Regret Bounds for (α, γ)-UCB

Finnian Lattimore

January 13, 2015

The basic setting is the stochastic bandit problem:

- We have a set of arms $i \in \{1...K\}$
- For each arm i, there is an unknown distribution of rewards $P_i(X)$
- Each time we select a given arm, i, the reward is sampled i.i.d from $P_i(X)$. This is a big assumption - it states that the reward at a given timestep depends only on the action selected at that timestep, not on the sequence of previous actions.
- Can the reward distributions be (fixed) functions of the timestep? (No I think - otherwise the idea of an optimal arm independent of t doesn't make sense)

Some notation:

- K the number of arms
- i identifies an arm, $i \in \{1...K\}$
- i^* the arm with the highest true expected reward
- I_t the arm selected by the algorithm at timestep t
- $\hat{\mu}_i$ an estimator for the expected reward of arm i based on the sample mean
- μ_i the true expected reward for an arm i
- μ^* the true expected reward of i^* (the best arm)
- $\Delta_i = \mu^* - \mu_i$ how much worse arm i is than the best arm
- $T_i(s) = \sum_{t=1}^{s} \mathbb{1}\{I_t = i\}$ the number of times arm i was selected upto timestep s
- $\hat{\epsilon}_{it} = (\psi^*)^{-1}\left(\frac{\alpha log(t)}{T_i(t-1)}\right)$ an estimate of the uncertainty in the empirical estimator for the expected reward

The goal is to get a bound on the pseudo-regret, defined as:

$$R_n = n\mu^* - E\left[\sum_{t=1}^{n} \mu_{It}\right] \tag{1}$$

$$= \sum_{i=1}^{K} \Delta_i E[T_i(n)] \tag{2}$$

For the UCB algorithm we make the additional assumption that for each $P(X)$ there exists a convex function ψ such that:

$$
\begin{aligned}
log(E[e^{\lambda(X-E[X]}]) &\leq \psi(\lambda) \\
log(E[e^{\lambda(E[X]-X}]) &\leq \psi(\lambda)
\end{aligned} \tag{3}
$$

This ensures that the moments of the distribution of X are defined and gives us a high probability bound on the how much our sample based estimate of the expected reward can be below the true expected reward. Let $\hat{\mu}_{is} = \frac{1}{s}\sum_{t=1}^{s} X_t$ be the sample average and $\mu_i = E[P_i(X)]$ if we select arm i a fixed number of times s:

$$
P(\mu_i - \hat{\mu}_{is} > \epsilon) \leq e^{-s\psi^*(\epsilon)} \tag{4}
$$

$$
\implies P\left(\mu_i - \hat{\mu}_{is} > (\psi^*)^{-1}\left(\frac{1}{s}\log\frac{1}{\delta}\right)\right) \leq \delta \tag{5}
$$

The UCB Algorithm
Define the upper confidence bound for each arm i at timestep t as:

$$
ucb_{it} = \hat{\mu}_{it} + (\psi^*)^{-1}\left(\frac{\alpha}{T_i(t-1)}\log t\right) \text{, where } \alpha \text{ is an input parameter} \tag{6}
$$

$$
= \hat{\mu}_{it} + \hat{\epsilon}_{it} \tag{7}
$$

At time t select arm I_t with the highest upper confidence bound:

$$
I_t = argmax_{i=1...K}\,(ucb_{it}) \tag{8}
$$

Then if $\alpha > 2$,

$$
R_n \leq \sum_{i:\Delta i>0}\left(\frac{\alpha\Delta i\log(n)}{\psi^*(\Delta i/2)} + \frac{\alpha}{\alpha-2}\right) \tag{9}
$$

For $[0,1]$ random variables, with $\psi(\lambda) = \frac{\lambda^2}{8}$ gives $\psi^*(\epsilon) = 2\epsilon^2$, $(\psi^*)^{-1}(x) = \sqrt{\frac{x}{2}}$ and:

$$
ucb_{it} = \hat{\mu}_{it} + \sqrt{\frac{\alpha}{2T_i(t-1)}\log t} \tag{10}
$$

$$
R_n \leq \sum_{i:\Delta i>0}\left(\frac{2\alpha\log(n)}{\Delta_i} + \frac{\alpha}{\alpha-2}\right) \tag{11}
$$

The confidence bound we use in the UCB algorithm is clearly related but not identical to the bound in equation 5. The difference is due to the fact that when we use the UCB algorithm, the number of times each arm is selected is not fixed in advance but depends on the results of previous actions.

We now want to prove the bound in equation 9 holds.

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, $T_i(t-1)$, is less than some bound (that grows logarithmically with n).* $T_i(t-1) < \frac{\alpha log(n)}{\psi^*(\Delta i/2)}$ $\longleftarrow$ *feels odd that this grows with n, not t. Consequence of union bound?*

Proof. Assume statements 1-3 are all false.

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

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

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

□

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.

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

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

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

From equation (5) we have:

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

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

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

Similarly, $P((2) \text{ is true}) \leq t^{1-\alpha}$ so we have proved:

$$E\left[T_i(n)\right] \leq \gamma + \sum_{t=\gamma+1}^{n} 2t^{1-\alpha} \tag{12}$$

Plugging this into the definition of pseudo-regret in equation 2 gives:

$$R_n \leq \sum_{i=1}^{K} \Delta_i \left(\left\lceil \frac{\alpha \log(n)}{\psi^*(\Delta i/2)} \right\rceil + \sum_{t=\gamma+1}^{n} 2t^{1-\alpha} \right) \tag{13}$$