, D. P., 2013. Abstract Dynamic Programming, Athena Scientific, Belmont, MA; a second edition appeared in 2017 on-line at http://web.mit.edu/dimitrib/www/home.html.

[Ber14] Bertsekas, D. P., 2014. "Robust Shortest Path Planning and Semicontractive Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-2915, MIT, Feb. 2014 (revised Jan. 2015 and June 2016); arXiv preprint arXiv:1608.01670; to appear in Naval Research Logistics.

[Ber15a] Bertsekas, D. P., 2015. "Value and Policy Iteration in Deterministic Optimal Control and Adaptive Dynamic Programming," arXiv preprint arXiv:1507.01026; IEEE Trans. on Neural Networks and Learning Systems, Vol. 28, 2017, pp. 500-509.

[Ber15b] Bertsekas, D. P., 2015. "Regular Policies in Abstract Dynamic Programming," Lab. for Information and Decision Systems Report LIDS-P-3173, MIT; arXiv preprint arXiv:1609.03115; to appear in SIAM J. on Optimization.

[Ber16] Bertsekas, D. P., 2016. "Affine Monotonic and Risk-Sensitive Models in Dynamic Programming", Lab. for Information and Decision Systems Report LIDS-3204, MIT, June 2016; arXiv preprint arXiv:1608.01393.

[Ber17] Bertsekas, D. P., 2017. Dynamic Programming and Optimal Control, Vol. I, 4th edition, Athena Scientific, Belmont, MA.

[Hey14] Heydari, A., 2014. "Stabilizing Value Iteration With and Without Approximation Errors," available at arXiv:1412.5675.

[JiJ14] Jiang, Y., and Jiang, Z. P., 2014. "Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems," IEEE Trans. on Neural Networks and Learning Systems, Vol. 25, pp. 882-893.

[LLL08] Lewis, F. L., Liu, D., and Lendaris, G. G., 2008. Special Issue on Adaptive Dynamic Programming and Reinforcement Learning in Feedback Control, IEEE Trans. on Systems, Man, and Cybernetics, Part B, Vol. 38, No. 4.

[LeL13] Lewis, F. L., and Liu, D., (Eds), 2013. Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, Wiley, Hoboken, N. J.

[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, pp. 779-789.

[Kle68] Kleinman, D. L., 1968. "On an Iterative Technique for Riccati Equation Computations," IEEE Trans. Aut. Control, Vol. AC-13, pp. 114-115.

[Kuc72] Kucera, V., 1972. "The Discrete Riccati Equation of Optimal Control," Kybernetika, Vol. 8, pp. 430-447.

[Kuc73] Kucera, V., 1973. "A Review of the Matrix Riccati Equation," Kybernetika, Vol. 9, pp. 42-61.

[LaR95] Lancaster, P., and Rodman, L., 1995. Algebraic Riccati Equations, Clarendon Press, Oxford, UK.

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

[SBP04] Si, J., Barto, A., Powell, W., and Wunsch, W., (Eds.), 2004. Learning and Approximate Dynamic Programming, IEEE Press, N. Y.

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

[VVL13] Vrabie, D., Vamvoudakis, K. G., and Lewis, F. L., 2013. Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles, The Institution of Engineering and Technology, London.

[Wil71] Willems, J., 1971. "Least Squares Stationary Optimal Control and the Algebraic Riccati Equation," IEEE Trans. on Automatic Control, Vol. 16, pp. 621-634.