Optimality of Markov Policies in MDPs

Marek Petrik

January 24, 2025

1 Probability Spaces and Expectation

In this section, we define the basic probability concepts on finite sets.

1.1 Definitions

Remark 1.1.1. This document omits basic intuitive definitions, such as the comparison or arithmetic operations on random variables. Arithmetic operations on random variables are performed element-wise for each element of the sample set. Please see the Lean file for complete details.

Definition 1.1.2. A *finite probability measure* $p \colon \Omega \to \mathbb{R}_+$ on a finite set Ω is any function that satisfies

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Definition 1.1.3. The set of *finite probability measures* $\Delta(\Omega)$ for a finite Ω is defined as

$$\Delta(\Omega) := \left\{ p \colon \Omega \to \mathbb{R}_+ \mid \sum_{\omega \in \Omega} p(\omega) = 1 \right\}.$$

Definition 1.1.4. A *finite probability space* is $P = (\Omega, p)$, where Ω is a finite set referred to as the *sample set*, $p \in \Delta(\Omega)$, and the σ-algebra is 2^Ω.

Definition 1.1.5. A *random variable* defined on a finite probability space P is a mapping $\tilde{x} \colon \Omega \to \mathbb{R}$.

For the remainder of Section 1, we assume that $P = (\Omega, p)$ is a *finite probability space*. All random variables are defined on the space P unless specified otherwise.

Definition 1.1.6. A *boolean* set is $\mathcal{B} = \{\text{false}, \text{true}\}$.

Definition 1.1.7. The *expectation* of a random variable $\tilde{x} \colon \Omega \to \mathbb{R}$ is

$$\mathbb{E}\left[\tilde{x}\right] := \sum_{\omega \in \Omega} p(\omega) \cdot \tilde{x}(\omega).$$

Definition 1.1.8. An *indicator* function $\mathbb{I} \colon \mathcal{B} \to \{0, 1\}$ is defined for $b \in \mathcal{B}$ as

$$\mathbb{I}(b) := \begin{cases} 1 & \text{if } b = \text{true}, \\ 0 & \text{if } b = \text{false}. \end{cases}$$

Definition 1.1.9. The *probability* of $\tilde{b}\colon \Omega \to \mathcal{B}$ is defined as

$$\mathbb{P}\left[\tilde{b}\right] := \mathbb{E}\left[\mathbb{I}(\tilde{b})\right].$$

Definition 1.1.10. The *conditional expectation* of $\tilde{x}\colon \Omega \to \mathbb{R}$ conditioned on $\tilde{b}\colon \Omega \to \mathcal{B}$ is defined as

$$\mathbb{E}\left[\tilde{x} \mid \tilde{b}\right] := \frac{1}{\mathbb{P}[\tilde{b}]}\mathbb{E}\left[\tilde{x} \cdot \mathbb{I} \circ \tilde{b}\right],$$

where we define that $x/0 = 0$ for each $x \in \mathbb{R}$.

Definition 1.1.11. The *conditional probability* of $\tilde{b}\colon \Omega \to \mathcal{B}$ on $\tilde{c}\colon \Omega \to \mathcal{B}$ is defined as

$$\mathbb{P}\left[\tilde{b} \mid \tilde{c}\right] := \mathbb{E}\left[\mathbb{I}(\tilde{b}) \mid \tilde{c}\right].$$

Remark 1.1.12. It is common to prohibit conditioning on a zero probability event both for expectation and probabilities. In this document, we follow the Lean convention, where the division by 0 is 0; see `div_zero`. However, even some basic probability and expectation results may require that we assume that the conditioned event does not have probability zero for it to hold.

Definition 1.1.13. The *random conditional expectation* of a random variable $\tilde{x}\colon \Omega \to \mathbb{R}$ conditioned on $\tilde{y}\colon \Omega \to \mathcal{Y}$ for a finite set $\mathcal{Y}$ is the random variable $\mathbb{E}[\tilde{x} \mid \tilde{y}] : \Omega \to \mathbb{R}$ is defined as

$$\mathbb{E}[\tilde{x} \mid \tilde{y}](\omega) := \mathbb{E}[\tilde{x} \mid \tilde{y} = \tilde{y}(\omega)], \quad \forall \omega \in \Omega.$$

Remark 1.1.14. The Lean file defines expectations more broadly for a data type ρ which is more general than just $\mathbb{R}$. The main reason to generalize to both $\mathbb{R}$ and $\mathbb{R}_+$. However, in principle, the definitions could be used to reason with expectations that go beyond real numbers and may include other algebras, such as vectors or matrices.

1.2 Basic Properties

Lemma 1.2.1. *Suppose that $\tilde{b}, \tilde{c}\colon \Omega \to \mathcal{B}$. Then:*

$$\mathbb{I}\left(\tilde{b} \wedge \tilde{c}\right) = \mathbb{I}(\tilde{b}) \cdot \mathbb{I}(\tilde{c}),$$

where the equality applies for all $\omega \in \Omega$.

Theorem 1.2.2. *Suppose that $\tilde{c}\colon \Omega \to \mathcal{B}$ such that $\mathbb{P}[\tilde{c}] = 0$. Then for any $\tilde{x}\colon \Omega \to \mathbb{R}$:*

$$\mathbb{E}[\tilde{x} \mid \tilde{c}] = 0.$$

Proof. Immediate from the definition and the fact that $0 \cdot x = 0$ for $x \in \mathbb{R}$. □

Theorem 1.2.3. *Suppose that $\tilde{c}\colon \Omega \to \mathcal{B}$ such that $\mathbb{P}[\tilde{c}] = 0$. Then for any $\tilde{b}\colon \Omega \to \mathbb{R}$:*

$$\mathbb{P}\left[\tilde{b} \mid \tilde{c}\right] = 0.$$

Proof. Immediate from Theorem 1.2.2. □

Theorem 1.2.4. *Suppose that $\tilde{b}, \tilde{c}\colon \Omega \to \mathcal{B}$, then*

$$\mathbb{P}\left[\tilde{b} \wedge \tilde{c}\right] = \mathbb{P}\left[\tilde{b} \mid \tilde{c}\right] \cdot \mathbb{P}\left[\tilde{c}\right].$$

Proof. The property holds immediately when $\mathbb{P}[\tilde{c}] = 0$. Assume that $\mathbb{P}[\tilde{c}] > 0$. Then:

$$\begin{aligned}
\mathbb{P}\left[\tilde{b} \wedge \tilde{c}\right] &= \mathbb{E}\left[\mathbb{I}(\tilde{b} \wedge \tilde{c})\right] && \text{[Definition 1.1.9]} \\
&= \mathbb{E}\left[\mathbb{I}(\tilde{b}) \cdot \mathbb{I}(\tilde{c})\right] && \text{[Lemma 1.2.1]} \\
&= \frac{1}{\mathbb{P}[\tilde{c}]}\mathbb{E}\left[\mathbb{I}(\tilde{b}) \cdot \mathbb{I}(\tilde{c})\right] \cdot \mathbb{P}[\tilde{c}] && \cdot 1 \\
&= \mathbb{E}\left[\mathbb{I}(\tilde{b}) \mid \tilde{c}\right] \cdot \mathbb{P}[\tilde{c}] && \text{[Definition 1.1.10]} \\
&= \mathbb{P}\left[\tilde{b} \mid \tilde{c}\right] \cdot \mathbb{P}[\tilde{c}] && \text{[Definition 1.1.11].}
\end{aligned}$$

□

Lemma 1.2.5. *Let $\tilde{y}\colon \Omega \to \mathcal{Y}$ with a finite $\mathcal{Y}$. Then*

$$\mathbb{P}[\tilde{y} = y(\omega)] \geq p(\omega), \quad \omega \in \Omega.$$

Proof.

$$\begin{aligned}
\mathbb{P}[\tilde{y} = y(\omega)] &= \sum_{\omega' \in \Omega} p(\omega) \cdot \mathbb{I}(\tilde{y}(\omega') = \tilde{y}(\omega)) && \text{[Definition 1.1.9]} \\
&\geq p(\omega) && \omega \in \Omega[\text{ and }]p(\omega') \geq 0, \forall \omega' \in \Omega.
\end{aligned}$$

□

Remark 1.2.6. Theorem 1.2.12 shows the equivalence of expectations for surely equal random variables.

Theorem 1.2.7. *Random variables $\tilde{x}, \tilde{y}\colon \Omega \to \mathbb{R}$ satisfy that*

$$\mathbb{E}[\tilde{x} + \tilde{y}] = \mathbb{E}[\tilde{x}] + \mathbb{E}[\tilde{y}].$$

Proof. From the distributive property of sums. □

Theorem 1.2.8. *A random variable $\tilde{x}\colon \Omega \to \mathbb{R}$ and $c \in \mathbb{R}$ satisfies that*

$$\mathbb{E}[c] = c.$$

Theorem 1.2.9. *Suppose that $\tilde{x}\colon \Omega \to \mathbb{R}$ and $c \in \mathbb{R}$. Then*

$$\mathbb{E}[c + \tilde{x}] = c + \mathbb{E}[\tilde{x}].$$

Proof. From Theorems 1.2.7 and 1.2.8. □

Theorem 1.2.10. *Suppose that $\tilde{x}, \tilde{y}\colon \Omega \to \mathbb{R}$ and $\tilde{z}\colon \Omega \to \mathcal{V}$ are random variables and $c \in \mathbb{R}$, such that $\tilde{y}(\omega) = c + \tilde{x}(\omega)$. Then*

$$\mathbb{E}[\tilde{y} \mid \tilde{z}](\omega) = c + \mathbb{E}[\tilde{x} \mid \tilde{z}](\omega), \quad \forall \omega \in \Omega.$$

Proof. From Theorem 1.2.9. □

Theorem 1.2.11. *Suppose that $\tilde{x}, \tilde{y}\colon \Omega \to \mathbb{R}$ satisfy that*

$$\forall \omega \in \Omega, p(\omega) > 0 \Rightarrow \tilde{x}(\omega) \geq \tilde{y}(\omega).$$

Then

$$\mathbb{E}[\tilde{x}] \geq \mathbb{E}[\tilde{y}].$$

Theorem 1.2.12 (Congruence of Expectation)**.** *Suppose that* $\tilde{x}, \tilde{z} \colon \Omega \to \mathbb{R}$ *satisfy that*

$$\forall \omega \in \Omega, p(\omega) > 0 \Rightarrow \tilde{x}(\omega) = \tilde{z}(\omega).$$

Then

$$\mathbb{E}[\tilde{x}] = \mathbb{E}[\tilde{z}].$$

Proof. Immediately from the congruence of sums. □

1.3 The Laws of The Unconscious Statisticians

Theorem 1.3.1. *Let* $\tilde{x} \colon \Omega \to \mathbb{R}$ *be a random variable. Then:*

$$\mathbb{E}[\tilde{x}] = \sum_{x \in \tilde{x}(\Omega)} \mathbb{P}[\tilde{x} = x] \cdot x.$$

Proof. Let $\mathcal{X} := \tilde{x}(\Omega)$, which is a finite set. Then:

$$\begin{aligned}
\mathbb{E}[\tilde{x}] &= \sum_{\omega \in \Omega} p(\omega) \cdot \tilde{x}(\omega) && \text{[Definition 1.1.7]} \\
&= \sum_{\omega \in \Omega} \sum_{x \in \mathcal{X}} p(\omega) \cdot \tilde{x}(\omega) \cdot \mathbb{I}(x = \tilde{x}(\omega)) && \text{[??]} \\
&= \sum_{\omega \in \Omega} \sum_{x \in \mathcal{X}} p(\omega) \cdot x \cdot \mathbb{I}(x = \tilde{x}(\omega)) && \text{[??]} \\
&= \sum_{x \in \mathcal{X}} x \cdot \sum_{\omega \in \Omega} p(\omega) \cdot \mathbb{I}(x = \tilde{x}(\omega)) && \text{[??]} \\
&= \sum_{x \in \mathcal{X}} x \cdot \mathbb{E}[\mathbb{I}(x = \tilde{x}(\omega))] && \text{[Definition 1.1.7]} \\
&= \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}[x = \tilde{x}(\omega)]. && \text{[Definition 1.1.9]}
\end{aligned}$$

□

The following theorem generalizes the theorem above.

Theorem 1.3.2. *Let* $\tilde{x} \colon \Omega \to \mathbb{R}$ *and* $\tilde{b} \colon \Omega \to \mathcal{Y}$ *be random variables. Then:*

$$\mathbb{E}\left[\tilde{x} \mid \tilde{b}\right] = \sum_{x \in \tilde{x}(\Omega)} \mathbb{P}\left[\tilde{x} = x \mid \tilde{b}\right] \cdot x.$$

Theorem 1.3.3. *Let* $\tilde{x} \colon \Omega \to \mathbb{R}$ *and* $\tilde{y} \colon \Omega \to \mathcal{Y}$ *be random variables with* $\mathcal{Y}$ *finite. Then:*

$$\mathbb{E}[\mathbb{E}[\tilde{x} \mid \tilde{y}]] = \sum_{y \in \mathcal{Y}} \mathbb{E}[\tilde{x} \mid \tilde{y} = y] \cdot \mathbb{P}[\tilde{y} = y].$$

1.4 Total Expectation and Probability

Theorem 1.4.1 (Law of Total Probability)**.** *Let* $\tilde{b} \colon \Omega \to \mathcal{B}$ *and* $\tilde{y} \colon \Omega \to \mathcal{Y}$ *be random variables with a finite set* $\mathcal{Y}$*. Then:*

$$\sum_{y \in \mathcal{Y}} \mathbb{P}\left[\tilde{b} \wedge (\tilde{y} = y)\right] = \mathbb{P}\left[\tilde{b}\right].$$

Theorem 1.4.2 (Law of Total Expectation)**.** *Let $\tilde{x}\colon \Omega \to \mathcal{X}$ and $\tilde{y}\colon \Omega \to \mathcal{Y}$ be random variables with a finite set $\mathcal{Y}$. Then:*

$$\mathbb{E}\left[\mathbb{E}\left[\tilde{x} \mid \tilde{y}\right]\right] = \mathbb{E}\left[\tilde{x}\right].$$

Proof. Recall that we are allowing the division by 0 and assume that $x/0 = 0$.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\tilde{x} \mid \tilde{y}\right]\right] &= \sum_{\omega\in\Omega} p(\omega)\cdot \mathbb{E}\left[\tilde{x} \mid \tilde{y}\right](\omega) && \text{[Definition 1.1.7]}\\
&= \sum_{\omega\in\Omega} p(\omega)\cdot \mathbb{E}\left[\tilde{x} \mid \tilde{y} = \tilde{y}(\omega)\right] && \text{[Definition 1.1.13]}\\
&= \sum_{\omega\in\Omega} \frac{p(\omega)}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega)\right]} \sum_{\omega'\in\Omega} p(\omega')\cdot \tilde{x}(\omega')\cdot \mathbb{I}\left(\tilde{y}(\omega') = \tilde{y}(\omega)\right) && \text{[Definition 1.1.10]}\\
&= \sum_{\omega'\in\Omega} p(\omega')\cdot \tilde{x}(\omega')\cdot \sum_{\omega\in\Omega} \frac{p(\omega)}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega)\right]} \mathbb{I}\left(\tilde{y}(\omega') = \tilde{y}(\omega)\right) && \text{[rearrange]}\\
&= \sum_{\omega'\in\Omega} p(\omega')\cdot \tilde{x}(\omega')\cdot \sum_{\omega\in\Omega} \frac{p(\omega)}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right]} \mathbb{I}\left(\tilde{y}(\omega') = \tilde{y}(\omega)\right) && \text{[equals when]}\tilde{y}(\omega') = \tilde{y}(\omega)\\
&= \sum_{\omega'\in\Omega} p(\omega')\cdot \tilde{x}(\omega') && \text{[see below]}\\
&= \mathbb{E}\left[\tilde{x}\right].
\end{aligned}$$

Above, we used the fact that

$$p(\omega')\cdot \sum_{\omega\in\Omega} \frac{p(\omega)}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right]} \mathbb{I}\left(\tilde{y}(\omega') = \tilde{y}(\omega)\right) = p(\omega'),$$

which follows by analyzing two cases. First, when $p(\omega') = 0$, then the equality holds immediately. If $p(\omega') > 0$ then by Lemma 1.2.5, $\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right] > 0$, we get from Definition 1.1.9 that

$$\sum_{\omega\in\Omega} \frac{p(\omega)}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right]} \mathbb{I}\left(\tilde{y}(\omega') = \tilde{y}(\omega)\right) = \frac{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right]}{\mathbb{P}\left[\tilde{y} = \tilde{y}(\omega')\right]} = 1,$$

which completes the step. □

2 Formal Decision Framework

2.1 Markov Decision Process

Definition 2.1.1. A *Markov decision process* $M := (\mathcal{S}, \mathcal{A}, P, r)$ consists of a finite nonempty set of states $\mathcal{S}$, a finite nonempty set of actions $\mathcal{A}$, transition function $P\colon \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, and a reward function $r\colon \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$.

2.2 Histories

We implicitly assume in the remainder of the section an MDP $M = (\mathcal{S}, \mathcal{A}, p, r)$.

Definition 2.2.1. A *history* h in a set of histories $\mathcal{H}$ is a sequence of states and actions defined for M recursively as

$$h := \langle s \rangle, \quad [\text{or}] \quad h := \langle h', s, a \rangle,$$

where $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h' \in \mathcal{H}$.

Definition 2.2.2. The *length* $l \colon \mathcal{H} \to \mathbb{N}$ of a history h is defined as

$$\begin{aligned} l(\langle s \rangle) &:= 0, \\ l(\langle h', s, a \rangle) &:= 1 + h', \qquad h' \in \mathcal{H}. \end{aligned}$$

Definition 2.2.3. The set $\mathcal{H}_{\text{NE}}$ of *non-empty histories* is

$$\mathcal{H}_{\text{NE}} := \{h \in \mathcal{H} \mid l(h) \geq 1\}.$$

Definition 2.2.4. *Following histories* $\mathcal{H}(h, t) \subseteq \mathcal{H}$ for $h \in \mathcal{H}$ of length $t \in \mathbb{N}$ are defined recursively as

$$\mathcal{H}(h,t) := \begin{cases} \{h\} & \text{if } t = 0, \\ \{\langle h', a, s \rangle \mid h \in \mathcal{H}(h', t-1), a \in \mathcal{A}, s \in \mathcal{S}\} & \text{otherwise.} \end{cases}$$

Definition 2.2.5. The set of *histories* $\mathcal{H}_t$ of *length* $t \in \mathbb{N}$ is defined recursively as

$$\mathcal{H}_t = \begin{cases} \{\langle s \rangle \mid s \in \mathcal{S}\} & \text{if } t = 0, \\ \{\langle h, a, s \rangle \mid h \in \mathcal{H}_{t-1}, a \in \mathcal{A}, s \in \mathcal{S}\} & \textit{textotherwise.} \end{cases}$$

Theorem 2.2.6. *For $h \in \mathcal{H}$:*

$$l_{\text{h}}(h') = l_{\text{h}}(h) + t, \qquad \forall h' \in \mathcal{H}(h, t).$$

Proof. The theorem follows by induction on t from the definition. □

Definition 2.2.7. We use $\tilde{s}_k \colon \mathcal{H} \to \mathcal{S}$ to denote the 0-based k-th state of each history.

Definition 2.2.8. We use $\tilde{a}_k \colon \mathcal{H} \to \mathcal{A}$ to denote the 0-based k-th action of each history.

Definition 2.2.9. The *history-reward* random variable $\tilde{r}^{\text{h}} \colon \mathcal{H} \to \mathbb{R}$ for $h = \langle h', a, s \rangle \in \mathcal{H}$ for $h' \in \mathcal{H}$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$ is defined recursively as

$$\tilde{r}^{\text{h}}(h) := r(s_{\text{l}}(h'), a, s) + r_{\text{h}}(h').$$

Definition 2.2.10. The *history-reward* random variable $\tilde{r}^{\text{h}}_k \colon \mathcal{H} \to \mathbb{R}$ for $h = \langle h', a, s \rangle \in \mathcal{H}$ for $h' \in \mathcal{H}$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$ is defined as the k-th reward (0-based) of a history.

Definition 2.2.11. The *history-reward* random variable $\tilde{r}^{\text{h}}_{\leq k} \colon \mathcal{H} \to \mathbb{R}$ for $h = \langle h', a, s \rangle \in \mathcal{H}$ for $h' \in \mathcal{H}$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$ is defined as the sum of all k-th or earlier rewards (0-based) of a history.

Definition 2.2.12. The *history-reward* random variable $\tilde{r}^{\text{h}}_{\geq k} \colon \mathcal{H} \to \mathbb{R}$ for $h = \langle h', a, s \rangle \in \mathcal{H}$ for $h' \in \mathcal{H}$, $a \in \mathcal{A}$, and $s \in \mathcal{S}$ is defined as the sum of k-th or later reward (0-based) of a history.

2.3 Policies

Definition 2.3.1. The set of *decision rules* $\mathcal{D}$ is defined as $\mathcal{D} := \mathcal{A}^{\mathcal{S}}$. A single action $a \in \mathcal{A}$ can also be interpreted as a decision rule $d := s \mapsto a$.

Definition 2.3.2. The set of *history-dependent policies* is $\Pi_{\text{HR}} := \Delta(\mathcal{A})^{\mathcal{H}}$.

Definition 2.3.3. The set of *Markov deterministic policies* Π_{MD} is $\Pi_{\text{MD}} := \mathcal{D}^{\mathbb{N}}$. A Markov deterministic policy $\pi \in \Pi_{\text{MD}}$ can also be interpreted as $\bar{\pi} \in \Pi_{\text{HR}}$:

$$\bar{\pi}(h) := \delta\left[\pi(l(h), s_{\text{l}}(h))\right],$$

where δ is the Dirac distribution, l is the length defined in Definition 2.2.2, and s_{l} is the history's last state.

Definition 2.3.4. The set of *stationary deterministic policies* Π_{SD} is defined as $\Pi_{\text{SD}} := \mathcal{D}$. A stationary policy $\pi \in \Pi_{\text{SD}}$ can be interpreted as $\bar{\pi} \in \Pi_{\text{HR}}$:

$$\bar{\pi}(h) := \delta\left[\pi(s_{\text{l}}(h))\right],$$

where δ is the Dirac distribution and s_{l} is the history's last state.

2.4 Distribution

Definition 2.4.1. The *history probability distribution* $p^{\text{h}}_T \colon \Pi_{\text{HR}} \to \Delta(\mathcal{H}(h,t))$ and $\pi \in \Pi_{\text{HR}}$ is defined for each $T \in \mathbb{N}$ and $h \in \mathcal{H}(\hat{h}, t)$ as

$$(p^{\text{h}}_T(\pi))(h) := \begin{cases} \mathbb{I}(h = \hat{h}) & \text{if } T = 0, \\ p^{\text{h}}_{T-1}(h', \pi) \cdot \pi(h', a) \cdot p(s_{\text{l}}(h'), a, s) & \text{if } T > 1 \wedge h = \langle h', a, s \rangle. \end{cases}$$

Moreover, the function p^{h} maps policies to correct probability distribution.

Definition 2.4.2. The *history-dependent expectation* is defined for each $t \in \mathbb{N}$, $\pi \in \Pi_{\text{HR}}$, $\hat{h} \in \mathcal{H}$ and a $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$ as

$$\mathbb{E}^{\hat{h},\pi,t}[\tilde{x}] := \mathbb{E}\left[\tilde{x}\right] = \sum_{h \in \mathcal{H}(\hat{h},t)} p^{\text{h}}(h, \pi) \cdot \tilde{x}(h).$$

In the $\mathbb{E}$ operator above, the random variable $\tilde{x}$ lives in a probability space (Ω, p) where $\Omega = \mathcal{H}(\hat{h}, t)$ and $p(h) = p^{\text{h}}(h, \pi), \forall h \in \Omega$. Moreover, if $\hat{h}$ is a state, then it is interpreted as a history with the single initial state.

Definition 2.4.3. The *history-dependent expectation* is defined for each $t \in \mathbb{N}$, $\pi \in \Pi_{\text{HR}}$, $\hat{h} \in \mathcal{H}$, $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$, $\tilde{b} \colon \mathcal{H} \to \mathcal{B}$ as

$$\mathbb{E}^{\hat{h},\pi,t}[\tilde{x} \mid \tilde{b}] := \mathbb{E}\left[\tilde{x} \mid \tilde{b}\right].$$

In the $\mathbb{E}$ operator above, the random variables $\tilde{x}$ and $\tilde{b}$ live in a probability space (Ω, p) where $\Omega = \mathcal{H}(\hat{h}, t)$ and $p(h) = p^{\text{h}}(h, \pi), \forall h \in \Omega$. Moreover, if $\hat{h}$ is a state, then it is interpreted as a history with the single initial state.

Definition 2.4.4. The *history-dependent expectation* is defined for each $t \in \mathbb{N}$, $\pi \in \Pi_{\text{HR}}$, $\hat{h} \in \mathcal{H}$, $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$, $\tilde{y} \colon \mathcal{H} \to \mathcal{V}$ as

$$\mathbb{E}^{\hat{h},\pi,t}[\tilde{x} \mid \tilde{y}](h) := \mathbb{E}\left[\tilde{x} \mid \tilde{y} = \tilde{y}(h)\right](h), \quad \forall h \in \mathcal{H}(\hat{h}, t).$$

In the $\mathbb{E}$ operator above, the random variables $\tilde{x}$ and $\tilde{h}$ live in a probability space (Ω, p) where $\Omega = \mathcal{H}(\hat{h}, t)$ and $p(h) = p^{\text{h}}(h, \pi), \forall h \in \Omega$. Moreover, if $\hat{h}$ is a state, then it is interpreted as a history with the single initial state.

2.5 Basic Properties

Theorem 2.5.1. *Assume $\tilde{x}\colon \mathcal{H} \to \mathbb{R}$ and $c \in \mathbb{R}$. Then $\forall h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}, t \in \mathbb{N}$:*

$$\mathbb{E}^{\hat{h},\pi,t}\left[c + \tilde{x}\right] = c + \mathbb{E}^{\hat{h},\pi,t}\left[\tilde{x}\right].$$

Proof. Directly from Theorem 1.2.12. □

Theorem 2.5.2. *Suppose that $\tilde{x}, \tilde{y}\colon \mathcal{H} \to \mathbb{R}$. Then $\forall h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}, t \in \mathbb{N}$:*

$$\mathbb{E}^{\hat{h},\pi,t}\left[\tilde{x} + \tilde{y}\right] = \mathbb{E}^{\hat{h},\pi,t}\left[\tilde{x}\right] + \mathbb{E}^{\hat{h},\pi,t}\left[\tilde{y}\right].$$

Proof. From Theorem 1.2.7. □

Theorem 2.5.3. *Suppose that $c \in \mathbb{R}$. Then $\forall h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}, t \in \mathbb{N}$:*

$$\mathbb{E}^{\hat{h},\pi,t}\left[c\right] = c.$$

Proof. From Theorem 1.2.8. □

Theorem 2.5.4. *Suppose that $\tilde{x}, \tilde{y}\colon \mathcal{H} \to \mathbb{R}$ satisfy that $\tilde{x}(h) = \tilde{y}(h), \forall h \in \mathcal{H}$. Then $\forall h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}, t \in \mathbb{N}$:*

$$\mathbb{E}^{\hat{h},\pi,t}\left[\tilde{x}\right] = c + \mathbb{E}^{\hat{h},\pi,t}\left[\tilde{y}\right].$$

Proof. From Theorem 1.2.9. □

Theorem 2.5.5. *For each $\hat{h} \in \mathcal{H}$, $\pi \in \Pi_{\mathrm{HR}}$, and $t \in \mathbb{N}$:*

$$\mathbb{E}^{\hat{h},\pi,t}\left[\tilde{r}^{\mathrm{h}}\right] = \mathbb{E}^{\hat{h},\pi,t}\left[\sum_{k=0}^{l_{\mathrm{h}}(\tilde{\mathrm{id}})-1} r(\tilde{s}_k, \tilde{a}_k, \tilde{s}_{k+1})\right],$$

where $\tilde{\mathrm{id}}(h)$ is the identity function, l_{h} is the length of a history (0-based), $\tilde{s}_k\colon \mathcal{H} \to \mathcal{S}$ and $\tilde{a}_k\colon \mathcal{H} \to \mathcal{A}$ are the 0-based k-th state and action, respectively of each history.

Proof. Follows from Theorem 2.5.1 and the equality of the reward function $\tilde{r}^{\mathrm{h}}$ and the sum in the expectation. □

Theorem 2.5.6. *For each $h \in \mathcal{H}$, $\pi \in \Pi_{\mathrm{HR}}$, and $t \in \mathbb{N}$:*

$$\mathbb{E}^{h,\pi,t}\left[\tilde{r}^{\mathrm{h}}\right] = \tilde{r}^{\mathrm{h}}(h) + \mathbb{E}^{h,\pi,t}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0}\right],$$

where $k_0 := l_{\mathrm{h}}(h)$.

Proof. Follows from Theorem 2.5.4. □

Theorem 2.5.7. *For each $\hat{h} \in \mathcal{H}$, $\pi \in \Pi_{\mathrm{HR}}$, $t \in \mathbb{N}$, $h \in \mathcal{H}$:*

$$\mathbb{P}^{\hat{h},\pi,t}[\tilde{s}_{k_0} = \tilde{s}_{k_0}(\omega) \wedge \tilde{a}_{k_0} = \tilde{a}_{k_0}(\omega)] > 0 \quad \Rightarrow \quad \mathbb{E}^{\hat{h},\pi,t}\left[\tilde{r}^{\mathrm{h}}_{k_0} \mid \tilde{s}_{k_0}, \tilde{a}_{k_0}\right](h) = \tilde{r}^{\mathrm{h}}_{k_0}(h), \forall h \in \mathcal{H}.$$

where $k_0 := l_{\mathrm{h}}(\hat{h})$.

Proof. From Theorem 1.2.8. □

Theorem 2.5.8. *Assume $h \in \mathcal{H}$ and $f \colon \mathcal{H} \to \mathbb{R}$ such that $s_0 := s_1(h)$*

$$f(\langle h, a, s \rangle) = f(\langle s_0, a, a \rangle), \forall a \in \mathcal{A}, s \in \mathcal{S}.$$

Then

$$\mathbb{E}^{h,\pi,1}\left[\tilde{f}\right] = \mathbb{E}^{s_0,\pi,1}\left[\tilde{f}\right].$$

Proof. Directly from the definition of the expectation. □

2.6 Total Expectation

Theorem 2.6.1 (Total Expectation)**.** *For each $h \in \mathcal{H}$, $\pi \in \Pi_{\mathrm{HR}}$, $t \in \mathbb{N}$, $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$ and $\tilde{y} \colon \mathcal{H} \to \mathcal{V}$:*

$$\mathbb{E}^{h,\pi,t}\left[\mathbb{E}^{h,\pi,t}\left[\tilde{x} \mid \tilde{y}\right]\right] = \mathbb{E}^{h,\pi,t}\left[\tilde{x}\right].$$

Proof. From Theorem 1.4.2. □

Theorem 2.6.2. *Suppose that the random variable $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$ satisfies for some $k, t \in \mathbb{N}$, with $k \leq t$, that*

$$\tilde{x}(h) = \tilde{x}(h_{\leq k}), \forall h \in \mathcal{H},$$

where $h_{\leq k}$ is the prefix of h of length k. Then for each $h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}$:

$$\mathbb{E}^{h,\pi,t}\left[\tilde{x}\right] = \mathbb{E}^{h,\pi,k}\left[\tilde{x}\right].$$

2.7 Conditional Properties

Theorem 2.7.1. *For each $h \in \mathcal{H}$, $k_0 := l_{\mathrm{h}}(h)$, $\pi \in \Pi_{\mathrm{HR}}$, $t \in \mathbb{N}$, $\tilde{x} \colon \mathcal{H} \to \mathbb{R}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$:*

$$\mathbb{E}^{h,\pi,t+1}[\tilde{x} \mid \tilde{a}_{k_0} = a, \tilde{s}_{k_0+1} = s] = \mathbb{E}^{\langle h,a,s \rangle,\pi,t}[\tilde{x}].$$

Proof. This should follow by algebraic manipulation. □

3 Dynamic Program: History-Dependent Finite Horizon

In this section, we derive dynamic programming equations for histories. We assume an MDP $M = (\mathcal{S}, \mathcal{A}, p, r)$ throughout this section.

The main idea of the proof is to:

1. Derive (exponential-size) dynamic programming equations for the history-dependent value function of history-dependent policies
   (a) Define the value function
   (b) Define an optimal value function
2. Show that value functions decompose to equivalence classes
3. Show that the value function for the equivalence classes can be computed efficiently

3.1 Definitions

Definition 3.1.1. A finite horizon objective definition is given by $O := (s_0, T)$ where $s_0 \in \mathcal{S}$ is the initial state and $T \in \mathbb{N}$ is the horizon.

In the reminder of the section, we assume an objective $O = (s_0, T)$.

Definition 3.1.2. The *finite horizon objective function* for and objective O is $\pi \in \Pi_{\mathrm{HR}}$ is defined as

$$\rho(\pi, O) := \mathbb{E}^{s_0,\pi,T}\left[\tilde{r}^{\mathrm{h}}\right].$$

Definition 3.1.3. A policy $\pi^\star \in \Pi_{\mathrm{HR}}$ is *return optimal* for an objective O if

$$\rho(\pi^\star, O) \geq \rho(\pi, O), \quad \forall \pi \in \Pi_{\mathrm{HR}}.$$

Definition 3.1.4. The *set of history-dependent value functions* $\mathcal{U}$ is defined as

$$\mathcal{U} := \mathbb{R}^{\mathcal{H}}.$$

Definition 3.1.5. A *history-dependent policy value function* $\hat{u}_t \colon \mathcal{H} \times \Pi_{\mathrm{HR}} \to \mathbb{R}$ for each $h \in \mathcal{H}$, $\pi \in \Pi_{\mathrm{HR}}$, and $t \in \mathbb{N}$ is defined as

$$\hat{u}_t(h; \pi) := \mathbb{E}^{h,\pi,t}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0}\right],$$

where $k_0 := l_{\mathrm{h}}(h)$. Note that this definition includes also the rewards accrued in the history h.

Definition 3.1.6. The optimal history-dependent value function $\hat{u}^\star_t \colon \mathcal{H} \to \mathbb{R}$ is defined for a horizon $t \in \mathbb{N}$ as

$$\hat{u}^\star_t(h) := \sup_{\pi \in \Pi_{\mathrm{HR}}} \hat{u}_t(h; \pi).$$

[It is not clear how to define this in Lean because the set of policies is not finite and it is not clear how to define the supremum or maximum.]

The following definition is another way of defining an optimal policy. Note that just because we define this value, it does not mean that it must exist.

Definition 3.1.7. An *optimal history-dependent policy* $\pi^\star \in \Pi_{\mathrm{HR}}$ is a policy that satisfies

$$\hat{u}_t(h; \pi^\star) \geq \hat{u}_t(h; \pi), \quad \forall \pi \in \Pi_{\mathrm{HR}}, h \in \mathcal{H}.$$

Definition 3.1.8. A policy $\pi^\star \in \Pi_{\mathrm{HR}}$ optimal in Definition 3.1.7 is also optimal in Definition 3.1.7 for any initial state s_0 and horizon T.

3.2 History-dependent Dynamic Program

The following definitions of history-dependent value functions use a dynamic program formulation.

Definition 3.2.1. The *history-dependent policy Bellman operator* $L^\pi_{\mathrm{h}} \colon \mathcal{U} \to \mathcal{U}$ is defined for $\pi \in \Pi_{\mathrm{HR}}$ as

$$(L^\pi_{\mathrm{h}} \tilde{u})(h) := \mathbb{E}^{h,\pi,1}\left[\tilde{r}^{\mathrm{h}}_{l_{\mathrm{h}}(h)} + \tilde{u}\right], \quad \forall \tilde{u} \in \mathcal{U},$$

where the value function $\tilde{u}$ is interpreted as a random variable on defined on the sample space $\Omega = \mathcal{H}$.

Definition 3.2.2. The *history-dependent optimal Bellman operator* $L_{\mathrm{h}}^{\star} \colon \mathcal{U} \to \mathcal{U}$ is defined as

$$(L_{\mathrm{h}}\tilde{u})(h) := \max_{a \in \mathcal{A}} \mathbb{E}^{h,a,1}\left[\tilde{r}^{\mathrm{h}}_{l_{\mathrm{h}}(h)} + \tilde{u}\right], \quad \forall \tilde{u} \in \mathcal{U},$$

where the value function $\tilde{u}$ is interpreted as a random variable on defined on the sample space $\Omega = \mathcal{H}$.

Definition 3.2.3. The history-dependent *DP value function* $u_t^{\pi} \in \mathcal{U}$ for a policy $\pi \in \Pi_{\mathrm{HR}}$ and $t \in \mathbb{N}$ is defined as

$$u_t^{\pi} := \begin{cases} 0 & \text{if } t = 0, \\ L_{\mathrm{h}}^{\pi} u_{t-1}^{\pi} & \text{otherwise.} \end{cases}$$

Definition 3.2.4. The history-dependent *DP value function* $u_t^{\star} \in \mathcal{U}$ for $t \in \mathbb{N}$ is defined as

$$u_t^{\star} := \begin{cases} 0 & \text{if } t = 0, \\ L_{\mathrm{h}}^{\star} u_{t-1}^{\star} & \text{otherwise.} \end{cases}$$

Lemma 3.2.5. *Suppose that* $u^1, u^2 \in \mathcal{U}$ *satisfy that* $u^1(h) \geq u^2(h), \forall h \in \mathcal{H}$. *Then*

$$(L_{\mathrm{h}}^{\star} u^1)(h) \geq (L_{\mathrm{h}}^{\pi} u^2)(h), \quad \forall \pi \in \Pi_{\mathrm{HR}}, h \in \mathcal{H}.$$

Proof. From Theorem 1.2.11. □

The following theorem shows the history-dependent value function can be computed by the dynamic program. The following theorem is akin to [Puterman, 2005, theorem 4.2.1].

Theorem 3.2.6. *For each* $\pi \in \Pi_{\mathrm{HR}}$ *and* $t \in \mathbb{N}$*:*

$$\hat{u}_t^{\pi}(h) = u_t^{\pi}(h), \quad \forall h \in \mathcal{H}.$$

Proof. By induction on t. The base case for $t = 0$ follows from the definition. The inductive case for $t + 1$ follows for each $h \in \mathcal{H}$ when $l_{\mathrm{h}}(h) = k_0$ as

$$\begin{aligned}
\hat{u}_{t+1}(h;\pi) &= \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0}\right] && \text{[Definition 3.1.5]} \\
&= \mathbb{E}^{h,\pi,t+1}\left[\mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0} \mid \tilde{a}_{k_0}, \tilde{s}_{k_0+1}\right]\right] && \text{[Theorem 2.6.1]} \\
&= \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{k_0} + \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0+1} \mid \tilde{a}_{k_0}, \tilde{s}_{k_0+1}\right]\right] && \text{[Theorem 2.5.7]} \\
&= \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{k_0} + \mathbb{E}^{\langle h, \tilde{a}_{k_0}, \tilde{s}_{k_0+1}\rangle,\pi,t}\left[\tilde{r}^{\mathrm{h}}_{\geq k_0+1}\right]\right] && \text{[Theorem 2.7.1]} \\
&= \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{k_0} + \hat{u}_t(\langle h, \tilde{a}_{k_0}, \tilde{s}_{k_0+1}\rangle;\pi)\right] && \text{[Definition 3.1.5]} \\
&= \mathbb{E}^{h,\pi,t+1}\left[\tilde{r}^{\mathrm{h}}_{k_0} + u_t^{\pi}(\langle h, \tilde{a}_{k_0}, \tilde{s}_{k_0+1}\rangle)\right] && \text{[inductive assm]} \\
&= \mathbb{E}^{h,\pi,1}\left[\tilde{r}^{\mathrm{h}} + \tilde{u}_t^{\pi}\right] && \text{[Theorem 2.6.2]} \\
&= L_{\mathrm{h}}^{\pi} u_t^{\pi} && \text{[Definition 3.2.1]} \\
&= u_t^{\pi}(h). && \text{[Definition 3.2.3]}
\end{aligned}$$

Also, we use $\tilde{u}_t^{\pi}$ to emphasize when we treat u_t^{π} as a random variable. □

The following theorem is akin to [Puterman, 2005, theorem 4.3.2].

Theorem 3.2.7. *For each* $t \in \mathbb{N}$*:*

$$u_t^{\star}(h) \geq \hat{u}_t(h;\pi), \quad \forall h \in \mathcal{H}, \pi \in \Pi_{\mathrm{HR}}.$$

Proof. By induction on t. The base case is immediate. The inductive case follows for $t+1$ as follows. For each $\pi \in \Pi_{\mathrm{HR}}$:

$$\begin{aligned}
u^\star_{t+1}(h) &= (L^\star_{\mathrm{h}} u^\star_t)(h) && \text{[Definition 3.2.2]}\\
&\geq (L^\pi_{\mathrm{h}} \hat{u}_t(\cdot;\pi))(h) && \text{[ind asm, Lemma 3.2.5]}\\
&= (L^\pi_{\mathrm{h}} u^\pi_t)(h) && \text{[Theorem 3.2.6]}\\
&= u^\pi_t(h) && \text{[Definition 3.1.5]}\\
&= \hat{u}_t(h;\pi). && \text{[Theorem 3.2.6]}
\end{aligned}$$

□

4 Expected Dynamic Program: Markov Policy

4.1 Optimality

We discuss results needed to prove the optimality of Markov policies.

Definition 4.1.1. The set of *independent value functions* is defined as $\mathcal{V} := \mathbb{R}^{\mathcal{S}}$.

Definition 4.1.2. A *Markov Bellman operator* $L^\star \colon \mathcal{V} \to \mathcal{V}$ is defined as

$$(L^\star v)(h) := \max_{a \in \mathcal{A}} \mathbb{E}^{h,a,1}\left[\tilde{r}^{\mathrm{h}} + v(\tilde{s}_1)\right], \quad \forall \tilde{u} \in \mathcal{U},$$

Definition 4.1.3. The *optimal value function* $v^\star_t \in \mathcal{V}, t \in \mathbb{N}$ is defined as

$$v^\star_t := \begin{cases} 0 & \text{if } t = 0\\ (L^\star v^\star_{t-1}) & \text{otherwise.}\end{cases}$$

Theorem 4.1.4. *Suppose that $t \in \mathbb{N}$. Then:*

$$v^\star_t(s_1(h)) = u^\star_t(h), \quad \forall h \in \mathcal{H}.$$

Proof. By induction on t. The base case follows immediately from the definition. The inductive step for $t+1$ follows as:

$$\begin{aligned}
u^\star_{t+1}(h) &= \max_{a \in \mathcal{A}} \mathbb{E}^{h,a,1}\left[\tilde{r}^{\mathrm{h}}_{l_{\mathrm{h}}(h)} + \tilde{u}^\star_t\right] && \text{[Definition 3.2.4]}\\
&= \max_{a \in \mathcal{A}} \mathbb{E}^{h,a,1}\left[\tilde{r}^{\mathrm{h}}_{l_{\mathrm{h}}(h)} + v^\star_t(\tilde{s}_l)\right] && \text{[inductive asm.]}\\
&= \max_{a \in \mathcal{A}} \mathbb{E}^{s_0,a,1}\left[\tilde{r}^{\mathrm{h}}_{l_{\mathrm{h}}(h)} + v^\star_t(\tilde{s}_l)\right] && \text{[Theorem 2.5.8]}\\
&= \max_{a \in \mathcal{A}} \mathbb{E}^{s_0,a,1}\left[\tilde{r}^{\mathrm{h}} + v^\star_t(\tilde{s}_l)\right] && \text{[Theorem 2.5.1]}\\
&= v^\star_{t+1}(s_1(h)) && \text{[Definition 4.1.3].}
\end{aligned}$$

□

Definition 4.1.5. The *optimal finite-horizon policy* $\pi^\star_t, t \in \mathbb{N}$ is defined as

$$\pi^\star_t(k,s) := \begin{cases} \arg\max_{a \in \mathcal{A}} \mathbb{E}^{s,a,1}\left[\tilde{r}^{\mathrm{h}} + v^\star_{t-k}(\tilde{s}_1)\right] & \text{if } k \leq t,\\ a_0 & \text{otherwise,}\end{cases}$$

where a_0 is an arbitrary action.

Theorem 4.1.6. *Assume a horizon $T \in \mathbb{N}$. Then:*

$$v^\star_{T-k}(s_1(h)) = u^{\pi^\star}_{T-k}(h), \quad \forall h \in \{h \in \mathcal{H} \mid l_{\mathrm{h}}(h) \leq T\},$$

$k := l_{\mathrm{h}}(h)$.

Proof. Fix some $T \in \mathbb{N}$. By induction on k from $k = T$ to $k = 0$. The base case is immediate from the definition. We prove the inductive case for $k-1$ from k as

$$\begin{aligned}
u^{\pi^\star_T}_{T-k+1}(h) &= \mathbb{E}^{h,\pi^\star_T,1}\left[\tilde{r}^{\mathrm{h}}_k + \tilde{u}^{\pi^\star_T}_{T-k}\right] && \text{[Definition 3.2.1]}\\
&= \mathbb{E}^{h,a^\star,1}\left[\tilde{r}^{\mathrm{h}}_k + \tilde{u}^{\pi^\star_T}_{T-k}\right] && \text{[???]}\\
&= \mathbb{E}^{h,a^\star,1}\left[\tilde{r}^{\mathrm{h}}_k + v^\star_{T-k}(\tilde{s}_1)\right] && \text{[ind asm]}\\
&= \mathbb{E}^{s_0,a^\star,1}\left[\tilde{r}^{\mathrm{h}} + v^\star_{T-k}(\tilde{s}_1)\right] && \text{[Theorem 2.5.8]}\\
&= \max_{a\in\mathcal{A}} \mathbb{E}^{s_0,a,1}\left[\tilde{r}^{\mathrm{h}} + v^\star_{T-k}(\tilde{s}_1)\right] && \text{[???]}\\
&= v^\star_{T-k+1}(s_0). && \text{[Definition 4.1.3]}
\end{aligned}$$

Here, $k := l_{\mathrm{h}}(h)$ $a^\star := \pi^\star_T(k, s_0)$ and $s_0 := s_1(h)$ □

4.2 Evaluation

We discuss results pertinent to the evaluation of Markov policies.

Markov value functions depend on the length of the history.

Definition 4.2.1. The set of *independent value functions* is defined as $\mathcal{V}_{\mathrm{M}} := \mathbb{R}^{\mathbb{N}\times\mathcal{S}}$.

Definition 4.2.2. A *Markov policy Bellman operator* $L^\pi \colon \mathcal{V}_{\mathrm{M}} \to \mathcal{V}_{\mathrm{M}}$ for $\pi \in \Pi$ is defined as

$$(L^\pi v)(k, s) := \max_{a\in\mathcal{A}} \mathbb{E}^{s,a,1}\left[\tilde{r}^{\mathrm{h}} + v(k+1, \tilde{s}_1)\right], \quad \forall v \in \mathcal{V}_{\mathrm{M}}, k \in \mathbb{N}, s \in \mathcal{S}.$$

Definition 4.2.3. The *Markov policy value function* $v^\pi_t \in \mathcal{V}_{\mathrm{M}}, t \in \mathbb{N}$ for $\pi \in \Pi_{\mathrm{MD}}$ is defined as

$$v^\pi_t := \begin{cases} 0 & \text{if } t = 0,\\ (L^\pi v^\pi_{t-1}) & \text{otherwise.}\end{cases}$$

References

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.