A Mixed Value and Policy Iteration Method for Stochastic Control with Universally Measurable Policies

Huizhen Yu* Dimitri P. Bertsekas†

Abstract

We consider the stochastic control model with Borel spaces and universally measurable policies. For this model the standard policy iteration is known to have difficult measurability issues and cannot be carried out in general. We present a mixed value and policy iteration method that circumvents this difficulty. The method allows the use of stationary policies in computing the optimal cost function, in a manner that resembles policy iteration. It can also be used to address similar difficulties of policy iteration in the context of upper and lower semicontinuous models. We analyze the convergence of the method in infinite horizon total cost problems, for the discounted case where the one-stage costs are bounded, and for the undiscounted case where the one-stage costs are nonpositive or nonnegative.

For the undiscounted total cost problems with nonnegative one-stage costs, we also give a new convergence theorem for value iteration, which shows that value iteration converges whenever it is initialized with a function that is above the optimal cost function and yet bounded by a multiple of the optimal cost function. This condition resembles Whittle's bridging condition and is partly motivated by it. The theorem is also partly motivated by a result of Maitra and Sudderth, which showed that value iteration, when initialized with the constant function zero, could require a transfinite number of iterations to converge. We use the new convergence theorem for value iteration to establish the convergence of our mixed value and policy iteration method for the nonnegative cost models.

*Lab. for Information and Decision Systems, M.I.T. janey_yu@mit.edu
†Lab. for Information and Decision Systems, M.I.T. dimitrib@mit.edu

Contents

1 Introduction

We consider discrete-time stochastic control problems with additive one-stage costs in a general framework that involves Borel state and control spaces and universally measurable policies. Historically, our framework traces back to the pioneering work on dynamic programming (DP) in Borel spaces by Blackwell [11, 12, 13, 14] and Strauch [47], which was developed further, along several directions, through a sequence of subsequent works. These include: the books by Hinderer [29], and Dynkin and Yushkevich [20], which considered a framework based on Borel measurable policies and the notion of almost-surely ϵ-optimality; the work of Maitra [32], Furukawa [25], Freedman [24] and Schäl [40], as well as Dynkin and Yushkevich [20], which studied Borel measurable policies and semicontinuous models; the work of Blackwell, Freedman and Orkin [16], which introduced a formulation involving analytic sets and analytically measurable policies; and the work of Shreve and Bertsekas [44, 45], and Bertsekas and Shreve [7, Part II], which considered universally measurable policies. Further research on alternative frameworks suitable for DP include: Shreve [41] and Bertsekas and Shreve [7, Part II] on C-sets and limit-measurable policies, Blackwell [15] on Borel-programmable functions and Shreve [43] on Borel-approachable functions. We refer to the monograph [7] and the papers [44, 42] for a discussion of the differences between these frameworks, along with a review of the literature for the early period of the subject. We refer to the books [38, 27, 28, 2, 23] and the survey paper [21] for more recent accounts and extensive references about the significant development of the field since then. In this paper, we will focus on the universally measurable policies framework of [44, 45, 7], and three types of classical infinite horizon total cost problems: the discounted case where the one-stage costs are bounded, and the undiscounted case where the one-stage costs are all nonpositive or all nonnegative.

The early works of Blackwell and Strauch showed that taking Borel measurable policies as the only admissible policies does not lead to desirable results that are comparable with the ones available for problems where measurability is not a concern. In particular, a Borel measurable policy need not exist even when the control constraint set is Borel [14]. Moreover, if we restrict attention to Borel measurable policies, there need not exist an everywhere ϵ-optimal policy even in discounted problems [12]. An important step toward a more satisfactory framework was taken by Blackwell, Freedman and Orkin [16]. Studying finite horizon nonnegative reward problems, they introduced an approach based on analytic sets and semi-analytic functions (a family of functions whose level sets are analytic sets), and obtained optimality results for analytically measurable policies (a larger class of policies that includes Borel measurable ones). Their model still does not admit the existence of everywhere optimal policies or the existence of everywhere ϵ-optimal nonrandomized policies among structured families of policies in general. Building upon analytic sets and semi-analytic functions as in [16], a fuller framework was developed in Shreve and Bertsekas [44, 45], Bertsekas and Shreve [7, Part II]. In this framework, the class of admissible policies is enlarged to be the class of universally measurable policies, structural properties of the optimal cost functions are derived, and selection theorems that stem from Jankov-von Neumann's theorem ensure the existence of everywhere ϵ-optimal or optimal policies among structured families of policies (e.g., stationary, Markov or semi-Markov policies), both for finite horizon problems and for infinite horizon problems that we consider.

However, with analytically or universally measurable policies, standard policy iteration has measurability-related difficulties, as noted in [16, p. 940] and [7, p. 232]. The selection of an admissible measurable policy can fail at the policy improvement step because the cost function of an analytically or universally measurable policy need not have the necessary structure for exact or ϵ-exact selection of an improved policy. This causes the policy iteration procedure to break down.

A similar difficulty occurs in upper and lower semicontinuous models. There the selection of a Borel measurable policy at the policy improvement step may fail because the cost function of the current Borel measurable policy does not have adequate semicontinuity structure.

One of the major purposes of this paper is to provide an approach to circumvent the difficulty

just discussed, and to allow stationary policies to be used in computing the optimal cost function, in a manner that resembles policy iteration (even when ϵ-optimal stationary policies do not exist). We refer to our approach as a mixed value and policy iteration method, as it combines characteristics of both value and policy iteration. Algorithmically, compared to standard policy iteration, the main difference of our method is in the policy evaluation phase: instead of computing the costs of a given policy, it solves exactly or approximately an optimal stopping problem defined by a stationary policy of interest and by a stopping cost that is an estimate of the optimal cost. The stopping costs are then adjusted and the procedure is repeated. To avoid measurability issues, we exploit the fact that every universally measurable stationary policy has Borel measurable portions (see Prop. 3.1(b)), and we define the optimal stopping problems accordingly so that the iterative method just mentioned can operate within the family of functions with the desired semi-analytic structure. Another critical feature of our approach results from the optimal-stopping formulation: for convergence, relying on an inherent value iteration character, it is not required that the policies involved improve successively over one another (this is generally impossible within our context). This feature allows us to operate the method with various policies and leads to algorithms of various forms. As a result, we obtain policy iteration-like algorithms if we choose policies in a way analogous to policy improvement, using the Jankov-von Neumann type of selection theorems.

Similarly, for semicontinuous models we exploit the fact that Borel measurable policies have continuous portions (Lusin's Theorem; see e.g., [19]). We use it to specialize our method to produce policy iteration-like algorithms that operate within the desired family of semicontinuous functions.

We establish the convergence of our method under certain initial conditions for the three types of infinite horizon total cost problems we consider. Our convergence results parallel those for standard value iteration for these problems.

The mixed value and policy iteration method of this paper evolved from the enhanced policy iteration algorithmic framework proposed and analyzed in our earlier works for finite-state and control problems [10, 57] and for abstract DP problems [9] under discounted and undiscounted total cost criteria (see also the book accounts of these works in [5, 6]). In the finite-spaces or abstract DP context, measurability is not an issue. Asynchronous distributed computation of the optimal cost function, by model-free stochastic approximation algorithms in certain cases, has been our main motivation for a policy iteration-like method that is convergent without relying strongly on the performance of the policies involved. The method in this paper is based on the same idea and shares many important features with its counterparts in our earlier works, although its form has been modified and extended, in order to overcome the measurability issues in the present general-spaces stochastic control context. By providing a Borel-space counterpart of the method, one of our purposes is also to demonstrate that the mixed value and policy iteration approach is useful for addressing issues of not only computational but also theoretical nature. Of course our method preserves the computational advantages of its predecessors. In particular, it is suitable for asynchronous distributed computation, although we do not discuss this possibility in detail in the present paper.

The convergence analysis of our mixed value and policy iteration method for nonnegative cost models relies on another main result of this paper, which is of independent interest. This is a new convergence theorem for value iteration. It is well-known that for nonnegative cost models, value iteration need not converge to the optimal cost function. Conditions for convergence from below, which involve compactness-type assumptions on the control constraint set, have been given by Bertsekas [3] for a related special case of minimax reachability problems, by Schäl [40] and Bertsekas [4] for cases where measurability issues are not a concern, and by Bertsekas and Shreve [7] for the universally measurable policies framework of this paper. Sufficient conditions have also been studied by Whittle [55, 56].

Our theorem shows that value iteration converges whenever it is initialized with a function that lies above the optimal cost function and yet is bounded by a multiple of the optimal cost function. This condition resembles Whittle's bridging condition [55, 26] and is partly motivated

by it. Whittle's condition, however, delineates a subset of nonnegative cost models in which value iteration converges when initialized with the constant function zero, whereas our theorem holds without model restrictions. In formulating the theorem, we were also partly motivated by a general convergence result of Maitra and Sudderth [33], which showed that starting from the constant function zero, value iteration could require a transfinite number of iterations to converge. Our proof of the new theorem for the convergence of value iteration (in the standard, non-transfinite form) uses, among others, Maitra and Sudderth's result.

Using the new convergence theorem for value iteration, we are also able to show that for certain nonnegative cost models (which include countable-spaces problems with finite optimal costs), convergence of our mixed value and policy iteration method is maintained if the optimal stopping problems involved are solved approximately by solving associated linear programs. This result can be contrasted with the fact that nonnegative cost models in general do not admit a linear programming formulation. It suggests that even when there are no measurability concerns, for the nonnegative cost models, the mixed value and policy iteration approach may provide computationally efficient algorithms that are based on linear programming.

The paper is organized as follows. In Section 2, we provide background. In Section 3, we introduce the mixed value and policy iteration method, and derive various algorithmic versions. We give greater attention to policy iteration-like algorithms, and we discuss their relation with standard policy iteration, as well as the application range of a special algorithm involving Borel measurable policies. In Section 4, we prove convergence results for the proposed method, for discounted problems with bounded one-stage costs and for total cost problems with nonpositive one-stage costs. In Section 5, we consider total cost problems with nonnegative one-stage costs. We first prove the new convergence theorem for value iteration in Section 5.1. We then derive convergence results for the proposed method in Section 5.2. In Section 6, we discuss the applications of our results in semicontinuous models, including the application of the mixed value and policy iteration approach, and a result on the structure of the optimal cost function and optimal policies for nonnegative cost upper semicontinuous models. In Section 7, we conclude the paper with remarks on extensions and future research directions. Appendices A-C collect some related formulations, proofs, and illustrative examples.

2 Background

In this section we describe the stochastic control framework with universally measurable policies. We give a brief summary of basic optimality results for infinite horizon, discounted and undiscounted total cost problems. We then explain the measurability issues that cause standard policy iteration to break down.

2.1 Preliminaries

In this subsection we introduce some concepts and terminologies, including universal σ-algebras, analytic sets and lower semi-analytic functions. We also highlight some properties that are important and provide the basis for the stochastic control framework.

Let us first introduce some notation. For a topological space X, we denote by $\mathscr{B}(X)$ the Borel σ-algebra. Let X and Y be two topological spaces. By a Borel measurable function (or mapping) from X to Y, we mean that the function is measurable from $(X, \mathscr{B}(X))$ to $(Y, \mathscr{B}(Y))$ (i.e., the preimage of any $B \in \mathscr{B}(Y)$ lies in $\mathscr{B}(X)$). Similarly, if $\mathcal{F}$ is a σ-algebra on X, by an $\mathcal{F}$-measurable function from X to Y, we mean that the function is measurable from $(X, \mathcal{F})$ to $(Y, \mathscr{B}(Y))$ (i.e., the preimage of any $B \in \mathscr{B}(Y)$ lies in $\mathcal{F}$). We define likewise $\mathcal{F}$-measurable functions from X' to Y, where X' is a subset of X and the σ-algebra on X' is the trace σ-algebra $\mathcal{F} \cap X' = \{D \cap X' \mid D \in \mathcal{F}\}$.

In this paper we will focus on separable and metrizable topological spaces, and besides the Borel σ-algebra $\mathscr{B}(X)$, we will need to consider σ-algebras on X that are finer than $\mathscr{B}(X)$. The universal σ-algebra on X is defined through the set $\mathscr{P}(X)$ of Borel probability measures on X (i.e., probability measures on $\mathscr{B}(X)$) as follows. A Borel probability measure p can be extended to a probability measure on the σ-algebra $\mathscr{B}_p(X)$ generated by $\mathscr{B}(X)$ and all the subsets of X that have p-outer measure zero, such that the extension agrees with the p-outer measure on $\mathscr{B}_p(X)$. This extension of p is called the *completion of* p [19, Sec. 3.3] and will also be denoted by p. The intersection of all the σ-algebras $\mathscr{B}_p(X)$ for $p \in \mathscr{P}(X)$ is called the *universal σ-algebra* $\mathscr{U}(X)$ [7, Def. 7.18]. Sets in $\mathscr{U}(X)$ and measurable functions on $(X, \mathscr{U}(X))$ are said to be *universally measurable*, and by the definition of $\mathscr{U}(X)$, they are measurable with respect to the completion of any Borel probability measure on X.

We consider subsets of a *Polish space* – a topological space that can be metrized by a metric under which it is separable and complete [19, p. 344]. In this paper, a *Borel space* refers to a Borel subset of a Polish space,[1] endowed with the relative topology and Borel σ-algebra. The Cartesian product of countably many Polish (Borel) spaces is also a Polish (Borel) space.

We now introduce analytic sets in a Polish space X. The empty set is an analytic set by definition. The nonempty analytic sets are the images of Borel sets under continuous or Borel measurable functions, roughly speaking. They were first discovered when studying projections of Borel sets, which are important also in the optimal control context since partial minimization can be viewed as projection. Analytic sets have several equivalent definitions (see e.g., [7, Prop. 7.41], [19, Sec. 13.2]). We mention one here. A nonempty set $A \subset X$ is *analytic* if $A = f(B)$ for some Borel set B in a Polish space and Borel measurable function $f : B \to X$ [19, Thm. 13.2.1(c')]. Every Borel set in a Polish space is analytic; the converse is not true ([7, Appendix B.3], [19, Prop. 13.2.5]). Every analytic set is universally measurable ([7, Cor. 7.42.1], [19, Thm. 13.2.6]).

For a Borel space X or an analytic set X, besides the Borel σ-algebra $\mathscr{B}(X)$ and the universal σ-algebra $\mathscr{U}(X)$, we also have the *analytic σ-algebra* $\mathscr{A}(X)$, the σ-algebra generated by the analytic subsets of X. A measurable function from $(X, \mathscr{A}(X))$ or $(X, \mathscr{U}(X))$ to $(Y, \mathscr{B}(Y))$, where Y is a topological space, is said to be *analytically measurable* or *universally measurable*, respectively. The three σ-algebras on X satisfy $\mathscr{B}(X) \subset \mathscr{A}(X) \subset \mathscr{U}(X)$ (the inclusions are strict if X is an uncountable Borel space) [7, p. 171]. Thus, every Borel measurable function is analytically measurable, and every analytically measurable function is universally measurable.

The class of analytic sets in a Polish space is closed under countable unions, countable intersections and Borel preimages ([7, Cor. 7.35.2, Prop. 7.40], [46, Chap. 4]). This gives rise to many nice properties of lower semi-analytic functions, functions whose lower level sets are analytic. More specifically, a function $f : D \to [-\infty, \infty]$ is said to be *lower semi-analytic* if D is an analytic set and for every $c \in \Re$, the level set $\{x \in D \mid f(x) < c\}$ of f is analytic [7, Def. 7.21]. (Equivalently, the epigraph of f, $\{(x, c) \mid x \in D, f(x) \le c, c \in \Re\}$, is analytic; cf. [7, p. 186].) Every lower semi-analytic function is universally measurable, since analytic sets are universally measurable. Moreover, based on the properties of analytic sets, the following operations on lower semi-analytic functions result in a lower semi-analytic function (see [7, Lemma 7.30]):

(i) If $f, g : D \to [-\infty, \infty]$ are lower semi-analytic functions, then $f + g$ is lower semi-analytic (here $\infty - \infty$ and $-\infty + \infty$ are defined to be ∞). In addition, if $f, g \ge 0$ or if g is Borel measurable and $g \ge 0$, then fg is lower semi-analytic. Note a particular implication of this: for a Borel subset B of D, $f \cdot \mathbb{1}_B$ is lower semi-analytic, where $\mathbb{1}_B$ denotes the indicator function for B.

(ii) If $g : X \to Y$ is Borel measurable, where X, Y are Borel spaces, and $f : g(X) \to [-\infty, \infty]$ is lower semi-analytic, then the composition $f \circ g$ is lower semi-analytic.

(iii) For a sequence of lower semi-analytic functions $f_n : D \to [-\infty, \infty]$, $n \ge 1$, the functions

[1] The definition of Borel spaces given in [7] is more general than the one we give here. The two definitions are, however, essentially equivalent, and the Borel spaces are now commonly called *standard Borel spaces* (see e.g. [46]).

$\inf_n f_n$, $\sup_n f_n$, $\liminf_n f_n$ and $\limsup_n f_n$ are all lower semi-analytic. (These are pointwise definitions.)

Several properties of analytic sets and lower semi-analytic functions play instrumental roles in the stochastic control framework we will introduce. They concern analytic sets in product spaces or functions involving two variables. The first property is closely related to value iteration and the structure of the optimal cost function in the stochastic control context. If A is an analytic set in $X \times Y$, where X, Y are Polish, the projection of A on X, $\mathrm{proj}_X(A) = \{x \mid (x,y) \in A \text{ for some } y\}$, is analytic [7, Prop. 7.39]. When applied to level sets of functions, an implication of this is that if $D \subset X \times Y$ is analytic and $f : D \to [-\infty, \infty]$ is lower semi-analytic, then the function $f^* : \mathrm{proj}_X(D) \to [-\infty, \infty]$ resulting from the partial minimization,

$$f^*(x) = \inf_{y \in D_x} f(x,y), \quad \text{where } D_x = \{y \mid (x,y) \in D\}, \tag{2.1}$$

is lower semi-analytic [7, Prop. 7.47].

The Jankov-von Neumann's selection theorem asserts that if A is an analytic set in $X \times Y$, where X, Y are Polish, then there exists an analytically measurable function $\phi : \mathrm{proj}_X(A) \to Y$ such that the graph of ϕ lies in A, i.e., $(x, \phi(x)) \in A$ for all $x \in \mathrm{proj}_X(A)$ [7, Prop. 7.49]. For minimization problems of the form (2.1), the theorem is applied to the level sets or epigraphs of lower semi-analytic functions, and together with other properties, it yields the existence of an analytically measurable ϵ-minimizer and the existence of a universally measurable ϵ-minimizer that attains the minimum $f^*(x)$ at every x where $f^*(x)$ is attained by some $y \in D_x$. For details, see the selection theorems given in [7, Prop. 7.50(a)-(b)]. In the stochastic control context, this is closely related to the existence of optimal or nearly optimal policies and their structures.

Another important property of lower semi-analytic functions involves integration and stochastic kernels. Let X and Y be Borel spaces. In this paper, a *Borel*, *analytically* or *universally measurable stochastic kernel* on Y given X is a mapping $\kappa(\cdot \mid \cdot) : \mathscr{B}(Y) \times X \to [0,1]$ such that:

(i) For each $B \in \mathscr{B}(Y)$, the function $\kappa(B \mid \cdot) : X \to [0,1]$ is $\mathcal{F}$-measurable with $\mathcal{F} = \mathscr{B}(X), \mathscr{A}(X)$ or $\mathscr{U}(X)$, respectively.

(ii) For each $x \in X$, $\kappa(\cdot \mid x)$ is a probability measure on $(Y, \mathscr{B}(Y))$.

Equivalently, the function $x \mapsto \kappa(\cdot \mid x)$ is $\mathcal{F}$-measurable from X to the space $\mathscr{P}(Y)$ endowed with the weak topology [7, Prop. 7.26, Lemma 7.28, Prop. 11.6]. If $f : X \times Y \to [0, \infty]$ is lower semi-analytic and $\kappa(dy \mid x)$ is a Borel measurable stochastic kernel on Y given X, then the integral

$$\int_Y f(x,y)\, \kappa(dy \mid x)$$

as a function of x is lower semi-analytic on X [7, Prop. 7.48], where for each x, the integration is defined to be with respect to the completion of the Borel probability measure $\kappa(dy \mid x)$. If instead $\kappa(dy \mid x)$ is analytically or universally measurable, then the integral as a function of x is universally measurable [7, Prop. 7.46, Sec. 11.2] and is *not* necessarily lower semi-analytic. These facts are closely related to the structure of the cost functions and the selection of measurable policies in the stochastic control context.

For more properties of analytic sets and lower semi-analytic functions, see the paper [16] and the monograph [7, Chap. 7]. (For general properties of analytic sets, see also the books [37, 46].)

2.2 Stochastic Control Model

Our stochastic control model involves a state space S and a control space C, which are assumed to be Borel spaces. We will write x for a state in S and u for a control in C. At each state $x \in S$,

one can apply a control from a nonempty subset $U(x) \subset C$. The set-valued function U given by $x \mapsto U(x)$ specifies the control constraint for all states. We assume that the graph of U,

$$\Gamma = \{(x, u) \mid x \in S, u \in U(x)\},$$

is an analytic subset of $S \times C$. Applying a control u at a state x incurs a possibly infinite one-stage cost and moves the system to another state x'. The one-stage cost is given by $g(x, u)$, where $g : \Gamma \to [-\infty, \infty]$ is assumed to be a lower semi-analytic function. The transition to state x' is according to a Borel measurable stochastic kernel $q(dx' \mid x, u)$ on S given $S \times C$.

2.2.1 Policies and Induced Stochastic Processes

A *policy* is a sequence of functions, $\pi = (\mu_0, \mu_1, \ldots)$, where for each k, μ_k maps $(x_0, u_0, \ldots, u_{k-1}, x_k) \in (S \times C)^k \times S$ to a Borel probability measure on C, denoted $\mu_k\big(du_k \,|\, x_0, u_0, \ldots, u_{k-1}, x_k\big)$, such that with respect to the completion of this measure,

$$\mu_k\big(U(x_k) \,|\, x_0, u_0, \ldots, u_{k-1}, x_k\big) = 1, \qquad \forall\, (x_0, u_0, \ldots, u_{k-1}, x_k). \tag{2.2}$$

The constraint (2.2) says that the set of non-admissible controls, $C \setminus U(x_k)$, has probability zero. This is meaningful since Γ is analytic: each vertical section $U(x)$ of Γ is universally measurable [7, Lemma 7.29] and hence measurable for the completion of any Borel probability measure on C.

A policy π is said to be *nonrandomized* if for every k and every $(x_0, u_0, \ldots, u_{k-1}, x_k)$, the probability measure $\mu_k\big(du_k \,|\, x_0, u_0, \ldots, u_{k-1}, x_k\big)$ is a Dirac measure that assigns probability one to some point in $U(x_k)$. A policy π is said to be *semi-Markov* if for every k, μ_k depends only on (x_0, x_k); *Markov* if for every k, μ_k depends only on x_k; *stationary* if π is Markov and $\mu_k = \mu$ for all k. For the stationary case, we simply write μ for $\pi = (\mu, \mu, \ldots)$. A nonrandomized stationary policy μ can be viewed as a mapping that maps $x \in S$ to a point in $U(x) \subset C$. We denote this mapping also by μ, and we will use both notations $\mu(x)$, $\mu(du \,|\, x)$ in the paper, depending on the context.

So far, no measurability conditions are placed on the functions μ_k of a policy $\pi = (\mu_0, \mu_1, \ldots)$. In this paper we will focus on measurable policies, in particular, universally measurable policies, defined as follows.

A policy π is said to be *universally measurable* if for each k, $\mu_k\big(du_k \mid x_0, u_0, \ldots, u_{k-1}, x_k\big)$ is a universally measurable stochastic kernel on C given $(S \times C)^k \times S$. Similarly, a policy π is said to be *Borel measurable* or *analytically measurable* if each stochastic kernel component of π is Borel measurable or analytically measurable; such a policy is by definition also universally measurable. Because Γ is analytic, by Jankov-von Neumann's selection theorem [7, Prop. 7.49], there exists at least one universally measurable (in fact, analytically measurable) nonrandomized stationary policy. A Borel measurable policy, however, may not exist [14].

We denote by Π' the set of universally measurable policies and by Π the set of universally measurable Markov policies. Both sets are nonempty, as just mentioned. In what follows, when no confusion arises, we will simply refer to universally measurable policies as policies.

Given a policy $\pi \in \Pi'$, the collection of stochastic kernels

$$\begin{aligned} &\mu_0(du_0 \mid x_0),\ q(dx_1 \mid x_0, u_0),\ \mu_1(du_1, \mid x_0, u_0),\ q(dx_2 \mid x_1, u_1),\ \ldots, \\ &\ldots,\ \mu_k\big(du_k \mid x_0, u_0, \ldots, u_{k-1}, x_k\big),\ q(dx_{k+1} \mid x_k, u_k),\ \ldots, \end{aligned}$$

uniquely determines, for each initial distribution p_0 of x_0, a probability measure $r(\pi, p_0)$ on the universal σ-algebra on $(S \times C)^\infty$ with the following property [7, Prop. 7.45]:[2] with respect to $r(\pi, p_0)$,

[2] It is worth noting that the universal σ-algebra on $(S \times C)^\infty$ is not a product σ-algebra, so the existence of a unique probability measure $r(\pi, p_0)$ here does not follow immediately from the Ionescu Tulcea theorem.

the expectation $\mathrm{E}f$ for any nonnegative, universally measurable function $f : (S \times C)^{k+1} \to [0, \infty]$ equals the iterated integral

$$\int_S \int_C \cdots \int_S \int_C f(x_0, u_0, \ldots, x_k, u_k) \mu_k(du_k \mid x_0, u_0, \ldots, x_k)\, q(dx_k \mid x_{k-1}, u_{k-1}) \cdots \mu_0(u_0 \mid x_0)\, p_0(dx_0).$$

Here and in what follows, an integral $\int f dp$ that involves a universally measurable function f and a Borel probability measure p, is defined to be the integral of f with respect to the completion of p.

In general, for a measurable, extended real-valued function f, $\mathrm{E}f$ is defined as usual to be $\mathrm{E}f^+ - \mathrm{E}f^-$, where $f^+ = \max\{0, f\}$, $f^- = -\min\{0, f\}$. The convention $\infty - \infty = -\infty + \infty = \infty$ will be adopted, although in the control problems we consider, we will not encounter such summations.

2.2.2 Infinite Horizon Total Cost Problems

We consider primarily three types of control problems under total expected cost criteria: discounted total cost problems with bounded one-stage costs (D), and undiscounted total cost problems with nonpositive one-stage costs (N) and with nonnegative one-stage costs (P). Specifically, let $\alpha \in [0, 1]$ be the discount factor.

(D) $\alpha < 1$ and $-b \le g(x, u) \le b$ for all $(x, u) \in \Gamma$, where $b \in \Re$.

(N) $\alpha = 1$ and $g \le 0$.

(P) $\alpha = 1$ and $g \ge 0$.

We mention that for reward maximization (instead of cost minimization), the reverse terminologies are used in the literature [13, 47, 33, 38]: case (N) here corresponds to the positive model and case (P) to the negative model considered there.

In each of the (D)(N)(P) cases, we define the cost of $\pi \in \Pi'$ for an initial state $x_0 = x \in S$ to be

$$J_\pi(x) = \mathrm{E}^\pi \left\{ \sum_{k=0}^{\infty} \alpha^k g(x_k, u_k) \right\},$$

the expectation of the universally measurable function $\sum_{k=0}^{\infty} \alpha^k g(x_k, u_k)$ with respect to the probability measure $r(\pi, \delta_x)$, which is induced by π and the initial distribution δ_x (a Dirac measure that assigns probability 1 to the point x), as described earlier. (Although g is only defined on Γ, π is a policy and satisfies the control constraint, so $(x_k, u_k) \in \Gamma$ for all k with probability one and the expectation is thus well-defined.) By the bounded convergence theorem (for case (D)) and the monotone convergence theorem (for cases (N)(P)), we can also write $J_\pi(x)$ as

$$J_\pi(x) = \sum_{k=0}^{\infty} \alpha^k \, \mathrm{E}^\pi \big\{ g(x_k, u_k) \big\},$$

where the expectation is with respect to the marginal of $r(\pi, \delta_x)$ on the space $S \times C$ of (x_k, u_k). For all $\pi \in \Pi'$, the cost functions J_π are universally measurable [7, p. 215].

The optimal cost function is defined by the minimal cost of universally measurable policies π for each state:

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

If $J_\pi(x) = J^*(x)$, π is optimal for state x. For $\epsilon > 0$, π is said to be *ϵ-optimal* if for all $x \in S$,

$$J_\pi(x) \le \begin{cases} J^*(x) + \epsilon & \text{if } J^*(x) > -\infty; \\ -1/\epsilon & \text{if } J^*(x) = -\infty. \end{cases}$$

Remark 2.1 (Special Cases with Transition-Dependent Discounting). In certain discounted problems, the discount factor at each stage can depend on the state transition. Problems of this type are considered in e.g., [40, 55]. Under mild model assumptions, we can convert such discounted problems to equivalent, discounted or undiscounted problems in the stochastic control framework we just gave, as we show below. Thus the existing results for (D)(N)(P) as well as the results of this paper are applicable to these problems as well.

Suppose that our objective is to minimize over π the expected total discounted cost for each initial state $x_0 = x$,

$$J_\pi(x) = \mathrm{E}^\pi \left\{ \hat{g}(x_0, u_0, x_1) + \sum_{k=1}^{\infty} \Big(\prod_{i=1}^{k} \beta(x_{i-1}, u_{i-1}, x_i) \Big) \cdot \alpha^k \, \hat{g}(x_k, u_k, x_{k+1}) \right\},$$

where $\alpha \in [0, 1]$ as earlier, $\beta : S \times C \times S \to [0, 1]$ is a Borel measurable function that describes the transition-dependent discount factors, and $\hat{g} : \Gamma \times S \to [-\infty, \infty]$ is a lower semi-analytic function that describes the transition costs.

We convert this problem to an equivalent one, by applying a simple transformation of the state transition dynamics as follows. In the equivalent problem, we introduce an additional cost-free and absorbing state, denoted ∞, and let the state space be $\tilde{S} = S \cup \{\infty\}$. We define the state transition kernel $\tilde{q}$ on C given $\tilde{S} \times C$ as follows: $\tilde{q}\big(\{\infty\} \,|\, \infty, u\big) = 1$ for all $u \in C$, and for each $(x, u) \in S \times C$,

$$\begin{aligned} \tilde{q}(B \,|\, x, u) &= \int_B \beta(x, u, x') \, q(dx' \,|\, x, u), \qquad B \subset S, \ B \text{ Borel measurable}, \\ \tilde{q}\big(\{\infty\} \,|\, x, u\big) &= 1 - \tilde{q}(S \,|\, x, u). \end{aligned}$$

We consider the standard discounted cost criterion ($\alpha < 1$) or undiscounted total cost criterion ($\alpha = 1$), with the one-stage cost function g given by $g(\infty, u) = 0$ for all $u \in C$ and

$$g(x, u) = \int_S \hat{g}(x, u, x') \, q(dx' \,|\, x, u), \qquad (x, u) \in \Gamma.$$

This problem is then of type (D) if in the original problem $\alpha < 1$ and $\hat{g}$ is bounded, of type (N) if $\alpha = 1$ and $\hat{g} \le 0$, and of type (P) if $\alpha = 1$ and $\hat{g} \ge 0$. □

2.3 Optimality Properties

Let $A(S)$ denote the set of functions $f : S \to [-\infty, \infty]$ that are lower semi-analytic, and let $\mathcal{M}(S)$ denote the set of functions $f : S \to [-\infty, \infty]$ that are universally measurable. In each of the (D)(N)(P) cases, the optimal cost function J^* is lower semi-analytic, and it satisfies the optimality equation

$$J^* = T(J^*),$$

where T maps $A(S)$ into $A(S)$ and is given by

$$T(J)(x) = \inf_{u \in U(x)} \left\{ g(x, u) + \alpha \int_S J(x') \, q(dx' \,|\, x, u) \right\}, \qquad x \in S. \tag{2.3}$$

We will refer to T as the optimal cost operator. We note that $T(J) \in A(S)$ for any function $J \in A(S)$, as just mentioned. This is a direct consequence of the preservation of lower semi-analyticity by the partial minimization operation (cf. Eq. (2.1)) and other properties of lower semi-analytic functions (cf. Section 2.1), combined with our model assumption that the graph Γ of the control constraint U is analytic, the one-stage cost g is lower semi-analytic, and the state transition kernel $q(dx' \,|\, x, u)$ is Borel measurable (cf. Section 2.2).

In each of the (D)(N)(P) cases, the cost function J_μ for a stationary policy μ is universally measurable. It satisfies a linear equation,

$$J_\mu = T_\mu(J_\mu),$$

where T_μ is a mapping from $\mathcal{M}(S)$ to $\mathcal{M}(S)$, given by

$$T_\mu(J)(x) = \int_C \left(g(x,u) + \alpha \int_S J(x')\, q(dx' \mid x,u) \right) \mu(du \mid x), \qquad x \in S. \tag{2.4}$$

In terms of the convergence properties of the value iteration sequence $T^k(J)$ and the structures of the optimal policies, the (D)(N)(P) cases differ. We consider primarily pointwise convergence. Throughout the paper, for a sequence of functions f_n converging pointwise to a function f, we write $f_n \to f$, and if the convergence is monotonically from above or from below, we write $f_n \downarrow f$ or $f_n \uparrow f$, respectively. Some convergence properties of value iteration under (D), (N) or (P) are:

(a) For (D)(N), value iteration converges pointwise to J^*. In particular, $T^k(J) \to J^*$ for any bounded lower semi-analytic function J in case (D), and $T^k(J) \downarrow J^*$ for $J \equiv 0$ in case (N).

(b) For (P), value iteration need not converge to J^*: for $J \equiv 0$,

$$T^k(J) \uparrow J_\infty \le J^*,$$

where the pointwise limit J_∞ of $\{T^k(J)\}$ satisfies $J_\infty \le T(J_\infty)$.

In all three cases, ϵ-optimal nonrandomized policies exist for each $\epsilon > 0$; however, they can be taken to be stationary for (D), *semi-Markov* for (N), and *Markov* for (P). An ϵ-optimal randomized Markov policy need not exist for (N) (a counterexample was given by van der Wal [51]; see also [38, p. 326]). If for each state x, an optimal policy exists, then:

(a) For (D)(P), an optimal nonrandomized stationary policy exists.

(b) For (N), an optimal randomized semi-Markov policy exists.

The readers can find in [7, Chap. 9] the optimality properties mentioned above, as well as finer characterizations of the optimal cost function and optimal policies, some of which we will mention later in the paper where they are needed.

2.4 Measurability Issues in Standard Policy Iteration

In the policy iteration scheme, we repeat the following two steps:

(i) Evaluate the cost function J_μ of a given stationary policy μ.

(ii) Find a stationary policy μ' with

$$T_{\mu'}(J_\mu) = T(J_\mu)$$

and go to step (i) with $\mu = \mu'$.

A variant of it is the modified policy iteration [38]:

(i') For a given stationary policy μ and a given function J, compute as an approximation of J_μ,

$$J' = T_\mu^m(J) \quad \text{for some positive integer } m.$$

(ii') Find a stationary policy μ' with

$$T_{\mu'}(J') = T(J')$$

and go to step (i') with $\mu = \mu'$ and $J = J'$.

Both schemes break down, however, for the stochastic control model with universally measurable policies, due to measurability issues (cf. [16, p. 940], [7, p. 232]). We explain the reasons below.

As defined in (2.3), T is also a mapping from $\mathcal{M}(S)$ to the space of functions on S: it maps a universally measurable function J to the function $T(J)$, possibly outside $\mathcal{M}(S)$. For a stationary policy μ, J_μ is universally measurable, so $T(J_\mu)$ is defined. But since J_μ need not be lower semi-analytic, even if $T(J_\mu)$ is universally measurable, a stationary, universally measurable policy μ' such that

$$T_{\mu'}(J_\mu) = T(J_\mu) \quad \text{or} \quad T_{\mu'}(J_\mu) \le T(J_\mu) + \epsilon, \quad \text{for some given } \epsilon > 0,$$

may not exist. When this happens, step (ii) of policy iteration cannot be carried out. The same issue also causes modified policy iteration to break down.

Blackwell et al. [16, Example (48)] gave an example of an analytically measurable function J on $[0,1]$ for which $T(J)$ is not Lebesgue measurable. If J_μ equals such J, then there is certainly no stationary policy μ' that can satisfy $T_{\mu'}(J) = T(J)$, because for all μ', $T_{\mu'}(J)$ is universally measurable, whereas $T(J)$ is not. Moreover, since $T(J)$ is not universally measurable, for some $p \in \mathscr{P}(S)$, $T(J)$ is not integrable with respect to the completion of p. Hence, $T^2(J)$ as well as $(T_\mu \circ T)(J)$ for a stationary policy μ can be undefined for some states x (cf. [16, Example (48)]). This means that variants of policy iteration of the form $J_{k+1} = T_k(J_k)$, where some of the T_k's equal T and others equal T_μ for some stationary policy μ, can also run into trouble.

3 A Mixed Value and Policy Iteration Method

Let $\mathcal{M}(\Gamma)$ (resp. $A(\Gamma)$) denote the set of all functions $f : \Gamma \to [-\infty, \infty]$ that are universally measurable (resp. lower semi-analytic). Denote the subset of bounded (resp. nonnegative and nonpositive) functions of $A(\Gamma)$ by $A_b(\Gamma)$ (resp. $A_+(\Gamma)$ and $A_-(\Gamma)$).

For (D)(N)(P), recall that the relation $J^* = T(J^*)$ holds:

$$J^*(x) = \inf_{x \in U(x)} \left\{ g(x,u) + \alpha \int_S J^*(x')\, q(dx' \mid x, u) \right\}, \qquad \forall\, x \in S.$$

We define $Q^* \in A(\Gamma)$ by

$$Q^*(x,u) = g(x,u) + \alpha \int_S J^*(x')\, q(dx' \mid x, u), \qquad (x,u) \in \Gamma. \tag{3.1}$$

For each $(x,u) \in \Gamma$, we may view $Q^*(x,u)$ as the result of cost minimization over controllers that start at state x, apply control u, and then choose some policy. This interpretation of $Q^*(x,u)$ is better revealed in the following equation, which is equivalent to (3.1) [7, Cor. 9.5.2]:

$$Q^*(x,u) = g(x,u) + \alpha \inf_{\pi \in \Pi'} \int_S J_\pi(x')\, q(dx' \mid x, u), \qquad (x,u) \in \Gamma. \tag{3.2}$$

(In the literature on learning and simulation-based DP, $Q^*(x,u)$ is known as the optimal Q-factor associated with (x,u); see e.g., [8, 48].) To simplify notation, for any function Q on Γ, let

$$M(Q)(x) = \inf_{u \in U(x)} Q(x,u), \qquad x \in S.$$

The mapping M maps $A(\Gamma)$ into $A(S)$ [7, Prop. 7.47]. With this notation, we can write the optimality equation in two equivalent ways:

$$J^* = T(J^*) \qquad \Longleftrightarrow \qquad J^* = M(Q^*). \tag{3.3}$$

We introduce in this section a mixed value and policy iteration method, which operates on the product space $A(S) \times A(\Gamma)$. The method combines characteristics of both value and policy iteration, and the combination has two crucial features. First, it uses portions of a universally measurable policy that are Borel, to preserve the lower semi-analytic properties of the functions involved, thereby overcoming the measurability issues in standard policy iteration. Second, thanks to its value iteration character, it does not rely strongly on the behavior of policies for convergence. In particular, the policies involved are not required to be successively improving – a requirement that in general cannot be met in our context or in the case where the policies involved are restricted to be Borel measurable [12]. Our method gives rise to various policy iteration-like algorithms, whose convergence we will analyze in Sections 4 and 5.

In what follows we introduce a family of mappings underlying the method and we discuss its relation to optimal stopping problems (Section 3.1). We then give various forms of algorithms (Section 3.2), followed by related discussions on the existence of Borel measurable policies (Section 3.3) in connection with one of our policy iteration-like algorithms.

3.1 Mappings Induced by Stationary Policies

First, we introduce a family of parametrized mappings F_θ, with parameter $\theta \in \Theta$. Let Θ denote the set of all pairs (μ, B), where μ is a stationary policy and B a Borel subset of S, such that the function $x \mapsto \mu(du \,|\, x)$ restricted to B is Borel measurable (equivalently, for every Borel subset D of C, $\mu(D \,|\, \cdot)$ is Borel measurable on B).

For each $\theta = (\mu, B) \in \Theta$, we define a mapping $F_\theta : \mathcal{M}(\Gamma) \times \mathcal{M}(S) \to \mathcal{M}(\Gamma)$ by

$$F_\theta(Q\,;J)(x,u) = g(x,u) + \alpha \int_{S\setminus B} J(x')\, q(dx' \,|\, x,u) + \alpha \int_B \int_C \min\big\{J(x')\,,\, Q(x',u')\big\}\, \mu(du' \,|\, x')\, q(dx' \,|\, x,u), \qquad (x,u) \in \Gamma, \tag{3.4}$$

for all $Q \in \mathcal{M}(\Gamma)$ and $J \in \mathcal{M}(S)$. Here the convention $\infty - \infty = -\infty + \infty = \infty$ is used. We also note that although Q is defined only on Γ, the inner integral in the third term in (3.4) is well-defined because μ satisfies the control constraint. (We could, for example, view this integral as an integral for the extension of Q to $S \times C$ with $Q(x',u') = \infty$ outside Γ.)

For any stationary policy μ, the trivial choice $B = \emptyset$ gives $\theta = (\mu, \emptyset) \in \Theta$, but the corresponding mapping F_θ does not depend on the policy μ at all. To introduce greater dependence of F_θ on μ, we desire "large" sets B. By the nature of universally measurable policies, one can indeed find "large" B with $(\mu, B) \in \Theta$ (see Prop. 3.1(b) below and see also Example 3.1, Section 3.2). If the policy μ is Borel measurable, then $(\mu, S) \in \Theta$.

An important property of F_θ is that it preserves the lower semi-analyticity of functions. This will allow us to overcome the measurability difficulties that hamper standard policy iteration.

Proposition 3.1.

(a) For any $\theta \in \Theta$ and $J \in A(S)$, $F_\theta(\cdot\,;J)$ maps $A(\Gamma)$ into $A(\Gamma)$.

(b) For each stationary policy μ, given any $p \in \mathscr{P}(S)$, there is a Borel set $B \subset S$ with $p(S \setminus B) = 0$ and $(\mu, B) \in \Theta$.

Proof. (a) Let $Q \in A(\Gamma)$. We show that the function $F_\theta(Q\,;J)(\cdot,\cdot)$ given by Eq. (3.4) is lower semi-analytic, by proving that each term in the right-hand side of Eq. (3.4) is lower semi-analytic. The first term is lower semi-analytic by definition. The second term equals $\int_S \mathbb{1}_{S\setminus B}(x') J(x')\, q(dx' \,|\, x,u)$. Here the set $S \setminus B$ is Borel, J is lower semi-analytic, and $q(dx' \,|\, x,u)$ is a Borel measurable stochastic kernel on S given $S \times C$ in our stochastic control model. Then by [7, Lemma 7.30(4) and Prop.

7.48], the second term is lower semi-analytic on $S \times C$ and hence lower semi-analytic on the analytic set Γ. We show now that the third term,

$$\alpha \int_B \int_C \min\left\{J(x')\,,\, Q(x',u')\right\} \mu(du' \mid x')\, q(dx' \mid x,u), \tag{3.5}$$

is lower semi-analytic. Let Q^e be an extension of Q to $S \times C$ with $Q^e(x,u) = \infty$ for $(x,u) \notin \Gamma$. Since Q is lower semi-analytic, Q^e is lower semi-analytic by definition. Using the fact that μ satisfies the control constraint, we can write the term in (3.5) equivalently as

$$\alpha \int_S \int_C f(x',u')\, \mu(du' \mid x')\, q(dx' \mid x,u), \tag{3.6}$$

where $f : S \times C \to [-\infty, \infty]$ is given by $f(x',u') = \mathbb{1}_B(x') \cdot \min\{J(x'), Q^e(x',u')\}$ for $(x',u') \in S \times C$. The function f is lower semi-analytic, since the functions J and Q^e are lower semi-analytic and the set B is Borel [7, Lemma 7.30(2),(4)]. It follows that f is lower semi-analytic on $B \times C$. Since $(\mu, B) \in \Theta$, the defining property of Θ implies that $\mu(du' \mid x')$ is a Borel measurable stochastic kernel on C given B. Hence $\int_C f(x',u')\, \mu(du' \mid x')$ is lower semi-analytic on B by [7, Prop. 7.48]. We also have $\int_C f(x',u')\, \mu(du' \mid x') = 0$ for $x' \notin B$. Therefore, $\int_C f(x',u')\, \mu(du' \mid x')$ is lower semi-analytic on S. Then, since $q(dx' \mid x,u)$ is a Borel measurable stochastic kernel on S given $S \times C$, the integral (3.6) as a function of (x,u) is lower semi-analytic on $S \times C$ by [7, Prop. 7.48] and hence lower semi-analytic on the analytic set Γ. Equivalently, the integral (3.5) as a function of (x,u) is lower semi-analytic on Γ. This proves part (a).

(b) Since $\mu(du \mid x)$ is a universally measurable stochastic kernel on C given S, by [7, Lemma 7.28], there is a Borel measurable stochastic kernel $\tilde{\mu}(du \mid x)$ with $\tilde{\mu}(du \mid x) = \mu(du \mid x)$ everywhere except on a set D with p-outer measure zero. Let $D' \supset D$ be a Borel set with $p(D') = 0$. Letting $B = S \setminus D'$ proves part (b). □

In the discounted case (D), we work with $J \in A_b(S), Q \in A_b(\Gamma)$, the subsets of bounded lower semi-analytic functions. In the nonpositive case (N), we work with $J \in A_-(S), Q \in A_-(\Gamma)$, the subsets of nonpositive lower semi-analytic functions, whereas in the nonnegative case (P), we work with $J \in A_+(S), Q \in A_+(\Gamma)$, the subsets of nonnegative lower semi-analytic functions. By Prop. 3.1 and the definition of $F_\theta(\cdot\,; J)$, we see that in each of the (D)(N)(P) cases, $F_\theta(\cdot\,; J)$ maps the sets $A_b(\Gamma)$, $A_-(\Gamma)$, and $A_+(\Gamma)$ into themselves, for $J \in A_b(S)$, $J \in A_-(S)$, and $J \in A_+(S)$, respectively.

For discrete spaces and abstract DP problems, where measurability is not a concern, we have considered in our earlier work [10, 57, 9] mappings of the form F_θ, $\theta = (\mu, S)$, without splitting the state space by a set $B \subset S$ according to the policy μ. In the present context, however, in order for F_θ to map lower semi-analytic functions to lower semi-analytic functions, it is important to introduce B as a parameter component in defining F_θ.

Optimal stopping problems corresponding to $F_\theta(\cdot\,; J)$

It is intuitive to relate $F_\theta(\cdot\,; J)$ to an optimal stopping problem defined by (θ, J) and the parameters of the original control problem, with J specifying the stopping costs. We give a precise mathematical formulation in Appendix A, where we will also show that $F_\theta(\cdot\,; J)$ can be viewed as a form of the optimal cost operator. Here we describe this optimal stopping problem intuitively. In the optimal stopping problem associated with $\theta = (\mu, B)$ and J, the states are the state-control pairs of the original control problem. Suppose we start from a state (x,u) in Γ at time 0; at this time we must pay $g(x,u)$ and choose to continue. (This corresponds to the first term in Eq. (3.4).) At time 1, we first land at x' according to $q(dx' \mid x,u)$. If $x' \in S \setminus B$, then we must pay $J(x')$ and immediately stop. (This corresponds to the second term in Eq. (3.4).) If $x' \in B$, then u' is generated and we land at (x',u') according to $\mu(du' \mid x')$, and there, we can either stop and pay $J(x')$, or continue with

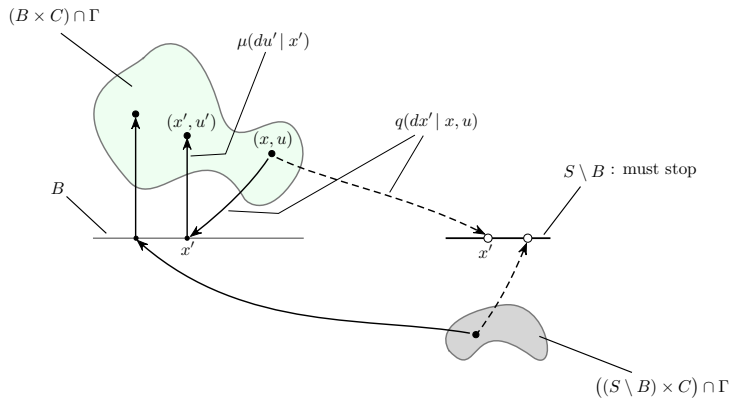

Figure 1: Illustration of the system dynamics of an optimal stopping problem corresponding to $F_\theta(\cdot\,; J)$ with $\theta = (\mu, B) \in \Theta$.

the continuation cost $g(x', u')$. (This corresponds to the third term in Eq. (3.4).) If we choose to continue, we repeat the process just described for time 1. Figure 1 illustrates this optimal stopping problem.

A special case of the function J provides further insight, by allowing us to relate the total cost in the optimal stopping problem to that in the original problem. Suppose $J = J_\pi$ for some $\pi = (\pi_0, \pi_1, \ldots) \in \Pi'$ (with J_π lower semi-analytic). Let $Q_{\theta, J_\pi}(x, u)$ be the minimal cost starting from (x, u) in the optimal stopping problem just described. We may interpret $Q_{\theta, J_\pi}(x, u)$ as the minimal cost of a set of policies (in an extended sense) in the original control problem, by constructing these policies from policies in the optimal stopping problem based on interpreting the action to stop as the decision to switching from applying μ to applying π forever in the original problem. More specifically, these policies apply control u at state x at time 0. From time 1 on, they either follow the stationary policy μ or use the policy π, which they must do if the state goes outside the set B. Once they start to use π at time τ, say, they apply $\pi_0, \pi_1, \ldots$ at time $\tau, \tau + 1, \ldots$, respectively, and continue in this way forever. (We do not include a formal proof for this interpretation of $Q_{\theta, J_\pi}(x, u)$ in the paper; but we note that it is similar to the analysis we give in Appendix B.)

Because of the correspondence between $F_\theta(\cdot\,; J)$ and an optimal stopping problem, some of the theories for (D)(N)(P) with a finite number of controls can be applied to analyze the properties of F_θ (see Appendices A and B).

Some basic properties of F_θ

We now discuss a few basic properties of the mappings F_θ and $F_\theta(\cdot\,; J)$, relating to monotonicity and fixed point properties, and their relation with (J^*, Q^*). Let $F_\theta^n(\cdot\,; J)$ denote the n-fold composition of $F_\theta(\cdot\,; J)$, i.e.,

$$F_\theta^n(Q\,; J) = \underbrace{F_\theta\big(\cdots F_\theta\big(F_\theta}_{n \text{ times}}(Q\,; J)\,; J\big)\cdots\,; J\big).$$

By definition F_θ is monotone:

$$J \geq J',\ Q \geq Q' \quad \Longrightarrow \quad F_\theta(Q\,; J) \geq F_\theta(Q'\,; J').$$

Applying this relation with $F_\theta(Q\,; J)$ in place of Q and $F_\theta(Q'\,; J')$ in place of Q', and repeating the argument n times, we see that

$$J \geq J',\ Q \geq Q' \quad \Longrightarrow \quad F_\theta^n(Q\,; J) \geq F_\theta^n(Q'\,; J'), \quad \forall\, n \geq 1. \tag{3.7}$$

Let $\mathbf{0}$ denote the constant function zero. We consider the pointwise limit

$$Q_{\theta,J} = \lim_{n\to\infty} F_\theta^n(\mathbf{0}\,;J),$$

which can be interpreted as the optimal cost function of the optimal stopping problem mentioned earlier (see Cor. A.1, Appendix A).

Proposition 3.2. (D)(N)(P) *Let $J \in A_b(S)$ for (D), $J \in A_-(S)$ for (N), and $J \in A_+(S)$ for (P). Then $Q_{\theta,J} = \lim_{n\to\infty} F_\theta^n(\mathbf{0}\,;J)$ is well-defined and lower semi-analytic, and satisfies*

$$Q_{\theta,J} = F_\theta(Q_{\theta,J}\,;J). \tag{3.8}$$

For (D), it is the only solution of $Q = F(Q\,;J)$ in $A_b(\Gamma)$.

Proof. For case (D), the proposition will be directly proved in Lemma 4.1(b), Section 4, after showing that $F_\theta(\cdot\,;J)$ has a contraction property. For case (N)(resp. (P)), $Q_{\theta,J}$ is the pointwise limit of a sequence of nonincreasing nonpositive functions (resp. nondecreasing nonnegative functions). Equation (3.8) then follows from the definition (3.4) of $F_\theta(\cdot\,;J)$ and the monotone convergence theorem. □

We can relate F_θ and $Q_{\theta,J}$ to Q^* as follows.

Proposition 3.3. (D)(N)(P) *Let $\theta \in \Theta$, $J \in A(S)$, $Q \in A(\Gamma)$.*

(a) $F_\theta(Q^*;J^*) = Q^*$.

(b) If $J \geq J^*, Q \geq Q^*$, then $F_\theta^n(Q;J) \geq Q^*$ for all $n \geq 1$.

(c) Let $J \geq J^*$ with $J \in A_b(S)$ for (D), $J \in A_-(S)$ for (N), and $J \in A_+(S)$ for (P). Then

$$Q_{\theta,J} \geq Q_{\theta,J^*} = Q^*.$$

Proof. Let $\theta = (\mu, B)$. We have $J^*(x) \leq Q^*(x,u)$ for all $(x,u) \in \Gamma$. Thus we can rewrite the iterated integral in the sum (3.4) defining $F_\theta(Q^*;J^*)(x,u)$ as

$$\int_B \int_C \min\left\{J^*(x')\,,\, Q^*(x',u')\right\} \mu(du' \mid x')\, q(dx' \mid x,u) = \int_B J^*(x')\, q(dx' \mid x,u).$$

and by combining it with the second term in (3.4), we obtain

$$F_\theta(Q^*\,;J^*)(x,u) = g(x,u) + \alpha \int_S J^*(x')\,(dx' \mid x,u) = Q^*(x,u), \qquad \forall\,(x,u) \in \Gamma.$$

This proves part (a). Part (b) then follows from part (a) and the monotonicity of F_θ (cf. Eq. (3.7)).

For part (c), since $J \geq J^*$, we have $F_\theta^n(\mathbf{0}\,;J) \geq F_\theta^n(\mathbf{0}\,;J^*)$ for every n, by the monotonicity of F_θ (cf. Eq. (3.7)). Then by Prop. 3.2,

$$Q_{\theta,J} = \lim_{n\to\infty} F_\theta^n(\mathbf{0}\,;J) \geq \lim_{n\to\infty} F_\theta^n(\mathbf{0}\,;J^*) = Q_{\theta,J^*}.$$

There remains to show $Q_{\theta,J^*} = Q^*$. For case (D), this is true because by part (a), Q^* is the solution of $F_\theta(Q\,;J^*) = Q, Q \in A_b(\Gamma)$, whereas we will show in Lemma 4.1 (Section 4) that this equation has Q_{θ,J^*} as its unique solution.

For case (N), we have $J^* \leq 0$ and consequently, $F_\theta(\mathbf{0}\,;J^*) = Q^*$ by the definitions of F_θ and Q^*. In view of part (a), this implies $F_\theta^n(\mathbf{0}\,;J) = Q^*$ for every n, and hence $Q_{\theta,J^*} = Q^*$ by Prop. 3.2.

For case (P), we will show that $Q_{\theta,J^*} = Q^*$ as Prop. B.1 in Appendix B (the proof is not as simple as in (D)(N)). □

3.2 Algorithms

We give first our mixed value and policy iteration algorithm in its basic form. The conditions needed for the convergence of the algorithm are different for each of the (D)(N)(P) cases, and will be given in the subsequent Sections 4 and 5. Our algorithm starts with a pair (J_0, Q_0), which depending on whether case (D), (N), or (P) holds, must belong to $A_b(S) \times A_b(\Gamma)$, or $A_-(S) \times A_-(\Gamma)$, or $A_+(S) \times A_+(\Gamma)$, respectively.

Algorithm I (basic form):

Iterate for each $k \geq 0$:

- Choose $\theta_k = (\mu_k, B_k) \in \Theta$, let

$$Q_{k+1} = F_{\theta_k}^{n_k}(Q_k; J_k) \quad \text{for some } n_k \geq 1, \quad \text{or} \quad Q_{k+1} = Q_{\theta_k, J_k}, \tag{3.9}$$

  and let

$$J_{k+1} = M(Q_{k+1}). \tag{3.10}$$

The above algorithm outputs a sequence of lower semi-analytic function pairs (J_k, Q_k). This can be seen from the inductive argument: by Props. 3.1 and 3.2, the function Q_k is lower semi-analytic if J_k is lower semi-analytic, whereas the infimization (3.10) results in a lower semi-analytic function J_{k+1} by [7, Prop. 7.47].

The algorithm (3.9)-(3.10) allows any choice of $\theta_k = (\mu_k, B_k) \in \Theta$. If we let $\theta_k = (\mu_k, \emptyset)$ for each iteration k, the policy μ_k has no effect on the iterates, and the algorithm reduces to value iteration $J_{k+1} = T(J_k)$. By Prop. 3.1(b), we can choose sets B_k that are not only nonempty but also large (cf. Example 3.1). In what follows, we consider choices of μ_k based on Q_k and a selection theorem of the Jankov-von Neumann type, and we derive policy iteration-like algorithms.

Recall that if $Q \in A(\Gamma)$, then by a selection theorem for lower semi-analytic functions [7, Prop. 7.50(b)], for any $\epsilon > 0$, we can select a universally measurable, nonrandomized stationary policy μ such that, with $I = \big\{x \in S \,\big|\, \arg\min_{u \in U(x)} Q(x,u) \neq \emptyset\big\}$,

$$\mu(x) \in \arg\min_{u \in U(x)} Q(x,u) \qquad \text{if } x \in I, \tag{3.11}$$

$$Q(x, \mu(x)) \leq \begin{cases} M(Q)(x) + \epsilon & \text{if } x \notin I,\ M(Q)(x) > -\infty, \\ -1/\epsilon & \text{if } x \notin I,\ M(Q)(x) = -\infty. \end{cases} \tag{3.12}$$

If we relax the condition (3.11), then by [7, Prop. 7.50(a)], we can find instead an analytically measurable policy μ such that for all states x,

$$Q(x, \mu(x)) \leq \begin{cases} M(Q)(x) + \epsilon & \text{if } M(Q)(x) > -\infty, \\ -1/\epsilon & \text{if } M(Q)(x) = -\infty. \end{cases} \tag{3.13}$$

Choosing the policies in the basic algorithm based on the above selection theorem, we obtain a special form of the basic algorithm that resembles to some degree the modified policy iteration:

Policy Iteration-Like Algorithm II:

In the basic algorithm I, for each $k \geq 1$:

- Let μ_{k+1} be a nonrandomized stationary policy satisfying Eqs. (3.11) and (3.12), or Eq. (3.13), with $Q = Q_{k+1}$ and a desired value of ϵ.

If there exists at least one Borel measurable policy, we can further specialize the above algorithm to use Borel measurable μ_k together with $B_k = S$ for every iteration or whenever this is desirable.

As an example, we give below a policy iteration-like algorithm with Borel measurable policies. When the set Γ is Borel, a nonrandomized Borel measurable policy is known to exist under fairly general conditions (see Section 3.3 for some useful facts). Thus algorithms of this kind can be applied to a large class of problems.

Policy Iteration-Like Algorithm III with Borel Measurable Policies:

Let μ_0 be a Borel measurable stationary policy (assumed to exist).

Iterate for each $k \geq 0$:

- For $\theta_k = (\mu_k, S)$, compute Q_{k+1}, J_{k+1} as in the basic algorithm I: let $Q_{k+1} = F_{\theta_k}^{n_k}(Q_k; J_k)$ for some $n_k \geq 1$ or let $Q_{k+1} = Q_{\theta_k, J_k}$, and then let $J_{k+1} = M(Q_{k+1})$.
- Let μ'_{k+1} be a stationary policy satisfying Eqs. (3.11) and (3.12), or Eq. (3.13), with $Q = Q_{k+1}$ and a desired value of ϵ.
- Select $p_{k+1} \in \mathscr{P}(S)$ and let $B \subset S$ be a Borel set such that $p_{k+1}(B) = 1$ and $(\mu'_{k+1}, B) \in \Theta$ (cf. Prop. 3.1(b)). Define a Borel measurable policy μ_{k+1} by

$$\mu_{k+1}(du \,|\, x) = \begin{cases} \mu'_{k+1}(du \,|\, x) & \text{on } B, \\ \bar{\mu}(du \,|\, x) & \text{on } S \setminus B, \end{cases} \tag{3.14}$$

  where $\bar{\mu}$ is some Borel measurable stationary policy.

In particular, if $\bar{\mu}$ can be chosen to be nonrandomized, then every $\mu_k, k \geq 1$, is a nonrandomized Borel measurable policy.

Remark 3.1. Let us contrast Algorithm III with standard policy iteration. Algorithm III involves mappings $F_{(\mu,S)}$ for Borel measurable policies μ. Such a mapping by its definition (3.4) is given by

$$F_{(\mu,S)}\big(Q\,;J\big)(x,u) = g(x,u) + \alpha \int_S \int_C \min\big\{J(x')\,, Q(x',u')\big\}\, \mu(du' \,|\, x')\, q(dx' \,|\, x,u), \quad (x,u) \in \Gamma,$$

and for a nonrandomized μ, reduces to

$$F_{(\mu,S)}\big(Q\,;J\big)(x,u) = g(x,u) + \alpha \int_S \min\big\{J(x')\,, Q\big(x',\mu(x')\big)\big\}\, q(dx' \,|\, x,u), \quad (x,u) \in \Gamma. \tag{3.15}$$

By contrast, standard policy evaluation of μ involves the affine mapping $T_\mu : \mathcal{M}(S) \to \mathcal{M}(S)$ (cf. Eq. (2.4)), which is given by

$$T_\mu(V)(x) = g\big(x,\mu(x)\big) + \alpha \int_S V(x')\, q\big(dx' \,|\, x,\mu(x)\big), \qquad x \in S. \quad \square$$

Remark 3.2. To further contrast Algorithm III with standard policy iteration, we discuss a property of the policies μ_k in the algorithm, which may be related to a notion of almost-surely ϵ-optimality. For simplicity, let us suppose that in Algorithm III, $Q_{k+1} = Q_{\theta_k, J_k}$ for all k and μ_k are nonrandomized policies. Denote

$$V_k(x) = \min\big\{J_k(x)\,, Q_{\theta_k,J_k}\big(x,\mu_k(x)\big)\big\}, \qquad x \in S.$$

Recall that with $\theta_k = (\mu_k, S)$, $Q_{\theta_k,J_k} = F_{(\mu_k,S)}(Q_{\theta_k,J_k}; J_k)$ by Prop. 3.2. From this relation and Eq. (3.15), we see that for all $x \in S$,

$$M\big(Q_{\theta_k,J_k}\big)(x) = \inf_{u \in U(x)} \left\{ g(x,u) + \alpha \int_S V_k(x')\, q(dx' \,|\, x,u) \right\} = T(V_k)(x).$$

Since μ_{k+1} is chosen based on either Eqs. (3.11)-(3.12) or Eq. (3.13) (cf. the definition (3.14) of μ_{k+1}), it follows that for $k \geq 0$,

$$p_{k+1}\Big(\big\{x \in S \mid T_{\mu_{k+1}}(V_k)(x) \leq T(V_k)(x) + \epsilon\big\}\Big) = 1, \tag{3.16}$$

where p_{k+1} is the probability measure in Algorithm III. Equation (3.16) says that μ_{k+1} is ϵ-optimal for a set of states with p_{k+1}-measure 1, in the two-stage problem with the terminal second-stage costs given by V_k. This property of the policies $\mu_k, k \geq 1$, bears similarity to the notion of "(p,ϵ)-optimal" policies [12, 47]. By contrast, standard policy iteration cannot operate with policies like μ_k, if they are not ϵ-optimal but only optimal in a "(p,ϵ)-sense" for the optimization problems involved in policy improvement.

It is also clear that we cannot obtain J^* by policy iteration with Borel measurable policies if for some state, there exists no *stationary*, ϵ-optimal Borel measurable policy. This can happen even in finite-state countable-control problems; see e.g., [47, Example 6.1]. Similarly, if J^* is not Borel measurable, we cannot obtain J^* by policy iteration or modified policy iteration operating with Borel measurable policies, since these algorithms keep the iterates J_k in the set of Borel measurable functions. For an example, see [47, Example 4.1]. By contrast, for Algorithm III we have $J_k \to J^*$ in case (D), as well as in cases (N)(P) under certain initial conditions. In fact, the convergence properties we will establish in Sections 4 and 5 hold for the basic algorithm I, regardless of the choices of μ_k. □

In Algorithms I-III, we repeatedly find, for a universally measurable policy μ, a Borel set $B \subset S$ such that as a function of x, $\mu(du \mid x)$ restricted to B is Borel measurable. As mentioned earlier, it is desirable to have a "large" set B so that a large portion of the policy can be taken into account in the algorithms. We may measure the "largeness" of B with respect to a chosen probability measure p on S (cf. Prop. 3.1(b)). The question is then how to choose the measure p. Let us discuss a natural possibility.

Example 3.1 (Choice of B based on the Markov chain induced by μ)**.** Consider the Markov chain $\{X_k\}$ on $(S, \mathscr{U}(S))$ with state transition kernel $\kappa(dx' \mid x)$ defined by

$$\kappa(D \mid x) = \int_C q(D \mid x, u)\, \mu(du \mid x), \qquad D \in \mathscr{U}(S),$$

where $q(D \mid x, u)$ is the measure of D with respect to the completion of $q(dx' \mid x, u)$. Define recursively the n-step transition kernels: $\kappa^0(dx' \mid x) = \delta_x(dx')$ and

$$\kappa^n(dx' \mid x) = \int_S \kappa^{n-1}(dx' \mid y)\, \kappa(dy \mid x), \qquad n \geq 1.$$

For some probability measure ρ on $(S, \mathscr{U}(S))$ and $\beta \in (0,1)$, let p be the probability measure on $(S, \mathscr{U}(S))$ given by

$$p(D) = (1-\beta) \sum_{n=0}^{\infty} \beta^n \int_S \kappa^n(D \mid x)\, \rho(dx), \qquad D \in \mathscr{U}(S).$$

We then let B be a Borel set in S with $(\mu, B) \in \Theta$ and $p(B) = 1$.

The measure p reflects which sets of states are visited with positive probability under the policy μ if the initial distribution is ρ. In particular, if μ induces a ψ-irreducible Markov chain $\{X_k\}$ with the maximal irreducibility probability measure ψ, then ψ is absolutely continuous with respect to p [34, Prop. 4.2.1(iii)]; if in addition the initial distribution ρ is an irreducibility measure of $\{X_k\}$, then $p = \psi$ [34, Prop. 4.2.2(iv)]. In both cases, $p(B) = 1$ implies that B contains a nonempty absorbing set of states [34, Prop. 4.2.3(ii)], and both the set $S \setminus B$ and the set of states from which $S \setminus B$ is reachable under μ have ψ-measure zero [34, Prop. 4.2.2(iii)]. □

3.3 Some Facts about the Existence of Borel Measurable Policies

In the rest of this section, we discuss some useful facts about the existence of Borel measurable policies, to show a broad application range of the policy iteration-like algorithm III given earlier, which uses Borel measurable policies. Recall that the graph of the control constraint U,

$$\Gamma = \big\{(x,u) \mid x \in S, u \in U(x)\big\} \subset S \times C,$$

is an analytic subset of the product of two Polish spaces (of which S and C are Borel subsets). The question whether a Borel measurable nonrandomized stationary policy exists in our control problem is equivalently whether the set Γ admits a *section* f – a function $f : S \to C$ whose graph lies in Γ (i.e., $f(x) \in U(x)$ for all x), such that f is Borel measurable. Measurable selection theorems concern questions of this type. The Jankov-von Neumman's selection theorem tells us that Γ admits an analytically measurable section.

Suppose Γ is a Borel subset of $S \times C$. It can still happen that Γ has no Borel measurable section [14]. Then, there exists no Borel measurable stationary policy, randomized or nonrandomized. (Because if a randomized Borel measurable stationary policy were to exist, a nonrandomized one must also exist by the selection theorem of Blackwell and Ryll-Nardzewski [17].) Nevertheless, when Γ is Borel, a number of selection theorems for Borel sets in the product of two Polish spaces can be applied to assert the existence of a Borel measurable section of Γ, under fairly general conditions on the control constraint U (see e.g., [46]). We give below several examples.

Let Y be the Polish space of which C is a Borel subset. Assume Γ is Borel. In each of the following cases, a Borel measurable nonrandomized stationary policy exists:

(a) For every x, $U(x)$ is a countable set (by a theorem of Lusin, [46, Theorem 5.8.11]).

(b) For every x, $U(x)$ contains a nonempty open set in Y (by theorems of Kechris and Sarbadhikari, [46, Theorem 5.8.5]).

(c) For every x, $U(x)$ is a σ-compact set in Y (by a theorem of Arsenin and Kunugui, [46, Theorem 5.12.1]), which is true, in particular when Y is σ-compact and each $U(x)$ is a countable union of open or closed sets in Y.

(d) U is a Borel measurable multifunction (i.e., set-valued function) and for every x, $U(x)$ is a closed set in Y (by Kuratowski and Ryll-Nardzewski's selection theorem, [46, Theorem 5.2.1]).

These examples illustrate that for many general classes of control constraints U, the policy iteration-like algorithm III, which operates with Borel measurable policies, can be applied.

4 Convergence Analysis for Discounted Case (D) and Non-positive Case (N)

In this section, we analyze the convergence of the mixed value and policy iteration algorithms given in Section 3.2 for cases (D) and (N). We state convergence results for the basic algorithm (3.9)-(3.10), since the two other policy iteration-like algorithms are its special cases.

4.1 Discounted Case (D)

In the discounted case (D), we work with bounded functions. Let $\mathcal{M}_b(S)$ and $\mathcal{M}_b(\Gamma)$ denote the vector spaces of bounded universally measurable functions on S and Γ respectively. With the supremum norm $\|\cdot\|_\infty$, defined for $f \in \mathcal{M}_b(S)$ or $\mathcal{M}_b(\Gamma)$ by $\|f\|_\infty = \sup_y |f(y)|$, $\mathcal{M}_b(S)$ and $\mathcal{M}_b(\Gamma)$ are Banach spaces. Note that $A_b(S)$, $A_b(\Gamma)$ (the sets of bounded, lower semi-analytic functions) are closed subsets of $\mathcal{M}_b(S)$, $\mathcal{M}_b(\Gamma)$, respectively, and endowed with the metric $d_{sup}(f,f') = \|f - f'\|_\infty$,

the spaces $(A_b(S), d_{sup})$ and $(A_b(\Gamma), d_{sup})$ are complete. Our mixed value and policy iteration algorithms work on the product space $A_b(S) \times A_b(\Gamma)$ endowed with the metric

$$d\big((J,Q),(J',Q')\big) = \|(J,Q)-(J',Q')\|_\infty := \max\big\{\|J-J'\|_\infty, \|Q-Q'\|_\infty\big\},$$

which is also a complete metric space. Our convergence analysis below parallels the one given in our earlier work [10] for discounted finite-state and control problems.

Lemma 4.1. (D) *Let $\theta \in \Theta$, $J, J' \in A_b(S)$, and $Q, Q' \in A_b(\Gamma)$.*

(a) We have

$$\|F_\theta(Q\,;J) - F_\theta(Q'\,;J')\|_\infty \le \alpha \max\big\{\|J-J'\|_\infty,\, \|Q-Q'\|_\infty\big\},$$
$$\|F_\theta(Q\,;J) - Q^*\|_\infty \le \alpha \max\big\{\|J-J^*\|_\infty,\, \|Q-Q^*\|_\infty\big\}.$$

(b) The function $Q_{\theta,J} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,;J)$ is the unique solution to $Q = F_\theta(Q\,;J), Q \in A_b(\Gamma)$. Moreover,

$$\|Q_{\theta,J} - Q^*\|_\infty \le \alpha \|J-J^*\|_\infty.$$

Proof. (a) For every $(x,u) \in \Gamma$,

$$J(x) \le J'(x) + \max\big\{\|J-J'\|_\infty,\, \|Q-Q'\|_\infty\big\}, \quad Q(x,u) \le Q'(x,u) + \max\big\{\|J-J'\|_\infty,\, \|Q-Q'\|_\infty\big\},$$

so

$$\min\big\{J(x), Q(x,u)\big\} - \min\big\{J'(x), Q'(x,u)\big\} \le \max\big\{\|J-J'\|_\infty, \|Q-Q'\|_\infty\big\}$$

and by symmetry,

$$\big|\min\big\{J(x), Q(x,u)\big\} - \min\big\{J'(x), Q'(x,u)\big\}\big| \le \max\big\{\|J-J'\|_\infty, \|Q-Q'\|_\infty\big\}.$$

Using the above inequality and the definition of F_θ given in Eq. (3.4), a direct calculation then shows that for each $(x,u) \in \Gamma$,

$$\begin{aligned}\Big|F_\theta(Q;J)(x,u) - F_\theta(Q';J')(x,u)\Big| &\le \alpha\,\|J-J'\|_\infty \cdot q(S\setminus B \mid x,u) \\ &\quad + \alpha \max\big\{\|J-J'\|_\infty,\, \|Q-Q'\|_\infty\big\} \cdot q(B \mid x,u) \\ &\le \alpha \max\big\{\|J-J'\|_\infty,\, \|Q-Q'\|_\infty\big\}.\end{aligned}$$

This proves the first inequality in part (a). The second inequality is proved by setting $Q' = Q^*$ and $J' = J^*$, and using the fact $F_\theta(Q^*;J^*) = Q^*$ (Prop. 3.3(a)).

(b) Part (a) implies $\|F_\theta(Q\,;J) - F_\theta(Q';J)\|_\infty \le \alpha\|Q-Q'\|_\infty$, so by Banach's contraction principle [39, p. 220], the equation $Q = F_\theta(Q\,;J), Q \in A_b(\Gamma)$, has a unique solution $\bar{Q}$, and $F_\theta^k(Q\,;J) \to \bar{Q}$ for any $Q \in A_b(\Gamma)$. This shows $Q_{\theta,J} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,;J) = \bar{Q}$ and $Q_{\theta,J} = F_\theta(Q_{\theta,J}\,;J)$. Letting $Q = Q_{\theta,J}$ in the second inequality in part (a), we then have

$$\|Q_{\theta,J} - Q^*\| \le \alpha \max\big\{\|J-J^*\|_\infty, \|Q_{\theta,J}-Q^*\|_\infty\big\}.$$

Since $\alpha < 1$, this is equivalent to $\|Q_{\theta,J} - Q^*\| \le \alpha\|J-J^*\|_\infty$. □

Theorem 4.1. (D) *For any $J_0 \in A_b(S)$ and $Q_0 \in A_b(\Gamma)$, the sequence $\{(J_k, Q_k)\}$ generated by the iteration (3.9)-(3.10) converges to (J^*, Q^*), and*

$$\big\|(J_k,Q_k) - (J^*,Q^*)\big\|_\infty \le \alpha^k \big\|(J_0,Q_0) - (J^*,Q^*)\big\|_\infty.$$

Proof. At iteration k, either $Q_{k+1} = F_\theta^n(Q_k\,;J_k)$ or $Q_{k+1} = Q_{\theta,J_k}$ for some $\theta \in \Theta, n \geq 1$. For the first case, applying the second inequality in Lemma 4.1(a) n times, we have

$$\big\|F_\theta^n(Q_k\,;J_k) - Q^*\big\|_\infty \leq \alpha \max\big\{\|J_k - J^*\|_\infty\,,\, \alpha^{n-1}\|Q_k - Q^*\|_\infty\big\},$$

whereas for the second case, $\|Q_{\theta,J_k} - Q^*\| \leq \alpha\|J_k - J^*\|_\infty$ by Lemma 4.1(b). Thus in either case,

$$\|Q_{k+1} - Q^*\|_\infty \leq \alpha \max\big\{\|J_k - J^*\|_\infty\,,\, \|Q_k - Q^*\|_\infty\big\}.$$

Since $J_{k+1} = M(Q_{k+1})$, $J^* = M(Q^*)$, and M is nonexpansive, i.e., $\|M(Q) - M(Q')\|_\infty \leq \|Q - Q'\|_\infty$, we have

$$\|J_{k+1} - J^*\|_\infty = \|M(Q_{k+1}) - M(Q^*)\|_\infty \leq \alpha \max\big\{\|J_k - J^*\|_\infty\,,\, \|Q_k - Q^*\|_\infty\big\}.$$

Combining the preceding two inequalities, we obtain

$$\big\|(J_{k+1}, Q_{k+1}) - (J^*, Q^*)\big\|_\infty \leq \alpha^{k+1}\big\|(J_0, Q_0) - (J^*, Q^*)\big\|_\infty,$$

which is the desired inequality and implies $(J_k, Q_k) \to (J^*, Q^*)$. □

Remark 4.1. Finally, let us note that given the sequence $\{J_k\}$ generated by the algorithm, we may extract an asymptotically near-optimal sequence of policies $\{\nu_k\}$ by using the selection theorem of [7, Prop. 7.50]: for some $\epsilon > 0$, choose universally measurable stationary policies ν_k such that

$$\big\|T_{\nu_k}(J_k) - T(J_k)\big\|_\infty \leq \epsilon, \qquad \forall\, k \geq 1.$$

Using the contraction property of T_{ν_k} and T, it can be shown (see e.g., [6, p. 45]) that

$$\|J_{\nu_k} - J^*\|_\infty \leq \frac{\epsilon}{1-\alpha} + \frac{2\alpha\,\|J_k - J^*\|_\infty}{1-\alpha}, \qquad \forall\, k \geq 1.$$

For the policy iteration-like algorithm II (resp. III) in particular, the sequence of policies $\{\mu_k\}$ (resp. $\{\mu'_k\}$) generated by the algorithm is asymptotically $\epsilon/(1-\alpha)$-optimal. □

4.2 Nonpositive Case (N)

In case (N) the one-stage cost function $g \leq 0$ and $J^* \leq 0$, $Q^* \leq 0$. The mixed value and policy iteration algorithms operate with nonpositive lower semi-analytic functions in $A_-(S)$ and $A_-(\Gamma)$. We will rely on the monotonicity and fixed point properties of F_θ to ensure their convergence.

First, we derive some simple upper and lower bounds on the iterates generated by the algorithms. To simplify notation, let

$$H(x,u,J) = g(x,u) + \int_S J(x')\, q(dx' \,|\, x,u), \qquad (x,u) \in \Gamma. \tag{4.1}$$

Expressed in these terms, $T(J)(x) = \inf_{u\in U(x)} H(x,u,J)$, the optimality equation $J^* = T(J^*)$ is

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

and by the definition of Q^* (cf. Eq. (3.1)),

$$Q^*(x,u) = H(x,u,J^*), \qquad (x,u) \in \Gamma. \tag{4.2}$$

For (D)(N)(P), the functions $F_\theta(Q\,;J)$ and $Q_{\theta,J}$ can be upper bounded simply by

$$F_\theta(Q\,;J)(x,u) \le H(x,u,J), \qquad Q_{\theta,J}(x,u) \le H(x,u,J), \quad \forall\,(x,u)\in\Gamma. \tag{4.3}$$

To derive the first inequality above, we upper bound the term $\min\{J(x'),Q(x',u')\}$ by $J(x')$ in the definition of $F_\theta(Q\,;J)(x,u)$. To derive the second inequality above, we apply the first one to $Q_{\theta,J}=F_\theta(Q_{\theta,J}\,;J)$ (Prop. 3.2). By minimizing over $U(x)$ for each x in Eq. (4.3), we see that

$$M\big(F_\theta(Q\,;J)\big) \le T(J), \qquad M\big(Q_{\theta,J}\big) \le T(J). \tag{4.4}$$

We use these bounds to bound the iterates of the algorithms. The next lemma applies also to (D)(P). For the algorithm that uses the second rule of (3.10) to set $Q_{k+1}=Q_{\theta_k,J_k}$ at some iterations, the second statement of the lemma will rely on Prop. 3.3(c), which in the case (P) will be proved in Appendix B as Prop. B.1.

Lemma 4.2. (N)(P) *Let $\{(J_k,Q_k)\}$ be iterates generated by the iteration (3.9)-(3.10) with $J_0\in A_-(S)$, $Q_0\in A_-(\Gamma)$ in case (N) and with $J_0\in A_+(S)$, $Q_0\in A_+(\Gamma)$ in case (P). Then for $k\ge 1$,*

$$J_k \le T^k(J_0), \qquad Q_k(x,u) \le H(x,u,J_{k-1}), \quad \forall\,(x,u)\in\Gamma. \tag{4.5}$$

If $J_0\ge J^*, Q_0\ge Q^*$, then we also have $J_k\ge J^*, Q_k\ge Q^*$.

Proof. For each $k\ge 0$, either $Q_{k+1}=F_\theta^n(Q_k\,;J_k)$ or $Q_{k+1}=Q_{\theta,J_k}$ for some $\theta\in\Theta, n\ge 1$. By Eq. (4.3), the right-hand side inequality for Q_k in Eq. (4.5) follows. Since $J_{k+1}=M(Q_{k+1})$, we have, by Eq. (4.4), $J_{k+1}\le T(J_k)$ for all k. This implies $J_k\le T^k(J_0)$ by the monotonicity of T.

Let $J_0\ge J^*$ and $Q_0\ge Q^*$. We show by induction that $J_k\ge J^*, Q_k\ge Q^*$ for every k. Suppose it holds for some $k\ge 0$. Then either $Q_{k+1}=F_\theta^n(Q_k\,;J_k)$, in which case, by the induction hypothesis, the monotonicity of F_θ (cf. Eq. (3.7)) and Prop. 3.3(a), we have

$$Q_{k+1}=F_\theta^n(Q_k\,;J_k)\ge F_\theta^n(Q^*\,;J^*)=Q^*;$$

or $Q_{k+1}=Q_{\theta,J_k}$, in which case $Q_{k+1}\ge Q^*$ by the induction hypothesis and Prop. 3.3(c) (proved as Prop. B.1 for (P)). Thus in either case, $Q_{k+1}\ge Q^*$. Hence $J_{k+1}=M(Q_{k+1})\ge M(Q^*)=J^*$. □

The relation $J^*\le J_k\le T^k(J_0)$ in Lemma 4.2, which holds when $J_0\ge J^*$, is the key to our convergence analysis for cases (N) and (P). It implies that our method converges to J^* from above whenever the ordinary value iteration method does. In case (N), we will exploit the generic convergence property of value iteration in the following theorem, whereas in case (P), we will derive sufficient conditions for convergence of value iteration from above in the next section.

Theorem 4.2. (N) *For any $J_0\in A_-(S)$ and $Q_0\in A_-(\Gamma)$ such that $J_0\ge J^*$ and $Q_0\ge Q^*$, the sequence $\{(J_k,Q_k)\}$ generated by the iteration (3.9)-(3.10) converges to (J^*,Q^*).*

Proof. We show first $J_k\to J^*$. We have $J^*\le J_k\le T^k(J_0)$ by Lemma 4.2. Since $J^*\le J_0\le 0$ by assumption and $T^k(\mathbf{0})\downarrow J^*$ under (N), we have $T^k(J_0)\to J^*$ and hence $J_k\to J^*$. Then, for each $(x,u)\in\Gamma$, by Fatou's lemma [19, p. 131] (applied to nonpositive functions),

$$\limsup_{k\to\infty} H(x,u,J_k) \le H\big(x,u,\limsup_{k\to\infty} J_k\big) = H(x,u,J^*) = Q^*(x,u)$$

(cf. Eqs. (4.1)-(4.2)). Since $Q^*(x,u)\le Q_{k+1}(x,u)\le H(x,u,J_k)$ by Lemma 4.2, this implies the convergence $Q_k\to Q^*$. □

Remark 4.2. Regarding near-optimal policies in case (N), recall that they are guaranteed to exist among semi-Markov policies, but not necessarily among stationary or Markov policies. The construction of an ϵ-optimal semi-Markov policy under (N) is much more involved than under (D)(P), and knowing the optimal cost function J^* alone is insufficient (see the proof of [7, Prop. 9.20]), even if it was available. Moreover, even if an optimal stationary policy exists, it is possible that a policy μ satisfies $T_\mu(J^*) = T(J^*)$ without being optimal.[3] Hence, we do not expect to have simple ways to obtain an asymptotically near-optimal sequence of policies from the iterate sequence $\{J_k\}$ generated by our algorithm. Intuitively, it seems possible to us to construct history-dependent or semi-Markov policies that are asymptotically near-optimal for each given state, by using the relations between the optimal stopping problems and the original problem. Due to its complexity, however, we do not discuss this subject in this paper. □

5 Convergence Analysis for Nonnegative Case (P)

In this section we consider the case (P) with nonnegative one-stage costs. We first prove a new convergence theorem for value iteration in Section 5.1. Using this theorem, we then derive in Section 5.2 convergence results for the mixed value and policy iteration algorithms discussed in Section 3.2, and for another variant algorithm which admits a linear programming implementation for a certain class of problems and thus has computational advantages.

Recall that $A_+(S)$ denotes the set of nonnegative, lower semi-analytic functions. The symbol $\mathbf{0}$ stands for the constant function zero.

5.1 A Convergence Theorem for Value Iteration

The nonpositive case (P) is more complex than (D)(N). Neither value iteration nor policy iteration are guaranteed to give us J^*, even if policy iteration encounters no measurability issues. For value iteration, as mentioned in Section 2.3, for some $J_\infty \in A_+(S)$, we have

$$T^k(\mathbf{0}) \uparrow J_\infty \leq J^*,$$

and it is possible that $J_\infty < J^*$. It is known that $J_\infty = J^*$ if $U(x)$ is a finite set for each $x \in S$, or more generally, if a compactness-type condition on the control constraint set holds [7, Prop. 9.17, Cor. 9.17.1]; but these conditions are restrictive. For policy iteration, it can happen that for a suboptimal stationary policy μ,

$$T_\mu(J_\mu) = T(J_\mu),$$

even in finite-state and control problems,[4] and the method terminates with the suboptimal policy μ.

We thus look for ways to mitigate the difficulties. Any condition forcing $T^k(\mathbf{0}) \uparrow J^*$, however, seems restrictive, in view of Maitra and Sudderth's result [33]. They showed that J^* can be obtained by applying T a transfinite number of times, starting from the function $J \equiv 0$, and in general, the number of times needed can be uncountably infinite [33, p. 930]. This led us to consider ways to make value iteration converge from above instead of from below, which is also natural when using policy costs, since $J_\mu \geq J^*$. We will modify Whittle's bridging condition [55, 26] to suit our purpose.

[3] As an example, let $S = \{0, 1\}$ with state 0 being cost-free and absorbing. At state 1, there are two controls: control 1 leads to state 1 with cost 0, and control 0 leads to state 0 with cost -1. Then $J^*(0) = 0$, $J^*(1) = -1$, and the suboptimal policy μ that makes self-transitions at state 1 satisfies $T_\mu(J^*) = T(J^*)$.

[4] For a simple example, consider a problem with two states $\{0, 1\}$. State 0 is cost-free and absorbing. State 1 has two controls $\{0, 1\}$: the control 1 leads to a zero-cost self-transition to state 1, and the control 0 leads to state 0 with cost 1. Then the nonrandomized stationary policy μ with $\mu(1) = 0$ is suboptimal but satisfies $T_\mu(J_\mu) = T(J_\mu)$. See [38, Example 7.3.4] for a similar example. We also note that total cost finite-state and control problems can be solved by using the policy iteration algorithms of Veinott [52] and of Miller and Veinott [35] based on the concept of sensitive optimality ([53]; see also [38, Sec. 10.3]).

Before proceeding, let us give a simple example to exemplify the behavior of value iteration just discussed. The example is from [7, p. 215]. In this example $J_\infty \equiv 0 < J^* \equiv \infty$. We illustrate how value iteration with transfinite recursion is able to obtain J^* in the end, after countably many iterations. This example falls into a special case analyzed in [33, Sec. 5], which predicted, for a broad class of problems, that the number of iterations required for value iteration to converge from below is at most countably infinite.

Example 5.1. The state and control spaces are $S = \{0, 1, 2, \ldots\}, C = \{1, 2, \ldots\}$, and the control constraint is $U(x) = C$ for every $x \in S$. State transitions are deterministic and uncontrolled except at state 0: applying control u at state x, the successor state is u if $x = 0$ and $x - 1$ if $x \geq 1$. The one-stage cost is zero except at state 1: $g(1, u) = 1$ for all u. Write a function J on S in vector form as $J = \big(J(0), J(1), \ldots\big)$. The optimal cost function is $J^* = (\infty, \infty, \ldots)$ because under any policy, the system will visit state 1 infinitely often and accumulate one more unit of cost at each visit.

The pointwise limit J_∞ of $\{T^k(\mathbf{0})\}$ is $J_\infty = (0, 1, 1, \ldots)$, since $T^k(\mathbf{0}) = (0, 1, 1, \ldots, 1, 0, 0, \ldots)$ with k 1's followed by all 0's. As in [30], set $J_{\infty 0} = J_\infty$ and initiate value iteration with it. This gives us $J_{\infty 1} = \lim_{k\to\infty} T^k(J_{\infty 0})$, which is $J_{\infty 1} = (1, 2, 2, \ldots)$. Continuing in this way, we define recursively $J_{\infty(m+1)} = \lim_{k\to\infty} T^k(J_{\infty m})$ and we get $J_{\infty(m+1)} = \big(m, m+1, m+1, \ldots\big) = J_{\infty m} + 1$. In the end, from the pointwise limit of the nondecreasing sequence $\{J_{\infty m}\}$ we obtain J^*. □

We now proceed to place a condition on the initial function J_0 for value iteration $T^k(J_0)$, to ensure the convergence of value iteration (from above, primarily) to J^*. This condition, given in the following theorem, is motivated by Whittle's bridging condition [55, 26] (cf. Remark 5.3) and its appealingly simple form. (The paper [55] called J_0 the "terminal function" instead of "initial function," for the reason that J_0 can be viewed as setting the terminal costs for finite horizon problems.) The implications of our theorem given below are, however, different from Whittle's [55, 26], as we will remark shortly.

Theorem 5.1. (P) *(a) For any $c > 1$, $T^k(cJ^*) \downarrow J^*$.*

(b) $T^k(J) \to J^*$ for all $J \in A_+(S)$ such that

$$\underline{J} \leq J \leq cJ^*, \quad \text{for some } c > 1,$$

where $\underline{J} \in A_+(S)$ satisfies $\underline{J} \leq J^*, T^k(\underline{J}) \to J^*$. In particular, if $T^k(\mathbf{0}) \uparrow J^*$, then $T^k(J) \to J^*$ for all $J \leq cJ^*, J \in A_+(S)$.

(c) J^* is the unique fixed point of T within the set $\{J \in A_+(S) \mid J \leq cJ^* \text{ for some } c > 1\}$.

We note that Theorem 5.1(b)-(c) follows directly from Theorem 5.1(a). To see this, suppose part (a) is proved. Then under the assumptions of part (b), we have $T^k(\underline{J}) \leq T^k(J) \leq T^k(cJ^*)$ by the monotonicity of T. Since $T^k(\underline{J}) \to J^*$ by assumption and $T^k(cJ^*) \downarrow J^*$ by part (a), part (b) follows. For part (c), by [7, Prop. 9.10(P)] we have the following implication,

$$J \in A_+(S),\ J = T(J) \qquad \Longrightarrow \qquad J \geq J^*,$$

which together with part (a) implies the conclusion of part (c). Thus to prove Theorem 5.1, it suffices to prove its part (a).

Before giving the proof, let us make several remarks about the implications of Theorem 5.1 and its relation with Whittle's bridging condition.

Remark 5.1. In Theorem 5.1(b), we can always let $\underline{J} = J^*$. Then Theorem 5.1(b) reads as:

$$T^k(J) \to J^*, \qquad \forall\, J \text{ s.t. } J^* \leq J \leq cJ^*,\ c > 1. \tag{5.1}$$

Indeed, in view of the result of Maitra and Sudderth [33] and the simple Example 5.1, among the functions obtainable with (transfinite) value iteration starting from the constant function 0, J^* may be the only function that can serve as $\underline{J}$ in Theorem 5.1(b). □

Remark 5.2. Theorem 5.1(a)-(b) roughly says that value iteration converges to J^* if the initial function J is "commensurate" with J^*. In particular, if $J \geq J^*$, then on the set of states x with finite $J^*(x)$, the shape of J must be "compatible" with that of J^*, with $J(x) = 0$ whenever $J^*(x) = 0$. The theorem also implies that whenever the policy iteration algorithm gets stuck at a suboptimal policy μ with $T_\mu(J_\mu) = T(J_\mu)$, J_μ must have a "wrong shape" relative to J^*.

Of course it can be difficult to know even the "shape" of J^*. In Example 5.1, for instance, $J^* \equiv \infty$, so the only function between J^* and cJ^*, $c > 1$, is J^* itself. In an example of Strauch [47, p. 881] (see also [33, p. 930]), J^* takes values in $\{0, 1\}$ and $T^k(\mathbf{0}) \not\to J^*$. The set $\{x \in S \mid J^*(x) = 0\}$ is rather intricate. If we know this set of states, then with any initial function J that takes the value 0 on this set and the value $a > 1$ elsewhere, value iteration turns out to converge in one iteration in this example (see Appendix C). □

Remark 5.3. Whittle's bridging condition is as follows: for some real c and stationary policy μ, either $J_\mu \leq cT^n(\mathbf{0})$ for some n, or $J_\mu \leq cJ_\infty$ and $J_\infty = T(J_\infty)$. The condition implies that $J_\infty = J^*$ and $T^k(J) \to J^*$ for all $J \in A_+(S)$ with $J \leq aJ^*$ for some real a [55, 26]. A similar, slightly weaker condition leading to the same conclusions is $J^* \leq cT^n(\mathbf{0})$ for some n (personal communication with E. Feinberg). The main difference between these results and Theorem 5.1 is that Theorem 5.1 does not place any condition on the model of the control problem. Instead, it restricts only the initial function for value iteration and it holds for all nonnegative control models. If the bridging condition or any other condition for $T^k(\mathbf{0}) \uparrow J^*$ holds, they can be used to set $\underline{J} \equiv 0$ in the theorem, as stated in Theorem 5.1(b). Then, the condition for J becomes $0 \leq J \leq cJ^*$, the same as in [55, 26]. □

We now proceed to prove Theorem 5.1. As discussed earlier, it suffices to prove Theorem 5.1(a). To this end, we start with two lemmas to characterize the pointwise limit of $\{T^k(cJ^*)\}$. The first lemma below is a basic fact; the second one is important for our proof.

Lemma 5.1. *If $J \in A_+(S)$ satisfies $T(J) \leq J$, then for some $J^\infty \in A_+(S)$, we have*

$$T^k(J) \downarrow J^\infty \quad \textit{and} \quad T(J^\infty) \leq J^\infty.$$

Proof. By the monotonicity of T, $T^k(J) \downarrow J^\infty$. For every k, since $J^\infty \leq T^k(J)$, we have, by the monotonicity of T, $T(J^\infty) \leq T^{k+1}(J)$. Hence $T(J^\infty) \leq J^\infty$. □

Lemma 5.2. *Let $c > 1$. Then we have $T(cJ^*) \leq cJ^*$ and for some $J^\infty \in A_+(S)$,*

$$T^k(cJ^*) \downarrow J^\infty, \qquad T(J^\infty) = J^\infty, \qquad J^* \leq J^\infty \leq cJ^*.$$

Proof. Since $c > 0$ and $J^* \in A_+(S)$, $cJ^* \in A_+(S)$. Since $c > 1$ and the one-stage costs are nonnegative, it follows from the definition of T that $T(cJ^*) \leq cJ^*$. Let $J^k = T^k(cJ^*)$. By Lemma 5.1,

$$cJ^* \geq J^k \downarrow J^\infty \geq J^* \quad \text{and} \quad T(J^\infty) \leq J^\infty,$$

where the inequality $J^\infty \geq J^*$ follows from the monotonicity of T and the fact $T(J^*) = J^*$. By rearranging the terms and using also the monotonicity of T, we have

$$cJ^* \geq J^\infty \geq T(J^\infty) \geq J^*.$$

To prove $T(J^\infty) = J^\infty$, we now show $T(J^\infty) \geq J^\infty$, using the monotone convergence theorem. Consider each $x \in S$. If $J^*(x) = \infty$, then $T(J^\infty)(x) = J^\infty(x) = \infty$. Suppose $J^*(x) < \infty$; we prove $T(J^\infty)(x) \geq J^\infty(x)$ below.

To simplify notation, for each $u \in U(x)$, denote

$$H(x, u, J) = g(x, u) + \int_S J(x')\, q(dx' \mid x, u).$$

Then $T(J^*)(x) = \inf_{u\in U(x)} H(x,u,J^*)$ (cf. Eq. (2.3)). Since $T(J^*)(x) = J^*(x) < \infty$, we have

$$D(x) := \big\{u \in U(x) \mid H(x,u,J^*) < \infty\big\} \neq \emptyset.$$

For $u \in D(x)$,

$$H(x,u,cJ^*) \le cH(x,u,J^*) < \infty$$

(because $c > 1$ and $g \ge 0$), so in view of the relation $cJ^* \ge J^k \downarrow J^\infty$, we have by the monotone convergence theorem [19, p. 131],

$$H(x,u,J^\infty) = \lim_{k\to\infty} H(x,u,J^k). \tag{5.2}$$

Consequently,

$$H(x,u,J^\infty) \ge \limsup_{k\to\infty} \left\{\inf_{u\in U(x)} H(x,u,J^k)\right\} = \lim_{k\to\infty} T(J^k)(x) = J^\infty(x), \qquad \forall\, u \in D(x). \tag{5.3}$$

For $u \in U(x) \setminus D(x)$,

$$H(x,u,J^\infty) \ge H(x,u,J^*) = \infty.$$

Combining this with Eq. (5.3), we have

$$T(J^\infty)(x) = \inf_{u\in U(x)} H(x,u,J^\infty) = \inf_{u\in D(x)} H(x,u,J^\infty) \ge J^\infty(x).$$

This completes the proof. □

We are now ready to prove Theorem 5.1. We will use a simple concavity property of T, which can be verified directly. (Hartley [26] also used it in an alternative proof of Whittle's bridging condition.) On the convex set $A_+(S)$, T has the property that for any $\beta \in [0,1]$ and $J_1, J_2 \in A_+(S)$,

$$T\big(\beta J_1 + (1-\beta)J_2\big) \ge \beta\, T(J_1) + (1-\beta)\, T(J_2). \tag{5.4}$$

We will also use Maitra and Sudderth's results [33]. Let ω_1 be the first uncountable ordinal. For ordinals $\xi < \omega_1$, define functions $J^\xi \in A_+(S)$ by transfinite recursion as follows. Let

$$J^0 = T(\mathbf{0}), \qquad J^\xi = T\Big(\sup_{\eta<\xi} J^\eta\Big), \quad \text{for } \xi > 0.$$

Also let

$$J^{\omega_1} = \sup_{\xi<\omega_1} J^\xi.$$

That all these functions are indeed in $A_+(S)$ is proved in [33]. Moreover, Maitra and Sudderth [33, Thm. 1.1] proved that

$$T(J^{\omega_1}) = J^{\omega_1} = J^*. \tag{5.5}$$

(For ordinals, transfinite induction and transfinite recursion, see e.g., [31, p. 27-28], [19, Secs. 1.3, A.3] or [46, Chap. 1].)

Proof of Theorem 5.1. Denote $J^\infty = \lim_{k\to\infty} T^k(cJ^*)$. By Lemma 5.2,

$$T(J^\infty) = J^\infty, \qquad J^* \le J^\infty \le cJ^*.$$

We now prove $J^\infty = J^*$. This will prove the theorem, as discussed earlier.

Let $\beta = 1/c < 1$. Since $J^\infty \le cJ^*$, $J^* \ge \beta J^\infty$, so by the monotonicity and concavity properties of T (Eq. (5.4)),

$$T(J^*) \ge T\big(\beta J^\infty + (1-\beta)\,\mathbf{0}\big) \ge \beta\, T(J^\infty) + (1-\beta)\, T(\mathbf{0}).$$

Since $T(J^\infty) = J^\infty$ and $T(J^*) = J^*$, using the definition $J^0 = T(\mathbf{0})$, we can write the above inequality equivalently as

$$J^* \ge \beta J^\infty + (1-\beta) J^0.$$

We now apply transfinite induction. Suppose that for an ordinal $\xi \le \omega_1$,

$$J^* \ge \beta J^\infty + (1-\beta) J^\eta, \qquad \forall\, \eta < \xi.$$

Then

$$J^* \ge \sup_{\eta<\xi} \Big\{ \beta J^\infty + (1-\beta) J^\eta \Big\} = \beta J^\infty + (1-\beta) \sup_{\eta<\xi} J^\eta.$$

Consequently, by the monotonicity and concavity properties of T,

$$J^* \ge \beta J^\infty + (1-\beta)\, T\Big(\sup_{\eta<\xi} J^\eta \Big) = \beta J^\infty + (1-\beta) J^\xi, \tag{5.6}$$

where in the first inequality we also used the fact $T(J^\infty) = J^\infty$ and $T(J^*) = J^*$, and in the equality we used the definition of J^ξ for $\xi < \omega_1$, and the definition of J^{ω_1} for $\xi = \omega_1$, together with the fact $T(J^{\omega_1}) = J^{\omega_1}$ ([33, Thm. 1.1]; cf. Eq. (5.5)). This proves, by transfinite induction, that the inequality (5.6) holds for all $\xi \le \omega_1$, and in particular,

$$J^* \ge \beta J^\infty + (1-\beta) J^{\omega_1}. \tag{5.7}$$

Since $J^{\omega_1} = J^*$ by [33, Thm. 1.1] (cf. Eq. (5.5)) and $J^* \ge 0$, we have $J^* \ge J^\infty$ by Eq. (5.7). We also have $J^\infty \ge J^*$. Therefore, $J^\infty = J^*$. □

We mention two immediate implications of Theorem 5.1. The first one follows from and sharpens slightly Theorem 5.1(b).

Corollary 5.1. (P) *Suppose $J \in A_+(S)$ is such that $\underline{J} \le T^n(J) \le cJ^*$ for some $c > 1$ and $n \ge 1$, where $\underline{J}$ is as in Theorem 5.1(b). Then $T^k(J) \to J^*$.*

Corollary 5.2. (P) *Suppose that the state space S is finite and J^* is real-valued. Let $\mathcal{V}$ be the set of nonnegative, real-valued functions J such that $J(x) = 0$ for all $x \in S$ with $J^*(x) = 0$. Then J^* is the unique fixed point of T within $\mathcal{V}$. Moreover, $T^k(J) \to J^*$ for all $J \in \mathcal{V}$.*

Proof. By Theorem 5.1(c), J^* is the unique fixed point of T in $\{J \ge 0 \mid \exists\, c \in \Re_+ \text{ s.t. } J \le cJ^*\} = \mathcal{V}$. By Theorem 5.1(b), we have $T^k(J) \to J^*$ for all $J \in \mathcal{V}$ if $T^k(\mathbf{0}) \uparrow J^*$. The latter holds when S is finite and J^* is finite everywhere (cf. [38, Thm. 7.3.10(a)]). The reason is that $T^k(\mathbf{0})$ converges to its limit J_∞ uniformly, i.e., for any $\epsilon > 0$, $\|T^k(\mathbf{0}) - J_\infty\|_\infty \le \epsilon$ for all k sufficiently large. Thus with $\mathbf{1}$ denoting the constant function 1, we have, by the monotonicity of T, that for all k sufficiently large,

$$T(J_\infty) \le T\big(T^k(\mathbf{0}) + \epsilon \mathbf{1}\big) \le T^{k+1}(\mathbf{0}) + \epsilon \mathbf{1}.$$

Since ϵ is arbitrary, this implies $T(J_\infty) \le J_\infty$. Since $J_\infty \le T(J_\infty)$ [7, Prop. 9.16], we have $J_\infty = T(J_\infty)$ and hence by [7, Prop. 9.16], $J_\infty = J^*$, i.e., $T^k(\mathbf{0}) \uparrow J^*$. □

In connection with Cor. 5.2, we note that even when S is finite, $J_\infty \neq J^*$ is possible if J^* is not real-valued. As an example, let $S = \{0, 1, 2\}$, $C = (0, 1) \cup \{t\}$ and $U(x) = C$ for all $x \in S$. States 0, 1 are absorbing; the self-transition costs are $g(0, u) = 0$ and $g(1, u) = 1$ for all u (so $J^*(0) = 0$, $J^*(1) = \infty$). At state 2, for control $u \in (0, 1)$, the one-stage cost is $g(2, u) = 0$ and the next state is state 1 with probability u and state 0 with probability $(1 - u)$; and for control $u = t$, $g(2, t) = 1$ and the next state is state 0. Then $T^k(\mathbf{0})(2) = 0$ for all k, so $J_\infty(2) = 0$; but $J^*(2) = 1$.

Finally, we remark that when the problem does not admit a near-optimal stationary policy for some state, one cannot hope to find an initial function J with the desired property $J \leq cJ^*$ for some $c > 1$, among the cost functions of stationary policies. This is shown in the following proposition.

Proposition 5.1. (P) *Suppose that for some $\bar{x} \in S$ and $\epsilon > 0$, an ϵ-optimal stationary policy for $\bar{x}$ does not exist. Then these exists no stationary policy μ such that $J_\mu \in A_+(S)$, $T^k(J_\mu) \to J^*$ (hence $\nexists\, J_\mu \in A_+(S)$ with $J_\mu \leq cJ^*$ for some $c > 1$).*

Proof. To arrive at a contradiction, suppose μ is a stationary policy with $J_\mu \in A_+(S)$, $T^k(J_\mu) \to J^*$. Let k be large enough so that $T^k(J_\mu)(\bar{x}) \leq J^*(\bar{x}) + \epsilon/2$. By the selection theorem of [7, Prop. 7.50], there exist stationary policies $\mu_1, \mu_2, \ldots, \mu_k$ satisfying

$$T_{\mu_i}\big(T^{i-1}(J_\mu)\big) \leq T^i(J_\mu) + \epsilon/(2k), \qquad i = 1, 2, \ldots, k.$$

Consider the Markov policy $\pi = (\mu_k, \mu_{k-1}, \ldots, \mu_1, \mu, \mu, \ldots)$. By a direct calculation we have

$$J_\pi(\bar{x}) = \big(T_{\mu_k} \circ T_{\mu_{k-1}} \circ \cdots T_{\mu_1}\big)(J_\mu) \leq T^k(J_\mu)(\bar{x}) + \epsilon/2 \leq J^*(\bar{x}) + \epsilon.$$

Now consider an associated subproblem where at every state x, there are only $k + 1$ "controls" corresponding to $\{\mu, \mu_1, \mu_2, \ldots, \mu_k\}$. Since the number of controls is finite, one can verify by a direct calculation that value iteration starting with the constant function zero does not have measurability issues and maintains the cost function iterates within the family of nonnegative universally measurable functions. Using a theorem for (P) [7, Props. 5.10, 5.4], we then obtain that for this subproblem, the optimal cost function is universally measurable and moreover, there exists an optimal nonrandomized universally measurable stationary policy $\tilde{\mu}$. Clearly, $\tilde{\mu}$ also corresponds to a universally measurable stationary policy in the original problem. By the optimality of $\tilde{\mu}$ in the subproblem, we have $J_{\tilde{\mu}}(\bar{x}) \leq J_\pi(\bar{x}) \leq J^*(\bar{x}) + \epsilon$, which contradicts the assumption that there exists no ϵ-optimal stationary policy for state $\bar{x}$. This proves the main part of the proposition; the rest then follows by Theorem 5.1(b). □

We illustrate Prop. 5.1 by an example based on [47, Example 6.1], in which the cost function of every stationary policy, although not nearly optimal, is a fixed point of T. Let $S = \{0, 1, 2\}$, $C = (0, 1)$ and $U(x) = C$ for all x. State 0 is cost-free and absorbing. From state 2, any control leads to state 1 with cost 1. For state 1, under control u, we have probability u to transition to state 0 with transition cost 1, and probability $(1 - u)$ to transition to state 1 with self-transition cost 0. The optimal costs are $J^*(0) = J^*(1) = 0, J^*(2) = 1$. An ϵ-optimal Markov policy, for example, is to apply at state 1 control u_k for the kth stage, with $\sum_k u_k \leq \epsilon$. No stationary policy is ϵ-optimal for states 1 and 2: for any stationary policy μ, transition from state 1 to state 0 occurs with probability one, so $J_\mu(0) = 0$, $J_\mu(1) = 1, J_\mu(2) = 2$, and moreover, J_μ is also a fixed point of T.

5.2 Convergence Properties of Mixed Value and Policy Iteration

We now consider the mixed value and policy iteration method in case (P). Unlike case (N) where it is natural to apply the method with the initial iterate $J_0 \equiv 0, Q_0 \equiv 0$, here, as can be shown by a direct calculation, doing so reduces the method to value iteration $T^k(\mathbf{0})$, and this is undesirable computationally even if $T^k(\mathbf{0}) \to J^*$, which is not guaranteed to hold. Our interest thus lies primarily

in applying the method with an initial (J_0, Q_0) above the optimal costs. We will apply Theorem 5.1 to analyze the convergence of the basic algorithm (3.9)-(3.10) given in Section 3.2. We will also discuss another variant algorithm that connects to linear programming, and prove its convergence.

Theorem 5.2. (P) *(a) Let $J_0 \in A_+(S)$ and $Q_0 \in A_+(\Gamma)$ be such that*

$$J^* \leq J_0 \leq cJ^* \quad \text{for some } c > 1, \quad \text{and} \quad Q_0 \geq Q^*. \tag{5.8}$$

Then the sequence $\{(J_k, Q_k)\}$ generated by the iteration (3.9)-(3.10) converges to (J^*, Q^*).

(b) If $T^k(\mathbf{0}) \uparrow J^*$, then the initial condition (5.8) in (a) on (J_0, Q_0) can be relaxed to $J_0 \leq cJ^*$.

(c) Suppose $T^k\big(\underline{J}\big) \uparrow J^*$ for some $\underline{J} \in A_+(S)$. Then the conclusion of (a) holds for the iteration (3.9)-(3.10) that always defines Q_k using the first rule in (3.9), under the initial condition that

$$\underline{J} \leq J_0 \leq cJ^* \quad \text{for some } c > 1, \quad \text{and} \quad Q_0(x,u) \geq \underline{J}(x) \quad \forall\, (x,u) \in \Gamma.$$

In either part of the theorem, it is assumed that $J_0 \leq cJ^*$ for some $c > 1$. We prove first that under this condition on J_0, the limits of the iterates (J_k, Q_k) can be upper bounded by (J^*, Q^*).

Lemma 5.3. (P) *Let $J_0 \in A_+(S)$ and $Q_0 \in A_+(\Gamma)$. If $J_0 \leq cJ^*$ for some $c > 1$, then the sequence $\{(J_k, Q_k)\}$ generated by the iteration (3.9)-(3.10) satisfies*

$$\limsup_{k\to\infty} J_k \leq J^*, \qquad \limsup_{k\to\infty} Q_k \leq Q^*.$$

Proof. Let $J^k = T^k(cJ^*)$. Since $J_0 \leq cJ^*$, we have $J_k \leq T^k(J_0) \leq J^k$ for every k, by Lemma 4.2 and the monotonicity of T. Since $J^k \downarrow J^*$ by Theorem 5.1(a), $\limsup_{k\to\infty} J_k \leq J^*$.

Consider now $Q_k(x,u)$ for each $(x,u) \in \Gamma$, and note that $Q^*(x,u) = H(x,u,J^*)$ by definition [cf. Eqs. (4.1), (4.2)]. By Lemma 4.2, for every $k \geq 0$,

$$Q_{k+1}(x,u) \leq H(x,u,J_k) \leq H(x,u,J^k).$$

If $Q^*(x,u) < \infty$, then we have

$$\lim_{k\to\infty} H(x,u,J^k) = H\big(x,u,\lim_{k\to\infty} J^k\big) = H(x,u,J^*) = Q^*(x,u),$$

where the first equality follows from the monotone convergence theorem as we showed with Eq. (5.2) in the proof of Lemma 5.2. By combining the preceding two relations, we obtain

$$\limsup_{k\to\infty} Q_{k+1}(x,u) \leq Q^*(x,u).$$

This inequality also holds, trivially, if $Q^*(x,u) = \infty$. Therefore, $\limsup_{k\to\infty} Q_k \leq Q^*$. □

We now proceed to prove the theorem by bounding the iterates from below.[5]

Proof of Theorem 5.2. (a) Since $J_0 \geq J^*$ and $Q_0 \geq Q^*$, we have $J_k \geq J^*, Q_k \geq Q^*$ by Lemma 4.2, and hence $J_k \to J^*$, $Q_k \to Q^*$ by Lemma 5.3.

(b) Starting with $J_0 \geq 0, Q_0 \geq 0$, let us prove by induction that for every $k \geq 0$,

$$J_k \geq T^k(\mathbf{0}), \qquad Q_k(x,u) \geq T^k(\mathbf{0})(x), \quad \forall\, (x,u) \in \Gamma. \tag{5.9}$$

[5] For part (a), we will use the lower bounds given in Lemma 4.2, which rely on the relation $Q_{\theta,J^*} = Q^*$ for all $\theta \in \Theta$ (cf. Prop. 3.3(c)). This relation will be proved as Prop. B.1 in Appendix B, and it is needed in the analysis for the algorithm that can set Q_{k+1} to be Q_{θ_k,J_k} at some iterations.

By Lemma 5.3 and the assumption $T^k(\mathbf{0}) \uparrow J^*$, the first inequality above will immediately imply that $J_k \to J^*$.

To simplify notation, let $\hat{J}_k = T^k(\mathbf{0})$ and define $\hat{J}_k^e \in A_+(\Gamma)$ by

$$\hat{J}_k^e(x,u) = \hat{J}_k(x), \quad \forall\, (x,u) \in \Gamma.$$

We will use the following facts. For any $\theta \in \Theta$, in view of $g \geq 0$ and the fact $\hat{J}_k \geq \hat{J}_{k-1} \geq \cdots \geq 0$, a direct calculation using the definition (3.4) of F_θ and its monotonicity shows that

$$F_\theta\big(\mathbf{0}\,;\hat{J}_k\big) \geq \hat{J}_1^e, \quad F_\theta\big(\hat{J}_1^e\,;\hat{J}_k\big) \geq \hat{J}_2^e, \quad \cdots \quad F_\theta\big(\hat{J}_{k-1}^e\,;\hat{J}_k\big) \geq \hat{J}_k^e, \tag{5.10}$$

and that for every $n \geq 1$,

$$F_\theta^n\big(\hat{J}_k^e\,;\hat{J}_k\big) \geq F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big). \tag{5.11}$$

In view of $g \geq 0$ and the definition of $H(x,u,J)$ (cf. Eq. (4.1)), a direct calculation shows that

$$F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big)(x,u) = H\big(x,u,\hat{J}_k\big) \geq T\big(\hat{J}_k\big)(x) = \hat{J}_{k+1}(x), \qquad \forall\, (x,u) \in \Gamma. \tag{5.12}$$

Now suppose Eq. (5.9) holds for some $k \geq 0$. Consider the kth iteration of the algorithm. We have either $Q_{k+1} = F_\theta^n(Q_k\,;J_k)$ or $Q_{k+1} = Q_{\theta,J_k}$ for some $\theta \in \Theta$ and $n \geq 1$. For the case $Q_{k+1} = F_\theta^n(Q_k\,;J_k)$, we have

$$F_\theta^n(Q_k\,;J_k) \geq F_\theta^n\big(\hat{J}_k^e\,;\hat{J}_k\big) \geq F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big),$$

where the first inequality follows from the monotonicity of F_θ (cf. Eq. (3.7)) and the induction hypothesis that $J_k \geq \hat{J}_k, Q_k \geq \hat{J}_k^e$, and the second inequality follows from Eq. (5.11). For the case $Q_{k+1} = Q_{\theta,J_k}$, we have

$$Q_{\theta,J_k} \geq F_\theta^{k+1}\big(\mathbf{0}\,;J_k\big) \geq F_\theta^{k+1}\big(\mathbf{0}\,;\hat{J}_k\big) \geq F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big),$$

where the first inequality holds because $F_\theta^n(\mathbf{0}\,;J_k) \uparrow Q_{\theta,J_k}$ as $n \to \infty$ (Prop. 3.2), the second inequality follows from the induction hypothesis $J_k \geq \hat{J}_k$ and the monotonicity of F_θ (cf. Eq. (3.7)), and the third inequality follows from Eq. (5.10) and the monotonicity of $F_\theta(\cdot\,;\hat{J}_k)$. Thus in either case, we have

$$Q_{k+1} \geq F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big) \geq \hat{J}_{k+1}^e, \qquad J_{k+1} = M(Q_{k+1}) \geq \hat{J}_{k+1},$$

where Eq. (5.12) is used in the second inequality of the first relation above. This completes the induction and establishes Eq. (5.9) for all k.

We can now conclude that $J_k \to J^*$, as discussed earlier. We prove $Q_k \to Q^*$ next. As we just proved, $Q_{k+1} \geq F_\theta\big(\hat{J}_k^e\,;\hat{J}_k\big)$ for every k. By Eq. (5.12), this is equivalent to

$$Q_{k+1}(x,u) \geq H\big(x,u,\hat{J}_k\big), \qquad \forall\, (x,u) \in \Gamma. \tag{5.13}$$

Since $\hat{J}_k \uparrow J^*$ and $\hat{J}_k \geq 0$, by the monotone convergence theorem,

$$H\big(x,u,\hat{J}_k\big) \uparrow H\big(x,u,J^*\big) = Q^*(x,u)$$

(cf. Eqs. (4.1), (4.2)). Together with Lemma 5.3, the preceding two relations imply that $Q_k \to Q^*$.

(c) By assumption $T^k(\underline{J}) \uparrow J^*$. For the algorithm stated in (c), if we define $\hat{J}_0 = \underline{J}$, $\hat{J}_k = T^k(\underline{J})$ for $k \geq 1$, then the same arguments in the preceding proof for part (b) go through to establish that Eqs. (5.11)-(5.12) hold, that for every k,

$$J_k \geq T^k(\underline{J}), \qquad Q_k(x,u) \geq T^k(\underline{J})(x), \quad \forall\, (x,u) \in \Gamma,$$

and that $J_k \to J^*, Q_k \to Q^*$. (Among the crucial facts used in the proof of part (b), the only one that does not hold under the present initial condition on J_0 is the first inequality $F_\theta\big(\mathbf{0}\,;\hat{J}_k\big) \geq \hat{J}_1^e$ in Eq. (5.10). This relation is needed in the convergence proof only when Q_{k+1} is generated by the second rule of (3.9) as $Q_{k+1} = Q_{\theta_k,J_k}$; but such cases are ruled out by the assumptions of part (c).) □

A Variation of the Basic Algorithm (3.9)-(3.10)

Let us consider a variation of the algorithm (3.9)-(3.10), whereby instead of (3.9), we use a different rule to update Q_{k+1}:

- Choose $\theta_k = (\mu_k, B_k) \in \Theta$, and find $Q_{k+1} \in A_+(\Gamma)$ such that

$$Q_{k+1} \le F_{\theta_k}(Q_{k+1}; J_k), \qquad Q_{k+1} \ge Q_{\theta_k, J_k}. \tag{5.14}$$

  Then let

$$J_{k+1} = M(Q_{k+1}). \tag{5.15}$$

This algorithm is motivated by a computational issue in case (P). Unlike (D)(N), control problems of type (P), even when the spaces S, C are discrete, do not admit a linear programming formulation in general (cf. [7, Prop. 9.10(P)], [38, Sec. 7.3.6]). Thus to calculate Q_{θ_k, J_k} in the algorithm (3.9)-(3.10) without iterating $F^n_{\theta_k}(\mathbf{0}\,; J_k)$ till convergence, we cannot solve the optimal stopping problem associated with (θ_k, J_k) by simply solving some linear program.

On the other hand, an upper bound on Q_{θ_k, J_k} will suffice if it also satisfies the first relation in (5.14), as we show in the theorem below. Unlike computing Q_{θ_k, J_k}, a solution to (5.14) can be computed by solving a linear program associated with the optimal stopping problem defined by (θ_k, J_k), under certain conditions that involve (θ_k, J_k), as we will show in Lemma A.3, Appendix A.3. These conditions are satisfied, for example, if S and C are countable and J_k is finite on B_k; see Remark A.1 in Appendix A.3. Given that if $J_0 \le cJ^*$ for some $c > 1$, the algorithm (5.14)-(5.15) will generate J_k with $J_k \le cJ^*$ throughout (see the theorem below), this means that the step (5.14) can be carried out by linear programming for countable-spaces problems where J^* is finite everywhere, in particular.

Theorem 5.3. (P) *Under the same conditions as in Theorem 5.2(a) or (b), the sequence* $\{(J_k, Q_k)\}$ *generated by the iteration (5.14)-(5.15) satisfies* $J_k \le cJ^*$ *for all* k, *and converges to* (J^*, Q^*).

Proof. The proof is similar to that for Theorem 5.2(a)-(b). We will bound (J_k, Q_k) from above and from below. As we derived in Eqs. (4.3)-(4.4), for any $\theta \in \Theta$, $J \in A_+(S)$ and $Q \in A_+(\Gamma)$,

$$F_\theta(Q\,; J)(x,u) \le H(x,u,J), \quad \forall\, (x,u) \in \Gamma, \qquad M\big(F_\theta(Q\,; J)\big) \le T(J).$$

From this and the upper bound on Q_{k+1} given in Eq. (5.14), we have

$$Q_{k+1}(x,u) \le H(x,u,J_k), \quad \forall\, (x,u) \in \Gamma, \qquad J_{k+1} = M(Q_{k+1}) \le T(J_k).$$

By the monotonicity of T, this implies that for every k, $J_k \le T^k(J_0)$ and hence $J_k \le T^k(cJ^*) \le cJ^*$.

The preceding upper bounds on J_k, Q_k are the same as the ones given in Lemma 4.2 for the basic algorithm. Using these bounds in place of Lemma 4.2 in the proof of Lemma 5.3, and using also the assumption that $J_0 \le cJ^*$ for some $c > 1$, we obtain that the conclusion of Lemma 5.3 holds for the iteration (5.14)-(5.15):

$$\limsup_{k\to\infty} J_k \le J^*, \qquad \limsup_{k\to\infty} Q_k \le Q^*. \tag{5.16}$$

Under the conditions of Theorem 5.2(a), we have $J_0 \ge J^*, Q_0 \ge Q^*$. Lemma 4.2 showed that if $Q_{k+1} = Q_{\theta_k, J_k}$ at every iteration of the algorithm, then $J_k \ge J^*, Q_k \ge Q^*$ for all k. Since here we have $Q_{k+1} \ge Q_{\theta_k, J_k}$ by Eq. (5.14), and the iteration (5.14)-(5.15) clearly has the monotonicity property, it follows that for the iteration (5.14)-(5.15), we have $J_k \ge J^*, Q_k \ge Q^*$ for all k as well. This together with Eq. (5.16) implies that $J_k \to J^*, Q_k \to Q^*$.

Similarly, under the conditions of Theorem 5.2(b), the proof of Theorem 5.2(b) established the lower bounds (5.9), (5.13) on J_k, Q_k for the case where $Q_{k+1} = Q_{\theta_k, J_k}$ at every iteration, and these lower bounds also hold for the iteration (5.14)-(5.15) since $Q_{k+1} \ge Q_{\theta_k, J_k}$. Together with Eq. (5.16), they imply that $J_k \to J^*, Q_k \to Q^*$, as the proof of Theorem 5.2(b) showed. □

Remark 5.4. We note that under (P), given the sequence $\{J_k\}$ generated by the algorithm (3.9)-(3.10) or (5.14)-(5.15), in general one cannot extract easily an asymptotically near-optimal sequence of policies in the manner of Remark 4.1. Even if J^* was available, an ϵ-optimal stationary policy may not exist (see the discussion after Prop. 5.1 or [6, p. 145] for an example). If an ϵ-optimal stationary policy exists, then under favorable circumstances it may be possible to extract such a sequence, based on the following observation. Let $\{J_k\} \subset A_+(S)$ be such that $J_k \to J^*$, and let $\{\nu_k\}$ be a sequence of universally measurable policies. Suppose $\{J_k\}$ and $\{\nu_k\}$ satisfy

$$T_{\nu_k}(J_k) = T(J_k) \leq J_k, \qquad \forall\, k \geq 1. \tag{5.17}$$

Then $J_{\nu_k} \to J^*$. (To see this, note that by [7, Prop. 9.11], J_{ν_k} is the "smallest" nonnegative function $J \in \mathcal{M}(S)$ satisfying $T_{\nu_k}(J) \leq J$, so the assumption implies that $J_{\nu_k} \leq J_k$. Since $J_k \to J^*$, the result follows.) The assumption (5.17), however, need not always hold for our algorithm. □

6 Applications in Semicontinuous Models

We discuss in this section some direct applications of our results for two special cases of the stochastic control model given in Section 2.2: the upper semicontinuous model and the lower semicontinuous model as defined in [7, Chap. 8]. To apply the mixed value and policy iteration method in these models, it is desirable to work with semicontinuous functions instead of lower semi-analytic functions. We will show that we can keep the function iterates within the set of semicontinuous functions by choosing properly the parameters of the mappings F_θ involved in the method, and we will use Lusin's theorem for this purpose. In this section we will also give a result about the structure of J^* and optimal policies for the upper semicontinuous model in case (P), as an application of Theorem 5.1.

We need some definitions. Let X be a metrizable topological space. A function $f : X \to [-\infty, \infty]$ is said to be *upper semicontinuous* (u.s.c.) if for every $c \in \Re$, its upper level set $\{x \in X \mid f(x) \geq c\}$ is closed in X. Equivalently, f is u.s.c. if and only if for any sequence $\{x_n\}$ in X converging to some $x \in X$, we have $\limsup_{n\to\infty} f(x_n) \leq f(x)$. A function $f : X \to [-\infty, \infty]$ is said to be *lower semicontinuous* (l.s.c.) if for every $c \in \Re$, its lower level set $\{x \in X \mid f(x) \leq c\}$ is closed in X. Equivalently, f is l.s.c. if and only if for any sequence $\{x_n\}$ in X converging to some $x \in X$, we have $\liminf_{n\to\infty} f(x_n) \geq f(x)$.

Let X and Y be separable metrizable topological spaces. Let the topology on the space $\mathscr{P}(Y)$ of Borel probability measures on Y be the weak topology. A stochastic kernel $\kappa(dy \mid x)$ on Y given X is *continuous* if the function $\kappa(dy \mid \cdot) : X \to \mathscr{P}(Y)$ is continuous. Similarly, if restricted to a subset $B \subset X$, the function $\kappa(dy \mid \cdot) : B \to \mathscr{P}(Y)$ is continuous, we say $\kappa(dy \mid x)$ is *continuous on* B.

6.1 Upper Semicontinuous Models

We consider the upper semicontinuous model as defined in [7, Def. 8.8]. Here, in addition to the model assumptions given in Section 2.2, we assume that:

(a) The control constraint set Γ is an open subset of $S \times C$.

(b) The state transition stochastic kernel $q(dx' \mid x, u)$ is continuous.[6]

(c) The one-stage cost function g is u.s.c. on Γ and bounded above.

It is known that under (D)(N), the optimal cost function J^* is u.s.c. Starting with $J \equiv 0$ for (N) and with any bounded u.s.c. function J for (D), value iteration generates u.s.c. functions $T^k(J)$

[6] Such state transition kernels are said to be weakly continuous in the literature to differentiate them from those that satisfy stronger continuity conditions; see e.g., [28, 22].

converging to J^*. There exists an ϵ-optimal, nonrandomized Borel measurable policy which is stationary under (D) and Markov under (P). (For these optimality results, see [7, Props. 8.7, 9.21].)

Consider the mixed value and policy iteration algorithm (3.9)-(3.10). By a selection theorem for u.s.c. functions [7, Prop. 7.34], if $Q : \Gamma \to [-\infty, \infty]$ is u.s.c., then the function resulting from partial minimization,

$$M(Q)(x) = \inf_{u \in U(x)} Q(x, u), \qquad x \in S,$$

is u.s.c., and for any $\epsilon > 0$, there exists a Borel measurable nonrandomized stationary policy μ such that for all $x \in S$,

$$Q\big(x, \mu(x)\big) \leq \begin{cases} M(Q)(x) + \epsilon & \text{if } M(Q)(x) > -\infty, \\ -1/\epsilon & \text{if } M(Q)(x) = -\infty. \end{cases} \tag{6.1}$$

Suppose we can maintain the iterates (J_k, Q_k) of the algorithm (3.9)-(3.10) within the family of u.s.c. functions. Then at each iteration k, we can choose the policy μ_k based on Q_k and the above selection theorem, thereby obtaining policy iteration-like algorithms[7] with Borel measurable policies μ_k. One way to keep the iterates (J_k, Q_k) within the family of u.s.c. functions is to choose, at each iteration, for a given stationary Borel measurable policy μ, an appropriate set $B \subset S$ to form the parameter $\theta = (\mu, B)$ in the mapping F_θ as follows.

Let μ be a Borel measurable stationary policy. Consider an open set $B \subset S$ such that restricted to B, the function $x \mapsto \mu(du \mid x)$ is continuous. We know from Lusin's theorem [19, Thm. 7.5.2] that there exists a closed subset $\bar{B}$ of S such that restricted to $\bar{B}$, the function $x \mapsto \mu(du \mid x)$ is continuous, and moreover, for any given $p \in \mathscr{P}(S)$, the set $\bar{B}$ can be chosen to have $p(\bar{B})$ arbitrarily close to 1. Then we can let $B = \text{int}(\bar{B})$, the interior of $\bar{B}$, for instance.[8]

Proposition 6.1 (Upper Semicontinuous Models)**.** *Let $\theta = (\mu, B)$ for an open subset B of S and a Borel measurable stationary policy μ such that $\mu(du \mid \cdot)$ is continuous on B. Then for any functions J, Q that are u.s.c. and bounded above, $F_\theta(Q\,; J)$ is u.s.c. and bounded above.*

Proof. Since g is u.s.c. and bounded above by our model assumption, to show that $F_\theta(Q\,; J)$ is u.s.c. and bounded above, it suffices to show that the sum of the two integral terms in the definition (3.4) of $F_\theta(Q\,; J)$ is u.s.c. and bounded above. To this end, let us rewrite this sum as

$$\alpha \int_S \big(\phi(x') \cdot \mathbb{1}_B(x') + J(x') \cdot \mathbb{1}_{S \setminus B}(x')\big)\, q(dx' \mid x, u), \tag{6.2}$$

where the function $\phi(x')$ is given by

$$\phi(x') = \int_C \min\{J(x'), Q(x', u')\}\, \mu(du' \mid x'), \quad x' \in S. \tag{6.3}$$

We prove first that $\phi(x')$ is u.s.c. on B. Since J, Q are u.s.c. and bounded above, the function $\min\{J(x), Q(x, u)\}$ is u.s.c. and bounded above on Γ. Note that since Γ is an open subset of $S \times C$, we may extend the function $\min\{J(x), Q(x, u)\}$ to an u.s.c. function on $S \times C$ that is bounded above, and view the integral defining $\phi(x')$ as the integral of this extension. This will not change the value

[7] Without stronger model assumptions, standard policy iteration has the same difficulties in the upper and lower semicontinuous models considered here as those explained in Section 2.4. For a Borel measurable policy, its cost function is Borel measurable and not necessarily u.s.c. or l.s.c., so the policy improvement step will generate an analytically or universally measurable policy. The subsequent iterations will then be subject to the measurability difficulties described in Section 2.4.

[8] We note that $\text{int}(\bar{B})$ may be empty. However, if the state space is continuous, e.g., $S = \Re^n$, then $\bar{B}$ can clearly be chosen to have a nonempty interior, by letting the probability measure be absolutely continuous with respect to Lebesgue measure, for example.

$\phi(x')$, since μ satisfies the control constraint. Then since the function $x \mapsto \mu(du \mid x)$ is continuous on B, we can apply [7, Prop. 7.31(b)] to conclude that $\phi(x')$ is u.s.c. on B.

Denote $\psi(x') = \phi(x') \cdot \mathbb{1}_B(x') + J(x') \cdot \mathbb{1}_{S \setminus B}(x')$ for $x' \in S$. We prove that $\psi(x')$ is u.s.c. on S. Consider any sequence $\{x_n\}$ in S converging to some $\bar{x} \in S$. By the definition of ϕ, we have $\phi(x') \leq J(x')$ for all $x' \in S$. Therefore, in the case $\bar{x} \notin B$, we have

$$\limsup_{n\to\infty} \psi(x_n) \leq \limsup_{n\to\infty} J(x_n) \leq J(\bar{x}) = \psi(\bar{x}),$$

where the second inequality follows from the u.s.c. property of J; whereas in the case $\bar{x} \in B$, since B is open, we have

$$\limsup_{n\to\infty} \psi(x_n) = \limsup_{n\to\infty} \phi(x_n) \leq \phi(\bar{x}) = \psi(\bar{x}),$$

where the inequality holds since ϕ restricted to B is u.s.c., as we proved earlier. This proves that the function ψ is u.s.c. Clearly ψ is bounded above. Then, using also the fact that the state transition kernel $q(dx' \mid x, u)$ is continuous, we have, by [7, Prop. 7.31(b)], that the integral (6.2) as a function of (x, u) is u.s.c. and clearly bounded above. This proves the proposition. □

Based on Prop. 6.1, we see that to keep the iterates J_k, Q_k of the iteration (3.9)-(3.10) within the set of functions that are u.s.c. and bounded above, we can start with J_0, Q_0 that are u.s.c. and bounded above, choose the parameters $\theta_k = (\mu_k, B_k)$ in the way described earlier, and update Q_{k+1} using always the first rule in (3.9), thereby resulting in u.s.c. functions Q_{k+1} and J_{k+1}. For cases (D)(N), it is not hard to show that the second rule in (3.9), $Q_{k+1} = Q_{\theta_k, J_k}$, also makes Q_{k+1} u.s.c. and therefore can be used. For case (P), however, we do not know if Q_{θ_k, J_k} is u.s.c. in general.

We conclude this subsection with an optimality result for the upper semicontinuous model in case (P). To our knowledge, here there is no guarantee that J^* is u.s.c.; however, an application of Theorem 5.1 shows the following.

Proposition 6.2 (Case (P) in Upper Semicontinuous Models)**.** *Suppose that J^* is bounded above and for some open set $B \subset S$ and $\delta > 0$, $B \supset \{x \in S \mid J^*(x) < \delta\}$ and J^* is u.s.c. on B. Then J^* is u.s.c. and for any $\epsilon > 0$, there exists an ϵ-optimal, Borel measurable Markov policy.*

Proof. Suppose $J^*(x) \leq a$ for all x. Let $J(x) = J^*(x)$ if $x \in B$ and $J(x) = a$ otherwise. Since J^* is u.s.c. on the open set B, J is by definition u.s.c. and bounded above. Consequently, for all k, $T^k(J)$ is u.s.c. and bounded above by [7, Props. 7.31, 7.34]. We also have $J^* \leq J \leq cJ^*$ for $c \geq \max\{1, a/\delta\}$, so by Theorem 5.1(b), $T^k(J) \to J^*$. Then, using the fact that $T^k(J)$ is u.s.c. and $T^k(J) \geq J^*$ for all k, it follows that J^* is u.s.c.[9] The assertion of the existence of ϵ-optimal, Borel measurable Markov policy then follows from a selection theorem for u.s.c. functions ([7, Prop. 7.34]; cf. Eq. (6.1)) and the same proof argument as that for [7, Prop. 9.19(P)]. □

6.2 Lower Semicontinuous Models

We now consider the lower semicontinuous model as defined in [7, Def. 8.7]. For simplicity, in addition to the model assumptions given in Section 2.2, let us assume that:

(a) The control space C is compact, and the control constraint set Γ is a closed subset of $S \times C$.

(b) The state transition stochastic kernel $q(dx' \mid x, u)$ is continuous.

(c) The one-stage cost function g is l.s.c. on Γ and bounded below.

[9] Here we used the fact that if $\{f_n\}$ is a sequence of u.s.c. functions on a metrizable space X converging pointwise to f with $f_n \geq f$ for all n, then f is u.s.c. To see this, let $\{x_k\}$ be a sequence in X converging to $x \in X$. We have for every n, $\limsup_{k\to\infty} f(x_k) \leq \limsup_{k\to\infty} f_n(x_k) \leq f_n(x)$, and hence $\limsup_{k\to\infty} f(x_k) \leq \lim_{n\to\infty} f_n(x) = f(x)$. This shows that f is u.s.c.

This is a special case of the model defined in [7, Def. 8.7], but our discussion below applies to that more general model. Let us also mention that there have been substantial efforts in the literature to weaken the assumptions (a) and (c) above. For these more general lower semicontinuous models and the most recent results, we refer to the paper by Feinberg, Kasyanov and Zadoianchuk [22]. In principle, the approach we describe here is applicable in these models as well to address the measurability issues in standard policy iteration (cf. Footnote 7), although the subject is beyond the scope of the present paper.

It is known that under the assumptions (a)-(c) above, the optimal cost function J^* is l.s.c. for the models (D)(P). Starting with $J \equiv 0$ for (P) and with any bounded l.s.c. function J for (D), value iteration generates l.s.c. functions $T^k(J)$ converging to J^*. There exists an optimal, Borel measurable nonrandomized stationary policy under (D)(P). (For these optimality results, see [7, Prop. 8.6 and Cor. 9.17.2].)

Consider the mixed value and policy iteration algorithm (3.9)-(3.10). In what follows, we apply arguments similar to those for the upper semicontinuous model, and we show that one can have policy iteration-like algorithms that keep iterates (J_k, Q_k) within the set of l.s.c. functions. More specifically, by a selection theorem for l.s.c. functions [7, Prop. 7.33], we have that if $Q : \Gamma \to [-\infty, \infty]$ is l.s.c., then the function $M(Q)(x) = \inf_{u \in U(x)} Q(x,u)$ is l.s.c. on S and for any $\epsilon > 0$, there exists a Borel measurable nonrandomized stationary policy μ such that

$$Q(x, \mu(x)) = M(Q)(x), \qquad x \in S. \tag{6.4}$$

Thus at the kth iteration of the algorithm (3.9)-(3.10), assuming Q_k is l.s.c., we can choose a Borel measurable policy μ_k based on Q_k and the above selection theorem, to obtain a policy iteration-like algorithm. In order for Q_{k+1}, J_{k+1} to be l.s.c. and bounded below, we can choose an appropriate set B_k when forming the parameter $\theta_k = (\mu_k, B_k)$ for the mapping F_{θ_k} in the algorithm, as follows.

Let μ be a Borel measurable stationary policy. There exists a closed subset $B \subset S$ such that restricted to B, the function $x \mapsto \mu(du \,|\, x)$ is continuous. Again, we know from Lusin's theorem [19, Thm. 7.5.2] that B can be chosen to be very "large," with its measure arbitrarily close to 1 for any given Borel probability measure on S.

Proposition 6.3 (Lower Semicontinuous Models)**.** *Let $\theta = (\mu, B)$ for a closed subset B of S and a Borel measurable stationary policy μ such that $\mu(du \,|\, \cdot)$ is continuous on B. Then for any functions J, Q that are l.s.c. and bounded below, $F_\theta(Q\,; J)$ is l.s.c. and bounded below.*

Proof. Similar to the proof of Prop. 6.1, it suffices to show that the integral (6.2) as a function of (x,u) is l.s.c. and bounded below on Γ. We prove first that the function $\phi(x')$ given by Eq. (6.3) is l.s.c. on B. Since J, Q are l.s.c. and bounded below, the function $\min\{J(x), Q(x,u)\}$ is l.s.c. and bounded below on Γ. We may extend the function $\min\{J(x), Q(x,u)\}$ to an l.s.c. function on $S \times C$ that is bounded below, by defining its values outside Γ to be ∞, and we can view the integral defining $\phi(x')$ as the integral of this extension. This will not change the value $\phi(x')$, since μ satisfies the control constraint. Then, since the function $x \mapsto \mu(du \,|\, x)$ is continuous on B, we can apply [7, Prop. 7.31(a)] to conclude that $\phi(x')$ is l.s.c. and bounded below on B.

Denote $\psi(x') = \phi(x') \cdot \mathbb{1}_B(x') + J(x') \cdot \mathbb{1}_{S \setminus B}(x')$ for $x' \in S$. We prove that $\psi(x')$ is l.s.c. on S. Consider any sequence $\{x_n\}$ in S converging to some $\bar{x} \in S$. If $\bar{x} \notin B$, then since $S \setminus B$ is open, we have

$$\liminf_{n\to\infty} \psi(x_n) = \liminf_{n\to\infty} J(x_n) \geq J(\bar{x}) = \psi(\bar{x}),$$

where the inequality follows from the l.s.c. property of J. Suppose now $\bar{x} \in B$. There exists a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ such that $\liminf_{n\to\infty} \psi(x_n) = \lim_{i\to\infty} \psi(x_{n_i})$ and either (i) $x_{n_i} \in B$ for all i or (ii) $x_{n_i} \notin B$ for all i. Then in case (i), we have

$$\liminf_{n\to\infty} \psi(x_n) = \lim_{i\to\infty} \psi(x_{n_i}) = \lim_{i\to\infty} \phi(x_{n_i}) \geq \phi(\bar{x}) = \psi(\bar{x}),$$

where the inequality holds since ϕ restricted to B is l.s.c., as we proved earlier. In case (ii), we have

$$\liminf_{n\to\infty} \psi(x_n) = \lim_{i\to\infty} \psi(x_{n_i}) = \lim_{i\to\infty} J(x_{n_i}) \geq J(\bar{x}) \geq \phi(\bar{x}) = \psi(\bar{x}),$$

where the first inequality holds since J is l.s.c., and the second inequality holds since by the definition of ϕ, we have $\phi(x') \leq J(x')$ for all $x' \in S$. Thus we have proved that the function ψ is l.s.c. Clearly ψ is bounded below. Then, using also the fact that the state transition kernel $q(dx' \mid x, u)$ is continuous, we have, by [7, Prop. 7.31(a)], that the integral (6.2) as a function of (x, u) is l.s.c. and bounded below. This proves the proposition. □

Based on Prop. 6.3, we see that to keep iterates J_k, Q_k of the iteration (3.9)-(3.10) within the set of functions that are l.s.c. and bounded below, we can start with J_0, Q_0 that are l.s.c. and bounded below, choose the parameters $\theta_k = (\mu_k, B_k)$ in the way described earlier, and update Q_{k+1} using always the first rule in (3.9), thereby resulting in l.s.c. functions Q_{k+1} and J_{k+1}. For cases (D)(P), it is not hard to show that the second rule in (3.9), $Q_{k+1} = Q_{\theta_k, J_k}$, also makes Q_{k+1} l.s.c. and therefore can be used. For case (N), however, we do not know if Q_{θ_k, J_k} is l.s.c. in general.

7 Concluding Remarks

In this paper we have addressed the long-standing open issue of constructing a valid policy iteration algorithm for total cost Borel-space stochastic DP with universally measurable policies. Our approach is based on a mixed value and policy iteration idea. It makes critical use of the fact that any universally measurable policy has Borel measurable portions, to maintain cost function iterates within the set of lower semi-analytic functions. It employs an algorithmic framework that combines the characteristics of both value and policy iteration, to allow stationary policies to be used in computing the optimal cost function. Our approach can also address similar policy iteration issues that arise in upper and lower semicontinuous models. By choosing algorithmic parameters accordingly, we have shown how to obtain policy iteration-like algorithms that can keep the cost function iterates within the desired family of semicontinuous functions.

The mixed value and policy iteration method was first proposed and studied in our earlier work for discrete spaces [10, 57] and abstract DP models [9], with the focus on asynchronous distributed computation. With this paper we have thus provided a Borel-space counterpart of the method, and broadened the algorithmic framework of our earlier work to deal with measurability or non-measurability related structural restrictions in stochastic DP problems.

For nonnegative DP models, however, the standard versions of value iteration and policy iteration may fail, even for discrete-state and other models where measurability issues are not a concern. In order to apply and analyze our mixed value and policy iteration method, we have derived a new sufficient condition for convergence of value iteration. This is a simple condition on the initial function only. It applies to all nonnegative models (countable space or uncountable Borel space models), and it provides, in addition, a new characterization of the set of functions within which the optimal cost function is the unique solution of the optimality equation. Using this condition, our method is shown to produce in the limit the optimal cost function when initialized properly. Obtaining useful initial functions satisfying this condition is generally an open question at present, which we aim to address in the future.

For nonnegative DP models, we have also proposed a variation of our method, where the optimal stopping problems in its "policy evaluation" phase can be approximately solved by using linear programming under certain conditions. To our knowledge, this is the first proposal of an algorithmic approach based on linear programming for nonnegative DP models.

Our approach yields function sequences that converge pointwise to the optimal cost function for discounted, nonpositive, and nonnegative cost DP models. It also yields asymptotically optimal

policy sequences for discounted cost, but not for nonpositive and nonnegative cost DP models. For the latter two models, extracting nearly optimal policies from the data produced by the algorithm is difficult in the absence of additional assumptions, since in general there may not exist ϵ-optimal stationary policies.

Further analyses of our algorithms and their variations, including stochastic asynchronous Q-learning versions (similar to those considered in [54, 49, 50, 18, 1, 10, 57, 58]), are important subjects for future investigation. We conclude the paper with a discussion about other applications of our approach and future research directions.

Asynchronous computation

One may consider asynchronous distributed computation in the framework of universally measurable policies, by combining the approach and analysis given in this paper with arguments used in our earlier works [10, 57, 9]. We discuss the subject briefly here, focusing on issues related to universal measurability in a simplified setting.

Suppose that instead of the basic algorithm (3.9)-(3.10), at each iteration k, we only compute $Q_{k+1}(x,u)$ for a subset Γ_k of state-control pairs in Γ and compute $J_{k+1}(x)$ for a subset S_k of states in S. (For the rest of states x or state-control pairs (x,u), we let $J_{k+1}(x) = J_k(x)$, $Q_{k+1}(x,u) = Q_k(x,u)$.) This is the type of operations that would be performed in a distributed computation environment, where a single processor handles only part of a computation task and processors share results with each other.

As before, with universally measurable policies, we need to keep the iterates within the set of lower semi-analytic functions. To meet this requirement, we can let S_k be a Borel subset of S and let $\Gamma_k = R_k \cap \Gamma$, where R_k is a *Borel subset of* $S \times C$. This will keep $Q_{k+1} \in A(\Gamma)$. The reason is that if $Q, Q' \in A(\Gamma)$ and R is a Borel set in $S \times C$, then the function

$$Q \cdot \mathbb{1}_{R \cap \Gamma} + Q' \cdot \mathbb{1}_{(S \times C \setminus R) \cap \Gamma}$$

is lower semi-analytic by [7, Lemma 7.30(4)], because $\mathbb{1}_{R \cap \Gamma}(x,u)$ and $\mathbb{1}_{(S \times C \setminus R) \cap \Gamma}(x,u)$ are nonnegative Borel measurable functions on Γ. Similarly, the reason for $J_{k+1} \in A(S)$ is that if $J, J' \in A(S)$ and $D \subset S$ is Borel, then the function $J \cdot \mathbb{1}_D + J' \cdot \mathbb{1}_{S \setminus D}$ is lower semi-analytic.

More elaborate variants

In this paper we have focused on the mappings $F_\theta, \theta \in \Theta$, defined by (3.4), where we partition the state space into two subsets. The same idea leads to more elaborate mappings, which can also be used in the mixed value and policy iteration approach. We give one such example here, in which we will partition the state-control space $S \times C$.

For a stationary universally measurable policy μ, let $R \subset S \times C$ be a Borel set such that $B = \mathrm{proj}_S(R)$ is Borel and the function $x \mapsto \mu(du \mid x)$ is Borel measurable on B. For any such pair $\hat{\theta} = (\mu, R)$, we may consider a mapping $F_{\hat{\theta}}$ defined by

$$\begin{aligned} F_{\hat{\theta}}(Q\,;J)(x,u) = g(x,u) &+ \alpha \int_{S \setminus B} J(x')\, q(dx' \mid x,u) + \alpha \int_B J(x') \cdot \mu\big(C \setminus R_{x'} \mid x'\big)\, q(dx' \mid x,u) \\ &+ \alpha \int_B \int_{R_{x'}} \min\big\{J(x')\,,\, Q(x',u')\big\}\, \mu(du' \mid x')\, q(dx' \mid x,u), \qquad (x,u) \in \Gamma, \end{aligned} \tag{7.1}$$

for all $J \in A(S), Q \in A(\Gamma)$, where $B = \mathrm{proj}_S(R)$ and $R_x = \{u \in C \mid (x,u) \in R\}$ is the vertical section of R at x. That the function $F_{\hat{\theta}}(Q\,;J)$ is lower semi-analytic can be established similar to Prop. 3.1(a), using the arguments in its proof, together with the fact that restricted to B, $\mu(C \setminus R_{x'} \mid x')$ is a nonnegative Borel measurable function [7, Cor. 7.26.1] and hence the term $\int_B J(x') \cdot \mu\big(C \setminus R_{x'} \mid x'\big)\, q(dx' \mid x,u)$ in (7.1) as a function of (x,u) is lower semi-analytic.

Extensions to other models

Finally, we note that while we have focused on the three classical total cost problems in this paper, the technique we used to handle the measurability issues in policy iteration can be applied to other types of stochastic control problems. These include, for instance, discounted problems with unbounded one-stage costs, and undiscounted total cost problems without sign constraints on the one-stage costs. Convergence properties of the mixed value and policy iteration method for these models are worthy of further study. Also among the important subjects for future research are extensions to average cost problems and partially observable problems.

Acknowledgments

We thank Prof. Steven Shreve for a helpful discussion and a suggestion about how to choose the probability measures for the algorithms in Section 3.2, which we described in Example 3.1. We also thank Prof. Eugene Feinberg, with whom our recent correspondence about Borel models stimulated this research. We appreciate Prof. Sanjoy Mitter's helpful feedback on our early draft. This work was supported by the Air Force Grant FA9550-10-1-0412.

References

[1] J. Abounadi, D. P. Bertsekas, and V. S. Borkar. Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms. *SIAM J. Control Opt.*, 41:1–22, 2002.

[2] E. Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, Roca Raton, 1999.

[3] D. P. Bertsekas. Infinite time reachability of state space regions by using feedback control. *IEEE Trans. Automatic Control*, AC-17:604–613, 1972.

[4] D. P. Bertsekas. Monotone mappings with application in dynamic programming. *SIAM J. Control Opt.*, 15:438–464, 1977.

[5] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, 4th edition, 2012.

[6] D. P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, Belmont, 2013.

[7] D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, New York, 1978.

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

[9] D. P. Bertsekas and H. Yu. Distributed asynchronous policy iteration in dynamic programming. In *Proc. 48th Allerton Conf. on Communication, Control and Computing*, pages 1368–1375, 2010.

[10] D. P. Bertsekas and H. Yu. Q-learning and enhanced policy iteration in discounted dynamic programming. *Math. Oper. Res.*, 37:66–94, 2012.

[11] D. Blackwell. Memoryless strategies in finite stage dynamic programming. *Ann. Math. Statist.*, 35:863–865, 1964.

[12] D. Blackwell. Discounted dynamic programming. *Ann. Math. Statist.*, 36:226–235, 1965.

[13] D. Blackwell. Positive dynamic programming. In *Proc. 5th Berkeley Sympos. Math. Satist. and Probability*, pages 415–418, 1965.

[14] D. Blackwell. A Borel set not containing a graph. *Ann. Math. Statist.*, 39:1345–1347, 1968.

[15] D. Blackwell. Borel-programmable functions. *Ann. Probability*, 6:321–324, 1978.

[16] D. Blackwell, D. Freedman, and M. Orkin. The optimal reward operator in dynamic programming. *Ann. Probability*, 2:926–941, 1974.

[17] D. Blackwell and C. Ryll-Nardzewski. Non-existence of everywhere proper conditional distributions. *Ann. Math. Statist.*, 34:223–225, 1963.

[18] V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Opt.*, 38:447–469, 2000.

[19] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 2002.

[20] E. B. Dynkin and A. A. Yushkevich. *Controlled Markov Processes*. Springer, New York, 1979.

[21] E. A. Feinberg. Total reward criteria. In E. A. Feinberg and A. Shwartz, editors, *Handbook of Markov Decision Processes*. Springer, New York, 2002.

[22] E. A. Feinberg, P. O. Kasyanov, and N. V. Zadoianchuk. Average cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res.*, 37:591–607, 2012.

[23] E. A. Feinberg and A. Shwartz, editors. *Handbook of Markov Decision Processes*. Springer, New York, 2002.

[24] D. Freedman. The optimal reward operator in special classes of dynamic programming problems. *Ann. Probability*, 2:942–949, 1974.

[25] N. Furukawa. Markovian decision processes with compact action spaces. *Ann. Math. Statist.*, 43:1612–1622, 1972.

[26] R. Hartley. A simple proof of Whittle's bridging condition in dynamic programming. *J. Appl. Prob.*, 17:1114–1116, 1980.

[27] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer, New York, 1996.

[28] O. Hernández-Lerma and J. B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer, New York, 1999.

[29] K. Hinderer. *Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter*. Springer, New York, 1970.

[30] D. M. Kreps and E. L. Porteus. On the optimality of structured policies in countable stage decision processes. II: positive and negative problems. *SIAM J. Appl. Math.*, 32:457–466, 1977.

[31] K. Kuratowski. *Topology I*. Academic Press, New York, 1966.

[32] A. Maitra. Discounted dynamic programming on compact metric spaces. *Sankhyā: The Indian Journal of Statistics, Series A*, 30:211–216, 1968.

[33] A. Maitra and W. Sudderth. The optimal reward operator in negative dynamic programming. *Math. Oper. Res.*, 17:921–931, 1992.

[34] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[35] B. L. Miller and A. F. Veinott. Discrete dynamic programming with a small interest rate. *Ann. Math. Statist.*, 40:366–370, 1969.

[36] J. Neveu. *Discrete-Parameter Martingales*. North-Holland, Amsterdam, 1975.

[37] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, New York, 1967.

[38] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, New York, 1994.

[39] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.

[40] M. Schäl. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 32:179–196, 1975.

[41] S. E. Shreve. Probability measures and the C-set of Selivanovskij. *Pacific J. Math.*, 79:189–196, 1978.

[42] S. E. Shreve. Resolution of measurability problems in discrete-time stochastic control. In *Stochastic Control Theory and Stochastic Differential Systems*, pages 580–587. Springer, Berlin, 1979.

[43] S. E. Shreve. Borel-approachable functions. *Fundamenta Mathematicae*, 112:17–24, 1981.

[44] S. E. Shreve and D. P. Bertsekas. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM J. Control Opt.*, 16:953–977, 1978.

[45] S. E. Shreve and D. P. Bertsekas. Universally measurable policies in dynamic programming. *Math. Oper. Res.*, 4:15–30, 1979.

[46] S. M. Srivastava. *A Course on Borel Sets*. Springer, New York, 1998.

[47] R. E. Strauch. Negative dynamic programming. *Ann. Math. Statist.*, 37:871–890, 1966.

[48] R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, 1998.

[49] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Mach. Learn.*, 16:185–202, 1994.

[50] J. N. Tsitsiklis and B. Van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing financial derivatives. *IEEE Trans. Automat. Contr.*, 44:1840–1851, 1999.

[51] J. van der Wal. *Stochastic Dynamic Programming*. The Mathematical Centre, Amserdam, 1981.

[52] A. F. Veinott. On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Statist.*, 37:1284–1294, 1966.

[53] A. F. Veinott. On discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.*, 40:1635–1660, 1969.

[54] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge Univ., England, 1989.

[55] P. Whittle. A simple condition for regularity in negative programming. *J. Appl. Prob.*, 16:305–318, 1979.

[56] P. Whittle. Stability and characterisation conditions in negative programming. *J. Appl. Prob.*, 17:635–645, 1980.

[57] H. Yu and D. P. Bertsekas. Q-learning and policy iteration algorithms for stochastic shortest path problems. *Ann. Oper. Res.*, 2012. Forthcoming; DOI: 10.1007/s10479-012-1128-z.

[58] H. Yu and D. P. Bertsekas. On boundedness of Q-learning iterates for stochastic shortest path problems. *Math. Oper. Res.*, 38:209–227, 2013.

Appendices

A Optimal Stopping Problems Associated with the Mappings F_θ

In this appendix, for a given $\theta = (\mu, B) \in \Theta$, $J \in A(S)$, and a control problem of type (D), (N) or (P), we formulate an associated optimal stopping problem of the same type. We establish the relation between its optimal cost function and the pointwise limit $Q_{\theta,J} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,; J)$, and we show that the mapping $F_\theta(\cdot\,; J)$ can be viewed as a form of the optimal cost operator and $F_\theta^k(\mathbf{0}\,; J)$ is related to the value iteration sequence for this problem. (Other formulations of the optimal stopping problem are also possible and equivalent for our purpose. We will focus only on one here.) In addition we describe a linear program in case (P) and show that under certain conditions, it yields an upper bound on $Q_{\theta,J}$ that can be used in a mixed value and policy iteration algorithm discussed in Section 5.2.

A.1 Formulation

As before we assume that the given function J is such that $J \in A_b(S)$ in case (D), $J \in A_-(S)$ in case (N), and $J \in A_+(S)$ in case (P). The function J will define the stopping costs, while the policy μ will be used to define the dynamics of the unstopped process.

Optimal Stopping Problem Associated with J and $(\mu, B) \in \Theta$

- State space $S^o = \big(S \times C\big) \cup \{\infty\}$, with ∞ representing an absorbing, cost-free state. (The topology of S^o consists of the open sets in $S \times C$, the set $\{\infty\}$ and their unions.)
- Control space $C^o = \{0, 1\}$, with 0 representing "to stop" and 1 "to continue."
- Control constraint: $U^o(\infty) = \{0, 1\}$ and

$$U^o\big((x,u)\big) = \{0,1\} \text{ on } B \times C, \qquad U^o\big((x,u)\big) = \{0\} \text{ on } (S \setminus B) \times C.$$

- One-stage costs: $g^o(\infty, 0) = g^o(\infty, 1) = 0$ and

$$\begin{aligned} &g^o\big((x,u),0\big) = J(x) && \forall\,(x,u) \in S \times C,\\ &g^o\big((x,u),1\big) = g(x,u) && \forall\,(x,u) \in (B \times C) \cap \Gamma,\\ &g^o\big((x,u),1\big) = K && \forall\,(x,u) \in (B \times C) \setminus \Gamma, \end{aligned}$$

  where $K = 0$ for (N), $K = +\infty$ for (P), and $K \geq \max\{\|g\|_\infty, \|J\|_\infty\}$ for (D).

- State transition stochastic kernel $q^o(\cdot \mid \cdot)$ on S^o given $S^o \times C^o$: for any Borel set $D \subset S^o$ and any $(x, u) \in S \times C$,

$$q^o\big(D \mid \infty, 0\big) = q^o\big(D \mid \infty, 1\big) = \delta_\infty(D), \qquad q^o\big(D \mid (x,u), 0\big) = \delta_\infty(D),$$

$$q^o\big(D \mid (x,u), 1\big) = \int_S \int_C \mathbb{1}_{D\setminus\{\infty\}}\big((x',u')\big)\, \tilde{\mu}(du' \mid x')\, q(dx' \mid x, u),$$

  where $\tilde{\mu}(du' \mid x')$ is a Borel measurable stochastic kernel on C given S such that

$$\tilde{\mu}(du' \mid x') = \mu(du' \mid x'), \quad \forall\, x' \in B.$$

(Such a kernel can be constructed by letting $\tilde{\mu}(du' \mid x') = \mu(du' \mid x')$ for $x' \in B$ and $\tilde{\mu}(du' \mid x') = p(du')$ for $x' \notin B$, where p is any Borel probability measure on C.) In particular, with the control 1, for any $(x,u) \in S \times C$ and Borel $D \subset S^o$,

$$q^o\big(D \mid (x,u),1\big) = \int_B \int_C \mathbb{1}_{D\setminus\{\infty\}}\big((x',u')\big)\mu(du' \mid x')\, q(dx' \mid x,u) + \int_{S\setminus B} \int_C \mathbb{1}_{D\setminus\{\infty\}}\big((x',u')\big)\, \tilde{\mu}(du' \mid x')\, q(dx' \mid x,u). \tag{A.1}$$

The above formulation fits the general stochastic control model described in Section 2.2. In particular, the graph of the control constraint U^o is an analytic set, the one-stage cost function g^o is lower semi-analytic, and the state transition kernel q^o is Borel measurable. For a state $z \in S^o$, we denote the cost of a universally measurable policy π^o by $V_{\pi^o}(z)$. It is as defined in Section 2.2 and can be expressed as follows. For $k \geq 0$, let (z_k, u_k^o) denote the state and control at time k, and let τ be the time when the process is stopped: $\tau = \min\big\{k \geq 0 \mid u_k^o = 0\big\}$ with $\tau = \infty$ if $\big\{k \geq 0 \mid u_k^o = 0\big\} = \emptyset$. For each k, let $(x_k, u_k) = z_k$ if $z_k \in S \times C$, and let (x_k, u_k) equal some fixed state in $S \times C$ if $z_k = \infty$. Then for $z_0 = (x,u) \in S \times C$, $V_{\pi^o}\big((x,u)\big)$ can be expressed as

$$V_{\pi^o}\big((x,u)\big) = \mathrm{E}^{\pi^o}\left\{\sum_{k=0}^{\infty} \alpha^k g^o\big(z_k, u_k^o\big)\right\} = \mathrm{E}^{\pi^o}\left\{\sum_{k=0}^{\tau-1} \alpha^k g^o\big((x_k,u_k),1\big) + \alpha^\tau J(x_\tau)\right\}. \tag{A.2}$$

Note that in the above, (x_k, u_k) is meaningfully defined on $\{\tau \geq k\}$.

Denote the optimal cost function by V^* and the optimal cost operator by T_o. The following lemma is a direct consequence of the theory for (D)(N)(P) in the case where the number of controls at each state is finite [7, Props. 9.8, 9.14, Cor. 9.17.1]. (We note that in case (N), an optimal policy need not exist even when the control space is finite. See [6, Ex. 4.1, p. 181] for such an example.)

Lemma A.1. (D)(N)(P) *The optimal cost function V^* is lower semi-analytic (bounded for (D), nonpositive for (N), and nonnegative for (P)), and satisfies*

$$T_o^k(\mathbf{0}) \to V^*, \qquad V^* = T_o\big(V^*\big).$$

For (N)(P), $T_o^k(\mathbf{0})$ converges monotonically. For (D), V^* is the unique solution to $V = T_o(V), V \in A_b(S^o)$. Furthermore, for (D)(P), there exists an optimal nonrandomized stationary policy.

Let $V_k = T_o^k(\mathbf{0})$, $k \geq 0$, be the optimal k-stage cost functions. To simplify notation we will write $V(x,u)$ for $V\big((x,u)\big)$. Clearly, for the absorbing state ∞ and for the states in $(S \setminus B) \times C$, where the only control is to stop, we have for all $k \geq 1$,

$$V^*(\infty) = V_k(\infty) = 0, \qquad V^*(x,u) = V_k(x,u) = J(x), \quad \forall\, (x,u) \in (S \setminus B) \times C. \tag{A.3}$$

Next we will calculate the optimal costs for states in the set $(B \times C) \cap \Gamma$ and relate the results to $Q_{\theta,J}$ and $F_\theta(\cdot\,; J)$. For our purposes, the set $(B \times C) \setminus \Gamma$ of states can be ignored, not only because they are outside the control constraint set of the original problem, but also because in the optimal stopping problem, they are formulated to be unreachable (as they should be) from the rest of the states. In particular, if the starting state (x,u) is in $(B \times C) \cap \Gamma$, then since the policy μ satisfies the control constraint of the original problem, we see from the first term in the expression (A.1) for the state transition probability $q^o(\cdot \mid (x,u),1)$ that the probability of the successor state being in $(B \times C) \setminus \Gamma$ is zero. If the starting state (x,u) is in $(S \setminus B) \times C$, then the control 1 (to continue) is not allowed according to the control constraint U^o, so the successor state is ∞. Therefore, the set $(B \times C) \setminus \Gamma$ is not reachable from the rest of the states.

Since at time k, the continuation cost is $g^o((x_k,u_k),1) = g(x_k,u_k)$ if $(x_k,u_k) \in (B\times C)\cap\Gamma$, the preceding discussion also shows that for each $(x,u)\in\Gamma$, the cost of π^o for the initial distribution $p^o(\cdot) = q^o(\cdot\,|\,(x,u),1)$ is

$$V_{\pi^o,p^o} = \mathrm{E}^{\pi^o,p^o}\left\{\sum_{k=0}^{\tau-1}\alpha^k g(x_k,u_k) + \alpha^\tau J(x_\tau)\right\}, \tag{A.4}$$

where the expectation is with respect to the probability measure induced by π^o and p^o (cf. Eq. (A.2)). We will use the expression (A.4) later to derive an expression for $Q_{\theta,J}$ (see Cor. A.1).

A.2 Relations with $F_\theta(\cdot\,;J), Q_{\theta,J}$

We will now express the operator T_o and calculate V_k, V^* for the states in $(B\times C)\cap\Gamma$. Consider the set of functions

$$\big\{V\in A(S^o)\mid V(\infty)=0, V(x,u)=J(x), (x,u)\in(S\setminus B)\times C\big\}, \tag{A.5}$$

which includes V^*, V_k (cf. Eq. (A.3)). For any V in this set, using the expression of $q^o(dz\,|\,(x,u),1)$ given in (A.1), we have that for any $(x,u)\in S\times C$,

$$\int_{S\times C} V(z)\,q^o\big(dz\,|\,(x,u),1\big) = \int_{S\setminus B} J(x')\,q(dx'\,|\,x,u) + \int_B\int_C V(x',u')\,\mu(du'\,|\,x')\,q(dx'\,|\,x,u), \tag{A.6}$$

and by a direct calculation we also have

$$T_o(V)(x,u) = \min\left\{J(x)\,,\; g(x,u)+\alpha\int_{S\times C} V(z)\,q^o\big(dz\,|\,(x,u),1\big)\right\}, \qquad (x,u)\in(B\times C)\cap\Gamma,$$

where the first term $J(x)$ is the stopping cost and the second term is associated with the continuation action. Therefore, for any V in the set (A.5),

$$T_o(V)(x,u) = \min\left\{J(x)\,,\; G_V(x,u)\right\}, \qquad (x,u)\in(B\times C)\cap\Gamma, \tag{A.7}$$

where

$$G_V(x,u) = g(x,u) + \alpha\int_{S\setminus B} J(x')\,q(dx'\,|\,x,u) + \alpha\int_B\int_C V(x',u')\,\mu(du'\,|\,x')\,q(dx'\,|\,x,u). \tag{A.8}$$

This yields the optimality equation $V = T_o(V)$ in a reduced form for V in the set (A.5).

Using the fact $V^* = T_o(V^*)$, we then obtain

$$V^*(x,u) = T_o\big(V^*\big)(x,u) = \min\left\{J(x)\,,\; f^*(x,u)\right\}, \qquad \forall\,(x,u)\in(B\times C)\cap\Gamma, \tag{A.9}$$

where $f^*(x,u)$ is the optimal expected future cost for continuation and can be expressed in several equivalent ways:

$$f^*(x,u) = g(x,u) + \alpha\int_{S\times C} V^*(z)\,q^o\big(dz\,|\,(x,u),1\big) \tag{A.10}$$

$$= g(x,u) + \alpha\int_{S\setminus B} J(x')\,q(dx'\,|\,x,u) + \alpha\int_B\int_C V^*(x',u')\,\mu(du'\,|\,x')\,q(dx'\,|\,x,u)$$

$$= g(x,u) + \alpha\int_{S\setminus B} J(x')\,q(dx'\,|\,x,u) + \alpha\int_B\int_C \min\big\{J(x')\,,\; f^*\big(x',u'\big)\big\}\,\mu(du'\,|\,x')\,q(dx'\,|\,x,u). \tag{A.11}$$

Here in deriving Eq. (A.11), we used the fact that for all $(x,u) \in S \times C$,

$$\int_B \int_C V^*(x',u')\,\mu(du' \mid x')\,q(dx' \mid x,u) = \int_B \int_C \min\big\{J(x')\,,\; f^*\big(x',u'\big)\big\}\,\mu(du' \mid x')\,q(dx' \mid x,u). \tag{A.12}$$

To see this, note that since μ satisfies the control constraint of the original problem, $\mu\big(U(x') \mid x'\big) = 1$ for $x' \in B$, and for $x' \in B$ and $u' \in U(x')$, $V^*(x',u')$ can be expressed as in (A.9).

Similar to the preceding derivation, we can calculate the optimal k-stage cost functions V_k, $k \geq 1$, and define functions f_k on $(B \times C) \cap \Gamma$ associated with the continuation action, for $k \geq 0$, by

$$f_k(x,u) = g(x,u) + \alpha \int_{S \times C} V_k(z)\,q^o\big(dz \mid (x,u),1\big), \qquad (x,u) \in (B \times C) \cap \Gamma,\ k \geq 0. \tag{A.13}$$

From the recursive relations,

$$V_{k+1}(x,u) = T_o\big(V_k\big)(x,u) = \min\{J(x)\,,\; f_k(x,u)\}\,, \qquad (x,u) \in (B \times C) \cap \Gamma,\ k \geq 0,$$

we obtain that the functions f_k, $k \geq 1$, satisfy the recursion (A.11) with f_k replacing f^* on the left-hand side and with f_{k-1} replacing f^* in the right-hand side.

We recognize the expression on the right-hand side of Eq. (A.11) as the same expression that defines $F_\theta(f^*;J)(x,u)$ (cf. Eq. (3.4)). To be more precise, since $F_\theta(\cdot\,;J)$ is a mapping on $A(\Gamma)$ and f^* is defined on $(B \times C) \cap \Gamma$, we will adopt the following convention: for any function f defined on $(B \times C) \cap \Gamma$, by $F_\theta(f\,;J)$ we mean any $F_\theta(f_e\,;J)$ where f_e is an (arbitrary) extension of f to Γ. This is valid because by definition $F_\theta(Q\,;J)$ is completely determined by the function Q restricted to $(B \times C) \cap \Gamma$. In other words, denoting $\Gamma_B = (B \times C) \cap \Gamma$, we have

$$Q\,|_{\Gamma_B} = Q'|_{\Gamma_B} \quad \Longrightarrow \quad F_\theta(Q\,;J) = F_\theta(Q';J). \tag{A.14}$$

Based on the equivalence between Eq. (A.11) and $F_\theta(f^*;J)(x,u)$, we can relate the optimal cost functions V^*, V_k of the optimal stopping problem to the mapping $F_\theta(\cdot\,;J)$ and the function $Q_{\theta,J} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,;J)$ as follows.

Lemma A.2. (D)(N)(P) *Let $\Gamma_B = (B \times C) \cap \Gamma$, and let $f^*, f_k : \Gamma_B \to [-\infty,\infty]$, $k \geq 0$, be the minimal future cost functions associated with continuation, given by Eqs. (A.10) and (A.13) respectively; in particular, $f_0 = g|_{\Gamma_B}$. Then*

$$f^* = F_\theta\big(f^*;J\big)\big|_{\Gamma_B}, \qquad f_k = F_\theta\big(f_{k-1};J\big)\,\big|_{\Gamma_B}, \quad k \geq 1,$$

and $f_k \to f^*$. Moreover,

$$Q_{\theta,J}\big|_{\Gamma_B} = f^*, \qquad Q_{\theta,J} = F_\theta\big(f^*;J\big). \tag{A.15}$$

Proof. The recursive relations for f^*, f_k were derived earlier. The fact $f_k \to f^*$ follows from Eqs. (A.10) and (A.13) by applying the bounded convergence theorem in case (D), and the monotone convergence theorem in cases (N)(P), using the convergence $V_k \to V^*$ in each of these cases (Lemma A.1).

We now prove the relation (A.15) between the function $Q_{\theta,J} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,;J)$ and f^*. Since $f_k \to f^*$, using the relation $f_k = F_\theta\big(f_{k-1};J\big)\,\big|_{\Gamma_B}$ and Eq. (A.14), we have $f_k = F_\theta^k(g\,;J)\big|_{\Gamma_B} \to f^*$. Suppose we have proved $F_\theta^k(g;J) \to Q_{\theta,J}$. Then it will follow that $Q_{\theta,J}\big|_{\Gamma_B} = f^*$. In turn, this will imply $F_\theta(Q_{\theta,J};J) = F_\theta(f^*;J)$ by Eq. (A.14), and hence $Q_{\theta,J} = F_\theta(f^*;J)$ since $Q_{\theta,J} = F_\theta(Q_{\theta,J};J)$ by Prop. 3.2. Thus it is sufficient to prove $F_\theta^k(g;J) \to Q_{\theta,J}$. For (D), this was proved by Lemma 4.1. For (N), we have $g \leq 0$ and $J \leq 0$. By a direct calculation, $F_\theta(\mathbf{0}\,;J) \leq g \leq 0$, so we have, by the monotonicity of $F_\theta(\cdot\,;J)$,

$$F_\theta^k(\mathbf{0}\,;J) \leq F_\theta^{k-1}(g;J) \leq F_\theta^{k-1}(\mathbf{0}\,;J), \quad k \geq 1.$$

Since $F_\theta^k(\mathbf{0}\,;J) \downarrow Q_{\theta,J}$ by Prop. 3.2, we have $F_\theta^k(g;J) \downarrow Q_{\theta,J}$. The convergence $F_\theta^k(g;J) \to Q_{\theta,J}$ in case (P) follows from a symmetrical argument. □

We see from Lemma A.2 that we may view $F_\theta(\cdot\,;J)$ as an optimal cost operator for the minimal future cost function f^* associated with the continuation action in the optimal stopping problem. For states $(x,u) \in \Gamma_B$, we can also interpret $Q_{\theta,J}(x,u)$ as the minimal costs at (x,u) with continuation at the first stage.

We now give several expressions of $Q_{\theta,J}(x,u)$ in terms of V^* and V_{π^o}, for all $(x,u) \in \Gamma$, in the following corollary. For each $(x,u) \in \Gamma$, we will consider the optimal stopping problem starting with an initial state distribution p^o given by $q^o(\cdot \mid (x,u),1)$, the transition distribution for (x,u) under the continuation action.

Corollary A.1. (D)(N)(P) *For all $(x,u) \in \Gamma$,*

$$\begin{aligned} Q_{\theta,J}(x,u) &= g(x,u) + \alpha \int_{S\times C} V^*(z)\, q^o(dz \mid (x,u),1) \\ &= g(x,u) + \alpha \inf_{\pi^o} \int_{S\times C} V_{\pi^o}(z)\, q^o(dz \mid (x,u),1). \end{aligned}$$

In particular, if in the optimal stopping problem associated with (θ,J), an optimal policy π^{o*} exists (as is true under (D)(P)), then for all $(x,u) \in \Gamma$,

$$Q_{\theta,J}(x,u) = g(x,u) + \alpha\, \mathrm{E}^{\pi^{o*},p^o} \left\{ \sum_{k=0}^{\tau-1} \alpha^k g(x_k,u_k) + \alpha^\tau J(x_\tau) \right\},$$

where $\tau = \min\{k \geq 0 \mid u_k^o = 0\}$ with $\tau = \infty$ if this set is empty, and the expectation is with respect to the probability measure induced by π^{o*} and the initial distribution p^o of (x_0,u_0), given by $p^o(\cdot) = q^o(\cdot \mid (x,u),1)$.

Proof. Since $Q_{\theta,J} = F_\theta(f^*;J)$ (Lemma A.2), using the definition of $F_\theta(\cdot\,;J)$ and Eq. (A.12), we have that for all $(x,u) \in \Gamma$,

$$Q_{\theta,J}(x,u) = g(x,u) + \alpha \int_{S\setminus B} J(x')\, q(dx' \mid x,u) + \alpha \int_B \int_C V^*(x',u')\, \mu(du' \mid x')\, q(dx' \mid x,u), \quad \text{(A.16)}$$

which together with (A.6) implies the first expression for $Q_{\theta,J}(x,u)$ in the corollary. The second expression for $Q_{\theta,J}$ in the corollary follows from the first one and [7, Cor. 9.5.2]. From the second expression and Eq. (A.4) for policy π^{o*}, we obtain the third expression for $Q_{\theta,J}$ in the corollary. □

A.3 A Useful Linear Program for Case (P)

As Cor. A.1 shows, we can obtain $Q_{\theta,J}$ from the optimal cost function V^* of the optimal stopping problem associated with (θ,J). For case (D) (resp. case (N)), the function V^* is the maximal solution to $V \leq T_o(V)$ among the set of bounded lower semi-analytic functions (resp. the set of nonpositive lower semi-analytic functions) [7, Props. 9.10, 9.15]. The inequality $V \leq T_o(V)$ can be expressed as a system of linear inequalities, so under suitable conditions, V^* can be obtained by solving a linear program. (See [27, Chap. 6] for standard linear programming formulations for DP problems with infinite state space.)

In case (P), however, V^* is the minimal nonnegative lower semi-analytic solution to $V \geq T_o(V)$ [7, Prop. 9.10(P)], and this in general does not admit a linear programming formulation. We consider below a linear program with linear constraints based on the inequality $V \leq T_o(V)$ instead. While

it does not yield V^* in general, under an assumption to be given shortly, we can use it to obtain an upper bound on V^* (in an almost-everywhere sense) and then an upper bound on $Q_{\theta,J}$ (see Lemma A.3). This bound on $Q_{\theta,J}$ can be used in a mixed value and policy iteration algorithm given in Section 5.2, which is convergent under certain initial conditions for case (P), as shown by Theorem 5.3.

Let $\Gamma_B = (B \times C) \cap \Gamma$ as earlier. Let $\mathscr{U}$ denote the universal σ-algebra on $S \times C$.

Assumption A.1. (P) *There exists a σ-finite measure ρ on $(S \times C, \mathscr{U})$ such that*

(i) $\int_{\Gamma_B} J(x)\rho\big(d(x,u)\big) < \infty$*; and*

(ii) for each $(x,u) \in \Gamma_B$, the measure $\rho_{x,u}$ on $(S \times C, \mathscr{U})$ given by

$$\rho_{x,u}(D) = \int_B \int_C \mathbb{1}_D(x',u')\,\mu(du' \mid x')\,q(dx' \mid x,u), \qquad D \in \mathscr{U},$$

is absolutely continuous with respect to ρ (i.e., $\rho(D) = 0 \Rightarrow \rho_{x,u}(D) = 0$).

Suppose Assumption A.1 holds (which is the case if S, C are countable and J is finite on B; see Remark A.1). Let $A_+(\Gamma_B)$ denote the set of nonnegative, lower semi-analytic functions on Γ_B. Let $\Gamma_{B,\rho} \subset \Gamma_B$ be such that $\rho(\Gamma_B \setminus \Gamma_{B,\rho}) = 0$. We consider a linear program on the space $A_+(\Gamma_B)$:

$$\begin{aligned} \underset{V \in A_+(\Gamma_B)}{\text{Maximize}} \quad & \int_{\Gamma_B} V(x,u)\,\rho\big(d(x,u)\big) \\ \text{Subject to:} \quad & V(x,u) \le J(x), \quad \forall\,(x,u) \in \Gamma_{B,\rho}, \\ & V(x,u) \le g(x,u) + \int_{S\setminus B} J(x')\,q(dx' \mid x,u) \\ & \qquad + \int_B \int_C V(x',u')\,\mu(du' \mid x')\,q(dx' \mid x,u), \quad \forall\,(x,u) \in \Gamma_{B,\rho}. \end{aligned} \tag{A.17}$$

As can be seen from the expression (A.7)-(A.8) for the operator T_o, this linear program corresponds to the following maximization problem:

$$\begin{aligned} \underset{V \in A_+(\Gamma_B)}{\text{Maximize}} \quad & \int_{\Gamma_B} V(x,u)\,\rho\big(d(x,u)\big) \\ \text{Subject to:} \quad & V(x,u) \le T_o(V^e)(x,u), \quad \rho\text{-almost every } (x,u) \in \Gamma_B, \end{aligned}$$

where V^e is the extension of V on S^o with $V^e(\infty) = 0$, $V^e(x,u) = J(x), (x,u) \in (S \setminus B) \times C$.

Corresponding to any optimal solution $\bar{V}$ of (A.17), we define $\bar{Q} \in A_+(\Gamma)$ by the expression on the right-hand side of the second constraint in (A.17), with $\bar{V}$ in place of V and for all (x,u) in Γ:

$$\bar{Q}(x,u) = g(x,u) + \int_{S\setminus B} J(x')\,q(dx' \mid x,u) + \int_B \int_C \bar{V}(x',u')\,\mu(du' \mid x')\,q(dx' \mid x,u), \quad (x,u) \in \Gamma. \tag{A.18}$$

The next lemma shows that $\bar{Q}$ satisfies a property needed for the convergence analysis of the mixed value and policy iteration algorithm (5.14)-(5.15) discussed in Section 5.2.

Lemma A.3. (P) *Let Assumption A.1 hold. Then an optimal solution $\bar{V}$ of the linear program (A.17) exists, and the function $\bar{Q} \in A_+(\Gamma)$ given by Eq. (A.18) satisfies*

$$\bar{Q} \le F_\theta\big(\bar{Q}\,; J\big), \qquad \bar{Q} \ge Q_{\theta,J}.$$

Proof. Since $V^* = T_o(V^*)$, the optimal cost function V^* restricted to Γ_B is a feasible solution of (A.17), so the feasible set of (A.17) is nonempty. By Assumption A.1(i), the optimal objective value v^* of (A.17) is finite. Let $\bar{V}_n, n \geq 1$, be a sequence of feasible solutions with their objective values approaching v^*. Then the function resulting from taking pointwise supremum, $\sup_n \bar{V}_n$, lies in $A_+(\Gamma_B)$ [7, Lemma 7.30(2)], satisfies the constraints of (A.17), and achieves the optimal value v^*. It is hence an optimal solution of (A.17). This shows that an optimal solution $\bar{V}$ of (A.17) exists.

The function $\max\{V^*, \bar{V}\}$ on Γ_B is then an optimal solution of (A.17) as well. This implies that

$$V^*(x,u) \leq \bar{V}(x,u) \qquad \text{for } \rho\text{-almost every } (x,u) \in \Gamma_B, \tag{A.19}$$

for otherwise, by Assumption A.1(i) we would have

$$\infty > \int_{\Gamma_B} \max\big\{V^*(x,u), \bar{V}(x,u)\big\}\, \rho\big(d(x,u)\big) > \int_{\Gamma_B} \bar{V}(x,u)\, \rho\big(d(x,u)\big),$$

a contradiction to the optimality of $\bar{V}$. We now show $\bar{Q} \geq Q_{\theta,J}$. By Eq. (A.16), for all $(x,u) \in \Gamma$, $Q_{\theta,J}(x,u)$ equals the right-hand side of Eq. (A.18) with V^* in place of $\bar{V}$. This, together with Assumption A.1(ii) and the relation (A.19), implies $Q_{\theta,J} \leq \bar{Q}$. To show $\bar{Q} \leq F_\theta\big(\bar{Q}\,;J\big)$, notice that by the feasibility of $\bar{V}$ for (A.17) and the definition of $\bar{Q}$,

$$\bar{V}(x,u) \leq \min\big\{J(x)\,,\, \bar{Q}(x,u)\big\}, \qquad \forall\,(x,u) \in \Gamma_{B,\rho}.$$

We use this relation to upper-bound $\bar{V}$ ρ-almost everywhere on Γ_B, in the integral on the right-hand side of (A.18), which defines $\bar{Q}$. Using also Assumption A.1(ii), we then obtain that for all $(x,u) \in \Gamma$,

$$\bar{Q}(x,u) \leq g(x,u) + \int_{S\setminus B} J(x')\, q(dx' \mid x,u) + \int_B \int_C \min\big\{J(x)\,,\, \bar{Q}(x',u')\big\}\; \mu(du' \mid x')\, q(dx' \mid x,u),$$

which is the inequality $\bar{Q} \leq F_\theta\big(\bar{Q}\,;J\big)$. This completes the proof. □

Remark A.1. Assumption A.1 holds in particular when the state and control spaces S and C are countable sets and the function J is finite on B. Without loss of generality, suppose $S = C = \{1,2,\ldots\}$. Denote by $\rho(x,u)$ the mass assigned to a point $(x,u) \in S \times C$ by the measure ρ in Assumption A.1. Then Assumption A.1 is satisfied by letting $\rho(x,u) = 2^{-(x+u)}/(J(x)+1)$ if $(x,u) \in \Gamma_B$, and $\rho(x,u) = 0$ otherwise, for instance.

In the case where μ is a nonrandomized policy, we may let $\rho(x,\mu(x)) = 2^{-x}/(J(x)+1)$ if $x \in B$ and let $\rho(x,u) = 0$ for all the other (x,u). Then, with $\Gamma_{B,\rho} = \{(x,\mu(x)) \mid x \in B\}$, the linear program (A.17) involves only the variables $V(x,\mu(x)), x \in B$, and with the change of variable $W(x) = V(x,\mu(x))$, it becomes:

$$\begin{aligned}
\underset{W \geq 0}{\text{Maximize}} \quad & \sum_{x\in B} W(x)\, \rho\big(x,\mu(x)\big) \\
\text{Subject to:} \quad & W(x) \leq J(x), \quad \forall\, x \in B, \\
& W(x) \leq g\big(x,\mu(x)\big) + \sum_{x'\in S\setminus B} J(x')\, q\big(x' \mid x,\mu(x)\big) + \sum_{x'\in B} W(x')\, q\big(x' \mid x,\mu(x)\big), \quad \forall\, x \in B.
\end{aligned}$$

Although S is countable, if B is a finite set, this is a finite-dimensional linear program. □

B Proof of $Q_{\theta,J^*} = Q^*$ for Nonnegative Case (P)

In this appendix we prove for the nonnegative case (P) that for any $\theta \in \Theta$, the function $Q_{\theta,J^*} = \lim_{k\to\infty} F_\theta^k(\mathbf{0}\,;J^*)$ is Q^* given in Eq. (3.1). This establishes Prop. 3.3(c) for (P), which is also used in the lower bound part of Lemma 4.2 for (P).

Proposition B.1. (P) *Let $\theta = (\mu, B) \in \Theta$. We have $Q_{\theta,J^*} = Q^*$.*

Since $Q^* \geq 0$ and $F_\theta(Q^*; J^*) = Q^*$ (Prop. 3.3(a)), we have by the monotonicity of F_θ (cf. Eq. (3.7)),

$$Q_{\theta,J^*} = \lim_{n\to\infty} F_\theta^n(\mathbf{0}\,; J^*) \leq Q^*.$$

Thus to prove Prop. B.1, we need to show $Q_{\theta,J^*} \geq Q^*$. We will prove this by showing that for each $(x,u) \in \Gamma$ and any $\epsilon > 0$,

$$Q_{\theta,J^*}(x,u) \geq Q^*(x,u) - \epsilon. \tag{B.1}$$

In the proof we will use the correspondence between the optimal stopping problem associated with $\theta = (\mu, B)$ and J^*, as defined in Appendix A.1, and a controller for the original problem.

We need some notations and an expression of Q_{θ,J^*} to be used in the proof. Fix $(\bar{x}, \bar{u}) \in \Gamma$. For the optimal stopping problem associated with $\theta = (\mu, B)$ and J^*, by [7, Cor. 9.17.1], there exists an optimal stationary nonrandomized (universally measurable) policy $\mu^o : S^o = (S \times C) \cup \{\infty\} \to \{0,1\}$. Let the optimal stopping problem start from time 1, and consider the stochastic process $(z_1, u_1^o), (z_2, u_2^o), \ldots$, where $z_k \in S^o$ and $u_k^o \in \{0,1\}$, induced by μ^o and the initial distribution of z_1 given by $q^o(\cdot \mid (\bar{x}, \bar{u}), 1)$ (cf. Eq. (A.1)). For each $k \geq 1$, define $(x_k, v_k) = z_k$ if $z_k \in S \times C$, and define (x_k, v_k) to be some fixed point in $S \times C$ if $z_k = \infty$ (the absorbing state). Here for clarity, we are using v_k instead of u_k to denote the component of z_k in C, since we will use u_k later for the controls applied in the original problem. By Cor. A.1 we have

$$Q_{\theta,J^*}(\bar{x}, \bar{u}) = g(\bar{x}, \bar{u}) + \mathrm{E}^{\mu^o}\left\{\sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau)\right\}, \tag{B.2}$$

where τ is the time the process is stopped: $\tau = \min\left\{k \geq 1 \mid \mu^o(x_k, v_k) = 0\right\}$ (∞ if the set is empty), and E^{μ^o} denotes expectation with respect to the probability measure induced by μ^o and the initial distribution of (x_1, v_1), given by $q^o(\cdot \mid (\bar{x}, \bar{u}), 1)$.

To simplify notation, let

$$D = \{(x,v) \in S \times C \mid \mu^o(x,v) = 0\}, \qquad D_x = \{v \in C \mid (x,v) \in D\}, \quad x \in S.$$

(D is the subset of $S \times C$ on which μ^o stops the process.) Since μ^o is universally measurable, D and hence D_x, $x \in S$, are universally measurable sets [7, Lemma 7.29]. Note that expressed in terms of these sets, $\tau = \min\left\{k \geq 1 \mid v_k \in D_{x_k}\right\}$ and

$$\tau = m \qquad \Longleftrightarrow \qquad v_1 \notin D_{x_1},\ \cdots,\ v_{m-1} \notin D_{x_{m-1}},\ v_m \in D_{x_m}, \tag{B.3}$$

$$\tau > m \qquad \Longleftrightarrow \qquad v_1 \notin D_{x_1},\ \cdots,\ v_{m-1} \notin D_{x_{m-1}},\ v_m \notin D_{x_m}. \tag{B.4}$$

We consider also the probability measure on the space of $(x_1, v_1, x_2, v_2, \ldots)$ induced by the policy μ and the initial distribution $q(dx_1 \mid \bar{x}, \bar{u})$ of x_1. Let τ be the same as defined earlier. Let us agree that in this appendix, the expectation $\mathrm{E}^{\mu}\left\{\sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau)\right\}$ is with respect to the induced probability measure just mentioned.

Lemma B.1. *We have*

$$\mathrm{E}^{\mu^o}\left\{\sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau)\right\} = \mathrm{E}^{\mu}\left\{\sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau)\right\}.$$

We first prove Prop. B.1, assuming that Lemma B.1 has been proved, and then give the proof of this lemma.

Proof of Prop. B.1. Fix $(\bar{x}, \bar{u}) \in \Gamma$ and let $\epsilon > 0$. Let $\pi^\epsilon = (\pi_0^\epsilon, \pi_1^\epsilon, \ldots) \in \Pi$ be an ϵ-optimal Markov policy for the original control problem; such a policy exists by [7, Prop. 9.19]. We use the policies μ^o, π^ϵ, and μ (the stationary policy defining F_θ and the associated optimal stopping problem) to define a controller $\hat{\pi}$ for the original problem, such that it applies control $\bar{u}$ at state $\bar{x}$ at the first stage, and its expected total cost for state $\bar{x}$ is no greater than $Q_{\theta,J^*}(\bar{x}, \bar{u}) + \epsilon$. From this the desired inequality (B.1) for establishing the proposition will be shown to follow.

Roughly speaking, the controller $\hat{\pi}$ follows the policy μ before it switches to following policy π^ϵ. To decide when to make the switch, it generates at each time $k \geq 1$, an auxiliary variable $v_k \in C$ to "simulate" a control that μ might apply at the current state and "query" the optimal-stopping policy μ^o about whether that control suggested by μ should be followed. The history at time $k \geq 1$ for the controller is

$$(x_0, u_0, x_1, v_1, u_1, \ldots, x_k, v_k) \in (S \times C) \times (S \times C^2)^{k-1} \times (S \times C),$$

including the auxiliary variables $v_j, 1 \leq j \leq k$, as well as the past states $x_j, j \leq k$, and past controls $u_j, j < k$. The controller is denoted $\hat{\pi} = (\hat{\mu}_0, \hat{\mu}_1, \ldots)$, where each $\hat{\mu}_k$ is a universally measurable stochastic kernel on C given the respective space of histories. We now define $\hat{\mu}_k, k \geq 0$.

For $k = 0$, let $\hat{\mu}_0$ be a universally measurable stochastic kernel on C given S, such that $\hat{\mu}_0$ satisfies the control constraint U and $\hat{\mu}_0(du_0 \,|\, \bar{x}) = \delta_{\bar{u}}$. For each $k \geq 1$, the auxiliary variable v_k is generated according to the stochastic kernel μ given the state x_k:

$$\mu\big(dv_k \,|\, x_k\big). \tag{B.5}$$

The stochastic kernels $\hat{\mu}_k$, $k \geq 1$, are defined as follows. For each $k \geq 1$, define a universally measurable function $\tau_k : (S \times C)^k \to \{0, 1, 2, \ldots\}$ by

$$\tau_k(x_1, v_1, x_2, v_2, \ldots, x_k, v_k) = \begin{cases} \min\big\{m \mid v_m \in D_{x_m},\, 1 \leq m \leq k\big\} & \text{if such } m \text{ exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{B.6}$$

Let $\hat{\mu}_k$ be a universally measurable stochastic kernel on C given $(S \times C) \times (S \times C^2)^{k-1} \times (S \times C)$, given by

$$\hat{\mu}_k(du_k \,|\, x_0, u_0, x_1, v_1, u_1, \ldots, x_k, v_k) = \begin{cases} \delta_{v_k} & \text{if } \tau_k(x_1, v_1, \ldots, x_k, v_k) = 0, \\ \pi^\epsilon_{k-m}(du_k \,|\, x_k) & \text{if } \tau_k(x_1, v_1, \ldots, x_k, v_k) = m. \end{cases} \tag{B.7}$$

(I.e., $\hat{\pi}$ "copies" the control v_k if it has not yet switched to applying policy π^ϵ, and the switch happens the first time $v_m \in D_{x_m}$.)

The controller $\hat{\pi} = (\hat{\mu}_0, \hat{\mu}_1, \ldots)$ induces a probability measure on the space $(S \times C) \times (S \times C^2)^\infty$ of $(x_0, u_0, x_1, v_1, u_1, x_2, v_2, u_2, \ldots)$ (with the universal σ-algebra). With respect to this probability, the expected total cost of $\hat{\pi}$ for state $\bar{x}$ is

$$\hat{J}_{\hat{\pi}}(\bar{x}) = g(\bar{x}, \bar{u}) + \mathrm{E}^{\hat{\pi}} \left\{ \sum_{k=1}^{\infty} g(x_k, u_k) \right\}.$$

Let $\tau = \min\big\{k \geq 1 \mid v_k \in D_{x_k}\big\}$ with $\tau = \infty$ if the set in the definition is empty. Using the definition of conditional expectation and a formula for conditional expectation given the sub-σ-algebra associated with the stopping time τ [36, Prop. II-1-3], it follows that

$$\begin{aligned} \hat{J}_{\hat{\pi}}(\bar{x}) &= g(\bar{x}, \bar{u}) + \mathrm{E}^{\hat{\pi}} \left\{ \sum_{k=1}^{\tau-1} g(x_k, u_k) + J_{\pi^\epsilon}(x_\tau) \right\} \\ &= g(\bar{x}, \bar{u}) + \mathrm{E}^{\hat{\pi}} \left\{ \sum_{k=1}^{\tau-1} g(x_k, v_k) + J_{\pi^\epsilon}(x_\tau) \right\}, \end{aligned} \tag{B.8}$$

where in (B.8) we used the fact $u_k = v_k$ for $k < \tau$ (cf. Eqs. (B.6), (B.7)). We have

$$J_{\pi^\epsilon}(x) \leq J^*(x) + \epsilon, \qquad \forall\, x \in S,$$

since π^ϵ is an ϵ-optimal policy of the original problem and $J^* \geq 0$. Then by Eq. (B.8),

$$\hat{J}_{\hat{\pi}}(\bar{x}) \leq g(\bar{x},\bar{u}) + \mathrm{E}^{\hat{\pi}}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\} + \epsilon.$$

On the other hand, based on the definition of $\hat{\pi}$, $\{v_k\}$ are generated according to μ (cf. Eq. (B.5)) and they are the controls applied before time τ (cf. Eqs. (B.6), (B.7)), so we have

$$\mathrm{E}^{\hat{\pi}}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\} = \mathrm{E}^{\mu}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\} = \mathrm{E}^{\mu^o}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\},$$

where the second equality follows from Lemma B.1. Combining the preceding two relations with Eq. (B.2), we obtain

$$\hat{J}_{\hat{\pi}}(\bar{x}) \leq g(\bar{x},\bar{u}) + \mathrm{E}^{\mu^o}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\} + \epsilon = Q_{\theta,J^*}(\bar{x},\bar{u}) + \epsilon. \tag{B.9}$$

Although the controller $\hat{\pi}$ uses the additional auxiliary variables $\{v_k\}$ for control, it does not have advantages over the set of policies in Π', in the sense that we can construct a semi-Markov randomized policy such that it applies control $\bar{u}$ at the first stage if $\bar{x}$ is the initial state, and it has the same expected total cost $\hat{J}_{\hat{\pi}}(\bar{x})$ for the state $\bar{x}$. (Such a construction is similar to that used to prove Props. 8.1, 9.1 in [7].) This means that $\hat{J}_{\hat{\pi}}(\bar{x}) \geq Q^*(\bar{x},\bar{u})$ (cf. Eq. (3.2)), so by Eq. (B.9),

$$Q^*(\bar{x},\bar{u}) \leq Q_{\theta,J^*}(\bar{x},\bar{u}) + \epsilon.$$

Since ϵ is arbitrary, we have $Q_{\theta,J^*}(\bar{x},\bar{u}) \geq Q^*(\bar{x},\bar{u})$. This proves the proposition, as discussed immediately after the proposition. □

We now establish Lemma B.1.

Proof of Lemma B.1. We need to prove

$$\mathrm{E}^{\mu^o}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\} = \mathrm{E}^{\mu}\left\{\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right\}.$$

First, we introduce some functions to rewrite the two expectations above. In view of (B.3) (i.e., $\tau = m \Leftrightarrow v_1 \notin D_{x_1}, \cdots, v_{m-1} \notin D_{x_{m-1}}, v_m \in D_{x_m}$), we have that for each $m \geq 1$,

$$\mathbb{1}_{\{\tau=m\}}(x_1,v_1,\ldots)\cdot\left(\sum_{k=1}^{\tau-1} g(x_k,v_k) + J^*(x_\tau)\right) = \phi_m(x_1,v_1,\ldots,x_m,v_m)$$

where $\phi_m : (S\times C)^m \to [0,\infty]$ is given by

$$\phi_m(x_1,v_1,\ldots,x_m,v_m) = \left(\prod_{i=1}^{m-1} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right)\cdot \mathbb{1}_{D_{x_m}}(v_m)\cdot\left(\sum_{k=1}^{m-1} g(x_k,v_k) + J^*(x_m)\right),$$

and for $m = \infty$,

$$\mathbb{1}_{\{\tau=\infty\}}(x_1, v_1, \ldots) \cdot \sum_{k=1}^{\infty} g(x_k, v_k) = \phi_\infty(x_1, v_1, x_2, v_2, \ldots)$$

where $\phi_\infty : (S \times C)^\infty \to [0, \infty]$ is given by

$$\phi_\infty(x_1, v_1, x_2, v_2, \ldots) = \left(\prod_{i=1}^{\infty} \mathbb{1}_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^{\infty} g(x_k, v_k).$$

Since $g \geq 0, J^* \geq 0$, we may write

$$\begin{aligned} \mathrm{E}^{\mu^o} \left\{ \sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau) \right\} &= \sum_{m=1}^{\infty} \mathrm{E}^{\mu^o} \left\{ \mathbb{1}_{\{\tau=m\}}(x_1, v_1, \ldots) \cdot \left(\sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right) \right\} \\ &\quad + \mathrm{E}^{\mu^o} \left\{ \mathbb{1}_{\{\tau=\infty\}}(x_1, v_1, \ldots) \cdot \sum_{k=1}^{\infty} g(x_k, v_k) \right\} \\ &= \sum_{m \in \{1,2,\ldots\} \cup \{\infty\}} \mathrm{E}^{\mu^o} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\}, \end{aligned} \tag{B.10}$$

and similarly,

$$\mathrm{E}^{\mu} \left\{ \sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau) \right\} = \sum_{m \in \{1,2,\ldots\} \cup \{\infty\}} \mathrm{E}^{\mu} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\}. \tag{B.11}$$

To prove that (B.10) and (B.11) are equal, we will proceed in four steps.

(i) First, we show that for each $m \geq 1$,

$$\mathrm{E}^{\mu^o} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\} = \mathrm{E}^{\mu} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\}. \tag{B.12}$$

Note that $\phi_m(x_1, v_1, \ldots, x_m, v_m) = 0$ on $\{\tau \neq m\}$, and $\tau \neq m$ if $x_i \notin B$ for some $i < m$ (since in the optimal stopping problem, the only control that μ^o can take for states in $(S \setminus B) \times C$ is to stop, $x_i \notin B$ implies $\tau \leq i$). Using these facts together with the definition of E^{μ}, we have that $\mathrm{E}^{\mu} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\}$ is equal to

$$\begin{aligned} &\int_B \int_C \cdots \int_B \int_C \left[\int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \mu(dv_m \mid x_m)\, q(dx_m \mid x_{m-1}, v_{m-1}) \right] \cdot \\ &\qquad \mu(dv_{m-1} \mid x_{m-1})\, q(dx_{m-1} \mid x_{m-2}, v_{m-2}) \cdots \mu(dv_1 \mid x_1)\, q(dx_1 \mid \bar{x}, \bar{u}). \end{aligned} \tag{B.13}$$

Using the same facts just mentioned, and using also the definition of the optimal stopping problem (Appendix A.1), we have that $\mathrm{E}^{\mu^o} \big\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \big\}$ is equal to

$$\begin{aligned} &\int_B \int_C \cdots \int_B \int_C \left[\int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \tilde{\mu}(dv_m \mid x_m)\, q(dx_m \mid x_{m-1}, v_{m-1}) \right] \cdot \\ &\qquad \tilde{\mu}(dv_{m-1} \mid x_{m-1})\, q(dx_{m-1} \mid x_{m-2}, v_{m-2}) \cdots \tilde{\mu}(dv_1 \mid x_1)\, q(dx_1 \mid \bar{x}, \bar{u}). \end{aligned}$$

Since $\tilde{\mu}(\cdot \mid x) = \mu(\cdot \mid x)$ for $x \in B$ by the definition of the optimal stopping problem, the above integral in turn equals

$$\begin{aligned} &\int_B \int_C \cdots \int_B \int_C \left[\int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \tilde{\mu}(dv_m \mid x_m)\, q(dx_m \mid x_{m-1}, v_{m-1}) \right] \cdot \\ &\qquad \mu(dv_{m-1} \mid x_{m-1})\, q(dx_{m-1} \mid x_{m-2}, v_{m-2}) \cdots \mu(dv_1 \mid x_1)\, q(dx_1 \mid \bar{x}, \bar{u}). \end{aligned} \tag{B.14}$$

Consider now the inner-most integral in (B.14). If $x_m \in S \setminus B$, then in view of the control constraint U^o of the optimal stopping problem (cf. Appendix A.1), we have $D_{x_m} = C$. Hence $\mathbb{1}_{D_{x_m}}(v_m) = 1$ for all $v_m \in C$, so

$$\phi_m(x_1, v_1, \ldots, x_m, v_m) = \left(\prod_{i=1}^{m-1} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \left(\sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m)\right)$$

does not depend on v_m. Consequently,

$$\int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \tilde{\mu}(dv_m \,|\, x_m) = \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \mu(dv_m \,|\, x_m), \quad x_m \in S \setminus B.$$

If $x_m \in B$, then since $\tilde{\mu}(\cdot \,|\, x) = \mu(\cdot \,|\, x)$ for $x \in B$, we have

$$\int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \tilde{\mu}(dv_m \,|\, x_m) = \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m)\, \mu(dv_m \,|\, x_m), \quad x_m \in B.$$

The preceding two equalities together imply that the value of the integral (B.14) remains unchanged if we replace $\tilde{\mu}(dv_m \,|\, x_m)$ in the inner-most integral in (B.14) by $\mu(dv_m \,|\, x_m)$. Hence the integral (B.14) is equal to the integral (B.13), and this proves the desired equality (B.12) for $m \geq 1$.

(ii) By arguments similar to the ones in the preceding proof, we have that for all $m \geq 1$ and $n \geq m$,

$$\mathrm{E}^{\mu^o}\left\{\left(\prod_{i=1}^{n} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\} = \mathrm{E}^{\mu}\left\{\left(\prod_{i=1}^{n} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\}. \tag{B.15}$$

In particular, observing that $\prod_{i=1}^{n} \mathbb{1}_{C\setminus D_{x_i}}(v_i) = 0$ if $x_i \notin B$ for some $i \leq n$, an analysis similar to the first half of the proof in (i) then shows that both sides of (B.15) are equal to

$$\int_B \int_C \cdots \int_B \int_C \left(\prod_{i=1}^{n} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \left(\sum_{k=1}^{m} g(x_k, v_k)\right) \mu(dv_n \,|\, x_n)\, q(dx_n \,|\, x_{n-1}, v_{n-1}) \cdot \\ \cdots \mu(dv_1 \,|\, x_1)\, q(dx_1 \,|\, \bar{x}, \bar{u}).$$

We will need (B.15) shortly in the proof.

(iii) Let us now consider the two terms corresponding to $m = \infty$ in Eqs. (B.10) and (B.11). We examine when they are equal, i.e., when

$$\mathrm{E}^{\mu^o}\big\{\phi_\infty(x_1, v_1, x_2, v_2, \ldots)\big\} = \mathrm{E}^{\mu}\big\{\phi_\infty(x_1, v_1, x_2, v_2, \ldots)\big\}. \tag{B.16}$$

From the definition of ϕ_∞, we have, by the monotone convergence theorem, that as $m \to \infty$,

$$\mathrm{E}^{\mu^o}\left\{\left(\prod_{i=1}^{\infty} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\} \uparrow \mathrm{E}^{\mu^o}\big\{\phi_\infty(x_1, v_1, x_2, v_2, \ldots)\big\},$$

$$\mathrm{E}^{\mu}\left\{\left(\prod_{i=1}^{\infty} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\} \uparrow \mathrm{E}^{\mu}\big\{\phi_\infty(x_1, v_1, x_2, v_2, \ldots)\big\}.$$

Thus Eq. (B.16) holds if for each $m \geq 1$,

$$\mathrm{E}^{\mu^o}\left\{\left(\prod_{i=1}^{\infty} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\} = \mathrm{E}^{\mu}\left\{\left(\prod_{i=1}^{\infty} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k)\right\}. \tag{B.17}$$

Now, for each m, we have the following pointwise convergence of functions as $n \to \infty$:

$$\left(\prod_{i=1}^{n} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k) \;\downarrow\; \left(\prod_{i=1}^{\infty} \mathbb{1}_{C\setminus D_{x_i}}(v_i)\right) \cdot \sum_{k=1}^{m} g(x_k, v_k).$$

We also have the equality (B.15) for all $n \geq m$. Hence, if for each m there exists some $n \geq m$ for which the quantity in (B.15) is less than ∞, then by the dominated convergence theorem [19, p. 132], (B.17) holds for each m, and hence the desired equality (B.16) holds, which together with (B.12) implies that (B.10) and (B.11) are equal.

(iv) The only case left now is that for some m and all $n \geq m$, the quantity in (B.15) is ∞. But in view of Eq. (B.4) (i.e., $\tau > m \Leftrightarrow v_1 \notin D_{x_1}, \cdots, v_{m-1} \notin D_{x_{m-1}}, v_m \notin D_{x_m}$), this would imply

$$\mathrm{E}^{\mu^o}\left\{\mathbb{1}_{\{\tau>m\}}(x_1, x_2, \ldots)\sum_{k=1}^{\tau-1} g(x_k, v_k)\right\} = \infty, \qquad \mathrm{E}^{\mu}\left\{\mathbb{1}_{\{\tau>m\}}(x_1, x_2, \ldots)\sum_{k=1}^{\tau-1} g(x_k, v_k)\right\} = \infty,$$

and hence both (B.10) and (B.11) equal ∞. This completes the proof. □

C An Illustrative Example for Value Iteration in Case (P)

In this appendix we use an example to illustrate Theorem 5.1(b) for the convergence of value iteration in case (P). This example is from Strauch [47, Example 6.2, p. 881] and also described in Maitra and Sudderth [33, p. 930]. Our description below closely follows [33].

Let $\mathcal{R}_{(0,1)}$ denote the set of rationals in $(0,1)$ with its usual ordering, and index its elements by $r_1, r_2, \ldots$ Let $\{W_r \mid r \in \mathcal{R}_{(0,1)}\}$ be a collection of Borel subsets of $[0,1]$ (called a Borel sieve). Correspondingly, define for each $z \in [0,1]$,

$$M_z = \{r \in \mathcal{R}_{(0,1)} \mid z \in W_r\}, \qquad D = \{z \in [0,1] \,|\, M_z \text{ is not well-ordered}\}.$$

Fix the sets $\{W_r\}$ such that the set D is not Borel measurable. Define the control problem as follows.

Let $S = \{(z,r) \mid 0 \leq z \leq 1,\ 0 \leq r \leq 1,\ r \text{ rational}\} \cup \{t\}$. Let $C = \{1, 2, \ldots\}$ and $U(x) = C$ for every state $x \in S$. State transitions are deterministic. The successor state $f(x,u)$ when applying control u at state x is given by

$$f(t,u) = t, \qquad f\big((z,r),u\big) = \begin{cases} (z, r_u) & \text{if } r_u < r \text{ and } z \in W_{r_u}, \\ t & \text{otherwise.} \end{cases}$$

The cost $\hat{g}(x,u,x')$ of transition from state x to state x' is given by

$$\hat{g}(t,u,t) = 0, \qquad \hat{g}\big((z,r),u,x'\big) = \begin{cases} 0 & \text{if } x' \neq t, \\ 1 & \text{otherwise.} \end{cases}$$

Equivalently, the one-stage costs are:

$$g(t,u) = 0, \qquad g\big((z,r),u\big) = \begin{cases} 0 & \text{if } r_u < r \text{ and } z \in W_{r_u}, \\ 1 & \text{otherwise.} \end{cases}$$

The optimal cost function J^* takes only values 0 or 1, and it is not Borel measurable [47]. In particular, $J^*(t) = 0$ and for states $(z,1)$ where $z \in [0,1]$, as shown by [47],

$$J^*(z,1) = \begin{cases} 0 & \text{if } z \in D, \\ 1 & \text{if } z \in [0,1] \setminus D. \end{cases} \tag{C.1}$$

Value iteration starting from the constant function zero requires uncountably many iterations to converge to J^*, as shown by Maitra and Sudderth [33, p. 930].

We have the convergence of value iteration $T^k(J) \to J^*$, if we let J be $J(t) = 0$ and for states $x = (z, r)$,

$$J(z,r) = \begin{cases} 0 & \text{if } (z,r) \in G, \\ v & \text{otherwise,} \end{cases} \qquad \text{for some constant } v \geq 1,$$

where

$$G = \big\{ x \in S \mid J^*(x) = 0 \big\}.$$

This function J satisfies the condition of Theorem 5.1(b) for the convergence of value iteration, since $J^* \leq J \leq vJ^*$.

Indeed $T(J) = J^*$, as can be verified directly. Consider each $(z, r) \in S$ where $z \in [0, 1]$. If $(z, r) \in G$, then by the definition of the control problem given above and by the relation $J^* = T(J^*)$, there must exists some $u \in C$ such that with $x' = f\big((z,r), u\big)$,

$$J^*(z,r) = \hat{g}\big((z,r), u, x'\big) + J^*(x') = 0,$$

which implies that $\hat{g}\big((z,r), u, x'\big) = 0$ and $x' \in G$. Consequently, we have $T(J)(z,r) = 0 = J^*(z,r)$.

Suppose $(z, r) \notin G$, i.e., $J^*(z,r) = 1$. Then, by the relation $J^* = T(J^*)$ and the binary nature of the costs, we must have that for each $u \in C$, either (i) or (ii) can happen:

(i) $f\big((z,r), u\big) = t$ and $\hat{g}\big((z,r), u, t\big) = 1$, in which case

$$\hat{g}\big((z,r), u, t\big) + J(t) = 1.$$

(ii) $x' = f\big((z,r), u\big) \neq t$, $\hat{g}\big((z,r), u, x'\big) = 0$, and $J^*(x') = 1$ (i.e., $x' \notin G$), in which case

$$\hat{g}\big((z,r), u, x'\big) + J(x') = v \geq 1.$$

Therefore, if there exists u satisfying (i), then $T(J)(z,r) = 1 = J^*(z,r)$.

Now, if $r = 0$, then only case (i) can happen, since $r_u > 0$ for all u. If $r \in (0, 1)$, then there exists u with $r_u = r$, and this u satisfies (i).

Suppose $r = 1$. Then the assumption $J^*(z,r) = 1$ implies that $z \notin D$ (cf. Eq. (C.1)). By the definition of D, this means that M_z is well-ordered and therefore has a smallest element $\bar{r}$. Then, there exists a rational number $r_u < \bar{r}$, and by the definition of M_z, $z \notin W_{r_u}$. The corresponding index u satisfies (i). Thus, we have shown $T(J) = J^*$.