# Exercise 2: Multi Armed Bandits

Distributed 7.5.20, Submission 28.5.20 (in pairs)

The aim of this exercise is to allow you to develop some intuition about basic features of bandit algorithms in a simple comparative setting. The knowledge required for this exercise consists of the basic elements of bandits and bandit algorithms as presented in the course notes.

**Preliminary requirement:** Read sections 1,2,3.1 of the paper titled "Exploration exploitation tradeoff using variance estimates in multi-armed bandits" by Audibert, Munos and Szepesvári. This paper, which can be found on the course webpage, improves upon UCB by estimating the variance of the reward in addition to its mean. This algorithm is termed UCB-V. There is no need to study the theoretical aspects of the paper, simply focus on the (very simple) algorithmic issues.

## I Empirical

Consider the stochastic MAB problem with Bernoulli rewards, namely for each arm $r \in \{0,1\}$ and $P(r=1)=\theta$. For the sub-optimal arms $\theta = 0.5$ while for the optimal arm $\theta = 0.5+\delta$. In the experiments we use

$$K = 2, 10, 100 \quad \text{and} \quad \delta = 0.1, 0.01\,.$$

Each setup is run for $10^7$ rounds and is repeated 10 times. The following algorithms will be used:

- $\epsilon$-greedy with a time-dependent step-size $\epsilon_t = \min\left(1, cK/d^2 t\right)$ where $d$ is the smallest gap, namely $d = \min_i(\mu^* - \mu_i)$. Note that this choice of $d$ is not appropriate in practice since the average rewards are unknown! Select a value of $c$ by running a relatively short experiment in each setting and choosing a reasonable value, and then sticking with it. Once choosing this optimal value, also test a value below and above this value to test robustness.
- UCB1
- Thompson sampling. Repeat this for 3 values of the parameters $\alpha, \beta$ of the prior beta distribution, which you can choose in any way.

- UCB-V based on the paper by Audibert *et al.* mentioned above. In the experiments below use $b = c = 1$ and $\mathcal{E}_{s,t} = \log t$ (the parameters are defined in the paper).

Perform the following analysis:

1. For each of the 6 $(K, \delta)$ combinations:
   (a) Compute the average empirical pseudo-regret and its standard deviation over time. For each case plot the value of the mean regret versus time, and on a separate graph plot the standard deviation versus time. For each setting of $(K, \delta)$ plot all algorithms on the same graph. Use a different color for each algorithm. Plot the 6 cases on the same page for ease of visual comparison.
   (b) Define the *exploration index* as the cumulative measure of the number of times the algorithm selected the non-greedy choice (where the greedy choice corresponds to the arm with the largest empirical reward). Plot the mean value of the exploration index versus time, and discuss possible implications on the results obtained in (a).
2. Study empirically the relative merits of UCB1 versus UCB-V. Specifically, construct a setup (not necessarily related to the above setup) where each of the two algorithms is expected to perform better, and test your intuition empirically. Plot your results similarly to the previous item, and discuss their meaning.

**Note:** In the comparisons 2 above, be sure that the results are statistically meaningful, by running the comparison a sufficient number of times, and computing the mean and std.

## II Theoretical

**Exercise T1** Regret bound for the $\epsilon$-greedy algorithm (Exercise 1.2 from Slivkins' book)

Consider the $\epsilon$-greedy algorithm below with $\epsilon_t = t^{-1/3} \left(K \log t\right)^{1/3}$. Show that $E[R(T)] \leq t^{2/3} O \left(K \log t\right)^{1/3}$ for each round $t$.

Hint: Fix round $t$ and analyze $E[\Delta(a_t)]$ for this round separately. Set up the "clean event" for rounds $1, \ldots, t$ much like in Section 1.3.1 (treating $t$ as the time horizon), but also include the number of exploration rounds up to time $t$.

```
Epsilon-greedy algorithm
for $t = 1, 2, \ldots$ do
  Toss a coin with success probbaility $\epsilon_t$
  if success then
    Explore: choose an unirmly at random
  else
    Exploit: choose te armn withhighest empirical reward
  end if
end for
```

**Exercise T2** Anytime algorithms (Exercise 1.5 from Slivkins' book)

Take any bandit algorithm $\mathcal{A}$ for fixed time horizon $T$. Convert it to an algorithm $\mathcal{A}_\infty$ which runs forever, in phases $i = 1, 2, 3, \ldots$of $2^i$ rounds each. In each phase the algorithm $A$ is restarted and run with time horizon $2^i$.

State and prove a theorem which converts an instance-independent upper bound on regret for $\mathcal{A}$ into similar bound for $\mathcal{A}_\infty$ (so that this theorem applies to both UCB1 and Explore-first).

**Exercise T3** Bayesian methods

1. Consider the standard Bayesian data generating process. Let $\theta \sim p_0(\theta)$ and $D_n = \{x_1, \ldots, x_n\} \overset{\text{iid}}{\sim} f(\cdot, \theta)$. Consider the Bayesian objective

$$L\left\{p(\cdot|D_n); p\right\} = \iiint \left(f(x;\theta) - p(x|D_n)\right)^2 p(D_n|\theta)p_0(\theta)dD_n d\theta dx\,.$$

   Show that the pdf $p(x|D_n)$ minimizing this objective is given by

$$\begin{aligned} p\left(x|D_n\right) &= \int f(x;\theta)p(\theta|D_n)d\theta \\ p(\theta|D_n) &= \frac{p(D_n|\theta)p_0(\theta)}{p(D_n)} \end{aligned}\,.$$

2. Let $D_n = \{x_i\}_{i=1}^n$ be i.i.d. Bernoulli random variables, $x_i \sim \text{Ber}(\theta)$. Let the prior distribution for $\theta \in (0, 1)$ be $\text{Beta}(\alpha, \beta)$ where $\alpha > 0$ and $\beta > 0$.

   (a) Compute the optimal Bayes estimator for $\theta$, based on $D_n$, and compare your results to the maximum likelihood estimator. In particular, compare the estimators for small and large sample sizes.

(b) Consider the posterior distribution of $\theta$ for $(\alpha, \beta) = (1, 1)$ and $(\alpha, \beta) = (2, 2)$. Consider two observation sets, $D^{(1)}$ with 4 observations of $x_i = 1$ and 1 observation with $x_i = 0$, and $D^{(2)}$ with 40 observations of $x_i = 1$ and 10 observations of $x_i = 0$. Plot the prior, likelihood and posterior pdfs for each of the 4 combinations of prior parameters and data sets. Explain how the sample size and prior parameters affect the results.