Revisiting the Softmax Bellman Operator: New Benefits and New Perspective

Zhao Song, Ronald Parr, and Lawrence Carin

June 11, 2019

Bellman Operators

- The Bellman equation

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q^*(s',a').$$

  - The optimal policy should be greedy w.r.t. actions

- The Bellman operator

$$\mathcal{T} Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$$

  - Widely used in reinforcement learning, e.g., deep Q-networks [Mnih et al., 2015, Nature]
  - A contraction

Bellman Operators (Cont.)

- The mellowmax operator [Asadi and Littman, 2017, ICML]

$$\max_{a'} Q(s', a') \quad \longrightarrow \quad \frac{\log\left(\frac{1}{m}\sum_{a'} \exp[\omega\, Q(s', a')]\right)}{\omega}$$

  - The `log-sum-exp` function has been extensively used [Todorov, 2007, Fox et al., 2016, Schulman et al., 2017, Haarnoja et al., 2017, Neu et al., 2017, Nachum et al., 2017]
  - A contraction

Bellman Operators (Cont.)

- The softmax operator

$$\max_{a'} Q(s', a') \longrightarrow \sum_{a'} \frac{\exp[\tau Q(s', a')]}{\sum_{\bar{a}} \exp[\tau Q(s', \bar{a})]} Q(s', a')$$

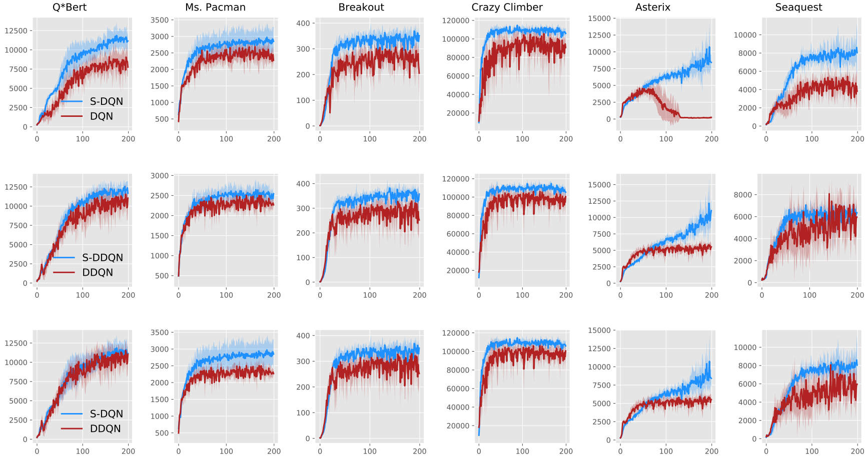
  - Can be a non-contraction [Littman, 1996]
  - Mainly used for Boltzmann exploration [Sutton and Barto, 1998]

Question: Is softmax really as bad as sour milk (credits go to Ron Parr), when updating the value function?

Experiments setup

- S-D(D)QN: Replacing the max function in the target network of D(D)QN with the softmax function
- All other steps same as [Mnih et al., 2015, Nature]
- Code available at `https://github.com/zhao-song/Softmax-DQN`, built upon Nathan Sprague's implementation.

Results on Atari games



A revisit of the softmax Bellman operator is warranted!

Main theorem

Definition: $\widehat{\delta}(s) \triangleq \sup_{Q} \max_{i,j} |Q(s,a_i) - Q(s,a_j)|$

Assumption: $\widehat{\delta}(s) > 0$

Performance bound: Assuming $\widehat{\delta}(s) > 0$, then $\forall (s,a)$,

- $\limsup_{k\to\infty} \mathcal{T}_{\text{soft}}^k Q_0(s,a) \leq Q^*(s,a)$ and
- $\liminf_{k\to\infty} \mathcal{T}_{\text{soft}}^k Q_0(s,a) \geq Q^*(s,a) - \frac{\gamma(m-1)}{(1-\gamma)} \max\left\{\frac{1}{\tau+2}, \frac{2Q_{\text{max}}}{1+\exp(\tau)}\right\}$

Convergence Rate: $\mathcal{T}_{\text{soft}}$ converges to $\mathcal{T}$ with an exponential rate, in terms of τ, i.e., the upper bound of $\mathcal{T}^k Q_0 - \mathcal{T}_{\text{soft}}^k Q_0$ decays exponentially fast, as a function of τ, the inverse temperature parameter.

Overestimation bias reduction

Assumptions: Same as DDQN [van Hasselt et al., 2016]

Smaller Error: The overestimation errors from $\mathcal{T}_{\mathsf{soft}}$ are smaller or equal to those of $\mathcal{T}$ using the max operator, for any $\tau \geq 0$;

Reduction Bound: The overestimation reduction by using $\mathcal{T}_{\mathsf{soft}}$ in lieu of $\mathcal{T}$ is within $[\frac{\widehat{\delta}(s)}{m \exp[\tau \widehat{\delta}(s)]}, (m-1)\max\{\frac{1}{\tau+2}, \frac{2Q_{\mathsf{max}}}{1+\exp(\tau)}\}]$;

Monotonicity: The overestimation error for $\mathcal{T}_{\mathsf{soft}}$ monotonically increases w.r.t. $\tau \in [0, \infty)$.

Simulated example

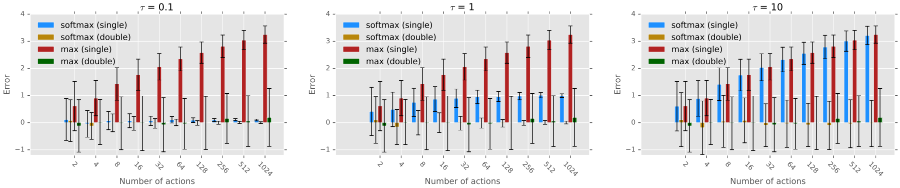
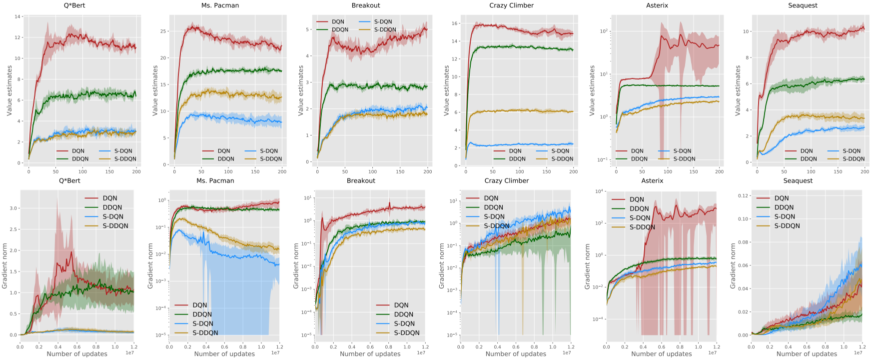


Figure 1: The mean and one standard deviation for the overestimation error for different values of τ.

Q-value and gradient norm for different methods



Comparison among different Bellman operators

Table 1: A comparison of different Bellman operators (B.O. Bellman optimality; O.R. overestimation reduction; P.R. policy representation; D.Q. double Q-learning).

	B.O.	Tuning	O.R.	P.R.	D.Q.
Max	Yes	No	-	Yes	Yes
Mellowmax	No	Yes	Yes	No	No
Softmax	No	Yes	Yes	Yes	Yes

Thank You

Asadi, K. and Littman, M. L. (2017).
An alternative softmax operator for reinforcement learning.
In *International Conference on Machine Learning*, pages 243–252.

Fox, R., Pakman, A., and Tishby, N. (2016).
Taming the noise in reinforcement learning via soft updates.
In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 202–211. AUAI Press.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017).
Reinforcement learning with deep energy-based policies.
In *International Conference on Machine Learning*, pages 1352–1361.

Littman, M. L. (1996).
Algorithms for Sequential Decision Making.
PhD thesis, Department of Computer Science, Brown University.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015).

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

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017).
Bridging the gap between value and policy based reinforcement learning.
In *Advances in Neural Information Processing Systems*, pages 2772–2782.

Neu, G., Jonsson, A., and Gómez, V. (2017).
A unified view of entropy-regularized Markov decision processes.
arXiv preprint arXiv:1705.07798.

Schulman, J., Chen, X., and Abbeel, P. (2017).
Equivalence between policy gradients and soft Q-learning.
arXiv preprint arXiv:1704.06440.

Sutton, R. S. and Barto, A. G. (1998).
Reinforcement Learning: An Introduction.
The MIT Press.

Todorov, E. (2007).

Linearly-solvable Markov decision problems.
In *Advances in neural information processing systems*, pages 1369–1376.

van Hasselt, H., Guez, A., and Silver, D. (2016).
Deep reinforcement learning with double Q-learning.
In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2094–2100. AAAI Press.