Measure-theoretic Bayesian Reinforcement Learning

Paulo Rauber

2024

These notes describe the fundamentals of Bayes-adaptive Markov decision processes [1] using measure-theoretic probability. For a less rigorous introduction, see the reinforcement learning notes by the same author.

1 Canonical Models

Definition 1.1. A set of states $\mathcal{S}$ is a non-empty subset of $\mathbb{N}$.

Definition 1.2. A set of actions $\mathcal{A}$ is a non-empty subset of $\mathbb{N}$.

Definition 1.3. A model p over a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$ is a function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ such that $\sum_{s'} p(s, a, s') = 1$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For convenience, let $p^a_{s,s'} = p(s, a, s')$.

Definition 1.4. A Markov decision process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$ is composed of:

- A set of states $\mathcal{S}$;
- A set of actions $\mathcal{A}$;
- A model p over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$;
- A reward function $r : \mathcal{S} \to \mathbb{R}$ such that $|r| \leq c$ for some $c \in (0, \infty)$;
- A discount factor $\gamma \in (0, 1)$.

Definition 1.5. For a set of states $\mathcal{S}$, an initial distribution μ is a probability measure on the measurable space $(\mathcal{S}, \mathcal{P}(\mathcal{S}))$, where $\mathcal{P}(\mathcal{S})$ is the set of all subsets of $\mathcal{S}$. For convenience, let $\mu_s = \mu(\{s\})$.

Definition 1.6. For a set of states $\mathcal{S}$ and a set of actions $\mathcal{A}$, an adaptive policy π is a sequence of functions $(\pi_t : \mathcal{S}^{t+1} \to \mathcal{A} \mid t \in \mathbb{N})$, where π_t is called a policy for time step t.

Proposition 1.1. For every Markov decision process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, initial distribution μ, and adaptive policy $\pi = (\pi_t \mid t \in \mathbb{N})$, there is a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a stochastic process $S = (S_t : \Omega \to \mathcal{S} \mid t \in \mathbb{N})$ such that, for every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$,

$$\mathbb{P}(S_0 = s_0, \dots, S_t = s_t) = \mu_{s_0} \prod_{k=1}^{t} p^{\pi_{k-1}(s_0,\dots,s_{k-1})}_{s_{k-1},s_k}.$$

Proof. By Kolmogorov's extension theorem, there is a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$ carrying a countable set of independent random variables $\{S_0 : \Omega \to \mathcal{S}\} \cup \{Z_{s_0,\dots,s_t} : \Omega \to \mathcal{S} \mid t \in \mathbb{N} \text{ and } (s_0, \dots, s_t) \in \mathcal{S}^{t+1}\}$ such that $\mathbb{P}(S_0 = s_0) = \mu_{s_0}$ for every $s_0 \in \mathcal{S}$ and $\mathbb{P}(Z_{s_0,\dots,s_t} = s_{t+1}) = p^{\pi_t(s_0,\dots,s_t)}_{s_t,s_{t+1}}$ for every $t \in \mathbb{N}$ and $(s_0, \dots, s_{t+1}) \in \mathcal{S}^{t+2}$.

For every $t \in \mathbb{N}$, let $S_{t+1} : \Omega \to \mathcal{S}$ be given by $S_{t+1} = Z_{S_0,\dots,S_t}$. By definition, for every $t \in \mathbb{N}$ and $\omega \in \Omega$,

$$S_{t+1}(\omega) = Z_{S_0(\omega),\dots,S_t(\omega)}(\omega) = \sum_{s_0} \cdots \sum_{s_t} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}(\omega) Z_{s_0,\dots,s_t}(\omega).$$

For every $t \in \mathbb{N}$, we know that S_{t+1} is a random variable because $S_0, \dots, S_t$ are random variables.

For every $t \in \mathbb{N}$ and $s_{t+1} \in \mathcal{S}$, since $\{S_{t+1} = s_{t+1}\} \cap \Omega = \{S_{t+1} = s_{t+1}\}$,

$$\{S_{t+1} = s_{t+1}\} = \{Z_{S_0,\dots,S_t} = s_{t+1}\} = \bigcup_{s_0} \cdots \bigcup_{s_t} \{S_0 = s_0, \dots, S_t = s_t\} \cap \{Z_{s_0,\dots,s_t} = s_{t+1}\}.$$

Using induction, we will now show that, for every $t \in \mathbb{N}^+$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$,

$$\bigcap_{k=0}^{t} \{S_k = s_k\} = \{S_0 = s_0\} \cap \bigcap_{k=1}^{t} \{Z_{s_0,\dots,s_{k-1}} = s_k\}.$$

Using the previous result, for every $(s_0, s_1) \in \mathcal{S}^2$,

$$\{S_0 = s_0\} \cap \{S_1 = s_1\} = \{S_0 = s_0\} \cap \bigcup_{s_0'} \{S_0 = s_0'\} \cap \{Z_{s_0'} = s_1\} = \{S_0 = s_0\} \cap \{Z_{s_0} = s_1\}.$$

Suppose that the inductive hypothesis is true for some $t \in \mathbb{N}^+$. For every $(s_0, \ldots, s_{t+1}) \in \mathcal{S}^{t+2}$,

$$\bigcap_{k=0}^{t+1} \{S_k = s_k\} = \left(\bigcap_{k=0}^{t} \{S_k = s_k\} \right) \cap \left(\bigcup_{s_0'} \cdots \bigcup_{s_t'} \{S_0 = s_0', \ldots, S_t = s_t'\} \cap \{Z_{s_0', \ldots, s_t'} = s_{t+1}\} \right).$$

By distributing the intersection over the unions and using the inductive hypothesis,

$$\bigcap_{k=0}^{t+1} \{S_k = s_k\} = \left(\bigcap_{k=0}^{t} \{S_k = s_k\} \right) \cap \{Z_{s_0, \ldots, s_t} = s_{t+1}\} = \{S_0 = s_0\} \cap \bigcap_{k=1}^{t+1} \{Z_{s_0, \ldots, s_{k-1}} = s_k\}.$$

For every $t \in \mathbb{N}^+$ and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$, the event $\bigcap_{k=0}^{t} \{S_k = s_k\}$ is the intersection of events from the σ-algebras of independent random variables. Therefore, using the previous result,

$$\mathbb{P}(S_0 = s_0, \ldots, S_t = s_t) = \mathbb{P}(S_0 = s_0) \prod_{k=1}^{t} \mathbb{P}(Z_{s_0, \ldots, s_{k-1}} = s_k) = \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1}, s_k}^{\pi_{k-1}(s_0, \ldots, s_{k-1})}.$$

□

Definition 1.7. For a set of states $\mathcal{S}$, the canonical space $(\Omega, \mathcal{F})$ that carries the state process $S = (S_t \mid t \in \mathbb{N})$ is a measurable space such that $\Omega = \mathcal{S}^\infty$. Furthermore, for every $t \in \mathbb{N}$, the function $S_t : \Omega \to \mathcal{S}$ is given by $S_t(\omega) = \omega_t$ and the σ-algebra $\mathcal{F}$ on Ω is given by $\mathcal{F} = \sigma(S_0, S_1, \ldots)$.

Theorem 1.1 (Existence and uniqueness of the canonical triple for a Markov decision process)**.** For every Markov decision process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$, initial distribution μ, and adaptive policy $\pi = (\pi_t \mid t \in \mathbb{N})$, there is a unique probability measure $\mathbb{P}^{\mu,\pi}$ on the canonical space $(\Omega, \mathcal{F})$ that carries the state process $S = (S_t \mid t \in \mathbb{N})$ such that, for every $t \in \mathbb{N}$ and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \ldots, S_t = s_t) = \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1}, s_k}^{\pi_{k-1}(s_0, \ldots, s_{k-1})}.$$

The probability triple $(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ is called the canonical triple for the Markov decision process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$ under the initial distribution μ and the adaptive policy π.

Proof. Proposition 1.1 ensures that there is a probability triple $(\tilde{\Omega}^{\mu,\pi}, \tilde{\mathcal{F}}^{\mu,\pi}, \tilde{\mathbb{P}}^{\mu,\pi})$ carrying the stochastic process $(\tilde{S}_t^{\mu,\pi} : \tilde{\Omega}^{\mu,\pi} \to \mathcal{S} \mid t \in \mathbb{N})$ such that, for every $t \in \mathbb{N}$ and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$,

$$\tilde{\mathbb{P}}^{\mu,\pi}(\tilde{S}_0^{\mu,\pi} = s_0, \ldots, \tilde{S}_t^{\mu,\pi} = s_t) = \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1}, s_k}^{\pi_{k-1}(s_0, \ldots, s_{k-1})}.$$

Consider the function $\tilde{S}^{\mu,\pi} : \tilde{\Omega}^{\mu,\pi} \to \Omega$ given by $\tilde{S}^{\mu,\pi}(\tilde{\omega}) = (\tilde{S}_t^{\mu,\pi}(\tilde{\omega}) \mid t \in \mathbb{N})$. By Proposition 8.1, the function $\tilde{S}^{\mu,\pi}$ is $\tilde{\mathcal{F}}^{\mu,\pi}/\mathcal{F}$-measurable, so that the function $\mathbb{P}^{\mu,\pi} : \mathcal{F} \to [0, 1]$ defined by

$$\mathbb{P}^{\mu,\pi}(F) = \tilde{\mathbb{P}}^{\mu,\pi}((\tilde{S}^{\mu,\pi})^{-1}(F)) = \tilde{\mathbb{P}}^{\mu,\pi}(\{\tilde{\omega} \in \tilde{\Omega}^{\mu,\pi} \mid \tilde{S}^{\mu,\pi}(\tilde{\omega}) \in F\})$$

is a probability measure on the measurable space $(\Omega, \mathcal{F})$.

Clearly, $\mathbb{P}^{\mu,\pi}(\Omega) = \tilde{\mathbb{P}}^{\mu,\pi}((\tilde{S}^{\mu,\pi})^{-1}(\Omega)) = \tilde{\mathbb{P}}^{\mu,\pi}(\tilde{\Omega}^{\mu,\pi}) = 1$ and $\mathbb{P}^{\mu,\pi}(\emptyset) = \tilde{\mathbb{P}}^{\mu,\pi}((\tilde{S}^{\mu,\pi})^{-1}(\emptyset)) = \tilde{\mathbb{P}}^{\mu,\pi}(\emptyset) = 0$. For any sequence of sets $(F_n \in \mathcal{F} \mid n \in \mathbb{N})$ such that $F_n \cap F_m = \emptyset$ for $n \neq m$,

$$\mathbb{P}^{\mu,\pi}\left(\bigcup_n F_n\right) = \tilde{\mathbb{P}}^{\mu,\pi}\left((\tilde{S}^{\mu,\pi})^{-1}\left(\bigcup_n F_n\right)\right) = \tilde{\mathbb{P}}^{\mu,\pi}\left(\bigcup_n (\tilde{S}^{\mu,\pi})^{-1}(F_n)\right) = \sum_n \tilde{\mathbb{P}}^{\mu,\pi}\left((\tilde{S}^{\mu,\pi})^{-1}(F_n)\right) = \sum_n \mathbb{P}^{\mu,\pi}(F_n),$$

where we have used the fact that $(\tilde{S}^{\mu,\pi})^{-1}(F_n) \cap (\tilde{S}^{\mu,\pi})^{-1}(F_m) = (\tilde{S}^{\mu,\pi})^{-1}(F_n \cap F_m) = \emptyset$ for $n \neq m$.

For every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$,

$$\mathbb{P}^{\mu,\pi}\left(S_0 = s_0, \dots, S_t = s_t\right) = \tilde{\mathbb{P}}^{\mu,\pi}(\{\tilde{\omega} \in \tilde{\Omega}^{\mu,\pi} \mid \tilde{S}^{\mu,\pi}(\tilde{\omega}) \in \{\omega \in \Omega \mid S_0(\omega) = s_0, \dots, S_t(\omega) = s_t\}\}).$$

Because $\Omega = \mathcal{S}^\infty$ and $S_t(\omega) = \omega_t$ for every $t \in \mathbb{N}$,

$$\mathbb{P}^{\mu,\pi}\left(S_0 = s_0, \dots, S_t = s_t\right) = \tilde{\mathbb{P}}^{\mu,\pi}(\{\tilde{\omega} \in \tilde{\Omega}^{\mu,\pi} \mid \tilde{S}_0^{\mu,\pi}(\tilde{\omega}) = s_0, \dots, \tilde{S}_t^{\mu,\pi}(\tilde{\omega}) = s_t\}) = \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})},$$

so that a probability measure on $(\Omega, \mathcal{F})$ with the desired properties exists.

Naturally, any two desired probability measures on $(\Omega, \mathcal{F})$ must agree on the π-system $\mathcal{I} \subseteq \mathcal{F}$ given by

$$\mathcal{I} = \{\emptyset\} \cup \{\{S_0 = s_0, \dots, S_t = s_t\} \mid t \in \mathbb{N} \text{ and } (s_0, \dots, s_t) \in \mathcal{S}^{t+1}\} \cup \{\Omega\}.$$

Since $\sigma(\mathcal{I}) = \mathcal{F}$ by Proposition 8.3, $\mathbb{P}^{\mu,\pi}$ is the unique probability measure with the desired properties. □

Definition 1.8. Let $\mathcal{M}$ be a set of models over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$. For every state $s \in \mathcal{S}$, action $a \in \mathcal{A}$, and state $s' \in \mathcal{S}$, the function $q_{s,s'}^a : \mathcal{M} \to [0,1]$ is given by $q_{s,s'}^a(p) = p_{s,s'}^a$.

Definition 1.9. The canonical space $(\mathcal{M}, \mathcal{G})$ for the set of models $\mathcal{M}$ over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$ is the measurable space such that $\mathcal{G} = \sigma\left(\cup_{(s,a,s')}\sigma\left(q_{s,s'}^a\right)\right)$.

Definition 1.10. A Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$ is composed of:

- A set of states $\mathcal{S}$;
- A set of actions $\mathcal{A}$;
- A non-empty set of models $\mathcal{M}$ over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$;
- A prior ψ, which is a probability measure on the canonical space $(\mathcal{M}, \mathcal{G})$ for the set of models $\mathcal{M}$;
- A reward function $r : \mathcal{S} \to \mathbb{R}$ such that $|r| \leq c$ for some $c \in (0, \infty)$;
- A discount factor $\gamma \in (0, 1)$.

Definition 1.11. Let $(\mathcal{M}, \mathcal{G})$ be the canonical space for the set of models $\mathcal{M}$ over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$. Let $(\Omega', \mathcal{F}')$ be the canonical space that carries the state process $S' = (S'_t \mid t \in \mathbb{N})$ for the set of states $\mathcal{S}$. The canonical space $(\Omega, \mathcal{F})$ that carries the model variable M and the state process $S = (S_t \mid t \in \mathbb{N})$ is given by $(\Omega, \mathcal{F}) = (\mathcal{M} \times \Omega', \mathcal{G} \times \mathcal{F}')$. The $\mathcal{F}/\mathcal{G}$-measurable function $M : \Omega \to \mathcal{M}$ is given by $M(p, \omega') = p$. For every $t \in \mathbb{N}$, the $\mathcal{F}$-measurable function $S_t : \Omega \to \mathcal{S}$ is given by $S_t(p, \omega') = S'_t(\omega')$.

Theorem 1.2 (Existence and uniqueness of the canonical triple for a Bayes-adaptive Markov decision process)**.** For every Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$, initial distribution μ, and adaptive policy π, there is a unique probability measure $\mathbb{P}^{\mu,\pi}$ on the canonical space $(\Omega, \mathcal{F}) = (\mathcal{M} \times \Omega', \mathcal{G} \times \mathcal{F}')$ that carries the model variable M and the state process $S = (S_t \mid t \in \mathbb{N})$ such that for every $G \in \mathcal{G}$, $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$,

$$\mathbb{P}^{\mu,\pi}(M \in G, S_0 = s_0, \dots, S_t = s_t) = \int_G \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})} \; \psi(dp).$$

The probability triple $(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ is called the canonical triple for the Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$ under the initial distribution μ and the adaptive policy π.

Proof. For every $p \in \mathcal{M}$, let $(\Omega', \mathcal{F}', \mathbb{P}^{\mu,\pi,p})$ denote the canonical triple for the Markov decision process $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$ under the initial distribution μ and the adaptive policy π.

Let $K^{\mu,\pi} : \mathcal{M} \times \mathcal{F}' \to [0,1]$ be a function given by $K^{\mu,\pi}(p, F') = \mathbb{P}^{\mu,\pi,p}(F')$. We will start by showing that $K^{\mu,\pi}$ is a probability kernel from $\mathcal{M}$ to Ω'.

For every $p \in \mathcal{M}$, note that the function $K^{\mu,\pi}(p, \cdot) : \mathcal{F}' \to [0,1]$ is a probability measure on $(\Omega', \mathcal{F}')$. For every $F' \in \mathcal{F}'$, it remains to show that the function $K^{\mu,\pi}(\cdot, F') : \mathcal{M} \to [0,1]$ is $\mathcal{G}$-measurable.

By Proposition 8.3, a π-system $\mathcal{I}' \subseteq \mathcal{F}'$ such that $\sigma(\mathcal{I}') = \mathcal{F}'$ is given by

$$\mathcal{I}' = \{\emptyset\} \cup \{\{S'_0 = s_0, \dots, S'_t = s_t\} \mid t \in \mathbb{N} \text{ and } (s_0, \dots, s_t) \in \mathcal{S}^{t+1}\} \cup \{\Omega'\}.$$

Since $K^{\mu,\pi}(\cdot,\emptyset)$ and $K^{\mu,\pi}(\cdot,\Omega')$ are $\mathcal{G}$-measurable, let $I' = \{S'_0 = s_0, \dots, S'_t = s_t\}$ for some $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$. In that case,

$$K^{\mu,\pi}(p, I') = \mathbb{P}^{\mu,\pi,p}(S'_0 = s_0, \dots, S'_t = s_t) = \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})} = \mu_{s_0} \prod_{k=1}^{t} q_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})}(p),$$

so that $K^{\mu,\pi}(\cdot, I')$ is $\mathcal{G}$-measurable for every $I' \in \mathcal{I}'$. Because $\mathcal{I}'$ is a π-system on Ω' such that $\sigma(\mathcal{I}') = \mathcal{F}'$, recall that $K^{\mu,\pi}$ is a probability kernel from $\mathcal{M}$ to Ω'.

Consider the unique probability measure $\mathbb{P}^{\mu,\pi}$ on $(\Omega, \mathcal{F})$ such that, for every $G \in \mathcal{G}$ and $F' \in \mathcal{F}'$,

$$\mathbb{P}^{\mu,\pi}(G \times F') = \int_G K^{\mu,\pi}(p, F')\psi(dp) = \int_G \mathbb{P}^{\mu,\pi,p}(F')\psi(dp).$$

We will show that $\mathbb{P}^{\mu,\pi}$ is the unique probability measure on $(\Omega, \mathcal{F})$ with the desired properties.

For every $G \in \mathcal{G}$, note that $\{M \in G\} = G \times \Omega'$. For every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, note that $\{S_0 = s_0, \dots, S_t = s_t\} = \mathcal{M} \times \{S'_0 = s_0, \dots, S'_t = s_t\}$. Therefore,

$$\mathbb{P}^{\mu,\pi}(M \in G, S_0 = s_0, \dots, S_t = s_t) = \mathbb{P}^{\mu,\pi}(G \times \{S'_0 = s_0, \dots, S'_t = s_t\}) = \int_G \mu_{s_0} \prod_{k=1}^{t} p_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})} \ \psi(dp).$$

Let $\mathcal{J} = \{G \times I' \mid G \in \mathcal{G} \text{ and } I' \in \mathcal{I}'\}$. Note that any two desired probability measures on $(\Omega, \mathcal{F})$ must agree on $\mathcal{J}$. Because $\mathcal{I}'$ is set of subsets of Ω' such that $\Omega' \in \mathcal{I}'$ and $\sigma(\mathcal{I}') = \mathcal{F}'$, recall that $\sigma(\mathcal{J}) = \mathcal{F}$. Because $\mathcal{J}$ is a π-system on Ω, $\mathbb{P}^{\mu,\pi}$ is the unique probability measure on $(\Omega, \mathcal{F})$ with the desired properties. □

For the remaining text, let $(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ denote the canonical triple for the Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$ under the initial distribution μ and the adaptive policy π. Recall that the measurable space $(\Omega, \mathcal{F})$ carries the model variable $M : \Omega \to \mathcal{M}$ and the state process $S = (S_t : \Omega \to \mathcal{S} \mid t \in \mathbb{N})$.

2 Conditional Expectations

Definition 2.1. For every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, the posterior predictive $\rho^{\mu,\pi}_{s_0,\dots,s_t} : \mathcal{S} \to [0,1]$ given the sequence of states $(s_0, \dots, s_t)$ under the initial distribution μ and the adaptive policy π is defined by

$$\rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}) = \begin{cases} \frac{\mathbb{P}^{\mu,\pi}(S_0=s_0,\dots,S_t=s_t,S_{t+1}=s_{t+1})}{\mathbb{P}^{\mu,\pi}(S_0=s_0,\dots,S_t=s_t)}, & \text{if } \mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_t = s_t) \neq 0, \\ 1, & \text{if } \mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_t = s_t) = 0 \text{ and } s_{t+1} = \min \mathcal{S}, \\ 0, & \text{if } \mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_t = s_t) = 0 \text{ and } s_{t+1} \neq \min \mathcal{S}, \end{cases}$$

where the last two cases help ensure that $\sum_{s_{t+1}} \rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}) = 1$.

Definition 2.2. For every $t \in \mathbb{N}$, the history $\mathcal{H}_t$ up to time t is defined by $\mathcal{H}_t = \sigma(S_0, \dots, S_t)$.

Proposition 2.1. For every $t \in \mathbb{N}$ and $s_{t+1} \in \mathcal{S}$, almost surely,

$$\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1}) = \mathbb{P}^{\mu,\pi}(S_{t+1} = s_{t+1} \mid \mathcal{H}_t).$$

Proof. Recall that $\mathbb{P}^{\mu,\pi}(S_{t+1} = s_{t+1} \mid \mathcal{H}_t) = \mathbb{E}^{\mu,\pi}(\mathbb{I}_{\{S_{t+1}=s_{t+1}\}} \mid \mathcal{H}_t)$. Clearly, $\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1}) \in \mathcal{L}^1(\Omega, \mathcal{H}_t, \mathbb{P}^{\mu,\pi})$. By Proposition 8.2, every $H_t \in \mathcal{H}_t$ is given by $H_t = \bigcup_{s \in A}\{S_0 = s_0, \dots, S_t = s_t\}$ for some $A \subseteq \mathcal{S}^{t+1}$, where $s = (s_0, \dots, s_t)$. Therefore,

$$\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1})\mathbb{I}_{H_t} = \sum_{s \in A} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}} \rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1}) = \sum_{s \in A} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}} \rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}).$$

Since the terms in the summation above are non-negative,

$$\mathbb{E}^{\mu,\pi}\left(\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1})\mathbb{I}_{H_t}\right) = \sum_{s \in A} \mathbb{P}^{\mu,\pi}\left(S_0 = s_0, \dots, S_t = s_t\right) \rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}).$$

By cancelling terms,

$$\mathbb{E}^{\mu,\pi}\left(\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1})\mathbb{I}_{H_t}\right) = \sum_{s \in A} \mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_t = s_t, S_{t+1} = s_{t+1}).$$

Since the terms in the summation above are non-negative,

$$\mathbb{E}^{\mu,\pi}\left(\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1})\mathbb{I}_{H_t}\right) = \mathbb{E}^{\mu,\pi}\left(\sum_{s \in A} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\mathbb{I}_{\{S_{t+1}=s_{t+1}\}}\right) = \mathbb{E}^{\mu,\pi}\left(\mathbb{I}_{\{S_{t+1}=s_{t+1}\}}\mathbb{I}_{H_t}\right).$$

□

Definition 2.3. For every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, the adaptive policies π and π' agree on the sequence of states $(s_0, \dots, s_t)$ if $\pi_k(s_0, \dots, s_k) = \pi'_k(s_0, \dots, s_k)$ for every $k \leq t$.

Proposition 2.2. For every $t \in \mathbb{N}$ and $(s_0, \dots, s_{t+1}) \in \mathcal{S}^{t+2}$, if the adaptive policies π and π' agree on the sequence of states $(s_0, \dots, s_t)$, then $\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mathbb{P}^{\mu,\pi'}(S_0 = s_0, \dots, S_{t'} = s_{t'})$ for every $t' \leq t+1$.

Proof. Since $\{M \in \mathcal{M}\} = \Omega$, and $\pi_{k-1}(s_0, \dots, s_{k-1}) = \pi'_{k-1}(s_0, \dots, s_{k-1})$ for every $k \leq t+1$,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \int_{\mathcal{M}} \mu_{s_0} \prod_{k=1}^{t'} p_{s_{k-1},s_k}^{\pi_{k-1}(s_0,\dots,s_{k-1})}\ \psi(dp) = \int_{\mathcal{M}} \mu_{s_0} \prod_{k=1}^{t'} p_{s_{k-1},s_k}^{\pi'_{k-1}(s_0,\dots,s_{k-1})}\ \psi(dp).$$

□

Proposition 2.3. For every $t \in \mathbb{N}$ and $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, if the adaptive policies π and π' agree on the sequence of states $(s_0, \dots, s_t)$, then $\rho^{\mu,\pi}_{s_0,\dots,s_t} = \rho^{\mu,\pi'}_{s_0,\dots,s_t}$.

Proof. This result is obtained by combining Proposition 2.2 with the definitions of $\rho^{\mu,\pi}_{s_0,\dots,s_t}$ and $\rho^{\mu,\pi'}_{s_0,\dots,s_t}$. □

Definition 2.4. For every $t \in \mathbb{N}$, sequence of states $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, and sequence of actions $(a_0, \dots, a_t) \in \mathcal{A}^{t+1}$, the posterior predictive $\rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t} : \mathcal{S} \to [0,1]$ given $(s_0, \dots, s_t)$ and $(a_0, \dots, a_t)$ under μ is defined by

$$\rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1}) = \rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}),$$

where π is an adaptive policy such that $\pi_k(s_0, \dots s_k) = a_k$ for every $k \leq t$, and well-defined by Proposition 2.3.

Proposition 2.4. Consider an adaptive policy π and let $A_k = \pi_k(S_0, \dots, S_k)$ for every $k \in \mathbb{N}$. For every $t \in \mathbb{N}$ and $s_{t+1} \in \mathcal{S}$, almost surely,

$$\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \mathbb{P}^{\mu,\pi}(S_{t+1} = s_{t+1} \mid \mathcal{H}_t).$$

Proof. Because $\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1})\mathbb{I}_\Omega$,

$$\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1}),$$

where $s = (s_0, \dots, s_t)$ and $a_k = \pi_k(s_0, \dots, s_k)$ for every $k \leq t$. From the definition of $\rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1})$,

$$\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\rho^{\mu,\pi}_{s_0,\dots,s_t}(s_{t+1}) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\rho^{\mu,\pi}_{S_0,\dots,S_t}(s_{t+1}).$$

By Proposition 2.1, almost surely,

$$\rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) = \sum_{s \in \mathcal{S}^{t+1}} \mathbb{I}_{\{S_0=s_0,\dots,S_t=s_t\}}\mathbb{P}^{\mu,\pi}(S_{t+1} = s_{t+1} \mid \mathcal{H}_t) = \mathbb{P}^{\mu,\pi}(S_{t+1} = s_{t+1} \mid \mathcal{H}_t).$$

□

Posterior predictive functions have a central role in many Bayesian reinforcement learning algorithms. These functions are provided for some Bayes-adaptive Markov decision processes in Section 7.

3 Discounted Return

Definition 3.1. The discounted return $U_{t:h}$ after time step $t \in \mathbb{N}$ up to the horizon $h \in \mathbb{N}$ is defined by

$$U_{t:h} = \sum_{k=t+1}^{h} \gamma^{k-t-1} r(S_k),$$

so that $U_{t:h} = 0$ if $t \geq h$.

Proposition 3.1. If $t \in \mathbb{N}$ and $h \in \mathbb{N}$, then $U_{t:h} \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ and $|U_{t:h}| \leq c/(1-\gamma)$.

Proof. The function $r(S_k)$ is bounded and $\mathcal{F}$-measurable for every $k \in \mathbb{N}$, so that $r(S_k) \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$. Since $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ is a vector space over the field $\mathbb{R}$, $U_{t:h} \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$. For $t < h$,

$$|U_{t:h}| \leq \sum_{k=t+1}^{h} \gamma^{k-t-1} |r(S_k)| \leq c \sum_{k=0}^{h-t-1} \gamma^k = c\left(\frac{1-\gamma^{h-t}}{1-\gamma}\right) \leq \frac{c}{1-\gamma}.$$

□

Proposition 3.2. For every $t, h', h \in \mathbb{N}$ such that $t \leq h' < h$, the discounted return $U_{t:h}$ is given by

$$U_{t:h} = U_{t:h'} + \gamma^{h'-t} U_{h':h}.$$

Proof. For every $t, h', h \in \mathbb{N}$ such that $t \leq h' < h$,

$$U_{t:h} = \sum_{k=t+1}^{h'} \gamma^{k-t-1} r(S_k) + \sum_{k=h'+1}^{h} \gamma^{k-t-1} r(S_k) = U_{t:h'} + \sum_{k=h'+1}^{h} \gamma^{k-t-1} r(S_k).$$

Because $\gamma^{h'}\gamma^{-h'} = 1$ for every $h' \in \mathbb{N}$,

$$U_{t:h} = U_{t:h'} + \gamma^{h'}\gamma^{-h'} \sum_{k=h'+1}^{h} \gamma^{k-1}\gamma^{-t} r(S_k) = U_{t:h'} + \gamma^{h'-t} \sum_{k=h'+1}^{h} \gamma^{k-h'-1} r(S_k).$$

□

Proposition 3.3. If $\omega \in \Omega$ and $t \in \mathbb{N}$, then $(U_{t:h}(\omega) \mid h \in \mathbb{N})$ is a Cauchy sequence.

Proof. For every $t, h', h \in \mathbb{N}$ such that $t \leq h' < h$,

$$|U_{t:h} - U_{t:h'}| = \left|U_{t:h'} + \gamma^{h'-t} U_{h':h} - U_{t:h'}\right| = \gamma^{h'-t}|U_{h':h}| \leq \gamma^{h'-t}\frac{c}{1-\gamma}.$$

Therefore, for every $t, h' \in \mathbb{N}$ such that $t \leq h'$,

$$0 \leq \sup_{h > h'} |U_{t:h} - U_{t:h'}| \leq \left(\frac{c\gamma^{-t}}{1-\gamma}\right)\gamma^{h'}.$$

By the squeeze theorem, for every $t \in \mathbb{N}$,

$$\lim_{h' \to \infty} \sup_{h > h'} |U_{t:h} - U_{t:h'}| = 0.$$

Therefore, for every $t \in \mathbb{N}$ and $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $h, h' > N$ implies $|U_{t:h} - U_{t:h'}| < \epsilon$. □

Definition 3.2. The discounted return $U_{t:\infty}$ after time step $t \in \mathbb{N}$ is defined by

$$U_{t:\infty} = \lim_{h \to \infty} U_{t:h} = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} r(S_k).$$

Proposition 3.4. If $t \in \mathbb{N}$, then $U_{t:\infty} \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ and $\mathbb{E}^{\mu,\pi}(U_{t:\infty}) = \lim_{h \to \infty} \mathbb{E}^{\mu,\pi}(U_{t:h})$.

Proof. For every $\omega \in \Omega$, recall that the Cauchy sequence $(U_{t:h}(\omega) \mid h \in \mathbb{N})$ converges to a real number, so that $U_{t:\infty}$ is well-defined and $\mathcal{F}$-measurable. By the dominated convergence theorem, $U_{t:\infty} \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}^{\mu,\pi})$ and

$$\mathbb{E}^{\mu,\pi}(U_{t:\infty}) = \lim_{h\to\infty} \mathbb{E}^{\mu,\pi}(U_{t:h}).$$

Additionally, because the absolute value is continuous,

$$|U_{t:\infty}| = \lim_{h\to\infty} |U_{t:h}| \leq \frac{c}{1-\gamma}.$$

□

Proposition 3.5. For every $t, h' \in \mathbb{N}$ such that $t \leq h'$, the discounted return $U_{t:\infty}$ is given by

$$U_{t:\infty} = U_{t:h'} + \gamma^{h'-t} U_{h':\infty}.$$

Proof. For every t, h' such that $t \leq h'$,

$$U_{t:\infty} = \lim_{h\to\infty} U_{t:h} = \lim_{h\to\infty} U_{t:h'} + \gamma^{h'-t} U_{h':h} = U_{t:h'} + \gamma^{h'-t} U_{h':\infty}.$$

□

4 Optimal Adaptive Policies

Definition 4.1. An adaptive policy π is optimal up to the horizon $h \in \mathbb{N} \cup \{\infty\}$ under the initial distribution μ if

$$\mathbb{E}^{\mu,\pi}\left(U_{0:h}\right) = \sup_{\pi'} \mathbb{E}^{\mu,\pi'}\left(U_{0:h}\right).$$

Proposition 4.1. Under an initial distribution μ, suppose that the adaptive policy π' is optimal up to the horizon $h' \in \mathbb{N}$ and that the adaptive policy π is optimal up to the horizon $h \in \mathbb{N}^+ \cup \{\infty\}$. If $h' < h$, then

$$0 \leq \mathbb{E}^{\mu,\pi}(U_{0:h}) - \mathbb{E}^{\mu,\pi'}(U_{0:h}) \leq 2c\left(\frac{\gamma^{h'} - \gamma^{h}}{1-\gamma}\right),$$

where γ^∞ is used to denote zero.

Proof. Because $\mathbb{E}^{\mu,\pi}(U_{0:h}) \geq \mathbb{E}^{\mu,\pi'}(U_{0:h})$, we know that $\mathbb{E}^{\mu,\pi}(U_{0:h}) - \mathbb{E}^{\mu,\pi'}(U_{0:h}) \geq 0$. By Propositions 3.2 and 3.5,

$$0 \leq \mathbb{E}^{\mu,\pi}(U_{0:h}) - \mathbb{E}^{\mu,\pi'}(U_{0:h}) = \mathbb{E}^{\mu,\pi}\left(U_{0:h'}\right) + \gamma^{h'}\mathbb{E}^{\mu,\pi}\left(U_{h':h}\right) - \mathbb{E}^{\mu,\pi'}\left(U_{0:h'}\right) - \gamma^{h'}\mathbb{E}^{\mu,\pi'}\left(U_{h':h}\right).$$

Because $\mathbb{E}^{\mu,\pi'}\left(U_{0:h'}\right) \geq \mathbb{E}^{\mu,\pi}\left(U_{0:h'}\right)$, we know that $\mathbb{E}^{\mu,\pi}\left(U_{0:h'}\right) - \mathbb{E}^{\mu,\pi'}\left(U_{0:h'}\right) \leq 0$. Therefore,

$$0 \leq \mathbb{E}^{\mu,\pi}(U_{0:h}) - \mathbb{E}^{\mu,\pi'}(U_{0:h}) \leq \gamma^{h'}\left(\mathbb{E}^{\mu,\pi}\left(U_{h':h}\right) - \mathbb{E}^{\mu,\pi'}\left(U_{h':h}\right)\right).$$

Because $\gamma^{h'} > 0$, we know that $\mathbb{E}^{\mu,\pi}\left(U_{h':h}\right) \geq \mathbb{E}^{\mu,\pi'}\left(U_{h':h}\right)$. From the proofs of Propositions 3.1 and 3.4,

$$-c\left(\frac{1-\gamma^{h-h'}}{1-\gamma}\right) \leq \mathbb{E}^{\mu,\pi'}\left(U_{h':h}\right) \leq \mathbb{E}^{\mu,\pi}\left(U_{h':h}\right) \leq c\left(\frac{1-\gamma^{h-h'}}{1-\gamma}\right),$$

where γ^∞ is used to denote zero. By subtracting the leftmost term above from the rightmost term above,

$$0 \leq \mathbb{E}^{\mu,\pi}(U_{0:h}) - \mathbb{E}^{\mu,\pi'}(U_{0:h}) \leq \gamma^{h'} 2c\left(\frac{1-\gamma^{h-h'}}{1-\gamma}\right).$$

□

Theorem 4.1 (Regret of truncated planning)**.** Suppose that the adaptive policy π is optimal up to the horizon ∞ under the initial distribution μ. For every $\epsilon > 0$ and $h' \in \mathbb{N}$ such that $h' > \log(\epsilon(1-\gamma)/2c)/\log(\gamma)$, if the adaptive policy π' is optimal up to the horizon h' under the initial distribution μ, then $\mathbb{E}^{\mu,\pi}(U_{0:\infty}) - \mathbb{E}^{\mu,\pi'}(U_{0:\infty}) < \epsilon$.

Proof. Proposition 4.1 ensures that $\mathbb{E}^{\mu,\pi}(U_{0:\infty}) - \mathbb{E}^{\mu,\pi'}(U_{0:\infty}) \leq 2c\gamma^{h'}/(1-\gamma) < \epsilon$. □

5 Policy Values

Definition 5.1. For every $t \in \mathbb{N}$ and $h \in \mathbb{N} \cup \{\infty\}$, the value $V^{\mu,\pi}_{t:h} : \Omega \to \mathbb{R}$ of time t up to the horizon h under the initial distribution μ and the adaptive policy π is defined such that, almost surely,

$$V^{\mu,\pi}_{t:h} = \mathbb{E}^{\mu,\pi}\left(U_{t:h} \mid \mathcal{H}_t\right) = \mathbb{E}^{\mu,\pi}\left(\sum_{k=t+1}^{h} \gamma^{k-t-1} r(S_k) \mid \mathcal{H}_t\right),$$

so that $V^{\mu,\pi}_{t:t+1} = \mathbb{E}^{\mu,\pi}(U_{t:t+1} \mid \mathcal{H}_t) = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t)$ almost surely.

Proposition 5.1. If $t \in \mathbb{N}$ and $h \in \mathbb{N} \cup \{\infty\}$, then $|V^{\mu,\pi}_{t,h}| \leq c/(1-\gamma)$ almost surely.

Proof. If $t \in \mathbb{N}$ and $h \in \mathbb{N} \cup \{\infty\}$, then $|U_{t:h}| \leq c/(1-\gamma)$. Therefore, almost surely,

$$|V^{\mu,\pi}_{t,h}| = |\mathbb{E}^{\mu,\pi}(U_{t:h} \mid \mathcal{H}_t)| \leq \mathbb{E}^{\mu,\pi}(|U_{t:h}| \mid \mathcal{H}_t) \leq \frac{c}{(1-\gamma)}.$$

□

Theorem 5.1 (Bellman equation)**.** For every $t \in \mathbb{N}$ and $h \in \mathbb{N}^+ \cup \{\infty\}$ such that $t+1 < h$, the value $V^{\mu,\pi}_{t:h}$ of time t up to the horizon h under the initial distribution μ and the adaptive policy π is almost surely given by

$$V^{\mu,\pi}_{t:h} = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) + \gamma \mathbb{E}^{\mu,\pi}\left(V^{\mu,\pi}_{t+1:h} \mid \mathcal{H}_t\right).$$

Proof. By the linearity of conditional expectation, almost surely,

$$V^{\mu,\pi}_{t:h} = \mathbb{E}^{\mu,\pi}\left(U_{t:t+1} + \gamma U_{t+1:h} \mid \mathcal{H}_t\right) = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) + \gamma \mathbb{E}^{\mu,\pi}\left(U_{t+1:h} \mid \mathcal{H}_t\right).$$

By the tower property, almost surely,

$$V^{\mu,\pi}_{t:h} = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) + \gamma \mathbb{E}^{\mu,\pi}\left(\mathbb{E}^{\mu,\pi}(U_{t+1:h} \mid \mathcal{H}_{t+1}) \mid \mathcal{H}_t\right).$$

□

Proposition 5.2. For every $t \in \mathbb{N}$, almost surely,

$$\mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) = \sum_{s_{t+1}} r(s_{t+1}) \mathbb{P}^{\mu,\pi}\left(S_{t+1} = s_{t+1} \mid \mathcal{H}_t\right).$$

Proof. For every $n \in \mathbb{N}$, let $X_n : \Omega \to \mathbb{R}$ be given by

$$X_n(\omega) = \sum_{s_{t+1} < n} r(s_{t+1}) \mathbb{I}_{\{S_{t+1} = s_{t+1}\}}(\omega) = \begin{cases} r(S_{t+1}(\omega)), & \text{if } S_{t+1}(\omega) < n, \\ 0, & \text{if } S_{t+1}(\omega) \geq n, \end{cases}$$

so that $r(S_{t+1}) = \lim_{n\to\infty} X_n$. By the conditional dominated convergence theorem, almost surely,

$$\mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) = \lim_{n\to\infty} \mathbb{E}^{\mu,\pi}(X_n \mid \mathcal{H}_t) = \lim_{n\to\infty} \sum_{s_{t+1} < n} r(s_{t+1}) \mathbb{E}^{\mu,\pi}\left(\mathbb{I}_{\{S_{t+1} = s_{t+1}\}} \mid \mathcal{H}_t\right).$$

□

Definition 5.2. For every $t \in \mathbb{N}$ and $h \in \mathbb{N}$ such that $t < h$, the function $v^{\mu,\pi}_{t:h} : \mathcal{S}^{t+1} \to \mathbb{R}$ is given by

$$v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t) = \sum_{s_{t+1}} \rho^{\mu,a_0,\ldots,a_t}_{s_0,\ldots,s_t}(s_{t+1}) \left(r(s_{t+1}) + \gamma v^{\mu,\pi}_{t+1:h}(s_0, \ldots, s_t, s_{t+1})\right),$$

where $a_k = \pi_k(s_0, \ldots, s_k)$ for every $k \leq t$. If $t \geq h$, let $v^{\mu,\pi}_{t:h} = 0$.

Proposition 5.3. If $t \in \mathbb{N}$ and $h \in \mathbb{N}$, then $|v^{\mu,\pi}_{t:h}| \leq c/(1-\gamma)$.

Proof. If $t \geq h$, then $|v^{\mu,\pi}_{t:h}| \leq c/(1-\gamma)$. If $t < h$, in order to employ backward induction, suppose that $|v^{\mu,\pi}_{t+1:h}| \leq c/(1-\gamma)$. In that case, for every $(s_0, \dots, s_{t+1}) \in \mathcal{S}^{t+2}$,

$$|r(s_{t+1}) + \gamma v^{\mu,\pi}_{t+1:h}(s_0, \dots, s_t, s_{t+1})| \leq |r(s_{t+1})| + \gamma |v^{\mu,\pi}_{t+1:h}(s_0, \dots, s_t, s_{t+1})| \leq c + \gamma \frac{c}{1-\gamma} = \frac{c}{1-\gamma},$$

so that $|v^{\mu,\pi}_{t:h}(s_0, \dots, s_t)| \leq c/(1-\gamma) \sum_{s_{t+1}} \rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1}) = c/(1-\gamma)$. □

Proposition 5.4. If $t \in \mathbb{N}$ and $h \in \mathbb{N}$, then $v^{\mu,\pi}_{t:h}(S_0, \dots, S_t) = V^{\mu,\pi}_{t:h}$ almost surely.

Proof. If $t \geq h$, then $v^{\mu,\pi}_{t:h}(S_0, \dots, S_t) = 0 = V^{\mu,\pi}_{t:h}$ almost surely. If $t = h-1$, by Propositions 2.4 and 5.2,

$$v^{\mu,\pi}_{t:h}(S_0, \dots, S_t) = \sum_{s_{t+1}} \rho^{\mu,A_0,\dots,A_t}_{S_0,\dots,S_t}(s_{t+1}) r(s_{t+1}) = \sum_{s_{t+1}} r(s_{t+1}) \mathbb{P}^{\mu,\pi}\left(S_{t+1} = s_{t+1} \mid \mathcal{H}_t\right) = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) = V^{\mu,\pi}_{t:h}$$

almost surely, where $A_k = \pi_k(S_0, \dots, S_k)$ for every $k \leq t$. If $t < h-1$, in order to employ backward induction, suppose that $v^{\mu,\pi}_{t+1:h}(S_0, \dots, S_{t+1}) = V^{\mu,\pi}_{t+1:h}$ almost surely. For every $n \in \mathbb{N}$, let $X_n : \Omega \to \mathbb{R}$ be given by

$$X_n(\omega) = \sum_{s_{t+1} < n} v^{\mu,\pi}_{t+1:h}(S_0(\omega), \dots, S_t(\omega), s_{t+1}) \mathbb{I}_{\{S_{t+1} = s_{t+1}\}}(\omega) = \begin{cases} v^{\mu,\pi}_{t+1:h}(S_0(\omega), \dots, S_{t+1}(\omega)), & \text{if } S_{t+1}(\omega) < n, \\ 0, & \text{if } S_{t+1}(\omega) \geq n, \end{cases}$$

so that $V^{\mu,\pi}_{t+1:h} = v^{\mu,\pi}_{t+1:h}(S_0, \dots, S_{t+1}) = \lim_{n \to \infty} X_n$ almost surely. By conditional dominated convergence,

$$\mathbb{E}^{\mu,\pi}(V^{\mu,\pi}_{t+1:h} \mid \mathcal{H}_t) = \lim_{n \to \infty} \sum_{s_{t+1} < n} v^{\mu,\pi}_{t+1:h}(S_0, \dots, S_t, s_{t+1}) \mathbb{E}^{\mu,\pi}\left(\mathbb{I}_{\{S_{t+1} = s_{t+1}\}} \mid \mathcal{H}_t\right)$$

almost surely, where we used the fact that $v^{\mu,\pi}_{t+1:h}(S_0, \dots, S_t, s_{t+1})$ is $\mathcal{H}_t$-measurable to take out what is known.

From the definition of $v^{\mu,\pi}_{t:h}$ and Proposition 2.4, almost surely,

$$v^{\mu,\pi}_{t:h}(S_0, \dots, S_t) = \sum_{s_{t+1}} \mathbb{P}^{\mu,\pi}\left(S_{t+1} = s_{t+1} \mid \mathcal{H}_t\right) r(s_{t+1}) + \gamma \sum_{s_{t+1}} \mathbb{P}^{\mu,\pi}\left(S_{t+1} = s_{t+1} \mid \mathcal{H}_t\right) v^{\mu,\pi}_{t+1:h}(S_0, \dots, S_t, s_{t+1}).$$

Almost surely, by Proposition 5.2 and Theorem 5.1,

$$v^{\mu,\pi}_{t:h}(S_0, \dots, S_t) = \mathbb{E}^{\mu,\pi}(r(S_{t+1}) \mid \mathcal{H}_t) + \gamma \mathbb{E}^{\mu,\pi}\left(V^{\mu,\pi}_{t+1:h} \mid \mathcal{H}_t\right) = V^{\mu,\pi}_{t:h}.$$

□

Theorem 5.2 (Value of an adaptive policy)**.** For every initial distribution μ, adaptive policy π, and horizon $h \in \mathbb{N}$,

$$\mathbb{E}^{\mu,\pi}(U_{0:h}) = \mathbb{E}^{\mu,\pi}(V^{\mu,\pi}_{0:h}) = \mathbb{E}^{\mu,\pi}(v^{\mu,\pi}_{0:h}(S_0)) = \sum_{s_0} \mu_{s_0} v^{\mu,\pi}_{0:h}(s_0).$$

The last result may enable evaluating an adaptive policy up to a finite horizon.

6 Optimal Policy Values

Definition 6.1. For every $t \in \mathbb{N}$ and $h \in \mathbb{N}$ such that $t < h$, the function $v^{\mu,*}_{t:h} : \mathcal{S} \times (\mathcal{A} \times \mathcal{S})^t \to \mathbb{R}$ is given by

$$v^{\mu,*}_{t:h}(s_0, a_0, s_1, \dots, a_{t-1}, s_t) = \sup_{a_t} \sum_{s_{t+1}} \rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1}) \left(r(s_{t+1}) + \gamma v^{\mu,*}_{t+1:h}(s_0, a_0, s_1, \dots, a_t, s_{t+1})\right).$$

If $t \geq h$, let $v^{\mu,*}_{t:h} = 0$.

Definition 6.2. For every $t \in \mathbb{N}$ and $h \in \mathbb{N}$ such that $t < h$, the function $q^{\mu,*}_{t:h} : (\mathcal{S} \times \mathcal{A})^{t+1} \to \mathbb{R}$ is given by

$$q^{\mu,*}_{t:h}(s_0, a_0, \dots, s_t, a_t) = \sum_{s_{t+1}} \rho^{\mu,a_0,\dots,a_t}_{s_0,\dots,s_t}(s_{t+1}) \left(r(s_{t+1}) + \gamma v^{\mu,*}_{t+1:h}(s_0, a_0, s_1, \dots, a_t, s_{t+1})\right).$$

If $t \geq h$, let $q^{\mu,*}_{t:h} = 0$. Note that $v^{\mu,*}_{t:h}(s_0, a_0, s_1, \dots, a_{t-1}, s_t) = \sup_a q^{\mu,*}_{t:h}(s_0, a_0, \dots, s_t, a)$.

Proposition 6.1. For every adaptive policy π, $t \in \mathbb{N}$, $h \in \mathbb{N}$, and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$,

$$v^{\mu,*}_{t:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t) \geq v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t),$$

where $a_k = \pi_k(s_0, \ldots, s_k)$ for every $k < t$.

Proof. If $t \geq h$, then $v^{\mu,*}_{t:h} = 0$ and $v^{\mu,\pi}_{t:h} = 0$. If $t < h$, in order to employ backward induction, suppose that

$$v^{\mu,*}_{t+1:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t, a_t, s_{t+1}) \geq v^{\mu,\pi}_{t+1:h}(s_0, \ldots, s_t, s_{t+1})$$

for every $(s_0, \ldots, s_{t+1}) \in \mathcal{S}^{t+2}$, where $a_k = \pi_k(s_0, \ldots, s_k)$ for every $k < t+1$. In that case,

$$v^{\mu,*}_{t:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t) = \sup_a q^{\mu,*}_{t:h}(s_0, a_0, \ldots, s_t, a) \geq q^{\mu,*}_{t:h}(s_0, a_0, \ldots, s_t, a_t).$$

By the inductive hypothesis,

$$v^{\mu,*}_{t:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t) \geq \sum_{s_{t+1}} \rho^{\mu, a_0, \ldots, a_t}_{s_0, \ldots, s_t}(s_{t+1}) \left(r(s_{t+1}) + \gamma v^{\mu,\pi}_{t+1:h}(s_0, \ldots, s_t, s_{t+1}) \right) = v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t).$$

□

Theorem 6.1 (Value of an optimal adaptive policy)**.** If $h \in \mathbb{N}$ and $v^{\mu,\pi}_{0:h}(S_0) = v^{\mu,*}_{0:h}(S_0)$ almost surely, then π is optimal up to the horizon h under the initial distribution μ.

Proof. For every adaptive policy π', using Proposition 6.1 and the fact that $\mathbb{P}^{\mu,\pi}$ and $\mathbb{P}^{\mu,\pi'}$ agree on $\mathcal{H}_0$,

$$\mathbb{E}^{\mu,\pi}(U_{0:h}) = \mathbb{E}^{\mu,\pi}(v^{\mu,\pi}_{0:h}(S_0)) = \mathbb{E}^{\mu,\pi}(v^{\mu,*}_{0:h}(S_0)) = \mathbb{E}^{\mu,\pi'}(v^{\mu,*}_{0:h}(S_0)) \geq \mathbb{E}^{\mu,\pi'}(v^{\mu,\pi'}_{0:h}(S_0)) = \mathbb{E}^{\mu,\pi'}(U_{0:h}).$$

□

Theorem 6.2 (Existence of an optimal adaptive policy)**.** Under every initial distribution μ, for every $h \in \mathbb{N}$, if the set of actions $\mathcal{A}$ is finite, then there is an adaptive policy that is optimal up to the horizon h.

Proof. Consider an adaptive policy $\pi = (\pi_t \mid t \in \mathbb{N})$ such that, for every $t \in \mathbb{N}$ and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$,

$$q^{\mu,*}_{t:h}(s_0, \pi_0(s_0), \ldots, s_t, \pi_t(s_0, \ldots, s_t)) = \sup_a q^{\mu,*}_{t:h}(s_0, \pi_0(s_0), \ldots, s_t, a),$$

which exists because the set of actions $\mathcal{A}$ is finite.

For every $t \in \mathbb{N}$ and $(s_0, \ldots, s_t) \in \mathcal{S}^{t+1}$, we will show that $v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t) = v^{\mu,*}_{t:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t)$, where $a_k = \pi_k(s_0, \ldots, s_k)$ for every $k < t$.

If $t \geq h$, then $v^{\mu,\pi}_{t:h} = 0$ and $v^{\mu,*}_{t:h} = 0$. If $t < h$, in order to employ backward induction, suppose that

$$v^{\mu,\pi}_{t+1:h}(s_0, \ldots, s_t, s_{t+1}) = v^{\mu,*}_{t+1:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t, a_t, s_{t+1})$$

for every $(s_0, \ldots, s_{t+1}) \in \mathcal{S}^{t+2}$, where $a_k = \pi_k(s_0, \ldots, s_k)$ for every $k < t+1$. By the inductive hypothesis,

$$v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t) = \sum_{s_{t+1}} \rho^{\mu, a_0, \ldots, a_t}_{s_0, \ldots, s_t}(s_{t+1}) \left(r(s_{t+1}) + \gamma v^{\mu,*}_{t+1:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t, a_t, s_{t+1}) \right).$$

By the definition of the adaptive policy π,

$$v^{\mu,\pi}_{t:h}(s_0, \ldots, s_t) = q^{\mu,*}_{t:h}(s_0, a_0, \ldots, s_t, a_t) = \sup_a q^{\mu,*}_{t:h}(s_0, a_0, \ldots, s_t, a) = v^{\mu,*}_{t:h}(s_0, a_0, s_1, \ldots, a_{t-1}, s_t).$$

Because $v^{\mu,\pi}_{0:h}(S_0) = v^{\mu,*}_{0:h}(S_0)$, π is optimal up to the horizon h under the initial distribution μ. □

The last result may enable finding an optimal adaptive policy up to a finite horizon given a finite set of actions.

7 Examples

7.1 Countable Bayes-adaptive Markov decision processes

Definition 7.1. A countable Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$ is composed of:

- A set of states $\mathcal{S}$;
- A set of actions $\mathcal{A}$;
- A countable non-empty set of models $\mathcal{M}$ over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$;
- A prior ψ, which is a probability measure on the canonical space $(\mathcal{M}, \mathcal{G})$ for the set of models $\mathcal{M}$;
- A reward function $r : \mathcal{S} \to \mathbb{R}$ such that $|r| \leq c$ for some $c \in (0, \infty)$;
- A discount factor $\gamma \in (0, 1)$.

For every model $p \in \mathcal{M}$, let $\psi(\{p\}) = \psi_p$.

Consider a countable Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$.

Proposition 7.1. For every $t \in \mathbb{N}$, sequence of states $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, and sequence of actions $(a_0, \dots, a_t) \in \mathcal{A}^{t+1}$, the posterior predictive $\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t} : \mathcal{S} \to [0, 1]$ given $(s_0, \dots, s_t)$ and $(a_0, \dots, a_t)$ under μ is given by

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \frac{\sum_p \psi_p \prod_{k=1}^{t+1} p^{a_{k-1}}_{s_{k-1}, s_k}}{\sum_p \psi_p \prod_{k=1}^{t} p^{a_{k-1}}_{s_{k-1}, s_k}}$$

whenever $\mu_{s_0} \sum_p \psi_p \prod_{k=1}^{t} p^{a_{k-1}}_{s_{k-1}, s_k} \neq 0$.

Proof. Let $\pi = (\pi_t \mid t \in \mathbb{N})$ be an adaptive policy such that $\pi_k = a_k$ for every $k \leq t$.

Since $\{M \in \mathcal{M}\} = \Omega$, for every $(s_0, \dots, s_{t+1}) \in \mathcal{S}^{t+2}$ and $t' \leq t + 1$,

$$\mathbb{P}^{\mu, \pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \int_{\mathcal{M}} \mu_{s_0} \prod_{k=1}^{t'} p^{a_{k-1}}_{s_{k-1}, s_k} \ \psi(dp) = \mu_{s_0} \sum_p \psi_p \prod_{k=1}^{t'} p^{a_{k-1}}_{s_{k-1}, s_k}.$$

Whenever $\mathbb{P}^{\mu, \pi}(S_0 = s_0, \dots, S_t = s_t) = \mu_{s_0} \sum_p \psi_p \prod_{k=1}^{t} p^{a_{k-1}}_{s_{k-1}, s_k} \neq 0$,

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \rho^{\mu, \pi}_{s_0, \dots, s_t}(s_{t+1}) = \frac{\mathbb{P}^{\mu, \pi}(S_0 = s_0, \dots, S_t = s_t, S_{t+1} = s_{t+1})}{\mathbb{P}^{\mu, \pi}(S_0 = s_0, \dots, S_t = s_t)} = \frac{\sum_p \psi_p \prod_{k=1}^{t+1} p^{a_{k-1}}_{s_{k-1}, s_k}}{\sum_p \psi_p \prod_{k=1}^{t} p^{a_{k-1}}_{s_{k-1}, s_k}}.$$

In particular, if $\psi_p = 1$ for some $p \in \mathcal{M}$, then $\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = p^{a_t}_{s_t, s_{t+1}}$. □

7.2 Dirichlet Bayes-adaptive Markov decision processes

Definition 7.2. The gamma function $\Gamma : (0, \infty) \to (0, \infty)$ is given by

$$\Gamma(a) = \int_{(0, \infty)} b^{a-1} e^{-b} \ \mathrm{Leb}(db),$$

where e is Euler's number. Remarkably, $a = \Gamma(a+1)/\Gamma(a)$ for every $a \in (0, \infty)$.

Definition 7.3. For every $n - 1 \in \mathbb{N}^+$, the simplex C^{n-1} is given by $C^{n-1} = \{\theta \in (0, 1)^{n-1} \mid \sum_{i=1}^{n-1} \theta_i < 1\}$.

Definition 7.4. For every $n - 1 \in \mathbb{N}^+$, the multivariate Beta function $B : (0, \infty)^n \to (0, \infty)$ is given by

$$B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^n \alpha_i\right)} = \int_{C^{n-1}} \prod_{i=1}^n \theta_i^{\alpha_i - 1} \, \mathrm{Leb}^{n-1}(d\theta),$$

where $\theta_n = 1 - \sum_{i=1}^{n-1} \theta_i$.

Definition 7.5. For every $n - 1 \in \mathbb{N}^+$, the joint probability density function $\text{Dir}(\cdot; \alpha) : \mathbb{R}^{n-1} \to [0, \infty]$ is given by

$$\text{Dir}(\theta; \alpha) = \mathbb{I}_{C^{n-1}}(\theta) \frac{1}{B(\alpha)} \prod_{i=1}^{n} \theta_i^{\alpha_i - 1},$$

where $\alpha \in (0, \infty)^n$ is a so-called pseudocount and $\theta_n = 1 - \sum_{i=1}^{n-1} \theta_i$.

Definition 7.6. For every $n - 1 \in \mathbb{N}^+$, the simplex space $(C^{n-1}, \mathcal{C}^{n-1})$ is given by restricting the measurable space $(\mathbb{R}^{n-1}, \mathcal{B}(\mathbb{R}^{n-1}))$ to the simplex C^{n-1}, so that $\mathcal{C}^{n-1} = \{B \in \mathcal{B}(\mathbb{R}^{n-1}) \mid B \subseteq C^{n-1}\}$.

Definition 7.7. A Dirichlet law $\mathcal{L} : \mathcal{C}^{n-1} \to [0, 1]$ on the simplex space $(C^{n-1}, \mathcal{C}^{n-1})$ is given by $\mathcal{L}(\Theta) = \mathcal{L}^*(\Theta)$, where $\mathcal{L}^* : \mathcal{B}(\mathbb{R}^{n-1}) \to [0, 1]$ is a probability measure on $(\mathbb{R}^{n-1}, \mathcal{B}(\mathbb{R}^{n-1}))$ such that, for some $\alpha \in (0, \infty)^n$,

$$\mathcal{L}^*(\Theta) = \int_{\Theta} \text{Dir}(\theta; \alpha) \ \text{Leb}^{n-1}(d\theta).$$

Definition 7.8. Let $\mathcal{M}$ be a set of models over the set of states $\mathcal{S} = \{1, 2, \dots, n\}$ and the set of actions $\mathcal{A}$. For every state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, the function $q_s^a : \mathcal{M} \to [0, 1]^{n-1}$ is given by

$$q_s^a(p) = (q_{s,1}^a(p), \dots, q_{s,n-1}^a(p)) = (p_{s,1}^a, \dots, p_{s,n-1}^a).$$

Definition 7.9. The set of positive models $\mathcal{M}$ over $\mathcal{S} = \{1, 2, \dots, n\}$ and $\mathcal{A}$ is given by

$$\mathcal{M} = \{p \in \mathcal{M}^* \mid p_{s,s'}^a > 0 \text{ for every } (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}\} = \{p \in \mathcal{M}^* \mid q_s^a(p) \in C^{n-1} \text{ for every } (s, a) \in \mathcal{S} \times \mathcal{A}\},$$

where $\mathcal{M}^*$ is the set of all models over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$.

Proposition 7.2. For some $n - 1 \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$, let $\mathcal{M}$ be the set of positive models over the set of states $\mathcal{S} = \{1, 2, \dots, n\}$ and the set of actions $\mathcal{A} = \{1, \dots, m\}$. For a given choice of pseudocounts $(\alpha_s^a \in (0, \infty)^n \mid (s, a) \in \mathcal{S} \times \mathcal{A})$, there is a unique probability measure ψ on the canonical space $(\mathcal{M}, \mathcal{G})$ for the set of models $\mathcal{M}$ such that

$$\psi\left(\bigcap_{(s,a)} \{q_s^a \in \Theta_s^a\}\right) = \prod_{(s,a)} \int_{\Theta_s^a} \text{Dir}(\theta_s^a; \alpha_s^a) \ \text{Leb}^{n-1}(d\theta_s^a)$$

for every sequence $(\Theta_s^a \in \mathcal{C}^{n-1} \mid (s, a) \in \mathcal{S} \times \mathcal{A})$. The probability measure ψ is called a Dirichlet prior on the canonical space $(\mathcal{M}, \mathcal{G})$ given the pseudocounts $(\alpha_s^a \mid (s, a) \in \mathcal{S} \times \mathcal{A})$.

Proof. For every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, consider the Dirichlet law $\mathcal{L}_s^a$ on the simplex space $(C^{n-1}, \mathcal{C}^{n-1})$ given by

$$\mathcal{L}_s^a(\Theta_s^a) = \int_{\Theta_s^a} \text{Dir}(\theta_s^a; \alpha_s^a) \ \text{Leb}^{n-1}(d\theta_s^a).$$

Furthermore, consider the product measure $\mathcal{L}$ on the measurable space $((C^{n-1})^{mn}, (\mathcal{C}^{n-1})^{mn})$ given by

$$\mathcal{L} = \mathcal{L}_1^1 \times \dots \times \mathcal{L}_1^m \times \mathcal{L}_2^1 \times \dots \times \mathcal{L}_2^m \times \ldots \times \mathcal{L}_n^1 \times \ldots \times \mathcal{L}_n^m.$$

Consider the invertible function $q : \mathcal{M} \to (C^{n-1})^{mn}$ given by

$$q(p) = (q_1^1(p), \dots, q_1^m(p), q_2^1(p), \dots, q_2^m(p), \dots, q_n^1(p), \dots, q_n^m(p)),$$

and let $u : (C^{n-1})^{mn} \to \mathcal{M}$ denote the inverse of q. Clearly, $\sigma(q) \subseteq \mathcal{G}$. Furthermore, $\mathcal{G} \subseteq \sigma(q)$, which relies on the fact that $\sigma(q_{s,s'}^a) \subseteq \sigma(q)$ for every $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $s' \in \mathcal{S}$. In particular, note that $q_{s,n}^a = 1 - \sum_{s' < n} q_{s,s'}^a$. Since q is invertible and $\sigma(q) = \mathcal{G}$, recall that $\sigma(u) = (\mathcal{C}^{n-1})^{mn}$.

Therefore, the function $\psi : \mathcal{G} \to [0, 1]$ given by $\psi(G) = \mathcal{L}(u^{-1}(G))$ is a probability measure on the canonical space $(\mathcal{M}, \mathcal{G})$. For every sequence $(\Theta_s^a \in \mathcal{C}^{n-1} \mid (s, a) \in \mathcal{S} \times \mathcal{A})$,

$$\psi\left(\bigcap_{(s,a)} \{q_s^a \in \Theta_s^a\}\right) = \mathcal{L}\left(\prod_{(s,a)} \Theta_s^a\right) = \prod_{(s,a)} \mathcal{L}_s^a(\Theta_s^a) = \prod_{(s,a)} \int_{\Theta_s^a} \text{Dir}(\theta_s^a; \alpha_s^a) \ \text{Leb}^{n-1}(d\theta_s^a).$$

Since $\mathcal{L}_s^a(\Theta_s^a) = \psi\left(q_s^a \in \Theta_s^a\right)$, note that $(\sigma(q_s^a) \mid (s, a) \in \mathcal{S} \times \mathcal{A})$ are independent.

Because $\mathcal{I} = \{\bigcap_{(s,a)} \{q_s^a \in \Theta_s^a\} \mid \Theta_s^a \in \mathcal{C}^{n-1} \text{ for every } (s, a) \in \mathcal{S} \times \mathcal{A}\}$ is a π-system on $\mathcal{M}$ such that $\sigma(\mathcal{I}) = \mathcal{G}$, ψ is the unique probability measure on the canonical space $(\mathcal{M}, \mathcal{G})$ with the desired properties. □

Definition 7.10. A Dirichlet Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$ is composed of:

- A set of states $\mathcal{S} = \{1, 2, \dots, n\}$, where $n - 1 \in \mathbb{N}^+$;
- A set of actions $\mathcal{A} = \{1, \dots, m\}$, where $m \in \mathbb{N}^+$;
- The set of positive models $\mathcal{M}$ over the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$;
- A Dirichlet prior ψ on the canonical space $(\mathcal{M}, \mathcal{G})$ given the pseudocounts $(\alpha_s^a \in (0, \infty)^n \mid (s, a) \in \mathcal{S} \times \mathcal{A})$;
- A reward function $r : \mathcal{S} \to \mathbb{R}$ such that $|r| \leq c$ for some $c \in (0, \infty)$;
- A discount factor $\gamma \in (0, 1)$.

Consider a Dirichlet Bayes-adaptive Markov decision process $(\mathcal{S}, \mathcal{A}, \mathcal{M}, \psi, r, \gamma)$.

Definition 7.11. For every $t \in \mathbb{N}$, $N^a_{s,s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)$ denotes the number of times that the triple (s, a, s') appears in the sequence $s_0, a_0, s_1, \dots, a_{t-1}, s_t \in \mathcal{S} \times (\mathcal{A} \times \mathcal{S})^t$ and $N^a_s(s_0, a_0, s_1, \dots, a_{t-1}, s_t) \in \mathbb{N}^n$ is given by

$$N^a_s(s_0, a_0, s_1, \dots, a_{t-1}, s_t) = (N^a_{s,1}(s_0, a_0, s_1, \dots, a_{t-1}, s_t), \dots, N^a_{s,n}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)).$$

Proposition 7.3. For every $t \in \mathbb{N}$, sequence of states $(s_0, \dots, s_t) \in \mathcal{S}^{t+1}$, and sequence of actions $(a_0, \dots, a_t) \in \mathcal{A}^{t+1}$, the posterior predictive $\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t} : \mathcal{S} \to [0, 1]$ given $(s_0, \dots, s_t)$ and $(a_0, \dots, a_t)$ under μ is given by

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \frac{\alpha^{a_t}_{s_t, s_{t+1}} + N^{a_t}_{s_t, s_{t+1}}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)}{\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)}$$

whenever $\mu_{s_0} \neq 0$.

Proof. Let $\pi = (\pi_t \mid t \in \mathbb{N})$ be an adaptive policy such that $\pi_k = a_k$ for every $k \leq t$.

Since $\{M \in \mathcal{M}\} = \Omega$, for every $(s_0, \dots, s_{t+1}) \in \mathcal{S}^{t+2}$ and $t' \leq t + 1$,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mu_{s_0} \int_{\mathcal{M}} \prod_{k=1}^{t'} q^{a_{k-1}}_{s_{k-1}, s_k} \, d\psi = \mu_{s_0} \int_{\mathcal{M}} \prod_{(s,a)} \prod_{s'} \left(q^a_{s,s'}\right)^{N^a_{s,s'}(s_0, a_0, s_1, \dots, a_{t'-1}, s_{t'})} \, d\psi.$$

Because $(\sigma(q^a_s) \mid (s, a) \in \mathcal{S} \times \mathcal{A})$ are independent,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mu_{s_0} \prod_{(s,a)} \int_{\mathcal{M}} \prod_{s'} \left(q^a_{s,s'}\right)^{N^a_{s,s'}(s_0, a_0, s_1, \dots, a_{t'-1}, s_{t'})} \, d\psi.$$

Since $\mathrm{Dir}(\cdot; \alpha^a_s)$ is a joint probability density function for q^a_s,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mu_{s_0} \prod_{(s,a)} \int_{\mathbb{R}^{n-1}} \mathrm{Dir}(\theta^a_s; \alpha^a_s) \prod_{s'} \left(\theta^a_{s,s'}\right)^{N^a_{s,s'}(s_0, a_0, s_1, \dots, a_{t'-1}, s_{t'})} \mathrm{Leb}^{n-1}(d\theta^a_s),$$

where $\theta^a_{s,n} = 1 - \sum_{s' < n} \theta^a_{s,s'}$. Therefore, by the definition of $\mathrm{Dir}(\cdot; \alpha^a_s)$,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mu_{s_0} \prod_{(s,a)} \frac{1}{B(\alpha^a_s)} \int_{C^{n-1}} \prod_{s'} \left(\theta^a_{s,s'}\right)^{N^a_{s,s'}(s_0, a_0, s_1, \dots, a_{t'-1}, s_{t'}) + \alpha^a_{s,s'} - 1} \mathrm{Leb}^{n-1}(d\theta^a_s).$$

From the definition of the multivariate Beta function,

$$\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t'} = s_{t'}) = \mu_{s_0} \prod_{(s,a)} \frac{B(\alpha^a_s + N^a_s(s_0, a_0, s_1, \dots, a_{t'-1}, s_{t'}))}{B(\alpha^a_s)}.$$

Whenever $\mu_{s_0} \neq 0$,

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \rho^{\mu,\pi}_{s_0, \dots, s_t}(s_{t+1}) = \frac{\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_{t+1} = s_{t+1})}{\mathbb{P}^{\mu,\pi}(S_0 = s_0, \dots, S_t = s_t)} = \prod_{(s,a)} \frac{B(\alpha^a_s + N^a_s(s_0, a_0, s_1, \dots, a_t, s_{t+1}))}{B(\alpha^a_s + N^a_s(s_0, a_0, s_1, \dots, a_{t-1}, s_t))}.$$

Note that $N^a_s(s_0, a_0, s_1, \dots, a_t, s_{t+1}) \neq N^a_s(s_0, a_0, s_1, \dots, a_{t-1}, s_t)$ if and only if $s_t = s$ and $a_t = a$. Therefore,

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \frac{B(\alpha^{a_t}_{s_t} + N^{a_t}_{s_t}(s_0, a_0, s_1, \dots, a_t, s_{t+1}))}{B(\alpha^{a_t}_{s_t} + N^{a_t}_{s_t}(s_0, a_0, s_1, \dots, a_{t-1}, s_t))}.$$

From the definition of the multivariate Beta function,

$$\rho^{\mu, a_0, \dots, a_t}_{s_0, \dots, s_t}(s_{t+1}) = \frac{\prod_{s'} \Gamma(\alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1}))}{\Gamma\left(\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1})\right)} \frac{\Gamma\left(\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)\right)}{\prod_{s'} \Gamma(\alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t))}.$$

Since $N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1}) \neq N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)$ if and only if $s' = s_{t+1}$,

$$\prod_{s'} \frac{\Gamma(\alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1}))}{\Gamma(\alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t))} = \alpha^{a_t}_{s_t, s_{t+1}} + N^{a_t}_{s_t, s_{t+1}}(s_0, a_0, s_1, \dots, a_{t-1}, s_t).$$

Since $\sum_{s'} N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1}) = 1 + \sum_{s'} N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)$,

$$\frac{\Gamma\left(\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)\right)}{\Gamma\left(\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_t, s_{t+1})\right)} = \frac{1}{\sum_{s'} \alpha^{a_t}_{s_t, s'} + N^{a_t}_{s_t, s'}(s_0, a_0, s_1, \dots, a_{t-1}, s_t)}.$$

□

8 Appendix

Proposition 8.1. Consider a measurable space $(\tilde{\Omega}, \tilde{\mathcal{F}})$ and a stochastic process $(\tilde{Y}_n : \tilde{\Omega} \to \mathbb{R} \mid n \in \mathbb{N})$. Let $\tilde{Y} : \tilde{\Omega} \to \mathbb{R}^\infty$ be given by $\tilde{Y}(\tilde{\omega}) = (\tilde{Y}_n(\tilde{\omega}) \mid n \in \mathbb{N})$. For every $n \in \mathbb{N}$, let $Y_n : \mathbb{R}^\infty \to \mathbb{R}$ be given by $Y_n(\omega) = \omega_n$ and let $\mathcal{F} = \sigma(\cup_n \sigma(Y_n))$. In that case, $\tilde{Y}$ is $\tilde{\mathcal{F}}/\mathcal{F}$-measurable.

Proof. For every $n \in \mathbb{N}$, note that $\tilde{Y}_n = Y_n \circ \tilde{Y}$, so that $\tilde{Y}_n^{-1}(B) = \tilde{Y}^{-1}(Y_n^{-1}(B))$ for every $B \in \mathcal{B}(\mathbb{R})$. Because $\tilde{Y}_n$ is $\tilde{\mathcal{F}}$-measurable for every $n \in \mathbb{N}$, we know that $\tilde{Y}^{-1}(C) \in \tilde{\mathcal{F}}$ for every $C \in \cup_n \sigma(Y_n)$. Since $(\mathbb{R}^\infty, \mathcal{F})$ is a measurable space, note that $\mathcal{E} = \{F \in \mathcal{F} \mid \tilde{Y}^{-1}(F) \in \tilde{\mathcal{F}}\}$ is a σ-algebra on $\mathbb{R}^\infty$. Because $\cup_n \sigma(Y_n) \subseteq \mathcal{F}$, we know that $\sigma(\cup_n \sigma(Y_n)) = \mathcal{F} \subseteq \mathcal{E}$, so that $\mathcal{E} = \mathcal{F}$. Therefore, $\tilde{Y}$ is $\tilde{\mathcal{F}}/\mathcal{F}$-measurable. □

Proposition 8.2. Consider a measurable space $(\Omega, \mathcal{F})$, a stochastic process $(Y_n : \Omega \to \mathbb{N} \mid n \in \mathbb{N})$, and let $\mathcal{F}_n = \sigma(Y_0, \dots, Y_n)$ for every $n \in \mathbb{N}$. Furthermore, for every $n \in \mathbb{N}$, let $\mathcal{G}_n$ be given by

$$\mathcal{G}_n = \left\{ \bigcup_{y \in A} \{Y_0 = y_0, \dots, Y_n = y_n\} \mid A \subseteq \mathbb{N}^{n+1} \right\},$$

where $y = (y_0, \dots, y_n)$. In that case, $\mathcal{F}_n = \mathcal{G}_n$.

Proof. For some $n \in \mathbb{N}$, consider a set given by

$$\bigcup_{y \in A} \{Y_0 = y_0, \dots, Y_n = y_n\} = \bigcup_{y \in A} \bigcap_{k=0}^{n} \{Y_k = y_k\}$$

for some $A \subseteq \mathbb{N}^{n+1}$, where $y = (y_0, \dots, y_n)$. For every $k \in \mathbb{N}$, recall that

$$\sigma(Y_k) = \left\{ \bigcup_{y_k \in A_k} \{Y_k = y_k\} \mid A_k \subseteq \mathbb{N} \right\}.$$

The set A is countable, since it is a subset of the countable set $\mathbb{N}^{n+1}$, which is a finite Cartesian product between countable sets. Because $\{Y_k = y_k\} \in \mathcal{F}_n$ for every $k \in \{0, \dots, n\}$ and $y_k \in A_k$, we know that $\mathcal{G}_n \subseteq \mathcal{F}_n$.

For some $n \in \mathbb{N}$, let $A = A_0 \times \cdots \times A_n$, where $A_k \subseteq \mathbb{N}$ for every $k \in \{0, \ldots, n\}$. In that case,

$$\bigcup_{y \in A} \bigcap_{k=0}^{n} \{Y_k = y_k\} = \bigcup_{y_0 \in A_0} \cdots \bigcup_{y_n \in A_n} \bigcap_{k=0}^{n} \{Y_k = y_k\} = \left(\bigcup_{y_0 \in A_0} \{Y_0 = y_0\} \right) \cap \cdots \cap \left(\bigcup_{y_n \in A_n} \{Y_n = y_n\} \right).$$

Since $\mathbb{N} \subseteq \mathbb{N}$, note that $\sigma(Y_k) \subseteq \mathcal{G}_n$ for every $k \in \{0, \ldots, n\}$. Because $\mathcal{F}_n = \sigma(\cup_{k=0}^{n} \sigma(Y_k))$ and $\mathcal{G}_n \subseteq \mathcal{F}_n$, showing that $\mathcal{F}_n = \mathcal{G}_n$ now only requires showing that $\mathcal{G}_n$ is a σ-algebra on Ω.

For some $n \in \mathbb{N}$, let $A = \mathbb{N}^{n+1}$. Using the previous result, we know that $\Omega \in \mathcal{G}_n$.

For some $n \in \mathbb{N}$, consider a sequence $(G_{n,m} \in \mathcal{G}_n \mid m \in \mathbb{N})$ where

$$G_{n,m} = \bigcup_{y \in A_m} \{Y_0 = y_0, \ldots, Y_n = y_n\}$$

for some sequence $(A_m \subseteq \mathbb{N}^{n+1} \mid m \in \mathbb{N})$. Clearly,

$$\bigcup_{m} G_{n,m} = \bigcup_{m} \bigcup_{y \in A_m} \{Y_0 = y_0, \ldots, Y_n = y_n\} = \bigcup_{y \in A} \{Y_0 = y_0, \ldots, Y_n = y_n\},$$

where $A = \cup_m A_m$. Because $A \subseteq \mathbb{N}^{n+1}$, we know that $\cup_m G_{n,m} \in \mathcal{G}_n$.

For some $n \in \mathbb{N}$ and every $A \subseteq \mathbb{N}^{n+1}$, note that $A^c \subseteq \mathbb{N}^{n+1}$ and $A \cup A^c = \mathbb{N}^{n+1}$, so that

$$\left(\bigcup_{y \in A} \{Y_0 = y_0, \ldots, Y_n = y_n\} \right) \cup \left(\bigcup_{y \in A^c} \{Y_0 = y_0, \ldots, Y_n = y_n\} \right) = \bigcup_{y \in \mathbb{N}^{n+1}} \{Y_0 = y_0, \ldots, Y_n = y_n\} = \Omega.$$

Since the leftmost sets above are disjoint, if $G_n \in \mathcal{G}_n$, then $G_n^c \in \mathcal{G}_n$, so that $\mathcal{G}_n$ is a σ-algebra on Ω. □

Proposition 8.3. Consider a measurable space $(\Omega, \mathcal{F})$ and a stochastic process $(Y_n : \Omega \to \mathbb{N} \mid n \in \mathbb{N})$. A π-system $\mathcal{I}$ on Ω such that $\sigma(\mathcal{I}) = \sigma(Y_0, Y_1, \ldots)$ is given by

$$\mathcal{I} = \{\emptyset\} \cup \{\{Y_0 = y_0, \ldots, Y_n = y_n\} \mid n \in \mathbb{N} \text{ and } (y_0, \ldots, y_n) \in \mathbb{N}^{n+1}\} \cup \{\Omega\}.$$

Proof. First, we will show that $\mathcal{I}$ is indeed a π-system on Ω. For every $I \in \mathcal{I}$, note that $I \cap \emptyset = \emptyset$ and $I \cap \Omega = I$. For some $n' \in \mathbb{N}$ and $(y'_0, \ldots, y'_{n'}) \in \mathbb{N}^{n'+1}$, let $I_1 = \{Y_0 = y'_0, \ldots, Y_{n'} = y'_{n'}\}$. For some $n \geq n'$ and $(y_0, \ldots, y_n) \in \mathbb{N}^{n+1}$, let $I_2 = \{Y_0 = y_0, \ldots, Y_n = y_n\}$. In that case,

$$I_1 \cap I_2 = \{\omega \in \Omega \mid Y_0(\omega) = y'_0 = y_0, \ldots, Y_{n'}(\omega) = y'_{n'} = y_{n'}, Y_{n'}(\omega) = y_{n'}, \ldots, Y_n(\omega) = y_n\},$$

so that

$$I_1 \cap I_2 = \begin{cases} I_2, & \text{if } y'_k = y_k \text{ for every } k \in \{0, \ldots, n'\}, \\ \emptyset, & \text{if } y'_k \neq y_k \text{ for some } k \in \{0, \ldots, n'\}. \end{cases}$$

Therefore, $I_1 \cap I_2 \in \mathcal{I}$, so that $\mathcal{I}$ is a π-system on Ω.

By Proposition 8.2, for every $n \in \mathbb{N}$, the σ-algebra $\sigma(Y_0, \ldots, Y_n)$ on Ω is given by

$$\sigma(Y_0, \ldots, Y_n) = \left\{ \bigcup_{y \in A} \{Y_0 = y_0, \ldots, Y_n = y_n\} \mid A \subseteq \mathbb{N}^{n+1} \right\},$$

where $y = (y_0, \ldots, y_n)$ and A is a countable set. For every $n \in \mathbb{N}$, because each $F_n \in \sigma(Y_0, \ldots, Y_n)$ is a countable union of elements of $\mathcal{I}$, we know that $F_n \in \sigma(\mathcal{I})$. Therefore, $\cup_n \sigma(Y_0, \ldots, Y_n) \subseteq \sigma(\mathcal{I})$ and $\sigma(Y_0, Y_1, \ldots) \subseteq \sigma(\mathcal{I})$. □

Acknowledgements

I would like to thank Daniel Valesin for the ideas behind some proofs found in these notes.

License

References

[1] Duff, Michael O'Gordon. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.