Increasing the Action Gap: New Operators for Reinforcement Learning

By M. Bellemare et. al.

Presented by Chaitanya Patel, Kulin Shah

Subject: Topics in ML
Prof.: Dr. Naresh Manwani
TA: Sahil Chelaramani

Table of contents

Notations

Notations

- Consider a MDP $M := (\mathcal{X}, \mathcal{A}, P, R, \gamma)$.
  - $\mathcal{X} =$ state space
  - $\mathcal{A} =$ finite action space
  - $P(x'|x, a) =$ = Transition probability
  - $R =$ reward function
  - γ = Discount factor
- $\mathcal{Q}$ = Space of Q state-action value functions over $\mathcal{X} \times \mathcal{A}$
- $\mathcal{V}$ = Space of V state value functions over $\mathcal{X}$.

Notations

- Bellman equation for deterministic policy π

$$Q^{\pi}(x, a) := R(x, a) + \gamma \mathbf{E}_P Q^{\pi}(x', \pi(x'))$$

where $\mathbf{E}_P = \mathbf{E}_{x' \sim P(.|x,a)}$.

- Bellman operator $\mathcal{T} : \mathcal{Q} \to \mathcal{Q}$

$$\mathcal{T} Q(x, a) := R(x, a) + \gamma \mathbf{E}_P \max_{b \in \mathcal{A}} Q(x', b)$$

- Q^* is a unique fixed point of Bellman operator $\mathcal{T}$.
- Optimal policy π^*:

$$\pi^*(x) := \arg\max_{a \in \mathcal{A}} Q^*(x, a)$$

Motivation

Motivation

- Authors argue that the optimal Q-function is *inconsistent*, in the sense that for any suboptimal action a in state x, Bellman equation for $Q^*(x, a)$ describes the value of *nonstationary* policy.

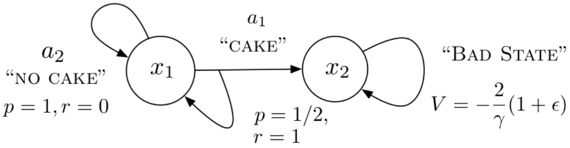


Figure 1: A two-state MDP illustrating the non-stationary aspect of the Bellman operator. Here, p and r indicate transition probabilities and rewards, respectively. In state x_1 the agent may either eat cake to receive a reward of 1 and transition to x_2 with probability $\frac{1}{2}$, or abstain for no reward. State x_2 is a low-value absorbing state with $\epsilon > 0$.

Motivation

- For this example,

$$Q^{\pi}(x_1, a_1) = \frac{\gamma}{2} V^{\pi}(x_1) - \epsilon$$
$$Q^{\pi}(x_1, a_2) = \gamma V^{\pi}(x_1)$$

- $Q^{\pi}(x_1, a_1) < Q^{\pi}(x_1, a_2)$ for any policy π. Thus a_2 is optimal.

$$Q^{*}(x_1, a_2) = V^{*}(x_1) = 0$$
$$Q^{*}(x_1, a_1) = -\epsilon$$

- Here, $Q^{*}(x_1, a_1)$ describes the value of a *nonstationary* policy which takes action a_1 in x_1 to start and then take action a_2 in subsequent turns.

Motivation

- When the MDP can be solved exactly, this nonstationarity is not an issue since only the Q-value for optimal actions matter.
- In the presence of approximation, small error in the Q-function may result in erroneously identifying the optimal action.
- To address this issue, authors propose new operator which incorporates stationarity.

Consistent Bellman Operator

Consistent Bellman Operator

- Authors describe a new Q-function,

$$Q^{\pi}_{stat}(x,a) := R(x,a) + \gamma \mathbf{E}_P \max_{b \in \mathcal{A}} Q^{\pi'}_{stat}(x',b)$$

where

$$\pi'(y) := \begin{cases} a & \text{if } y = x \\ \pi(y) & \text{otherwise.} \end{cases}$$

- As a practical approximation, authors propose the consistent Bellman operator, which preserves a local stationarity:

$$\mathcal{T}_c Q(x,a) := R(x,a) + \gamma \mathbf{E}_P \left[\mathbb{I}_{[x \neq x']} \max_{b \in \mathcal{A}} Q(x',b) + \mathbb{I}_{[x = x']} Q(x,a) \right]$$

- This operator is both *optimality-preserving* and *gap-increasing*.

Optimality Preserving and Gap Increasing Operator

- **Optimality-preserving:** An operator $\mathcal{T}'$ is optimality-preserving if, for any $Q_0 \in \mathcal{Q}$ and $x \in \mathcal{X}$, letting $Q_{k+1} := \mathcal{T}' Q_k$,

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

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

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

- **Gap Increasing:** Let M be an MDP. An operator $\mathcal{T}'$ for M is gap-increasing if for all $Q_0 \in \mathcal{Q}$, $x \in \mathcal{X}$, $a \in \mathcal{A}$, letting $Q_{k+1} := \mathcal{T}' Q_k$ and $V_k(x) := \max_b Q_k(x, b)$,

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

Use of Consistent Bellman Operator in Aggregation Schemes

- An aggregation scheme for MDP M is $(\mathcal{Z}, A, D)$
  - $\mathcal{Z}$ is a set of aggregate state
  - A is a mapping from $\mathcal{X}$ to distributions over $\mathcal{Z}$
  - D is a mapping from $\mathcal{Z}$ to distributions over $\mathcal{X}$
- Define $\mathsf{E}_D := \mathsf{E}_{x \sim D(.|z)}$ and $\mathsf{E}_A := \mathsf{E}_{z' \sim A(.|x')}$. Define the aggregation Bellman operator $\mathcal{T}_A$ as

$$\mathcal{T}_A Q(z, a) := \mathsf{E}_D \left[R(x, a) + \gamma \, \mathsf{E}_P \, \mathsf{E}_A \max_{b \in \mathcal{A}} Q(z', b) \right]$$

Use of Consistent Bellman Operator in Aggregation Schemes

- Authors define the Consistent Bellman operator for aggregation schemes as follows:

$$\mathcal{T}_c Q(z,a) := \mathsf{E}_D \left[R(x,a) + \gamma \mathsf{E}_P \mathsf{E}_A \left[\mathbb{I}_{[z \neq z']} \max_{b \in \mathcal{A}} Q(z',b) + \mathbb{I}_{[z=z']} \, Q(z,a) \right] \right]$$

- To get $Q-$values over $\mathcal{X}$ from $Q-$value from $\mathcal{Z}$, one needs to invert D which is practically infeasible.
- Therefore, authors propose $Q-$value interpolation and corresponding Consistent Bellman operator.

$$Q(x,a) := \mathsf{E}_{z' \sim A(.|x)} \, Q(z',a)$$

Family of Convergent Operators

Family of Convergent Operators

- Authors describe the family of operators which are applicable to arbitrary Q-value approximation schemes.
- These operators are *optimality-preserving* and *gap-increasing*.
- More specifically, authors derive sufficient conditions for an operator to be *optimality-preserving*.
- They show that these operators need not be contractive, nor even guarantee convergence of the Q-values for suboptimal actions.

Main Result

Theorem

Let $\mathcal{T}$ be the Bellman operator. Let $\mathcal{T}'$ be an operator with the property that there exists an $\alpha \in [0, 1)$ such that for all $Q \in \mathcal{Q}, x \in \mathcal{X}, a \in \mathcal{A}$ and letting $V(x) := \max_b Q(x, b)$,

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

Then $\mathcal{T}'$ is both optimality-preserving and gap-increasing.

Consistent Bellman Operator satisfies these conditions and hence it is a part of this family of operators.

Baird's Advantage Learning

- Baird's Advantage Learning is a method of increasing the gap between the optimal and suboptimal actions.
- From Consistent Bellman Operator equation,

$$\mathcal{T}_C Q(x,a) = \mathcal{T} Q(x,a) - \gamma P(x|x,a)[V(x) - Q(x,a)]$$

- Approximating $\gamma P(x|x,a)$ as constant α, we get

$$\mathcal{T}_{AL} Q(x,a) := \mathcal{T} Q(x,a) - \alpha[V(x) - Q(x,a)]$$

- It is similar to the operator of Baird's Advantage Learning and shares the same fixed point.

Persistent Advantage Learning

- In domains with high temporal resolution, it may be advantageous to encourage greedy policies which don't switch between actions too frequently.
- To achieve this *persistent* behaviour, authors define an operator which favours repeated actions,

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

Experiments

Experiments

- Experiments are carried out using normal Bellman operator, advantage learning operator(AL) and persistent advantage learning operator(PAL) on Atari 2600 games.
- Gradient descent on the sample squared error on Q-function is performed as follows:

$$\Delta Q(x,a) := R(x,a) + \gamma V(x') - Q(x,a)$$

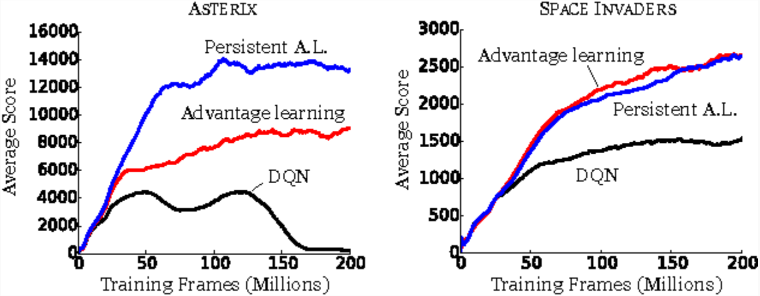
where (x, a, x') is observed transition.

- The gradient for new operators are defined as following

$$\begin{aligned} \Delta Q_{AL}(x,a) &:= \Delta Q(x,a) - \alpha[V(x) - Q(x,a)], \\ \Delta Q_{PAL}(x,a) &:= \max\{\, \Delta Q_{AL}(x,a),\ \Delta Q(x,a) - \alpha[V(x') - Q(x',a)]\,\} \end{aligned}$$

Authors' Results

- Authors have shown improved performance of AL and PAL as compared to DQN with normal Bellman operator.

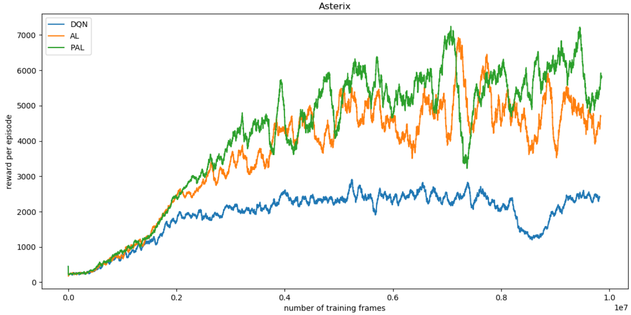


Our experiments

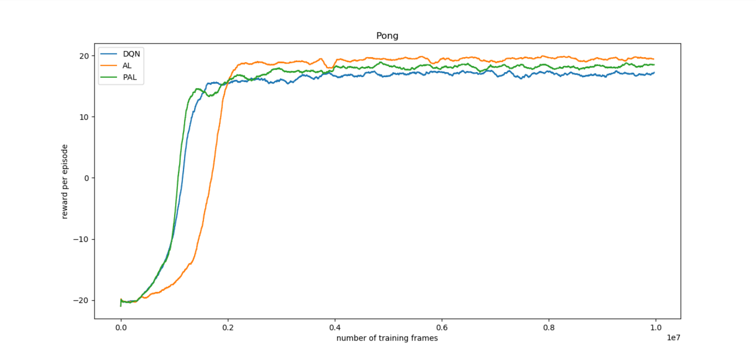
- We implemented DQN with 3 operators
  - Normal Bellman operator
  - Advantage Learning (AL)
  - Persistent Advantage Learning (PAL)
- We trained our agent on 10 million frames(Time $\sim$ 20-25 hours) (relatively less training than author's 200 million frames).
- Results on 5 games: Pong, Asterix, Phoenix , Breakout and SpaceInvaders.

Results on Pong, Asterix and Phoenix

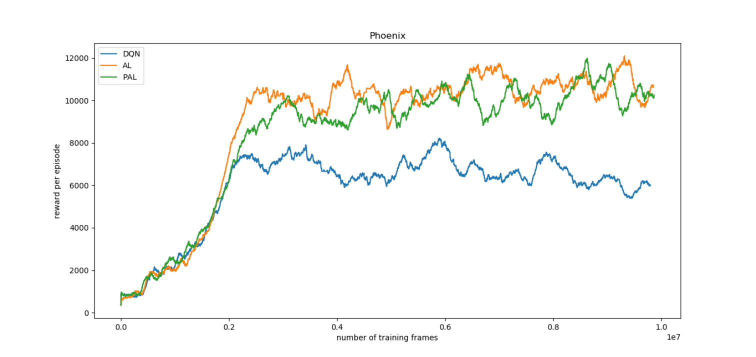
- In Pong, Asterix and Phoenix game, the exploration parameters decays from 1 to 0.05 in 2 million iterations.



Results on Pong, Asterix and Phoenix

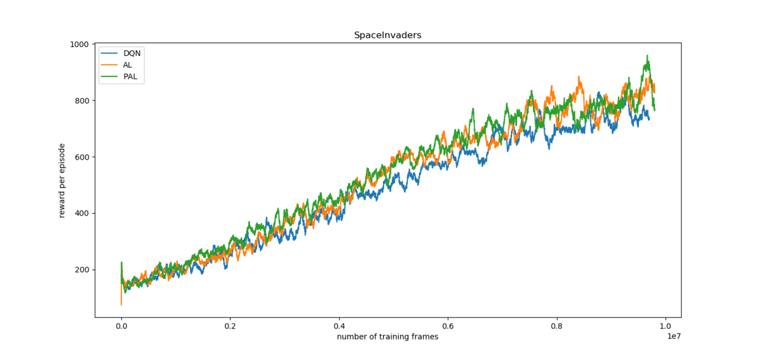


Results on Pong, Asterix and Phoenix

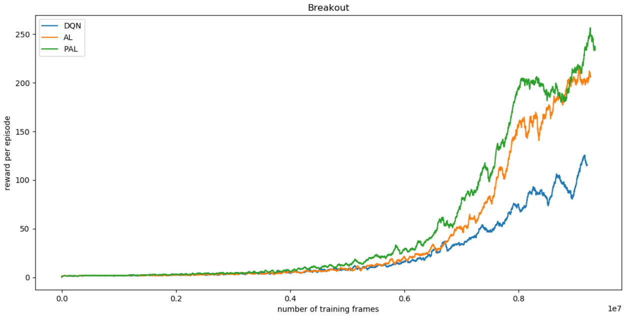


Results on Breakout, SpaceInvaders

- In Breakout and SpaceInvaders game, the exploration parameters decays from 1 to 0.05 in 8 million iterations.

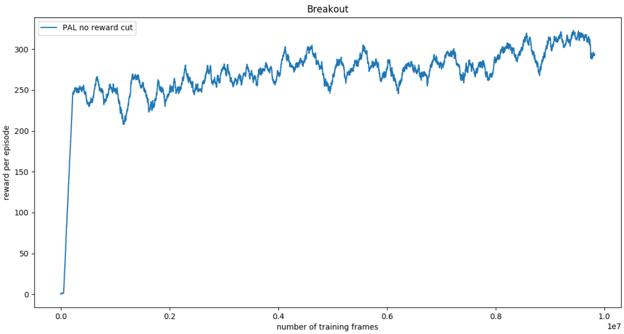


Results on Breakout, SpaceInvaders

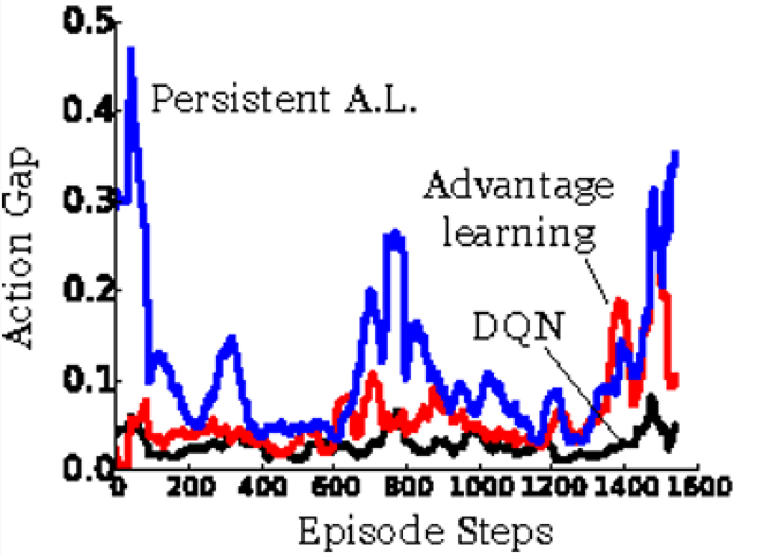


Results Without Reward Clamping

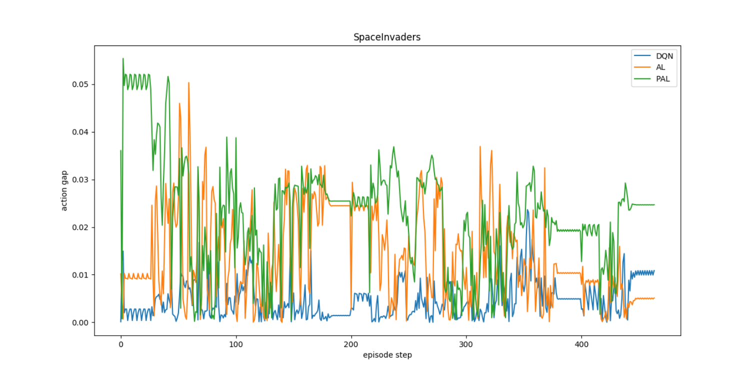
- To back-propagate error, authors clamp the reward between -1 to 1 because the scale of reward can differ a lot.
- To see the effect of clamping in learning agent, we did an experiment with out clamping.
- **Observation:** Unstable learning with very high variance.



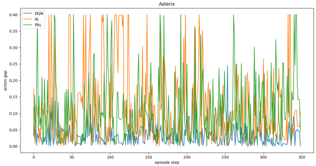
Action Gap Analysis : Authors' Results

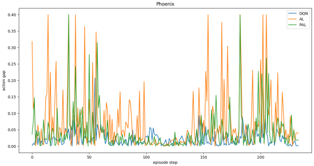


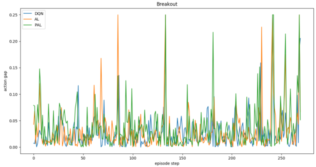
Action Gap Analysis : Our Experiments

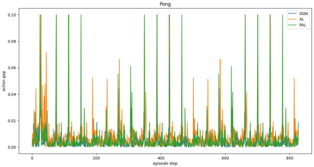


Action Gap Analysis : Our Experiments









Project Scope and Tools

Project Scope and Tools

- Phase 1
  - Understood the problem in original Bellman Operator
  - Understood the proposed solution-Consistent Bellman Operator
  - Understood the sufficient conditions for *optimality-preserving* and *gap-increasing* operators proposed in main theorem
- Phase 2
  - Implemented *DQN*, *advantage learning*, *persistent advantage learning* operator
  - Evaluated performance on 5 Atari-2600 games for all the algorithms (Asterix, Phoenix, Pong, SpaceInvaders, Breakout)
  - Evaluated action gap on all 5 Atari-2600 games for all the algorithms
- Tools : PyTorch, OpenAI Gym

Proof of Main Theorem

Lemma 1

Lemma

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

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

Consider the sequence $Q_{k+1} := \mathcal{T}'Q_k$ with $Q_0 \in \mathcal{Q}$, and let $V_k(x) := \max_a Q_k(x, a)$. Then the sequence $(V_k : k \in \mathbb{N})$ converges, and furthermore, for all $x \in \mathcal{X}$,

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

Lemma

Let $\mathcal{T}'$ be an operator satisfying the conditions of Lemma 1, and let $\|R\|_\infty := \max_{x,a} R(x,a)$. Then for all $x \in \mathcal{X}$ and all $k \in \mathbb{N}$,

$$|V_k(x)| \leq \frac{1}{1-\gamma}\left[2\|V_0\|_\infty + \|R\|_\infty\right].$$

Theorem 2

Theorem

Let $\mathcal{T}$ be the Bellman operator. Let $\mathcal{T}'$ be an operator with the property that there exists an $\alpha \in [0, 1)$ such that for all $Q \in \mathcal{Q}$, $x \in \mathcal{X}$, $a \in \mathcal{A}$, and letting $V(x) := \max_b Q(x, b)$,

1. $\mathcal{T}'Q(x, a) \leq \mathcal{T}Q(x, a)$, *and*
2. $\mathcal{T}'Q(x, a) \geq \mathcal{T}Q(x, a) - \alpha\,[V(x) - Q(x, a)]$.

Consider the sequence $Q_{k+1} := \mathcal{T}'Q_k$ with $Q_0 \in \mathcal{Q}$, and let $V_k(x) := \max_a Q_k(x, a)$. Then $\mathcal{T}'$ is optimality-preserving and gap-increasing.

Proof Idea for Theorem 2

1. Note that given conditions imply the conditions of Lemma 1. Thus for all $x \in \mathcal{X}$, $(V_k(x) : k \in \mathbb{N})$ converges to the limit $\tilde{V}(x) \leq V^*(x)$.
2. We can prove,

$$\tilde{Q}(x, a) = \limsup_{k \to \infty} \mathcal{T}' Q_k(x, a) \leq \limsup_{k \to \infty} \mathcal{T} Q_k(x, a) \leq \mathcal{T} \tilde{Q}(x, a)$$

$$\tilde{V}(x) \geq \max_{a \in \mathcal{A}} \mathcal{T} \tilde{Q}(x, a)$$

   From above 2 equations, we can conclude that $\tilde{V}(x) = V^*(x)$.
3. Proof of gap increasing and optimality preserving from $\tilde{V}(x) = V^*(x)$.