Introduction to MDPs and Reinforcement Learning II

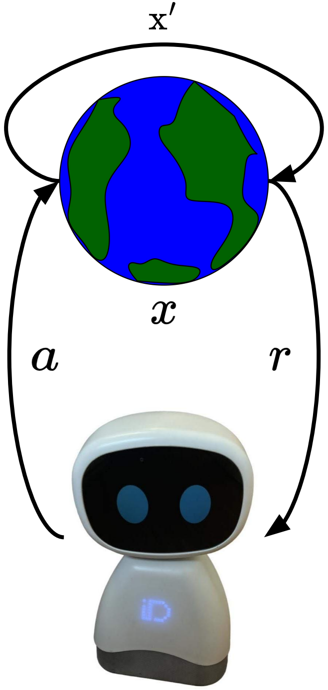
Some Recent Results and Open Problems

Pedro A. Santos (w/ Diogo S. Carvalho,
Francisco S. Melo)

TÉCNICO LISBOA

inesc id lisboa

Reinforcement learning



For each policy π there is the state value function v_π and the state-action value function q_π

There is an optimal policy π^* with state value function v^* and state-action value function q^*

Reinforcement learning

We have a space of functions $\mathcal{Q} = \{q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}\}$, with finite $\mathcal{X} \times \mathcal{A}$

We have an operator $\mathrm{H} : \mathcal{Q} \to \mathcal{Q}$ such that

$$(\mathrm{H}q)(x, a) := \mathbb{E}\left[r(x, a) + \gamma \cdot \mathbf{max}_{a' \in \mathcal{A}} q(\mathrm{x}', a')\right]$$

Goal: We want to solve $q = \mathrm{H}q$, i.e., to find a fixed-point of H

Reinforcement learning

Theorem: H is contractive with contraction factor γ in the infinity-norm

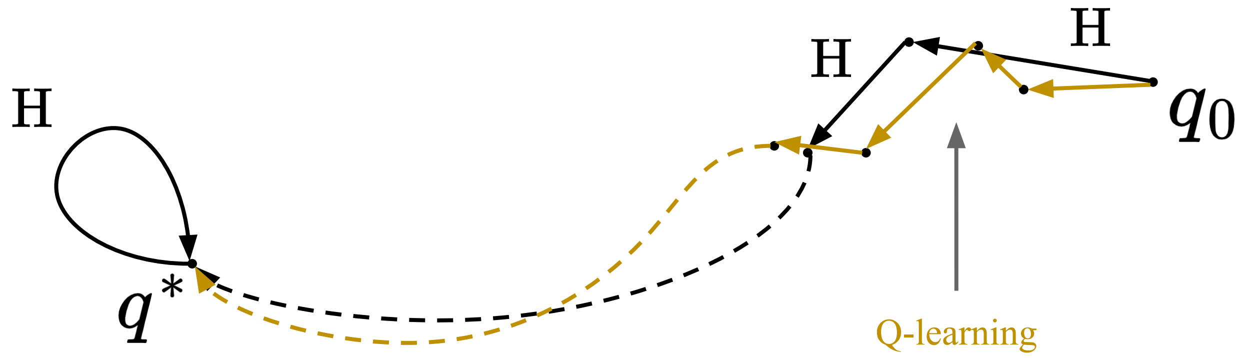
Proof:

$$
\begin{aligned}
\|\mathrm{H}q - \mathrm{H}p\|_\infty &= \max_{x,a\in\mathcal{X}\times\mathcal{A}} |(\mathrm{H}q)(x,a) - (\mathrm{H}p)(x,a)| \\
&= \max_{x,a\in\mathcal{X}\times\mathcal{A}} |\mathbb{E}[r(x,a) + \gamma\cdot\max_{a'\in\mathcal{A}} q(x',a')] - \mathbb{E}[r(x,a) + \gamma\cdot\max_{a'\in\mathcal{A}} p(x',a')]| \\
&= \gamma\cdot\max_{x,a\in\mathcal{X}\times\mathcal{A}} |\mathbb{E}[\max_{a'\in\mathcal{A}} q(x',a') - \max_{a'\in\mathcal{A}} p(x',a')]| \\
&\leq \gamma\cdot\max_{x,a\in\mathcal{X}\times\mathcal{A}} \mathbb{E}[|\max_{a'\in\mathcal{A}} q(x',a') - \max_{a'\in\mathcal{A}} p(x',a')|] \\
&\leq \gamma\cdot\max_{x,a\in\mathcal{X}\times\mathcal{A}} \mathbb{E}[\max_{a'\in\mathcal{A}} |q(x',a') - p(x',a')|] \\
&= \gamma\cdot\max_{x,a\in\mathcal{X}\times\mathcal{A}} \mathbb{E}[\|q-p\|_\infty] \\
&= \gamma\cdot\|q-p\|_\infty
\end{aligned}
$$

Corollary: There exists a unique fixed-point q^* of H

Reinforcement learning

$$q^* = \mathbf{H}q^*$$



Q-learning:

$$q_{t+1} = q_t + \alpha_t(r_t + \gamma \cdot \max_{a' \in \mathcal{A}} q_t(x'_t, a') - q_t(x_t, a_t))\mathbb{1}_{(x_t, a_t)}$$

Stochastic approximation

$$w_{t+1} = w_t + \alpha_t\big(f(w_t) + \mathbf{noise}_t\big)$$

$$\mathrm{H}q_t - q_t$$

Assumptions:

$f : \mathbb{R}^k \to \mathbb{R}$ is Lipschitz

$\sum_{t=0}^{\infty} \alpha_t$ is infinite, $\sum_{t=0}^{\infty} \alpha_t^2$ is finite

$\mathbf{noise}_t : \Omega \to \mathbb{R}^k$ has zero-mean given the past and bounded variance

Theorem:

If the dynamical system governed by the o.d.e. $\dot{w} = f(w)$ has a unique globally asymptotically stable equilibrium w^* $\longleftarrow$ q^* then $w_t \to w^*$ almost surely

Dynamical systems

$\frac{1}{2}\|q^* - q\|_2^2$

Theorem:

$\dot{w} = f(w)$ has a unique globally asymptotically stable equilibrium w^* such that $f(w^*) = 0$ if there is a Lyapunov function $l : \mathbb{R}^k \to \mathbb{R}$ such that

$$l(w) > 0, \quad \text{for all } w \neq w^*, \quad \text{and} \quad l(w^*) = 0$$

$$\dot{l}(w) = (\nabla l(w)) \cdot f(w) < 0, \quad \text{for all } w \neq w^*$$

$$\lim_{\|w\| \to \infty} l(w) = \infty$$

Linear function approximation

We have a linear function approximation space generated by features $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^k$

$$\mathcal{Q} = \{q_w : \mathcal{X} \times \mathcal{A} \to \mathbb{R} \mid q_w(x,a) = \phi(x,a) \cdot w, w \in \mathbb{R}^k\}$$

We have a distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$

Goal: We want to find the function in $\mathcal{Q}$ that is closest to q^*, i.e.,

$$\text{Proj } q^* := \text{argmin}_{q \in \mathcal{Q}} \frac{1}{2} \|q^* - q\|_{2,\mu}^2$$

Where the norm $\|\cdot\|_{2,\mu}$ is $\|q\|_{2,\mu} := \sqrt{\mathbb{E}_\mu[q^2(x,a)]}$

Linear function approximation

Proposition:

The projection is a unique and given by

$$(\text{Proj } q)(x,a) = \phi(x,a) \cdot \mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})\phi^\top(x,a)]^{-1} \cdot \mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})q(\mathrm{x},\mathrm{a})]$$

Proof:

We have that $\|q - q_w\|_{2,\mu}^2 = \mathbb{E}_\mu[(q(\mathrm{x},\mathrm{a}) - \phi(\mathrm{x},\mathrm{a})^\top w)^2]$

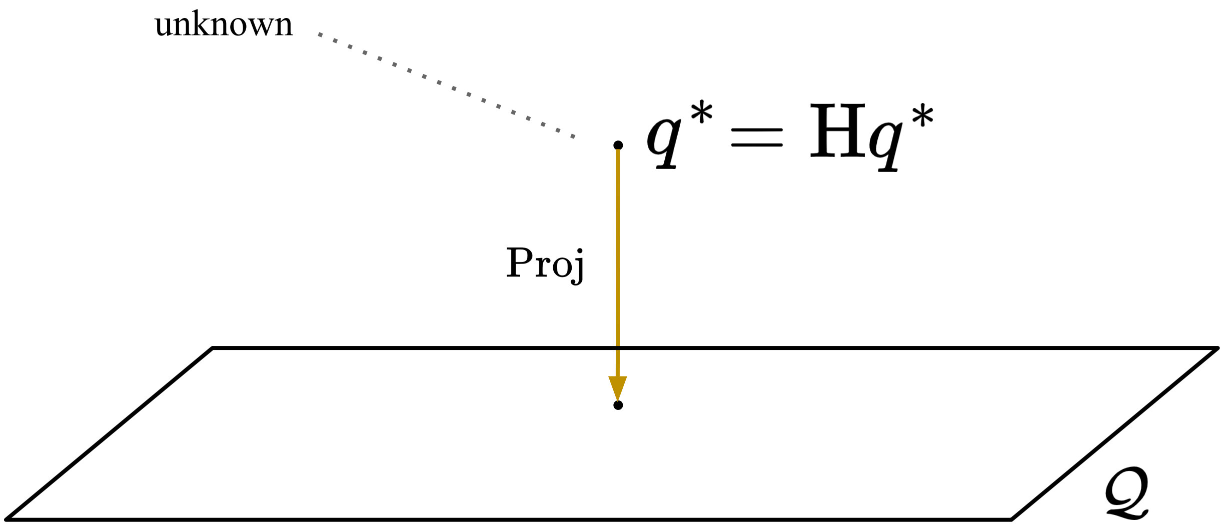
and so $\nabla_w \frac{1}{2}\|q - q_w\|_{2,\mu}^2 = \mathbb{E}_\mu[(q(\mathrm{x},\mathrm{a}) - \phi(\mathrm{x},\mathrm{a})^\top w)\phi(\mathrm{x},\mathrm{a})]$

We want to solve $\nabla_w \|q - q_w\|_{2,\mu}^2 = 0$

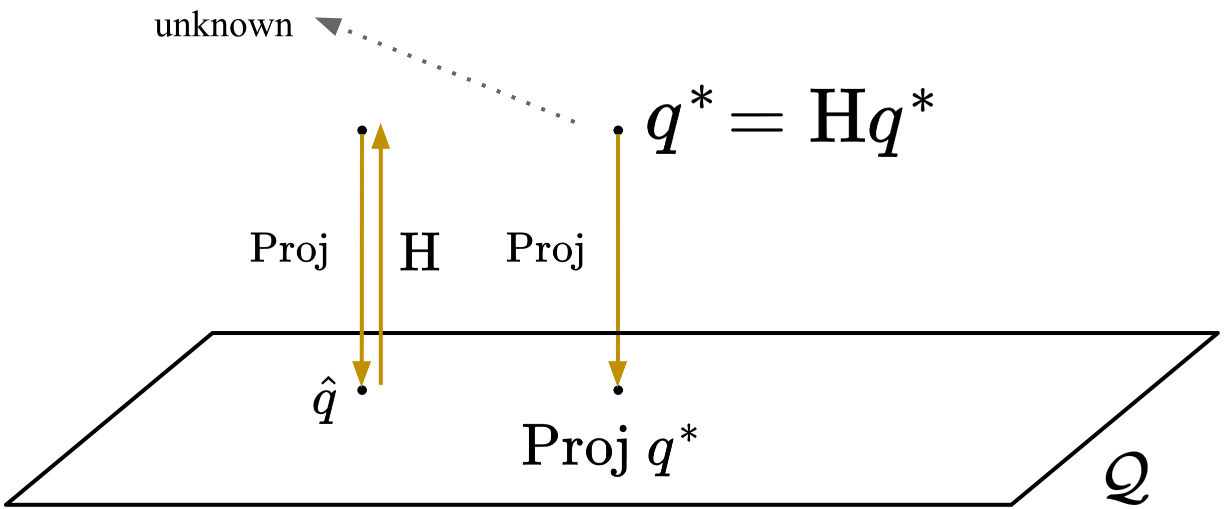
and we obtain $\mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})\phi^\top(\mathrm{x},\mathrm{a})]w = \mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})q(\mathrm{x},\mathrm{a})] \iff$

$$w = \mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})\phi^\top(\mathrm{x},\mathrm{a})]^{-1}\mathbb{E}_\mu[\phi(\mathrm{x},\mathrm{a})q(\mathrm{x},\mathrm{a})]$$

Linear function approximation



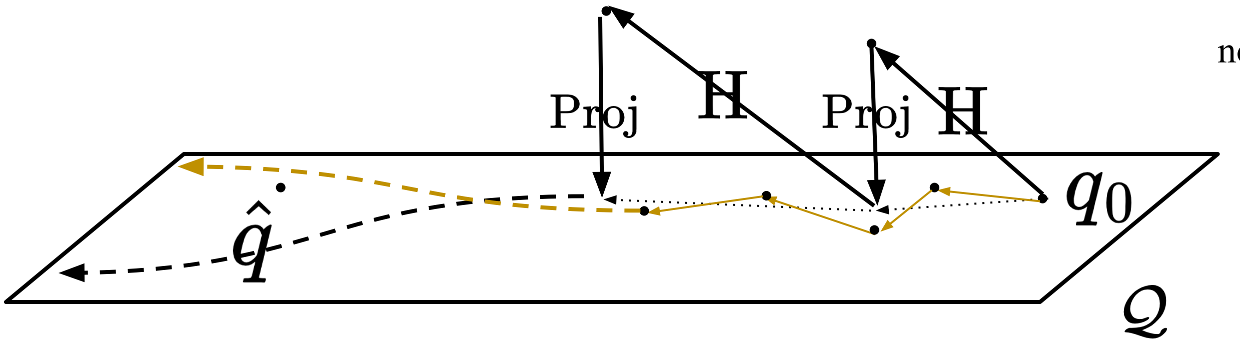
Linear function approximation



Linear function approximation

$$\hat{q} = (\mathrm{Proj}\ \mathrm{H})\hat{q}$$

not contractive



Q-learning with linear function approximation:

$$w_{t+1} = w_t + \alpha_t(r_t + \gamma \cdot \max_{a' \in \mathcal{A}} \phi^\top(x'_t, a')w - \phi(x_t, a_t)^\top w)\phi(x_t, a_t)$$

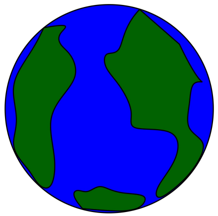
Linear function approximation

$\mathcal{X} = \{s_1, s_2\}$

$\mathcal{A} = \{a\}$

$r(x, a) = 0$

$\mu = \text{uniform}(\mathcal{X} \times \mathcal{A})$



$q^* = 0$

Proposition:

Proj H is expansive in any norm if $\gamma > \frac{5}{6}$

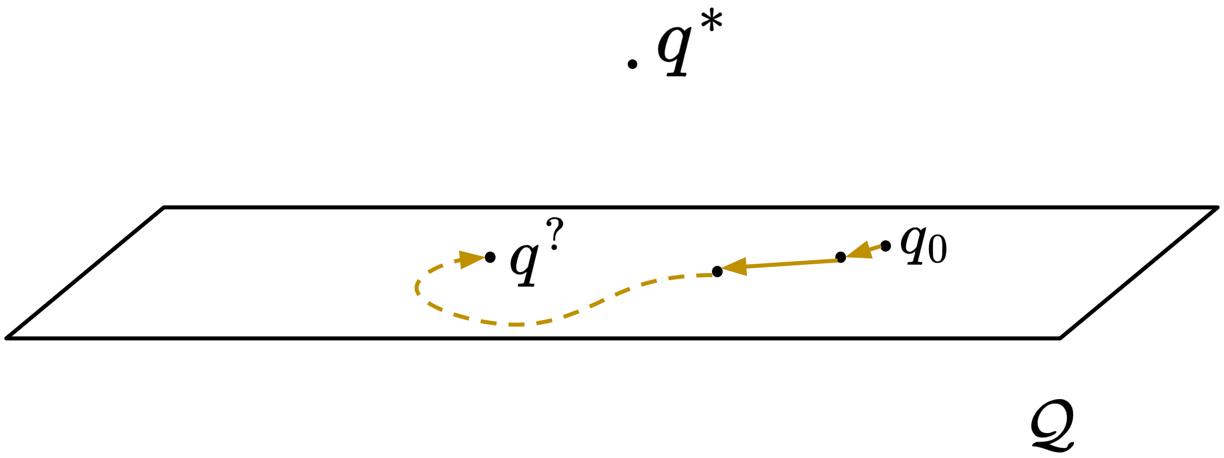
Proof sketch:

$$(\text{Proj H})q_w = \frac{6}{5}\gamma\, q_w$$

so $$\|(\text{Proj H})q_u - (\text{Proj H})q_v\| = \frac{6}{5}\gamma\|q_u - q_v\|$$

Research problem

How can we approximate the optimal value function?



Contributions

How can we approximate the optimal value function?

2020		2022		2023		2024		2025	
	Coupled Q-learning		Non-stationary environments		Non-stationary features		Multi-Bellman Q-learning		Full-gradient Q-learning
	NeurIPS + Under review		ECAI (outstanding paper)		Under review		TMLR		Open problem

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

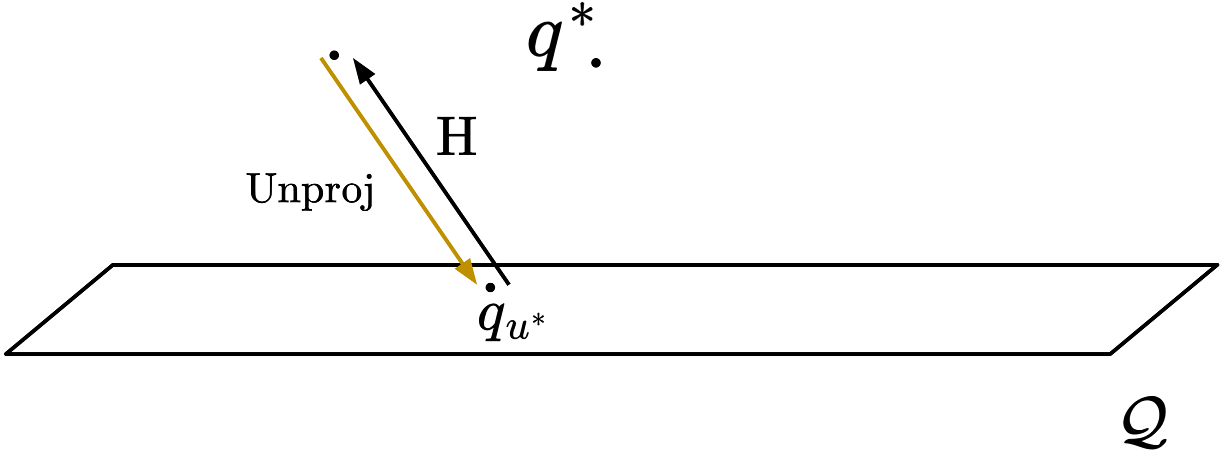
Diogo S. Carvalho, Pedro A. Santos, and Francisco S. Melo. Theoretical remarks on feudal hierarchies and reinforcement learning. European Conference on Artificial Intelligence, 26, 2023.

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Coupled Q-learning

$$q_{u^*} = (\text{Unproj H})q_{u^*}$$

contractive



Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Coupled Q-learning

$$q_{u^*} = (\text{Unproj H})q_{u^*}$$

$$(\text{Proj } q)(x,a) = \phi(x,a) \cdot \mathbb{E}[\phi(x,a)\phi^\top(x,a)]^{-1} \cdot \mathbb{E}[\phi(x,a)q(x,a)]$$

$$(\text{Unproj } q)(x,a) = \phi(x,a) \cdot \mathbb{E}[\phi(x,a)q(x,a)]$$

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Coupled Q-learning

$$q_{u^*} = (\text{Unproj H})q_{u^*}$$

Theorem:

If $\|\phi(x,a)\| \leq 1 \quad \forall x,a$, the operator Unproj is non-expansive

Proof:

$$\begin{aligned}
\|\text{Unproj } q - \text{Unproj p}\|_\infty &= \max_{x,a \in \mathcal{X}\times\mathcal{A}} |\phi(x,a)\mathbb{E}[\phi(\text{x},\text{a})q(\text{x},\text{a})] - \phi(x,a)\mathbb{E}[\phi(\text{x},\text{a})p(\text{x},\text{a})]| \\
&= \max_{x,a \in \mathcal{X}\times\mathcal{A}} |\phi(x,a)\mathbb{E}[\phi(\text{x},\text{a})(q(\text{x},\text{a}) - p(\text{x},\text{a}))]| \\
&\leq \max_{x,a \in \mathcal{X}\times\mathcal{A}} \|\phi(x,a)\| \cdot \|\mathbb{E}[\phi(\text{x},\text{a})(q(\text{x},\text{a}) - p(\text{x},\text{a}))]\| \\
&\leq \mathbb{E}[\|\phi(\text{x},\text{a})(q(\text{x},\text{a}) - p(\text{x},\text{a}))\|] \\
&\leq \mathbb{E}[\|\phi(\text{x},\text{a})\| \cdot |(q(\text{x},\text{a}) - p(\text{x},\text{a}))|] \\
&\leq \|q - p\|_\infty
\end{aligned}$$

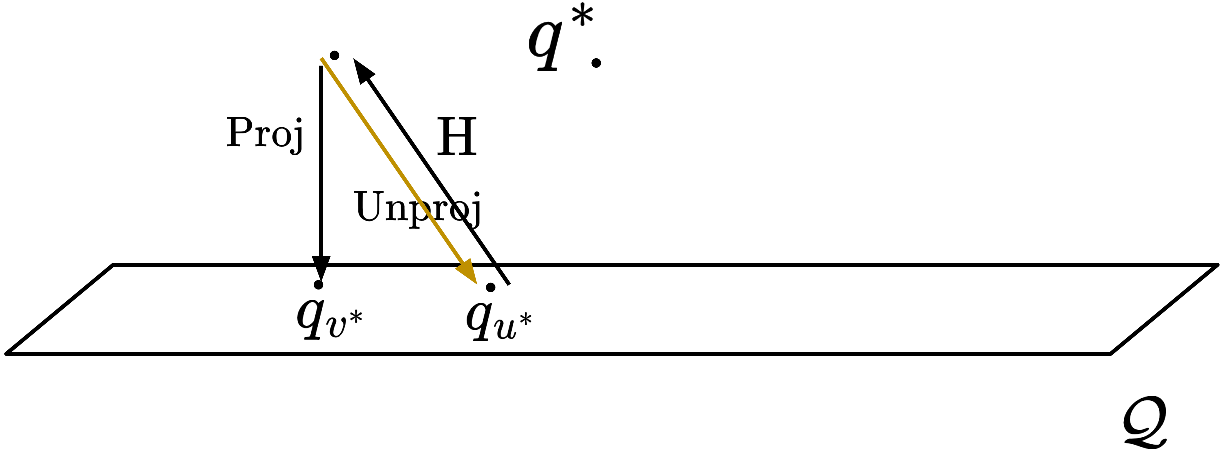
Corollary:

The combined operator Unproj H has a unique fixed-point q_{u^*}

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Coupled Q-learning

$$q_{v^*} = (\text{Proj H})q_{u^*}$$
$$q_{u^*} = (\text{Unproj H})q_{u^*}$$



Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Two-time-scale stochastic approximation

$$v_{t+1} = v_t + \alpha_t(f^{\text{fast}}(v_t, u_t) + \text{noise}_t^{\text{fast}})$$

$$u_{t+1} = u_t + \beta_t(f^{\text{slow}}(v_t, u_t) + \text{noise}_t^{\text{slow}})$$

Additional assumption: $\beta_t = o(\alpha_t)$

Theorem:

If the dynamical system governed by the o.d.e. $\dot{v} = f^{\text{fast}}(v, u)$
has a unique globally asymptotically stable equilibrium $\lambda(u)$
for all u and the dynamical system governed by the o.d.e. $\dot{u} = f^{\text{slow}}(\lambda(u), u)$
has a unique globally asymptotically stable equilibrium u^*
then $u \to u^*$ and $v \to v^* = \lambda(u^*)$ almost surely

Coupled Q-learning

Coupled Q-learning with linear function approximation:

$$v_{t+1} = v_t + \alpha_t(r_t + \gamma \cdot \max_{a' \in \mathcal{A}} \phi^\top(x'_t, a')u - \phi(x_t, a_t)^\top v)\phi(x_t, a_t)$$
$$u_{t+1} = u_t + \beta_t(\phi(x, a)\phi^\top(x, a)v - u)$$

Theorem:

We have that $v_t \to v^*$ and $u_t \to u^*$

such that $q_{v^*} = (\text{Proj H})q_{u^*}$ and $q_{u^*} = (\text{Unproj H})q_{u^*}$

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

Coupled Q-learning

Additional Assumption:

The features are orthogonal and uniformly excited, i.e, $\mathbb{E}[\phi(x,a)\phi^\top(x,a)] = \sigma\mathbb{I}$

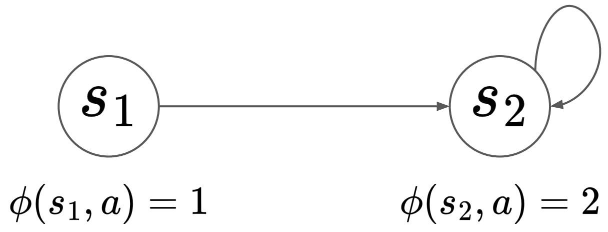
Theorem:

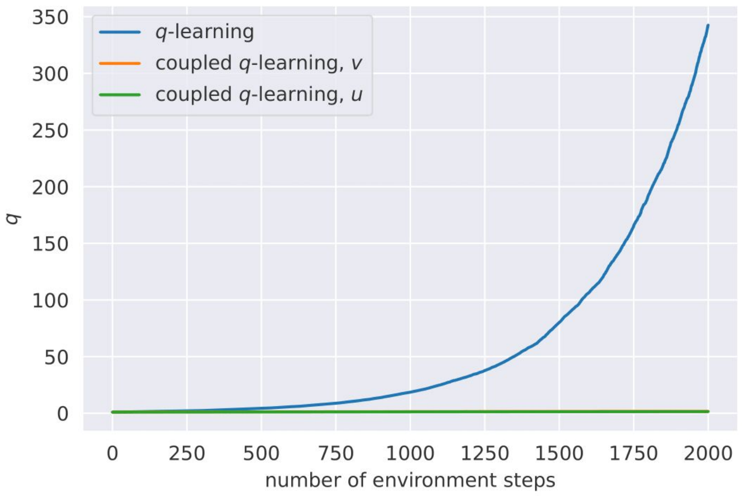
We have that $\|q^* - q_{v^*}\|_\infty \leq \frac{1}{1-\sigma^{-1}\gamma}\|q^* - \text{Proj } q^*\|_\infty + \gamma\frac{1-\sigma}{\sigma}\frac{r_{\max}}{(1-\gamma)^2}$

and $\|q^* - q_{u^*}\|_\infty \leq \frac{1}{1-\sigma^{-1}\gamma}\|q^* - \text{Proj } q^*\|_\infty + \frac{1-\sigma}{\sigma}\frac{r_{\max}}{(1-\gamma)^2}$

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

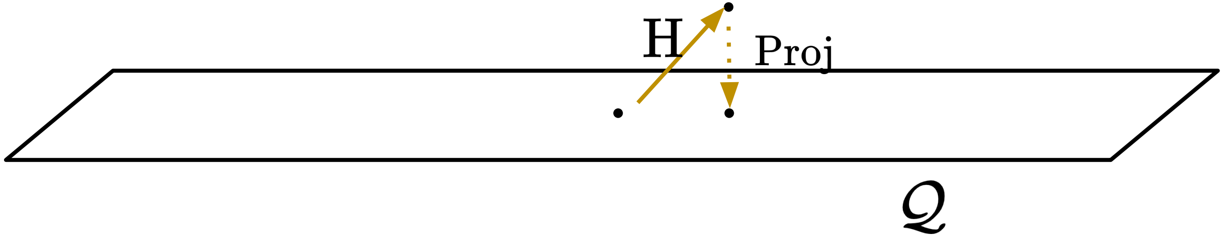
Coupled Q-learning





Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. Advances in Neural Information Processing Systems, 33, 2020.

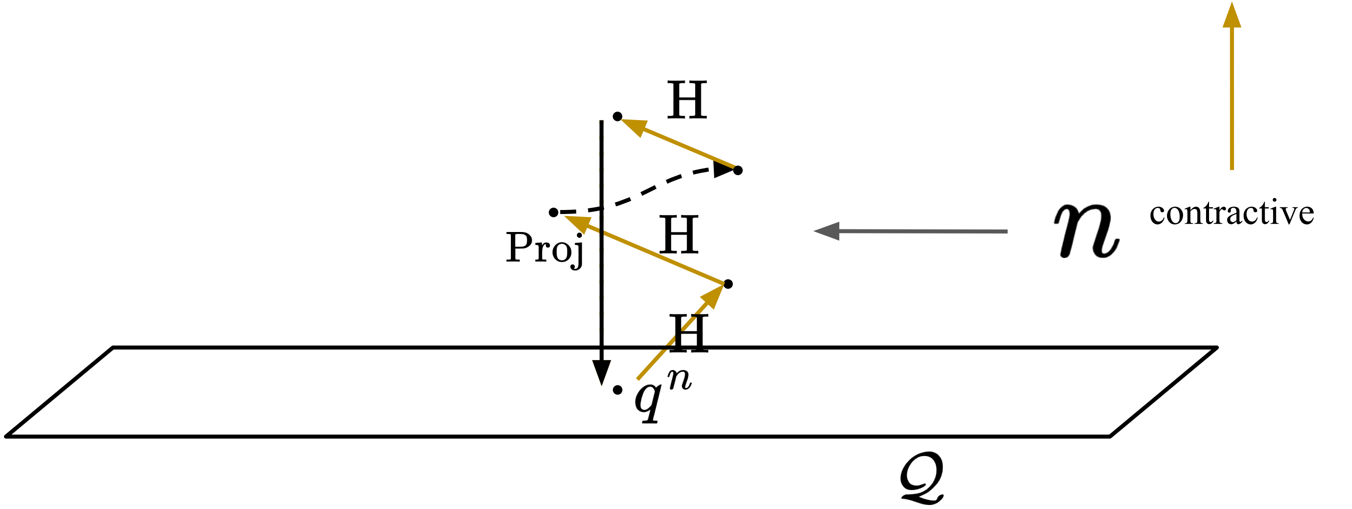
Multi-Bellman Q-learning



Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathrm{Proj}\,\mathrm{H}^n)\tilde{q}_n$$



$$\left(\mathrm{H}^{n+1} q\right)(x, a) = \mathbb{E}\left[r(x, a) + \gamma \cdot \max_{a'} \mathrm{H}^n(\mathrm{x}', a')\right]$$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathrm{Proj}\, \mathrm{H}^n)\tilde{q}_n$$

Proposition:

The operator H^n is contractive with contraction factor γ^n

Proof:

$$\begin{aligned}
\|\mathrm{H}^{n+1}q - \mathrm{H}^{n+1}p\|_\infty &= \|\mathrm{H}^n(\mathrm{H}q) - \mathrm{H}^n(\mathrm{H}p)\|_\infty \\
&\leq \gamma^n \|\mathrm{H}q - \mathrm{H}p\|_\infty \\
&\leq \gamma^{n+1} \|q - p\|_\infty
\end{aligned}$$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning $\tilde{q}_n = (\mathbf{Proj}\,\mathbf{H}^n)\tilde{q}_n$

Assumption: $\|\phi(x,a)\| \leq 1 \quad \forall x, a$ and $\|\mathbb{E}[\phi(x,a)\phi^\top(x,a)]^{-1}\|_2 = \sigma^{-1}$

Proposition:

The operator $\mathbf{Proj}$ is Lipschitz with factor σ^{-1}

Proof:

$$
\begin{aligned}
\|\mathrm{Proj}\,q - \mathrm{Proj}\,p\|_\infty &= \max_{x,a\in\mathcal{X}\times\mathcal{A}} |\phi(x,a)\mathbb{E}[\phi(\mathrm{x},\mathrm{a})\phi^\top(\mathrm{x},\mathrm{a})]^{-1}\mathbb{E}[\phi(\mathrm{x},\mathrm{a})(q(x,a) - p(x,a))]| \\
&\leq \max_{x,a\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|_2 \cdot \|\mathbb{E}[\phi(\mathrm{x},\mathrm{a})\phi^\top(\mathrm{x},\mathrm{a})]^{-1}\|_2 \cdot \mathbb{E}[\|\phi(\mathrm{x},\mathrm{a})\|_2 |q(x,a) - p(x,a)|] \\
&\leq \max_{x,a\in\mathcal{X}\times\mathcal{A}} \|\phi(x,a)\|_2 \cdot \|\mathbb{E}[\phi(\mathrm{x},\mathrm{a})\phi^\top(\mathrm{x},\mathrm{a})]^{-1}\|_2 \cdot \mathbb{E}[\|\phi(\mathrm{x},\mathrm{a})\|_2]\|q - p\|_\infty \\
&\leq \sigma^{-1}\|q - p\|_\infty
\end{aligned}
$$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathrm{Proj}\,\mathrm{H}^n)\tilde{q}_n$$

Theorem:

The operator $\mathrm{Proj}\,\mathrm{H}^n$ is contractive for $n > \log_\gamma \sigma$

Proof:

$$\begin{aligned}\|\mathrm{Proj}(\mathrm{H}^n q) - \mathrm{Proj}(\mathrm{H}^n p)\|_\infty &\leq \sigma^{-1}\|\mathrm{H}^n q - \mathrm{H}^n p\|_\infty \\ &\leq \sigma^{-1}\gamma^n\|q-p\|_\infty\end{aligned}$$

Corollary:

There exists a unique fixed point $\tilde{q}_n = (\mathrm{Proj}\,\mathrm{H}^n)\tilde{q}_n$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathrm{Proj}\,\mathrm{H}^n)\tilde{q}_n$$

Proposition:

We have that $\|q^* - \tilde{q}_n\|_\infty \leq \frac{1}{1-\sigma^{-1}\gamma^n}\|q^* - \mathrm{Proj}\;q^*\|_\infty$

Proof:

$$\|q^* - \tilde{q}_n\|_\infty \leq \|q^* - \mathrm{Proj}\;q^*\|_\infty + \|\mathrm{Proj}\;q^* - \tilde{q}_n\|_\infty$$

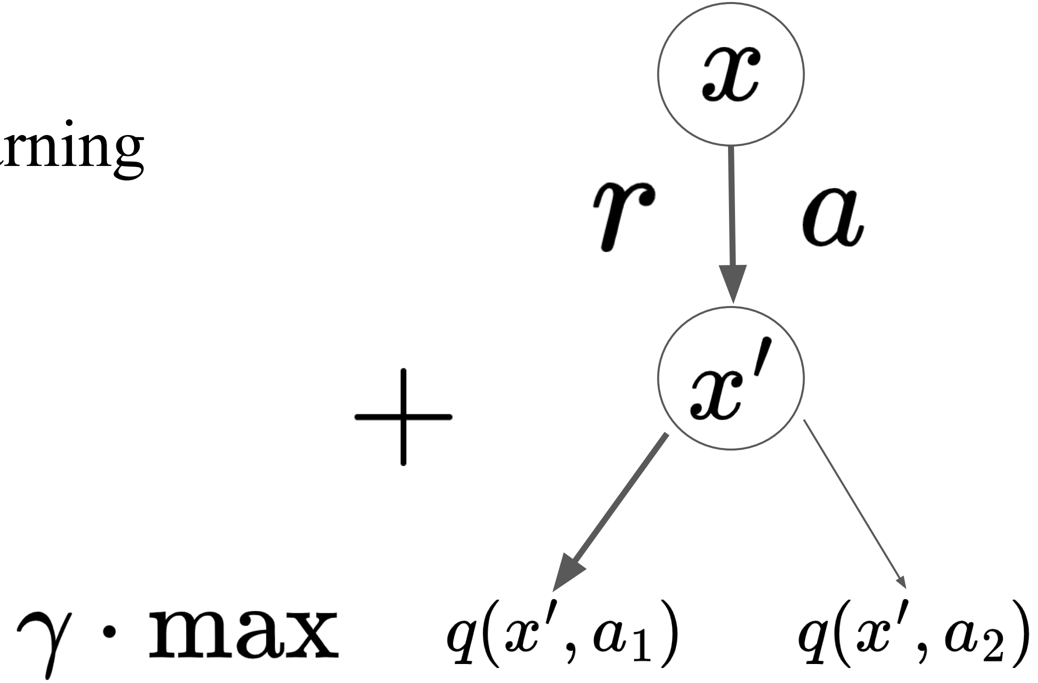
and
$$\begin{aligned}\|\mathrm{Proj}\;q^* - \tilde{q}_n\|_\infty &= \|\mathrm{Proj}\;\mathrm{H}^n q^* - \mathrm{Proj}\;\mathrm{H}^n\tilde{q}_n\|_\infty \\ &\leq \sigma^{-1}\gamma^n\|q^* - \tilde{q}_n\|_\infty\end{aligned}$$

Corollary:

$$\lim_{n\to\infty}\tilde{q}_n = \mathrm{Proj}\;q^*$$

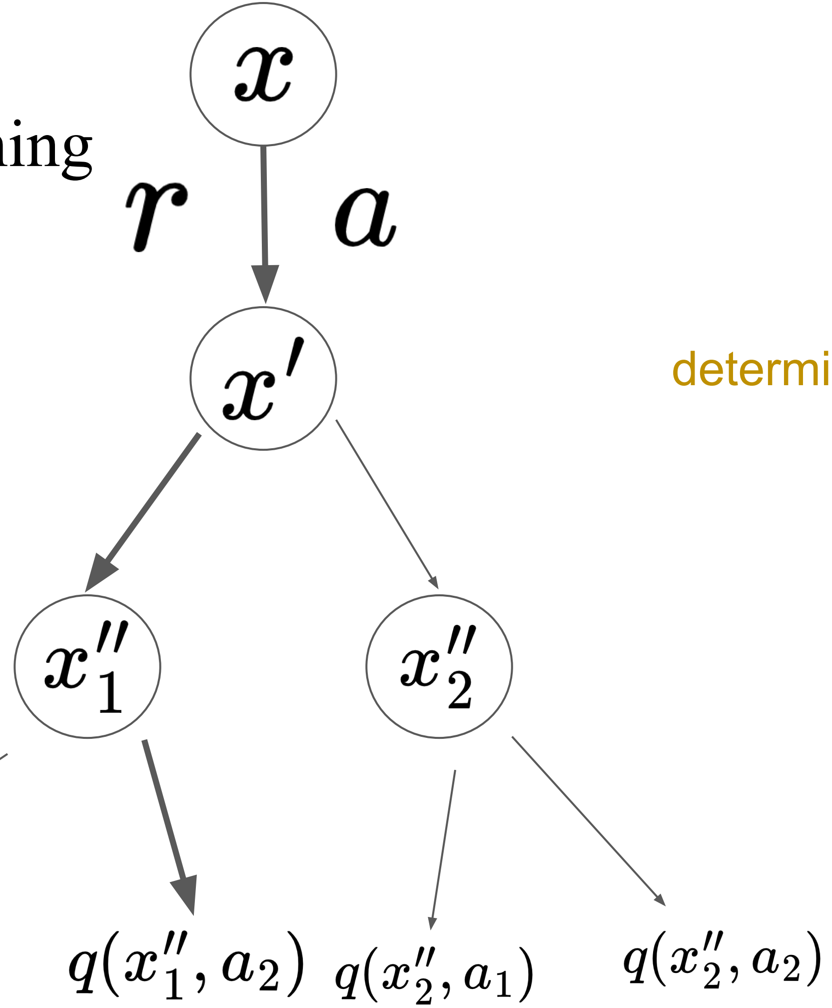
Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Q-learning



Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning



deterministic case

$$+$$

$$\gamma \cdot \max(r' + \dots)$$

$$+$$

$$\gamma \cdot \max$$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathrm{Proj}\,\mathrm{H}^n)\tilde{q}_n$$

Theorem:

In deterministic environments we have that $q_{w_t} \to \tilde{q}_n$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

$$\tilde{q}_n = (\mathbf{Proj}\ \mathbf{H}^n)\tilde{q}_n$$

Non-deterministic case

$$\mathbf{max}_i\ \mathbb{E}\left[\mathbf{Z}_i\right] \neq \mathbb{E}\left[\mathbf{max}_i\ \mathbf{Z}_i\right]$$

maximization of
expected values

$$\left(\mathrm{H}^{n+1}q\right)(x,a) = \mathbb{E}\left[r(x,a) + \gamma \cdot \mathrm{max}_{a'}\mathrm{H}^n(\mathrm{x}',a')\right]$$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

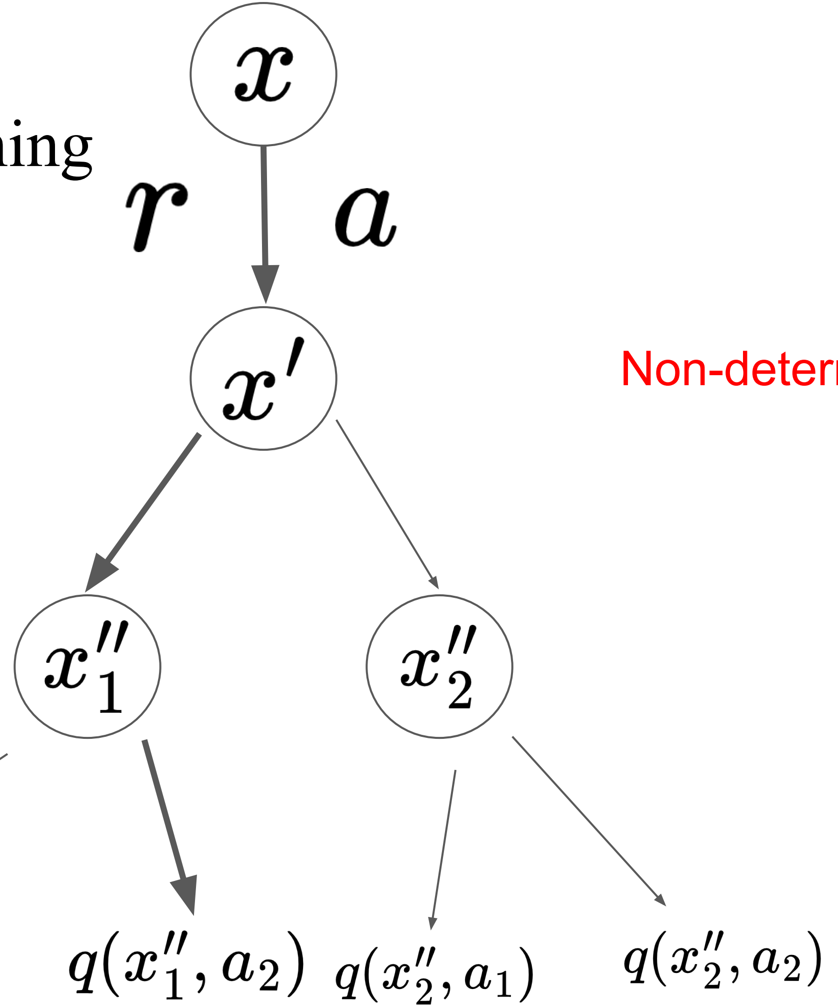
Non-deterministic case

$+$

$\gamma \cdot \max(r' + \ldots)$

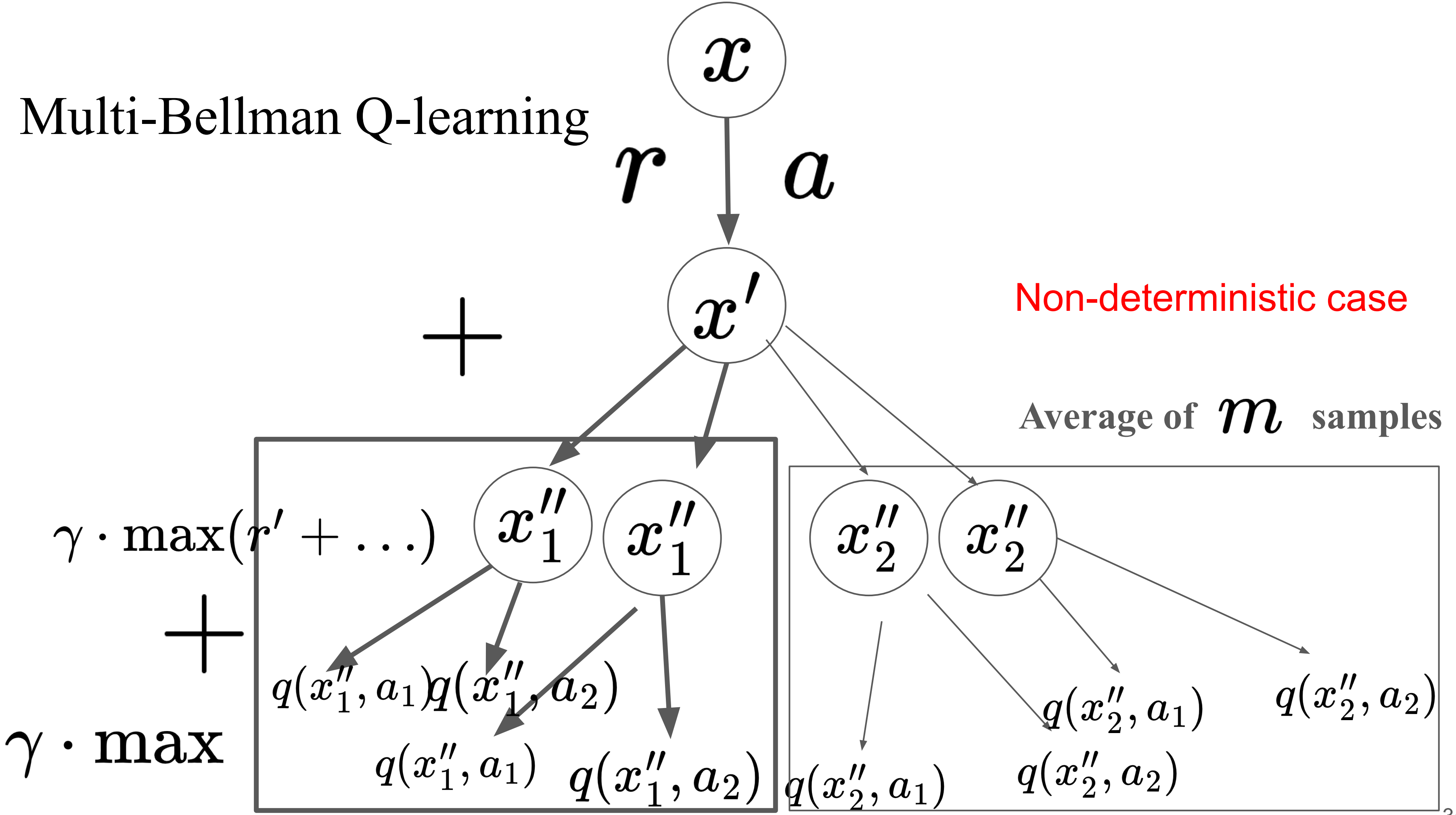
$+$

$\gamma \cdot \max$



Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning



Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

Proposition:

The time complexity of the algorithm is $\mathcal{O}((m \cdot |\mathcal{A}|)^n)$

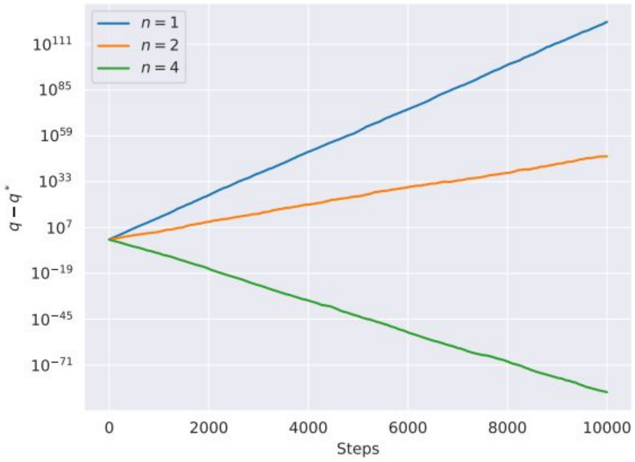
Proposition:

In non-deterministic environments, we have that $q_{w_t} \to \hat{q}_{n,m}$

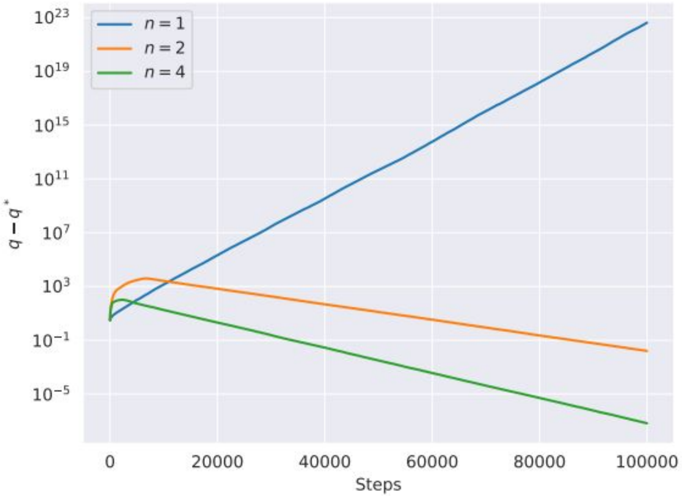
We have, however, that $\lim_{m\to\infty} \hat{q}_{n,m} = \tilde{q}_n$

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning



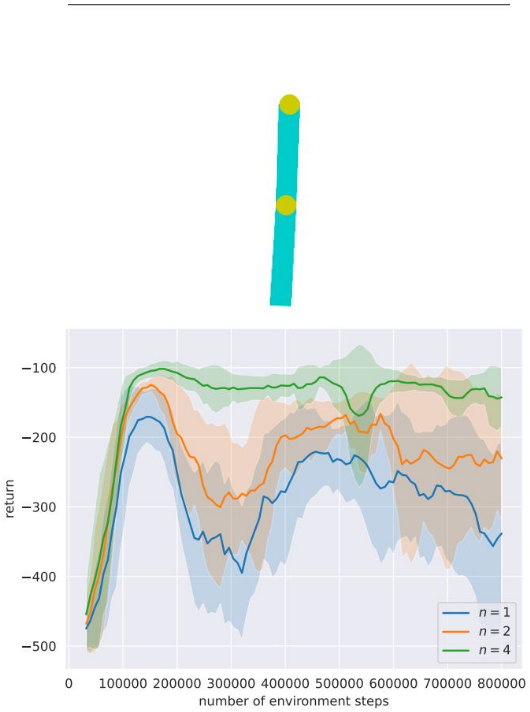
(a) $\omega \to 2\omega$.



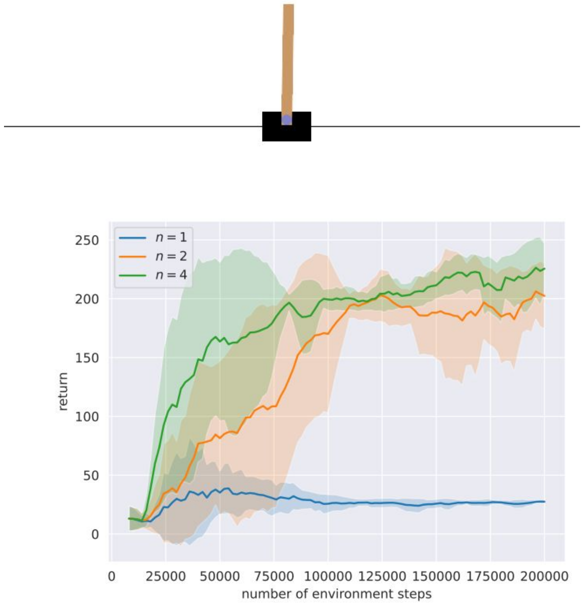
(b) Star.

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

Multi-Bellman Q-learning

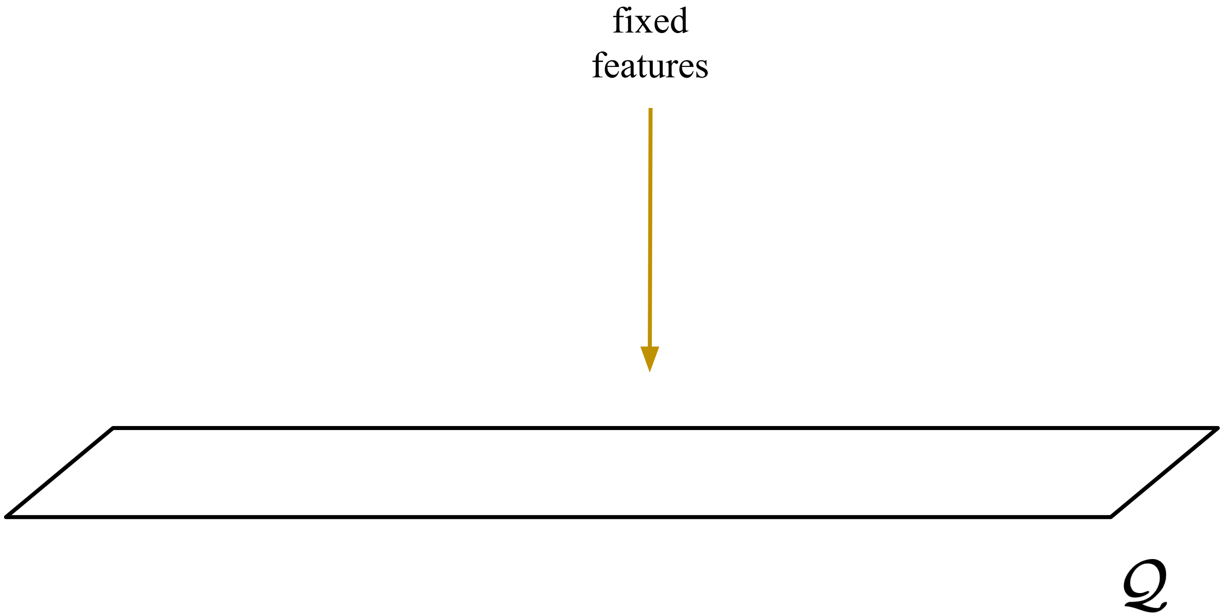


(a) Acrobot.

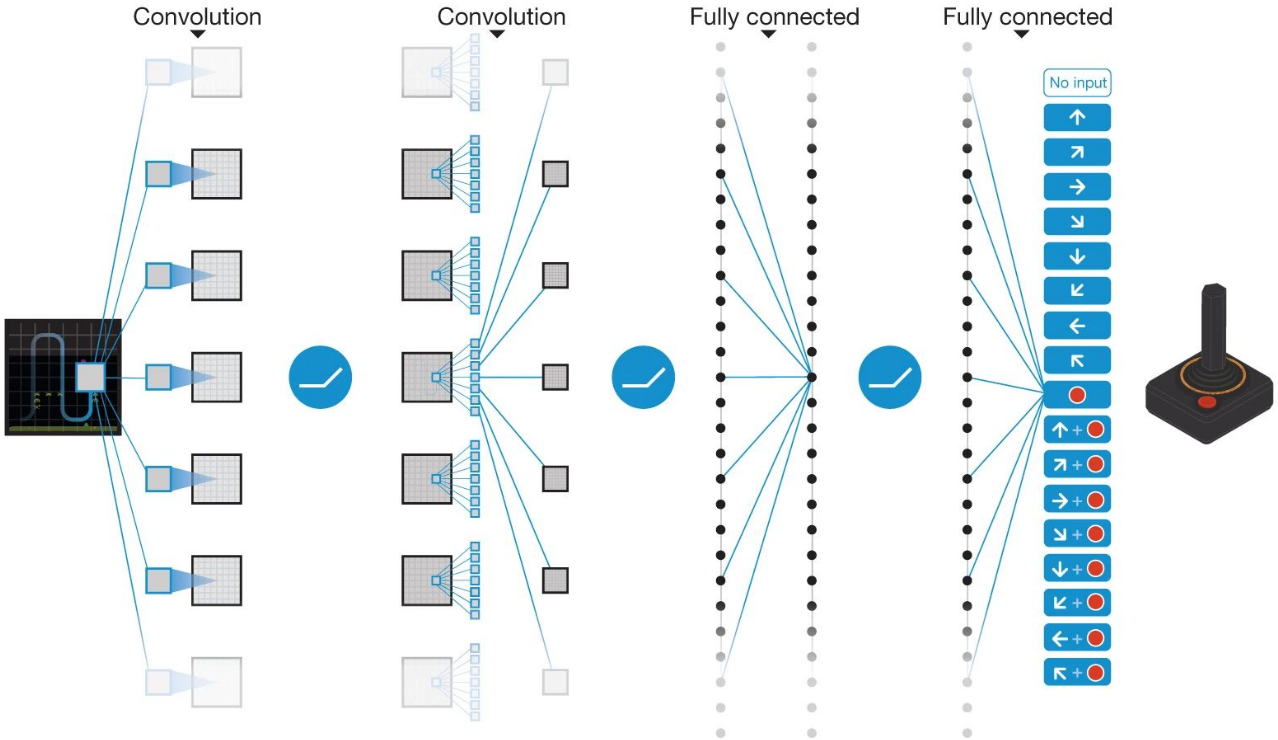


(b) Cartpole.

Diogo S. Carvalho, Pedro A. Santos,, and Francisco S. Melo. Multi-Bellman operator for convergence of Q-learning with linear function approximation, Transactions on Machine Learning Research, 2025

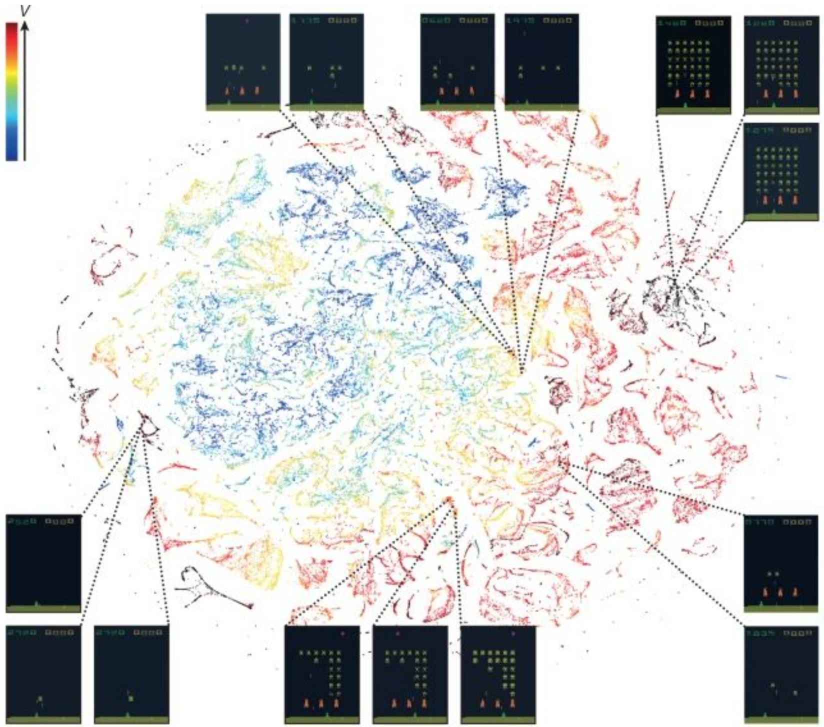

fixed
features
$\mathcal{Q}$

Deep Q-Learning in Practice



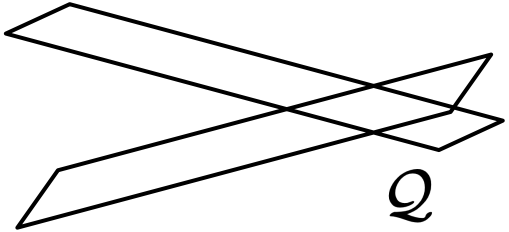
Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). https://doi.org/10.1038/nature14236

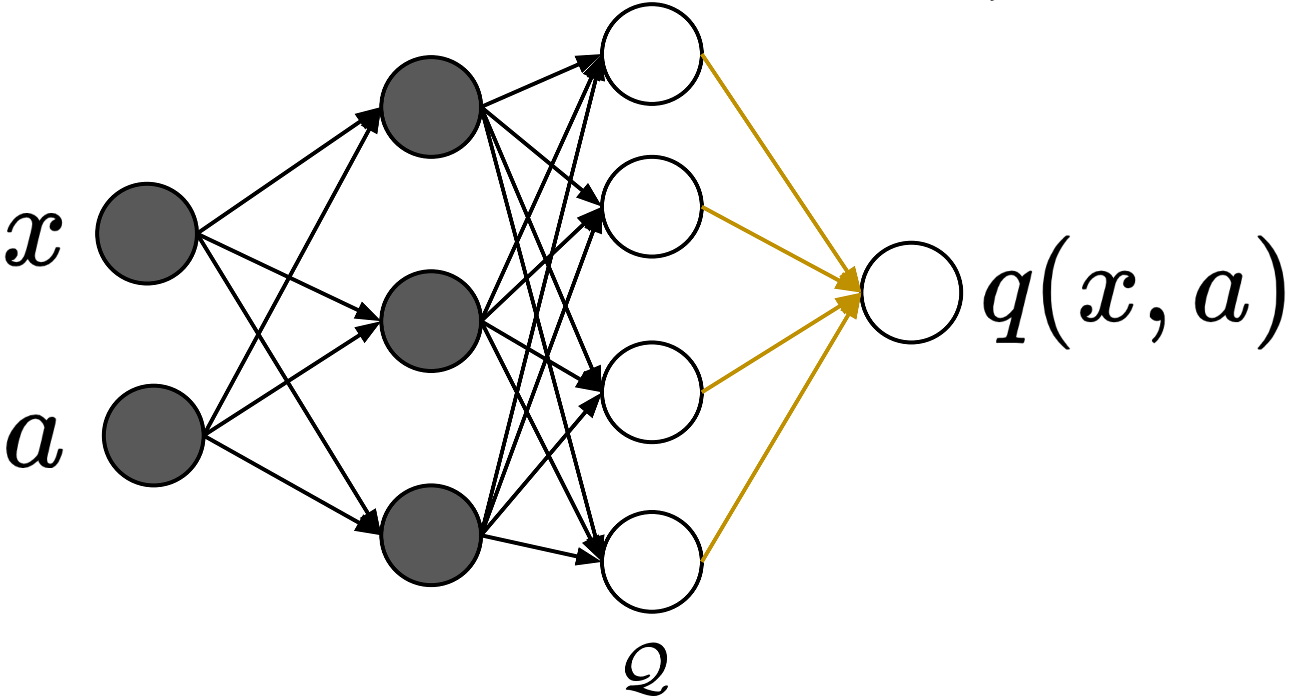
Deep Q-Learning in Practice



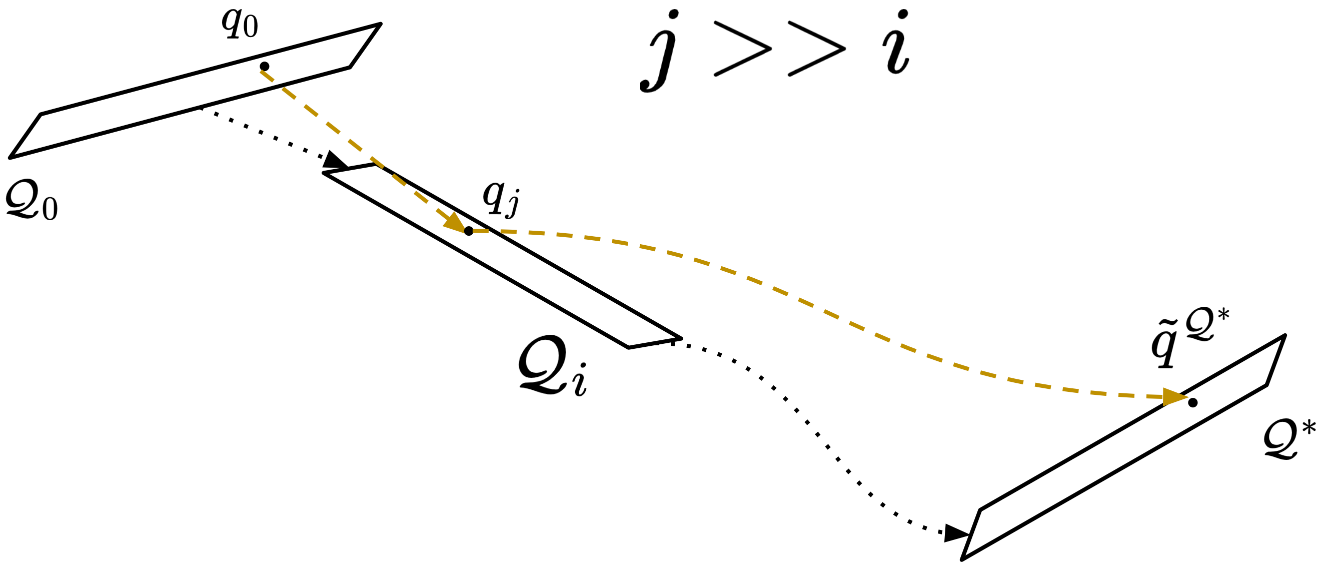
Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). https://doi.org/10.1038/nature14236

Deep reinforcement learning





Non-stationary features



Non-stationary features

We have parameterized features $\phi_u : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^{k_1}, \quad u \in \mathbb{R}^{k_2}$
for instance the inner layers of a neural network $q_{u,v} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}, \quad v \in \mathbb{R}^{k_1}$
such that $q_{u,v}(x,a) = \phi_u^\top(x,a)v$

Assumptions:

ϕ_u is Lipschitz with respect to the parameters u

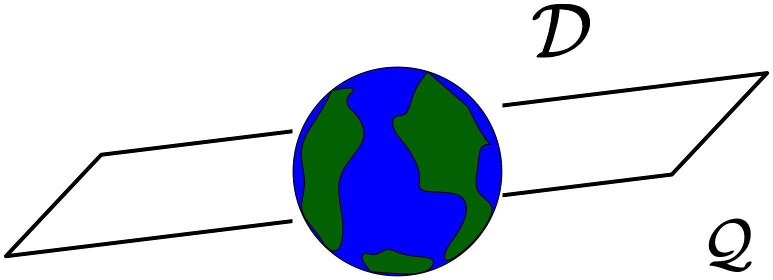
u follows a **convergent** stochastic approximation update along a slower time-scale $\beta_t = o(\alpha_t)$

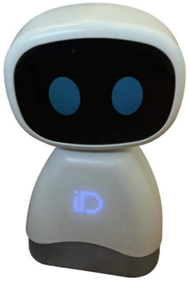
Theorem:

The updates

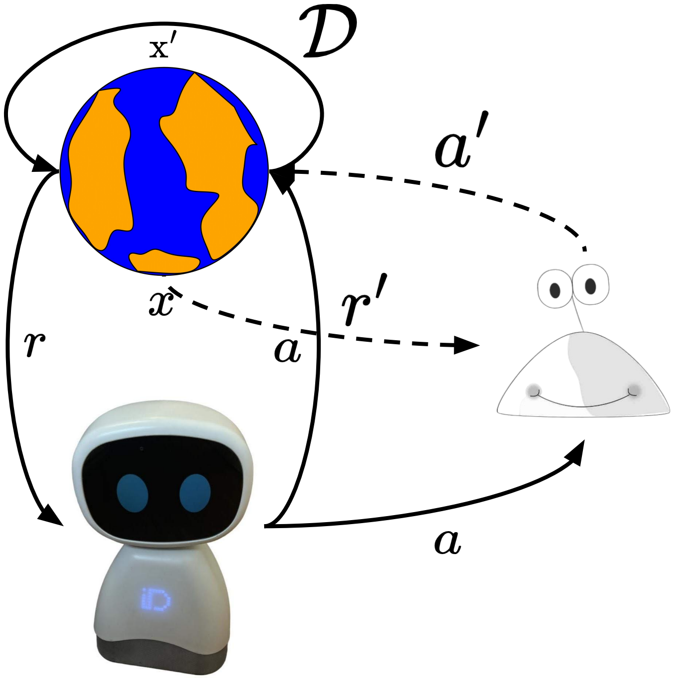
$$v_{t+1} = v_t + \alpha_t((r_t + \gamma \cdot \max_{a' \in \mathcal{A}} q_{u_t,v_t}(x'_t, a') - \xi \cdot q_{u_t,v_t}(x_t, a_t))\phi(x_t, a_t)) - \alpha_t \eta v_t$$

where $\eta \geq 0$ and $\xi \geq 1$ are regularizers
converge for sufficiently large η and ξ

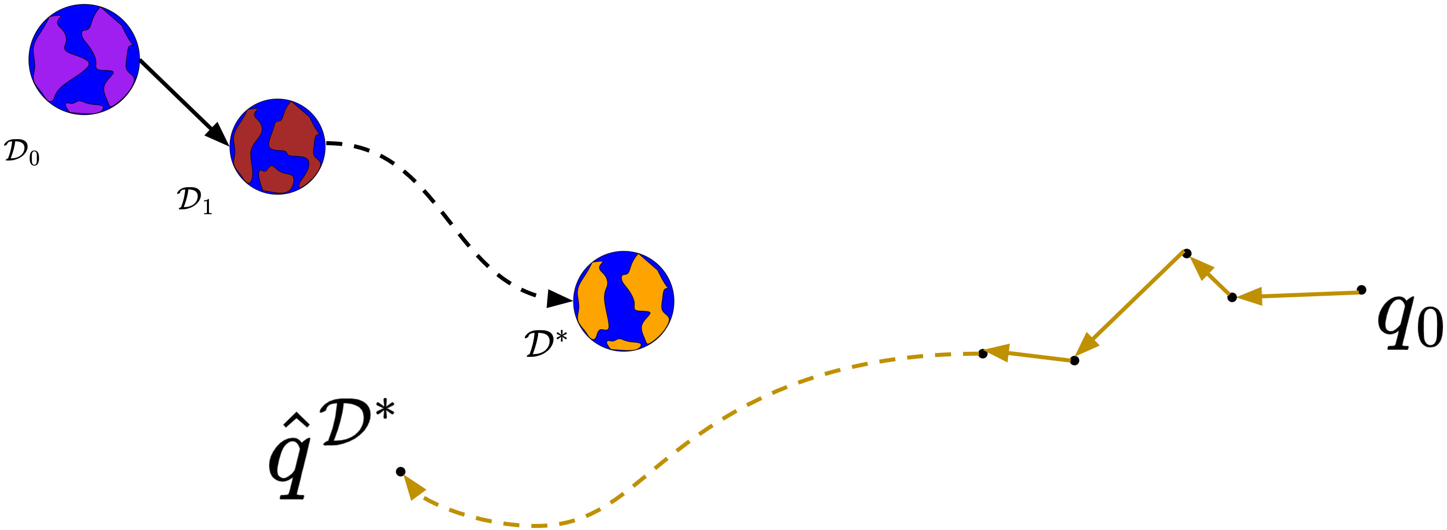

$\mathcal{D}$
$\mathcal{Q}$



Hierarchical reinforcement learning



Non-stationary environments



Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. Theoretical remarks on feudal hierarchies and reinforcement learning. European Conference on Artificial Intelligence, 26, 2023.

Non-stationary environments

The low level is an MDP $\mathcal{M}^{\text{low}} = (\mathcal{X}^{\text{low}} \times \mathcal{A}^{high}, \mathcal{A}^{\text{low}}, P^{\text{low}}, r^{\text{low}}, \gamma^{\text{low}})$

$$\mathcal{M}^{\text{high}} = (\mathcal{X}^{\text{high}}, \mathcal{A}^{\text{high}}, P^{\text{high}}(\pi^{\text{low}}), r^{\text{high}}(\pi^{\text{low}}), \gamma^{\text{high}})$$

non-stationarity arises

Assumption:

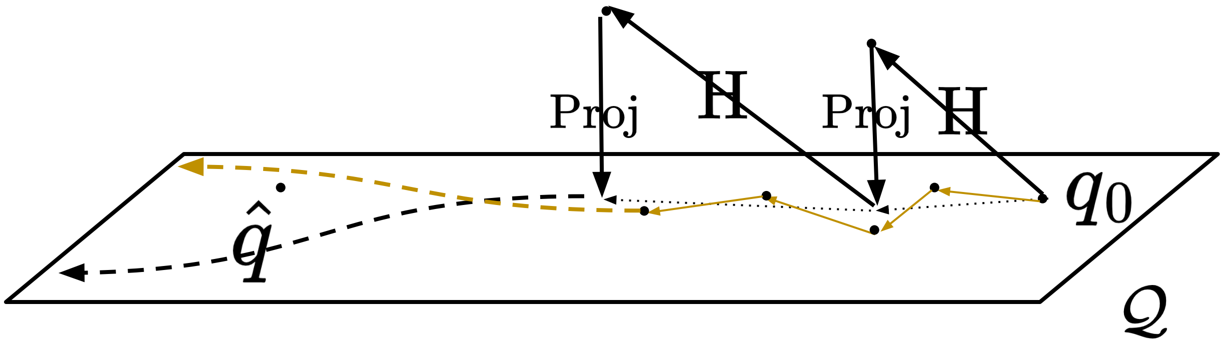
Low-level episodes duration follows a geometric distribution

Theorem:

Despite the non-stationarity on the high level, $\mathcal{M}^{\text{high}}$ converges
Furthermore, Q-learning converges on convergent MDPs
Therefore, Q-learning converges in hierarchical MDPs.

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. Theoretical remarks on feudal hierarchies and reinforcement learning. European Conference on Artificial Intelligence, 26, 2023.

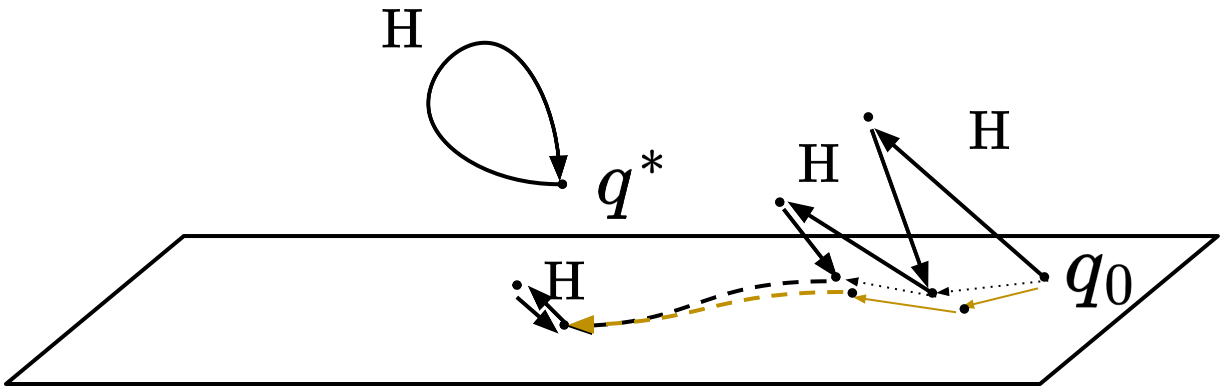
Full-gradient Q-learning



Remark:

Q-learning takes steps in the direction of the "semi-gradient" of the loss $l(w) = \|\mathrm{H}q_w - q_w\|^2_{2,\mu}$
i.e., fixes the target $\mathrm{H}q_w$ and only takes the derivative of the current estimate q_w

Full-gradient Q-learning



Open problem:

Perform "full-gradient" descent on the loss function, $l(w) = \|\mathrm{H}q_w - q_w\|_{2,\mu}^2$
Convergence could be guaranteed, but local minima could be a problem

Full-gradient Q-learning

$$\mathbb{E}\left[f\left(\mathrm{Z}\right)\right] \cdot \mathbb{E}\left[g\left(\mathrm{Z}\right)\right] \neq \mathbb{E}\left[f\left(\mathrm{Z}\right) \cdot g\left(\mathrm{Z}\right)\right]$$

product of
expected values

double-sampling

$$(\nabla l)(q) = \mathbb{E}\left[\left((\mathrm{H}q)(\mathrm{x},\mathrm{a}) - q(\mathrm{x},\mathrm{a})\right) \cdot \left(\left(\nabla\left(\mathrm{H}q\right)\right)(\mathrm{x},\mathrm{a}) - (\nabla q)(\mathrm{x},\mathrm{a})\right)\right]$$