Control Frequency Adaptation via Action Persistence in Batch Reinforcement Learning

Alberto Maria Metelli   Flavio Mazzolini
Lorenzo Bisi   Luca Sabbioni   Marcello Restelli

July 2020
Thirty-seventh International Conference on Machine Learning

Motivations

- **Problem**: How to select the *control frequency* for a system?

Lower Frequencies

Higher Frequencies

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency?

Motivations

- **Problem**: How to select the *control frequency* for a system?

Lower Frequencies

Higher Frequencies

Control Opportunities

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency?

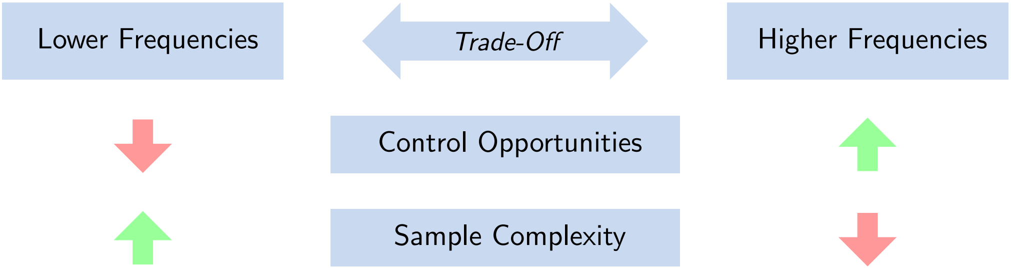
- **Problem**: How to select the *control frequency* for a system?

Lower Frequencies		Higher Frequencies
↓	Control Opportunities	↑
↑	Sample Complexity	↓

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency?

Motivations

- **Problem**: How to select the *control frequency* for a system?



- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency?

Control Frequency and Action Persistence

- **Idea**: *persisting* each action for k steps

	continuous-time MDP		discrete-time MDP		k-persistent MDP
	$\mathcal{M}_0$	*time discretization*	$\mathcal{M}_{\Delta t}$	*action persistence k*	$\mathcal{M}_{k\Delta t}$
control time-step	0		Δt		$k\Delta t$
control frequency	∞		$f = \frac{1}{\Delta t}$		$\frac{f}{k}$

- Action persistence as form of environment *configurability* (Metelli et al., 2018)

Control Frequency and Action Persistence

- **Idea**: *persisting* each action for k steps

	continuous-time MDP		discrete-time MDP		k-persistent MDP
	$\mathcal{M}_0$	*time discretization*	$\mathcal{M}_{\Delta t}$	*action persistence k*	$\mathcal{M}_{k\Delta t}$
control time-step	0		Δt		$k\Delta t$
control frequency	∞		$f = \frac{1}{\Delta t}$		$\frac{f}{k}$

- Action persistence as form of environment *configurability* (Metelli et al., 2018)

Control Frequency and Action Persistence

- **Idea**: *persisting* each action for k steps

	continuous-time MDP		discrete-time MDP		k-persistent MDP
	$\mathcal{M}_0$	*time discretization* →	$\mathcal{M}_{\Delta t}$	*action persistence k* →	$\mathcal{M}_{k\Delta t}$
control time-step	0		Δt		$k\Delta t$
control frequency	∞		$f = \frac{1}{\Delta t}$		$\frac{f}{k}$

- Action persistence as form of environment *configurability* (Metelli et al., 2018)

Control Frequency and Action Persistence

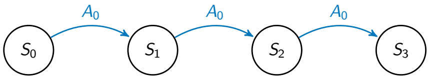
- **Idea**: *persisting* each action for k steps

	continuous-time MDP		discrete-time MDP		k-persistent MDP
	$\mathcal{M}_0$	*time discretization* →	$\mathcal{M}_{\Delta t}$	*action persistence* k →	$\mathcal{M}_{k\Delta t}$
control time-step	0		Δt		$k\Delta t$
control frequency	∞		$f = \frac{1}{\Delta t}$		$\frac{f}{k}$

- Action persistence as form of environment *configurability* (Metelli et al., 2018)

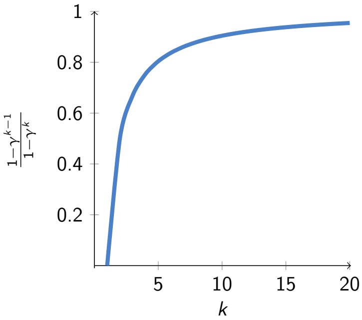
Outline

1. Action persistence formalization
2. Performance loss due to persistence
3. Persistent Fitted Q-Iteration



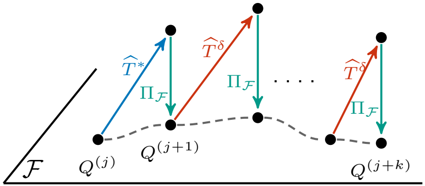
Outline

1. Action persistence formalization
2. Performance loss due to persistence
3. Persistent Fitted Q-Iteration



Outline

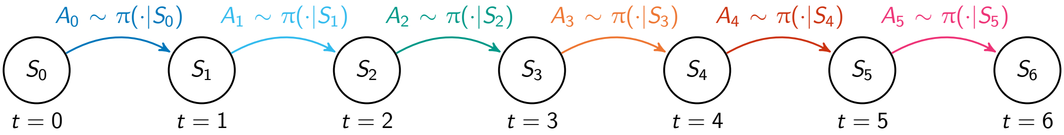
1. Action persistence formalization
2. Performance loss due to persistence
3. Persistent Fitted Q-Iteration



No Action Persistence

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \quad \text{and} \quad \pi$$

- $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$ is Markovian and Stationary (Puterman, 2014; Sutton and Barto, 2018)



Change the policy → k-persistent policy

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \quad \text{and} \quad \pi_k$$

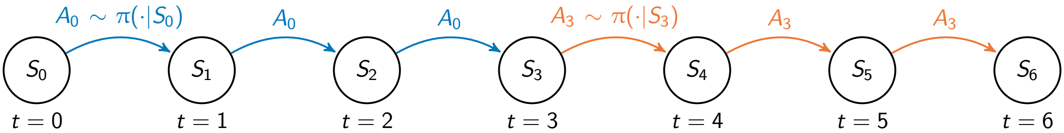
$$\pi_{t,k}(a|h_t) = \begin{cases} \pi(a|s_t) & \text{if } t \bmod k = 0 \\ \delta_{a_{t-1}}(a) & \text{otherwise} \end{cases}$$

- History $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$
- π_k is Non-Markovian and Non-Stationary

Change the policy $\rightarrow$ **k-persistent policy**

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma) \quad \text{and} \quad \pi_k \qquad \pi_{t,k}(a|h_t) = \begin{cases} \pi(a|s_t) & \text{if } t \bmod k = 0 \\ \delta_{a_{t-1}}(a) & \text{otherwise} \end{cases}$$

- History $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$
- π_k is Non-Markovian and Non-Stationary

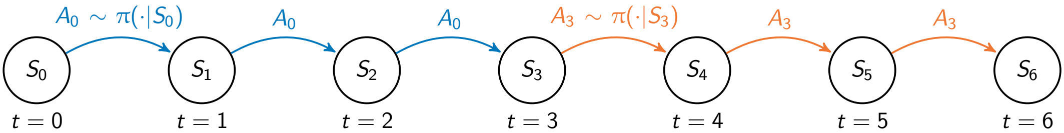


Change the policy $\rightarrow$ k-**persistent policy**

$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and π_k

$$\pi_{t,k}(a|h_t) = \begin{cases} \pi(a|s_t) & \text{if } t \bmod k = 0 \\ \delta_{a_{t-1}}(a) & \text{otherwise} \end{cases}$$

- History $h_t = (s_0, a_0, \dots, s_{t-1}, a_{t-1}, s_t)$
- π_k is Non-Markovian and Non-Stationary



Change the MDP → k-persistent MDP

$$\mathcal{M}_k = (\mathcal{S}, \mathcal{A}, P_k, R_k, \gamma^k) \quad \text{and} \quad \pi$$

$$P_k(s'|s,a) = \left((P^\delta)^{k-1} P\right)(s'|s,a)$$

$$R_k(s'|s,a) = \sum_{i=0}^{k-1} \gamma^i \left((P^\delta)^i R\right)(s'|s,a)$$

- *Persistent* state-action kernel $P^\delta(s', a'|s,a) = \delta_{a'}(a) P(s'|s,a)$
- $\mathcal{M}_k$ has *smaller* discount factor γ^k

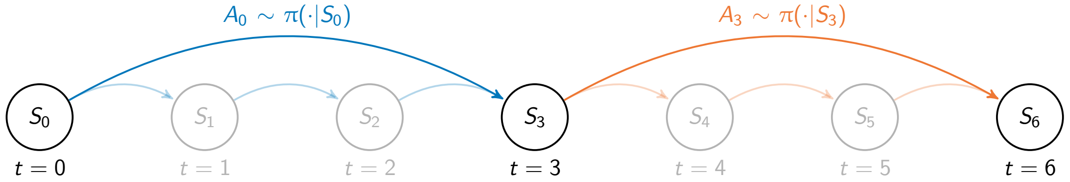
Change the MDP → k-**persistent MDP**

$$\mathcal{M}_k = \left(\mathcal{S}, \mathcal{A}, P_k, R_k, \gamma^k\right) \quad \text{and} \quad \pi$$

$$P_k(s'|s,a) = \left((P^\delta)^{k-1} P\right)(s'|s,a)$$

$$R_k(s'|s,a) = \sum_{i=0}^{k-1} \gamma^i \left((P^\delta)^i R\right)(s'|s,a)$$

- *Persistent* state-action kernel $P^\delta(s', a'|s,a) = \delta_{a'}(a) P(s'|s,a)$
- $\mathcal{M}_k$ has *smaller* discount factor γ^k



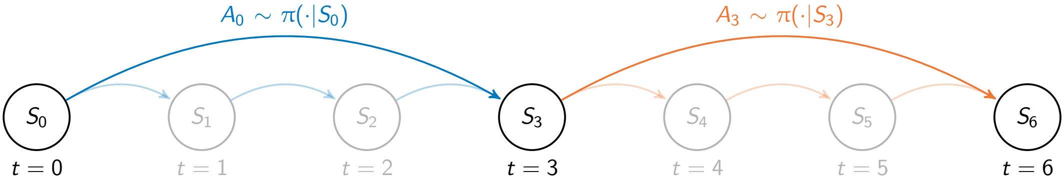
Change the MDP $\rightarrow$ **k-persistent MDP**

$$\mathcal{M}_k = \left(\mathcal{S}, \mathcal{A}, P_k, R_k, \gamma^k\right) \quad \text{and} \quad \pi$$

$$P_k(s'|s,a) = \left((P^\delta)^{k-1} P\right)(s'|s,a)$$

$$R_k(s'|s,a) = \sum_{i=0}^{k-1} \gamma^i \left((P^\delta)^i R\right)(s'|s,a)$$

- *Persistent* state-action kernel $P^\delta(s', a'|s,a) = \delta_{a'}(a) P(s'|s,a)$
- $\mathcal{M}_k$ has *smaller* discount factor γ^k



MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_k$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T_{\boldsymbol{k}}^* = \left(T^\delta\right)^{\boldsymbol{k}-1} T^*$$

- $T_{\boldsymbol{k}}^*$ is a γ^k-contraction in L_∞-norm
- $Q_{\boldsymbol{k}}^*$ is the unique fixed point of $T_{\boldsymbol{k}}^*$

$$T_{\boldsymbol{k}}^* Q_{\boldsymbol{k}}^* = Q_{\boldsymbol{k}}^*$$

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_k$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T_{\boldsymbol{k}}^* = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T_{\boldsymbol{k}}^*$ is a γ^k-contraction in L_∞-norm
- $Q_{\boldsymbol{k}}^*$ is the unique fixed point of $T_{\boldsymbol{k}}^*$

$$T_{\boldsymbol{k}}^* Q_{\boldsymbol{k}}^* = Q_{\boldsymbol{k}}^*$$

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_k$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T^*_{\boldsymbol{k}} = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T^*_{\boldsymbol{k}}$ is a γ^k-contraction in L_∞-norm
- $Q^*_{\boldsymbol{k}}$ is the unique fixed point of $T^*_{\boldsymbol{k}}$

$$T^*_{\boldsymbol{k}} Q^*_{\boldsymbol{k}} = Q^*_{\boldsymbol{k}}$$

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_{\boldsymbol{k}}$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T^*_{\boldsymbol{k}} = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T^*_{\boldsymbol{k}}$ is a γ^k-contraction in L_∞-norm
- $Q^*_{\boldsymbol{k}}$ is the unique fixed point of $T^*_{\boldsymbol{k}}$

$$T^*_{\boldsymbol{k}} Q^*_{\boldsymbol{k}} = Q^*_{\boldsymbol{k}}$$

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_{\boldsymbol{k}}$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T^*_{\boldsymbol{k}} = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T^*_{\boldsymbol{k}}$ is a γ^k-contraction in L_∞-norm
- $Q^*_{\boldsymbol{k}}$ is the unique fixed point of $T^*_{\boldsymbol{k}}$

$$T^*_{\boldsymbol{k}} Q^*_{\boldsymbol{k}} = Q^*_{\boldsymbol{k}}$$

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_{\boldsymbol{k}}$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T^*_{\boldsymbol{k}} = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T^*_{\boldsymbol{k}}$ is a γ^k-contraction in L_∞-norm
- $Q^*_{\boldsymbol{k}}$ is the unique fixed point of $T^*_{\boldsymbol{k}}$

$$T^*_{\boldsymbol{k}} Q^*_{\boldsymbol{k}} = Q^*_{\boldsymbol{k}}$$

Persistent Bellman Operators

MDP $\mathcal{M}$

- Bellman Operator (Bertsekas, 2005)

$$(T^* f)(s,a) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) \max_{a' \in \mathcal{A}} f(s',a')$$

- T^* is a γ-contraction in L_∞-norm
- Q^* is the unique fixed point of T^*

$$T^* Q^* = Q^*$$

k-persistent MDP $\mathcal{M}_{\boldsymbol{k}}$

- Persistence Operator

$$(T^\delta f)(s,\boldsymbol{a}) = r(s,a) + \gamma \int_{\mathcal{S}} P(\mathrm{d}s'|s,a) f(s',\boldsymbol{a})$$

- k-persistent Bellman Operator

$$T^*_{\boldsymbol{k}} = (T^\delta)^{\boldsymbol{k}-1} T^*$$

- $T^*_{\boldsymbol{k}}$ is a γ^k-contraction in L_∞-norm
- $Q^*_{\boldsymbol{k}}$ is the unique fixed point of $T^*_{\boldsymbol{k}}$

$$T^*_{\boldsymbol{k}} Q^*_{\boldsymbol{k}} = Q^*_{\boldsymbol{k}}$$

Bounding the Performance Loss

- $Q_k^* \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions of policy π?

$$\|Q^\pi - Q_k^\pi\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \quad \frac{1-\gamma^{k-1}}{1-\gamma^k} \quad \left\| d(P^\pi, P^\delta) \right\|_{p,\mu}$$

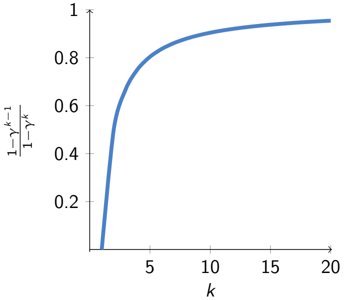
- Increasing with k
- $d(P^\pi, P^\delta)$: discrepancy between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)

Bounding the Performance Loss

- $Q_k^* \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions of policy π?

$$\|Q^\pi - Q_k^\pi\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \boxed{\frac{1-\gamma^{k-1}}{1-\gamma^k}} \left\| d(P^\pi, P^\delta) \right\|_{p,\mu}$$

- Increasing with k
- $d(P^\pi, P^\delta)$: discrepancy between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)



Bounding the Performance Loss

- $Q_k^* \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions of policy π?

$$\|Q^\pi - Q_k^\pi\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \frac{1-\gamma^{k-1}}{1-\gamma^k} \boxed{\left\| d(P^\pi, P^\delta) \right\|_{p,\mu}}$$

- Increasing with k
- $d(P^\pi, P^\delta)$: discrepancy between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)

$$P^{\boldsymbol{\pi}}(s', a'|s, a) = \pi(a'|s') P(s'|s, a)$$

$$P^{\boldsymbol{\delta}}(s', a'|s, a) = \delta_{a'}(a) P(s'|s, a)$$

Bounding the Performance Loss

- $Q_k^* \leqslant Q^*$ for all $k \geqslant 1$
- How much do we lose by persisting k times the actions of policy π?

$$\|Q^\pi - Q_k^\pi\|_{p,\mu} \leqslant \frac{\gamma}{1-\gamma} \quad \frac{1-\gamma^{k-1}}{1-\gamma^k} \quad \left\| d(P^\pi, P^\delta) \right\|_{p,\mu}$$

- Increasing with k
- $d(P^\pi, P^\delta)$: discrepancy between transition kernels
  - Can be bounded under Lipschitz conditions (Rachelson and Lagoudakis, 2010)

$$\left\| d(P^\pi, P^\delta) \right\|_{p,\mu} \leqslant L\left[(L_\pi + 1)L_T + \sigma_p\right]$$

Fitted Q-Iteration
(Ernst et al., 2005)

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

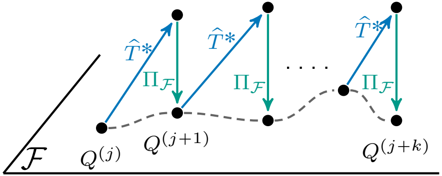
$$Q^{(j+1)} = \Pi_{\mathcal{F}} \widehat{T}^* Q^{(j)}$$

- $Q^{(j)} \rightsquigarrow Q^*$
- What about Q_k^*?

Empirical Bellman Operators

$$(\widehat{T}^* f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$

$$T^* \simeq \Pi_{\mathcal{F}} \widehat{T}^*$$



Persistent Fitted Q-Iteration (PFQI)

Persistent Fitted Q-Iteration

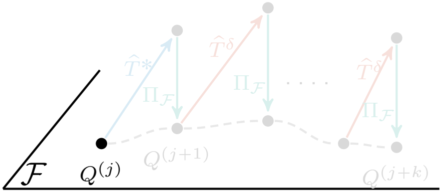
- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q_k^{*}$

Empirical Bellman Operators



Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

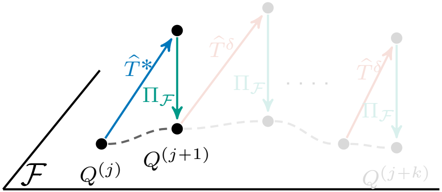
$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^* Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

- $Q^{(j)} \rightsquigarrow Q_k^*$

Empirical Bellman Operators

$$(\widehat{T}^* f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$



Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

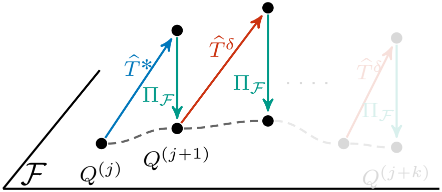
- $Q^{(j)} \rightsquigarrow Q^{*}_{k}$

Empirical Bellman Operators

$$(\widehat{T}^{*} f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$

$$(\widehat{T}^{\delta} f)(S_i, A_i) = R_i + \gamma f(S_{i+1}, A_i)$$

$$T^{*}_{k} = (T^{\delta})^{k-1} T^{*} \simeq (\Pi_{\mathcal{F}} \widehat{T}^{\delta})^{k-1} \Pi_{\mathcal{F}} \widehat{T}^{*}$$



Persistent Fitted Q-Iteration

- Approximation space $\mathcal{F}$
- Initial estimate $Q^{(0)}$
- Dataset

$$\mathcal{D} = \{(S_i, A_i, S_{i+1}, R_i)\}_{i=1}^{n} \sim \nu$$

$$Q^{(j+1)} = \begin{cases} \Pi_{\mathcal{F}} \widehat{T}^{*} Q^{(j)} & \text{if } j \bmod k = 0 \\ \Pi_{\mathcal{F}} \widehat{T}^{\delta} Q^{(j)} & \text{otherwise} \end{cases}$$

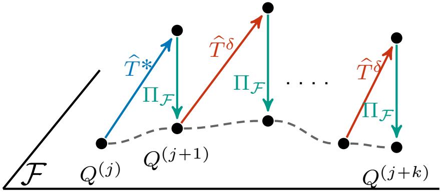
- $Q^{(j)} \rightsquigarrow Q_k^{*}$

Empirical Bellman Operators

$$(\widehat{T}^{*} f)(S_i, A_i) = R_i + \gamma \max_{a \in \mathcal{A}} f(S_{i+1}, a)$$

$$(\widehat{T}^{\delta} f)(S_i, A_i) = R_i + \gamma f(S_{i+1}, A_i)$$

$$T_k^{*} = (T^{\delta})^{k-1} T^{*} \simeq (\Pi_{\mathcal{F}} \widehat{T}^{\delta})^{k-1} \Pi_{\mathcal{F}} \widehat{T}^{*}$$



PFQI Analysis

- **Computational Complexity**: monotonically decreasing with k

$$\mathcal{O}\left(Jn\left(1+\frac{|\mathcal{A}|-1}{k}\right)\right) \quad \text{for } J \text{ iterations}$$

- **Error propagation**

$$\left\|Q_k^* - Q_k^{\pi^{(J)}}\right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \quad \frac{\gamma^k}{1-\gamma^k} \quad \mathcal{E}(J,\mu,\nu,p)$$

- Decreasing with k
- Approximation errors $\epsilon^{(j)}$ and concentrability coefficients (Farahmand, 2011)

$$\epsilon^{(j)} = \begin{cases} T^* Q^{(j)} - Q^{(j+1)} & \text{if } j \bmod k = 0 \\ T^\delta Q^{(j)} - Q^{(j+1)} & \text{otherwise} \end{cases}$$

PFQI Analysis

- **Computational Complexity**: monotonically decreasing with k

$$\mathcal{O}\left(Jn\left(1+\frac{|\mathcal{A}|-1}{k}\right)\right) \quad \text{for } J \text{ iterations}$$

- **Error propagation**

$$\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \quad \frac{\gamma^k}{1-\gamma^k} \quad \mathcal{E}(J,\mu,\nu,p)$$

- Decreasing with k
- Approximation errors $\epsilon^{(j)}$ and concentrability coefficients (Farahmand, 2011)

$$\epsilon^{(j)} = \begin{cases} T^* Q^{(j)} - Q^{(j+1)} & \text{if } j \bmod k = 0 \\ T^{\delta} Q^{(j)} - Q^{(j+1)} & \text{otherwise} \end{cases}$$

PFQI Analysis

- **Computational Complexity**: monotonically decreasing with k

$$\mathcal{O}\left(Jn\left(1+\frac{|\mathcal{A}|-1}{k}\right)\right) \quad \text{for } J \text{ iterations}$$

- **Error propagation**

$$\left\|Q_k^* - Q_k^{\pi^{(J)}}\right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \boxed{\frac{\gamma^k}{1-\gamma^k}} \mathcal{E}(J,\mu,\nu,p)$$

- Decreasing with k
- Approximation errors $\epsilon^{(j)}$ and concentrability coefficients (Farahmand, 2011)

$$\epsilon^{(j)} = \begin{cases} T^* Q^{(j)} - Q^{(j+1)} & \text{if } j \bmod k = 0 \\ T^\delta Q^{(j)} - Q^{(j+1)} & \text{otherwise} \end{cases}$$

PFQI Analysis

- **Computational Complexity**: monotonically decreasing with k

$$\mathcal{O}\left(Jn\left(1+\frac{|\mathcal{A}|-1}{k}\right)\right) \quad \text{for } J \text{ iterations}$$

- **Error propagation**

$$\left\|Q_k^* - Q_k^{\pi^{(J)}}\right\|_{p,\mu} \leqslant \frac{2}{1-\gamma} \quad \frac{\gamma^k}{1-\gamma^k} \quad \boxed{\mathcal{E}(J,\mu,\nu,p)}$$

  - Decreasing with k
  - Approximation errors $\epsilon^{(j)}$ and concentrability coefficients (Farahmand, 2011)

$$\epsilon^{(j)} = \begin{cases} T^* Q^{(j)} - Q^{(j+1)} & \text{if } j \bmod k = 0 \\ T^\delta Q^{(j)} - Q^{(j+1)} & \text{otherwise} \end{cases}$$

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \left\| Q^* - Q_k^* \right\|_{p,\mu} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$

- Control Opportunities
- Algorithm-independent
- **Increasing with k**

- Sample Complexity
- Algorithm-dependent
- **Decreasing with k**

- How to identify the optimal persistence?

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \boxed{\left\| Q^* - Q_k^* \right\|_{p,\mu}} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$

- Control Opportunities
- Algorithm-independent
- **Increasing with** k

- Sample Complexity
- Algorithm-dependent
- **Decreasing with** k

- How to identify the optimal persistence?

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \left\| Q^* - Q_k^* \right\|_{p,\mu} + \boxed{\left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}}$$

- Control Opportunities
- Algorithm-independent
- **Increasing with k**

- Sample Complexity
- Algorithm-dependent
- **Decreasing with k**

- How to identify the optimal persistence?

Control Frequency Trade-Off

$$\left\| Q^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu} \leqslant \left\| Q^* - Q_k^* \right\|_{p,\mu} + \left\| Q_k^* - Q_k^{\pi^{(J)}} \right\|_{p,\mu}$$

- Control Opportunities
- Algorithm-independent
- **Increasing with** k

- Sample Complexity
- Algorithm-dependent
- **Decreasing with** k

- How to identify the optimal persistence?

Persistence Selection

- How to identify the optimal persistence in a *batch* setting?
- Given estimated Q-function $\{Q_k : k \in \mathcal{K}\}$

$$\widetilde{k} \in \arg\max_{k \in \mathcal{K}} B_k = \widehat{J}_k - \frac{1}{1-\gamma^k} \left\| \widetilde{Q}_k - Q_k \right\|_{\mathcal{D}}$$

- estimated performance derived from Q_k
- $\simeq$ Bellman residual ($\widetilde{Q}_k \simeq T_k^* Q_k$) (Farahmand and Szepesvári, 2011)

Persistence Selection

- How to identify the optimal persistence in a *batch* setting?
- Given estimated Q-function $\{Q_k : k \in \mathcal{K}\}$

$$\widetilde{k} \in \arg\max_{k \in \mathcal{K}} B_k = \boxed{\widehat{J}_k} - \frac{1}{1-\gamma^k} \left\| \widetilde{Q}_k - Q_k \right\|_{\mathcal{D}}$$

- estimated performance derived from Q_k
- $\simeq$ Bellman residual ($\widetilde{Q}_k \simeq T_k^* Q_k$) (Farahmand and Szepesvári, 2011)

Persistence Selection

- How to identify the optimal persistence in a *batch* setting?
- Given estimated Q-function $\{Q_k : k \in \mathcal{K}\}$

$$\widetilde{k} \in \arg\max_{k \in \mathcal{K}} B_k = \widehat{J}_k - \frac{1}{1-\gamma^k} \boxed{\left\| \widetilde{Q}_k - Q_k \right\|_{\mathcal{D}}}$$

- estimated performance derived from Q_k
- $\simeq$ Bellman residual ($\widetilde{Q}_k \simeq T_k^* Q_k$) (Farahmand and Szepesvári, 2011)

- PFQI & ExtraTrees (Geurts et al., 2006)

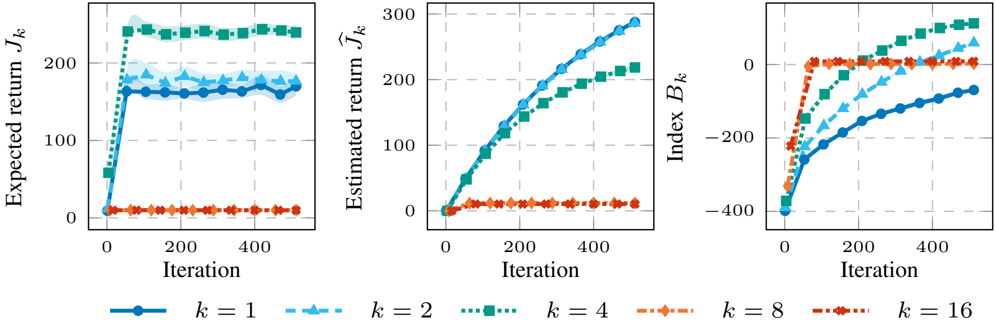
Environments	*Best Persistence*
Cartpole	4
Mountain Car	8, 16, 32
LunarLander	4, 8
Pendulum	1, 2, 4
Acrobot	2, 4
Swimmer	2, 4, 8
Hopper	64
Walker2D	8, 16, 32, 64

- The best persistence **is usually not 1**
- **Excessive increase** of the persistence prevents control at all

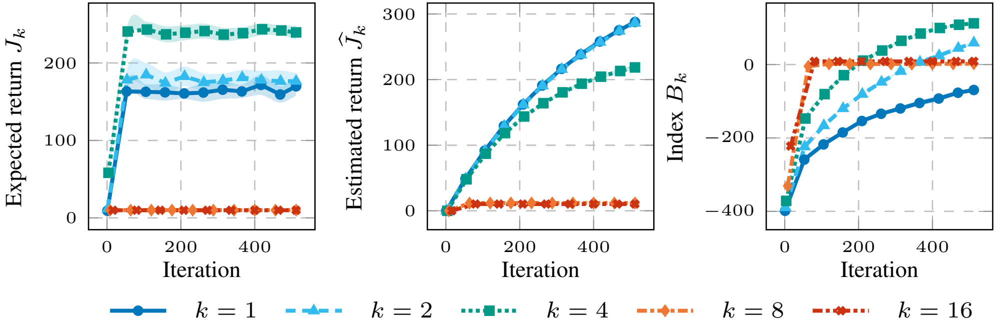
- PFQI & ExtraTrees (Geurts et al., 2006)

Environments	*Best Persistence*
Cartpole	4
Mountain Car	8, 16, 32
LunarLander	4, 8
Pendulum	1, 2, 4
Acrobot	2, 4
Swimmer	2, 4, 8
Hopper	64
Walker2D	8, 16, 32, 64

- The best persistence **is usually not 1**
- **Excessive increase** of the persistence prevents control at all



- **Overestimated** lower persistence Q-functions
- The persistence selection heuristic correctly selects $k = 4$



- **Overestimated** lower persistence Q-functions
- The persistence selection heuristic correctly selects $k = 4$

Conclusions

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency?

- **Open Questions**
  1. Can persistence improve **exploration**?
  2. Persistence in **on–line RL**
  3. **Dynamic** persistent selection

Conclusions

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency? **Yes!**

- **Open Questions**
  1. Can persistence improve **exploration**?
  2. Persistence in **on–line RL**
  3. **Dynamic** persistent selection

Conclusions

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency? **Yes!**

- **Open Questions**
  1. Can persistence improve **exploration**?
  2. Persistence in **on–line RL**
  3. **Dynamic** persistent selection

Conclusions

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency? **Yes!**

- **Open Questions**
  1. Can persistence improve **exploration**?
  2. Persistence in **on–line RL**
  3. **Dynamic** persistent selection

Conclusions

- **Research Question**: Can we exploit this *trade-off* to find an *optimal* control frequency? **Yes!**

- **Open Questions**
  1. Can persistence improve **exploration**?
  2. Persistence in **on–line RL**
  3. **Dynamic** persistent selection

Thank You for Your Attention!

References

Dimitri P. Bertsekas. *Dynamic programming and optimal control, 3rd Edition*. Athena Scientific, 2005. ISBN 1886529264.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6:503–556, 2005.

Amir Massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011.

Amir Massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine Learning*, 85(3):299–332, 2011. doi: 10.1007/s10994-011-5254-7.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

Alberto Maria Metelli, Mirco Mutti, and Marcello Restelli. Configurable markov decision processes. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3488–3497. PMLR, 2018.

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

Emmanuel Rachelson and Michail G. Lagoudakis. On the locality of action domination in sequential decision making. In *International Symposium on Artificial Intelligence and Mathematics, ISAIM 2010, Fort Lauderdale, Florida, USA, January 6-8, 2010*, 2010.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.