Sufficiency of Deterministic, Stationary, Markovian Policies for MDPs with Infinite Discounted Reward

Shivaram Kalyanakrishnan

September 2021

Abstract

In class, we have defined MDPs and taken infinite discounted reward as the criterion to optimise. We have claimed based on intuition, rather than proof, that to realise optimal behaviour in this setting, it suffices for an agent to consider policies that are deterministic, stationary, and Markovian. In this note, we provide a formal proof of the same claim. In particular, we show that history and stochasticity do not offer any additional benefit in this setting. Central to our proof is the *finite horizon* discounted reward criterion, on which we establish comparisons between different classes of policies. The infinite discounted reward setting is treated as a limiting case of the finite horizon setting.

1 Four Policy Classes

Let us anchor our discussion around a fixed MDP (S, A, T, R, γ), with notations as usual. We shall assume, as is commonly done, that the individual rewards are bounded in $[-R_{\max}, R_{\max}]$ for some known $R_{\max} > 0$. When an agent interacts with the MDP, starting from some state s^0 at time step $t = 0$, it encounters a state-action-reward sequence over time: $s^0, a^0, r^0, s^1, a^1, r^1, s^2, \ldots$. It is this sequence, up to t steps, that is treated as the *history* of the first t steps, $t \geq 0$. However, to simplify our discussion, we leave out the rewards from histories[1], instead taking a t-step history h as

$$h = s^0, a^0, s^1, a^1, s^2, a^2, \ldots, s^t.$$

For history h, let $s(h)$ denote its last entry (a state). Hence, for the t-length history h given above (length t means t actions have been taken), we have $s(h) = s^t$.

Although we are interested in the infinite discounted reward setting, we build out our proof by first considering the finite horizon discounted reward setting. Let us fix a horizon $K \geq 1$ for our discussion. Let $\mathcal{H}^K$ denote the set of all histories of length $0 \leq k \leq K - 1$. In its interaction up to K steps, the agent at any time step will have encountered a history from $h \in H^K$, and must decide which action to take based on h. We consider some natural choices for determining the agent's decision based on the information present in h.

[1] If rewards are generated deterministically, our choice results in no loss of information. Even if rewards are generated stochastically, our eventual claims shall still hold; the student is encouraged to take up this extension as an exercise.

1.1 History-dependent Stochastic Policies

The most general class of policies we can imagine is $\Lambda = \{\lambda : \mathcal{H}^K \to \mathrm{PD}(A)\}$, where $PD(A)$ denotes the set of probability distributions over A. If the agent follows policy $\lambda \in \Lambda$, then from history $h \in \mathcal{H}^K$, the agent takes action $a \in A$ with probability $\lambda(h, a)$. We adopt the convention of subscripting values with the number of steps left up to the horizon, setting the value to 0 when there are no more steps. Thus, for policy $\lambda \in \Lambda$ and $h \in \mathcal{H}^K$, we obtain $V_0^\lambda(h) = 0$, and for $1 \leq k \leq K$,

$$V_k^\lambda(h) = \sum_{a \in A} \lambda(h, a) \sum_{s' \in S} T(s(h), a, s')\{R(s(h), a, s') + \gamma V_{k-1}^\lambda(h.(a, s'))\},$$

where $h.(a, s')$ is the history obtained by appending h with action a and state s'. Although we have gone ahead with our definition above for all histories h and time steps k from their respective sets, note that $V_k^\lambda(h)$ will only be of interest to us when h is of length $K - k$, and moreover, h is a history that can be realised by the underlying MDP. Nonetheless, we lose nothing by extending our definition to a larger set of histories and time steps.

1.2 History-dependent Deterministic Policies

If we constrain the agent to choose actions deterministically for each history, we obtain a class of policies $M = \{\mu : \mathcal{H}^K \to A\}$. As before, for policy $\mu \in M$ and $h \in \mathcal{H}^K$, we set $V_0^\mu(h) = 0$, and for $1 \leq k \leq K$, take

$$V_k^\mu(h) = \sum_{s' \in S} T(s(h), \mu(h), s')\{R(s(h), \mu(h), s') + \gamma V_{k-1}^\mu(h.(\mu(h), s'))\}.$$

1.3 Time-dependent Deterministic Policies

We may further restrict our agent to choose an action deterministically solely based on the current state s and the number of time steps t that have elapsed so far (or equivalently and more conveniently, the number of steps to termination, since the horizon K can be treated as a constant). The resulting set of policies is $\Theta = \{\theta : S \times \{1, 2, \ldots, K\} \to A\}$. While writing down values of such policies, we continue subscripting by the number of steps to go to the horizon, assuming that the policy is passed this same number for action-selection. In summary, we have for $\theta \in \Theta, s \in S$: $V_0^\theta(s) = 0$, and for $1 \leq k \leq K$,

$$V_k^\theta(s) = \sum_{s' \in S} T(s, \theta(s, k), s')\{R(s(h), \theta(s, k), s') + \gamma V_{k-1}^\theta(s')\}.$$

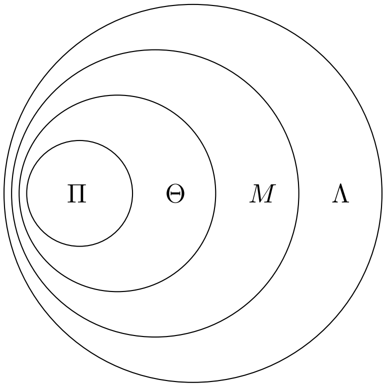
Recall that the property of being time-varying is also called being *nonstationary*, which is another commonly-used term to denote the policies in Θ.

1.4 Deterministic, Stationary, Markovian Policies

Our final class of policies is the familiar one from class: $\Pi = \{\pi : S \to A\}$. For these policies—which are deterministic, Markovian, and stationary—we have for $s \in S$, $V_0^\pi(s) = 0$, and for $1 \leq k \leq K$,

$$V_k^\Pi(s) = \sum_{s' \in S} T(s, \pi(s), s')\{R(s, \pi(s), s') + \gamma V_{k-1}^\pi(s'))\}.$$

It is easy to see that $\Pi \subset \Theta \subset M \subset \Lambda$, which is also shown diagrammatically below.



We shall show that in spite of being a restricted class, Π suffices to express optimal behaviour in MDPs with infinite discounted reward. We begin with Λ and go through M, Θ, and Π in sequence.

2 Results

For stating our results, let us fix an arbitrary start state $s^0 \in S$ (which also doubles up as a starting history). Since the value of any policy at s^0 is a scalar, we can be sure that there will be an optimal policy in each class; define these policies as below.

$$\begin{aligned}
\lambda^\star &= \operatorname*{argmax}_{\lambda \in \Lambda} V_K^{\lambda}(s^0).\\
\mu^\star &= \operatorname*{argmax}_{\mu \in M} V_K^{\mu}(s^0).\\
\theta^\star &= \operatorname*{argmax}_{\theta \in \Theta} V_K^{\theta}(s^0).\\
\pi^\star &= \operatorname*{argmax}_{\pi \in \Pi} V_K^{\pi}(s^0).
\end{aligned}$$

In the next sections, we prove the following results.

$$V_K^{\lambda^\star}(s^0) = V_K^{\mu^\star}(s^0) = V_K^{\theta^\star}(s^0). \tag{1}$$

$$V_K^{\pi^\star}(s^0) \geq V_K^{\theta^\star}(s^0) - 2\gamma^K \frac{R_{\max}}{1-\gamma}. \tag{2}$$

From (1), we see that deterministic time-dependent are as powerful as stochastic, history-dependent ones for the finite horizon setting. Since $\Pi \subset \Theta$, it is self-evident that $V_K^{\pi^\star}(s^0) \leq V_K^{\theta^\star}(s^0)$. Since $\gamma < 1$. we have from (2) that

$$\lim_{K \to \infty} V_K^{\pi^\star}(s^0) = \lim_{K \to \infty} V_K^{\theta^\star}(s^0),$$

which implies $\pi^\star$ gives as high a value from s_0 as any of $\theta^\star$, $\mu^\star$, and $\lambda^\star$ as $K \to \infty$ (the infinite discounted reward setting). We already know that there is an optimal policy in Π that yields the highest value from every start state (among all policies in Π). If we choose such as policy as our $\pi^\star$, then we have effectively shown that in spite of belonging to Π, $\pi^\star$ cannot be improved upon by any policy in Λ in the infinite discounted reward setting, regardless of starting state.

All that is left to do now is to get through with our proofs of (1) and (2).

3 Proof of (1)

We set out to prove the following statement for $1 \leq k \leq K$: for $h \in \mathcal{H}^K$,

$$V_k^{\lambda^\star}(h) = V_k^{\mu^\star}(h) = V_k^{\theta^\star}(s(h)).$$

We use induction, taking as base case our definition for $k = 0$: for $h \in \mathcal{H}^K$,

$$V_0^{\lambda^\star}(h) = V_0^{\mu^\star}(h) = V_0^{\theta^\star}(h(s)) = 0.$$

Assuming the statement to be true for $k-1$, we shall show it is true for k, for $k \in \{1, 2, \ldots, K\}$. The Bellman equations for $\lambda^\star : \mathcal{H}^K \to PD(A)$ give us for $h \in \mathcal{H}^K$:

$$V_k^{\lambda^\star}(h) = \sum_{a \in A} \lambda^\star(h, a) \sum_{s' \in S} T(s(h), a, s')\{R(s(h), a, s') + \gamma V_{k-1}^{\lambda^\star}(h.(a, s'))\},$$

which, because of our induction hypothesis, can be written as

$$V_k^{\lambda^\star}(h) = \sum_{a \in A} \lambda^\star(h, a) \sum_{s' \in S} T(s(h), a, s')\{R(s(h), a, s') + \gamma V_{k-1}^{\theta^\star}(s')\}.$$

The RHS is of the form $\sum_a \lambda^\star(h, a) f(a)$, where $\lambda^\star(h, \cdot)$ is a probability distribution over A. It is clear that putting the entire probability mass on some action $a \in A$ maximising f will not decrease the k-step value. In other words, we may choose a policy $\mu : \mathcal{H}^T \to A$ such that for $h \in \mathcal{H}^K$, $\mu(h) = a \implies \lambda^\star(h, a) > 0$. Then we observe that

$$\begin{aligned}
V_k^{\mu}(h) &= \sum_{s' \in S} T(s(h), \mu(s(h)), s')\{R(s(h), \mu(s(h)), s') + \gamma V_{k-1}^{\mu}(h.(\mu(s(h)), s'))\} \\
&= \sum_{s' \in S} T(s(h), \mu(s(h)), s')\{R(s(h), \mu(s(h)), s') + \gamma V_{k-1}^{\theta^\star}(s')\} \\
&= \max_{a \in A} \sum_{s' \in S} T(s(h), a, s')\{R(s(h), a, s') + \gamma V_{k-1}^{\theta^\star}(s')\} \\
&\geq V_k^{\lambda^\star}(h).
\end{aligned}$$

Now, consider a policy $\theta \in \Theta$ such that for $s \in S$, $j \in \{1, 2, \ldots, k\}$,

$$\theta(s, j) = \operatorname*{argmax}_{a \in A} \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V_{j-1}^{\theta}(s')\}.$$

From the working shown above, we verify that for $h \in \mathcal{H}^K$, $V_k^{\theta}(s(h)) = V_k^{\mu}(h)$.

Our inductive proof has thereby established that for any $h \in \mathcal{H}^K$, we can choose policies $\mu \in M$ and $\theta \in \Theta$ such that they yield at least as high k-step values as $\lambda^\star \in \Lambda$. Recall that $\lambda^\star$ itself is a policy that maximises the k-step value from s^0. By applying our result on s^0, it is clear that there exist $\mu^\star \in M$ and $\theta^\star \in \Theta$ such that

$$V_K^{\lambda^\star}(s^0) = V_K^{\mu^\star}(s^0) = V_K^{\theta^\star}(s^0).$$

Now, observe that there is no need for $\theta^\star$ to depend on s^0; in fact, it can be defined as: for $s \in S, k \in \{1, 2, \ldots, K\}$:

$$\theta^\star(s, k) = \operatorname*{argmax}_{a \in A} \sum_{s' \in S} T(s, a, s')\{R(s, a, s') + \gamma V_{k-1}^{\theta^\star}(s')\},$$

with arbitrary tie-breaking. In vector notation, we obtain a succinct formula for computing the value function of $\theta^\star$: $V_K^{\theta^\star} = (B^\star)^K(\mathbf{0})$, where $B^\star$ is the Bellman optimality operator, and $\mathbf{0}$ is the zero vector.

The main contribution of the finite horizon setting to our proof is in allowing us to eliminate the need for full histories—easily argued since there is no decision-making to be done beyond the horizon, and so no use carrying any history beyond the horizon. It is a separate matter that even time-dependence is not needed when the number of steps is infinite; this is the proof coming up in the next section.

4 Proof of (2)

We have just seen that for $K \geq 1$,

$$V_K^{\theta^\star} = (B^\star)^K(\mathbf{0}).$$

Now, let π be an optimal policy from Π for the infinite discounted reward setting that we have discussed in class. We must be careful to distinguish π from $\pi^\star \in \Pi$ introduced in Section 2, which is maximal (from s^0) for a horizon of K, and hence could be different from π. If B^π is the Bellman operator for π, it is easy to see that for $K \geq 1$,

$$V_K^\pi = (B^\pi)^K(\mathbf{0}).$$

We know that $B^\star$ and B^π both have the same unique fixed point $V^\star$, which is the optimal value function for the infinite discounted reward setting. We obtain the result below by applying the triangle inequality and Banach's Fixed Point Theorem, while noting that values in the infinite discounted reward setting must lie bounded in $[-\frac{R_{\max}}{1-\gamma}, \frac{R_{\max}}{1-\gamma}]$.

$$\begin{aligned}
\|V_K^\pi - V_K^{\theta^\star}\|_\infty &= \|V_K^\pi - V^\star + V^\star - V_K^{\theta^\star}\|_\infty \\
&\leq \|V_K^\pi - V^\star\|_\infty + \|V_K^{\theta^\star} - V^\star\|_\infty \\
&= \|(B^\pi)^K(\mathbf{0}) - V^\star\|_\infty + \|(B^\star)^K(\mathbf{0}) - V^\star\|_\infty \\
&\leq \gamma^K\|\mathbf{0} - V^\star\|_\infty + \gamma^K\|\mathbf{0} - V^\star\|_\infty \\
&= 2\gamma^K\|V^\star\|_\infty \\
&\leq 2\gamma^K\frac{R_{\max}}{1-\gamma}.
\end{aligned}$$

We have

$$V_K^{\pi^\star}(s^0) \geq V_K^{\pi}(s^0) \geq V_K^{\theta^\star}(s^0) - 2\gamma^K \frac{R_{\max}}{1-\gamma},$$

which is our claim in (2).