doi:10.20944/preprints202004.0426.v1

UCB with an optimal inequality

Mark Burgess MARK.BURGESS@ANU.EDU.AU
Australian National University, Canberra

Abstract

Upper confidence bound multi-armed bandit algorithms (UCB) typically rely on concentration inequalities (such as Hoeffding's inequality) for the creation of the upper confidence bound. Intuitively, the tighter the bound is, the more likely the respective arm is or isn't judged appropriately for selection. Hence we derive and utilise an optimal inequality.

Usually the sample mean (and sometimes the sample variance) of previous rewards are the information which are used in the bounds which drive the algorithm, but intuitively the more information that taken from the previous rewards, the tighter the bound could be. Hence our inequality explicitly considers the values of each and every past reward into the upper bound expression which drives the method.

We show how this UCB method fits into the broader scope of other information theoretic UCB algorithms, but unlike them is free from assumptions about the distribution of the data, We conclude by reporting some already established regret information, and give some numerical simulations to demonstrate the method's effectiveness.

Keywords: Bandit Algorithm; Upper Confidence Bounds; Kullback-Leibler divergence

1. Introduction

The multi-armed bandit problem (MAB) is a classic example of a scenario that encodes the general dynamic trade-off between exploration and exploitation in artificial intelligence systems.

The scenario consists of a repeated single decision, who's various choices lead the learner to stochastic rewards. Over time more information is gathered about the rewards of the decisions that are made, until eventually the learner is in a position of choosing the machine that yields the highest average reward to achieve the greatest long-term expected payoff. The visceral picture of the situation is of a gambler choosing which poker-machine is best to play over some time frame.

There are several methods of learning in this situation such as the classic ϵ-greedy algorithm (and other ϵ type methods), Softmax technique (aka. Boltzmann exploration), Pursuit Algorithms, Thomson sampling and Reinforcement Comparison (aka Gradient) methods; but we focus specifically on the UCB family of bandit algorithms.

Upper confidence bound (UCB) bandit algorithms are a popular class of methods where the bandit arm (ie. decision) with the highest upper confidence bound (for some confidence level) on its mean is selected each turn, Lai and Robbins (1985). Typical methods of generating confidence bounds rely on various concentration inequalities; and these function as methods of generating conservative confidence intervals that inform the selection process.

One of the interesting developments among UCB methods is the improvements which are possible by incorporating tighter concentration inequalities. Classic concentration inequalities such as Hoeffding's inequality, give confidence interval widths on the mean, based simply on the number of reward samples taken and the size of the support of the data. Whereas some newer concentration

doi:10.20944/preprints202004.0426.v1

inequalities include reward variance information to create a tighter bound. For instances see Maurer and Pontil (2009); Audibert et al. (2009); Carpentier et al. (2011).

It is understood that the more information about the rewards are used in the creation of an upper confidence bound, the more tightly that bound could be created and the more powerful the resulting UCB method might be. Indeed, there has been experimentation and positive results incorporating more information (such as variance information) into the UCB method; Mukherjee et al. (2018).

This leads naturally to the question of how to develop more powerful concentration inequalities. Unfortunately there are an array of powerful and divergent approaches - including is the historical class of Chernoff bounds (including Hoeffding's inequalities, Bernstein's inequality, Bennett's inequality etc.) which arise from an application of Markov's inequality (otherwise more directly leading to Chebyshev's inequality) about the moment generating function, there are developments surrounding Entropy methods (Boucheron et al. (2003)), and the use of the Efron-Stein inequality (or broader 'jackknife' method - Efron and Stein (1981)). What is notable is that all of these techniques use some loosening approximations to develop their respective concentration inequality expressions.

However, some developments (particularly around Optimal Uncertainty Quantification (OUQ)) has motivated research into developing optimal concentration inequalities - and these don't use any approximations whatsoever - and are optimally tight as a result from finding the worst-case probability density functions directly; Owhadi et al. (2013); Han et al. (2013). We take inspiration from these approaches, and in this paper we develop an optimal concentration inequality and show its performance in the context of the UCB method.

The resulting probability bound not only incorporates information about the sample mean and sample variance but all sample central moments implicitly, as it accounts for all the specific values (and their multiplicities) of all historic rewards from the arms. Our bound directly incorporates all the reward values (without loosening simplifications) into the creation of a probability bound that is perfectly tight and which utilises all available information available to the learner.

The probability bound that we develop is tight over all possible probability distributions of rewards on bounded support, and we show how it fits into a framework of UCB algorithms developed by Cappé et al. (2013) called KL−UCB .

We finish by giving a quick numerical case study where we compare the effectiveness of our method against some others.

The paper is divided into the following sections: In section 2 we introduce the kind of probability bound we intend to derive, in Sections 3 and 4 we give summary statements of the optimal bound and a UCB algorithm that uses it, in Section 4.1 we show how the UCB algorithm is identical to KL−UCB in context of Bernoulli data, where-after in Section 4.2 we talk about the relationship with KL−UCB more generally, and Section 5 concludes with some numerical simulations.

The contributions of this paper are:

- We derive an optimal probability inequality for the mean of a distribution from the sample values, via the empirical likelihood method, and build it into a UCB algorithm.
- We show how that resulting algorithm is a computable, performant and lossless instance of the KL−UCB framework that is free from the assumption that the rewards belong to any parametric family of distributions.

doi:10.20944/preprints202004.0426.v1

2. Making an optimal concentration inequality

Historical UCB algorithms have relied on the usage of concentration inequalities such as Hoeffding's inequality. And these concentration inequalities can be interpreted as analytic unconditioned probability statements about the relationship between sample statistics and population statistics or alternatively about sample statistics as conditioned by population statistics, but not conversely (as is sometimes misunderstood) as about population statistics conditioned by sample statistics. (see appendix C)

We can see in Appendix B (or derivations elsewhere), that the derivation of Hoeffding's inequality begins with the assumption of the mean, and then proceeds to develop a probabilistic bounds the sample mean; and this is a legitimate process of develop bounds on the sample mean conditioned on the assumption of the population mean.

To develop an optimal concentration inequality to replace Hoeffding's inequality in UCB algorithms it is therefore legitimate that we ask the same question that Hoeffding's inequality answers: for a specific possible mean of the data distribution, what is the maximum probability of receiving the relevant sample statistics? And the way in which we answer this question is by solving for the maximising probability density function directly.

3. An optimal concentration inequality for finite values

For a specific possible mean of the data distribution, what is the maximum probability of receiving the relevant sample statistics?

We begin with the assumption that the distribution of the data only features a finite number of values. And since we are interested in using more than just the sample mean (and sample variance) we ask the question about receiving all sample information.

Our question therefore becomes: what probabilities can the values in our distribution have (consistent with a specified distribution mean) to maximise the probability of receiving all the samples that we did (with their values and multiplicities)? This question is fundamentally a task of maximising an Empirical Likelihood subject to a constraint, and hence an optimisation problem.

For this problem, the qualities of its solution can be intuited, particularly that non-zero probability must be allocated to all the values of the samples (or otherwise the likelihood would straightforwardly be zero), and then perhaps one other extreme point may have some probability to balance the distribution with the specified mean most effectively. A graph illustrating this intuition is shown in Figure 1.

In Theorem 4 of the appendix A we solve the optimisation problem analytically to show exactly what probabilities the values have, and when and where the other extreme point occurs. This process of maximising an empirical likelihood is a well known statistical technique - called the Empirical Likelihood method, and our derivation is reminiscent of this classic technique; see Owen (1988).

4. The optimal concentration inequality for finite support

By maximising the empirical likelihood of receiving the samples, we saw that there was an upper limit on the number of values in the distribution that will be allocated non-zero probability, specifically the number of sample values plus one. And this fact remains true irrespective of how many values we limit our distribution to support, and irrespective of how those values are arranged or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2020 doi:10.20944/preprints202004.0426.v1

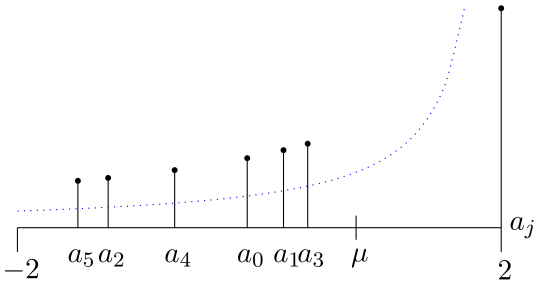


Figure 1: For distributions bounded between -2 and 2, with a specified mean μ, the distribution that maximises the likelihood of drawing sample values $a_0, \ldots, a_5$ (each with single multiplicity) including the unsampled value 2. We also see the hyperbolic weighting of probabilities centered at 2.

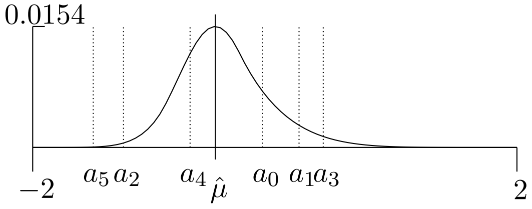


Figure 2: For distributions bounded between -2 and 2 and with the same sample values $a_0, \ldots, a_5$ as in figure 1, the maximum probability that the mean $\mu = x$ consistant with thoes sample values per Theorem 1. We see that the maximum height is at the sample mean $\hat{\mu}$ and is very small, and we can see that there is a slightly heavier right side tail as the data samples are left heavy.

densely packed. Therefore we propose that the same behaviour occurs in the limit of continuous distributions.

Hence the continuous distribution that maximises the likelihood of drawing specific samples (for a specific distribution mean) is given by a finite summation of delta functions.

If we assume that the distribution only contains values in a specific range (ie. has finite support) then the maximising likelihood is given by the following theorem (see Appendix A for full derivation):

Theorem 1 *Over all distributions (bounded between d^+ and d^-) that have mean $\mu = x$, the likelihood of drawing samples with values $s_1, s_2, s_3, \ldots, s_n$ with multiplicities $n_1, n_2, n_3, \ldots, n_n$ is least upper bounded by a positive $p_{max}(s_{1..n}, n_{1..n}|\mu = x)$.*
if the sample mean equals the mean (ie. $\hat{\mu} = \sum_i n_i s_i / (\sum_i n_i) = \mu$) then:

$$p_{max}(s_{1..n}, n_{1..n}|\mu = x) = \frac{N!}{\prod_j (n_j!)} \prod_{i=1}^{n} \left(\frac{n_i}{N}\right)^{n_i} \tag{1}$$

otherwise:

$$p_{max}(s_{1..n}, n_{1..n}|\mu = x) = \frac{N!}{\prod_j (n_j!)} \prod_i \left(\frac{n_i}{N} \frac{G}{G - (s_i - x)} \right)^{n_i} \tag{2}$$

Where G is constrained such that, $G < \min_i(s_i - \mu)$ *or* $G > \max_i(s_i - \mu)$ *and is given by:*

$$\begin{cases} d^+ - x & \text{if } \sum_i \frac{n_i(s_i - x)}{s_i - d^+} \geq 0 \\ d^- - x & \text{if } \sum_i \frac{n_i(s_i - x)}{s_i - d^-} \geq 0 \\ \text{solution of } \sum_i \frac{(s_i - x)n_i}{(s_i - x) - G} = 0 & \text{otherwise} \end{cases}$$

As this function p_{max} is parameterised by the mean value $\mu = x$ it is possible to solve and plot it against x. An example plot against x is shown in figure 2 where we can see that the probability profile of the mean indicates that it is most likely about the sample mean, which is what we would expect.

One inconvenience of using this inequality in UCB methods is that the maximum height of the probability profile for μ does not reach a maximum of 1. In the traditional case of Hoeffding's inequality it does reach a maximum of 1 since it is always perfectly possible that the data distribution has a variance of zero, in which case the mean and sample mean must coincide. Whereas our inequality considers the spread of sample points more directly, and it is true, that the probability that there is any spread in sample points *and* (or *given that*) the mean has any particular value (or is in any range of values) must be less than one.

However for practical purposes (and this is quite optional) it is convenient to do some scaling to make the probability profile have a height of 1, and we incorporate this scaling in subsequent Theorem 4. Additionally, we are primarily interested in the region which the mean exceeds the sample mean ($\mu \geq \hat{\mu}$) by some factor, and one of the easy things to notice is that the maximum point of the profile corresponds to the case $\mu = \hat{\mu}$, and that for $\mu > \hat{\mu}$ that the function is decreasing. And so we convert our function to be an inequality over a range of mean values greater than a parameter x, with $\mu \geq x \geq \hat{\mu}$:[1]

$$p_{max}(s_{1..n}, n_{1..n}, \mu \geq x) = \int_{y \geq x} p_{max}(s_{1..n}, n_{1..n}|\mu = y) p(\mu = y) dy$$

$$\leq \max_{y \geq x} p_{max}(s_{1..n}, n_{1..n}|\mu = y) = p_{max}(s_{1..n}, n_{1..n}|\mu = x)$$

Thus we have converted our inequality which is conditioned by a specific value of the mean, into an unconditioned inequality that is over the range of mean values that we are interested in. Thus we can restate our Theorem for unit-scaled probability over range of mean values - which coincidentally doesn't need a lower bound anymore:

Theorem 2 *Over all distributions bounded above by* d^+*, that have mean* $\mu \geq x$*, the unit-scaled likelihood of drawing samples with values* $s_1, s_2, s_3, \ldots, s_n$ *with multiplicities* $n_1, n_2, n_3, \ldots, n_n$*, given that* $x > \hat{\mu}$*, is least upper bounded by:*

$$p^{scaled}_{max}(s_{1..n}, n_{1..n}, \mu \geq x) = \prod_i \left(\frac{G}{G - (s_i - x)} \right)^{n_i} \tag{3}$$

[1] In the same line of reasoning as in Appendix C

Where G is constrained such that $G > \max_i(s_i - \mu)$, and is given by:

$$\begin{cases} d^+ - x & \textit{if } \sum_i \frac{n_i(s_i-x)}{s_i-d^+} \geq 0 \\ \textit{solution of } \sum_i \frac{(s_i-x)n_i}{(s_i-x)-G} = 0 & \textit{otherwise} \end{cases}$$

Working with this measure p^{scaled}_{max} in the context of UCB suggests an like Algorithm 1:

Algorithm 1 UCB operating with p^{scaled}_{max} - `UCBOI`

```
Require: A decreasing function $g : \mathbb{N} \to \mathbb{R} < 1$
Pull each of arms $\{1, \ldots, K\}$ once
thus setting number of pulls for for each arm $a$, $N_a(t) = 1$
for $t = K$ to $T - 1$ do
    for each arm $a$ compute its average historic reward $\hat{\mu}_a$,
    and compute the $x_a > \hat{\mu}_a$ such that:
        $$p^{scaled}_{max}(s^a_{1..n}, n^a_{1..n}, \mu_a \geq x_a) = g(t)$$
    pull an arm $A_{t+1} \in \arg\max_{a \in \{1,\ldots,K\}} x_a$
end for
```

Let us give an example application:

4.1. The case of Bernoulli data

If we are dealing with Bernoulli data then things become a bit simpler and Theorem 4 reduces to:

Theorem 3 *for Bernoulli distributions that have mean $\mu \geq x$, the scaled likelihood of drawing n_0 samples of value 0 and n_1 samples of 1, if $x > \hat{\mu}$ (with $N = n_0 + n_1$), is least upper bounded by:*

$$p^{scaled}_{max,bernoulli}(n_0, n_1, \mu \geq x) = \left(\frac{1-x}{1-\hat{\mu}}\right)^{n_0} \left(\frac{x}{\hat{\mu}}\right)^{n_1} \tag{4}$$

Proof for $s_0, s_1 = 0, 1$ solving $\sum_i \frac{(s_i-x)n_i}{(s_i-x)-G} = 0$ gives $G = \frac{\mu(1-x)}{x-\hat{\mu}}$ which is put into equation 3. ■

As Algorithm 1 runs with a process of finding x such that $p^{scaled}_{max}(\ldots, \mu \geq x) = g(t)$, this is much the same thing as finding x such that $\frac{-1}{N} \log p^{scaled}_{max}(\ldots, \mu \geq x) = \frac{-1}{N} \log g(t)$ and

$$\frac{-1}{N} \log p^{scaled}_{max,bernoulli} = \hat{\mu} \log\left(\frac{\hat{\mu}}{x}\right) + (1-\hat{\mu}) \log\left(\frac{1-\hat{\mu}}{1-x}\right)$$

And this expression might look slightly familiar ...
as it is the Kullback-Leibler divergence between the Bernoulli distributions suggested by $\hat{\mu}$ and that is consistent with $\mu = x$. And if we ran Algorithm 1 with Bernoulli data and $g(t) = (t(\log(t))^3)^{-1}$ for $t \geq 3$ with $g(1) = g(2) = g(3)$ we would be running a process equivalent to an instance of Cappé et al. (2013)'s `KL-UCB` algorithm which has been demonstrated to be quite performant. If we denote the divergence between Bernoulli distributions with means μ and μ' as:

$$d_{BER}(\mu, \mu') = \mu \log\left(\frac{\mu}{\mu'}\right) + (1-\mu) \log\left(\frac{1-\mu}{1-\mu'}\right)$$

and if $\mu^\star$ is the maximum mean of the arms (ie. the optimal one) then the regret of the algorithm instance can be characterised by Cappé et.al's proof that the expected number of pulls of any sub-optimal arm a is bounded by:

$$\mathbb{E}[N_a(t)] \leq \left(4e + \tfrac{3}{d_{BER}(\mu_a,\mu^\star)}\right)\log(\log(T)) + \tfrac{\log(T)}{d_{BER}(\mu_a,\mu^\star)} + \tfrac{2(\log(\mu^\star(1-\mu_a)/(\mu_a(1-\mu^\star))))^2}{(d_{BER}(\mu_a,\mu^\star))^2} + 6 \\ + \tfrac{\sqrt{2\pi}\log(\mu^\star(1-\mu_a)/(\mu_a(1-\mu^\star)))}{(d_{BER}(\mu_a,\mu^\star))^{\frac{3}{2}}}\sqrt{\log(T) + 3\log(\log(T))}$$

4.2. A Kullback-Leibler reformulation

The occurrence of the Kullback-Leibler divergence is not coincidence, and if we rearrange equation 3 we get:

$$-\frac{1}{N}\log p^{scaled}_{max}(s_{1..n}, n_{1..n}, \mu \geq x) = \sum_i \frac{n_i}{N}\log\left(\frac{\frac{n_i}{N}}{\frac{n_i}{N}\frac{G}{G-(s_i-x)}}\right) \tag{5}$$

Which is the Kullback-Leibler divergence between our derived distribution that maximises likelihood of the samples given a specific mean x, and the distribution that is most consistent with those samples, and maximises likelihood irrespective of a specific mean (and occurs when $\hat{\mu} = \mu$).

This then raises the question of exactly where our algorithm and KL−UCB coincide - and the answer is that we believe our algorithm is an unexplored special instance of their more abstract and general framework. In their more abstract framework they consider a mapping (or a projection operator) $\Pi_{\mathcal{D}}$ which maps the empirical distribution of the historic rewards of a given arm $\hat{v}_a(t) = \frac{1}{N_a(t)}\sum_{s=1}^{t}\delta_{Y_s}\mathbb{I}_{A_s=a}$ (which is a sum of Dirac deltas corresponding to rewards) to an element of class $\mathcal{D}$, which is the specific class of distributions that they are considering and comparing via Kullback-Leibler divergences (and potentially needs appropriate selection). In the application cases that they consider, this mapping mostly occurs via the sample mean - with loss of potential information. Whereas we have created a rather unique case where the this mapping can effectively be the identity - and is therefore lossless as it utilises all available information.

Not only is our algorithm a lossless instance of the KL−UCB framework, but is is also free from any assumptions about the distribution of the data (excepting that is bounded above) whereas all other promoted cases of KL−UCB that are developed into a computable algorithms, assume that the data distribution must be one of a parameterised family of distributions (eg. Bernoulli, Exponential, Poisson, etc) within which the Kullback-Leibler divergences can be computed and expressed. This is a restriction which we seem to have bypassed.

However since our UCB algorithm is an instance of KL−UCB it also inherits some favourable regret bounds, particularly that regret bounds can be derived from the fact that the number of pull of sub-optimal arm a is bounded by:

$$\mathbb{E}[N_a(t)] \leq \frac{\log(T)}{\mathcal{K}_{inf}(\nu, \mu^\star)}(1 + o(1))$$

where $\mathcal{K}_{inf}(\nu, \mu) = \inf\{KL(\nu, \nu') : \nu' \in \mathcal{D} \text{ and } E(\nu') > \mu\}$ and reads as the infimum of Kullback-Leibler divergences between the arm distribution ν and distributions in the model $\mathcal{D}$ that have expectations larger than μ. Which in our case is given by equation 5.

This expression has been identified as a formula for lower and upper bounds on optimal asymptotic regret of UCB algorithms, see Maillard et al. (2011). Additionally there are some other regret expressions that our algorithm tacitly inherits, such as given by Garivier et al. (2018), as the analysis and performance of information theoretic UCB algorithms is an area of research. However in order witness how effective our algorithm is in practice, it is simple enough to run some experiments.

5. Concluding experiments

Let us conclude with some numerical experiments.[2] In Figure 3, we plot the average cumulative regret over bandit simulations with 5 arms. We considered:

- a UCB1 algorithm (developed in Auer et al. (2002)) which chooses the arm with the highest sample mean of n historic rewards plus $\sqrt{2\log(t)/n}$
- an untuned UCB-V algorithm (given in Audibert et al. (2009)) which chooses the arm with highest sample mean plus $\sqrt{2\hat{\sigma}^2\log(t)/n} + \sqrt{\log(t)/n}$ (where $\hat{\sigma}^2$ is the biased sample variance of past rewards)
- our own algorithm 1 with $g(t) = (t(\log(t))^3)^{-1}$ and
- also a method of random arm pulling.

In Figure 3 we considered different distributions of rewards for the arms. particularly, we considered:

1. Beta distributed rewards parameterised by α and β parameters were uniformly randomly generated between 0.1 and 3.1
2. We also considered uniform distributed rewards where the two poles of the reward were uniformly generated to be within the interval $[0, 1]$
3. We also considered Triangularly distributed rewards, where mode of the triangle was randomly generated to be within the interval $[0, 1]$ and where its bounds were $0, 1$
4. We also considered Trapezoidally distributed data, where the data spans the range $[0, 1]$ and the density is flat between two poles uniformly generated within the range $[0, 1]$

From these graphs we can see our method performs well across a range of possible data.

Acknowledgments

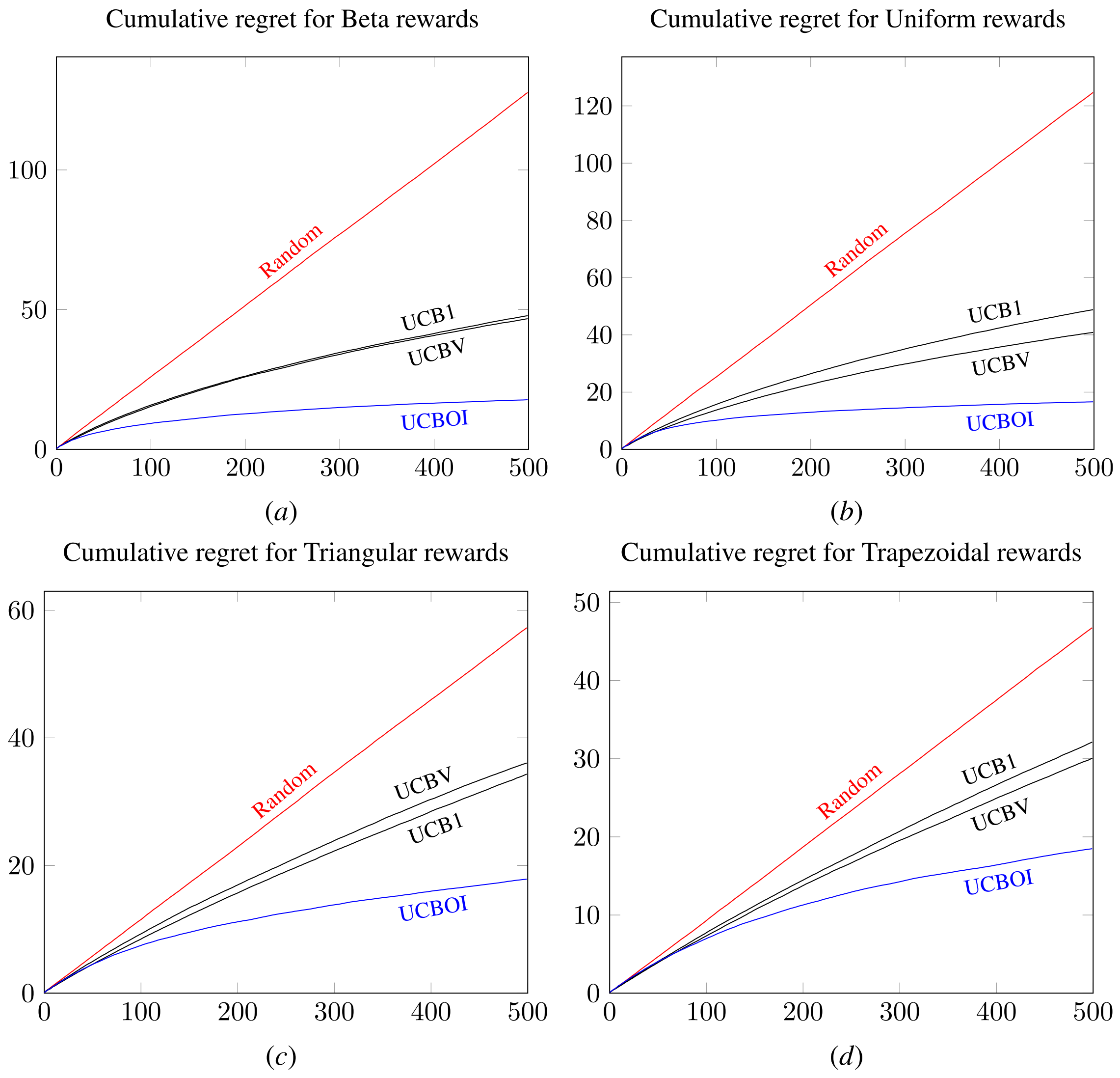
We would like to thank Sylvie Thiébaux, Paul Scott and Archie Chapman for their patience and support.

References

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, May 2002. ISSN 1573-0565.

[2] Github URL to be inserted here in the final version, for reproducible results



(*a*)

(*b*)

(*c*)

(*d*)

Figure 3: The cumulative average regret for simulations of 5-armed bandits with Beta (a), Uniform(b), Triangular(c), and Trapezoidal (d) distributed rewards, in each case the x-axis is the number of pulls of 4 different methods trialed, UCB1, UCB-V, UCBOI, and Random. details of the simulations are provided in Section 5

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31, 07 2003.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 06 2013.

doi:10.20944/preprints202004.0426.v1

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, ALT'11, pages 189–203, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-24411-7.

V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285 – 287, 1979.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 05 1981.

Aurélien Garivier, Hédi Hadiji, Pierre Ménard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv,stat.ML*, 1805.05071, 2018.

Shuo Han, Molei Tao, Ufuk Topcu, Houman Owhadi, and Richard Murray. Convex optimal uncertainty quantification. *SIAM Journal on Optimization*, 25, 11 2013.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22, 1985.

Odalric-ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multiarmed bandits problems with kullback-leibler divergences. In *In Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2011.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization, 2009.

Subhojyoti Mukherjee, K. P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-ucbv: An almost optimal algorithm using variance estimates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6417–6424, 2018.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75 (2):237–249, 06 1988.

H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal uncertainty quantification. *SIAM Review*, 55(2):271–345, 2013.

Appendix A. Derivation of the Optimum bound on the mean

In this section we will show with one (rather large) proof that among distributions constrained to finite values, that the maximising probability of a specific mean μ together with drawing a set of samples has a concise statement.

The proof encodes the intuition that probability will be assigned to each of the samples in proportion to their multiplicity, and weighted such as to make it consistent with the specific μ, in addition to possibly allocating probability to at-most one other extreme unsampled value - the image of this dynamic is easily seen in the image of Figure 1, but the proof is required and is here. Since the dynamic that probability only to value points which are sampled (plus potentially one other point) holds true irrespective of how many other points exist in the distribution or how densely they are packed, we also assert that this same dynamic occurs in the limit of continuous distributions.

We conduct the optimisation explicitly here, to yield a calculation methodology for the maximum likelihood as a function of the specified mean and sample values with their multiplicities:

Theorem 4 *Supposing that the distribution of our data has a mean $\mu = 0$ and takes unique finite values $a_1, a_2, a_3, \ldots, a_n$. if we sample those values with multiplicities $n_1, n_2, n_3, \ldots, n_n$ (some of which are non-zero, $N = \sum_{i=1}^{n} n_i$), then the probabilities of each of those values $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ which maximise the likelihood of our having sampled per those multiplicities is given by:*

$$\textit{for all } n_i > 0: \qquad \alpha_i = \frac{n_i}{N}\frac{G}{G - a_i}$$

and hence the likelihood of our sampling is:

$$p_{max}(\alpha_{1 \ldots n}) = \frac{N!}{\prod_{i=1}^{n}(n_i!)} \prod_{i=1}^{n} \left(\frac{n_i}{N}\frac{G}{G - a_i} \right)^{n_i}$$

Where there is a point G which is constrained to be in the region:

$$\Big(-\infty, \min(0, \min_{i, n_i > 0} a_i) \Big) \cup \Big(\max(0, \max_{i, n_i > 0} a_i), \infty \Big]$$

Where G is either equal to a data value a_j where $n_j = 0$ and where $\alpha_j = \sum_{i=1}^{n} \frac{n_i}{N}\frac{a_i}{a_i - a_j} > 0$, or G solves $\alpha^* = \sum_{i=1}^{n} \frac{n_i}{N}\frac{a_i}{a_i - G} = 0$.
(If there is no qualifying G value, the maximum likelihood is zero.)
All other values α_k with $n_k = 0$ (and for $k \neq j$ if it applies) are zero.
In this way there are no more non-zero values of $\alpha_{i=1,\ldots,n}$ then there are non-zero multiplicities of $n_{i=1,\ldots,n}$ plus (potentially) one.

Proof The probability of drawing sample values $a_{1 \ldots n}$ with multiplicities $n_{1 \ldots n}$ is given by:

$$p(\alpha_{1 \ldots n}) = \frac{N!}{\prod_{i=1}^{n}(n_i!)} \prod_{i=1}^{n} \alpha_i^{n_i}$$

Which is the optimisation objective function we seek to maximise (is the probability of drawing the samples in any specific sequence multiplied by the combinatorial number of ways that sequence of draws could occur) Subject to the constraints that the population has a mean of zero, that the probabilities add to one, and the probabilities are between zero and one:

$$\sum_{i=1}^{n} a_i \alpha_i = 0, \qquad \sum_{i=1}^{n} \alpha_i = 1, \qquad \forall i \quad 0 \leq \alpha_i \leq 1$$

Now the constraint that each of the probabilities are less than one is straightforwardly redundant (as the probabilities are non-negative and add to one). To make the constraint that the probabilities are non-negative redundant, we set $\beta_i^2 = \alpha_i$ and use β_i in the algebra instead. And since the n and n_i are constant, to make the optimisation more tractable we consider maximising the substitute function:

$$q(\alpha_{1...n}) = \log\left(\frac{\prod_{i=1}^n (n_i!)}{N!} p(\alpha_{1...n})\right) = 2\sum_{i=1}^n n_i \log(\beta_i)$$

Subject to re-written constraints that:

$$\sum_{i=1}^n a_i\beta_i^2 = 0, \quad \text{and} \quad \sum_{i=1}^n \beta_i^2 = 1 \tag{6}$$

If λ_1 and λ_2 are the respective Lagrange multipliers associated with these two constraints respectively then the KKT conditions associated with this optimisation problem are:

$$\text{for } \ i \in 1,\dots,n \qquad \frac{n_i}{\beta_i} - \lambda_1 a_i \beta_i - \lambda_2 \beta_i = 0$$

By assuming that for all samples points with non-zero multiplicity (ie $n_i > 0$) that the corresponding $\beta_i \neq 0$ (since if such a $\beta_i = 0$ for multiplicity points then the objective would trivially be zero), we get:

$$\begin{aligned} &\text{for all } \ n_i > 0 \qquad \beta_i^2(\lambda_1 a_i + \lambda_2) = n_i \\ &\text{for all } \ n_i = 0 \qquad \beta_i(\lambda_1 a_i + \lambda_2) = 0 \end{aligned}$$

Therefore, eliminating the possibility of imaginary β_i (and arbitrarily assuming β_i are positive) gives:

$$\begin{aligned} &\text{for all } \ n_i > 0 \qquad \beta_i = \sqrt{\frac{n_i}{\lambda_1 a_i + \lambda_2}} \quad \text{and} \quad \lambda_1 a_i + \lambda_2 > 0 \\ &\text{for all } \ n_i = 0 \qquad \beta_i = 0 \quad \text{or} \quad \lambda_1 a_i + \lambda_2 = 0 \end{aligned}$$

And since each of the a_i are distinct then at most one of these β_i for which $n_i = 0$ can be nonzero. And this demonstrates the third clause of our proof (that there are at most as many non-zero α's as there are non-zero multiplicities plus one).
Now there is a split in the path, particularly if λ_1 is zero or not zero.

Case 1: (If $\lambda_1 = 0$) - then we have an easy time, then $\lambda_2 > 0$ (as we assume there are some points with non-zero multiplicity) and hence $\beta_i = 0$ for any points with zero multiplicity, and for those with positive multiplicity have $\beta_i = \sqrt{n_i/\lambda_2}$; substituting into our constraints (6) gives: $\lambda_2 = N$ and hence $\hat{\mu} = \sum_{i=1}^n a_i n_i/N = 0$ Which identifies the convenient case where the average of our samples and population mean coincide, and hence the constraint that the mean is zero bears no pressure on the solution and would be the result if that constraint did not exist (in this way this solution is maximal), and where:

$$p_{max}(\alpha_{1...n}) = \frac{N!}{\prod_{i=1}^n (n_i!)} \prod_{i=1}^n \left(\frac{n_i}{N}\right)^{n_i} \tag{7}$$

End Case 1

However, if $\lambda_1 \neq 0$ then we can sensibly substitute a ratio G between our Lagrange multipliers as $\lambda_2 = -\lambda_1 G$ and hence:

$$\begin{aligned} &\text{for all } n_i > 0 \qquad \beta_i = \sqrt{\frac{n_i}{\lambda_1(a_i - G)}} \quad \text{and} \quad \lambda_1(a_i - G) > 0 \\ &\text{for all } n_i = 0 \qquad \beta_i = 0 \quad \text{or} \quad a_i = G \end{aligned}$$

from this it is easy to see that if $\lambda_1 > 0$ then $G < \min_{i:n_i>0} a_i$ and if $\lambda_i < 0$ then $G > \max_{i:n_i>0} a_i$. Now there is a second split in the path, if there exists a non-zero β_j where $n_j = 0$, or not:

Case 2: (If $\lambda_1 \neq 0$ and there exists an a_j where $n_j = 0$ and $\beta_j > 0$) - then $G = a_j$, then we can re-write our constraints (6) as:

$$0 = \sum_{i=1}^{n} \frac{a_i n_i}{\lambda_1(a_i - a_j)} + a_j\beta_j^2, \quad \text{and} \quad a_j = \sum_{i=1}^{n} \frac{a_j n_i}{\lambda_1(a_i - a_j)} + a_j\beta_j^2 \tag{8}$$

We subtract one from the other and to get $\lambda_1 a_j = -N$.

From this fact it must be true that $a_j \neq 0$ and also that λ_1 and a_j must have opposite sign, hence we can also propagate the earlier constraint, that if $a_j < 0$ then $a_j < \min_{i:n_i>0} a_i$ and if $a_j > 0$ then $a_j > \max_{i:n_i>0} a_i$. Now we can substitute our relation between λ_1 and a_j back in to our constraints (8) to get

$$\beta_j = \sqrt{\sum_{i=1}^{n} \frac{n_i}{N} \frac{a_i}{a_i - a_j}} \tag{9}$$

The requirement that β_j is not complex leads to real constraints on what could count as plausible candidate for a_j.

But supposing all these requirements are met for a candidate a_j (per our hypothesis in case 2), then our solution is then given by:

$$p_{max}(\alpha_{1\ldots n}) = \frac{N!}{\prod_{i=1}^{n}(n_i!)} \prod_{i=1}^{n} \left(\frac{n_i}{N} \frac{a_j}{a_j - a_i} \right)^{n_i} \tag{10}$$

End Case 2

And by exclusion between cases 1 and 2, there is only the remaining case:

Case 3: (If $\lambda_1 \neq 0$ and there does not exists an a_j where $n_j = 0$ and $\beta_j > 0$) - In this context the only non-zero β_i are those which have some multiplicity $n_i > 0$, and we can re-write our second constraint from (6) as:

$$1 = \sum_{i=1}^{n} \frac{n_i}{\lambda_1(a_i - G)} \qquad \text{and therefore} \qquad \lambda_1 = \sum_{i=1}^{n} \frac{n_i}{a_i - G}$$

From this it is clear that if $G < \min_{i:n_i>0} a_i$ then $\lambda_1 > 0$ and also if $G > \max_{i:n_i>0} a_i$ then $\lambda_i < 0$, and so our earlier constraint is redundant. Thus we can re-write our first constraint from (6) as:

$$0 = \sum_{i=1}^{n} \frac{a_i n_i}{a_i - G} \tag{11}$$

doi:10.20944/preprints202004.0426.v1

In this context, a suitable G may or may not exist, and solving for it is a bit more difficult, but insofar as it does exist we can progress.

Consequently:

$$p_{max}(\alpha_{1\ldots n}) = \frac{N!}{\prod_{i=1}^{n}(n_i!)} \prod_{i=1}^{n} \left(\frac{\frac{n_i}{a_i - G}}{\sum_{k=1}^{n} \frac{n_k}{a_k - G}} \right)^{n_i} \tag{12}$$

End Case 3

From these cases we notice that the results of the three cases overlap at their limits. Particularly if $\hat{\mu} = \sum_{i=1}^{n} a_i n_i / N = 0 = \mu$ (ie. if the sample mean is equal to the mean) then the conditions of case1 are satisfied, then the condition of case3 (equation 11) cannot be satisfied by any real G, but is satisfied by G in the limit that G goes to infinity (positive or negative), at which point λ_1 in case3 equals λ_1 in case1, and consequently the solutions of case1 and case3 (equations 7 and 12) are the same. Equally if $G \to \infty$ then β_j of case2 is zero (per equation 9 with $a_j = G$), and solutions of case2 (equation 10) match those of cases 3 and 1. Thus if the mean of the samples is zero, then all three cases overlap (or atleast at their limits), and the solution is comfortably given by equation (7), or equivalently by equations 10 or 12 with $a_j, G \to \infty$.

We also see that cases 2 and 3 overlap in a particularly interesting way, that case3 is case2 in the context where a hypothetical point $a^* = G$ is selected such as to allocate it with probability of zero. So for instance if hypothetical β^* in equation (9) is zero, then the associated $a^* = G$ satisfies the requirement for case3's equation 11 and then the expressions of maximum probability for cases 2 and 3 coincide (equations 10 and 12).

All of this simply identifies that cases 1 and 3 are actually just interesting limit instances of case2 - and that case2's equation 10 constitutes the more general expression of the solution; all of these considerations come to yield in the theorem statement.

This completes the proof. ■

In this Theorem statement, we have a G that is left uncharacterised, we will now show that a G that maximises the likelihood is unqiue.

A.1. Characterising the G

In Theorem 4 that we have a total expression of the solution conditioned on a selection of an appropriate G value, hence our analysis turns around the function that determines the existence of an appropriate G, specifically the function:

$$f(x) = \sum_i \frac{n_i a_i}{a_i - x} \tag{13}$$

And we are interested in characterising the points in the region $x > \max(0, \max(a_i))$ and $x < \min(0, \min(a_i))$. For simplicity we will only characterise the positive region $x > \max(0, \max(a_i))$, in the knowledge that it also implicitly characterises the negative region by symmetry (ie. the logic for negative region simply follows by reversing the signs of the a_i and x in every step).

Theorem 5 *For function (13) for arbitrary finite numbers a_i and corresponding positive integers n_i (some of which we assume are non-zero pairs), in the region $x > \max(0, \max(a_i))$. If there exists a positive a_i, then there exists exactly one finite zero-crossing point of f iff $\sum_i a_i n_i < 0$, and* *only points x greater than this zero-crossing point will have positive $f(x)$, if $\sum_i a_i n_i = 0$ then $f(x)$ will cross in the limit of $x \to +\infty$ and in all other cases $f(x)$ will be negative. Otherwise if there does not exist a positive a_i all points x in the region will be positive.*

Proof Supposing that there exists a positive a_i then there exists a maximally positive a_k and thence we are constrained $x > a_k$. In this case we consider the function $g(x) = xf(x)$ and identify that g will have the same zero crossing points as f and will be the same sign as f in our region of interest ($x > a_k$). We hence aim to show that if $\sum_i a_i n_i < 0$ then g has one zero crossing point and is positive for all points greater than this, and if $\sum_i a_i n_i \geq 0$ then there is no zero crossing point and there are no positive points in the region.

We begin by proving that g is monotonically increasing:

$$g'(x) = \frac{d}{dx}\left(\sum_i \frac{n_i a_i x}{a_i - x}\right) = \sum_i \frac{n_i a_i^2}{(a_i - x)^2} > 0$$

and that in our region of interest g begins at $-\infty$ and goes to $\sum_i a_i n_i$.

$$\text{ie.} \lim_{x \to a_k^+} g(x) \to -\infty \quad \text{and} \quad \lim_{x \to \infty} g(x) = -\sum_i a_i n_i$$

This is sufficient to prove that if $\sum_i a_i n_i < 0$ then there exists exactly one zero-crossing point of f and all the points more positive than this have f positive, otherwise there exists no zero-crossing points of f and f is entirely negative in the region $x > a_k$.

Conversly supposing there is not positive a_i, then in region $x > 0$, $f(x) = \sum_i \frac{n_i a_i}{a_i - x} > 0$ and all points are positive. ■

The result of this Theorem 5 (applied both positively and negatively) inform us that in the context of our big Theorem 4 that there are 5 cases.

1. that the sample mean $\hat{\mu} \propto \sum a_i n_i = 0$, therefore the only G that satisfies is ∞

2. that the sample mean $\hat{\mu} \propto \sum a_i n_i < 0$ and there does exist a positive a_k, hence there is a minimum satisfying G which is positive and all points more positive than it will also satisfy.

3. that the sample mean $\hat{\mu} \propto \sum a_i n_i < 0$ and there does not exist a positive a_k, hence all positive G values satisfy.

4. that the sample mean $\hat{\mu} \propto \sum a_i n_i > 0$ and there does exist a negative a_k, hence there is a maximum satisfying G which is negative and all points more negative than it will also satisfy.

5. that the sample mean $\hat{\mu} \propto \sum a_i n_i > 0$ and there does not exist a negative a_k, hence all negative G values satisfy.

The last piece of the puzzle is to prove that where there exists a selection of satisfying values of G, that it always results in a greater likelihood (ie. is more optimum) to choose the one most extreme (that is positive or negative wrt zero). The case where $\hat{\mu} \propto \sum a_i n_i = 0$ is trivial that $G \to \infty$ is the most extreme possible point, however it remains to be proven for the other cases. We will explicitly treat the case where $\sum a_i n_i < 0$ and hence there is a continuous range of positive G where $G > \max(0, \max_{i, n_i > 0} a_i)$ that satisfies requirements for Theorem 4; we do this in the knowledge that it also implicitly characterises the negative region by symmetry.

Theorem 6 *in the context of Theorem 4 where there is a selection of appropriate positive G values, that the maximum likelihood always occurs with the selection of one that is most positive.*

Proof In the statement of Theorem 4 the maximum likelihood is given by:

$$p_{max}(\alpha_{1\ldots n}) = \frac{N!}{\prod_{i=1}^{n}(n_i!)}\prod_{i=1}^{n}\left(\frac{n_i}{N}\frac{G}{G-a_i}\right)^{n_i}$$

It suffices to show that the derivative of this with respect to G is non-negative. We show this by a substitute function:

$$M(G) = \log\left(\frac{\prod_{i=1}^{n}(n_i!)}{N!}p_{max}(\alpha_{1\ldots n})\right) = \sum_{i=1}^{n} n_i\left(\log(\frac{n_i}{N}) + \log(G) - \log(G-a_i)\right)$$

$$\text{thus: } \frac{dM(G)}{dG} = \frac{N}{G} - \sum_{i=1}^{n}\frac{n_i}{G-a_i}$$

Since we are in the region of satisfying values of G per Theorem 4 then $\sum_{i=1}^{n}\frac{n_i}{N}\frac{a_i}{a_i-G} \geq 0$ Manipulating this expression we get: $\sum_{i=1}^{n}\frac{n_i}{G-a_i} \leq \frac{N}{G}$ and thus $\frac{dM(G)}{dG} \geq 0$ ■

Taking Theorem 4 and combining with the 5 cases that result from Theorem 5, and the consideration of Theorem 6, taking the limit of continuous functions, offsetting for non-zero mean μ gives the result of Theorem 1.

Appendix B. Summary Derivation of Hoeffding's inequality

We begin with a Lemma that categorises the root of all Chernoff bounds, which is then directly used in the derivation of Hoeffding's inequality:

Lemma 7 (Chernoff Bound) *If $\hat{\mu}$ is sample mean of n independent and identically distributed samples of random variable X then for any $s > 0$ and t:*

$$\mathbb{P}(\hat{\mu} \geq t) \leq \mathbb{E}\left[\exp(sX)\right]^n \exp(-snt)$$

Proof $\mathbb{P}(\hat{\mu} \geq t) = \mathbb{P}\left(\exp\left(s\sum_{i=1}^{n} x_i\right) \geq \exp(snt)\right)$
$\leq \mathbb{E}\left[\exp\left(s\sum_{i=1}^{n} x_i\right)\right]\exp(-snt) \leq \mathbb{E}\left[\exp\left(sX\right)\right]^n \exp(-snt)$
using Markov's inequality and i.i.d of the samples respectively. ■

Theorem 8 (Hoeffding's inequality for mean zero) *Let X be a real-valued random variable that is bounded $a \leq X \leq b$, with a mean μ of zero. Then for $D = b - a$ and any $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} \geq t) \leq \exp\left(\frac{-2nt^2}{D^2}\right) \tag{14}$$

Proof To prove Hoeffding's inequality we develop an upper bound for $\mathbb{E}[\exp(sX)]$, if we assume variable X has a probability density function $f(x)$, then we can linearize $\exp(sx)$ as:
$\mathbb{E}[\exp(sX)] = \int_a^b f(x)\exp(sx)dx \leq \int_a^b f(x)\left(\frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}\right)dx$
Using the fact that the mean $\mu = \int_a^b f(x)xdx = 0$ thus: $\mathbb{E}[\exp(sX)] \leq \frac{1}{sb-sa}\left(sb\exp(sa) - sa\exp(sb)\right)$
Given the fact that for any $\kappa > 0, \gamma < 0$ that $\kappa\exp(\gamma) - \gamma\exp(\kappa) \leq (\kappa-\gamma)\exp\left(\frac{1}{8}(\kappa-\gamma)^2\right)$ (15)
thus: $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{1}{8}s^2(b-a)^2\right)$
Applying our Chernoff bound lemma 7 we get: $\mathbb{P}(\hat{\mu} \geq t) \leq \exp\left(\frac{1}{8}s^2(b-a)^2 n - snt\right)$
And minimising with respect to s yields the required result. ■

At a first glance the most limiting feature of the derivation given is the requirement that the mean is zero, however this is ultimately immaterial and intentionally used to simplify the derivation. Because any data distribution can be shifted such that its expectation value becomes zero, leaving D unchanged. hence:

Theorem 9 (Hoeffding's inequality) *Let X be a real-valued random variable that is bounded $a \leq X \leq b$. Then for $D = b - a$ and any $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} - \mu \geq t) \leq \exp\left(\frac{-2nt^2}{D^2}\right) \qquad \textit{or rearranged} \qquad \mathbb{P}\left(\hat{\mu} - \mu \geq \sqrt{\frac{D^2\log(1/t)}{2n}}\right) \leq t \quad (16)$$

However we can do without simplifications to develop more powerfull inequality:

Theorem 10 *Let X be a real-valued random variable that is bounded $a \leq X \leq b$, with a mean μ of zero. Then for $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} - \mu \geq t) \leq \left(\frac{b}{b-a}\left(\frac{b(a-t)}{a(b-t)}\right)^{\frac{a-t}{b-a}} - \frac{a}{b-a}\left(\frac{b(a-t)}{a(b-t)}\right)^{\frac{b-t}{b-a}}\right)^n \quad (17)$$

Proof Similar to the proof in Theorem 8 we follow the same steps except do not apply Equation 15, leading to: $\mathbb{P}(\hat{\mu} \geq t) \leq (b\exp(sa) - a\exp(sb))^n \exp(-snt)(b-a)^{-n}$
And minimising with respect to s occurs at $s = \frac{1}{b-a}\log\left(\frac{b(a-t)}{a(b-t)}\right)$ which is the required result. ■

This concentration inequality is more powerful but more difficult to manipulate, it is also more commonly stated for variable X with non-zero mean and bounded $0 < X < 1$. See Chvátal (1979); Hoeffding (1963):

Theorem 11 (Also called Hoeffding's inequality) *Let X be a real-valued random variable that is bounded $0 \leq X \leq 1$, with mean μ. Then for $t > 0$, the mean $\hat{\mu}$ of n independent samples of X is probability bounded by:*

$$\mathbb{P}(\hat{\mu} - \mu \geq t) \leq \left[\left(\frac{1-\mu}{1-t-\mu}\right)^{1-t-\mu}\left(\frac{\mu}{t+\mu}\right)^{t+\mu}\right]^n \quad (18)$$

Proof Follows from substitution $a = -\mu$ and $b = 1 - \mu$. ■

doi:10.20944/preprints202004.0426.v1

Appendix C. A note on what concentration inequalities are not

The derivation of Hoeffdings inequality begins with the assumption of the mean, and then proceeds to develop a probabilistic bounds the sample mean; and this is a legitimate process of develop bounds on the sample mean conditioned on the assumption of knowing the population mean. However this process is fundamentally different from a process that begins with information on the sample mean and then infers probabilities on the population mean.

For instance, Hoeffding's inequality can be expressed as (see Theorem 9):

$$\mathbb{P}(\hat{\mu} - \mu \geq t | \mu = x, n, t, D) \leq \exp\left(\frac{-2nt^2}{D^2}\right)$$

Which identifies that the the probability that the sample mean underestimates the true population mean by more than or equal to t, given the assumption of the population mean μ has a value (say x), given the parameter t, the number of samples n and the width of the data support D. And this can be converted into a more unconditioned probability statement via integration of Bayes' theorem with a maximisation:

$$\begin{aligned}\mathbb{P}(\hat{\mu} - \mu \geq t | n, t, D) &= \int \mathbb{P}(\hat{\mu} - \mu \geq t | \mu = x, n, t, D)\, \mathbb{P}(\mu = x | n, t, D) dx \\ &\leq \max_x \left(\mathbb{P}(\hat{\mu} - \mu \geq t | \mu = x, n, t, D)\right) \\ &\leq \max_x \left(\exp\left(\frac{-2nt^2}{D^2}\right)\right) = \exp\left(\frac{-2nt^2}{D^2}\right) \qquad (19)\end{aligned}$$

The first line here identifies that the probability that the sample mean underestimates the true population mean by more than t is equal to the same probability conditioned on $\mu = x$ times the probability that $\mu = x$, then integrated over all possible values of x. The second line identifies that since $\mathbb{P}(\mu = x | n, t, D)$ are positive and sum to one, the upper bound is created by the $\mu = x$ which maximises $\mathbb{P}(\hat{\mu} - \mu \geq t | \mu = x, n, t, D)$. And the third line identifies that as the expression is bound above by the exponential expression which does not depend on x anyway.

By this process we have arrived at the unconditioned probability statement of Hoeffding's inequality which many people are familiar with - as a bound that the sample mean underestimates (or reversely overestimates) the mean by more than t, given t,n and D.

Although it is a bit mundane to point all this out, this process is not specifically trivial - for instance, Theorem 11 is an unambiguously more powerful form of Hoeffding's inequality, that seems far less well known and used, even though they are built on identical assumptions. To use Theorem 11 as an statement of probability that is unconditioned by the mean, one would have to conduct a similar maximisation procedure as in equations (19).

However what these concentration inequalities are not (and this is a subtle misconception) are statements of probability conditioned on the knowledge of sample statistics. And this is true even if the UCB method seems to operate by using concentration inequalities to create upper bounds on the mean of the reward distributions in light of past rewards.

It is easy to construct a counterexample: If Hoeffding's inequality works, then it must work across a range of distributions with different means. we might imagine that the output from a machine is either going to be Rademacher (values of -1 and 1 with equal probability) or Bernoulli

doi:10.20944/preprints202004.0426.v1

distributed (values of 0 and 1 with equal probability) on a hidden coin-toss, supposing we take three samples and receive sample mean of a -1/3, we hence should be absolutely certain that the distribution is Rademacher and that the mean is zero, even though a naive application of Hoeffding's inequality would suggest that $\mathbb{P}(\mu - \hat{\mu} \geq 1/3) \leq \exp\left(\frac{-1}{6}\right) \approx 0.85$

Concentration inequalities gain validity in UCB primarily by being unconditioned probability statements between population statistics and sample statistics, that can be used to give expected regret bounds prior to the occurrence of sample rewards. The intention of this paper is to take the concept of what Hoeffding's inequality is, in the context of UCB, and then make it a tad stronger by using more than the just the sample mean.