A11.1 - Modifying the Bandits Problem and the ε-Greedy Solver

Jaan Tollander de Balsch - 452056

December 13, 2018

Multi-Armed Bandit

A tuple of $\langle A, R \rangle$, where:

- A is a set of **actions**.
- R is a **reward function**.

At each time step t, we take an action $a \in A$ on one slot machine and receive a rewards $r \in R$.

Reward and Regret

Action-Value is the mean reward for action a

$$Q(a) = \mathbb{E}[r \mid a]$$

Optimal reward of the optimal action a^*

$$Q(a^*) = \max_{a \in A} Q(a)$$

The **regret** is the opportunity loss for one step

$$l_t = \mathbb{E}[Q(a^*) - Q(a_t)]$$

Epsilon-Greedy

- Eplore with probability ε
- Eploit with probability $1 - \varepsilon$
- Estimated action-value for arm a is

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{t=1}^{T} r_t \mathbb{1}[a_t = a]$$

- Arm to exploit:

$$a_t^* = \text{argmax}_{a \in A} \widehat{Q}_t(a)$$

Bernoulli-Uniform Bandit

The **reward** r is a random number sampled from a probability distribution. The distributions are assumed to be known and independent, but the parameters unknown. For example, the reward for *Bernoulli-Uniform* bandit is

$$r \sim \text{Bernoulli}(\theta) \cdot \text{Uniform}(l, h)$$

where $\theta \in [0, 1]$ and $l, h \in \mathbb{R}, l \leq h$. The expected reward for action a is

$$\begin{aligned} \mathbb{E}[r \mid a] &= \mathbb{E}[\text{Bernoulli}(\theta)] \cdot \mathbb{E}[\text{Uniform}(l, h)] \\ &= \theta \cdot (l + h)/2. \end{aligned}$$

Other distributions could be used in similar way.

Epsilon-Greedy with Decreasing Epsilon

The Epsilon-Greedy should be modified such that initially it explores more in order to estimate the action values and as time increases and the evidence for action value estimates increase, the epsilon should decrease such that the best arm is exploited more often.

A naive formulation of epsilon ε_t is to formulate it as a decreasing function $\varepsilon_{t+1} < \varepsilon_t$ of time t starting from initial epsilon ε_0 decreasing towards zero as time tends towards infinity $t \to \infty$. For example,

$$\varepsilon_t = \varepsilon_0 / t^c$$

where $c \in \mathbb{R}$ is a parameter that controls the rate of the decrease. For larger number of bandits the rate of decrease should be set lower. If $c = 0$ the approach reduces to normal Epsilon-Greedy.