Exercise 1: softmax Gibbs-policy

Younesse Kaddar

- PDF Version
- Online Version

Static action choice and rewards

We assume that there are two types of flowers:

- blue flowers (which we give the index 1)
- yellow flowers (with index 2)

The flowers carry nectar rewards r_1 and r_2, and we assume that the bee's interal 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,

$$p(c=i) = \frac{\exp(\beta m_i)}{\exp(\beta m_1) + \exp(\beta m_2)} \qquad \text{for } i = 1, 2$$

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

It's a straight-forward calculation:

$$\begin{aligned}\sum_{c=1}^{2} p(c) &= \frac{\exp(\beta m_1)}{\exp(\beta m_1) + \exp(\beta m_2)} + \frac{\exp(\beta m_2)}{\exp(\beta m_1) + \exp(\beta m_2)} \\ &= \frac{\exp(\beta m_1) + \exp(\beta m_2)}{\exp(\beta m_1) + \exp(\beta m_2)} = 1\end{aligned}$$

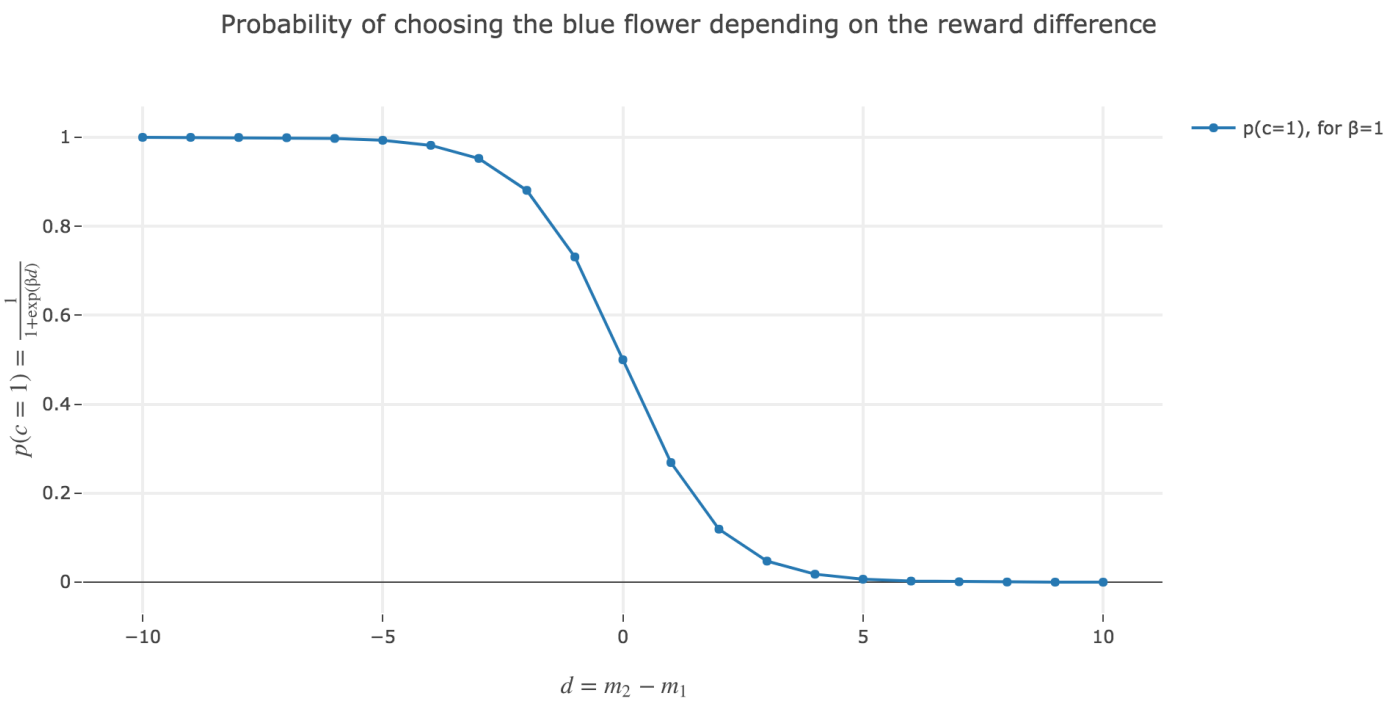
b). Show that you can rewrite $p(c=1)$ as $p(c=1) = \frac{1}{1+\exp(\beta(m_2 - m_1))}$

By divinding the numerator and the denomator of $p(c=1) = \frac{\exp(\beta m_1)}{\exp(\beta m_1) + \exp(\beta m_2)}$ by $\exp(\beta m_1) \neq 0$ (which amounts to multiplying both by $\exp(-\beta m_1)$), the result follows immediately (under the hood, we also use the fact that exp is a group homomorphism between $(\mathbb{R}, +)$ and $(\mathbb{R}_+^*, \times)$).

Thus:

> $p(c = i)$ *is a sigmoid of* $(m_i - m_{1-i})$

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 d is very small (= negative)?

As $p(c = 1)$ is a sigmoid of $-d$:

- $$p(c = 1) \xrightarrow[d \to +\infty]{} 0$$
- $$p(c = 1) \xrightarrow[d \to -\infty]{} 1$$

What does that say about the bee's choice?

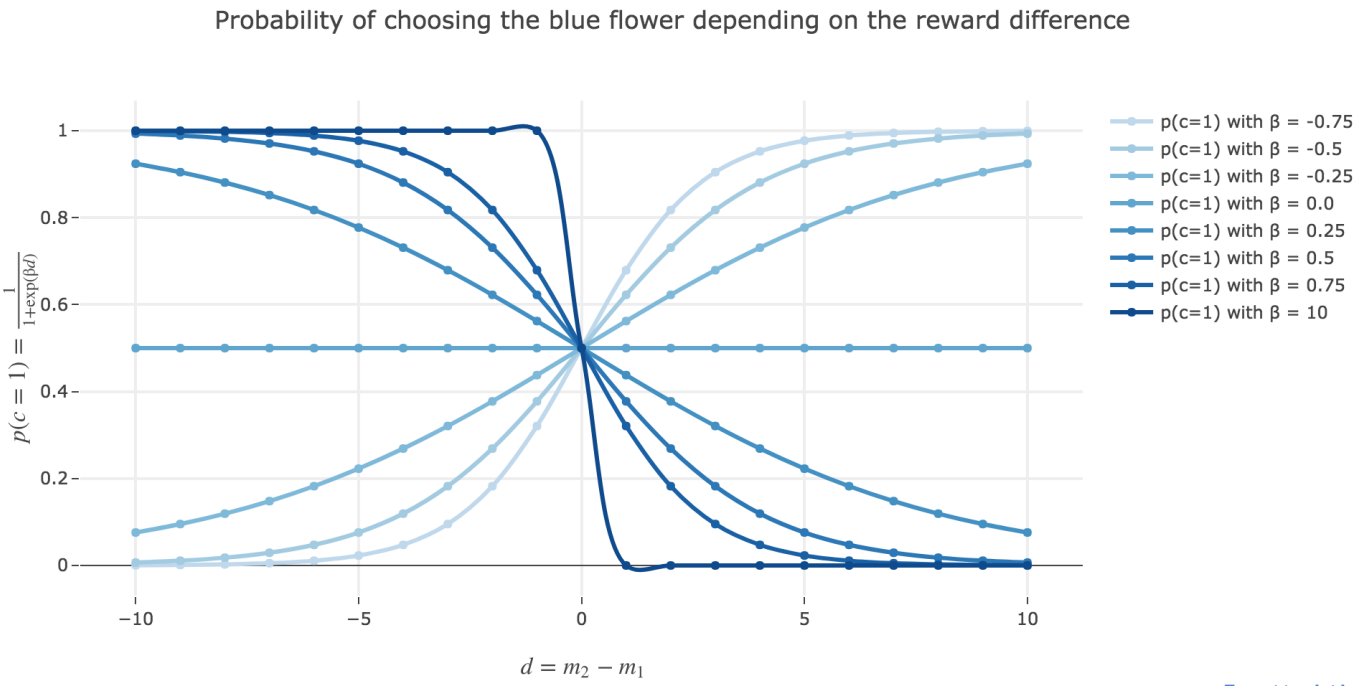
The larger the reward difference $d \stackrel{\text{def}}{=} m_2 - m_1$, the better the bee estimates the yellow flower (number 2) is compared to the blue (first) one. As a result, the lower the probability $p(c = 1)$ of the bee landing on the blue one.

Analogously, the lower the reward difference, the worse the yellow flower (from the bee's point of vue), and the greater the probability $p(c = 1)$ of the bee landing on the blue flower.

And in between, the bee's behavior is a mix between exploiting the most nutritious flower (according to the bee) and exploring the other one (which is what is expected).

d). Investigate the meaning of the parameter β.

For different values of β



How does β influence the exploitation-exploration tradeoff? What happens if you increase β and make it very large, e.g $\beta = 10$? What happens if you let it go to zero?

The parameter β plays a role analogous to the inverse temperature $\beta \overset{\text{def}}{=} \frac{1}{k_B T}$ in statistical physics. That is, for $\beta \geq 0$:

- **For $\beta \longrightarrow +\infty$:** the bigger the parameter β (corresponding, in physics, *to a low temperature/entropy*), the more the bee tends to **exploit** the most seemingly most nutritious flower. As a result:
    - when $d > 0$ (i.e. the yellow is avantageous), the probability of the bee landing on the blue flower rapidly decreases to 0
    - when $d < 0$ (i.e. the blue flower is more advantageous), the probability of the bee landing on the blue flower rapidly increases to 1

- **For $\beta \longrightarrow 0$:** the lower the parameter β (*high temperature/entropy in physics*), the more the bee tends to **explore** the flowers. Indeed:
    - as $\beta \longrightarrow 0$, $p(c = 1)$ becomes less and less steep, to such a point that it ends up being the constant function $1/2$ when $\beta = 0$ (at this point, the bee does nothing but exploring, since landing on either of the flowers is equiprobable, no matter how nutritious the bee deems the flowers to be)

What happens if it becomes negative? Do negative β make any sense?

As

$$p(-\beta, d) = \frac{1}{1 + \exp(-\beta d)} = \frac{1}{1 + \exp(\beta(-d))} = p(\beta, -d)$$

The graphs for $-\beta < 0$ are symmetric to those for $\beta > 0$ with respect to the y-axis, **which makes no sense from a behavioral point of vue:** it means that: the most nutritious a flower appears to the bee, the less likely the bee is to land on it (and conversely)!

e). Imagine that there are N flowers instead of just two. How can you extend the above action choice strategy to N flowers?

To continue drawing a comparison with statistical physics: in physics, the probability that a system occupies a microstate s is given by

$$p(s) = \frac{1}{Z}\exp(-\beta \underbrace{E_s}_{\text{microstate energy}})$$

Where the partition function $Z \overset{\text{def}}{=} \sum_{s'} \exp(-\beta E_{s'})$ is a normalization constant.

Likewise, here, the natural generalization is given by:

$$p(c) = \text{Const} \times \exp(\beta m_c)$$

where the **normalization constant** Const is equal to

$$\text{Const} = \sum_{c'} \exp(\beta m_{c'})$$

so that the probabilities add to 1.

> *Therefore, overall, the action choice strategy can be extended to N flowers as follows:*
>
> $$\forall c \in \{1, \cdots, N\}, \quad p(c) \overset{\text{def}}{=} \frac{\exp(\beta m_c)}{\sum_{c'=1}^{N} \exp(\beta m_{c'})}$$

How can you trade off exploration and exploitation for the N-flower case ?

As before (which was a particular case with $N = 2$), the exploration-exploitation tradeoff depends on the value of β:

- the bigger the parameter β, the more the bee exploits the flower it currently considers as the most nutritious
- the lower the parameter β, the more the bee explores all the flowers

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

When the time is discrete: we have seen, in class, three ways for the bee to update its internal estimates:

1. *The greedy update*:

$$m_c \overset{\text{def}}{=} \underbrace{r_{c,i}}_{\text{lastest reward received from this flower}}$$

2. *The batch update*: for $M \in \mathbb{N}^*$:

$$m_c = \frac{1}{M}\sum_{i=1}^{M} r_{c,i}$$

3. *The online update/the delta-rule*: for a learning rate $\varepsilon > 0$:

$$m_c \overset{\text{is updated as}}{\longrightarrow} m_c + \varepsilon(r_{c,i} - m_c)$$

If the time is now continuous:

By denoting by $T_c(t)$ the time that elapsed, at time t, since the last time (strictly before t) the bee visited the flower $c \in \{1, \cdots, N\}$, the analogous continuous-time updates are:

1. *The greedy update*:

$$m_c \overset{\text{def}}{=} r_c(t)$$

2. *The batch update*: for $M \in \mathbb{N}^*$, and if
   - $t_0 = 0$
   - $t_{n+1} = T_c(t - t_n)$
   - $\sum_{i=0}^{M} t_i = t$

$$m_c(t) = \frac{1}{M}\sum_{i=1}^{M} r_c(t - t_i)$$

3. *The online update/the delta-rule*: for a learning rate $\varepsilon > 0$:

$$m_c(t) = m_c(t - T_c(t)) + \varepsilon\Big(r_c(t) - m_c(t - T_c(t))\Big)$$

e). Given the learning rules you developed in f), what will happen to the bee's internal estimates $m_c(t)$, if the rewards stay constant for all c?

When it comes to the batch and greedy updates: if the rewards stay constant, so do the internal estimates, as soon as the bee lands on the corresponding flower.

The situation is trickier for the online update (so we will focus on this update from now on).

As we have seen in class:

$$m_c(t) \xrightarrow[t \to +\infty]{} r_c$$

How does that depends on the parameter β?

As the update of m_c is done whenever the flower c is visited by the bee:

- If β is big (exploitation mode): the most nutritious a flower c seems to the bee, the more often the bee will visit it, and the faster the convergence of $m_c(t)$ toward r_c
- If β is low (exploration mode): the convergence speed is as independent of the appeal of a flower c as β is low, since the lower the parameter β, the more often the bee explores (leading to the update of the explored flowers).

What is the characteristic time constant of convergence for the learning rules, i.e. how fast do the estimates converge to their real values?

For the sake of convenience, let's assume $T_c(t) = T_c$ is constant. We have

$$m_c(t) = (1-\varepsilon)\, m_c(t - T_c) + \varepsilon r_c$$

As a result, if $t \overset{\text{def}}{=} n\, T_c$:

$$m_c(t) = (1-\varepsilon)^n \Big(m_c(0) - r_c \Big) = (1-\varepsilon)^{t/T_c} \Big(m_c(0) - r_c \Big)$$

So for $t = \tau$ the characteristic time constant, we have, as in physics:

$$\frac{1}{\mathrm{e}} \Big(m_c(0) - r_c \Big) = m_c(\tau) = (1-\varepsilon)^{\tau/T_c} \Big(m_c(0) - r_c \Big)$$

And finally:

$$\begin{aligned} & \frac{1}{\mathrm{e}} = (1-\varepsilon)^{\tau/T_c} \\ \Longrightarrow \quad & \tau = -\frac{T_c}{\ln(1-\varepsilon)} \end{aligned}$$

Therefore:

> *Assuming that $T_c(t) = T_c$ is constant for the flower c, the characteristic time constant of convergence for the online update rule is equal to*
>
> $$\tau \overset{\text{def}}{=} -\frac{T_c}{\ln(1-\varepsilon)}$$