Cognitive Neuroscience II

Prof. Dr. Andreas Wendemuth

Lehrstuhl Kognitive Systeme

Institut für Elektronik, Signalverarbeitung und Kommunikationstechnik

Fakultät für Elektrotechnik und Informationstechnik

Otto-von-Guericke Universität Magdeburg

http://iesk.et.uni-magdeburg.de/ko/

Lecture 14

- Instrumental conditioning:
  Actions of the animal determines which reinforcement is provided

  – Static Action Choice (direct rewards)
  – Sequential Action Choice (delayed rewards)

Static Action Choice

- Animals develop policies (plans of action that increase reward)
- Example: foraging bee, blue and yellow flowers:
- Reward rb from probability density function (pdf) p[rb], reward ry from p[ry]
- Stochastic policy P[b], P[y]=1-P[b], parametrized as softmax functions with *action values* mb, my and *exploration parameter* β.
- Exploration-exploitation dilemma.

Stochastic policy

- Sigmoids $$P[b] = \frac{\exp(\beta m_b)}{\exp(\beta m_b) + \exp(\beta m_y)}$$
- Adjusting the parameters:
  - Indirect actor: estimate nectar volume by delta rule
  - Direct Actor: maximize expected average reward
- as follows:

Indirect actor

- Estimate nectar volume mb = <rb>
- Delta rule (Rescola-Wagner): on blue flower, rb is received and mb was expected, so change $mb \rightarrow mb + \varepsilon\delta$ with $\delta = rb - mb$, the same on yellow flower. I.e. if the pdfs p[rb], p[ry] change slowly relative to learning rate, this converges.
- *exploration parameter* β not changed.

Indirect actor-model

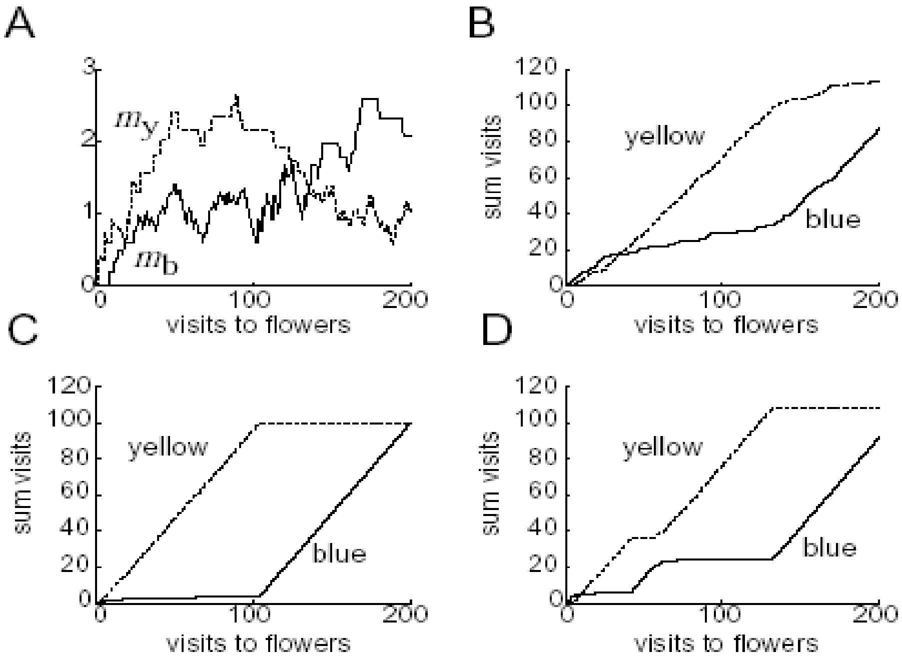


Figure 9.4: The indirect actor. Rewards were $\langle r_b \rangle = 1$, $\langle r_y \rangle = 2$ for the first 100 flower visits, and $\langle r_b \rangle = 2$, $\langle r_y \rangle = 1$ for the second 100 flower visits. Nectar was delivered stochastically on half the flowers of each type. A) Values of m_b (solid) and m_y (dashed) as a function of visits for $\beta = 1$. Because a fixed value of $\epsilon = 0.1$ was used, the weights do not converge perfectly to the corresponding average reward, but they fluctuates around these values. B-D) Cumulative visits to blue (solid) and yellow (dashed) flowers. B) When $\beta = 1$, learning is slow, but ultimately the change to the optimal flower color is made reliably. C;D) When $\beta = 50$, sometimes the bee performs well (C), and other times it performs poorly (D).

Indirect actor-experiments

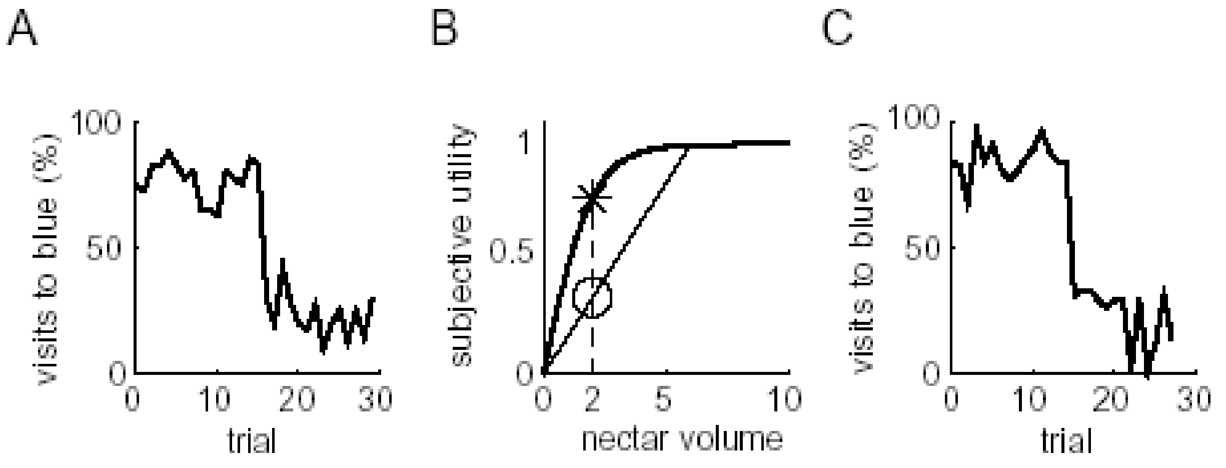


Figure 9.5: Foraging in bumble bees. A) The mean preference of five real bumble bees for blue flowers over 30 trials involving 40 flower visits. There is a rapid switch of flower preference following the interchange of characteristics after trial 15. Here, $\epsilon = 3/10$ and $\beta = 23/8$. B) Concave subjective utility function mapping nectar volume (in μl) to the subjective utility. The circle shows the average utility of the variable flowers, and the star shows the utility of the constant flowers. C) The preference of a single model bee on the same task as the bumble bees. (Data in A from Real, 1991; B & C adapted from Montague *et al*, 1995.)

Direct actor

- maximize expected average reward:

$$r = P[b]rb + P[y]ry$$

- Use $\frac{\partial}{\partial mb} P[b] = \frac{\partial}{\partial mb} \frac{\exp(\beta mb)}{\exp(\beta mb) + \exp(\beta my)} = \beta P[b]P[y]$

$$\frac{\partial r}{\partial mb} = \beta P[b]P[y](rb - ry) = \beta P[b]P[y](rb - r^*) - \beta P[y]P[b](ry - r^*)$$

- Interpret 2 terms: choice of b/y flowers with P[b], P[y].
- Change m_b by $\delta[b] = P[y](rb - r^*)$ if b is selected, and $\delta[b] = -P[b](ry - r^*)$ if y is selected. For m_y equiv.
- r* = mean reward

Direct actor-model

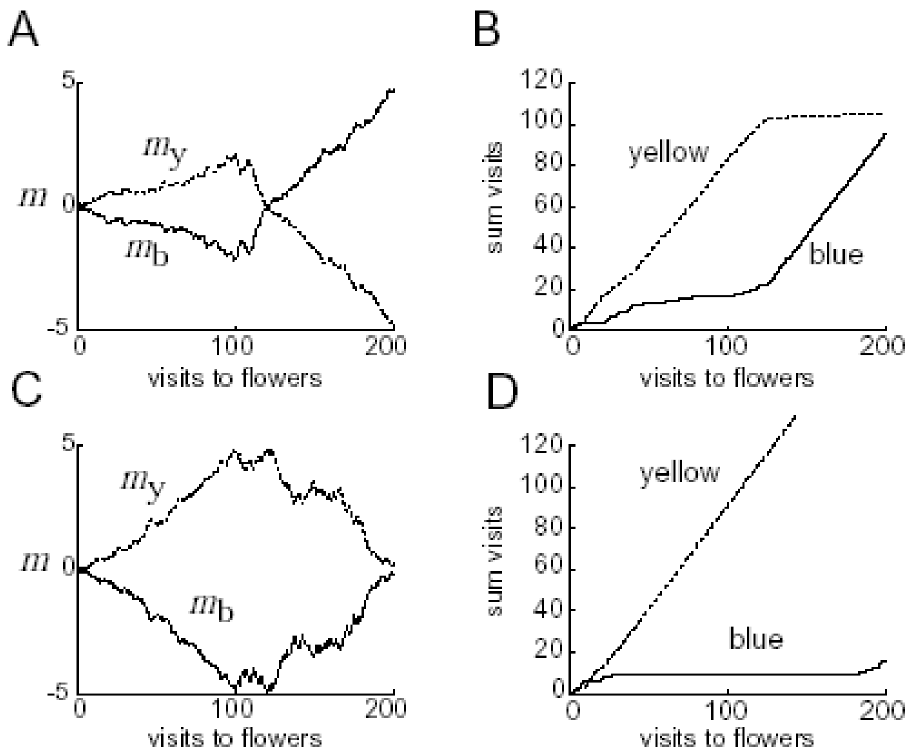


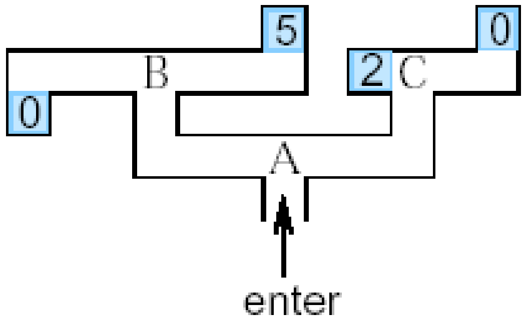
Figure 9.6: The direct actor. The statistics of the delivery of reward are the same as in figure 9.4, and $\epsilon = 0.1$, $\bar{r} = 1.5$, and $\beta = 1$. The evolution of the weights and cumulative choices of flower type (with yellow dashed and blue solid) are shown for two sample sessions, one with good performance (A & B) and one with poor performance (C & D).

Ex 4

- Study indirect and direct actors on a simple two-flower model where reward is given as in fig. 9.4.
- Why do the models sometimes not converge? What can be done to prevent this?

Sequential action choice

- (delayed rewards). Example: maze task



- Policy evaluation

$$v(\mathrm{B}) = \frac{1}{2}(0+5) = 2.5, \quad v(\mathrm{C}) = \frac{1}{2}(0+2) = 1, \quad \text{and}$$

$$v(\mathrm{A}) = \frac{1}{2}(v(\mathrm{B}) + v(\mathrm{C})) = 1.75.$$

Critic: Learning rule

- The rat chooses action a at location u and ends up at location u‘:

$$w(u) \rightarrow w(u) + \epsilon\delta \quad \text{with} \quad \delta = r_a(u) + v(u') - v(u)\,.$$

- Result of policy evaluation:

Policy evaluation: Model

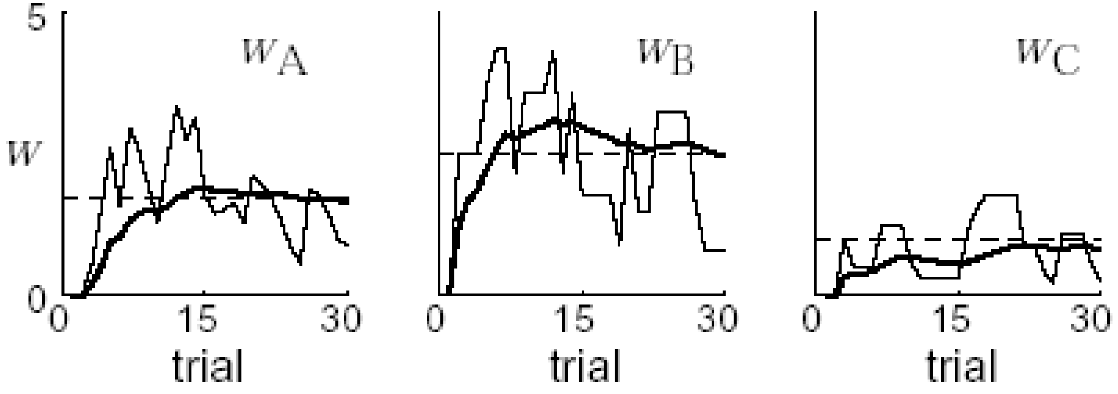


Figure 9.8: Policy evaluation. The thin lines show the course of learning of the weights $w(\mathrm{A})$, $w(\mathrm{B})$ and $w(\mathrm{C})$ over trials through the maze in figure 9.7 using a random unbiased policy ($\mathbf{m}(u) = 0$). Here $\epsilon = 0.5$, so learning is fast but noisy. The dashed lines show the correct weight values from equation 9.23. The thick lines are running averages of the weight values.

Actor: Policy Improvement

- Compare to direct actor: use rb-r*, here:
  rb = worth of action = ra(u) + v(u‘)
  r* = average worth = v(u)
- So use softmax with $\delta = r_a(u) + v(u') - v(u)$

$$m_{a'}(u) \rightarrow m_{a'}(u) + \epsilon\left(\delta_{aa'} - P[a';u]\right)\delta \qquad (9.25)$$

for all a', where $P[a';u]$ is the probability of taking action a' at location u given by the softmax distribution of equation 9.11 or 9.12 with action value $m_{a'}(u)$.

- $\delta = 0 + v(\mathrm{B}) - v(\mathrm{A}) = \ \ 0.75$ for a left turn
  $\delta = 0 + v(\mathrm{C}) - v(\mathrm{A}) = -0.75$ for a right turn.

Actor: experiments

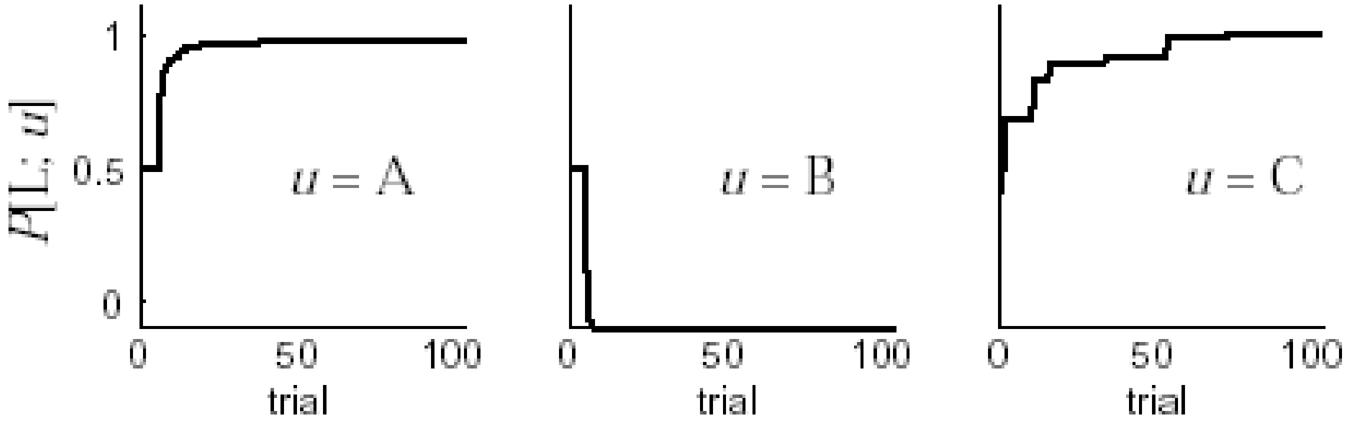


Figure 9.9: Actor-critic learning. The three curves show $P[\mathrm{L}; u]$ for the three starting locations $u = \mathrm{A}$, B, and C in the maze of figure 9.7. These rapidly converge to their optimal values, representing left turns and A and C and a right turn at B. Here, $\epsilon = 0.5$ and $\beta = 1$.

Ex 5

- Study critic and actor on the simple maze task model which was given in the lecture.
- Do the models always converge?

Resumé of Chapter 9

- Classical conditioning: fixed rewards
  - Rescorla Wagner Rule
  - Temporal Difference Learning (Analytical Treatment)
  - Linear rules and updates.
- Instrumental conditioning: animal determined rewards
  - Static Action Choice (indirect/direct actor)
  - Sequential Action Choice (delayed rewards, critic/actor)
  - Stochastic rules and updates. Animals chooses policy.
- In all cases, rewards / policies are learnt by rules.