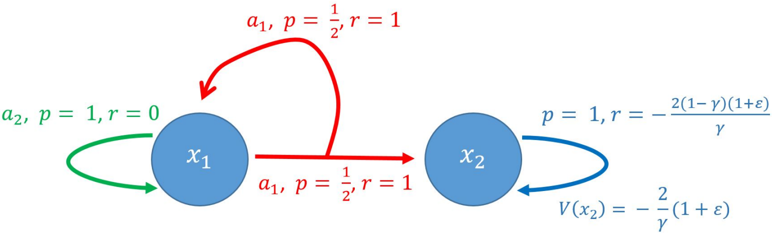
Increasing the Action Gap

Mark Gluzman

iDDA, CUHK (Shenzhen)

March 18, 2019

Motivation Example



- Let $\pi \in \Pi$ be a stationary deterministic policy, $Q^\pi(x,a) = r(x,a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)} V^\pi(x')$.
- $Q^\pi(x_1, a_1) = 1 + \gamma \left[\frac{1}{2} V^\pi(x_1) + \frac{1}{2} V^\pi(x_2) \right] = 1 + \frac{\gamma}{2} V^\pi(x_1) - (1+\varepsilon) = \frac{\gamma}{2} V^\pi(x_1) - \varepsilon$
- $Q^\pi(x_1, a_2) = 0 + \gamma V^\pi(x_1)$
- Note that for any $\pi \in \Pi$, we have $Q^\pi(x_1, a_2) > Q^\pi(x_1, a_1)$, therefore $V^*(x_1) = 0$. The value difference between optimal and second best action, *action gap*, is

$$Q^*(x_1, a_2) - Q^*(x_1, a_1) = \varepsilon$$

Why the size of action gap is important?

In Q-learning methods, we usually use a greedy policy w.r.t. Q-function:

$$\pi(x) = \arg\max_{a \in A} Q(x, a).$$

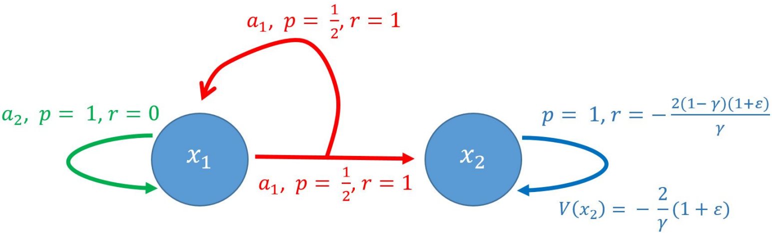
- When the MDP can be solved exactly, there is no issue.
- When we cannot solve the MDP exactly, we use some approximation methods, for example:

  sample-based estimations $Q(x, a) \approx r(x, a) + \gamma \frac{1}{N} \sum_{n=1}^{N} \max_{b} Q(x'_n, b')$ or

  low-dimensional representation $Q(x, a) \approx \sum_{k=1}^{K} \omega_k \psi_k(x, a)$.

Small perturbations in the Q-function may result in identifying a wrong action to be the optimal!

Nonstationarity



- Let Π be a set of all stationary deterministic policies.
- Note $\Pi = \{\pi_1, \pi_2 : \ \pi_1(x_1) = a_1;\ \pi_2(x_1) = a_2\}$.
- From the Bellman eq., $V^{\pi_1}(x_1) = 1 + \gamma\left[\frac{1}{2}V^{\pi_1}(x_1) + \frac{1}{2}V^{\pi_1}(x_2)\right] = \frac{\gamma}{2}V^{\pi_1}(x_1) - \varepsilon$

$$V^{\pi_1}(x_1) = -\frac{\varepsilon}{1-\gamma/2}, \quad V^{\pi_2}(x_1) = 0$$

Why $Q^*(x_1, a_2) - Q^*(x_1, a_1) = \varepsilon$?

$$Q^*(x_1, a_1) = r(x_1, a_1) + \gamma \mathbb{E}_{x' \sim P(\cdot|x, a_1)} V^{\pi_2}(x') = -\varepsilon$$

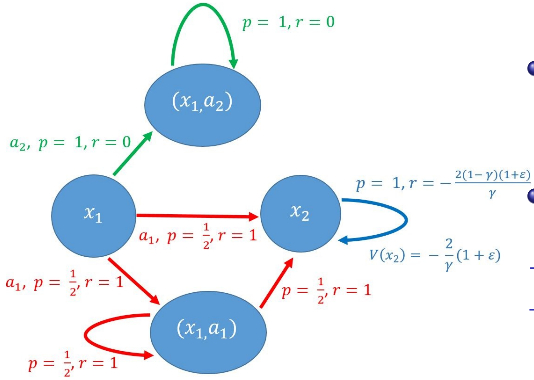
does not describe the value of any stationary policy!

Standard approach

- In the algorithms based on value iterations, we update $Q-$factors according to the Bellman equation.
  The Bellman operator:
  $$TQ(x,a) := r(x,a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)} \left[\max_{b \in A} Q(x', b) \right].$$
- Iterations
  $$Q_{k+1} = TQ_k$$
  converge to the optimal $Q^*(x,a)$ from which one can obtain an optimal policy
  $$\pi^*(x) = \arg\max_{a \in A} Q^*(x,a).$$
- Can we modify the algorithm so that its iterations will converge to $\tilde{Q}^*(x,a)$ s.t.
  - $\pi^*(x) = \arg\max_{a \in A} \tilde{Q}^*(x,a)$.
  - $\tilde{Q}^*(x, \pi^*(x)) - \tilde{Q}^*(x,a) \geq Q^*(x, \pi^*(x)) - Q^*(x,a)$?

Extended state space

Idea: let do not change the action if we return to the state after one time-step.



- If $P(x|x,a) = p > 0$ add a state (x,a) to a state space, $P_{ext}(x|x,a) := 0$, $P_{ext}((x,a)|x,a) := p$.
- $P_{ext}(y|(x,a)) := P(y|x,a)$, $P_{ext}((x,a)|(x,a)) := P(x|x,a)$.

\+ $Q^*(x_1, a_2) - Q^*(x_1, a_1) = \frac{\varepsilon}{1-\gamma/2}$

– We doubled the state space

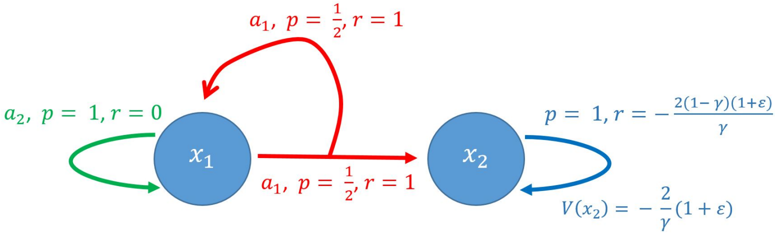
Consistent Bellman operator

The consistent Bellman operator (CB operator):

$$T_{CB}Q(x,a) := r(x,a) + \gamma\mathbb{E}\left[\mathbb{I}_{x\neq x'} \max_{b\in A} Q(x',b) + \mathbb{I}_{x=x'}Q(x,a)\right]. \tag{1}$$

- The consistent Bellman operator is *optimality-preserving* and *gap-increasing*!

Example



Let $Q_k(x_1, a_1) = 0$, $Q_k(x_1, a_2) = 1$ and, for extended MC $Q_k((x_1, a_1), \times) = 10$. Consider a transition $(x, a, y, r) = (x_1, a_1, x_1, 1)$.

- The Bellman Operator: $Q_{k+1}(x_1, a_1) = TQ(x_1, a_1) = r + \gamma \max_b Q(x_1, b) = 1 + \gamma$
- The consistent Bellman Operator: $Q_{k+1}(x_1, a_1) = T_{CB}Q(x_1, a_1) = r + \gamma Q(x_1, a_1) = 1$
- The Bellman Operator on Extended MC: $Q_{k+1}(x_1, a_1) = TQ(x_1, a_1) = r + \gamma Q_k((x_1, a_1), \times) = 1 + 10\gamma$

Optimality-preserving operator

Definition

An operator T' is *optimality-preserving* if, for any $Q_0 \in \mathcal{Q}$ and $x \in X$, for iterations

$$Q_{k+1} = T'Q_k$$

the limit

$$\tilde{V}(x) := \lim_{k\to\infty} \max_{a\in A} Q_k(x,a)$$

exists, is unique s.t. $\tilde{V}(x) = V^*(x)$, and for all $a \in A$,

$$Q^*(x,a) < V^*(x) \Longrightarrow \limsup_{k\to\infty} Q_k(x,a) < V^*(x),$$

where $Q^*(x,a) = r(x,a) + \gamma \mathbb{E}_{x'\sim P(\cdot|x,a)} V^*(x')$.

- At least one optimal action remains optimal
- Suboptimal actions remain suboptimal

Gap-increasing operator

Definition

An operator T' is *gap-increasing* if, for all $Q_0 \in \mathcal{Q}$, $x \in X$, $a \in A$ letting

$$Q_{k+1} = T'Q_k \quad \text{and} \quad V_k(x) := \max_b Q_k(x, b)$$

we have

$$\liminf_{k \to \infty} \left[V_k(x) - Q_k(x, a) \right] \geq V^*(x) - Q^*(x, a)$$

Main Result

Theorem

Let T be the Bellman operator. Let T' be an operator with the property that there exists $\alpha \in [0,1)$ s.t. for all $Q \in \mathcal{Q}$, $x \in X$, $a \in A$

1. $T'Q(x,a) \leq TQ(x,a)$
2. $T'Q(x,a) \geq TQ(x,a) - \alpha\Big[\max_{a \in A} Q(x,a) - Q(x,a)\Big].$

Then T' is both optimality-preserving and gap-increasing.

Consistent Bellman Operator

Theorem

Let T be the Bellman operator. Let T' be an operator with the property that there exists $\alpha \in [0, 1)$ s.t. for all $Q \in \mathcal{Q}$, $x \in X$, $a \in A$

1. $T'Q(x,a) \leq TQ(x,a)$
2. $T'Q(x,a) \geq TQ(x,a) - \alpha\left[\max_{b \in A} Q(x,b) - Q(x,a)\right]$.

Then T' is both optimality-preserving and gap-increasing.

The consistent Bellman operator:

$$\begin{aligned} T_{CB}Q(x,a) &= r(x,a) + \gamma \mathbb{E}\left[\mathbb{I}_{x \neq x'} \max_{b \in A} Q(x', b) + \mathbb{I}_{x = x'} Q(x,a)\right] \\ &= TQ(x,a) - \gamma P(x|x,a)\left[\max_{b \in A} Q(x,b) - Q(x,a)\right]. \end{aligned}$$

1. Obvious
2. $1 > \alpha \geq \max_{x,a} \gamma P(x|x,a)$, for example $\alpha = \gamma$.

Family of Convergent Operators

- The advantage learning (AL) operator:

$$\mathcal{T}_{AL}Q(x,a) := \mathcal{T}Q(x,a) - \alpha\left[\max_{b\in A} Q(x,b) - Q(x,a)\right] \tag{2}$$

Intuition:

We may subtract up to $\max_b Q_k(x,b) - Q_k(x,a)$ from $Q_k(x,a)$ at each iteration. $\max_b Q_k(x,b) - Q_k(x,a)$ is the action gap for Q_k, not Q^*.

- The persistent advantage learning (PAL) operator:

$$\mathcal{T}_{PAL}Q(x,a) := \max\left\{\mathcal{T}_{AL}Q(x,a), r(x,a) + \gamma\mathbb{E}Q(x',a)\right\} \tag{3}$$

Intuition:

We encourage greedy policies which infrequently switch between actions.

$\alpha-$Lazy Operator

- The AL (2) and PAL (3) operators are not contractions, but cannot have more than 1 fixed point. The CB operator (1) is a contraction map.
- The $\alpha-$Lazy Operator may have multiple fixed points:

$$T_{\alpha-Lazy}Q(x,a) := \begin{cases} Q(x,a), & \text{if } Q(x,a) \leq TQ(x,a) \text{ and} \\ & TQ(x,a) \leq \alpha V(x) + (1-\alpha)Q(x,a) \\ TQ(x,a), & \text{otherwise} \end{cases}$$

  - $T_{\alpha-Lazy}$ is optimality-preserving and gap-increasing
  - $T_{\alpha-Lazy}$ is not a contraction map and may have multiple fixed points.

Experimental Results on Atari games

Deep Q-learning (Mnih et al. 2015):

- Initialize function Q_θ with random weights, replay memory D to capacity N.
- for episode $= 1, ..., M$
  - for $t = 1, .., T$
    - Choose action a_t according to $\epsilon-$greedy policy w.r.t. $\max_a Q_{\theta_t}(x_t, a)$
    - Observe (x_{t+1}, r_t). Store (x_t, a_t, r_t, s_{t+1}) in D.
    - Sample minibatch $\left\{(x_j, a_j, r_j, x_{j+1})\right\}_{j \in M}$ from D.

      Set $y_j = r_j + \gamma \max_{a'} Q_{\theta_t}(x_{j+1}, a')$
    - Find θ_{t+1} that minimizes $\frac{1}{|M|} \sum_{j \in M} \Big(y_j - Q_\theta(x_j, a_j)\Big)^2$

We will compare:

- Standard DQL: $y_j = r_j + \gamma \max_{a'} Q_{\theta_t}(x_{j+1}, a')$
- AL-DQL: $y_j = r_j + \gamma \max_{a'} Q_{\theta_t}(x_{j+1}, a') - \alpha[\max_b Q_{\theta_t}(x_j, b) - Q_{\theta_t}(x_j, a_j)]$
- PAL-DQL: $y_j = r_j + \gamma \max_{a'} Q_{\theta_t}(x_{j+1}, a') -$

  $\alpha \min \left[\max_b Q_{\theta_t}(x_j, b) - Q_{\theta_t}(x_j, a_j), \max_b Q_{\theta_t}(x_{j+1}, b) - Q_{\theta_t}(x_{j+1}, a_j)\right]$

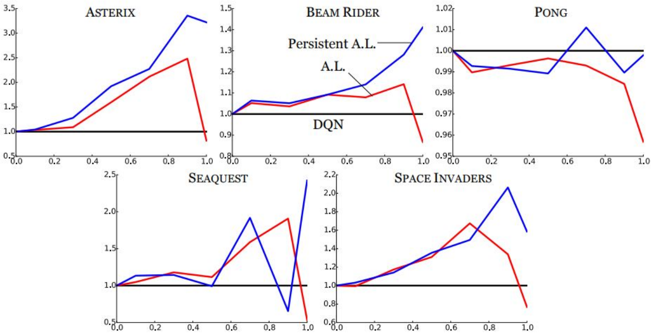


Figure 7: Performance of trained agents in function of the α parameter. Note that $\alpha = 1.0$ does not satisfy our theorem's conditions. We attribute the odd performance of Seaquest agents using Persistent Advantage Learning with $\alpha = 0.9$ to a statistical issue.

Over 60 games for $\alpha = 0.9$:

Operator	DQL	AL-DQL	PAL-DQL
Best score amount 60 games	12*	21*	31*
The median score improvement	0%	8.4%	9.1%
The average score improvement	0%	27%	32.5%

* For 2 games the score was equal for all three settings.

Action gap and value function estimation

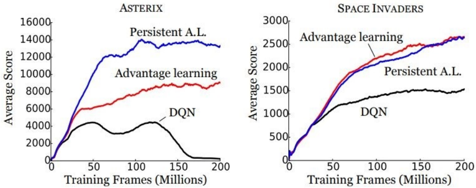


Figure: Learning curves for two Atari games: Asterix and Space Invaders

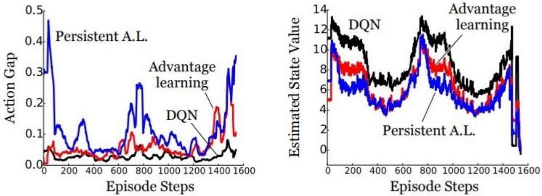


Figure: Action gap and estimated value function (see, van Hasselt 2010) for Space Invaders

Conclusion and open questions

1. The results of the article indicate that there are many practical *optimality-preserving* operators which do not preserve suboptimal $Q-$values and are not contraction.

   Is it possible to find weaker conditions on operators to be optimality-preserving?

2. The consistent Bellman operator was proposed:

$$\begin{aligned} T_{CB}Q(x,a) &= r(x,a) + \gamma \mathbb{E}\Big[\mathbb{I}_{x\neq x'} \max_{b\in A} Q(x',b) + \mathbb{I}_{x=x'} Q(x,a)\Big] \\ &= TQ(x,a) - \gamma P(x|x,a)\Big[V(x) - Q(x,a)\Big] \end{aligned}$$

   Then the authors generalized it to the advantage learning (AL) operator:

$$T_{AL}Q(x,a) := TQ(x,a) - \alpha[V(x) - Q(x,a)], \quad \alpha \in [0,1). \tag{4}$$

   What is the probabilistic interpretation of α and advantage learning?

3. The existence of a broad family of optimality-preserving operators have been revealed: CB, AL, PAL, $\alpha-$Lazy.

   Which of these operators, if any, should we preferred to the Bellman operator?
   Is it possible to find a "maximal efficient" optimality-preserving operator?

Proofs

Lemma

Let $Q \in \mathcal{Q}$ *and* π^Q *be the policy greedy with respect to* Q*:* $\pi^Q(x) := \arg\max_a Q(x,a)$*.*
Let T' *be an operator with the properties that, for all* $x \in X$ *and* $a \in A$*:*

1. $T'Q(x,a) \leq TQ(x,a)$, *and*
2. $T'Q(x,\pi^Q(x)) = TQ(x,\pi^Q(x))$.

Consider the sequence

$$Q_{k+1} := T'Q_k$$

with $Q_0 \in \mathcal{Q}$*, and let*

$$V_k := \max_a Q_k(x,a),$$

Then
the sequence $(V_k : k \in \mathbb{N})$ *converges, and, for all* $x \in X$

$$\lim_{k\to\infty} V_k(x) \leq V^*(x).$$

Proof of Lemma

For an arbitrary $x \in X$, consider a sequence $\{V_k(x)\}_{k=0}^{\infty}$

- The sequence $\{V_k(x)\}_{k=0}^{\infty}$ is bounded:

$$\limsup_{k\to\infty} Q_k(x,a) = \limsup_{k\to\infty} (T')^k Q_0(x,a) \leq \limsup_{k\to\infty} T^k Q_0(x,a) = Q^*(x,a)$$

- Fact: if we have a bounded sequence of real numbers $\{b_0, b_1, ..., b_k, ...\}$ s.t.

$$b_{k+1} \geq b_k - c\gamma^k, \qquad \gamma \in [0,1) \text{ and } c > 0,$$

then the sequence $\{b_k\}_{k=0}^{\infty}$ converges.

- Let $a_k := \arg\max_a Q_k(x,a)$, $P_k := P(\cdot|x,a_k)$, $P_{1:k} = P_k P_{k-1} ... P_1$.

$$\begin{aligned}
V_{k+1}(x) &\geq r(x,a_k) + \gamma \mathbb{E}_{P_k} V_k(x') \\
&= TQ_{k-1}(x,a_k) + \gamma \mathbb{E}_{P_k}[V_k(x') - V_{k-1}(x')] \\
&\geq T'Q_{k-1}(x,a_k) + \gamma \mathbb{E}_{P_k}[V_k(x') - V_{k-1}(x')] \\
&= V_k(x) + \gamma \mathbb{E}_{P_k}[V_k(x') - V_{k-1}(x')] \\
&\geq V_k(x) + \gamma \mathbb{E}_{P_{1:k}}[V_1(x'') - V_0(x'')] \\
&\geq V_k(x) - \gamma^k ||V_1 - V_0||_\infty .
\end{aligned}$$

Proof of the main theorem

Theorem

Let T be the Bellman operator. Let T' be an operator with the property that there exists $\alpha \in [0, 1)$ s.t. for all $Q \in \mathcal{Q}$, $x \in X$, $a \in A$

1. $T'Q(x, a) \leq TQ(x, a)$
2. $T'Q(x, a) \geq TQ(x, a) - \alpha[V(x) - Q(x, a)]$.

Then T' is both

1. *optimality-preserving:*
   1.1 $\lim_{k\to\infty} V_k(x) = V^*(x)$
   1.2 $Q^*(x, a) < V^*(x) \implies \limsup_{k\to\infty} Q_k(x, a) < V^*(x)$.
2. *gap-increasing:* $\liminf_{k\to\infty} \left[V_k(x) - Q_k(x, a)\right] \geq V^*(x) - Q^*(x, a)$

Proof of the main theorem

- Note that $V_k(x) - Q_k(x, \pi_k^Q(x)) = 0$ and we can apply the previous lemma: $\lim\limits_{k\to\infty} V_k(x) = \tilde{V}(x)$ exists, where

$$\begin{cases} Q_k(x,a) = T'Q_{k-1}(x,a) \\ V_k(x) = \max_a Q_k(x,a) \end{cases}$$

- We want to get $\tilde{V}(x) = V^*(x)$.
  Let's show that $\tilde{V}(x) = \max\limits_{a\in A} T\tilde{Q}(x,a)$, where $\tilde{Q}(x,a) = \limsup\limits_{k\to\infty} Q_k(x,a)$.
- $\tilde{Q}(x,a) \leq T\tilde{Q}(x,a)$.

$$\begin{aligned} \tilde{Q}(x,a) &= \limsup_{k\to\infty} T'Q_k(x,a) \leq \limsup_{k\to\infty} TQ_k(x,a) \\ &= \limsup_{k\to\infty} \left[r(x,a) + \gamma E[\max_b Q_k(x',b) \right] \\ &\leq r(x,a) + \gamma E[\max_b \limsup_{k\to\infty} Q_k(x',b) \\ &= T\tilde{Q}(x,a). \end{aligned}$$

Proof of the main theorem: $\tilde{Q}(x,a) \geq T\tilde{Q}(x,a)$.

$\tilde{Q}(x,a) \geq T\tilde{Q}(x,a)$.

- $Q_{k+1}(x,a) \geq TQ_k(x,a) - \alpha[V_k(x) - Q_k(x,a)] =$
  $r(x,a) + \gamma \mathbb{E} V_k(x') - \alpha V_k(x) + \alpha Q_k(x,a)$.
- Taking $\limsup$ of both sides:
  $\tilde{Q}(x,a) \geq r(x,a) + \gamma \mathbb{E}\tilde{V}(x') - \alpha \tilde{V}(x) + \alpha \tilde{Q}(x,a) = T\tilde{Q}(x,a) - \alpha \tilde{V}(x) + \alpha \tilde{Q}(x,a)$.
- $\tilde{Q}(x,a) \geq \frac{1}{1-\alpha}\left[T\tilde{Q}(x,a) - \alpha \tilde{V}(x)\right]$
- Taking $\max\limits_{a \in A}$ of both sides:
  $\tilde{V}(x) \geq \frac{1}{1-\alpha}\left[\max\limits_{a \in A} T\tilde{Q}(x,a) - \alpha \tilde{V}(x)\right] \Longrightarrow \tilde{V}(x) \geq \max\limits_{a \in A} T\tilde{Q}(x,a)$.

Proof of the main theorem: gap-increasing

Observe that the statement:

$$\liminf_{k\to\infty}\Big[V_k(x)-Q_k(x,a)\Big]\geq V^*(x)-Q^*(x,a)$$

is equivalent for the following one for optimality-preserving operators:

$$\limsup_{k\to\infty} Q_k(x,a)\leq Q^*(x,a). \tag{5}$$

The statement (5) has already been proved in the Lemma.

References

- Bellemare, Marc G., Ostrovski, Georg, Guez, Arthur, Thomas, Philip S., and Munos, Rémi. Increasing the action gap: New operators for reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2016.
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Humanlevel control through deep reinforcement learning. Nature 518(7540):529–533.
- van Hasselt, H. 2010. Double Q-learning. In Advances in Neural Information Processing Systems 23.
- Some insights how to interpreted and choose α for the AL operator:

  Greg Farquhar, The Advantage Learning Operator, 2016, http://aims.robots.ox.ac.uk/wp-content/uploads/2015/07/RL_minproject_v1-Greg.pdf
- New optimality-preserving operators on Q-functions:

  Yingdong Lu, Mark S. Squillante, Chai Wah Wu, "A General Family of Robust Stochastic Operators for Reinforcement Learning", 2018, https://arxiv.org/abs/1805.08122