APPENDIX

A MDPS WITH DELAY: A DEGRADATION EXAMPLE

A.1 PROOF OF PROPOSITION 3.1

Without loss of generality, assume $p \in [0.5, 1]$. It is easy to see than in our 2-state MDP, the optimal policy selects a_0 if the most likely state of the system is s_0, and a_1 if it is s_1. Since $p \geqslant 0.5$, the most-likely state of the system when observing s_0 is s_0 if m is even, and s_1 if m is odd. The same logic holds when observing s_1. Therefore, if m is even, $\pi^*(s_0) = a_0, \pi^*(s_1) = a_1$. Otherwise, $\pi^*(s_0) = a_1, \pi^*(s_1) = a_0$. Note that re-iterating the rest of the proof with a randomized policy (of the form $\pi(a|s) \in (0, 1),\ \forall a \in \{a_0, a_1\},\ \forall s \in \{s_0, s_1\}$) yields sub-optimal return. Hence, in this example it is enough to consider deterministic policies.

The expected reward at time $t + m$ with action $\pi^*(s_t)$ selected at s_t is

$$\begin{aligned} R^*_{t+m}(s_t) :&= \mathbb{E}_{s_{t+m}|s_t}[r(s_{t+m}, \pi^*(s_t))] \\ &= r(s_0, \pi^*(s_t))\mathbb{P}(s_{t+m} = s_0|s_t) + r(s_1, \pi^*(s_t))\mathbb{P}(s_{t+m} = s_1|s_t). \end{aligned} \tag{2}$$

From here on, we inspect the case where $s_t = s_0$ for brevity. By symmetry, identical arguments apply if $s_t = s_1$. If m is even, $r(s_0, \pi^*(s_0)) = 1$ and $r(s_1, \pi^*(s_0)) = 0$. If m is odd, $r(s_0, \pi^*(s_0)) = 0$ and $r(s_1, \pi^*(s_0)) = 1$. Thus, using (2),

$$R^*_{t+m}(s_0) = \begin{cases} \mathbb{P}(s_{t+m} = s_0|s_t = s_0) & \text{if } m \text{ is even,} \\ \mathbb{P}(s_{t+m} = s_1|s_t = s_0) & \text{if } m \text{ is odd.} \end{cases} \tag{3}$$

Note that, by construction, the transition probabilities are independent of the actions. Specifically, if m is even,

$$\mathbb{P}(s_{t+m} = s_0|s_t = s_0) = \sum_{k\ even}^{m} \binom{m}{k} p^k (1-p)^{m-k}, \tag{4}$$

since we count the possibilities of an even number of jumps between the two states. Similarly, if m is odd,

$$\mathbb{P}(s_{t+m} = s_1|s_t = s_0) = \sum_{k\ odd}^{m} \binom{m}{k} p^k (1-p)^{m-k}. \tag{5}$$

Also note that the same applies for $s_t = s_1$, i.e.,

$$R^*_{t+m}(s_0) = R^*_{t+m}(s_1) \quad \forall t, \tag{6}$$

and that these probabilities are independent of t, i.e.,

$$R^*_{t+m}(s_0) = R^*_{t+m+k}(s_0) \quad \forall k \in \mathbb{N}. \tag{7}$$

Next, we compute the optimal return starting from s_0 :

$$\begin{aligned} v^*_m(s_0) :&= \mathbb{E}^{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+m}, \pi^*(s_t))|s_{t=0} = s_0\right] \\ &= R^*_m(s_0) + \gamma\left[\mathbb{P}(s_{t=1} = s_0)R^*_{m+1}(s_0) + \mathbb{P}(s_{t=1} = s_1))R^*_{m+1}(s_1)\right] \\ &\quad + \gamma^2\left(\mathbb{P}(s_{t=2} = s_0)R^*_{m+2}(s_0) + \mathbb{P}(s_{t=2} = s_1))R^*_{m+2}(s_1)\right) + \dots \\ &= R^*_m(s_0) + \gamma R^*_{m+1}(s_0)\left(\mathbb{P}(s_{t=1} = s_0) + \mathbb{P}(s_{t=1} = s_1))\right) \\ &\quad + \gamma^2 R^*_{m+2}(s_0)\left(\mathbb{P}(s_{t=2} = s_0) + \mathbb{P}(s_{t=2} = s_1))\right) + \dots \\ &= \frac{1}{1-\gamma} R^*_m(s_0), \end{aligned} \tag{8}$$

where in the second relation we used (6), and in the last relation (7) as well as $\mathbb{P}(s_t = s_0) = 1 - \mathbb{P}(s_t = s_1) \quad \forall t$.

Plugging (4) and (5) into (3), together with (7), (8) and (6) gives the optimal return

$$v^*_m(s_0) = v^*_m(s_1) = \begin{cases} \frac{1}{1-\gamma}\sum_{k\ even}^{m} \binom{m}{k} p^k (1-p)^{m-k}, & \text{if } m \text{ is even,} \\ \frac{1}{1-\gamma}\sum_{k\ odd}^{m} \binom{m}{k} p^k (1-p)^{m-k}, & \text{if } m \text{ is odd.} \end{cases} \tag{9}$$

This concludes the first part of the proof.

In the second part, we shall now derive a simpler expression for (9) which can then be analyzed to determine monotonicity w.r.t. m and p. Observe that

$$(1-2p)^m = (-p+1-p)^m = \sum_k^m \binom{m}{k}(-p)^k(1-p)^{m-k}$$
$$= \sum_{k\ even}^m \binom{m}{k}p^k(1-p)^{m-k} - \sum_{k\ odd}^m \binom{m}{k}p^k(1-p)^{m-k}.$$

Since

$$\sum_{k\ even}^m \binom{m}{k}p^k(1-p)^{m-k} + \sum_{k\ odd}^m \binom{m}{k}p^k(1-p)^{m-k} = 1,$$

we have that

$$\sum_{k\ even}^m \binom{m}{k}p^k(1-p)^{m-k} = \frac{1}{2}\left(1+(1-2p)^m\right), \tag{10}$$

$$\sum_{k\ odd}^m \binom{m}{k}p^k(1-p)^{m-k} = \frac{1}{2}\left(1-(1-2p)^m\right). \tag{11}$$

Denote $a := -(1-2p)$, remember that $0 \leqslant a \leqslant 1$, and let $m = 2n$ (resp. $m = 2n+1$) with $n \in \mathbb{N}$ when m is even (resp. odd). Then

$$\frac{1}{2}\left(1+(1-2p)^m\right) = \frac{1}{2}\left(1+(a^2)^n\right) \tag{12}$$

and

$$\frac{1}{2}\left(1-(1-2p)^m\right) = \frac{1}{2}\left(1+a(a^2)^n\right). \tag{13}$$

Both (12) and (13) obviously monotonically decrease with n, so the even and odd subsequences are monotone. Also, since $a \leqslant 1$, (13) $\leqslant$ (12), which gives that the whole sequence itself is monotone in m. Lastly, as p increases a increases. This obviously causes both (12) and (13) to increase as well.

B The Standard Approach: Augmentation

B.1 The augmented MDP

Let the augmented state space $\mathcal{X}_m := \mathcal{S} \times \mathcal{A}^m$. Then, $x_t := (s_t, a_t^{-1}, \cdots, a_t^{-m}) \in \mathcal{X}_m$ is an extended state, where a_t^{-i} is the i-th pending action at time t. It means that in the following step, $t+1$, action a_t^{-m} will be executed independently of the present action selection. Accordingly, a new transition function for $\mathcal{X}_m$ is induced by the original transition matrix P and m-step delay. More explicitly, for $(x, a, x') \in \mathcal{X}_m \times \mathcal{A} \times \mathcal{X}_m$ we have

$$F(x'|x,a) = \begin{cases} P(e_1^\top x'|e_1^\top x, e_{m+1}^\top x) & \text{if } e_2^\top x' = a \text{ and } e_{i+1}^\top x' = e_i^\top x\ \forall i \in [2:m], \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

where $e_i \in \{0,1\}^{m+1}$ is the elementary vector with 1 only in its i-th coordinate.[3] Similarly, the reward function on the augmented state-space is:

$$g(x,a) = r(e_1^\top x, e_{m+1}^\top x). \tag{15}$$

Note that g does not depend on the newly decided action $a \in \mathcal{A}$, but rather on the first and last coordinates of the current state $x \in \mathcal{X}_m$. This leads us to the following definition.

[3] Throughout this work, we assume without loss of generality that $s \in \mathcal{S}$ is a scalar, to simplify notation of inner products with e_i. This assumption is non-limiting since any multi-dimensional state space can be easily transformed to single-dimensional via enumeration as it is finite.

B.2 mA-PI ALGORITHM

Let the set of greedy policies w.r.t. $v \in \mathbb{R}^{|\mathcal{X}_m|}$: $\bar{\mathcal{G}}(v) := \{\bar{\pi} \in \bar{\Pi}_m : \bar{T}^{\bar{\pi}} v = \bar{T} v\}$.

Algorithm 1 mA-PI

Initialize: $\bar{\pi}_0 \in \bar{\Pi}_m, k = 0$
while $\bar{\pi}_k$ is changing **do**
  $v_k \leftarrow v^{\bar{\pi}_k}$
  $\bar{\pi}_{k+1} \leftarrow$ any element of $\bar{\mathcal{G}}(v_k)$
  $k \leftarrow k + 1$
Return: $\bar{\pi}_k, v_k$

B.3 CONVERGENCE OF mA-PI

Convergence of mA-PI directly follows from the improvement property of greedy policies, which we prove below.

Proposition (mA Evaluation and Improvement)**.** *(i) For any $x \in \mathcal{X}_m$ and $\bar{\pi} \in \bar{\Pi}_m$, the augmented value function $v^{\bar{\pi}}$ satisfies the Bellman recursion $v^{\bar{\pi}}(x) = \bar{T}^{\bar{\pi}} v^{\bar{\pi}}(x)$.*
(ii) The optimal augmented value $\bar{v}^*$ is the unique fixed point of $\bar{T}$. Furthermore, if $\bar{\pi}^*$ is preserving, i. e., $\bar{\pi}^* \in \arg\max_{\bar{\pi}} \{g^{\bar{\pi}} + \gamma F^{\bar{\pi}} \bar{v}^*\}$, then $\bar{\pi}^*$ is optimal and thus, $\bar{v}^* = \bar{v}^{\bar{\pi}^*}$.

Proof. Using standard Bellman recursion on the augmented MDP, we can write

$$
\begin{aligned}
v^{\bar{\pi}}(x) &= \mathbb{E}^{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t g(x_t, a_t) | x_0 = x\right] \\
&= \mathbb{E}^{\bar{\pi}}\left[g(x_0, a_0) + \sum_{t=1}^{\infty} \gamma^t g(x_t, a_t) | x_0 = x\right] \\
&= g(e_1^\top x, e_{m+1}^\top x) + \gamma \mathbb{E}^{\bar{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t g(x_{t+1}, a_{t+1}) | x_0 = x\right] \\
&= g(e_1^\top x, e_{m+1}^\top x) + \gamma \sum_{(x', a) \in \mathcal{X}_m \times \mathcal{A}} \bar{\pi}(a|x) F(x'|x, a) v^{\bar{\pi}}(x') \\
&= \bar{T}^{\bar{\pi}} v^{\bar{\pi}}(x)
\end{aligned}
$$

which ends the proof of Claim (i).

Note that by definition of g and F as in Equations (15) and (14) respectively, the sum can be reformulated as follows:

$$
\begin{aligned}
v^{\bar{\pi}}(x) &= r(e_1^\top x, e_{m+1}^\top x) + \gamma \sum_{\substack{(x', a) \in \mathcal{X}_m \times \mathcal{A}: \\ e_{i+1}^\top x' = e_i^\top x \text{ for } i \in [2:m]; \\ e_2^\top x' = a}} \bar{\pi}(a|x) P(e_1^\top x' | e_1^\top x, e_{m+1}^\top x) v^{\bar{\pi}}(x') \\
&= r(e_1^\top x, e_{m+1}^\top x) + \gamma \sum_{(s', a) \in \mathcal{S} \times \mathcal{A}} \bar{\pi}(a|x) P(s' | e_1^\top x, e_{m+1}^\top x) v^{\bar{\pi}}(s', a, e_2^\top x, \cdots, e_m^\top x)
\end{aligned}
$$

Claim (ii) relies on classical theory of discounted MDPs, applied to the augmented MDP (Puterman, 2014). □

B.4 PROOF OF THEOREM 4.1

First, we give a general lower bound to the classic PI algorithm by Howard (Howard, 1960) for non-delayed MDPs, that immediately confirms the exponential complexity of mA-PI.

Proposition (Lower Bound for Howard's PI)**.** *The number of iterations required for Howard's PI to converge in standard MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is $\Omega(|\mathcal{S}|)$.*

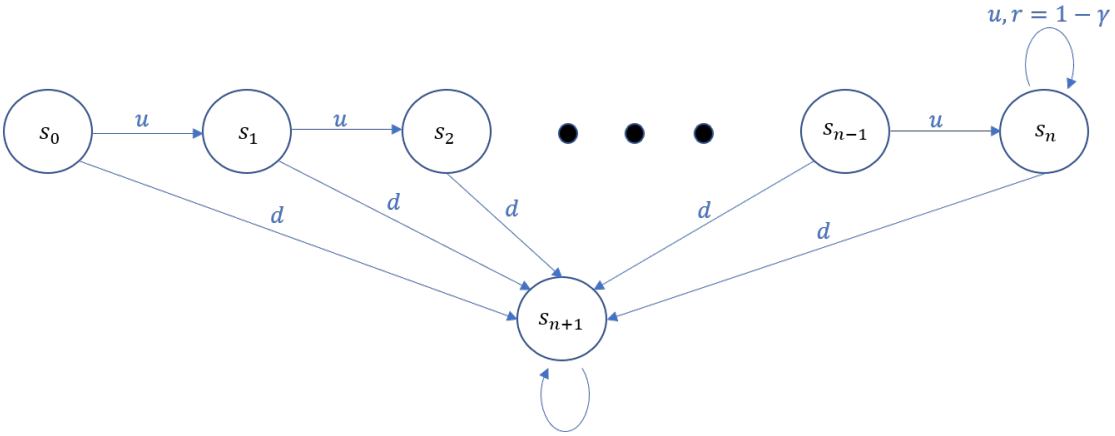

Figure 11: MDP example: the transitions are deterministic and the rewards are 0 everywhere except for $r(s_n, u) = 1 - \gamma$.

Proof. To prove the lower bound, we construct an example infinite-horizon MDP in which Howard's PI updates exactly one state at each iteration, and the number of updates is $|\mathcal{S}| - 1$.

The example MDP is given in Fig. 11. It contains a row of $n+1$ states $(s_0, s_1, \ldots, s_n)$, and a single absorbing state s_{n+1}. The transitions are deterministic. From each state except for s_{n+1} there are two actions, u and d, which respectively lead to the next state in the sequence or to s_{n+1}. The last state in the row, s_n, leads to itself or s_{n+1} by respectively taking actions u or d. Any action leads s_{n+1} to itself. The rewards are 0 everywhere except for $r(s_n, u) := 1 - \gamma$. We denote by (v_t, π_{t+1}) the value-policy pair at iteration t of Howard's PI. We shall now describe the convergence process to the optimal policy, which is obviously $\pi^*(s) = u \; \forall s \in \mathcal{S} \setminus \{s_{n+1}\}$.

Initialization: Set $\pi_0(s) = d \; \forall s \in \mathcal{S} \setminus \{s_{n+1}\}$.

Iteration 0: Clearly, $v_0 = \mathbf{0}$. Then, for all $s \in \mathcal{S} \setminus \{s_n, s_{n+1}\}$, $\pi_1(s) = \arg\max_a \{r(s,a) + \gamma v_0(s')\} = \arg\max_a \{0, 0\} = d$ [4]. Also, $\pi_1(s_n) = u$ since $1 - \gamma > 0$.

Iteration t $(t = 1, \ldots, n)$: We have

$$v_t(s_i) = 0 \;\; \text{for } i \in \{0, \ldots, n-t\}$$

and

$$v_t(s_i) = \gamma^{n-i} \frac{1-\gamma}{1-\gamma} = \gamma^{n-i} \;\; \text{for } i \in \{n-t+1, \ldots, n\}.$$

The policy output is thus

$$\pi_{t+1}(s_i) = d \;\; \text{for } i \in \{0, \ldots n-t-1\}$$

and

$$\pi_{t+1}(s_i) = u \;\; \text{for } i \in \{n-t, \ldots n\}.$$

To summarize, at each iteration a single state updates its action to the optimal one such that at iteration t, the policy stabilizes on $\pi(s) = u$ for all $s \in \{s_{n-t}, \ldots, s_n\}$. Therefore, the total number of iterations until convergence is $n + 1 = |\mathcal{S}| - 1$.

□

The exponential complexity of mA-PI follows, as stated in Thm. 4.1 that we recall below:

Proposition (Lower Bound for mA-PI)**.** *The number of iterations required for mA-PI to converge in m-EDMDP $\mathcal{M}_m$ is $\Omega(|\mathcal{X}_m|) = \Omega(|\mathcal{S}||\mathcal{A}|^m)$.*

[4] The policy improvement step needs to choose between two actions that both yield values 0. Without loss of generality, in such case, it simply chooses according to the lowest index, giving d here.

B.5 PROOF OF THEOREM 4.2

Theorem (mA-PI Convergence)**.** *The mA-PI algorithm as given in Alg.1 converges to the optimal value-policy pair $(\bar{v}^*, \bar{\pi}^*)$ in at most*

$$|\mathcal{S}||\mathcal{A}|^m(|\mathcal{A}|-1)\left\lceil \log\left(\frac{1}{\gamma}\right)^{-1}\log\left(\frac{1}{1-\gamma}\right)\right\rceil$$

iterations.

Proof. The proof proceeds in three steps which follow the same lines as in (Scherrer et al., 2016) except that here, we adapt that method to the augmented MDP $\mathcal{M}_m$ with its corresponding Bellman operators $\bar{T}^{\bar{\pi}}$ and $\bar{T}$ instead. For completeness, we recall these three steps whose proofs can be found in (Scherrer et al., 2016).

Given policy $\bar{\pi}_t$ output at iteration t, define the advantage of $\bar{\pi}'$ w.r.t. $\bar{\pi}$ as:

$$a_{\bar{\pi}}^{\bar{\pi}'} := \bar{T}^{\bar{\pi}'}v^{\bar{\pi}} - v^{\bar{\pi}}$$

and the maximal advantage w.r.t. $\bar{\pi}$ as

$$a_{\bar{\pi}} := \max_{\bar{\pi}'\in\bar{\Pi}_m} a_{\bar{\pi}}^{\bar{\pi}'} = \max_{\bar{\pi}'\in\bar{\Pi}_m} \bar{T}^{\bar{\pi}'}v^{\bar{\pi}} - v^{\bar{\pi}} = \bar{T}v^{\bar{\pi}} - v^{\bar{\pi}}.$$

Step 1 (Scherrer et al., 2016)[Lemma 10]. For all augmented policies $\bar{\pi}, \bar{\pi}' \in \bar{\Pi}_m$, $v^{\bar{\pi}'} - v^{\bar{\pi}} = (\bar{I} - \gamma F^{\bar{\pi}'})^{-1}a_{\bar{\pi}}^{\bar{\pi}'} = (\bar{I} - \gamma F^{\bar{\pi}})^{-1}(-a_{\bar{\pi}'}^{\bar{\pi}})$, with $\bar{I}$ being the identity matrix in $\mathbb{R}^{|\mathcal{X}_m|\times|\mathcal{A}|}$.

Step 2 (Scherrer et al., 2016)[Lemma 2]. Define as $\bar{v}^* = v^{\bar{\pi}^*}$ the optimal value function of the augmented MDP $\mathcal{M}_m$ as defined in Def. 4.1. Then, the sequence $(\|\bar{v}^* - v^{\bar{\pi}_t}\|_\infty)_{t\geqslant 0}$ built by the mA-PI algorithm as given in Alg.1 is a γ-contraction w.r.t. the max-norm.

Step 3 (Scherrer et al., 2016)[Section 7]. Let $x_0 \in \mathcal{X}_m$ be such that $-a_{\bar{\pi}^*}^{\bar{\pi}_0}(x_0) = \|a_{\bar{\pi}^*}^{\bar{\pi}_0}\|_\infty$. Then, for all $t \geqslant 0$ we have

$$-a_{\bar{\pi}^*}^{\bar{\pi}_t}(x_0) \leqslant \|a_{\bar{\pi}^*}^{\bar{\pi}_t}\|_\infty \leqslant \frac{\gamma^t}{1-\gamma}\|a_{\bar{\pi}^*}^{\bar{\pi}_0}\|_\infty = \frac{\gamma^t}{1-\gamma}(-a_{\bar{\pi}^*}^{\bar{\pi}_0}(x_0)).$$

From there it results that $\bar{\pi}_t(x_0)$ must be different from $\bar{\pi}_0(x_0)$ whenever $\frac{\gamma^t}{1-\gamma} < 1$, that is, for all iterations

$$t > \left\lceil \frac{\log(1/(1-\gamma))}{\log(1/\gamma)} \right\rceil =: t^*.$$

Therefore, one sub-optimal action is eliminated in favor of a better one within t^* iterations. There are at most $|\mathcal{X}_m|(|\mathcal{A}|-1)$ of them, which ends the proof. □

C EXECUTION-DELAY MDP: A NEW FORMULATION

Let μ be the initial state distribution. Then policy $\pi \in \Pi^{\text{HR}}$ induces a probability measure on $(\Omega, \mathcal{B}(\Omega))$ denoted by $\mathbb{P}_m^\pi$ and defined through the following:

$$\mathbb{P}_m^\pi(\tilde{s}_0 = s_0) = \mu(s_0); \tag{16}$$

$$\mathbb{P}_m^\pi(\tilde{a}_t = a|\tilde{h}_t = h_t) = \delta_{\bar{a}_t}(a), \qquad \forall t < m; \tag{17}$$

$$\mathbb{P}_m^\pi(\tilde{a}_t = a|\tilde{h}_{t-m} = h_{t-m}) = q_{d_{t-m}(h_{t-m})}(a), \qquad \forall t \geqslant m; \tag{18}$$

$$\mathbb{P}_m^\pi(\tilde{s}_{t+1} = s|\tilde{h}_t = (h_{t-1}, a_{t-1}, s_t), \tilde{a}_t = a_t) = P(s|s_t, a_t). \tag{19}$$

C.1 PROOF OF PROPOSITION 5.1

Proof. We first state the following, which holds by definition of conditional probability. For all measurable sets $A_1, \cdots, A_n \in \mathcal{B}(\Omega)$, we have

$$\mathbb{P}_m^\pi(\cap_{i=1}^n A_i) = \left(\prod_{i=1}^{n-1}\mathbb{P}_m^\pi(A_i|\cap_{j=i+1}^n A_j)\right)\mathbb{P}_m^\pi(A_n). \tag{20}$$

Applying (20) to $n = 2t + 1$ on the following events:

$$\begin{aligned} A_{2t+1} &:= \{\tilde{s}_0 = s_0\} \\ A_{2t} &:= \{\tilde{a}_0 = a_0\} \\ &\vdots \\ A_2 &:= \{\tilde{a}_{t-1} = a_{t-1}\} \\ A_1 &:= \{\tilde{s}_t = s_t\}, \end{aligned}$$

we obtain that

$$\begin{aligned} &\mathbb{P}^\pi_m(\tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t) \\ &= \mathbb{P}^\pi_m(\tilde{s}_0 = s_0) \prod_{i=0}^{t-1} \mathbb{P}^\pi_m(\tilde{a}_i = a_i | \tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{s}_i = s_i) \mathbb{P}^\pi_m(\tilde{s}_{i+1} = s_{i+1} | \tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_i = a_i) \\ &= \mathbb{P}^\pi_m(\tilde{s}_0 = s_0) \prod_{i=0}^{t-1} \mathbb{P}^\pi_m(\tilde{a}_i = a_i | \tilde{h}_i = h_i) \mathbb{P}^\pi_m(\tilde{s}_{i+1} = s_{i+1} | \tilde{h}_i = (h_{i-1}, a_{i-1}, s_i), \tilde{a}_i = a_i) \end{aligned}$$

If $t \leqslant m$, then $0 \leqslant i < m$ and by Eqs. (16), (17) and (19),

$$\mathbb{P}^\pi_m(\tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t) = \mu(s_0) \left(\prod_{i=0}^{t-1} \delta_{\bar{a}_i}(a_i) P(s_{i+1} | s_i, a_i) \right).$$

Otherwise, by Eq. (18),

$$\begin{aligned} &\mathbb{P}^\pi_m(\tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t) \\ &= \mathbb{P}^\pi_m(\tilde{s}_0 = s_0) \prod_{i=0}^{m-1} \mathbb{P}^\pi_m(\tilde{a}_i = a_i | \tilde{h}_i = h_i) \mathbb{P}^\pi_m(\tilde{s}_{i+1} = s_{i+1} | \tilde{h}_i = (h_{i-1}, a_{i-1}, s_i), \tilde{a}_i = a_i) \\ &\quad \prod_{k=m}^{t-1} \mathbb{P}^\pi_m(\tilde{a}_k = a_k | \tilde{h}_k = h_k) \mathbb{P}^\pi_m(\tilde{s}_{k+1} = s_{k+1} | \tilde{h}_k = (h_{k-1}, a_{k-1}, s_k), \tilde{a}_k = a_k) \\ &= \mu(s_0) \left(\prod_{i=0}^{m-1} \delta_{\bar{a}_i}(a_i) P(s_{i+1} | s_i, a_i) \right) \left(\prod_{k=m}^{t-1} q_{d_{k-m}(h_{k-m})}(a_k) P(s_{k+1} | s_k, a_k) \right), \end{aligned}$$

which concludes the proof. □

C.2 Remark regarding the Markov property

For $T > t \geqslant m$, the conditional probability can be evaluated through:

$$\begin{aligned} &\mathbb{P}^\pi_m(\tilde{a}_t = a_t, \tilde{s}_{t+1} = s_{t+1}, \cdots, \tilde{a}_{T-1} = a_{T-1}, \tilde{s}_T = s_T | \tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t) \\ &= \frac{\mathbb{P}^\pi_m(\tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{T-1} = a_{T-1}, \tilde{s}_T = s_T)}{\mathbb{P}^\pi_m(\tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t)} \\ &= q_{d_{t-m}(h_{t-m})}(a_t) P(s_{t+1} | s_t, a_t) \cdots q_{d_{T-m-1}(h_{T-m-1})}(a_{T-1}) P(s_T | s_{T-1}, a_{T-1}). \end{aligned}$$

For a stationary policy $\pi := (d, d, \cdots) \in \Pi^{\text{SR}}$, this simplifies to

$$\begin{aligned} &\mathbb{P}^\pi_m(\tilde{a}_t = a_t, \tilde{s}_{t+1} = s_{t+1}, \cdots, \tilde{a}_{T-1} = a_{T-1}, \tilde{s}_T = s_T | \tilde{s}_0 = s_0, \tilde{a}_0 = a_0, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t) \\ &= q_{d(s_{t-m})}(a_t) P(s_{t+1} | s_t, a_t) \cdots q_{d(s_{T-m-1})}(a_{T-1}) P(s_T | s_{T-1}, a_{T-1}). \end{aligned}$$

Observing that the resulting conditional probability is a function of past observations when $m > 0$, we conclude that even under a stationary policy, the induced stochastic process is not a Markov chain. This is different from the standard MDP setting in which any Markov policy induces a discrete time Markov chain (Puterman, 2014)[Sec. 2.1.6].

C.3 PROOF OF THEOREM 5.1

We first prove the following lemma, which will be used in the theorem's proof.

Lemma C.1. *For all $m > 0$, $t \geqslant 0$,*

$$\mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{a}_{t+1} = a', \tilde{s}_t = s, \tilde{a}_t = a) = \mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{s}_t = s, \tilde{a}_t = a) \tag{21}$$

Proof. First, note that for all delay value $m > 0$, $\tilde{a}_{t+1}$ only depends on the history up to $t - m + 1$, which is $h_{t-m+1} = (h_{t-m}, a_{t-m}, s_{t-m+1})$, as Eq. (18) suggests. Thus, since $t - m + 1 < t + 1$, we have that $\tilde{a}_{t+1}$ is independent of $\tilde{s}_{t+1}$. Using Bayes rule, it follows that

$$\begin{aligned}
&\mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{a}_{t+1} = a', \tilde{s}_t = s, \tilde{a}_t = a) \\
&= \frac{\mathbb{P}_m^\pi(\tilde{a}_{t+1} = a' | \tilde{s}_{t+1} = s', \tilde{s}_t = s, \tilde{a}_t = a) \mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{s}_t = s, \tilde{a}_t = a)}{\mathbb{P}_m^\pi(\tilde{a}_{t+1} = a' | \tilde{s}_t = s, \tilde{a}_t = a)} \\
&= \frac{\mathbb{P}_m^\pi(\tilde{a}_{t+1} = a' | \tilde{s}_t = s, \tilde{a}_t = a) \mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{s}_t = s, \tilde{a}_t = a)}{\mathbb{P}_m^\pi(\tilde{a}_{t+1} = a' | \tilde{s}_t = s, \tilde{a}_t = a)} \\
&= \mathbb{P}_m^\pi(\tilde{s}_{t+1} = s' | \tilde{s}_t = s, \tilde{a}_t = a).
\end{aligned}$$

□

Theorem. *Let $\pi := (d_0, d_1, \cdots) \in \Pi^{\text{HR}}$ be a history dependent policy. For all $s_0 \in \mathcal{S}$, there exists a Markov policy $\pi' := (d'_0, d'_1, \cdots) \in \Pi^{\text{MR}}$ that yields the same process distribution as π, i. e., for all $a \in \mathcal{A}, s' \in \mathcal{S}, t \geqslant m$,*

$$\mathbb{P}_m^{\pi'}(\tilde{s}_{t-m} = s', \tilde{a}_t = a | \tilde{s}_0 = s_0) = \mathbb{P}_m^\pi(\tilde{s}_{t-m} = s', \tilde{a}_t = a | \tilde{s}_0 = s_0). \tag{22}$$

Proof. When $m = 0$, the result holds true by standard RL theory (Puterman, 2014)[Thm 5.5.1]. Thus, assume that $m > 0$. Fix $s \in \mathcal{S}$. Let $\pi' := (d'_0, d'_1, \cdots)$ with $d'_0 : \{s\} \to \Delta_{\mathcal{A}}$ defined as

$$q_{d'_0(s)}(a) := \mathbb{P}_m^\pi(\tilde{a}_m = a | \tilde{s}_0 = s) \tag{23}$$

and for all $t > m$,

$$q_{d'_{t-m}(s')}(a) := \mathbb{P}_m^\pi(\tilde{a}_t = a | \tilde{s}_{t-m} = s', \tilde{s}_0 = s), \quad \forall s' \in \mathcal{S}, a \in \mathcal{A}\,. \tag{24}$$

For the policy π' defined as in Eqs. (23)-(24), we prove Eq. (22) by induction on $t \geqslant m$. By construction of π', the induction base is satisfied at $t = m$. By construction of π' again, for all $t > m$ we have

$$\begin{aligned}
\mathbb{P}_m^{\pi'}(\tilde{a}_t = a | \tilde{s}_{t-m} = s', \tilde{s}_0 = s) &= \mathbb{P}_m^{\pi'}(\tilde{a}_t = a | \tilde{s}_{t-m} = s') \\
&= q_{d'_{t-m}(s')}(a) \\
&= \mathbb{P}_m^\pi(\tilde{a}_t = a | \tilde{s}_{t-m} = s', \tilde{s}_0 = s).
\end{aligned} \tag{25}$$

Assume that Eq. (22) holds up until $t = n - 1$. Further let the Euclidean division $n - 1 = km + r$ of $n - 1$ by m, so that $k, r \in \mathbb{N}$ with $0 \leqslant r < m$. Then, we can write

$$\begin{aligned}
&\mathbb{P}_m^\pi(\tilde{s}_n = s' | \tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r} \in \mathcal{S}, \\ a_{km+r} \in \mathcal{A}}} \mathbb{P}_m^\pi(\tilde{s}_n = s', \tilde{s}_{km+r} = s_{km+r}, \tilde{a}_{km+r} = a_{km+r} | \tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r} \in \mathcal{S}, \\ a_{km+r} \in \mathcal{A}}} \mathbb{P}_m^\pi(\tilde{s}_n = s' | \tilde{a}_{km+r} = a_{km+r}, \tilde{s}_{km+r} = s_{km+r}, \tilde{s}_0 = s) \\
&\qquad \mathbb{P}_m^\pi(\tilde{a}_{km+r} = a_{km+r} | \tilde{s}_{km+r} = s_{km+r}, \tilde{s}_0 = s) \mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r} | \tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r} \in \mathcal{S}, \\ a_{km+r} \in \mathcal{A}}} P(s' | s_{km+r}, a_{km+r}) \mathbb{P}_m^\pi(\tilde{a}_{km+r} = a_{km+r} | \tilde{s}_{km+r} = s_{km+r}, \tilde{s}_0 = s) \mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r} | \tilde{s}_0 = s).
\end{aligned}$$

By Eq. (18), $\tilde{a}_{km+r}$ only depends on history up to $(k-1)m + r$. Thus, $\mathbb{P}_m^\pi(\tilde{a}_{km+r} = a_{km+r} | \tilde{s}_{km+r} = s_{km+r}, \tilde{s}_0 = s) = \mathbb{P}_m^\pi(\tilde{a}_{km+r} = a_{km+r} | \tilde{s}_0 = s)$ and

$$\begin{aligned}
&\mathbb{P}_m^\pi(\tilde{s}_n = s' | \tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r} \in \mathcal{S}, \\ a_{km+r} \in \mathcal{A}}} P(s' | s_{km+r}, a_{km+r}) \mathbb{P}_m^\pi(\tilde{a}_{km+r} = a_{km+r} | \tilde{s}_0 = s) \mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r} | \tilde{s}_0 = s).
\end{aligned}$$

Since $km + r = n - 1$, by the induction hypothesis we can rewrite

$$\begin{aligned}\mathbb{P}_m^{\pi}(\tilde{a}_{km+r} = a_{km+r}|\tilde{s}_0 = s) &= \sum_{s_{(k-1)m+r} \in \mathcal{S}} \mathbb{P}_m^{\pi}(\tilde{a}_{km+r} = a_{km+r}, \tilde{s}_{(k-1)m+r} = s_{(k-1)m+r}|\tilde{s}_0 = s) \\ &= \sum_{s_{(k-1)m+r} \in \mathcal{S}} \mathbb{P}_m^{\pi'}(\tilde{a}_{km+r} = a_{km+r}, \tilde{s}_{(k-1)m+r} = s_{(k-1)m+r}|\tilde{s}_0 = s) \\ &= \mathbb{P}_m^{\pi'}(\tilde{a}_{km+r} = a_{km+r}|\tilde{s}_0 = s),\end{aligned}$$

so that

$$\begin{aligned}&\mathbb{P}_m^{\pi}(\tilde{s}_n = s'|\tilde{s}_0 = s) \\ &= \sum_{\substack{s_{km+r} \in \mathcal{S}, \\ a_{km+r} \in \mathcal{A}}} P(s'|s_{km+r}, a_{km+r}) \mathbb{P}_m^{\pi'}(\tilde{a}_{km+r} = a_{km+r}|\tilde{s}_0 = s) \mathbb{P}_m^{\pi}(\tilde{s}_{km+r} = s_{km+r}|\tilde{s}_0 = s).\end{aligned}$$

We now study the last term in the above equation, $\mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r}|\tilde{s}_0)$. We have

$$
\begin{aligned}
&\mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r}, \tilde{s}_{km+r-1} = s_{km+r-1}, \tilde{a}_{km+r-1} = a_{km+r-1}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \mathbb{P}_m^\pi(\tilde{s}_{km+r} = s_{km+r}|\tilde{s}_{km+r-1} = s_{km+r-1}, \tilde{a}_{km+r-1} = a_{km+r-1}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}, \tilde{s}_0 = s) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km+r-1} = s_{km+r-1}, \tilde{a}_{km+r-1} = a_{km+r-1}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} P(s_{km+r}|s_{km+r-1}, a_{km+r-1}) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km+r-1} = s_{km+r-1}, \tilde{a}_{km+r-1} = a_{km+r-1}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} P(s_{km+r}|s_{km+r-1}, a_{km+r-1}) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km+r-1} = s_{km+r-1}|\tilde{a}_{km+r-1} = a_{km+r-1}, \tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}, \tilde{s}_0 = s) \\
&\quad \mathbb{P}_m^\pi(\tilde{a}_{km+r-1} = a_{km+r-1}, \tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&\overset{\text{Lemma C.1}}{=} \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} P(s_{km+r}|s_{km+r-1}, a_{km+r-1}) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km+r-1} = s_{km+r-1}|\tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}, \tilde{s}_0 = s) \\
&\quad \mathbb{P}_m^\pi(\tilde{a}_{km+r-1} = a_{km+r-1}, \tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} P(s_{km+r}|s_{km+r-1}, a_{km+r-1}) P(s_{km+r-1}|s_{km+r-2}, a_{km+r-2}) \\
&\quad \mathbb{P}_m^\pi(\tilde{a}_{km+r-1} = a_{km+r-1}, \tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} P(s_{km+r}|s_{km+r-1}, a_{km+r-1}) P(s_{km+r-1}|s_{km+r-2}, a_{km+r-2}) \\
&\quad \mathbb{P}_m^\pi(\tilde{a}_{km+r-1} = a_{km+r-1}|\tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2} \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km+r-2} = s_{km+r-2}, \tilde{a}_{km+r-2} = a_{km+r-2} \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \\
&\vdots \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Big(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1}, a_{km+i-1})\Big) \\
&\quad \Big(\prod_{j=1}^{r-1} \mathbb{P}_m^\pi(\tilde{a}_{km+j} = a_{km+j}|\tilde{s}_{km+j-1} = s_{km+j-1}, \tilde{a}_{km+j-1} = a_{km+j-1}, \cdots, \tilde{s}_{km} = s_{km}, \tilde{a}_{km} = a_{km}, \tilde{s}_0 = s)\Big) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km} = s_{km}|\tilde{a}_{km} = a_{km}, \tilde{s}_0 = s)\mathbb{P}_m^\pi(\tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&\overset{(1)}{=} \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Big(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1}, a_{km+i-1})\Big)\Big(\prod_{j=1}^{r-1} \mathbb{P}_m^\pi(\tilde{a}_{km+j} = a_{km+j}|\tilde{s}_0 = s)\Big) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km} = s_{km}|\tilde{a}_{km} = a_{km}, \tilde{s}_0 = s)\mathbb{P}_m^\pi(\tilde{a}_{km} = a_{km}|\tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Big(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1}, a_{km+i-1})\mathbb{P}_m^\pi(\tilde{a}_{km+i-1} = a_{km+i-1}|\tilde{s}_0 = s)\Big) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km} = s_{km}|\tilde{a}_{km} = a_{km}, \tilde{s}_0 = s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S} \\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Big(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1}, a_{km+i-1}) \\
&\quad \Big(\sum_{s'_{(k-1)m+i-1}\in\mathcal{S}} \mathbb{P}_m^\pi(\tilde{a}_{km+i-1} = a_{km+i-1}, \tilde{s}_{(k-1)m+i-1} = s'_{(k-1)m+i-1}|\tilde{s}_0 = s)\Big)\Big) \\
&\quad \mathbb{P}_m^\pi(\tilde{s}_{km} = s_{km}|\tilde{a}_{km} = a_{km}, \tilde{s}_0 = s)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(2)}{=} \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S}\\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Bigg(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1},a_{km+i-1}) \\
&\Bigg(\sum_{s'_{(k-1)m+i-1}\in\mathcal{S}} \mathbb{P}_m^{\pi'}(\tilde{a}_{km+i-1}=a_{km+i-1},\tilde{s}_{(k-1)m+i-1}=s'_{(k-1)m+i-1}|\tilde{s}_0=s)\Bigg)\Bigg) \\
&\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{a}_{km}=a_{km},\tilde{s}_0=s) \\
&= \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S}\\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Bigg(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1},a_{km+i-1})\mathbb{P}_m^{\pi'}(\tilde{a}_{km+i-1}=a_{km+i-1}|\tilde{s}_0=s)\Bigg) \\
&\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{a}_{km}=a_{km},\tilde{s}_0=s) \\
&\overset{(3)}{=} \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S}\\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Bigg(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1},a_{km+i-1})\mathbb{P}_m^{\pi'}(\tilde{a}_{km+i-1}=a_{km+i-1}|\tilde{s}_0=s)\Bigg) \\
&\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s).
\end{aligned}
$$

In (1), we use Eq. (18) to establish that $\tilde{a}_{km+j}$ only depends on history up to $(k-1)m+j$. Since $m-1>r-1\geqslant j\geqslant 1$, we have $km>(k-1)m+j$, and

$$
\begin{aligned}
&\mathbb{P}_m^{\pi}(\tilde{a}_{km+j}=a_{km+j}|\tilde{s}_{km+j-1}=s_{km+j-1},\tilde{a}_{km+j-1}=a_{km+j-1},\cdots,\tilde{s}_{km}=s_{km},\tilde{a}_{km}=a_{km},\tilde{s}_0=s) \\
&=\mathbb{P}_m^{\pi}(\tilde{a}_{km+j}=a_{km+j}|\tilde{s}_0=s).
\end{aligned}
$$

In (2), we use the induction hypothesis. In (3) we use Bayes rule and Eq. (18) again to obtain:

$$
\begin{aligned}
\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{a}_{km}=a_{km},\tilde{s}_0=s) &= \frac{\mathbb{P}_m^{\pi}(\tilde{a}_{km}=a_{km}|\tilde{s}_{km}=s_{km},\tilde{s}_0=s)\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s)}{\mathbb{P}_m^{\pi}(\tilde{a}_{km}=a_{km}|\tilde{s}_0=s)} \\
&= \frac{\mathbb{P}_m^{\pi}(\tilde{a}_{km}=a_{km}|\tilde{s}_0=s)\mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s)}{\mathbb{P}_m^{\pi}(\tilde{a}_{km}=a_{km}|\tilde{s}_0=s)} \\
&= \mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s).
\end{aligned}
$$

Thus, it results that

$$
\begin{aligned}
&\mathbb{P}_m^{\pi}(\tilde{s}_n=s'|\tilde{s}_0=s) \\
&= \sum_{\substack{s_{km+r}\in\mathcal{S},\\ a_{km+r}\in\mathcal{A}}} P(s'|s_{km+r},a_{km+r})\mathbb{P}_m^{\pi'}(\tilde{a}_{km+r}=a_{km+r}|\tilde{s}_0=s) \\
&\quad \sum_{\substack{s_{km+r-1},\cdots,s_{km}\in\mathcal{S}\\ a_{km+r-1},\cdots,a_{km}\in\mathcal{A}}} \Bigg(\prod_{i=1}^{r} P(s_{km+i}|s_{km+i-1},a_{km+i-1})\mathbb{P}_m^{\pi'}(\tilde{a}_{km+i-1}=a_{km+i-1}|\tilde{s}_0=s)\Bigg) \\
&\quad \mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s) \\
&= \sum_{\substack{s_{km+r},\cdots,s_{km}\in\mathcal{S}\\ a_{km+r},\cdots,a_{km}\in\mathcal{A}}} \Bigg(\prod_{i=1}^{r+1} P(s_{km+i}|s_{km+i-1},a_{km+i-1})\mathbb{P}_m^{\pi'}(\tilde{a}_{km+i-1}=a_{km+i-1}|\tilde{s}_0=s)\Bigg) \\
&\quad \mathbb{P}_m^{\pi}(\tilde{s}_{km}=s_{km}|\tilde{s}_0=s),
\end{aligned}
$$

where we used the convention $s_{km+r+1}=s_n=s'$. We similarly use backward induction until the remaining term that depends on π becomes

$$
\begin{aligned}
&\mathbb{P}_m^{\pi}(\tilde{s}_m=s_m|\tilde{s}_0=s) \\
&= \sum_{\substack{s_{m-1},\cdots,s_1\in\mathcal{S}\\ a_{m-1},\cdots,a_0\in\mathcal{A}}} \mathbb{P}_m^{\pi}(\tilde{s}_m=s_m,\tilde{s}_{m-1}=s_{m-1},\tilde{a}_{m-1}=a_{m-1},\cdots,\tilde{s}_1=s_1,\tilde{a}_1=a_1,\tilde{a}_0=a_0|\tilde{s}_0=s) \\
&= \sum_{\substack{s_{m-1},\cdots,s_1\in\mathcal{S}\\ a_{m-1},\cdots,a_0\in\mathcal{A}}} \frac{1}{\mathbb{P}_m^{\pi}(\tilde{s}_0=s)}\mathbb{P}_m^{\pi}(\tilde{s}_m=s_m,\tilde{s}_{m-1}=s_{m-1},\tilde{a}_{m-1}=a_{m-1},\cdots,\tilde{s}_1=s_1,\tilde{a}_1=a_1,\tilde{a}_0=a_0,\tilde{s}_0=s) \\
&\overset{(4)}{=} \sum_{\substack{s_{m-1},\cdots,s_1\in\mathcal{S}\\ a_{m-1},\cdots,a_0\in\mathcal{A}}} \frac{1}{\mu(s)}\mu(s)\left(\prod_{i=0}^{m-1} P(s_{i+1}|s_i,a_i)\delta_{\bar{a}_i}(a_i)\right) = \sum_{\substack{s_{m-1},\cdots,s_1\in\mathcal{S}\\ a_{m-1},\cdots,a_0\in\mathcal{A}}} \left(\prod_{i=0}^{m-1} P(s_{i+1}|s_i,a_i)\delta_{\bar{a}_i}(a_i)\right),
\end{aligned}
$$

where (4) results from Prop. 5.1. Since the obtained quantity is independent of π, we have

$$\mathbb{P}_m^{\pi}(\tilde{s}_m = s_m | \tilde{s}_0 = s) = \mathbb{P}_m^{\pi'}(\tilde{s}_m = s_m | \tilde{s}_0 = s).$$

Thus, if we decompose $\mathbb{P}_m^{\pi'}(\tilde{s}_n = s' | \tilde{s}_0 = s)$ according to the exact same derivation as we did for $\mathbb{P}_m^{\pi}(\tilde{s}_n = s' | \tilde{s}_0 = s)$, we obtain that at $t = n$,

$$\mathbb{P}_m^{\pi}(\tilde{s}_n = s' | \tilde{s}_0 = s) = \mathbb{P}_m^{\pi'}(\tilde{s}_n = s' | \tilde{s}_0 = s). \tag{26}$$

As a result, at $t = n$ we have

$$\begin{aligned}
\mathbb{P}_m^{\pi'}(\tilde{s}_{n-m} = s', \tilde{a}_n = a | \tilde{s}_0 = s) &= \mathbb{P}_m^{\pi'}(\tilde{a}_n = a | \tilde{s}_{n-m} = s', \tilde{s}_0 = s) \mathbb{P}_m^{\pi'}(\tilde{s}_{n-m} = s' | \tilde{s}_0 = s) \\
&\overset{(a)}{=} \mathbb{P}_m^{\pi'}(\tilde{a}_n = a | \tilde{s}_{n-m} = s', \tilde{s}_0 = s) \mathbb{P}_m^{\pi}(\tilde{s}_{n-m} = s' | \tilde{s}_0 = s) \\
&\overset{(b)}{=} \mathbb{P}_m^{\pi}(\tilde{a}_n = a | \tilde{s}_{n-m} = s', \tilde{s}_0 = s) \mathbb{P}_m^{\pi}(\tilde{s}_{n-m} = s' | \tilde{s}_0 = s) \\
&\overset{(c)}{=} \mathbb{P}_m^{\pi}(\tilde{s}_{n-m} = s', \tilde{a}_n = a | \tilde{s}_0 = s),
\end{aligned}$$

where (b) follows from Eq. (26); (c) from Eq. (25). Finally, assuming it is satisfied at $t = n - 1$, the induction step is proved for $t = n$, which ends the proof. □

C.4 Degradation due to Stationarity

Prop. 5.2 follows from computing the optimal return on an execution-delay MDP (EDMDP) using simulation. Specifically, we use Example 3.1 which we analytically studied in Sec. 3. We exhaustively search over the deterministic policy spaces Π^{SD} and Π^{MD} to find the optimum. We stress that limiting our search to deterministic policies is sufficient for this MDP. Indeed, as shown in Appx. A.1, optimal return is attained for a deterministic policy when maximizing over Π^{SR}. Regarding the non-stationary Markov policy space Π^{MR}, as we have proved in Thm. 5.2, there exists an optimal deterministic policy in Π^{MD}. We show with this experiment that this optimal Markov deterministic policy attains better return than any stationary policy. We set $p = 0.8$. Since the search-space of non-stationary policies is exponential in the simulation horizon ($T = 10$ here), we choose $\gamma = 0.5$ to have low approximation error.

Policy-Type	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
Stationary (theoretical)	2	1.6	1.36	1.216	1.129	1.077
Stationary	1.99 ± 0.01	1.59 ± 0.03	1.32 ± 0.05	1.22 ± 0.08	1.11 ± 0.09	1.02 ± 0.13
Non-stationary Markov	1.99 ± 0.01	1.82 ± 0.05	1.67 ± 0.08	1.59 ± 0.12	1.46 ± 0.15	1.38 ± 0.2

Table 1: Optimal return for different delay values and policy types.

The results are summarized in Table 1. Apart from demonstrating sub-optimality of the stationary policy, they also confirm that our theoretical return for the stationary policy $\frac{1+(2p-1)^m}{2(1-\gamma)}$ from Prop. 3.1 matches closely with simulation.

C.5 The Delayed Value Function

Given a random variable W over $(\Omega, \mathcal{B}(\Omega), \mathbb{P}_m^{\pi})$, its expectation is $\mathbb{E}_m^{\pi}[W] = \sum_{\omega \in \Omega} W(\omega) \mathbb{P}_m^{\pi}(\omega)$, where $\omega = (s_0, a_0, s_1, \cdots)$ is a sample path. A typical W to consider is the discounted sum of rewards $W(s_0, a_0, s_1, \cdots) := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$. Thus, the expectation conditioned on initial state s_0 is given by $\mathbb{E}_m^{\pi}[W | s_0] = \sum_{\omega \in \Omega} W(s_0, a_0, \cdots) \mathbb{P}_m^{\pi}(s_0, a_0, s_1, \cdots | s_0)$. Let the *delayed value function*

$$v_m^{\mu_0 : \mu_{m-1}, \pi}(s_0) := \mathbb{E}_m^{\pi} \left[\sum_{t=m}^{\infty} \gamma^{t-m} r(\tilde{s}_t, \tilde{a}_t) \middle| \tilde{s}_0 = s_0 \right], \tag{27}$$

where $\mu_0 : \mu_{m-1} := (\mu_0, \cdots, \mu_{m-1})$ denotes some fixed queue of action distributions according to which the initial m actions should be executed. Note that the definition of W does not change w.r.t. the delay value m: it always denotes the discounted sum of rewards. However, its distribution

does depend on the delay value m through the process distribution $\mathbb{P}_m^\pi$ and, as a result, so does its expectation $\mathbb{E}_m^\pi[W|s_0]$.

Consider a Markov policy $\pi := (d_k)_{k\geqslant 0} \in \Pi^{\text{MR}}$. For all $s, s' \in \mathcal{S}, k \in \mathbb{N}$ and $u \in \Delta_{\mathcal{A}}$, let $P_u(s,s') := \sum_{a\in\mathcal{A}} u(a)P(s'|s,a)$ and $R_{d_k}(s',s) := \sum_{a\in\mathcal{A}} q_{d_k(s)}(a)r(s',a)$. We then have the following result.

Theorem C.1. *For a Markov policy $\pi \in \Pi^{\text{MR}}$ given by $\pi := (d_0, d_1, \cdots)$, the delayed value function satisfies the following relation:*

$$v_m^{\mu_0:\mu_{m-1},\pi}(s_0) = \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R_{d_0}\right)(s_0,s_0) + \gamma \sum_{s_1,\cdots,s_m\in\mathcal{S}} \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k,s_{k+1})\right) v_m^{d_0(s_0):d_{m-1}(s_{m-1}),\pi_{m:}}(s_m),$$

where $\pi_{m:} := (d_m, d_{m+1}, \cdots)$ denotes the policy π starting from its $m+1$-th decision rule.

In addition, this relation becomes a recursion when the policy is m-periodic. Its proof is omitted since the result immediately follows from Thm. C.1.

Corollary C.1. *For an m-periodic Markov policy $\pi \in \Pi^{\text{MR}}$ given by $\pi := (d_0, \cdots, d_{m-1}, d_0, \cdots, d_{m-1}, \cdots)$, the delayed value function satisfies the following recursion:*

$$v_m^{\mu_0:\mu_{m-1},\pi}(s_0) = \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R_{d_0}\right)(s_0,s_0) + \gamma \sum_{s_1,\cdots,s_m\in\mathcal{S}} \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k,s_{k+1})\right) v_m^{d_0(s_0):d_{m-1}(s_{m-1}),\pi}(s_m).$$

Proof of Theorem C.1. By definition of the delayed value function we have:

$$\begin{aligned}
v_m^{\mu_0:\mu_{m-1},\pi}(s_0) &= \mathbb{E}_m^\pi\left[\sum_{t=m}^{\infty}\gamma^{t-m}r(\tilde{s}_t,\tilde{a}_t)\middle|\tilde{s}_0 = s_0\right] \\
&\overset{(1)}{=} \sum_{t=m}^{\infty}\gamma^{t-m}\mathbb{E}_m^\pi\left[r(\tilde{s}_t,\tilde{a}_t)\middle|\tilde{s}_0 = s_0\right] \\
&\overset{(2)}{=} \sum_{t=m}^{\infty}\gamma^{t-m}\sum_{\substack{s_1,\cdots,s_t\in\mathcal{S}\\ a_0,\cdots,a_t\in\mathcal{A}}} r(s_t,a_t)\mathbb{P}_m^\pi(\tilde{a}_0 = a_0, \tilde{s}_1 = s_1, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t, \tilde{a}_t = a_t|\tilde{s}_0 = s_0),
\end{aligned}$$

where (1) results from the dominated convergence theorem and (2) by the definition of expectation. Using Prop. 5.1 and the fact that π is a Markov policy, we can write the probability of a sample path conditioned on the initial state as:

$$\begin{aligned}
&\mathbb{P}_m^\pi(\tilde{a}_0 = a_0, \tilde{s}_1 = s_1, \cdots, \tilde{a}_{t-1} = a_{t-1}, \tilde{s}_t = s_t, \tilde{a}_t = a_t|\tilde{s}_0 = s) \\
&= \left(\prod_{k=0}^{m-1}\mu_k(a_k)P(s_{k+1}|s_k,a_k)\right)\left(\prod_{k=m}^{t-1}q_{d_{k-m}(s_{k-m})}(a_k)P(s_{k+1}|s_k,a_k)\right)q_{d_{t-m}(s_{t-m})}(a_t),
\end{aligned}$$

so that:

$$\begin{aligned}
v_m^{\mu_0:\mu_{m-1},\pi}(s_0) = \sum_{t=m}^{\infty}\gamma^{t-m}\sum_{\substack{s_1,\cdots,s_t\in\mathcal{S}\\ a_0,\cdots,a_t\in\mathcal{A}}} r(s_t,a_t)&\left(\prod_{k=0}^{m-1}\mu_k(a_k)P(s_{k+1}|s_k,a_k)\right) \\
&\cdot\left(\prod_{k=m}^{t-1}q_{d_{k-m}(s_{k-m})}(a_k)P(s_{k+1}|s_k,a_k)\right)q_{d_{t-m}(s_{t-m})}(a_t).
\end{aligned}$$

Then, we can rewrite the delayed value function as:

$$
\begin{aligned}
&v_m^{\mu_0:\mu_{m-1},\pi}(s_0) \\
&= \sum_{t=m}^{\infty} \gamma^{t-m} \sum_{\substack{s_1,\cdots,s_t\in\mathcal{S} \\ a_0,\cdots,a_{t-1}\in\mathcal{A}}} R_{d_{t-m}}(s_t, s_{t-m}) \cdot \left(\prod_{k=0}^{m-1} \mu_k(a_k) P_{a_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-1} q_{d_{k-m}(s_{k-m})}(a_k) P_{a_k}(s_k, s_{k+1})\right). \\
&= \sum_{t=m}^{\infty} \gamma^{t-m} \sum_{\substack{s_1,\cdots,s_t\in\mathcal{S} \\ a_m,\cdots,a_{t-2}\in\mathcal{A}}} R_{d_{t-m}}(s_t, s_{t-m}) \cdot \left(\prod_{k=0}^{m-1} \mu_k(a_k) P_{a_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-2} q_{d_{k-m}(s_{k-m})}(a_k) P_{a_k}(s_k, s_{k+1})\right) \cdot \\
&\quad \sum_{a_{t-1}\in\mathcal{A}} q_{d_{t-1-m}(s_{t-1-m})}(a_{t-1}) P_{a_{t-1}}(s_{t-1}, s_t) \\
&= \sum_{t=m}^{\infty} \gamma^{t-m} \sum_{\substack{s_1,\cdots,s_t\in\mathcal{S} \\ a_m,\cdots,a_{t-2}\in\mathcal{A}}} R_{d_{t-m}}(s_t, s_{t-m}) \cdot \left(\prod_{k=0}^{m-1} \mu_k(a_k) P_{a_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-2} q_{d_{k-m}(s_{k-m})}(a_k) P_{a_k}(s_k, s_{k+1})\right) \cdot \\
&\quad P_{d_{t-1-m}(s_{t-1-m})}(s_{t-1}, s_t) \\
&\vdots \\
&= \sum_{t=m}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \cdot \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right),
\end{aligned}
$$

and the following can be derived:

$$
\begin{aligned}
v_m^{\mu_0:\mu_{m-1},\pi}(s_0) &= \sum_{t=m}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{s_1,\cdots,s_m} R_{d_0}(s_m, s_0) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) + f(s_0) \\
&= \sum_{s_m} \left(P_{\mu_0} \cdots P_{\mu_{m-1}}\right)(s_0, s_m) R_{d_0}(s_m, s_0) + f(s_0) \\
&= \left(P_{\mu_0} \cdots P_{\mu_{m-1}} R_{d_0}\right)(s_0, s_0) + f(s_0) \\
&= \left(P_{\mu_0} \cdots P_{\mu_{m-1}} R_{d_0}\right)(s_0, s_0) + f(s_0), \qquad (28)
\end{aligned}
$$

where

$$
\begin{aligned}
f(s_0) :&= \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_m} \sum_{s_{m+1},\cdots,s_t} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \\
&\quad \cdot \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_m} \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \sum_{s_{m+1},\cdots,s_t} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right). \qquad (29)
\end{aligned}
$$

In fact, the last part of the sum corresponds to the following expectation:

$$
\sum_{s_{m+1},\cdots,s_t} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) = \mathbb{E}_m^{\pi}\left[r(\tilde{s}_t, \tilde{a}_t) | \tilde{s}_0 = s_0, \tilde{s}_1 = s_1, \cdots, \tilde{s}_m = s_m\right],
$$

so

$$
f(s_0) = \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_m} \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \mathbb{E}_m^{\pi}\left[r(\tilde{s}_t, \tilde{a}_t) | \tilde{s}_0 = s_0, \tilde{s}_1 = s_1, \cdots, \tilde{s}_m = s_m\right]
$$

and

$$
\begin{aligned}
v_m^{\mu_0:\mu_{m-1},\pi}(s_0) &= \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R_{d_0}\right)(s_0,s_0) \\
&\quad + \sum_{t=m+1}^{\infty}\gamma^{t-m}\sum_{s_1,\cdots,s_m}\left(\prod_{k=0}^{m-1}P_{\mu_k}(s_k,s_{k+1})\right)\mathbb{E}_m^{\pi}\left[r(\tilde{s}_t,\tilde{a}_t)|\tilde{s}_0=s_0,\tilde{s}_1=s_1,\cdots,\tilde{s}_m=s_m\right] \\
&= \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R^{\pi_0}\right)(s_0,s_0) \\
&\quad + \gamma\sum_{s_1,\cdots,s_m}\left(\prod_{k=0}^{m-1}P_{\mu_k}(s_k,s_{k+1})\right)\mathbb{E}_m^{\pi}\left[\sum_{t=m+1}^{\infty}\gamma^{t-m-1}r(\tilde{s}_t,\tilde{a}_t)|\tilde{s}_0=s_0,\tilde{s}_1=s_1,\cdots,\tilde{s}_m=s_m\right]
\end{aligned}
$$

Finally, fixing initial states $s_0,\cdots,s_{m-1}$ implies fixing a queue of m action distributions $d_0(s),\cdots,d_{m-1}(s_{m-1})$. Therefore, denoting by $\pi_{m:}:=(d_m,d_{m+1},\cdots)$ the original policy π starting from its $m+1$-th decision rule, we have

$$
\mathbb{E}_m^{\pi}\left[\sum_{t=m+1}^{\infty}\gamma^{t-m-1}r(\tilde{s}_t,\tilde{a}_t)|\tilde{s}_0=s_0,\tilde{s}_1=s_1,\cdots,\tilde{s}_m=s_m\right] = v_m^{d_0(s_0):d_{m-1}(s_{m-1}),\pi_{m:}}(s_m),
$$

and

$$
\begin{aligned}
&v_m^{\mu_0:\mu_{m-1},\pi}(s_0) \\
&= \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R^{\pi_0}\right)(s_0,s_0) + \gamma\sum_{s_1,\cdots,s_m}\left(\prod_{k=0}^{m-1}P_{\mu_k}(s_k,s_{k+1})\right)v_m^{d_0(s_0):d_{m-1}(s_{m-1}),\pi_{m:}}(s_m),
\end{aligned}
$$

which concludes the proof. □

C.6 Proof of Theorem 5.2

Theorem. *For any action distribution queue $\mu_0:\mu_{m-1}:=(\mu_0,\ldots,\mu_{m-1})$ and $s_0\in\mathcal{S}$,*

$$
\max_{\pi\in\Pi^{\mathrm{MD}}} v_m^{\mu_0:\mu_{m-1},\pi} = \max_{\pi\in\Pi^{\mathrm{MR}}} v_m^{\mu_0:\mu_{m-1},\pi}. \tag{30}
$$

Proof. First, since $\Pi^{\mathrm{MD}}\subset\Pi^{\mathrm{MR}}$, the RHS of (30) must be at least as great as the LHS. We now establish the reverse inequality. Recall from Eqs.(28)-(29)[Proof of Thm. C.1] that

$$
\begin{aligned}
&v_m^{\mu_0:\mu_{m-1},\pi}(s_0) = \left(P_{\mu_0}\cdots P_{\mu_{m-1}}R_{d_0}\right)(s_0,s_0) \\
&+\sum_{t=m+1}^{\infty}\gamma^{t-m}\sum_{s_1,\cdots,s_t\in\mathcal{S}}R_{d_{t-m}}(s_t,s_{t-m})\cdot\left(\prod_{k=0}^{m-1}P_{\mu_k}(s_k,s_{k+1})\right)\left(\prod_{k=m}^{t-1}P_{d_{k-m}(s_{k-m})}(s_k,s_{k+1})\right).
\end{aligned} \tag{31}
$$

We shall now rewrite the value in two forms, to respectively show its dependence on $d_0(s_0)$, and $d_i(s_i)$ for $i\geqslant 1$. For each of these two forms, we will prove a deterministic decision is at least as good as a random one in terms of value.

We begin with rewriting the first term in (31) as

$$
\begin{aligned}
\left(P_{\mu_0}\cdots P_{\mu_{m-1}}R_{d_0}\right)(s_0,s_0) &= \left(\left(P_{\mu_0}\cdots P_{\mu_{m-1}}\right)R_{d_0}\right)(s_0,s_0) \\
&= \sum_{s_m\in\mathcal{S}}(P_{\mu_0}\cdots P_{\mu_{m-1}})(s_0,s_m)R_{d_0}(s_m,s_0) \\
&= \sum_{s_m\in\mathcal{S}}(P_{\mu_0}\cdots P_{\mu_{m-1}})(s_0,s_m)\left(\sum_{a_0\in\mathcal{A}}q_{d_0(s_0)}(a_0)r(s_m,a_0)\right) \\
&= \sum_{a_0\in\mathcal{A}}q_{d_0(s_0)}(a_0)\sum_{s_m\in\mathcal{S}}(P_{\mu_0}\cdots P_{\mu_{m-1}})(s_0,s_m)r(s_m,a_0) \\
&= \sum_{a_0\in\mathcal{A}}q_{d_0(s_0)}(a_0)(P_{\mu_0}\cdots P_{\mu_{m-1}}r_{a_0})(s_0),
\end{aligned}
$$

where for all $a \in \mathcal{A}$, $r_a := (r(s,a))_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ is the reward vector corresponding to a given action. Next, we rewrite the second term in (31) as

$$
\begin{aligned}
&\sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \left(\prod_{k=m}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \\
&\qquad P_{d_0(s_0)}(s_m, s_{m+1}) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) \\
&\qquad \left(\sum_{a_0\in\mathcal{A}} q_{d_0(s_0)}(a_0) P(s_{m+1}|s_m, a_0)\right) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right) \\
&= \sum_{a_0\in\mathcal{A}} q_{d_0(s_0)}(a_0) \left[\sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right)\right. \\
&\qquad \left. P(s_{m+1}|s_m, a_0) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right)\right]
\end{aligned}
$$

Putting the two expressions together gives

$$
\begin{aligned}
v_m^{\mu_0:\mu_{m-1},\pi}(s_0) &= \sum_{a_0\in\mathcal{A}} q_{d_0(s_0)}(a_0) \left[(P_{\mu_0}\cdots P_{\mu_{m-1}} r_{a_0})(s_0) + \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m})\right. \\
&\qquad \left.\left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) P(s_{m+1}|s_m, a_0) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right)\right] \\
&\overset{(1)}{\leqslant} \max_{a_0\in\mathcal{A}} \left\{(P_{\mu_0}\cdots P_{\mu_{m-1}} r_{a_0})(s_0) + \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t\in\mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m})\right. \\
&\qquad \left.\left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1})\right) P(s_{m+1}|s_m, a_0) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1})\right)\right\} \\
&\overset{(2)}{:=} v_m^{\mu_0:\mu_{m-1},\pi_{\mathrm{D}}^0}(s_0),
\end{aligned}
$$

where (1) holds by applying (Puterman, 2014)[Lemma 4.3.1], and (2) by defining policy $\pi_{\mathrm{D}}^0 := (d_0^{\mathrm{D}}, d_1, d_2 \cdots) \in \Pi^{\mathrm{MR}}$ such that the first decision rule is deterministic $d_0^{\mathrm{D}} := \delta_{a_0^*}$ with a_0^* the argmax of (1), while $(d_1, d_2, \cdots)$ are the same as in the original policy $\pi \in \Pi^{\mathrm{MR}}$.

We now continue to showing the dependence of the value on $d_i(s_i)$ for $i \geqslant 1$, by continuing with the second term in (31):

$$\sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t \in \mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*) \left(\prod_{k=m+1}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1}) \right)$$

$$= \gamma \sum_{s_1,\cdots,s_{m+1} \in \mathcal{S}} R_{d_1}(s_{m+1}, s_1) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$+ \sum_{t=m+2}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t \in \mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$P_{d_1(s_1)}(s_{m+1}, s_{m+2}) \left(\prod_{k=m+2}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1}) \right)$$

$$= \sum_{a_1 \in \mathcal{A}} q_{d_1(s_1)}(a_1) \Bigg[\gamma \sum_{s_1,\cdots,s_{m+1} \in \mathcal{S}} r(s_{m+1}, a_1) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$+ \sum_{t=m+2}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t \in \mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$P(s_{m+2}|s_{m+1}, a_1) \left(\prod_{k=m+2}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1}) \right) \Bigg]$$

$$\leqslant \max_{a_1 \in \mathcal{A}} \Bigg\{ \gamma \sum_{s_1,\cdots,s_{m+1} \in \mathcal{S}} r(s_{m+1}, a_1) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$+ \sum_{t=m+2}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t \in \mathcal{S}} R_{d_{t-m}}(s_t, s_{t-m}) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) P(s_{m+1}|s_m, a_0^*)$$
$$P(s_{m+2}|s_{m+1}, a_1) \left(\prod_{k=m+2}^{t-1} P_{d_{k-m}(s_{k-m})}(s_k, s_{k+1}) \right) \Bigg\}$$

$$\vdots$$

$$\leqslant \sum_{t=m+1}^{\infty} \gamma^{t-m} \sum_{s_1,\cdots,s_t \in \mathcal{S}} r(s_t, a_{t-m}^*) \left(\prod_{k=0}^{m-1} P_{\mu_k}(s_k, s_{k+1}) \right) \left(\prod_{k=m+1}^{t-1} P(s_{k+1}|s_k, a_{k-m}^*) \right).$$

Let the deterministic decision rule $d_i^{\mathrm{D}} := \delta_{a_i^*}$ with a_i^* being the optimal action per each maximization above for every $i \geqslant 1$, and the resulting deterministic policy $\pi_{\mathrm{D}} := (d_0^{\mathrm{D}}, d_1^{\mathrm{D}}, \cdots)$. Then,

$$v_m^{\mu_0:\mu_{m-1},\pi}(s_0) \leqslant v_m^{\mu_0:\mu_{m-1},\pi_{\mathrm{D}}}(s_0),$$

i.e.,

$$\max_{\pi \in \Pi^{\mathrm{MR}}} v_m^{\mu_0:\mu_{m-1},\pi}(s_0) \leqslant \max_{\pi \in \Pi^{\mathrm{MD}}} v_m^{\mu_0:\mu_{m-1},\pi}(s_0).$$

□

D EXPERIMENTS

D.1 NUMERICAL SUMMARY OF ATARI RESULTS

	Environment	$m = 5$			$m = 15$			$m = 25$		
		Del.	Aug.	Obl.	Del.	Aug.	Obl.	Del.	Aug.	Obl.
Tabular	Maze	$\mathbf{0.50} \pm 0.49$	0.18 ± 0.54	-0.61 ± 0.26	$\mathbf{-0.25} \pm 0.43$	-1 ± 0	-0.76 ± 0.16	$\mathbf{-0.45} \pm 0.36$	N/A	-0.71 ± 0.19
	Noisy Maze	$\mathbf{0.40} \pm 0.44$	-0.29 ± 0.45	-0.64 ± 0.21	$\mathbf{-0.50} \pm 0.26$	-1 ± 0	-0.78 ± 0.15	$\mathbf{-0.49} \pm 0.34$	N/A	-0.99 ± 0
Physical	Cartpole	$\mathbf{489} \pm 11$	453 ± 16	27 ± 4	$\mathbf{414} \pm 14$	192 ± 15	30 ± 3	$\mathbf{324} \pm 7$	41 ± 2	41 ± 3
	Noisy Cartpole	$\mathbf{435} \pm 8$	379 ± 17	26 ± 3	$\mathbf{251} \pm 22$	129 ± 24	30 ± 3	$\mathbf{60} \pm 7$	36 ± 3	40 ± 3
	Acrobot	$\mathbf{-131} \pm 32$	-463 ± 18	-467 ± 47	$\mathbf{-211} \pm 53$	-481 ± 21	-467 ± 34	$\mathbf{-351} \pm 57$	-493 ± 5	-465 ± 20
	Noisy Acrobot	$\mathbf{-134} \pm 37$	-491 ± 2	-445 ± 11	$\mathbf{-329} \pm 24$	-425 ± 41	-399 ± 41	$\mathbf{-361} \pm 62$	-471 ± 12	-438 ± 39
Atari	Enduro	16 ± 6	29 ± 4	$\mathbf{33} \pm 2$	$\mathbf{1.4} \pm 0.4$	0.6 ± 0.7	0.5 ± 0.2	$\mathbf{1.1} \pm 0.6$	0.2 ± 0.1	0.2 ± 0.4
	MsPacman	$\mathbf{1354} \pm 86$	1083 ± 60	1319 ± 35	$\mathbf{1034} \pm 124$	691 ± 272	701 ± 123	$\mathbf{959} \pm 77$	450 ± 84	612 ± 23
	NameThisGame	$\mathbf{2476} \pm 96$	2278 ± 167	2153 ± 152	$\mathbf{2122} \pm 132$	1573 ± 43	2013 ± 300	$\mathbf{1887} \pm 204$	1510 ± 210	1775 ± 96
	Qbert	367 ± 19	372 ± 177	$\mathbf{402} \pm 152$	$\mathbf{304} \pm 15$	245 ± 29	254 ± 34	$\mathbf{253} \pm 29$	154 ± 77	200 ± 74
	RoadRunner	$\mathbf{2975} \pm 237$	1790 ± 255	1152 ± 430	$\mathbf{1294} \pm 472$	1153 ± 119	360 ± 204	$\mathbf{1056} \pm 698$	668 ± 268	485 ± 451
	StarGunner	902 ± 74	838 ± 104	$\mathbf{919} \pm 44$	$\mathbf{801} \pm 38$	622 ± 68	643 ± 50	$\mathbf{712} \pm 49$	649 ± 47	635 ± 20
	TimePilot	$\mathbf{1941} \pm 133$	1844 ± 599	1616 ± 474	$\mathbf{2695} \pm 418$	2049 ± 665	2341 ± 72	$\mathbf{2690} \pm 201$	2671 ± 127	1980 ± 623
	Zaxxon	$\mathbf{1418} \pm 148$	431 ± 77	605 ± 66	$\mathbf{461} \pm 185$	97 ± 65	225 ± 19	$\mathbf{130} \pm 42$	72 ± 22	67 ± 35

Table 2: Experiment summary: episodic return mean and std for all domains. Delayed-Q outperforms the alternatives in 39 of 42 experiments.

D.2 Comparison to RNN-Based Policy

In one environment, we compared Delayed-Q with a fourth algorithm which uses an RNN-based policy that is unaware of the delay value. Specifically, we tested A2C, which managed to converge on Atari's Frostbite. As can be seen in Fig. 12, using a recurrent policy does not improve upon Augmented-Q or Oblivious-Q. This result is not surprising though: as stated in Thm. 5.1, the sequence of states $s_{t-m}, s_{t-m-1}, \ldots$ does not aid the policy any further than only using s_{t-m}. An additional deficiency of RNN-policies that are oblivious to the delay value is that, similarly to Oblivious-Q, they target the wrong Q-value without accounting for delayed execution. Notice that this is not the case in both Augmented-Q and Delayed-Q.

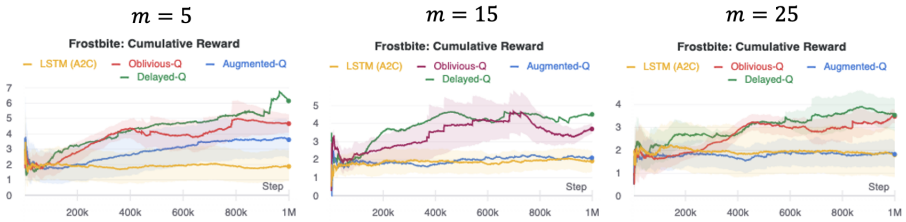

Figure 12: Comparison to RNN-based policy on Atari "Frostbite".

REFERENCES

Avner Bar-Ilan and Agnès Sulem. Explicit solution of inventory problems with delivery lags. *Mathematics of Operations Research*, 20(3):709–720, 1995.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. arXiv:1606.01540v1, 2016.

Benjamin Bruder and Huyên Pham. Impulse control problem on finite horizon with execution delay. *Stochastic Processes and their Applications*, 119(5):1436–1469, 2009.

Jeffrey S Campbell, Sidney N Givigi, and Howard M Schwartz. Multiple model q-learning for stochastic asynchronous rewards. *Journal of Intelligent & Robotic Systems*, 81(3-4):407–422, 2016.

Baiming Chen, Mengdi Xu, Liang Li, and Ding Zhao. Delay-aware model-based reinforcement learning for continuous control. *arXiv preprint arXiv:2005.05440*, 2020a.

Baiming Chen, Mengdi Xu, Zuxin Liu, Liang Li, and Ding Zhao. Delay-aware multi-agent reinforcement learning. *arXiv preprint arXiv:2005.05441*, 2020b.

Luc Dugard and Erik I Verriest. *Stability and control of time-delay systems*, volume 228. Springer, 1998.

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

John Fearnley. Exponential lower bounds for policy iteration. In *International Colloquium on Automata, Languages, and Programming*, pp. 551–562. Springer, 2010.

Vlad Firoiu, Tina Ju, and Josh Tenenbaum. At human speed: Deep reinforcement learning with action delay. *arXiv preprint arXiv:1810.07286*, 2018.

Emilia Fridman. *Introduction to time-delay systems: Analysis and control*. Springer, 2014.

Thomas Dueholm Hansen and Uri Zwick. Lower bounds for howard's algorithm for finding minimum mean-cost cycles. In *International Symposium on Algorithms and Computation*, pp. 415–426. Springer, 2010.

Todd Hester and Peter Stone. Texplore: real-time sample-efficient reinforcement learning for robots. *Machine learning*, 90(3):385–429, 2013.

Romain Hollanders, Jean-Charles Delvenne, and Raphaël M Jungers. The complexity of policy iteration is exponential for discounted markov decision processes. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 5997–6002. IEEE, 2012.

Ronald A Howard. Dynamic programming and Markov processes. *John Wiley*, 1960.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pp. 1453–1461, 2013.

Konstantinos V Katsikopoulos and Sascha E Engelbrecht. Markov decision processes with delays and asynchronous cost collection. *IEEE transactions on automatic control*, 48(4):568–574, 2003.

Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with gamegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1231–1240, 2020.

Wei Niu, Xiaolong Ma, Yanzhi Wang, and Bin Ren. 26ms inference time for resnet-50: Towards real-time execution of all dnns on smartphone. *arXiv preprint arXiv:1905.00571*, 2019.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grünewälder. Bandits with delayed anonymous feedback. *stat*, 1050:20, 2017.

Martin L Puterman. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Simon Ramstedt and Chris Pal. Real-time reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 3073–3082, 2019.

Jean-Pierre Richard. Time-delay systems: an overview of some recent advances and open problems. *automatica*, 39(10):1667–1694, 2003.

Bruno Scherrer et al. Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41(3):758–774, 2016.

Alessandro Toschi, Mustafa Sanic, Jingwen Leng, Quan Chen, Chunlin Wang, and Minyi Guo. Characterizing perception module performance and robustness in production-scale autonomous driving system. In *IFIP International Conference on Network and Parallel Computing*, pp. 235–247. Springer, 2019.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *arXiv preprint arXiv:1509.06461*, 2015.

Thomas J Walsh, Ali Nouri, Lihong Li, and Michael L Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83, 2009.

Ted Xiao, Eric Jang, Dmitry Kalashnikov, Sergey Levine, Julian Ibarz, Karol Hausman, and Alexander Herzog. Thinking while moving: Deep reinforcement learning with concurrent control. *arXiv preprint arXiv:2004.06089*, 2020.

Hengyu Zhao, Yubo Zhang, Pingfan Meng, Hui Shi, Li Erran Li, Tiancheng Lou, and Jishen Zhao. Towards safety-aware computing system design in autonomous vehicles. *arXiv preprint arXiv:1905.08453*, 2019.