Regret Bounds for UCB

Finnian Lattimore

November 14, 2014

Assume $\exists$ a convex function, ψ, such that:

$$\begin{aligned} log(E[e^{\lambda(X-E[X]}]) &\leq \psi(\lambda) \\ log(E[e^{\lambda(E[X]-X}]) &\leq \psi(\lambda) \end{aligned} \tag{1}$$

This ensures that the moments of the distribution of X are defined. Assume that at time t we select arm I_t with the highest upper confidence bound:

$$I_t = argmax_{i=1\ldots K}\left[\hat{\mu}_{it} + (\psi^*)^{-1}\frac{\alpha log(t)}{T_{it}}\right] \tag{2}$$

Then if $\alpha > 2$,

$$R_n \leq \sum_{i:\Delta i>0}\left(\frac{\alpha\Delta i}{\psi^*(\Delta i/2)}log(n) + \frac{\alpha}{\alpha-2}\right) \tag{3}$$

Theorem 1. *If $I_t = i \neq i^*$ at least one of the following statements is true:*

1. *The estimated UCB on the best arm, i^*, is less than or equal to the actual reward for that arm: $\hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} \leq \hat{\mu}$*
2. *The estimated reward for arm i is greater than or equal to the estimated CI higher than the true reward for that arm: $\hat{\mu}_{it} \geq \mu_i + \hat{\epsilon}_{it}$*
3. *The number of times we have selected arm i in previous timesteps, T_{it}, is less than some bound (that grows logarithmically with n). $T_{it} < \frac{\alpha log(n)}{\psi^*(\Delta i/2)}$*

Proof. Assume statements 1-3 are all false.

$$\begin{aligned}
3.\ &\implies T_{it} > \frac{\alpha log(n)}{\psi^*(\Delta i/2)} \\
&\implies \Delta i > 2(\psi^*)^{-1}\frac{\alpha log(n)}{T_{it}} \geq 2(\psi^*)^{-1}\frac{\alpha log(t)}{T_{it}} = 2\hat{\epsilon}_{it} \\
&\implies \Delta i > 2\hat{\epsilon}_{it} \\
1.\ &\implies \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} > \hat{\mu} = \mu_i + \Delta i > \mu_i + 2\hat{\epsilon}_{it} \\
&\implies \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} > \mu_i + 2\hat{\epsilon}_{it} \\
2.\ &\implies \hat{\mu}_{it} < \mu_i + \hat{\epsilon}_{it} \\
&\implies \hat{\mu}_{it} + \hat{\epsilon}_{it} < \mu_i + 2\hat{\epsilon}_{it} \\
&\implies \hat{\mu}_{it} + \hat{\epsilon}_{it} < \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} \longleftarrow \text{ UCB for arm } i < \text{UCB for arm } i^*, \text{ which contradicts equation (2)}
\end{aligned}$$

□

If statements 1. and 2. are both false, then statement 3. places a bound on the number of times we can previously have selected the incorrect arm i in order to select it in this timestep.