An introduction to the mathematics of reinforcement learning theory

Sebastian Scherer

February 10, 2019

1 Preliminaries

In reinforcement learning we concern ourselves with optimising the behaviour of an agent acting in a given environment to maximise some reward handed out by said 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 and issuing a reward. The process of obtaining this new state and determining the value of said reward can also be probabilistic. In summary, our problem setting will be the (finite) repetition of the following steps:

1. Determine current state s_t. For $t = 0$ this will be a given starting state. For $t \geq 1$, this will be the environment probabilistically reacting to
   (a) The state s_{t-1} in the previous time step.
   (b) The action a_{t-1} as chosen by the agent in the previous time step.
2. Determine the agent's action a_t. This is done via the control law, which only considers the current state s_t, and nothing else, and is usually probabilistic.
3. Based on s_t and a_t, a probabilistic reward is handed out by the environment. For the last time step of the finite horizon problem, the reward only depends on s_t since no further action will be taken.

1.1 The agent, the environment and the reward

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 $t = 0, \dots, i$ (we ignore the rewards for the time being) 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.

The constraints on the environment's rewards are as follows. At any given time step t, the reward r_t issued by the environment is distributed according to a distribution that only takes into account the present time step's state s_t and agent action a_t. That is, the rewards awarded are instantaneous in nature and reward the current configuration of both agent and environment, but does not take into account the past. Furthermore, we assume that it is uniformly bounded, i.e. that

$$0 < r_t < M \ \forall t \in \mathbb{N}_0 \tag{4}$$

for some $M \in \mathbb{R}$. These considerations amount to the uniform boundedness of the conditioned expectations

$$0 < \mathbb{E}[r_t | s_t = s, a_t = a] =: R(a, s)_t < M \tag{5}$$

for any $s \in S, a \in A,\ t \in \mathbb{N}_0$. We investigate these expectations more closely in the following section. For now, let it be mentioned that it is with respect to these, more precisely, trying to maximise these expected rewards, that we will try and optimise the control laws gouverning our agent's behaviour.

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

$$Pr(s_{i+1} = s' | (s_0, a_0, \dots, s_i, a_i)) = Pr(s_{i+1} = s' | (s_i, a_i)) =: P_{s_i}^{a_i}(s') \tag{6}$$

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

1.2 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 μ 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 μ at time i to make our agent's decision a_i, for any possible environment state s μ generates a probability distribution over the action space A, assigning a probability

$$\begin{aligned} & Pr(\text{Choosing action } a_i | \text{The environment is in state } s_i \text{ while following control law } \mu) \\ = \quad & Pr(\text{Choosing action } a_i | s_i, \mu) \\ =: \quad & \mu(s_i, a_i). \end{aligned} \tag{7}$$

For completeness, we note that for any such control law μ clearly

$$\mu(s, a) \geq 0 \tag{8}$$

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

$$\sum_{a \in A} \mu(s, a) = 1 \tag{9}$$

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 μ is a collection of $|S|$ discrete probability distributions over A, we can identify μ with an element from $\mathbb{R}^{|S| \times |A|}$ via the canonical representation

$$\mu = \begin{pmatrix} \mu(s^1, a^1) & \mu(s^1, a^2) & \cdots & \mu(s^1, a^{|A|}) \\ \mu(s^2, a^1) & \mu(s^2, a^2) & \cdots & \mu(s^2, a^{|A|}) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(s^{|S|}, a^1) & \mu(s^{|S|}, a^2) & \cdots & \mu(s^{|S|}, a^{|A|}) \end{pmatrix}. \tag{10}$$

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

$$\begin{aligned} \Big\{\mu|\mu \text{ is a control law}\Big\} &= \Big\{\mu \in \mathbb{R}^{|S|\times|A|}\Big|\mu_{ij} \geq 0\,\forall i, j, \sum_{j=1}^{|A|} \mu_{ij} = 1\,\forall i = 1, \ldots, |S|\Big\} \\ &=: \Pi(S, A). \end{aligned} \tag{11}$$

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.

1.3 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 μ, 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 μ over time. It is much more desirable for our agent to be able to follow a sequence of different control laws, i.e. a *policy*, say,

$$\pi(i) = (\mu_0, \mu_1, \ldots, \mu_i), \tag{12}$$

where μ_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 policy $\pi(i)$ 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 $\pi(i)$ 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 $\pi(i)$ is given by

$$\prod_{t=0}^{i-1} \left(\mu_t(s_t, a_t) \cdot P^{a_t}_{s_t}(s_{t+1})\right) \cdot \mu_i(s_i, a_i). \tag{13}$$

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

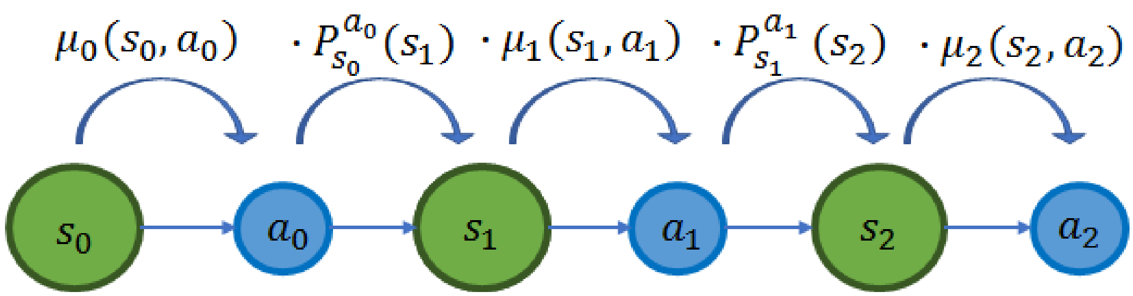

Figure 1: A trajectory starting from s_0 while following $\mu_0, \dots, \mu_i$ for $i = 2$

$$\begin{aligned}
& \Pr\{(s_0, a_0, s_1, a_1)| \text{ starting at } s_0 \text{ and following } \pi(1) = (\mu_0, \mu_1)\} \\
= \; & \Pr\{(s_0, a_0, s_1, a_1)|s_0,\, \pi(1)\} = \Pr\{(s_0, a_0) \cap (s_1, a_1)|s_0,\, \pi(1)\} \\
= \; & \Pr\{(s_0, a_0, s_1)|s_0,\, \pi(1)\} \cdot \Pr\{a_1|(s_0, a_0, s_1),\, s_0,\, \pi(1)\} \\
= \; & \Pr\{(s_0, a_0, s_1)|s_0,\, \mu_0\} \cdot \Pr\{a_1|s_1,\, \mu_1)\} \\
= \; & \Pr\{(s_0, a_0)|s_0,\, \mu_0\} \cdot \Pr\{s_1|(s_0, a_0),\, \mu_0\} \cdot \Pr\{a_1|s_1,\, \mu_1)\} \\
= \; & \mu_0(s_0, a_0) \cdot\, P_{s_0}^{a_0}(s_1) \cdot \mu_1(s_1, a_1).
\end{aligned} \tag{14}$$

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

$$\begin{aligned}
& \Pr\{(s_0, a_0, \dots, s_i, a_i)| \text{ starting at } s_0 \text{ and following } \pi(i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_i, a_i)|s_0,\, \pi(i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1}) \cap (s_i, a_i)|s_0,\, \pi(i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1}, s_i)|s_0,\, \pi(i)\} \cdot \Pr\{a_i|(s_0, a_0, \dots, s_{i-1}, a_{i-1}, s_i),\, s_0,\, \pi(i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1}, s_i)|s_0,\, \pi(i-1)\} \cdot \Pr\{a_i|s_i,\, \mu_i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1})|s_0,\, \pi(i-1)\} \\
& \cdot \Pr\{s_i|(s_0, a_0, \dots, s_{i-1}, a_{i-1}),\, s_0,\, \pi(i-1)\} \cdot \Pr\{a_i|s_i,\, \mu_i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1})|s_0,\, \pi(i-1)\} \\
& \cdot \Pr\{s_i|(s_{i-1}, a_{i-1})\} \cdot \Pr\{a_i|s_i,\, \mu_i)\} \\
= \; & \Pr\{(s_0, a_0, \dots, s_{i-1}, a_{i-1})|s_0,\, \pi(i-1)\} \cdot P_{s_{i-1}}^{a_{i-1}}(s_i) \cdot \mu_i(s_i, a_i) \\
= \; & \textstyle\prod_{t=0}^{i-2} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_{i-1}(s_{i-1}, a_{i-1}) \cdot P_{s_{i-1}}^{a_{i-1}}(s_i) \cdot \mu_i(s_i, a_i) \\
= \; & \textstyle\prod_{t=0}^{i-1} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_i(s_i, a_i).
\end{aligned} \tag{15}$$

$\square$

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

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

For some $i \in \mathbb{N}_0$, let $\pi(i-1)$ 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 $\pi(i-1)$ is given by

$$\prod_{t=0}^{i-1} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right). \tag{16}$$

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 $\pi(i)$ 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 $s_0 = s' \in S$ be some fixed but arbitrary starting state. Then the probability of following any of the state-action trajectories $(s', a_0, ..., s, a)$, $a_t \in A$ for $t = 0, ..., i-1$, $s_t \in S$ for $t = 1, ..., i-1$, while following $\pi(i)$ is given by

$$\sum_{\substack{s_0 = s' \\ a_0, \ldots, a_{i-1} \in A \\ s_1, \ldots, s_{i-1} \in S}} \left[\prod_{t=0}^{i-2} \left(\mu_t(s_t, a_t) \cdot P_{s_t}^{a_t}(s_{t+1})\right) \cdot \mu_{i-1}(s_{i-1}, a_{i-1}) \cdot P_{s_{i-1}}^{a_{i-1}}(s)\right] \cdot \mu_i(s, a). \tag{17}$$

Proof. Using Lemma 1, we immediately arrive at the expression in Eq. 17 by summing over the set of relevant trajectories.

$$Tr_{s'}^{s,a} = \left\{ (s', a_0, s_1, a_1, \ldots, s_{i-1}, a_{i-1}, s, a) \,\middle|\, \begin{array}{l} a_0, \ldots, a_{i-1} \in A, \\ s_1, \ldots, s_{i-1} \in S \end{array} \right\}. \tag{18}$$

$\square$

Before we turn to the rewards in the next section, let us view this section's results from a functional point of view.

In our theoretical considerations, we usually assume the environment's transitional probabilities $P_{ss'}^a, a \in A, s, s' \in S$ to be both constant and known. Since the finding of a policy that is in some way optimal will be our main goal, it is intuitive to view all formulae derived in Lemma **??**, Corollary **??**nd Corollary **??** as functions of some policy $\pi = (\mu_0, \ldots, \mu_i)$. Recollecting our embedding of individual policies μ into (subsets) of $\mathbb{R}^{|S| \times |A|}$ in section **??**, Eq. 11, we can see that the space of all policies of length i can be seen as

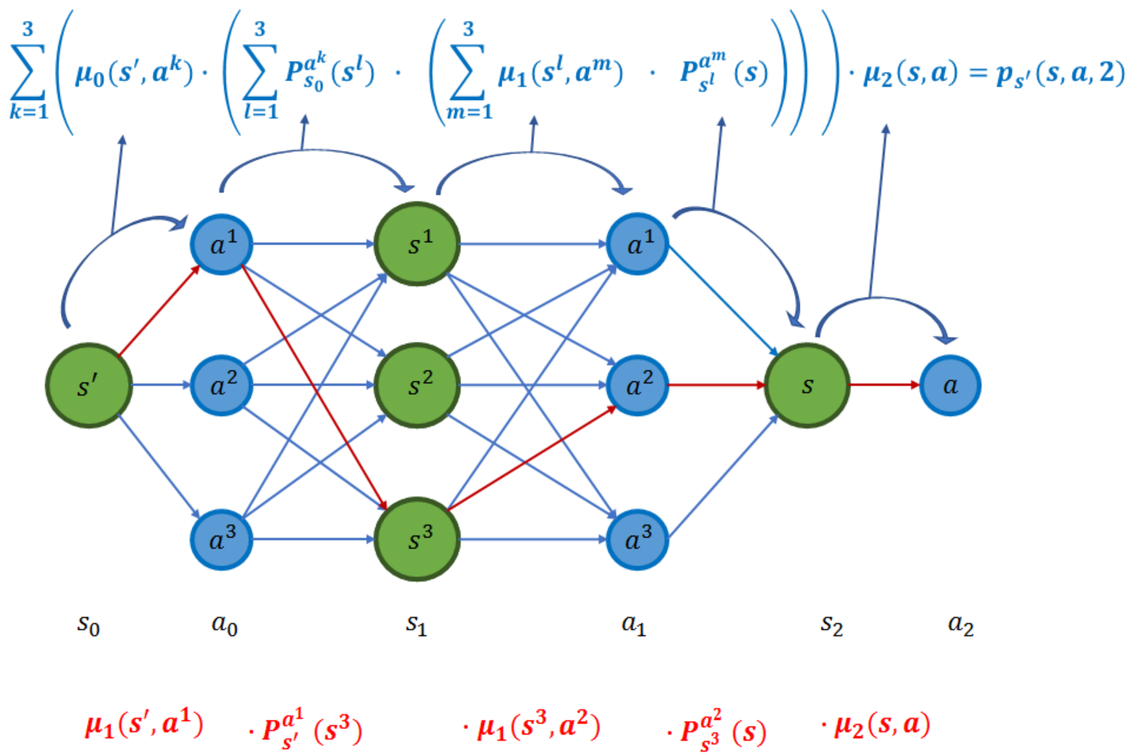

Figure 2: All the possible trajectories $(s', \dots, s, a)$ starting from $s_0 = s'$ and ending on $(s_3, a_3) = (s, a)$ while following $\mu_0, \dots, \mu_i$ for $i = 2$, $|S| = |A| = 3$. A sample trajectory is highlighted in magenta.

$$\begin{aligned}\Big\{\pi(i)|\pi(i) \text{ is a policy of length } i+1\Big\} &= \Big\{\pi(i) = (\mu_t)_{t=0,\dots,i}\Big|\mu_t \textit{ is a control law }\Big\}\\ &\cong \Pi(S,A)^{i+1},\end{aligned} \tag{19}$$

the three aforementioned results induce 3 continuous (the control laws' matrix representations' coefficients are being added and multiplied only - continuous operations) functions on the *compact* set $\Pi(S,A)^i$. We spell this out explicitly for the most important result, Corollary 2.

Corollary 3. *Let $s' \in S$ be a fixed but arbitrary starting state, and let $i \in \mathbb{N}_0$. Let further $(s,a) \in S \times A$ be a fixed but arbitrary state-action pair. Then the function*

$$\begin{aligned} p_{s'}(s,a,i): \quad \Pi(S,A)^i &\to [0,1] \\ (\mu_t)_{t=0,\dots,i} &\mapsto \Pr\left\{ (s', a_0, \dots, s, a), \left| \begin{array}{l} a_0, \dots, a_{i-1} \in A, \\ s_1, \dots, s_{i-1} \in S, \\ \pi(i) \end{array} \right. \right\} \end{aligned} \tag{20}$$

mapping policies onto their conditional probabilities of following any trajectory ending in (s,a), conditioned on starting in state s' at time $t=0$, is continuous on the space of permissable policies.

Proof. The proof consists solely in realizing that, for any policy $\pi(i) = (\mu_0, \dots, \mu_i) \in \Pi(S,A)^i$, the image of $\pi(i)$ under $p_{s'}(s,a)$ is of course the expression appearing in 17, where s_0 understood to be fixed at s'. This in turn is merely a sum of products of all of the guiding policy $\pi(i)$'s components' coefficients, and hence continuous in $\pi(i) \in \Pi(S,A)^i$, endowed with its canonical $|\cdot|_{\mathbb{R}^i}$ norm. □

A direct consequence of this is that for any choice of starting state s' and ending state-action pair $(s,a) \in S \times A$, the function $p_{s'}(s,a)$ as defined in 20 attains its maximum over the space of permissable policies of length $\Pi(S,A)^i$.

1.4 Rewards revisited

We have gathered enough preliminary results to return to a closer inspection our main object of focus: the rewards issued by the environment, depending on the state-action trajectories along which our agent travels.

Recall that at any time step t, we are given the immediate reward r_t's expectation conditioned only s_t and a_t (i.e. ignoring *all* previous elements of the state-action trajectory) as

$$\mathbb{E}[r_t|(s_0, a_0, \dots, s_{t-1}, a_{t-1}, s, a)] = \mathbb{E}[r_t|s_t = s, a_t = a] =: R(s,a)_t. \tag{21}$$

If we are interested in the expectation of r_t at time t in *general*, i.e. without conditioning on s_t and a_t, can use the above to see that generally

$$\begin{aligned}
\mathbb{E}[r_t] &= \sum_r \Pr\{r_t = r\} \cdot r \\
&= \sum_r \left(\sum_{a\in A} \sum_{s\in S} \Pr\{r_t = r | s_t = s, a_t = a\} \cdot \Pr\{a_t = a \cap s_t = s\} \right) \cdot r \\
&= \sum_{a\in A} \sum_{s\in S} \Pr\{a_t = a \cap s_t = s\} \sum_r \Pr\{r_t = r | a_t = a, s_t = s\} \cdot r \\
&= \sum_{a\in A} \sum_{s\in S} \Pr\{a_t = a \cap s_t = s\} \cdot \mathbb{E}[r_t | a_t = a, s_t = s] \\
&= \sum_{a\in A} \sum_{s\in S} \Pr\{a_t = a \cap s_t = s\} \cdot R(s,a)_t
\end{aligned} \tag{22}$$

It is natural to ask about the rather generic expression $\Pr(a_t = a \cap s_t = s)$ appearing in the above equation and how it might be connected to the control law sequences we discussed in the previous section. What Eq. 23 tells us is that obtaining the expectation of r_t requires the knowledge of the probability of observing the state action pair (s, a) at t for *all* $s \in S, a \in A$. This in turn implies that for any finite policy $\pi(i) = (\mu_0, \dots, \mu_i)$, $i \geq t$, and starting state s', we must have

$$\begin{aligned}
\mathbb{E}[r_t | s_0 = s', \pi(i)] &= \sum_{a\in A} \sum_{s\in S} \Pr\{a_t = a \cap s_t = s | s_0 = s', \pi(i)\} \cdot R(s,a)_t \\
&= \sum_{a\in A} \sum_{s\in S} p_{s'}(s,a,t) \cdot R(s,a)_t
\end{aligned} \tag{23}$$

Our knowledge of $p_{s'}(s, a)$ then ensures that the mapping

$$\begin{aligned}
\mathbb{E}[r_t | s_0 = s', \pi(i)] : \quad \Pi(S,A)^i &\to [0, M] \\
\pi(i) &\mapsto \sum_{a\in A} \sum_{s\in S} p'_s(s,a,t) \cdot R(s,a)_t
\end{aligned} \tag{24}$$

is continuous as a finite sum of continuous functions for any $s' \in S$, provided $i \geq t$. This result is worth repeating in cursive.

Corollary 4. *(Continuity of policy induced rewards w.r.t guiding policy) For each $t \in \mathbb{N}_0$, $i \geq t$, and starting state $s' \in S$, the reward r_t's expectation $\mathbb{E}[r_t | s' = s_0, \pi(i)]$, taken over all trajectories starting at s' and sampled according to some $\pi(i) \in \Pi(S,A)^i$, depends continuously on the guiding policy $\pi(i)$. As such, it attains its maximum on $\Pi(S,A)^i$.*

2 The discounted finite horizon problem

We dedicate this section to the formulation and solution of the so-called discounted finite horizon problem. We will use the results obtained in this section to solve the infinite horizon counterpart problem later.

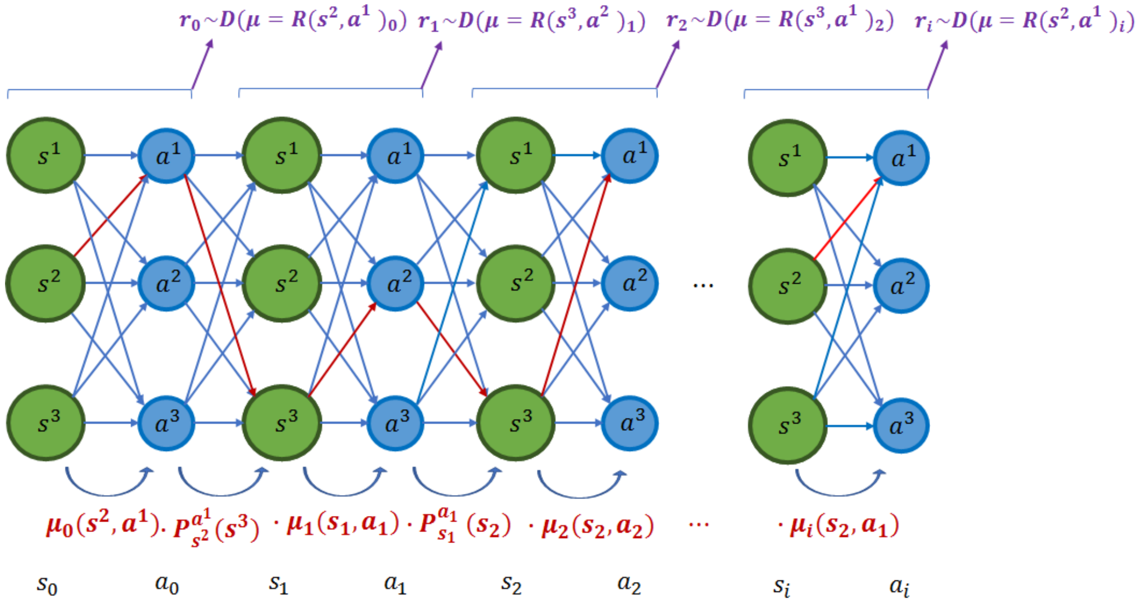

Figure 3: Having arbitrarily ordered our state space into $\{s^1, s^2, \ldots, s^{|S|}\} = S$, we show a sample trajectory, the associated probabilities and rewards while following $\mu_0, \ldots, \mu_i$ for $i = 3$, $|S| = |A| = 3$. The sample trajectory is highlighted in magenta.

Let $0 < \gamma < 1$, and let $\leq j \leq i$, $j, i \in \mathbb{N}$ be given. Let further d_{s_j} be an arbitrary distribution on the state space S at time $t = j$.

The discounted finite horizon problem is the searching of a policy π as discussed in section **??** that, conditioned on the starting state s_j being distributed according to $s_j \sim d_{s_j}$ at $t = j$, maximizes the induced finite sum of remaining discounted expected rewards. To be more precise, putting

$$J_{j,i}(d_{s_j}, \pi(i)) := \sum_{t=j}^{i} \mathbb{E}[\gamma^{t-j} \cdot r_t | s_j \sim d_{s_j}, \pi(i)] = \sum_{t=j}^{i} \gamma^{t-j} \cdot \mathbb{E}[r_t | s_j \sim d_{s_j}, \pi(i)], \tag{25}$$

our discounted finite horizon problem reduces to, given an arbitrary but fixed starting state distribution $s_j \sim d_{s_j}$, finding

$$\pi(i)^* := \underset{\pi(i) \in \Pi(S,A)^{i+1}}{\arg\max} J_{j,i}(d_{s_j}, \pi(i)). \tag{26}$$

For $j = 0$, we usually refer to the above problem simply as the *finite horizon problem*. For $j = 1, \ldots, i$ we usually refer to it as the *reward-to-go from time j*.

Note that since the distribution at time $t = j$ is specified and states, actions and rewards at times $t \geq j$ do not depend on any part of the state action trajectory prior to time $t = j$, the reward-to-go problem effectively is solved by finding appropriate control laws $\mu_j, \ldots, \mu_i$, leaving the first j control laws $\mu_0, \ldots, \mu_{j-1}$ of the solution $\pi(i)$ unspecified and therefore variable.

The existence of a solution to problem 26 follows from the continuity results found in the previous section.

Proposition 1. *(Existence of a solution) For any $j, i \in \mathbb{N}_0$ with $j \leq i$, initial distribution d_{s_j} over the state space S, the reward-to-go problem has a solution in $\Pi(S, A)^{i+1}$. That is, there exists a policy $\pi^*(i) \in \Pi(S, A)^{i+1}$ such that*

$$\sum_{t=j}^{i} \gamma^{t-j} \cdot \mathbb{E}[r_t | s_j \sim d_{s_j}, \pi^*(i)] = \max_{\pi(i) \in \Pi(S,A)^{i+1}} J_{j,i}(d_{s_j}, \pi(i)). \tag{27}$$

Proof. We can shift the environment index to create an environment that has been 'fast-forwarded' by j time steps by putting $s'_t = s_{t+j}$ and $r'_t = r_{t+j}$. The same is done for corresponding control laws, i.e. $\mu'_t = \mu'_{t+j}$. Thus, applying the t-th control law μ'_t of a policy π' to state s'_t in the fast forwarded environment is equivalent to applying $(t + j)$-th control law μ_t of a policy $\pi(i)$ to state s_t in the original environment.

Applying Corollary 4 to our fast forwarded environment, the mapping

$$\begin{aligned} J_{i-j}(s', \cdot) : \quad \Pi(S, A)^{i-j+1} &\rightarrow [0, (i - j)M] \\ \pi'(i) &\mapsto \sum_{t=0}^{i-j} \gamma^t \cdot \mathbb{E}[r'_t | s'_0 = s', \pi'(i - j)] \end{aligned} \tag{28}$$

is continuous as the sum of continuous functions. We can explicitly write out the specified initial state distribution d_{s_j}'s probabilities to see that

$$\sum_{t=0}^{i-j}\gamma^t\cdot\mathbb{E}\Big[r'_t\Big|\begin{array}{c}s'_0\sim d_{s_j},\\ \mu'_0,\dots,\mu'_{i-j}\end{array}\Big]=\sum_{t=0}^{i-j}\Big(\sum_{s'\in S}\Big(\Pr\{s'_0=s'|s'_0\sim d_{s_j}\}\cdot\gamma^t\cdot\mathbb{E}\Big[r'_t\Big|\begin{array}{c}s'_0=s',\\ \mu'_0,\dots,\mu'_{i-j}\end{array}\Big]\Big)\Big) \tag{29}$$

is continuous in μ'_0,μ'_{i-j} and therefore attains its maximum ${\mu'}^*_0,\dots,{\mu'}^*_{i-j}$ on the compact set $\Pi(S,A)^{i-j+1}$. But since by construction

$$\sum_{t=0}^{i}\gamma^t\cdot\mathbb{E}\Big[r_t\Big|\begin{array}{c}s_j\sim d_{s_j},\\ \mu_j,\dots,\mu_i\end{array}\Big]=\sum_{t=j}^{i-j}\gamma^t\cdot\mathbb{E}\Big[r'_t\Big|\begin{array}{c}s'_0\sim d_{s_j},\\ \mu'_0,\dots,\mu'_{i-j}\end{array}\Big], \tag{30}$$

it is clear that

$$\pi^*(i)=(\mu_0,\dots,\mu_{j-1},\mu^*_j,\dots,\mu^*_i)=(\mu_0,\dots,\mu_{j-1},{\mu'}^*_0,\dots,{\mu'}^*_{i-j}) \tag{31}$$

is a policy of length i that achieves the max of the reward-to-go from time j problem for any control laws $\mu_0,\dots,\mu_{j-1}\in\Pi(S,A)$. Note that the first $j+1$ control laws remain unspecified since they do not impact on the reward terms considered by the reward-to-go from time j, and can therefore be chosen at wish.

□

Now that our problem is well-defined and guaranteed to have a solution, we take a closer look at finding the optimal policies. In particular, because of the environment's and the agent's markov property, the reward-to-go from time j is closely connected with the reward-to-go from time $j-1$. In fact, a series of control laws forming the solution to the latter is also a solution to the former, plus an additional optial control law at time $j-1$. This is the so-called principle of optimality: An optimal path from A to C via B must also contain an optimal path from B to C. We make formalize this in the next

Theorem 1. *(Principle of optimality) Consider the discounted finite horizon problem as defined in 26. For all $s^1,\dots,s^{|S|}\in S$, define*

$$\mu^*_j(s^k,\cdot)\quad:=\quad\underset{\mu_a\in\Pi(\{s^k\},A)}{\arg\max}\Big(\mathbb{E}\Big[r_j\Big|\begin{array}{c}s_j=s^k,\\ \mu_j(s',\cdot)=\mu_a\end{array}\Big]+\max_{(\mu_{j+1},\dots,\mu_i)\in Pi(S,A)^{i-j}}\Big(\mathbb{E}\Big[\sum_{l=1}^{i-j}\gamma^l r_{j+l}\Big|\begin{array}{c}s_j=s^k,\\ \mu_j(s^k,\cdot)=\mu_a,\\ (\mu_j,\dots,\mu_i)\end{array}\Big]\Big)\Big), \tag{32}$$

$k=1,\dots,|S|$. Define the optimal control law at time j as

$$\mu^*_j=\begin{pmatrix}\mu_j(s^1,\cdot)\\ \vdots\\ \mu_j(s^{|S|},\cdot)\end{pmatrix}. \tag{33}$$

Let $0 \leq j \leq i$ be fixed but arbitrary, and let d_{s_j} be any distribution on the state space S. Then any policy with arbitrary first j policies of the form

$$\pi^*(i) = (\mu_0, \dots, \mu_{j-1}, \mu_j^*, \dots, \mu_i^*) \tag{34}$$

satisfies

$$\mathbb{E}\Big[\sum_{l=0}^{i-j} \gamma^l r_{j+l} \Big| \begin{array}{c} s_j \sim d_{s_j} \\ \pi^*(i) \end{array} \Big] = \max_{\pi(i) \in \Pi(S,A)^{i+1}} J_{j,i}(d_{s_j}, \pi(i)). \tag{35}$$

In other words, given any initial distribution d_{s_j} over state s_j at time $t = j$, the control laws $\mu_j^*, \mu_{j+1}^*, \dots, \mu_i^*$ as defined above achieve the optimal cost-to-go reward from time j.

Note that the above claim implies that the rewards collected from point $t = j$ onwards until the end $t = i$ can be maximised by following a fixed set of control laws that don't depend on your starting state distribution d_{s_j} of s_j. In particular, these control laws are the same for any fixed starting state $s_j = s'$, $s' \in S$. As the following proof will show, this is achieved by making use of the markov property - specifically, that rewards only depend on the most recent state and action, respectively, but not on the more distant past. It is this property that allows us to define the the optimal control laws μ_j^* iteratively, effectively rolling up the rewards process from the back.

Proof. We will show the claim by induction over the delayed start time index j. Let $s_k \in \{s^1, \dots, s^{|S|}\} = S$ be a fixed but arbitrary state from the (arbitrarily) ordered state space S, and let d_{s_i} be any distribution on s_i. For $j = i$, we clearly have

$$\mu_i^*(s^k, \cdot) := \underset{\mu_a \in \Pi(\{s^k\}, A)}{\arg\max} \, (\mathbb{E}[r_i | s_i \sim d_{s_i}, \mu_i]). \tag{36}$$

Since the specified distribution d_{s_i} at time $t = i$ renders the first i control laws of $\pi(i)$ irrelevant when considering the expectation of r_i, we can see that

$$
\begin{aligned}
\max_{\pi(i)\in\Pi(S,A)^{i+1}} \Big(J_{i,i}(d_{s_i},\pi(i))\Big) &:= \max_{\pi(i)\in\Pi(S,A)^{i+1}} (\mathbb{E}[r_i|s_i\sim d_{s_i},\pi(i)]) \\
&= \max_{\mu_i\in\Pi(S,A)} (\mathbb{E}[r_i|s_i\sim d_{s_i},\mu_i]) \\
&= \max_{\mu_i\in\Pi(S,A)} \Big(\sum_{s\in S} \Pr\{s_i=s|s_i\sim d_{s_i}\}\cdot\mathbb{E}[r_i|s_i=s,\mu_i]\Big) \\
&\leq \sum_{s\in S} \Big(\Pr\{s_i=s|s_i\sim d_{s_i}\}\cdot \max_{\mu_i\in\Pi(S,A)} (\mathbb{E}[r_i|s_i=s,\mu_i])\Big) \\
&= \sum_{s\in S} \Big(\Pr\{s_i=s|s_i\sim d_{s_i}\}\cdot \max_{\mu_a\in\Pi(\{s\},A)} \Big(\mathbb{E}\Big[r_i\Big| \begin{array}{c} s_i=s, \\ \mu_i(s,\cdot)=\mu_a \end{array}\Big]\Big)\Big) \\
&= \sum_{s\in S} \Big(\Pr\{s_i=s|s_i\sim d_{s_i}\}\cdot \mathbb{E}\Big[r_i\Big| \begin{array}{c} s_i=s,\mu_i \\ \mu_i(s,\cdot)=\mu_i^*(s,\cdot) \end{array}\Big]\Big) \\
&= \sum_{s\in S} \Big(\Pr\{s_i=s|s_i\sim d_{s_i}\}\cdot \mathbb{E}[r_i|s_i=s,\mu_i^*]\Big) \\
&= \mathbb{E}[r_i|s_i\sim d_{s_i},\mu_i^*].
\end{aligned}
\tag{37}
$$

Since the max of the reward-to-go over all partial control law sequences can not be smaller than the expectation achieved by one specific sequence, we naturally have $\geq$ as well. From this follows equality, marking our induction start.

For the induction step, let's assume Eq. **??** holds for $j=k$ for some k, $1\leq k\leq i$. We will show that it then also holds for $j=k-1$. Let again d_{s_k} be any distribution over the state s_k at time $t=k$, the time we start accumulating rewards in this Proposition's cost-to-go scenario. Denote by $\mu_k,\ldots\mu_i$ the control laws achieving optimal reward-to-go expectations as per our induction step's assumption. Since the specified distribution $d_{s_{k-1}}$ at time $t=k-1$ renders the first $k-1$ control laws of $\pi(i)$ irrelevant when considering the expectations of $r_{k-1},\ldots,r_i$, we can see that

$$
\begin{aligned}
& \max_{\pi(i)\in\Pi(S,A)^{i+1}} \Big(J_{k-1,i}(d_{s_{k-1}},\pi(i))\Big) \\
:= & \max_{\pi(i)\in\Pi(S,A)^{i+1}} \Big(\mathbb{E}\Big[\sum_{l=0}^{i-k+1}\gamma^l r_{k-1+l}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \pi(i) \end{array}\Big]\Big) \\
= & \max_{(\mu_{k-1},\ldots,\mu_i)\in\Pi(S,A)^{i-k+2}} \Big(\mathbb{E}\Big[\sum_{l=0}^{i-k+1}\gamma^l r_{k-1+l}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1},\ldots,\mu_i \end{array}\Big]\Big) \\
= & \max_{\mu_{k-1}\in\Pi(S,A)} \Big(\max_{(\mu_k,\ldots,\mu_i)\in\Pi(S,A)^{i-k+1}} \Big(\sum_{l=0}^{i-k+1}\gamma^l\mathbb{E}\Big[r_{k-1+1}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1},\ldots,\mu_i \end{array}\Big]\Big)\Big)
\end{aligned}
\tag{38}
$$

making use of Lemma **??** in the appendix.

Since no reward can be affected by a control law applied later in time, we can rewrite

$$\begin{aligned} &\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Big(\sum_{l=0}^{i-k+1}\gamma^l\mathbb{E}\Big[r_{k-1+l}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1},\dots,\mu_i\end{array}\Big]\Big)\Big) \\ =&\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\mathbb{E}\Big[r_{k-1}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1}\end{array}\Big]+\gamma\max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Big(\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1},\dots,\mu_i\end{array}\Big]\Big)\Big). \end{aligned} \tag{39}$$

An initial state distribution $s_{k-1}\sim d_{s_{k-1}}$ at time $t=k-1$ together with a control law $a_{k-1}\sim\mu_{k-1}(s_{k-1},\cdot)$ induces a state distribution which we will denote by $s_k\sim d_{s_k}(d_{s_{k-1}},\mu_{k-1})$ at time $t=k$. Since the distribution of r_t only depends on s_t and a_t but not on previous parts of the state-action trajectory, we can rewrite that last line to see that

$$\begin{aligned} &\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Big(\sum_{l=0}^{i-k+1}\gamma^l\mathbb{E}\Big[r_{k-1+l}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1},\dots,\mu_i\end{array}\Big]\Big)\Big) \\ =&\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\mathbb{E}\Big[r_{k-1}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1}\end{array}\Big]+\gamma\max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Big(\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big|\begin{array}{c} s_k\sim d_{s_k}(d_{s_{k-1}},\mu_{k-1}), \\ \mu_k,\dots,\mu_i\end{array}\Big]\Big)\Big) \end{aligned} \tag{40}$$

Due to the nested maxima, one would assume that the control laws $\mu_k,\dots,\mu_i$ achieving the inner maxima would be dependendent on the outer control law μ_{k-1} applied first. However, our induction assumption assures us that the inner maxima is achieved by the very control laws maximising the reward-to-go starting at $t=k$, that is, the control laws $\mu_k^*,\dots,\mu_i^*$ - *simultaneously* for all initial state distributions $d_{s_k}=d_{s_k}(d_{s_{k-1}},\mu_{k-1})$ induced by varying $d_{s_{k-1}}$ and μ_{k-1}:

$$\begin{aligned} &\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\mathbb{E}\Big[r_{k-1}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1}\end{array}\Big]+\gamma\max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Big(\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big|\begin{array}{c} s_k\sim d_{s_k}(d_{s_{k-1}},\mu_{k-1}), \\ \mu_k,\dots,\mu_i\end{array}\Big]\Big)\Big) \\ =&\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\mathbb{E}\Big[r_{k-1}\Big|\begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1}\end{array}\Big]+\gamma\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big|\begin{array}{c} s_k\sim d_{s_k}(d_{s_{k-1}},\mu_{k-1}), \\ \mu_k^*,\dots,\mu_i^*\end{array}\Big]\Big) \end{aligned} \tag{41}$$

Further decomposing the expectation over the initial state distribution $d_{s_{k-1}}$ at time $t=k-1$ and making use of Lemma **??** from the appendix yields

$$
\begin{aligned}
& \max_{\mu_{k-1}\in\Pi(S,A)} \Big(\mathbb{E}\Big[r_{k-1}\Big| \begin{matrix} s_{k-1}\sim d_{s_{k-1}}, \\ \mu_{k-1} \end{matrix}\Big] && + \quad \gamma\cdot\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{matrix} s_k\sim d_{s_k}(d_{s_{k-1}},\mu_{k-1}), \\ \mu_k^*,\dots,\mu_i^* \end{matrix}\Big]\Big) \\
= \quad & \max_{\mu_{k-1}\in\Pi(S,A)} \Big(\sum_{s\in S}\Big(\Big(\mathbb{E}\Big[r_{k-1}\Big| \begin{matrix} s_{k-1}=s, \\ \mu_{k-1} \end{matrix}\Big] && + \quad \gamma\cdot\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{matrix} s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}), \\ \mu_k^*,\dots,\mu_i^* \end{matrix}\Big]\Big) \\
& && \quad\cdot\Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_{k-1}}\}\Big)\Big) \\
\leq \quad & \sum_{s\in S}\Big(\max_{\mu_{k-1}\in\Pi(S,A)}\Big(\mathbb{E}\Big[r_{k-1}\Big| \begin{matrix} s_{k-1}=s, \\ \mu_{k-1} \end{matrix}\Big] && + \quad \gamma\cdot\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{matrix} s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}), \\ \mu_k^*,\dots,\mu_i^* \end{matrix}\Big]\Big) \\
& && \quad\cdot\Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_{k-1}}\}\Big) \\
= \quad & \sum_{s\in S}\Big(\max_{\mu_a\in\Pi(\{s\},A)}\Big(\mathbb{E}\Big[r_{k-1}\Big| \begin{matrix} s_{k-1}=s, \\ \mu_{k-1}(s,\cdot)=\mu_a \end{matrix}\Big] && + \quad \gamma\cdot\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{matrix} s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}(s,\cdot)=\mu_a), \\ \mu_k^*,\dots,\mu_i^* \end{matrix}\Big]\Big) \\
& && \quad\cdot\Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_{k-1}}\}\Big).
\end{aligned}
\tag{42}
$$

Note that in the step yielding $\leq$ we moved the maximisation inside the sum, maximising over each term individually. The last step is justified by the expectations inside the sum being constrained to the state s_{k-1} being fixed at some arbitrary state $s^k \in S$. Thus, effectively, only the 'k-th row' of μ_{k-1}, namely $\mu_{k-1}(s^k,\cdot)$, is used in each expectation. The somewhat clumsy expression

$$s_k \sim d_{s_k}(s_{k-1}=s, \mu_{k-1}(s,\cdot)=\mu_a) \tag{43}$$

expresses the fact that the distribution on the state s_k at time $t=k$ is the one induced by starting in state $s_{k-1}=s\in S$ at the previous time step $t=k-1$ and then applying a control law μ_{k-1} whose conditional probabilites of choosing actions $a\in A$ are required to be given by $\mu_{k-1}(s,\cdot)=\mu_a(\cdot)$. With that in mind, we reapply our induction step assumption to obtain

$$
\begin{aligned}
&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\}\\
&\qquad \cdot\Bigg(\max_{\mu_a\in\Pi(\{s\},A)}\Bigg(\mathbb{E}\Bigg[r_{k-1}\Bigg|\begin{array}{c}s_{k-1}=s,\\ \mu_{k-1}(s,\cdot)=\mu_a\end{array}\Bigg]\\
&\qquad\quad +\gamma\cdot\max_{(\mu_k,\ldots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Bigg(\mathbb{E}\Bigg[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Bigg|\begin{array}{c}s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}(s,\cdot)=\mu_a),\\ \mu_k,\ldots,\mu_i\end{array}\Bigg]\Bigg)\Bigg)\Bigg)\\
=&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\}\\
&\qquad \cdot\Bigg(\max_{\mu_a\in\Pi(\{s\},A)}\Bigg(\mathbb{E}\Bigg[r_{k-1}\Bigg|\begin{array}{c}s_{k-1}=s,\\ \mu_{k-1}(s,\cdot)=\mu_a\end{array}\Bigg]\\
&\qquad\quad +\gamma\cdot\max_{(\mu_k,\ldots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Bigg(\mathbb{E}\Bigg[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Bigg|\begin{array}{c}s_{k-1}=s,\\ \mu_{k-1}(s,\cdot)=\mu_a),\\ \mu_k,\ldots,\mu_i\end{array}\Bigg]\Bigg)\Bigg)\Bigg)
\end{aligned}
\tag{44}
$$

Closer inspection of the nested max reveals them to be the exact terms maximised by $\mu^*_{k-1}(s,\cdot)$, enabling us to get to simplify

$$
\begin{aligned}
&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\}\\
&\qquad \cdot\Bigg(\mathbb{E}\Bigg[r_{k-1}\Bigg|\begin{array}{c}s_{k-1}=s,\\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot)\end{array}\Bigg]\\
&\qquad\quad +\gamma\cdot\max_{(\mu_k,\ldots,\mu_i)\in\Pi(S,A)^{i-k+1}}\Bigg(\mathbb{E}\Bigg[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Bigg|\begin{array}{c}s_{k-1}=s,\\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot)),\\ \mu_k,\ldots,\mu_i\end{array}\Bigg]\Bigg)\Bigg)\Bigg)
\end{aligned}
\tag{45}
$$

Rewriting the conditions in the second expectation like before as a requirement on the distribution on the state s_k at time $t=k$ and applying our induction step assumption one last time, we get

$$
\begin{aligned}
&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\} \\
&\qquad \cdot \Bigg(\mathbb{E}\Big[r_{k-1}\Big| \begin{array}{c} s_{k-1}=s, \\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot) \end{array}\Big] \\
&\qquad\quad + \gamma \cdot \max_{(\mu_k,\dots,\mu_i)\in\Pi(S,A)^{i-k+1}} \Bigg(\mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{array}{c} s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot)), \\ \mu_k,\dots,\mu_i \end{array}\Big]\Bigg)\Bigg)\Bigg) \\
&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\} \\
&\qquad \cdot \Bigg(\mathbb{E}\Big[r_{k-1}\Big| \begin{array}{c} s_{k-1}=s, \\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot) \end{array}\Big] \\
&\qquad\quad + \gamma \cdot \mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{array}{c} s_k\sim d_{s_k}(s_{k-1}=s,\mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot)), \\ \mu^*_k,\dots,\mu^*_i \end{array}\Big]\Bigg)\Bigg) \\
&\sum_{s\in S}\Bigg(\quad \Pr\{s_{k-1}=s|s_{k-1}\sim d_{s_k}\} \\
&\qquad \cdot \Bigg(\mathbb{E}\Big[r_{k-1}\Big| \begin{array}{c} s_{k-1}=s, \\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot) \end{array}\Big] + \gamma \cdot \mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{array}{c} s_{k-1}=s, \\ \mu_{k-1}(s,\cdot)=\mu^*_{k-1}(s,\cdot) \end{array}\Big]\Bigg)\Bigg).
\end{aligned} \tag{46}
$$

Collapsing the sum into the initial distribution over s_{k-1} at time $t=k-1$, we finally have

$$
\begin{aligned}
\mathbb{E}\Big[r_{k-1}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ (\mu^*_{k-1},\dots,\mu^*_i) \end{array}\Big] + \mathbb{E}\Big[\sum_{l=0}^{i-k}\gamma^l r_{k+l}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ (\mu^*_{k-1},\dots,\mu^*_i) \end{array}\Big] \\
= \mathbb{E}\Big[\sum_{l=0}^{i-(k-1)}\gamma^l r_{k+l}\Big| \begin{array}{c} s_{k-1}\sim d_{s_{k-1}}, \\ (\mu^*_{k-1},\dots,\mu^*_i) \end{array}\Big].
\end{aligned} \tag{47}
$$

This shows $\leq$ in the induction step. Since the *max* of the reward-to-go starting at time $t=k-1$ taken over control laws $\mu_{k-1},\dots,\mu_i$ can not be smaller than the value achieved by $\mu^*_{k-1},\dots,\mu^*_i$, we have shown that equality holds. This completes the induction step and proves the claim for all $j=0,\dots,i$.

□

While the proof of this principle was somewhat technical, we hope that the following picture illustrates the intuitive idea behind it.

INSERT OPTIMAL POLICY FOR REWARD-TO-GO SCENARIOS HERE.

We will now reformulate the above results in a more operator friendly way. More precisely, we will define value functions for each point in time t that will assign the maximal achievable reward-to-go to each potential initial state $s\in S$. The knowledge of optimal policies for these sub-problems will be of use.

Proposition 2. *(Maixmal reward-to-go value functions) Let $i\in\mathbb{N}_0$ and let $j=-1,0,\dots,i$.*

Let further $d_{s_{i-j}}$ be a distribution on the state space S. For $j=-1$, set

$$\hat{J}_{j,i}(d_{s_j}) := 0 \tag{48}$$

for all policies $\pi(i) \in \Pi(S,A)^{i+1}$. For $j = 0, \dots, i$, set

$$\hat{J}_{j,i}(d_{s_{i-j}}) := \max_{\mu_{i-j} \in \Pi(S,A)} \left(\mathbb{E}\left[r_{i-j} + \gamma \cdot \hat{J}_{j-1,i}(d_{s_{i-j+1}}) \middle| \begin{array}{c} s_{i-j} \sim d_{s_{i-j}}, \mu_{i-j} \\ d_{s_{i-j+1}} = d_{s_{i-j+1}}(d_{s_{i-j}}, \mu_{i-j}) \end{array} \right] \right). \tag{49}$$

Then for all $j = i, \dots, 0$, we have

$$\hat{J}_{i-j,i}(d_{s_j}) = \max_{\pi(i) \in \Pi(S,A)^{i+1}} J_{j,i}(d_{s_j}, \pi(i)). \tag{50}$$

In other words, the functional $\hat{J}_{i-j,i}(d_{s_j})$ gives the maximal achievable reward-to-go from time $t = j$ given an initial state distribution $s_j \sim d_{s_j}$. The proof will show that this maximal reward-to-go is achieved by precisely the optimal control laws $(\mu_0, \dots, \mu_i)$ constructed in Theorem **??**.

Proof. We will use an induction argument for the start index j. For the induction start, let $j = 0$, and let d_{s_i} be any arbitrary but fixed distribution on the space state S. $\square$