Reinforcement Learning: Multi-Armed Bandits

Lecturers: Diederik M. Roijers & Ann Nowé

TA: Denis Steckelmacher

AI laboratory
Vrije Universiteit Brussel

Fall, 2017

Updates: about the course

- Grading
  - Assignments: 50%
  - Research project: 50%
- Assignents until February
  - MDP assignments part 1: planning (due: *November 27, 11:59pm*)
  - Multi-armed bandits (due: *November 8, 11:59pm*)
  - MDP assignments part 2: RL (due: *December 22, 11:59pm*)

This lecture is based on

- Sutton and Barto: chapters 1 and 2

Reinforcement learning

- No model of the environment given (a priori)
- The agent must learn by trial and error
- No examples of correct or incorrect behavior; instead only *rewards* for actions tried
- The agent has (partial) control over what data it will obtain for learning

K-armed bandit problem

- A slot machine (bandit) with K arms (or K, 1-armed machines)
- Each arm has an unknown stochastic reward or *payoff*
- Goal is to maximise cumulative payoff over some period

K-armed bandit problem

- A slot machine (bandit) with K arms (or K, 1-armed machines)
- Each arm has an unknown stochastic reward or *payoff*
- Goal is to maximise cumulative payoff over some period

Formalising the multi-armed bandit (MAB) problem

- K *actions* (or arms)
- Each time-step (also called *play* or *pull*) an arm, a_t is pulled
- The agent receives reward $r_t \sim R_{a_t}$
- Common probability distributions for R_a are e.g., Bernoulli and Gaussian distributions
- Arm values: $\mu_a = \mathbb{E}_{R_a}[r] = \int_{-\infty}^{\infty} r \; R_a(r) dr$
- Optimal arm / value: $\arg\max_a / \max_a \mu_a$
- Expected regret of pulling an arm once: $\Delta_a = \mu^* - \mu_a$

Formalising MAB learning objective

- Common in MDPs as well:
  - *Finite-horizon* return maximisation: maximise total reward $\sum_{t=1}^{T} r_t$ (online)
  - *Infinite-horizon* return maximisation: maximise *discounted* total reward $\sum_{t=0}^{\infty} \gamma^t r_t$, where $\gamma \in [0, 1)$ (online)
  - The *discount factor* γ can be interpreted as the probability of the game continuing after each step
- Common in MAB literature:
  - Regret minimisation: minimise $\sum_{t=0}^{T} \Delta_{a_t}$ (online)
  - Best-arm identification: within T pulls, identify the best arm (pure exploration / offline)

Exploration and exploitation

- The agent's ability to get reward in the future depends on what it knows about the arms. Thus, it must *explore* the arms in order to learn about them and improve its chances of getting future reward

- But the agent must also use what it already knows in order to maximise its total reward; Thus it must *exploit* by pulling the arms it expects to give the largest rewards

Balancing exploration and exploitation

- The main challenge in (online) RL is how to balance the competing needs of exploration and exploitation
- If the horizon is finite, exploration should decrease as the horizon gets closer
- If the horizon is infinite but $\gamma < 1$, exploration should decrease as the agent's uncertainty about expected rewards goes down
- If the horizon is infinite and $\gamma = 1$, there is an *infinitely delayed splurge*.

Human behaviour: hyperbolic discounting

Would you prefer to receive 50 euros today or 100 euros a year from now?

Human behaviour: hyperbolic discounting

Would you prefer to receive 50 euros today or 100 euros a year from now?

Would you prefer to receive 50 euros 5 years from now or 100 euros six years from now?

Human behaviour: hyperbolic discounting

Would you prefer to receive 50 euros today or 100 euros a year from now?

Would you prefer to receive 50 euros 5 years from now or 100 euros six years from now?

- *Exponential discounting*: γ is fixed
- Exponential discounting is common in RL
- But human behaviour is typically *not* rational for any fixed γ
- Instead we tend to use *hyperbolic discounting*
- The *decline* itself *decreases* over time

Ainslie, G. (2001). Breakdown of will. Cambridge University Press.

Action-value methods

- Based on observed rewards, maintain estimates of the expected value of each arm: $Q_t(a)$
- Estimates are based on the sample average; If action a has been chosen k_a times, yielding rewards $r_1, r_2, \ldots, r_{k_a}$, then:

$$Q_t(a) = \frac{\sum_{i=1}^{k_a} r_i}{k_a}$$

- Incremental implementation if a is selected at time $t+1$:

$$Q_{t+1}(a) = Q_t(a) + \frac{1}{k_a}\Big[r_{t+1} - Q_t(a)\Big]$$

- Example of an *update rule*:

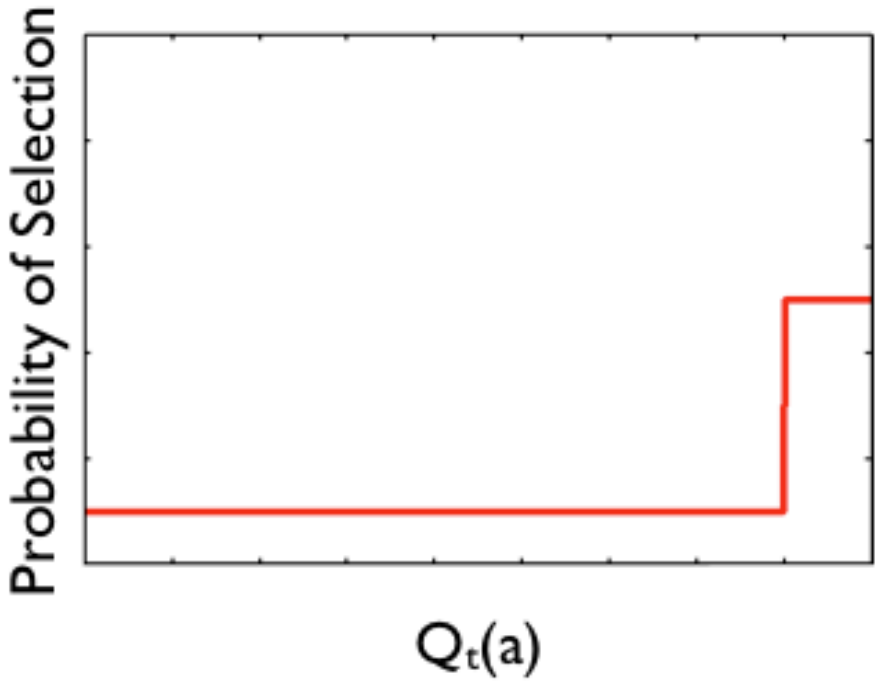
$$NewEstimate \leftarrow OldEstimate + StepSize[Target - OldEstimate]$$

Defining exploration vs. exploitation

- When using action-value methods, exploration and exploitation are easy to define.
- Exploiting means taking the *greedy* action: $a^* = \arg\max_a Q_t(a)$
- Exploring means taking any other action

Epsilon-greedy exploration

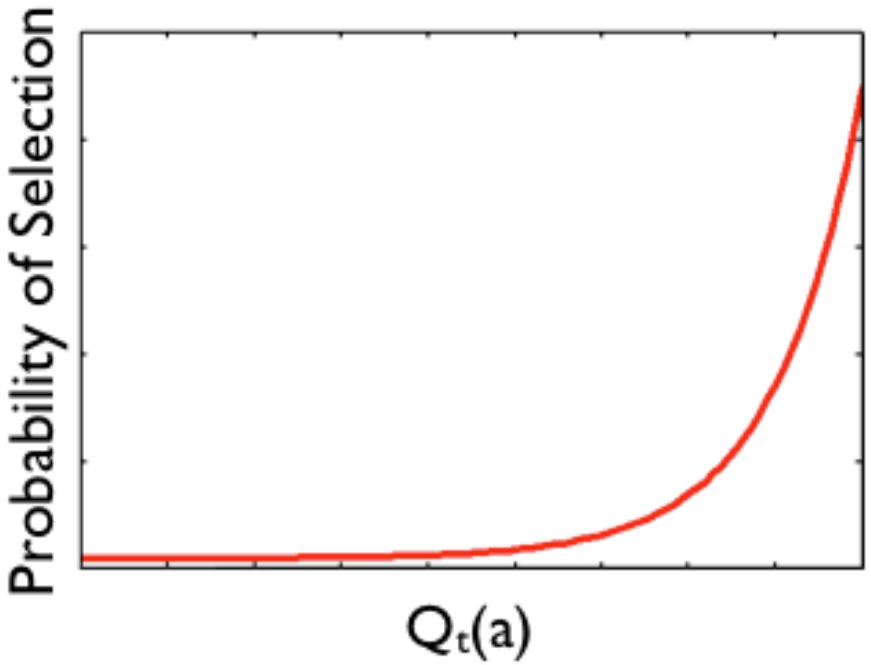
In *ϵ-greedy* exploration, the agent selects a random action with probability ϵ, and the greedy action otherwise



Softmax exploration

In *softmax* exploration, the agent chooses actions according to a *Boltzmann* distribution

$$p(a) = \frac{e^{Q(a)/\tau}}{\sum_{a'} e^{Q(a')/\tau}}$$

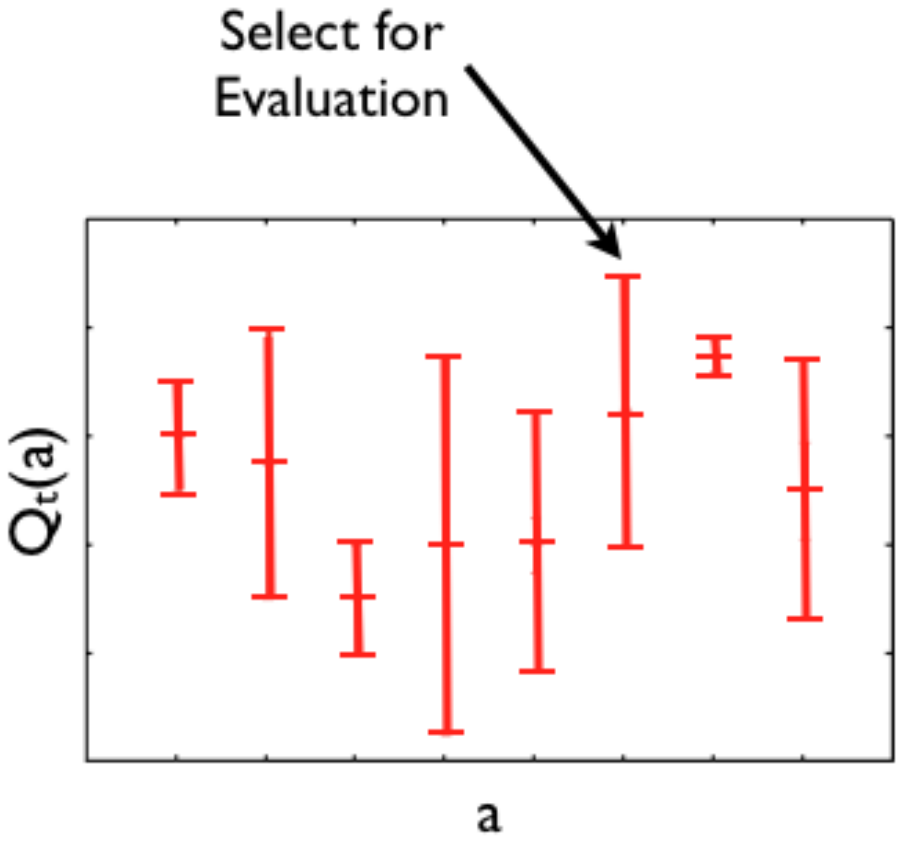


Optimistic initialisation

- In *optimistic initialisation*, the agent initialises its action-value estimates higher than the largest possible reward
- The agent always selects the greedy action
- Rewards are always disappointing, directing the agent to the least explored arms

Upper confidence bounds

- Neither ϵ-greedy nor softmax consider uncertainty in action-value estimates
- Goal of exploration is to reduce uncertainty
- So focus exploration on most uncertain actions
- Compute confidence intervals for each action
- Always take action with highest upper bound



Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multi- armed bandit problem." Machine Learning, 47:235–256, 2002.
Peter Auer, and Ronald Ortner. "UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem." Periodica Mathematica Hungarica 61.1–2: 55–65. 2010.

Thompson sampling

- $p_t(Q(a))$ is a distribution over true action values (μ_a)
- Start with a *prior* belief, $p_0(Q(a))$
- Observe r_t, compute a *posterior* belief using *Bayes rule*:

$$p_{t+1}(Q(a)) = p_t(Q(a)|r_t) = \frac{p_t(r_t|Q(a))p_t(Q(a))}{p(r_t)}$$

- For each timestep, sample $\hat{\mu}_{a,t}$ from $p_t(Q(a))$ for each a
- Pull arm that corresponds to best arm in set of samples: $\arg\max_a \hat{\mu}_{a,t}$
- Thompson sampling works very well in practice
- Algorithm from the 1930-ies, but only recently have there been proofs regarding regret bounds

William R. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." Biometrika, 25(3–4):285–294, 1933.
Kaufmann, Emilie, Nathaniel Korda, and Rémi Munos. "Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis." In ALT, pp. 199–213. 2012.
Agrawal, Shipra, and Navin Goyal. "Analysis of Thompson sampling for the multi-armed bandit problem." In Conference on Learning Theory, pp. 39.1–39.26 2012.

Bayes-optimal exploration

- $p(Q(a))$ is a distribution over true action values (μ_a)
- Start with a *prior* belief
- Observe r_t, compute a *posterior* belief using *Bayes rule*
- For each action, compute the probability of each r_t and the resulting posterior belief
- Repeat at each timestep, expanding *search tree* to horizon
- Always select action resulting in largest sum of rewards
- Optimal in a Bayesian sense but almost always computationally intractable!

Bandit application: preventive bandits

- Pandemic starts spreading
- Limited number of vaccines
- Find best allocation strategy
- Computationally expensive simulator
- Limited computation time

Pieter Libin, Timothy Verstraeten, Kristof Theys, Diederik M. Roijers, Peter Vrancx, and Ann Nowé. "Efficient evaluation of influenza mitigation strategies using preventive bandits." ALA workshop at AAMAS, 2017. (Innovative paper award)

Ongoing research at VUB.

Contextual bandit problem

- Also called *associative search*
- At each play, agent receives a *state signal*, also called an *observation* or *side-information*
- Expected payoffs depend on that observation
- Suppose there are many bandits, each a different color; after each play, you are randomly transported to another bandit
- In principle, can be treated as multiple simultaneous bandit problems and estimate $Q(s, a)$

Contextual bandit applications

- Ad placement
- News recommendation
- Information retrieval

Li, Wei, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. "Exploitation and exploration in a performance based contextual advertising system." In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 27–36. 2010.

Li, Lihong, Wei Chu, John Langford, and Robert E. Schapire. "A contextual-bandit approach to personalized news article recommendation." In Proceedings of the 19th international conference on World wide web, pp. 661–670. 2010.

Hofmann, Katja, Shimon Whiteson, and Maarten de Rijke. "Contextual bandits for information retrieval." In NIPS 2011 Workshop on Bayesian Optimization, Experimental Design, and Bandits, Granada, 2011.