October 9, 2020

Errata
On the Convergence of Optimistic Policy Iteration
J. N. Tsitsiklis
Journal of Machine Learning Research, Vol. 3, July 2002, pp. 59-72.

Proof of Proposition 1. The proof involves a certain time $t(\epsilon)$ that was defined (near the end of p. 64) in terms of the entire history of the process J_t. Thus, the initialization of the sequence Z_t, defined in Eq. (3), depends on the future. For this reason, stochastic approximation results cannot be used directly to establish the convergence of Z_t. This difficulty can be bypassed using the argument that follows.

For every positive integer N, we define a sequence of random variables Z_t^N, $t \geq N$, by setting $Z_N^N = J_N$ and

$$Z_{t+1}^N = (1-\gamma_t)Z_t^N + \gamma_t T Z_t^N + \gamma_t \frac{\epsilon\alpha}{1-\alpha} e + \gamma_t w_t, \qquad \forall\ t \geq N.$$

Let A_N be the event that $t(\epsilon) = N$. An easy inductive argument shows that for any N, for any sample path in A_N, and for all $t \geq N$, we have $J_t \leq Z_t^N$.

In contrast to the process Z_t, each of the processes Z_t^N *does* satisfy the standard stochastic approximation assumptions, and the argument in the paper shows that Z_t^N converges to Z_δ^*, almost surely.

Now, for sample paths in A_N, we have

$$\limsup_{t\to\infty} J_t \leq \limsup_{t\to\infty} Z_t^N = Z_\delta^*.$$

Since the union of the events A_N is the entire sample space, we conclude that the inequality

$$\limsup_{t\to\infty} J_t \leq Z_\delta^*$$

holds for (almost) all sample paths.

Proof of Proposition 3. The end of the proof of Proposition 3 states that "... the rest of the proof is identical to the last part of the proof of Prop. 1." This is indeed the case, up to and including the point where the inequality $\limsup_{t\to\infty} J_t \leq J^*$ is established. However, the reverse inequality requires a different argument. (The reason is that the inequality $J_{t+1} \geq (1-\gamma_t)J_t + \gamma_t J^* + \gamma_t w_t$ does not hold for the case of TD(λ).)

The argument goes as follows. Recall that in the proof of Proposition 3 it is shown that $\limsup_t X_t \leq 0$, where $X_t = TJ_t - J_t$. We focus on a single sample path of the process. Let us fix some $\epsilon > 0$. It follows that, for all t large enough, we will have $TJ_t \leq J_t + \epsilon e$. (Recall that e is the vector of all ones.) Let us fix

such a time t. By applying the operator T to both sides of the earlier inequality, we have

$$T^2 J_t \leq T J_t + \alpha\epsilon \leq J_t + \epsilon e + \epsilon\alpha e.$$

By repeatedly applying T and then taking the limit, we obtain

$$J^* = \lim_{k\to\infty} T^k J_t \leq J_t + \frac{\epsilon}{1-\alpha} e,$$

Since this is true for every large enough t, we have

$$J^* \leq \liminf_{t\to\infty} J_t + \frac{\epsilon}{1-\alpha} e.$$

Since ϵ can be taken arbitrarily small, we conclude that

$$J^* \leq \liminf_{t\to\infty} J_t.$$

Acknowledgments. The author is grateful to R. Srikant and Anna Winnicki for pointing out the gap in the proof of Proposition 1, and to Bruno Scherrer and Yuanlong Chen for pointing out the gap in the proof of Proposition 3.