CO6 2019 Exercise 1

Lecturer: J. Lussange
29 rue d'Ulm, 2nd floor
johann'lussange@ens.fr

Please submit solutions March 19 2019.

(1) **Static action choice and rewards.** We assume that there are two types of flowers, blue flowers (which we give the index 1) and yellow flowers (with index 2). The flowers carry nectar rewards r_1 and r_2, and we assume that the bee's internal estimates for the rewards are m_1 and m_2. The bee chooses flowers according to a softmax-policy based on its internal reward estimates,

$$p(c=1) = \frac{\exp(\beta m_1)}{\exp(\beta m_1) + \exp(\beta m_2)}$$

$$p(c=2) = \frac{\exp(\beta m_2)}{\exp(\beta m_1) + \exp(\beta m_2)}$$

where c denotes the choice — $c = 1$ meaning it chooses blue, and $c = 2$ meaning it chooses yellow.

(a) Show that $\sum_{c=1}^{2} p(c) = 1$.

(b) Show that you can rewrite $p(c = 1)$ as

$$p(c=1) = \frac{1}{1 + \exp(\beta(m_2 - m_1))} \tag{1}$$

(c) Plot the formula in (b) as a function of the reward difference, $d = m_2 - m_1$. Choose $\beta = 1$ and choose the range of differences d yourself. What happens if d gets very large? What happens if it gets very small (=negative)? What does that say about the bee's choices?

(d) Investigate the meaning of the parameter β. What happens if you increase β and make it very large, e.g., $\beta = 10$? What happens if you let it go to zero? What happens if it becomes negative? Do negative β make any sense? How does β influence the exploitation-exploration tradeoff?

(e) Imagine that there are N flowers instead of just two. How can you extend the above action choice strategy to N flowers? How can you trade off exploration and exploitation for the N-flower case?

(f) Imagine that there are N flowers, yet the rewards on these flowers, $r_i(t)$, change as a function of time. How should the bee adapt its internal estimates $m_i(t)$?

(g) **Advanced:** Given the learning rules you developed in (f), what will happen to the bee's internal estimates $m_i(t)$, if the rewards stay constant, i.e., $r_i(t) = \text{const}$ for all i? How does that depend on the parameter β? What is the characteristic time constant of convergence for the learning rules, i.e., how fast do the estimates converge to their real values?