ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

**Unsupervised and reinforcement learning in neural networks**
Week 9

Lecturer: Marc-Oliver Gewaltig

Assistants: Dane Corneil
Berat Denizdurduran

# Corrections: RL in Continuous States

## Exercise 1 (in class): Q-values for Continuous States

Using the definition of $Q(s, a)$ given, we find the gradient:

$$\frac{dQ(s,a)}{dw_{aj}} = \Phi(s - s_j)$$

i.e., the direction of the gradient is the basis function itself.

## Exercise 2: Eligibility trace

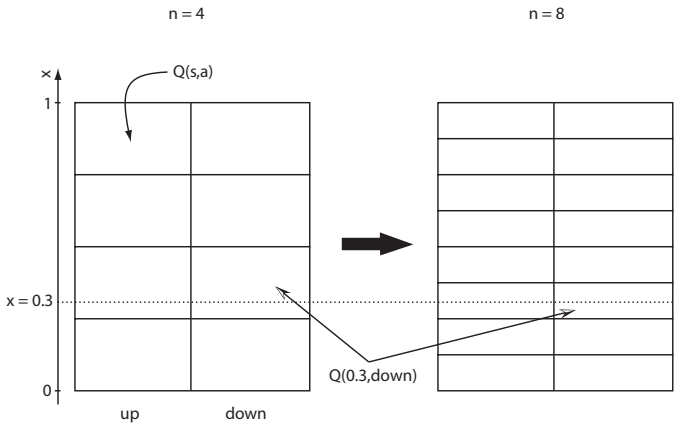
The table below shows the evolution of the $Q$ values for each relevant state action pair during the first 2 trials, starting at the first step when there is a non-zero update $\Delta = \eta\,[r - (Q(s^*, a^*) - Q(s', a'))]$. We assume that the agent goes straight down in the first trial, that it always picks the best action, and that the eligibility traces are reset when the agent picks the reward, and the agent is put back to the starting position.

| trial | transition | | $(s, a_1)$ | $(s', a_1)$ | $(s'', a_1)$ | $\Delta$ |
|---|---|---|---|---|---|---|
| 1 | $s'' \to s'''$ | $Q$ | $0$ | $0$ | $0$ | $\eta$ |
| | | $e$ | $\lambda^2$ | $\lambda$ | $1$ | – |
| 2 | $s \to s'$ | $Q$ | $\eta\lambda^2$ | $\eta\lambda$ | $\eta$ | $\eta^2(\lambda - \lambda^2)$ |
| | | $e$ | $1$ | $0$ | $0$ | – |
| 2 | $s' \to s''$ | $Q$ | $\eta\lambda^2 + \eta^2(\lambda - \lambda^2)$ | $\eta\lambda$ | $\eta$ | $\eta^2(1 - \lambda)$ |
| | | $e$ | $\lambda$ | $1$ | $0$ | – |
| 2 | $s'' \to s'''$ | $Q$ | $\eta\lambda^2 + 2\eta^2\lambda - 2\eta^2\lambda^2$ | $\eta\lambda + \eta^2(1 - \lambda)$ | $\eta$ | $\eta - \eta^2$ |
| | | $e$ | $\lambda^2$ | $\lambda$ | $1$ | – |
| 3 | $s \to s'$ | $Q$ | $2\eta\lambda^2 + 2\eta^2\lambda - 3\eta^2\lambda^2$ | $2\eta\lambda + \eta^2 - 2\eta^2\lambda$ | $2\eta - \eta^2$ | $\cdots$ |
| | | $e$ | $1$ | $0$ | $0$ | – |

...

Although the $Q$-values for $s''$ are the same as without the eligibility trace (see exercise from last week), the $Q$-values for $s$ and $s'$ already start to approach their asymptotical value (i. e. 1) in the first trial.

## Exercise 3: Eligibility Trace for Continuous States




*How should we rescale the parameter $\Delta t$, so that the speed $v = \Delta x/\Delta t$ remains constant?*

If $\Delta x/\Delta t$ has to remain constant, then $\Delta t$ should vary like $\Delta x$, i.e., $\Delta t \propto 1/n$.

*How should we rescale $\lambda$, in order that the "speed of information propagation" in SARSA($\lambda$) remains constant?*

A point "sitting" at $x$ is $x/\Delta x = x \cdot n$ steps away from the reward (assuming we always choose the "down" action). As we have seen in exercise 2, on the first trial, the $Q$-value of a state $d$ steps from the trial is updated proportional to $\lambda^d$. Thus, if we want the update to stay constant under rescaling, we need

$$\Delta Q(s,a) \propto \lambda^d = \lambda^{x \cdot n} = cst$$

This holds if we replace $\lambda$ by a "rescaled" $\tilde{\lambda} = \lambda^{\frac{1}{n}}$.

## Exercise 4: Gradient-Based Learning of the Q-values

**4.1.** Let's start by computing the derivative of $Q(s,a)$ with respect to $w_{\tilde{k}}^{\tilde{a}}$ (we'll use this later):

$$\frac{\mathrm{d}Q(s,a)}{\mathrm{d}w_{\tilde{k}}^{\tilde{a}}} = \delta_{a\tilde{a}}\, \Phi(s - s_{\tilde{k}})\,,$$

where $\delta_{a\tilde{a}}$ is the Kroneker, i.e., it is 1 if $a = \tilde{a}$, and 0 otherwise (not to be confused with $\delta_t$).
We then compute the gradient, i.e., the derivative of $E_t$ with respect to $w_{\tilde{k}}^{\tilde{a}}$, using the chain rule a few times and the result above:

$$\begin{aligned}\frac{\mathrm{d}E_t}{\mathrm{d}w_{\tilde{k}}^{\tilde{a}}} &= \delta_t \left[\gamma \frac{\mathrm{d}Q(s',a')}{\mathrm{d}w_{\tilde{k}}^{\tilde{a}}} - \frac{\mathrm{d}Q(s,a)}{\mathrm{d}w_{\tilde{k}}^{\tilde{a}}}\right] \\ &= \delta_t \left[\gamma \delta_{a'\tilde{a}}\, \Phi(s' - s_{\tilde{k}}) - \delta_{a\tilde{a}}\, \Phi(s - s_{\tilde{k}})\right].\end{aligned}$$

To turn this into a learning rule, we have to move the weights in the direction that *minimizes* the error, i.e.

$$\Delta w_{\tilde{k}}^{\tilde{a}} = -\eta \frac{\mathrm{d}E_t}{\mathrm{d}w_{\tilde{k}}^{\tilde{a}}} = \eta\, \delta_t \, \left[\delta_{a\tilde{a}}\, \Phi(s - s_{\tilde{k}}) - \gamma \delta_{a'\tilde{a}}\, \Phi(s' - s_{\tilde{k}})\right]. \tag{1}$$

In the case where $a = a'$ (i.e., the action taken is the same in the two consecutive steps):

$$\Delta w_{\tilde{k}}^{\tilde{a}} = \eta\, \delta_t \, \left(\Phi(s - s_{\tilde{k}}) - \gamma \Phi(s' - s_{\tilde{k}})\right) \delta_{a\tilde{a}}. \tag{2}$$

**4.2.** If we further assume that $s \simeq s'$, the term in Eq. (2), when $a = a'$, is a (modulated) Hebbian rule: a change of the weights happens if the postsynaptic neuron is active (action $a$ was taken and $a = \tilde{a}$) and is proportional to the activation $\Phi(s - s_{\tilde{k}})$ of the presynaptic neuron. The term $\delta_t$ modulates the sign and amplitude of the weight change. Without these assumption however, the rule in Eq. (1) does not have an elegant interpretation in neural terms.

**4.3.** Since we have continuous states, the difference between two consecutive states $s$ and $s'$ is $s' - s = v(a) \cdot \Delta t$, where $v(a)$ is the velocity (in the state space) resulting from the choice of the action $a$. $v(a)$ does not depend on the choice of $\Delta t$, and so, as , $\lim_{\Delta t \to 0} s' - s = 0$. So, as the time steps decrease, the approximation $s \simeq s'$ becomes more accurate. Similarly, if we assume that the agent follows smooth trajectories, the finer the time resolution becomes, the likelier it is that $a \simeq a'$. Thus, for small time steps, an online gradient descent on the error term $E_t$ becomes closer to a $\delta_t$-modulated Hebbian rule.