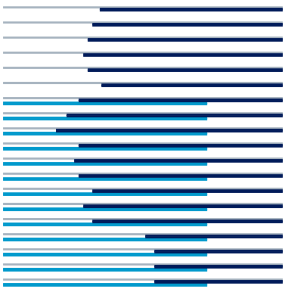
Ergodic MDPs Admit Self-Optimising Policies

Shane Legg
Marcus Hutter



Ergodic MDPs Admit Self-Optimising Policies No. IDSIA-21-04
November 9, 2004

IDSIA / USI-SUPSI
Dalle Molle Institute for Artificial Intelligence
Galleria 2, 6928 Manno, Switzerland

Ergodic MDPs Admit Self-Optimising Policies*

Shane Legg†
Marcus Hutter‡

November 9, 2004

Abstract

Markov decision processes (MDPs) are an important class of dynamic systems with many applications. Intuitively it seems clear that if an MDP is ergodic then it should admit self-optimising policies. This is because ergodicity ensures that an MDP's state-transition space can be freely explored which should allow a sufficiently accurate model of the MDP to be constructed. In this paper we prove that this intuition is indeed correct, though the full analysis is surprisingly complex.

1 Introduction

Intuitively it seems that if an MDP is finite, stationary and ergodic then it should be possible for an adaptive policy to eventually achieve an optimal level of expected reward per cycle. This is because ergodicity ensures that the MDP's transition space can be freely explored, thus allowing an agent to construct an accurate model of the MDP after a sufficient number of cycles. From such a model it should then be possible to discover the optimal policy for the environment or at least something very close.

In this paper we prove that this intuition is indeed correct and ergodic MDPs do admit so called *self-optimising policies*. While the intuition described above is quite simple and, as we will see, more or less correct, the analysis needed to prove this is quite significant.

The paper assumes a familiarity with the idea of an agent interacting with an environment in order to receive reward. This is known as the agent–environment model in the context of reinforcement learning, or the controller–plant model in the context of control theory. Here we use the reinforcement learning terminology, though the two are equivalent. For an excellent introduction to reinforcement learning see [7].

The mathematical notation used in this paper is fairly standard though a few things should be clarified. We will represent real valued matrices with capital letters, for example $A \in \mathbb{R}^{n \times m}$. By A_{ij} we mean the single scalar element of A on the i^{th} row and j^{th} column.

*This work was supported by SNF grant 2100-67712.02.

†shane@idsia.ch
‡marcus@idsia.ch

By A_{*j} we mean the j^{th} column of A and similarly for A_{i*}. We represent vectors with a bold lowercase variable, for example $\mathbf{a} \in \mathbb{R}^{\mathbf{n}}$. Similar to the case for matrices, by $\mathbf{a}_i$ we mean the i^{th} element of $\mathbf{a}$. In some situations a matrix or vector may already have other indexes, in this case we place square brackets around it and then index so as to avoid confusion, for example $[\mathbf{r}_\pi]_i$ is the i^{th} element of the vector $\mathbf{r}_\pi$. We represent the classical adjoint of a matrix A by $\text{adj}(A)$ and the determinant by $\det(A)$.

2 Markov Decision Processes

The formal model of agent–environment interaction that we use is that of chronological systems developed in [6] and [4], however almost all of the results employ nothing more than standard matrix algebra and probability measures defined over spaces of strings and so a familiarity with these should suffice. Here we will very briefly summarise the essential definitions.

2.1 Definition. An **Agent** is a tuple $(\mathcal{X}, \mathcal{Y}, \pi)$ where $\mathcal{X}$ is a recursive prefix free language called the **perception space**, $\mathcal{Y}$ is a recursive prefix free language called the **action space** and π is a chronological system $\pi : (\mathcal{Y} \times \mathcal{X})^* \times \mathcal{Y} \to [0, 1]$ called the **policy**.

2.2 Definition. An **Environment** is a tuple $(\mathcal{Y}, \mathcal{X}, \mu)$ where $\mathcal{Y}$ is a recursive prefix free language called the **action space**, $\mathcal{X}$ is a recursive prefix free language called the **perception space**, and $\mu : (\mathcal{Y} \times \mathcal{X})^* \to [0, 1]$ is chronological system. We further define $\mathcal{X} := \mathcal{R} \times \mathcal{O}$ where $\mathcal{R}$ is the **reward space** and $\mathcal{O}$ is the **observation space**. If the action and perception spaces are finite we call the system a **Finite Environment**.

Whenever in the text we write $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{R}$ or $\mathcal{O}$ we will assume that they fit the above definitions. We will also use lower case variables such as y_k, x_k, r_k and o_k to denote elements of these sets. In particular whenever we have $x_k \in \mathcal{X}$ we take this to be defined $x_k := r_k o_k$ and simply write r_k or o_k to mean the corresponding components of x_k, that is, the components of the perception in the k^{th} cycle.

Strictly speaking r_k is an element of the reward space $\mathcal{R}$ which is a recursive prefix free language. Thus r_k is a string. However we are really only interested in the value that r_k represents when mapped to $\mathbb{R}$ by some simple recursive injective function $R : \mathcal{R} \to \mathbb{R}$. As R performs only a technical role, rather than always writing $R(r_k)$ we will interpret r_k to be its associated real value were convenient.

In the agent–environment model we entangle the agent $(\mathcal{X}, \mathcal{Y}, \pi)$ and the environment $(\mathcal{Y}, \mathcal{X}, \mu)$ so that the output of one forms the input to the other and vice versa. Our analysis in this paper will revolve around various agents and environments considered either separately or entangled to form a single system.

2.3 Definition. A **finite stationary MDP** is a finite environment $(\mathcal{Y}, \mathcal{X}, \mu)$ such that $\forall k \in \mathbb{N}$, $\forall yx_{1:k} \in (\mathcal{Y} \times \mathcal{X})^k$ we have

$$\mu(x_k | yx_{<k}y_k) = \mu(x_k | x_{k-1}, y_k).$$

Thus the next perception only ever depends on the most recent action and perception. These might seem limited and simple but they are actually very general with many real world and theoretical applications. Indeed some have proposed that the physical universe may be no more than an enormous stationary finite MDP!

Of particular interest to us will be MDPs which are *ergodic*. There are many definitions of *ergodicity* (see for example [5]) and the word has different meanings in different contexts. The following simple definition is relevant to our work and comes from [4]:

2.4 Definition. An environment $(\mathcal{Y}, \mathcal{X}, \mu)$ is **ergodic** if there exists an agent $(\mathcal{X}, \mathcal{Y}, \pi)$ such that under policy π every possible observation $o \in \mathcal{O}$ occurs infinitely often with probability 1.

Intuitively this means that the environment never becomes restricted to some subset of possibilities, instead everything that is possible in the environment initially always remains possible. This is important for adaptive agents because it means that the agent can never make a "mistake" which is impossible to recover from. Without this condition an agent could take some action after which the optimal average reward per cycle for that environment is unattainable. This certainly is a restrictive assumption and few real world situations are truly ergodic. However many are almost ergodic, in particular students or children are usually given environments in which the mistakes they make while they learn do not carry long term negative consequences. Naturally our adaptive agents require something similar if we want them to be able to always achieve optimal behaviour after sufficient experience.

It is clear from the definition that we can represent a stationary finite MDP as a three dimensional Cartesian tensor $D \in \mathbb{R}^{n_1 \times n_2 \times n_1}$ defined $\forall x, x' \in \mathcal{X}$, $\forall y \in \mathcal{Y}$,

$$D_{xyx'} := \mu(x'|x, y),$$

where we have assumed, without loss of generality, that $\mathcal{X} := \{1, \ldots, n_1\}$ and $\mathcal{Y} := \{1, \ldots, n_2\}$ for $n_1, n_2 \in \mathbb{N}$. Because $\mathcal{X}$ and $\mathcal{Y}$ are finite recursive prefix free languages we could have allowed them to be arbitrary and associated with each element of $\mathcal{X}$ and $\mathcal{Y}$ an integer by means of a recursive coding function, however this just adds unnecessary complexity to the notation.

Let $(\mathcal{X}, \mathcal{Y}, \pi)$ be a stationary agent such that $\forall k \in \mathbb{N}$, $\forall yx_{<k}y_k \in (\mathcal{Y} \times \mathcal{X})^{k-1} \times \mathcal{Y}$: $\pi(y_k|yx_{<k}) = \pi(y_k|x_{k-1})$. That is, under the policy π the distribution of actions depends on only the last perception. It follows that the equation for the k^{th} perception x_k given history $yx_{<k} \in (\mathcal{Y} \times \mathcal{X})^{k-1}$ is,

$$\pi(y_k|yx_{<k})\mu(x_k|yx_{<k}, y_k) = \pi(y_k|x_{k-1})\mu(x_k|x_{k-1}, y_k) = \pi(y_k|x_{k-1})D_{x_{k-1}y_kx_k}.$$

Thus, for a given μ and π the next perception x_k depends on only the previous perception x_{k-1} in a way that is independent of k, that is, π and μ together form a stationary Markov chain.

This allows us to express the interaction of the agent and the environment as a square stochastic matrix $T \in \mathbb{R}^{n_1 \times n_1}$ defined

$$T_{xx'} := \sum_{y \in \mathcal{Y}} \pi(y|x)\mu(x'|x, y) = \sum_{y \in \mathcal{Y}} \pi(y|x)D_{xyx'} \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are sets of integers as defined previously.

We say that a Markov chain is ergodic if all observations are visited infinitely often with probability 1. Thus a stationary MDP is ergodic if and only if there exists a stationary policy such that the agent–environment interaction forms an ergodic Markov chain.

It should be noted that this characterisation of the interaction between μ and π as a stochastic matrix T is only possible if μ is a stationary MDP and π is a stationary policy. Fortunately this is all we will need for optimality, though we will briefly have to consider non-stationary policies in Section 5 in order to prove this. It is worth keeping in mind that when we see a matrix T this represents a specific environment and agent interacting rather than just an environment. Perhaps we should indicate this by always writing $T_{\pi\mu}$ however that would get messy as we will often take powers of T and index its elements.

While matrix notation has its limitations, one important advantage is that the probability of transiting between any two states can be easily computed by taking powers of T. For example, if we have an observation i then the probability that exactly k cycles later the observation will be j, is given by $[T^k]_{ij}$. The other advantage is, of course, that a full range of linear algebra techniques are now available to aid us.

3 Expected Value

3.1 Definition. For a chronological environment $(\mathcal{Y}, \mathcal{X}, \mu)$ and an agent $(\mathcal{X}, \mathcal{Y}, \pi)$ the **expected total value** in cycles $k \in \mathbb{N}$ to $m \in \mathbb{N}$ given history $yx_{<k} \in (\mathcal{Y} \times \mathcal{X})^{k-1}$ is defined to be

$$\begin{aligned} V_{km}^{\pi\mu}(yx_{<k}) &:= \mathbf{E}\left(\sum_{i=k}^{m} r_i \middle| yx_{<k}\right) \\ &= \sum_{yx_{k:m}} \pi(y_{k:m}|yx_{<k}, x_{k:m-1})\mu(x_{k:m}|yx_{<k}, y_{k:m}) \sum_{i=k}^{m} r_i, \end{aligned}$$

for $m \geq k$ and 0 otherwise. In the expected value we have taken the expectation with respect to the distribution over sequences defined by combining the policy π and the environmental distribution μ.

In the case of a deterministic policy the equation reduces to

$$V_{km}^{\pi\mu}(yx_{<k}) = \sum_{yx_{k:m} \in P(yx_{<k})} \mu(x_{k:m}|yx_{<k}, y_{k:m}) \sum_{i=k}^{m} r_i$$

where $P(yx_{<k}) := \{\hat{y}\hat{x}_{k:k+n} \in (\mathcal{Y} \times \mathcal{X})^n : \pi(\hat{y}_{k:k+n}|yx_{<k}, \hat{x}_{k:k+n}) = 1, n \in \mathbb{N}\}$ is the set of all futures consistent with the policy π given history $yx_{<k}$.

Often we are interested in the average expected reward per cycle rather than the total. In some cases this value will also exist in the limit:

3.2 Definition. For a chronological environment $(\mathcal{Y}, \mathcal{X}, \mu)$ and an agent $(\mathcal{X}, \mathcal{Y}, \pi)$ the **expected average value** in cycles $k \in \mathbb{N}$ to $m \in \mathbb{N}$ given history $yx_{<k} \in (\mathcal{Y} \times \mathcal{X})^{k-1}$ is defined to be

$$\bar{V}_{km}^{\pi\mu}(yx_{<k}) := \frac{1}{m} V_{km}^{\pi\mu}(yx_{<k}).$$

Additionally we define the **expected long run average value** to be

$$\bar{V}_{k\infty}^{\pi\mu}(yx_{<k}) := \lim_{m\to\infty} \frac{1}{m} V_{km}^{\pi\mu}(yx_{<k})$$

when the limit exists.

Other than considering the total value and related average, another possibility is to discount by some $\gamma_i \in [0, 1]$ for each cycle.

3.3 Definition. For a chronological environment $(\mathcal{Y}, \mathcal{X}, \mu)$ and an agent $(\mathcal{X}, \mathcal{Y}, \pi)$ the **discounted future value** is defined to be

$$\begin{aligned}
\hat{V}_{k\gamma}^{\pi\mu}(yx_{<k}) &:= \frac{1}{\Gamma_k} \mathbf{E}\left(\sum_{i=k}^{\infty} \gamma_i r_i \middle| yx_{<k} \right) \\
&= \frac{1}{\Gamma_k} \lim_{m\to\infty} \sum_{yx_{k:m}} \pi(y_{k:m}|yx_{<k}, x_{k:m-1}) \mu(x_{k:m}|yx_{<k}, y_{k:m}) \sum_{i=k}^{m} \gamma_i r_i,
\end{aligned}$$

where $\forall i \in \mathbb{N}_0 : \gamma_i \in [0, 1]$ and $\Gamma_k := \sum_{i=k}^{\infty} \gamma_i < \infty$.

One common form of discounting is *geometric discounting* where we set $\forall i \in \mathbb{N}_0 : \gamma_i = \alpha^i$ for some $\alpha \in (0, 1)$. In this case the value of $\hat{V}_{k\alpha}^{\pi\mu}$ is bounded as r_i is bounded by some $r^+ \in \mathbb{R}$ and so $\forall m \in \mathbb{N}$,

$$\sum_{i=k}^{m} \alpha^i r_i \leq r^+ \sum_{i=k}^{\infty} \alpha^i \leq \frac{\alpha r^+}{1-\alpha}.$$

It then follows that,

$$\begin{aligned}
\hat{V}_{k\alpha}^{\pi\mu}(yx_{<k}) &\leq \frac{\alpha r^+}{1-\alpha} \frac{1}{\Gamma_k} \lim_{m\to\infty} \sum_{yx_{k:m}} \pi(y_{k:m}|yx_{<k}, x_{k:m-1}) \mu(x_{k:m}|yx_{<k}, y_{k:m}) \\
&\leq \frac{\alpha r^+}{\Gamma_k (1-\alpha)}.
\end{aligned}$$

As the series for $\hat{V}_{k\alpha}^{\pi\mu}$ is also strictly non-decreasing it follows that the series converges, that is, $\hat{V}_{k\alpha}^{\pi\mu}$ exists.

The intuition behind our choice of notation for the value function is that the flat bar in the symbol $\bar{V}$ represents both an average and indicates that the weighting over the rewards considered is even. The peaked hat in the symbol $\hat{V}$ indicates that some rewards may be more heavily weighted than others.

For each of the above definitions, when $k = 1$ there is no history, that is, $yx_{<k} = \epsilon$, the null string. In this case we simplify the notation slightly by defining $V_{1m}^{\pi\mu} := V_{1m}^{\pi\mu}(\epsilon)$ and similarly for $\bar{V}_{1m}^{\pi\mu}$, $\bar{V}_{1\infty}^{\pi\mu}$ and $\hat{V}_{1\gamma}^{\pi\mu}$.

In a similar fashion to the transition matrix T defined above we can also express the reward under the stationary policy π as a vector $\mathbf{r}_\pi$ defined

$$[\mathbf{r}_\pi]_x := \mathbf{E}(R_{xy}|x) = \sum_{y\in\mathcal{Y}} \pi(y|x) R_{xy}$$

where $R_{xy} \in \mathbb{R}^{n_1 \times n_2}$ is the expected reward matrix when choosing action y after perception x.

This allows us to express the geometric discounted expected value for each initial perception $x_1 \in \mathcal{X} = \{1, \ldots, n_1\}$ as a vector of value functions,

$$\hat{\mathbf{V}}_{1\alpha}^{\pi\mu} := \begin{pmatrix} \hat{V}_{1\alpha}^{\pi\mu}(1) \\ \vdots \\ \hat{V}_{1\alpha}^{\pi\mu}(n_1) \end{pmatrix} = \sum_{k=0}^{\infty} \alpha^k T^k \mathbf{r}_\pi = \left(\sum_{k=0}^{\infty} \alpha^k T^k \right) \mathbf{r}_\pi = (I - \alpha T)^{-1} \mathbf{r}_\pi \tag{2}$$

where $\alpha \in (0, 1)$ is the geometric discount factor.

Similarly we can express the expected long run average value vector in this stationary case as,

$$\bar{\mathbf{V}}_{1\infty}^{\pi\mu} = \begin{pmatrix} \bar{V}_{1\infty}^{\pi\mu}(1) \\ \vdots \\ \bar{V}_{1\infty}^{\pi\mu}(n_1) \end{pmatrix} = \left(\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} T^k \right) \mathbf{r}_\pi, \tag{3}$$

if the value of the limit exists.

3.4 Definition. For a chronological environment $(\mathcal{Y}, \mathcal{X}, \mu)$ the **optimal policy**, denoted π^μ, is defined as:

$$\pi^\mu := \arg\max_\pi \bar{V}_{1\infty}^{\pi\mu},$$

where the maximum is taken over all policies, including non-stationary ones.

In some sense the optimal policy is the ideal policy. However the optimal policy is usually only optimal with respect to the specific environment for which it was designed. If we don't know the specific details of the environment that the agent will face in advance the best we can do is to have a policy which will adapt to the environment based on experience. In such a situation the policy is unlikely to be optimal as it will probably make some non-optimal actions as it learns about the environment it faces. In this situation the following concept is useful.

3.5 Definition. We say that a policy π is **self-optimising** in an environment $(\mathcal{Y}, \mathcal{X}, \mu)$ if its expected average value converges to the optimal expected average value as $m \to \infty$, that is,

$$\bar{V}_{1m}^{\pi\mu} \longrightarrow \bar{V}_{1m}^{\pi^\mu\mu}.$$

Intuitively this means that the expected performance of the policy in the long run is as good as an optimal policy which was designed with complete knowledge of the environment in advance. Classes of environments which admit self-optimising policies are important because they are environments in which it is possible for general purpose agents to adapt their behaviour until eventually their actions become optimal.

4 Analysis of Stationary Markov Chains

In this section we will establish some of the properties of Markov chains that we will require. The results are technical in nature and thus the proofs may be skipped on a first reading.

Our first lemma shows that the term $(I - \alpha T)^{-1}$, which appears in the definition of $\hat{\mathbf{V}}^{\pi\mu}_{1\alpha}$, can be expanded using a Taylor series. The proof of this lemma and the following lemma and theorem are based on the proof of Proposition 1.1 from Section 4.1 of [2].

4.1 Lemma. *For a stochastic matrix $T \in \mathbb{R}^{n\times n}$ and scalar $\alpha \in (0,1)$ there exist stochastic matrices $T^* \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{n\times n}$ such that*

$$(I - \alpha T)^{-1} = (1-\alpha)^{-1}T^* + H + O(|1-\alpha|)$$

where $\lim_{\alpha\to 1} O(|1-\alpha|) = 0$.

Proof. Define the $n \times n$ matrix,

$$M(\alpha) := (1-\alpha)(I - \alpha T)^{-1}.$$

Applying the matrix inversion formula we see that

$$M(\alpha) = (1-\alpha)\frac{\operatorname{adj}(I-\alpha T)}{\det(I-\alpha T)},$$

where the determinant $\det(I - \alpha T)$ is an n^{th} order polynomial in α and the classical adjoint $\operatorname{adj}(I - \alpha T)$ is an $n \times n$ matrix of $n-1^{th}$ order polynomials in α. Therefore $M(\alpha)$ can be expressed as an $n \times n$ matrix where each element is either zero or a fraction of two polynomials in α that have no common factors.

We know that the denominator polynomials of $M(\alpha)$ cannot have 1 as a root as this would imply that the corresponding element of $M(\alpha) \to \infty$ as $\alpha \to 1$. This cannot happen because, following from Equation (2),

$$(1-\alpha)^{-1}M(\alpha)\mathbf{r}_\pi = (I - \alpha T)^{-1}\mathbf{r}_\pi = \hat{\mathbf{V}}^{\pi\mu}_{1\alpha}$$

where $\forall i \in \{1,\dots,n\} : \left|\left[\hat{\mathbf{V}}^{\pi\mu}_{1\alpha}\right]_i\right| \leq (1-\alpha)^{-1}\max_k \left|[\mathbf{r}_\pi]_k\right|$. Clearly then the absolute value of the elements of $M(\alpha)\mathbf{r}_\pi$ are bounded by $\max_k \left|[\mathbf{r}_\pi]_k\right|$ for $\alpha < 1$. Therefore we can express the ij^{th} element of $M(\alpha)$ as

$$M_{ij}(\alpha) = \frac{\gamma(\alpha-\zeta_1)\cdots(\alpha-\zeta_p)}{(\alpha-\xi_1)\cdots(\alpha-\xi_q)}$$

where $\gamma, \zeta_i, \xi_j \in \mathbb{R}$ for all $i \in \{1,\dots p\}$ and $j \in \{1,\dots q\}$.

Using this expression we can take a Taylor expansion of $M(\alpha)$ about 1 as follows. Firstly, define the matrix $T^* \in \mathbb{R}^{n\times n}$ as

$$T^* := \lim_{\alpha\to 1} M(\alpha)$$

and the matrix $H \in \mathbb{R}^{n\times n}$ as

$$H_{ij} := -\frac{\partial}{\partial\alpha} M_{ij}(\alpha)]\Big|_{\alpha=1}. \tag{4}$$

That is, H is a matrix having as it ij^{th} element the first derivative of $-M_{ij}(\alpha)$ with respect to α evaluated at $\alpha = 1$.

From the equation for a first order Taylor expansion,

$$M(\alpha) = T^* + (1-\alpha)H + O((1-\alpha)^2)$$

where $O((1-\alpha)^2)$ is an α-dependent matrix such that

$$\lim_{\alpha\to 1} \frac{O((1-\alpha)^2)}{1-\alpha} = 0.$$

Dividing through by $(1-\alpha)$ we get

$$(1-\alpha)^{-1}M(\alpha) = (1-\alpha)^{-1}T^* + H + O(|1-\alpha|)$$

where $\lim_{\alpha\to 1} O(|1-\alpha|) = 0$. The result then follows as $(I-\alpha T)^{-1} = (1-\alpha)^{-1}M(\alpha)$ by definition. □

We will soon show that T^* as defined above plays a significant role in the analysis. Before looking at this more closely, we will firstly prove some useful identities.

4.2 Lemma. *It follows from the definitions of T^* and $M(\alpha)$ that*

$$T^* = T^*T = TT^* = T^*T^*$$

and for $k \in \mathbb{N}$

$$(T - T^*)^k = T^k - T^*.$$

Proof. By subtracting the identity $\alpha I = \alpha(I-\alpha T)(I-\alpha T)^{-1}$ from the identity $I = (I-\alpha T)(I-\alpha T)^{-1}$ we see that

$$(1-\alpha)I = (I-\alpha T)(1-\alpha)(I-\alpha T)^{-1}$$

and thus

$$\alpha T(1-\alpha)(I-\alpha T)^{-1} = (1-\alpha)(I-\alpha T)^{-1} + (\alpha-1)I.$$

Taking $\alpha \to 1$ gives

$$\lim_{\alpha\to 1} \alpha T \cdot \lim_{\alpha\to 1}(1-\alpha)(I-\alpha T)^{-1} = \lim_{\alpha\to 1}(1-\alpha)(I-\alpha T)^{-1} + \lim_{\alpha\to 1}(\alpha-1)I$$

which, using the definition of $M(\alpha)$, becomes

$$T \cdot \lim_{\alpha \to 1} M(\alpha) = \lim_{\alpha \to 1} M(\alpha).$$

Finally using the definition of T^* this reduces to just

$$TT^* = T^*.$$

Using essentially the same argument it can also be shown that $T^*T = T^*$. It then immediately follows that $\forall k \in \mathbb{N} : T^k T^* = T^* T^k = T^*$.

From the relation $TT^* = T^*$ it follows that $T^* - \alpha TT^* = T^* - \alpha T^*$ and so $(I - \alpha T)T^* = (1 - \alpha)T^*$ and thus

$$T^* = (1 - \alpha)(I - \alpha T)^{-1} T^*.$$

Taking $\alpha \to \infty$ gives

$$\lim_{\alpha \to 1} T^* = \lim_{\alpha \to 1} (1 - \alpha)(I - \alpha T)^{-1} \cdot \lim_{\alpha \to 1} T^*$$

which by the definition of T^* is just

$$T^* = T^* T^*.$$

This establishes the first result.

The second result will be proven by induction. Trivially $(T - T^*)^1 = T^1 - T^*$ which establishes the case $k = 1$. Now assume that the induction hypothesis holds for the k^{th} case and consider the $(k+1)^{th}$ case:

$$\begin{aligned} (T - T^*)^{k+1} &= (T - T^*)^k (T - T^*) \\ &= (T^k - T^*)(T - T^*) \\ &= T^{k+1} - T^k T^* - T^* T^k + T^* T^* \\ &= T^{k+1} - T^*. \end{aligned}$$

The second line follows from the induction assumption and the final line from the results above. □

We will use these simple relations frequently in the proofs that follow without further comment. Now we can prove an important result about the structure of T^*: It is the limiting average distribution for the matrix T.

4.3 Theorem. *For a stochastic matrix $T \in \mathbb{R}^{n \times n}$ and $\forall m \in \mathbb{N}$,*

$$T^* = \frac{1}{m} \sum_{k=0}^{m-1} T^k + \frac{1}{m}(T^m - I)H.$$

where $H \in \mathbb{R}^{n \times n}$ is the matrix that satisfies Lemma 4.1.

Proof. As T is a stochastic matrix, from Lemma 4.1 we see that there exist matrices $H \in \mathbb{R}^{n\times n}$ and $T^* \in \mathbb{R}^{n\times n}$ such that

$$H = (I - \alpha T)^{-1} - (1-\alpha)^{-1}T^* - O(|1-\alpha|) \tag{5}$$

where $\alpha \in (0,1)$ and $\lim_{\alpha\to 1} O(|1-\alpha|) = 0$.

However from the geometric series equations it follows that

$$\begin{aligned}
(I-\alpha T)^{-1} - (1-\alpha)^{-1}T^* &= \sum_{k=0}^{\infty}\alpha^k T^k - T^*\sum_{k=0}^{\infty}\alpha^k = \sum_{k=0}^{\infty}\alpha^k(T^k - T^*) \\
&= I - T^* + \sum_{k=1}^{\infty}(\alpha(T-T^*))^k \\
&= I - T^* + \frac{\alpha(T-T^*)}{I-\alpha(T-T^*)} \\
&= (I-\alpha(T-T^*))^{-1} - T^*.
\end{aligned}$$

Substituting this result into Equation (5) and taking $\alpha \to 1$,

$$\begin{aligned}
H &= \lim_{\alpha\to 1}\left[(I-\alpha(T-T^*))^{-1} - T^* - O(|1-\alpha|)\right] \\
&= (I - T + T^*)^{-1} - T^*.
\end{aligned}$$

Multiplying by $(I - T + T^*)$ and then T^* we see that

$$\begin{aligned}
(I-T+T^*)H &= I - (I-T+T^*)T^* \\
H - TH - T^*H &= I - T^* + TT^* - T^*T^* = I - T^* \\
T^*H - T^*H - T^*H &= T^* - T^* \\
T^*H &= 0.
\end{aligned}$$

It now also follows that $H - TH = I - T^*$ and so $T^* + H = I + TH$.

Multiplying by T^k on the left for $k \in \mathbb{N}_0$ now gives

$$T^* + T^kH = T^k + T^{k+1}H.$$

Summing over $k = 0, 1, \ldots, m-1$ and cancelling equal terms and dividing through by m produces

$$\begin{aligned}
mT^* + \sum_{k=0}^{m-1}T^kH &= \sum_{k=0}^{m-1}T^k + \sum_{k=0}^{m-1}T^{k+1}H \\
mT^* + H &= \sum_{k=0}^{m-1}T^k + T^mH
\end{aligned}$$

from which the result follows as $m \neq 0$. □

This establishes bounds on the convergence of $\frac{1}{m}\sum_{k=0}^{m-1}T^k$ to T^* that we will need. As H is bounded, simply taking $m \to \infty$ yields the following:

4.4 Corollary. *For a stochastic matrix* $T \in \mathbb{R}^{n\times n}$

$$T^* = \lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} T^k.$$

By applying this result to Equation (3) we can now express the expected long run average value very simply in terms of T^*,

$$\bar{\mathbf{V}}_{1\infty}^{\pi\mu} := \left(\lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} T^k \right) \mathbf{r}_\pi = T^* \mathbf{r}_\pi. \tag{6}$$

Thus by the existence of T^* we can infer that the expected long run average value also exists in this case.

4.5 Corollary. *Let* $(\mathcal{Y}, \mathcal{X}, \mu)$ *be a stationary MDP environment. For any agent* $(\mathcal{X}, \mathcal{Y}, \pi)$ *with a stationary policy* π, $\forall m \in \mathbb{N}$

$$\left| \bar{V}_{1\infty}^{\pi\mu} - \bar{V}_{1m}^{\pi\mu} \right| = O\left(\frac{1}{m}\right).$$

Proof. Let $T \in \mathbb{R}^{n\times n}$ represent the Markov chain formed by the interaction of μ and π. From Theorem 4.3 we see that $\forall m \in \mathbb{N}$,

$$T^* - \frac{1}{m} \sum_{k=0}^{m-1} T^k = \frac{1}{m}(T^m - I)H.$$

Multiplying by $\mathbf{r}_\pi$ on the right gives

$$T^* \mathbf{r}_\pi - \left(\frac{1}{m} \sum_{k=0}^{m-1} T^k \right) \mathbf{r}_\pi = \frac{1}{m} \left(T^m - I\right) H \mathbf{r}_\pi,$$

and thus the result follows as the elements of both T^m and H are bounded. □

Of course this result is not surprising as we would expect the expected average value to converge to its limit in a reasonable way when both the environment and policy are stationary.

Finally let us note some technical results on the relationship between T and T^*.

4.6 Lemma. *For an ergodic stochastic matrix* $T \in \mathbb{R}^{n\times n}$ *the row vectors of* T^* *are all the same and define a stationary distribution under* T.

This is a standard result in the theory of ergodic Markov chains. See for example Chapter V of [3] or any book on stochastic processes for a proof.

The following result shows that the limiting matrix T^* is in some sense continuous with respect to small changes in T. This will be important because it means that if we have an estimate of T that converges in the limit then our estimate of T^* will also converge.

4.7 Theorem. *For an ergodic stochastic matrix $T \in \mathbb{R}^{n\times n}$ the matrix $T^* \in \mathbb{R}^{n\times n}$ is continuous in T in the following sense: If $\hat{T} \in \mathbb{R}^{n\times n}$ is a stochastic matrix where $\max_{ij} \left|T_{ij} - \hat{T}_{ij}\right|$ is small, then $\exists c_T > 0$ which depends on T, such that $\max_{ij} \left|T^*_{ij} - \hat{T}^*_{ij}\right| \leq c_T \max_{ij} \left|T_{ij} - \hat{T}_{ij}\right|$.*

Proof. For an ergodic square matrix $T \in \mathbb{R}^{n\times n}$ the row vectors of T^* are all the same and correspond to the stationary distribution row vector $\mathbf{t}^* \in \mathbb{R}^{1\times n}$. That is,

$$T^* = \begin{pmatrix} \mathbf{t}^* \\ \vdots \\ \mathbf{t}^* \end{pmatrix} \tag{7}$$

where $\mathbf{t}^* T = \mathbf{t}^*$ and for all distribution vectors $\mathbf{t} \in \mathbb{R}^{1\times n}$ we have $\mathbf{t}T^* = \mathbf{t}^*$. Thus T has an eigenvalue of 1 with $\mathbf{t}^*$ being the corresponding left eigenvector.

From linear algebra we know that $\forall T \in \mathbb{R}^{n\times n}$,

$$\operatorname{adj}(I - T)\,(I - T) = \det(I - T) I.$$

However as T has an eigenvalue of 1, $\det(I - T) = 0$ and thus,

$$\operatorname{adj}(I - T)\,T = \operatorname{adj}(I - T),$$

or equivalently, $\forall i \in \{1, \ldots, n\}$,

$$[\operatorname{adj}(I - T)]_{i*} T = [\operatorname{adj}(I - T)]_{i*}.$$

Because $\{\mathbf{t}^*\}$ is a basis for the eigenspace corresponding to the eigenvalue 1, $[\operatorname{adj}(I - T)]_{i*}$ must be in this eigenspace. That is, $\forall i \in \{1, \ldots, n\}$, $\exists c_i \in \mathbb{R}$:

$$[\operatorname{adj}(I - T)]_{i*} = \big(\operatorname{cof}_{1i}(I - T), \ldots, \operatorname{cof}_{ni}(I - T)\big) = c_i \mathbf{t}^*$$

where the cofactor is defined $\operatorname{cof}_{ji}(I - T) := (-1)^{j+i} \det(\operatorname{minor}_{ji}(I - T))$.

As T has an eigenvalue of 1 with geometric multiplicity 1 it follows that $I - T$ as an eigenvalue of 0 also with geometric multiplicity 1. Thus the nullity of $I - T$ is 1 and so $\operatorname{rank}(I - T) = n - 1$. While we define the rank of a matrix to be the dimension of its column or row space, it also can be defined as the size of the largest non-zero minor and the two defintions can be proven to be equivalent. As the adjoint is composed of order $n-1$ minors it immediately follows that $\operatorname{adj}(I - T) \neq 0$ and thus $\exists k$, which depends on T, such that $c_k > 0$.

As $\operatorname{minor}_{jk}(I - T)$ is an $(n-1) \times (n-1)$ sub-matrix of $(I - T)$ the determinant of this is an order $n-1$ polynomial in the elements of T. Thus, by the continuity of polynomials, $\exists c' > 0$ such that for a sufficiently small $\varepsilon > 0$ change in any element of T we will get at most a $c'\varepsilon$ change in each $\operatorname{cof}_{jk}(I - T)$. However we know that $\mathbf{t}^* = \frac{1}{c_k}[\operatorname{adj}(I - T)]_{k*}$, and so an ε change in the elements of T results in at most a $\frac{c'}{c_k}\varepsilon$ change in the elements of $\mathbf{t}^*$ and thus T^*. Define $c_T := \frac{c'}{c_k}$ to indicate that this constant depends on T and we are done. □

5 An Optimal Stationary Policy

We now turn our attention to optimal policies. While our analysis so far has only dealt with stationary policies, in general optimal policies need not be stationary. As non-stationary policies are more difficult to analyse our preference is to deal with only stationary policies if possible. In this section we prove that for the class of ergodic finite stationary MDP environments an optimal policy can indeed be chosen so that it is stationary. This will simplify our analysis in later sections. However in order to show this result we will need to briefly consider policies which are potentially non-stationary. The proofs in this section follow those of Section 4.2 in [2].

Let us assume that the policy π is deterministic but not necessarily stationary, that is, $\pi := \{\pi_1, \pi_2, \ldots\}$. Thus in the k^{th} cycle we apply π_k. Define $p_i(x) := \arg\max_{y \in \mathcal{Y}} \pi_i(y|x)$ to be the action chosen by policy π in cycle i. Clearly this is unique for deterministic π.

In order to make some of the equations that follow more manageable we need to define the following two mappings. For any function $f : \mathcal{X} \to \mathbb{R}$ and deterministic policy $\pi := \{\pi_1, \pi_2, \ldots\}$ we define the mapping B_{π_k} for any $k \in \mathbb{N}$ to be $\forall x \in \mathcal{X}$,

$$(B_{\pi_k} f)(x) := R_{x p_k(x)} + \sum_{x' \in \mathcal{X}} \mu(x'|x, p_k(x)) f(x').$$

Of interest will be the policy that simply selects the action which maximises this expression in each cycle for any given x. For this we define for any function $f : \mathcal{X} \to \mathbb{R}$ and $\forall x \in \mathcal{X}$,

$$(Bf)(x) := \max_{y \in \mathcal{Y}} \left[R_{xy} + \sum_{x' \in \mathcal{X}} \mu(x'|x, y) f(x') \right].$$

Clearly this policy is stationary as the maximising y depends only on x and is independent of which cycle the system is in. By $(B^2 f)(x)$ we mean $(B(Bf))(x)$ and similarly higher powers such as $(B^i f)(x)$ and $(B^i_{\pi_k} f)(x)$. The equation $Bf = f$ is the well known Bellman equation (see [1]).

An elementary property of the mappings B_{π_k} and B is their monotonicity in the following sense.

5.1 Lemma. *For any $f, f' : \mathcal{X} \to \mathbb{R}$ such that $\forall x \in \mathcal{X} : f(x) \leq f'(x)$ and for any possibly non-stationary deterministic policy $\pi := \{\pi_1, \pi_2, \ldots\}$, we have $\forall x \in \mathcal{X}$, $\forall i, k \in \mathbb{N}$,*

$$(B^i_{\pi_k} f)(x) \leq (B^i_{\pi_k} f')(x)$$

and

$$(B^i f)(x) \leq (B^i f')(x).$$

Proof. Clearly the cases B^1 and $B^1_{\pi_k}$ are true from their definitions. A simple induction argument establishes the general result. □

Define the column vector $\mathbf{e} := (1, \ldots, 1)^{\mathrm{t}} \in \mathbb{R}^{n \times 1}$. Using these mappings we can now prove that the optimal policy can be chosen stationary if certain conditions hold.

5.2 Theorem. *Let $(\mathcal{Y}, \mathcal{X}, \mu)$ be a finite stationary MDP environment. If $\lambda \in \mathbb{R}$ is a scalar and $\mathbf{h} \in \mathbb{R}^{n\times 1}$ a column vector such that $\forall x \in \mathcal{X}$,*

$$\lambda + [\mathbf{h}]_x = \max_{y\in\mathcal{Y}} \left[R_{xy} + \sum_{x'\in\mathcal{X}} \mu(x'|x,y)[\mathbf{h}]_{x'} \right] \tag{8}$$

or equivalently,

$$\lambda\mathbf{e} + \mathbf{h} = B\mathbf{h},$$

then

$$\lambda = \bar{V}^{*\mu}_{1\infty} := \max_{\pi} \bar{V}^{\pi\mu}_{1\infty}.$$

Furthermore, if a stationary policy π^μ attains the maximum in Equation (8) for each x then this policy is optimal, that is, $\bar{V}^{\pi^\mu\mu}_{1\infty} = \lambda$.

Proof. We have $\lambda \in \mathbb{R}$ and $\mathbf{h} \in \mathbb{R}^{n\times 1}$ such that for any (possibly non-stationary) policy $\pi = \{\pi_1, \pi_2, \ldots\}$ and cycle $m \in \mathbb{N}$ and $\forall x_m \in \mathcal{X}$,

$$\lambda + [\mathbf{h}]_{x_m} \geq R_{x_m p_m(x_m)} + \sum_{x_{m+1}\in\mathcal{X}} \mu(x_{m+1}|x_m, p_m(x_m))[\mathbf{h}]_{x_{m+1}}.$$

Furthermore, if π_m attains the maximum in Equation (8) for each $x_m \in \mathcal{X}$ then equality holds in the m^{th} cycle and $p_m(x_m)$ is optimal for this single cycle. The main idea of this proof is to extend this result so that we get a policy which is optimal across all cycles.

Using the mapping B_{π_m} we can express the above equation more compactly as,

$$\lambda\mathbf{e} + \mathbf{h} \geq B_{\pi_m}\mathbf{h}.$$

Applying now $B_{\pi_{m-1}}$ to both sides and using the monotonicity property from Lemma 5.1 we see that,

$$\lambda\mathbf{e} + B_{\pi_{m-1}}\mathbf{h} \geq B_{\pi_{m-1}}B_{\pi_m}\mathbf{h}.$$

However we also know that $\lambda\mathbf{e} + \mathbf{h} \geq B_{\pi_{m-1}}\mathbf{h}$ and so it follows that,

$$2\lambda\mathbf{e} + \mathbf{h} \geq B_{\pi_{m-1}}B_{\pi_m}\mathbf{h}.$$

Repeating this m times we get

$$m\lambda\mathbf{e} + \mathbf{h} \geq B_{\pi_1}B_{\pi_2}\cdots B_{\pi_m}\mathbf{h},$$

where equality continues to hold in the case where π_k attains the maximum in Equation (8) in each cycle $k \in \{1, \ldots, m\}$. When this is the case we see that π is optimal for the cycles 1 to m.

From the definition of B_{π_k} we see that,

$$[B_{\pi_1}B_{\pi_2}\cdots B_{\pi_m}\mathbf{h}]_{x_1} = \mathbf{E}\left\{ [\mathbf{h}]_{x_{m+1}} + \sum_{k=1}^{m} R_{x_k p_k(x_k)} \middle| x_1, \pi, \mu \right\}$$

is the total expected reward over m cycles from the initial perception x_1 to the final perception x_{m+1} under policy π and environment μ. Thus $\forall x_1 \in \mathcal{X}$,

$$m\lambda + [\mathbf{h}]_{x_1} \geq \mathbf{E}\left\{ [\mathbf{h}]_{x_{m+1}} + \sum_{k=1}^{m} R_{x_k p_k(x_k)} \middle| x_1, \pi, \mu \right\}$$

where equality holds if π_k attains the maximum in Equation (8) in each cycle.

Dividing by m, gives $\forall x_1 \in \mathcal{X}$,

$$\lambda + \frac{1}{m}[\mathbf{h}]_{x_1} \geq \frac{1}{m}\mathbf{E}\left\{ [\mathbf{h}]_{x_{m+1}} | x_1, \pi, \mu \right\} + \frac{1}{m}\mathbf{E}\left\{ \sum_{k=1}^{m} R_{x_k p_k(x_k)} \middle| x_1, \pi, \mu \right\}. \tag{9}$$

Taking $m \to \infty$ this reduces to $\forall x_1 \in \mathcal{X}$,

$$\lambda \geq \lim_{m\to\infty} \frac{1}{m}\mathbf{E}\left\{ \sum_{k=1}^{m} R_{x_k p_k(x_k)} \middle| x_1, \pi, \mu \right\},$$

or equivalently,

$$\lambda \geq \bar{V}_{1\infty}^{\pi\mu},$$

where equality holds if π_k attains the maximum in Equation (8) for each cycle. When this is the case, π is optimal and thus $\bar{V}_{1\infty}^{\pi\mu} = \max_\pi \bar{V}_{1\infty}^{\pi\mu} = \lambda$. Furthermore, we see that this optimal policy is stationary because in Equation (8) the action y only depends on the current perception x and is independent of the cycle number. We call this optimal stationary policy π^μ. □

The above result only guarantees the existence of an optimal stationary policy π^μ for a stationary MDP environment $(\mathcal{Y}, \mathcal{X}, \mu)$ in the case where there is a solution to the Bellman equation $\lambda\mathbf{e} + \mathbf{h} = B\mathbf{h}$. Fortunately for ergodic MDPs it can be shown that such a solution always exists (our definition of ergodicity implies condition (2) of Proposition 2.6 in [2] where the existence of a solution is proven). It now follows that:

5.3 Theorem. *For any ergodic finite stationary MDP environment $(\mathcal{Y}, \mathcal{X}, \mu)$ there exists an optimal stationary policy π^μ.*

This is a useful result because the interaction between a stationary MDP environment and a stationary policy is much simpler to analyse than the non-stationary case. We will refer back to this result a number of times when we need to assert the existence of an optimal stationary policy.

One thing that we have not shown is that the optimal policy with respect to a given MDP can be computed. Given that our MDP is finite and therefore the number of possible stationary deterministic policies is also finite we might expect that this problem should be solvable. Indeed it can be shown that the Policy Iteration algorithm is able to compute an optimal stationary policy in this situation (see Section 4.3 of [2]).

6 Convergence of Expected Average Value

Our goal is to find a good policy for an unknown stationary MDP $(\mathcal{X}, \mathcal{Y}, \mu)$. Because we don't know the structure of the MDP, that is μ, we create an estimate $\hat{\mu}$ and then find the optimal policy with respect to this estimate, which we will call $\pi^{\hat{\mu}}$. Our hope is that if our estimate $\hat{\mu}$ is sufficiently close to μ, then $\pi^{\hat{\mu}}$ will perform well compared to the true optimal policy π^{μ}. Specifically we would like $\left|\bar{V}_{1\infty}^{\pi^{\hat{\mu}}\mu} - \bar{V}_{1\infty}^{\pi^{\mu}\mu}\right| = 0$.

In the analysis that follows we will need to be careful about whether we are talking about the true environment μ, our estimate of this $\hat{\mu}$, or various combinations of environments interacting with various policies. Sometimes policies will be optimal with respect to the environment that they are interacting with, sometimes they will only be optimal with respect to an estimate of the environment that they are actually interacting with, and in some cases the policy may be arbitrary. Needless to say that care is required to avoid mixing things up.

As defined previously, let $D \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}}$ represent the chronological system μ and $\hat{D} \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}\times\mathcal{X}}$ the chronological system $\hat{\mu}$. From Equation (1) we know that the matrix $T \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$ representing the Markov chain formed by a policy π interacting with μ is defined by,

$$T_{xx'} := \sum_{y\in\mathcal{Y}} \pi(y|x) D_{xyx'}.$$

We can similarly define $\hat{T}$ from π and $\hat{D}$. It now follows that if $\hat{D}$ is close to D, in the sense that $\varepsilon := \max_{xyx'} |D_{xyx'} - \hat{D}_{xyx'}|$ is small, then for *any* stationary policy π the associated matrices T and $\hat{T}$ are close:

$$\begin{aligned}\max_{xx'} \left|T_{xx'} - \hat{T}_{xx'}\right| &= \max_{xx'} \left|\sum_{y\in\mathcal{Y}} \pi(y|x)\big(D_{xyx'} - \hat{D}_{xyx'}\big)\right| \\ &\leq \max_{x} \left|\sum_{y\in\mathcal{Y}} \pi(y|x)\varepsilon\right| = \varepsilon.\end{aligned}$$

This is important as it means that we can take bounds on the accuracy of our estimate of the true MDP and imply from this bounds on the accuracy of the estimate $\hat{T}$ for any stationary policy.

For any given stationary policy this bound also carries over to the associated expected long run average value functions in a straightforward way:

6.1 Lemma. *For stationary finite MDPs such that $\varepsilon := \max_{xyx'} |D_{xyx'} - \hat{D}_{xyx'}|$ it follows that for any stationary policy π,*

$$\left|\bar{V}_{1\infty}^{\pi\mu} - \bar{V}_{1\infty}^{\pi\hat{\mu}}\right| = O(\varepsilon).$$

Proof. Let T and $\hat{T}$ be the Markov chains defined by D and $\hat{D}$ interacting with a stationary policy π. By the argument above we see that $\max_{xx'} |T_{xx'} - \hat{T}_{xx'}| \leq \varepsilon$. Thus by Theorem 4.7 we know that there exists c_T such that $\max_{xx'} |T^*_{xx'} - \hat{T}^*_{xx'}| \leq c_T\varepsilon$, where c_T depends on T. By Equation (6) we see that,

$$\left|\bar{V}_{1\infty}^{\pi\mu} - \bar{V}_{1\infty}^{\pi\hat{\mu}}\right| = \left|(T^* - \hat{T}^*)\mathbf{r}_\pi\right| = O(\varepsilon). \tag{10}$$

□

From this lemma we can show that the optimal policies with respect to μ and $\hat{\mu}$ are bounded:

6.2 Theorem. *For a stationary finite MDP such that* $\varepsilon := \max_{xyx'} |D_{xyx'} - \hat{D}_{xyx'}|$ *it follows that,*

$$\left|\bar{V}_{1\infty}^{\pi^\mu \mu} - \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \hat{\mu}}\right| = O(\varepsilon),$$

where π^μ *and* $\pi^{\hat{\mu}}$ *are optimal policies that are also stationary.*

Proof. For any two functions $f, f' : \mathcal{D} \to \mathbb{R}$ such that $\forall x \in \mathcal{D} : |f(x) - f'(x)| \leq \delta$ it follows that $|\max_{x\in\mathcal{D}} f(x) - \max_{x'\in\mathcal{D}} f'(x')| \leq \delta$. From Lemma 6.1 it then follows that,

$$\left|\max_{\pi} \bar{V}_{1\infty}^{\pi\mu} - \max_{\pi'} \bar{V}_{1\infty}^{\pi'\hat{\mu}}\right| = O(\varepsilon)$$

where π and π' belong to the set of stationary policies. However by Theorem 5.3 we know that the optimal policies for μ and $\hat{\mu}$ can be chosen stationary and thus the result follows. □

We now have all the necessary results to show that if $\hat{\mu}$ is a good estimate of μ then our policy $\pi^{\hat{\mu}}$ that is based on $\hat{\mu}$ will perform near optimally with respect to the true environment in the limit.

6.3 Theorem. *Let* $(\mathcal{Y}, \mathcal{X}, \mu)$ *and* $(\mathcal{Y}, \mathcal{X}, \hat{\mu})$ *be two ergodic stationary finite MDP environments that are close in the sense that* $\varepsilon := \max_{xyx'} |D_{xyx'} - \hat{D}_{xyx'}|$ *is small. It can be shown that for* $m \in \mathbb{N}$*,*

$$\left|\bar{V}_{1m}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1m}^{\pi^\mu \mu}\right| = O\left(\frac{1}{m}\right) + O(\varepsilon)$$

where π^μ *is an optimal policy for the true distribution* μ*, and* $\pi^{\hat{\mu}}$ *is an optimal policy with respect to the estimate of the true distribution* $\hat{\mu}$*.*

Proof.

From Theorem 5.3 we see that $\pi^{\hat{\mu}}$ can be chosen stationary. From the triangle inequality and the results of Corollary 4.5 (with $\pi \rightsquigarrow \pi^{\hat{\mu}}$), Lemma 6.1 (with $\pi \rightsquigarrow \pi^{\hat{\mu}}$) and Theorem 6.2 we see that,

$$\begin{aligned}
\left|\bar{V}_{1m}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1\infty}^{\pi^\mu \mu}\right| &= \left|\bar{V}_{1m}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \mu} + \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \hat{\mu}} + \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \hat{\mu}} - \bar{V}_{1\infty}^{\pi^\mu \mu}\right| \\
&\leq \left|\bar{V}_{1m}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \mu}\right| + \left|\bar{V}_{1\infty}^{\pi^{\hat{\mu}} \mu} - \bar{V}_{1\infty}^{\pi^{\hat{\mu}} \hat{\mu}}\right| + \left|\bar{V}_{1\infty}^{\pi^{\hat{\mu}} \hat{\mu}} - \bar{V}_{1\infty}^{\pi^\mu \mu}\right| \\
&= O\left(\frac{1}{m}\right) + O(\varepsilon).
\end{aligned}$$

Now from Corollary 4.5 (with $\pi \rightsquigarrow \pi^\mu$) and the triangle inequality again,

$$\begin{aligned}\left|\bar{V}_{1m}^{\pi^{\hat{\mu}}\mu} - \bar{V}_{1m}^{\pi^\mu\mu}\right| &= \left|\bar{V}_{1m}^{\pi^{\hat{\mu}}\mu} - \bar{V}_{1\infty}^{\pi^\mu\mu} + \bar{V}_{1\infty}^{\pi^\mu\mu} - \bar{V}_{1m}^{\pi^\mu\mu}\right| \\ &\leq \left|\bar{V}_{1m}^{\pi^{\hat{\mu}}\mu} - \bar{V}_{1\infty}^{\pi^\mu\mu}\right| + \left|\bar{V}_{1\infty}^{\pi^\mu\mu} - \bar{V}_{1m}^{\pi^\mu\mu}\right| = O\left(\frac{1}{m}\right) + O(\varepsilon).\end{aligned}$$

□

7 Ergodic MDPs Admit Self-Optimising Policies

The final step now is to exhibit a policy which starts without any knowledge of the structure of the MDP but is able to converge to having an optimal average reward per cycle. Firstly we will look at how an agent can build an estimate of its environment based on experience and the way in which this estimate converges to the true environment.

For a given a perception x and action y we can think of the system transitioning to a perception x' as being a Bernoulli trial as the system either transistions to x' or it does not. Thus we can view the environment as being a collection of $|\mathcal{X}| \times |\mathcal{Y}| \times |\mathcal{X}|$ Bernoulli distributions: one for each element of D. These distributions can be estimated by simple frequency estimates as follows.

Let $n^k_{xyx'}$ be the number of times the system has transistioned from x to x' via action y after the first k cycles and define $n^k_{xy} := \sum_{\bar{x}\in\mathcal{X}} n^k_{xy\bar{x}}$ to be the total number of times action y has been taken after perception x. From this we can define an estimate $\hat{D}^k$ of D after the first k cycles:

$$\hat{D}^k_{xyx'} := \begin{cases} \frac{n^k_{xyx'}}{n^k_{xy}} & \text{if } n^k_{xy} > 0, \\ \frac{1}{|\mathcal{X}|} & \text{otherwise.} \end{cases}$$

From statistics we know that a frequency estimate of a Bernoulli distribution is unbiased and thus $\forall k, x, y, x'$ whenever $n^k_{xy} > 0$ we have $\mathbf{E}(\hat{D}^k_{xyx'}) = D_{xyx'}$. Furthermore $\forall k, x, y, x'$ the variance of each estimator is given by,

$$var(\hat{D}^k_{xyx'}) = \frac{D_{xyx'}(1 - D_{xyx'})}{n^k_{xy}} \leq \frac{1}{4n^k_{xy}}.$$

This if $n^k_{xy} \to \infty$ we have $var(\hat{D}^k_{xyx'}) \to 0$ and so $\hat{D}^k_{xyx'} \to \mathbf{E}(\hat{D}^k_{xyx'}) = D_{xyx'}$ with probability 1. By the Chebychev inequality it now follows that $\forall k, x, y, x'$ and $\forall h \in \mathbb{N}$,

$$\Pr\left(\left|\hat{D}_{xyx'} - D_{xyx'}\right| \leq \frac{h}{4n^k_{xy}}\right) \geq 1 - \frac{1}{h^2}.$$

Thus we can bound the elements of $\hat{D}$ to be within any required distance of their corresponding elements in D with arbitrarily high probability by taking n^k_{xy} sufficiently large.

Consider now a policy π^r that explores the environment by performing uniformly random actions and building an estimate $\hat{D}$ of the environment as described above. Formally $\forall y \in$

$\mathcal{Y} : \pi^r(y) = 1/|\mathcal{Y}|$. Let $d_{min} := \min\{D_{xyx'} \neq 0 : x, x' \in \mathcal{X}, y \in \mathcal{Y}\}$ be the minimum non-zero transition probability for a single cycle. Because the environment is ergodic there must exist a policy $\tilde{\pi}$ such that for every pair of perceptions $x, x' \in \mathcal{X}$ there exists a sequence of cycles of length at most $|\mathcal{X}| - 1$ such that perception x' can be reached starting from perception x with non-zero probability. However under the uniformly random policy π^r the probability of selecting the same action as $\tilde{\pi}$ after a given perception is always $|\mathcal{Y}|^{-1}$ and the probability of the required transition then occuring is at least d_{min}. Thus the probability of reaching a specific x' starting from any x in $|\mathcal{X}| - 1$ cycles or less under policy π^r is at least $(d_{min}/|\mathcal{Y}|)^{|\mathcal{X}|-1}$.

After perception x' is reached, action y' is taken with probability $|\mathcal{Y}|^{-1}$. Thus the probability of action y' following perception x' during the first $|\mathcal{X}|$ cycles under policy π^r can be bounded: $\forall x', y'$

$$\Pr\left(n_{x'y'}^{|\mathcal{X}|} \geq 1\right) \geq \frac{1}{|\mathcal{Y}|}\left(\frac{d_{min}}{|\mathcal{Y}|}\right)^{|\mathcal{X}|-1} > 0.$$

As $k \to \infty$ the number of contiguous blocks of $|\mathcal{X}|$ cycles goes to infinity and so $n_{x'y'}^k \to \infty$ with probability 1. It follows then that if we estimate D by frequency estimate $\hat{D}$, as described above, then for a finite ergodic stationary MDP environment and policy π^r we get $\hat{D}^k \to D$ as $k \to \infty$ with probability 1.

While this shows that it is possible for a policy to build an accurate estimate of the environment with probability 1, the average reward per cycle under policy π^r is normally far from being optimal. What we require is a policy that is able to estimate D well and utilise this estimate at the same time in such a way that the optimal average expected reward per cycle is achieved.

7.1 Theorem. *There exists a policy π which is self-optimising with respect to all finite stationary ergodic MDP environments $(\mathcal{Y}, \mathcal{X}, \mu)$ in the sense that,*

$$\left|\bar{V}_{1k}^{\pi\mu} - \bar{V}_{1k}^{\pi^\mu\mu}\right| \to 0,$$

with probability 1 as $k \to \infty$.

Before detailing the proof let us first outline the general idea. We know from Theorem 6.3 that if a policy has a good estimate $\hat{\mu}$ of μ, or equivalently $\hat{D}$ of D, then in the limit it can achieve close to optimal expected reward per cycle by acting optimally with respect to this estimate of the environment. Because the policy isn't provided with this estimate it must somehow create it based on its experience.

One solution might be for the policy to perform random actions for some period of time, as π^r does, until its estimated model of the environment appears to be sufficiently accurate and then switch to acting optimally with respect to this estimate. This would allow us to claim that with arbitrarily high probability the ε term in Theorem 6.3 is below some value and thus with arbitrarily high probability the expected reward per cycle after the initial random exploratory phase is within some ε term of being optimal. This is essentially what is done in Theorem 5.38 of [4].

The problem is that this only demonstrates that there exists a sequence of policies of increasing expected reward per cycle. In particular the limit of this sequence has expected reward per cycle that is optimal. While this is sufficient for some purposes it doesn't demonstrate the existence of any single policy that is self-optimising. Here we would like to establish this stronger result.

The difficulty is that we require a policy where the estimate of the environment converges (with probability 1) to the true environment and thus the ε term in Theorem 6.3 goes to zero with high probability. However in order for this to happen the policy must keep on exploring its environment and thus we cannot cleanly seperate the exploration and expliotation into two phases. Put another way: If a policy ever stopped exploring it's possible that its model never sufficently converges to the true model and thus its expected reward per cycle is close to optimal but it never actually converges.

This complicates things somewhat because when a policy is exploring its environment it is not usually following actions that are optimal with respect to some estimate of the environment and so the optimality bound in Theorem 6.3 cannot be applied to these cycles. Furthermore these non-optimal exploritary actions reduce the expected reward per cycle making the desired convergence more difficult. Thus we need to find a way to keep the policy exploring the environment so that the ε term in Theroem 6.3 goes to 0 with probability 1 and yet the cost of this infinite amount of exploration is spread out in such a way that in the limit it becomes insignificant.

We also need to be careful with the other free variable in Theorem 6.3, the m in this $O(\frac{1}{m})$ term. In order for this to go to zero we require that the system spends increasingly long periods of time exploiting estimates of the optimal policy. We need to ensure that these periods of increasingly large m and small ε dominate the average long run expected reward per cycle.

Our solution is to set up a policy that forever alternates between increasingly long periods of exploration and expliotation taking care to make sure that the ratio goes to zero. Essentially the formal proof that follows is just a demonstation that such a policy does indeed exist.

Proof. We break the cycles of the system up into contiguous blocks that we will call "episodes". Over time the episodes increase in length with the e^{th} episode being e^2 cycles long. Thus cycle 1 belongs to episode 1, cycles 2, 3, 4 and 5 belong to episode 2 and so on. It can be shown that $k_e := 1^2 + 2^2 + \cdots + (e-1)^2 = \frac{1}{6}(2e^3 - 3e^2 + e)$ is the number of cycles that have occurred prior to the e^{th} episode.

Now define a non-stationary policy π as follows: During the e^{th} episode π explores the environment by taking uniformly random actions for the first e cycles of the episode, that is it behaves like π^r analysed previously. When this has finished the system has just completed cycle $k_e + e$. The policy π then computes the optimal stationary policy, which we call π^e, based on its current estimate of D, that is $\hat{D}^{k_e+e}$, and then utilises π^e for the remaining $e^2 - e$ cycles in this episode. In this way π continually alternates between exploring and exploiting with the length of the explore and exploit periods increasing without bound over time. This is really all there is to our policy π, the task now is to show that π is in fact self-optimsing.

For notational purposes we will refer to the explore phases as "search" and the exploit phases as "utilise". Because the policy π spends e cycles in the e^{th} episode searching we see

that $s_e := 1 + 2 + \cdots + (e-1) = \frac{1}{2}(e^2 - e)$ is the total number of search cycles have occurred prior to the e^{th} episode. It now follows that $u_e := k_e - s_e = \frac{1}{6}(e^3 - 6e^2 + 4e)$ is the total number of utilise cycles have occurred prior to the e^{th} episode.

Consider now $\bar{V}_{1\infty}^{\pi\mu}$, the long run average expected reward per cycle under the policy π. Our task is to show that this is equal to the optimal long run average expected reward per cycle. Firstly we bound this quantify for all cycles k in terms of episodes by noting that the average expected reward per cycle during any episode is lower bounded by the average expected reward per cycle at the start of the episode adjusted as if all rewards in the current episode were 0. If we break the long run average expected reward into episodes we can achieve the same effect by simply extending the denominator that weights the average reward for each episode one episode into the future, thus we get,

$$\bar{V}_{1\infty}^{\pi\mu} = \lim_{k\to\infty} \bar{V}_{1k}^{\pi\mu} \geq \lim_{l\to\infty} \sum_{e=1}^{l} \frac{e^2}{k_{l+2}} \bar{V}_{k_e+1,k_{e+1}}^{\pi\mu}$$

remembering that k_{l+1} is the total number of cycles that have occured before the $(l+1)^{th}$ episode and thus k_{l+2} is the total number of cycles in $l+1$ episodes.

If we now separate out the search and utilise periods within each episode we see that,

$$\begin{aligned} \bar{V}_{1\infty}^{\pi\mu} &= \lim_{l\to\infty} \sum_{e=1}^{l} \frac{e}{k_{l+2}} \bar{V}_{k_e+1,k_e+e}^{\pi^r\mu} + \lim_{l\to\infty} \sum_{e=1}^{l} \frac{e^2-e}{k_{l+2}} \bar{V}_{k_e+e+1,k_{e+2}}^{\pi^e\mu} \\ &\geq \lim_{l\to\infty} \sum_{e=1}^{l} \frac{e^2-e}{k_{l+2}} \bar{V}_{k_e+e+1,k_{e+2}}^{\pi^e\mu}. \end{aligned}$$

The inequality follows as the reward per cycle is non-negative and bounded. Because $\bar{V}_{1\infty}^{\pi\mu}$ by definition cannot be better than optimal it's sufficient to bound from below to show convergence.

Now consider the terms $\bar{V}_{k_e+e+1,k_{e+1}}^{\pi^e\mu}$ in the above sum. As each π^e is a stationary policy that is optimal with respect to an estimate of the environment we can apply Theorem 6.3 to these cycles,

$$\left| \bar{V}_{k_e+e+1,k_{e+1}}^{\pi^e\mu} - \bar{V}_{1\infty}^{\pi^\mu\mu} \right| = O\left(\frac{1}{m_e}\right) + O(\varepsilon_e),$$

where $m_e := e^2 - e$ is the length of the utilise phase in the e^{th} episode and $\varepsilon_e := \max_{xyx'} \left| \hat{D}_{xyx'}^k - D_{xyx'} \right|$ is the bound on the accuracy of the estimate of the environment at the end of the search phase of the e^{th} episode. Thus it follows that $\exists c_1, c_2 > 0$ such that

$$\begin{aligned} \bar{V}_{1\infty}^{\pi\mu} &\geq \lim_{l\to\infty} \sum_{e=1}^{l} \frac{e^2-e}{k_{l+2}} \left(\bar{V}_{1\infty}^{\pi^\mu\mu} - c_1 \frac{1}{e^2-e} - c_2\varepsilon_e \right) \\ &= \bar{V}_{1\infty}^{\pi^\mu\mu} \lim_{l\to\infty} \sum_{e=1}^{l} \frac{(e^2-e)}{k_{l+2}} - c_1 \lim_{l\to\infty} \frac{l}{k_{l+2}} - c_2 \lim_{l\to\infty} \sum_{e=1}^{l} \frac{(e^2-e)}{k_{l+2}} \varepsilon_e. \end{aligned}$$

However $k_{l+2} = \frac{1}{6}\big(2l^2 - 3l + l\big) + l^2 + (l+1)^2$ and so in the last line above the second limit is 0 and for the first limit we get,

$$\lim_{l\to\infty} \frac{\sum_{e=1}^{l}(e^2 - e)}{k_{l+2}} = \lim_{l\to\infty} \frac{\frac{1}{6}(2l^3 + 3l^2 + l) - \frac{1}{2}(l^2 - l)}{\frac{1}{6}(2l^3 - 3l^2 + l) + l^2 + (l+1)^2} = 1.$$

Thus,

$$\bar{V}_{1\infty}^{\pi\mu} \geq \bar{V}_{1\infty}^{\pi^\mu\mu} - c_2 \lim_{l\to\infty} \frac{1}{k_{l+2}} \sum_{e=1}^{l} (e^2 - e)\varepsilon_e.$$

We know that $\forall e : |\varepsilon_e| \leq 1$ and so all the terms in the sum are bounded. Therefore for the limiting average we can ignore any finite number of terms, that is, $\forall n \in \mathbb{N}$

$$\lim_{l\to\infty} \frac{1}{k_{l+2}} \sum_{e=1}^{l} (e^2 - e)\varepsilon_e = \lim_{l\to\infty} \frac{1}{k_{l+2}} \sum_{e=n}^{l} (e^2 - e)\varepsilon_e.$$

We already know that by choosing n sufficiently large we can ensure that $\forall e \geq n : \varepsilon_e < \delta$ for any $\delta > 0$ with probability 1. Thus for any $\delta > 0$, $\exists n$ such that

$$c_2 \lim_{l\to\infty} \frac{1}{k_{l+2}} \sum_{e=1}^{l} (e^2 - e)\varepsilon_e \leq c_2\delta \lim_{l\to\infty} \sum_{e=n}^{k} \frac{(e^2 - e)}{k_{l+2}} = c_2\delta$$

with arbitrarly high probability. Thus with probability 1 this limit goes to zero and so with probability 1 we have $\bar{V}_{1\infty}^{\pi\mu} \geq \bar{V}_{1\infty}^{\pi^\mu\mu}$ from which the result follows as π cannot be better than π^μ by definition. □

References

[1] R. Bellman. *Dynamic Programming.* Princeton University Press, New Jersey, 1957.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. (II).* Athena Scientific, Belmont, Massachusetts, 1995.

[3] J. L. Doob. *Stochastic Processes.* John Wiley & Sons, New York, 1953.

[4] M. Hutter. Optimal sequential decisions based on algorithmic probability, 2003. 288 pages, draft, http://www.idsia.ch/∼marcus/ai/habil.htm.

[5] Ulrich Krengel. *Ergodic Theorems.* Walter de Gruyter, 1985.

[6] S. Legg and M. Hutter. A taxonomy for abstract environments. Technical report, IDSIA, 2004.

[7] R. Sutton and A. Barto. *Reinforcement learning: An introduction.* Cambridge, MA, MIT Press, 1998.