Consider a 2-armed bandit instance $B$ in which the rewards from the arms come from uniform distributions. The rewards of arm 1 are drawn uniformly at random from $[a, b]$, and the rewards of arm 2 are drawn uniformly at random from $[c, d]$, where $0 < a < c < b < d < 1$: Observe that an overlap in this means: both arms produce some rewards from the interval $[c, b]$. An algorithm $L$ proceeds as follows. First it pulls arm 1; then it pulls arm 2; whichever of these arms produced a higher reward (or arm 1 in case of a tie) is then pulled a further 20 times. In other words, the algorithm performs round-robin exploration for 2 steps and greedily picks an arm for the subsequent exploitation phase, during which that arm is blindly pulled 20 times.

(a) What is the expected cumulative regret of $L$ on $B$ after 22 pulls? Give theoretical derivation of the problem.

(b) Consider a small program to simulate $L$ and run it many times for fixed $a, b, c, d$. Is the average regret from these runs close to your answer?

Submit both the theoretical analysis and program.