Regret Bounds for UCB

Finnian Lattimore

November 26, 2014

Assume for each arm $i \in \{1...K\}$ there is an unknown distribution of rewards $P(X)$ and 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. If we select arm i a fixed number of times s:

$$P(|\hat{\mu}_{is} - \mu_i| > \epsilon) \leq 2e^{-s\psi^*(\epsilon)} \tag{2}$$

$$\implies P(|\hat{\mu}_{is} - \mu_i| > (\psi^*)^{-1}\frac{\log(\frac{2}{\delta})}{s}) \leq \delta \tag{3}$$

Assume that at time t we select arm I_t with the highest upper confidence bound:

$$I_t = argmax_{i=1...K}\left[\hat{\mu}_{it} + (\psi^*)^{-1}\frac{\alpha\log(t)}{T_{it}}\right] \tag{4}$$

Then if $\alpha > 2$,

$$R_n \leq \sum_{i:\Delta i>0}\left(\frac{\alpha\Delta i}{\psi^*(\Delta i/2)}log(n) + \frac{\alpha}{\alpha - 2}\right) \tag{5}$$

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: $\hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} \leq \mu^*$*
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). $T_{it} < \frac{\alpha log(n)}{\psi^*(\Delta i/2)}$* $\longleftarrow$ *feels odd that this grows with n, not t*

Proof. Assume statements 1-3 are all false.

$$
\begin{aligned}
3. \implies & T_{it} > \frac{\alpha log(n)}{\psi^*(\Delta i/2)} \\
\implies & \Delta i > 2(\psi^*)^{-1}\frac{\alpha log(n)}{T_{it}} \geq 2(\psi^*)^{-1}\frac{\alpha log(t)}{T_{it}} = 2\hat{\epsilon}_{it} \\
\implies & \Delta i > 2\hat{\epsilon}_{it}
\end{aligned}
$$

$$
\begin{aligned}
1. \implies & \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} > \mu^* = \mu_i + \Delta i > \mu_i + 2\hat{\epsilon}_{it} \\
\implies & \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} > \mu_i + 2\hat{\epsilon}_{it}
\end{aligned}
$$

$$
\begin{aligned}
2. \implies & \hat{\mu}_{it} < \mu_i + \hat{\epsilon}_{it} \\
\implies & \hat{\mu}_{it} + \hat{\epsilon}_{it} < \mu_i + 2\hat{\epsilon}_{it} \\
\implies & \hat{\mu}_{it} + \hat{\epsilon}_{it} < \hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} \longleftarrow \text{ UCB for arm } i < \text{UCB for arm } i^*\text{, which contradicts } i \neq i^*
\end{aligned}
$$

□

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. We can write the regret in terms of the number of times we select each arm and its sub-optimality:

$$
\begin{aligned}
\bar{R}_n =& n\mu^* - \sum_{t=1}^{n} E[\mu_{I_t}] \\
=& \sum_{i=1}^{K} \Delta_i E[T_{in}] \\
=& \sum_{i=1}^{K} \Delta_i E\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i\}\right] \quad \leftarrow \text{Expected number of times selected arm } I_t \text{ is } i
\end{aligned}
$$

Let $\gamma = \left\lceil \frac{\alpha \log(n)}{\psi^*(\Delta i/2)} \right\rceil$ and suppose we had selected arm i in all timesteps until γ. In the remaining timesteps, we can only select i if statement 3) is false

$$
\begin{aligned}
\implies E\left[T_{in}\right] &\leq \gamma + E\left[\sum_{t=1}^{n} \mathbb{1}\{I_t = i \text{ and (3) is false}\}\right] \\
&\leq \gamma + E\left[\sum_{t=\gamma+1}^{n} \mathbb{1}\{\text{(1) or (2) is true}\}\right] \leftarrow \text{since if (3) is false, (1) or (2) must be true} \\
&\leq \gamma + \sum_{t=\gamma+1}^{n} \left[\mathbb{P}(\text{(1) is true}) + \mathbb{P}(\text{(2) is true})\}\right] \leftarrow \text{Bubeck has} = \text{here but are (1) and (2) disjoint?}
\end{aligned}
$$

$$
\begin{aligned}
P((1)\text{ is true}) =& P\left(\hat{\mu}_{i^*t} + (\psi^*)^{-1}\left(\frac{\alpha\log t}{t}\right) \leq \mu^*\right) \\
\leq& P\left(\exists s \in \{1...t\} : \hat{\mu}_{i^*s} + (\psi^*)^{-1}\left(\frac{\alpha\log t}{s}\right) \leq \mu^*\right) \leftarrow \text{to get around the problem that t is random} \\
\leq& \sum_{s=1}^{t} P\left(\hat{\mu}_{i^*s} + (\psi^*)^{-1}\left(\frac{\alpha\log t}{s}\right) \leq \mu^*\right) \leftarrow \text{union bound}
\end{aligned}
$$

From equation (3) we have:

$$
P\left(\hat{\mu}_{i^*s} + (\psi^*)^{-1}\left(\frac{\log\frac{1}{\delta}}{s}\right) \leq \mu^*\right) < \delta
$$

$$
\text{Let } \delta = t^{-\alpha} \implies P\left(\hat{\mu}_{i^*s} + (\psi^*)^{-1}\left(\frac{\alpha\log t}{s}\right) \leq \mu^*\right) < t^{-\alpha}
$$

$$
\implies P((1)\text{ is true}) \leq \sum_{s=1}^{t} t^{-\alpha} = t * t^{-\alpha} = t^{1-\alpha}
$$

Similarly, $P((2)\text{ is true}) \leq t^{1-\alpha} \implies E[T_{in}] \leq \gamma + \sum_{t=\gamma+1}^{n} 2t^{1-\alpha}$