Exercise 06: Policy learning

Theoretical Neuroscience II

Johannes Gätjen Lorena Morton

June 9, 2016

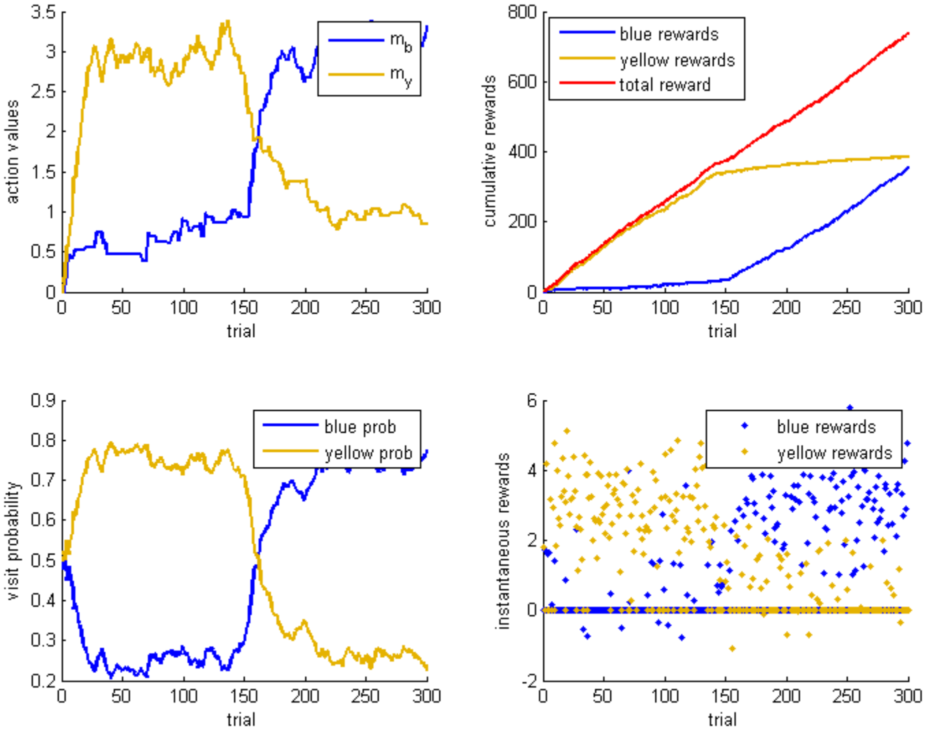

Figure 1: Indirect actor, example variable development with $\beta = 1$, $\epsilon = 0.1$ over $2 \cdot 150$ trials. Top left: action values for blue and yellow. Top right: Cumulative rewards from yellow and blue flowers, as well as the total reward. Bottom left: Visit probabilities P_b and P_y. Bottom right: Instantaneous rewards for blue and yellow flowers. Soon after the rewards are switched, the new action values are learned and the policy is adjusted. The action values in every trial relax towards the amount of instantaneous reward received.

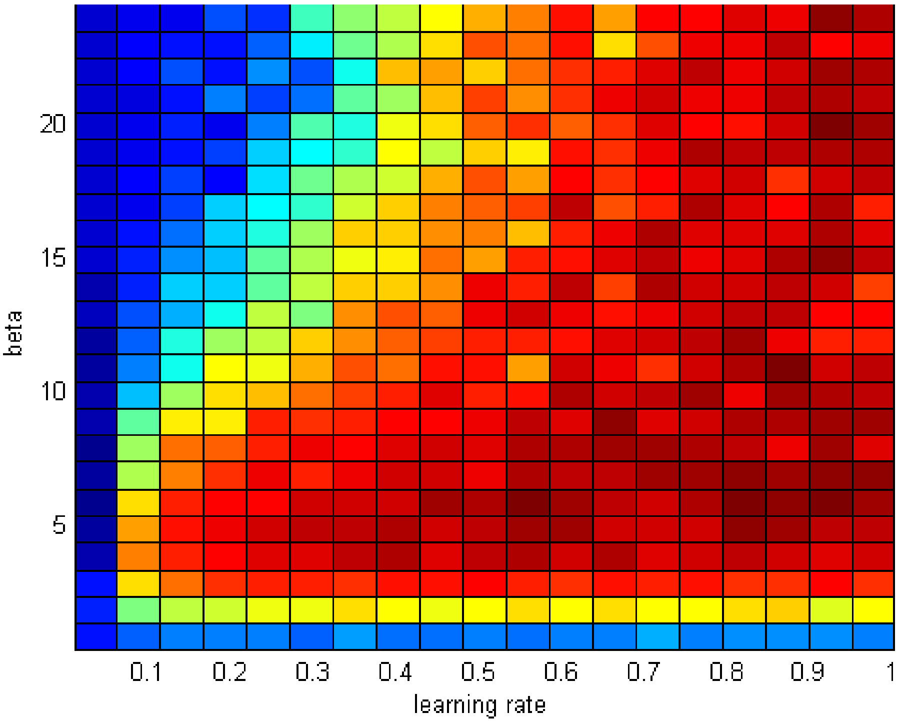

Figure 2: Indirect actor: Dependence of average total reward in 300 trials on learning rate and exploration parameter β. Neither ϵ nor β may be too small. A higher learning rate means higher β values are possible, otherwise there are not large differences, with the average total reward often being around 900.

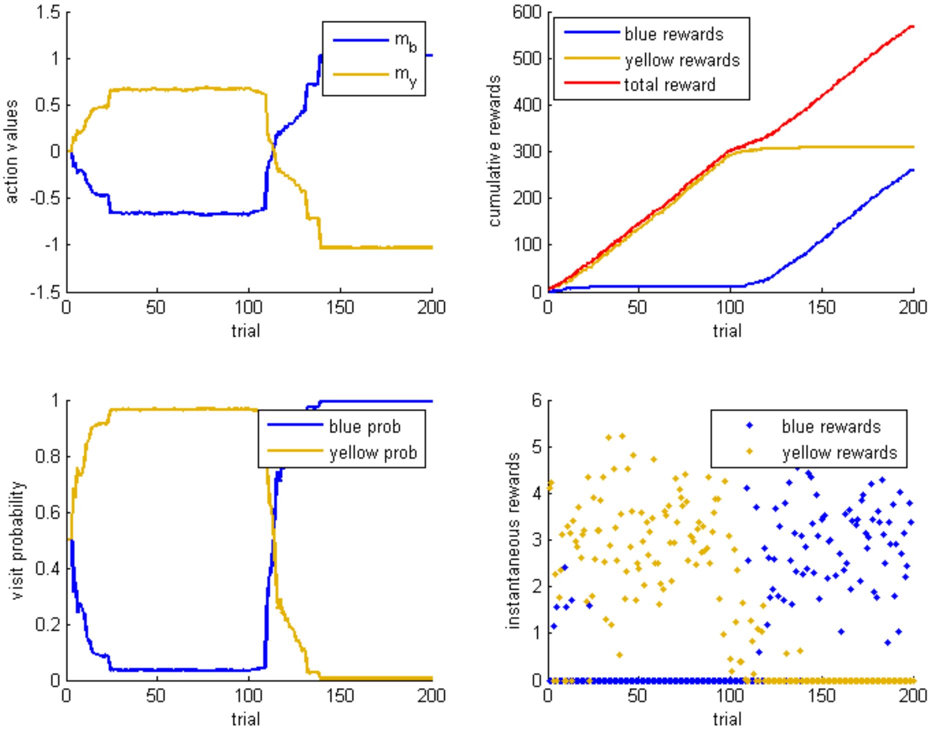

Figure 3: Indirect actor, example variable development with $\beta = 5$, $\epsilon = 0.15$ over $2 \cdot 100$ trials. Top left: action values for blue and yellow. Top right: Cumulative rewards from yellow and blue flowers, as well as the total reward. Bottom left: Visit probabilities P_b and P_y. Bottom right: Instantaneous rewards for blue and yellow flowers. In this case the new rewards are learned quickly, but often the bee is does not switch the policy at all. There is no direct relation between instantaneous rewards and action values.

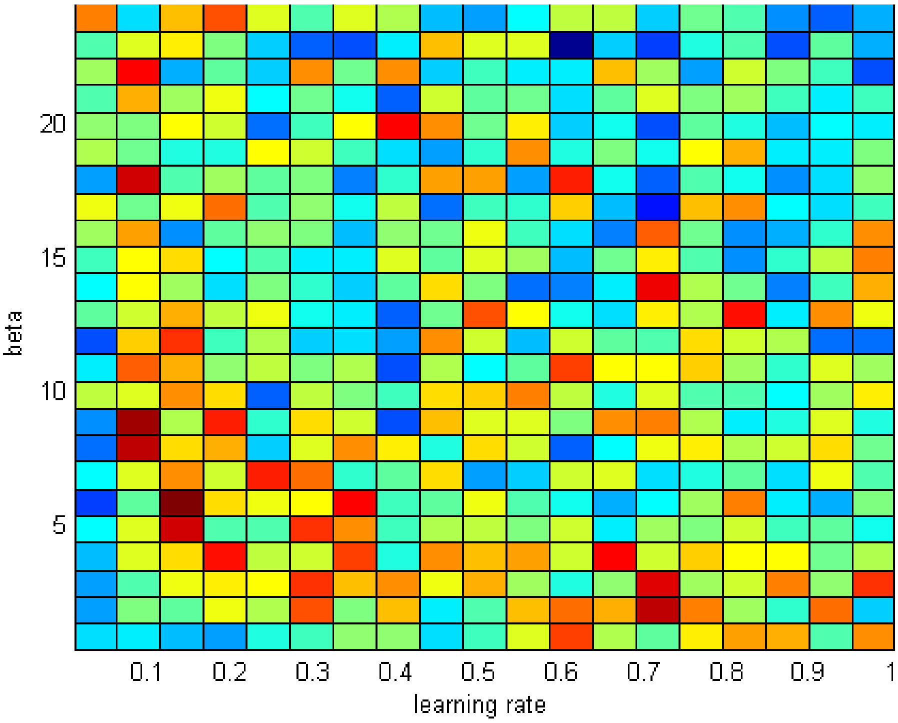

Figure 4: Direct actor: Dependence of average total reward in 300 trials on learning rate and exploration parameter β. Results overall are much worse and much less reliable than in the case of the indirect actor, with the average only reaching a value of around 700. The best parameter settings appear to be around $\beta = 5$, $\epsilon = 0.2$.