Dueling

June 22, 2014

1 Notation

Let $k \in \mathbb{Z}$. At each time step t the algorithm pulls two arms $x_t, y_t \in [k]$ and observes a binary outcome b_t distributed as follows:

$$\Pr[b_t = 1] = \frac{\mu(y_t) - \mu(x_t) + 1}{2} ,$$

where $\mu(1) \dots \mu(k) \in [0, 1]$. (In a more elaborate version, μ is the expectation of a utility expectation. We stick to the "fixed utility" here.)

The reward of the algorithm at time t is $(\mu(x_t)+\mu(y_t))/2$. If $\mu^* = \max_{i\in[k]} \mu(i)$, then the total regret after T steps is $T\mu^* - \sum_{t=1}^{T}(\mu(x_t) + \mu(y_t))/2$.

2 An Algorithm

The algorithm is an "improvement" of the Doubler algorithm in [], which works in epoques, where the p'th epoque is of length 2^p. Let T_p denote the time iterations in the t'th epoque, so that $T_0 = \{1\}, T_1 = \{2, 3\}, T_3 = \{4, 5, 6, 7\}, \dots$. The left arm at each epoque is a random draw from the history of the right arm's draws in the previous epoque. The right arm is drawn from a UCB but with a "fix" that will be described below. The improvement compared to Doubler is that instead of resetting the right arm's UCB at each epoque (giving rise to $\log^2 T$ regret, we will combine all the historical information in a clever way.

Let $\mathcal{D}_p$ denote the distribution of the left arm in the p'th epoque. The main obsrevation is that if at each epoque we knew $f_p := \mathbb{E}_{x\in\mathcal{D}_p}\mu(x)/2$, then we would be able to run a normal UCB on the right side, by using $b_t + f_p$ as feedback (where $t \in T_p$). Indeed, since $\mathbb{E}[b_t] = \frac{\mu(y_t)-\mu(x_t)+1}{2}$, then by conditional expecatation $\mathbb{E}[b_t + f_p] = \frac{\mu(y_t)+1}{2}$. In other words, we have separated y_t from x_t in the feedback to UCB.

The question is, how do we estimate f_p, and with what confidence interval? By the defintion of the algorithm ,

$$f_{p+1} = |T_p|^{-1} \sum_{t \in T_p} \mu(y_t)/2 \ .$$

Hence

$$f_{p+1} = |T_p|^{-1} \sum_{t \in T_p} \mathbb{E}\left[b_t - \frac{1-\mu(x_t)}{2}\right] \ .$$

$$\mathbb{E} f_{p+1} = |T_p|^{-1} \sum_{t \in T_p} \mathbb{E}[b_t] + f_{p-1} - 1/2 \ .$$

As an approximation $\hat{f}_p$ of f_p, we can recursively take

$$\hat{f}_{p+1} := |T_p|^{-1} \sum_{t \in T_p} b_t + \hat{f}_{p-1} - 1/2 \ .$$

By Hoeffding, the estimate $|T_p|^{-1} \sum_{t \in T_p} b_t$ of the quantity $|T_p|^{-1} \sum_{t \in T_p} \mathbb{E}[b_t]$ has a confidence interval of $\Delta_p = \sqrt{\frac{\log(1/\delta)}{2^p}}$ in the p'th epoque, for success probability $1-\delta$. Assuming we are "ok" if Δ_p is smaller than the gap between the arms' utilities, then it is enough for δ to be 2^{-cp} for some small $c > 0$. But that's good, because $\sum_{p=1}^{\infty} 2^{-cp}$ converges, and hence we get a constant success probability for the all the estimates $\hat{f}_p$ (uniformly, by union bound).

Does this make sense, and is it possible to "incorporate" the $\hat{f}_p$'s in UCB somehow?