Phase 2 report
Topics in Machine Learning
Prof. Naresh Manwani
TA: Sahil Chelaramani

Kulin Shah (201501234)
Chaitanya Patel (201501071)

Paper

Increasing the Action Gap: New Operators for Reinforcement Learning

Notations

Consider a MDP $M := (\mathcal{X}, \mathcal{A}, P, R, \gamma)$ where

- $\mathcal{X}$ is the state space.
- $\mathcal{A}$ is the finite action space.
- $P(x'|x, a)$ is the transition probability of next state x' from state x with action a.
- R is the reward function.
- γ is the discount factor.

Note that $\mathcal{Q}$ is the space of Q state-action value functions over $\mathcal{X} \times \mathcal{A}$ and $\mathcal{V}$ is the Space of V state value functions over $\mathcal{X}$. The Bellman equation for deterministic policy $\pi : \mathcal{X} \times \mathcal{A}$

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

where $\mathbf{E}_P = \mathbf{E}_{x' \sim P(.|x,a)}$. The Bellman operator $\mathcal{T} : \mathcal{Q} \to \mathcal{Q}$

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

$\mathcal{T}$ is a contraction mapping in supremum norm whose unique fixed point is the optimal Q-function

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

which induces the optimal policy π^*

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

Motivation

In this paper, authors shows that the optimal $Q-$function is inconsistent, in the sense that for any action a which is suboptimal in state x, Bellman's equation for $Q^*(x; a)$ describes the value of a nonstationary policy. This means that upon returning to x, this policy selects $\pi^*(x)$ rather than a.

To illustrate inconsistency of Bellman operator, the authors gave an example. Consider a two-state MDP as following. In figure, 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

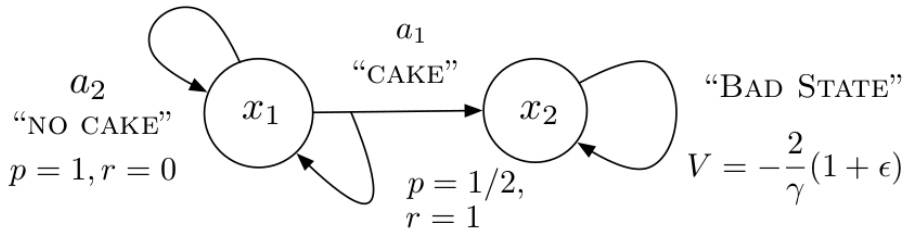

Figure 1: A MDP illustrating the non-stationary aspect of the Bellman operator.

x_2 with probability 0.5, or abstain for no reward. State x_2 is a low-value absorbing state with $\epsilon > 0$.
For a policy $\pi \in \Pi$,

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

For any policy π, $Q^\pi(x_1, a_1) < Q^\pi(x_1, a_2)$. 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. 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.

Proposed Solution

Consistent Bellman Operator

To address the issue of inconsistency, authors describe a new Q-function,

$$Q^\pi_{stat}(x, a) := R(x, a) + \gamma \mathbf{E}_P \max_{b \in \mathcal{A}} Q^{\pi'}_{stat}(x', b) \tag{1}$$

where

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

We can see that (1) explicitly incorporates stationarity. Under this new definition, the action gap of the optimal policy for above discussed example, is $\frac{\epsilon}{1-\gamma/2} > Q(x_1, a_2) - Q(x_1, a_1)$. Unfortunately, it doesn't yield a useful operator on Q. As a practical approximation, authors proposed the Consistent Bellman operator, which preserves a local form of stationarity:

$$\mathcal{T}_c Q(x, a) := R(x, a) + \gamma \mathbf{E}_P \left[\mathbb{I}_{[x \neq x']} \max_{b \in \mathcal{A}} Q(x', b) + \mathbb{I}_{[x = x']} Q(x, a) \right] \tag{2}$$

This operator is both *optimality-preserving* and *gap-increasing*. The property of the operator being *optimality-preserving* and *gap-increasing* can be formalized as follows:

Definition 1. *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 $a \in \mathcal{A}$,

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

Definition 2. *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 $V_k(x) := \max_b Q_k(x,b)$,*

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

The proof is given in the appendix section.

Aggregation Schemes

Authors then show that consistent bellman operator is well-defined over 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 $\mathbf{E}_D := \mathbf{E}_{x\sim D(.|z)}$ and $\mathbf{E}_A := \mathbf{E}_{z'\sim A(.|x')}$. The aggregation Bellman operator $\mathcal{T}_A$ is defined as:

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

Authors define the Consistent Bellman operator $\mathcal{T}_c$ over $\mathcal{Q}_{\mathcal{Z},\mathcal{A}}$:

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

Aggregation schemes as defined above do not immediately yield a Q-function over $\mathcal{X}$ so authors perform Q-value interpolation to get consistent Q-value interpolation Bellman operator($\mathcal{T}_{\text{CQVI}}Q$). Experiments show the improved performance and stability on the bicycle domain experiments.

Family of Convergent Operators

Authors propose sufficient condition for any operator to be *optimality-preserving* and *gap-increasing* which leads to a family of operators including the Consistent Bellman operator. These operators are applicable to arbitrary Q-value approximation schemes. They need not be contractive, nor even guarantee convergence of the Q-values for suboptimal actions. The theorem describing the family of convergent operators is as follows:

Theorem 1. *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.

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 operator with $K := C\Delta_t$, $\Delta_t = 1$ and $\alpha := 1 - K$

$$\mathcal{T}'Q(x,a) = \frac{1}{K}\left[R(x,a) + \gamma^{\Delta_t}\mathbf{E}_P V(x') + (K-1)V(x)\right]$$

Baird's Advantage Learning also shares the same fixed point as Bellman operator.

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,

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

The Lazy Operator

The authors also considered following operator with $\alpha \in [0,1)$

$$\mathcal{T}'Q(x,a) := \begin{cases} Q(x,a) & \text{if } Q(x,a) \le \mathcal{T}Q(x,a) \text{ and } \mathcal{T}Q(x,a) \le \alpha V(x) + (1-\alpha)Q(x,a) \\ \mathcal{T}Q(x,a) & \text{otherwise} \end{cases}$$

This α-lazy operator only updates values when this would affect the greedy policy i.e. this operator will only update when Q value increases but still it satisfies condition of theorem 1 therefore $\mathcal{T}'$ is optimality preserving and gap-increasing.

Experiments

Authors evaluated proposed new operators on the Arcade Learning Environment, a reinforcement learning interface to Atari 2600 games. Experiments are carried out using normal Bellman operator, advantage learning operator(AL) and persistent advantage learning operator(PAL). 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. Corresponding gradient for new operators are defined as:

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

Authors show that AL and PAL performs better than normal Bellman operator on majority of the games. Here we reproduce the authors' results for Asterix and SpaceInvaders.

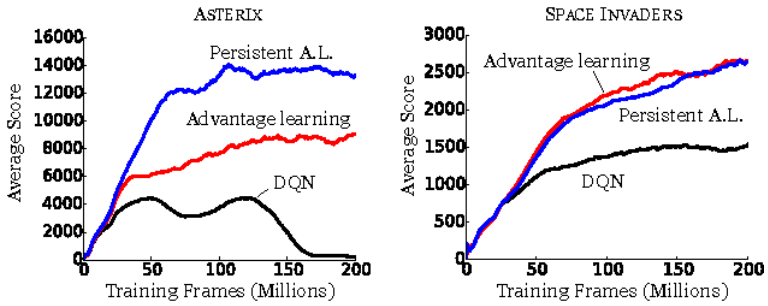

Figure 2: Author's results for Asterix and SpaceInvaders game

Experimental setup

We implemented DQN with three operators : normal Bellman operator, advantage learning operator(AL) and persistent advantage learning operators(PAL). Authors have trained their agents for 200 millions frames for all games and for all operators. Due to limited computational resources, we trained all three operators for 10 million frames on 5 games : Pong, Asterix, Phoenix , Breakout and SpaceInvaders. To train one agent for 10 million frames for any game, our code atleast took around 20 to 25 hours.

First, we trained Asterix, Pong and Phoenix for 10 million frames with exploration parameter epsilon linearly decaying from 1.0 to 0.05 for first 2 million frames. Our results align with authors' results, that is, we observe increased performance on PAL and AL as compared to normal Bellman operator.

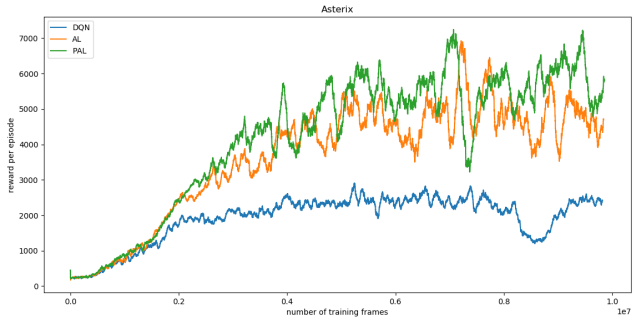

Figure 3: Rewards for Asterix game for all three operators

Observations

Note that the learning is not saturated at the end of training except Pong. If we would have continued the training longer, just like authors, we would have obtained the performance similar to their results. But we found that training for 10 million frames is sufficient for the proof of concept.

We observed that after exploration period, the learning was saturating with high variance. This was because of less exploration period. So we trained Breakout and SpaceInvaders with longer amount of exploration, epsilon decaying from 1.0 to 0.05 for first 8 million frames. Again, we observed similar relative results i.e. AL and PAL performed better than normal Bellman

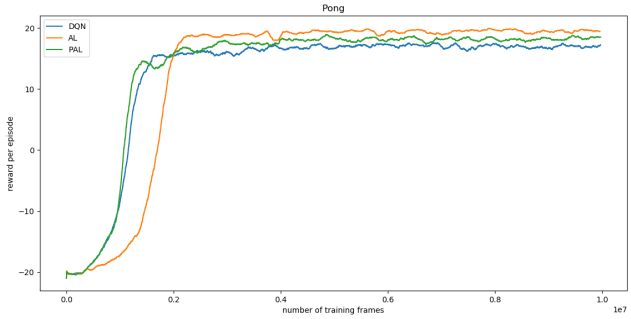

Figure 4: Rewards for Pong game for all three operators

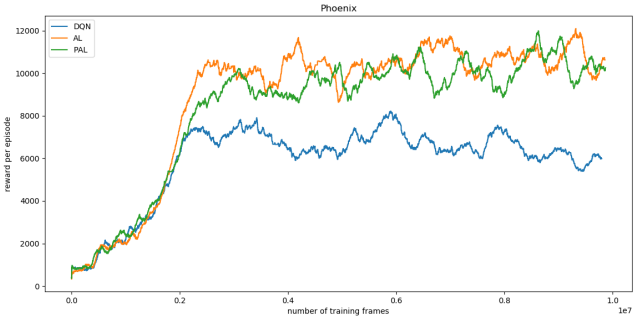

Figure 5: Rewards for Phoenix game for all three operators

operator. Performance was still increasing, so we could have obtained much higher performance with extensive training.

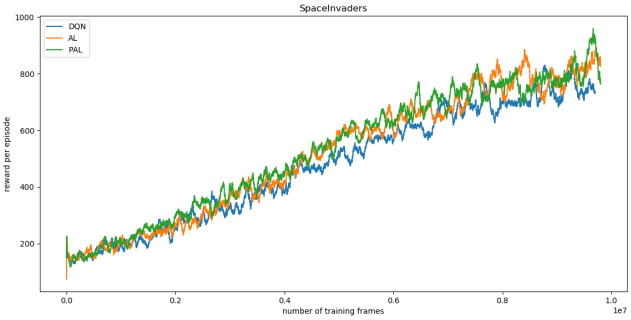

Figure 6: Rewards for Breakout game for all three operators

As we discussed earlier, main idea of advantage learning is to increase the action gap. Authors show that the action gap over one episode for AL and PAL is greater than that for normal Bellman operator. Here we reproduce the figure of action gap over an episode of SpaceInvaders from the paper.

We carried out the similar analysis for action gap for all of our experiments. The action gap for AL and PAL is, on average, greater than normal Bellman operator.

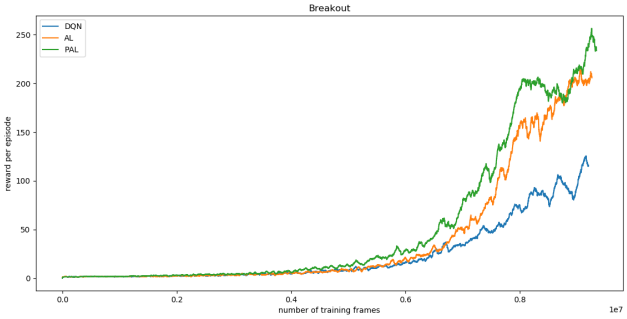

Figure 7: Rewards of Breakout game for all three operators

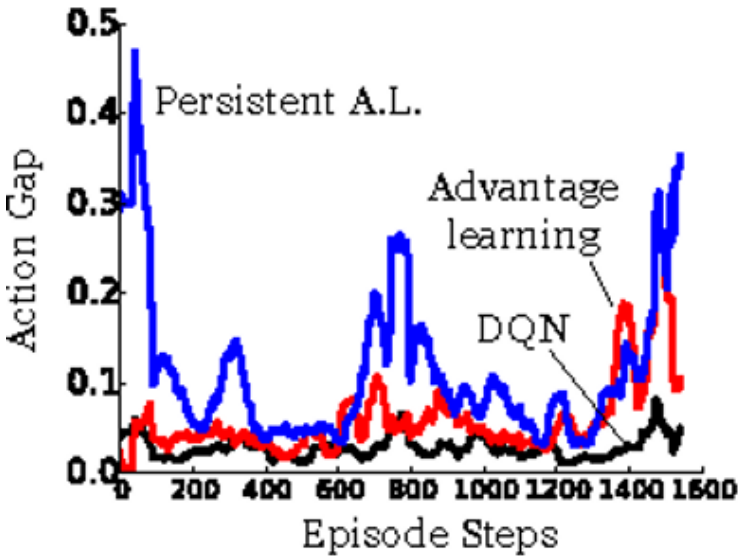

Figure 8: Authors results for action gap for SpaceInvaders game of all three operators

In general, to back propagate the error in the learning phase, authors clamped the reward between -1 to 1 because the scale of reward for each game differ can by a great amount. To see the effect of reward clamping between -1 to 1, we did an experiment where we did not clamped the reward between -1 to 1. We observed that with out clamping, the learning is very unstable. It is quite evident from following figure.

breakout without reward clamping

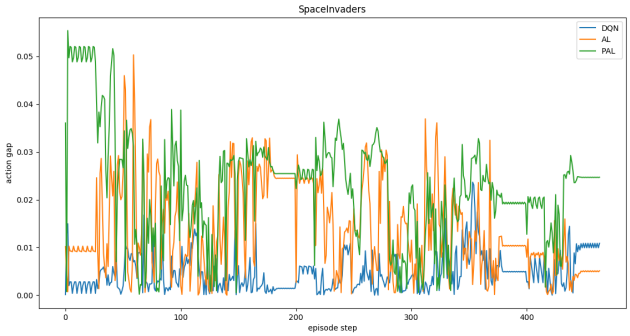

Figure 9: Action gap for SpaceInvaders game over one episode of all three operators

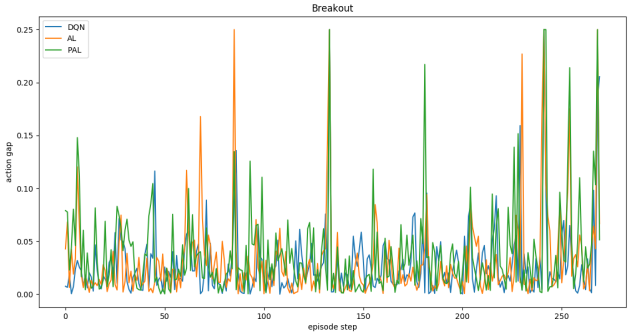

Figure 10: Action gap for Breakout game over one episode of all three operators

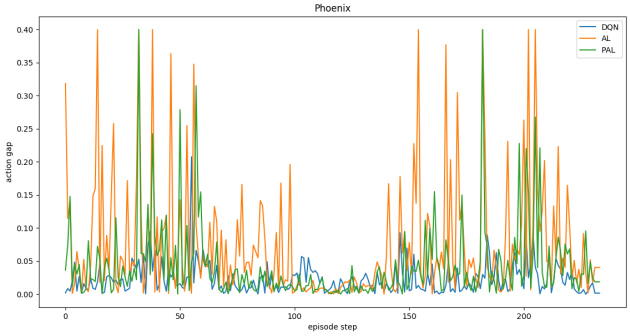

Figure 11: Action gap for Phoenix game over one episode of all three operators

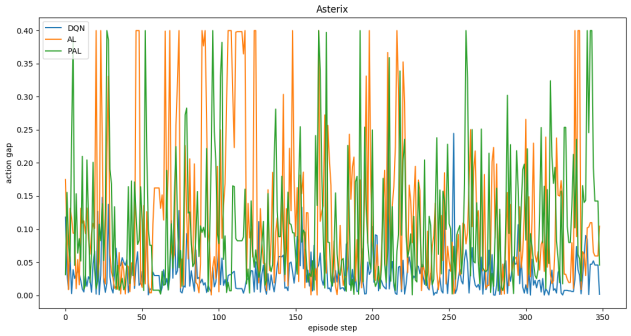

Figure 12: Action gap for Asterix game over one episode of all three operators

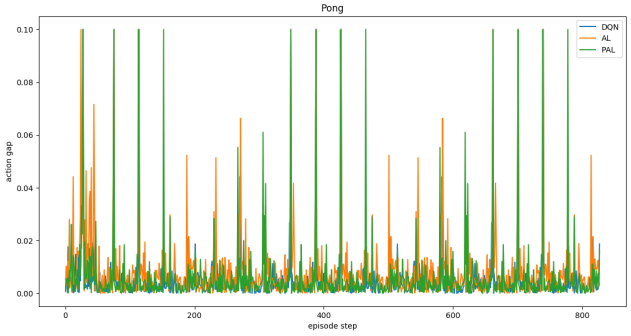

Figure 13: Action gap for Pong game over one episode of all three operators

Proof of Main Theorem

Proof of main theorem describing the family of convergent operators will be followed here. First we'll prove two lemmas, finally proving the main theorem.

Lemma 1. *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 $x \in \mathcal{X}$,

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

Proof. By condition 1, we have

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

Since $\mathcal{T}'$ has a unique fixed point. From this, we can deduce the second claim.
For the first claim of convergence, let $P_k := P(.|x,a_k)$ and $a_k := \pi_k(x) := argmax_a Q_k(x,a)$. We can get,

$$V_{k+1}(x) - V_k(x) \geq \gamma\, \mathbf{E}_{P_k}[V_k(x') - V_{k-1}(x')],$$

Thus by induction, using $P_{1:k} = P_k P_{k-1} \dots P_1$, we get,

$$V_{k+1}(x) - V_k(x) \geq \gamma^k\, \mathbf{E}_{P_{1:k}}[V_1(x') - V_0(x')].$$

Let $\tilde{V}(x) := \limsup_{k\to\infty} V_k(x)$. We can show $\liminf_{k\to\infty} V_k(x) = \tilde{V}(x)$ also. Since $P_{1:k}$ is a nonexpansion in ∞-norm, we have,

$$V_{k+1}(x) - V_k(x) \geq -\gamma^k \|V_1 - V_0\|_\infty \tag{3}$$

We can choose large k to make R.H.S. small. Thus $\liminf_{k\to\infty} V_k(x) = \tilde{V}(x)$ and sequence $V_k(x)$ converges.

□

Lemma 2. *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 $k \in \mathbb{N}$,*

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

Proof. Using eq. (3), we have

$$V_{k+1}(x) - V_0(x) \geq -\frac{1}{1-\gamma}\|V_1 - V_0\|_\infty \tag{4}$$

Now, we can bound $V_{k+1}(x)$ as

$$V_{k+1}(x) = \max_a Q_{k+1}(x,a) \leq \frac{1}{1-\gamma}\left[2\|V_0\|_\infty + \|R\|_\infty\right]$$

Combining above equation with eq. (4) and statement of lemma for $k = 0$, we will get the required result. □

Theorem 2. *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: for all $x \in \mathcal{X}$, $(V_k(x) : k \in \mathbb{N})$ converges,

$$\lim_{k\to\infty} V_k(x) = V^*(x),$$

and

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

Furthermore, $\mathcal{T}'$ is also gap-increasing:

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

Proof. 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 $\tilde{V}(x) \leq V^*(x)$.
Now let $\tilde{Q}(x,a) := \limsup_k Q_k(x,a)$. We can prove,

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

Then we can prove,

$$Q_{k+1}(x,a) \geq R(x,a) + \gamma\, \mathbf{E}_P V_k(x') - \alpha V_k(x) + \alpha Q_k(x,a).$$

Taking lim sup both side,

$$\begin{aligned}
\tilde{Q}(x,a) &\geq \mathcal{T}\tilde{Q}(x,a) - \alpha\tilde{V}_k(x) + \alpha\tilde{Q}(x,a) \\
\tilde{Q}(x,a) &\geq \frac{1}{1-\alpha}\left[\mathcal{T}\tilde{Q}(x,a) - \alpha\tilde{V}(x)\right] \\
\tilde{V}(x) &\geq \max_{a\in\mathcal{A}} \mathcal{T}\tilde{Q}(x,a).
\end{aligned}$$

From above two results, we can conclude that $\tilde{V}(x) = V^*(x)$. Now, to prove optimality preserving,

$$Q_k(x,a) \leq Q^*(x,a) - \gamma\mathbb{E}_P[V^*(x') - V_{k-1}(x')]$$

Now, combining above equation with $Q^*(x,a) < V^*(x)$, we will get

$$\lim_{k\to\infty} \sup Q_k(x,a) < V^*(x)$$

Proving $\mathcal{T}'$ is gap-increasing is equivalent to

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

which is immediate from the conditions of Lemma 1.

□