In-class Assignment 2

Machine Learning, Summer term 2015, Norman Hendrich

To be discussed on June 29 / July 1

Action-Selection for RL (20 points)

The exploration/exploitation problem lies at the heart of reinforcement-learning. During the class, we will try to implement and play with the three fundamental action-selection strategies discussed in the lecture, namely

1. greedy action selection (with optimistic initialization),
2. ϵ-greedy action selection,
3. softmax action selection.

For simplicity, the n-armed bandit problem will be used as the testbed, with $n = 10$ and $r_i(t) = \mathcal{N}(B_i, 1)$. That is, when bandit i is played, the reward generated is a random number with mean B_i and variance 1.

One episode lasts $M = 2000$ actions, and the collected reward as well as the Q_i approximations for the different actions can then be plotted vs. time for each episode, or averaged (and plotted) over many episodes.

Write Matlab or Python/C/Java code to initialize the mean-values of the bandits, to initialize (and update) the Q-values for each bandit, and to generate normal-distribution random values.

ϵ-greedy action selection Assuming that the bandits are initialized with $B_i = 0.1 \cdot i$ for $i = 1, \ldots, 10$ (that is, $\{0.1, 0.2, 0.3, \ldots, 0.9, 1.0\}$), what is the expected long-term return (after the Q_i values have converged) for the ϵ-greedy strategy with $\epsilon = 0.1$ and $\epsilon = 0.01$?

softmax action selection Assuming that the bandits are initialized with $B_i = 0.1 \cdot i$ for $i = 1, \ldots, 10$, what is the expected long-term return (after the Q_i values have converged) for the softmax action selection with (a) very high temperature, and (b) very low temperature?

Implementation and play Implement the three action selection scheme and check their behaviour. Which scheme performs best for short episodes (e.g. 1000 or 2000 actions)?

Initialize the bandits bandits with $B_i = 0.1 \cdot i$ for $i = 1, \ldots, 10$ for comparison with the values calculated above. Next, initialize the bandits with $B_i = \mathcal{N}(1, 1)$ similar to the lecture. Finally, initialize one of the bandits with a very low value, e.g. $B_{10} = -100$.

Try to play with ϵ, the temperature τ and optionally optimistic initial values (e.g. $Q_i = 5$) to obtain the highest return.