Robust Risk-averse Stochastic Multi-Armed Bandits

Odalric-Ambrym Maillard

Robust Risk-averse Stochastic Multi-Armed Bandits

Odalric-Ambrym Maillard

Technion, Faculty of Electrical Engineering
Haifa, Israel
odalricmaillard@ee.technion.ac.il

Abstract. *We study a variant of the standard stochastic multi-armed bandit problem when one is not interested in the arm with the best mean, but instead in the arm maximising some coherent risk measure criterion. Further, we are studying the deviations of the regret instead of the less informative expected regret. We provide an algorithm, called* `RA-UCB` *to solve this problem, together with a high probability bound on its regret.*

Keywords: Multi-armed bandits, coherent risk measure, cumulant generative function, concentration of measure.

1 Introduction

The setting of the stochastic multi-armed bandit problem is an old and well-known problem (see [30], [31] and [26]), with a simple formalization that is nevertheless extremely powerful, leading to a large range of beautiful theoretical developments as well as important practical questions (medical treatment strategies, web advertisement, economy, etc.). The standard stochastic setting considers an agent facing a finite number of distributions (arms) that she can sample one at a time. Each sample is considered as a reward and the goal is to maximize after T trials the cumulative sum of rewards in some sense. Generally one measures performance with the *expected regret* criterion, that compares the expected cumulative reward of the learner to that of the strategy that constantly pulls the arm with highest *mean*. However in a number of applications, this criterion is not sufficient. For instance a medical treatment that is very effective on average may still have a high variability and may potentially endanger the patients. Thus, we are interested in this paper in a so-called *risk-averse* rather than *expected* measure of performance.

Risk-aversion is an old notion, however with no consensus about its definition (see [24, 23, 1]). However, any risk measure that is *coherent* (see [27]) is considered to be a good measure. We here use a standard risk measure that is coherent, defined in equation (3). In the relevant field of reinforcement learning, risk-aversion is also quite old (see [18]), and a number of works try to solve risk-averse problems [25, 2, 9, 21], although generally on the algorithmic and not on the theoretical side. Risk-aversion has been more closely looked in the online learning setting (where one sees all the rewards after pulling an arm and

not only that of the chosen arm), with tight positive and negative results (see e.g. [12, 32]) on what can be done. Now in the bandit literature, the *expected regret* criterion has been extensively looked over the past, with recent, extremely tight, non-asymptotic results for various algorithms [16, 17, 13, 22, 19]. However, much less attention has been put on the risk-averse problem. One can cite [11] about optimality of index policies, for exponential utility and with the Gittins index perspective, which is however quite different than our goal. More recently, the work of [28] goes in the direction we target as it analyses the deviations of the regret for algorithms that compete with the best expected arm, however no risk-aversion measure is considered. Of special interest is [29] that explicitly considers the risk-aversion problem in multi-armed bandits, targeting finite time performance guarantees. They use the very standard risk-measure called the *mean-variance* (see [23]), and show that it is possible to get sub-linear regret for such a setting, for a specific definition of regret that they introduce. However, the regret analysis is not completely satisfactory, since their notion of regret takes into account not only the variability of each arms but also the variability of the learning algorithm itself (that is, an algorithm is somehow penalized for switching between arms). Due to that, the considered regret is difficult to interpret and debatable (it is not clear whether penalizing an algorithm for switching is a desired feature).

The present work is inspired by the works of [27, 28, 22] and [29]. We consider a notion of regret different from [29], that we believe to be more natural and easier to interpret, where only the variability of the distribution of arms defines the regret on the one hand, while a control on the tail of the regret is provided on the other hand, similarly to [28]. We consider a coherent risk measure that generalizes the mean-variance criterion in the sense that the two measures coincide in the special case of Gaussian distributions, while the former takes into account the entire tail distribution of the random variables, not only the first two moments. We introduce the `RA-UCB` algorithm, inspired from [22] and provide a regret analysis in Proposition 1, Theorem 1 that we believe to be tight up to constant factor (note that the focus of this work is not on optimising the leading constants, such as in [17, 13, 22], which would require a much more technical and uninformative analysis).

The paper is organized as follows. In Section 2, we introduce the regret and the coherent risk measure that is considered here, together with intuition about its meaning. In Section 3, we provide a generic robust (high-probability) non-asymptotic upper bound on the regret of any algorithm (Proposition 1) that depends only on the risk measure and on a control of the number of pulls of suboptimal arms by the algorithm. Section 4 introduces the `RA-UCB` algorithm, that is inspired by the $\mathcal{K}_{\inf}$ strategy of [6, 22], together with a dual formulation that enables effective implementation. Section 5 concludes the paper with a high-probability bound on the regret of the `RA-UCB` (Theorem 1, Corollary 1). The analysis makes use of adaptation of concentration tools that are detailed in the Appendix.

2 Setup and Notations

We consider a standard multi-armed bandit game with A many unknown distributions $\{\nu_a\}_{a=1,\dots,A}$ where for each a, $\nu_a \in \mathfrak{M}_1^+(\mathbb{R})$. At each time step, a learning algorithm $\mathfrak{A}$ must choose an arm $A_t^{\mathfrak{A}} \in \{1,\dots,A\}$, and then receives one new sample (the *reward*) from the corresponding distribution $\nu_{A_t^{\mathfrak{A}}}$. We write $Y_t \sim \nu_{A_t^{\mathfrak{A}}}$ for the random reward received when the strategy $\mathfrak{A}$ pulls the arm $A_t^{\mathfrak{A}}$ at time t, $X_{i,a}$ to refer to the i^{th} random variable sampled from arm a from the beginning of the game, and we finally introduce the quantity $N_{T,a}^{\mathfrak{A}} \overset{\text{def}}{=} \sum_{t=1}^{A} \mathbb{I}\{A_t^{\mathfrak{A}} = a\}$. Using these notations, the cumulated reward received by algorithm $\mathfrak{A}$ up to time T is given by

$$\sum_{t=1}^{T} Y_t = \sum_{a=1}^{A} \sum_{i=1}^{N_{T,a}^{\mathfrak{A}}} X_{i,a}\,.$$

2.1 Measure of risk-aversion

As mentioned in the introduction, there exists many possible ways to define risk-aversion. From a practical point of view, being risk-averse generally implies avoiding situations when we receive too bad reward (think of a medical treatment strategy, where the actions are the possible treatments, and the reward correspond to the health state of a patient). That is, we want to have a control on the tails, and more specifically on the lower-tail (the mass below the mean).

More formally, let us recall that for arbitrary random variable X admitting a finite cumulant generative function around 0, then the two following properties hold (this is by a simple application of Markov's inequality)

$$\mathbb{P}\Big[X \geq \inf\Big\{\frac{1}{\lambda}\log\mathbb{E}\exp(\lambda X) + \frac{\log(1/\delta)}{\lambda} \;:\; \lambda > 0\Big\}\Big] \leq \delta\;, \tag{1}$$

$$\mathbb{P}\Big[X \leq \sup\Big\{-\frac{1}{\lambda}\log\mathbb{E}\exp(-\lambda X) - \frac{\log(1/\delta)}{\lambda} \;:\; \lambda > 0\Big\}\Big] \leq \delta\,. \tag{2}$$

Note that (1) measures the probability that X is big, while (2) measures the probability that X is small, which is what we want to be protected against. Now, for the sake of clarity, it makes sense to introduce the value of the cumulant generative function of the variable X at point λ, rescaled by λ, that we denote

$$\kappa_{\lambda,\nu} = \frac{1}{\lambda}\log\mathbb{E}_\nu\exp\Big(\lambda X\Big), \tag{3}$$

and similarly we denote $\kappa_{-\lambda,\nu}$ the value of $\kappa_{\lambda',\nu}$ for $\lambda' = -\lambda$. This quantity is at the heart of many key-results and tools of concentration of measure (e.g. the Cramer-Chernoff method, the Chernoff transform, the log-Laplace transform). More importantly here, $\kappa_{-\lambda,\nu}$ is a key quantity to control the probability that X is small. We now provide more intuition for people unfamiliar with that quantity.

Example: To understand (1) and (2), let us consider t Gaussian random variables $\{Z_k\}_{k=1,\dots,t}$ i.i.d. from a distribution ν with mean μ and variance σ^2, then $X = \sum_{k=1}^{t} Z_k$ is Gaussian with mean μt and variance $\sigma^2 t$, and simple

computations show that $\kappa_{\lambda,\nu} = \mu t + \frac{\lambda\sigma^2 t}{2}$, which yields, after optimizing the previous bounds in λ, to the optimal value $\lambda = \sqrt{\frac{2\log(1/\delta)}{\sigma^2 t}}$ and the familiar concentration bounds for Gaussian random variables

$$\mathbb{P}\Big(\frac{1}{t}\sum_{k=1}^{t} Z_k - \mu \geq \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\Big) \leq \delta \quad \text{and} \quad \mathbb{P}\Big(\mu - \frac{1}{t}\sum_{k=1}^{t} Z_k \geq \sigma\sqrt{\frac{2\log(1/\delta)}{t}}\Big) \leq \delta\,.$$

Let us comment on this example. First, the quantity $\kappa_{-\lambda,\nu} = \mu t - \frac{\lambda\sigma^2 t}{2}$ takes the form of an operator that measures the mean of a random variable, penalized by some higher moment (the variance in that case). This is actually a general property, since by the variational formula for the Kullback-Leibler divergence, we have for a random variable X distributed according to $\nu \in \mathfrak{M}_1^+(\mathbb{R})$ that

$$\kappa_{-\lambda,\nu} \stackrel{\text{def}}{=} \inf\left\{\mathbb{E}_{\nu'}(X) + \frac{1}{\lambda}\,KL(\nu'||\nu) \,:\, \nu' \in \mathfrak{M}_1^+(\mathbb{R})\right\} \leq \mathbb{E}_\nu\big[X\big]\,. \tag{4}$$

where $KL(\nu'||\nu)$ denotes the Kullback-Leibler divergence between two distributions ν and ν'. Using $\kappa_{-\lambda,\nu}$ as a measure of risk-aversion is natural for several reasons: Additionally to the formulation (4) and the control (2) that are important for interpretability it is also a standard *coherent* risk-measure (see [27]). Also, due to its deep link for concentration of measure, it is especially natural for analysis.

Mixability gaps Finally, for completeness, we also introduce the two fundamental quantities $m^+_{\lambda,\nu}\big[X\big]$ and $m^-_{\lambda,\nu}\big[X\big]$ that we call here the upper (and respectively lower) mixability gap and that are defined by

$$m^+_{\lambda,\nu} = \kappa_{\lambda,\nu} - \mathbb{E}_\nu\Big[X\Big] \text{ and } m^-_{\lambda,\nu} = \mathbb{E}_\nu\Big[X\Big] - \kappa_{-\lambda,\nu}\,.$$

Note that the mixability gaps are always non negative by Jensen's inequality, and that an upper bound on them immediately provides a high probability confidence interval. Indeed, with these notations, the previous equations (1) and (2), can thus be rewritten more compactly as

$$\mathbb{P}\Big[X - \mathbb{E}_\nu\Big[X\Big] \geq \inf_{\lambda>0}\Big\{m^+_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leq \delta\,, \tag{5}$$

$$\mathbb{P}\Big[\mathbb{E}_\nu\Big[X\Big] - X \geq \inf_{\lambda>0}\Big\{m^-_{\lambda,\nu} + \frac{\log(1/\delta)}{\lambda}\Big\}\Big] \leq \delta\,. \tag{6}$$

2.2 Regrets for risk-averse multi-armed bandits

Optimal arm We now naturally define the optimal arm $a^\star$ as the one maximizing the risk aversion at some fixed level λ, that is we define

$$a^\star \in \operatorname*{argmax}_{a=1,\dots,A} \kappa_{-\lambda,\nu_{a^\star}}\,.$$

Note again that in the case of Gaussian distributions with mean μ_a and variance σ_a^2, we simply have $\kappa_{-\lambda,\nu_{a^\star}} = \mu_a - \frac{\lambda\sigma_a^2}{2}$, and that in general we always have $\kappa_{-\lambda,\nu_{a^\star}} \leq \mathbb{E}_{\nu_a}\big[X\big]$.

Regret Now we define the *empirical regret* $\mathfrak{R}_T(\lambda)$ of the strategy $\mathfrak{A}$ with respect to the strategy $\star$ that constantly pulls the same arm $a^\star \in \{1,\dots,A\}$ by the difference between the cumulated reward received by algorithm $\mathfrak{A}$ and

the cumulated reward that the strategy $\star$ would have received during the same game, that is, by introducing the fictitious plays $\{X_{i,a^\star}\}_{N^{\mathfrak{A}}_{T,a^\star}<i\leq T}$,

$$\mathfrak{R}_T(\lambda) \stackrel{\text{def}}{=} \sum_{i=1}^{T} X_{i,a^\star} - \sum_{a=1}^{A}\sum_{i=1}^{N^{\mathfrak{A}}_{T,a}} X_{i,a} = \sum_{i=N^{\mathfrak{A}}_{T,a^\star}+1}^{T} X_{i,a^\star} - \sum_{a\neq a^\star}\sum_{i=1}^{N^{\mathfrak{A}}_{T,a}} X_{i,a}\,. \quad (7)$$

Note that we are not interested here in controlling the *expected regret* $\overline{\mathfrak{R}}_T$ as it gives no information on the risk of the strategy $\mathfrak{A}$ and of pulling one arm. Indeed, we have the following standard decomposition

$$\overline{\mathfrak{R}}_T = T\mathbb{E}_{\nu_{a^\star}}\big[X\big] - \mathbb{E}\bigg[\sum_{s=1}^{T} Y_s\bigg] = \sum_{a\in\mathcal{A}}\Big(\mathbb{E}_{\nu_{a^\star}}[X] - \mathbb{E}_{\nu_a}[X]\Big)\mathbb{E}\Big[N_{T,a}\Big]\,, \quad (8)$$

while one would prefer to have a more informative measure, taking into account for instance the variance of the arms or some control of the tails. For this purpose, an other natural notion of regret is the *risk-averse regret* $\overline{\mathfrak{R}}_T(\lambda)$ defined by

$$\overline{\mathfrak{R}}_T(\lambda) = \sum_{a\in\mathcal{A}}\Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}\Big)\mathbb{E}\Big[N_{T,a}\Big]\,. \quad (9)$$

In the sequel, we control both (7) and (9) as they both offer interesting interpretation.

2.3 The price for risk-aversion

At a high-level, there is obviously a trade-off between trying to get maximal rewards and being risk-averse. Being too cautious (such as, arguably, `Exp3` see [4]) avoids getting linear regret, but prevents from getting high rewards as well. On the other hand, simply targeting the maximal mean (such as `UCB` see [3]) enables to get close to optimal rewards on average, but possibly very bad rewards in difficult environments (when sub-optimal arms have fat lower tails).

A similar situation appears in the standard expected regret setting for the class of `UCB-`ρ algorithms as shown by [28]: for, $\rho > \rho'$, `UCB-`ρ can compete with a larger class of environments than `UCB-`ρ'. However `UCB-`ρ' will beat `UCB-`ρ on simpler environments.

Simple and complex environments The risk-averse regret (9) captures the sub-optimality of an algorithm in terms of risk-aversion at some fixed level λ. As such, it is the direct equivalent of the expected regret in multi-armed bandits, and we control this regret for our `RA-UCB` procedure in Theorem 1. If such control may seem satisfactory for many reasons, it has also some drawback. Namely, the level of risk-aversion is here not related in any way to the actual distribution of rewards, since it is some parameter chosen a priori by the practitioner who wants to be protected against sampling possibly very low rewards. As a result, in easy situations when the rewards distributions have very light tails, a high risk-averse algorithm will be too cautious, and will get lower cumulative rewards than a less risk-averse algorithm, such as `UCB`. Similarly, if the actual distributions have very fat lower tails, a low risk-averse algorithm may not be cautious enough and thus get bad rewards compared to a more risk-averse algorithm, such as `Exp3`. See also figure 2.3.

Since such situations, that are of immediate practical interest, are not captured by the risk-averse regret (9) defined for some level λ, this motivates the

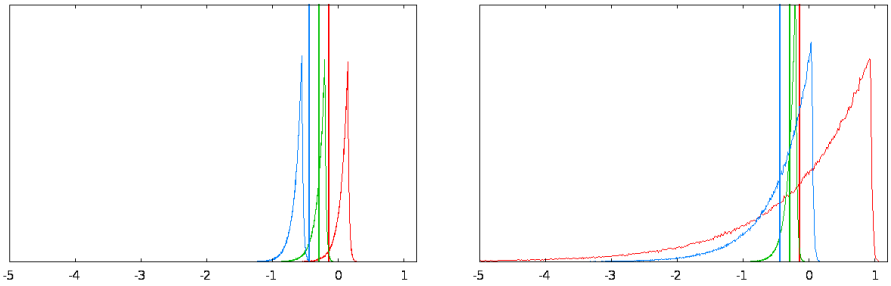

Fig. 1. Plot of arms' densities and their mean: left) an environment where no arm has fat lower tail. right) an environment where for some λ, the best arm (green) does not have best mean, and sub-optimal arms (red, blue) have fat lower tails.

study of the empirical risk-aversion regret (7) as this one is able to capture such behaviors (this is because it makes appear the empirical rewards coming from the actual distribution explicitly).

Note that this also raises the question of automatically adapting the level of risk-aversion to some bandit problem, or equivalently getting the best of all RA-UCB-λ algorithms (in terms of cumulated reward), which is very hard, (or even impossible, see [28] for impossibility results regarding UCB-ρ in the related problem of adaptivity in bandit problems). Since this involves orthogonal ideas that would worsen readability and interpretation, add a difficult layer of complexity, and is little justified in practice (where the level of risk-aversion is often simply fixed), we do not study this question in the present work.

Contribution The difficult situation for risk-aversion appears when the sub-optimal arms produce rewards much lower than their mean (heavy lower tail) while the best arm produces rewards much higher than its mean (heavy upper tail): this creates maximal regret. We introduce in section 4 the RA-UCB algorithm that guarantees a low regret in such difficult environments (contrary to e.g. UCB).

For clarity purpose, we clearly separate in two dedicated sections the analysis that is tighten and intrinsic to the risk-aversion problem (Proposition 1 that holds for any algorithm) from the more standard techniques used in stochastic bandits (Theorem 1). We derive on the way some non-trivial concentration results needed for the proof (Lemma 1, Lemma 2, equation (14)). The regret of RA-UCB essentially scales as $O(\log(T))$ with the time horizon T up to a distribution-dependent complexity factor.

3 A Generic Decomposition of the Empirical Regret

We now introduce a generic decomposition of the regret, valid for any strategy $\mathfrak{A}$, that is the direct equivalent of (9) for the empirical regret.

Proposition 1. *Let us define, for some non negative constants* $\{u_a\}_{a=1,\dots,A}$ *the event that sub-optimal arms are pulled too much*

$$\Omega \stackrel{\text{def}}{=} \left\{ \exists a \neq a^\star \, : \, N_{T,a}^{\mathfrak{A}} > u_a \right\},$$

and let us fix some value of λ *such that* $\kappa_{-\lambda,\nu_a}$ *exists for all* $a = 1, \dots, A$. *Then, for all* $\delta \in (0,1)$, *with probability higher than* $1 - \delta - \mathbb{P}(\Omega)$, *the regret of the strategy* $\mathfrak{A}$ *is upper bounded by*

$$\mathfrak{R}_T(\lambda) \leq \sum_{a \neq a^\star} u_a \left(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \right) + \left(m^-_{-\lambda,\nu_{a^\star}} \sum_{a \neq a^\star} u_a + \frac{(A-1)\log(2A/\delta)}{\lambda} \right) + \inf_{\lambda' > 0} \left\{ m^+_{\lambda',\nu_{a^\star}} \sum_{a \neq a^\star} u_a + \frac{\log(2A/\delta)}{\lambda'} \right\}. \tag{10}$$

The first term of (10) makes appear a quantity very similar to that of the optimal regret bounds for the expected regret in the stochastic setting, where the standard optimality gaps $\mathbb{E}_{\nu_{a^\star}}\big[X\big] - \mathbb{E}_{\nu_a}\big[X\big]$ are replaced by $\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}$, as expected. Now the second and third terms involve the mixability gaps of the optimal arm. The third term is intuitive: indeed, a regret minimizing algorithm will try to understand $\kappa_{-\lambda,\nu_a}$ for each arm, and prevent from large deviations below the mean (bad rewards). However, this does not prevent the optimal arm to have large deviations above the mean (that is, unexpected good rewards), which is precisely captured by the third term. Now the presence of the second term comes from another phenomenon: λ is a parameter of the algorithm that tries to pull the arm with highest risk-aversion at level λ. As such, this goal may be successful or not depending on intrinsic properties of the environment. We say that λ is well-adapted to the environment if it is such that the second term in (10) is negligible before the first term.

So as to provide some intuition, let us now specialize Proposition 1 to the case of Example 1 for illustration purpose. In this case, the mixability gaps of the optimal arm $a^\star$ equal $\frac{\lambda}{2}\sigma^2_{a^\star}$ and $\frac{\lambda'}{2}\sigma^2_{a^\star}$, so that if we introduce for convenience the quantity $u \stackrel{\text{def}}{=} \sum_{a \neq a^\star} u_a$, one can rewrite (10) as

$$\mathfrak{R}_T(\lambda) \leq \sum_{a \neq a^\star} u_a \left(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \right) + \left(\frac{u\lambda}{2}\sigma^2_{a^\star} + \frac{(A-1)\log(A/\delta)}{\lambda} \right) + \sqrt{2u\log(A/\delta)}\sigma_{a^\star}\,. \tag{11}$$

Thus λ is well-adapted to the environment for instance when $\lambda = \Omega(u^{-1/2})$. Since any reasonable algorithm will pull sub-optimal arms only $u_a = O(\log(T))$ times with high probability, this indicates that a well-adapted level of risk aversion for a Gaussian game of length T is of order[1] $\lambda = \Omega(\log(T)^{-1/2})$. A similar reasoning holds for the sub-Gaussian and thus the bounded case as well, since we only need an upper-bound on the mixability gaps rather than an equality here. In the sequel, we consider such a case, disregarding the extremely challenging question of defining and estimating a distribution-dependent optimally-adapted value of λ (it also conveys difficult interpretation since the optimal arm depends on λ). Note finally that contrary to the empirical regret, the risk-averse regret (9) is completely blind to such situations, as it basically corresponds to the first term in (10).

[1] Such (weak) dependency with T is intuitive: if we only have 10 trials do to something, we would be much more risk-averse (big λ) than with 1000 trials.

Proof. We begin the proof with the following decomposition. For non negative values λ and λ', using the property that $T - N^{\mathfrak{A}}_{T,a^\star} = \sum_{a\neq a^\star} N^{\mathfrak{A}}_{T,a}$, one has by a simple rewriting

$$\mathfrak{R}_T(\lambda) = \sum_{i=N^{\mathfrak{A}}_{T,a^\star}+1}^{T} \Big(X_{i,a^\star} - \kappa_{\lambda',\nu_{a^\star}} \Big) + \sum_{a\neq a^\star} N^{\mathfrak{A}}_{T,a} \Big(\kappa_{\lambda',\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \Big) + \sum_{a\neq a^\star} \sum_{i=1}^{N^{\mathfrak{A}}_{T,a}} \Big(\kappa_{-\lambda,\nu_a} - X_{i,a} \Big).$$

Now, we have on the one hand that $\mathfrak{R}_T(\lambda) \leq T$ under the event Ω, while on the other hand, under its complement Ω^c, we have

$$\mathfrak{R}_T(\lambda) \leq \sum_{a\neq a^\star} u_a \Big| \kappa_{\lambda',\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \Big| + \max_{s\leq \sum_{a\neq a^\star} u_a} \sum_{i=T-s+1}^{T} \Big(X_{i,a^\star} - \kappa_{\lambda',\nu_{a^\star}} \Big) + \sum_{a\neq a^\star} \max_{s\leq u_a} \sum_{i=1}^{s} \Big(\kappa_{-\lambda,\nu_a} - X_{i,a} \Big).$$

In this decomposition, the two last terms are controlled by means of concentration of measure, following suitable adaptations of (1) and (2), and the first term can be further decomposed due to the inequality

$$\Big| \kappa_{\lambda',\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \Big| \leq m^+_{\lambda',\nu_{a^\star}} + m^-_{-\lambda,\nu_{a^\star}} + \Big| \kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \Big|,$$

where we made appear the mixability gaps of the optimal arm $a^\star$. More precisely, the generic bound on the regret of a learning algorithm $\mathfrak{A}$ now relies on the previous decomposition and on the following two concentration results, whose proof is in the Appendix.

Lemma 1. *Let $\tau^\star \in \mathbb{N}$ be some positive constant, and $\lambda' > 0$ be such that $\kappa_{\lambda',\nu_{a^\star}}$ is finite. Then we have the property that for all $\epsilon > 0$, then*

$$\mathbb{P}\Big(\max_{s\leq \tau^\star} \sum_{i=T-s+1}^{T} \Big(X_{i,a^\star} - \kappa_{\lambda',\nu_{a^\star}} \Big) \geq \epsilon \Big) \leq \exp(-\lambda'\epsilon)$$

Lemma 2. *Let $\tau \in \mathbb{N}$ be some positive constant, and $\lambda > 0$ be such that κ_{λ,ν_a} is finite. Then we have the property that for all $\epsilon > 0$, then*

$$\mathbb{P}\Big(\max_{s\leq \tau} \sum_{i=1}^{s} \Big(\kappa_{-\lambda,\nu_a} - X_{i,a} \Big) \geq \epsilon \Big) \leq \exp(-\lambda\epsilon)$$

In order to conclude the proof of Proposition 1, we then apply Lemma 1 to the value $\tau^\star = \sum_{a\neq a^\star} u_a$ with $\epsilon = \frac{\log(2A/\delta)}{\lambda'}$ and then Lemma 2 to $\tau = u_a$ with $\epsilon = \frac{\log(2A/\delta)}{\lambda}$. We deduce, by a union bound, that with probability higher than $1 - \delta - \mathbb{P}(\Omega)$, then

$$\mathfrak{R}_T(\lambda) \leq \sum_{a\neq a^\star} u_a \Big(\kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a} \Big) + \sum_{a\neq a^\star} \frac{\log(2A/\delta)}{\lambda} + \inf_{\lambda'>0} \Big\{ \sum_{a\neq a^\star} u_a \Big(m^+_{\lambda',\nu_{a^\star}} + m^-_{-\lambda,\nu_{a^\star}} \Big) + \frac{\log(2A/\delta)}{\lambda'} \Big\}.$$

4 The Risk-Averse Upper Confidence Bound algorithm

We introduce in this section a strategy $\mathfrak{A}$ that we call the RA-UCB algorithm. From now on, we restrict to the case when all distributions belong to $\mathfrak{M}_1^+(\mathbb{R}_B)$, where $\mathbb{R}_B = (-\infty, B]$ for some known value of B. Thus, let us introduce for all $a \in \mathcal{A}$, the empirical distribution $\hat{\nu}_t(a) \in \mathfrak{M}_1^+(\mathbb{R}_B)$ associated to ν_a, built using the past observations $Y_1, \ldots, Y_t$; we define

$$\hat{\nu}_t(a) \stackrel{\text{def}}{=} \frac{1}{N_{t,a}^{\mathfrak{A}}} \sum_{s=1}^{t} \delta_{Y_s} \mathbb{I}\{A_s = a\} \text{ where } N_{t,a}^{\mathfrak{A}} \stackrel{\text{def}}{=} \sum_{s=1}^{t} \mathbb{I}\{A_s = a\}.$$

Further, for clarity purpose, we now use the notation $\hat{\nu}_{n,a}$ (with a in subscript) in order to denote the empirical distribution built from the n first samples drawn from ν_a, while we reserve the functional notation $\hat{\nu}_t(a)$ for the empirical distribution built from the samples received from arm a up to time t. Naturally, we have that $\hat{\nu}_t(a) = \hat{\nu}_{N_t^{\mathfrak{A}}(a),a}$. More generally, for some distribution ν, we also write $\hat{\nu}_n$ for its empirical distribution built from n samples.

The RA-UCB algorithm is inspired from the strategies introduced by [20, 6, 22, 13, 7] as it selects at time $t+1$ the arm $A_{t+1} = \operatorname{argmax}_{a \in \mathcal{A}} U_t(a)$, where $U_t(a)$ is an upper confidence bound on the risk aversion of arm a at level λ, defined by

$$U_t(a) \stackrel{\text{def}}{=} \sup \left\{ \kappa_{-\lambda,\nu} \,:\, \mathbf{K}(\hat{\nu}_t(a), \kappa_{-\lambda,\nu}) \leq \frac{f(t)}{N_{t,a}} \right\}, \tag{12}$$

and where we introduced the following quantity

$$\mathbf{K}(\hat{\nu}_t(a), r) \stackrel{\text{def}}{=} \inf \left\{ KL(\hat{\nu}_t(a) || \nu) \,:\, \nu \in \mathfrak{M}_1^+(\mathbb{R}_B), \kappa_{-\lambda,\nu} \geq r \right\}. \tag{13}$$

Note that UCB-like algorithms are unnatural in this setting: they are based on empirical *means* only, while we really need to control the tail distributions here. KL-based algorithm are more suitable, and produce much stronger results. Note also that the parameter λ is here the same that defines the level of risk aversion used in the definition of the regret. The algorithm requires another parameter, that is a non-decreasing function of the time f. A typical choice is such that $f(t) = O(\log(t))$, as mentioned in Theorem 1.

A Useful Formulation with Dual Optimality Conditions The definition of the bound (12) may seem quite abstract. In order to make it more computable and explicit, we now provide the following result, that is a dual formulation of the optimization problem appearing in the definition of $\mathbf{K}(\hat{\nu}_t(a), r)$ (see the proof in the appendix).

Lemma 3. *Let $\hat{\nu}_n$ denote with a finite number n of atoms. Then the following dual formulation holds*

$$\mathbf{K}(\hat{\nu}_n, r) = \max \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \left(1 - \frac{\gamma^\star}{\lambda} \left(1 - e^{-\lambda(x_i - r)} \right) \right) : 0 \leq \gamma^\star \leq \frac{\lambda}{1 - e^{-\lambda(B-r)}} \right\}.$$

This result shows that the optimization problem (12) can actually be solved numerically. and is deeply linked to the numerically efficient dual formulation considered for instance in [5], [14], or re-derived more recently in [15] for the related problem of optimal regret bounds in the stochastic multi-armed bandit with expected regret criterion. For completeness, it makes sense to introduce the quantity for distributions $\nu \in \mathfrak{M}_1^+(\mathbb{R}_B)$

$$\tilde{\mathbf{K}}(\nu, r) = \sup \left\{ \mathbb{E}\left[\log \left(1 - \frac{\gamma^\star}{\lambda} \left(1 - e^{-\lambda(X-r)} \right) \right) \right] \,:\, 0 \leq \gamma^\star \leq \frac{\lambda}{1 - e^{-\lambda(B-r)}} \right\}.$$

5 Regret Analysis of the RA-UCB Algorithm

By the generic decomposition result of Proposition 1, we only have to provide a high-probability upper bound on the number of pulls of any sub-optimal arm a, more precisely on the event

$$\Omega \stackrel{\text{def}}{=} \left\{ \exists a \neq a^\star \,:\, N_{T,a}^{\mathfrak{A}} > u_a \right\},$$

In order to control the probability of such on event, let us introduce, for all $a \neq a^\star$, the random time t_a corresponding to the last round when a is chosen, that is we have $N_{t_a,a}^{\mathfrak{A}} = N_{T,a}^{\mathfrak{A}} - 1$ and $N_{t_a+1,a}^{\mathfrak{A}} = N_{T,a}^{\mathfrak{A}}$. For such a t_a, we also have by definition $A_{t_a+1}^{\mathfrak{A}} = a$.

Decomposition of Events (step 1) We start by considering the event $A_{t+1}^{\mathfrak{A}} = a$ for a sub-optimal arm. By definition of the algorithm, we have the property that $U_t(a) \geq U_t(\star)$. This event can be decomposed as

$$\begin{aligned}
\left\{ A_{t+1}^{\mathfrak{A}} = a \right\} &\subset \left\{ U_t(\star) \leq \kappa^\star \right\} \cup \left\{ U_t(\star) > \kappa^\star \text{ and } A_{t+1}^{\mathfrak{A}} = a \right\} \\
&\subset \left\{ U_t(\star) \leq \kappa^\star \right\} \cup \left\{ U_t(a) > \kappa^\star \text{ and } A_{t+1}^{\mathfrak{A}} = a \right\} \\
&\subset \left\{ \mathbf{K}(\hat{\nu}_t(\star), \kappa^\star) \geq \frac{f(t)}{N_{t,\star}^{\mathfrak{A}}} \right\} \cup \left\{ \mathbf{K}(\hat{\nu}_t(a), \kappa^\star) \leq \frac{f(t)}{N_{t,a}^{\mathfrak{A}})} \text{ and } A_{t+1}^{\mathfrak{A}} = a \right\},
\end{aligned}$$

where we introduced here some quantity $\kappa^\star$. We now make use of this decomposition in order to show that, for all choice of constant $\{u_a\}_{a=1,\dots,A}$ with $u_a > 1$, we have

$$\begin{aligned}
&\mathbb{P}\left(\exists a \neq a^\star \,:\, N_{T,a}^{\mathfrak{A}} > u_a \right) \leq \mathbb{P}\left(\exists a \neq a^\star \,:\, N_{t_a,a}^{\mathfrak{A}} > u_a - 1 \text{ and } A_{t_a+1} = a \right) \\
&\leq \mathbb{P}\left(\exists a \neq a^\star \,:\, N_{t_a,a}^{\mathfrak{A}} \geq u_a \text{ and } \mathbf{K}(\hat{\nu}_{t_a}(\star), \kappa^\star) \geq \frac{f(t_a)}{N_{t_a,\star}^{\mathfrak{A}}} \right) \\
&\quad + \mathbb{P}\left(\exists a \neq a^\star \,:\, N_{t_a,a}^{\mathfrak{A}} \geq u_a \text{ and } \mathbf{K}(\hat{\nu}_{t_a}(a), \kappa^\star) \leq \frac{f(t_a)}{N_{t_a,a}^{\mathfrak{A}}} \text{ and } A_{t_a+1}^{\mathfrak{A}} = a \right) \\
&\leq \mathbb{P}\left(\exists n \leq T \,:\, \mathbf{K}(\hat{\nu}_{n,\star}, \kappa^\star) \geq \frac{f(n)}{n} \right) + \sum_{a \neq a^\star} \sum_{n=u_a}^{T} \mathbb{P}\left(\mathbf{K}(\hat{\nu}_{n,a}, \kappa^\star) \leq \frac{f(T)}{n} \right),
\end{aligned}$$

where we used in the last line that $f(t_a) \geq f(N_{t_a,\star}^{\mathfrak{A}})$ for the first term, since f is non-decreasing, and similarly that $f(t_a) \leq f(T)$ together with a union bound for the second term.

Concentration Inequalities (step 2) We now make use of concentration inequalities. More precisely, we first use that for the optimal arm it holds that for the value $\kappa^\star = \kappa_{-\lambda,\nu_{a^\star}}$ and for all $\epsilon > 0$, then

$$\mathbb{P}\left(\mathbf{K}(\hat{\nu}_{n,\star}, \kappa^\star) \geq \epsilon \right) \leq e(n+2)\exp(-n\epsilon)\,, \tag{14}$$

(the proof of which is provided in the Appendix, Proposition 2) and that for any suboptimal arm a, for all $\epsilon > 0$ one can resort to an application of non-asymptotic Sanov's lemma. Indeed, under some conditions that we detail below, an easy consequence of [10, Exercise 2.2.38] is that

$$\mathbb{P}\Big(\mathbf{K}(\hat{\nu}_{n,a},\kappa^\star)\leq\epsilon\Big)\leq\exp\Big(-n\chi_a(\epsilon)\Big)\,, \tag{15}$$

where the quantity $\chi_a(\epsilon)$ is intrinsic to the complexity of the problem of testing whether a is a good arm and is defined by

$$\chi_a\big(\epsilon\big)\stackrel{\text{def}}{=}\inf\Big\{\,KL(\nu||\nu_a)\,:\,\tilde{\mathbf{K}}(\nu,\kappa^\star)\leq\epsilon\Big\}\,. \tag{16}$$

Note that this term satisfies that $\chi_a\big(\tilde{\mathbf{K}}(\nu_a,\kappa^\star)\big)=0$ and that for all $\epsilon<\tilde{\mathbf{K}}(\nu_a,\kappa^\star)$ then $\chi_a\big(\epsilon\big)>0$. In particular, the inequality (15) is non trivial only for such $\epsilon<\tilde{\mathbf{K}}(\nu_a,\kappa^\star)$.

Important remark Note at this point that there may be some topological difficulty in order to meet the conditions needed for (15). As a reminder, these conditions are that $\chi_a(\epsilon)<\infty$, that the set $\{\nu\,:\,\tilde{\mathbf{K}}(\nu,\kappa^\star)\leq\epsilon\}$ is convex and moreover that it is close (for the weak topology). The first condition is not restrictive (the case when it is not met is actually even a favorable situation). Now, the convexity of the considered set easily follows from the biconvexity of the Kullback-Leibler divergence (see [8]) and of the convexity of the set of distributions with high risk-aversion. The latter follows from the concavity of the log function. Now, in order to show that this set is close, it is sufficient to show that $\tilde{\mathbf{K}}(\cdot,\kappa^\star)$ is lower semi-continuous on a set including the limit distribution ν_a, which is also not difficult to prove (following [15] for instance).

Final Control of the Number of Pulls (step 3) So far, by combining the initial decomposition with (14) and (15) together with a union bound, we have shown

$$\mathbb{P}\Big(\exists a\neq a^\star\,:\,N^{\mathfrak{A}}_{T,a}>u_a\Big)\leq\sum_{n=1}^{T}e(n+2)e^{-f(n)}+\sum_{a\neq a^\star}\sum_{n=u_a}^{T}\exp\Big(-n\chi_a\Big(\frac{f(T)}{n}\Big)\Big).$$

The first sum in the right hand side of this inequality is easily controlled. For instance if the parameter function f satisfies that $f(n)\geq\log(2e(n+2)n^2/\delta)$, then it is less than δ. The last sum can be made more explicit. Let us now define the quantity $u_a=(1+\epsilon_a)\big(\frac{f(T)}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}+1\big)$ for all $a\neq a^\star$ and some $\epsilon_a>0$, and further define ϵ_a by

$$\epsilon_a=\inf\left\{\epsilon>0:\frac{1+\epsilon}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}\chi_a\Big(\frac{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}{1+\epsilon}\Big)\geq 1\right\}. \tag{17}$$

Note that such an ϵ_a exists and is finite since χ_a is a non increasing function, so that $\frac{1+\epsilon}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}\chi_a\Big(\frac{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}{1+\epsilon}\Big)$ is an increasing function of ϵ. With such notation, we deduce

$$\begin{aligned}
&\sum_{n=u_a}^{T}\exp\Big(-n\chi_a\Big(\frac{f(T)}{n}\Big)\Big)\leq\exp\Big(-u_a\chi_a\Big(\frac{f(T)}{u_a}\Big)\Big)\Big(1+\sum_{n=1}^{\infty}\exp\Big(-n\chi_a\Big(\frac{f(T)}{u_a}\Big)\Big)\Big)\\
&\leq\exp\Big(-f(T)\frac{1+\epsilon_a}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}\chi_a\Big(\frac{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}{1+\epsilon_a}\Big)\Big)\Big(1+\sum_{n=1}^{\infty}\exp\Big(-n\chi_a\Big(\frac{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}{1+\epsilon_a}\Big)\Big)\Big)\\
&\leq\frac{\delta}{2e(T+2)T^2}\Big(1+\frac{1}{1-\exp\Big(-\chi_a\big(\frac{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)}{1+\epsilon_a}\big)\Big)}\Big).
\end{aligned}$$

where we used in the first line that $n > u_a$ and that χ_a is non increasing, in the second line that

$$\frac{u_a}{f(T)} = (1+\epsilon_a)\Big(\frac{1}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)} + \frac{1}{f(T)}\Big) > \frac{1+\epsilon_a}{\tilde{\mathbf{K}}(\nu_a,\kappa^\star)},$$

and finally in the third line the bound on $f(T)$ and the definition of ϵ_a.

The previous analysis together with the result of Proposition 1 for the empirical risk-averse regret (7) on the one hand, and the definition of the risk-averse regret (9) on the other hand enable us to deduce the following bound on the regret of the `RA-UCB` algorithm, which is the main result of this paper. We further provide Corollary 1 for illustration purpose.

Theorem 1. *Assume that for all $a = 1, \dots, A$, then $\nu_a \in \mathfrak{M}_1^+(\mathbb{R}_B)$, and define ϵ_a by equation (17). Let us define $f(t) = \log(2e(t+e)t^2/\gamma)$ for some $\gamma \in (0,1)$. Then, the risk-averse regret of the `RA-UCB` algorithm is upper bounded at time T by*

$$\overline{\mathfrak{R}}_T(\lambda) \le \sum_{a\neq a^\star} \frac{(1+\epsilon_a)\Delta_a}{\mathbf{K}_a} \log\Big(2e(T+e)T^2/\gamma\Big) + \sum_{a\neq a^\star}(1+\epsilon_a)\Delta_a + \sum_{a\neq a^\star}\gamma\Delta_a T + \frac{\gamma\Delta_a}{2e(T+2)T}\Bigg(1 + \frac{1}{1-\exp\Big(-\chi_a\big(\frac{\mathbf{K}_a}{1+\epsilon_a}\big)\Big)}\Bigg).$$

Further, the empirical regret of the `RA-UCB` algorithm at time T is upper bounded for all $\delta \in [0,1]$ by

$$\mathfrak{R}_T(\lambda) \le \sum_{a\neq a^\star} \frac{(1+\epsilon_a)\Delta_a}{\mathbf{K}_a} \log\Big(2e(T+e)T^2/\gamma\Big) + \sum_{a\neq a^\star}(1+\epsilon_a)\Delta_a + \sum_{a\neq a^\star}\frac{\log(2A/\delta)}{\lambda} + \inf_{\lambda'>0}\Bigg\{\sum_{a\neq a^\star}(1+\epsilon_a)\Big(\frac{\Delta_a}{\mathbf{K}_a}\log\Big(2e(T+e)T^2/\gamma\Big)+1\Big)\Big(m^+_{\lambda',\nu_{a^\star}} + m^-_{-\lambda,\nu_{a^\star}}\Big) + \frac{\log(2A/\delta)}{\lambda'}\Bigg\},$$

where we introduced the optimality gaps $\Delta_a = \kappa_{-\lambda,\nu_{a^\star}} - \kappa_{-\lambda,\nu_a}$, and $\mathbf{K}_a = \tilde{\mathbf{K}}(\nu_a, \kappa_{-\lambda,\nu_{a^\star}})$, with probability higher than

$$1 - \delta - \gamma - \frac{\gamma}{2e(T+2)T^2}\sum_{a\neq a^\star}\Bigg(1 + \frac{1}{1-\exp\Big(-\chi_a\big(\frac{\mathbf{K}_a}{1+\epsilon_a}\big)\Big)}\Bigg).$$

Note that the quantity $1-\exp\Big(-\chi_a\big(\frac{\mathbf{K}_a}{1+\epsilon_a}\big)\Big)$ appearing in the high probability bound is problem dependent but is actually a constant, and that similarly ϵ_a is also a problem dependent constant. In particular, both quantities are independent on T and on the algorithm. Taking into account these remarks together with the discussion after Proposition 1 regarding well-adapted values of λ leads to the more readable corollary

Corollary 1. *Under the same assumptions as Theorem 1, for* $\gamma = \Theta(T^{-1})$*, then*

$$\overline{\mathfrak{R}}_T(\lambda) \leq 5 \sum_{a \neq a^\star} \frac{(1+\epsilon_a)\Delta_a}{\mathbf{K}_a} \log\left(T\right) + O\Big(1\Big). \tag{18}$$

Further, assuming moreover that the distributions of rewards are all sub-Gaussian, and that the level of risk-aversion is $\lambda = \Theta(\log(T)^{-1/2})$*, then for a choice of* $\gamma = \Theta(T^{-\beta})$ *for some* $\beta \geq 0$ *the empirical regret of the* `RA-UCB` *algorithm at time* T *is bounded as*

$$\mathfrak{R}_T(\lambda) \leq c \sum_{a \neq a^\star} \frac{(1+\epsilon_a)\Delta_a}{\mathbf{K}_a} \log\left(T\right) + O\Big(\sqrt{\log(T)}\Big), \tag{19}$$

with probability of order $1 - \delta - o(1)$ *for some constant* $c \leq 4 + \beta$*.*

Note also that assuming $\lambda = \Omega(\log(T)^{-1/2})$ only makes the second to last terms in Theorem 1 $o(\log(T))$ instead of $O(\log(T))$. Thus even for others value of λ (that is a constant), we still have $O(\log(T))$ empirical regret (with possibly larger constants).

Discussion The bound (19) makes appear a first order term scaling with a $\log(T)$, which looks very much like the results for the standard multi-armed bandit with expected regret. Note that that such a dependency is achievable is not obvious since working with risk-aversion is usually considered as much more difficult than working with expectation. This should also be compared to the result of [29], although they consider a different, trickier to interpret, setting. We show however that this is possible.

The constant before the logarithmic term consists of the ratio $\frac{\Delta_a}{\mathbf{K}_a}$ that is also very similar to the known bounds for the expected regret ([6, 22]), up to the constant c, that could definitely be reduced by a more careful analysis and parameter tuning (this is not the main focus of this work), and more importantly the constant $1 + \epsilon_a$. Theorem 1 holds for a larger class of distributions than the one considered e.g. in [22]. The reason for this is precisely because we accept to loose the constant $1 + \epsilon_a$ (as opposed to 1 in their work). This term is not entirely intrinsic: one could easily change the threshold 1 in the definition (17) to a smaller constant at the price of an increased probability term (in the $o(1)$), so that one may a priori optimize this term further. It also more complex than the quantities Δ_a and $\mathbf{K}_a$ whose interpretation is immediate. However, understanding the function $x \to x\chi_a(1/x)$, which is related, by the definition of χ_a, to understanding how $\tilde{\mathbf{K}}(\nu, \kappa_{-\lambda,\nu_{a^\star}})$ varies when ν moves from ν_a to distributions having higher risk-aversion, is definitely needed, and can not be done simply using Δ_a and $\mathbf{K}_a$ (one could use for instance the derivative of $\tilde{\mathbf{K}}(\cdot, \kappa_{-\lambda,\nu_{a^\star}})$ in the direction of high risk-aversion, but this is not a lot more easier to interpret either and still difficult to handle in specific cases). Thus, we here prefer to let the term $1 + \epsilon_a$ as it is, since it is anyway fully *explicit*, and captures such distribution-dependent behavior in a fairly concise way. Note also that a similar difficulty appears in the expected regret setting (see [16], [22], [13]).

Conclusion

In this work, the variant of the stochastic multi-armed bandit problem when one considers looking for the maximally risk-averse arm for a user-defined level of risk-aversion (instead of the mean) is considered. We first provide a generic decomposition of the regret (Proposition 1) for any algorithm that enables to focus on the number of plays of sub-optimal arms only. We make use of a coherent risk measure based on the cumulant generative function and show that it is possible to achieve optimal performance up to a regret that is logarithmic (Corollary 1) in the time horizon (with distribution dependent constants). This logarithmic regret is achieved by some adaptation of existing algorithms designed for the expected regret, together with new concentration results precisely introduced for the control of risk-aversion rather than of the mean, which are of independent interest.

References

1. A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.
2. P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and bellman's principle. *Annals of Operations Research*, 152(1):5–22, 2007.
3. P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
4. Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32:48–77, January 2003.
5. J.M. Borwein and A.S. Lewis. Duality relationships for entropy-like minimization problem. *SIAM Journal on Computation and Optimization*, 29(2):325–338, 1991.
6. A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.
7. O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 2013.
8. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.
9. Boris Defourny, Damien Ernst, and Louis Wehenkel. Risk-aware decision making and dynamic programming. In *NIPS Workshop on Model Uncertainty and Risk in RL*, 2008.
10. A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, second edition, 1998.
11. E.V. Denardo and U.G. Rothblum. Optimal stopping, exponential utility and linear programming. *Mathematical Programming*, 16:228–244, 1979.
12. Eyal Even-Dar, Michael Kearns, and Jennifer Wortman. Risk-sensitive online learning. In *Proceedings of the 17th international conference on Algorithmic Learning Theory*, 2006.
13. A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

14. Hugo Harari-Kermadec. *Vraisemblance empirique généralisée et estimation semi-paramétrique*. PhD thesis, Université Paris–Ouest, December 2006.
15. J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Haifa, Israel, 2010.
16. J. Honda and A. Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85:361–391, 2011.
17. J. Honda and A. Takemura. Finite-time regret bound of a bandit algorithm for the semi-bounded support model. arXiv:1202.2277, 2012.
18. Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18:356–369, 1972.
19. E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. *Lecture Notes in Computer Science, Proceedings of the Algorithmic Learning Theory conference*, 7568:199–213, 2012.
20. T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
21. Yaxin Liu and Sven Koenig. An exact algorithm for solving mdps under risk-sensitive planning objectives with one-switch utility functions. pages 453–460, 2008.
22. O-A. Maillard, R. Munos, and G. Stoltz. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences. In *Proceedings of the 23rd Annual Conference on Learning Theory*, Budapest, Hungary, 2011.
23. Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.
24. J. Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University, 1947.
25. Stephen D. Patek. On terminating markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.
26. H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.
27. R. Tyrrell Rockafellar. Coherent approaches to risk in optimization under uncertainty. *Tutorials in operation Research*, pages 38–61, 2007.
28. A. Salomon and J.-Y. Audibert. Robustness of stochastic bandit policies. *Theoretical Computer Science, Special issue*, 2012.
29. A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Proceedings of Advancezs in neural information processing system*, 2012.
30. W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.
31. W.R. Thompson. On the theory of apportionment. *American Journal of Mathematics*, 57:450–456, 1935.
32. Manfred K. Warmuth and Dima Kuzmin. Online variance minimization. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 514–528, 2006.

A Generic Decomposition of the Regret

In this section, we provide a proof for Lemma 1 and for Lemma 2.

Lemma 1 *Let $\tau^\star \in \mathbb{N}$ be some positive constant, and $\lambda' > 0$ be such that $\kappa_{\lambda',\nu_{a^\star}}$ is finite. Then we have the property that for all $\epsilon > 0$, then*

$$\mathbb{P}\Big(\max_{s\leq\tau^\star}\sum_{i=T-s+1}^{T}\Big(X_{i,a^\star}-\kappa_{\lambda',\nu_{a^\star}}\Big)\geq\epsilon\Big)\leq\exp(-\lambda'\epsilon)$$

Proof. The proof follows by an application of Doob's inequality for submartingales. Let us introduce the quantity $M_s = \sum_{i=T-s+1}^{T}\Big(X_{i,a^\star}-\kappa_{\lambda',\nu_{a^\star}}\Big)$, which is a sum of s iid random variables. Note that M_s is not centered, and is a supermartingale. Indeed, it satisfies that

$$\mathbb{E}\Big[M_{s+1}\Big|M_1,..,M_s\Big] = M_s + \mathbb{E}\Big[X_{T-s,a^\star}\Big] - \kappa_{\lambda',\nu_{a^\star}} \leq M_s$$

On the other hand, we have the property that for suitable values of $\gamma > 0$, then $\exp(\gamma M_s)$ is a submartingale. Indeed, we have the property that

$$\begin{aligned}\log\mathbb{E}\Big[\exp(\gamma M_{s+1})\Big|M_1,..,M_s\Big] &= \gamma M_s + \log\mathbb{E}\Big[\exp\big(\gamma X_{T-s,a^\star} - \frac{\gamma}{\lambda'}\log\mathbb{E}\exp(\lambda' X)\big)\Big]\\ &= \gamma M_s + \gamma\Big(\frac{1}{\gamma}\log\mathbb{E}\exp(\gamma X) - \frac{1}{\lambda'}\log\mathbb{E}\exp(\lambda' X)\Big)\\ &= \gamma M_s + \gamma\Big(\kappa_{\gamma,\nu_{a^\star}} - \kappa_{\lambda',\nu_{a^\star}}\Big)\,.\end{aligned}$$

Thus, for all $\gamma > 0$ such that $\kappa_{\gamma,\nu_{a^\star}} \geq \kappa_{\lambda',\nu_{a^\star}}$, then $\exp(\gamma M_s)$ is a submartingale, and Doob's maximal inequality applies. This shows that for all $\epsilon > 0$, then

$$\begin{aligned}\mathbb{P}\Big(\max_{s\leq\tau^\star} M_s \geq \epsilon\Big) &= \mathbb{P}\Big(\max_{s\leq\tau^\star}\exp(\gamma M_s)\geq\exp(\gamma\epsilon)\Big)\\ &\leq \mathbb{E}\Big[\exp(\gamma M_s)\Big]\exp(-\gamma\epsilon)\\ &= \Big(\mathbb{E}\Big[\exp(\gamma X)\Big]\Big)^{\tau^\star}\Big(\mathbb{E}\Big[\exp(\lambda' X)\Big]\Big)^{-\frac{\gamma}{\lambda'}\tau^\star}\exp(-\gamma\epsilon)\\ &= \exp\Big(\gamma\tau^\star\big(\kappa_{\gamma,\nu_{a^\star}} - \kappa_{\lambda',\nu_{a^\star}}\big)\Big)\exp(-\gamma\epsilon)\,.\end{aligned}$$

We conclude by simply taking $\gamma = \lambda'$ (though this is a suboptimal choice).

Lemma 2 *Let $\tau \in \mathbb{N}$ be some positive constant, and $\lambda > 0$ be such that κ_{λ,ν_a} is finite. Then we have the property that for all $\epsilon > 0$, then*

$$\mathbb{P}\Big(\max_{s\leq\tau}\sum_{i=1}^{s}\Big(\kappa_{-\lambda,\nu_a}-X_{i,a}\Big)\geq\epsilon\Big)\leq\exp(-\lambda\epsilon)$$

Proof. The proof is very similar do that of Lemma 1. We have the property that for $s = 1, \dots, \tau$, $M_s = \sum_{i=1}^{s} \Big(\kappa_{-\lambda,\nu_a} - X_{i,a} \Big)$ is a supermartingale since $\kappa_{-\lambda,\nu_a} \leq \mathbb{E}_{\nu_a}\Big[X\Big]$, and then that $\exp(\gamma M_s)$ is a submartingale for values of γ such that $\kappa_{-\lambda,\nu_a} \geq \kappa_{-\gamma,\nu_a}$.

B Dual Formulation of the Minimization Algorithm

Lemma 3 *Let $\hat{\nu}_n$ denote the empirical distribution build from $\nu \in \mathfrak{M}_1^+(\mathbb{R}_B)$ with n i.i.d samples. We have the following rewriting*

$$\mathbf{K}(\hat{\nu}_n, r) = \max \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \left(1 - \frac{\gamma^\star}{\lambda} \Big(1 - \exp\big(-\lambda(x_i - r)\big) \Big) \right) \,:\, 0 \leq \gamma^\star \leq \frac{\lambda}{1 - \exp(-\lambda(B-r))} \right\}.$$

Proof. Let $x_1 < \cdots < x_n$ denote the support of the empirical distribution $\hat{\nu}_n$ and note that the optimization (13) can be reduced to distributions with support in $\{x_i\}_{i \leq n}$ up maybe to one extra point x_{n+1} that will receive extra weight if needed. For $\nu \ll \hat{\nu}_n$, let p_i be the probability weight associated to x_i, and let q be the one associated to x_{n+1}. The optimization problem can be rewritten in the following form

$$\begin{aligned} \text{minimize over } \{p_i\}_{i \leq n}, q : \quad & \sum_{i=1}^{n} \frac{1}{n} \log \Big(\frac{1/n}{p_i} \Big) \\ \text{subject to} \quad & -\frac{1}{\lambda} \log \Big(\sum_{i=1}^{n} \exp(-\lambda x_i) p_i + \exp(-\lambda x_{n+1}) q \Big) \geq r \,, \end{aligned}$$

and it is immediate to see that $x_{n+1} < x_n$ does not help. Now by introducing appropriate Lagrange-multipliers, this corresponds to minimizing the following quantity over $\{p_i\}_{i \leq n}, q \in \mathbb{R}^+$, $\gamma, \alpha, \{\xi_i\}_{i \leq n}, \xi \in \mathbb{R}$:

$$\begin{aligned} \mathcal{V} \stackrel{\text{def}}{=} & \sum_{i=1}^{n} \frac{1}{n} \log \Big(\frac{1/n}{p_i} \Big) + \gamma \Big(r + \frac{1}{\lambda} \log \Big(\sum_{i=1}^{n} \exp(-\lambda x_i) p_i + \exp(-\lambda x_{n+1}) q \Big) \Big) \\ & + \alpha \Big(1 - \sum_{i=1}^{n} p_i - q \Big) - \sum_{i=1}^{n} \xi_i p_i - \xi q \,, \end{aligned}$$

with Karush-Kuhn Tucker optimality conditions corresponding to a distribution $\nu^\star$ being:

- $\gamma^\star \Big(r + \frac{1}{\lambda} \log \Big(\sum_{i=1}^{n} \exp(-\lambda x_i) p_i^\star + \exp(-\lambda x_{n+1}) q^\star \Big) \Big) = 0$, with the conditions $\gamma^\star \geq 0$ and $-\frac{1}{\lambda} \log \Big(\sum_{i=1}^{n} \exp(-\lambda x_i) p_i^\star + \exp(-\lambda x_{n+1}) q^\star \Big) \geq r$,
- $\sum_{i=1}^{n} p_i^\star + q^\star = 1$,

- $\xi_i^\star p_i^\star = 0$, $\xi_i^\star \geq 0$, $p_i^\star \geq 0$ and $\xi^\star q^\star = 0$, $\xi^\star \geq 0$, $q^\star \geq 0$,
- $-\frac{1}{np_i^\star} + \frac{\gamma^\star}{\lambda}\Big[\sum_{j\neq i=1}^n \exp(\lambda(x_i - x_j))p_j^\star + p_i^\star + \exp(\lambda(x_i - x_{n+1}))q^\star\Big]^{-1} - \xi_i^\star - \alpha^\star = 0$,
- $\frac{\gamma^\star}{\lambda}\Big[\sum_{j=1}^n \exp(\lambda(x_{n+1} - x_j))p_j^\star + q^\star\Big]^{-1} - \xi^\star - \alpha^\star = 0$.

Now, in order for $\mathcal{K}(\hat{\nu}_n, \nu^\star)$ to be finite, we need $p_i^\star > 0$ i.e. $\xi_i^\star = 0$ for all $i \leq n$.

If $\gamma^\star = 0$, then we deduce that $\alpha^\star = -\frac{1}{np_i^\star} = -\xi^\star$ for all $i \leq n$. In particular, this means that $\xi^\star > 0$ and thus that $q^\star = 0$. Thus, since $p_i^\star = \frac{1}{n\xi^\star}$ and that $\sum_{i=1}^n p_i^\star = 1$, then $\xi^\star = 1$ and $p_i^\star = \frac{1}{n}$ for all $i \leq n$. This correspond to the cases when $\mathcal{V} = 0$.

Now let us consider that $\gamma^\star > 0$. In that case, we deduce that we must have for all $i \leq n$

$$\exp(\lambda(x_i - r)) = \sum_{j\neq i=1}^n \exp(\lambda(x_i - x_j))p_j^\star + p_i^\star + \exp(\lambda(x_i - x_{n+1}))q^\star$$

$$\text{and also} \quad \exp(\lambda(x_{n+1} - r)) = \sum_{j=1}^n \exp(\lambda(x_{n+1} - x_j))p_j^\star + q^\star\,,$$

from which we deduce that we must have

- $-\frac{1}{np_i^\star} + \frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star = 0$,
- $\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_{n+1} - r)) - \xi^\star - \alpha^\star = 0$.

Thus, on the one hand we have $p_i^\star = \frac{1/n}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star}$ and $q^\star = 1 - \frac{1}{n}\sum_{i=1}^n \frac{1}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star}$, and on the other hand, we have $q^\star = \exp(\lambda(x_{n+1} - r)) - \frac{1}{n}\sum_{i=1}^n \frac{\exp(\lambda(x_{n+1} - x_i))}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star}$, and it remains to determine $\alpha^\star$ and $\gamma^\star$.

If $q^\star > 0$, then $\xi^\star = 0$ and thus $\alpha^\star = \frac{\gamma^\star}{\lambda}\exp(-\lambda(x_{n+1} - r))$. Thus, this entails that

$$1 - \frac{\lambda}{\gamma^\star}\frac{1}{n}\sum_{i=1}^n \frac{1}{\exp(-\lambda(x_i - r)) - \exp(-\lambda(x_{n+1} - r))} =$$

$$\exp(\lambda(x_{n+1} - r)) - \frac{\lambda}{\gamma^\star}\frac{1}{n}\sum_{i=1}^n \frac{\exp(\lambda(x_{n+1} - x_i))}{\exp(-\lambda(x_i - r)) - \exp(-\lambda(x_{n+1} - r))}\,,$$

i.e. that the optimal value of $\gamma^\star$ is given by

$$\gamma^\star = \frac{\lambda}{1 - \exp(\lambda(x_{n+1} - r))}\left(\frac{1}{n}\sum_{i=1}^n \frac{1 - \exp(-\lambda(x_i - x_{n+1}))}{\exp(-\lambda(x_i - r)) - \exp(-\lambda(x_{n+1} - r))}\right)$$

$$= \frac{\lambda}{1 - \exp(-\lambda(x_{n+1} - r))}\,.$$

We thus have shown that the optimal weights are given by

$$p_i^\star = \frac{1}{n}\frac{1 - \exp(-\lambda(x_{n+1} - r))}{\exp(-\lambda(x_i - r)) - \exp(-\lambda(x_{n+1} - r))}\,,$$

and that the corresponding value is

$$\mathcal{V} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(-\lambda(x_i - x_{n+1})) - 1}{\exp(-\lambda(r - x_{n+1})) - 1}\right).$$

Now if $\xi^\star > 0$, then $\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_{n+1} - r)) > \alpha^\star$, thus since we have

$$\exp(-\lambda(x_{n+1} - r)) - 1 = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-\lambda(x_{n+1} - r)) - \exp(-\lambda(x_i - r))}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star},$$

we deduce that

$$\exp(-\lambda(x_{n+1} - r)) - 1 \geq \frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-\lambda(x_{n+1} - r)) - \exp(\lambda(x_i - r))}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \frac{\gamma^\star}{\lambda}\exp(-\lambda(x_{n+1} - r))},$$

i.e. that $\gamma^\star \leq \frac{\lambda}{1-\exp(-\lambda(x_{n+1}-r))}$.

On the other hand, we must have $q^\star = 0$ and thus

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star} = 1 \text{ and}$$
$$\frac{1}{n}\sum_{i=1}^{n} \frac{\exp(-\lambda(x_i - r))}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star} = 1$$

Thus combining these two equations we deduce that $\frac{\gamma^\star}{\lambda} - \alpha^\star = 1$ and thus that

$$p_i^\star = \frac{1/n}{\frac{\gamma^\star}{\lambda}\exp(-\lambda(x_i - r)) - \alpha^\star} = \frac{1/n}{\frac{\gamma^\star}{\lambda}\Big(\exp(-\lambda(x_i - r)) - 1\Big) + 1}.$$

The value is thus given in that case by

$$\mathcal{V} = \max\left\{\frac{1}{n}\sum_{i=1}^{n} \log\Big(1 - \frac{\gamma^\star}{\lambda}\Big(1 - \exp(-\lambda(x_i - r))\Big)\Big) \,:\, 0 < \gamma^\star < \frac{\lambda}{1 - \exp(-\lambda(x_{n+1} - r))}\right\}.$$

C Concentration Inequalities for $\mathbf{K}(\hat{\nu}_n, r)$

Let us first show the following key intermediate lemma

Lemma *We have the property that there exists some discrete set Γ_c with at most $2 + 1/c$ elements such that*

$$\mathbf{K}(\hat{\nu}_n, r) \leq c + \max_{\gamma \in \Gamma_c} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 - \frac{\gamma}{\lambda}\Big(1 - \exp\big(-\lambda(x_i - r)\big)\Big)\right),$$

Proof. Using the change of variable $\gamma = \frac{\lambda}{1-\exp(-\lambda(x_{n+1}-r))}u$, with $x_{n+1} = B$ this corresponds to showing

$$\max_{u\in[0,1]} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 - u\frac{1-\exp\big(-\lambda(x_i-r)\big)}{1-\exp\big(-\lambda(x_{n+1}-r)\big)}\right) \leq c + \max_{u'\in U_c} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 - u\frac{1-\exp\big(-\lambda(x_i-r)\big)}{1-\exp\big(-\lambda(x_{n+1}-r)\big)}\right).$$

This follows from the fact that for all $u \in [0,1]$ there exists $u' \in U_c$ such that for all $x \leq x_{n+1}$, then

$$\log\left(1 - u\frac{1-\exp\big(-\lambda(x-r)\big)}{1-\exp\big(-\lambda(x_{n+1}-r)\big)}\right) \leq c + \log\left(1 - u'\frac{1-\exp\big(-\lambda(x-r)\big)}{1-\exp\big(-\lambda(x_{n+1}-r)\big)}\right) \quad (20)$$

More precisely we use the fact that for all $u, u' \in [0,1]$ such that $u \leq u' \leq 1/2$ or $u \geq u' \geq 1/2$, then for all $y \leq 1$

$$\log\Big(1-uy\Big) \leq \log\Big(1-u'y\Big) + 2|u'-u|\,.$$

The proof of (20) then follows from simple algebra. We then apply this result with the following grid that has at most $2 + 1/c$ elements

$$U_c \stackrel{\text{def}}{=} \{1/2, 1\} \cup \{1/2 + c, \ldots, 1/2 + \lfloor 1/(2c)\rfloor c\} \cup \{1/2 - c, \ldots, 1/2 - \lfloor 1/(2c)\rfloor c\}\,.$$

We then prove the following proposition, that is the main result of this section.

Proposition 2. *Let $\hat{\nu}_n$ be the empirical distribution built from ν with n i.i.d samples. Then, for all $\epsilon > 0$, it holds that*

$$\mathbb{P}\Big(\mathbf{K}(\hat{\nu}_n, \kappa_{-\lambda,\nu}) \geq \epsilon\Big) \leq e(n+2)\exp(-n\epsilon)\,.$$

Proof. Let $r = \kappa_\nu$. By Lemma C, we know that for all $c > 0$, there exist a finite set $\Gamma_c \subset [0, \gamma^\star \leq \frac{\lambda}{1-\exp(-\lambda(B-r))}]$ with at most $2 + 1/c$ points such that

$$\mathbf{K}(\hat{\nu}_n, r) \leq c + \max_{\gamma\in\Gamma_c} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 - \frac{\gamma}{\lambda}\Big(1-\exp\big(-\lambda(x_i-r)\big)\Big)\right).$$

Then, we thus have for any $\beta > 0$

$$\begin{aligned}
\log \mathbb{E}\exp\Big[\beta\mathbf{K}(\hat{\nu}_n, r)\Big] &\leq \log \mathbb{E}\exp\left[\beta c + \beta\max_{\gamma\in\Gamma_c}\frac{1}{n}\sum_{i=1}^{n}\log\left(1-\frac{\gamma}{\lambda}\Big(1-\exp\big(-\lambda(x_i-r)\big)\Big)\right)\right] \\
&\leq \log\left(\sum_{\gamma\in\Gamma_c}\mathbb{E}_\nu\exp\left[\beta c + \frac{\beta}{n}\sum_{i=1}^{n}\log\left(1-\frac{\gamma}{\lambda}\Big(1-\exp\big(-\lambda(x_i-r)\big)\Big)\right)\right]\right) \\
&\leq \log\left(\exp\Big(\beta c\Big)\sum_{\gamma\in\Gamma_c}\prod_{i=1}^{n}\mathbb{E}_\nu\left[\left(1-\frac{\gamma}{\lambda}\Big(1-\exp\big(-\lambda(x_i-r)\big)\Big)\right)^{\frac{\beta}{n}}\right]\right).
\end{aligned}$$

Thus, for $\beta = n$, we get, since $r = -\frac{1}{\lambda}\log\Big(\mathbb{E}_\nu\Big[\exp\big(-\lambda X\big)\Big]\Big)$,

$$\mathbb{E}_\nu\bigg[\Big(1-\frac{\gamma}{\lambda}\Big(1-\exp\big(-\lambda(X-r)\big)\Big)\Big)^{\frac{\beta}{n}}\bigg] = 1-\frac{\gamma}{\lambda}+\gamma\frac{1}{\lambda}\mathbb{E}_\nu\Big[\exp\big(-\lambda X\big)\Big]\exp(\lambda r) = 1\,,$$

from which we deduce that

$$\log\mathbb{E}\exp\Big[\beta\mathbf{K}(\hat{\nu}_n, r)\Big] \leq \inf_{c>0}\log\bigg(\exp\Big(nc\Big)|\mathit{\Gamma}_c|\bigg)\,.$$

In particular, for $c = 1/n$, we get the bound

$$\log\mathbb{E}\exp\Big[\beta\mathbf{K}(\hat{\nu}_n, r)\Big] \leq \log\Big(e|\mathit{\Gamma}_{1/n}|\Big) \leq \log\Big(e(n+2)\Big)\,,$$

and the result then follows via a simple Markov's inequality.