Increasing The Action Gap: New Operators For Reinforcement Learning

Supplemental

This appendix is divided into three sections. In Section 1 we present the proofs of our theoretical results. In Section 2 we provide experimental details and additional results for the Bicycle domain. Finally in Section 3 we provide details of our experiments on the Arcade Learning Environment, including results on 60 games.

1 Theoretical Results

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 that

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

since $\mathcal{T}$ has a unique fixed point. From this we deduce the second claim. Now, for a given $x \in \mathcal{X}$, let $a_k := \pi_k(x) := \arg\max_a Q_k(x,a)$ and $P_k := P(\cdot \mid x, a_k)$. We have

$$\begin{aligned}V_{k+1}(x) \geq Q_{k+1}(x,a_k) &= \mathcal{T}'Q_k(x,a_k) \\ &= \mathcal{T}Q_k(x,a_k) \\ &= \mathcal{T}Q_{k-1}(x,a_k) + \gamma \mathbf{E}_{P_k}\left[V_k(x') - V_{k-1}(x')\right] \\ &\geq \mathcal{T}'Q_{k-1}(x,a_k) + \gamma \mathbf{E}_{P_k}\left[V_k(x') - V_{k-1}(x')\right] \\ &= V_k(x) + \gamma \mathbf{E}_{P_k}\left[V_k(x') - V_{k-1}(x')\right],\end{aligned}$$

where in the second line we used Condition 2 of the lemma, and in the third the definition of $\mathcal{T}$ applied to Q_k. Thus we have

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

and by induction

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

where $P_{1:k} := P_k P_{k-1} \dots P_1$ is the k-step transition kernel at x derived from the nonstationary policy $\pi_k \pi_{k-1} \dots \pi_1$. Let $\tilde{V}(x) := \limsup_{k\to\infty} V_k(x)$. We now show that $\liminf_{k\to\infty} V_k(x) = \tilde{V}(x)$ also. First note that Conditions 1 and 2, together with the boundedness of V_0, ensure that V_1 is also bounded and thus $\|V_1 - V_0\|_\infty < \infty$. By definition, for any $\delta > 0$ and $n \in \mathbb{N}$, $\exists k \geq n$ such that $V_k(x) > \tilde{V}(x) - \delta$. Since $P_{1:k}$ is a nonexpansion in ∞-norm, we have

$$\begin{aligned}V_{k+1}(x) - V_k(x) &\geq -\gamma^k \|V_1 - V_0\|_\infty \\ &\geq -\gamma^n \|V_1 - V_0\|_\infty =: -\epsilon,\end{aligned}$$

and for all $t \in \mathbb{N}$,

$$V_{k+t}(x) - V_k(x) \geq -\sum_{i=0}^{t-1} \gamma^i \epsilon \geq \frac{-\epsilon}{1-\gamma},$$

such that

$$\inf_{t\in\mathbb{N}} V_{k+t}(x) \geq \tilde{V}(x) - \delta - \frac{\epsilon}{1-\gamma}.$$

It follows that for any $x \in \mathcal{X}$ and $\delta' > 0$, we can choose an $n \in \mathbb{N}$ to make ϵ small enough such that for all $k \geq n$, $V_k(x) > \tilde{V}(x) - \delta'$. Hence

$$\liminf_{k\to\infty} V_k(x) = \tilde{V}(x),$$

and thus $V_k(x)$ 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 $k \in \mathbb{N}$,*

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

Proof. Following the derivation of Lemma 1, we have

$$\begin{aligned}V_{k+1}(x) - V_0(x) &\geq -\sum_{i=1}^{k} \gamma^i \|V_1 - V_0\|_\infty \\ &\geq \frac{-1}{1-\gamma}\|V_1 - V_0\|_\infty.\end{aligned}$$

By the same derivation, for $a_0 := \arg\max_a Q_0(x,a)$ we have

$$V_1(x) \geq \mathcal{T}Q_0(x,a_0).$$

But then

$$V_1(x) - V_0(x) \geq R(x,a_0) + \gamma \mathbf{E}_{P_0} V_0(x') - V_0(x),$$

from which the lower bound follows. Now let P_k be defined as in the proof of Lemma 1, and assume the upper bound of (2) holds up to $k \in \mathbb{N}$. Then

$$\begin{aligned}
V_{k+1}(x) &= \max_a Q_{k+1}(x,a) = \max_a \mathcal{T}'Q_k(x,a) \\
&\leq \max_a \mathcal{T}Q_k(x,a) \\
&= \max_a \left[R(x,a) + \gamma \mathbf{E}_{P_k} V_k(x')\right] \\
&\leq \|R\|_\infty + \gamma \|V_k\|_\infty \\
&\leq \|R\|_\infty + \frac{\gamma}{1-\gamma}\left[2\|V_0\|_\infty + \|R\|_\infty\right] \\
&\leq \frac{1}{1-\gamma}\left[2\|V_0\|_\infty + \|R\|_\infty\right],
\end{aligned}$$

and combined with the fact that (2) holds for $k = 0$ this proves the upper bound. □

Theorem 1. *Let $\mathcal{T}$ be the Bellman operator ((1) in the main text). 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 $V_k(x) := \max_a Q_k(x,a)$. 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

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

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

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

Proof. Note that these 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 have

$$\begin{aligned}
\tilde{Q}(x,a) &= \limsup_{k\to\infty} \mathcal{T}'Q_k(x,a) \\
&\leq \limsup_{k\to\infty} \mathcal{T}Q_k(x,a) \\
&= \limsup_{k\to\infty} \left[R(x,a) + \gamma \mathbf{E}_P \max_{b\in\mathcal{A}} Q_k(x',b)\right] \\
&\leq R(x,a) + \gamma \mathbf{E}_P \limsup_{k\to\infty} \max_{b\in\mathcal{A}} Q_k(x',b) \quad (3) \\
&= R(x,a) + \gamma \mathbf{E}_P \max_b \limsup_{k\to\infty} Q_k(x',b) \quad (4) \\
&= \mathcal{T}\tilde{Q}(x,a), \quad (5)
\end{aligned}$$

where in (3) we used Jensen's inequality, and (4) follows from the commutativity of $\max$ and $\limsup$. Now

$$\begin{aligned}
Q_{k+1}(x,a) &= \mathcal{T}'Q_k(x,a) \\
&\geq \mathcal{T}Q_k(x,a) - \alpha\left[V_k(x) - Q_k(x,a)\right] \\
&= R(x,a) + \gamma \mathbf{E}_P V_k(x') - \alpha V_k(x) \quad + \\
&\quad \alpha Q_k(x,a). \quad (6)
\end{aligned}$$

Now, by Lemma 1 $V_k(x)$ converges to $\tilde{V}(x)$. Furthermore, using Lemma 2 and Lebesgue's dominated convergence theorem, we have

$$\lim_{k\to\infty} \mathbf{E}_P V_k(x') = \mathbf{E}_P \tilde{V}(x'). \quad (7)$$

We now take the $\limsup$ of both sides of (6), which Lemma 2 guarantees exists, and obtain

$$\begin{aligned}
\tilde{Q}(x,a) &\geq R(x,a) + \gamma \mathbf{E}_P \tilde{V}(x') - \alpha\tilde{V}(x) + \alpha\tilde{Q}(x,a) \\
&= \mathcal{T}\tilde{Q}(x,a) - \alpha\tilde{V}(x) + \alpha\tilde{Q}(x,a).
\end{aligned}$$

Thus

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

Combining the above with (5), we deduce that

$$\tilde{V}(x) = \max_{a\in\mathcal{A}} \mathcal{T}\tilde{Q}(x,a) = \max_{a\in\mathcal{A}} \left[R(x,a) + \gamma \mathbf{E}_P \tilde{V}(x')\right]$$

and, by uniqueness of the fixed point of the Bellman operator over $\mathcal{V}$, it must be that $\tilde{V} = V^*$.

Now suppose that for some $x \in \mathcal{X}$, $\tilde{a} \in \mathcal{A}$, we have

$$Q^*(x,\tilde{a}) < V^*(x).$$

By Condition 1

$$\begin{aligned}
Q_k(x,\tilde{a}) &= \mathcal{T}'Q_{k-1}(x,\tilde{a}) \\
&\leq \mathcal{T}Q_{k-1}(x,\tilde{a}) \\
&= \mathcal{T}Q^*(x,\tilde{a}) - \gamma \mathbf{E}_{P_{\tilde{a}}}\left[V^*(x') - V_{k-1}(x')\right] \\
&= Q^*(x,\tilde{a}) - \gamma \mathbf{E}_{P_{\tilde{a}}}\left[V^*(x') - V_{k-1}(x')\right],
\end{aligned}$$

where $P_{\tilde{a}} := P(\cdot \mid x, \tilde{a})$. Using (7) we take the $\limsup$ on both sides and find that

$$\begin{aligned}
\limsup_{k\to\infty} Q_k(x,\tilde{a}) &\leq Q^*(x,\tilde{a}) - \gamma \mathbf{E}_{P_{\tilde{a}}}\left[V^*(x') - \tilde{V}(x')\right] \\
&= Q^*(x,\tilde{a}) \\
&< V^*(x).
\end{aligned}$$

We conclude that

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

Hence, $\mathcal{T}'$ is optimality-preserving. To prove that $\mathcal{T}'$ is gap-increasing, observe that the statement

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

is now equivalent to

$$\limsup_{k\to\infty} Q_k(x,a) \leq Q^*(x,a) \tag{8}$$

since $\lim_k V_k(x) = V^*(x)$. But we know (8) to be true from Condition 1 (see the proof of Lemma 1). □

Corollary 1. *The consistent Bellman operator $\mathcal{T}_{\text{C}}$ ((5) in the main text) and consistent Q-value interpolation Bellman operator $\mathcal{T}_{\text{CQVI}}$ ((9) in the main text) are optimality-preserving and gap-increasing.*

2 Experimental Details: Bicycle

We used the bicycle simulator described by Randlov and Alstrom (1998) with a reward function which encourages driving towards the goal. Recall that Randlov and Alstrom's reward function is

$$R(x,a) := \begin{cases} -1 & \text{if bicycle falls} \\ 0.01 & \text{if goal is reached} \\ (4-\psi^2)\times 0.00004 & \text{otherwise} \end{cases}$$

As noted by Randlov and Alstrom themselves, this reward function is unsuitable for value iteration methods, since it rewards driving away from the goal. Instead we use the following related reward function

$$R(x,a) := \begin{cases} -c & \text{if fallen} \\ 1.0 & \text{if goal reached} \\ (\pi^2/4 - \psi^2 - 1)\times 0.001 & \text{otherwise} \end{cases}$$

with $c := (\frac{3}{4}\pi^2 - 1)\times 0.001$ the largest negative reward achievable by the agent. Empirically, we found this reward function easier to work with, while our results remained qualitatively similar for similar reward functions. We further use a discount factor of $\gamma = 0.99$.

We consider two sample-based operators on $\mathcal{Q}_{\mathcal{Z},\mathcal{A}}$, the space of Q-functions over representative states. The sample-based Q-value interpolation Bellman operator is defined as

$$\mathcal{T}_{\text{QVI}}Q(z,a) := R(z,a) + \gamma \frac{1}{k}\sum_{i=1}^{k} \max_{b\in\mathcal{A}} Q(x_i', b),$$

with $k \in \mathbb{N}$ and $x_i' \sim P(\cdot \,|\, z,a)$. The sample-based consistent Q-value interpolation Bellman operator $\mathcal{T}_{\text{CQVI}}$ is similarly defined by sampling x' from P:

$$\mathcal{T}'_{\text{QVI}}Q(z,a) := R(z,a) + \frac{\gamma}{k}\sum_{i=1}^{k} \max_{b\in\mathcal{A}} \left[Q(x',b) - A(z\,|\,x')\left(Q(z,b) - Q(z,a)\right)\right]$$

$$\mathcal{T}_{\text{CQVI}}Q(z,a) := \min\left\{\mathcal{T}_{\text{QVI}}Q(z,a), \mathcal{T}'_{\text{QVI}}Q(z,a)\right\}.$$

In both cases, we use Q-value interpolation to define a Q-function over $\mathcal{X}$:

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

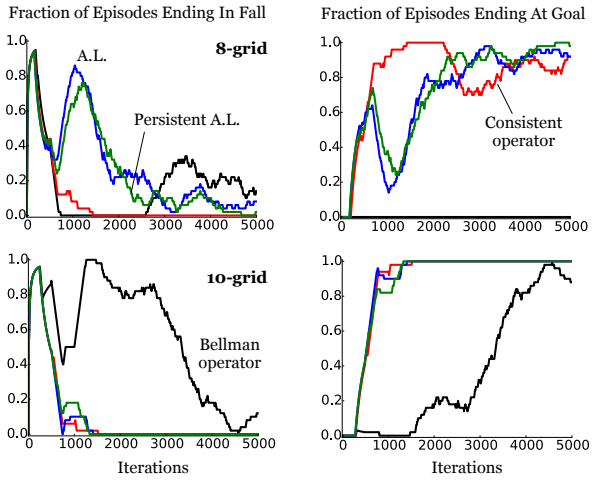

Figure 1: **Top.** Falling and goal-reaching frequency for greedy policies derived from value iteration on a $8\times\cdots\times 8$ grid. **Bottom.** The same, for a $10\times\cdots\times 10$ grid.

For each operator $\mathcal{T}'$, we computed a sequence of Q-functions $Q_k \in \mathcal{Q}_{\mathcal{Z},\mathcal{A}}$ using an averaging form of value iteration:

$$Q_{k+1}(z,a) = (1-\eta)Q_k(z,a) + \eta\mathcal{T}'Q_k(z,a),$$

applied simultaneously to all $z \in \mathcal{Z}$ and $a \in \mathcal{A}$. We chose this averaging version because it led to faster convergence, and lets us take $k = 1$ in the definition of both operators. From a parameter sweep we found $\eta = 0.1$ to be a suitable step-size.

Our multilinear grid was defined over the six state variables. As done elsewhere in the literature, we defined our grid over the following bounded variables:

$$\begin{aligned} \omega &\in \left[-\frac{4}{9}\pi, \frac{4}{9}\pi\right], \\ \dot\omega &\in [-2, 2], \\ \theta &\in \left[-\frac{\pi}{15}, \frac{\pi}{15}\right], \\ \dot\theta &\in [-0.5, 0.5], \\ \psi &\in [-\pi, \pi], \\ d &\in [10, 1200]. \end{aligned}$$

Values outside of these ranges were accordingly set to the range's minimum or maximum.

For completeness, Figure 1 compares the performance of the Bellman and consistent Bellman operators, as well as advantage learning and persistent advantage learning (with $\alpha = 0.1$), on $8\times\cdots\times 8$ and $10\times\cdots\times 10$ grids. Here, the usual Bellman operator is unable to find a solution to the goal, while the consistent Bellman operator successfully does so. The two other operators also achieve superior performance compared to Bellman operator, although appear slightly more unstable in the smaller grid setting.

3 Experimental Details: ALE

We omit details of the DQN architecture, which are provided in Mnih et al. (2015). A *frame* is a single emulation step within the ALE, while a *time step* consists of four consecutive frames which are treated atomically by the agent.

Our first Atari 2600 experiment (*Stochastic Minimal* setting) used stochastic controls, which operate as follows: at each frame (not time step), the environment *accepts* the agent's action with probability $1-p$, or *rejects* it with probability p. If an action is rejected, the previous frame's action is repeated. In our setting, the agent selects a new action every four frames; in this situation, the stochastic controls approximate a form of reaction delay. This particular setting is part of the latest Arcade Learning Environment. For our experiments we use the ALE 0.5 standard value of $p = 0.25$, and trained agents for 100 million frames.

Our second Atari 2600 experiment (*Original DQN* setting) was averaged over three different trials, ran for 200 million frames (instead of 100 million), defined a lost life as a termination signal, and did not use stochastic controls. This matches the experimental setting of Mnih et al. (2015). A full table of our results is provided in Table 1.

Our last experiment took place in the Original DQN setting. We generated a trajectory from a trained DQN agent playing an ϵ-greedy policy with $\epsilon = 0.05$. The full trajectory (up to the end of the episode) was recorded in this way. We then queried the value functions of the trained agents, including the DQN used to generate the trajectory, in order to generate Figure 4 of the main text. For clarity we report action gaps averaged according to a rolling window of length 50.

Out of the 60 games for which we report results, 5 are new when compared to the table of results provided by Bellemare et al. (2013). These five games are identified with a † in Table 1.

DQN Implementation Details

Recall that DQN maintains two networks in parallel: a *policy* network, which is used to select actions and is updated at every time step, and a *target* network. The target network is used to compute the error term ΔQ, and is only updated every 10,000 time steps(Mnih et al., 2015). In our experiments we also used this target network to compute the $\Delta_{\text{AL}} Q$ and $\Delta_{\text{PAL}} Q$, including the added correction term. Our operators performed worse when the correction term was instead computed from the policy network.

Parameter Selection

We used five training games (ASTERIX, BEAM RIDER, PONG, SEAQUEST, SPACE INVADERS) to select the α parameter for both of our operators. Specifically, we trained agents using our second experimental setup with parameters $\alpha \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$, evaluated them according to the highest score achieved, and manually selected the α value which seemed to achieve the best performance. Note that $\alpha = 0.0$ corresponds to DQN in both cases. Figure 2 depicts the results of this parameter sweep.

References

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47:253–279.

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. Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Randlov, J., and Alstrom, P. 1998. Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the Fifteenth International Conference on Machine Learning*.

Game	Bellman	Advantage Learning	Persistent A.L.
Asterix	6074.98	12852.08	**19564.90**
Beam Rider	9316.10	10054.58	**13145.34**
Pong	**19.80**	19.66	19.76
Seaquest	5458.17	8670.50	**13230.74**
Space Invaders	2067.19	**3460.79**	3277.59
Alien	3154.67	4990.91	**5699.81**
Amidar	969.88	**1557.43**	1451.65
Assault	**4573.67**	3661.51	3304.33
Asteroids	1827.97	**1924.42**	1673.52
Atlantis	636657.62	553591.67	**1465250.00**
Bank Heist	511.00	633.63	**874.99**
Battle Zone	28082.91	28789.29	**34583.07**
Berzerk	667.61	747.26	**1328.25**
Bowling	**74.62**	57.41	71.59
Boxing	88.66	93.94	**94.30**
Breakout	378.69	425.32	**431.89**
Carnival	**5238.14**	5111.40	4679.93
Centipede	**5719.11**	4225.18	4539.55
Chopper Command	**8195.88**	5431.36	5734.93
Crazy Climber	114105.56	123410.71	**130002.71**
Defender†	16746.68	30643.59	**32038.93**
Demon Attack	23212.19	27153.48	**70908.17**
Double Dunk	-6.23	**-0.15**	-2.51
Elevator Action	26675.00	27088.89	**29100.00**
Enduro	776.14	1252.70	**1343.10**
Fishing Derby	11.65	21.32	**28.13**
Freeway	31.14	31.72	**32.30**
Frostbite	1485.42	2305.82	**3248.96**
Gopher	8479.98	**11912.68**	10611.81
Gravitar	**448.74**	417.65	446.92
H.E.R.O.	18490.97	**24788.86**	24175.79
Ice Hockey	-2.13	-1.24	**-0.25**
James Bond	**867.84**	848.46	772.09
Kangaroo	9157.98	10809.16	**11478.46**
Krull	8500.48	**9548.92**	8689.81
Kung-Fu Master	25977.53	32182.99	**34650.91**
Montezuma's Revenge	0.64	0.42	**1.72**
Ms. Pac-Man	3081.29	**4065.80**	3917.55
Name This Game	8585.03	**11025.26**	10431.33
Phoenix†	14278.95	**22038.27**	14495.56
Pitfall!†	**0.00**	**0.00**	**0.00**
Pooyan	4736.79	4801.27	**5858.84**
Private Eye	957.83	**5276.16**	339.15
Q*Bert	10840.83	**14368.03**	14254.78
River Raid	7315.20	10585.12	**12813.27**
Road Runner	38042.07	**52351.23**	37856.16
Robotank	61.97	69.31	**70.53**
Skiing	-13049.42	-13264.51	**-12173.35**
Solaris†	4638.85	**4785.16**	3274.70
Star Gunner	55558.27	61353.59	**61521.87**
Surround	-5.79	-4.15	**0.72**
Tennis	**0.00**	**0.00**	**0.00**
Time Pilot	5788.96	**8969.12**	8749.26
Tutankham	200.17	**245.22**	197.33
Up and Down	12831.57	**13909.74**	13542.07
Venture	**373.79**	198.69	243.75
Video Pinball	**611840.72**	543504.00	542052.00
Wizard Of Wor	2410.47	9541.14	**10254.01**
Yar's Revenge†	21440.45	**24240.03**	17141.56
Zaxxon	6416.06	**9129.61**	8155.60
Times Best	12	21	31

Table 1: Highest performance achieved by each of our operators. For each game, the score of the best operator is highlighted. Games with a † were not used by Bellemare et al. (2013). See Section 4 of the main text for more details.

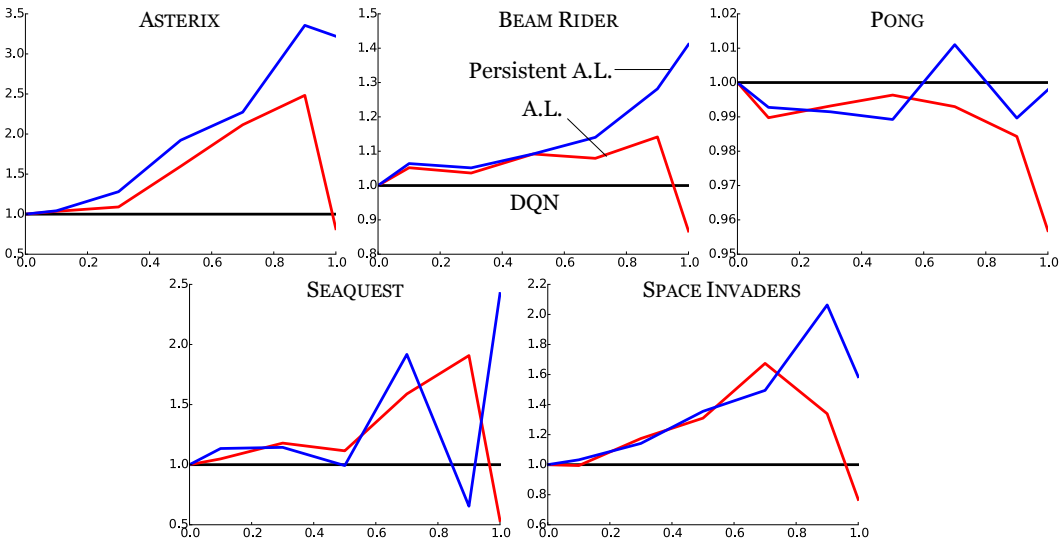

Figure 2: Performance of trained agents in function of the α parameter. Note that $\alpha = 1.0$ does not satisfy our theorem's conditions. We attribute the odd performance of Seaquest agents using Persistent Advantage Learning with $\alpha = 0.9$ to a statistical issue.