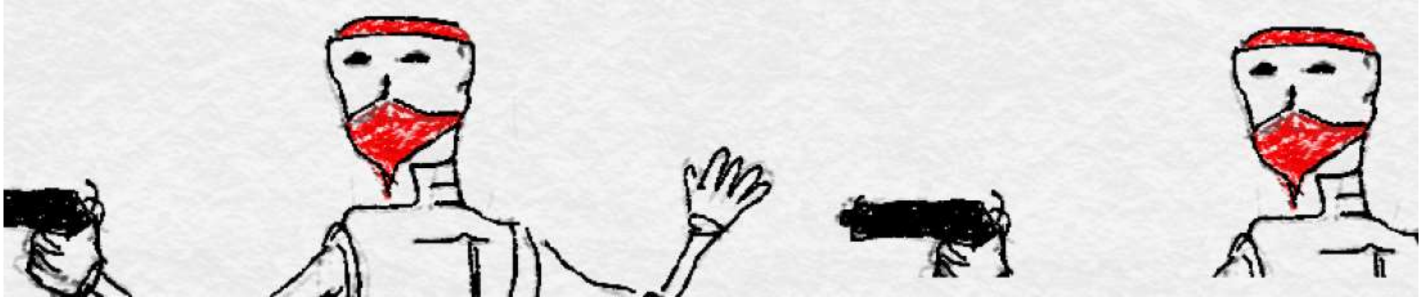
Bandit Algorithms

Home About View all posts



The Upper Confidence Bound Algorithm

Posted on September 18, 2016 by Tor Lattimore

We now describe the celebrated Upper Confidence Bound (UCB) algorithm that overcomes all of the limitations of strategies based on exploration followed by commitment, including the need to know the horizon and sub-optimality gaps. The algorithm has many different forms, depending on the distributional assumptions on the noise.

Optimism in the face of uncertainty but on overdose: Not recommended!

The algorithm is based on the principle of **optimism in the face of uncertainty**, which is to choose your actions as if the environment (in this case bandit) is as nice as is **plausibly possible**. By this we mean that the unknown mean payoffs of each arm is as large as plausibly possible based on the data that has been observed (unfounded optimism will not work — see the illustration on the right!). The intuitive reason that this works is that when acting optimistically one of two things happens. Either the optimism was justified, in which case the learner is acting optimally, or the optimism was not justified. In the latter case the agent takes some action that they believed might give a large reward when in fact it does not. If this happens sufficiently often, then the learner will learn what is the true payoff of this action and not choose it in the future. The careful reader may notice that this explains why this rule will eventually get things right (it will be "consistent" in some sense), but the argument does not quite explain why an optimistic algorithm should actually be a good algorithm among all consistent ones. However, before getting to this, let us clarify what we mean by **plausible**.

Recall that if $X_1, X_2, \dots, X_n$ are independent and 1-subgaussian (which means that $\mathbb{E}[X_i] = 0$) and $\hat\mu = \sum_{t=1}^n X_t/n$, then

$$\mathbb{P}\left(\hat\mu \geq \varepsilon\right) \leq \exp\big(-n\varepsilon^2/2\big)\,.$$

Equating the right-hand side with δ and solving for ε leads to

$$\mathbb{P}\left(\hat\mu \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta\,. \tag{1}$$

This analysis immediately suggests a definition of "as large as plausibly possible". Using the notation of the previous post, we can say that when the learner is deciding what to do in round t it has observed $T_i(t-1)$

samples from arm i and observed rewards with an empirical mean of $\hat{\mu}_i(t-1)$ for it. Then a good candidate for the largest plausible estimate of the mean for arm i is

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)}\log\left(\frac{1}{\delta}\right)}\,.$$

Then the algorithm chooses the action i that maximizes the above quantity. If δ is chosen very small, then the algorithm will be more optimistic and if δ is large, then the optimism is less certain. We have to be very careful when comparing the above display to (1) because in one the number of samples is the constant n and in the other it is a *random variable* $T_i(t-1)$. Nevertheless, this is in some sense a technical issue (that needs to be taken care of properly, of course) and the intuition remains that δ is approximately an upper bound on the probability of the event that the above quantity is an underestimate of the true mean.

The value of $1-\delta$ is called the *confidence level* and different choices lead to different algorithms, each with their pros and cons, and sometimes different analysis. For now we will choose $1/\delta = f(t) = 1 + t\log^2(t)$, $t = 1, 2, \dots$. That is, δ is time-dependent, and is decreasing to zero slightly faster than $1/t$. Readers are not (yet) expected to understand this choice whose pros and cons we will discuss later. In summary, in round t the UCB algorithm will choose arm A_t given by

$$A_t = \begin{cases} \operatorname{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}}\right), & \text{if } t > K\,; \\ t\,, & \text{otherwise}\,. \end{cases} \tag{2}$$

The reason for the cases is that the term inside the square root is undefined if $T_i(t-1) = 0$ (as it is when $t = 1$), so we will simply have the algorithm spend the first K rounds choosing each arm once. The value inside the argmax is called the **index** of arm i. Generally speaking, an **index** algorithm chooses the arm in each round that maximizes some value (the index), which usually only depends on current time-step and the samples from that arm. In the case of UCB, the index is the sum of the empirical mean of rewards experienced and the so-called *exploration bonus*, also known as the *confidence width*.

Besides the slightly vague "optimism guarantees optimality or learning" intuition we gave before, it is worth exploring other intuitions for this choice of index. At a very basic level, we should explore arms more often if they are (a) promising (in that $\hat{\mu}_i(t-1)$ is large) or (b) not well explored ($T_i(t-1)$ is small). As one can plainly see from the definition, the UCB index above exhibits this behaviour. This explanation is unsatisfying because it does not explain why the form of the functions is just so.

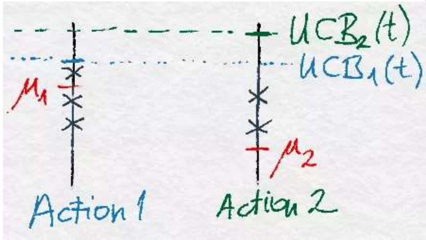


Illustration of UCB with 2 actions. The true means are shown in red ink. The observations are shown by crosses. Action 2 received fewer observations than Action 1. Hence, although its empirical mean is about the same as that of Action 1, Action 2 will be chosen in the next round.

An alternative explanation comes from thinking of what we expect from any reasonable algorithm. Suppose in some round we have played some arm (let's say arm 1) much more frequently than the others. If we did a good job designing our algorithm we would hope this is the optimal arm. Since we played it so much we can expect that $\hat{\mu}_1(t-1) \approx \mu_1$. To confirm the hypothesis that arm 1 is indeed optimal the algorithm better be highly confident about that other arms are indeed worse. This leads very naturally to confidence intervals and the requirement that $T_i(t-1)$ for other arms $i \neq 1$ better be so large that

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)}\log\left(\frac{1}{\delta}\right)} \le \mu_1\,, \tag{3}$$

because, at a confidence level of $1-\delta$ this guarantees that μ_i is smaller than μ_1 and if the above inequality did not hold, the algorithm would not be justified in choosing arm 1 much more often than arm i. Then, planning for (3) to hold makes it reasonable to follow the UCB rule as this will eventually guarantee that this inequality holds when arm 1 is indeed optimal and arm i is suboptimal. But how to choose δ? If the confidence interval fails, by which we mean, if actually it turns out that arm i is optimal and by unlucky chance it holds that

$$\hat{\mu}_i(t-1) + \sqrt{\frac{2}{T_i(t-1)}\log\left(\frac{1}{\delta}\right)} \le \mu_i\,,$$

then arm i can be disregarded even though it is optimal. In this case the algorithm may pay linear regret (in n), so it better be the case that the failure occurs with about $1/n$ probability to fix the upper bound on the expected regret to be constant for the case when the confidence interval fails. Approximating $n \approx t$ leads then (after a few technicalities) to the choice of $f(t)$ in the definition of UCB given in (2). With this much introduction, we state the main result of this post:

> ***Theorem (UCB Regret)****: The regret of UCB is bounded by*
>
> $$R_n \le \sum_{i:\Delta_i>0} \inf_{\varepsilon\in(0,\Delta_i)} \Delta_i \left(1 + \frac{5}{\varepsilon^2} + \frac{2}{(\Delta_i - \varepsilon)^2}\left(\log f(n) + \sqrt{\pi \log f(n)} + 1\right)\right)\,. \tag{4}$$
>
> *Furthermore,*
>
> $$\limsup_{n\to\infty} R_n/\log(n) \le \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}\,. \tag{5}$$

Note that in the first display, $\log f(n) \approx \log(n) + 2\log\log(n)$. We thus see that this bound scales logarithmically with the length of the horizon and is able to essentially reproduce the bound that we obtained for the unfeasible version of ETC with $K=2$ (when we tuned the exploration time based on the knowledge of Δ_2). We shall discuss further properties of this bound later, but now let us present a simpler version of the above bound, avoiding all these epsilons and infimums that make for a confusing theorem statement. By choosing $\varepsilon = \Delta_i/2$ inside the sum leads to the following corollary:

> ***Corollary (UCB Simplified Regret)****: The regret of UCB is bounded by*
>
> $$R_n \le \sum_{i:\Delta_i>0} \left(\Delta_i + \frac{1}{\Delta_i}\left(8\log f(n) + 8\sqrt{\pi \log f(n)} + 28\right)\right)\,.$$
>
> *and in particular there exists some universal constant $C>0$ such that for all $n \ge 2$, $R_n \le \sum_{i:\Delta_i>0}\left(\Delta_i + \frac{C\log n}{\Delta_i}\right)$.*

Note that taking the limit of the ratio of the bound above and $\log(n)$ does not result in the same rate as in the theorem, which is the main justification for introducing the epsilons in the first place. In fact, as we shall see the asymptotic bound on the regret given in (5), which is derived from~(4) by choosing $\varepsilon = \log^{-1/4}(n)$, is **unimprovable** in a strong sense.

The proof of the theorem relies on the basic regret decomposition identity that expresses the expected regret as the weighted sum of the expected number of times the suboptimal actions are chosen. So why will $\mathbb{E}\left[T_i(n)\right]$ be small for a suboptimal action i? This is based on a couple of simple observations: First, (disregarding the initial period when all arms are chosen once) the suboptimal action i can only be chosen if its UCB index is higher than that of an optimal arm. Now, this can only happen if the UCB index of action i is "too high", i.e., higher than $\mu^* - \varepsilon > \mu_i$ **or** the UCB index of that optimal arm is "too low", i.e., if it is below $\mu^* - \varepsilon < \mu^*$. Since the UCB index of any arm is with reasonably high probability an upper bound on the arm's mean, we don't expect the index of any arm to be below its mean. Hence, the total number of times when the optimal arm's index is "too low" (as defined above) is expected to be negligibly small. Furthermore, if the sub-optimal arm i is played sufficiently often, then its exploration bonus becomes small and simultaneously the empirical estimate of its mean converges to the true value, making the expected total number of times when its index stays above $\mu^* - \varepsilon$ small.

We start with a useful lemma that will help us quantify the *last* argument.

> ***Lemma*** *Let $X_1, X_2, \dots$ be a sequence of independent 1-subgaussian random variables, $\hat\mu_t = \sum_{s=1}^t X_s/t$, $\varepsilon > 0$ and*
>
> $$\kappa = \sum_{t=1}^n \mathbb{I}\left\{\hat\mu_t + \sqrt{\frac{2a}{t}} \ge \varepsilon\right\}\,.$$
>
> *Then, $\mathbb{E}[\kappa] \le 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1)$.*

Because the X_i are 1-subgaussian and independent we have $\mathbb{E}[\hat\mu_t] = 0$, so we cannot expect $\hat\mu_t + \sqrt{2a/t}$ to be smaller than ε until t is at least $2a/\varepsilon^2$. The lemma confirms that this is indeed of the right order as an estimate for $\mathbb{E}\left[\kappa\right]$.

Proof
Let $u = 2a\varepsilon^{-2}$. Then, by the concentration theorem for subgaussian variables,

$$\begin{aligned}
\mathbb{E}[\kappa] &\le u + \sum_{t=\lceil u\rceil}^n \mathbb{P}\left(\hat\mu_t + \sqrt{\frac{2a}{t}} \ge \varepsilon\right) \\
&\le u + \sum_{t=\lceil u\rceil}^n \exp\left(-\frac{t\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2}{2}\right) \\
&\le 1 + u + \int_u^\infty \exp\left(-\frac{t\left(\varepsilon - \sqrt{\frac{2a}{t}}\right)^2}{2}\right) dt \\
&= 1 + \frac{2}{\varepsilon^2}(a + \sqrt{\pi a} + 1)\,.
\end{aligned}$$

QED

Before the proof of the UCB regret theorem we need a brief diversion back to the bandit model. We have defined $\hat\mu_i(t)$ as the empirical mean of the ith arm after the tth round, which served us well enough for the analysis of the explore-then-commit strategy where the actions were chosen following a deterministic rule. For UCB it is very useful also to have $\hat\mu_{i,s}$, the empirical average of the ith arm *after s observations from that* *arm*, which occurs at a random time (or maybe not at all). To define $\hat{\mu}_{i,s}$ rigorously, we argue that without the loss of generality one may assume that the reward X_t received in round t comes from choosing the $T_i(t)$th element from the reward sequence $(Z_{i,s})_{1\le s\le n}$ associated with arm i, where $(Z_{i,s})_s$ is an i.i.d. sequence with $Z_{i,s} \sim P_i$. Formally,

$$X_t = Z_{A_t, T_{A_t}(t)}\,. \tag{6}$$

The advantage of introducing $(Z_{i,s})_s$ is that it allows a clean definition (without $Z_{i,s}$, how does one even define $\hat{\mu}_{i,s}$ if $T_i(n) \le s$?). In particular, we let

$$\hat{\mu}_{i,s} = \frac{1}{s}\sum_{u=1}^{s} Z_{i,u}\,.$$

Note that $\hat{\mu}_{i,s} = \hat{\mu}_i(t)$ when $T_i(t) = s$ (formally: $\hat{\mu}_{i,T_i(t)} = \hat{\mu}_i(t)$).

Proof of Theorem

As in the analysis of the explore-then-commit strategy we start by writing the regret decomposition.

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)]\,.$$

The rest of the proof revolves around bounding $\mathbb{E}[T_i(n)]$. Let i be some sub-optimal arm (so that $\Delta_i > 0$). Following the suggested intuition we decompose $T_i(n)$ into two terms. The first measures the number of times the index of the optimal arm is less than $\mu_1{-}\varepsilon$. The second term measures the number of times that $A_t = i$ and its index is larger than $\mu_1{-}\varepsilon$.

$$\begin{aligned}
T_i(n) &= \sum_{t=1}^{n} \mathbb{I}\{A_t = i\} \\
&\le \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_1(t-1) + \sqrt{\frac{2\log f(t)}{T_1(t-1)}} \le \mu_1{-}\varepsilon\right\} + \\
&\qquad \sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t-1) + \sqrt{\frac{2\log f(t)}{T_i(t-1)}} \ge \mu_1{-}\varepsilon \text{ and } A_t = i\right\}\,.
\end{aligned} \tag{7}$$

The proof of the first part of the theorem is completed by bounding the expectation of each of these two sums. Starting with the first, we again use the concentration guarantee.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{\hat\mu_1(t-1)+\sqrt{\frac{2\log f(t)}{T_1(t-1)}}\leq \mu_1-\varepsilon\right\}\right] &= \sum_{t=1}^n \mathbb{P}\left(\hat\mu_1(t-1)+\sqrt{\frac{2\log f(t)}{T_1(t-1)}}\leq \mu_1-\varepsilon\right) \\
&\leq \sum_{t=1}^n\sum_{s=1}^n \mathbb{P}\left(\hat\mu_{1,s}+\sqrt{\frac{2\log f(t)}{s}}\leq \mu_1-\varepsilon\right) \\
&\leq \sum_{t=1}^n\sum_{s=1}^n \exp\left(-\frac{s\left(\sqrt{\frac{2\log f(t)}{s}}+\varepsilon\right)^2}{2}\right) \\
&\leq \sum_{t=1}^n \frac{1}{f(t)}\sum_{s=1}^n \exp\left(-\frac{s\varepsilon^2}{2}\right) \\
&\leq \frac{5}{\varepsilon^2}\,.
\end{aligned}$$

The first inequality follows from the union bound over all possible values of $T_1(t-1)$. This is an important point. The concentration guarantee cannot be applied directly because $T_1(t-1)$ is a random variable and not a constant. The last inequality is an algebraic exercise. The function $f(t)$ was chosen precisely so this bound would hold. If $f(t)=t$ instead, then the sum would diverge. Since $f(n)$ appears in the numerator below we would like f to be large enough that its reciprocal is summable and otherwise as small as possible. For the second term in (7) we use the previous lemma.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{\hat\mu_i(t-1)+\sqrt{\frac{2\log f(t)}{T_i(t-1)}}\geq \mu_1-\varepsilon \text{ and } A_t=i\right\}\right] \\
&\quad\leq \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\left\{\hat\mu_i(t-1)+\sqrt{\frac{2\log f(n)}{T_i(t-1)}}\geq \mu_1-\varepsilon \text{ and } A_t=i\right\}\right] \\
&\quad\leq \mathbb{E}\left[\sum_{s=1}^n \mathbb{I}\left\{\hat\mu_{i,s}+\sqrt{\frac{2\log f(n)}{s}}\geq \mu_1-\varepsilon\right\}\right] \\
&\quad= \mathbb{E}\left[\sum_{s=1}^n \mathbb{I}\left\{\hat\mu_{i,s}-\mu_i+\sqrt{\frac{2\log f(n)}{s}}\geq \Delta_i-\varepsilon\right\}\right] \\
&\quad\leq 1+\frac{2}{(\Delta_i-\varepsilon)^2}\left(\log f(n)+\sqrt{\pi\log f(n)}+1\right)\,.
\end{aligned}$$

The first part of the theorem follows by substituting the results of the previous two displays into (7). The second part follows by choosing $\varepsilon=\log^{-1/4}(n)$ and taking the limit as n tends to infinity.
QED

Next week we will see that UCB is close to optimal in several ways. As with the explore-then-commit strategy, the bound given in the previous theorem is not meaningful when the gaps Δ_i are small. Like that algorithm it is possible to prove a *distribution-free* bound for UCB by treating the arms i with small Δ_i differently. Fix $\Delta>0$ to be chosen later. Then, from the proof of the bound on the regret of UCB we can derive that $\mathbb{E}\left[T_i(n)\right]\leq\frac{C\log(n)}{\Delta_i^2}$ holds for all $n\geq 2$ with some universal constant $C>0$. Hence, the regret can be bounded without dependence on the sub-optimality gaps by

$$
\begin{aligned}
R_n &= \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[T_i(n)] = \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i\geq\Delta} \Delta_i \mathbb{E}[T_i(n)] \\
&< n\Delta + \sum_{i:\Delta_i\geq\Delta} \Delta_i \mathbb{E}[T_i(n)] \leq n\Delta + \sum_{i:\Delta_i\geq\Delta} \frac{C\log n}{\Delta_i} \\
&\leq n\Delta + K\frac{C\log n}{\Delta} = \sqrt{CKn\log(n)}\,,
\end{aligned}
$$

where in the last step we chose $\Delta = \sqrt{KC\log(n)/n}$, which optimizes the upper bound.

There are many directions to improve or generalize this result. For example, if more is known about the noise model besides that it is subgaussian, then this can often be exploited to improve the regret. The main example is the Bernoulli case, where one should make use of the fact that the variance is small when the mean is close to zero or one. Another direction is improving the worst-case regret to match the lower bound of $\Omega(\sqrt{Kn})$ that we will see next week. This requires a modification of the confidence level and a more complicated analysis.

Notes

Note 1: Here we argue that there is no loss in generality in assuming that the rewards experienced satisfy (6). Indeed, let $T' = (A'_1, X'_1, \dots, A'_n, X'_n)$ be any sequence of random variables satisfying that $A'_t = f_t(A'_1, X'_1, \dots, A'_{t-1}, X'_{t-1})$ and that for any $U \subset \mathbb{R}$ open interval

$$
\mathbb{P}\left(X'_t \in U \,|\, A'_1, X'_1, \dots, A'_{t-1}, X'_{t-1}, A'_t\right) = P_{A'_t}(U)\,,
$$

where $1 \leq t \leq n$. Then, choosing $(Z_{i,s})_s$ as described in the paragraph before (6), we let $T = (A_1, X_1, \dots, A_n, X_n)$ be such that $A_t = f_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ and X_t be so that it satisfies (6). It is not hard to see then that the distributions of T and T' agree. Hence, there is indeed no loss of generality by assuming that the rewards are indeed generated by (6).

Note 2: The view that n rewards are generated ahead of time for each arm and the algorithm consumes these rewards as it chooses an action was helpful in the proof as it reduced the argument to the study of averages of independent random variables. The analysis could also have been done directly without relying on the "virtual" rewards $(Z_{i,s})_s$ with the help of martingales, which we will meet later.
A third model of how X_t is generated could have been that $X_t = Z_{A_t,t}$. We will meet this "skipping model" later when studying adversarial bandits. For the stochastic bandit models we study here, all these models coincide (they are indistinguishable in the sense described in the first note above).

Note 3: So is the optimism principle universal? Does it always give good algorithms, even in more complicated settings? Unfortunately, the answer is no. The optimism principle leads to reasonable algorithms when using an action gives feedback that informs the learner about how much the action is worth. If this is not true (i.e., in models where you have to choose action B to learn about the rewards of action A, and choosing action A would not give you information about the reward of action A), the principle fails! (Why?) Furthermore, even if all actions give information about their own value, the optimistic principle may give rise to algorithms whose regret is overly large compared to what could be achieved with more clever algorithms. Thus, in a way, finite-armed stochastic bandits is a perfect fit for optimistic algorithms. While the more complex feedback models may not make much sense at the moment, we will talk about them later.

References

The idea of using upper confidence bounds appeared in '85 in the landmark paper of Lai and Robbins. In this paper they introduced a strategy which plays the leader of the "often sampled" actions except that for any action j in every Kth round the strategy is checking whether the UCB index of arm j is higher than the estimated reward of the leader. They proved that this strategy, when appropriately tuned, is asymptotically unimprovable the same way UCB as we defined it is asymptotically unimprovable (we still owe the definition of this and a proof, which will come soon). The cleaner UCB idea must have been ready to be found in ’95 because Agrawal and Katehakis & Robbins discovered this idea independently in that year. Auer et al. later modified the strategy slightly and proved a finite-time analysis.

- Tzu L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules, 1985
- Rajeev Agrawal. Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem, 1995
- Michael N Katehakis and Herbert Robbins. Sequential choice from several populations, 1995
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem, 2002

Share this:

Related

Bandits: A new beginning September 4, 2016 In "Bandits"	Stochastic Linear Bandits and UCB October 19, 2016 In "Bandits"	Lower Bounds for Stochastic Linear Bandits October 20, 2016 In "Bandits"

Posted on September 18, 2016 by Tor Lattimore in Bandits, Finite-armed bandits. Bookmark the permalink.

← First steps: Explore-then-Commit

Optimality concepts and information theory →

17 thoughts on “The Upper Confidence Bound Algorithm”

1. *Aaron* says:
October 15, 2016 at 2:10 am
Sorry that I just can’t understand
why at a confidence level of 1−δ this guarantees that μi is smaller than μ1 and if the above inequality did not hold (after inequality (3))

Reply

2. *Tor Lattimore* says:
October 17, 2016 at 11:58 am
The statement is slightly informal, but roughly $\hat{\mu}_i(t-1)$ is an empirical estimate of μ_i based on $T_i(t-1)$ samples. Since we assumed that the rewards are 1-subgaussian we know that for $T_i(t-1) = u$ that $\mathbb{P}\left(\hat{\mu}_i + \sqrt{\frac{2}{u}\log\left(\frac{1}{\delta}\right)} \leq \mu_i\right) \leq \delta$. The informality comes from the fact that $T_i(t-1)$ is usually also a random variable, which makes the analysis a little trickier, but does not change much the intuition.

Note that we treat the concentration of subgaussian random variables in a previous post (http://banditalgs.com/2016/09/14/first-steps-explore-then-commit/)

Reply

- *Sidak Pal Singh* says:
  January 20, 2018 at 9:33 pm
  Hi! Thanks for this amazing series of blogs.

  I have a small question. I think it might be better to say that the value of 1-δ is called the confidence level (instead of saying δ is the confidence level). δ is sort of like the upper bound on the probability of error that we allow. Also, using 1-δ will possibly make it more consistent with confidence interval terminology used in statistics. Please correct me if I am wrong. Thanks! 🙂

  Reply

  - *Csaba Szepesvari* says:
    February 10, 2018 at 2:44 am
    You are right! I'll fix this:) Thanks for the comment!

    Reply

3. *Chris* says:
   November 9, 2016 at 8:30 pm
   Hey Tor, do you mind expanding on inequality (7)? I sort of understand that the first and second terms in the inequality represent the events that the UCB index for the optimal arm is "too low" and that the UCB index for the sub-optimal arm is "too high", respectively. However, I'm confused as to how either of these events imply that the UCB index for the sub-optimal arm is less than the UCB index for the optimal arm in a given round.

   Reply

   -

     Csaba Szepesvari says:
     November 21, 2016 at 1:27 pm
     Hi,
     I'll dub as Tor. The implication is easiest to see by inverting things. We want to see that $A_t = i$ implies that UCB_i is high, or UCB_1 is low. Well, if UCB_i was low and UCB_1 was high, then arm 1 would have been preferred to arm i, so it must be that if arm i is selected then either UCB_i is high, or UCB_1 is low. Does this make sense?
     Cheers,
     Csaba
     PS: Sorry for the slow reply.

     Reply

- *Janos Divenyi* says:
  October 30, 2017 at 8:03 am
  I am confused too about this inequality

  In the first case (UCB_1 is low) the left hand side of the inequality is the index of arm 1, right? Shall we not use $T_1(t-1)$ in the denominator of the square root?

  Reply

  - *Janos Divenyi* says:
    October 30, 2017 at 8:05 am
    And why we compare the indices to $\mu_1-\varepsilon$? Why not simply $\mu_1?

    Reply

    -

      Csaba Szepesvari says:
      January 12, 2018 at 9:28 am
      Another good point. With $\varepsilon=0$, what we would need to bound is the probability of the index of the optimal arm smaller than the optimal mean. The way the index is defined, if the optimal arm is pulled a fixed, say, s number of times, this probability happens to be constant. This is too large; it would render the bound vacuous. I hope this helps. And sorry for the slow response; somehow I did not get a notification of the comment, or I just missed it.

  -

    Csaba Szepesvari says:
    January 12, 2018 at 9:20 am
    Hmm, I have not caught this before. True. We should have had $T_1(t-1)$ in the denominator. I have corrected this now, thanks!

    Reply

4. *Xiang Wang* says:

May 8, 2017 at 7:48 pm

Hi, in the UCB Simplified Regret, does the universal constant C rely on suboptimality gaps? I suspect that if the suboptimality gaps are not bounded, we cannot find such a constant C.

Also, in the UCB regret (3), should the last constant be '1', but not '3'?

Thank you!

Reply

- *Tor Lattimore* says:
  May 9, 2017 at 12:30 am
  Hi Xiang. You're right on all counts. See the updated theorem for the (hopefully) correct statement of this kind of result.

  Thanks for pointing out the bugs!

  Tor

  Reply

  - *Xiang Wang* says:
    May 9, 2017 at 4:15 am
    Great thanks for the quick response!

    However, after the correction, there still exists a flaw in the final distribution-free bound for UCB. This bound also requires the suboptimality gaps be bounded, right?

    Reply

    - *Csaba Szepesvari* says:
      August 21, 2017 at 3:48 pm

      Hi Xiang! Sorry for the slow response. Where is the bug? The universal constant just relies on bounding constant + $\log \log n$ by $C \log n$, it seems to me.

      Reply

5. *Hairi* says:
   June 11, 2017 at 10:09 pm
   Hi, professor, when reasoning arm 1 is optimal and not the arm j (j \neq 1), we say that we have a 1-\delta level of confidence. But, should we also say arm 1 is optimal

compared to all the other K-1 arms, so the confidence level would be (1-\delta)^{K-1}?

Reply

-

  Csaba Szepesvari says:
  August 21, 2017 at 3:50 pm
  Where is this? The trick we use is that we bound the *expected* number of pulls of suboptimal arms. Hence, each suboptimal arm is compared to the optimal arm, one by one, separately, avoiding the need to argue about multiple suboptimal arms at the same time. I hope this clarifies things.

  Reply

6. *Tim* says:
   February 24, 2018 at 11:17 am
   Hey!
   It would be a big help for me if you could explain where the infimum condition in the equation for the UCB regret (Eq. (4)) comes from. It is comprehensible in the sense that one wants to keep the upper bound as small as possible, but why is the range of epsilon chosen like this?

   By the way, thanks for creating this blog – I think this is a really nice medium to get into this topic!

   Reply

Leave a Reply

Your email address will not be published.

Comment

□

Notify me of followup comments via e-mail

Name

Email

Website

☑ Sign me up for the newsletter!

Post Comment

☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

Pages

About

View all posts

Recent Posts

Bandit tutorial slides and update on book

Adversarial linear bandits and the curious case of the unit ball

Adversarial linear bandits

Sparse linear bandits

Ellipsoidal Confidence Sets for Least-Squares Estimators

Recent Comments

Park on Optimality concepts and information theory

Csaba Szepesvari on Optimality concepts and information theory

Park on Optimality concepts and information theory

Csaba Szepesvari on Optimality concepts and information theory

Park on Optimality concepts and information theory

Archives

February 2018

November 2016

October 2016

September 2016

August 2016

Categories

Bandits

Finite-armed bandits

Lower bound

Probability

Meta

Log in

Entries RSS

Comments RSS

WordPress.org

Newsletter

Email address:

Your email address

First Name

Last Name

Sign up

Form action

● Subscribe

○ Unsubscribe

Proudly powered by WordPress

Clean Content by On Edge
(Way of the future)