# Mathematics of Reinforcement Learning: Homework #1

Deadline: 4 December 2022, 23:59

## 1 Theoretical part

**Problem 1** (1 point). Consider a version of the $K$-armed bandit problem with all rewards supported in $[1/2; 1/2 + \varepsilon]$ for some $\varepsilon \in (0, 1/2)$. Define version of UCB-1 algorithm that attains improved regret bounds that depend on the $\varepsilon$ parameter.

**Problem 2** (2 points). Consider epsilon-greedy exploration algorithm with time-dependent exploration parameter

$$\varepsilon_t = t^{-1/3}(K \log t)^{1/3}\,,$$

where $K$ is the number of bandit arms. Show that $\mathbb{E}[\mathfrak{R}^T] = \mathcal{O}\left(T^{2/3} \cdot (K \log T)^{1/3}\right)$.

**Hint.** *Fix round $t$ and analyze $\mathbb{E}[\Delta(a_t)]$ separately. Set up the clean event for rounds $1, \ldots, t$, but also include the number of exploration rounds up to time $t$.*

**Problem 3** (2 points). Consider UCB-1 algorithm. Show that for any time $t$ and any sub-optimal arm $a$ ($\Delta(a) > 0$) it holds

$$\mathbb{E}[n_t(a)] = \mathcal{O}\left(\frac{\log t}{\Delta^2(a)}\right).$$

**Definition 1.** Random variable $X$ is called *sub-gaussian* with parameter $\sigma^2$ if

$$\log\left(\mathbb{E}[\exp(\lambda(X - \mathbb{E}X))]\right) \leq \frac{\lambda^2\sigma^2}{2}\,, \forall \lambda \in \mathbb{R}\,.$$

**Problem 4** (1 point). Propose the version of UCB-1 algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{TK})$ regret for sub-gaussian $K$-armed bandits: distributions $\mathcal{D}_a$ that corresponds to each arm $a$ are sub-gaussian with parameter $\sigma^2$.

## 2 Practical part

For this part you may find a template jupyter notebook and more details on wiki page.

**Problem 5** (4 points). Implement several baselines (UCB, Thompson Sampling, KL-UCB) for the Gaussian bandits problem and analyze them.