DYNAMIC PROGRAMS ON PARTIALLY ORDERED SETS

JOHN STACHURSKI AND THOMAS J. SARGENT

Abstract. We represent a dynamic program as a family of operators acting on a partially ordered set. We provide an optimality theory based on order-theoretic assumptions and show how many applications of dynamic programming fit into this framework. These range from traditional dynamic programs to those involving nonlinear recursive preferences, desire for robustness, function approximation, Monte Carlo sampling and distributional dynamic programs. We apply our framework to establish new optimality and algorithmic results for specific applications.

1. Introduction

Dynamic programs occur in many fields including operations research, artificial intelligence, economics, and finance (Bäuerle and Rieder, 2011; Bertsekas, 2021; Kochenderfer et al., 2022). They are used to price products, control aircraft, sequence DNA, route shipping, recommend information, and solve frontier research problems. Within economics, applications of dynamic programming range from monetary and fiscal policy to asset pricing, unemployment, firm investment, wealth dynamics, inventory control, commodity pricing, sovereign default, natural resource extraction, retirement decisions, portfolio choice, and dynamic pricing.

The key idea behind dynamic programming is to reduce an intertemporal problem with many stages into a two-period problem by assigning appropriate values to future states (Bellman, 1957). While optimality theory for conventional dynamic programs—often called Markov decision processes (MDPs)—is well understood (see, e.g., Puterman (2005), Bäuerle and Rieder (2011), Hernández-Lerma and Lasserre (2012), Bertsekas (2012)), many recent applications fall outside this framework. These include models with nonlinear discounting Bäuerle et al. (2021), Epstein–Zin preferences (Epstein and Zin, 1989), risk-sensitive preferences (Hansen and Sargent,

Date: December 18, 2024.

The authors thank Dimitri Bertsekas, Anmol Bhandari, Shu Hu, Yuchao Li, Rami Tabri, Jingni Yang, and Junnan Zhang for valuable comments and suggestions.

1995; Bäuerle and Jaśkiewicz, 2018), adversarial agents (Cagetti et al., 2002; Hansen and Sargent, 2011), and ambiguity (Iyengar, 2005; Yu et al., 2024).

Mathematicians have constructed frameworks that include both standard MDPs and the growing list of nonstandard applications discussed in the last paragraph. Bertsekas (2022) contributed important work in this direction. His "abstract dynamic programming" framework extends earlier ideas dating back to Denardo (1967) and Bertsekas (1977) and generalizes the traditional Bellman equation in ways that can represent many different models. Recent applications of the framework include Ren and Stachurski (2021), Bloise et al. (2024) and Toda (2023).

Nevertheless, many classes of dynamic programs lie outside the framework of Bertsekas (2022). One such class is dynamic programs that reverse the order of expectation and maximization in the Bellman equation, with well known examples in optimal stopping (Jovanovic, 1982; Hubmer et al., 2020), Q-learning (see, e.g., Watkins (1989); Kochenderfer et al. (2022)), and structural estimation (Rust (1987), Rust (1994), and Mogensen (2018), etc.). Another such class involves dynamic programs with value functions that are not real-valued. Examples can be found in distributional dynamic programming (Bellemare et al., 2017), where states are mapped to distributions, and in empirical dynamic programming, where states are mapped to random elements taking values in a function space (Munos and Szepesvári, 2008; Haskell et al., 2016; Bertsekas, 2021; Kalathil et al., 2021; Rust et al., 2002; Sidford et al., 2023). A related situation occurs with dynamic programs cast in L_p spaces, where value "functions" are actually equivalence classes of functions.

Another motivation for extending the framework of Bertsekas (2022) is to study approximate dynamic programming, which replaces exact value and policy functions with parametric or nonparametric approximations (Powell, 2016; Bertsekas, 2021, 2022). The framework developed here facilitates viewing at least some of these approximate dynamic programs as dynamic programs in their own right. One advantage is acquiring the ability to analyze whether an approximate dynamic program under study is a well-defined Bellman-type optimization problem, in the sense that a single policy function is optimal starting from all possible initial states (see, e.g., Naik et al. (2019)).

In order to capture the broad class of applications described above, the framework developed in this paper replaces traditional value functions with elements of an abstract partially ordered set V. Policies that are traditionally understood as mappings from states to actions now become abstract indices over a family $\{T_\sigma : \sigma \in \Sigma\}$ of "policy operators." Each policy operator T_σ is an order preserving self-map on V. Whenever it exists, the lifetime value v_σ of a policy $\sigma \in \Sigma$ is identified with the unique fixed point of T_σ. A policy σ^* is defined to be optimal when v_{σ^*} is a greatest element of $\{v_\sigma : \sigma \in \Sigma\}$. In this setting we provide an order-theoretic treatment of dynamic programming and describe conditions under which some classical dynamic programming optimality results hold, e.g., the value function is a unique fixed point of the Bellman equation, Bellman's principle of optimality is valid. We also provide conditions under which standard dynamic programming algorithms converge.

Our framework brings three significant benefits. First, policy operators are general enough to represent both standard and many nonstandard dynamic programs. Second, because we work in an abstract partially ordered space, it is possible to handle Bellman equations defined not only over spaces of real-valued functions, but also over spaces of distributions and spaces of random functions. Third, its high level of abstraction simplifies analysis and isolates roles of various assumptions. We illustrate our framework by applying it to establish new results in the context of several applications. .

Section 2 introduces terminology and essential concepts. Section 3 introduces our abstract dynamic programming framework and provides examples that help illustrate the definitions. Section 4 defines optimality and Section 5 provides fundamental optimality results that are then proved in Section 6. Section 7 presents applications and Section 8 concludes.

2. Preliminaries

Let $V = (V, \preceq)$ be a partially ordered set. We use the symbol $\bigvee$ to represent suprema; for example, if $\{v_\alpha\}_{\alpha \in \Lambda}$ is a subset of V, then $\bigvee_\alpha v_\alpha$ is the least element of the set of upper bounds of $\{v_\alpha\}_{\alpha \in \Lambda}$. V is called *bounded* when V has a least and greatest element. A subset C of V is called a *chain* if it is totally ordered by $\preceq$. A sequence $(v_n)_{n \in \mathbb{N}}$ in V is called *increasing* if $v_n \preceq v_{n+1}$ for all $n \in \mathbb{N}$. If (v_n) is increasing and $\bigvee_n v_n = v$ for some $v \in V$, then we write $v_n \uparrow v$. A partially ordered set V is called *chain complete* (resp., *countably chain complete*) if V is bounded and every chain (resp., every increasing sequence) in V has a supremum. V is called *countably*

Dedekind complete if every countable subset of V that is bounded above (i.e., the set of upper bounds is nonempty) has a supremum in V.

A self-map S on V is called *order preserving* on V when $v, w \in V$ and $v \preceq w$ imply $Sv \preceq Sw$, and *order continuous* on V if, for any (v_n) in V with $v_n \uparrow v$, we have $Sv_n \uparrow Sv$.[1] Simple arguments show that if S is order continuous on V then S is order preserving.

In the theorem below, S is a self-map on partially ordered set V and $\mathrm{fix}(S)$ is the set of all fixed points of S in V. While the result is well-known, our version is slightly nonstandard so we include a partial proof.

Theorem 2.1. *The set* $\mathrm{fix}(S)$ *is nonempty if either*

(i) V *is chain complete and* S *is order preserving, or*
(ii) V *is countably chain complete and* S *is order continuous.*

Moreover, in the second case, $v \in V$ *and* $v \preceq Sv$ *implies* $\bigvee_n S^n v \in \mathrm{fix}(S)$.

Proof. For a proof of case (i), see, for example, Theorems 8.11 and 8.22 of Davey and Priestley (2002). As for (ii), let S, V be as stated in (ii) and fix $v \in V$ with $v \preceq Sv$. S is order continuous and hence order preserving, so $(v_n) \coloneqq (S^n v)$ is increasing. As V is countably chain complete, the suprema $\bigvee_{n \geqslant 1} v_n$ and $\bigvee_{n \geqslant 1} Sv_n$ exist in V. If $\bar{v} \coloneqq \bigvee_n v_n$, then $S\bar{v} = S \bigvee_{n \geqslant 1} v_n = \bigvee_{n \geqslant 1} Sv_n = \bigvee_{n \geqslant 2} v_n = \bar{v}$, where the second equality is by order continuity. Hence $\bar{v} \in \mathrm{fix}(S)$. Finally, countable chain completeness implies that V has a least element $\bot$. We then have $\bot \preceq S\bot$, so $\mathrm{fix}(S)$ is nonempty. □

We call a self-map S on V *upward stable* if S has a unique fixed point $\bar{v}$ in V and $v \preceq Sv$ implies $v \preceq \bar{v}$, *downward stable* if S has a unique fixed point $\bar{v}$ in V and $Sv \preceq v$ implies $\bar{v} \preceq v$, and *order stable* if S is both upward and downward stable. Order stability is a weak and purely order-theoretic version of asymptotic stability.

Example 2.1. Let V be a metric space endowed with a closed partial order $\preceq$, so that $u_n \preceq v_n$ for all n implies $\lim_n u_n \preceq \lim_n v_n$ whenever the limits exist, and let $S \colon V \to V$ be order preserving and globally stable under the metric on V, so that S has a unique fixed point $\bar{v}$ in V and $\lim_n S^n v = \bar{v}$ for all $v \in V$. If $v \preceq Sv$, then $v \preceq S^n v$ for all n. Taking the limit and using the fact that $\preceq$ is closed yields $v \preceq \bar{v}$. Hence S is upward stable. A similar argument shows that S is downward stable.

[1] The definition of order continuity varies across subfields of mathematics but the one just given suffices for our purposes.

Lemma 2.2. *Let S be a self-map on V with at most one fixed point in V. If either*

(i) *V is chain complete and S is order preserving, or*
(ii) *V is countably chain complete and S is order continuous,*

then S is order stable on V.

Proof. First suppose that S is order preserving and V is chain complete, with least element $\bot$ and greatest element $\top$. By Theorem 2.1, S has a fixed point $\bar{v} \in V$. By assumption, $\bar{v}$ is the only fixed point of S in V. Now fix $v \in V$ with $v \preceq Sv$. Since $I := [v, \top]$ is itself chain complete, and since S maps I to itself, S has a fixed point in I. Hence $\bar{v} \in I$, which yields $v \preceq \bar{v}$. This proves upward stability. The proof of downward stability is similar.

Now suppose that V is countably chain complete and S is order continuous. Let $\bar{v}$ be the unique fixed point of S in V (existence of which follows from Theorem 2.1). Pick any $v \in V$ with $v \preceq Sv$. Since $I := [v, \top]$ is itself countably chain complete, and since S is an order continuous map from I to itself, Theorem 2.1 implies that S has a fixed point in I. Hence $\bar{v} \in I$, which yields $v \preceq \bar{v}$. This proves upward stability. The proof of downward stability is similar. □

3. Abstract Dynamic Programs

We define an *abstract dynamic program* (ADP) to be a pair $(V, \mathbb{T})$, where $V = (V, \preceq)$ is a partially ordered set and $\mathbb{T} = \{T_\sigma : \sigma \in \Sigma\}$ is a family of order preserving self-maps on V. The set V is called the *value space*. The operators in $\mathbb{T}$ are called *policy operators*. Σ is an arbitrary index set and elements of Σ will be referred to as *policies*. In applications we impose conditions under which each T_σ has a unique fixed point. In these settings, the significance of T_σ is that its fixed point, denoted below by v_σ, represents the lifetime value (or cost) of following policy σ.

Example 3.1. In some settings, V is a set of functions and $\preceq$ is the pointwise partial order $\leqslant$ (i.e., $v \leqslant w$ if $v(x) \leqslant w(x)$ for all x in X). The value $v_\sigma(x)$ represents the lifetime value of following policy σ when the initial state is X.

Let $(V, \mathbb{T})$ be an ADP with policy set Σ. Given $v \in V$, a policy σ in Σ is called *v-greedy* if $T_\sigma v \succeq T_\tau v$ for all $\tau \in \Sigma$. We call $(V, \mathbb{T})$ *regular* when each $v \in V$ has at least one

v-greedy policy. Given $v \in V$, we set

$$Tv = \bigvee_{\sigma \in \Sigma} T_\sigma v \tag{1}$$

whenever the supremum exists. We call T the *Bellman operator* generated by $(V, \mathbb{T})$. We say that $v \in V$ *satisfies the Bellman equation* if $Tv = v$.

For a given ADP $(V, \mathbb{T})$, we define three sets:

- $V_G := \{v \in V : \text{ at least one } v\text{-greedy policy exists}\}$,
- $V_\Sigma := \{v \in V : v \text{ is a fixed point of } T_\sigma \text{ for some } \sigma \in \Sigma\}$, and
- $V_U := \{v \in V : v \preceq Tv\}$.

The next lemma shows that T has attractive properties on V_G.

Lemma 3.1. *The Bellman operator T has the following properties:*

(i) *T is well-defined and order preserving on V_G.*
(ii) *For $v \in V_G$ we have $T_\sigma v = Tv$ if and only if $\sigma \in \Sigma$ is v-greedy.*

Proof. We begin with part (ii). Fix $v \in V_G$ and let σ be v-greedy. Then, by definition, $T_\sigma v$ is the greatest element of $\{T_\tau v\}_{\tau \in \Sigma}$. A greatest element is also a supremum, so $Tv = T_\sigma v$. Conversely, if $Tv = T_\sigma v$, then $T_\tau v \preceq T_\sigma v$ for all $\tau \in \Sigma$. Hence (ii) holds. As for (i), fixing $v \in V_G$, a v-greedy policy exists, so Tv is well-defined by part (ii). As for the order preserving claim, fix $v, w \in V_G$ with $v \preceq w$. Let $\sigma \in \Sigma$ be v-greedy. Since T_σ is order preserving, we have $Tv = T_\sigma v \preceq T_\sigma w \preceq Tw$. □

Lemma 3.2. *If $(V, \mathbb{T})$ is regular, then $V_\Sigma \subset V_U$.*

Proof. Fix $v_\sigma \in V_\Sigma$ and let T_σ be such that v_σ is a fixed point. Since T is well-defined on all of V, we have $v_\sigma = T_\sigma v_\sigma \preceq Tv_\sigma$. In particular, $v_\sigma \in V_U$. □

3.1. **Example: MDPs.** We begin with a simple and familiar example involving Markov decision processes (MDPs, see, e.g., Bäuerle and Rieder (2011)). The objective is to maximize $\mathbb{E} \sum_{t \geqslant 0} \beta^t r(X_t, A_t)$ where X_t takes values in finite set X (the state space), A_t takes values in finite set A (the action space), Γ is a nonempty correspondence from X to A (the feasible correspondence), $\mathsf{G} := \{(x, a) \in \mathsf{X} \times \mathsf{A} : a \in \Gamma(x)\}$ denotes the feasible state-action pairs, r is a reward function defined on G, $\beta \in (0, 1)$

is a discount factor, and $P\colon \mathsf{G} \times \mathsf{X} \to [0,1]$ provides transition probabilities. The Bellman equation is

$$v(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x'} v(x') P(x,a,x') \right\} \qquad (x \in \mathsf{X}). \tag{2}$$

The set of feasible policies is the finite set $\Sigma := \{\sigma \in \mathsf{A}^{\mathsf{X}} : \sigma(x) \in \Gamma(x) \text{ for all } x \in \mathsf{X}\}$. We combine $\mathbb{R}^{\mathsf{X}}$ (the set of all real-valued functions on X) with the pointwise partial order $\leqslant$ and, for $\sigma \in \Sigma$ and $v \in \mathbb{R}^{\mathsf{X}}$, define the MDP policy operator

$$(T_\sigma v)(x) = r(x, \sigma(x)) + \beta \sum_{x'} v(x') P(x, \sigma(x), x') \qquad (x \in \mathsf{X}). \tag{3}$$

Let $V := \mathbb{R}^{\mathsf{X}}$ and $\mathbb{T} = \{T_\sigma : \sigma \in \Sigma\}$. Since each T_σ is an order preserving self-map on V, the pair $(V, \mathbb{T})$ is an ADP. Given $v \in V$, let $\sigma \in \Sigma$ be such that

$$\sigma(x) \in \operatorname*{arg\,max}_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \sum_{x'} v(x') P(x,a,x') \right\} \quad \text{for all } x \in \mathsf{X}. \tag{4}$$

Such a σ satisfies $T_\sigma v \geqslant T_\tau v$ for all $\tau \in \Sigma$, implying σ is v-greedy. Moreover, since Γ is nonempty, at least one policy obeying (4) exists. This proves that $(V, \mathbb{T})$ is regular.

We stated above that, in applications, the fixed point of each $T_\sigma \in \mathbb{T}$ is the lifetime value of policy σ. To see the idea in the MDP setting, fix $\sigma \in \Sigma$ and let r_σ and P_σ be defined by $P_\sigma(x, x') := P(x, \sigma(x), x')$ and $r_\sigma(x) := r(x, \sigma(x))$. In the present setting, the lifetime value of σ given $X_0 = x$ is understood to be $v_\sigma(x) = \mathbb{E} \sum_{t \geqslant 0} \beta^t r_\sigma(X_t)$, where $(X_t)_{t \geqslant 0}$ is a Markov chain generated by P_σ with initial condition $X_0 = x \in \mathsf{X}$. Pointwise on X, we can express v_σ as $v_\sigma = \sum_{t \geqslant 0} (\beta P_\sigma)^t r_\sigma = (I - \beta P_\sigma)^{-1} r_\sigma$ (see, e.g., Puterman (2005), Theorem 6.1.1). This implies that v_σ is the unique solution to the equation $v = r_\sigma + \beta P_\sigma v$. From the definition of T_σ in (3), this is equivalent to the statement that v_σ is the unique fixed point of T_σ.

For the ADP $(V, \mathbb{T})$, the ADP Bellman equation (1) reduces to the MDP Bellman equation (2). This connection is important because it allows to use optimality properties of $(V, \mathbb{T})$ to study optimality properties of the MDP. To see that it holds observe that, since V is endowed with the pointwise partial order, for given $v \in V$ and $x \in \mathsf{X}$, the ADP Bellman operator (1) reduces to

$$(T v)(x) = \sup_{\sigma \in \Sigma} (T_\sigma v)(x) = \sup_{\sigma \in \Sigma} \left\{ r(x, \sigma(x)) + \beta \sum_{x'} v(x') P(x, \sigma(x), x') \right\}.$$

By the definition of Σ, we can also write this as

$$(T\,v)(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \sum_{x'} v(x') P(x, a, x') \right\}. \tag{5}$$

Evidently v satisfies the ADP Bellman equation $Tv = v$ if and only if the traditional MDP Bellman equation (2) holds.

3.2. **Example: Risk-Sensitive Q-learning.** Some dynamic programs reverse the order of maximization and mathematical expectation. One example is Q-factor risk-sensitive decision processes (see, e.g., Fei et al. (2021)), where the Bellman equation takes the form

$$f(x, a) = r(x, a) + \frac{\beta}{\theta} \ln \left\{ \sum_{x'} \exp\left[\theta \max_{a' \in \Gamma(x')} f(x', a') \right] P(x, a, x') \right\} \tag{6}$$

for $(x, a) \in \mathsf{G}$ and nonzero $\theta \in \mathbb{R}$. (We take X, A, Γ, Σ and G as in our discussion of MDPs in Section 3.1.) Given $\sigma \in \Sigma$, the corresponding policy operator is

$$(T_\sigma f)(x, a) = r(x, a) + \frac{\beta}{\theta} \ln \left[\sum_{x'} \exp\left[\theta f(x', \sigma(x')) \right] P(x, a, x') \right] \tag{7}$$

where $f \in \mathbb{R}^{\mathsf{G}} :=$ all real-valued functions on G. With $\mathbb{T} := \{T_\sigma : \sigma \in \Sigma\}$ and G endowed with the pointwise order, the pair $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is an ADP. The ADP Bellman operator is $Tf := \bigvee_\sigma T_\sigma f$ (by (1)) and the ADP Bellman equation is $Tf = f$. By replicating arguments in Section 3.1, one can show that $f \in \mathsf{G}$ solves $Tf = f$ if and only if it solves (6). This allows us to study optimality properties of the original model (as characterized by (6)) through the ADP $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$.

3.3. **Distributional Dynamic Programming.** Distributional dynamic programming focuses on an entire distribution of lifetime returns, not just its expected value (Bellemare et al., 2017). This falls outside frameworks such as Puterman (2005) or Bertsekas (2022) because elements of the value space $\mathcal{P}^{\mathsf{X}}$ are not real-valued functions. In this section we show how distributional dynamic programming can be represented in the setting of ADPs.

In what follows, $\mathcal{P}$ is the set of all probability distributions on $\mathbb{R}$ and $\mathcal{P}^{\mathsf{X}}$ is the set of all functions from X into $\mathcal{P}$. A typical element is written as $\eta(x, \mathrm{d}g)$, indicating that

$\eta(x,\cdot)$ is a distribution on $\mathbb{R}$ for each $x \in \mathsf{X}$. For $\eta, \hat{\eta} \in \mathcal{P}^{\mathsf{X}}$, we write $\eta \preceq \hat{\eta}$ when $\eta(x)$ lies below $\hat{\eta}(x)$ in the sense of stochastic dominance for every $x \in \mathsf{X}$. Thus,

$$\eta \preceq \hat{\eta} \quad \iff \quad \int h(g)\eta(x, \mathrm{d}g) \leqslant \int h(g)\hat{\eta}(x, \mathrm{d}g) \quad \forall\, h \in ib\mathbb{R},\; x \in \mathsf{X},$$

where $ib\mathbb{R}$ is the set of increasing bounded measurable functions from $\mathbb{R}$ to itself.

Maintaining the MDP setting in Section 3.1 but switching to a distributional perspective, the distributional policy operator, written here as D_σ, maps $\eta \in \mathcal{P}^{\mathsf{X}}$ to $D_\sigma \eta$, where $(D_\sigma \eta)(x)$ is the distribution of the random variable $G' := r_\sigma(x) + \beta G_{X'}$ when $G_{X'}$ is sampled by first drawing the next period state X' from $P_\sigma(x,\cdot)$ and then drawing G from $\eta(X',\cdot)$. So the expectation of $h \in ib\mathbb{R}$ under the distribution $(D_\sigma\eta)(x)$ can be expressed as

$$\mathbb{E}h(G') := \langle h, (D_\sigma \eta)(x) \rangle := \sum_{x'} \int h(r_\sigma(x) + \beta g)\eta(x', \mathrm{d}g) P_\sigma(x, x').$$

If we now take $\eta \preceq \hat{\eta}$ and $h \in ib\mathbb{R}$, we get $\int h(r_\sigma(x) + \beta g)\eta(x', \mathrm{d}g) \leqslant \int h(r_\sigma(x) + \beta g)\hat{\eta}(x', \mathrm{d}g)$ for any x' and hence $\langle h, (D_\sigma \eta)(x)\rangle \leqslant \langle h, (D_\sigma \hat{\eta})(x)\rangle$. Since this holds for any x we have $D_\sigma \eta \preceq D_\sigma \hat{\eta}$, so D_σ is order preserving. In particular, $(\mathcal{P}^{\mathsf{X}}, \{D_\sigma\}_{\sigma\in\Sigma})$ is an ADP.

3.4. **Empirical Dynamic Programming.** Monte Carlo estimates are sometimes used to approximate mathematical expectations in dynamic programs with large state spaces. For example, in Haskell et al. (2016), the MDP Bellman operator (5) is replaced by

$$(\hat{T}\, v)(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \frac{1}{n} \sum_{i=1}^{n} v(F(x,a,\xi_i)) \right\},$$

where $(\xi_i)_{i=1}^n$ is a collection of random variables on probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and each $F(x,a,\xi_i)$ has distribution $P(x,a,\cdot)$. The corresponding policy operators $\hat{\mathbb{T}} := \{\hat{T}_\sigma : \sigma \in \Sigma\}$ are

$$(\hat{T}_\sigma\, v)(x) = r(x, \sigma(x)) + \beta \frac{1}{n} \sum_{i=1}^{n} v(F(x, \sigma(x), \xi_i)). \tag{8}$$

Following Haskell et al. (2016), we take $\mathcal{V}$ to be the set of random elements defined on probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in the function space $\mathbb{R}^{\mathsf{X}}$. To make the dependence on $\omega \in \Omega$ explicit we write a realization as $v(\omega, \cdot)$, so that $x \mapsto v(\omega, x)$ is a function in $\mathbb{R}^{\mathsf{X}}$ assigning values to states. The policy operator (8) then becomes

$$(\hat{T}_\sigma v)(\omega, x) = r(x, \sigma(x)) + \beta \frac{1}{n} \sum_{i=1}^{n} v(\omega, F(x, \sigma(x), \xi_i(\omega))). \tag{9}$$

A partial order can be introduced on $\mathcal{V}$ by writing $v \leqslant w$ when $v(\omega, x) \leqslant w(\omega, x)$ for all $\omega \in \Omega$ and $x \in \mathsf{X}$. It is clear that $\hat{T}_\sigma v \leqslant \hat{T}_\sigma w$ whenever $v \leqslant w$, so $(\mathcal{V}, \hat{\mathbb{T}})$ is an ADP.

3.5. **Example: Approximate Dynamic Programming.** In practice, solution methods for a vast range of dynamic programs involve some form of function approximation to simplify update steps and generate representations of value and policy functions (see, e.g., Powell (2016); Bertsekas (2021, 2022)). For example, the MDP policy operator T_σ from (3) might be replaced by $A \circ T_\sigma$, where A implements an approximation architecture such as kernel averaging or projection onto a space of basis functions.

When function approximation is added to policy and Bellman operators, properties needed to pose a well-defined dynamic program may not be satisfied. For example, function approximations may transmute a well-behaved MDP into a dynamic program for which no optimal stationary policy exists (Naik et al., 2019). Our framework facilitating addressing these issues. For example, if the approximation operator A is order preserving, then setting $V = \mathbb{R}^{\mathsf{X}}$ and $\mathbb{T}_A = \{A \circ T_\sigma : \sigma \in \Sigma\}$ yields an ADP $(V, \mathbb{T}_A)$. Below we provide results under which ADPs such as $(V, \mathbb{T}_A)$ have well-defined optimal policies.

3.6. **Example: Structural Estimation.** Rust (1987) and many subsequent authors study discrete choice problems with Bellman equations of the form

$$g(x, a) = \sum_{x'} \int \left\{ \max_{a' \in \mathsf{A}} \left[r(x', a', e') + \beta g(x', a') \right] \right\} \nu(\mathrm{d}e') P(x, a, x'). \tag{10}$$

Here $(x, a) \in \mathsf{G} := \mathsf{X} \times \mathsf{A}$ where A and X are the action and state spaces respectively. The set A is finite (hence discrete choice) and we take X to be finite for simplicity (although other settings can also be handled). We assume that r is a bounded reward function, $\beta \in (0, 1)$ and $P(x, a, \cdot)$ is a probability distribution (probability mass function) on X for each $(x, a) \in \mathsf{G}$. The component e' in the reward function takes values in some measurable space and is IID with distribution ν.

The function g can be interpreted as an "expected post-action value function." Advantages of working with this version of the Bellman equation are discussed in Rust (1994), Kristensen et al. (2021) and other sources. Because the max operator is inside the expectation, frameworks such as Puterman (2005) and Bertsekas (2022) do not directly apply. Nevertheless, we can set this problem up as an ADP by taking Σ to be the set of maps from X to A and, for each $\sigma \in \Sigma$, introducing the policy operator

$$(T_\sigma g)(x, a) = \sum_{x'} \int \{r(x', \sigma(x'), e') + \beta g(x', \sigma(x'))\} \, \nu(\mathrm{d}e') P(x, a, x'). \tag{11}$$

Clearly T_σ maps $\mathbb{R}^{\mathsf{G}}$ into itself and is order preserving on $\mathbb{R}^{\mathsf{G}}$ under the pointwise partial order. Hence, with $\mathbb{T} := \{T_\sigma : \sigma \in \Sigma\}$, the pair $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is an ADP. Moreover, if we take $M \in \mathbb{N}$ such that $|r| \leqslant M$ and set W to all $g \in \mathbb{R}^{\mathsf{G}}$ with $|g| \leqslant M/(1-\beta)$, then straightforward calculations show that T_σ maps W to itself. Hence $(W, \mathbb{T})$ is also an ADP.

4. Properties of ADPs

In this section we define optimality for ADPs. We also categorize ADPs with the aim of determining properties that lead to strong optimality results.

4.1. Basic Properties.

Let $(V, \mathbb{T})$ be an ADP with policy set Σ. When it exists, we denote the unique fixed point of T_σ in V by v_σ. We call $(V, \mathbb{T})$ *finite* if $\mathbb{T}$ is finite, *well-posed* if each $T_\sigma \in \mathbb{T}$ has a unique fixed point v_σ in V, *order stable* if each $T_\sigma \in \mathbb{T}$ is order stable on V, *order continuous* if each $T_\sigma \in \mathbb{T}$ is order continuous on V, and *bounded above* if there exists a $u \in V$ with $T_\sigma u \preceq u$ for all $T_\sigma \in \mathbb{T}$.

In our applications, the lifetime value of a policy σ coincides with the fixed point v_σ of its policy operator T_σ. Well-posedness is a minimal condition because without it we cannot be sure that policies have well-defined lifetime values. Maximizing lifetime values (or, equivalently, minimizing lifetime costs) is the objective of dynamic programming.

Example 4.1. The distributional ADP $(\mathcal{P}^{\mathsf{X}}, \{D_\sigma\}_{\sigma \in \Sigma})$ described in Section 3.3 is well-posed, since, given boundedness of the reward function r and $\beta \in (0, 1)$, each policy operator D_σ has a unique fixed point in $\mathcal{P}^{\mathsf{X}}$. This existence and uniqueness result follows from Theorem 1 of Gerstenberg et al. (2023).

Example 4.2. The ADP $(V, \mathbb{T})$ generated by the MDP model in Section 3.1 is finite, well-posed, order stable, order continuous, and bounded above. Finiteness holds because X and A are finite. $(V, \mathbb{T})$ is well-posed because, under the stated assumptions, each $T_\sigma v = r_\sigma + \beta P_\sigma v$ has a unique fixed point in V given by $v_\sigma := (I - \beta P_\sigma)^{-1} r_\sigma$. Order stability follows from Lemma 2.2. $(V, \mathbb{T})$ is order continuous because $v_n \uparrow v$ is equivalent to $v_n \to v$ pointwise when $V = \mathbb{R}^{\mathsf{X}}$. Hence $v_n \uparrow v$ implies $T_\sigma v_n \uparrow T_\sigma v$. Finally, $(V, \mathbb{T})$ is bounded above because, for any $\sigma \in \Sigma$,

$$u = \frac{\max r}{1 - \beta} \quad \implies \quad T_\sigma u = r_\sigma + \beta P_\sigma u \leqslant \max r + \beta u = u.$$

Example 4.3. The ADP $(W, \mathbb{T})$ generated by the dynamic structural model in Section 3.6 is finite, well-posed, order stable, and order continuous. The proof is almost identical to that given in Example 4.2.

Below we use order stability as a condition for optimality. The next lemma shows that order continuity passes from the policy operators to the Bellman operator.

Lemma 4.1. *If $(V, \mathbb{T})$ is order continuous and V is countably chain complete, then T is order continuous on V.*

Proof. Fix $(v_n) \subset V$ with $v_n \uparrow \bar v \in V$. Since, T is order preserving, $(Tv_n)_{n \geqslant 1}$ is also increasing. Hence $\bigvee_n Tv_n$ exists in V. We claim that $\bigvee_n Tv_n = T\bar v$. On one hand, $T\bar v$ is an upper bound for (Tv_n). On the other hand, if $w \in V$ is such that $Tv_n \preceq w$ for all n, then $T_\sigma v_n \preceq w$ for all n and σ. Fixing $\sigma \in \Sigma$, taking the supremum over n and using order-continuity of T_σ gives $T_\sigma \bar v \preceq w$. Hence $T\bar v \preceq w$, which means that $T\bar v$ is a least upper bound of (Tv_n). This confirms that $\bigvee_n Tv_n = T\bar v$. Hence T is order continuous. □

4.2. Defining Optimality.

Next we define optimality for ADPs using concepts that are direct generalizations of dynamic program optimality from existing frameworks. To begin, we recall that, for a well-posed ADP $(V, \mathbb{T})$ with policy set Σ, the symbol V_Σ represents the set of lifetime values $\{v_\sigma\}_{\sigma \in \Sigma}$ generated by $(V, \mathbb{T})$. A policy $\sigma \in \Sigma$ is called *optimal* for $(V, \mathbb{T})$ if v_σ is a greatest element of V_Σ.

Example 4.4. The ADP $(V, \mathbb{T})$ generated by the MDP model in Section 3.1 uses the pointwise partial order on V, so v_σ is optimal if and only if $v_\sigma(x) = \max_{\tau \in \Sigma} v_\tau(x)$ for all $x \in \mathsf{X}$. This is the standard definition of optimality of MDP policies (see, e.g., Puterman (2005), Ch. 6).

We say that *Bellman's principle of optimality holds* if V_Σ has a greatest element v^* and, for $\sigma \in \Sigma$,

$$\sigma \text{ is optimal} \iff \sigma \text{ is } v^*\text{-greedy.} \tag{12}$$

Example 4.5. The ADP generated by the MDP model in Section 3.1 satisfies Bellman's principle of optimality. See, for example, Bertsekas (2022), Lemma 2.1.1 (c).

We say that the *fundamental ADP optimality results* hold for $(V, \mathbb{T})$ if

(B1) V_Σ has a greatest element v^*,
(B2) v^* is the unique solution to the Bellman equation in V, and
(B3) Bellman's principle of optimality holds.

When (B1) holds we call the greatest element v^* the *value function*. Clearly (B1) is equivalent to the statement that at least one optimal policy exists.

Properties (B1)–(B3) are not independent, as the next lemma shows.

Lemma 4.2. *If V_Σ has greatest element v^*, then v^* satisfies the Bellman equation if and only if Bellman's principle of optimality holds.*

Proof. Let V_Σ have greatest element v^*. Suppose first that $Tv^* = v^*$. Fixing $\sigma \in \Sigma$, we claim that that (12) holds. As for $\Rightarrow$, if $\sigma \in \Sigma$ is optimal, then $v_\sigma = v^*$. Since $T_\sigma v_\sigma = v_\sigma$, this implies $T_\sigma v^* = v^*$. But $T v^* = v^*$, so $T_\sigma v^* = Tv^*$. It follows that σ is v^*-greedy (by Lemma 3.1). As for $\Leftarrow$, if σ is v^*-greedy, then $T_\sigma v^* = Tv^* = v^*$. But v_σ is the unique fixed point of T_σ in V, so $v_\sigma = v^*$. Hence σ is an optimal policy. As for the converse implication, the definition of greatest elements implies existence of a $\sigma \in \Sigma$ such that $v_\sigma = v^*$. By Bellman's principle of optimality, the policy σ is v^*-greedy. As a result, $Tv^* = T_\sigma v^* = T_\sigma v_\sigma = v_\sigma = v^*$. □

4.3. **Algorithms.** Let $(V, \mathbb{T})$ be a regular well-posed ADP with Bellman operator T and σ-lifetime value functions V_Σ. Suppose that the fundamental optimality properties (B1)–(B3) hold and let v^* denote the value function. We consider three major algorithms for computing v^*: value function iteration, optimistic policy iteration and Howard policy iteration. To this end, we define the *Howard policy operator* $H \colon V \to V_\Sigma$ corresponding to $(V, \mathbb{T})$ via $Hv = v_\sigma$ where σ is v-greedy. So that H is well-defined, we always select the same v-greedy policy when applying H to v. Also, fixing arbitrary

$m \in \mathbb{N}$, we define the *optimistic policy operator* W_m from V_G to V via $W_m v := T_\sigma^m v$ where σ is v-greedy. As was the case for H, we always select a fixed v-greedy policy when applying W_m to v. We say that

- *value function iteration (VFI) converges* if $T^n v \uparrow v^*$ for all $v \in V_U$,
- *Howard policy iteration (HPI) converges* if $H^n v \uparrow v^*$ for all $v \in V_U$, and
- *optimistic policy iteration (OPI) converges* if $W_m^n v \uparrow v^*$ for all $v \in V_U$.

5. Optimality Results

In this section we present our main theoretical results. Proofs are deferred to Section 6. *Throughout this section, $(V, \mathbb{T})$ is a regular well-posed ADP.* First, we state a high-level result that assumes the existence of a fixed point for the Bellman operator.

Theorem 5.1. *If $(V, \mathbb{T})$ is downward stable and T has at least one fixed point in V, then the fundamental ADP optimal results hold.*

Now we drop the assumption that T has a fixed point and suppose instead that V has some form of completeness.

Theorem 5.2. *If the value space V is chain complete, then the fundamental ADP optimal results hold.*

The next result weakens chain completeness to a milder condition on the space, while adding continuity properties on the policy operators.

Theorem 5.3. *If $(V, \mathbb{T})$ is order continuous and V is countably chain complete, then*

(i) *the fundamental ADP optimal results hold, and*
(ii) *VFI, OPI, and HPI all converge.*

In the previous two theorems, it is assumed that V is order bounded. The next two theorems drop this assumption. In the first, we state a result for the case where $(V, \mathbb{T})$ is finite, which is relatively common in applications.

Theorem 5.4. *If $(V, \mathbb{T})$ is finite and order stable, then*

(i) *the fundamental ADP optimal results hold, and*

(ii) *HPI converges in finitely many steps.*

Finally, we consider a setting where $(V, \mathbb{T})$ is not finite and the value space can be unbounded.

Theorem 5.5. *Let $(V, \mathbb{T})$ be order continuous and order stable. If V is countably Dedekind complete and $(V, \mathbb{T})$ is bounded above, then*

(i) *the fundamental ADP optimality properties hold and*
(ii) *VFI, OPI and HPI all converge.*

6. Proofs of Section 5 Results

In this section we prove optimality results from Section 5. Throughout this section, $(V, \mathbb{T})$ is a regular well-posed ADP with Howard policy operator H, optimistic policy operator W and Bellman operator T.

6.1. Preliminaries.

We begin with some lemmas.

Lemma 6.1. *The following statements hold.*

(L1) *If $v \in V$ with $Hv = v$, then $Tv = v$.*
(L2) *The operators T, W and H all map V_U to itself.*
(L3) *If $v \in V_U$, then $Tv \preceq Wv \preceq T^m v$.*

Proof. As for (L1), fix $v \in V$ with $H\,v = v$ and let σ be a v-greedy policy such that $H\,v = v_\sigma$. Then $v_\sigma = H\,v = v$. Since σ is v-greedy, $T_\sigma\,v = T\,v$. Since v_σ is fixed for T_σ, we also have $T_\sigma\,v = v$. Combining the last two equalities proves (L1).

As for (L2), fix $v \in V_U$. Since $v \preceq Tv$ and T is order preserving on V_U, we have $Tv \preceq TTv$. Hence $Tv \in V_U$. As for W, let σ be v-greedy with $W = T_\sigma^m v$. Since T and T_σ are order preserving and $v \preceq Tv$, we have $Wv = T_\sigma T_\sigma^{m-1} v \preceq TT_\sigma^{m-1} v \preceq TT_\sigma^{m-1} Tv = TT_\sigma^m v = TWv$. Hence $Wv \in V_U$. Finally, regarding H, we observe that $Hv \in V_\Sigma$ and, by Lemma 3.2, $V_\Sigma \subset V_U$.

To prove (L3) we fix $v \in V_U$. Letting σ be v-greedy, we have $v \preceq Tv = T_\sigma\,v$. Iterating on this inequality with T_σ proves that $(T_\sigma^k\,v)$ is increasing. In particular, $Tv = T_\sigma\,v \preceq Wv$. For the second inequality in (L3) we use the fact that $T_\sigma \preceq T$ on V and T and T_σ are both order preserving to obtain $Wv = T_\sigma^m v \preceq T^m v$. □

The next lemma adds upward stability and derives additional implications.

Lemma 6.2. *If* $(V, \mathbb{T})$ *is upward stable, then for every* $v \in V_U$*,*

$$T^n v \preceq W^n v \qquad \text{and} \qquad T^n v \preceq H^n v \tag{13}$$

for all $n \in \mathbb{N}$*. Moreover, the VFI sequence* $(T^n v)$*, the HPI sequence* $(H^n v)$ *and the OPI sequence* $(W^n v)$ *are all increasing.*

Proof. Our first claim is that

$$u, v \in V_U \text{ with } u \preceq v \implies Tu \preceq Wv \text{ and } Tu \preceq Hv. \tag{14}$$

To show this we fix such u, v and take σ to be v-greedy. Let v_σ be the σ-value function, so that $T_\sigma v_\sigma = v_\sigma$ and $v_\sigma = Hv$. Since $v \in V_U$ we have

$$v \preceq Tv = T_\sigma v \preceq T_\sigma^m v = Wv \preceq v_\sigma = Hv. \tag{15}$$

The second inequality is by iterating on $v \preceq T_\sigma v$, while the third is by upward stability. Since $Tu \preceq Tv$, we can use (15) to obtain (14). Iterating on (14) produces (13). The last claim in Lemma 6.2 follows from (15), which tells us that elements of V_U are mapped up by T, W, and H. □

Corollary 6.3. *If* $(V, \mathbb{T})$ *is upward stable and an optimal policy exists, then convergence of VFI implies convergence of OPI and convergence of HPI.*

Proof. Assume the conditions of the corollary and fix $v \in V_U$. Since an optimal policy exists, v^* exists and is the greatest element of V_Σ. Lemma 6.2 yields $v \preceq T^n v \preceq W^n v \preceq v^*$ for all n, where the last inequality follows from (15) and the fact that $v_\sigma \preceq v^*$ for all σ. Hence convergence of VFI implies convergence of OPI. The proof for HPI is similar. □

6.2. **Remaining Proofs.** We now prove the main optimality results from Section 5.

Proof of Theorem 5.1. Let $(V, \mathbb{T})$ be downward stable and suppose that T has at least one fixed point $\bar v$ in V. Since $(V, \mathbb{T})$ is regular, there exists a $\sigma \in \Sigma$ with $T_\sigma \bar v = T \bar v$ (Lemma 3.1). Since $(V, \mathbb{T})$ is well-posed, this last equality and $T\bar v = \bar v$ imply that $\bar v$ is the unique fixed point of T_σ. Thus, $\bar v \in V_\Sigma$. Moreover, if $\tau \in \Sigma$, then $T_\tau \bar v \preceq T \bar v = \bar v$, so, by downward stability, $v_\tau \preceq \bar v$. Hence $\bar v$ is both the greatest element of V_Σ and a solution to the Bellman equation in V. The fundamental ADP optimality properties now follow from Lemma 4.2. □

Proof of Theorem 5.2. Let V be chain complete. $(V, \mathbb{T})$ is order stable by Lemma 2.2 and, by Theorem 2.1, T has at least one fixed point in V. Hence the conditions of Theorem 5.1 hold, which implies the fundamental ADP optimality properties. □

Proof of Theorem 5.3. Suppose $(V, \mathbb{T})$ is order continuous and V is countably chain complete. Since T is order continuous (Lemma 4.1), Theorem 2.1 implies that T has at least one fixed point in V. Also, by Lemma 2.2, $(V, \mathbb{T})$ is order stable. Hence, by Theorem 5.1, the fundamental ADP optimality properties hold. Moreover, for $v \in V_U$ the sequence $v_n := T^n v$ is increasing. Since T is order continuous, the supremum is a fixed point of T (Theorem 2.1). But, by (B2) of the fundamental ADP optimality properties, the value function v^* is the only fixed point of T in V. Hence VFI converges. Convergence of OPI and HPI now follows from Corollary 6.3. □

Proof of Theorem 5.4. Let $(V, \mathbb{T})$ be order stable and finite. Fix $v \in V$ and let $v_n = H^n v$ for all $n \in \mathbb{N}$. By Lemma 6.2, $v_n \preceq v_{n+1}$ for all $n \in \mathbb{N}$. Since (v_n) is contained in the finite set V_Σ, it must be that $v_{n+1} = v_n$ for some $n \in \mathbb{N}$. But then $H v_n = v_n$, so, by Lemma 6.1, we have v_n is a fixed point of T. Hence, by Theorem 5.1, the fundamental ADP optimality properties hold. By these same properties, the fixed point v_n equals the value function v^*. Thus, we have also shown that HPI converges in finitely many steps. □

Proof of Theorem 5.5. In view of Theorem 5.1, the fundamental ADP optimality properties will hold when T has a fixed point in V. To see that this is true, fix any $v \in V_U$ (which is nonempty by Lemma 3.2) and set $v_n := T^n v$. Since $(V, \mathbb{T})$ is bounded above and V is countably Dedekind complete, there exists a $\bar{v} \in V$ with $v_n \uparrow \bar{v}$. We claim that $T\bar{v} = \bar{v}$. Indeed, $v_{n+1} = T v_n \preceq T\bar{v}$ for all n, so, taking the supremum, $\bar{v} \preceq T\bar{v}$. For the reverse inequality we take σ to be $\bar{v}$-greedy and use order continuity of T_σ to obtain

$$T\bar{v} = T_\sigma \bar{v} = T_\sigma \bigvee_n v_n = \bigvee_n T_\sigma v_n \preceq \bigvee_n T v_n = \bigvee_n v_{n+1} = \bar{v}.$$

The fundamental ADP optimality properties are now proved. In view of these properties, the only fixed point of T in V is v^*. Hence $T^n v = v_n \uparrow \bar{v} = v^*$. This proves convergence of VFI. Convergence of OPI and HPI follow from Corollary 6.3. □

7. Applications

Next we illustrate how the ADP optimality results stated above can be applied.

7.1. **Non-EU Discrete Choice.**

Some studies have found incompatibilities between data and predictions of utility maximization problems founded on additively separable preferences (see, e.g, Lu et al. (2023)). To further this line of analysis, we return to the discrete choice Bellman equation in Section 3.6, which was motivated by structural estimation, while replacing ordinary conditional expectation with a general certainty equivalent operator (so that preferences can fail to be additively separable). In particular, we adopt the setting and assumptions of Section 3.6 while modifying the policy operator (11) to $(T_\sigma g)(x, a) = (\mathcal{E} H_\sigma g)(x, a)$, where

$$(H_\sigma g)(x') := \int \{r(x', \sigma(x'), e') + \beta g(x', \sigma(x'))\} \, \nu(\mathrm{d}e')$$

and $\mathcal{E}$ is a certainty equivalent operator mapping $\mathbb{R}^{\mathsf{X}}$ into $\mathbb{R}^{\mathsf{G}}$. This means that $\mathcal{E}$ is order preserving with respect to the pointwise order and $\mathcal{E}\lambda = \lambda$ whenever λ is constant. (Thus, $\mathcal{E}$ is a generalization of a conditional expectations operator.) We call $\mathcal{E}$ *constant subadditive* if $\mathcal{E}(f + \lambda) \leqslant \mathcal{E}f + \lambda$ for all $f \in \mathbb{R}^{\mathsf{X}}$ and $\lambda \in \mathbb{R}_+$.

Let $\mathbb{T} = \{T_\sigma\}_{\sigma \in \Sigma}$. Each T_σ is order preserving on $\mathbb{R}^{\mathsf{G}}$ under the usual pointwise order, so $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is an ADP. Moreover, given $g \in \mathbb{R}^{\mathsf{G}}$, a policy $\sigma \in \Sigma$ is g-greedy whenever

$$\sigma(x) \in \operatorname*{arg\,max}_{a' \in \mathsf{A}} \int [r(x', a', e') + \beta g(x', a')] \, \nu(\mathrm{d}e') \quad \text{for all } x \in \mathsf{X}.$$

Since A is finite, such a policy always exists. Hence $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is regular.

Proposition 7.1. *If $\mathcal{E}$ is constant subadditive, then the fundamental ADP optimality properties hold and HPI converges in finitely many steps.*

Proof. Since $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is regular and finite, it suffices to show that $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is also order stable (by Theorem 5.4). To this end, fix $f, g \in \mathbb{R}^{\mathsf{G}}$. Since $\mathcal{E}$ and H_σ are order preserving, we have $T_\sigma f = \mathcal{E} H_\sigma (g + f - g) \leqslant \mathcal{E} H_\sigma (g + \|f - g\|) \leqslant \mathcal{E}(H_\sigma g + \beta \|f - g\|)$. Using constant subadditivity of $\mathcal{E}$ and rearranging gives $T_\sigma f - T_\sigma g \leqslant \mathcal{E} \beta \|f - g\| = \beta \|f - g\|$. Reversing the roles of f and g gives $|T_\sigma f - T_\sigma g| \leqslant \beta \|f - g\|$, so each $T_\sigma \in \mathbb{T}$ is a contraction on $\mathbb{R}^{\mathsf{G}}$. Since $\mathbb{R}^{\mathsf{G}}$ is complete under the supremum norm, $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is well-posed. Moreover, from the argument in Example 2.1, $(\mathbb{R}^{\mathsf{G}}, \mathbb{T})$ is order stable. This completes the proof of Proposition 7.1. □

As an illustration, suppose that $\mathcal{E}$ is the risk-sensitive certainty equivalent

$$(\mathcal{E}f)(x,a) := \frac{1}{\theta} \ln \left\{ \int \exp\left[\theta f(x')\right] P(x,a,\mathrm{d}x') \right\} \qquad ((x,a) \in \mathsf{G}),$$

where P is a stochastic kernel from G to X and θ is a nonzero constant. This choice of certainty equivalent is constant subadditive, so, among other things, Proposition 7.1 tells us that $\sigma \in \Sigma$ is optimal if and only if $\sigma(x) \in \arg\max_{a' \in \mathsf{A}} [r(x',a') + \beta g^*(x',a')]$ for all $x \in \mathsf{X}$, where g^* is the unique solution to the functional equation

$$g(x,a) = \frac{1}{\theta} \ln \left\{ \int \exp\left\{ \theta \max_{a' \in \mathsf{A}} \left[r(x',a') + \beta g(x',a') \right] \right\} P(x,a,\mathrm{d}x') \right\}$$

in the value space $\mathbb{R}^{\mathsf{G}}$.[2]

7.2. **Firm Valuation.** We consider a firm valuation problem studied by Jovanovic (1982) with the following extensions: (i) firm profits depend on an aggregate shock, as well as a firm-specific shock and a cross-sectional distribution, (ii) the interest rate is allowed to vary over time, (iii) the outside option of the firm is permitted to depend on aggregates and the cross-section, and shocks and rewards are allowed to be discontinuous and unbounded.

In this version of the problem, a firm that receives current profit $\pi(s,\mu,z)$ and then transitions to the next period, where management will choose to either exit and receive $q(\mu',z')$ or continue. Thus, the maximal expected firm value $v(s,\mu,z)$ obeys

$$v(s,\mu,z) = \pi(s,\mu,z) + \beta(\mu,z)\,\mathbb{E}_{(s,\mu,z)} \max\{q(\mu',z'), v(s',\mu',z')\}. \tag{16}$$

Here s is an idiosyncratic state for the firm that takes values in set S, μ is a cross-sectional distribution taking values in a space D, z is an aggregate shock taking values in set Z, and $\pi(s,\mu,z)$ is current profit. Primes denote next period values. The discount factor β depends on cost of capital and hence the current state. Let $x := (s,\mu,z)$ take values in $\mathsf{X} := S \times D \times Z$. Let $\mathcal{B}$ be a σ-algebra over X that makes π, β and the transition probabilities measurable. We rewrite the dynamics as $x' \sim P(x,\cdot)$,

[2]Another example of a nonlinear certainty equivalent operator is the quantile operator studied in de Castro and Galvao (2019), which allows for separation of intertemporal elasticity of substitution and risk aversion. This certainty equivalent is also constant subadditive, so Proposition 7.1 extends the results in de Castro and Galvao (2019).

meaning that P is a stochastic kernel on $(\mathsf{X}, \mathcal{B})$ and the next period composite state x' is drawn from distribution $P(x, \cdot)$. With this notation, (16) becomes

$$v(x) = \pi(x) + \beta(x) \int \max\{q(x'), v(x')\} P(x, \mathrm{d}x') \qquad (x \in \mathsf{X}). \tag{17}$$

A v that solves (17) gives firm valuation at each state under optimal management.

Let K be the *discount operator* defined by

$$(Kv)(x) := \beta(x) \int v(x') P(x, \mathrm{d}x').$$

We suppose there exists a σ-finite measure φ on $(\mathsf{X}, \mathcal{B})$ such that π, q and β are nonnegative elements of $L_1 := L_1(\mathsf{X}, \mathcal{B}, \varphi)$, and that K maps L_1 to itself. We endow L_1 with the φ-a.e. pointwise order $\leqslant$, so that $f \leqslant g$ means $\varphi\{f > g\} = 0$. In what follows, for any linear operator A on L_1, we use $\rho(A)$ to represent the spectral radius of A. Also, A is called positive if $0 \leqslant v$ implies $0 \leqslant Av$.

Assumption 7.1. The discount operator obeys $\rho(K) < 1$.

Assumption 7.1 is weaker than that found in Hansen and Scheinkman (2012) and related sources, since we impose no irreducibility or compactness conditions on K. (Later, in Proposition 7.3, we show that, when such conditions *are* in force, Assumption 7.1 is both necessary and sufficient for optimality.)

Let Σ be the set of policies, each of which is a $\mathcal{B}$-measurable map σ from X to $\{0, 1\}$. Here $\sigma(x) = 1$ indicates the decision to exit at state x and $\sigma(x) = 0$ indicates the decision to continue. To each $\sigma \in \Sigma$ we assign the policy operator

$$T_\sigma v = \pi + K(\sigma q + (1 - \sigma) v) \tag{18}$$

Since K is positive and hence order preserving, T_σ is order preserving on L_1. Hence $(L_1, \mathbb{T})$ is an ADP when $\mathbb{T} := \{T_\sigma : \sigma \in \Sigma\}$.

Let V be all $v \in L_1$ such that $0 \leqslant v \leqslant \bar{v}$, where $\bar{v} := (I - K)^{-1}(\pi + Kq)$ and I is the identity map ($\bar{v}$ is well-defined by Assumption 7.1). Straightforward arguments show that every T_σ maps V to itself. Hence $(V, \mathbb{T})$ is also an ADP. Since $0 \leqslant K(1 - \sigma) \leqslant K$ we have $\rho(K(1 - \sigma)) \leqslant \rho(K) < 1$, so each T_σ is has a unique fixed point v_σ in V. In particular, $(V, \mathbb{T})$ is well-posed. By definition, its ADP Bellman operator obeys $Tv := \bigvee_{\sigma \in \Sigma} T_\sigma v = \pi + K(q \vee v)$, which coincides with (17). This means that solving the ADP optimization problem is equivalent to solving the original dynamic program with Bellman operator (17).

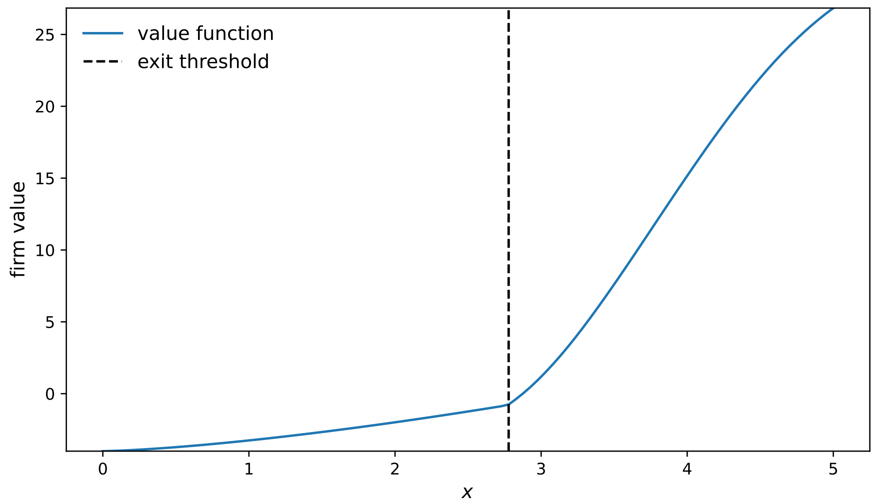

FIGURE 1. Firm value function and exit threshold

Proposition 7.2. *If Assumption 7.1 holds, then the fundamental ADP optimality properties hold and VFI, OPI, and HPI all converge.*

Proof. Given $v \in V$, the set $\{T_\sigma v : \sigma \in \Sigma\}$ has greatest element $\pi + K(q \vee v)$, which is attained by the v-greedy policy $\sigma = \mathbb{1}\{q \geqslant v\}$. Hence $(V, \mathbb{T})$ is regular. Also, K is order continuous because positive linear operators on L_1 are order continuous (see, e.g., Zaanen (2012), Example 21.6). It then follows that each T_σ is order continuous (since the order limit $\uparrow$ is preserved under basic arithmetic operations – see, e.g., Theorem 10.2 of Zaanen (2012)) and, in particular, $(V, \mathbb{T})$ is order continuous. Because $(V, \mathbb{T})$ is regular, well-posed and order continuous, and because V is chain complete (see, e.g., Example 12.5 of Zaanen (2012)), Theorem 5.3 applies. This yields the conclusions of Proposition 7.2. □

Figure 1 shows an approximation of the value function v^* computed by VFI, as well as a representation of a v^*-greedy policy σ in the form of an exit threshhold. For x below the threshold, $\sigma(x) = 0$, indicating that exit is optimal. In this example, the state space is just $\mathbb{R}_+$ and x can be thought of as productivity. The function π is given by

$$\pi(x) = \max_{\ell \geqslant 0} \left\{ p x \ell^\theta - c - \ell \right\},$$

where p is the output price, ℓ represents labor input, θ is a productivity parameter, and c is a fixed cost. The dynamics of x are given by $P(x, \mathrm{d}x) \stackrel{\mathrm{d}}{=} Ax$, where A is lognormal $(-0.012, 0.1)$. We set $\beta = 0.95$, $\theta = 0.3$, and $c = 4$. The outside option q is set to zero.

To show that our assumptions are weak, we now prove that, under some mild conditions, well-posedness fails whenever Assumption 7.1 fails. This means that Assumption 7.1 is necessary for the dynamic program to be well-defined.

Proposition 7.3. *Let π be nonzero and let K be weakly compact and irreducible on L_1. In this setting, if $(V, \mathbb{T})$ is well-posed, then Assumption 7.1 holds.*

Proof. Let π be nonzero and let K be weakly compact and irreducible. Let K' be the adjoint of K and let λ be the spectral radius. By the Krein–Rutman theorem (see, in particular, Lemma 4.2.11 of Meyer-Nieberg (2012)), there exists an $e \in L_\infty$ such that $K'e = \lambda e$ and, in addition, $\langle e, f \rangle > 0$ for all nonzero nonnegative $f \in L_1$. Consider the policy $\sigma \equiv 0$. Under this policy we have $T_\sigma v = \pi + Kv$. If $(V, \mathbb{T})$ is well-posed, then there exists a solution $v \in V$ to $v = \pi + Kv$ in V. Since π is nonzero and $v = \pi + Kv \geqslant \pi$, the same is true for v. Now observe that $\langle e, v \rangle = \langle e, \pi \rangle + \langle e, Kv \rangle = \langle e, \pi \rangle + \langle K'e, v \rangle = \langle e, \pi \rangle + \lambda \langle e, v \rangle$. Since v and π are nonnegative and nonzero, it must be that $\langle e, \pi \rangle > 0$ and $\langle e, v \rangle > 0$. Therefore λ satisfies $(1 - \lambda)\alpha = \beta$ for $\alpha, \beta > 0$. Hence $\lambda < 1$. □

8. Conclusion

The framework constructed in this paper represents dynamic programs as operators over partially ordered sets and allows us to acquire a range of new optimality results that include many existing results as special cases. These methods are suitable for applications with a number of challenging features.

A limitation of our results is that we assumed our dynamic programs are regular. Some dynamic programs do not have this property because certain policies lead to infinite loss (see, e.g., Li and Rantzer (2024), Pates and Rantzer (2024), or Chapters 3–4 of Bertsekas (2022)). Others lack this property due to nonstandard discounting (Balbus et al., 2020; Jaśkiewicz and Nowak, 2021). Extensions of the results in this paper to such problems would be valuable.

Another of our assumptions that could be altered is well-posedness, i.e.,that each T_σ has a unique fixed point v_σ. It could be replaced by generalizing the approach in Bertsekas (2022), where v_σ is defined by $v_\sigma(x) := \limsup_k T_\sigma^k \bar{v}(x)$ for some fixed reference point $\bar{v} \in V$. The limsup could be generalized to an abstract partially ordered set environment by setting $v_\sigma := \wedge_{n \geqslant 1} \vee_{k \geqslant n} T_\sigma^k \bar{v}$. The element v_σ would always be well-defined if, say, V is a complete lattice. We have not yet explored this modification of our framework, but think it would be worthwhile.

We have focused on applications and theoretical settings where optimal policies always exist. If one wishes to consider approximately optimal policies, then some metric on the value space must be added in order to measure approximations. A promising path forward would be to replace the assumption that V is an arbitrary partially ordered set with the assumption that V is a partially ordered space; that is, a metric space with partial order $\preceq$ such that the order $\preceq$ is preserved under limits.

Many further extensions could be built on top of our framework. One example is average-cost optimality for dynamic programs, which we have not considered. Another is continuous time models. These topics are also left for future work.

References

Balbus, Ł., Jaśkiewicz, A., and Nowak, A. S. (2020). Markov perfect equilibria in a dynamic decision model with quasi-hyperbolic discounting. *Annals of Operations Research*, 287:573–591.

Bäuerle, N. and Jaśkiewicz, A. (2018). Stochastic optimal growth model with risk sensitive preferences. *Journal of Economic Theory*, 173:181–200.

Bäuerle, N., Jaśkiewicz, A., and Nowak, A. S. (2021). Stochastic dynamic programming with non-linear discounting. *Applied Mathematics & Optimization*, 84(3):2819–2848.

Bäuerle, N. and Rieder, U. (2011). *Markov decision processes with applications to finance*. Springer Science & Business Media.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR.

Bellman, R. (1957). *Dynamic programming*. American Association for the Advancement of Science.

Bertsekas, D. (2012). *Dynamic programming and optimal control*, volume 1. Athena Scientific.

Bertsekas, D. (2021). *Rollout, policy iteration, and distributed reinforcement learning*. Athena Scientific.

Bertsekas, D. P. (1977). Monotone mappings with application in dynamic programming. *SIAM Journal on Control and Optimization*, 15(3):438–464.

Bertsekas, D. P. (2022). *Abstract dynamic programming*. Athena Scientific, 3 edition.

Bloise, G., Le Van, C., and Vailakis, Y. (2024). Do not blame Bellman: It is Koopmans' fault. *Econometrica*, 92(1):111–140.

Cagetti, M., Hansen, L. P., Sargent, T. J., and Williams, N. (2002). Robustness and pricing with uncertain growth. *The Review of Financial Studies*, 15(2):363–404.

Davey, B. A. and Priestley, H. A. (2002). *Introduction to lattices and order*. Cambridge University Press.

de Castro, L. and Galvao, A. F. (2019). Dynamic quantile models of rational behavior. *Econometrica*, 87(6):1893–1939.

Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *Siam Review*, 9(2):165–177.

Epstein, L. G. and Zin, S. E. (1989). Risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, 57(4):937–969.

Fei, Y., Yang, Z., Chen, Y., and Wang, Z. (2021). Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 34:20436–20446.

Gerstenberg, J., Neininger, R., and Spiegel, D. (2023). On solutions of the distributional Bellman equation.

Hansen, L. P. and Sargent, T. J. (1995). Discounted linear exponential quadratic Gaussian control. *IEEE Transactions on Automatic control*, 40(5):968–971.

Hansen, L. P. and Sargent, T. J. (2011). *Robustness*. Princeton university press.

Hansen, L. P. and Scheinkman, J. A. (2012). Recursive utility in a Markov environment with stochastic growth. *Proceedings of the National Academy of Sciences*, 109(30):11967–11972.

Haskell, W. B., Jain, R., and Kalathil, D. (2016). Empirical dynamic programming. *Mathematics of Operations Research*, 41(2):402–429.

Hernández-Lerma, O. and Lasserre, J. B. (2012). *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media.

Hubmer, J., Krusell, P., and Smith, Jr, A. A. (2020). Sources of us wealth inequality: Past, present, and future. *NBER Macroeconomics Annual 2020, volume 35*.

Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280.

Jaśkiewicz, A. and Nowak, A. S. (2021). Markov decision processes with quasi-hyperbolic discounting. *Finance and Stochastics*, 25(2):189–229.

Jovanovic, B. (1982). Selection and the evolution of industry. *Econometrica*, pages 649–670.

Kalathil, D., Borkar, V. S., and Jain, R. (2021). Empirical q-value iteration. *Stochastic Systems*, 11(1):1–18.

Kochenderfer, M. J., Wheeler, T. A., and Wray, K. H. (2022). *Algorithms for decision making*. The MIT Press.

Kristensen, D., Mogensen, P. K., Moon, J. M., and Schjerning, B. (2021). Solving dynamic discrete choice models using smoothing and sieve methods. *Journal of Econometrics*, 223(2):328–360.

Li, Y. and Rantzer, A. (2024). Exact dynamic programming for positive systems with linear optimal cost. *IEEE Transactions on Automatic Control*.

Lu, J., Luo, Y., Saito, K., and Xin, Y. (2023). Does Harold Zurcher have time separable preferences? Technical report, North American Summer Meeting of the Econometric Society.

Meyer-Nieberg, P. (2012). *Banach lattices*. Springer Science & Business Media.

Mogensen, P. K. (2018). Solving dynamic discrete choice models: Integrated or expected value function? *arXiv preprint arXiv:1801.03978*.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Naik, A., Shariff, R., Yasui, N., Yao, H., and Sutton, R. S. (2019). Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*.

Pates, R. and Rantzer, A. (2024). Optimal control on positive cones. Technical report, arXiv:2407.18774.

Powell, W. B. (2016). Perspectives of approximate dynamic programming. *Annals of Operations Research*, 241:319–356.

Puterman, M. L. (2005). *Markov decision processes: discrete stochastic dynamic programming*. Wiley Interscience.

Ren, G. and Stachurski, J. (2021). Dynamic programming with value convexity. *Automatica*, 130:109641.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, pages 999–1033.

Rust, J. (1994). Structural estimation of Markov decision processes. *Handbook of Econometrics*, 4:3081–3143.

Rust, J., Traub, J. F., and Wozniakowski, H. (2002). Is there a curse of dimensionality for contraction fixed points in the worst case? *Econometrica*, 70(1):285–329.

Sidford, A., Wang, M., Wu, X., and Ye, Y. (2023). Variance reduced value iteration and faster algorithms for solving Markov decision processes. *Naval Research Logistics (NRL)*, 70(5):423–442.

Toda, A. A. (2023). Unbounded Markov dynamic programming with weighted supremum norm Perov contractions.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Technical report, PhD Thesis, King's College, Cambridge United Kingdom.

Yu, Z., Dai, L., Xu, S., Gao, S., and Ho, C. P. (2024). Fast Bellman updates for Wasserstein distributionally robust MDPs. *Advances in Neural Information Processing Systems*, 36.

Zaanen, A. C. (2012). *Introduction to operator theory in Riesz spaces*. Springer.

Research School of Economics, Australian National University

Email address: john.stachurski@anu.edu.au

Department of Economics, New York University

Email address: ts43@nyu.edu