Sequential Learning: Homework #2

What I care about (again). I care about well-written proofs: with sufficient details, with calculations worked out and leading to pleasant and readable bounds. I favor quality of the writing over the quantity of questions answered. I give bonus points for elegant solutions.

Formats of your submission, deadline. Please send your solutions in a sequential manner, one exercise after the other. Wait for my OK to send a new solution, as I may request you to re-work a solution badly written. I may take 1 or 2 business days to get back to you, please take this into account when trying to abide by the deadline.

I expect to receive PDF files, with answers either handwritten and neatly scanned (as I do for my weekly lecture notes) or typed in LaTeX.

The PDF file must be named `YourName-HW2-ExN.pdf`, where `YourName` is to be replaced by your family name, and `N` by the exercise number. E.g., my submission for the second exercise of this homework would be named `Stoltz-HW2-Ex2.pdf`.

Deadline is Friday, April 1, at 6pm. This is a strict deadline. Please start early to allow for the iterations, do not wait for the last minute.

Beware: Typos. Most likely the statement comes with typos. This is part of the job. Try to correct them on your own!

Exercise 1: Explore then commit (can be solved after Course #4)

Consider a stochastic bandit setting with $K = 2$ arms only, each associated with a probability distribution ν_1, ν_2 over $[0, 1]$, with respective expectations μ_1, μ_2. Assume that you have to play for a given horizon $T \geqslant 4$. Explore each arm by pulling it m times, where $1 \leqslant m \leqslant T/2$. Compute the empirical averages $\widehat{\mu}_{1,m}$ and $\widehat{\mu}_{2,m}$ obtained. For the remaining $T - 2m$ steps, play only the arm j with maximal empirical average $\widehat{\mu}_{j,m}$ (ties broken arbitrarily). What is the regret of this strategy (called "explore then commit")?

For the analysis, we will assume with no loss of generality that arm 1 is the optimal arm and we will denote by $\Delta = \mu_1 - \mu_2$ the gap between the expectations associated with the two arms.

1. Show that $\mathbb{P}\big\{\widehat{\mu}_{1,m} < \widehat{\mu}_{2,m}\big\} \leqslant \exp\big(-m\Delta^2/c\big)$ where c is a constant (provide a numerical value).

2. Conclude that the regret is bounded by $m\Delta + (T - 2m)\Delta \exp\big(-m\Delta^2/c\big)$.

3. Assume that T and the range $[0, 1]$ are known. How should we choose m? Show a distribution-free bound on the regret that is a $o(T)$ — but it does not need to be of the typical $\sqrt{T}$ order of magnitude, it can be (much) larger. *Reminder: "distribution-free" means that the bound should only depend on* T *and on* $[0, 1]$*, not on the specific bandit problem considered, e.g., not on* Δ.