Explore then Commit

Lucas Berry

Comp 767

March 17th, 2017

Strategy

Explore then Commit (ETC):

- First stage: sample each arm m times.
- Second stage: choose the arm which produced the highest sample average and pull it for the remainder of the time horizon.

The strategy is one of the simplest and has nice properties.

Regret

Recall that total expected regret is defined as,

$$R_n = n\mu^* - E\left[\sum_{t=1}^{n} X_t\right]$$
$$= \sum_{k=1}^{K} \Delta_k E[T_k(n)],$$

where μ^* is the optimal arm, X_t is the sample at time t, K is the number of arms, $\Delta_k = (\mu^* - \mu_k)$ and $T_k(n)$ is the number of times arm k is pulled up to time n.

Regret

The ETC regret has the following form,

$$R_n = m \sum_{i=1}^{K} \Delta_i + (n - mK) \sum_{i=1}^{K} \Delta_i P(i = argmax_j \hat{\mu}_j(mK)).$$

Worst case, the wrong arm is chosen after the sampling period and the regret grows linearly. If the correct arm is chosen after the sampling period the expected regret should not grow anymore.

Subgaussian

Theorem: If X is σ^2-subgaussian, then $P(X \geq \epsilon) \leq exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$.

Using this theorem we can bound the regret of our ETC strategy. WLOG let $\mu_1 = \mu^*$ then

$$\begin{aligned} P(i = argmax_j \hat{\mu}_j(mK)) &\leq P(\hat{\mu}_i(mK) - \hat{\mu}_1(mK) \leq 0) \\ &= P(\hat{\mu}_i(mK) - mu_i - \hat{\mu}_1(mK) + \mu_1 \geq \Delta_i) \end{aligned}$$

Bound

Then by noticing that $\hat{\mu}_i(mK) - mu_i - \hat{\mu}_1(mK) + \mu_1$ is $\frac{2}{m}$-subgaussian,

$$P(i = argmax_j \hat{\mu}_j(mK)) \leq exp\left(-\frac{m\Delta_i^2}{4}\right)$$

and

$$R_n \leq m\sum_{i=1}^{K}\Delta_i + (n - mK)\sum_{i=1}^{K} exp\left(-\frac{m\Delta_i^2}{4}\right).$$

2-Arm Bandit

If we are given a 2-arm bandit problem we can rewrite our previous regret bound as such,

$$R_n \leq m\Delta + (n-2m)\Delta exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta exp\left(-\frac{m\Delta^2}{4}\right).$$

Taking the derivative and setting equal to zero we can see for quantity of m our regret is minimized,

$$m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil.$$

2-Arm Bandit

Plugging in our m into the bound for regret produces,

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right).$$

This is not the only bound we can derive for two arms. Let us assume the worst case scenario, the agent chooses the wrong arm each time. We can write this as,

$$R_n \leq n\Delta.$$

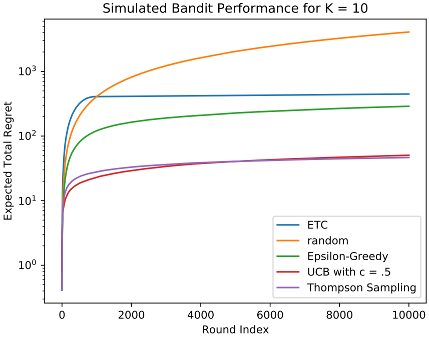
2-Arm Bandit

Thus the proper bound can be written as,

$$R_n \leq min \left\{ n\Delta, \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right)\right\}.$$

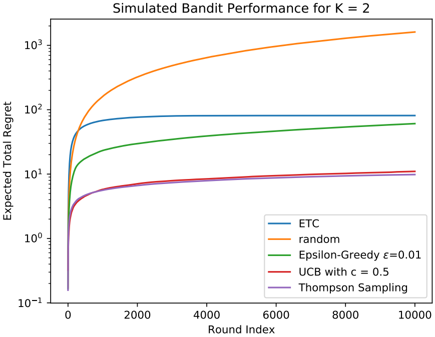
10-Arm Bandit Simulation

Below is a graph depicting the expected total regret for different strategies given 10 arms and binary rewards. Each strategy was run 100 times and then averaged.



2-Arm Bandit Simulation

Below is a graph depicting the expected total regret for different strategies given 2 arms and binary rewards. Each strategy was run 100 times and then averaged.



Notes

For the ETC strategy with 10 arms m was set 100 and for 2 arms $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$. In particular the latter m is impractical because normally one would not Δ before hand, also one might not know n. Another fact that makes $m = \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil$ impractical is that it can spit out negative values. If $n\Delta^2 < 4$ then $m < 0$.

References

- www.banditalgs.com/2016/09/14/first-steps-explore-then-commit/
- www.bandits.wikischolars.columbia.edu/file/view/Lecture+4.pdf