# An introduction to the mathematics of reinforcement learning theory

Sebastian Scherer

January 15, 2019

## 1 The agent and the environment

In reinforcement learning we concern ourselves with optimising the behaviour of an agent acting in a given environment. The agent's behaviour is governed by what we call control laws which act on the environment's current state, and can be probabilistic in nature. The environment responds to the agent's actions by assuming a new state. The process of obtaining this new state can also be probabilistic.

There are some constrains to this very general setting, which we will outline in this section.

Firstly, we assume that at any time $t \in \mathbb{N}_0$, the environment can only assume one of finite states $s \in S$, where $S$ is the finite set of states possible.

Similarly, we demand that our agent has only a finite set of actions $a \in A$, where $A$ is the finite set of actions, at his disposal at any given time $t \in \mathbb{N}_0$.

For an arbitrary but fixed starting state $s_0 \in S$, the continuous back-and-forth between the agent choosing an action $a_t$ and the environment assuming a subsequent state $s_{t+1}$ for $t = 0, \dots, i$ leads to *state-action trajectories* of the form

$$(s_0, a_0, s_1, a_1, \cdots, s_{i-1}, a_{i-1}, s_i, a_i). \tag{1}$$

We will also sometimes refer to *state trajectories*

$$(s_0, s_1, \cdots, s_{i-1}, s_i, ) \tag{2}$$

and *action trajectories*

$$(a_0, a_1, \cdots, a_{i-1}, a_i, ) \tag{3}$$

as needed. For any arbtrarily given but fixed end point in time $i \in \mathbb{N}_0$, we can imbue all of these three trajectory spaces with probability distributions depending on the control laws governing the actions (where plausible), and the probabilistic behaviour of the environment. We will do this in the next section.

Lastly, we require our environment has no memory when evolving from one state to the next, be it in response to our agent's chosen action or otherwise. We demand that the environment's state at time $i+1$, $s_{i+1} \in S$, only depends on the previous time step's state, $s_i \in S$, and the agent's chosen action $a_i \in A$ at time $i$, but *not* on any other preceding states and actions $s_t, a_t, t < i$ forming the state-action trajectory leading up to the state $s_i$ and action $a_i$ at time $i$. To be more precise, we require the *transitional probabilities* of our environment to satisfy

$$Pr(s_{i+1} = s'|(s_0, a_0, \dots, s_i, a_i)) = Pr(s_{i+1} = s'|(s_i, a_i)) =: P_{s_i}^{a_i}(s') \tag{4}$$

for all $s' \in S$ and $i \in \mathbb{N}_0$. This property is often referred to as the *Markov* property.

## 2 Probabilistic control laws

In this section we will formalize our understanding of a *control law*, which can be regarded as the decision making process of our agent at a fixed time $i\mathbb{N}_0$. A control law $\mu$ is a set of probability distributions over the action space $A$, one conditional distribution $\mu(s, \cdot)$ for each possible state $s \in S$. The idea is that, using the control law $\mu$ at time $i$ to make our agent's decision $a_i$, for any possible environment state $s$ $\mu$ generates a probability distribution over the action space $A$, assigning a probability

$$\begin{aligned} & Pr(\text{Choosing action } a_i|\text{The environment is in state } s_i \text{ while following control law } \mu) \\ = \quad & Pr(\text{Choosing action } a_i|s_i, \mu) \\ =: \quad & \mu(s_i, a_i). \end{aligned} \tag{5}$$

For completeness, we note that for any such control law $\mu$ clearly

$$\mu(s, a) \geq 0 \tag{6}$$

for every state action pair $(s, a) \in S \times A$, as well as

$$\sum_{a \in A} \mu(s, a) = 1 \tag{7}$$

for all $s \in S$, must hold.

It is worth noting that by the above interpretation we are only allowing control laws and distributions conditioned on *only the immediate state* $s_i$, and nothing else. In this sense, the control laws considered have no memory of past environmental or agent behaviour either.

Before we conclude this section, let us develop a slightly more abstract but, as we shall see later, highly useful perspective on the set of control laws just outlined. We first order the finite state and action sets arbitrarily: $s^1, \dots, s^{|S|}$ and $a^1, \dots, a^{|A|}$. Since any control law $\mu$ is a collection of $|S|$ discrete probability distributions over $A$, we can identify $\mu$ with an element from $\mathbb{R}^{|S| \times |A|}$ via the canonical representation

$$\mu = \begin{pmatrix} \mu(s^1, a^1) & \mu(s^1, a^2) & \cdots & \mu(s^1, a^{|A|}) \\ \mu(s^2, a^1) & \mu(s^2, a^2) & \cdots & \mu(s^2, a^{|A|}) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(s^{|S|}, a^1) & \mu(s^{|S|}, a^2) & \cdots & \mu(s^{|S|}, a^{|A|}) \end{pmatrix}. \tag{8}$$

Here, the $i$-th row of the right hand side encodes the conditional probability distribution $\mu(s^i, \cdot)$ conditioned on state $s^i \in S$. We can thus see that the set of control laws can be identified with a closed and bounded, and therefore *compact*, subset of the $\mathbb{R}^{|S|\times|A|}$ with its canonical norm via

$$\Big\{\mu | \mu \text{ is a control law}\Big\} = \Big\{\mu \in \mathbb{R}^{|S|\times|A|} \Big| \mu_{ij} \geq 0\, \forall i, j, \sum_{j=1}^{|A|} \mu_{ij} = 1\, \forall i = 1, \ldots, |S|\Big\}. \tag{9}$$

The identification of the set of control laws with a compact set will be crucial in maximization arguments further down the line. Before that, however, let us next see how we can use these control laws to formalize the behaviour of our agent.

## 3 Control law sequences and state-action trajectories

In the previous section we have formalized the nature of our agent's decision making process at any given time $i$: Given that the environment is in state $s_i$ and our agent follows the control law $\mu$, it will pick any action $a \in A$ with probability $\mu(s_i, a)$. There is no reason for us to constrain our agent to keep using the same control law $\mu$ over time. It is much more desirable for our agent to be able to follow a sequence of different control laws, say,

$$\vec{\mu}(i) = (\mu_0, \mu_1, \ldots, \mu_i), \tag{10}$$

where $\mu_t$ is the control law employed at time $t = 0, \ldots i$ by our agent to pick action $a_i$. As referred to earlier, an environment with known transition probabilities $P^a_{ss'}, a \in A, s, s' \in S$ together with a sequence $\vec{\mu}(i)$ of control laws of length $i$ induces probability distributions on the sets of state-action, state and action trajectories. The environment's Markov property and the control laws' lack of memory allow for a nice factorization of these probabilities. The following Lemma makes this claim more precise.

**Lemma 1.** *(State-action trajectory distribution under policies)*

*For some $i \in \mathbb{N}_0$, let $\vec{\mu}(i)$ be a finite series of probabilistic control laws. Then for any fixed starting state $s_0 \in S$ and state-action trajectory $(s_0, a_0, \ldots, s_i, a_i)$, the probability of obtaining said state-action trajectory up to time i while following $\vec{\mu}(i)$ is given by*

$$\prod_{t=0}^{i-1} \big(\mu_t(s_t, a_t) \cdot P^{a_t}_{s_t}(s_{t+1})\big) \cdot \mu_i(s_i, a_i). \tag{11}$$

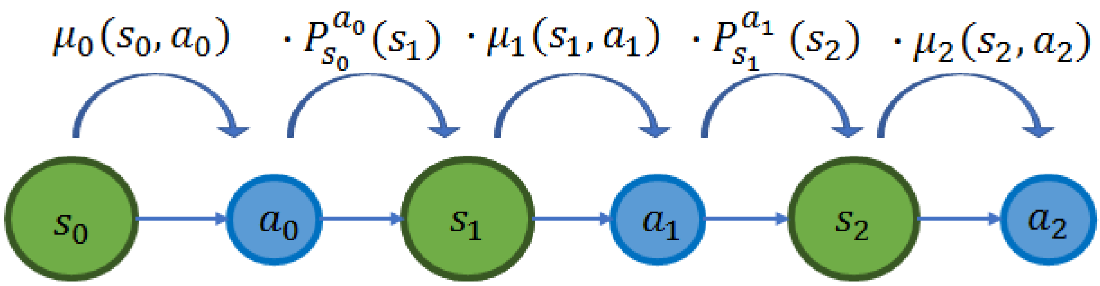


Figure 1: A trajectory starting from $s_0$ while following $\mu_0, \ldots, \mu_i$ for $i = 2$

*Proof.* We prove this claim via induction over the control law sequence length parameter $i$. Since $i = 0$ is somewhat trivial, we start our induction with $i = 1$. We see that

$$\begin{aligned} & \Pr\{(s_0, a_0, s_1, a_1)| \text{ starting at } s_0 \text{ and following } \vec{\mu}(1) = (\mu_0, \mu_1)\} \\ = \ & \Pr\{(s_0, a_0, s_1, a_1)|s_0,\ \vec{\mu}(1)\} = \Pr\{(s_0, a_0) \cap (s_1, a_1)|s_0,\ \vec{\mu}(1)\} \\ = \ & \Pr\{(s_0, a_0, s_1)|s_0,\ \vec{\mu}(1)\} \cdot \Pr\{a_1|(s_0, a_0, s_1),\ s_0,\ \vec{\mu}(1)\} \\ = \ & \Pr\{(s_0, a_0, s_1)|s_0,\ \mu_0\} \cdot \Pr\{a_1|s_1,\ \mu_1)\} \\ = \ & \Pr\{(s_0, a_0)|s_0,\ \mu_0\} \cdot \Pr\{s_1|(s_0, a_0),\ \mu_0\} \cdot \Pr\{a_1|s_1,\ \mu_1)\} \\ = \ & \mu_0(s_0, a_0) \cdot\ P_{s_0}^{a_0}(s_1) \cdot \mu_1(s_1, a_1). \end{aligned} \tag{12}$$

Now assume this claim holds for some $i - 1 \in \mathbb{N}_0$. The exact same argument applied above then yields

$$\begin{aligned} & \Pr\{(s_0, a_0, \ldots, s_i, a_i)| \text{ starting at } s_0 \text{ and following } \vec{\mu}(i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_i, a_i)|s_0,\ \vec{\mu}(i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1}) \cap (s_i, a_i)|s_0,\ \vec{\mu}(i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1}, s_i)|s_0,\ \vec{\mu}(i)\} \cdot \Pr\{a_i|(s_0, a_0, \ldots, s_{i-1}, a_{i-1}, s_i),\ s_0,\ \vec{\mu}(i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1}, s_i)|s_0,\ \vec{\mu}(i-1)\} \cdot \Pr\{a_i|s_i,\ \mu_i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1})|s_0,\ \vec{\mu}(i-1)\} \\ & \cdot \Pr\{s_i|(s_0, a_0, \ldots, s_{i-1}, a_{i-1}),\ s_0,\ \vec{\mu}(i-1)\} \cdot \Pr\{a_i|s_i,\ \mu_i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1})|s_0,\ \vec{\mu}(i-1)\} \\ & \cdot \Pr\{s_i|(s_{i-1}, a_{i-1})\} \cdot \Pr\{a_i|s_i,\ \mu_i)\} \\ = \ & \Pr\{(s_0, a_0, \ldots, s_{i-1}, a_{i-1})|s_0,\ \vec{\mu}(i-1)\} \cdot P_{s_{i-1}}^{a_{i-1}}(s_i) \cdot \mu_i(s_i, a_i) \\ = \ & \textstyle\prod_{t=0}^{i-2} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_{i-1}(s_{i-1,}a_{i-1}) \cdot P_{s_{i-1}}^{a_{i-1}}(s_i) \cdot \mu_i(s_i, a_i) \\ = \ & \textstyle\prod_{t=0}^{i-1} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_i(s_i, a_i). \end{aligned} \tag{13}$$

□

Similarly, without executing the last action at time $t = i$ and thus effectively only following $\vec{\mu}(i-1) = (\mu_0, \ldots, \mu_{i-1})$ we obtain the corollary result

**Corollary 1.** *(State-action trajectory distribution under policies II)*

*For some $i \in \mathbb{N}_0$, let $\vec{\mu}(i-1)$ be a finite series of probabilistic control laws. Then for any fixed starting state $s_0 \in S$ and state-action trajectory $(s_0, a_0, \ldots, s_i)$, the probability of obtaining said state-action trajectory up to time $i-1$ while following $\vec{\mu}(i-1)$ is given by*

$$\prod_{t=0}^{i-1} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right). \tag{14}$$

Given a some starting state $s_0$, what about the chances of following *any* state-action trajectory ending with some specified state-action pair $(s_i, a_i) \in S \times A$? Clearly, the answer is to simply add over all relevant state-action trajectory probabilities.

**Corollary 2.** *(State-action trajectory distribution under policies III)*
*For some $i \in \mathbb{N}_0$, let $\vec{\mu}(i)$ be a finite series of probabilistic control laws, and let $(s, a) \in S \times A$ be any fixed but arbitrary state-action pair.Let finally $s_0 \in S$ be some fixed but arbitrary starting state. Then the probability of following any of the state-action trajectories $(s_0, a_0, ..., s, a)$, $a_t \in A$ for $t = 0, ..., i-1$, $s_t \in S$ for $t = 1, ..., i-1$, while following $\vec{\mu}(i)$ is given by*

$$\sum_{\substack{a_0, \ldots, a_{i-1} \in A \\ s_1, \ldots, s_{i-1} \in S}} \left[\prod_{t=0}^{i-2} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_{i-1}(s_{i-1}, a_{i-1}) \cdot P_{s_{i-1}}^{a_{i-1}}(s)\right] \cdot \mu_i(s, a). \tag{15}$$

*Proof.* Using Lemma 1, we immediately arrive at the expression in Eq. 15 by summing over the set of relevant trajectories.

$$Tr_{s_0}^{s,a} = \left\{ (s', a_0, s_1, a_1, \ldots, s_{i-1}, a_{i-1}, s, a) \;\middle|\; \begin{array}{l} a_0, \ldots, a_{i-1} \in A, \\ s' = s_0, \\ s_1, \ldots, s_{i-1} \in S \end{array} \right\}. \tag{16}$$

□

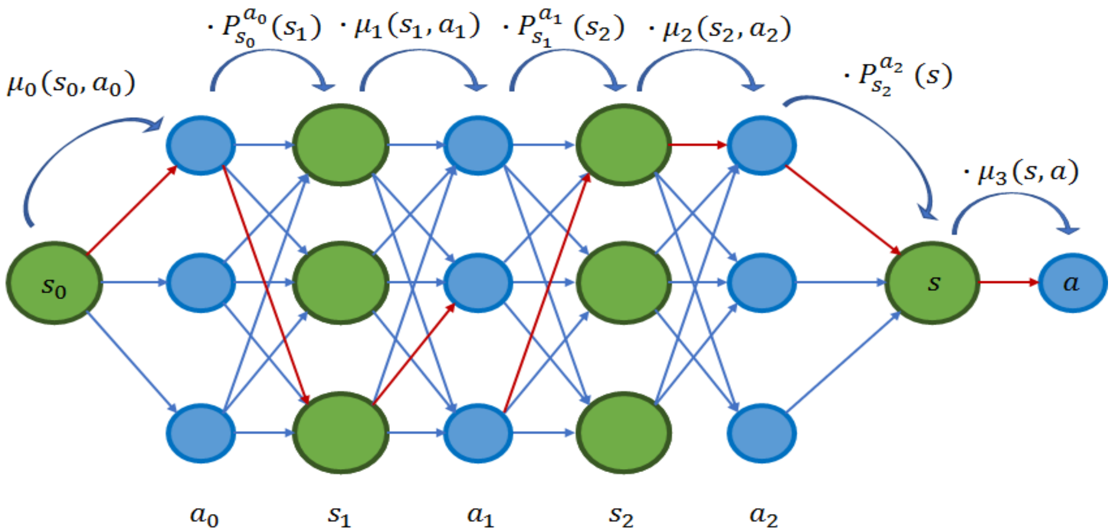


Figure 2: All the possible trajectories $(s_0, \ldots, s, a)$ starting from $s_0$ while following $\mu_0, \ldots, \mu_i$ for $i = 3$, $|S| = |A| = 3$. A sample trajectory is highlighted in magenta.