From Reflexes to Actor-Critic

May 16, 2015

We have developed a robot system that avoids collisions using a fixed stochastic policy with a given representation. The representation x has four components indexed from 0, (0) a bias unit, (1) a learned policy-contingent prediction (GVF) of the bumper, (2) the previous action was forward, (3) the previous action was to turn counterclockwise.

The policy is originally crafted by formalizing the real-world characteristics desired from the behavior. We have two action forward (F) and turn (T). The prediction $p0$ anticipates when the robot bumper will be pressed, presumably from a collision with a wall. The ideal value of $p0$ is 1 when the bumper will be pressed on the next timestep if the robot drives forward, and it is 0 when the robot is infiinitely far away from a wall. The robot interacts with the world $H = 100/3$ times per second. The prediction $p0$ has a gamma of 0 when the bumper is pressed, and a gamma of 0.95 otherwise. Thus in the absence of any bumper presses, the prediction extends for $20(= (1-\gamma)^{-1})$ timesteps in expectation (i.e. 2/3 of a second). We want the robot to have an equal probability selecting a drive action or a turn action when the ideal prediction has a value of $eq = 0.5$ (what is the timescale? $0.95^{timescale} = 0.5$).

Now as we don't want to be limited by the decision rate of the robot, we instead specify the desired number of switches per second if the prediction was at this equilibrium point, and we set this value to be 5. Let TT= decisions-per-second/ expected-number-of-switches-per-second. After some algebraic manipulation, we find that there is a constant $K_s = log_e((TT-1)(number-of-actions-1)$ which would govern switching at equilibirum in the absence of more information from the prediction.

There is one more constant used to scale the relative influence of the prediction on the action selection, we chose the number 4 (so p0=.75 means turning is e times more likely than going forward).

We can write the preferences and the policy in an intuitive form.

$$pref[F] = -4K_s(p0 - eq) + K_s 1(LastAction == F)$$
$$pref[T] = K_s 1(LastAction == T)$$
$$\pi(a = F|x) = \exp(pref[F])/(\sum_{i=F,T} \exp(pref[i]))$$
$$\pi(a = T|x) = \exp(pref[T])/(\sum_{i=F,T} \exp(pref[i]))$$

We rewrite the stochastic policy in a canonical quasi-linear form, with action selection probabilities given by exponentiated preferences for each action, where each preference is a linear function of the feature vector x, and the policy parameters u are initialized to the constants from above.

$$pref[F] = u_0x_0 + u_1x_1 + u_2x_2$$
$$pref[T] = u_2x_3$$
$$\pi(a = F|x) = \exp(pref[F])/(\sum_{i=F,T} \exp(pref[i]))$$
$$\pi(a = T|x) = \exp(pref[T])/(\sum_{i=F,T} \exp(pref[i]))$$

Now an average-reward actor-critic algorithm can be used to tune the parameters u so as to maximize a given reward signal over a long term, but it first requires us to calculate gradients, namely ${}_u\pi(a = k|x)/\pi(a = k|x)$. Given the above equations, the gradients can be found through the simple, tedious calculations shown below.

$$\frac{\nabla_u \pi}{\pi} = ((\partial_{u_0}\pi)/\pi, (\partial_{u_1}\pi)/\pi, (\partial_{u_2}\pi)/\pi)$$

To compute these, it is useful to precompute $\partial_{u_j} \exp(pref[i])$ for all i and j.

$$\partial_{u_0} \exp(pref[F]) = \partial_{u_0} \exp(u_0x_0+u_1x_1+u_2x_2) = \exp(u_0x_0+u_1x_1+u_2x_2)\partial_{u_0}(u_0x_0+u_1x_1+u_2x_2) = \exp(pref[F])x_0$$

$$\partial_{u_1} \exp(pref[F]) = \exp(pref[F])x_1$$
$$\partial_{u_2} \exp(pref[F]) = \exp(pref[F])x_2$$

$$\partial_{u_0} \exp(pref[T]) = 0$$
$$\partial_{u_1} \exp(pref[T]) = 0$$
$$\partial_{u_2} \exp(pref[T]) = \exp(pref[T])x_3$$

Let $W = \sum_i \exp(pref(i))$.

$$\partial_{u_0} W = \exp(pref(F))x_0$$
$$\partial_{u_1} W = \exp(pref(F))x_1$$
$$\partial_{u_2} W = \exp(pref[F])x_2 + \exp(pref(T)x_3$$

Now we can write the relevant gradients directly, recalling that $\pi(a|x) = \exp(pref(a))/W$.

$$\begin{aligned}
\frac{\partial_{u_0}\pi(F|x)}{\pi(F|x)} &= \frac{W}{\exp(pref(F))}\partial_{u_0}\frac{\exp(pref(F))}{W} \\
&= \frac{W}{\exp(pref(F))}\frac{1}{W^2}((\partial_{u_0}\exp(pref(F)))W - \exp(pref(F))\partial_{u_0}W) \\
&= \frac{W}{\exp(pref(F))}\frac{1}{W^2}(\exp(pref(F))x_0W - \exp(pref(F))\exp(pref(F))x_0) \\
&= \frac{1}{1}\frac{1}{W}(x_0W - \exp(pref(F))x_0) \\
&= x_0(1 - \frac{\exp(pref(F))}{W}) \\
&= x_0(1 - \pi(F|x))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial_{u_1}\pi(F|x)}{\pi(F|x)} &= \frac{W}{\exp(pref(F))}\partial_{u_1}\frac{\exp(pref(F))}{W} \\
&= \frac{W}{\exp(pref(F))}\frac{1}{W^2}((\partial_{u_1}\exp(pref(F)))W - \exp(pref(F))\partial_{u_1}W) \\
&= \frac{W}{\exp(pref(F))}\frac{1}{W^2}(\exp(pref(F))x_1W - \exp(pref(F))\exp(pref(F))x_1) \\
&= x_1(1 - \pi(F|x))
\end{aligned}$$

$$\frac{\partial_{u_2}\pi(F|x)}{\pi(F|x)} = x_2(1-\pi(F|x))$$

$$\begin{aligned}
\frac{\partial_{u_0}\pi(T|x)}{\pi(T|x)} &= \frac{W}{\exp(pref(T))}\partial_{u_0}\frac{\exp(pref(T))}{W} \\
&= \frac{W}{\exp(pref(T))}\frac{1}{W^2}((\partial_{u_0}\exp(pref(T)))W - \exp(pref(T))\partial_{u_0}W) \\
&= \frac{1}{\exp(pref(T))}\frac{1}{W}(0 - \exp(pref(T))\exp(pref(F))x_0) \\
&= -\frac{\exp(pref(F))}{W}x_0 \\
&= -\pi(F|x)x_0
\end{aligned}$$

$$\frac{\partial_{u_1}\pi(T|x)}{\pi(T|x)} = -\pi(F|x)x_1$$

$$\begin{aligned}
\frac{\partial_{u_2}\pi(T|x)}{\pi(T|x)} &= \frac{W}{\exp(pref(T))}\partial_{u_2}\frac{\exp(pref(T))}{W} \\
&= \frac{W}{\exp(pref(T))}\frac{1}{W^2}((\partial_{u_2}\exp(pref(T)))W - \exp(pref(T))\partial_{u_2}W) \\
&= \frac{1}{\exp(pref(T))}\frac{1}{W}((\exp((T))x_3)W - \exp(pref(T))(\exp((F))x_2 + exp(pref(T))x_3)) \\
&= \frac{1}{1}\frac{1}{W}(x_3W - (\exp((F))x_2 + exp(pref(T))x_3)) \\
&= (x_3 - \pi(F)x_2 - \pi(T)x_3)) \\
&= (1-\pi(T))x_3 - \pi(F)x_2
\end{aligned}$$

These gradient terms are used in the actor-critic algorithm, evaluated for a given action (F or T) and the feature vector x.

The average reward formulation is given in the Actor-Critic algorithms in practice paper, and it is listed below.

Average Reward Actor-Critic Algorithm

For each step with feature vector $x(S)$

Choose A according to $\pi_u(\cdot|S)$

Take action A

Observe $x(S')$, R

$\delta \leftarrow R - \bar{R} + v^\top x(S') - v^\top x(S)$

$\bar{R} \leftarrow \alpha_1 \delta$

$e_v \leftarrow \lambda e_v + x(S)$

$v \leftarrow v + \alpha_2 \delta e_v$

$e_u \leftarrow \lambda e_u + \frac{\nabla_u \pi(A|x(S))}{\pi(A|x(S))}$

$u \leftarrow u + \alpha_3 \delta e_u$