DYNAMIC PROGRAMMING: FROM LOCAL OPTIMALITY TO GLOBAL OPTIMALITY

JOHN STACHURSKI*, JINGNI YANG†, ZIYUE YANG‡

ABSTRACT. In the theory of dynamic programming, an optimal policy is a policy whose lifetime value dominates that of all other policies from every possible initial condition in the state space. This raises a natural question: when does optimality from a single state imply optimality from every state? Working in a general setting, we provide sufficient conditions for this property that relate to reachability and irreducibility. Our results are of theoretical interest and also have implications for modern policy-based algorithms used to solve large-scale dynamic programs.

1. INTRODUCTION

Dynamic programming is a major branch of optimization theory with a vast and growing range of applications (see, e.g., Hernández-Lerma and Lasserre (1996), Bertsekas (2012), or Bertsekas (2022)). Within economics and finance, applications extend from monetary and fiscal policy to asset pricing, unemployment, firm investment, wealth dynamics, commodity pricing, sovereign default, the division of labor, environmental economics, natural resource extraction, human capital accumulation, retirement decisions, portfolio choice, and dynamic pricing (see, e.g., Stokey and Lucas (1989), Rust (1996), Bäuerle and Rieder (2011), Ljungqvist and Sargent (2018), or Bloise et al. (2023)). Recently there has been a surge in interest in dynamic programming fueled by artificial intelligence, where such programs form the core of reinforcement learning techniques (see, e.g., Bertsekas (2021) or Kochenderfer et al. (2022)).

In this paper we address a question on the foundations of dynamic programming. To state it, we consider a dynamic program with state space X and set of feasible policies Σ. Each σ in Σ is a map from the state space X to the action space A. Associated to σ is a function v_σ on X, where $v_\sigma(x)$ denotes the lifetime value of applying the

*Research School of Economics, Australian National University. john.stachurski@anu.edu.au.

†School of Economics, University of Sydney. jingni.yang@sydney.edu.au.

‡Research School of Economics, Australian National University. humphrey.yang@anu.edu.au.

policy σ in every period when starting at initial state x. (For example, consider a simple household problem, where σ maps wealth into consumption. If x is initial wealth, then $v_\sigma(x)$ indicates expected discounted utility conditional on starting with wealth x and always following policy σ.) Recall that a policy σ is called *optimal* if $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for every $x \in \mathsf{X}$; that is, if σ is such that following this policy in every period leads to maximum lifetime value from every initial state x.

We now ask the following question: when is it true that $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for some $x \in \mathsf{X}$ implies $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for every $x \in \mathsf{X}$? In other words, when is it true that $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for some $x \in \mathsf{X}$ implies that σ is an optimal policy?[1]

This implication does not always hold. To see this, consider the shortest path problem in Figure 1, where the aim is to traverse the graph to destination node G. States are nodes and actions are arrows. The number on each arrow is a flow reward from choosing that action. The blue lines represent a fixed policy σ. Under this policy, the lifetime value $v_\sigma(A)$ of σ starting from A is -8 (the sum of -3, -4 and -1, assuming no discounting). This cannot be improved upon, conditional on starting at A, so $v_\sigma(A) = \max_{s\in\Sigma} v_s(A)$ holds. At the same time, σ is not optimal — for example, σ performs poorly when starting at node B.

Nevertheless, we can envisage dynamic programs when local optimality might imply (global) optimality. For example, suppose we freeze x and solve $\max_\sigma v_\sigma(x)$. This value depends on present and future rewards, so maximizing it over the set of feasible policies requires choosing a σ that performs well in states we are likely to visit over the lifetime of the program. If we travel "all over" the state space X under σ, then we are incentivized to choose a policy σ that performs well everywhere. This leads us to the conjecture that local optimality implies global optimality whenever the dynamics under σ exhibit sufficient mixing.

(The phrase "all over" from the previous paragraph is in quotation marks because the state spaces we deal with are, in many cases, uncountably infinite, and the probability of the state hitting any one given point is often zero. Hence, to obtain general results, it is necessary for us to consider how to implement such ideas in ways that are sufficient for our purposes.)

[1] In some instances in the paper, we refer to optimality of a policy as "global optimality" in order to strengthen the contrast between this property and local optimality (i.e., policies that are maximal in terms of lifetime value at some point in the state space).

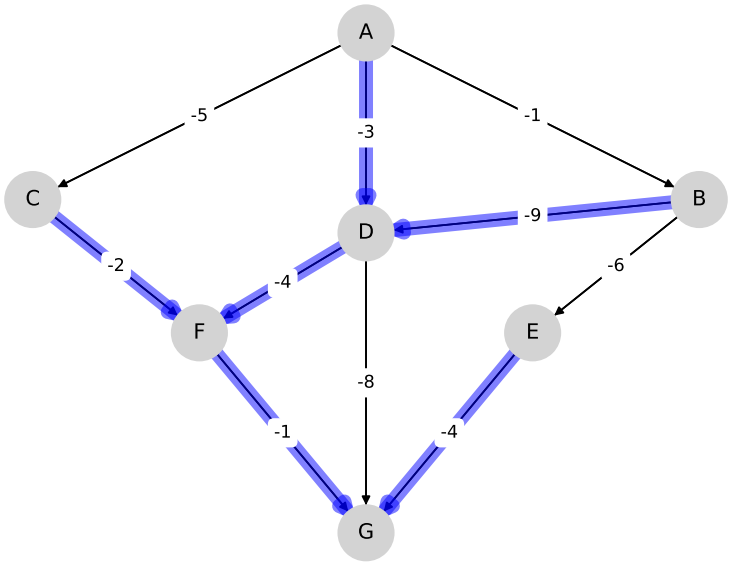

FIGURE 1. A shortest path problem with destination node G

The results in our paper address the conjecture stated above. More specifically, we address the following questions:

(a) Under what conditions does $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for some given $x \in \mathsf{X}$ imply that σ is an optimal policy?

(b) Under what conditions does $\int v_\sigma(x)\rho(\mathrm{d}x) = \max_{s\in\Sigma} \int v_s(x)\rho(\mathrm{d}x)$ for some distribution ρ on X imply that σ is an optimal policy?

We show that, for standard programs on general state spaces, irreducibility of the Markov dynamics generated by σ is sufficient for these implications. For example, if a policy is maximal at a single state and has an irreducible transition kernel, then this optimality propagates throughout the entire state space, making the policy globally optimal. In addition, we show that weaker forms of irreducibility are sufficient for this result when σ is also continuous. Finally, we provide economic applications that illustrate our results, including inventory problems, optimal savings problems, and optimal stopping problems.

Regarding motivation, the question of when local optimality implies (global) optimality is important for three reasons. First, it illuminates the foundations of a core branch of optimization theory. Second, for a number of foundational economic models, equilibria can be realized as solutions to dynamic programs, so the results in the

paper can be used to address global properties of endogenous objects such as prices from purely local properties. Third, the question has relevance to modern methods for solving dynamic programs (and associated reinforcement learning) routinely employed for high-dimensional problems.

To understand the third point, consider the fact that many researchers solving large-scale dynamic programs have, in recent years, moved away from traditional value-based methods and towards policy-based methods, such as policy gradient ascent (see, e.g., Murphy (2024)). These policy-based methods seek to maximize a real-valued objective, such as

$$m(\sigma) \coloneqq \int v_\sigma(x)\rho(\mathrm{d}x) \qquad (\sigma \in \Sigma). \tag{1}$$

for some choice of "initial distribution" ρ. Policy gradient methods take a current guess σ_n for the maximizer in (1) and iterate on

$$\sigma_{n+1} = \sigma_n + \lambda_n \nabla m(\sigma_n), \qquad n = 0, 1, \ldots,$$

where λ_n is a step size and $\nabla m(\sigma_n)$ is the gradient of m evaluated at the current guess σ_n. (In practice, each σ is represented by a neural network and the method amounts to training a neural net when σ is the parameters of the network and $-m(\sigma)$ is the loss function. The basic gradient ascent method is typically modified to current best-practice for training neural nets.) Policy-based methods often handle large problems with continuous state and action spaces more efficiently than more traditional value-based methods. This is at least partly because many researchers and engineering teams now have access to enormously powerful tools for training neural nets. (Closely related ideas have been applied to high-dimensional economic problems by Han and Yang (2021), Friedl et al. (2023), and Payne et al. (2025).)

The gradient method described above begins by choosing the initial distribution ρ that defines the integrated lifetime value function m in (1). This step is essential to the methodology because it shifts the problem from finding the greatest element in a partially ordered set (finding the σ such that $v_s \leqslant v_\sigma$ for all $s \in \Sigma$) to maximizing the real-valued function m (which is amenable to gradient ascent after approximating policy functions with neural nets). At the same time, it has the following disadvantage: when one maximizes the objective $m(\sigma) \coloneqq \int v_\sigma(x)\rho(\mathrm{d}x)$, the outcome depends on the initial conditional ρ. Our research question (b) above seeks to provide conditions under which maximizing m leads to an optimal policy.

Papers examining theoretical properties of policy gradient methods have some connection to our work. For example, Bhandari and Russo (2024) examine when policy gradient ascent (actually descent) yields a globally optimal policy. To step from local to global optimality, they restrict the classes of dynamic programs under consideration and require that the initial distribution ρ dominates the discounted state occupancy measure under the optimal policy (i.e., the measure is absolutely continuous with respect to ρ). This condition is difficult to verify in practice, since the optimal policy is *a priori* unknown (and, in fact, the object of the whole computational procedure). Here we avoid both large support restrictions on ρ and explicit restrictions on the optimal policy. (At the same time, Bhandari and Russo (2024) supply many valuable new results concerning the policy gradient algorithm, and we make no direct contribution to this analysis.)

The paper is structured as follows. Section 2 provides background on dynamic programming. Section 3 states our main results in a general setting. Section 4 examines how the irreducibility condition from Section 3 can be weakened while still obtaining some transmission of optimality across states. Section 6.2 illustrates our theoretical results in the context of a benchmark optimal savings problem. Section 7 outlines avenues for future work.

2. Set Up

In this section, we review essential properties of Markov decision processes and state a technical lemma that will be applied in our main results.

2.1. **Preliminaries.**

Let X and A be metric spaces, let $\mathscr{B}$ be the Borel subsets of X, let $b\mathsf{X}$ be the set of bounded Borel measurable functions from X to $\mathbb{R}$, and let $cb\mathsf{X}$ be the continuous functions in $b\mathsf{X}$. Both $b\mathsf{X}$ and $cb\mathsf{X}$ are paired with the supremum norm $\|\cdot\|$ and the pointwise partial order $\leqslant$. For example, $f \leqslant g$ indicates that $f(x) \leqslant g(x)$ for all $x \in \mathsf{X}$. A map M from $b\mathsf{X}$ to itself is called *order preserving* if $f \leqslant g$ implies $Mf \leqslant Mg$. Absolute values are applied pointwise, so that $|f|$ is the function $x \mapsto |f(x)|$.

A *transition kernel* on X is a function Q from $\mathsf{X} \times \mathscr{B}$ to $[0, 1]$ such that $x \mapsto Q(x, B)$ is Borel measurable for all $B \in \mathscr{B}$ and $B \mapsto Q(x, B)$ is a Borel probability measure for

all $x \in \mathsf{X}$. To any such kernel Q we associate a bounded linear operator on $b\mathsf{X}$, often referred to as its *Markov operator* and also denoted by Q, via

$$f \mapsto Qf, \qquad (Qf)(x) = \int f(x')Q(x, \mathrm{d}x'). \tag{2}$$

Typically, $(Qf)(x)$ represents the expectation of $f(X_{t+1})$ given that $X_t = x$ and X_{t+1} is drawn from $Q(x, \mathrm{d}x')$.

As usual, the positive cone of $b\mathsf{X}$, denoted here by $b\mathsf{X}_+$, is all $v \in b\mathsf{X}$ with $v \geqslant 0$. Let $b\mathsf{X}'$ be the dual space of $b\mathsf{X}$ and let $b\mathsf{X}'_+$ be the positive cone of $b\mathsf{X}'$. The set $b\mathsf{X}'_+$ contains, among other objects, the set $\mathcal{D}(\mathsf{X})$ of Borel probability measures on X. For simplicity, elements of $\mathcal{D}(\mathsf{X})$ are referred to as *distributions*. For $\rho \in \mathcal{D}(\mathsf{X})$ and $f \in b\mathsf{X}$ we set

$$\langle f, \rho \rangle := \int f \, \mathrm{d}\rho.$$

Following standard terminology, given $\mu \in \mathcal{D}(\mathsf{X})$, the *support* of μ is the intersection of all closed sets F such that $\mu(F) = 1$. We denote the support of μ by $\operatorname{supp}(\mu)$.

For each $x \in \mathsf{X}$, the *point evaluation functional* generated by x is the map δ_x that sends $w \in b\mathsf{X}$ into $w(x) \in \mathbb{R}$. Below it will be convenient for us to write this in dual notation, so that $\langle w, \delta_x \rangle = w(x)$ for every $w \in b\mathsf{X}$. We will make use of the following lemma.

Lemma 2.1. *Every point evaluation functional on $b\mathsf{X}$ is a nonzero element of $b\mathsf{X}'_+$.*

Proof. Fix $x \in \mathsf{X}$. Linearity of δ_x is obvious: given $a, b \in \mathbb{R}$ and $v, w \in b\mathsf{X}$, we have

$$\langle av + bw, \delta_x \rangle = (av + bw)(x) = av(x) + bw(x) = a \langle v, \delta_x \rangle + b \langle w, \delta_x \rangle .$$

Regarding continuity, if $w_n \to w$ in $b\mathsf{X}$, then $w_n \to w$ pointwise on X, so $\langle w_n, \delta_x \rangle = w_n(x) \to w(x) = \langle w, \delta_x \rangle$. Regarding positivity, it suffices to show that $\langle w, \delta_x \rangle \geqslant 0$ whenever $w \geqslant 0$. This clearly holds, since $w \geqslant 0$ implies $w(x) = \langle w, \delta_x \rangle \geqslant 0$. Finally, δ_x is not the zero element of $b\mathsf{X}'$ because $w = \mathbf{1}$ is in $b\mathsf{X}$ and $\langle w, \delta_x \rangle = w(x) = 1 \neq 0$. □

2.2. **Markov Decision Process.** Let X and A be metric spaces, as in Section 2.1. A *Markov decision process* (MDP) with state space X and action space A is a tuple (r, Γ, β, P), where r is a reward function, $x \mapsto \Gamma(x) \subset \mathsf{A}$ is a feasible correspondence, $\beta \in \mathbb{R}$ is a discount factor and $P(x, a, \mathrm{d}x')$ is a distribution over next period states given current state x and action a. Let $\mathsf{G} := \{(x, a) \in \mathsf{X} \times \mathsf{A} : a \in \Gamma(x)\}$. We consider

a relatively standard environment, as considered in, say, Bäuerle and Rieder (2011) and Hernández-Lerma and Lasserre (1996), where

(a) $\beta \in (0,1)$,
(b) Γ is nonempty, continuous, and compact-valued on X,
(c) r is bounded and continuous on G, and
(d) the map $(x,a) \mapsto \int v(x')P(x,a,dx')$ is continuous on G whenever $v \in cb\mathsf{X}$.

(The case of unbounded r is discussed in Section 7.)

Let Σ denote the set of feasible policies, by which we mean all Borel measurable functions σ mapping X to A with $\sigma(x) \in \Gamma(x)$ for all $x \in \mathsf{X}$. For each $\sigma \in \Sigma$ and $x \in \mathsf{X}$, we set

$$r_\sigma(x) := r(x,\sigma(x)) \quad \text{and} \quad P_\sigma(x,\mathrm{d}x') := P(x,\sigma(x),\mathrm{d}x').$$

Thus, $r_\sigma(x)$ is the reward at x under policy σ and P_σ is the transition kernel on X generated by σ. Using the corresponding Markov operator P_σ, as defined in (2), the *lifetime value* of a policy σ, denoted by v_σ, can be expressed as

$$v_\sigma = \sum_{t=0}^{\infty} (\beta P_\sigma)^t r_\sigma \tag{3}$$

(see, e.g., Puterman (2014), Theorem 6.1.1). We will use the fact that v_σ is also the unique fixed point in $b\mathsf{X}$ of the *policy operator* T_σ defined by $T_\sigma v = r_\sigma + \beta P_\sigma v$. This operator can be written more explicitly as

$$(T_\sigma v)(x) = r(x,\sigma(x)) + \beta \int v(x')P(x,\sigma(x),\mathrm{d}x') \qquad (v \in b\mathsf{X},\ x \in \mathsf{X}).$$

(The lifetime value v_σ is the unique fixed point of T_σ because the spectral radius of the linear operator βP_σ is β, so, by the geometric series lemma, $\beta < 1$ implies that $v = r_\sigma + \beta P_\sigma v$ has the unique solution given by the right-hand side of (3).) Iterating on the definition $T_\sigma v = r_\sigma + \beta P_\sigma v$, we find that

$$T_\sigma^n v = r_\sigma + \beta P_\sigma r_\sigma + \cdots + (\beta P_\sigma)^{n-1} r_\sigma + (\beta P_\sigma)^n v \quad \text{for all } n \in \mathbb{N}. \tag{4}$$

This expression will be useful in the theory below.

The *value function* is denoted v^* and defined at each $x \in \mathsf{X}$ by $v^*(x) := \sup_{\sigma\in\Sigma} v_\sigma(x)$. A policy σ is called *optimal* if $v_\sigma(x) = v^*(x)$ for all $x \in \mathsf{X}$.

We define the Bellman operator by

$$(Tv)(x) = \max_{a \in \Gamma(x)} \left\{ r(x,a) + \beta \int v(x') P(x,a,\mathrm{d}x') \right\} \qquad (v \in b\mathsf{X},\ x \in \mathsf{X}). \tag{5}$$

We will use the following facts:

Proposition 2.2. *Under the stated assumptions,*

(a) *the value function v^* is the unique fixed point of the Bellman operator in $b\mathsf{X}$,*
(b) *the value function v^* is well-defined and contained in $cb\mathsf{X}$, and*
(c) *at least one optimal policy exists.*

Proof. See Theorem 7.3.5 and Theorem 7.3.6 of Bäuerle and Rieder (2011) and Theorem 4.2.3 of Hernández-Lerma and Lasserre (1996). □

We also make use of the following technical lemma, which shows one implication of local optimality at some given $x \in \mathsf{X}$.

Lemma 2.3. *If $\sigma \in \Sigma$ and $v_\sigma(x) = v^*(x)$, then*

$$\sum_{n \in \mathbb{N}} \int (v^*(x') - v_\sigma(x')) P_\sigma^n(x, \mathrm{d}x') = 0. \tag{6}$$

Proof. Fix $n \in \mathbb{N}$. Applying the expression for $T_\sigma^n v$ from (4) twice, first with $v = v_\sigma$ and then with $v = v^*$, we get

$$T_\sigma^n v_\sigma - T_\sigma^n v^* = \beta^n (P_\sigma^n v_\sigma - P_\sigma^n v^*). \tag{7}$$

In addition, we have

$$v_\sigma = T_\sigma^n v_\sigma \leqslant T_\sigma^n v^* \leqslant T^n v^* = v^*. \tag{8}$$

In (8), the first inequality is due to the fact T_σ is order preserving and $v_\sigma \leqslant v^*$, while the second follows from the fact that $T_\sigma v \leqslant Tv$ for all $v \in b\mathsf{X}$. (Since $T_\sigma v \leqslant T v$ for all v, $T_\sigma^2 v^* \leqslant T_\sigma T v^* \leqslant T^2 v^*$ and so $T_\sigma^2 v^* \leqslant T^2 v^*$. By induction, the second inequality holds.) The claim in Lemma 2.3 follows from (7) and (8). To see this, fix $x \in \mathsf{X}$ with $v_\sigma(x) = v^*(x)$. From this equality and (8) we get $(T_\sigma^n v_\sigma)(x) = (T_\sigma^n v^*)(x)$. Since $\beta > 0$, combining this result with (7) yields $(P_\sigma^n v_\sigma)(x) = (P_\sigma^n v^*)(x)$. Hence (6) holds. □

3. From Local to Global Optimality

The standard definition of optimality, which was given in Section 2.2, is global in nature, since it concerns the lifetime value of the policy at every $x \in \mathsf{X}$. We seek conditions under which local optimality implies global optimality. In particular, we seek conditions under which the following three statements are equivalent:

(E1) there exists an $x \in \mathsf{X}$ such that $v_s(x) \leqslant v_\sigma(x)$ for all $s \in \Sigma$,
(E2) there exists a $\rho \in \mathcal{D}(\mathsf{X})$ such that $\langle v_s, \rho \rangle \leqslant \langle v_\sigma, \rho \rangle$ for all $s \in \Sigma$,
(E3) σ is an optimal policy.

In studying these conditions, we always work in the setting of Section 2.2. In particular, conditions (a)–(d) are always in force.

3.1. **Preliminary Results.** We begin with the following observation, which is always valid in the MDP set up we have described.

Lemma 3.1. *For all $\sigma \in \Sigma$, the statements (E1) and (E2) are equivalent.*

Proof. To show (E1) implies (E2), assume (E1) and fix $x \in \mathsf{X}$ with $v_\sigma(x) \geqslant v_s(x)$ for all $s \in \Sigma$. Let δ_x be the point evaluation functional generated by x. Since $\delta_x \in \mathcal{D}(\mathsf{X})$ and $\langle v_\sigma, \delta_x \rangle = v_\sigma(x) \geqslant v_s(x) = \langle v_s, \delta_x \rangle$ for all $s \in \Sigma$, and so (E2) holds. To show (E2) implies (E1), fix $\rho \in \mathcal{D}(\mathsf{X})$ with $\langle v_\sigma, \rho \rangle \geqslant \langle v_s, \rho \rangle$ for all $s \in \Sigma$. Since at least one $s \in \Sigma$ is optimal (Proposition 2.2), this yields $\langle v_\sigma, \rho \rangle \geqslant \langle v^*, \rho \rangle$. Now suppose to the contrary that for each $x \in \mathsf{X}$, we can find a $\tau \in \Sigma$ such that $v_\tau(x) > v_\sigma(x)$. Since $v^*(x) \geqslant v_s(x)$ for all $s \in \Sigma$ and $x \in \mathsf{X}$, we have $v^*(x) > v_\sigma(x)$ for all $x \in \mathsf{X}$. Hence $\langle v^*, \rho \rangle > \langle v_\sigma, \rho \rangle$. This contradiction proves (E1). □

3.2. **Irreducibility.** Recalling the discussion in the introduction, our conjecture is that (E1)–(E3) are equivalent under sufficient mixing. The natural way to consider mixing in a general Markov environment is through irreducibility. We begin here with the most standard notion of irreducibility in our general setting. We prove that (E1)–(E3) are indeed equivalent under this standard notion.

To state this result, we recall that a linear subspace I of $b\mathsf{X}$ is called an *ideal* in $b\mathsf{X}$ when $f \in I$ and $|g| \leqslant |f|$ implies $g \in I$. An ideal I is said to be *invariant* for a linear operator K if $KI \subset I$. A linear operator K from $b\mathsf{X}$ to itself is called *positive* when

$Kf \geqslant 0$ for all $f \geqslant 0$. A positive linear operator K is called *irreducible* if the only invariant ideals under K are the trivial subspace $\{0\}$ and the whole space $b\mathsf{X}$. By Proposition 8.3 (c) of Schaefer (1974), such a K is irreducible if and only if, for each nonzero $f \in b\mathsf{X}_+$ and each nonzero $\mu \in b\mathsf{X}'_+$, there exists an $m \in \mathbb{N}$ with $\langle \mu, K^m f \rangle > 0$.

In what follows, we call a transition kernel P on X *irreducible* if its Markov operator (see (2)) is irreducible on $b\mathsf{X}$ in the sense just defined. (When X is finite, this definition reduces to the more familiar one involving communication of all states, as discussed in Section 6.1.) We now state our main result for the irreducible case. In the statement, σ is any feasible policy.

Theorem 3.2. *If P_σ is irreducible, then (E1)–(E3) are equivalent.*

For example, Theorem 3.2 tells us that, under the stated conditions, we can obtain an optimal policy by fixing an arbitrary initial state $x \in \mathsf{X}$ and maximizing the real-valued function $s \mapsto v_s(x)$ over Σ. Alternatively, we can fix any distribution ρ and maximize $s \mapsto \langle v_s, \rho \rangle$.

Proof of Theorem 3.2. By Lemma 3.1, it suffices to show that (E1) and (E3) are equivalent. That (E3) implies (E1) is immediate from the definition of optimal policies. Hence we need only show that (E1) implies (E3). In line with the conditions of Theorem 3.2, we assume that σ is a feasible policy and P_σ is irreducible. Using (E1), we take $x \in \mathsf{X}$ with $v_\sigma(x) = v^*(x)$.

Let $h := v^* - v_\sigma$. By the definition of v^* we have $0 \leqslant h$. We claim in addition that $h = 0$. To see this, suppose to the contrary that h is nonzero. In this case, by irreducibility, for each nonzero μ in the positive cone of $b\mathsf{X}'$ we can find an $m \in \mathbb{N}$ such that $\langle \mu, P_\sigma^m h \rangle > 0$. Because δ_x is a nonzero element of the positive cone of $b\mathsf{X}'$ (Lemma 2.1), we can set $\mu = \delta_x$ to obtain an $m \in \mathbb{N}$ with $(P_\sigma^m h)(x) > 0$. This contradicts (6), so $h = 0$ holds. In other words, $v_\sigma = v^*$. This proves (E3). □

4. Topological Conditions

The discussion in Section 6.1 shows that irreducibility cannot be dropped without either (a) weakening the conclusions of Theorems 3.2, or (b) adding some side conditions. In this section, we investigate both scenarios. In particular, we show that

(a) even when irreducibility fails, optimality can pass across *some* states under a continuity condition and a type of "local irreducibility," and
(b) when seeking the full global conclusions of Theorem 3.2, we can drop strong irreducibility if we assume a weaker form of irreducibility and pair it with continuity.

The first topic is treated in Section 4.1. The second is treated in Sections 4.2 and 4.3.

4.1. **Reachable States.** To begin, we return to the general MDP setting from Section 2.2, where X and A are arbitrary metric spaces. Letting Q be any transition kernel on X, a point $y \in \mathsf{X}$ is called *Q-reachable* from $x \in \mathsf{X}$ when, for each open neighborhood G of y, there exists an $n \in \mathbb{N}$ with $Q^n(x, G) > 0$. In addition, we will say that y is *Q-reachable* from $D \subset \mathsf{X}$ if y is Q-reachable from some $x \in D$.

Lemma 4.1. *Let σ be any continuous policy. If $v_\sigma(x) = v^*(x)$ and y is P_σ-reachable from x, then $v_\sigma(y) = v^*(y)$.*

Proof. As a preliminary step, we show that $h := v^* - v_\sigma$ is continuous under the stated assumptions. Since σ is continuous, our conditions on (r, Γ, β, P) imply that the mappings $x \mapsto \int v(x')P(x, \sigma(x), dx')$ and $x \mapsto r(x, \sigma(x))$ are continuous on X whenever $v \in cb\mathsf{X}$. This implies that T_σ is invariant on $cb\mathsf{X}$. Moreover, $cb\mathsf{X}$ is a closed subset of the complete metric space $b\mathsf{X}$ under the supremum norm (since uniform limits of continuous functions are continuous). In addition, given $v, w \in b\mathsf{X}$, we have

$$\begin{aligned} \|T_\sigma v - T_\sigma w\| &\leqslant \beta \sup_{x \in \mathsf{X}} \int |v(x') - w(x')| P(x, \sigma(x), \mathrm{d}x') \\ &\leqslant \beta \sup_{x \in \mathsf{X}} \int \|v - w\| P(x, \sigma(x), \mathrm{d}x') = \beta \|v - w\|. \end{aligned}$$

Since $\beta < 1$, the contraction mapping theorem implies that $T_\sigma^n w \to v_\sigma$ for every $w \in b\mathsf{X}$. If we now fix $w \in cb\mathsf{X}$ and use the fact that T_σ is invariant on this set, we obtain a sequence $(T^n w)_{n \in \mathbb{N}}$ converging to v_σ and entirely contained in $cb\mathsf{X}$. As $cb\mathsf{X}$ is closed in $b\mathsf{X}$, this implies that v_σ is in $cb\mathsf{X}$. In particular, v_σ is continuous. As v^* is also continuous (see Proposition 2.2), we see that h is continuous.

Now fix $x \in \mathsf{X}$ with $v_\sigma(x) = v^*(x)$. Seeking a contradiction, we suppose that y is P_σ-reachable from x and yet h obeys $h(y) > 0$. By this continuity and $h(y) > 0$, there

exists an open neighborhood G of y with $h > 0$ on G. Because y is P_σ-reachable from x, there exists an $n \in \mathbb{N}$ with $P_\sigma^n(x, G) > 0$. As a result, we have

$$\int (v^*(x') - v_\sigma(x'))P_\sigma^n(x, \mathrm{d}x') \geqslant \int_G h(x')P_\sigma^n(x, \mathrm{d}x') > 0.$$

But $v_\sigma(x) = v^*(x)$, so this inequality contradicts Lemma 2.3. The contradiction proves Lemma 4.1. □

4.2. **Open Set Irreducibility.** The results in Section 4.1 discussed forms of "local" irreducibility and their implications. In this section, we analyze settings where these local conditions extend across the whole space and policies are continuous.

In general, a transition kernel Q from X to itself is called *open set irreducible* if every $y \in \mathsf{X}$ is reachable from every $x \in \mathsf{X}$. For continuous policies that generate open set irreducible transitions, we have the following result.

Theorem 4.2. *If P_σ is open set irreducible and σ is continuous, then (E1)–(E3) are equivalent.*

Proof. By Lemma 3.1, it suffices to show (E1) implies (E3). So fix $x \in \mathsf{X}$ and suppose that $v_\sigma(x) = v^*(x)$. For any $y \in \mathsf{X}$, open set irreducibility implies that y is P_σ-reachable from x. Hence, by Lemma 4.1, we have $v_\sigma(y) = v^*(y)$. In particular, (E3) holds. □

4.3. **π-Irreducibility.** We treat one more form of irreducibility, due to its importance in the literature on Markov dynamics. In general, given a nontrivial measure π on $(\mathsf{X}, \mathscr{B})$, a transition kernel Q on X is called *π-irreducible* if, for each $x \in \mathsf{X}$ and every Borel set $B \subset \mathsf{X}$ with $\pi(B) > 0$, there exists an $n \in \mathbb{N}$ such that $Q^n(x, B) := (Q^n \mathbb{1}_B)(x) > 0$. (See, e.g., (Meyn and Tweedie, 2012, p.82).) Here, we will say that Q is *weakly irreducible* if there exists a measure π on $(\mathsf{X}, \mathscr{B})$ such that

(ii) π assigns positive measure to all nonempty open sets, and
(i) Q is π-irreducible.

Lemma 4.3. *The following implications hold for any transition kernel Q on X.*

$$\textit{irreducibility} \implies \textit{weak irreducibility} \implies \textit{open set irreducibility.}$$

Proof. Regarding the first implication, let Q be irreducible and let π be any distribution on X such that $\pi(G) > 0$ whenever $G \subset \mathsf{X}$ is open and nonempty.[2] Fix $B \in \mathscr{B}$ with $\pi(B) > 0$ and fix $x \in B$. We recall from Lemma 2.1 that δ_x is a nonzero element of the dual space $b\mathsf{X}'$. Also, B is not the empty set because $\pi(B) > 0$, so $\mathbb{1}_B$ is a nonzero element of $b\mathsf{X}$. Hence, by irreducibility, there exists an $n \in \mathbb{N}$ with $\langle \delta_x, Q^n \mathbb{1}_B \rangle > 0$. We can rewrite this as $Q^n(x, B) = (Q^n \mathbb{1}_B)(x) > 0$. This proves that Q is π-irreducible. We conclude that Q is weakly irreducible.

Regarding the second implication, let Q be weakly irreducible and let π be the measure in (i)–(ii) of the definition of weak irreducibility. Pick any $x, y \in \mathsf{X}$ and let G be any open neighborhood of y. By (i), we have $\pi(G) > 0$. By (ii), we can find an $m \in \mathbb{N}$ with $P^m(x, G) > 0$. Hence y is reachable from x. Since x and y were chosen arbitrarily, we conclude that Q is open set irreducible. □

Now we state a result for the weakly irreducible case. In the statement, σ is any feasible policy.

Theorem 4.4. *If P_σ is weakly irreducible and σ is continuous, then (E1)–(E3) are equivalent.*

Proof. In view of Lemma 3.1, it suffices to show (E1) implies (E3). This is true by Theorem 4.2 and Lemma 4.3. □

5. Beyond Irreducibility

Next we present results for the case where irreducibility fails. In such cases, equality of v_σ and v^* at a point will not imply optimality everywhere. At the same time, we can obtain partial local optimality results by sampling initial conditions across suitable subsets of the state space. The details are given below. Throughout, we continue to work in the setting of Section 2.2.

In the next theorem, ρ is a given element of $\mathcal{D}(\mathsf{X})$ and its support $\operatorname{supp}(\rho)$ is as defined in Section 2.1. Also, σ is a fixed policy in Σ. We set

$$R_{\sigma,\rho} := \{x \in \mathsf{X} : x \text{ is } P_\sigma\text{-reachable from } \operatorname{supp}(\rho)\}.$$

[2]Such a measure exists in many settings, such as when X is a locally compact topological group – in which case we can take π to be the Haar measure. In many applications, X will be a subset of $\mathbb{R}^n$ and π will be Lebesgue measure.

Theorem 5.1. *If σ is continuous and $R_{\sigma,\rho} = \mathsf{X}$, then σ is optimal whenever*

$$\langle v_\sigma, \rho \rangle = \max_{s \in \Sigma} \langle v_s, \rho \rangle \,. \tag{9}$$

In proving Theorem 5.1, we use the fact that (9) is equivalent to $\langle v_\sigma, \rho \rangle = \langle v^*, \rho \rangle$.

Proof of Theorem 5.1. Fix a distribution $\rho \in \mathcal{D}(\mathsf{X})$ and a continuous policy $\sigma \in \Sigma$. We first prove the following claim: If $\langle v_\sigma, \rho \rangle = \langle v^*, \rho \rangle$, then $v_\sigma(x) = v^*(x)$ at all $x \in \operatorname{supp}(\rho)$. To see that this is so, suppose instead that $v_\sigma(x) < v^*(x)$ at some $x \in \operatorname{supp}(\rho)$. Since both of these functions are continuous (see the proof of Lemma 4.1), it follows that there exists an open neighborhood G of x with $v_\sigma(y) < v^*(y)$ for all $y \in G$. Note that $\rho(G) > 0$ (because $\rho(G) = 0$ implies $\rho(G^c) = 1$, and since G^c is closed, the definition of support then implies that $\operatorname{supp}(\rho) \subset G^c$, which contradicts $x \in \operatorname{supp}(\rho) \cap G$). As a result,

$$\langle v_\sigma, \rho \rangle = \int_G v_\sigma \, \mathrm{d}\rho + \int_{G^c} v_\sigma \, \mathrm{d}\rho < \int_G v^* \, \mathrm{d}\rho + \int_{G^c} v_\sigma \, \mathrm{d}\rho \leqslant \langle v^*, \rho \rangle \,.$$

This contradicts the hypothesis $\langle v_\sigma, \rho \rangle = \langle v^*, \rho \rangle$, so the claim above is valid.

Now suppose that $R_{\sigma,\rho} = \mathsf{X}$ and $\langle v_\sigma, \rho \rangle = \langle v^*, \rho \rangle$. Fix any $y \in \mathsf{X}$. Since $R_{\sigma,\rho} = \mathsf{X}$, there exists an x in the support of ρ such that y is P_σ-reachable from x. By the result from the previous paragraph, we also have $v_\sigma(x) = v^*(x)$. Combining these facts with Lemma 4.1 proves the claim in Theorem 5.1. □

6. Applications

In this section, we consider three applications. The first application, found in Section 6.1, demonstrates that the irreducibility assumptions in Theorems 3.2, 4.2, and 4.4 cannot be removed without changing the conclusions. The second involves an optimal savings problem with stochastic returns on wealth. The third application provides an extension showing how results in the paper can be applied outside the MDP framework.

6.1. **An Inventory Problem.**

In this section, we consider optimal order strategies for a simple inventory problem with fixed costs. Our main objective is to provide intuition for the connection between irreducibility and the equivalence of (E1)–(E3). The problem is helpful for intuition because the state space is finite and the implications of irreducibility are easy to understand.

The problem has the following structure. The size of a firm's inventory takes values in the set of integers $\mathsf{X} := \{0, \dots, K\}$ and updates according to

$$X_{t+1} = \max\{X_t - D_{t+1}, 0\} + A_t.$$

The action A_t is the current order quantity and D_{t+1} is a demand shock realized after A_t is chosen. Expected rewards in the current period at state x and action a are

$$r(x, a) = \sum_{d \geqslant 0} \min\{x, d\}\varphi(d) - ca - \kappa \mathbb{1}\{a > 0\},$$

where φ is the density of the demand shock, taking values in the nonnegative integers, c is the unit cost of ordering, κ is a fixed cost, and the unit price is 1.

In this setting we endow X with the discrete topology (and metric), so that all subsets of X are open and $\mathscr{B}$ is the set of all subsets of X. We also impose the discrete metric on A. As a result, every map and hence every policy from X to A is continuous. A transition kernel Q on X can be understood as a mapping $Q \colon \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ satisfying $\sum_{x'} Q(x, x') = 1$ for all x in X. Since individual points are open under the discrete topology, a point y is Q-reachable from another point x if and only if there exists an $m \in \mathbb{N}$ such that $Q^m(x, y) > 0$. The usual definition of irreducibility of a transition kernel Q on finite X is that all states communicate (every $x \in \mathsf{X}$ is Q-reachable from every y in X). In the current setting, this agrees with our previous notion of irreducibility, stated in Section 3.2. (The proof is straightforward and can be obtained from the authors on request.)

We first solve the model with $K = 6$, $\beta = 0.97$, $c = 0.2$, $\kappa = 0.8$, and φ set to the geometric distribution $\varphi(d) = (1 - p)^d p$, where $p = 0.6$. We apply Howard policy iteration (HPI), which produces an exact optimal policy σ^* in finitely many steps. This optimal policy σ^* is shown in Figure 2. It is characterized by large orders in low states and zero orders in high states, which follows from the existence of the fixed cost. In particular, $\sigma^*(0) = 6 = K$, so firms with zero inventory completely restock to the maximum value K.

Next we solve the model using the same parameters by maximizing $v_\sigma(0)$ over the set of all $\sigma \in \Sigma$. The resulting policy $\hat{\sigma}$, which we call the locally maximal policy, is also shown in Figure 2. This locally maximal policy turns out to agree with the optimal policy, as can be seen in the figure. This is as predicted by Theorem 3.2. Indeed, the state process is irreducible under the locally optimal policy $\hat{\sigma}$, so maximality of lifetime value at state 0 implies maximality everywhere.

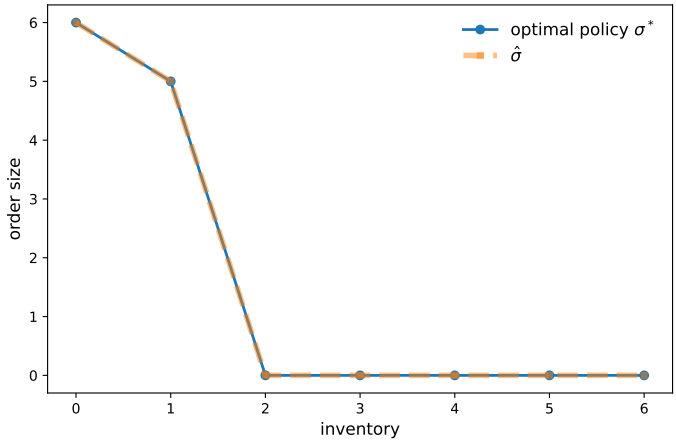

FIGURE 2. Policies under geometric demand

To investigate what happens without irreducibility, we set $\kappa = 1$, hold the other parameters constant, and change the demand shock to $\hat{D}_t = 2D_t$, where, as before, D has the geometric distribution. Thus, $\hat{D}_t$ has a scaled geometric distribution and only takes even values. Now the locally maximal policy we obtain by solving $\max_{\sigma\in\Sigma} v_\sigma(0)$ does not agree with the optimal policy, as shown in Figure 3. This is because, under the resulting locally maximal policy $\hat{\sigma}$, and conditional on starting at state 0, the state only takes even values (because $\hat{\sigma}(x)$ is even at every x and the demand shock takes even values). As a consequence, the state dynamics are not irreducible, and maximizing $v_\sigma(0)$ over σ does not generate an optimal policy. Figure 4 confirms that the locally maximal policy $\hat{\sigma}$ is not optimal.

6.2. **Optimal Savings.** In this section, we use a canonical optimal savings problem with stochastic returns and labor income as a one–dimensional laboratory for our results in continuous state–action MDPs. With full support shocks, the transition kernel is open set irreducible and the conditions (E1)–(E3) are equivalent. With bounded support shocks, the kernel is reducible and the equivalence can fail.

Consider an agent who seeks to maximize lifetime utility by choosing optimal savings and consumption. The evolution of wealth is governed by the equation

$$w_{t+1} = \eta_{t+1}(w_t - c_t) + y_{t+1}, \qquad t = 0, 1, \ldots, \tag{10}$$

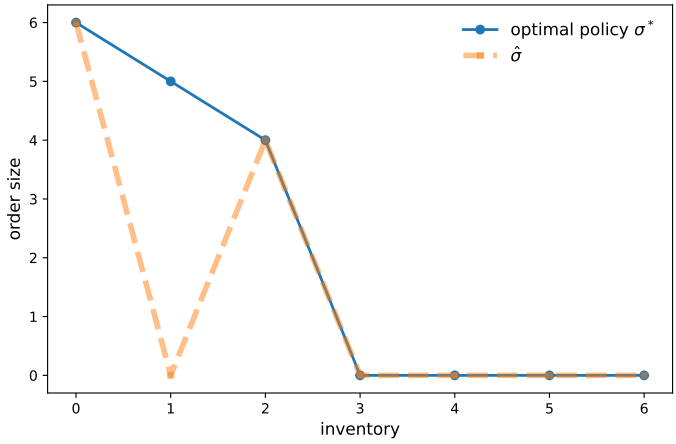

FIGURE 3. Policies under scaled geometric demand

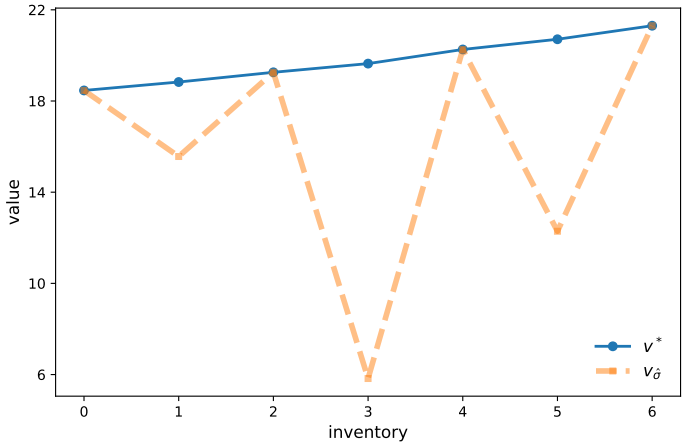

FIGURE 4. Lifetime values under scaled geometric demand

where $w_t \in \mathbb{R}_+$ is time-t wealth, c_t is the current-period consumption, y_{t+1} is the next-period labor income, and η_{t+1} represents the stochastic return on savings. The sequences (y_t) and (η_t) are drawn from φ and ψ respectively.

We formulate this problem as an MDP. The state space is $\mathbb{R}_+$ and the set of feasible actions at wealth level w is $\Gamma(w) = \{c \in \mathbb{R}_+ : c \leqslant w\}$. A feasible policy in this setup is a Borel measurable function σ from $\mathbb{R}_+$ to itself satisfying $\sigma(w) \leqslant w$ for all $w \in \mathbb{R}_+$. The reward function is $r(w, c) \coloneqq u(c)$, where $u(c)$ is the utility derived from consumption and u is continuous, differentiable, and strictly concave on $\mathbb{R}_+$. The

transition kernel P is given by

$$P(w, c, B) = \int \mathbb{1}_B(\eta'(w - c) + y')\psi(d\eta')\varphi(dy'), \tag{11}$$

where $0 \leqslant c \leqslant w$ and B is a Borel set in $\mathbb{R}_+$. Given $\sigma \in \Sigma$, the corresponding policy operator T_σ is given by

$$(T_\sigma v)(w) = u(\sigma(w)) + \beta(P_\sigma v)(w) := u(\sigma(w)) + \beta \int v(w')P(w, \sigma(w), \mathrm{d}w'),$$

where $\beta \in (0, 1)$ is the discount factor. Let

$$B(w, c, v) = u(c) + \beta \int \int v(\eta'(w - c) + y')\psi(d\eta')\varphi(dy').$$

Since u is strictly concave, the map $c \mapsto B(w, c, v)$ is strictly concave whenever v is concave on $\mathbb{R}_+$. One can also show that v^* is concave on $\mathbb{R}_+$. Combining these facts with the Bellman equation, it is straightforward to show that the optimal policy is both unique and continuous. We record this in the proposition below. More details on the arguments can be found in Chapter 12 of Stachurski (2022).

Proposition 6.1. *Under the assumptions stated above, the optimal policy of the optimal savings model is unique and continuous on* $\mathbb{R}_+$.

To cross-check our theory, we conduct two computational experiments: one with an irreducible transition kernel (Section 6.2.1) and another with a reducible transition kernel (Section 6.2.2). In both cases, we use a neural network to approximate the optimal policy by maximizing $v_\sigma(\bar{w})$ at a fixed point $\bar{w} \in \mathbb{R}_+$. Our approach employs a simple gradient ascent algorithm to maximize a real-valued criterion in the form of (1) by adjusting the parameters of the policy network. We denote the resulting policy as $\hat{\sigma}$ and the corresponding $\hat{\sigma}$–value function as $v_{\hat{\sigma}}$. The algorithmic details are provided in the online supplement.

We benchmark $v_{\hat{\sigma}}$ and $\hat{\sigma}$ against optimistic policy iteration (OPI), a variant of value function iteration (VFI) known to converge globally to the optimal policy in this model-based setting (Sargent and Stachurski, 2025a, §5.1.4.4). The OPI algorithm operates on a discretized state space and serves as our ground truth comparison. We compute the value function v^* by applying OPI over a fine grid of wealth levels, yielding a global approximation of both the optimal value function and the corresponding optimal policy σ^*. Throughout our experiments, we employ the CRRA

utility function $u(c) = c^{1-\gamma}/(1-\gamma)$ with risk aversion parameter $\gamma = 2$ and discount factor $\beta = 0.98$.

6.2.1. *An Irreducible Case.* For now, we assume both distributions have full support on $\mathbb{R}_+$. Specifically, we let (η_t) and (y_t) be independent IID sequences

$$\eta_t \overset{\text{IID}}{\sim} \psi \coloneqq \text{LogNormal}(0, \sigma_\eta^2), \qquad y_t \overset{\text{IID}}{\sim} \varphi \coloneqq \text{LogNormal}(\mu_y, \sigma_y^2),$$

where the numerical values of σ_η^2 and σ_y^2 are chosen to match the stationary variances implied by AR(1) processes. In particular, we set $\sigma_\eta^2 = \frac{\nu_\eta^2}{1-\rho_\eta^2}$ and $\sigma_y^2 = \frac{\nu_y^2}{1-\rho_y^2}$ with parameters $\rho_\eta = 0.9$, $\nu_\eta = 0.02$, $\rho_y = 0.9$, $\nu_y = 0.3$, $\mu_y = 1.5$. Obviously, $\text{supp}(\psi) = \text{supp}(\varphi) = \mathbb{R}_+$.

Using this setup, we have the following result:

Lemma 6.2. *Under the stated assumptions, the optimal savings transition kernel P_σ is open set irreducible for every $\sigma \in \Sigma$.*

Proof. Let Σ be the set of feasible policies. Fix $w \in \mathbb{R}_+$ and an open set $B \subseteq \mathbb{R}_+$ with positive Lebesgue measure. Let $\alpha = w - \sigma(w) > 0$. By a change of variable, we obtain

$$\begin{aligned} P_\sigma(w, B) &= \int \mathbb{1}_B(w')\psi(\eta')\varphi(w' - \alpha\eta')\, d\eta'\, dw' \\ &= \int_B \left(\int_0^\infty \psi(\eta')\varphi(w' - \alpha\eta')\, d\eta' \right) dw'. \end{aligned}$$

We treat the inner integral first. Since η' and y' are independent random variables on $\mathbb{R}_+$ with strictly positive densities ψ and φ respectively and $w' = \alpha\eta' + y'$, we have the probability density function f of w' at any point $w' > 0$ is given by the convolution

$$f(w') = \int_0^\infty \psi(\eta')\varphi(w' - \alpha\eta')d\eta'$$

as in the inner integral. Since $w' = \eta'(w - c) + y'$ and $y' > 0$, $w' - \alpha\eta' > 0$, which implies that $0 < \eta' < \frac{w'}{\alpha}$. Therefore, we have

$$f(w') = \int_0^{\frac{w'}{\alpha}} \psi(\eta')\varphi(w' - \alpha\eta')d\eta'.$$

Note that $\psi(\eta') > 0$ for $\eta' > 0$ and $\varphi(w' - \alpha\eta') > 0$ with $\eta' \in (0, \frac{w'}{\alpha})$. Moreover, $|\frac{w'}{\alpha}| > 0$ for every $w' > 0$. Hence $f(w') > 0$ for all $w' \in \mathbb{R}_+$. For $\alpha = 0$, $w' = y'$ implies that $f(w') = \varphi(w') > 0$ for all $w' \in \mathbb{R}_+$. Since B is open, there is a nonempty open

interval $(l, m) \subset B$. Thus, $P_\sigma(w, B) = \int_B f(w')\, dw' \geqslant \int_l^m f(w')\, dw' > 0$. That is, P_σ is open set irreducible. □

Given the open set irreducibility of the transition kernel at any feasible policy stated in Lemma 6.2, Theorem 4.2 implies that we can compute a (globally) optimal policy by maximizing $v_\sigma(\bar{w})$ at any fixed $\bar{w} \in \mathbb{R}_+$.

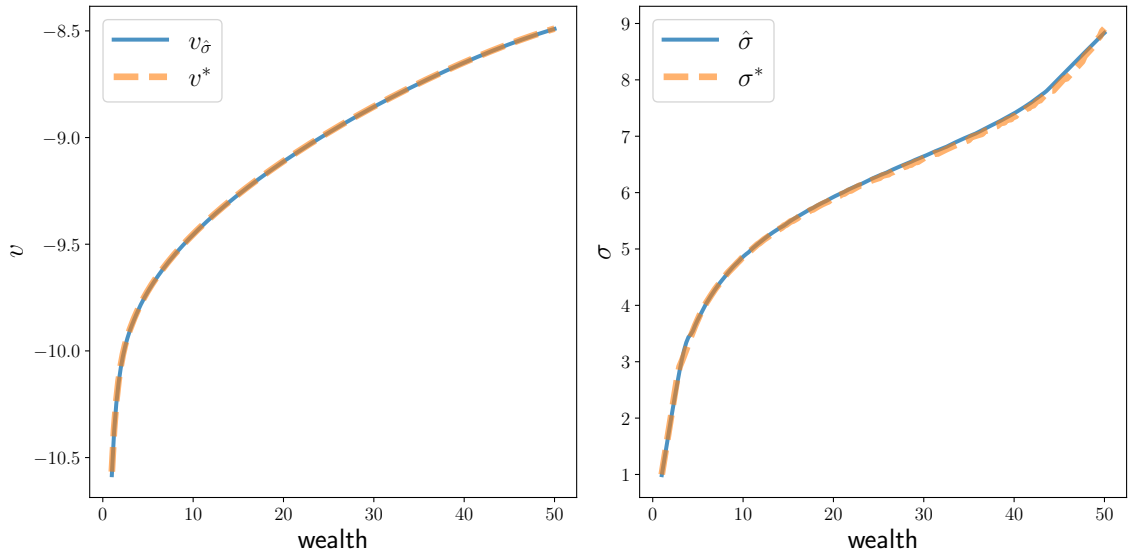

FIGURE 5. $v_{\hat{\sigma}}$ and $\hat{\sigma}$ with $\bar{w} = 1$ vs OPI solutions

Figure 5 shows the result of these computations when $\bar{w} = 1$. The function $v_{\hat{\sigma}}$, shown in blue, closely matches the globally optimal value function v^* computed via OPI (red dotted line). Similarly, in the second panel, the approximated policy $\hat{\sigma}$ (blue line) closely matches the optimal policy σ^* (red dotted line). This convergence demonstrates that both the $\hat{\sigma}$-value function $v_{\hat{\sigma}}$ and the policy $\hat{\sigma}$ successfully recover their globally optimal counterparts v^* and σ^*, respectively. This is consistent with the result in Theorem 4.2, which states that we can compute a globally optimal policy by solving $\max_{\sigma \in \Sigma} v_\sigma(\bar{w})$ at any fixed $\bar{w} \in \mathbb{R}_+$. As predicted by Theorem 4.2, optimizing at any single point $\bar{w}$ yields a globally optimal policy. In the online supplement, we present the result of optimizing at $\bar{w} = 50$ which also yields a globally optimal policy.

6.2.2. *A Reducible Case.* Now consider a modified version of the optimal savings MDP where both returns and labor income are bounded:

$$w' = \eta'(w - c) + y', \quad \eta' \in [\underline{\eta}, \overline{\eta}], \quad y' \in [\underline{y}, \overline{y}] \tag{12}$$

with $0 < \underline{\eta} < \overline{\eta} < 1$ and $0 \leqslant \underline{y} < \overline{y} < \infty$. For $w \in \mathbb{R}_+$ and a Borel set $B \subseteq \mathbb{R}_+$, the transition kernel P_σ is

$$P_\sigma(w, B) = \int \mathbb{1}_B(\eta'(w - \sigma(w)) + y')\psi(d\eta')\varphi(dy') \tag{13}$$

where ψ and φ have support on $[\underline{\eta}, \overline{\eta}]$ and $[\underline{y}, \overline{y}]$ respectively. In this case, we let ψ and φ be uniform distributions. For the rest of this section, we always set $[\underline{\eta}, \overline{\eta}] = [0.5, 0.8]$ and $[\underline{y}, \overline{y}] = [1, 8]$. At the same time, we deliberately continue to take $\mathbb{R}_+$ as the state space. This leads to a failure of irreducibility, as clarified in the next proposition.

Proposition 6.3. *For any $\sigma \in \Sigma$, the transition kernel P_σ is neither weakly irreducible nor open set irreducible.*

Proof. Fix $\sigma \in \Sigma$. In view of Lemma 4.3, we only need to show that P_σ is not weakly irreducible. To this end, let ν be a distribution on $\mathbb{R}_+$ that assigns positive probability to open sets. Fix initial wealth $w_0 \in \mathbb{R}_+$. For any t-step transition, let $\alpha_t := w_t - \sigma(w_t)$ be savings at step t. Then $w_{t+1} \leqslant \overline{\eta}\alpha_t + \overline{y} \leqslant \overline{\eta}w_t + \overline{y}$. Iterating this inequality n times from w_0

$$w_n \leqslant \overline{\eta}^n w_0 + \overline{y}(1 + \overline{\eta} + \cdots + \overline{\eta}^{n-1}) = \overline{\eta}^n w_0 + \overline{y}\frac{1 - \overline{\eta}^n}{1 - \overline{\eta}}.$$

Hence, there exists an $M \in \mathbb{R}_+$ such that

$$w_n < \overline{\eta}^n w_0 + \overline{y}\frac{1}{1 - \overline{\eta}} < M \quad \forall n \in \mathbb{N}.$$

Let $B = (M, \infty)$. Then $\nu(B) > 0$, and $P_\sigma^n(w_0, B) = 0$ for all $n \in \mathbb{N}$. This shows that P_σ is not weakly irreducible. □

The preceding proposition is partly illustrated in Figure 6. The figure shows the 45 degree line and an upper bound law of motion for wealth, obtained by setting consumption to zero and both shocks to their upper bound. States above the steady state are not reachable from states below the steady state.

For reducible MDPs, maximizing the objective at a single initial state $\bar{w} \in \mathbb{R}_+$ no longer guarantees convergence to the optimal policy. This limitation is clearly demonstrated in Figure 7, which shows significant discrepancies between the algorithm's output and the OPI solution.

Figure 7 is consistent with Lemma 4.1. The policy network achieves near-optimal performance at lower wealth levels because these levels are reachable from the initial

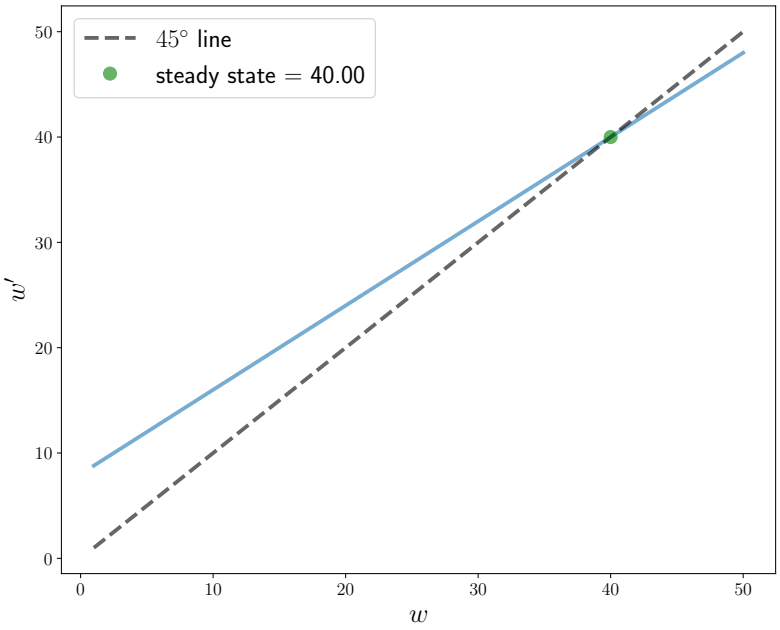

FIGURE 6. The upper bound law of motion for wealth

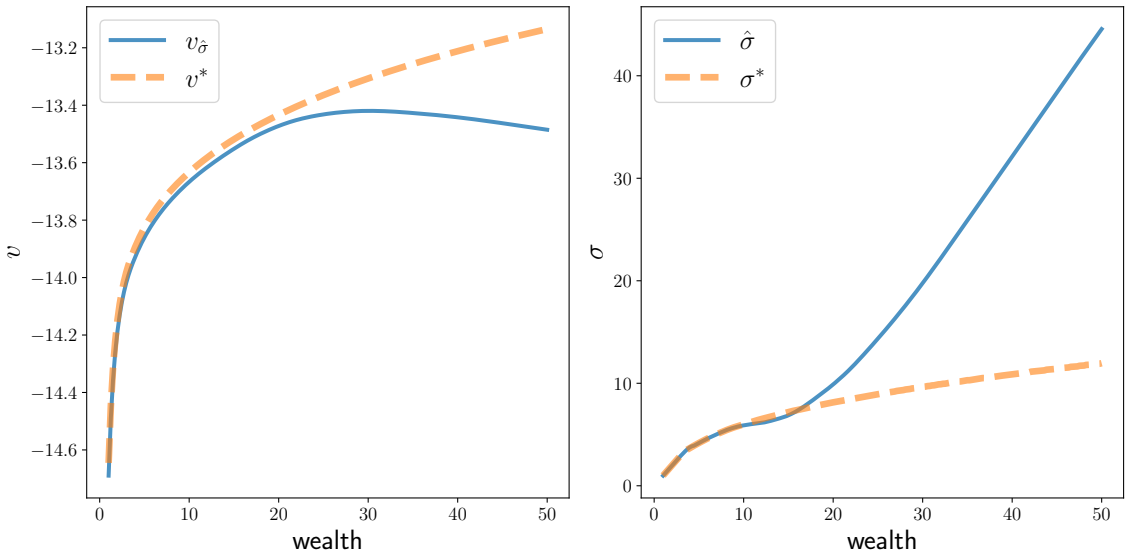

FIGURE 7. $v_{\hat{\sigma}}$ and $\hat{\sigma}$ with $\bar{w} = 1$ vs OPI solutions

state under the learned policy $\hat{\sigma}$. However, the performance deteriorates for wealth levels that are not reachable from the initial state given the policy $\hat{\sigma}$.[3]

These results confirm that, when irreducibility fails, the policy $\hat{\sigma}$ computed to maximize local optimality is not guaranteed to be optimal, and the performance of the policy is sensitive to the initial state and the reachable subset of the state space starting from the initial state.

The ideas in this section can be extended to more sophisticated methods, where the policy network is trained to maximize other variants of the real-valued objective in (1). One example is actor-critic methods (Silver et al., 2014; Lillicrap et al., 2019), which are used for solving challenging heterogeneous agent models in Han and Yang (2021). Further discussion is provided in the online supplement.

6.3. **Firm Entry.** The next example provides an extension showing how the results in the paper can be used outside the MDP setting. This is significant because many recent examples of dynamic programs are not MDPs (see, e.g., Sargent and Stachurski (2025a), Chapter 8). The example concerns an optimal stopping problem with state-dependent discounting. We prove that, under the assumptions stated below, local optimality implies global optimality.

Specifically, we consider a stopping problem for a firm choosing if and when to enter a market. Stopping corresponds to entering the market and continuing corresponds to waiting until the next period and then deciding again. If the firm enters at time t it receives one-off profit $\pi(X_t)$, where $(X_t)_{t\geqslant 0}$ is a Markov sequence on possibly unbounded interval $\mathsf{X} \subset \mathbb{R}$ with transition kernel Q. If the firm continues then it pays a fixed cost c. We admit the possibility that discounting is not constant. (This allows for the fact that firms evaluate future profit opportunities based on cost-of-capital, which fluctuates over time.) Thus, the Bellman equation for the firm is

$$v(x) = \max\left\{\pi(x),\ -c + \beta(x)\int v(x')Q(x,\mathrm{d}x')\right\}, \tag{14}$$

where $\beta(x) \in (0,\infty)$ is the discount factor associated with cost-of-capital in state x. (Discount factor realizations above one are admitted in order to accommodate occasionally negative interest rates.)

[3]The reachable range of wealth levels in the graph is smaller than in Figure 6 because $\hat{\sigma}(w) > 0$ for all $w \in \mathbb{R}_+$.

Let $ib\mathsf{X}$ be the increasing functions in $b\mathsf{X}$ and let $icb\mathsf{X}$ be the continuous functions in $ib\mathsf{X}$. We assume that the discount operator K defined by

$$(Kf)(x) = \beta(x) \int f(x')Q(x, \mathrm{d}x') \qquad (x \in \mathsf{X},\ f \in b\mathsf{X})$$

preserves monotonicity and continuity, in the sense that K is invariant on both $ib\mathsf{X}$ and $icb\mathsf{X}$, and that $r(K) < 1$, where $r(K)$ is the spectral radius of K. For simplicity, we also assume that Q always has full support on $\mathbb{R}$, so that $Q(x, B) > 0$ for any Borel set $B \subset \mathbb{R}$ with positive Lebesgue measure. (For example, (X_t) might be driven by an AR(1) sequence with normal shocks.) Finally, we assume that the profit function π is bounded, increasing, and continuous.

A policy is a map σ from X to $\{0, 1\}$, with $\sigma(x) = 1$ indicating the decision to stop. The lifetime value v_σ of policy σ is the unique fixed point of the operator T_σ on $b\mathsf{X}$ given by

$$(T_\sigma v)(x) = \sigma(x)\pi(x) + (1 - \sigma(x)) \left[-c + \beta(x) \int v(x')Q(x, \mathrm{d}x') \right]. \tag{15}$$

We let Σ be the set of all policies σ such that T_σ is invariant on $ib\mathsf{X}$.

As for the MDP case, the *value function* is denoted v^* and defined at each $x \in \mathsf{X}$ by $v^*(x) := \sup_{\sigma \in \Sigma} v_\sigma(x)$. A policy σ is called *optimal* if $v_\sigma(x) = v^*(x)$ for all $x \in \mathsf{X}$. In the proofs below we will make use of the Bellman operator for this problem, which takes the form

$$(Tv)(x) = \max \left\{ \pi(x),\ -c + \beta(x) \int v(x')Q(x, \mathrm{d}x') \right\} \qquad (x \in \mathsf{X},\ v \in b\mathsf{X}).$$

The model has standard optimality properties, which we detail in the theorem below.

Theorem 6.4. *Under the stated assumptions, the following results hold:*

(a) *The value function is the unique solution to the Bellman equation in* $b\mathsf{X}$,
(b) *the value function is increasing and continuous, and*
(c) *at least one optimal policy exists.*

We will prove in addition the following result, which holds for any $\sigma \in \Sigma$:

Theorem 6.5. *If there exists an* $x \in \mathsf{X}$ *such that* $v_\sigma(x) = v^*(x)$*, then* σ *is optimal.*

We begin the proof of Theorem 6.4 and Theorem 6.5 with a series of lemmas. In what follows, an operator S on bX is called *globally stable* when S has a unique fixed point $\bar{v}$ in bX and $S^n v \to \bar{v}$ as $n \to \infty$ for all $v \in bX$.

Lemma 6.6. *For each $\sigma \in \Sigma$, the operator T_σ is globally stable on bX.*

Proof. Fix $\sigma \in \Sigma$ and $v, w \in bX$. Applying the triangle inequality, combined with the fact that K is a positive linear operator, we have $|T_\sigma v - T_\sigma w| \leqslant |K(v-w)| \leqslant K|v-w|$. Since $\rho(K) < 1$ and bX is a Banach lattice, by Proposition 4.1 and Theorem 2.1 in Stachurski and Zhang (2021), there exists some $N \in \mathbb{N}$ such that T_σ^N is a contraction. By Banach fixed point theorem, T_σ^N is globally stable and it follows that T_σ is globally stable on bX. □

The next proof uses several concepts from Sargent and Stachurski (2025b). We refer to that paper for the definitions, which are omitted here for brevity.

Proof of Theorem 6.4. To prove Theorem 6.4 we apply Theorem 5.5 of Sargent and Stachurski (2025b), which applies to the abstract dynamic program $(bX, \mathbb{T}) := (bX, \{T_\sigma : \sigma \in \Sigma\})$. First, in bX with its usual partial order, the statement $v_n \uparrow v$ is equivalent to $v_n(x) \uparrow v(x)$ in $\mathbb{R}$ for each $x \in X$. From this fact and the monotone convergence theorem, we obtain $Tv_n \uparrow Tv$ whenever $(v_n) \subset bX$ and $v_n \uparrow v \in bX$. Thus, $(bX, \mathbb{T})$ is order continuous. Also, bX is countably-Dedekind complete because the set of bounded Borel measurable functions is closed under pointwise convergence. In addition, each T_σ is globally stable (Lemma 6.6) and hence order stable (see Sargent and Stachurski (2025b), Example 2.1). Finally, let $\hat{T}$ be the operator given by $\hat{T}v = |\pi| + Kv$. Since $\rho(K) < 1$, there is a $u \in bX_+$ with $\hat{T}u = u$. For this u we have $T_\sigma u \leqslant \hat{T}u \leqslant u$, so $(bX, \mathbb{T})$ is bounded above. It now follows from Theorem 5.5 of Sargent and Stachurski (2025b) that the value function is the unique solution to the Bellman equation and at least one optimal policy exists.

To prove the final claim we first note that T is globally stable on bX, since, given $v, w \in bX$, we have $|Tv - Tw| \leqslant |K(v-w)| \leqslant K|v-w|$, so the argument in the proof of Lemma 6.6 applies. Now fix $v \in icbX$. Since K is invariant on $icbX$, the element $-c + Kv$ is in $icbX$. Moreover, $\pi \in icbX$ and so $\max\{\pi, -c + Kv\} \in icbX$. This shows that T is invariant on $icbX$. Since $icbX$ is nonempty and closed under uniform limits, and T is globally stable on bX, we conclude that $v^* \in icbX$. The proof of Theorem 6.4 is now done. □

Lemma 6.7. *If $v_\sigma(x) = v^*(x)$ for some $x \in \mathsf{X}$, then $v_\sigma = v^*$ almost everywhere.*

Proof. Since $v_\sigma \leqslant v^*$ and $T_\sigma v \leqslant Tv$ for all $v \in b\mathsf{X}$, we have $v_\sigma = T_\sigma v_\sigma \leqslant T_\sigma v^* \leqslant Tv^* = v^*$. As a result, $v_\sigma(x) = v^*(x)$ implies $(T_\sigma v_\sigma)(x) = (T_\sigma v^*)(x)$, which, using the definition of T_σ in (15), yields

$$\beta(x) \int v^*(x')Q(x, \mathrm{d}x') = \beta(x) \int v_\sigma(x')Q(x, \mathrm{d}x').$$

As $\beta(x) > 0$, we obtain $\int (v^*(x') - v_\sigma(x'))Q(x, \mathrm{d}x') = 0$. If $v^* > v_\sigma$ on a set E of positive Lebesgue measure, then, at the same time, we have

$$\int (v^*(x') - v_\sigma(x'))Q(x, \mathrm{d}x') \geqslant \int_E (v^*(x') - v_\sigma(x'))Q(x, \mathrm{d}x') > 0,$$

where the last inequality is due to the assumption that $Q(x, \mathrm{d}x')$ assigns positive measure to any such E. From this contradiction, we conclude that $v^* = v_\sigma$ almost everywhere. □

Lemma 6.8. *Let f, g be two increasing functions in $b\mathsf{X}$ with $f \leqslant g$. If g is continuous and $f = g$ almost everywhere, then $f = g$.*

Proof. If f is continuous, the result follows from the fact that any two continuous measurable functions equal almost everywhere on an interval of $\mathbb{R}$ are equal. Now suppose instead that f is not continuous at $x_0 \in \mathsf{X}$. Let $a = \lim_{x \uparrow x_0} f(x)$ and $b = \lim_{x \downarrow x_0} f(x)$. Monotonicity and discontinuity of f implies $a < b$. Since g is continuous, $g((a, b))^{-1}$ is open. We claim that $g((a, b))^{-1}$ is nonempty. Indeed, since $f = g$ almost everywhere, there are $y, y' \in \mathsf{X}$ such that $g(y) \leqslant a$ and $g(y') \geqslant b$. Since g is continuous and increasing, there exist $y'' \in \mathsf{X}$ such that $g(y'') \in (a, b)$. Thus $g((a, b))^{-1}$ is a nonempty open set. This proves the existence of a set of nonzero Lebesgue measure on which $f < g$. Contradiction. □

Proof of Theorem 6.5. Suppose there exists an $x \in \mathsf{X}$ and a policy σ such that $v_\sigma(x) = v^*(x)$. Then, by Lemma 6.7, we have $v_\sigma = v^*$ almost everywhere. By Theorem 6.4, the function v^* is increasing and continuous. By Lemma 6.6, the operator T_σ is invariant on the closed set $ib\mathsf{X}$, so the fixed point is in $ib\mathsf{X}$; in particular, v_σ is also increasing. By definition, $v_\sigma \leqslant v^*$. Hence, by Lemma 6.8, we have $v_\sigma = v^*$. □

7. Extensions and Future Work

Using MDP optimality results from Bäuerle and Rieder (2011) or Bertsekas (2022), it should be possible to extend our results to the case of unbounded rewards by replacing the ordinary supremum norm on $b\mathsf{X}$ with a weighted supremum norm. Also, while our results have focused on standard MDPs with constant discount factors, one useful variation of this model is MDPs with state-dependent discount factors, so that β becomes a map from X to $\mathbb{R}_+$ (see, e.g., Stachurski and Zhang (2021)). We conjecture that the main results we obtained for MDPs will extend to generalized ADPs with state-dependent discounting under suitable stability and irreducibility assumptions. Finally, it seems likely that results similar to Theorem 3.2 will be valid for some continuous time MDPs, as well as at least some of the nonstandard discrete time dynamic programs discussed in Bertsekas (2022) and Sargent and Stachurski (2025a). These topics are also left for future work.

References

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

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

——— (2021): *Rollout, policy iteration, and distributed reinforcement learning*, Athena Scientific.

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

Bhandari, J. and D. Russo (2024): "Global optimality guarantees for policy gradient methods," *Operations Research*.

Bloise, G., C. Le Van, and Y. Vailakis (2023): "Do not blame Bellman: It is Koopmans' fault," *Econometrica*, in press.

Friedl, A., F. Kübler, S. Scheidegger, and T. Usui (2023): "Deep uncertainty quantification: with an application to integrated assessment models," Tech. rep., Working Paper University of Lausanne.

Han, J. and Y. Yang (2021): "Deepham: A global solution method for heterogeneous agent models with aggregate shocks," *arXiv preprint arXiv:2112.14377*.

Hernández-Lerma, O. and J. B. Lasserre (1996): *Discrete-time Markov control processes: basic optimality criteria*, vol. 30, Springer Science & Business Media.

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

Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra (2019): "Continuous control with deep reinforcement learning," .

Ljungqvist, L. and T. J. Sargent (2018): *Recursive macroeconomic theory*, MIT press, 4 ed.

Meyn, S. P. and R. L. Tweedie (2012): *Markov chains and stochastic stability*, Springer Science & Business Media.

Murphy, K. (2024): "Reinforcement Learning: An Overview," .

Payne, J., A. Rebei, and Y. Yang (2025): "Deep learning for search and matching models," *Available at SSRN 5123878*.

Puterman, M. L. (2014): *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons.

Rust, J. (1996): "Numerical Dynamic Programming in Economics," in *Handbook of Computational Economics*, ed. by H. Amman, D. Kendrick, and J. Rust, Elsevier, vol. 1, 619–729.

Sargent, T. J. and J. Stachurski (2025a): *Dynamic Programming: Finite States*, Cambridge University Press.

——— (2025b): "Dynamic Programs on Partially Ordered Sets," *SIAM Journal on Control and Optimization*, 63, 778–795.

Schaefer, H. H. (1974): *Banach Lattices and Positive Operators*, Springer.

Silver, D., G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller (2014): "Deterministic policy gradient algorithms," in *International conference on machine learning*, Pmlr, 387–395.

Stachurski, J. (2022): *Economic dynamics: theory and computation*, MIT Press, 2 ed.

Stachurski, J. and J. Zhang (2021): "Dynamic programming with state-dependent discounting," *Journal of Economic Theory*, 192, 105190.

Stokey, N. L. and R. E. Lucas (1989): *Recursive methods in dynamic economics*, Harvard University Press.