Taming the Noise in Reinforcement Learning via Soft Updates — Supplementary Material —

Roy Fox* Hebrew University

Ari Pakman* Columbia University

Naftali Tishby Hebrew University

CONVERGENCE OF G-LEARNING

In this section we prove the convergence of G to the optimal G^*, with probability 1, under the G-learning update rule

$$G(s_t,a_t) \leftarrow (1-\alpha_t)G(s_t,a_t) + \alpha_t\left(c_t - \tfrac{\gamma}{\beta}\log\left(\sum_{a'}\rho(a'|s_{t+1})e^{-\beta G(s_{t+1},a')}\right)\right). \tag{1}$$

Recall that the supremum norm is defined as $|x|_\infty = \max_i |x_i|$, and that the optimal G function satisfies

$$G^*(s,a) = \mathrm{E}_\theta[c|s,a] - \tfrac{\gamma}{\beta}\,\mathrm{E}_p\left[\log\sum_{a'}\rho(a'|s')e^{-\beta G^*(s',a')}\right] \tag{2}$$

$$\equiv \mathbf{B}^*[G^*]_{(s,a)}. \tag{3}$$

The convergence proof relies on the following Lemma.

Lemma 1. *The operator $\mathbf{B}^*[G]_{(s,a)}$ defined in (3) is a contraction in the supremum norm.*

Proof. Let us define

$$\mathbf{B}^\pi[G]_{(s,a)} = k^\pi(s,a) + \gamma\sum_{s',a'}p(s'|s,a)\pi(a'|s')G(s',a'), \tag{4}$$

where

$$k^\pi(s,a) = \mathrm{E}_\theta[c|s,a] + \tfrac{\gamma}{\beta}\sum_{s',a'}p(s'|s,a)\pi(a'|s')\log\tfrac{\pi(a'|s')}{\rho(a'|s')}. \tag{5}$$

Now, for any policy π, the operator (4) is a contraction under the supremum norm [1], i.e. for any G_1 and G_2

$$|\mathbf{B}^\pi[G_1] - \mathbf{B}^\pi[G_2]|_\infty \le \gamma|G_1 - G_2|_\infty. \tag{6}$$

Also note that

$$\mathbf{B}^*[G_i]_{(s,a)} = \min_\pi \mathbf{B}^\pi[G_i]_{(s,a)}, \tag{7}$$

and that the optimum is achieved for

$$\pi_{G_i}(a|s) = \frac{\rho(a|s)e^{-\beta G_i(s,a)}}{\sum_{a'}\rho(a'|s)e^{-\beta G_i(s,a')}}. \tag{8}$$

The Lemma now follows from

$$\begin{aligned}\big|\mathbf{B}^*[G_1] - \mathbf{B}^*[G_2]\big|_\infty &= \max_{(s,a)}\big|\mathbf{B}^*[G_1]_{(s,a)} - \mathbf{B}^*[G_2]_{(s,a)}\big| \\ &= \max_{(s,a)}\big|\mathbf{B}^{\pi_{G_1}}[G_1]_{(s,a)} - \mathbf{B}^{\pi_{G_2}}[G_2]_{(s,a)}\big|\end{aligned} \tag{9}$$

(choose $i = \arg\min \mathbf{B}^{\pi_{G_i}}[G_i]_{(s,a)}$)

$$\begin{aligned}&\le \max_{(s,a)}\max_{i=1,2}\big|\mathbf{B}^{\pi_{G_i}}[G_1]_{(s,a)} - \mathbf{B}^{\pi_{G_i}}[G_2]_{(s,a)}\big| \\ &= \max_{i=1,2}\big|\mathbf{B}^{\pi_{G_i}}[G_1] - \mathbf{B}^{\pi_{G_i}}[G_2]\big|_\infty \\ &\le \gamma\big|G_1 - G_2\big|_\infty.\end{aligned}$$

□

The update equation (1) of the algorithm can be written as a stochastic iteration equation

$$G_{t+1}(s_t,a_t) = (1-\alpha_t)G_t(s_t,a_t) + \alpha_t(\mathbf{B}^*[G_t]_{(s_t,a_t)} + z_t(c_t,s_{t+1})) \tag{10}$$

where the random variable z_t is

$$z_t(c_t,s_{t+1}) \equiv -\mathbf{B}^*[G_t]_{(s_t,a_t)} + c_t - \tfrac{\gamma}{\beta}\log\sum_{a'}\rho(a'|s_{t+1})e^{-\beta G_t(s_{t+1},a')}. \tag{11}$$

Note that z_t has expectation 0. Many results exist for iterative equations of the type (10). In particular, given conditions

$$\sum_t \alpha_t = \infty; \qquad \sum_t \alpha_t^2 < \infty, \tag{12}$$

the contractive nature of $\mathbf{B}^*$, infinite visits to each pair (s_t,a_t) and assuming that $|z_t| < \infty$, G_t is guaranteed to converge to the optimal G^* with probability 1 [1, 2].

*These authors contributed equally to this work.

References

[1] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1,2. Athena Scientific Belmont, MA, 1995.

[2] Vivek S Borkar. Stochastic approximation. *Cambridge Books*, 2008.