Revisiting the Softmax Bellman Operator: New Benefits and New Perspective Supplemental Material

Zhao Song [1 *] **Ronald E. Parr** [1] **Lawrence Carin** [1]

1. Proof for Performance Bound

We first show that for all Q-functions that occur during Q-iteration with $\mathcal{T}_{\text{soft}}$, their corresponding Q-values are bounded.

Lemma A1. *Assuming $\forall(s,a)$, the initial Q-values $Q_0(s,a) \in [R_{min}, R_{max}]$, the Q-values during Q-iteration with $\mathcal{T}_{soft}$ are within $[Q_{min}, Q_{max}]$, with $Q_{min} = \frac{R_{min}}{1-\gamma}$ and $Q_{max} = \frac{R_{max}}{1-\gamma}$.*

Proof. The upper bound can be obtained by showing $\forall(s,a)$, the Q-values at the ith iteration are bounded as

$$Q_i(s,a) \leq \sum_{j=0}^{i} \gamma^j R_{\max}. \tag{A1}$$

We then prove Eq. (A1) by induction as follows. The lower bound can be proven similarly.

(i) When $i = 1$, we start from the definition of $\mathcal{T}_{\text{soft}}$ in Eq. (3) and the assumption of Q_0 to have

$$\begin{aligned}Q_1(s,a) &= \mathcal{T}_{\text{soft}}\, Q_0(s,a) \\ &\leq R_{\max} + \gamma \sum_{s'} P(s'|s,a)\, \max_{a'} Q_0(s',a') \\ &\leq R_{\max} + \gamma \sum_{s'} P(s'|s,a)\, R_{\max} \\ &= (1+\gamma) R_{\max}.\end{aligned}$$

(ii) Assuming Eq. (A1) holds when $i = k$, i.e., $Q_k(s,a) \leq \sum_{j=0}^{k} \gamma^j R_{\max}$. Then,

$$\begin{aligned}Q_{k+1}(s,a) &= \mathcal{T}_{\text{soft}} Q_k(s,a) \\ &\leq R_{\max} + \gamma \sum_{s'} P(s'|s,a)\, \max_{a'} Q_k(s',a') \\ &\leq R_{\max} + \gamma \sum_{s'} P(s'|s,a) \sum_{j=0}^{k} \gamma^j R_{\max} \\ &= \sum_{j=0}^{k+1} \gamma^j R_{\max}.\end{aligned}$$

□

Corollary A2. *Assuming $R_{max} \geq -R_{min} \geq 0$ WLOG, we have $|Q(s,a_i) - Q(s,a_j)| \leq 2\frac{R_{max}}{1-\gamma}, \forall Q$ and $\forall s$.*

Proof. This follows by using the assumption and the results in Lemma A1. □

Proof of Lemma 2. We first sort the sequence $\{Q(s,a_i)\}$ such that $Q(s,a_{[1]}) \geq \ldots \geq Q(s,a_{[m]})$. Then, $\forall Q$ and $\forall s$, we have

$$\begin{aligned}&\max_a Q(s,a) - f_\tau^T\left(Q(s,)\right) Q(s,) \\ &= Q(s,a_{[1]}) - \frac{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right] Q(s,a_{[i]})}{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right]} \\ &= \frac{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right] \left[Q(s,a_{[1]}) - Q(s,a_{[i]})\right]}{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right]}.\end{aligned} \tag{A2}$$

By introducing $\delta_i(s) = Q(s,a_{[1]}) - Q(s,a_{[i]})$, and noting $\delta_i(s) \geq 0$ and $\delta_1(s) = 0$, we can proceed from Eq. (A2) as

$$\begin{aligned}&\frac{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right] \left[Q(s,a_{[1]}) - Q(s,a_{[i]})\right]}{\sum_{i=1}^m \exp\left[\tau Q(s,a_{[i]})\right]} \\ &= \frac{\sum_{i=1}^m \exp[-\tau\delta_i(s)]\, \delta_i(s)}{\sum_{i=1}^m \exp[-\tau\delta_i(s)]} \\ &= \frac{\sum_{i=2}^m \exp[-\tau\delta_i(s)]\, \delta_i(s)}{1 + \sum_{i=2}^m \exp[-\tau\delta_i(s)]}.\end{aligned} \tag{A3}$$

[1]Duke University *Work performed as a graduate student at Duke University; now at Baidu Research. Correspondence to: Zhao Song <zsong@alumni.duke.edu>, Ronald E. Parr <parr@cs.duke.edu>, Lawrence Carin <lcarin@duke.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Now, we can proceed from Eq. (A3) to prove each direction separately as follows.

(i) Upper bound: First note that for any two non-negative sequences $\{x_i\}$ and $\{y_i\}$,

$$\frac{\sum_i x_i}{1+\sum_i y_i} \le \sum_i \frac{x_i}{1+y_i}. \tag{A4}$$

We then apply Eq. (A4) to Eq. (A3) as

$$\begin{aligned}\frac{\sum_{i=2}^m \exp[-\tau\delta_i(s)]\,\delta_i(s)}{1+\sum_{i=2}^m \exp[-\tau\delta_i(s)]} &\le \sum_{i=2}^m \frac{\exp[-\tau\delta_i(s)]\,\delta_i(s)}{1+\exp[-\tau\delta_i(s)]}\\ &= \sum_{i=2}^m \frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]}.\end{aligned} \tag{A5}$$

Next, we bound each term in Eq. (A5), by considering the following two cases:

1) $\delta_i(s) > 1$: $\frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]} \le \frac{\delta_i(s)}{1+\exp(\tau)} \le \frac{2Q_{\max}}{1+\exp(\tau)}$, where we apply Corollary A2 to bound $\delta_i(s)$.

2) $0 \le \delta_i(s) \le 1$: $\frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]} = \frac{1}{\frac{2}{\delta_i(s)}+\tau+0.5\tau^2\delta_i(s)+\cdots} \le \frac{1}{\tau+2}$, where we first expand the denominator using Taylor series for the exponential function.

By combining these two cases with Eq. (A5), we achieve the upper bound.

(ii) Lower bound:

$$\begin{aligned}&\frac{\sum_{i=2}^m \exp[-\tau\delta_i(s)]\,\delta_i(s)}{1+\sum_{i=2}^m \exp[-\tau\delta_i(s)]}\\ \ge\ &\frac{\sum_{i=2}^m \exp[-\tau\delta_i(s)]\,\delta_i(s)}{m}\\ \ge\ &\frac{\sum_{i=2}^m \delta_i(s)}{m\exp[\tau\,\widehat{\delta}(s)]}\\ \ge\ &\frac{\widehat{\delta}(s)}{m\exp[\tau\,\widehat{\delta}(s)]}.\end{aligned} \tag{A6}$$

□

Proof of Theorem 3. We first prove the upper bound by induction as follows.

(i) When $i=1$, we start from the definitions for $\mathcal{T}$ and $\mathcal{T}_{\text{soft}}$ in Eq. (2) and Eq. (3), and proceed as

$$\begin{aligned}&\mathcal{T}Q_0(s,a) - \mathcal{T}_{\text{soft}}\,Q_0(s,a)\\ =&\gamma\sum_{s'} P(s'|s,a)\left[\max_{a'} Q_0(s',a') - f_\tau^T\big(Q_0(s',)\big)Q_0(s',)\right]\\ \ge&0.\end{aligned}$$

(ii) Suppose this claim holds when $i=l$, i.e., $\mathcal{T}^l Q_0(s,a) \ge \mathcal{T}_{\text{soft}}^l\,Q_0(s,a)$. When $i=l+1$, we have

$$\begin{aligned}&\mathcal{T}^{l+1}Q_0(s,a) - \mathcal{T}_{\text{soft}}^{l+1}\,Q_0(s,a)\\ =&\mathcal{T}\mathcal{T}^l Q_0(s,a) - \mathcal{T}_{\text{soft}}\mathcal{T}_{\text{soft}}^l\,Q_0(s,a)\\ \ge&\mathcal{T}\mathcal{T}_{\text{soft}}^l\,Q_0(s,a) - \mathcal{T}_{\text{soft}}\mathcal{T}_{\text{soft}}^l\,Q_0(s,a)\\ \ge&0.\end{aligned}$$

Since Q^* is the fixed point for $\mathcal{T}$, we know $\lim_{k\to\infty}\mathcal{T}^k\,Q_0(s,a) = Q^*(s,a)$. Therefore, $\limsup\limits_{k\to\infty}\mathcal{T}_{\text{soft}}^k\,Q_0(s,a) \le Q^*(s,a)$.

To prove the lower bound, we first conjecture that

$$\mathcal{T}^k\,Q_0(s,a) - \mathcal{T}_{\text{soft}}^k\,Q_0(s,a) \le \sum_{j=1}^k \gamma^j\,\zeta, \tag{A7}$$

where $\zeta = \sup_Q \max_s[\max_a Q(s,a) - f_\tau^T\big(Q(s,)\big)\,Q(s,)]$ denotes the supremum of the difference between the max and softmax operators, over all Q-functions that occur during Q-iteration, and state s. Eq. (A7) is proven using induction as follows.

(i) When $i=1$, we start from the definitions for $\mathcal{T}$ and $\mathcal{T}_{\text{soft}}$ in Eq. (2) and Eq. (3), and proceed as

$$\begin{aligned}&\mathcal{T}Q_0(s,a) - \mathcal{T}_{\text{soft}}\,Q_0(s,a)\\ =&\gamma\sum_{s'} P(s'|s,a)\left[\max_{a'} Q_0(s',a') - f_\tau^T\big(Q_0(s',)\big)Q_0(s',)\right]\\ \le&\gamma\sum_{s'} P(s'|s,a)\,\zeta = \gamma\zeta.\end{aligned}$$

(ii) Suppose the conjecture holds when $i=l$, i.e., $\mathcal{T}^l Q_0(s,a) - \mathcal{T}_{\text{soft}}^l\,Q_0(s,a) \le \sum_{j=1}^l \gamma^j\zeta$, then

$$\begin{aligned}&\mathcal{T}^{l+1}Q_0(s,a) - \mathcal{T}_{\text{soft}}^{l+1}\,Q_0(s,a)\\ =&\mathcal{T}\,\mathcal{T}^l Q_0(s,a) - \mathcal{T}_{\text{soft}}^{l+1}\,Q_0(s,a)\\ \le&\mathcal{T}\left[\,\mathcal{T}_{\text{soft}}^l\,Q_0(s,a) + \sum_{j=1}^l \gamma^j\zeta\,\right] - \mathcal{T}_{\text{soft}}^{l+1}\,Q_0(s,a)\\ =&\sum_{j=1}^l \gamma^{j+1}\zeta + (\mathcal{T}-\mathcal{T}_{\text{soft}})\,\mathcal{T}_{\text{soft}}^l\,Q_0(s,a)\\ \le&\sum_{j=1}^l \gamma^{j+1}\zeta + \gamma\zeta = \sum_{j=1}^{l+1}\gamma^j\zeta,\end{aligned}$$

where the last inequality follows from the definition of ζ. By using the fact that $\lim_{k\to\infty}\mathcal{T}^k\,Q_0(s,a) = Q^*(s,a)$ again and applying Lemma 2 to bound ζ, we finish the proof for Part (I).

To prove part (II), note that as a byproduct of Eq. (A5) in the proof of Lemma 2, Eq. (A7) can be bounded as

$$\mathcal{T}^k Q_0(s,a) - \mathcal{T}_{\text{soft}}^k Q_0(s,a) \le \frac{\gamma(1-\gamma^k)}{1-\gamma} \sum_{i=2}^{m} \frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]}. \quad \text{(A8)}$$

From the definition of $\delta_i(s)$, we know $\delta_m(s) \ge \delta_{m-1}(s) \ge \ldots \ge \delta_2(s) \ge 0$. Furthermore, there must exist an index $i^* \le m$ such that $\delta_i > 0, \forall i^* \le i \le m$ (otherwise the upper bound becomes zero). Subsequently, we can proceed from Eq. (A8) as

$$\begin{aligned}
&\frac{\gamma(1-\gamma^k)}{1-\gamma} \sum_{i=2}^{m} \frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]} \\
=& \frac{\gamma(1-\gamma^k)}{1-\gamma} \sum_{i=i^*}^{m} \frac{\delta_i(s)}{1+\exp[\tau\delta_i(s)]} \\
\le& \frac{\gamma(1-\gamma^k)}{1-\gamma} \sum_{i=i^*}^{m} \frac{\delta_i(s)}{\exp[\tau\delta_i(s)]} \\
\le& \frac{\gamma(1-\gamma^k)}{1-\gamma} \sum_{i=i^*}^{m} \frac{\delta_i}{\exp[\tau\delta_{i^*}(s)]} \\
=& \frac{\gamma(1-\gamma^k)}{1-\gamma} \exp[-\tau\delta_{i^*}(s)] \sum_{i=i^*}^{m} \delta_i(s),
\end{aligned}$$

which implies an exponential convergence rate in terms of τ and hence proves part (II). □

2. Proofs for Overestimation Reduction

Lemma A3. $g_{\mathbf{x}}(\tau) = \frac{\sum_{i=1}^{m}[\exp(\tau x_i)x_i]}{\sum_{i=1}^{m}\exp(\tau x_i)}$ *is a monotonically increasing function for* $\tau \in [0,\infty)$.

Proof. The gradient of $g_{\mathbf{x}}(\tau)$ can be computed as

$$\frac{\partial g_{\mathbf{x}}(\tau)}{\partial \tau} = \Big\{ \big[\sum_{i=1}^{m} \exp(\tau x_i)\, x_i^2\big] \big[\sum_{i=1}^{m} \exp(\tau x_i)\big] - \big[\sum_{i=1}^{m} \exp(\tau x_i)\, x_i\big]^2 \Big\} \Big/ \big[\sum_{j=1}^{m} \exp(\tau x_j)\big]^2 \ge 0,$$

where the last step holds because of the Cauchy-Schwarz inequality. □

The overestimation bias due to the max operator can be observed by plugging assumption $(A2)$ in Theorem 4 into Eq. (2) as

$$\begin{aligned}
&\mathbb{E}\big[\max_a \big(Q_t(s,a)\big) - \max_a \big(Q_*(s,a)\big)\big] \\
=&\mathbb{E}\big[\max_a \big(Q_t(s,a) - V_*(s)\big)\big] \\
=&\mathbb{E}\big[\max_a(\epsilon_a)\big],
\end{aligned}$$

and $\max_a(\epsilon_a)$ is typically positive for a large action set and the noise satisfying a normal distribution, or a uniform distribution with the symmetric support.

Proof of Theorem 4 . First, the overestimation error from $\mathcal{T}_{\text{soft}}$ can be represented as

$$\begin{aligned}
&\mathbb{E}\Big\{ \sum_a \frac{\exp[\tau Q_t(s,a)]}{\sum_{\bar{a}} \exp[\tau Q_t(s,\bar{a})]} Q_t(s,a) - V^*(s) \Big\} \\
=& \mathbb{E}\Big\{ \sum_a \frac{\exp[\tau V^*(s) + \tau\epsilon_a]}{\sum_{\bar{a}} \exp[\tau V^*(s) + \tau\epsilon_{\bar{a}})]} [V^*(s) + \epsilon_a] - V^*(s) \Big\} \\
=& \mathbb{E}\Big\{ \sum_a \frac{\exp[\tau\epsilon_a]}{\sum_{\bar{a}} \exp[\tau\epsilon_{\bar{a}}]} \epsilon_a \Big\} \quad \text{(A9)} \\
\le& \mathbb{E}\big[\max_a(\epsilon_a)\big].
\end{aligned}$$

To prove Part (II), note that the overestimation reduction of $\mathcal{T}_{\text{soft}}$ from $\mathcal{T}$ can then be represented as

$$\begin{aligned}
&\mathbb{E}\big[\max_a(\epsilon_a) - \sum_a \frac{\exp[\tau\epsilon_a]}{\sum_{\bar{a}} \exp[\tau\epsilon_{\bar{a}}]} \epsilon_a\big] \\
=&\mathbb{E}\Big\{ \max_a \big[\epsilon_a + V^*(s)\big] - \sum_a \frac{\exp[\tau\epsilon_a]}{\sum_{\bar{a}} \exp[\tau\epsilon_{\bar{a}}]} \big[\epsilon_a + V^*(s)\big] \Big\} \\
=&\mathbb{E}\Big\{ \max_a \big[Q_t(s,a)\big] - \sum_a \frac{\exp[\tau\epsilon_a]}{\sum_{\bar{a}} \exp[\tau\epsilon_{\bar{a}}]} \big[Q_t(s,a)\big] \Big\} \\
=&\mathbb{E}\Big\{ \max_a \big[Q_t(s,a)\big] - \sum_a \frac{\exp[\tau\epsilon_a + \tau V^*(s)]}{\sum_{\bar{a}} \exp[\tau\epsilon_{\bar{a}} + \tau V^*(s)]} \\
&\times \big[Q_t(s,a)\big] \Big\} \\
=&\mathbb{E}\Big\{ \max_a \big[Q_t(s,a)\big] - \sum_a \frac{\exp[\tau Q_t(s,a)]}{\sum_{\bar{a}} \exp[\tau Q_t(s,\bar{a})]} \big[Q_t(s,a)\big] \Big\}.
\end{aligned}$$

Subsequently, we can employ Lemma 2 to obtain the range.

Finally, the monotonicity for the overestimation error in terms of τ follows, by noting the term inside the expectation of Eq. (A9) can be represented as $g_\epsilon(\tau)$, which is a monotonic function of τ, according to Lemma A3. □

3. Additional Plots and Setups

Figures A1 and A2 are the full version of the corresponding figures in the main text, by plotting all six games. The corresponding values for τ in S-DQN and S-DDQN are provided in Table A1.

Figures A3, A4, and A5 show the scores, Q-values, and gradient norm, for different values of τ, for S-DDQN.

Table A1. Values of τ used for S-DQN and S-DDQN in Figures A1 and A2.

	Q	M	B	C	A	S
S-DQN	1	1	5	1	5	5
S-DDQN	1	5	5	5	5	10

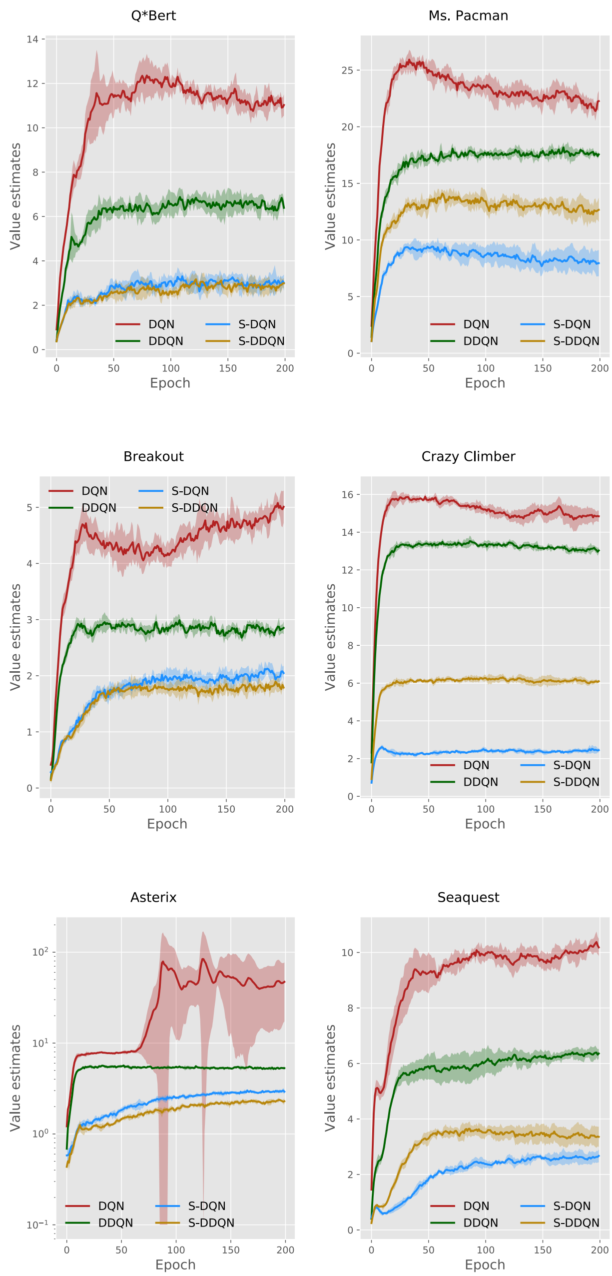


Figure A1. Mean and one standard deviation of the estimated Q-values on the Atari games, for different methods.

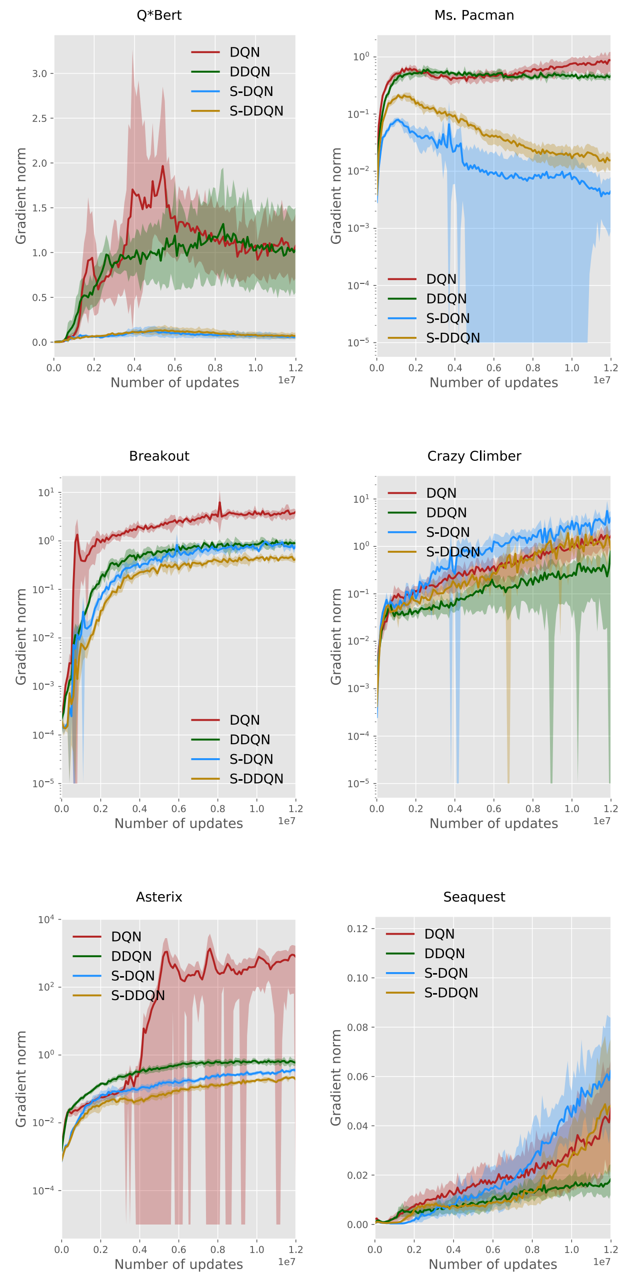
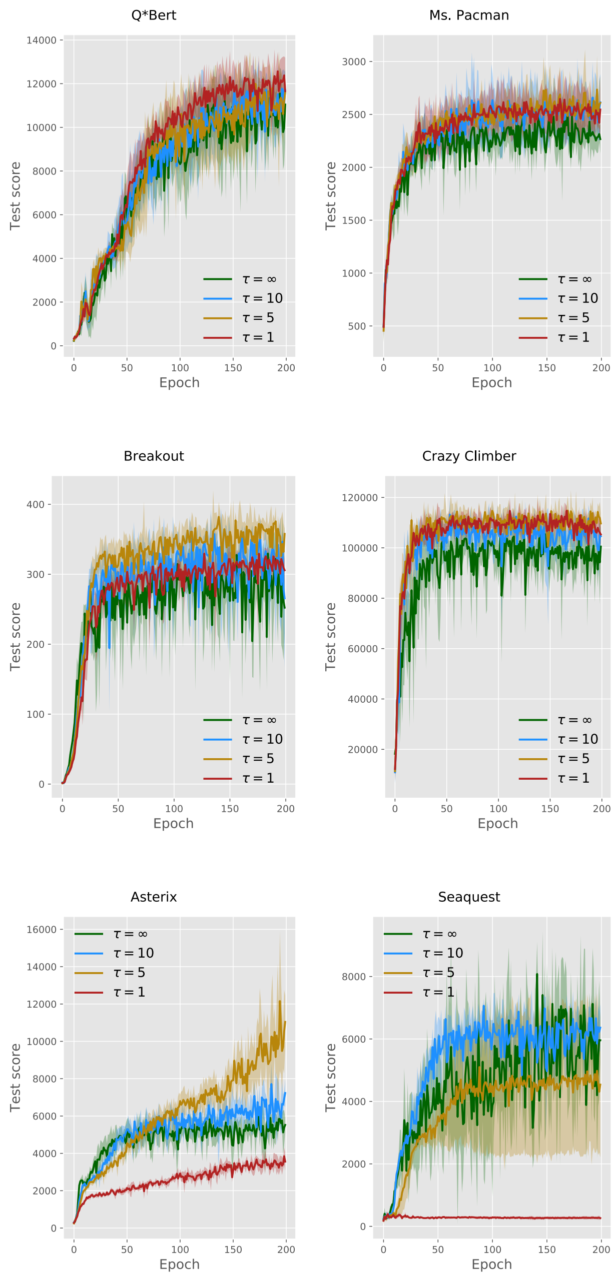


Figure A2. Mean and one standard deviation of the gradient norm on the Atari games, for different methods.



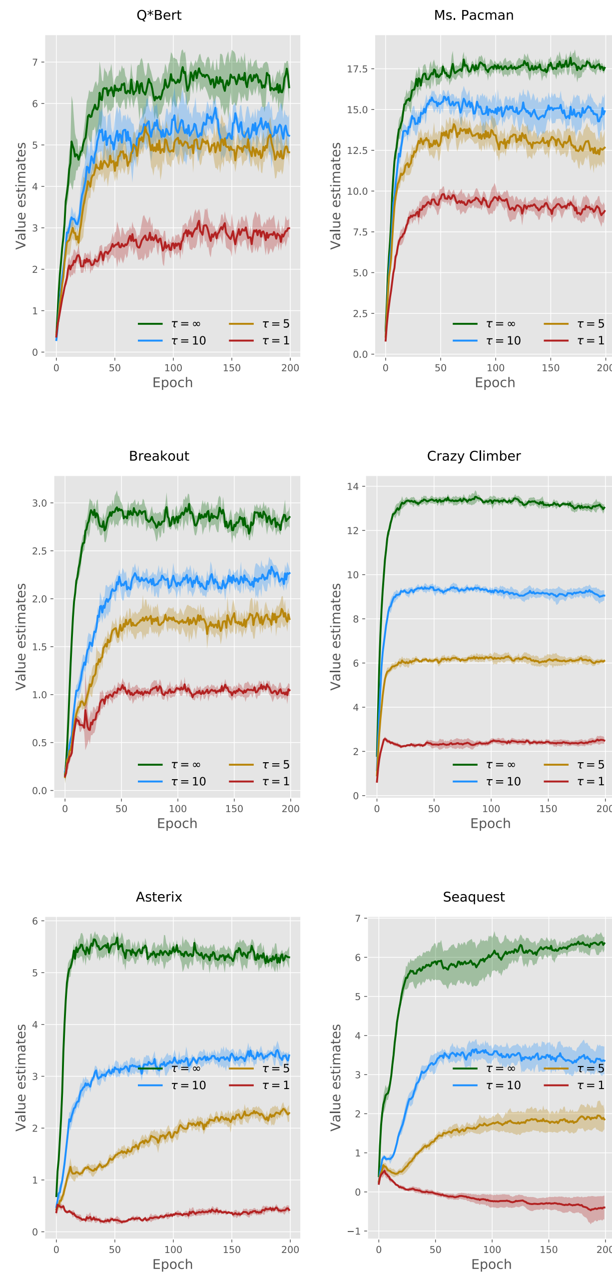


Figure A3. Mean and one standard deviation of test scores on the Atari games, for different values of τ in S-DDQN.

Figure A4. Mean and one standard deviation of the estimated Q-values on the Atari games, for different values of τ in S-DDQN.

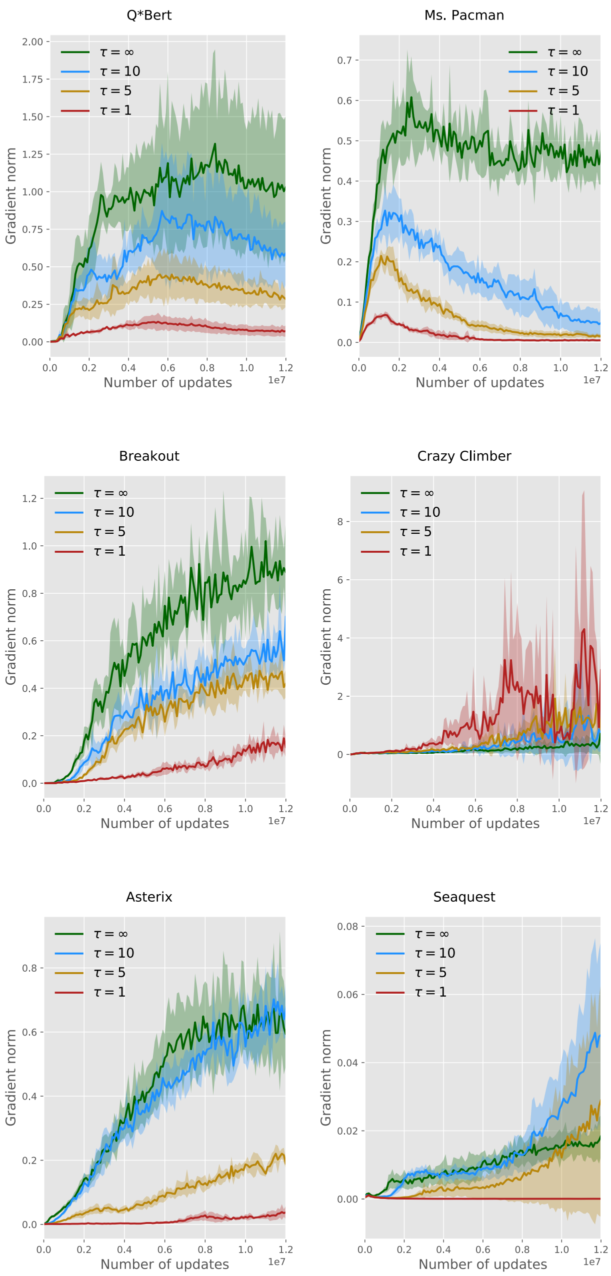


Figure A5. Mean and one standard deviation of the gradient norm on the Atari games, for different values of τ in S-DDQN.