Strategies in POMDPs with Stage Duration

Ivan Novikov

Université Paris 1 Panthéon-Sorbonne

30/03/2026

Introduction

- We study POMDPs: MDPs in which the state is not fully observable, but the decision maker observes a signal on the state.
- Given a base POMDP G_1, we consider the POMDP G_h with stage duration h representing the time interval between two consecutive actions. We denote by $V_\lambda(h)$ the λ-discounted value of G_h. $V_\lambda(h)$ was studied in the literature, especially its limit as $h \to 0$.
- Our goal is to study the global behavior of the asymptotic value $V(h) := \lim_{\lambda \to 0} V_\lambda(h)$ as a function of h.

Table of contents

POMDPs (1)

A partially observable Markov decision process (POMDP) is a 6-tuple (S, A, Ω, f, g, P), where:

- S is the finite set of signals;
- Ω is the finite set of states;
- A is the finite set of actions;
- $g : \Omega \times A \to \mathbb{R}$ is the stage payoff function;
- $P : \Omega \times A \to \Delta(\Omega)$ is the transition probability function;
- $f : \Omega \to S$ is the signaling function.

POMDPs (2)

The POMDP (S, A, Ω, f, g, P) proceeds in stages as follows. Before the first stage, an initial state $\omega_1 \in \Omega$ is chosen according to some probability law $p_0 \in \Delta(\Omega)$. At each stage $n \in \mathbb{N}^*$:

1. The current state is ω_n, and is unobserved. The decision maker observes the signal $s_n = f(\omega_n)$ and remembers the previous signals and actions;
2. The decision maker chooses a mixed action $x_n \in \Delta(A)$. Pure action $a_n \in A$ is drawn according to x_n;
3. The decision maker obtains a payoff $g(\omega_n, a_n)$. The new state ω_{n+1} is chosen according to the probability law $P(\cdot \mid \omega_n, a_n)$.

POMDPs (3)

- A history of length t is $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t) \in (S \times A)^{t-1} \times S$;
- H_t is the set of all histories of length t;
- A strategy σ is a function

$$\sigma : \bigcup_t H_t \to \Delta(A).$$

- Let $\lambda \in (0, 1]$. The λ-discounted payoff:

$$\mathbb{E}^p_\sigma \left(\sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1} g(\omega_i, a_i) \right).$$

- Value $v_\lambda : \Delta(\Omega) \to \mathbb{R}$:

$$v_\lambda(p) = \sup_\sigma \mathbb{E}^p_\sigma \left(\sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1} g(\omega_i, a_i) \right).$$

Table of contents

POMDPs with stage duration (1)

- Fix a base POMDP $G_1 = (S, A, \Omega, f, g, P)$.
- In the POMDP G_h with stage duration h, the leaving probabilities and the discount factor are scaled by $h \in (0, 1)$.
- $P_h(\cdot \mid \omega, a) = hP(\cdot \mid \omega, a) + (1 - h)\delta_\omega(\cdot)$, where

$$\delta_\omega(\omega') = \begin{cases} 1, & \text{if } \omega' = \omega; \\ 0, & \text{otherwise.} \end{cases}$$

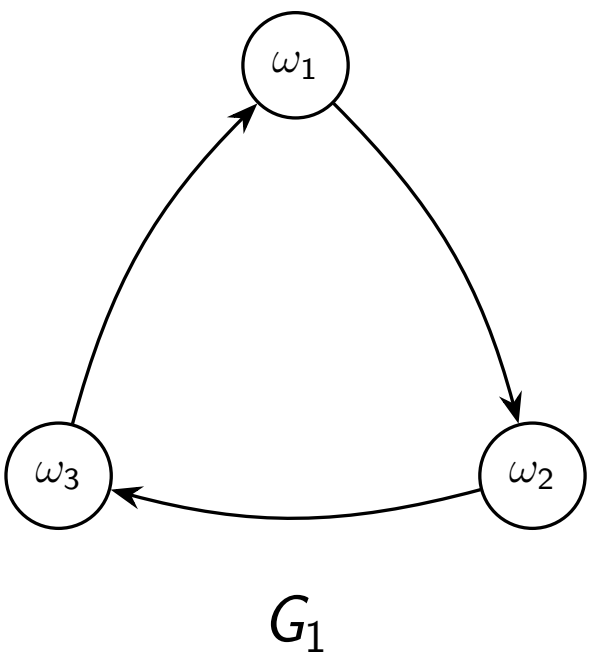
- The λ-discounted payoff of G_h is

$$\mathbb{E}_\sigma^h \left(\sum_{i=1}^{\infty} \lambda h (1 - \lambda h)^{i-1} g(\omega_i, a_i) \right).$$

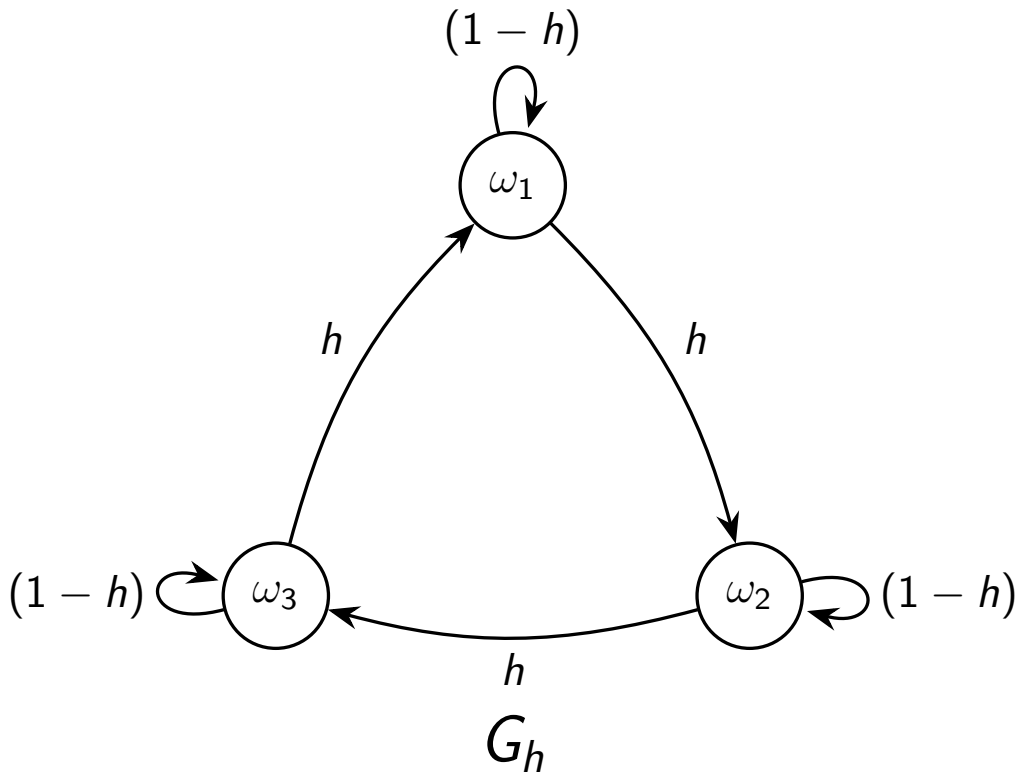
- It depends on the stage duration h, the strategy σ, and the initial probability distribution p.
- $V_\lambda(h)$ is the discounted value of G_h.

POMDPs with stage duration (2)

- h can be thought of as a time period between actions: when h is small, G_h approximates the continuous-time behavior, and $\lim_{h\to 0} V_\lambda(h)$ can be seen as the λ-discounted value of the continuous-time MDP.



G_1



G_h

Papers on the stage duration

- Neyman A, "Stochastic games with short-stage duration" (2013);
- Sorin S and Vigeral G, "Operator approach to values of stochastic games with varying stage duration" (2016).
- I. N., Zero-Sum State-Blind Stochastic Games with Vanishing Stage Duration (2025);
- I.N., Asymptotic Value in Zero-Sum Stochastic Games with Vanishing Stage Duration and Public Signals (2024).

Similar model:

- Sorin S, "Limit Value of Dynamic Zero-Sum Games with Vanishing Stage Duration" (2018);
- Cardaliaguet P, Rainer C, Rosenberg D, Vieille N, "Markov Games with Frequent Actions and Incomplete Information—The Limit Case" (2016);
- Gensbittel F, "Continuous-time limit of dynamic games with incomplete information and a more informed player" (2016).

Our goal

- We consider the asymptotic value of G_h:

$$V(h) = \lim_{\lambda \to 0} V_\lambda(h)$$
$$= \lim_{\lambda \to 0} \sup_{\sigma} \left(\mathbb{E}_\sigma^h \left(\sum_{i=1}^{\infty} \lambda h (1 - \lambda h)^{i-1} g(\omega_i, a_i) \right) \right).$$

- The limit exists by Rosenberg, Solan and Vieille 2002.
- We aim to study $V(h)$ as a function of h.

Table of contents

The case of full state observation

- $V(h) = \lim_{\lambda \to 0} V_\lambda(h)$.
- Sorin and Vigeral (2016) have proven that

$$V_\lambda(h) = V_{\frac{\lambda}{1+\lambda-\lambda h}}(1).$$

- Hence

$$V(h) = \lim_{\lambda \to 0} V_\lambda(h) = \lim_{\lambda \to 0} V_{\frac{\lambda}{1+\lambda-\lambda h}}(1) = \lim_{\lambda \to 0} V_\lambda(1) = V(1).$$

- Thus $V(h)$ is a constant in the case of full state observation.

General case with signals.

- $V(h) = \lim_{\lambda \to 0} V_\lambda(h)$.
- $V(h)$ is not a constant anymore.
- In that case, $V(h)$ is a nondecreasing function.
- As a consequence, the continuous-time limit $\lim_{h \to 0} V(h)$ exists.

Introduction of $R(\sigma, h)$.

- The *expected long-run average payoff* is

$$R(\sigma, h) := \liminf_{T\to\infty} \mathbb{E}^h_\sigma \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right).$$

- We have

$$\begin{aligned} V(h) &= \lim_{\lambda\to 0} \left[\sup_\sigma \mathbb{E}^h_\sigma \left(\lambda h \sum_{i=1}^{\infty} (1-\lambda h)^{i-1} g(\omega_i, a_i) \right) \right] \\ &= \lim_{\lambda\to 0} \left[\sup_\sigma \mathbb{E}^h_\sigma \left(\lambda \sum_{i=1}^{\infty} (1-\lambda)^{i-1} g(\omega_i, a_i) \right) \right] \\ &= \sup_\sigma \left[\liminf_{T\to\infty} \mathbb{E}^h_\sigma \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right) \right], \end{aligned}$$

where the last equality follows from Rosenberg, Solan and Vieille 2002. Hence $V(h) = \sup_\sigma R(\sigma, h)$.

Theorem 1

$$R(\sigma, h) := \liminf_{T\to\infty} \mathbb{E}_\sigma^h \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right).$$

Theorem

Let $h \in (0,1)$ and let σ be a strategy in G_h. Then there exists a strategy $\widehat{\sigma}$ in G_1 such that

$$R(\sigma, h) = R(\widehat{\sigma}, 1).$$

- This means that if some strategy in G_h produces the asymptotic payoff v, there is a strategy in G_1 that also produces v.

Proof of Theorem 1: preparation (1)

- In G_h, at the end of each stage, the next state is drawn according to P with probability h; otherwise, the state remains unchanged (drawn according to δ_ω) with probability $1-h$.
- X_i is the random variable with

$$X_i \sim \mathit{Bernoulli}(h) \iff \Big(\mathbb{P}(X_i = 1) = h \quad \text{and} \quad \mathbb{P}(X_i = 0) = 1 - h\Big)$$

- X_i indicates whether the transition at the end of the i-th stage is governed by P (if $X_i = 1$) or by δ_ω (if $X_i = 0$).
- T_i is the random variable

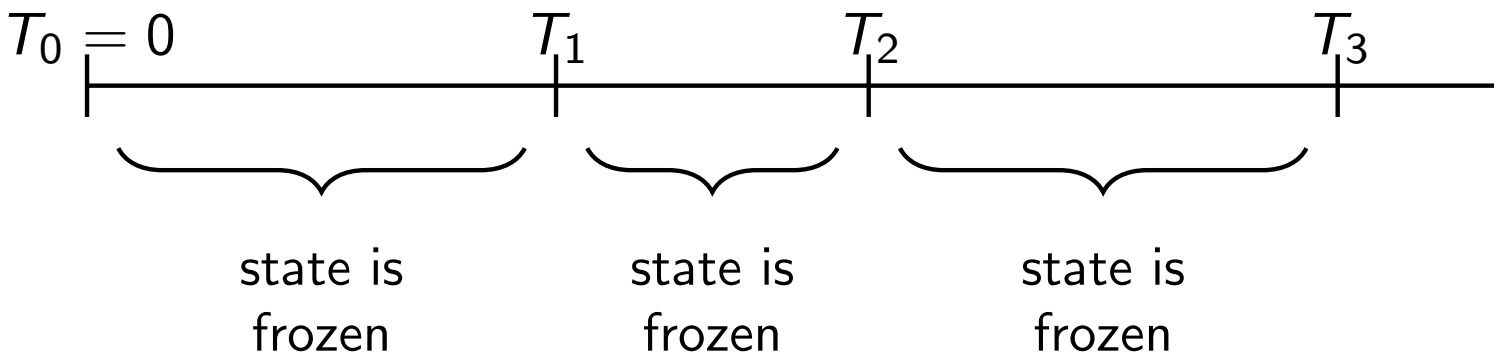
$$T_0 = 0;$$
$$T_i = \inf\{n > T_{i-1} \mid X_n = 1\} \quad \text{for } i > 0.$$

Proof of Theorem 1: preparation (2)

- T_i is the random variable

$$T_0 = 0;$$
$$T_i = \inf\{n > T_{i-1} \mid X_n = 1\} \quad \text{for } i > 0.$$

- T_i is the stage number at which the state transition is governed by P for the i-th time.



Proof of Theorem 1: definition of $\widehat{\sigma}$

Fix $k \in \mathbb{N}^*$. For a fixed infinite history $\mathcal{H} = (s'_1, a'_1, s'_2, \ldots)$ in G_h, the filtered history $\mathcal{H}_k^{fil}$ of length k is the random vector

$$\mathcal{H}_k^{fil} = (s'_1, a'_{T_1}, s'_{T_1+1}, a'_{T_2}, \ldots, a'_{T_{k-1}}, s'_{T_{k-1}+1}).$$

Definition

Let $\eta_k = (s_1, a_1, s_2, \ldots, a_{k-1}, s_k)$ be a history of length k in G_1, and let $\widehat{a} \in A$. The strategy $\widehat{\sigma}$ is defined by

$$\widehat{\sigma}(\eta_k)(\widehat{a}) := \mathbb{P}_\sigma^h(a'_{T_k} = \widehat{a} \mid \mathcal{H}_k^{fil} = \eta_k).$$

If $\mathbb{P}_\sigma^h(\mathcal{H}_k^{fil} = \eta_k) = 0$, then we define $\widehat{\sigma}(\eta_k)$ as an arbitrary fixed mixed action.

Technical Lemma 1

Lemma

For each $k \in \mathbb{N}^*$, we have

$$\mathbb{E}^1_{\widehat{\sigma}} g(\omega_k, a_k) = \mathbb{E}^h_{\sigma} g(\omega_{T_k}, a_{T_k}).$$

- The expected k-th stage payoff (under $\widehat{\sigma}$) in G_1 coincides with the expected T_k-th stage payoff (under σ) in G_h.
- $\widehat{\sigma}$ is constructed specifically for this lemma to be true.

Technical Lemma 2

Lemma

For each $k \in \mathbb{N}^*$, we have

$$\mathbb{E}_\sigma^h \left(\sum_{j=T_{k-1}+1}^{T_k} g(\omega_j, a_j) \right) = \frac{1}{h} \mathbb{E}_\sigma^h g(\omega_{T_k}, a_{T_k}).$$

- In G_h, the expected cumulative payoff during the waiting period between stages T_{k-1} and T_k coincides with the expected payoff at stage T_k, multiplied by the expectation of the waiting period $T_k - T_{k-1}$:

$$\mathbb{E}_\sigma^h \left(\sum_{j=T_{k-1}+1}^{T_k} g(\omega_j, a_j) \right) = \mathbb{E}(T_k - T_{k-1}) \cdot \mathbb{E}_\sigma^h g(\omega_{T_k}, a_{T_k}).$$

Technical Lemma 3

Lemma

Let $t_k = \lfloor k/h \rfloor$. For each $k \in \mathbb{N}^*$, we have

$$\liminf_{k\to+\infty} \mathbb{E}_\sigma^h \left(\frac{1}{t_k} \sum_{j=1}^{t_k} g(\omega_j, a_j) \right) = \liminf_{k\to+\infty} \mathbb{E}_\sigma^h \left(\frac{h}{k} \sum_{j=1}^{T_k} g(\omega_j, a_j) \right).$$

- We can replace the random variable T_k with

$$t_k = \lfloor \mathbb{E}(T_k) \rfloor = \lfloor k/h \rfloor.$$

Technical Lemma 4

Lemma

Let $M \in \mathbb{N}^*$. Let $\{x_n\}_{n=1}^{\infty}$ be a sequence and $\{x_{n_k}\}_{k=1}^{\infty}$ be a subsequence of $\{x_n\}_{n=1}^{\infty}$. Suppose that

$$\lim_{j\to\infty}(x_{j+1} - x_j) = 0 \qquad \textit{and} \qquad |n_{j+1} - n_j| \leq M \textit{ for all } j \in \mathbb{N}^*.$$

We then have

$$\liminf_{n\to\infty} x_n = \liminf_{k\to\infty} x_{n_k}.$$

Proof of Theorem 1 – Part 1

$$R(\sigma, h) := \liminf_{T\to\infty} \mathbb{E}^h_\sigma \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right).$$

Theorem

$R(\sigma, h) = R(\widehat{\sigma}, 1)$.

Proof. By Lemma 1, we have

$$\begin{aligned} R(\widehat{\sigma}, 1) &= \liminf_{T\to+\infty} \mathbb{E}^1_{\widehat{\sigma}} \left(\frac{1}{T} \sum_{k=1}^{T} g(\omega_k, a_k) \right) \\ &= \liminf_{T\to+\infty} \left(\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}^1_{\widehat{\sigma}} g(\omega_k, a_k) \right) \\ &= \liminf_{T\to+\infty} \left(\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}^h_{\sigma} g(\omega_{T_k}, a_{T_k}) \right). \end{aligned}$$

Proof of Theorem 1 – Part 2

We now aim to compute $R(\sigma, h)$.
Consider the sequence $\{x_t\}_{t=1}^{\infty}$ and the subsequence $\{x_{t_k}\}_{k=1}^{\infty}$ of $\{x_t\}_{t=1}^{\infty}$, where

$$x_t := \mathbb{E}_\sigma^h \left(\frac{1}{t} \sum_{i=1}^{t} g(\omega_i, a_i) \right) \quad \text{and} \quad t_k = \lfloor k/h \rfloor = \lfloor \mathbb{E}(T_k) \rfloor.$$

By Lemma 4

$$\begin{aligned} R(\sigma, h) &= \liminf_{T \to +\infty} \mathbb{E}_\sigma^h \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right) \\ &= \liminf_{k \to +\infty} \mathbb{E}_\sigma^h \left(\frac{1}{t_k} \sum_{j=1}^{t_k} g(\omega_j, a_j) \right), \end{aligned}$$

Proof of Theorem 1 – Part 3

By Lemmas 2 and 3, we have

$$\begin{aligned}
\liminf_{k\to+\infty} \mathbb{E}_\sigma^h \left(\frac{1}{t_k} \sum_{j=1}^{t_k} g(\omega_j, a_j) \right) &= \liminf_{k\to+\infty} \mathbb{E}_\sigma^h \left(\frac{h}{k} \sum_{j=1}^{T_k} g(\omega_j, a_j) \right) \\
&= \liminf_{K\to+\infty} \mathbb{E}_\sigma^h \left(\frac{h}{K} \sum_{k=1}^{K} \sum_{j=T_{k-1}+1}^{T_k} g(\omega_j, a_j) \right) \\
&= \liminf_{K\to+\infty} \left(\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_\sigma^h g(\omega_{T_k}, a_{T_k}) \right).
\end{aligned}$$

We have proved that

$$R(\sigma, h) = R(\widehat{\sigma}, 1).$$

Corollaries of Theorem 1 – Part 1

$$R(\sigma, h) := \liminf_{T\to\infty} \mathbb{E}^h_\sigma \left(\frac{1}{T} \sum_{i=1}^{T} g(\omega_i, a_i) \right).$$

Theorem

Let $h \in (0, 1)$ and let σ be a strategy in G_h. Then there exists a strategy $\widehat{\sigma}$ in G_1 such that

$$R(\sigma, h) = R(\widehat{\sigma}, 1).$$

Corollary

Let $0 < h_1 < h_2 \leq 1$ and let σ be a strategy in G_{h_1}. Then there exists a strategy $\widehat{\sigma}$ in G_{h_2} such that

$$R(\sigma, h_1) = R(\widehat{\sigma}, h_2).$$

Corollaries of Theorem 1 – Part 2

Corollary

Let $0 < h_1 < h_2 \leq 1$ and let σ be a strategy in G_{h_1}. Then there exists a strategy $\widehat{\sigma}$ in G_{h_2} such that

$$R(\sigma, h_1) = R(\widehat{\sigma}, h_2).$$

- G_{h_1} is the POMDP with stage duration h_1 relative to the base POMDP G_1.
- G_{h_1} is also the POMDP with stage duration h_1/h_2 relative to the base POMDP G_{h_2}.
- This is true because

$$P_{h_1} = (1-h_1)Id + h_1 P_1 = \left(1 - \frac{h_1}{h_2}\right) Id + \frac{h_1}{h_2}((1-h_2)\,Id + h_2 P_1)$$

$$= \left(1 - \frac{h_1}{h_2}\right) Id + \frac{h_1}{h_2} P_{h_2}.$$

- The corollary follows from Theorem 1.

Corollaries of Theorem 1 – Part 3

Corollary

The function $h \mapsto V(h)$ is nondecreasing.

Proof.

Follows from the previous corollary: for any $0 < h_1 \leq h_2 \leq 1$

$$\begin{aligned} V(h_1) &= \sup\{R(\sigma, h_1) \mid \sigma \text{ is a strategy in } G_{h_1}\} \\ &\leq \sup\{R(\sigma, h_2) \mid \sigma \text{ is a strategy in } G_{h_2}\} = V(h_2). \end{aligned}$$ □

Corollary

The limit $\lim_{h \to 0} V(h)$ exists.

Proof.

Follows from the previous corollary. □

Table of contents

Lower-semicontinuity on $(0,1)$

- $V(h) = \lim_{\lambda \to 0} V_\lambda(h)$.
- Natural question: is $V(h)$ continuous in h?

Proposition

The function $h \mapsto V(h)$ is lower semi-continuous on $(0,1)$.

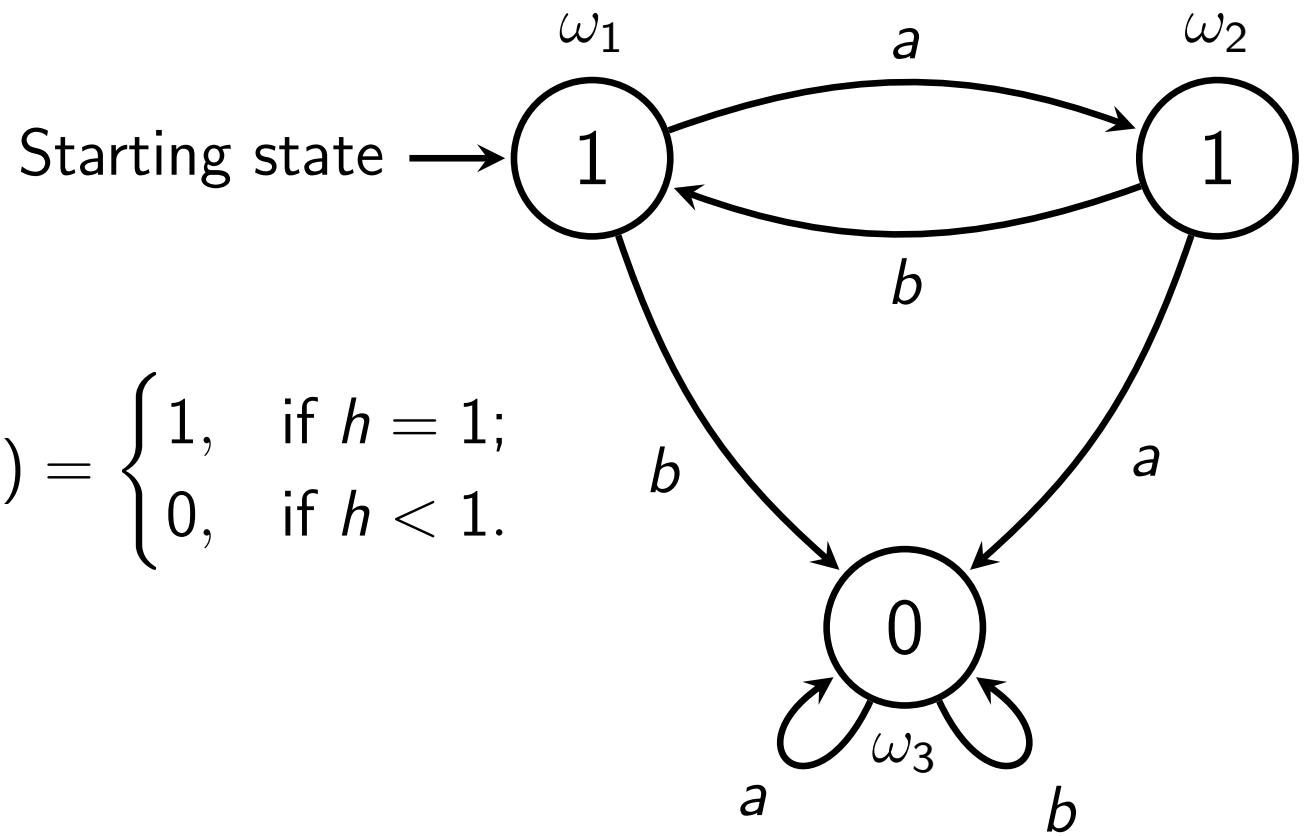
Proof.

If $h' \in (0,1)$ and $\varepsilon > 0$ is sufficiently small, then the support of the transition probabilities of the POMDPs in the family $\{G_h\}_{h \in (h'-\varepsilon, h'+\varepsilon)}$ is identical. The proposition now follows directly from the article "Finite-Memory Strategies in POMDPs with Long-Run Average Objectives" by Krishnendu Chatterjee, Raimundo Saona and Bruno Ziliotto (2022). □

No lower-semicontinuity at 1

However, it is possible to introduce new transitions when moving from $h = 1$ to $h < 1$. In fact, it is possible that $V(h)$ is not lower semi-continuous at 1, even in the state-blind case. Example:

$$V(h) = \begin{cases} 1, & \text{if } h = 1; \\ 0, & \text{if } h < 1. \end{cases}$$



Continuity on $(0, 1)$

It is not yet known whether $h \mapsto V(h)$ is continuous on $(0, 1)$. In general, it is possible for the asymptotic value to be discontinuous even if no new transitions are introduced, as shown in the article by Krishnendu Chatterjee, Raimundo Saona and Bruno Ziliotto (2022). However, I was unable to adapt the counterexample to construct one for $V(h)$.

Conclusion

- Given a base POMDP G_1, we considered the POMDP G_h with stage duration h representing the time interval between two consecutive actions. We denoted by $V(h) := \lim_{\lambda \to 0} V_\lambda(h)$ the asymptotic value of G_h.
- We showed that any strategy in G_h can be mimicked by a strategy in G_1.
- As a consequence, $V(h)$ is nondecreasing.
- As another consequence, the continuous-time limit $\lim_{h \to 0} V(h)$ exists.
- The paper: `https://arxiv.org/pdf/2603.16055`

This is all.

Thank you!