A multiplicative UCB strategy for Gamma rewards

Matthieu Geist

A multiplicative UCB strategy for Gamma rewards

Matthieu Geist MATTHIEU.GEIST@CENTRALESUPELEC.FR
IMS-MaLIS Research Group, CentraleSupélec
& UMI 2958 (GeorgiaTech-CNRS)

Abstract

We consider the stochastic multi-armed bandit problem where rewards are distributed according to Gamma probability measures (unknown up to a lower bound on the form factor). To handle this problem, we propose an UCB-like strategy where indexes are multiplicative (sampled mean times a scaling factor). An upper-bound for the associated regret is provided and the proposed strategy is illustrated on some simple experiments.
Keywords: Stochastic multi-armed bandits, Gamma-distributed rewards

1. Introduction

We consider the stochastic multi-armed bandit problem where rewards are distributed according to Gamma laws. A random variable X follows a Gamma law, $X \sim \Gamma(a, b)$ with $a > 0$ the form parameter and $b > 0$ the scale parameter, if its probability density function satisfies

$$f(x) = \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} \mathbb{I}_{\mathbb{R}_+}(x),$$

where Γ is the Gamma function and $\mathbb{I}$ the indicator function. This encompasses as special cases the exponential distribution (which is of interest in a bandit setting for cognitive radio (Jouini, 2012)), the Erlang distribution or the (scaled) Chi-squared distribution. Therefore, studying stochastic bandits in this setting might be of interest.

A classical approach for stochastic bandits is the so-called UCB (Upper Confidence bound) strategy (Auer et al., 2002), derived from the Hoeffding (1963) inequality. This idea is generalized with the (α,ψ)-UCB strategy, which assumes that the cumulant generating functions (associated to rewards) are bounded by a convex function ψ (Bubeck and Cesa-Bianchi, 2012). We follow a similar path in this paper. However, we take advantage of the specific structure of the cumulant generating function of a Gamma distribution to introduce a multiplicative form of the Hoeffding inequality, and we take advantage of the fact that the left tail is lighter than the right tail, from which the Γ-UCB(α,a_0) strategy is derived.

The proposed approach is actually a (slight) generalization of the MUCB(α) (Multiplicative UCB) strategy of Jouini and Moy (2012), that handles exponentially distributed rewards (this being a special case of Gamma rewards, with a form factor $a = 1$). For this (exponential) case, the KL-UCB (Kullbach-Leibler UCB) strategy (Garivier and Cappé, 2011) or Thompson sampling (Korda et al., 2013) could be applied too, as they can generally be adapted to the setting where the rewards belong to the one-parameter canonical exponential family. However, Gamma distributions with unknown shape parameter do not belong to this family, so extension to the case studied in this paper is not direct.

2. Γ-UCB(α,a_0)

We consider the basic stochastic multi-armed bandit problem. Each arm $i \in \{1, \dots, K\}$ is associated to a Gamma probability measure ν_i, $\nu_i \sim \Gamma(a_i, b_i)$. The coefficients a_i and b_i are unknown (otherwise, the means would be known), but we assume that a lower bound a_0 on the form factors is known: $a_0 \leq \min_{1 \leq i \leq K} a_i$. At each time step $t = 1, 2, \dots$, the player chooses an arm $I_t \in \{1, \dots, K\}$ (based on past choices and observations) and receives a reward $X_{I_t,t}$ drawn from ν_{I_t}, independently from the past. We write μ_i the expectation of arm i and define: $\mu_* = \max_{1 \leq i \leq k} \mu_i$ and $i_* \in \operatorname{argmax}_{1 \leq i \leq K} \mu_i$. The ideal (but unreachable) strategy would consist in choosing systematically $I_t = i_*$. Therefore, the quality of a strategy can be measured with the regret, defined as the cumulative difference (in expectation) between the optimal arm and chosen arms:

$$R_n = n\mu_* - \mathbb{E}[\sum_{t=1}^{n} \mu_{I_t}].$$

Write $T_i(s) = \sum_{t=1}^{s} \mathbb{I}_{I_t=i}$ the number of times the player selected arm i on the first s rounds and $\Delta_i = \mu_* - \mu_i$ the suboptimality of arm i. The regret can be written equivalently as $R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i(n)]$. Therefore, a good strategy should control $\mathbb{E}[T_i(n)]$, the (expected) number of time a suboptimal arm is played. Write $\hat{\mu}_{i,s}$ the sample mean of rewards obtained by pulling arm i for s times: $\hat{\mu}_{i,s} = \frac{1}{s} \sum_{t=1}^{s} X_{i,t}$. Let also $\alpha > 0$ be an input parameter. At time t, the proposed Γ-UCB(α,a_0) selects

$$I_t \in \operatorname*{argmax}_{1 \leq i \leq K} \frac{\hat{\mu}_{i,T_i(t-1)}}{\left(1 - \sqrt{\frac{2\alpha \ln t}{a_0 T_i(t-1)}}\right)_+},$$

where $(\cdot)_+ = \max(\cdot, 0)$ and with the convention $\frac{1}{0} = \infty$. With $a_0 = 1$ (the form factor of an exponential distribution), we retrieve the MUCB(α) strategy of Jouini and Moy (2012).

3. Regret

In this section, we study the regret encountered by the proposed strategy.

Theorem 1 (Regret bound for Γ-UCB(α,a_0)) *Define ρ_i as the ratio between expectations of arms i and i_*:*

$$\rho_i = \frac{\mu_i}{\mu_*}.$$

For $\rho \in (0, 1)$, define the function g as

$$g(\rho) = \frac{3\rho + 1 - (\rho + 1)\sqrt{1 + 4\rho - \rho^2}}{\rho^2}. \tag{1}$$

Assume that $a_0 \leq \min_{1 \leq i \leq K} a_i$ and that $\alpha > 2$. Then, the Γ-UCB(α,a_0) satisfies

$$R_n \leq \mu_* \sum_{i:\rho_i < 1} \left(\frac{\alpha(1 - \rho_i)}{a_0 g(\rho_i)} \ln n + \frac{\alpha(1 - \rho_i)}{\alpha - 2} \right).$$

To prove this, we will need a multiplicative Hoeffding bound for Gamma random variables.

Lemma 2 (A multiplicative bound for the Gamma case) *Let $X_1, \dots, X_n$ be i.i.d. (independent and identically distributed) random variables such that for all i, $X_i \sim \Gamma(a, b)$. Let $a_0 \leq a$. Write $\mu = \mathbb{E}[X_1]$ the common expectation and $\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ the sample mean. Then, for any $\delta \in (0, 1)$, we have*

$$P\left(\frac{\mu_n}{(1-\epsilon)_+} < \mu\right) \leq \delta \text{ with } \epsilon = \sqrt{\frac{2\ln\frac{1}{\delta}}{a_0 n}} \text{ and } P\left(\frac{\mu_n}{1+\epsilon} > \mu\right) \leq \delta \text{ with } \epsilon = \sqrt{\frac{2\ln\frac{1}{\delta}}{a_0 n}} + \frac{\ln\frac{1}{\delta}}{a_0 n}.$$

Proof We start by studying the cumulant generating function of a Gamma random variable[1], before applying a standard Chernoff argument. Let $X \sim \Gamma(a, b)$. The cumulant generative function of the centered random variable $X - \mu$ satisfies:

$$\forall \lambda < \frac{1}{b}, \quad L_{X-\mu}(\lambda) = \ln \mathbb{E}[e^{\lambda(X-\mu)}] = -a(\ln(1-\lambda b) + \lambda b) \leq \frac{\lambda^2 a b^2}{2(1-\lambda b)}. \tag{2}$$

In other words, $X - \mu$ is sub-gamma on the right tail with variance factor ab^2 and with scale factor b. As X is Gamma, its expectation satisfies $\mu = \mathbb{E}[X] = ab$. Therefore, we have

$$\forall \lambda < \frac{1}{b}, \quad L_{X-\mu}(\frac{\lambda}{\mu}) \leq \frac{1}{a}\frac{\lambda^2}{2(1-\frac{\lambda}{a})} \leq \frac{1}{a_0}\frac{\lambda^2}{2(1-\frac{\lambda}{a_0})} \doteq \psi(\lambda). \tag{3}$$

On the other hand, $X - \mu$ is sub-gaussian on the left tail with variance factor ab^2:

$$\forall \lambda > 0, L_{\mu-X}(\lambda) = \ln \mathbb{E}[e^{\lambda(\mu-X)}] = -a(\ln(1+\lambda b) - \lambda b) \leq \frac{\lambda^2 a b^2}{2}.$$

Given that $\mu = ab$, we have

$$\forall \lambda > 0, L_{\mu-X}(\frac{\lambda}{\mu}) \leq \frac{1}{a}\frac{\lambda^2}{2} \leq \frac{1}{a_0}\frac{\lambda^2}{2} \doteq \varphi(\lambda). \tag{4}$$

We can now provide the desired bounds using a standard Chernoff argument. For the second one, we have for any $0 < \lambda < \frac{1}{b}$:

$$\begin{aligned}
P(\frac{\mu_n - \mu}{\mu} > \epsilon) &= P(e^{\lambda \sum_{i=1}^n \frac{X_i - \mu}{\mu}} > e^{\lambda n \epsilon}) \\
&\leq e^{-\lambda n \epsilon} \mathbb{E}[e^{\lambda \sum_{i=1}^n \frac{X_i - \mu}{\mu}}] \text{ (by the Markov's inequality)} \\
&= e^{-\lambda n \epsilon} \left(\mathbb{E}[e^{\lambda \frac{X_1 - \mu}{\mu}}]\right)^n \text{ (by the i.i.d. assumption)} \\
&= e^{-n(\lambda\epsilon - L_{X_1-\mu}(\frac{\lambda}{\mu}))} \leq e^{-n(\lambda\epsilon - \psi(\lambda))} \text{ (by Eq. (3)).}
\end{aligned}$$

Write $\psi_*(\epsilon)$ the Legendre-Fenchel transform of ψ,

$$\psi_*(\epsilon) = \sup_{0<\lambda<\frac{1}{b}} (\lambda\epsilon - \psi(\lambda)) = a_0\left(1 + \epsilon - \sqrt{1+2\epsilon}\right).$$

1. This is largely inspired from Boucheron et al. (2013, Ch. 2).

We therefore have:

$$P(\frac{\mu_n - \mu}{\mu} > \epsilon) = P(\frac{\mu_n}{1+\epsilon} > \mu) \leq e^{-n\psi_*(\epsilon)}.$$

Writing $\delta = e^{-n\psi_*(\epsilon)}$ and using the fact that for $x > 0$,

$$\psi_*^{-1}(x) = \sqrt{\frac{2x}{a_0}} + \frac{x}{a_0}, \text{ we get } P\left(\frac{\mu_n}{1 + \psi_*^{-1}(\frac{1}{n}\ln\frac{1}{\delta})} > \mu\right) \leq \delta,$$

which is the stated second bound.

For the first bound, the proof is similar. We have with the same Chernoff argument:

$$P(\frac{\mu - \mu_n}{\mu} > \epsilon) \leq e^{-n\varphi_*(\epsilon)},$$

where $\varphi_*(\epsilon)$ is the Fenchel-Legendre transform of $\varphi(\lambda)$, the upper-bound of the cumulant generating function (on the left tail) defined in Eq. (4):

$$\varphi_*(\epsilon) = \sup_{\lambda > 0}(\lambda\epsilon - \varphi(\lambda)) = \frac{a_0}{2}\epsilon^2.$$

On the other hand, $P(\frac{\mu-\mu_n}{\mu} > \epsilon) = P(\mu_n < \mu(1-\epsilon))$. If $1-\epsilon < 0$, then $P(\mu_n < \mu(1-\epsilon)) = P(\mu_n < 0) = 0 \leq e^{-n\varphi_*(\epsilon)}$. Therefore:

$$P(\mu_n < \mu(1-\epsilon)_+) = P(\frac{\mu_n}{(1-\epsilon)_+} < \mu) \leq e^{-n\varphi_*(\epsilon)}.$$

Writing $\delta = e^{-n\varphi_*(\epsilon)}$ and using the fact that $\varphi_*^{-1}(x) = \sqrt{\frac{2x}{a_0}}$, we get the first stated bound:

$$P\left(\frac{\mu_n}{\left(1 - \varphi_*^{-1}(\frac{1}{n}\ln\frac{1}{\delta})\right)_+} > \mu\right) \leq \delta.$$

■

The first bound of Lemma 2 states that with probability at least $1-\delta$, we have

$$\frac{\mu_n}{\left(1 - \sqrt{\frac{2\ln\frac{1}{\delta}}{a_0 n}}\right)_+} > \mu,$$

which is the proposed strategy with $\delta = t^{-\alpha}$ (optimism in the face of uncertainty based on the upper confidence bound given by Lemma 2). We can now prove the main theorem.

Proof [Theorem 1] To bound the regret R_n, we will bound the quantity $\mathbb{E}[T_i(n)]$. Assume without loss of generality that $I_t = i \neq i_*$. At least one of the tree following inequalities must be true :

$$\frac{\hat{\mu}_{i_*,T_{i_*}(t-1)}}{\left(1 - \sqrt{\frac{2\alpha\ln t}{a_0 T_{i_*}(t-1)}}\right)_+} \leq \mu_* \tag{5}$$

$$\frac{\mu_{i,T_i(t-1)}}{1 + \sqrt{\frac{2\alpha\ln t}{a_0 T_i(t-1)}} + \frac{\alpha\ln t}{a_0 T_i(t-1)}} > \mu_i \tag{6}$$

$$T_i(t-1) < \frac{\alpha\ln t}{a_0 g(\rho_i)}, \tag{7}$$

with g defined in Eq. (1). Actually, if all equations were false, we would have

$$\begin{aligned}
\frac{\hat{\mu}_{i_*,T_{i_*}(t-1)}}{\left(1-\sqrt{\frac{2\alpha\ln t}{a_0 T_{i_*}(t-1)}}\right)_+} &> \mu_* && \text{(by (5) false)}\\
&= \frac{\mu_i}{\rho_i} && \text{(by definition)}\\
&> \frac{\mu_i\left(1+\sqrt{\frac{2\alpha\ln t}{a_0 T_i(t-1)}}+\frac{\alpha\ln t}{a_0 T_i(t-1)}\right)}{\left(1-\sqrt{\frac{2\alpha\ln t}{a_0 T_i(t-1)}}\right)_+} && \text{(by (7) false)}\\
&\geq \frac{\mu_{i,T_i(t-1)}}{\left(1-\sqrt{\frac{2\alpha\ln t}{a_0 T_i(t-1)}}\right)_+} && \text{(by (6) false).}
\end{aligned}$$

This would mean that $I_t = i_* \neq i$, which is a contradiction. Now, define u as

$$u = \left\lceil \frac{\alpha \ln n}{a_0 g(\rho_i)} \right\rceil.$$

We have that

$$\begin{aligned}
\mathbb{E}[T_i(n)] = \mathbb{E}[\sum_{t=1}^{n} \mathbb{I}_{I_t=i}] &\leq u + \sum_{t=u+1}^{n} \mathbb{I}_{I_t=i \text{ and (7) is false}}\\
&\leq u + \sum_{t=u+1}^{n} \mathbb{I}_{\text{(5) or (6) is true}}\\
&\leq u + \sum_{t=u+1}^{n} \left(P(\text{(5) is true}) + P(\text{(6) is true})\right).
\end{aligned}$$

Thus, we have to bound the probabilities of events (5) and (6), which can be done thanks to Lemma 2. Indeed:

$$\begin{aligned}
P(\text{(5) is true}) &= P\left(\exists s \leq t : \frac{\hat{\mu}_{i_*,s}}{\left(1-\sqrt{\frac{2\alpha\ln t}{a_0 s}}\right)_+} < \mu_*\right)\\
&\leq \sum_{s=1}^{t} P\left(\frac{\hat{\mu}_{i_*,s}}{\left(1-\sqrt{\frac{2\alpha\ln t}{a_0 s}}\right)_+} < \mu_*\right)\\
&\leq \sum_{s=1}^{t} \frac{1}{t^\alpha} = \frac{1}{t^{\alpha-1}}.
\end{aligned}$$

Similarly, one can show that:

$$P(\text{(6) is true}) \leq \frac{1}{t^{\alpha-1}}.$$

Therefore, we can bound the number of time arm i has been played:

$$\mathbb{E}[T_i(n)] \leq u + 2\sum_{t=u+1}^{n} \frac{1}{t^{\alpha-1}} \leq \frac{\alpha\ln n}{a_0 g(\rho_i)} + 1 + \int_1^\infty \frac{2dt}{t^{\alpha-1}} = \frac{\alpha\ln n}{a_0 g(\rho_i)} + \frac{\alpha}{\alpha-2}.$$

	KL-UCB	MUCB($\alpha > 4$)	($\alpha > 1$,ψ)-UCB	Γ-UCB($\alpha > 1$,1)
$\lim_{n\to\infty} \frac{R_n}{\mu_* \ln n}$	$\sum_{i:\rho_i<1} \frac{1-\rho_i}{\rho_i - 1 - \ln \rho_i}$	$\sum_{i:\rho_i<1} \frac{4\alpha}{1-\rho_i}$	$\sum_{i:\rho_i<1} \frac{\alpha(1-\rho_i)}{h(\rho_i)}$	$\sum_{i:\rho_i<1} \frac{\alpha(1-\rho_i)}{g(\rho_i)}$

Table 1: Comparison of bounds (exponential rewards).

Using the fact that $R_n = \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)]$ and that $\Delta_i = \mu_*(1-\rho_i)$, we obtain the desired bound:

$$R_n \leq \mu_* \sum_{i:\rho_i<1} \left(\frac{\alpha(1-\rho_i)}{a_0 g(\rho_i)} \ln n + \frac{\alpha(1-\rho_i)}{\alpha - 2} \right).$$

■

This proof is very similar to the one of (α,ψ)-UCB (Bubeck and Cesa-Bianchi, 2012, Ch. 2). The difference is that we consider different convex bounding functions on the left and right tail (regarding the cumulant generating function) and we use a multiplicative concentration inequality instead of an additive one.

Remark 3 (About the constant α) *We have kept the proof simple. However, using a peeling argument instead of a simple union bound (as done by Bubeck (2010) for UCB) for bounding the probabilities of events* (5) *and* (6) *in the preceding proof should give the same result for any $\alpha > 1$ (instead of $\alpha > 2$, and with a different constant term).*

4. Discussion

As far as we know, the sole alternative to Γ-UCB in the studied case is (α,ψ)-UCB (Bubeck and Cesa-Bianchi, 2012, Ch. 2), that we instantiate now for Gamma rewards. Recall the bound (2) on the cumulant generating function of a centered Gamma reward. We assume that upper-bounds on the two parameters are known: $a_\infty \geq \max_{1\leq i\leq K} a_i$ and $b_\infty \geq \max_{1\leq i\leq k} b_i$. A convex function bounding cumulant generating functions of all rewards (on both the left and right tails) is

$$\psi(\lambda) = \frac{\lambda^2 a_\infty b_\infty^2}{2(1-\lambda b_\infty)}.$$

The associated Legendre-Fenchel transform and its inverse are

$$\psi_*(\epsilon) = a_\infty \left(1 + \frac{\epsilon}{a_\infty b_\infty} - \sqrt{1 + 2\frac{\epsilon}{a_\infty b_\infty}} \right) \text{ and } \psi_*^{-1}(x) = b_\infty(x + \sqrt{2a_\infty x}).$$

The (α,ψ)-UCB is therefore

$$I_t \in \operatorname*{argmax}_{1\leq i\leq K} \left(\hat{\mu}_{i,T_i(t-1)} + b_\infty \left(\frac{\alpha \ln t}{T_i(t-1)} + \sqrt{\frac{2\alpha a_\infty \ln t}{T_i(t-1)}} \right) \right).$$

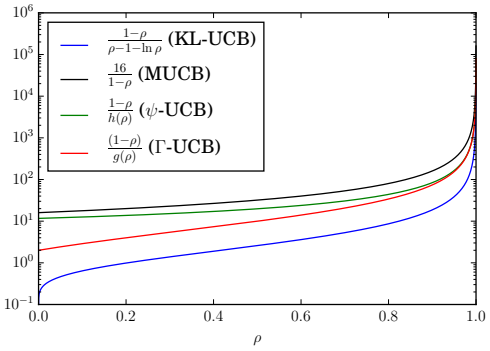

Figure 1: Main term of each bound.

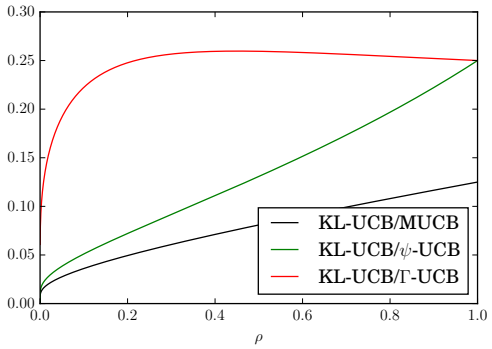

Figure 2: Ratio between the terms of Fig. 1.

For any $\alpha > 2$ ($\alpha > 1$ being possible, see remark 3), its regret satisfies

$$R_n \leq \sum_{i:\Delta_i>0} \left(\frac{\alpha\Delta_i}{a_\infty \left(1 + \frac{\Delta_i}{2a_\infty b_\infty} - \sqrt{1 + \frac{\Delta_i}{a_\infty b_\infty}}\right)} \ln n + \frac{\alpha\Delta_i}{\alpha - 2} \right).$$

More strategies deal with the exponential case (that is, Gamma case with $a = 1$) such as KL-UCB (Garivier and Cappé, 2011) with a proper divergence or MUCB (Jouini and Moy, 2012). Therefore, we also compare our bound to their. More precisely, we compare their asymptotic rates in Table 1. For (α,ψ)-UCB, we take $a_\infty = 1$ and $b_\infty = \max_{1\leq i\leq K} b_k = \mu_*$ (which is an optimistic setting, the largest mean is exactly known beforehand). In this case, the bound simplifies as

$$R_n \leq \mu_* \sum_{i:\rho_i<1} \left(\frac{\alpha\rho_i}{h(\rho_i)} \ln n + O(1) \right), \text{ with } h(\rho) = \frac{3-\rho}{2} - \sqrt{2-\rho}.$$

As comparing these analytical expressions is not straightforward, we plot on Fig. 1 (in log-scale) the main term of each bound, as a function of ρ, the ratio between expectations. One can see that KL-UCB provides the better bound. This is not a big surprise, as it matches the lower bound of Lai and Robbins (1985) (Garivier and Cappé, 2011). Γ-UCB has a slightly better bound than MUCB (which is a consequence of the proof, as both strategies are identical for exponential rewards) and (α,ψ)-UCB. We also compare the ratio of these terms on Fig. 2, to see how much better is KL-UCB. This shows that for large enough values of ρ, the regret of Γ-UCB is no more than four times the one of KL-UCB. Notice also that we considered the exponential case to ease the comparison, but Γ-UCB and (α,ψ)-UCB apply more generally to Gamma rewards.

5. Experiments

In this section, we consider stochastic bandits with exponential rewards. We compare the Γ-UCB(α,1) strategy (therefore, the MUCB strategy) to KL-UCB (with the divergence

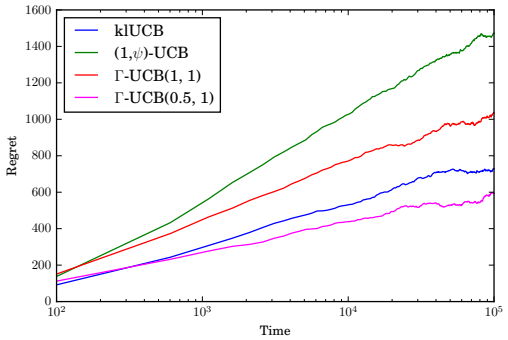

Figure 3: Exponential rewards with means $\{1, 2, 3, 4, 5\}$.

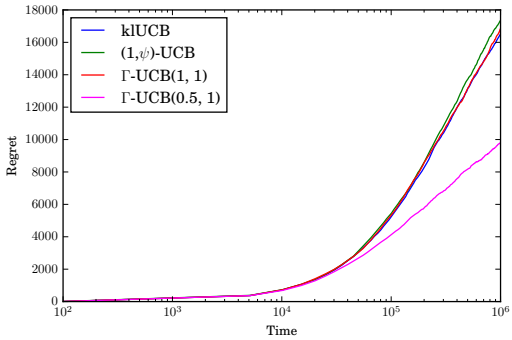

Figure 4: Exponential rewards with means $\{5, 5, 5.1, 5, 5\}$.

corresponding to exponential rewards) and (1,ψ)-UCB. For Γ-UCB, we consider two values of α: $\alpha = 1$ (limit of the theory) and $\alpha = 0.5$ (beyond). For (1,ψ)-UCB, we take $a_\infty = 1$, $b_\infty = \mu_*$ (the most optimistic setting). We consider two experiments. In the first one, arms have expectations $\{1, 2, 3, 4, 5\}$ (results are provided in Fig. 3). In the second one, arms have expectations $\{5, 5, 5.1, 5, 5\}$ (results are provided in Fig. 4). This case is harder, as the ratio of expectations ρ is closer to one. Each provided result is averaged over 10^3 independent runs.

KL-UCB has a smaller regret than Γ-UCB(1,1), which is better than (1,ψ)-UCB, as predicted in Sec. 3. However, on these experiments, Γ-UCB(0.5, 1) has a smaller regret. If general conclusions cannot be drawn (as this is beyond the theory), this might be a viable empirical alternative.

6. Conclusion

In this paper, we have proposed a multiplicative UCB strategy for Gamma rewards, extending the original MUCB idea of Jouini and Moy (2012). We have derived a regret bound, rather close to the one of KL-UCB in the exponential case. We also experimented the strategy on two simple bandit problems with exponential rewards. With an optimistic choice of the α parameter, Γ-UCB provides a lower regret than KL-UCB on these examples.

Compared to (α,ψ)-UCB (the sole alternative in the Gamma case, as far as we know), the proposed strategy requires a lower bound on the form factor instead of upper bounds on both the form and scale factors, and it provides a better regret, both theoretically and empirically (on the studied exponential cases).

An interesting perspective would be to study if such a multiplicative strategy could be of interest for other kind of distributions (which would depend on the structure of the related cumulant generating function).

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille 1, 2010.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and trends in machine learning*, 5(1):1–122, 2012.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Annual Conference on Learning Theory (COLT)*, 2011.

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

Wassim Jouini. *Contribution to learning and decision making under uncertainty for Cognitive Radio*. PhD thesis, Supélec, 2012.

Wassim Jouini and Christophe Moy. Channel selection with Rayleigh fading: A multi-armed bandit framework. In *Signal Processing Advances in Wireless Communications (SPAWC)*, pages 299–303. IEEE, 2012.

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.