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 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 case.

∗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** 3

2. **Background** 6
   - 2.1 Preliminaries 6
   - 2.2 Stochastic Control Model 8
   - 2.3 Optimality Properties 11
   - 2.4 Measurability Issues in Standard Policy Iteration 11

3. **A Mixed Value and Policy Iteration Method** 12
   - 3.1 Mappings Induced by Stationary Policies 13
   - 3.2 Algorithms 17

4. **Convergence Analysis for Discounted Case (D) and Nonpositive Case (N)** 22
   - 4.1 Discounted Case (D) 22
   - 4.2 Nonpositive Case (N) 24

5. **Convergence Analysis for Nonnegative Case (P)** 26
   - 5.1 A Convergence Theorem for Value Iteration 26
   - 5.2 Convergence Properties of Mixed Value and Policy Iteration 32

6. **Applications in Semicontinuous Models** 36
   - 6.1 Upper Semicontinuous Models 37
   - 6.2 Lower Semicontinuous Models 39

7. **Concluding Remarks** 40

References 43

Appendices 46

A. **Optimal Stopping Problems Associated with \( F_\theta \)** 46
   - A.1 Formulation 46
   - A.2 Relations with \( F_\theta(\cdot; J) \) and \( Q_{\theta,J} \) 47
   - A.3 A Useful Linear Program for Case (P) 50

B. **Proof of \( Q_{\theta,J}^* = Q^* \) for Nonnegative Case (P)** 52

C. **About the Existence of Borel Measurable Policies** 57

D. **An Illustrative Example for Value Iteration in Case (P)** 58

E. **An Alternative Proof of Theorem 5.1** 59
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 [12, 13, 14, 15] and Strauch [50], which was developed further, along several directions, through a sequence of subsequent works. These include: the books by Hinderer [31], and Dynkin and Yushkevich [22], which considered a framework based on Borel measurable policies and the notion of almost-sure $\epsilon$-optimality; the work of Maitra [34], Furukawa [27], Freedman [26] and Schäl [43], as well as Dynkin and Yushkevich [22], which studied Borel measurable policies and semi-continuous models; the work of Blackwell, Freedman and Orkin [17], which introduced a formulation involving analytic sets and analytically measurable policies; and the work of Shreve and Bertsekas [47, 48], and Bertsekas and Shreve [8, Part II], which considered universally measurable policies. Further research on alternative frameworks suitable for DP include: Shreve [44] and Bertsekas and Shreve [8, Part II] on C-sets and limit-measurable policies, Blackwell [16] on Borel-programmable functions and Shreve [46] on Borel-approachable functions. We refer to the monograph [8] and the papers [17, 45] 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 [11, 29, 30, 2, 25], [36, Part III], and the survey papers [3, 23] 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 [47, 48, 8, Part II], 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 [15]. Moreover, if we restrict attention to Borel measurable policies, there need not exist an everywhere $\epsilon$-optimal policy even in discounted problems [19]. An important step toward a more satisfactory framework was taken by Blackwell, Freedman and Orkin [17]. 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 $\epsilon$-optimal nonrandomized policies among structured families of policies in general. Building upon analytic sets and semi-analytic functions as in [17], a fuller framework was developed in Shreve and Bertsekas [47, 48], Bertsekas and Shreve [8, 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 the Jankov-von Neumann theorem ensure the existence of everywhere $\epsilon$-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 [17, p. 940] and [8, 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 $\epsilon$-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 $\epsilon$-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 and its inherent value iteration character: for convergence, it does not require the policies involved to improve successively on 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 the lower semicontinuous model, we exploit the fact that Borel measurable policies have continuous portions (Lusin’s Theorem; see e.g., [21]), and we use it to specialize our method to produce policy iteration-like algorithms that operate within functions having the desired lower semicontinuity structure. For the upper semicontinuous model, a similar, albeit less general, result is derived: with additional assumptions which may hold only for certain problems, we show how to construct policy iteration-like algorithms that operate within upper 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.

For nonnegative cost models, the convergence analysis of our mixed value and policy iteration method 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 [4] for a related special case of minimax reachability problems, by Schäl [43] and Bertsekas [5] for cases where measurability issues are not a concern, and by Bertsekas and Shreve [8] for the universally measurable policies framework of this paper. Sufficient conditions have also been studied by Whittle [58, 59].

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 [58, 59] 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 [35], which showed that starting from the constant function zero, value iteration could require a transfinite number of iterations to converge. We will give two alternative proofs of the new theorem for the convergence of value iteration (in the standard, non-transfinite form), one of which uses, among others, Maitra and Sudderth’s result.

We also note that Meyn in his book [36, Chap. 9] uses systematically the space of functions bounded in absolute value by a multiple of the optimal cost function (or an optimal relative cost function) to analyze average cost and total cost problems. The other proof we give for the new convergence theorem is motivated by some arguments used there.

Using the new convergence theorem for value iteration, we are also able to show that for cer-
tain 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 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 [11, 61] and for abstract DP problems [10] under discounted and undiscounted total cost criteria (see also the book accounts of these works in [6, 7]). In the finite-spaces or abstract DP context of these papers, measurability is not an issue. Instead, our main motivation has been the development of an asynchronous distributed, possibly model-free, policy iteration-like method that involves a flexible form of policy evaluation. A stopping problem formulation of policy evaluation was first proposed in [11] for that purpose. It has led to computationally efficient policy iteration-like algorithms with guaranteed convergence for the finite-spaces and abstract DP problems considered in [11, 61, 10], even in the context of an asynchronous distributed computing environment. 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.

Although the focus of this paper is not so much on computation as on analytical issues, we note that our results suggest several computational benefits of the proposed method. First, some of its policy iteration-like variants may in practice converge significantly faster than value iteration (which our method contains as a special case). As will be discussed in Section 3.2, these variants resemble modified policy iteration, a well-known method that combines value and policy iteration, which many computational studies and some analyses have found to be more efficient than both value iteration and policy iteration (see [41, Chap. 6]). Our own computational experiments with our method for finite-spaces problems [11] are also consistent with this assessment. Second, compared to standard policy iteration, our method offers better convergence properties, thanks to its inherent value-iteration character. This is evidenced by the convergence results of this paper, especially in the cases of nonpositive and nonnegative models, for which it is well-known that standard policy iteration can have convergence difficulties even when it can be carried out. This advantage was also found earlier for certain total cost finite-spaces problems in our earlier work [61]. Third, like its predecessors in [11, 61, 10], the proposed method is suitable for asynchronous distributed computation (see the discussion in Section 7), although we do not discuss this possibility in the present paper. In a distributed computing environment, based on the analyses in [11, 61, 10] and the convergence results of this paper, we expect some of our policy iteration-like algorithms to be more efficient than asynchronous value iteration and at the same time, to converge more reliably than asynchronous policy iteration. Finally, as another distinctive feature of the proposed method, it can solve problems that cannot be handled by standard policy iteration. These include, e.g., problems in which exact policy improvement is impossible or $\epsilon$-optimal stationary policies do not exist. Potentially one may extract near-optimal policies from the iterates produced by the method. For such problems, however, the overall efficiency of our method relative to value iteration still requires further study.

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. 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 application of the mixed value and policy iteration approach in semicontinuous models, and we also give a result on the structure of the optimal cost function and optimal policies for nonnegative upper semicontinuous models. In Section 7, we conclude the paper with remarks on extensions and future research directions. Appendices A-D collect some related formulations, proofs and illustrative examples; Appendix E gives an alternative proof of the new convergence theorem for value iteration.

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 in standard policy iteration.

2.1 Preliminaries

In this subsection we introduce some concepts and terminologies, including universal $\sigma$-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. Several definitions and notation that we will use throughout the rest of the paper are given at the end of this subsection.

In this paper we will need to consider $\sigma$-algebras finer than the Borel $\sigma$-algebra, one of which is the universal $\sigma$-algebra defined as follows. Any Borel probability measure $\mu \in \mathcal{P}(X)$ can be extended to a probability measure on the $\sigma$-algebra $\mathcal{B}_\mu(X)$ generated by $\mathcal{B}(X)$ and all the subsets of $X$ that have $\mu$-outer measure zero, such that the extension agrees with the $\mu$-outer measure on $\mathcal{B}_\mu(X)$. This extension of $\mu$ is called the completion of $\mu$ [21, Sec. 3.3] and will also be denoted by $\mu$. The intersection $\mathcal{U}(X) = \cap_{\mu \in \mathcal{P}(X)} \mathcal{B}_\mu(X)$ is called the universal $\sigma$-algebra [8, Def. 7.18]. Sets in $\mathcal{U}(X)$ and measurable functions on $(X, \mathcal{U}(X))$ are said to be universally measurable, since they are measurable with respect to the completion of any $\mu \in \mathcal{P}(X)$.

We consider subsets of a Polish space – a topological space that can be metrized by a metric under which it is separable and complete. In this paper, a Borel space refers to a Borel subset of a Polish space [46] endowed with the relative topology and Borel $\sigma$-algebra. Analytic sets in a Polish space can be defined in several equivalent ways (see e.g., [8, Prop. 7.41], [21, Sec. 13.2]), and roughly speaking, they are the images of Borel subsets of some Polish space under continuous or Borel measurable functions. More specifically, in a Polish space $Y$, the empty set is analytic by definition, and a nonempty set $D$ is analytic if $D = f(B)$ for some Borel set $B$ in a Polish space and Borel measurable function $f : B \to Y$ [21, Thm. 13.2.1(c')]. In a Polish space every Borel set is analytic and every analytic set is universally measurable (8, Cor. 7.42.1, [21, Thm. 13.2.6]).

\footnote{The definition of Borel spaces given in [8] is more general than and yet essentially equivalent to the one we give here. Borel spaces, as defined here, are now commonly called standard Borel spaces (see e.g. [19]).}
§2. Background

If $X$ is a Borel space or an analytic set, the analytic $\sigma$-algebra $\mathcal{A}(X)$ is the $\sigma$-algebra generated by the analytic subsets of $X$, and it lies between the Borel $\sigma$-algebra and the universal $\sigma$-algebra; $\mathcal{B}(X) \subseteq \mathcal{A}(X) \subseteq \mathcal{P}(X)$ (the inclusions are strict if $X$ is an uncountable Borel space) [8, p. 171].

Thus on $X$, a Borel measurable function is functionally measurable (i.e., $\mathcal{A}(X)$-measurable) and an 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 ([8, Cor. 7.35.2, Prop. 7.40], [89, Chap. 4]). This gives rise to many nice properties of lower semi-analytic functions, that is, 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 $c \in \mathbb{R}$, the level set $\{ x \in D | f(x) < c \}$ of $f$ is analytic ([8, Def. 7.21]). (Equivalently, the epigraph of $f$, $\{(x, c) | x \in D, f(x) \leq c, c \in \mathbb{R}\}$, is analytic; cf. [8, 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 [8, Lemma 7.30]):

(i) If $f, g : D \to [-\infty, \infty]$ are lower semi-analytic functions, then $f + g$ is lower semi-analytic (here we define $-\infty - \infty = -\infty + \infty = \infty$). In addition, if $f, g \geq 0$ or if $g$ is Borel measurable and $g \geq 0$, then $f g$ is lower semi-analytic (here we define $0 \cdot \infty = 0 : (-\infty) = \infty \cdot 0 = (-\infty) : 0 = 0$).

Note a particular implication of this: for any $B \in \mathcal{B}(D)$, $f \cdot 1_B$ is lower semi-analytic, where $1_B$ denotes the indicator function for $B$: $1_B(x) = 1$ if $x \in B$; $1_B(x) = 0$ if $x \notin B$.

(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 $f_n : D \to [-\infty, \infty]$, $n \geq 1$, the functions $\inf_n f_n, \sup_n f_n, \lim \inf_n f_n$ and $\lim \sup_n f_n$ 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 $D$ is an analytic set in $X \times Y$, where $X, Y$ are Polish, the projection of $D$ on $X$, $\text{proj}_X(D) = \{ x | (x, y) \in D \text{ for some } y \}$, is analytic ([8, 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 after partial minimization of $f$ over the vertical sections $D_x$ of $D$ for each $x$, the resulting function $f^* : \text{proj}_X(D) \to [-\infty, \infty]$ given by

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

is also lower semi-analytic ([8, Prop. 7.47]).

The Jankov-von Neumann 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 $\phi : \text{proj}_X(D) \to Y$ such that the graph of $\phi$ lies in $D$, i.e., $(x, \phi(x)) \in D$ for all $x \in \text{proj}_X(D)$ ([8, 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 $\epsilon$-minimizer and the existence of a universally measurable $\epsilon$-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 [8, 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.

---

\[ ^{2} \text{By an analytic subset of } X \text{ we mean a subset of } X \text{ which, as a subset of the Polish space } Y \text{ that contains } X, \text{ is analytic. The } \sigma\text{-algebras } \mathcal{B}(X), \mathcal{A}(X), \mathcal{P}(X) \text{ here are the trace } \sigma\text{-algebras } X \cap \mathcal{B}(Y), X \cap \mathcal{A}(Y), X \cap \mathcal{P}(Y), \text{ respectively.} \]
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 $\kappa(\cdot|\cdot): \mathcal{B}(Y) \times X \to [0, 1]$ such that:

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

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

Equivalently, the function $x \mapsto \kappa(\cdot|x)$ is $\mathcal{F}$-measurable from $X$ to $\mathcal{B}(Y)$ [8, Prop. 7.26, Lemma 7.28, Prop. 11.6]. If $f : X \times Y \to [0, \infty]$ is lower semi-analytic and $\kappa(dy|x)$ is a Borel measurable stochastic kernel on $Y$ given $X$, then the integral

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

as a function of $x$ is lower semi-analytic on $X$ [8, Prop. 7.48], where for each $x$, the integration is defined to be with respect to the completion of the Borel probability measure $\kappa(dy|x)$. If instead $\kappa(dy|x)$ is analytically or universally measurable, then the above integral as a function of $x$ is universally measurable [8, Prop. 7.46, Sec. 11.2] but 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 [17] and the monograph [8, Chap. 7]. (For general properties of analytic sets, see also the books [40, 49].) Later in this paper, we will also encounter special classes of semicontinuous functions and continuous stochastic kernels when we consider semicontinuous control models, and we will give the definitions of these objects then in Section 6.

Let us introduce now some notation to use throughout the paper. For a Borel space or analytic set $X$, we denote by $\mathcal{M}(X)$ (resp. $A(X)$) the set of functions $f : X \to [-\infty, \infty]$ that are universally measurable (resp. lower semi-analytic), and we denote by $A_b(X)$, $A_+(X)$ and $A_-(X)$ the subsets of functions in $A(X)$ that are bounded, nonnegative and nonpositive, respectively. The set $\mathcal{M}_b(X)$ of bounded universally measurable functions on $X$ is a Banach space with the norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$ for $f \in \mathcal{M}_b(X)$. It is worth noting that the set $A_b(X)$ is a closed subset of $\mathcal{M}_b(X)$ and hence for the metric $d_{\text{sup}}(f, f') = \|f - f'\|_\infty$, the space $(A_b(X), d_{\text{sup}})$ is a complete metric space. Except in one occasion, however, we will consider primarily pointwise convergence of functions on $X$. We write $f_n \to f$ for a sequence of functions $f_n$ converging pointwise to a function $f$, and if the convergence is monotonically from above or from below, we write $f_n \downarrow f$ or $f_n \uparrow f$, respectively. For $x \in X$, we denote by $\delta_x$ the Dirac measure that assigns probability 1 to the point $x$. We use the symbol $\mathbb{R}$ for the set of reals, and the symbol $\mathbf{0}$ for the constant function zero on any space given in the context of discussion.

## 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 $q(dx' | x, u)$ on $S$ given $S \times C$. 


We consider measurable policies defined as follows. A universally measurable policy is a sequence of stochastic kernels, \( \pi = (\mu_0, \mu_1, \ldots) \), where for each \( k \), \( \mu_k(du_k \mid x_0, u_0, \ldots, u_{k-1}, x_k) \) is a universally measurable stochastic kernel on \( C \) given \( (S \times C)^k \times S \) such that

\[
\mu_k(U(x_k) \mid x_0, u_0, \ldots, u_{k-1}, x_k) = 1, \quad \forall (x_0, u_0, \ldots, u_{k-1}, x_k) \in (S \times C)^k \times S. \tag{2.2}
\]

Here the constraint (2.2) says that the set of non-admissible controls, \( C \setminus U(x_k) \), has probability zero, and this is meaningful since \( \Gamma \) is analytic: each vertical section \( U(x) \) of \( \Gamma \) is universally measurable [S Lemma 7.29] and hence measurable with respect to the completion of the Borel probability measure \( \mu_k(\cdot \mid x_0, u_0, \ldots, u_{k-1}, x_k) \). In what follows, when no confusion arises, we will simply refer to universally measurable policies as policies.

A policy \( \pi \) is nonrandomized if \( \mu_k(\cdot \mid x_0, u_0, \ldots, u_{k-1}, x_k) \) is a Dirac measure that assigns probability one to some point in \( U(x_k) \) for every \( k \) and every \( (x_0, u_0, \ldots, u_{k-1}, x_k) \). A policy \( \pi \) is semi-Markov if for every \( k \), the function \( (x_0, u_0, \ldots, u_{k-1}, x_k) \mapsto \mu_k(\cdot \mid x_0, u_0, \ldots, u_{k-1}, x_k) \) depends only on \( (x_0, x_k) \); Markov if for every \( k \), the latter function depends only on \( x_k \); stationary if \( \pi \) is Markov and \( \mu_k = \mu \) for all \( k \). For the stationary case, we simply write \( \mu \) for \( \pi = (\mu, \mu, \ldots) \). A nonrandomized stationary policy \( \mu \) 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 \( \mu \), and we will use both notations \( \mu(x), \mu(du \mid x) \) in the paper, depending on the context.

A policy \( \pi \) is Borel measurable or analytically measurable if each component \( \mu_k \) of \( \pi \) is a Borel measurable or analytically measurable stochastic kernel; such a policy is by definition also universally measurable. Because \( \Gamma \) is analytic, by the Jankov-von Neumann selection theorem [S 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 [S].

We denote by \( \Pi \) the set of universally measurable policies. Given a policy \( \pi \in \Pi \), the collection of stochastic kernels

\[
\mu_0(du_0 \mid x_0), \quad q(dx_1 \mid x_0, u_0), \quad \mu_1(du_1 \mid x_0, u_0), \quad q(dx_2 \mid x_1, u_1), \quad \ldots,
\]

\[
..., \quad \mu_k(du_k \mid x_0, u_0, \ldots, u_{k-1}, x_k), \quad q(dx_{k+1} \mid x_k, u_k), \quad \ldots
\]

uniquely determines, for each initial distribution \( p_0 \) of \( x_0 \), a probability measure \( r(\pi, p_0) \) on the universal \( \sigma \)-algebra on \( (S \times C)^\infty \) with the following property [S Prop. 7.45] with respect to \( r(\pi, p_0) \), the expectation \( \mathbb{E} f \) for any nonnegative, universally measurable function \( f : (S \times C)^{k+1} \to [0, \infty] \) equals the iterated integral

\[
\int_S \int_C \cdots \int_S \int_C f(x_0, u_0, \ldots, x_k, u_k) \mu_k(du_k \mid x_0, u_0, \ldots, x_k) q(dx_{k+1} \mid x_k, u_k) \cdots \mu_0(du_0 \mid x_0) p_0(dx_0).
\]

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 \( f, \mathbb{E} f \) is defined as usual to be \( \mathbb{E} f^+ - \mathbb{E} f^- \), 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.

### Infinite Horizon Total Cost Problems

In this paper we formulate various stochastic control problems as cost minimization problems. We consider primarily three 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.

---

3It is worth noting that the universal \( \sigma \)-algebra on \( (S \times C)^\infty \) is not a product \( \sigma \)-algebra, so the existence of a unique probability measure \( r(\pi, p_0) \) here does not follow immediately from the Ionescu Tulcea theorem.
We mention that for reward maximization (instead of cost minimization), the reverse terminologies are used in the literature [14, 50, 39, 41]: 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

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

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 \( \pi \) and the initial distribution \( \delta_x \) as described earlier. (Although \( g \) is only defined on \( \Gamma \), \( \pi \) 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

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

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 \( J_\pi \) are universally measurable \([8, \text{p. 215}]\).

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

\[
J^*(x) = \inf_{\pi \in \Pi} J_\pi(x), \quad \forall x \in S.
\]

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

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

We mention that certain nonnegative or nonpositive discounted problems where the discount factor at each stage depends on the state transition (see e.g., \([43, 58]\)), can be converted to the (N)(P) models given above\(^4\). Results for (N)(P) including those given in this paper are thus applicable to these problems with transition-dependent discounting as well.

---

\(^4\) Suppose the total cost of \( \pi \) for each initial state \( x_0 = x \) is defined as

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

where \( \beta : S \times C \times S \to [0, 1] \) is the transition-dependent discount factor, and \( \hat{g} : \Gamma \times S \to [-\infty, 0] \) or \([0, \infty]\) is the transition cost function. Assume that \( \beta \) is Borel measurable and \( \hat{g} \) is lower semi-analytic. This problem can be converted to an equivalent problem of type (N) or (P) by introducing an absorbing cost-free state \( \infty \), defining the new state transitional kernel \( \hat{q}(\cdot | \cdot, \cdot) \) by

\[
\hat{q}(B | x, u) = \int_B \beta(x, u, x') \hat{q}(dx' | x, u), \quad B \in \mathcal{B}(S), \quad \hat{q}(\{\infty\} | x, u) = 1 - \hat{q}(S | x, u),
\]

for \((x, u) \in \Gamma\), and letting the one-stage cost be \( g(x, u) = \int_{\mathcal{D}} \hat{g}(x, u, x') \hat{q}(dx' | x, u) \) on \( \Gamma \).
2.3 Optimality Properties

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

\[
T(J)(x) = \inf_{u \in U(x)} \left\{ g(x, u) + \alpha \int_S J(x') q(dx' \mid x, u) \right\}, \quad x \in S. \tag{2.3}
\]

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 properties of lower semi-analytic functions discussed in Section 2.1 and the stochastic control model given in Section 2.2. More specifically, since the state transition kernel \( q(dx' \mid x, u) \) is Borel measurable and the one-stage cost \( g \) is lower semi-analytic, \( g(x, u) + \alpha \int_S J(x') q(dx' \mid x, u) \) as a function of \((x, u)\) is lower semi-analytic on \( \Gamma \) by [8 Prop. 7.48 and Lemma 7.30(4)]. Then, since partial minimization preserves lower semi-analyticity [8 Prop. 7.47], it follows that \( T(J) \) is lower semi-analytic.

In each of the (D)(N)(P) cases, the cost function \( J_\mu \) for a stationary policy \( \mu \) is universally measurable. It satisfies a linear equation, \( J_\mu = T_\mu(J_\mu) \), where \( T_\mu \) is a mapping from \( M(S) \) to \( M(S) \), given by

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

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.

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

(b) For (P), value iteration need not converge to \( J^* \): \( T^k(0) \nrightarrow J^* \), where the pointwise limit \( J_\infty \) of \( \{T^k(0)\} \) satisfies \( J_\infty \leq T(J_\infty) \).

In all three cases, \( \epsilon \)-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 \( \epsilon \)-optimal randomized Markov policy need not exist for (N) (a counterexample was given by van der Wal [54]; see also [41 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 [8 Chap. 9] the optimality properties mentioned above, as well as finer characterizations of the optimal cost function and optimal policies.

2.4 Measurability Issues in Standard Policy Iteration

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

(i) Evaluate the cost function \( J_\mu \) of a given stationary policy \( \mu \).

(ii) Find a stationary policy \( \mu' \) 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 [41]:

(i) For a given stationary policy \( \mu \) and a given function \( J \), compute as an approximation of \( J_\mu \), \( J' = T^m_\mu(J) \) for some positive integer \( m \).

(ii) Find a stationary policy \( \mu' \) 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. [17, p. 940], [8, p. 232]). We explain the reasons below.

As defined in [23], $T$ maps a universally measurable function $J \in \mathcal{M}(S)$ to the function $T(J)$, possibly outside $\mathcal{M}(S)$. For a stationary policy $\mu$, $J_\mu$ is universally measurable, so $T(J_\mu)$ is defined. But since $J_\mu$ need not be lower semi-analytic, even if $T(J_\mu)$ is universally measurable, a stationary, universally measurable policy $\mu'$ such that

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

may not exist ([17, p. 940], [8, p. 232]). 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. [17 Example (48)] gave an example of an analytically measurable function $J$ on $[0,1]$ for which $T(J)$ is not Lebesgue measurable. If $J_\mu$ equals such $J$, then there is certainly no stationary (universally measurable) policy $\mu'$ that can satisfy $T_{\mu'}(J) = T(J)$, because $T_{\mu'}(J)$ is universally measurable, whereas $T(J)$ is not. Moreover, since $T(J)$ is not universally measurable, for some $p \in \mathcal{P}(S)$, $T(J)$ 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 $\mu$ can be undefined for some states $x$ (cf. [17 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 $T_\mu$ for some stationary policy $\mu$, can also run into trouble.

The measurability issues in policy iteration can be assumed away by imposing stronger model assumptions. Most of these assumptions involve continuity and semicontinuity conditions to ensure that policy iteration can operate within the class of Borel measurable policies (see e.g., [29]). For an interesting case where policy iteration can operate with universally measurable policies without measurability difficulty, see the subsequent Example 3.1). Nevertheless, without strong model assumptions, even upper and lower semicontinuous models (see Section 6) have the measurability issues discussed above. Such issues arise because the cost function $J_\mu$ of a Borel measurable policy $\mu$ is Borel measurable but not necessarily (upper or lower) semicontinuous in those models, so for policy improvement, special selection theorems for semicontinuous functions (see Section 6) cannot be applied to ensure the existence of a Borel measurable $\mu'$ that satisfies (2.5). Then the policy improvement step may generate an analytically or universally measurable policy, subjecting the subsequent iterations to the measurability difficulties described above.

3 A Mixed Value and Policy Iteration Method

In this section we introduce formally our mixed value and policy iteration approach and discuss some specific algorithms. For (D)(N)(P), recall that the relation $J^* = T(J^*)$ holds:

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

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

$$Q^*(x,u) = g(x,u) + \alpha \int_S J^*(x') q(dx'| x,u), \quad (x,u) \in \Gamma. \quad (3.1)$$

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 [9, Cor. 9.5.2]:

$$Q^*(x,u) = g(x,u) + \alpha \inf_{\pi \in \Pi} \int_S J_\pi(x') q(dx'| x,u), \quad (x,u) \in \Gamma. \quad (3.2)$$
(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., [9 [51].] To simplify notation, for any function \( Q \) on \( \Gamma \), let
\[
M(Q)(x) = \inf_{u \in U(x)} Q(x, u), \quad x \in S.
\]
The mapping \( M \) maps \( A(\Gamma) \) into \( A(S) \) [8 Prop. 7.47]. With this notation, we can write the optimality equation in two equivalent ways:
\[
J^* = T(J^*) \iff J^* = M(Q^*).
\] (3.3)

Our mixed value and policy iteration method operates on the product space \( A(S) \times A(\Gamma) \) and aims to compute \((J^*, Q^*)\). 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 [13]. Our method gives rise to various policy iteration-like algorithms, whose convergence we will analyze in Sections 4 and 5.

In what follows we first introduce a family of mappings underlying the method and discuss their basic properties and their relation to optimal stopping problems (Section 3.1). We then give various forms of algorithms (Section 3.2), and we emphasize particularly policy iteration-like algorithms and contrast them with standard policy iteration.

### 3.1 Mappings Induced by Stationary Policies

We define a family of parametrized mappings \( F_\theta \), which will be used later in a step of the mixed value and policy iteration algorithms that is analogous to policy evaluation. The parameters \( \theta \) include a policy component and a set component paired with that policy, introduced for overcoming measurability issues. Specifically, let \( \Theta \) denote the set of all pairs \((\mu, B)\), where \( \mu \) is a stationary policy and \( B \) a Borel subset of \( S \), such that the function \( x \mapsto \mu(du \mid x) \) restricted to \( B \) is Borel measurable (equivalently, \( \mu(D \mid \cdot) \) is Borel measurable on \( B \) for every \( D \in \mathcal{B}(C) \)). For each \( \theta = (\mu, B) \in \Theta \), we define \( F_\theta : \mathcal{M}(\Gamma) \times \mathcal{M}(S) \to \mathcal{M}(\Gamma) \) by
\[
F_\theta(Q; J)(x, u) = g(x, u) + \alpha \int_{S \setminus B} J(x') q(dx' \mid x, u)
+ \alpha \int_B \int_C \min \{J(x'), Q(x', u')\} \mu(du' \mid x') q(dx' \mid x, u), \quad (x, u) \in \Gamma,
\] (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 \( \Gamma \), the inner integral in the third term in (3.4) is well-defined because \( \mu \) 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 \( \Gamma \).

For any stationary policy \( \mu \), the trivial choice \( B = \emptyset \) gives \( \theta = (\mu, \emptyset) \in \Theta \), but the corresponding mapping \( F_{\theta} \) does not depend on the policy \( \mu \) at all. To introduce greater dependence of \( F_\theta \) on \( \mu \), 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 Examples 3.1, 3.2 in Section 3.2). If the policy \( \mu \) is Borel measurable, then \((\mu, S) \in \Theta \), so one may let \( B \) be the entire space.

An important property of \( F_\theta \) is that it preserves the lower semi-analyticity of functions. This will allow us to overcome the measurability difficulties that hamper standard policy iteration.
§3. Mixed Value and 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 \( \mu \), given any \( p \in \mathcal{P}(S) \), there is a Borel set \( B \subset S \) with \( p(S \setminus B) = 0 \) and \( (\mu, B) \in \Theta \).

Proof. (a) Let \( Q \in A(\Gamma) \). To show \( F_\theta(Q; J) \in A(\Gamma) \), we show that each term in its definition \((3.4)\) is a lower semi-analytic function on \( \Gamma \). The first term \( g(x, u) \) is lower semi-analytic by definition. The second term equals \( \alpha \int_B \mathbf{1}_{S \setminus B}(x')J(x')q(dx'| x, u) \). Here the function \( \mathbf{1}_{S \setminus B} \) is lower semi-analytic given that \( S \setminus B \) is a Borel set and \( J \) is lower semi-analytic \([8]\text{ Lemma 7.30(4)}\), and the stochastic kernel \( q(dx'| x, u) \) is by definition Borel measurable. So by \([8]\text{ Prop. 7.48}\) this integral as a function of \( (x, u) \) is lower semi-analytic on \( S \times C \) and hence lower semi-analytic on the analytic set \( \Gamma \). For the third term, \( \alpha \int_B \min \{ J(x'), Q(x', u') \} \mu(du'| x')q(dx'| x, u) \), since \( \mu \) satisfies the control constraint, we can write it equivalently as

\[
\alpha \int_S \int_C f(x', u') \mu(du'| x')q(dx'| x, u),
\]

where \( f : S \times C \to [-\infty, \infty] \) is given by \( f(x', u') = 1_B(x') \cdot \min \{ J(x'), Q(x', u') \} \) for \( (x', u') \in S \times C \), with \( Q^e \) being a lower semi-analytic extension of \( Q \) to \( S \times C \) defined as \( Q^e(x, u) = \infty \) for \( (x, u) \notin \Gamma \). The function \( f \) is lower semi-analytic, since \( J \) and \( Q^e \) are lower semi-analytic functions and \( B \) is a Borel set \([8] \text{ Lemma 7.30(2),(4)}\); thus \( f \) is also lower semi-analytic on \( B \times C \). The fact \( (\mu, B) \in \Theta \) implies that \( \mu(du'| x') \) is a Borel measurable stochastic kernel \( C \) given \( B \). Consequently, the inner integral \( \int_C f(x', u') \mu(du'| x') \) is lower semi-analytic on \( B \) by \([8] \text{ Prop. 7.48}\). We also have \( \int_C f(x', u') \mu(du'| x') = 0 \) for \( x' \notin B \). Therefore, \( \int_B f(x', u') \mu(du'| x') \) is lower semi-analytic on \( S \). Then, since \( q(dx'| x, u) \) is a Borel measurable stochastic kernel \( S \) given \( S \times C \), the integral \((3.5)\) as a function of \( (x, u) \) is lower semi-analytic on \( S \times C \) by \([8] \text{ Prop. 4.78}\) and hence lower semi-analytic on the analytic set \( \Gamma \). This proves part (a).

(b) Since \( \mu(du| x) \) is a universally measurable stochastic kernel \( C \) given \( S \), by \([8] \text{ Lemma 7.28}\), there is a Borel measurable stochastic kernel \( \tilde{\mu}(du| x) \) with \( \tilde{\mu}(du| x) = \mu(du| x) \) for \( p \)-almost every \( x \). Part (b) then follows. \( \square \)

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 \([11][12][13]\) mappings of the form \( F_\theta, \theta = (\mu, S) \), without splitting the state space into two parts, \( B \) and \( S \setminus B \), according to the policy \( \mu \). In the present context, however, in order for \( F_\theta \) to map lower semi-analytic functions to lower semi-analytic functions, it is important to introduce \( B \) as a parameter component in defining \( F_\theta \).

Some Basic Properties of \( F_\theta \)

We now discuss a few basic properties of the mappings \( F_\theta \) 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.,

\[
F_\theta^n(Q; J) = F_\theta \left( \cdots F_\theta \left( F_\theta(Q; J); J \right) \cdots ; J \right).
\]
From its definition $F_\theta$ is monotone: $F_\theta(Q; J) \geq F_\theta(Q'; J')$ if $J \geq J'$ and $Q \geq Q'$. Hence
\[ J \geq J', \ Q \geq Q' \quad \Rightarrow \quad F_\theta^n(Q; J) \geq F_\theta^n(Q'; J'), \quad \forall n \geq 1. \tag{3.6} \]

We consider the pointwise limit
\[ Q_{\theta, J} = \lim_{n \to \infty} F_\theta^n(0; J), \]
the existence of which will be shown as part of the following proposition.

**Proposition 3.2.** (D)(N)(P) Let $J \in A_0(S)$ for (D), $J \in A_-(S)$ for (N), and $J \in A_+(S)$ for (P).

Then $Q_{\theta, J} = \lim_{n \to \infty} F_\theta^n(0; J)$ is well-defined, is lower semi-analytic, and satisfies
\[ Q_{\theta, J} = F_\theta(Q_{\theta, J}; J). \tag{3.7} \]

For (D), it is the only solution of $Q = F(Q; J)$ in $A_0(\Gamma)$.

**Proof.** In case (D), $F_\theta(\cdot; J)$ is a contraction mapping on the complete metric space $A_0(\Gamma)$: for any $Q, Q' \in A_0(\Gamma)$, we have
\[ \| \min \{ J(x), Q(x, u) \} - \min \{ J(x), Q'(x, u) \} \| \leq \| Q - Q' \|_\infty, \quad \forall (x, u) \in \Gamma, \]
and hence by direct calculation, $\| F_\theta(Q; J) - F_\theta(Q'; J) \|_\infty \leq \alpha \| Q - Q' \|_\infty$. It then follows from Banach's contraction principle [12, p. 220] that $F_\theta^n(0; J)$ converges to $Q_{\theta, J}$, the unique solution of $Q = F(Q; J)$, $Q \in A_0(\Gamma)$.

For case (N) (resp. (P)), $Q_{\theta, J}$ is the pointwise limit of a sequence of nonincreasing nonpositive functions (resp. nondecreasing nonnegative functions). Equation (3.7) then follows by the definition of $F_\theta(\cdot; J)$ and the monotone convergence theorem. \hfill \Box

We can relate $F_\theta$ 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 $J \in A_0(S)$ for (D), $J \in A_-(S)$ for (N), and $J \in A_+(S)$ for (P). Then $Q_{\theta, J} \geq Q_{\theta, J^*}$.

(d) $Q_{\theta, J^*} = Q^*$.

**Proof.** Let $\theta = (\mu, B)$. Since $J^*(x) = \inf_{v \in U(x)} Q^*(x, v) \leq Q^*(x, u)$ for all $(x, u) \in \Gamma$, we can rewrite the iterated integral in the sum $3.4$ defining $F_\theta(Q^*; J^*)(x, u)$ as
\[ \int_B \int_C \min \{ J^*(x'), Q^*(x', u') \} \mu(du' | x') q(dx' | x, u) = \int_B J^*(x') q(dx' | x, u). \]

Combining with the second term in $3.4$, we obtain
\[ F_\theta(Q^*; J^*)(x, u) = g(x, u) + \alpha \int_S J^*(x') q(dx' | x, u) = Q^*(x, u), \quad \forall (x, u) \in \Gamma. \]

This proves part (a). Part (b) then follows from part (a) and the monotonicity of $F_\theta$ (cf. Eq. 3.6).

For part (c), since $J \geq J^*$, we have $F_\theta^n(0; J) \geq F_\theta^n(0; J^*)$ for every $n$, by the monotonicity of $F_\theta$ (cf. Eq. 3.6). Then by Prop. 3.2 $Q_{\theta, J} = \lim_{n \to \infty} F_\theta^n(0; J) \geq \lim_{n \to \infty} F_\theta^n(0; J^*) = Q_{\theta, J^*}$.

There remains to show $Q_{\theta, J^*} = Q^*$ in part (d). For case (D), this is true because by part (a), $Q^*$ is the solution of $F_\theta(Q; J^*) = Q, Q \in A_0(\Gamma)$, whereas this equation has $Q_{\theta, J^*}$ as its unique solution by Prop. 3.2 (D). For case (N), we have $J^* \leq 0$ and consequently, $F_\theta(0; J^*) = Q^*$ by the definitions of $F_\theta$ and $Q^*$. In view of part (a), this implies $F_\theta^n(0; J^*) = Q^*$ 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)). \hfill \Box
§3. Mixed Value and Policy Iteration

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

Relation to Optimal Stopping Problems

In the mixed value and policy iteration algorithms, we will use mappings $F_\theta(\cdot;J)$, $\theta \in \Theta$, $J \in A(S)$, and their fixed points $Q_{\theta,J}$ in a step that is analogous to policy evaluation. Let us first discuss the relations of these objects with certain optimal stopping problems and with the original control problem, to provide more insights and intuition.

Remark 3.1 (Optimal stopping problem corresponding to $F_\theta(\cdot;J)$). The form of $F_\theta(\cdot;J)$ (cf. Eq. (3.4)) suggests a connection to an optimal stopping problem defined by $\theta = (\mu, B)$, $J$ 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 (Lemma A.1). Here we give an intuitive description (see Figure 1 for an illustration). 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 $\Gamma$ 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 $q(dx' | x,u)$. 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 according to $\mu(du' | x')$ and we land at $(x',u')$. There, we can either stop and pay $J(x')$, or continue with the continuation cost $g(x',u')$ and then repeat the process just described for time 1. (This corresponds to the third term in Eq. (3.4), and the minimization inside that term reflects the two choices, to stop or to continue.) 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 $F_\theta$ (see Appendices A and B).

Remark 3.2 (Interpreting $Q_{\theta,J}$ in the original control problem). The function $Q_{\theta,J}$ is the optimal cost function of the optimal stopping problem associated with $J$ and $\theta = (\mu, B)$ mentioned in the preceding remark (see Cor. A.1 Appendix A). We discuss a special case of the stopping costs $J$. It allows us to relate $Q_{\theta,J}$ to the total costs in the original problem.

Suppose $J = J_\pi$ for some $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$ in the original problem. Then for each $(x,u) \in \Gamma$, we may interpret $Q_{\theta,J_\pi}(x,u)$ as the minimal cost over the subset of policies for the original problem which switch from $\mu$ to $\pi$ at most once according to the following restrictions. At time 0 we must apply control $u$ at state $x$. From time 1 on, we can either follow the stationary policy $\mu$ or use the policy $\pi$, which we must do if the state goes outside the set $B$. Once we start to use $\pi$ at time $\tau$, say, we must apply the randomized control rules $\pi_k(\cdot | x_\tau, u_\tau, x_{\tau+1}, u_{\tau+1}, \ldots, x_{\tau+k})$ for time $\tau+k, k \geq 0$, and continue in this way forever. In other words, if we can only apply policy $\mu$ or $\pi$ in the original
problem in the manner just described, then \( Q_{\theta,J_k}(x,u) \) equals the minimal cost of switching from following \( \mu \) to following \( \pi \). Furthermore, an optimal or near-optimal time to switch to \( \pi \) can be found from an optimal or near-optimal solution of the associated optimal stopping problem, if we interpret the action to stop as the decision to switch policies in the original problem.

We do not include a formal proof for the above interpretation of \( Q_{\theta,J_k}(x,u) \) in the paper; but we note that it is similar to the analysis we give in Appendix 3 and that the fact \( Q_{\theta,J_\ast} = Q^\ast \) (Prop. 3.3(d)) is also a manifestation of this interpretation.

### 3.2 Algorithms

We give first our mixed value and policy iteration algorithm in its basic form. The conditions needed for the convergence of this form 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 \( A_0(S) \times A_0(\Gamma) \), or \( A_-(S) \times A_+(\Gamma) \), or \( A_+(S) \times A_+(\Gamma) \), respectively. It then uses a sequence \( \theta_k = (\mu_k, B_k) \in \Theta \) to calculate iteratively functions \( (J_k, Q_k) \in A(S) \times A(\Gamma), k \geq 1 \), as follows, with each iteration consisting of two steps, analogous to the policy evaluation and improvement steps in standard policy iteration. The sequence \( \theta_k \) can be arbitrary here, but specific choices will be given in subsequent algorithms.

**Algorithm I (\( k \)th iteration; basic form):**

- Choose \( \theta_k = (\mu_k, B_k) \in \Theta \), let
  \[
  Q_{k+1} = F_{\theta_k}^{n_k}(Q_k; J_k) \quad \text{for some} \ n_k \geq 1, \quad \text{or} \quad Q_{k+1} = Q_{\theta_k,J_k},
  \]
  and let
  \[
  J_{k+1} = M(Q_{k+1}).
  \]

The fact that for all \( k \), \( (J_k, Q_k) \in A(S) \times A(\Gamma) \) can be seen from the inductive argument: by Props. 3.1 and 3.2, the function \( Q_{k+1} \) is lower semi-analytic if \( J_k, Q_k \) are lower semi-analytic, whereas the minimization \( (3.9) \) results in a lower semi-analytic function \( J_{k+1} \) by \( [\text{5}] \text{ Prop. 7.47} \).

The algorithm \( (3.8)-(3.9) \) 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 \( \mu_k \) 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. the subsequent Examples 3.1-3.2). In what follows, we consider choices of \( \mu_k \) based on \( Q_k \) and a selection theorem of the Jankov-von Neumann type, and we derive policy iteration-like algorithms.

Let \( Q \in A(\Gamma) \) and \( \epsilon > 0 \). By a selection theorem for lower semi-analytic functions \( [\text{5}] \text{ Prop. 7.50(b)} \), we can select a universally measurable, nonrandomized stationary policy \( \mu \) such that, with
\[
I = \{ x \in S \mid \min_{u \in U(x)} Q(x,u) \neq \emptyset \},
\]

\[
\mu(x) \in \arg \min_{u \in U(x)} Q(x,u) \quad \text{if} \ x \in I,
\]

\[
Q(x,\mu(x)) \leq \begin{cases} M(Q)(x) + \epsilon & \text{if} \ x \not\in I, \ M(Q)(x) > -\infty, \\ -1/\epsilon & \text{if} \ x \not\in I, \ M(Q)(x) = -\infty. \end{cases}
\]

If we relax the condition \( (3.10) \), then by \( [\text{5}] \text{ Prop. 7.50(a)} \), we can find instead an analytically measurable policy \( \mu \) such that for all states \( x \),
\[
Q(x,\mu(x)) \leq \begin{cases} M(Q)(x) + \epsilon & \text{if} \ M(Q)(x) > -\infty, \\ -1/\epsilon & \text{if} \ M(Q)(x) = -\infty. \end{cases}
\]
By choosing the policies in the basic algorithm by “approximate policy improvement” based on the above selection theorem and for a desired value of $\epsilon$, we can obtain special forms of the basic algorithm that resemble to some degree standard forms of modified policy iteration. Here is an example. The algorithm starts with a pair of functions $(J_0, Q_0)$ and a stationary policy $\mu_0$ as in Algorithm I. In the process of computing $J_{k+1}$ by partial minimization of $Q_{k+1}$, it also obtains the policy $\mu_{k+1}$ for the next iteration using the selection theorem above.

**Policy Iteration-Like Algorithm II ($k$th iteration):**

- Choose a Borel set $B_k \subset S$ so that $\theta_k = (\mu_k, B_k) \in \Theta$, and compute $Q_{k+1}$ by Eq. (3.8).
- Let $J_{k+1} = M(Q_{k+1})$ as in Eq. (3.9), and let $\mu_{k+1}$ be a nonrandomized stationary policy satisfying Eqs. (3.10)-(3.11), or Eq. (3.12), with $Q = Q_{k+1}$ and a desired value of $\epsilon$.

If there exists at least one Borel measurable policy, we can further specialize the basic algorithm to use Borel measurable $\mu_k$ 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 $\Gamma$ is Borel, a nonrandomized Borel measurable policy is known to exist under fairly general conditions (see Appendix C). Thus algorithms of this kind can be applied to a large class of problems.

**Policy Iteration-Like Algorithm III with Borel Measurable Policies:**

Let $\mu_0$ be a Borel measurable stationary policy (assumed to exist).

Iterate for each $k \geq 0$:

- For $\theta_k = (\mu_k, S)$, compute $Q_{k+1}$ by Eq. (3.8).
- Let $J_{k+1} = M(Q_{k+1})$ as in Eq. (3.9), and let $\mu'_{k+1}$ be a stationary policy satisfying Eqs. (3.10)-(3.11), or Eq. (3.12), with $Q = Q_{k+1}$ and a desired value of $\epsilon$.
- Select $p_{k+1} \in \mathcal{P}(S)$ 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 $\mu_{k+1}$ by

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

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.)

**Comparison with Classical Value and Policy Iteration**

The convergence of the preceding algorithms will be the subject of the next two sections. Here, for a better understanding of the theoretical and computational properties of these algorithms, let us compare them to the classical value iteration and policy iteration. First, let us try to relate informally the structure of the preceding Algorithms I-III to the classical algorithms, leaving aside for the moment the measurability and other mathematical issues that may prevent policy iteration from being carried out.

There are two basic choices at each iteration of the form (3.8)-(3.9): the pair $(\mu_k, B_k)$ and the integer $n_k$. Both of these choices determine how closely our algorithms resemble value iteration and forms of policy iteration. In particular, as noted earlier, the choice $B_k = \emptyset$ makes Algorithms I-II identical to value iteration, regardless of the choice of $n_k$ and $\mu_k$. It can also be seen that if we let $n_k = 1$ so that $Q_{k+1} = F_{\theta_k}(Q_k \mid J_k)$, and if in addition $\mu_{k+1}$ satisfies Eq. (3.10) with $Q = Q_{k+1}$ and $I = S$, Algorithms I-III also become identical to value iteration.

To show the similarities and differences between Algorithms I-III and policy iteration, let us rewrite the latter in terms of mappings $F_\theta$ and functions on $\Gamma$. Denote the constant function $+\infty$
simply by $+\infty$. The modified or exact policy iteration algorithm (cf. Section 2.4) can be equivalently stated as follows. Iterate for each $k \geq 0$ and a given nonrandomized stationary policy $\mu_k$:

(i) (Evaluation step) Let $Q_{k+1} = F^{n_k} \mu_k(S, Q_k : +\infty)$ for some $n_k \geq 1$ (in the case of modified policy iteration) or let $Q_{k+1} = Q \mu_k(S, \cdot, +\infty)$ (in the case of exact policy iteration).

(ii) (Improvement step) Let $\mu_{k+1}$ be such that $\mu_{k+1}(x) \in \arg \min_{u \in U(x)} Q_{k+1}(x, u)$ for all $x \in S$.

We now compare this description of standard policy iteration to Algorithms I-III, and make a few observations regarding its similarities, and its theoretical and computational differences from our methods.

**Remark 3.3 (Relation to standard policy iteration).** From the above description, we see that as $B_k$ is made larger and approaches $S$, and $\mu_k$ is chosen by “approximate policy improvement” based on selection of the form (3.10)-(3.11), our algorithms approach the exact form of policy iteration (if $Q_{k+1} = Q_{\theta_k} \mu_k, J_k$) or modified policy iteration (if $Q_{k+1} = F^{n_k} \mu_k(S, Q_k : J_k)$). Indeed, suppose for all $k$, $\mu_k$ can be chosen by exact policy improvement (i.e., $J = S$ in Eq. (3.10)) and $B_k$ can be chosen to be $S$. Then our algorithms become *almost identical* to standard policy iteration, except for a difference in the evaluation step where the “stopping costs” are set to be $J_k$ instead of $+\infty$.

Thus, if for all $k$, $J_k$ lies above the functions against which it is compared in our algorithms so that it acts like $+\infty$, our algorithms become identical to the standard policy iteration algorithms, under the preceding assumptions on $(\mu_k, B_k)$. As an example, under (D), the conditions on $J_k$ just mentioned are satisfied if $\mu_0$ is nonrandomized and the initial condition $J_0 \geq J_{\mu_0}$, $J_0 \geq V_0 \geq T_{\mu_0}(V_0)$, where $V_0 = (0, \mu_0(\cdot))$, holds; then, implemented with $Q_{k+1} = Q_{\theta_k} \mu_k$ or $Q_{k+1} = F^{n_k} \mu_k (S, Q_k : J_k)$, our method becomes identical to the classical policy iteration or modified policy iteration, respectively.

Being equivalent to standard policy iteration is, however, not always the goal of our algorithms. Indeed, in some cases of (N)(P), they can never be the same because of the $\{J_k\}$ sequence used by our algorithms. For example, under (N) exact policy iteration may cycle even for discrete-spaces problems [11] Example 7.2.4, p. 296], and for convergence under (P), exact or modified policy iteration need conditions on the initial policy which may be unsatisfiable in some problems (cf. Cor. 5.2 Prop. 5.1 in Section 5.1). By contrast, our algorithms converge under general initial conditions that guarantee the convergence of value iteration (see Sections 5.2 and 5.3). This more reliable convergence behavior can be attributed to the use of the “stopping costs” sequence $\{J_k\}$ in the evaluation step, which brings a value-iteration character to our algorithms. (Such advantages of our method have also been shown in an asynchronous distributed setting for discounted and total cost finite-spaces problems and abstract DP problems; see [11] [61] [10].)

Finally, since with nonrandomized policies $\mu_k$, our algorithms have similar forms as standard policy iteration, they have comparable per-iteration computation overhead. In particular, compared to exact policy iteration, computing $Q_{\theta_k} \mu_k$ involves solving an optimal stopping problem that has at most two controls at each state. Compared to modified policy iteration, computing $F^{n_k}_{\theta_k} (Q_k : J_k)$ involves only an extra comparison of the functions $J_k$ and $Q_k$ before computing the expected value in the mapping $F_{\theta_k}$.

---

4Here we look at the functions on the space $\Gamma$ produced by standard policy iteration just before the minimization in the policy improvement step takes place, and we write the policy evaluation and improvement steps equivalently in terms of these functions. Specifically, the modified policy iteration calculates the next cost function iterate $V'$ and policy $\mu'$ with $V' = T_{\mu'}(V)$, $T_{\mu'}(V') = T(V')$ for some integer $n \geq 1$. The improvement step, finding $\mu'$ with $T_{\mu'}(V') = T(V')$ (assuming such $\mu'$ exists), is to solve the partial minimization problem $\mu'(x) \in \arg \min_{u \in U(x)} Q'(x, u), x \in S$, for the function $Q'$ obtained by policy evaluation: $Q'(x, u) = \int_G F_{\theta_k} \mu(S, V')(\gamma dx | x, u)$, $(x, u) \in \Gamma$. Expressed in terms of $F_{\theta_k} \mu_k$, we have $Q'(x, u) = \int S T_{\mu_k}(V')(x') q(dx' | x, u), (x, u) \in \Gamma$. From the definition of $\mu'$, we put $\mu'(x) = T_{\mu'}(V')$, $x \in S$. Putting these relations for each iteration of the algorithm together, we obtain the alternative description of modified policy iteration given in the discussion, where the integer $n_k$ corresponds to the integer $n$ here for iteration $k \geq 1$ and to $n + 1$ for iteration $k = 0$. The equivalent form for the exact policy iteration is derived similarly.
As noted in the preceding discussion, when exact policy improvement can take place, the choice of \((\mu_k, B_k)\) and the use of \(\{J_k\}\) in the evaluation step may cause differences between our algorithms and standard policy iteration. The use of \(\{B_k\}\) lets our method overcome measurability issues, whereas the use of \(\{J_k\}\) gives our method better convergence properties. We now discuss further the high-level differences between our methods and standard policy iteration.

**Remark 3.4** (Differences from standard policy iteration). We first note a generic difference that has to do with approximate policy improvement and can be related to a notion of almost-sure \(\epsilon\)-optimality. For simplicity, we consider Algorithm III and assume that it generates nonrandomized policies \(\mu_k\) and sets \(Q_{k+1} = Q_{\theta_k, J_k}\) for all \(k\), where \(\theta_k = (\mu_k, S)\). The mappings \(F_{\theta_k}\) have the form

\[
F_{\theta_k}(Q; J)(x, u) = g(x, u) + \alpha \int_S \min \{J(x'), Q(x', \mu(x'))\} q(dx' | x, u), \quad (x, u) \in \Gamma, \tag{3.14}
\]

for the nonrandomized Borel measurable policies \(\mu = \mu_k\). Denote

\[
V_k(x) = \min \{J_k(x), Q_{\theta_k, J_k}(x, \mu_k(x))\}, \quad x \in S.
\]

From Eq. (3.14) and the relation \(Q_{\theta_k, J_k} = F_{\theta_k, S}(Q_{\theta_k, J_k}; J_k)\) (Prop. 3.2), we see that for all \(x \in S\),

\[
J_{k+1}(x) = M(Q_{\theta_k, J_k})(x) = \inf_{u \in U(x)} \left\{ g(x, u) + \alpha \int_S V_k(x') q(dx' | x, u) \right\} = T(V_k)(x),
\]

\[
Q_{\theta_k, J_k}(x, \mu_{k+1}(x)) = T_{\mu_{k+1}}(V_k)(x).
\]

Assume \(T(V_k)(x) > -\infty\) for all \(x\), for simplicity. Since \(\mu_{k+1}\) is chosen based on either Eqs. (3.10)-(3.11) or Eq. (3.12) (cf. the definition (3.13) of \(\mu_{k+1}\)), it follows that for \(k \geq 0\),

\[
p_{k+1}\left(\{x \in S \mid T_{\mu_{k+1}}(V_k)(x) \leq T(V_k)(x) + \epsilon\}\right) = 1, \tag{3.15}
\]

where \(p_{k+1}\) is the probability measure in Algorithm III. In other words, the generated policy \(\mu_{k+1}\) is \(\epsilon\)-optimal for the optimization problem \(T(V_k)(x)\) except on a \(p\)-null set of states with \(p = p_{k+1}\) — a property that bears similarity to the notion of \("(p, \epsilon)\)-optimal\) policies [13] [50].

Equation (3.15) shows that Algorithm III includes as a special case a generalized version of policy iteration, which involves \((p, \epsilon)\)-optimal policy improvement as defined by Eq. (3.15), and is able to deal with situations where policy improvement cannot be exact either because in Eq. (3.10) \(I \neq S\) or because of Borel measurability issues. By contrast, standard policy iteration does not allow such generalized forms of policy improvement.

Algorithm III also demonstrates important theoretical differences between our policy iteration-like algorithms and standard policy iteration, despite the many similarities the two share in their forms. Like Algorithms I-II, Algorithm III yields \(J^*\) in the limit in cases (D)(N)(P) under certain initial conditions (see Sections 4 and 5), although it operates with Borel measurable policies only. By comparison, 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 [13 Example 2] or [50 Example 4.1]. It is also impossible to obtain \(J^*\) by exact policy iteration if for some state, there exists no stationary, \(\epsilon\)-optimal policy (even if policy improvement can be carried out exactly). This can happen even in countable-state and control problems; see e.g., [50 Example 6.1]. Thus our algorithms overcome some fundamental theoretical limitations of standard policy iteration.

For many discrete-spaces problems, policy iteration has been preferred to value iteration because of its computational efficiency [11]. Let us discuss now the potential computational benefits of using our policy iteration-like algorithms in place of value iteration.
Remark 3.5 (Computational efficiency and the role of the parameter \(n_k\)). We noted earlier that certain cases of \(n_k = 1\) make our algorithms identical to value iteration. The choice of \(n_k\) in our algorithms plays a similar role as in modified policy iteration, and determines how frequently the high overhead minimization in the policy improvement step \((3.9)\) is performed. In practice, values of \(n_k\) that are larger than 1 are recommended in order to keep this overhead low, and experience has shown that modified policy iteration with well-designed choice of \(n_k\) is computationally more efficient than both value iteration and exact policy iteration in finite spaces (see e.g., [11 Chap. 6.5]). An intuition behind this is that when the generated policies \(\mu_k\) are close to an optimal policy, each relatively cheap iteration \(T_\mu(V)\) in the evaluation step can act like the far more expensive value iteration \(T(V)\). Based on their similarities to modified policy iteration, we expect our policy iteration-like algorithms, with a judicious choice of \(n_k\) and other parameters, to have similar computational advantages over value iteration. This is also consistent with our observations from computational experimentation with discounted finite-spaces problems [11]. For undiscounted problems under (P) or (N) and infinite-spaces problems in general, our algorithms are new and still need to be tested computationally, however.

Choosing the Set \(B\)

In Algorithms I-III, we repeatedly find, for a universally measurable policy \(\mu\), a Borel set \(B \subset S\) such that \(\mu(du)\) 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 \(p\) in a way coupled to the policy \(\mu\). Let us discuss two natural possibilities. The first example describes a fortunate situation where \(\theta (\cdot, B, \mu)\) with state transition kernel \(q(dx' | x, u)\) has a density with respect to some \(p \in \mathcal{P}(S)\), i.e., it can be expressed as

\[
q(dx' | x, u) = f(x, u, x') p(dx'), \quad \forall (x, u) \in S \times C,
\]

for a nonnegative Borel measurable function \(f(x, u, x')\). Let us choose the set \(B\) based on \(p\), with \(\theta = (\mu, B) \in \Theta\) and \(p(B) = 1\). Since \(p(S \setminus B) = 0, q(S \setminus B | x, u) = 0\) and hence the second term in the definition \((3.4)\) of \(F_\theta(Q; J)(x, u)\) vanishes, giving a simplified expression of \(F_\theta\) as

\[
F_\theta(Q; J)(x, u) = g(x, u) + \alpha \int_B \int_C \min \{ J(x'), Q(x', u') \} \mu(du') q(dx' | x, u), \quad (x, u) \in \Gamma,
\]

where the integration over \(B\) can also be replaced by the integration over \(S\).

It is worth mentioning that in this example, standard policy iteration has no measurability difficulty. Indeed, for any stationary policy \(\mu\), one can find a Borel set \(B \subset S\) with \(p(B) = 1\) such that restricted to \(B\), the universally measurable cost function \(J_\mu\) is Borel measurable [8] Lemma 7.27]. This implies that the function \(g(x, u) + \alpha \int_B J_\mu(x') q(dx' | x, u) = g(x, u) + \alpha \int_B J_\mu(x') q(dx' | x, u)\) is lower semi-analytic, and hence at a policy improvement step a stationary policy satisfying Eq. (2.5) can be found.

Example 3.2 (Choice of \(B\) based on the Markov chain induced by \(\mu\)). Consider a Markov chain on \((S, \mathcal{F}(S))\) with state transition kernel \(\kappa(dx' | x)\) defined by

\[
\kappa(D | x) = \int_C q(D | x, u) \mu(du | x), \quad D \in \mathcal{F}(S),
\]
where \( q(D \mid x, u) \) is the measure of \( D \) with respect to the completion of \( q(dx' \mid x, u) \)\(^6\). The \( n \)-step transition kernels are: \( \kappa^n(dx' \mid x) = \delta_0(dx'), \kappa^n(dx' \mid x) = \int_S \kappa^{n-1}(dx' \mid y) \kappa(dy \mid x), \) \( n \geq 1 \). Given a probability measure \( \rho \) on \((S, \mathcal{W}(S))\) and \( \beta \in (0, 1) \), define a probability measure \( p \) on \((S, \mathcal{W}(S))\) by
\[
p(D) = (1 - \beta) \sum_{n=0}^{\infty} \beta^n \int_S \kappa^n(D \mid x) \rho(dx), \quad D \in \mathcal{W}(S).
\]
The measure \( p \) reflects which sets of states are visited with positive probability under the policy \( \mu \) if the initial distribution is \( \rho \). This choice of \( p \) was suggested by S. Shreve.

Let us choose \( B \) such that \( \theta = (\mu, B) \in \Theta \) and \( p(B) = 1 \). Using the definition of \( p \), it can be shown that \( B \) contains a nonempty absorbing set of states, \( \bar{D} = \{ x \in B \mid \kappa^n(S \setminus B \mid x) = 0, n \geq 1 \} \); i.e., \( \kappa(\bar{D} \mid x) = 1 \) for \( x \in \bar{D} \) (the proof is similar to the proof of [37, Prop. 4.2.3(ii)]). It can also be shown that \( p(\bar{D}) = 1 \). To make \( \bar{D} \) “large,” we can let \( \rho \) have a large support. A nice case is when the state space \( S \) is indecomposable for the above Markov chain (i.e., there do not exist two disjoint absorbing sets). Then \( \bar{D} \) is indeed large regardless of our choice of the initial distribution \( \rho \)\(^7\).

To have a large absorbing set inside \( B \) is an appealing property of the preceding way of choosing \( p \) and \( B \). To elaborate this point, suppose \( \mu \) is nonrandomized. Then on the absorbing set \( \bar{D} \),
\[
F_\theta(Q ; J)(x, \mu(x)) = g(x, \mu(x)) + \alpha \int_{\bar{D}} \min \{ J(x'), Q(x', \mu(x')) \} q(dx' \mid x, \mu(x)), \quad \forall x \in \bar{D}.
\]
(Because for \( x \in \bar{D} \), \( q(B \mid x, \mu(x)) = \kappa(B \mid x) = 1 \) and consequently, the second term in the definition (3.4) of \( F_\theta(Q ; J)(x, \mu(x)) \) vanishes, similar to Example 3.1.\(^8\) The above expression shows that for the state-control pairs \((x, \mu(x))\), \( x \in \bar{D} \), the evaluation step (3.8) with the mapping \( F_\theta(\cdot ; J) \) acts like standard policy evaluation when \( J \) is sufficiently large (cf. the alternative description of standard policy evaluation preceding Remark 3.3). This is desirable if we believe that the policy \( \mu \) achieves better costs than the estimated costs \( J \), for in that case our choice of \( B \) allows the step (3.8) to evaluate \( \mu \) faithfully on the absorbing set \( \bar{D} \) of the Markov chain induced by \( \mu \).

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 (3.8)-(3.9), since the two other policy iteration-like algorithms are its special cases.

4.1 Discounted Case (D)

In case (D) we consider the sets \( A_b(S) \), \( A_b(\Gamma) \) of bounded, lower semi-analytic functions. As mentioned in Section 2.1, they are closed subsets of the Banach spaces \( M_b(S) \) and \( M_b(\Gamma) \) respectively, and endowed with the metric \( d_{\sup}(f, f') = \| f - f' \|_\infty \). The spaces \((A_b(S), d_{\sup})\) and \((A_b(\Gamma), d_{\sup})\)

---

\(^6\)By definition \( \kappa(dx' \mid x) \) has the following two properties, which make it a state transition kernel on \((S, \mathcal{W}(S))\). First, for any fixed state \( x \in S \), \( \kappa(\cdot \mid x) \) is a probability measure on \((S, \mathcal{W}(S))\). Second, for any fixed set \( D \in \mathcal{W}(S) \), by [8, Prop. 7.46], \( q(D \mid x, u) = \int_D \chi_D(dx') q(dx' \mid x, u) \) as a function of \((x, u)\) is universally measurable on \( S \times C \), and therefore, \( \kappa(D \mid x) \) as a function of \( x \) is universally measurable on \( S \) by [8, Prop. 7.46].

\(^7\)In particular, \( S \) is indecomposable if the Markov chain is \( \psi \)-irreducible (with \( \psi \) being a maximal irreducibility measure). Then \( \psi \) is absolutely continuous with respect to the probability measure \( p \) given above [37, Prop. 4.2.1(ii)], and moreover, \( p \) is itself a maximal irreducibility probability measure of the Markov chain (so \( p, \psi \) are mutually absolutely continuous) if the initial distribution \( \rho \) is an irreducibility measure [37, Prop. 4.2.2(iv)].
are complete. The algorithm \( (3.8)-(3.9) \) works on the product space \( A_b(S) \times A_b(\Gamma) \) endowed with the metric
\[
d((J, Q), (J', Q')) = ||(J, Q) - (J', Q')||_\infty := \max \{||J - J'||_\infty, ||Q - Q'||_\infty\},
\]
which is also a complete metric space. The convergence results below use a contraction property of \( F_\theta \) and parallel those given in our earlier work \[11\] for discounted finite-state and control problems.

**Lemma 4.1.** (D) Let \( \theta \in \Theta \), \( J, J' \in A_b(S) \), and \( Q, Q' \in A_b(\Gamma) \). Then
\[
\|F_\theta(Q; J) - F_\theta(Q'; J')\|_\infty \leq \alpha \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\},
\]
\[
\|F_\theta(Q; J) - Q^*\|_\infty \leq \alpha \max \{\|J - J^*\|_\infty, \|Q - Q^*\|_\infty\},
\]
\[
\|Q_{\theta,J} - Q^*\|_\infty \leq \alpha \|J - J^*\|_\infty.
\]

**Proof.** For every \( (x, u) \in \Gamma \),
\[
J(x) \leq J'(x) + \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\}, \quad Q(x, u) \leq Q'(x, u) + \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\},
\]
so
\[
\min \{J(x), Q(x, u)\} - \min \{J'(x), Q'(x, u)\} \leq \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\}
\]
and by symmetry,
\[
\left| \min \{J(x), Q(x, u)\} - \min \{J'(x), Q'(x, u)\} \right| \leq \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\}.
\]

Using the above inequality and the definition of \( F_\theta \) given in Eq. \[3.4\], a direct calculation then shows that for each \( (x, u) \in \Gamma \),
\[
\left| F_\theta(Q; J)(x, u) - F_\theta(Q'; J')(x, u) \right| \leq \alpha \|J - J'\|_\infty \cdot q(S \setminus B \mid x, u)
\]
\[
+ \alpha \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\} \cdot q(B \mid x, u)
\]
\[
\leq \alpha \max \{\|J - J'\|_\infty, \|Q - Q'\|_\infty\}.
\]

This proves the first inequality in the lemma, from which the second desired inequality follows by setting \( (J', Q') = (J^*, Q^*) \) and using the fact \( F_\theta(Q^*; J^*) = Q^* \) (Prop. \[3.3\]a). To show the third inequality in the lemma, set \( Q = Q_{\theta,J} \) in the second inequality just proved. Since \( Q_{\theta,J} = F_\theta(Q_{\theta,J}; J) \) (Prop. \[3.2\]D)), we obtain \( \|Q_{\theta,J} - Q^*\|_\infty \leq \alpha \max \{\|J - J^*\|_\infty, \|Q_{\theta,J} - Q^*\|_\infty\} \), and since \( \alpha < 1 \), this is equivalent to \( \|Q_{\theta,J} - Q^*\|_\infty \leq \alpha \|J - J^*\|_\infty \).

**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 \( (3.8)-(3.9) \) converges to \((J^*, Q^*)\), and
\[
\|\langle J_k, Q_k \rangle - (J^*, Q^*)\|_\infty \leq \alpha^k \|\langle J_0, Q_0 \rangle - (J^*, Q^*)\|_\infty.
\]

**Proof.** At iteration \( k \), either \( Q_{k+1} = F_\theta^k(Q_k; J_k) \) 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 \[4.1\] \( n \) times, we have
\[
\|F_\theta^k(Q_k; J_k) - Q^*\|_\infty \leq \alpha \max \{\|J_k - J^*\|_\infty, \alpha^{n-1}\|Q_k - Q^*\|_\infty\},
\]
whereas for the second case, \( \|Q_{\theta,J_k} - Q^*\| \leq \alpha \|J_k - J^*\|_\infty \) by the third inequality in Lemma \[4.1\]. Thus in either case,
\[
\|Q_{k+1} - Q^*\|_\infty \leq \alpha \max \{\|J_k - J^*\|_\infty, \|Q_k - Q^*\|_\infty\}.\]
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
\[
\|J_{k+1} - J^*\|_\infty = \|M(Q_{k+1}) - M(Q^*)\|_\infty \leq \alpha \max \{ \|J_k - J^*\|_\infty, \|Q_k - Q^*\|_\infty \}.
\]
Combining the preceding two inequalities, we obtain
\[
\|(J_{k+1}, Q_{k+1}) - (J^*, Q^*)\|_\infty \leq \alpha^{k+1}\|(J_0, Q_0) - (J^*, Q^*)\|_\infty,
\]
which is the desired inequality and implies \( (J_k, Q_k) \to (J^*, Q^*) \).

**Remark 4.1** (Finding near-optimal policies). From the iterate sequence \( \{(J_k, Q_k)\} \) generated by the algorithm, we may extract an asymptotically near-optimal sequence of nonrandomized stationary policies \( \{\nu_k\} \) as follows. Specify \( \epsilon > 0 \) and let \( \nu_k \) be such that \( Q_k(x, \nu_k(x)) \leq M(Q_k)(x) + \epsilon \) for all \( x \in S \) (cf. Eq. (3.12)). The policies \( \{\mu_k\} \) (resp. \( \{\mu'_k\} \)) generated in the policy iteration-like algorithm II (resp. III), for example, already satisfy these conditions. To analyze \( J_{\nu_k} \), note that by Theorem 4.1,
\[
\|Q_k - Q^*\|_\infty \leq \alpha^k \Delta \text{ where } \Delta = \|(J_0, Q_0) - (J^*, Q^*)\|_\infty.
\]
Hence for all \( x \in S \),
\[
Q^*(x, \nu_k(x)) \leq Q_k(x, \nu_k(x)) + \alpha^k \Delta \leq M(Q^*)(x) + 2\alpha^k \Delta + \epsilon = J^*(x) + 2\alpha^k \Delta + \epsilon.
\]
Since \( Q^*(x, \nu_k(x)) = T_{\nu_k}(J^*)(x) \) (cf. Eqs. (3.11), (2.4)), the preceding inequality shows
\[
T_{\nu_k}(J^*)(x) \leq J^*(x) + 2\alpha^k \Delta + \epsilon, \quad \forall x \in S.
\]
By the monotonicity and contraction properties of \( T_{\nu_k} \), this implies that \( \|J_{\nu_k} - J^*\|_\infty \leq (2\alpha^k \Delta + \epsilon)/(1 - \alpha) \) and hence
\[
\limsup_{k \to \infty} \|J_{\nu_k} - J^*\|_\infty \leq \epsilon/(1 - \alpha). \tag{4.1}
\]
Alternatively, we may choose \( \nu_k \) such that \( \|T_{\nu_k}(J_k) - T(J_k)\|_\infty \leq \epsilon \) (this is possible by the selection theorem [8, Prop. 7.50]). Then using the contraction property of \( T \) and \( T_{\nu_k} \), it can be shown (see e.g., [7, p. 45]) that \( \|J_{\nu_k} - J^*\|_\infty \leq (\epsilon + 2\alpha \|J_k - J^*\|_\infty)/(1 - \alpha) \) for all \( k \), so by Theorem 4.1 Eq. (4.1) holds and the sequence \( \{\nu_k\} \) is asymptotically \( \epsilon/(1 - \alpha) \)-optimal in that sense.

### 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 algorithm [3.8], [3.9] operates with nonpositive lower semi-analytic functions in \( A_-(S) \) and \( A_-(\Gamma) \). We will rely on the monotonicity and fixed point properties of \( F_\theta \) to ensure its convergence.

First, we derive some simple upper and lower bounds on the iterates generated by the algorithm. To simplify notation, let
\[
H(x, u, J) = g(x, u) + \int_{S'} J(x') q(dx' | x, u), \quad (x, u) \in \Gamma.
\]
Expressed in these terms, \( T(J)(x) = \inf_{u \in U(x)} H(x, u, J) \), the optimality equation \( J^* = T(J^*) \) is
\[
J^*(x) = \inf_{u \in U(x)} H(x, u, J^*), \quad x \in S,
\]
and by the definition of \( Q^* \) (cf. Eq. (3.1)),
\[
Q^*(x, u) = H(x, u, J^*), \quad (x, u) \in \Gamma. \tag{4.3}
\]
§4. Convergence Analysis for Cases (D)(N)

**Lemma 4.2.** (D)(N)(P) Let $J \in A(S)$ be as in Prop. 3.2, and let $Q \in A(\Gamma)$. Then

\begin{align}
F_\theta(Q; J)(x, u) & \leq H(x, u, J), & Q_{\theta, J}(x, u) & \leq H(x, u, J), & \forall (x, u) \in \Gamma, \\
M(F_\theta(Q; J)) & \leq T(J), & M(Q_{\theta, J}) & \leq T(J).
\end{align}

**Proof.** Since $\min\{J(x'), Q(x', u')\} \leq J(x')$, we have that for all $(x, u) \in \Gamma$,

\begin{align}
F_\theta(Q; J)(x, u) &= g(x, u) + \alpha \int_{S \setminus B} J(x') q(dx' | x, u) \\
&\quad + \alpha \int_{B} \int_{C} \min\{J(x'), Q(x', u')\} \mu(du' | x') q(dx' | x, u) \\
&\leq g(x, u) + \alpha \int_{S \setminus B} J(x') q(dx' | x, u) + \alpha \int_{B} J(x') q(dx' | x, u) = H(x, u, J).
\end{align}

From this inequality and Prop. 3.2 it follows $Q_{\theta, J} = F_\theta(Q_{\theta, J}; J) \leq H(x, u, J)$. This establishes the inequalities in Eq. (4.4). Minimizing over $U(x)$ for each $x$ in Eq. (4.4), we obtain Eq. (4.5).

We use the above bounds to upperbound the iterates of the algorithms. The next lemma applies also to (D)(P). For the algorithm that uses the second rule of (3.8) to set $Q_{k+1} = Q_{\theta, J}$ at some iterations, the second statement of the lemma will rely on Prop. 3.3(d), which in the case (P) will be proved in Appendix B as Prop. 3.1.

**Lemma 4.3.** (N)(P) Let $\{(J_k, Q_k)\}$ be iterates generated by the iteration (3.8)-(3.9) 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 $k \geq 1$,

\begin{align}
J_k & \leq T^k(J_0), & Q_k(x, u) & \leq H(x, u, J_{k-1}), & \forall (x, u) \in \Gamma.
\end{align}

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

**Proof.** For each $k \geq 0$, either $Q_{k+1} = F^\theta_n(Q_k; J_k)$ or $Q_{k+1} = Q_{\theta, J_k}$ for some $\theta \in \Theta, n \geq 1$. By Eq. (4.4), the right-hand side inequality for $Q_k$ in Eq. (4.6) follows. Since $J_{k+1} = M(Q_{k+1})$, we have, by Eq. (4.5), $J_{k+1} \leq T(J_k)$ for all $k$. This implies $J_k \leq T^k(J_0)$ by the monotonicity of $T$.

Let $J_0 \geq J^*$ and $Q_0 \geq Q^*$. We show by induction that $J_k \geq J^*, Q_k \geq Q^*$ for every $k$. Suppose this is true for some $k \geq 0$. Consider the $(k+1)$-th iterate. If $Q_{k+1} = F^\theta_n(Q_k; J_k)$, then by the induction hypothesis, the monotonicity of $F_\theta$ (cf. Eq. (4.4)) and Prop. 3.3(a), we have

\[ Q_{k+1} = F^\theta_n(Q_k; J_k) \geq F^\theta_n(Q^*(J^*)); J^* \geq Q^*. \]

If $Q_{k+1} = Q_{\theta, J_k}$, then since $J_k \geq J^*$ by the induction hypothesis and $Q_{\theta, J_k} \geq Q_{\theta, J^*}$ by Prop. 3.3(c), we have $Q_{k+1} \geq Q_{\theta, J^*} = Q^*$ by Prop. 3.3(d) (proved as Prop. 3.1 for (P)). Thus in either case, $Q_{k+1} \geq Q^*$, and consequently, $J_{k+1} = M(Q_{k+1}) \geq M(Q^*) = J^*$. This completes the induction.

The relation $J^* \leq J_k \leq T^k(J_0)$ in Lemma 4.3, which holds when $J_0 \geq J^*$, 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 $J_0 \geq J^*$ and $Q_0 \geq Q^*$, the sequence $\{(J_k, Q_k)\}$ generated by the iteration (3.8)-(3.9) converges to $(J^*, Q^*)$. 


Proof. We show first $J_k \to J^*$. We have $J^* \leq J_k \leq T^k(J_0)$ by Lemma 4.3. Since $J^* \leq J_0 \leq 0$ by assumption and $T^k(0) \downarrow J^*$ 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 [21 p. 131] (applied to nonpositive functions),
\[
\limsup_{k \to \infty} H(x,u,J_k) \leq H(x,u,\limsup_{k \to \infty} J_k) = H(x,u,J^*) = Q^*(x,u)
\]
(cf. Eqs. (4.2)-(4.3)). Since $Q^*(x,u) \leq Q_{k+1}(x,u) \leq H(x,u,J_k)$ by Lemma 4.3, this implies the convergence $Q_k \to Q^*$.

\[\square\]

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 $\epsilon$-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 [8] Prop. 9.20), even if it were available. Moreover, even if an optimal stationary policy exists, it is possible that a policy $\mu$ satisfies $T_{\mu}(J^*) = T(J^*)$ without being optimal. Hence, we do not expect to have simple ways to obtain near-optimal policies from the iterate sequence $\{(J_k,Q_k)\}$ generated by our algorithm. Intuitively, if we start the algorithm with $J_0 = 0$, $Q_0 = 0$ or cost functions of some policies, it seems possible to construct for each given state, history-dependent or semi-Markov policies that asymptotically become near-optimal for that state, by using the relations between the optimal stopping problems associated with the step [3.5] and the original problem (cf. Remarks 5.1, 3.2). Due to its complexity, however, we do not discuss this subject in this paper.

§. Convergence Analysis for 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.

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 $T^k(0) \uparrow J_\infty \leq J^*$, 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 [8] Prop. 9.17, Cor. 9.17.1]; but these conditions are restrictive. For policy iteration, it can happen that for a suboptimal stationary policy $\mu$, $J_\mu = T_\mu(J_\mu) = T(J_\mu)$, even in finite-state and control problems[9] and the method terminates with the suboptimal policy $\mu$.

We thus look for ways to mitigate the difficulties. Any condition forcing $T^k(0) \uparrow J^*$, however, seems restrictive, in view of Maitra and Sudderth’s result [15]. They showed that $J^*$ can be obtained

---

8As 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 $\mu$ that makes self-transitions at state 1 satisfies $T_\mu(J^*) = T(J^*)$.

9For 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 leads 1 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 $\mu$ with $\mu(1) = 0$ is suboptimal but satisfies $T_\mu(J_\mu) = T(J_\mu)$. See [11] 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 [55] and of Miller and Veinott [58] based on the concept of sensitive optimality [55; see also [11] Sec. 10.3]).
by applying $T$ a transfinite number of times, starting from the function $J = 0$, and in general, the number of times needed can be uncountably infinite [35, 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 [35, 28] to suit our purpose.

Before proceeding, let us give a simple example to elucidate the behavior of value iteration just discussed. The example is from [8, p. 215]. In this example $J_\infty < J^* \equiv \infty$. We illustrate how value iteration with transfinite recursion is able to obtain its relation with Whittle’s bridging condition. The pointwise limit of the nondecreasing sequence $\{T^k(0)\}$ is $J_\infty = (0,1,1,\ldots)$, since $T^k(0) = (0,1,1,\ldots,1,0,0,\ldots)$ with $k$ 1’s followed by all 0’s. As in [32], 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 $J_{\infty m+1} = J_{\infty m} + 1$. 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 [35, 28] (cf. Remark 5.3) and its appealingly simple form. (The paper [58] 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 [58, 28], as we will remark shortly.

Theorem 5.1. (P) The following hold:

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

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

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

where $J \in A_+(S)$ satisfies $J \leq J^*, T^k(J) \to J^*$. In particular, if $T^k(0) \uparrow J^*$, 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 $T^k(J) \leq T^k(J^*) \leq T^k(cJ^*)$ by the monotonicity of $T$. Since $T^k(J) \to J^*$ by assumption and $T^k(cJ^*) \downarrow J^*$ by part (a), part (b) follows. For part (c), by [8, Prop. 9.10(P)] we have the following implication,

$$J \in A_+(S), \quad J = T(J) \implies J \geq J^*,$$

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 \( J = J^* \). Then Theorem 5.1(b) reads as:
\[
T^k(J) \to J^*, \quad \forall J \in A_+(S) \text{ s.t. } J^* \leq J \leq cJ^*, \ c > 1.
\]
(5.1)
Indeed, in view of the result of Maitra and Sudderth \( [35] \) and the simple Example 5.1, \( J^* \) may be the only function that can serve as an initial function \( J \) lying below \( J^* \) from which VI converges as 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 \( \mu \) with \( T_\mu(J_\mu) = T(J_\mu) \), \( J_\mu \) 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 \( [50] \) p. 881 (see also \( [35] \) p. 930), \( J^* \) takes values in \( \{0, 1\} \) and \( T^k(0) \neq J^* \). 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 \( a \geq 1 \) elsewhere, value iteration turns out to converge in one iteration in this example (see Appendix D). \( \square \)

Remark 5.3. Whittle’s bridging condition is as follows: for some real \( c \) and Markov policy \( \pi \), either \( J^*_\pi \leq cT^n(0) \) for some \( n \), or \( J^*_\pi \leq cJ^*_\infty \) 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 \( a \) \( [58] \) \( [28] \). A similar, slightly weaker condition leading to the same conclusions is \( J^* \leq cT^n(0) \) 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 \( T^k(0) \uparrow J^* \) holds, they can be used to set \( J = 0 \) 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 \( [58] \) \( [28] \). \( \square \)

We now proceed 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
\[
T^k(J) \downarrow J^\infty \quad \text{and} \quad T(J^\infty) \leq J^\infty.
\]

Proof. By the monotonicity of \( T \), \( T^k(J) \downarrow J^\infty \); by \( [3] \) Lemma 7.30(2)], \( J^\infty \in A_+(S) \). 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 \( T(J^\infty) \leq J^\infty \). \( \square \)

Lemma 5.2. Let \( c > 1 \). We have \( T(cJ^*) \leq cJ^* \) and for some \( J^\infty \in A_+(S) \),
\[
T^k(cJ^*) \downarrow J^\infty, \quad T(J^\infty) = J^\infty, \quad J^* \leq J^\infty \leq cJ^*.
\]

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 \( T \) and the fact \( T(J^*) = J^* \) 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 an arbitrary \( x \in S \). If \( J^*(x) = \infty \), then \( T(J^\infty)(x) = J^\infty(x) = \infty \) by the preceding relation. Suppose \( J^*(x) < \infty \); we prove \( T(J^\infty)(x) \geq J^\infty(x) \) below.

To simplify notation, consider the function

\[
H(x, u, J) = g(x, u) + \int_S \int_{x'} J(x') q(\text{d}x' \mid x, u), \quad u \in U(x).
\]

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

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

For \( u \in D(x) \),

\[
H(x, u, cJ^*) \leq cH(x, u, J^*) < \infty
\]

(because \( c > 1 \) and \( g \geq 0 \)), so in view of the relation \( cJ^* \geq J^k \downarrow J^\infty \), we have by the monotone convergence theorem [21, p. 131],

\[
H(x, u, J^\infty) = \lim_{k \to \infty} H(x, u, J^k).
\]

Consequently,

\[
H(x, u, J^\infty) = \limsup_{k \to \infty} \left\{ \inf_{u \in U(x)} H(x, u, J^k) \right\} = \lim_{k \to \infty} T(J^k)(x) = J^\infty(x), \quad \forall u \in D(x).
\]

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

\[
H(x, u, J^\infty) \geq H(x, u, J^*) = \infty.
\]

Combining this with Eq. (5.3), we have

\[
T(J^\infty)(x) = \inf_{u \in U(x)} H(x, u, J^\infty) = \inf_{u \in D(x)} H(x, u, J^\infty) \geq J^\infty(x).
\]

This completes the proof.

We are now ready to prove Theorem 5.1. We will give two different proofs, one here, and the other in Appendix B. In the proof that follows, we will use a simple concavity property of \( T \), which can be verified directly. (Hartley [22] 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 \( J_1, J_2 \in A_+(S) \),

\[
T(\beta J_1 + (1 - \beta)J_2) \geq \beta T(J_1) + (1 - \beta) T(J_2).
\]

We will also use Maitra and Sudderth’s results [35]. Let \( \omega_1 \) be the first uncountable ordinal. For ordinals \( \xi < \omega_1 \), define functions \( J^\xi \in A_+(S) \) by transfinite recursion as follows. Let

\[
J^0 = T(0), \quad J^\xi = T\left( \sup_{\eta < \xi} J^\eta \right), \quad \text{for} \ \xi > 0.
\]

Also let

\[
J^{\omega_1} = \sup_{\xi < \omega_1} J^\xi.
\]

That all these functions are indeed in \( A_+(S) \) is proved in [35]. Moreover, Maitra and Sudderth [35] Thm. 1.1] proved that

\[
T(J^{\omega_1}) = J^{\omega_1} = J^*.
\]

(For ordinals, transfinite induction and transfinite recursion, see e.g., [33] p. 27-28, [21] Secs. 1.3, A.3 or [49] Chap. 1.)
Proof of Theorem 5.1. Denote $J^* = \lim_{k \to \infty} T^k(cJ^*)$. By Lemma 5.2,

$$T(J^\infty) = J^\infty, \quad J^* \leq J^\infty \leq cJ^*.$$

We now prove $J^\infty = J^*$. This will prove part (a) and hence the entire theorem, as discussed earlier.

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

$$T(J^\infty) \geq T(\beta J^\infty + (1 - \beta)0) \geq \beta T(J^\infty) + (1 - \beta)T(0).$$

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

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

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

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

Then

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

Consequently, by the monotonicity and concavity properties of $T$,

$$J^* \geq \beta J^\infty + (1 - \beta) T\left( \sup_{\eta<\xi} J^\eta \right) = \beta J^\infty + (1 - \beta) J^\xi,$$

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 $J^\xi$ for $\xi < \omega_1$, and the definition of $J^{\omega_1}$ for $\xi = \omega_1$, together with the fact $T(J^{\omega_1}) = J^{\omega_1}$ (35, Thm. 1.1); cf. Eq. (5.5). This proves, by transfinite induction, that the inequality (5.6) holds for all $\xi \leq \omega_1$, and in particular,

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

Since $J^{\omega_1} = J^*$ by 35, Thm. 1.1 (cf. Eq. 5.5) and $J^* \geq 0$, we have $J^* \geq J^\infty$ by Eq. (5.7). We also have $J^\infty \geq J^*$. Therefore, $J^\infty = J^*$.

We mention two immediate implications of Theorem 5.1.

Corollary 5.1. (P) Suppose that the state space $S$ is finite and $J^*$ is real-valued. Let $V$ 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 $V$. Moreover, $T^k(J) \to J^*$ for all $J \in V$.

Proof. By Theorem 5.1(c), $J^*$ is the unique fixed point of $T$ in $\{ J \geq 0 \mid \exists c \in \mathbb{R}_+ \text{ s.t. } J \leq cJ^* \} = V$. By Theorem 5.1(b), we have $T^k(J) \to J^*$ for all $J \in V$ if $T^k(0) \uparrow J^*$. The latter holds when $S$ is finite and $J^*$ is finite everywhere (cf. 11, Thm. 7.3.10(a))]. The reason is that $T^k(0)$ converges to its limit $J_\infty$ uniformly, i.e., for any $\epsilon > 0$, $\|T^k(0) - J_\infty\| \leq \epsilon$ for all $k$ sufficiently large. Thus with 1 denoting the constant function 1, we have, by the monotonicity of $T$, that for all $k$ sufficiently large,

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

Since $\epsilon$ is arbitrary, this implies $T(J_\infty) \leq J_\infty$. Since $J_\infty \leq T(J_\infty)$ 5 Prop. 9.16, we have $J_\infty = T(J_\infty)$ and hence by 5 Prop. 9.16, $J_\infty = J^*$, i.e., $T^k(0) \uparrow J^*$.
In connection with Cor. 5.1, 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 $J^*_1(1) = \infty$). 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 $T^k(0)(2) = 0$ for all $k$, so $J^*_\infty(2) = 0$; but $J^*_2(1) = 1$.

The next corollary relates to the convergence of the value functions generated by exact policy iteration (assuming it can be carried out). First, note that Theorem 5.1 implies also the convergence of value iteration for certain initial functions that cannot be bounded by a multiple of $J^*$. Specifically, by the theorem, if $J \in A_+(S)$ is such that $J \leq T^n(J) \leq cJ^*$ for some $n \geq 0$ and $c > 1$, where $J$ is as in Theorem 5.1(b), then $T^k(J) \rightarrow J^*$.

**Corollary 5.2** (Convergence result for policy iteration under (P)). Suppose $\{\mu_k\}$ is a sequence of stationary policies that satisfy $J_{\mu_k} \in A_+(S)$, $T_{\mu_k+1}(J_{\mu_k}) = T(J_{\mu_k})$ for $k \geq 0$. Then if $\mu_0$ is such that $T^n(\mu_0) \leq cJ^*$ for some $n \geq 0$ and $c > 1$, we have $J_{\mu_k} \rightarrow J^*$.

**Proof.** By Theorem 5.1, the condition on $\mu_0$ implies that $T^k(\mu_0) \rightarrow J^*$. Since $T_{\mu_k+1}(J_{\mu_k}) = T(\mu_k) \leq T_{\mu_k}(J_{\mu_k}) = J_{\mu_k}$, by the monotonicity of $T_{\mu_k+1}$, we have that $T_{\mu_k+1}(J_{\mu_k})$ is non-increasing as $n \rightarrow \infty$, so $J_{\mu_k+1} = \lim_{n \rightarrow \infty} T^n_{\mu_k+1}(0) \leq \lim_{n \rightarrow \infty} T^n_{\mu_k+1}(J_{\mu_k}) \leq T(J_{\mu_k})$. It then follows that $J^* \leq J_{\mu_k+1} \leq T^{k+1}(J_{\mu_k})$, and hence $J_{\mu_k} \rightarrow J^*$.

A similar conclusion holds for modified policy iteration (assuming it can be carried out).

Note also that Cor. 5.2 does not imply $J^*_\infty = J^*$. Neither does it imply the existence of an $\epsilon$-optimal stationary policy because the convergence of $J_{\mu_k}$ to $J^*$ is only pointwise, but it does imply the existence of an $\epsilon$-optimal stationary policy for each state.

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 $\epsilon$-optimal stationary policy for $\bar{x}$ does not exist. Then there exists no stationary policy $\mu$ such that $J_\mu \in A_+(S)$, $T^k(J_\mu) \rightarrow J^*$ (hence there exists no $J_\mu \in A_+(S)$ with $T^n(J_\mu) \leq cJ^*$ for some $n \geq 0$ and $c > 1$).

**Proof.** To arrive at a contradiction, suppose $\mu$ is a stationary policy with $J_\mu \in A_+(S)$, $T^k(J_\mu) \rightarrow J^*$. Let $k$ be large enough so that $T^k(J_\mu)(\bar{x}) \geq J^*(\bar{x}) + \epsilon/2$. By the selection theorem of [8] Prop. 7.50, there exist stationary policies $\mu_1, \mu_2, \ldots, \mu_k$ satisfying

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

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

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

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) [8 Props. 5.10, 5.4], we then obtain that for this subproblem, the optimal cost function is universally measurable and there exists an optimal nonrandomized stationary policy $\tilde{\mu}$. Clearly, $\tilde{\mu}$ corresponds to a universally measurable stationary policy in

---

10Specifically, suppose $\{ J_{\mu_k}, \mu_k \}$ is a sequence of cost function and stationary policy pairs that satisfy $J_{\mu_k} \in A_+(S)$, $J_{\mu_k+1} = T_{\mu_k}(J_{\mu_k})$ for some $\mu_k \geq 1$, $T_{\mu_k+1}(J_{\mu_k+1}) = T(J_{\mu_k+1})$ for all $k \geq 0$. Suppose in addition that $T_{\mu_0}(J_0) \leq J_0$ and $T^n(J_0) \leq cJ^*$ for some $n \geq 0$ and $c > 1$. Then $J^* \leq J_{\mu_k} \leq J_0 \rightarrow J^*$. 
the original problem. By the optimality of \( \hat{\mu} \) in the subproblem, we have \( J_{\hat{\mu}}(\bar{x}) \leq J_{\mu}(\bar{x}) \leq J^*(\bar{x}) + \epsilon \), which contradicts the assumption that there exists no \( \epsilon \)-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 [50, 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 \( \epsilon \)-optimal Markov policy, for example, is to apply at state 1 control \( u_k \) for the \( k \)th stage, with \( \sum_k u_k \leq \epsilon \). No stationary policy is \( \epsilon \)-optimal for states 1 and 2: for any stationary policy, 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, \( J_\mu \) 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 \( J_0 = 0, Q_0 = 0 \), here, as can be shown by a direct calculation, doing so reduces the method to value iteration \( T^k(0) \), and this is undesirable computationally even if \( T^k(0) \rightarrow J^* \), 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 (3.8)-(3.9) 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) Let \( J_0 \in A_+(S) \) and \( Q_0 \in A_+(\Gamma) \).

(a) Suppose

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

Then the sequence \( \{(J_k, Q_k)\} \) generated by the iteration (3.8)-(3.9) converges to \( (J^*, Q^*) \).

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

(c) Suppose \( T^k(J) \uparrow J^* \) for some \( J \in A_+(S) \). Then the conclusion of (a) holds for the iteration (3.8)-(3.9) that always defines \( Q_{k+1} \) using the first rule in (3.8), under the initial condition that \( J \leq J_0 \leq cJ^* \) for some \( c > 1 \), and \( Q_0(x, u) \geq \hat{J}(x) \quad \forall (x, u) \in \Gamma \).

In either part of the theorem, we assume 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 (3.8)-(3.9) satisfies that

\[
\limsup_{k \to \infty} J_k \leq J^*, \quad \limsup_{k \to \infty} Q_k \leq Q^*.
\]

Proof. Let \( J^* = T^k(cJ^*) \). Since \( J_0 \leq cJ^* \), we have \( J_k \leq T^k(J_0) \leq J^k \) for every \( k \), by Lemma 4.3 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. 4.2, 4.3). By Lemma 4.3 for every \( k \geq 0 \),

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

If \( Q^*(x, u) \) is finite, then we have

\[
\lim_{k \to \infty} H(x, u, J_k) = H(x, u, \lim_{k \to \infty} J_k) = H(x, u, J^*) = Q^*(x, u),
\]
§5. Convergence Analysis for Case (P) 33

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

\[ \lim_{k \to \infty} Q_{k+1}(x, u) \leq Q^*(x, u). \]

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.\(^\text{11}\)

**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 4.3 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 \( k \geq 0 \),

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

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

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

\[ \hat{J}^*_k(x, u) = \hat{J}_k(x), \quad \forall (x, u) \in \Gamma. \]

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 \( F_\theta \) and its monotonicity shows that

\[ F_\theta(0; \hat{J}_k) \geq \hat{J}^*_1, \quad F_\theta(\hat{J}^*_1; \hat{J}_k) \geq \hat{J}^*_2, \quad \cdots \quad F_\theta(\hat{J}^*_{k-1}; \hat{J}_k) \geq \hat{J}^*_k, \tag{5.10} \]

and that for every \( n \geq 1 \),

\[ F^n_\theta(\hat{J}^*_k; \hat{J}_k) \geq F_\theta(\hat{J}^*_k; \hat{J}_k). \tag{5.11} \]

In view of \( g \geq 0 \) and the definition of \( H(x, u, J) \) (cf. Eq. (4.2)), a direct calculation shows that

\[ F_\theta(\hat{J}^*_k; \hat{J}_k)(x, u) = H(x, u, \hat{J}_k) \geq T(\hat{J}_k)(x) = \hat{J}_{k+1}(x), \quad \forall (x, u) \in \Gamma. \tag{5.12} \]

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

\[ F^n_\theta(Q_k; J_k) \geq F^n_\theta(\hat{J}^*_k; \hat{J}_k) \geq F_\theta(\hat{J}^*_k; \hat{J}_k), \]

where the first inequality follows from the monotonicity of \( F_\theta \) (cf. Eq. (3.6)) and the induction hypothesis that \( J_k \geq \hat{J}_k, Q_k \geq \hat{J}^*_k \), and the second inequality follows from Eq. (5.11). For the case \( Q_{k+1} = Q_{\theta, J_k} \), we have

\[ Q_{\theta, J_k} \geq F^{k+1}_\theta(0; J_k) \geq F^{k+1}_\theta(0; \hat{J}_k) \geq F_\theta(\hat{J}^*_k; \hat{J}_k), \]

where the first inequality holds because \( F^n_\theta(0; J_k) \uparrow Q_{\theta, J_k} \) 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 \( F_\theta \) (cf. Eq. (3.6)), and the third inequality follows from Eq. (5.10) and the monotonicity of \( F_\theta(\cdot; J_k) \). Thus in either case, we have

\[ Q_{k+1} \geq F_\theta(\hat{J}^*_k; \hat{J}_k) \geq \hat{J}^*_{k+1}, \quad J_{k+1} = M(Q_{k+1}) \geq \hat{J}_{k+1}, \]

\(^{11}\)For part (a), we will use the lower bounds given in Lemma 4.3 which rely on the relation \( Q_{\theta, J^*} = Q^* \) for all \( \theta \in \Theta \) (cf. Prop. 3.3(d)). 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 \( Q_{\theta, J_k} \) at some iterations.
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, \( Q_{k+1} \geq F_\Theta(J_k^*; J_k) \) for every \( k \). By Eq. (5.12), this is equivalent to

\[
Q_{k+1}(x, u) \geq H(x, u, J_k), \quad \forall (x, u) \in \Gamma.
\]

(5.13)

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

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

(cf. Eqs. (4.2), (4.3)). Together with Lemma 5.3 the preceding two relations imply that \( Q_k \to Q^* \).

(c) By assumption \( T^k(J) \uparrow J^* \). For the algorithm stated in (c), if we define \( \hat{J}_0 = J, \hat{J}_k = T^k(J) \) 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 \( k \),

\[
J_k \geq T^k(J), \quad Q_k(x, u) \geq T^k(J)(x), \quad \forall (x, u) \in \Gamma,
\]

and that \( J_k \to J^*, Q_k \to Q^* \). The reason that the proof arguments for part (b) are applicable here is as follows. 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 \( F_\Theta(0; \hat{J}_k) \geq \hat{J}_k^* \) in Eq. (5.10). This relation is needed in the convergence proof only when \( Q_{k+1} \) is generated by the second rule of (3.8) 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 (3.8)-(3.9)**

Let us consider a variation of the algorithm (3.8)-(3.9), whereby instead of (3.8), 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

\[
Q_{k+1} \leq F_{\theta_k}(Q_{k+1}; J_k), \quad Q_{k+1} \geq Q_{\theta_k, J_k}.
\]

(5.14)

Then let

\[
J_{k+1} = M(Q_{k+1}).
\]

(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. [8] Prop. 9.10(P), [11] Sec. 7.3.6). Thus to calculate \( Q_{\theta_k, J_k} \) in the algorithm (3.8)-(3.9) without iterating \( F_{\theta_k}(0; J_k) \) up to convergence, we cannot solve the optimal stopping problem associated with \( (\theta_k, J_k) \) by simply solving some linear program.

On the other hand, an upper bound on \( Q_{\theta_k, J_k} \) will suffice if it also satisfies the first relation in (5.14), as we show in the theorem below. Unlike computing \( Q_{\theta_k, J_k} \), a solution to (5.14) can be computed by solving a linear program associated with the optimal stopping problem defined by \( (\theta_k, J_k) \), under certain conditions that involve \( (\theta_k, J_k) \), as we will show in Lemma A.2, 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 \( J_k \leq cJ^* \) for some \( c > 1 \), the algorithm (5.14)-(5.15) will generate \( J_k \) with \( J_k \leq cJ^* \) 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 \( J_k \leq cJ^* \) 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. By Lemma 4.2
\[
F_{\theta_k}(Q_{k+1}; J_k)(x,u) \leq H(x,u,J_k), \quad \forall (x,u) \in \Gamma,
\]
From this and the upper bound on \(Q_{k+1}\) given in Eq. (5.14), we have
\[
Q_{k+1}(x,u) \leq H(x,u,J_k), \quad \forall (x,u) \in \Gamma, \quad J_{k+1} \leq T(J_k).
\]
The second inequality above implies, by the monotonicity of \(T\), that \(J_k \leq T^k(J_0)\) for every \(k\). Since \(J_0 \leq cJ^*\) by assumption, we obtain that the conclusion of Lemma 5.3 holds for the iteration \((5.14)-(5.15)\):
\[
\limsup_{k \to \infty} J_k \leq J^*, \quad \limsup_{k \to \infty} Q_k \leq Q^*.
\]

Under the conditions of Theorem 5.2(a), we have \(J_0 \geq J^*, Q_0 \geq Q^*\). Lemma 4.3 shows that if \(Q_{k+1} = Q_{\theta_k,J_k}\) at every iteration of the algorithm, then \(J_k \geq J^*, Q_k \geq Q^*\) for all \(k\). Since here we have \(Q_{k+1} \geq Q_{\theta_k,J_k}\) by Eq. (5.14), the same induction-based proof for Lemma 4.3 shows that for the iteration \((5.14)-(5.15)\), we have \(J_k \geq J^*, Q_k \geq Q^*\) 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 \(Q_{k+1} \geq Q_{\theta_k,J_k}\). 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 (Finding near-optimal policies). Under (P), from the sequence \(\{J_k\}\) generated by the algorithm \((5.8)-(5.9)\) or \((5.14)-(5.15)\), in general one cannot extract easily near-optimal policies in the manner of Remark 4.1. Even if \(J^*\) was available, an \(\epsilon\)-optimal stationary policy may not exist (see the discussion after Prop. 5.1 or [7, p. 145] for an example). Nevertheless, let us describe first a favorable case where one can find stationary policies \(\nu_k\) with \(J_{\nu_k} \to J^*\). In particular, let \(\{J_k\} \subset A_+(S)\) be such that \(J_k \to J^*\). Suppose \(J_k\) and \(\nu_k\) satisfy
\[
T_{\nu_k}(J_k) = T(J_k) \leq J_k, \quad \forall k \geq 1.
\]
Then \(J_{\nu_k} \to J^*\). (To see this, note that by [8, Prop. 9.11], \(J_{\nu_k}\) is the “smallest” nonnegative function \(J \in M(S)\) 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 algorithms.

Consider now a more general case. Suppose \(J_k \to J^*\) and for some \(c > 1\), \(J^* \leq J_k \leq cJ^*\) for all \(k \geq 1\). These conditions hold for our algorithms under the initial condition given in Theorem 5.2(a). With such \(\{J_k\}\), one can construct asymptotically near-optimal policies, even though these policies are not necessarily stationary or Markov and the constructions can be complicated. To simply discuss just the idea of the construction, let us assume in addition that \(T^k(cJ^*)\) converges to \(J^*\) uniformly, i.e.,
\[
e_k = \sup_{x \in S} (T^k(cJ^*)(x) - J^*(x)) \to 0.
\]
Then given \(J_k\), for any \(\delta > 0\), we can find \(n \geq 1\) such that \(T^n(J_k) \leq J_k + \delta/2\) (because \(T^n(J_k) \leq T^n(cJ^*) \leq J^* + \epsilon_n \leq J_k + \epsilon_n\) and \(\epsilon_n \to 0\)), together with an \(n\)-stage policy \((\mu_1, \mu_2, \ldots, \mu_n)\) such that
\[
(T_{\mu_1} \circ T_{\mu_2} \circ \cdots \circ T_{\mu_n})(J_k) \leq J_k + \delta.
\]
Using this inequality, for any \(\epsilon' > 0\), an \((e_k + \epsilon')\)-optimal Markov policy can be constructed as follows (similar to how a near-optimal Markov policy is formed when \(J^*\) is given [8, p. 238]). Let \(\delta_m, m \geq 1\),
be a sequence of positive numbers with \( \sum_{m} \delta_{m} = \epsilon' \). For each \( m \), let \( \pi^{(m)} = (\mu_{1}^{(m)}, \mu_{2}^{(m)}, \ldots, \mu_{N_{m}}^{(m)}) \) be an \( n_{m} \)-stage policy satisfying the inequality (5.18) for \( \delta = \delta_{m} \). Then the Markov policy \( \pi = (\pi^{(1)}, \pi^{(2)}, \ldots) \) is \((\epsilon_{k} + \epsilon')\)-optimal because \( J_{\pi} \leq J_{k} + \epsilon' \leq J^{*} + (\epsilon_{k} + \epsilon') \).

In the above, uniform convergence of \( T^{k}(cJ^{*}) \) and \( J_{k} \) to \( J^{*} \) was assumed. If the convergence is only pointwise, then instead of a Markov policy, one can similarly construct a history-dependent such policies directly from some of our algorithms, without computing the value iteration \( T^{\theta}(J_{k}) \), based on the optimal stopping interpretation of \( F_{\theta} \) for the step (3.8) (cf. Remarks 3.1, 3.2). We do not discuss this possibility in this paper, however, due to its complexity and space limit.

\[ \text{6 Applications in Semicontinuous Models} \]

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. However, our results are applicable to the model as defined in [8, Chap. 8].

To apply the mixed value and policy iteration method in these semicontinuous models, it is desirable to work with semicontinuous functions instead of lower semi-analytic functions. Our strategy is to keep the function iterates within the set of semicontinuous functions is to choose properly the \( \kappa(\cdot) \)-involved in the method. For the lower semicontinuous model, we will use Lusin’s theorem for this purpose. For the upper semicontinuous model, our result is weaker in the sense that we will need additional conditions in order to obtain nontrivial policy iteration-like algorithms that operate within the set of upper semicontinuous functions. 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 \( f : X \rightarrow \mathbb{R} \) is said to be \textit{upper semicontinuous} (u.s.c.) if for every \( c \in \mathbb{R} \), 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 \( \limsup_{n \rightarrow \infty} f(x_{n}) \leq f(x) \). A function \( f : X \rightarrow \mathbb{R} \) is said to be \textit{lower semicontinuous} (l.s.c.) if for every \( c \in \mathbb{R} \), 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 \( \liminf_{n \rightarrow \infty} f(x_{n}) \geq f(x) \).

Let \( X \) and \( Y \) be separable metrizable topological spaces. Let the topology on the space \( \mathcal{P}(Y) \) of Borel probability measures on \( Y \) be the weak topology. A stochastic kernel \( \kappa(dy \mid x) \) on \( Y \) given \( X \) is \textit{continuous} if the function \( \kappa(dy \mid \cdot) : X \rightarrow \mathcal{P}(Y) \) is continuous. (In some literature such a kernel is said to be weakly continuous; see e.g., [24].) Similarly, if restricted to a subset \( B \subset X \), the function \( \kappa(dy \mid \cdot) : B \rightarrow \mathcal{P}(Y) \) is continuous, we say \( \kappa(dy \mid x) \) is \textit{continuous on} \( B \).

\[ 12 \text{To see this, write} \ V_{\pi}(n, J) \text{for the cost function of} \pi \text{in the} n \text{-stage problem, with the terminal cost at stage} n + 1 \text{given by} J_{t} \text{in particular,} V_{\epsilon}(0, J) = J. \text{Consider the} n \text{-stage cost function of} \pi \text{, denoted} J_{\pi,n} \text{, for stages} n = N_{1}, N_{2}, \ldots, \text{where} N_{i} = \sum_{m=1}^{i} n_{m}. \text{Using the inequality (5.18) for each segment} \pi^{(m)} \text{of the policy} \pi \text{, we have} \]

\[ J_{\pi,N_{i}} \leq V_{\epsilon}(N_{i}, J_{k}) \leq V_{\epsilon}(N_{i-1}, J_{k}) + \delta_{i} \leq \cdots \leq V_{\epsilon}(0, J_{k}) + \sum_{m=1}^{i} \delta_{m}. \]

Since \( J_{\pi,N_{i}} \rightarrow J_{\pi} \) as the sequence \( \{ N_{i} \} \) of stages increases to infinity, it follows that \( J_{\pi} \leq J_{k} + \sum_{m} \delta_{m} \leq J^{*} + \epsilon_{k} + \epsilon' \).
6.1 Upper Semicontinuous Models

We consider the upper semicontinuous model as defined in [8, 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 $q(dx' | x,u)$ is continuous.
(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 $J = 0$ for (N) and with any bounded u.s.c. function $J$ for (D), value iteration generates u.s.c. functions $T^k(J)$ converging to $J^*$. There exists an $\epsilon$-optimal, nonrandomized Borel measurable policy which is stationary under (D) and semi-Markov under (N). (For these optimality results, see [8, Props. 8.7, 9.21].)

Consider the mixed value and policy iteration algorithm (3.8)-(3.9). By a selection theorem for u.s.c. functions [8, Prop. 7.34], if $Q : \Gamma \to [-\infty, \infty]$ is u.s.c., then the function resulting from partial minimization,

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

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

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

We are interested in applying the above selection theorem in place of the one given in Eqs. (3.10)-(3.11) to derive policy iteration-like algorithms that are similar to Algorithm II in Section 3.2 but operate within the family of u.s.c. functions. One way to achieve this is to choose, at each iteration of (3.8)-(3.9), for a given stationary Borel measurable policy $\mu$, an appropriate set $B \subset S$ to form the parameter $\theta = (\mu, B)$ in the mapping $F_\theta$, as the following result shows.

**Proposition 6.1 (Upper Semicontinuous Models).** Let $\theta = (\mu, B)$ for an open subset $B$ of $S$ and a Borel measurable stationary policy $\mu$ such that $\mu(du | \cdot)$ 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

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

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

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

13That $g$ being u.s.c. on $\Gamma$ means that $g$ is u.s.c. on the space $\Gamma$ endowed with the relative topology—relative to the topology on $S \times C$.

14As discussed at the end of Section 2.4 in the upper (or lower) semicontinuous model we consider, standard policy iteration also encounters measurability difficulty. Because under the model conditions here, a Borel measurable policy need not have an u.s.c. (or l.s.c.) cost function, so at the policy improvement step, we cannot apply the selection theorem just mentioned (or the one in the next subsection) to ensure that policy iteration can avoid the measurability difficulty discussed in Section 2.4 by generating only Borel measurable policies.
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 \( \Gamma \). Note that since \( \Gamma \) 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 \( \phi(x') \), since \( \mu \) satisfies the control constraint. Then since the function \( x \mapsto \mu(du \mid x) \) is continuous on \( B \), we can apply \( \text{[8 Prop. 7.31(b)]} \) to conclude that \( \phi(x') \) is u.s.c. on \( B \).

Denote \( \psi(x') = \phi(x') \cdot 1_B(x') + J(x') \cdot 1_S \cdot B(x') \) 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 \( \phi \), we have \( \phi(x') \leq J(x') \) for all \( x' \in S \). Therefore, in the case \( \bar{x} \not\in B \), 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 \( J \). In the other case \( \bar{x} \in B \), since \( B \) is open, \( x_n \in B \) for \( n \) sufficiently large, so 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 \( \phi \) restricted to \( B \) is u.s.c., as we proved earlier. This proves that the function \( \psi \) is u.s.c. Clearly \( \psi \) is bounded above. Then, using also the model assumption that the stochastic kernel \( q(dx \mid x, u) \) is continuous, we have, by \( \text{[8 Prop. 7.31(b)]} \), that the integral (6.2) as a function of \( (x, u) \) is u.s.c. and clearly bounded above. This proves the proposition.

Remark 6.1. In the preceding proposition, we need an open set \( B \subset S \) such that restricted to \( B \), the function \( x \mapsto \mu(du \mid x) \) is continuous. The empty set \( B = \emptyset \) trivially fulfills this requirement, but then \( F_\theta(Q; J) \) does not depend on the policy \( \mu \) at all, and this does not serve our purpose of using policy iteration-like algorithms as an alternative to value iteration. As emphasized earlier, it is desirable to have a “large” set \( B \). Lusin’s theorem [21 Thm. 7.5.2] tells us that there exists a closed set \( \bar{B} \subset S \) restricted to which, the function \( \mu(du \mid \cdot) \) is continuous, where \( \bar{B} \) can be chosen to be very “large,” with its measure arbitrarily close to 1 for any given Borel probability measure on \( S \). We can let \( B = \text{int}(\bar{B}) \), the interior of \( \bar{B} \); however, despite the “large” size of \( \bar{B} \), \( \text{int}(\bar{B}) \) may still be empty. Because of this, the preceding proposition is not as powerful as its counterpart, the subsequent Prop. 6.3, for the lower semicontinuous model to be discussed shortly, where Lusin’s theorem can be directly applied to yield a large set \( \bar{B} \) with the desired continuity property.

Based on Prop. 6.1, we see that to keep the iterates \( J_k, Q_k \) of the basic algorithm (3.8)-(3.9) 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 \( \theta_k = (\mu_k, B_k) \) to satisfy the condition in Prop. 6.1 and update \( Q_{k+1} \) using always the first rule in (3.8), 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 (3.8), \( Q_{k+1} = Q_{\theta_k} \cdot J_k \), also makes \( Q_{k+1} \) u.s.c. and therefore can be used. (For case (P), however, we do not know if \( Q_{\theta_k} \cdot J_k \) is u.s.c. in general.) Then, combined with the selection of \( \mu_k \) based on Eq. (6.1) for \( Q = Q_k \), we obtain, similar to Algorithm II in Section 3.2, policy iteration-like algorithms that operate with Borel measurable policies \( \{\mu_k\} \) and u.s.c. functions \( \{(J_k, Q_k)\} \).

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 6.1 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 \( \epsilon \)-optimal, Borel measurable nonrandomized Markov policy.
Proof. Let \( a = \sup_{x \in S} J^*(x) < \infty \). 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 \([8, \text{Props. 7.31, 7.34}]\). We also have \( J^* \leq J \leq c J^* \) for \( c \geq \max\{1, a/\delta\} \), so by Theorem \([5.1 b]\), \( T^k(J) \rightarrow J^* \). 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.\(^{15}\) The assertion of the existence of \( \epsilon \)-optimal, Borel measurable Markov policy then follows from a selection theorem for u.s.c. functions (\([8, \text{Prop. 7.34}]\); cf. Eq. (6.1)) and the same proof argument as that for \([8, \text{Prop. 9.19(P)}]\). \( \square \)

### 6.2 Lower Semicontinuous Models

We now consider the lower semicontinuous model from \([8, \text{Def. 8.7}]\). 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 \( \Gamma \) is a closed subset of \( S \times C \).

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

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

This is a special case of the model defined in \([8, \text{Def. 8.7}]\), 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 \([24]\).

In principle, the approach we describe here is applicable to these models as well to address the measurability issues in standard policy iteration (cf. the discussion at the end of Section 2.4 and Footnote 14), 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 \( J = 0 \) 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 \([8, \text{Props. 8.6 and Cor. 9.17.2}]\).)

Consider the mixed value and policy iteration algorithm \((3.8)-(3.9)\). In what follows, we apply arguments similar to those for the upper semicontinuous model, and we obtain policy iteration-like algorithms that keep iterates \( (J_k, Q_k) \) within the set of l.s.c. functions. Specifically, by a selection theorem for l.s.c. functions \([8, \text{Prop. 7.33}]\), we have that if \( Q : \Gamma \rightarrow [-\infty, \infty] \) 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 \( \mu \) such that

\[
Q(x, \mu(x)) = M(Q)(x), \quad x \in S. \tag{6.4}
\]

Thus, if at the \( k \)th iteration of the algorithm \((3.8)-(3.9)\), the function \( Q_k \) is l.s.c., then we can choose a Borel measurable policy \( \mu_k \) according to the above selection theorem with \( Q = Q_k \), similar to the policy iteration-like Algorithm II in Section 3.2. In order for the subsequent iterates \( Q_{k+1}, J_{k+1} \) to be l.s.c., we can choose an appropriate set \( B_k \) based on the following result, when forming the parameter \( \theta_k = (\mu_k, B_k) \) for the mapping \( F_{\theta_k} \) in the algorithm.

Let \( \mu \) be a Borel measurable stationary policy. We know from Lusin’s theorem \([21, \text{Thm. 7.5.2}]\) that there exists a closed subset \( B \subset S \) such that restricted to \( B \), the function \( x \mapsto \mu(dx | x) \) is continuous, and moreover, for any given \( p \in \mathcal{P}(S) \), the set \( B \) can be chosen to have \( p(B) \) arbitrarily close to 1.

\(^{15}\)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.
Proposition 6.3 (Lower Semicontinuous Models). Let \( \theta = (\mu, B) \) for a closed subset \( B \) of \( S \) and a Borel measurable stationary policy \( \mu \) 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 \( \Gamma \). 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 \( \Gamma \). 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 \( \Gamma \) to be \( \infty \), and we can view the integral defining \( \phi(x') \) as the integral of this extension. This will not change the value \( \phi(x') \), since \( \mu \) satisfies the control constraint. Then, since the function \( x \mapsto \mu(du | x) \) is continuous on \( B \), we can apply [8] Prop. 7.31(a) to conclude that \( \phi(x') \) is l.s.c. and bounded below on \( B \).

Denote \( \psi(x') = \phi(x') \cdot 1_B(x') + J(x') \cdot 1_{S \backslash B}(x') \) 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 \( S \backslash B \) is open, we have \( x_n \in S \backslash B \) for all \( n \) sufficiently large, so

\[
\lim_{n \to \infty} \inf_{i \to \infty} \psi(x_n) = \lim_{n \to \infty} \inf_{i \to \infty} J(x_n) \geq J(\bar{x}) = \psi(\bar{x}),
\]

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 \( \lim_{i \to \infty} \inf_{n \to \infty} \psi(x_n) = \lim_{i \to \infty} \psi(x_{n_i}) \) 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

\[
\lim_{n \to \infty} \inf_{i \to \infty} \psi(x_{n_i}) = \lim_{i \to \infty} \inf_{n \to \infty} J(x_{n_i}) \geq J(\bar{x}) = \psi(\bar{x}),
\]

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

\[
\lim_{n \to \infty} \inf_{i \to \infty} \psi(x_{n_i}) = \lim_{i \to \infty} \psi(x_{n_i}) = \lim_{i \to \infty} J(x_{n_i}) \geq J(\bar{x}) = \psi(\bar{x}),
\]

where the first inequality holds since \( J \) is l.s.c., and the second inequality holds since by the definition of \( \phi \), we have \( \phi(x') \leq J(x') \) for all \( x' \in S \). Thus we have proved that the function \( \psi \) is l.s.c. Clearly \( \psi \) is bounded below. Then, using also the fact that the stochastic kernel \( q(dx'|x,u) \) is continuous, we have, by [8] 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, to obtain policy iteration-like algorithms that keep the iterates \( J_k,Q_k \) within the set of functions that are l.s.c. and bounded below, we can specialize the basic algorithm (3.8)–(3.9) as follows: Start with \( J_0,Q_0 \) that are l.s.c. and bounded below. For \( k \geq 0 \), form the parameter \( \theta_k = (\mu_k,B_k) \) by first choosing a Borel measurable \( \mu_k \) based on the selection theorem (6.4) with \( Q = Q_{k-1} \), and next choosing a closed set \( B_k \) based on Lusin’s theorem as described earlier. Then update \( Q_{k+1} \) using the first rule in (3.8). This results in l.s.c. functions \( Q_{k+1} \) and \( J_{k+1} \). For cases (D)(P), the second rule in (3.8), \( Q_{k+1} = Q_{\theta_k,J_k} \), can also be shown to make \( Q_{k+1} \) l.s.c. and therefore can be used. For case (N), however, we do not know if \( Q_{\theta_k,J_k} \) 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.

Besides its ability to handle measurability or non-measurability related structural restrictions in stochastic DP problems, our mixed value and policy iteration approach has attractive computational features. We have discussed them in Sections 1 and 3.2 here we emphasize again two specific advantages. The approach yields policy iteration-like algorithms that have comparable computation complexity to standard policy iteration (which is usually much more efficient than value iteration) and yet have much more reliable convergence properties. It works naturally with approximate policy improvement, and is capable of solving difficult problems that cannot be handled by standard policy iteration, such as problems whose optimal or near-optimal policies are non-stationary.

A number of issues are still not fully understood and deserve further research, however. One important question is how to choose the parameters \((\mu, B)\), especially the set \(B\), for the policy iteration-like algorithms. We have discussed in the paper two possible choices of \(B\) for a given policy \(\mu\). It is still unclear how to choose \(B\) to facilitate fast convergence of the algorithms. It is also desirable to be able to find such sets \(B\) easily so that the computation involved does not tamper with the overall efficiency of the method.

Another question relates to finding near-optimal policies, especially in problems with non-stationary near-optimal policies. Our approach yields function sequences that converge pointwise to the optimal cost function for discounted, nonpositive and nonnegative cost models. It also yields asymptotically optimal policy sequences for discounted cost models. For nonpositive and nonnegative cost models, it seems plausible that we can extract nearly optimal policies from the data produced by the algorithms, but the details still need to be worked out.

For nonnegative cost models, we have derived a new sufficient condition for convergence of value iteration, which involves 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 also 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, however. Moreover, in the literature on average cost problems, it has been shown that one can speed up the convergence of (relative) value iteration significantly, by choosing an initial function that qualitatively matches the shape of the optimal (relative) value function (see Chen and Meyn [20] and Meyn [36, Example 9.5.7]). Whether the same can be said for total cost problems is another interesting open question, which we aim to address in the future.

We conclude the paper with a discussion of 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 [11, 61, 10]. We discuss the subject briefly here, focusing on issues related to universal measurability in a simplified setting.

Suppose that instead of the basic algorithm (3.8)-(3.9), at each iteration \(k\), we only compute \(Q_{k+1}(x, u)\) for a subset \(\Gamma_k\) of state-control pairs in \(\Gamma\) and compute \(J_{k+1}(x)\) for a subset \(S_k\) of states in \(S\). (For the rest of the 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

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

is lower semi-analytic by [8, Lemma 7.30(4)], because $1_{R \cap \Gamma}(x, u)$ and $1_{(S \times C \setminus R) \cap \Gamma}(x, u)$ are nonnegative Borel measurable functions on $\Gamma$. 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 $J \cdot 1_D + J' \cdot 1_{S \setminus D}$ is lower semi-analytic.

Further studies of such asynchronous algorithms, in particular, stochastic asynchronous Q-learning versions (similar to those considered in [57, 52, 53, 19, 1, 11, 61, 60]), are important subjects for future investigation.

**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 partition the state-control space $S \times C$.

For a stationary universally measurable policy $\mu$, let $R \subset S \times C$ be a Borel set such that $B = \text{proj}_S(R)$ is Borel and the function $x \mapsto \mu(du \mid x)$ is Borel measurable on $B$. For any such pair $\theta = (\mu, R)$, we may consider a mapping $F_\theta$ defined as: for all $J \in A(S), Q \in A(\Gamma)$,

$$F_\theta(Q; J)(x, u) = g(x, u) + \alpha \int_S \int_C 1_{S \times C \setminus R}(x', u') \cdot J(x') \mu(du' \mid x') \, q(dx' \mid x, u)$$

$$+ \alpha \int_S \int_C 1_R(x', u') \cdot \min \{ J(x'), Q(x', u') \} \mu(du' \mid x') \, q(dx' \mid x, u), \quad (x, u) \in \Gamma.$$  

It can be equivalently written as

$$F_\theta(Q; J)(x, u) = g(x, u) + \alpha \int_{S \setminus B} J(x') \, q(dx' \mid x, u) + \alpha \int_B J(x') \cdot \mu(C \setminus R_{x'} \mid x') \, q(dx' \mid x, u)$$

$$+ \alpha \int_B \int_{R_{x'}} \min \{ J(x'), Q(x', u') \} \mu(du' \mid x') \, q(dx' \mid x, u), \quad (x, u) \in \Gamma, \quad (7.1)$$

where $B = \text{proj}_S(R)$ and $R_x = \{ u \in C \mid (x, u) \in R \}$ is the vertical section of $R$ at $x$. That the function $F_\theta(Q; J)$ 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 [8, Cor. 7.26.1] and hence the term $\int_B J(x') \cdot \mu(C \setminus R_{x'} \mid x') \, q(dx' \mid x, u)$ in (7.1) as a function of $(x, u)$ is lower semi-analytic.

Let us also mention that the algorithms we gave earlier all use the formula $J_{k+1} = M(Q_{k+1})$ to compute the cost function iterates. Suppose in a given problem one knows that $J \leq J^* \leq J$ for certain $J, J' \in A(S)$. Then it is natural and more efficient to constrain $J_{k+1}$ to be within these upper and lower bounds of the optimal cost function by setting $J_{k+1} = \max \{ J, \min \{ J, M(Q_{k+1}) \} \}$.

**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 partially observable problems, since they can be reduced to equivalent completely observable problems [8, Chap. 10]. These also 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 such models are worthy of further study. Also among the important subjects for future research are extensions to average cost problems and constrained 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 3.2. We also thank Prof. Eugene Feinberg, with whom our recent correspondence about Borel models stimulated this research, and Prof. Sean Meyn, who pointed us to related work in his book [36], which motivated the alternative proof for Theorem 5.1 we gave in Appendix E. We appreciate Prof. Sanjoy Mitter's helpful feedback on our early draft, and we owe many thanks to the anonymous reviewers for their critical and suggestive comments about our manuscript, including the suggestion of Example 3.1.

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] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM J. Control Opt.*, 31:282–344, 1993.

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

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

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

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

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

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

[10] 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.

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

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

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

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

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

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

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

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

[19] 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.
[20] R. R. Chen and S. Meyn. Value iteration and optimization of multiclass queueing networks. *Queueing Systems*, 32:65–97, 1999.

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

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

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

[24] 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.

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

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

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

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

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

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

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

[32] 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.

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

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

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

[36] S. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, Cambridge, 2008.

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

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

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

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

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

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

[43] 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.

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

[45] 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.

[46] S. E. Shreve. Borel-approachable functions. *Fundamenta Mathematicae*, 112:17–24, 1981.
[47] 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.

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

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

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

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

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

[53] 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.

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

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

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

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

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

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

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

[61] H. Yu and D. P. Bertsekas. Q-learning and policy iteration algorithms for stochastic shortest path problems. *Ann. Oper. Res.*, 208:95–132, 2013.
Appendices

A Optimal Stopping Problems Associated with $F_\theta$

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 $Q_{\theta,J} = \lim_{k \to \infty} F^{k}_{\theta}(0; J)$, and we show that the mapping $F_{\theta}(\cdot; J)$ can be viewed as a form of the optimal cost operator 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 $J \in A_{\theta}(S)$ 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, whereas the policy $\mu$ will be used to define the dynamics of the unstopped process.

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

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

\[ U^0((x,u)) = \{0\} \text{ on } (S \setminus B) \times C. \]

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

\[
\begin{align*}
g^0((x,u), 0) &= J(x) \quad \forall (x,u) \in S \times C, \\
g^0((x,u), 1) &= g(x,u) \quad \forall (x,u) \in (B \times C) \cap \Gamma, \\
g^0((x,u), 1) &= K \quad \forall (x,u) \in (B \times C) \setminus \Gamma,
\end{align*}
\]

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

- State transition stochastic kernel $q^0(\cdot; \cdot)$ on $S^0 \times C^0$: for any Borel set $D \subset S^0$ and any $(x,u) \in S \times C$, $q^0(D|\infty,0) = q^0(D|\infty,1) = q^0(D|(x,u), 0) = \delta_{\infty}(D)$, and

\[
q^0(D|(x,u), 1) = \int_{S} \int_{C} 1_{D\setminus\{\infty\}}((x',u')) \tilde{\mu}(du'|x') q(dx'|x,u),
\]

where $\tilde{\mu}(du'|x')$ is a Borel measurable stochastic kernel on $C$ given $S$ such that $\tilde{\mu}(du'|x') = \mu(du'|x')$ for all $x' \in B$.

The state transition kernel required in the above can be chosen, for instance, by letting $\tilde{\mu}(du'|x') = \mu(du'|x')$ for $x' \in B$ and $\tilde{\mu}(du'|x') = p(du')$ for $x' \notin B$, where $p$ is any Borel probability measure on $C$. Note also that with the control 1, for any $(x,u) \in S \times C$ and Borel $D \subset S^0$, we have

\[
q^0(D|(x,u), 1) = \int_{B} \int_{C} 1_{D\setminus\{\infty\}}((x',u')) \mu(du'|x') q(dx'|x,u) \\
+ \int_{S \setminus B} \int_{C} 1_{D\setminus\{\infty\}}((x',u')) \tilde{\mu}(du'|x') q(dx'|x,u).
\]
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 $\pi^o$ 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 $\tau$ be the time when the process is stopped: $\tau = \min \{k \geq 0 \mid u_k^o = 0\}$ with $\tau = \infty$ if $\{k \geq 0 \mid u_k^o = 0\} = \emptyset$. 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 $z_0 = (x, u) \in S \times C$, $V_{\pi^o}(x, u)$ can be expressed as

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

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 $T_o$. By the theory for (D)(N)(P) in the case where the number of controls at each state is finite [8 Props. 9.8, 9.14, Cor. 9.17.1], we have

$$T^o_k(0) \to V^*, \quad V^* = T_o(V^*),$$

where the convergence is monotonic for (N)(P), and furthermore, for (D)(P), there exists an optimal nonrandomized stationary policy. (In case (N), an optimal policy need not exist even when the control space is finite; see [7 Ex. 4.1, p. 181] for such an example.)

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

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

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 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 $\mu$ satisfies the control constraint of the original problem, we see from the expression (A.1) for the state transition probability $q^o(\cdot \mid x, u, 1)$ that the probability of the next 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 next state is $\infty$. 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 expected total cost of $\pi^o$ for the initial distribution $p^o(\cdot) = q^o(\cdot \mid x, u, 1)$, with respect to the probability measure induced by $\pi^o$ and $p^o$, is

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

(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 $F_{\theta}(\cdot; J)$ and $Q_{\theta, J}$

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

$$\{ V \in A(S^o) \mid V(\infty) = 0, V(x, u) = J(x), (x, u) \in (S \setminus B) \times C \}, \tag{A.5}$$
which includes $V^*, V_k$ (cf. Eq. (A.3)). For any $V$ in this set, we can express $T_\alpha(V)$ as

$$T_\alpha(V)(x, u) = \min \{ J(x), G_V(x, u) \}, \quad (x, u) \in (B \times C) \cap \Gamma,$$

(A.6)

where the first term $J(x)$ is the stopping cost, and the second term $G_V(x, u)$ is associated with the continuation action, given by

$$G_V(x, u) = g(x, u) + \alpha \int_{S \times C} V(z) \, q^\alpha(\,dz \mid (x, u), 1)$$

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

(A.7)

In the above, to obtain Eq. (A.7) from the first equality, we used the expression of $q^\alpha(\,dz \mid (x, u), 1)$ given in (A.1), which implies that for any $(x, u) \in S \times C$ and function $V$ in the set (A.5),

$$\int_{S \times C} V(z) \, q^\alpha(\,dz \mid (x, u), 1) = \int_{S \setminus B} J(x') \, q(dx' \mid x, u) + \int_B \int_C V(x', u') \, \mu(du' \mid x') \, q(dx' \mid x, u),$$

(A.8)

Equations (A.6)-(A.7) express the optimality equation $V = T_\alpha(V)$ for $V$ in the set (A.5).

Using the fact $V^* = T_\alpha(V^*)$, we then obtain

$$V^*(x, u) = T_\alpha(V^*)(x, u) = \min \{ J(x), f^*(x, u) \}, \quad \forall (x, u) \in (B \times C) \cap \Gamma,$$

(A.9)

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

$$f^*(x, u) = g(x, u) + \alpha \int_{S \times C} V^*(z) \, q^\alpha(\,dz \mid (x, u), 1)$$

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

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

(A.10)

(A.11)

Here in deriving Eq. (A.11), we used the fact that for any $x' \in B$,

$$\int_C V^*(x', u') \, \mu(du' \mid x') = \int_C \min \{ J(x'), f^*(x', u') \} \, \mu(du' \mid x'),$$

(A.12)

which holds since $\mu$ satisfies the control constraint of the original problem: $\mu(U(x') \mid x') = 1$, 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

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

(A.13)

From the recursive relations,

$$V_{k+1}(x, u) = T_\alpha(V_k)(x, u) = \min \{ J(x), f_k(x, u) \}, \quad (x, u) \in (B \times C) \cap \Gamma, \quad k \geq 0,$$

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. (5.4)). 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 $F_\theta(f_c;J)$ where $f_c$ is an (arbitrary) extension of $f$ to $\Gamma$. 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

$$Q|_{\Gamma_B} = Q'|_{\Gamma_B} \implies F_\theta(Q;J) = F_\theta(Q';J).$$  \hfill (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 $Q_{\theta,J} = \lim_{k \to \infty} F_\theta^k(0;J)$ as follows.

**Lemma A.1.** (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

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

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

$$Q_{\theta,J}|_{\Gamma_B} = f^*, \quad Q_{\theta,J} = F_\theta(f^*;J).$$ \hfill (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.

We now prove the relation (A.15) between $Q_{\theta,J}$ and $f^*$. Since $f_k \to f^*$, using the relation $f_k = F_\theta(f_{k-1};J)|_{\Gamma_B}$ and Eq. (A.14), we have $f_k = F_\theta^k(g;J)|_{\Gamma_B} \to f^*$. Suppose we have proved $F_\theta^k(g;J) \to Q_{\theta,J}$. Then it will follow that $Q_{\theta,J}|_{\Gamma_B} = f^*$. 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 follows from the proof of Prop. 3.2 (D), which established the contraction property of $F_\theta(\cdot;J)$. For (N), we have $g \leq 0$ and $J \leq 0$. By a direct calculation, $F_\theta(0;J) \leq g \leq 0$, so we have, by the monotonicity of $F_\theta(\cdot;J)$,

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

Since $F_\theta^k(0;J) \downarrow Q_{\theta,J}$ 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.

Based on Lemma A.1, we may view $F_\theta(\cdot;J)$ as an optimal cost operator for the minimal future costs associated with the continuation action in the optimal stopping problem. For states $(x,u) \in \Gamma_B$, based on the relation $Q_{\theta,J}|_{\Gamma_B} = f^*$ in (A.15), 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 $Q_{\theta,J}$ in terms of $V^*$ in the following corollary. For each $(x,u) \in \Gamma$, we will consider the optimal stopping problem starting with an initial state distribution $p^\circ$ given by $q^\circ(\cdot | (x,u), 1)$, the transition distribution for $(x,u)$ under the continuation action.

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

$$Q_{\theta,J}(x,u) = g(x,u) + \alpha \int_{S \times C} V^*(z) q^\circ(dz | (x,u), 1).$$
In particular, if in the optimal stopping problem associated with \((\theta, J)\), an optimal policy \(\pi^{o*}\) exists (as is true under (D)(P)), then for all \((x, u) \in \Gamma\),

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

where \(\tau = \min\{k \geq 0 \mid u^*_k = 0\}\) with \(\tau = \infty\) if this set is empty, and the expectation is with respect to the probability measure induced by \(\pi^{o*}\) 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.1), using the definition of \(F_\theta(\cdot; J)\) and Eq. (A.12), we have that for all \((x, u) \in \Gamma\),

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

which together with (A.8) implies the first expression for \(Q_{\theta,J}(x, u)\) in the corollary. From this expression and Eq. (A.4) for policy \(\pi^{o*}\), we obtain the second expression for \(Q_{\theta,J}\) in the corollary. \(\square\)

### 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. For case (D) (resp. case (N)), the function \(V^*\) is the maximal solution to \(V \leq T_\theta(V)\) among the set of bounded lower semi-analytic functions (resp. the set of nonpositive lower semi-analytic functions) \([8, \text{Prop. 9.10}(P)]\). The inequality \(V \leq T_\theta(V)\) can be expressed as a system of linear inequalities, so under suitable conditions, \(V^*\) can be obtained by solving a linear program. (See [29, 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 \(V \geq T_\theta(V)\) \([8, \text{Prop. 9.10}(P)]\), and this in general does not admit a linear programming formulation. We consider below a linear program with linear constraints based on the inequality \(V \leq T_\theta(V)\) 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.2). 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 \(\mathcal{U}\) denote the universal \(\sigma\)-algebra on \(S \times C\).

**Assumption A.1.** (P) There exists a \(\sigma\)-finite measure \(\rho\) on \((S \times C, \mathcal{U})\) such that

(i) \(\int_{\Gamma_B} J(x) \rho(dx) < \infty;\) and

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

\[
\rho_{x,u}(D) = \int_B \int_C \mathbf{1}_D(x', u') \mu(du' \mid x') q(dx' \mid x, u), \quad D \in \mathcal{U},
\]

is absolutely continuous with respect to \(\rho\) (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 \(\Gamma_B\). 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 \( A_+(\Gamma_B) \):

\[
\begin{align*}
\text{Maximize} & \quad \int_{\Gamma_B} V(x, u) \rho(d(x, u)) \\
\text{Subject to:} & \quad V(x, u) \leq J(x), \quad \forall (x, u) \in \Gamma_{B,\rho}, \\
& \quad V(x, u) \leq g(x, u) + \int_{S \setminus B} J(x') q(dx' \mid x, u) \\
& \quad \quad \quad + \int_B \int_C V(x', u') \mu(du' \mid x') q(dx' \mid x, u), \quad \forall (x, u) \in \Gamma_{B,\rho}.
\end{align*}
\]

(A.17)

As can be seen from the expression (A.6)-(A.7) for the operator \( T_\theta \), this linear program corresponds to the following maximization problem:

\[
\begin{align*}
\text{Maximize} & \quad \int_{\Gamma_B} V(x, u) \rho(d(x, u)) \\
\text{Subject to:} & \quad V(x, u) \leq T_\theta(V^c)(x, u), \quad \rho\text{-almost every } (x, u) \in \Gamma_B,
\end{align*}
\]

where \( V^c \) is any extension of \( V \) on \( S^0 \) satisfying Eq. (A.5) (i.e., \( V^c(\infty) = 0 \), \( V^c(x, u) = J(x) \) for \( (x, u) \in (S \setminus B) \times C \)).

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 \( \Gamma \):

\[
\bar{Q}(x, u) = g(x, u) + \int_{S \setminus B} J(x') q(dx' \mid x, u) + \int_B \int_C \bar{V}(x', u') \mu(du' \mid x') q(dx' \mid x, u).
\]

(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.2.** (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

\[
\bar{Q} \leq F_\theta(\bar{Q} ; J), \quad \bar{Q} \geq Q_{\theta, J}.
\]

**Proof.** Since \( V^* = T_\theta(V^*) \), the optimal cost function \( V^* \) restricted to \( \Gamma_B \) is a feasible solution of (A.17), so the feasible set of (A.17) is nonempty. By Assumption [A.1i], the optimal objective value \( v^* \) of (A.17) is finite. Let \( V_n, n \geq 1 \), be a sequence of feasible solutions with their objective values approaching \( v^* \). Then the function resulting from taking pointwise supremum, \( \sup_n V_n \), lies in \( A_+(\Gamma_B) \) (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 \( \Gamma_B \) is then an optimal solution of (A.17) as well. This implies that

\[
V^*(x, u) \leq \bar{V}(x, u) \quad \text{for } \rho\text{-almost every } (x, u) \in \Gamma_B,
\]

(A.19)

for otherwise, by Assumption [A.1i] we would have

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

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 \( V \). This, together with Assumption [A.1i] and the relation (A.19), implies \( Q_{\theta, J} \leq \bar{Q} \). To show \( \bar{Q} \leq F_\theta(\bar{Q} ; J) \), notice that by the feasibility of \( \bar{V} \) for (A.17) and the definition of \( Q \),

\[
\bar{V}(x, u) \leq \min \{J(x), \bar{Q}(x, u)\}, \quad \forall (x, u) \in \Gamma_{B,\rho}.
\]
We use this relation to upper-bound $\bar{V}$ $\rho$-almost everywhere on $\Gamma_B$, 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 $(x,u) \in \Gamma$,

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

which is the inequality $\bar{Q} \leq F_0(\bar{Q} ; J)$. 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 $\rho$ in Assumption A.1. Then Assumption A.1 is satisfied by letting $\rho(x,u) = 2^{-x}/(J(x)+1)$ if $(x,u) \in \Gamma_B$, and $\rho(x,u) = 0$ otherwise, for instance.

If $\mu$ 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:

Maximize $\sum_{x \in B} W(x) \rho(x,\mu(x))$

Subject to: $W(x) \leq J(x), \forall x \in B,$

$W(x) \leq g(x,\mu(x)) + \sum_{x' \in S \times B} J(x') q(x' | x,\mu(x)) + \sum_{x' \in B} W(x') q(x' | x,\mu(x)), \forall x \in B.$

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 $Q_{\theta,J^*} = \lim_{k \to \infty} F_k^\theta(0; J^*)$ is $Q^*$ given in Eq. (3.1). This establishes Prop. 3.3(d) for (P), which is also used in the lower bound part of Lemma 4.3 for (P).

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

Since $Q^* \geq 0$ and $F_0(Q^* ; J^*) = Q^*$ (Prop. 3.3(a)), we have by the monotonicity of $F_0$ (cf. Eq. (3.6)),

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

(B.1)

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 $\epsilon > 0$,

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

(B.2)

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 $Q_{\theta,J^*}$ to be used in the proof. Fix $(\check{x}, \check{u}) \in \Gamma$. For the optimal stopping problem associated with $\theta = (\mu,B)$ and $J^*$, by [3] Cor. 9.17.1, there exists an optimal stationary nonrandomized (universally measurable) policy $\mu^* : S^0 = (S \times C) \cup \{\infty\} \to \{0, 1\}$. Let the optimal stopping problem start from time 1, and consider the stochastic process $(z_1, u_1^0, z_2, u_2^0, \ldots)$, where $z_k \in S^0$ and $u_k^0 \in [0, 1]$, induced by $\mu^*$ and the initial distribution of $z_1$ given by $q^0(\cdot \mid \{\check{x}, \check{u}\})$ (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

\[
Q_{θ,J∗}(\bar{x}, \bar{u}) = g(\bar{x}, \bar{u}) + \mathbb{E}^{µ^o}\left\{\sum_{k=1}^{τ-1} g(x_k, v_k) + J^∗(x_τ)\right\},
\]

(B.3)

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

To simplify notation, let

\[
D = \{(x, v) \in S \times C \mid µ^o(x, v) = 0\}, \quad D_x = \{v \in C \mid (x, v) \in D\}, \quad x \in S.
\]

(\( D \) 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 [8, Lemma 7.29]. Note that expressed in terms of these sets, \( τ = \min \{ k \geq 1 \mid v_k \in D_{x_k} \} \) and

\[
τ = m \iff v_1 \notin D_{x_1}, \ldots, v_{m-1} \notin D_{x_{m-1}}, v_m \in D_{x_m},
\]

(B.4)

\[
τ > m \iff v_1 \notin D_{x_1}, \ldots, v_{m-1} \notin D_{x_{m-1}}, v_m \notin D_{x_m}.
\]

(B.5)

We will then give the proof of this lemma.

**Lemma B.1.** We have

\[
\mathbb{E}^{µ^o}\left\{\sum_{k=1}^{τ-1} g(x_k, v_k) + J^∗(x_τ)\right\} = \mathbb{E}^{µ}\left\{\sum_{k=1}^{τ-1} g(x_k, v_k) + J^∗(x_τ)\right\}.
\]

We first prove Prop. B.1 using Lemma B.1. We will then give the proof of this lemma.

**Proof of Prop. B.1.** Fix \((\bar{x}, \bar{u}) \in Γ\) and let \( ε > 0 \). Let \( π^∗ = (π^∗_0, π^∗_1, \ldots) \in Π \) be an \( ε \)-optimal Markov policy for the original control problem; such a policy exists by [8, Prop. 9.19]. We use the policies \( µ^o, π^∗, \) and \( µ \) (the stationary policy defining \( F_θ \) and the associated optimal stopping problem) to define a controller \( \hat{π} \) 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 \( Q_{θ,J∗}(\bar{x}, \bar{u}) + ε \). From this the desired inequality [B.2] for establishing the proposition will be shown to follow.

Roughly speaking, the controller \( \hat{π} \) follows the policy \( µ \) before it switches to following policy \( π^∗ \). 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 \( \hat{π} = (\hat{µ}_0, \hat{µ}_1, \ldots) \), where each \( \hat{µ}_k \) is a universally measurable stochastic kernel on \( C \) given the respective space of histories. We now define \( \hat{µ}_k, k \geq 0 \).

For \( k = 0 \), let \( \hat{µ}_0 \) be a universally measurable stochastic kernel on \( C \) given \( S \), such that \( \hat{µ}_0 \) satisfies the control constraint \( U \) and \( \hat{µ}_0(dv_0 | \bar{x}) = δ_{\bar{u}} \). For each \( k \geq 1 \), the auxiliary variable \( v_k \) is generated according to the stochastic kernel \( µ \) given the state \( x_k \):

\[
µ(dv_k | x_k).
\]

(B.6)
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

$$\tau_k(x_1, v_1, x_2, v_2, \ldots, x_k, v_k) = \begin{cases} \min \{m | v_m \in D_{x_m}, 1 \leq m \leq k\} & \text{if such } m \text{ exists,} \\ 0 & \text{otherwise.} \end{cases} \quad (B.7)$$

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

$$\hat{\mu}_k(du \mid x_0, u_0, x_1, u_1, \ldots, x_k, v_k) = \begin{cases} \delta_{v_k} & \text{if } \tau_k(x_1, v_1, \ldots, x_k, v_k) = 0, \\ \pi_k^{-m}(du \mid x_k) & \text{if } \tau_k(x_1, v_1, \ldots, x_k, v_k) = m \geq 1. \end{cases} \quad (B.8)$$

(Thus, $\hat{\pi}$ “copies” the control $v_k$ if it has not yet switched to applying policy $\pi^*$, 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 $\sigma$-algebra). With respect to this probability, the expected total cost of $\hat{\pi}$ for state $\bar{x}$ is

$$\hat{J}_\pi(\bar{x}) = g(\bar{x}, \bar{u}) + \mathbb{E}^\pi \left\{ \sum_{k=1}^{\infty} g(x_k, u_k) \right\}. \quad (B.9)$$

Let $\tau = \min \{k \geq 1 \mid v_k \in D_{x_k} \}$ 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-$\sigma$-algebra associated with the stopping time $\tau$ [39] Prop. III-1-3], it follows that

$$\hat{J}_\pi(\bar{x}) = g(\bar{x}, \bar{u}) + \mathbb{E}^\pi \left\{ \sum_{k=1}^{\tau-1} g(x_k, u_k) + J^*(x_\tau) \right\} = g(\bar{x}, \bar{u}) + \mathbb{E}^\pi \left\{ \sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau) \right\}, \quad (B.9)$$

where the second equality follows from the fact $u_k = v_k$ for $k < \tau$ (cf. Eqs. (B.7), (B.8)). We have

$$J^*(x) \leq J^*(x) + \epsilon, \quad \forall x \in S,$$

since $\pi^*$ is an $\epsilon$-optimal policy of the original problem and $J^* \geq 0$. Then by Eq. (B.9),

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

On the other hand, based on the definition of $\hat{\pi}$, $\{v_k\}$ are generated according to $\mu$ (cf. Eq. (B.6)) and they are the controls applied before time $\tau$ (cf. Eqs. (B.7), (B.8)), so we have

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

where the second equality follows from Lemma [B.1]. Combining the preceding two relations with Eq. (B.3), we obtain

$$\hat{J}_\pi(\bar{x}) \leq g(\bar{x}, \bar{u}) + \mathbb{E}^\mu \left\{ \sum_{k=1}^{\tau-1} g(x_k, v_k) + J^*(x_\tau) \right\} + \epsilon = Q_{\theta, J^*}(\bar{x}, \bar{u}) + \epsilon. \quad (B.10)$$

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 II, in the sense that we can construct a semi-Markov
randomized policy in $I$ 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 $\bar{J}_2(\bar{x})$ for the state $\bar{x}$. (Such a construction is similar to that used to prove Props. 8.1, 9.1 in [8].) This means that $\bar{J}_2(\bar{x}) \geq Q^*(\bar{x}, \bar{u})$ (cf. Eq. (3.2)), so by Eq. (B.10),

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

Since $\epsilon$ is arbitrary, we have $Q_{\theta, J^*}(\bar{x}, \bar{u}) \geq Q^*(\bar{x}, \bar{u})$. Then by Eq. (B.11), we have $Q_{\theta, J^*} = Q^*$.

We now establish Lemma B.1.

**Proof of Lemma B.1.** To prove $E^\mu \left\{ \sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right\} = E^\mu \left\{ \sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right\}$, we first introduce some functions to rewrite the two expectations. In view of (B.4) (i.e., $\tau = m \Rightarrow v_1 \notin D_{x_1}, \ldots, v_{m-1} \notin D_{x_{m-1}}, v_m \in D_{x_m}$), we define for each $m \geq 1$, $\phi_m : (S \times C)^m \to [0, \infty]$ by

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

We define for $m = \infty$, $\phi_\infty : (S \times C)\infty \to [0, \infty]$ by

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

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

\[ E^\mu \left\{ \sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right\} = \sum_{m=1}^{\infty} E^\mu \left\{ \mathbb{1}_{\{\tau = m\}}(x_1, v_1, \ldots) \cdot \left( \sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right) \right\} \]

\[ = \sum_{m=1}^{\infty} E^\mu \left\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \right\}. \]

and similarly,

\[ E^\mu \left\{ \sum_{k=1}^{m-1} g(x_k, v_k) + J^*(x_m) \right\} = \sum_{m=1}^{\infty} E^\mu \left\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \right\}. \]

To prove that (B.11) and (B.12) are equal, we will proceed in four steps.

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

\[ E^\mu \left\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \right\} = E^\mu \left\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \right\}. \]

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 $\mu^\tau$ 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 $E^\mu$, we have that $E^\mu \left\{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \right\}$ is equal to

\[ \int_B \int_C \cdots \int_B \int_C \int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m) \mu(dv_m \mid x_m) q(dx_m \mid x_{m-1}, v_{m-1}) \cdots \mu(dv_1 \mid x_1) q(dx_1 \mid x). \]

(B.14)
Using the same facts just mentioned, and using also the definition of the optimal stopping problem (Appendix A.1), we have that
\[ \mathbb{E}^\mu \{ \phi_m(x_1, v_1, \ldots, x_m, v_m) \} \]
is equal to
\[ \int_B \int_C \cdots \int_B \int_C \left[ \int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m) \mu(dv_m | x_m) \right. \]
\[ \left. \cdot \mu(dv_{m-1} | x_{m-1}) \cdots \mu(dv_1 | x_1) q(dx_1 | \bar{x}, \bar{u}) \right]. \]

Since \( \tilde{\mu}(\cdot | x) = \mu(\cdot | x) \) for \( x \in B \) by the definition of the optimal stopping problem, the above integral in turn equals
\[ \int_B \int_C \cdots \int_B \int_C \left[ \int_S \int_C \phi_m(x_1, v_1, \ldots, x_m, v_m) \tilde{\mu}(dv_m | x_m) \right. \]
\[ \left. \cdot \tilde{\mu}(dv_{m-1} | x_{m-1}) \cdots \tilde{\mu}(dv_1 | x_1) q(dx_1 | \bar{x}, \bar{u}) \right]. \]

Consider now the inner-most integral in (B.15). If \( x_m \in S \setminus B \), then in view of the control constraint \( U^* \) 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.15) remains unchanged if we replace \( \tilde{\mu}(dv_m | x_m) \) in the inner-most integral in (B.15) by \( \mu(dv_m | x_m) \). Hence the integral (B.15) is equal to the integral (B.14), and this proves the desired equality (B.13) for \( m \geq 1 \).

(ii) By arguments similar to the ones in the preceding proof, we have that for all \( m \geq 1 \) and \( n \geq m \),
\[ \mathbb{E}^\mu \left\{ \left( \prod_{i=1}^n \mathbb{1}_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\} = \mathbb{E}^\mu \left\{ \left( \prod_{i=1}^n \mathbb{1}_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\}. \quad \text{(B.16)} \]

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

We will need (B.16) shortly in the proof.

(iii) Let us now consider the two terms corresponding to \( m = \infty \) in Eqs. (B.11) and (B.12). We examine when they are equal, i.e., when
\[ \mathbb{E}^\mu \{ \phi_{\infty}(x_1, v_1, x_2, v_2, \ldots) \} = \mathbb{E}^\mu \{ \phi_{\infty}(x_1, v_1, x_2, v_2, \ldots) \}. \quad \text{(B.17)} \]
From the definition of \( \phi_\infty \), we have, by the monotone convergence theorem, that as \( m \to \infty \),
\[
\mathbb{E}^\mu \left\{ \left( \prod_{i=1}^\infty 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\} \uparrow \mathbb{E}^\mu \{ \phi_\infty(x_1, v_1, x_2, v_2, \ldots) \},
\]
\[
\mathbb{E}^\mu \left\{ \left( \prod_{i=1}^\infty 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\} \uparrow \mathbb{E}^\mu \{ \phi_\infty(x_1, v_1, x_2, v_2, \ldots) \}.
\]
Thus Eq. (B.17) holds if for each integer \( m \geq 1 \),
\[
\mathbb{E}^\mu \left\{ \left( \prod_{i=1}^n 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\} = \mathbb{E}^\mu \left\{ \left( \prod_{i=1}^\infty 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \right\}.
\] (B.18)

Now, for each positive integer \( m \), we have the following pointwise convergence of functions as \( n \to \infty \):
\[
\left( \prod_{i=1}^n 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k) \downarrow \left( \prod_{i=1}^\infty 1_{C \setminus D_{x_i}}(v_i) \right) \cdot \sum_{k=1}^m g(x_k, v_k).
\]

We also have the equality (B.16) for all \( n \geq m \). Hence, if for each \( m \) there exists some \( n \geq m \) for which the quantity in (B.16) is less than \( \infty \), then by the dominated convergence theorem [21 p. 132], (B.18) holds for each \( m \), and hence the desired equality (B.17) holds, which together with (B.13) implies that (B.11) and (B.12) are equal.

(iv) The only case left now is that for some integer \( m \geq 1 \) and all \( n \geq m \), the quantity in (B.16) is \( \infty \). But in view of Eq. (B.5) (i.e., \( \tau > m \Leftrightarrow v_1 \notin D_{x_1}, \ldots, v_{m-1} \notin D_{x_{m-1}}, v_m \notin D_{x_m} \)), this would imply
\[
\mathbb{E}^\mu \left\{ 1_{\{\tau > m\}}(x_1, v_1, \ldots) \sum_{k=1}^{\tau-1} g(x_k, v_k) \right\} = \infty, \quad \mathbb{E}^\mu \left\{ 1_{\{\tau > m\}}(x_1, v_1, \ldots) \sum_{k=1}^{\tau-1} g(x_k, v_k) \right\} = \infty,
\]
and hence both (B.11) and (B.12) equal \( \infty \). This completes the proof.

C About the Existence of Borel Measurable Policies

In this appendix, we mention some useful facts about the existence of Borel measurable policies, to demonstrate the application range of Algorithm III given in Section 3.2, which uses Borel measurable policies.

Recall that the graph of the control constraint \( U, \Gamma = \{(x, u) \mid x \in S, u \in U(x)\} \), 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 \( \Gamma \) admits a Borel measurable function \( f : S \to C \) whose graph lies in \( \Gamma \) (i.e., \( f(x) \in U(x) \) for all \( x \)).

Suppose \( \Gamma \) is a Borel subset of \( S \times C \). It can still happen that such \( f \) does not exist [15]. 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 [18].) Nevertheless, when \( \Gamma \) is Borel, a number of selection theorems for Borel sets in the product of two Polish spaces can be applied for a Borel measurable selection in \( \Gamma \), under fairly general conditions on the control constraint \( U \) (see e.g., [49]). Below are some examples.

Let \( Y \) be the Polish space of which \( C \) is a Borel subset. Assume \( \Gamma \) is Borel. Then 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, [49, Theorem 5.8.11]).

(b) For every $x$, $U(x)$ contains a nonempty open set in $Y$ (by theorems of Kecharis and Saradhi, [49, Theorem 5.8.5]).

(c) For every $x$, $U(x)$ is a $\sigma$-compact set in $Y$ (by a theorem of Arsenin and Kunugui, [49, Theorem 5.12.1]), which is true in particular when $Y$ is $\sigma$-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 the Kuratowski and Ryll-Nardzewski selection theorem, [49, Theorem 5.2.1]).

Thus for many general control constraints, the policy iteration-like algorithm III can be applied.

D 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 [50, Example 6.2, p. 881] and also described in Maitra and Sudderth [35, p. 930]. Our description below closely follows [35]. As mentioned in Remark 5.2, in this example value iteration converges in one iteration if we know the set \{ $x \in S$ | $J^*(x) = 0$ \} and hence can choose an initial function $J$ to satisfy $J \leq J^* \leq cJ$ for some constant $c$, fulfilling the convergence condition in Theorem 5.1(b). In what follows we give a detailed description and calculation to supplement this remark.

Let $R(0,1)$ denote the set of rationals in $(0,1)$ with its usual ordering, and index its elements by $r_1, r_2, \ldots$ (in an arbitrary way). Let \{ $W_r$ | $r \in R(0,1)$ \} be a collection of Borel subsets of $[0,1]$ (called a Borel sieve). Correspondingly, define for each $z \in [0,1]$, $M_z = \{ r \in R(0,1) | z \in W_r \}$, and let $D = \{ z \in [0,1] | M_z \text{ is not well-ordered} \}$.

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

Let $S = \{ \langle z, r \rangle | 0 \leq z \leq 1, 0 \leq r \leq 1, r \text{ rational} \} \cup \{ t \}$. 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

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

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

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

(Equivalently, the one-stage costs are given by \( g(t,u) = 0, g((z,r),u) = 0 \) if \( r_u < r \) and \( z \in W_{r_u} \), and \( g((z,r),u) = 1 \) otherwise.) The optimal cost function \( J^* \) takes only values 0 or 1, and it is not Borel measurable [50]. As shown by [50], \( J^*(t) = 0 \) and for states $(z,1)$ where $z\in [0,1]$,

$$J^*(z,1) = \begin{cases} 0 & \text{if } z \in D, \\ 1 & \text{if } z \in [0,1] \setminus D. \end{cases} \quad (D.1)$$
Value iteration starting from the constant function zero requires uncountably many iterations to converge to \( J^* \), as shown by Maitra and Sudderth [35, p. 930].

Denote \( G = \{ x \in S \mid J^*(x) = 0 \} \). Then \( T^k(J) \rightarrow J^* \) if we let \( J(t) = 0 \) and for states \( x = (z, r) \),

\[
J(z, r) = \begin{cases} 
0 & \text{if } (z, r) \in G, \\
u & \text{otherwise,}
\end{cases}
\]

for some constant \( a \geq 1 \).

Since \( J^* \leq J \leq aJ^* \), this function \( J \) satisfies the condition of Theorem 5.1(b) for the convergence of value iteration.

Indeed \( T(J) = J^* \), as can be verified directly. Consider any \((z, r) \in S\) 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 exist some \( u \in C \) such that with \( x' = f((z, r), u), \)

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

which implies that \( \hat{g}((z, r), u, x') = 0 \) 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) \( f((z, r), u) = t \) and \( \hat{g}((z, r), u, t) = 1 \), in which case \( \hat{g}((z, r), u, t) + J(t) = 1 \).

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

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). Finally, suppose \( r = 1 \). Then the assumption \( J^*(z, r) = 1 \) implies that \( z \notin D \) (cf. Eq. (4.1)). By the definition of \( D \), this means that \( M_z \) is well-ordered and therefore has a smallest element \( \bar{\tau} \in R_{(0,1)} \) if \( M_z \neq \emptyset \). Then, there exists a positive rational number \( r_u \) such that \( r_u < \bar{\tau} \) if \( M_z \neq \emptyset \), and by the definition of \( M_z \), \( z \notin W_{M_z} \). The corresponding index \( u \) satisfies (i). Thus, we have shown \( T(J) = J^* \).

\[\text{Appendix E. Alternative Proof of Theorem 5.1}\]

In this Appendix, we give an alternative proof of Theorem 5.1 about the convergence of value iteration in the nonnegative cost case (P). This proof combines Lemma 5.2 with arguments that are motivated by similar arguments used in the book by Meyn [36, Chap. 9.4] to analyze average cost and total cost problems.

It is also worth mentioning that there are two alternative proofs for Whittle’s bridging condition, Whittle’s original proof [35], and Hartley’s proof [28] based on a concavity argument. For Theorem 5.1 in terms of the arguments involved, the proof given in Section 5 can be compared with Hartley’s proof, and the alternative proof given below can be compared with Whittle’s original proof of the bridging condition.

**Lemma E.1.** (P) For a given \( x \in S \), let \( \pi \) be a policy with \( J_\pi(x) < \infty \). Then with \( x_0 = x \), we have

\[
\lim_{n \to \infty} \mathbb{E}^\pi \{J^*(x_n)\} = 0.
\]

**Proof.** For any \( n \geq 0 \), we have

\[
J_\pi(x) = \mathbb{E}^\pi \left\{ \sum_{k=0}^{n-1} g(x_k, u_k) \right\} + \mathbb{E}^\pi \left\{ \sum_{k=n}^{\infty} g(x_k, u_k) \right\} \geq \mathbb{E}^\pi \left\{ \sum_{k=0}^{n-1} g(x_k, u_k) \right\} + \mathbb{E}^\pi \{J^*(x_n)\}.
\]
Since $\mathbb{E}_\pi \left\{ \sum_{k=0}^{n-1} g(x_k, u_k) \right\} \uparrow J_\pi(x) < \infty$ as $n \to \infty$, it follows that $\limsup_{n \to \infty} \mathbb{E}_\pi \{ J_\pi(x_n) \} = 0$ and hence $\lim_{n \to \infty} \mathbb{E}_\pi \{ J_\pi(x_n) \} = 0$.

**Proof of Theorem 5.1.** As discussed immediately after the theorem, it is sufficient to prove part (a). Denote $J_\infty = \lim_{k \to \infty} T^k(cJ^*)$. By Lemma 5.2, $T(J_\infty) = J_\infty$ and $J^* \leq J_\infty \leq cJ^*$. We need to show $J_\infty = J^*$. The equality $J_\infty(x) = J^*(x)$ holds certainly for all states $x$ with $J^*(x) = \infty$. So we consider an arbitrary $x \in S$ with $J^*(x) < \infty$, and we show that for an arbitrary $\epsilon > 0$, $J_\infty(x) \leq J^*(x) + \epsilon$.

Let $\pi = (\mu_0, \mu_1, \ldots)$ be an $\epsilon$-optimal Markov policy for the state $x$. Since $J_\infty = T(J_\infty)$, we have for any $n \geq 1$, with $x_0 = x$,

\[
J_\infty(x) = T^n(J_\infty)(x) \leq (T_{\mu_0} \circ T_{\mu_1} \circ \cdots \circ T_{\mu_{n-1}})(J_\infty)(x) \\
= \mathbb{E}_\pi \left\{ \sum_{k=0}^{n-1} g(x_k, u_k) \right\} + \mathbb{E}_\pi \{ J_\infty(x_n) \} \\
\leq J^*(x) + \epsilon + c \mathbb{E}_\pi \{ J^*(x_n) \},
\]

where we used the $\epsilon$-optimality of $\pi$ for $x$ and the fact $J_\infty \leq cJ^*$ in the last inequality. Since $J_\pi(x) \leq J^*(x) + \epsilon < \infty$, we have $\lim_{n \to \infty} \mathbb{E}_\pi \{ J^*(x_n) \} = 0$ by Lemma E.1, so it follows from the inequality (E.1) that $J_\infty(x) \leq J^*(x) + \epsilon$. Since $\epsilon$ is arbitrary, we obtain $J_\infty(x) \leq J^*(x)$ for all $x$ with $J^*(x) < \infty$. In view of the fact $J_\infty \geq J^*$, equality must hold and this completes the proof. \[\square\]