A Reinforcement Learning Theory for Homeostatic Regulation: Supplementary Methods

Mehdi Keramati
Group for Neural Theory, LNC, ENS
Paris, France
mohammadmahdi.keramati@ens.fr

Boris Gutkin
Group for Neural Theory, LNC, ENS
Paris, France
boris.gutkin@ens.fr

1 Performance measure of instrumental responses

Definition 1.1: A "homeostatic pathway", denoted by p, is an ordered sequence of transitions in the homeostatic space, like $\{K_1, K_2, ...\}$. We also define $\mathcal{P}(H_0)$ as the set of all pathways that if start from H_0, will end up at H^*. ■

Definition 1.2: For each homeostatic pathway p that starts from the initial motivational state H_0 and consists of n elements, we define $SDD_p(H_0)$ as the "sum of discounted drives" through that pathway:

$$SDD_p(H_0) = \sum_{t=0}^{n-1} \gamma^t d(H_{t+1}) \tag{1}$$

Where H_t is the motivational state at time t. ■

Definition 1.3: Similarly, for each homeostatic pathway p that starts from the initial motivational state H_0 and consists of n elements, we define $SDR_p(H_0)$ as the "sum of discounted rewards" through that pathway:

$$SDR_p(H_0) = \sum_{t=0}^{n-1} \gamma^t r_t = \sum_{t=0}^{n-1} \gamma^t (d(H_t) - d(H_{t+1})) \tag{2}$$

■

Proposition 1: For any initial state H_0, if $\gamma < 1$, we will have:

$$\underset{p \in \mathcal{P}(H_0)}{\operatorname{argmin}} SDD_p(H_0) = \underset{p \in \mathcal{P}(H_0)}{\operatorname{argmax}} SDR_p(H_0) \tag{3}$$

Roughly, this means that a policy that minimizes deviation from the setpoint, also maximizes acquisition of reward, and vice versa.

Proof: Assume that $p_i \in \mathcal{P}(H_0)$ is a sample pathway consisting of n_i transitions. As a result of these transitions, the internal state will take a sequence like: $\{H_{0,i} = H_0, H_{1,i}, H_{2,i}, .., H_{n,i} = H^*\}$. Denoting $D(H_x)$ by d_x for the sake of simplicity in notation, the drive value will take the following sequence: $\{d_{0,i} = d_0, d_{1,i}, d_{2,i}, .., d_{n,i} = d^* = 0\}$. We have:

$$SDD_{p_i}(H_0) = d_{i,1} + \gamma d_{i,2} + \gamma^2 d_{i,3} + .. + \gamma^{n-2} d_{i,n-1} + \gamma^{n-1} d^* \tag{4}$$

We also have:

$$\begin{aligned} SDR_{p_i}(H_0) &= r_{i,0} + \gamma r_{i,1} + \gamma^2 r_{i,2} + .. + +\gamma^{n-1} r_{i,n-1} \\ &= (d_0 - d_{i,1}) + \gamma(d_{i,1} - d_{i,2}) + \gamma^2(d_{i,2} - d_{i,3}) + .. + \gamma^{n-1}(d_{i,n-1} - d^*) \\ &= d_0 + (\gamma - 1)(d_{i,1} + \gamma d_{i,2} + \gamma^2 d_{i,3} + .. + \gamma^{n-2} d_{i,n-1}) \\ &= d_0 + (\gamma - 1).SDD_{p_i}(H_0) \end{aligned} \tag{5}$$

Since d_0 has a fixed value and $\gamma - 1 < 0$, it can be concluded that if a certain pathway from $\mathcal{P}(H_0)$ maximizes $SDR(H_0)$, it will also minimize $SDD(H_0)$. and vice versa. Thus, the pathways that satisfy these two objective are identical. ■

2 Energizing effect of motivational manipulation on habitual responses

We show the energizing effect for a two-dimensional homeostatic space. The proposition, however, can be easily extended to higher dimensions.

Proposition 2: For the two internal states $H_0 = (x_0, y_0)$, $H_1 = (x_1, y_1)$, and the drive-reducing outcome $K = (\varepsilon_X), \varepsilon_Y)$, we will have:

$$r(H_1, K) = \frac{d(H_1)}{d(H_0)}.r(H_0, K) \tag{6}$$

if $\varepsilon_X, \varepsilon_Y \to 0$, and $\dfrac{x^* - x_0}{y^* - y_0} = \dfrac{x^* - x_1}{y^* - y_1} = \dfrac{\varepsilon_X}{\varepsilon_Y}$.

Proof: Figure 1 shows a sample case that satisfies the assumptions of the proposition. The assumptions can be rewritten as:

$$\begin{aligned} \varepsilon_X &= k(x^* - x_0) = l(x^* - x_1) \\ \varepsilon_Y &= k(y^* - y_0) = l(y^* - y_1) \end{aligned} \tag{7}$$

Using the drive-reduction definition of reward, we will have:

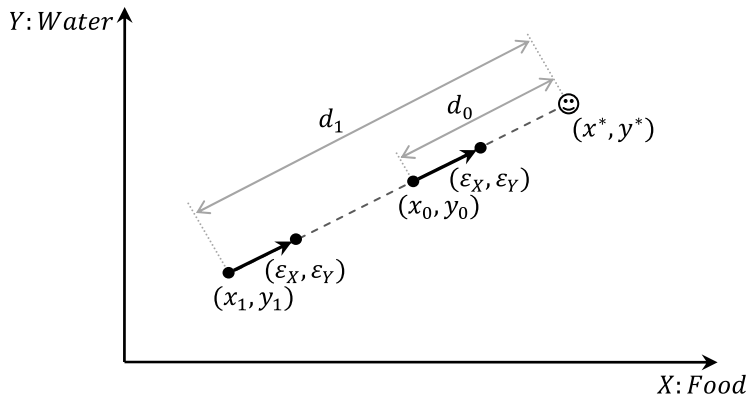

Figure 1: A sample for motivational shift in a two dimensional homeostatic space

$$
\begin{aligned}
\frac{r(H_1,K)}{r(H_0,K)} &= \frac{d(h_1)-d(h_1+K)}{d(h_0)-d(h_0+K)} \\
&= \frac{d(h_1)-d((x_1+\varepsilon_X,y_1+\varepsilon_Y))}{d(h_0)-d((x_0+\varepsilon_X,y_0+\varepsilon_Y))} \\
&= \frac{d(h_1)-\sqrt[m]{(x^*-x_1-l(x^*-x_1))^n+(y^*-y_1-k(y^*-y_1))^n}}{d(h_0)-\sqrt[m]{(x^*-x_0-l(x^*-x_0))^n+(y^*-y_0-k(y^*-y_0))^n}} \\
&= \frac{d(h_1)-\sqrt[m]{((x^*-x_1)(1-l))^n+((y^*-y_1)(1-l))^n}}{d(h_0)-\sqrt[m]{((x^*-x_0)(1-k))^n+((y^*-y_0)(1-k))^n}} \\
&= \frac{d(h_1)(1-\sqrt[m]{(1-l)^n})}{d(h_0)(1-\sqrt[m]{(1-k)^n})} \qquad (8)
\end{aligned}
$$

Assuming that $k,l \to 0$, which is equivalent to assuming that $\varepsilon_X,\varepsilon_Y \to 0$, and using the hopital rule, we will have:

$$
\lim_{\varepsilon_X,\varepsilon_Y\to 0}\frac{r(H_1,K)}{r(H_0,K)} = \lim_{\varepsilon_X,\varepsilon_Y\to 0}\frac{d(h_1)(1-\sqrt[m]{(1-l)^n})}{d(h_0)(1-\sqrt[m]{(1-k)^n})} = \frac{d(H_1)}{d(H_0)} \qquad (9)
$$

■

By assuming that the animals motivational state during the training trials has been H_0, and has shifted to H_1 for test trials, proposition 2 shows that under some assumptions, the rewarding value of an outcome that could be acquired in the training phase will be multiplied by $\frac{d(H_1)}{d(H_0)}$ in the test phase. Thus, as cached value of each state-action pair is expected to have converged to the sum of discounted rewards, and as all the rewards are multiplied by the factor $\frac{d(H_1)}{d(H_0)}$, the value of state-action pairs in the new motivational state can be approximated by:

$$
Q_1(s,a) = \frac{d(H_1)}{d(H_0)}.Q_0(s,a) \qquad (10)
$$

The above argument requires the animal to be in a single physiological state during the course of learning. Although this assumption is consistent with most of the behavioral conditioning experiments where the motivational state of the animal is tried to be kept constant at different blocks of the experiment, the internal state is so variable in real life conditions. To handle this variability it can be assumed that the habitual system uses a fixed physiological state, denoted by $\bar{H}$, as a common currency that everything else can be translated into. To see how this works, let's assume that at time t, the animal is in internal state H_t and after taking action a in state s, receives an outcome that has a rewarding value equal to r_t. This rewarding value is computed by the drive reduction equation and thus, is a function of H_t. The habitual system, however, rather that using r_t for updating the value of the performed action, uses $\bar{r}_t$ computed by the below equation:

$$
\bar{r}_t = \frac{d(\bar{H})}{d(H_t)}.r_t \qquad (11)
$$

This means that the habitual system learns the value of that action as if the outcome has been received when the internal state had been $\bar{H}$, rather than H_t. Again, the above equation gives a perfect approximation only when $H_t = c.\bar{H}$, where $c \geq 0$. Using this mechanism, variations in the internal state during training doesn't disrupt value learning, because the rewarding effect of outcomes has been evaluated as if the animal has been in a fixed internal state, $\bar{H}$. At the time of performance, t', the common-currency-based values, denote by $\bar{Q}(s,a)$, can be translated back, given the current internal state of the animal, $H_{t'}$:

$$
Q_{t'}(s,a) = \frac{d(H_{t'})}{d(\bar{H})}.\bar{Q}(s,a) \qquad (12)
$$

3 Anticipatory responses

Proposition 3: Assume that at the drive state x^*, the animal can voluntarily choose k_X and change its drive state to $(x^* + k_X)$. After one time delay the drive state moves to $(x^* + k_X - l_X)$, and after one more delay it will again go back to the setpoint, x^*. If $\gamma \to 1$, then the optimal strategy in this scenario in order to maximize the sum of discounted rewards, or minimize the sum of discounted deviations, is when $k_X = l_X/2$.

Proof: Figure 2 shows the trajectory of the internal state in the scenario described above. Denoting

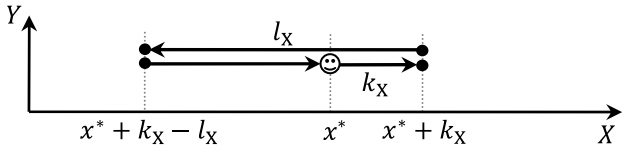

Figure 2: Predictive homeostasis for temperature regulation

the sum of discounted rewards by SDR, we will have:

$$\begin{aligned} SDR &= r(x^*, k_X) + \gamma.r(x^* + k_X, -l_X) + \gamma^2.r(x^* + k_X - l_X, l_X - k_X) \\ &= [d(x^*) - d(x^* + k_X)] + \gamma.[d(x^* + k_X) - d(x^* + k_X - l_X)] + \gamma^2.[d(x^* + k_X - l_X) - d(x^*)] \\ &= 0 - \sqrt[m]{|k_X|^n} + \gamma.\left[\sqrt[m]{|k_X|^n} - \sqrt[m]{|k_X - l_X|^n}\right] + \gamma^2.\left[\sqrt[m]{|l_X - k_X|^n} - 0\right] \end{aligned} \quad (13)$$

Thus,

$$\frac{d(SDR)}{dl_X} = 0 \;\Rightarrow\; (k_X)^{\frac{n}{m}-1}(\gamma - 1) = (l_X - k_X)^{\frac{n}{m}-1}\gamma(\gamma - 1) \quad (14)$$

If $\gamma \neq 1$, then

$$k_X = l_X.\frac{1}{1 + \gamma^{\frac{m}{m-n}}} \quad (15)$$

Thus, for $\gamma \to 1$ we have $k_X = l_X/2$. ■