Exercises
Data Mining: Learning from Large Data Sets
HS 2016

Series 6, Dec 15th, 2016 (Bandits)

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/teaching/dm-f16/`
Email questions to:
Johannes Kirschner, jkirschner@inf.ethz.ch

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise6` containing a PDF (LaTeXor scan) to `jkirschner@inf.ethz.ch` until Thursday, Dec 22th 2016.

Problem 1 (Analysis of UCB1):

In this exercise, we will prove a regret bound of the UCB1 algorithm, under the assumption that we know the total number of rounds T beforehand. We assume that there are k arms with random payoffs in $[0,1]$ and means $\mu_1, \mu_2, \dots, \mu_k$. We denote the optimal mean by $\mu^* = \max_{i=1}^k \mu_i$. Furthermore, let $\hat{\mu}_i^t$ be the empirical estimate of the mean μ_i at time t and denote by $\Delta_i = \mu^* - \mu_i$ the sub-optimality gaps. The full algorithm is given below.

Algorithm 1 UCB1 Policy for k-armed bandits with fixed T

---

function UCB1(T)
Initialize: $\hat{\mu}_i^0 = 0$, $n_i^0 = 0$ for each $i = 1, 2, \dots k$
Play each arm once for initialization purpose and update $\hat{\mu}_i^t$ and n_i^t
for $t = k+1, \dots, T$ **do**
  pick arm $j \leftarrow \arg\max_i \hat{\mu}_i^t + \sqrt{\frac{\ln T}{n_i^t}}$
  update count $n_j^{t+1} \leftarrow n_j^t + 1$ and mean estimate $\hat{\mu}_j^{t+1} \leftarrow \hat{\mu}_j^t + \frac{y^t - \hat{\mu}_j^t}{n_j^t}$
end for

---

1. We will prove that the expected regret $\mathbb{E}[R_T] = T\mu^* - \mathbb{E}[\sum_{t=1}^T y_t]$ of UCB1 after T rounds is at most
$$\mathbb{E}[R_T] \leq 4 \sum_{\Delta_i > 0} \frac{\ln(T)}{\Delta_i} + 2 \sum_{\Delta_i > 0} \Delta_i = O\left(\frac{k \ln(T)}{\min_i \Delta_i}\right) . \quad (1)$$
   (a) Denote by n_i^t the number of times arm i has been played until round t (note that this is a random variable). Show that the total expected regret can be written as $\mathbb{E}[R_T] = \sum_{i=1}^k \mathbb{E}[n_i^T]\Delta_i$.
   (b) Next, we define a confidence set $\mathcal{C}_i^t = \{\mu : |\mu - \hat{\mu}_i^t| \leq \sqrt{\frac{\ln(T)}{n_i^t}}\}$ for each arm i. Note that the UCB1 policy plays the arm with the largest upper bound of the confidence set. Use Hoeffding's inequality to show that
$$\mathbb{P}[\mu_i \notin \mathcal{C}_i^t] \leq \frac{2}{T^2} \quad \text{for any } t = 1, \dots, T. \quad (2)$$
   (c) Let i^* denote the index of an optimal arm, ie $\mu_{i^*} = \mu^*$. Consider any suboptimal arm i and show that if $\mu_i \in \mathcal{C}_i^t$ and $\mu_{i^*} \in \mathcal{C}_{i^*}^t$ for all $t = 1, \dots, T$, then $n_i^T \leq \frac{4\ln(T)}{\Delta_i^2}$.
   (d) Use the probabilistic bounds above to bound the expected number of times $\mathbb{E}[n_i^T]$ a suboptimal arm i is played, and put everything together to obtain the desired regret bound.
2. The bound we derived in the first part of the exercise is called an *instance dependent* regret bound, as it contains the sub-optimality gaps Δ_i. In particular the bound degrades as $\Delta_i \to 0$. Use the regret decomposition and that $\mathbb{E}[n_i^T] \in O\left(\frac{\log(T)}{\Delta^2}\right)$ to prove the *worst-case* regret bound $\mathbb{E}[R_T] = O(\sqrt{kT\ln(T)})$.