DYNAMIC PROGRAMMING: OPTIMALITY AT A POINT IMPLIES OPTIMALITY EVERYWHERE

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 at every point in the state space. This raises a natural question: under what conditions does optimality at a single state imply optimality at every state? We show that, in a general setting, the irreducibility of the transition kernel under a feasible policy is a sufficient condition for extending optimality from one state to all states. These results have important implications for dynamic optimization algorithms based on gradient methods, which are routinely applied in reinforcement learning and other large scale applications.

1. Introduction

Dynamic programming is a major branch of optimization theory, with applications ranging from supply chain management and fleet maintenance to option pricing, DNA sequencing, and air traffic management. Dynamic programs that include uncertainty are often called Markov decision processes (MDPs) and the theory of such processes has been extensively developed (see, e.g., Bäuerle and Rieder (2011), Hernández-Lerma and Lasserre (2012), Bertsekas (2012), or Bertsekas (2022)). Much of the recent surge in interest in MDPs has been fueled by artificial intelligence and reinforcement learning (see, e.g., Bertsekas (2021) or Kochenderfer et al. (2022)).

Let Σ be the set of all policies for a given MDP, each of which is a map σ from a state space X into an action space A. Let $v_\sigma(x)$ represent the lifetime value of policy σ given initial state x. A policy σ is called optimal when $v_\sigma \geqslant v_s$ for all $s \in \Sigma$. Here functions are ordered pointwise, so the statement $v_\sigma \geqslant v_s$ means that $v_\sigma(x) \geqslant v_s(x)$ for all $x \in \mathsf{X}$. One fundamental result of dynamic programming theory is that, for standard MDPs, optimal policies always exist. Further theory provides characterizations of

*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.

optimal policies—typically via the Bellman equation—and algorithms for computing either exact or approximately optimal policies.

We consider the following question: under what conditions does optimality at a single state imply optimality at all states? In other words, when does $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ at some x imply $v_\sigma(x) = \max_{s\in\Sigma} v_s(x)$ for all x? In this paper, we show that, for standard MDPs on general state spaces, irreducibility of the Markov dynamics generated by σ is sufficient for this property. Specifically, if a policy is optimal at a single state and has an irreducible transition kernel, then this optimality propagates throughout the entire state space, making the policy globally optimal. Similarly, if irreducibility holds and there exists a distribution ρ such that $\int v_\sigma \,\mathrm{d}\rho \geqslant \int v_s \,\mathrm{d}\rho$ for all $s \in \Sigma$, then σ is an optimal policy. Some extensions are provided, as well as a sharper result for finite state MDPs.

Our results have particular significance for policy gradient methods, which have become increasingly popular for solving large-scale MDPs (Sutton et al., 1999; Lan et al., 2023; Kumar et al., 2023). Because this technique uses gradient ascent rather than more standard dynamic programming algorithms, it can only maximize a real-valued criterion such as $v_\sigma(x)$ for some fixed x or $\int v_\sigma \,\mathrm{d}\rho$ for some specified distribution ρ, rather than maximizing v_σ at all x simultaneously. Our results show that, under irreducibility, maximizing one of these real-valued criteria is sufficient for global optimality. In addition, our result in finite state setting show that, even when irreducibility does not hold, optimality still holds for an accessible subset of the state space.

Other papers have looked at theoretical properties of gradient policy methods, where an expression such as $\int v_\sigma \,\mathrm{d}\rho$ is maximized over all $\sigma \in \Sigma$ for some specified distribution ρ. Examples include Khodadadian et al. (2021), Agarwal et al. (2021), and Xiao (2022). However, in these papers, focus is on proving the convergence of $\int v_\sigma \,\mathrm{d}\rho$ to the maximal value $\int v^* \,\mathrm{d}\rho$, rather than proving global convergence from local convergence. At the same time, these papers provide rates of convergence for specific algorithms, which we do not discuss.

A related line of research focuses on average-optimal policies in finite-state MDPs by leveraging specific state space structures under some policies. This includes exploring unichain, multichain, communicating, and weakly communicating MDPs to study algorithmic convergence (Bartlett and Tewari, 2009; Puterman, 2014). Our result demonstrates that, in finite-state MDPs, optimality can extend from a single state to

all accessible states. Our result is applicable to various classes of MDPs in this line of research and supports the development of more efficient algorithms.

2. Main Result

In this section, we present our main result, characterizing the conditions under which optimality at a single state implies optimality at all states.

2.1. **Markov Decision Process.** Let X and A be metric spaces, let $b\mathsf{X}$ be the set of bounded Borel measurable functions from X to $\mathbb{R}$, and let $bc\mathsf{X}$ be the continuous functions in $b\mathsf{X}$. Both $b\mathsf{X}$ and $bc\mathsf{X}$ 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 $x \in \mathsf{X}$. Absolute values are applied pointwise, so that $|f| \in b\mathsf{X}$ is the function $x \mapsto |f(x)|$. In all of what follows, $\mathcal{D}(\mathsf{X})$ is the set 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

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

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 M if $MI \subset I$. A linear operator M from $b\mathsf{X}$ to itself is called *positive* when $Mf \geqslant 0$ for all $f \geqslant 0$. A positive linear operator M is called *irreducible* if the only invariant ideals under M are the trivial subspace $\{0\}$ and the whole space $b\mathsf{X}$ (see, e.g., Zaanen (2012)).

We consider an MDP (r, Γ, β, P) with state space X and action space A. Here r is the reward function, Γ is a feasible correspondence, β 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 G be the graph of Γ; that is, $\mathsf{G} \coloneqq \{(x, a) \in \mathsf{X} \times \mathsf{A} : a \in \Gamma(x)\}$. We assume that

(a) $\beta \in (0, 1)$,
(b) Γ is a nonempty, continuous and compact-valued,
(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 $v \in bc\mathsf{X}$.

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

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

Thus, $r_\sigma(x)$ is rewards at x under policy σ and P_σ is the Markov dynamics associated with σ. We can view P_σ as a linear operator $f \mapsto P_\sigma f$ on $b\mathsf{X}$, where $(P_\sigma f)(x) := \int f(x')P_\sigma(x, \mathrm{d}x')$ is the expectation of $f(X_{t+1})$ given policy σ and current state $X_t = x$. Using this operator, the *lifetime value* of a policy σ, denoted by v_σ, can be expressed as

$$v_\sigma := \sum_{t=0}^{\infty} (\beta P_\sigma)^t r_\sigma = (I - \beta P_\sigma)^{-1} r_\sigma. \tag{1}$$

(See, e.g., Puterman (2014), Theorem 6.1.1.) 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 $x \in \mathsf{X}$.

Theorem 2.1. *Let σ be a feasible policy and suppose that P_σ is irreducible. In this setting, the following statements are equivalent.*

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

For example, Theorem 2.1 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 $s \mapsto v_s(x)$ over Σ. Alternatively, we can fix any distribution ρ and maximize $s \mapsto \langle v_s, \rho \rangle$. The proof of Theorem 2.1 is given in Section 2.2.

2.2. **Proof of Theorem 2.1.** We denote the positive cone of $b\mathsf{X}$ by $b\mathsf{X}_+$, the set of all $v \in b\mathsf{X}$ with $v \geqslant 0$. We take $b\mathsf{X}'_+$ to be the set of all positive linear functionals on $b\mathsf{X}$, the set of all $\mu \in b\mathsf{X}'_+$ with $\langle \mu, f \rangle \geqslant 0$ for all $f \in b\mathsf{X}_+$. Let $b\mathsf{X}' = b\mathsf{X}'_+ - b\mathsf{X}'_+$ and $b\mathsf{X}'$ is the order dual of $b\mathsf{X}$. Proposition 5.5 of Schaefer (1974) implies $b\mathsf{X}'$ is the topological dual of $b\mathsf{X}$, the set of all bounded linear functionals on $b\mathsf{X}$. Also, Proposition 8.3 (c) of Schaefer (1974) states the following for a positive linear operator K mapping $b\mathsf{X}$ into itself:

Proposition 2.2. *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 $\langle \mu, K^m f \rangle > 0$.*

For each $x \in \mathsf{X}$, the *point evaluation functional* on $b\mathsf{X}$ is the map δ_x that sends each $w \in b\mathsf{X}$ into $w(x)$ (i.e., $\langle w, \delta_x \rangle = w(x)$ for every $w \in b\mathsf{X}$).

Lemma 2.3. *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 we can always take a $w = \mathbf{1} \in b\mathsf{X}$ with $\langle w, \delta_x \rangle = w(x) = 1 \neq 0$. □

By the Neumann series lemma, the lifetime value v_σ defined in (1) is the unique fixed point in $b\mathsf{X}$ of the policy operator $T_\sigma \, v = r_\sigma + \beta P_\sigma \, v$. More explicitly,

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

We define the Bellman operator by

$$(Tv)(x) = \max_{a \in \Gamma(x)} \left\{ r(x, a) + \beta \int v(x') P(x, a, \mathrm{d}x') \right\} \qquad (v \in b\mathsf{X}, \; x \in \mathsf{X}). \tag{2}$$

In the current setting,

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

See, for example, Hernández-Lerma and Lasserre (2012) or Bäuerle and Rieder (2011).

To prove Theorem 2.1 we first show that (a)–(b) are equivalent. To show (a) implies (b), assume (a) and fix $x \in \mathsf{X}$ with $v_\sigma(x) \geqslant v_s(x)$ for all $s \in \Sigma$. Then $\delta_x \in \mathcal{D}(\mathsf{X})$ and $\langle v_\sigma, \delta_x \rangle = v_\sigma(x) \geqslant v_s(x) = \langle v_s, \delta_x \rangle$ for all $s \in \Sigma$, so (b) holds. To show (b) implies (a), fix $\rho \in \mathcal{D}(\mathsf{X})$ with $\langle v_\sigma, \rho \rangle \geqslant \langle v_s, \rho \rangle$ for all $s \in \Sigma$. 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 $x \in \mathsf{X}$ and so $\langle v^*, \rho \rangle > \langle v_\sigma, \rho \rangle$. This contradiction proves (a).

To complete the proof of Theorem 2.1, it suffices to show that (a) and (c) are equivalent. That (c) implies (a) is immediate from the definition of optimal policies. Hence we need only show that (a) implies (c). To this end, fix $\bar{x} \in \mathsf{X}$ and $\sigma \in \Sigma$ such that $v_s(x) \leqslant v_\sigma(x)$ for all $s \in \Sigma$. Then $v_\sigma(x) = v^*(x)$. For all $n \in \mathbb{N}$ we have

$$v_\sigma = T_\sigma^n \, v_\sigma \leqslant T_\sigma^n \, v^* \leqslant T^n \, v^* = v^*.$$

The first inequality is due to the fact T_σ is order preserving. 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. Since at $\bar{x}$ we have $v_\sigma(\bar{x}) = v^*(\bar{x})$, it follows that $(T_\sigma^n\, v_\sigma)(\bar{x}) = (T_\sigma^n\, v^*)(\bar{x})$ for all $n \in \mathbb{N}$. Expanding this expression out by $T_\sigma v = r_\sigma + \beta P_\sigma v$ and canceling r_σ and β gives $(P_\sigma^n\, v_\sigma)(\bar{x}) = (P_\sigma^n\, v^*)(\bar{x})$ for all $n \in \mathbb{N}$ and so

$$\int (v^*(x') - v_\sigma(x')) P_\sigma^n(\bar{x}, \mathrm{d}x') = 0 \qquad \text{for all } n \in \mathbb{N}. \tag{3}$$

Let $w \coloneqq v^* - v_\sigma$ and note that $0 \leqslant w$. We claim that $w = 0$. To see this, suppose to the contrary that w is nonzero. In this case, by irreducibility and Proposition 2.2, for each nonzero μ in the positive cone of $b\mathsf{X}'$ we can find an $m \in \mathbb{N}$ such that $\langle \mu, P_\sigma^m\, w \rangle > 0$. Because $\delta_{\bar{x}}$ is a nonzero element of the positive cone of $b\mathsf{X}'$, we can set $\mu = \delta_{\bar{x}}$ to obtain an $m \in \mathbb{N}$ with $(P_\sigma^m\, w)(\bar{x}) > 0$. This contradicts (3), so $w = 0$ holds. In other words, $v_\sigma(x) = v^*(x)$ for all $x \in \mathsf{X}$, as was to be shown.

3. Special Case: Discrete States

We now move to MDPs with finite state spaces, so that $|\mathsf{X}| < \infty$. We say that x in X is *P-accessible* from $\bar{x} \in \mathsf{X}$ when there exists an $m \in \mathbb{N}$ such that $P^m(\bar{x}, x) > 0$. The following result provides additional information in the finite state case.

Theorem 3.1. *If $v_\sigma(\bar{x}) = v^*(\bar{x})$ and x is P_σ-accessible from $\bar{x}$, then $v_\sigma(x) = v^*(x)$*

Proof. Let $v_\sigma(\bar{x}) = v^*(\bar{x})$ and x is P_σ-accessible from $\bar{x}$. Then there exists an $n \in \mathbb{N}$ such that $P^n(\bar{x}, x) > 0$. Moreover, using (3) from the proof of Theorem 2.1 (which does not require the irreducibility condition in the theorem), we have

$$\sum_{x' \in X} (v^*(x') - v_\sigma(x')) P_\sigma^n(\bar{x}, x') = 0. \tag{4}$$

Since $v^*(x') - v_\sigma(x') \geqslant 0$ for all $x' \in \mathsf{X}$ and $P_\sigma^n(\bar{x}, x) > 0$ we must have $v^*(x) - v_\sigma(x) = 0$. As a result, $v_\sigma(x) = v^*(x)$. □

3.1. Three-State Example.

The following example shows that the accessibility assumption in Theorem 3.1 cannot be dropped. The example involves a simple infinite horizon job search model (McCall, 1970), where the job seeker faces wage offer $w \in \mathsf{W} = \{1, 2, 3\}$. The job seeker accepts or rejects at each offer, so her action space

A can be represented by $\{0, 1\}$. Her task is to choose a policy σ mapping W to A that maximizes her income. The Bellman equation is given by

$$(Tv)(w) = \max\left\{\frac{w}{1-\beta}, c + \beta \sum_{w' \in \mathsf{W}} v(w')P(w, w')\right\} \qquad (w \in \mathsf{W}), \tag{5}$$

where c is the unemployment compensation satisfying $0 < c < \min \mathsf{W}$ and $P(w, w')$ is the transition probability from wage offer w to offer w' (see, e.g., Ch.4 of Sargent and Stachurski (2025)). Consider the transition matrix

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

Note that states 2 and 3 are not accessible from state 1. Let $\beta = 0.9$ and $c = 0.5$. Using value function iteration we obtain the value function $v^* = (10, 25.4545, 30)$ and an optimal policy $\sigma = (1, 0, 1)$. Consider an alternative policy $\pi = (1, 1, 0)$. By computing the fixed point of the policy operator T_π, which is given by

$$(T_\pi v)(w) = \pi(w)\frac{w}{1-\beta} + (1 - \pi(w))\left(c + \beta \sum_{w' \in \mathcal{W}} v(w')P(w, w')\right) \qquad (w \in \mathcal{W}),$$

we obtain the lifetime value function v_π for π:

$$v_\pi = (10, 20, 17.2727) \leqslant (10, 25.4545, 30) = v^*.$$

We see that optimality at the point $w = 1$ does not guarantee global optimality when irreducibility fails.

4. Extensions and Future Work

Using MDP optimality results from Bäuerle and Rieder (2011), it is possible to extend our results to the case of unbounded rewards by defining a weight function b on X that is continuous and Borel measurable, with $b(x) \geqslant 1$ for all $x \in \mathsf{X}$. The weighted supremum norm is given by $\|v\|_b = \sup_{x \in \mathsf{X}} |v(x)|/b(x)$. The weight function is chosen so that $\|r\|_b$ is finite. After replacing the supremum with the weighted supremum norm, both the space of bounded measurable functions and the space of continuous bounded functions remain complete. Our main results hold without significant modification.

So far 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 $\mathbb{R}_+$. We define an operator $K = \beta \circ P$ and the corresponding $K_\sigma = \beta_\sigma \circ P_\sigma$. The same result goes through under certain stability assumptions, provided that K_σ is irreducible.

It seems likely that results similar Theorem 2.1 will be valid for standard continuous time MDPs, as well as some of the nonstandard dynamic programs discussed in Bertsekas (2022) and Sargent and Stachurski (2025). We leave these topics for future research.

References

Agarwal, A., S. M. Kakade, J. D. Lee, and G. Mahajan (2021): "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *Journal of Machine Learning Research*, 22, 1–76.

Bartlett, P. L. and A. Tewari (2009): "REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, AUAI Press, 35–42.

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.

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

Khodadadian, S., P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri (2021): "On the Linear Convergence of Natural Policy Gradient Algorithm," *2021 60th IEEE Conference on Decision and Control (CDC)*, 3794–3799.

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

Kumar, N., E. Derman, M. Geist, K. Y. Levy, and S. Mannor (2023): "Policy Gradient for Rectangular Robust Markov Decision Processes," in *Neural Information Processing Systems*.

Lan, G., H. Wang, J. Anderson, C. G. Brinton, and V. Aggarwal (2023): "Improved Communication Efficiency in Federated Natural Policy Gradient via ADMM-based Gradient Updates," *ArXiv*, abs/2310.19807.

McCall, J. J. (1970): "Economics of Information and Job Search," *The Quarterly Journal of Economics*, 84, 113–126.

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

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

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

Sutton, R. S., D. A. McAllester, S. Singh, and Y. Mansour (1999): "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *Neural Information Processing Systems*.

Xiao, L. (2022): "On the convergence rates of policy gradient methods," *Journal of Machine Learning Research*, 23, 1–36.

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