Optimal Thompson Sampling strategies for support-aware CVaR bandits

Dorian Baudry [1] **Romain Gautron** [2 3] **Emilie Kaufmann** [1] **Odalric-Ambrym Maillard** [1]

Abstract

In this paper we study a multi-arm bandit problem in which the quality of each arm is measured by the *Conditional Value at Risk* (CVaR) at some level α of the reward distribution. While existing works in this setting mainly focus on *Upper Confidence Bound* algorithms, we introduce a new *Thompson Sampling* approach for CVaR bandits on bounded rewards that is flexible enough to solve a variety of problems grounded on physical resources. Building on a recent work by Riou and Honda (2020), we introduce B-CVTS for continuous bounded rewards and M-CVTS for multinomial distributions. On the theoretical side, we provide a non-trivial extension of their analysis that enables to theoretically bound their CVaR regret minimization performance. Strikingly, our results show that these strategies are the first to provably achieve *asymptotic optimality* in CVaR bandits, matching the corresponding asymptotic lower bounds for this setting. Further, we illustrate empirically the benefit of Thompson Sampling approaches both in a realistic environment simulating a use-case in agriculture and on various synthetic examples.

1. Introduction

Over the past few years, a number of works have focused on adapting multi-armed bandit strategies (see e.g. Lattimore and Szepesvari (2019)) to optimize an other criterion than the *expected* cumulative reward. Sani et al. (2012), Vakili and Zhao (2015), Vakili and Zhao (2016), Zimin et al. (2014) consider a mean-variance criterion, (Szorenyi et al., 2015) studies a quantile (Value-at-Risk) criterion, (Maillard, 2013) focuses on Entropic-value-at-risk. The *Conditional Value at Risk* (CVaR) as well as more generic *coherent spectral risk measures* (Acerbi and Tasche, 2002) have received specific attention from the bandit community (Galichet et al. (2013); Galichet (2015); Cassel et al. (2018); Zhu and Tan (2020); Tamkin et al. (2020); Prashanth et al. (2020) to cite a few). Indeed, in a large number of application domains (healthcare, agriculture, marketing,...), one needs to take into account personalized *preferences* of the practitioner that are not captured by the *expected* reward. We consider an illustrative use-case in agriculture in section 4, where an algorithm recommends planting dates to farmers.

The *Conditional Value at Risk* (CVaR) at level $\alpha \in [0,1]$ (see Mandelbrot (1997), Artzner et al. (1999)) is easily interpretable as the expected reward in the worst α-fraction of the outcomes, and hence captures different preferences, from being neutral to the shape of the distribution ($\alpha = 1$, mean criterion) to trying to maximize the reward in the worst-case scenarios (α close to 0, typically in finance or insurance). It is further a coherent spectral measure in the sense of Rockafellar et al. (2000), see Acerbi and Tasche (2002)). Several definitions of the CVaR exist in the literature, depending on whether the samples are considered as *losses* or as *rewards*. Brown (2007), Thomas and Learned-Miller (2019) and Agrawal et al. (2020) consider the *loss* version of CVaR. We here follow Galichet et al. (2013) and Tamkin et al. (2020) who use the *reward* version, defined for arm k with distribution ν_k as

$$\mathrm{CVaR}_\alpha(\nu_k) = \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{\alpha}\mathbb{E}_{X\sim\nu_k}\left[(x-X)^+\right]\right\} . \quad (1)$$

This implies that the best arm is the one with the *largest* CVaR. To simplify the notation we write $c_k^\alpha = \mathrm{CVaR}_\alpha(\nu_k)$ in the sequel. Following e.g. Tamkin et al. (2020), for unknown arm distributions $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ we measure the CVaR regret at time T for some risk-level α of a sequential sampling strategy $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ as

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(T) = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^T \left(\max_k c_k^\alpha - c_{A_t}^\alpha\right)\right] = \sum_{k=1}^K \Delta_k^\alpha \mathbb{E}_{\boldsymbol{\nu}}[N_k(T)], \quad (2)$$

where $\Delta_k^\alpha = \max_{k'} c_{k'}^\alpha - c_k^\alpha$ is the gap in CVaR between arm k and the best arm, and $N_k(t) = \sum_{s=1}^t \mathbb{1}(A_s = k)$ is the number of selections of arm k up to round t. Other notions of regret have been studied for risk-averse bandits, e.g.

[1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9198-CRIStAL, F-59000 Lille, France [2]CIRAD, UPR AIDA, F-34398 Montpellier, France [3]CGIAR Platform for Big Data in Agriculture, Alliance of CIAT and Bioversity International, Km 17 Recta Cali-Palmira, Apartado Aéreo 6713, Cali, Colombia. Correspondence to: Dorian Baudry <dorian.baudry@inria.fr>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

computing the risk metric of the full trajectory of observed rewards (Sani et al. (2012); Cassel et al. (2018); Maillard (2013)), but are less interpretable.

Related work At a high level, the multi-armed bandit literature on the CVaR is largely inspired from adapting the popular Upper Confidence Bounds (UCB) algorithms (Auer et al. (2002)) for bounded distributions to work under this criterion, hence rely on concentration tools for the CVaR. Two main approaches can be distinguished: using an empirical CVaR estimate plus a confidence bound as considered in MaRaB (Galichet et al. (2013); Galichet (2015), U-UCB (Cassel et al., 2018), or exploiting the link between the CVaR and the CDF to build an optimistic CDF as in CVaR-UCB (Tamkin et al., 2020), resorting to the celebrated Dvoretzky–Kiefer–Wolfowitz (DKW) concentration inequality (see Massart (1990)). Indeed DKW inequality has been used for example by Brown (2007) and Thomas and Learned-Miller (2019) to develop concentration inequalities for the empirical CVaR of bounded distributions. These strategies provably achieve a logarithmic CVaR regret in bandit models with bounded distributions[1], with a scaling in $\frac{K \log T}{\alpha^2 \Delta}$ where Δ is the smallest (positive) CVaR gap Δ_k^α. However, the asymptotic optimality of these strategies is not established. Strikingly, few works have tried to adapt to the CVaR setting the *asymptotically optimal* bandit strategies for the mean criterion that provably match the lower bound on the regret given by (Lai and Robbins, 1985), such as KL-UCB (Cappé et al., 2013), Thompson Sampling (TS) (Thompson, 1933; Agrawal and Goyal, 2013; Kaufmann et al., 2012) or IMED (Honda and Takemura, 2015). We note that Zhu and Tan (2020) adapts TS to the slightly different *risk-constrained setting* introduced by Kagrecha et al. (2020) for which the goal is to maximize the mean rewards under the constraint that arms with a small CVaR are not played too often. Unfortunately the analysis is limited to Gaussian distributions and does not target optimality. (A TS algorithm was also proposed by Zhu and Tan (2020) for the mean-variance criterion.)

We believe the reason is two-fold: First, despite asymptotic optimal strategies being appealing to improve practical performances, such strategies were, until recently, relying on assuming known parametric family (Honda and Takemura (2010; 2015); Korda et al. (2013); Cappé et al. (2013) to name a few), such as one-parameter exponential families, deriving one specific algorithm for each family. Unfortunately, assuming a simple parametric distributions may not be meaningful to model complex, realistic situations. Rather, the most accessible information to the practitioner is often whether or not the distribution is discrete, and for the continuous case how it is bounded. That is typically the case in applications such as agriculture, healthcare, or resource management, when the reward distributions are grounded on physical realities. Indeed the practitioner can realistically assume that the support of the distributions is known and bounded, with bounds that can be either natural or provided by experts. For instance, in the use-case we consider in section 4 the algorithm recommends planting dates to farmers to maximize the yield of a maize field, that is naturally bounded. Further, distributions in these settings can have shapes that are not well captured by standard parametric families of distributions, as for instance they can be multi-modal with an unknown number of modes that depend on external factors unknown at the decision time (weather conditions, illness, pests, ...). This suggests one may prefer algorithms that can cover a variety of possible shapes for the distributions, rather than targeting a specific known family. UCB-type strategies assuming only boundedness are thus handy even though not optimal.

Second, targeting asymptotic optimality for CVaR bandits is challenging: Massart's bound for DKW-inequality was already a non-trivial result, solving a long-lasting open question back at the time, and yet only provides a "Hoeffding version" of the CDF concentration. Adapting this to work e.g. with Kullback-Leibler, plus considering that the CVaR writes as an optimization problem, makes the quest for a tight analysis even more challenging, and providing regret guarantees for a CVaR equivalent of kl-ucb and empirical KL-UCB (Cappé et al., 2013) is an interesting direction for future work. Looking at the CVaR community, recent works (Kagrecha et al., 2019; Holland and Haress, 2020; Prashanth et al., 2020) have developed new tools for CVaR concentration. Unfortunately, they may not be adapted for this purpose since they aim at capturing properties of heavy-tail distributions in a highly risk-averse setup. The setting considered in this paper is different, and applying the optimistic principle for CVaR bandits to achieve asymptotic optimality may be a daunting task. This suggests the idea to turn towards alternative methods, such as e.g. TS strategies.

As it turns out, two powerful variants of TS were introduced recently by Riou and Honda (2020) for the mean criterion, that enable to overcome the "parametric" limitation, in the sense that these approaches reach the minimal achievable regret given by the lower bound of Burnetas and Katehakis (1996), respectively for discrete and bounded distributions. This timely contribution opens the room to overcome the two previous limitations and achieve the first provably optimal strategy for CVaR bandit for such practitioner-friendly assumptions.

Remark 1. *In finance CVaR is often associated to heavy-tail distributions. Other variants of bandits have been considered to deal with possibly heavy-tail distributions, or*

[1] Cassel et al. (2018) gives an upper bound on the proxy regret of U-UCB, which is also valid for the smaller CVaR regret. For completeness, we provide in Appendix F an analysis of U-UCB specifically tailored to the CVaR regret.

weak moment conditions: In (Carpentier and Valko, 2014), the authors study regret minimization for extreme statistics (the maximum), for Weibull of Frechet-like distributions. In (Lattimore, 2017), a median-of-mean estimator is studied to minimize regret for distributions with bounded kurtosis. A CVaR strategy has been proposed for the different pure exploration setting (Kagrecha et al., 2019; Agrawal et al., 2020), under weak moment conditions. These works consider a different setup and objective.

Contributions In this paper, we purposely focus on minimizing the CVaR regret considering either distributions with discrete, finite support, or with continuous and bounded support, as we believe this has great practical relevance and is still a relatively unexplored topic in the literature. More precisely, we target first-order *asymptotic optimality* for these (sometimes called "non-parametric") families and first derive in Theorem 1 a lower-bound on the CVaR regret, adapting that of (Lai and Robbins, 1985; Burnetas and Katehakis, 1996) to the CVaR criterion. This simple result highlights the right complexity term that should appear when deriving regret upper bounds. We then introduce in Section 2 B-CVTS for CVaR bandits with bounded support, and M-CVTS for CVaR bandits with multinomial arms, adapting the strategies proposed by Riou and Honda (2020) for the CVaR. We provide in Theorem 2 and Theorem 3 the regret bound of each algorithm, proving asymptotic optimality of these strategies. Up to our knowledge, these are the first results showing asymptotic optimality of a Thompson Sampling based CVaR regret minimization strategy. As expected, adapting the regret analysis from Riou and Honda (2020) is non-trivial; we highlight the main challenges of this adaption in section 3.3. For instance, one of the key challenge was to handle boundary crossing probability for the CVaR, and another difficulty comes in the analysis of the non-parametric B-CVTS due to regularity properties of the Kulback-Leibler projection. In Section 4, we provide a case study in agriculture, making the well-established DSSAT agriculture simulator (Hoogenboom et al., 2019) available to the bandit community, and highlight the benefits of using strategies based on Thompson Sampling in this CVaR bandit setting against state-of-the-art baselines: We compare to U-UCB and CVaR-UCB[2] as they showcase two fundamentally different approaches to build a UCB strategy for a non-linear utility function. The first one is closely related to UCB, the second one exploits properties of the underlying CDF, which may generalize to different risk metrics. As claimed in Tamkin et al. (2020), our experiments confirm that CVaR-UCB generally performs better than U-UCB. However, both TS strategies outperform UCB algorithms that tend to suffer from non-optimized confidence bounds. We complete this study with more classical experiments on synthetic data that also confirm the benefit of TS.

[2]MaRaB is similar to U-UCB but enjoys weaker guarantees.

2. Thompson Sampling Algorithms

We present two novel algorithms based on Thompson Sampling and targeting the lower bound of Theorem 1 on the CVaR-regret, for any specified value of $\alpha \in (0, 1]$. These algorithms are inspired by the first algorithms based on Thompson Sampling matching the Burnetas and Katehakis lower bound for bounded distributions in the expectation setting, recently proposed by Riou and Honda (2020).

Notations We introduce the notation $C_\alpha(\mathcal{X}, p)$ for the CVaR of the distribution of support $\mathcal{X}$ and probability $p \in \mathcal{P}^{|\mathcal{X}|}$, where $\mathcal{P}^n$ denotes the probability simplex of size n. For a multinomial arm k we denote its known support $\mathcal{X}_k = (x_k^1, \dots, x_k^{M_k})$ for some $M_k \in \mathbb{N}$, and its true probability vector p_k. We also define $N_k^i(t)$ as the number of times the algorithm has observed x_k^i for arm k before the time t. For general bounded distributions we denote ν_k the distribution of arm k and introduce $\mathcal{X}_{k,t}$ the set of its observed rewards before time t, augmented with a known upper bound B_k for the support of ν_k. We further introduce $\mathcal{D}_n$ as the uniform distribution on the simplex $\mathcal{P}^n$, corresponding to the Dirichlet distribution $\mathrm{Dir}((1, ..., 1))$.

M-CVTS Thompson Sampling (or posterior sampling) is a general Bayesian principle that can be traced back to the work of Thompson (1933), and that is now investigated for many sequential decision making problems (see Russo et al. (2018) for a survey). Given a prior distribution on the bandit model, Thompson Sampling is a randomized algorithm that selects each arm according to its posterior probability of being optimal. This can be implemented by drawing a possible model from the posterior distribution, and acting optimally in the sampled model. For multinomial distribution M-CVTS (Multinomial-CVaR-Thompson-Sampling), described in Algorithm 1, follows this principle. For each arm k, p_k is assumed to be drawn from $\mathcal{D}_{M_k}$, the uniform prior on $\mathcal{P}^{M_k}$. The posterior distribution at a time t is $\mathrm{Dir}(\beta_{k,t})$, with $\beta_{k,t} = (N_k^i(t) + 1)_{i \in \{1,\dots,M_k\}}$. At time t, M-CVTS draws a sample $w_{k,t} \sim \mathrm{Dir}(\beta_{k,t})$ for each arm k and computes $c_{k,t}^\alpha = C_\alpha(\mathcal{X}_k, w_{k,t})$. Then, it selects $A_t = \mathrm{argmax}_k c_{k,t}^\alpha$. For $\alpha = 1$, this algorithm coincides with the Multinomial Thompson Sampling algorithm of Riou and Honda (2020).

B-CVTS We further introduce the B-CVTS algorithm (for Bounded-CVaR-Thompson-Sampling) for general bounded distributions. B-CVTS, stated as Algorithm 2, bears some similarity with a Thompson Sampling algorithm, although it *does not* explicitly use a prior distribution. The algorithm retains the idea of using a noisy version of ν_k, obtained by a *random re-weighting* of the previous observations. Hence,

Algorithm 1 M-CVTS

Input: Level α, horizon T, K, supports $\mathcal{X}_1, \dots, \mathcal{X}_K$
Init.: $t \leftarrow 1$, $\forall k \in \{1, ..., K\}$, $\beta_k = \underbrace{(1, \dots, 1)}_{M_k}$
for $t \in \{2, \dots, T\}$ **do**
  for $k \in \{1, \dots, K\}$ **do**
    Draw $w_k \sim Dir(\beta_k)$.
    Compute $c_{k,t} = C_\alpha(\mathcal{X}_k, w_k)$.
  Pull arm $A_t = \text{argmax}_{k \in \{1,\dots,K\}} c_{k,t}$.
  Receive reward r_{t,A_t}.
  Update $\beta_{A_t}(j) = \beta_{A_t}(j) + 1$, for j as $r_{t,A_t} = x_k^j$

at a time t the index used by the algorithm for an arm k is simply $c_{k,t} = C_\alpha(\mathcal{X}_{k,t}, w_{k,t})$, where $w_{k,t} \sim \mathcal{D}_{N_k(t)}$ is drawn uniformly at random in the simplex $\mathcal{P}^{|\mathcal{X}_{k,t}|}$. B-CVTS then selects the arm $A_t = \text{argmax}_k c_{k,t}$. For $\alpha = 1$, this algorithm coincides with the Non Parametric Thompson Sampling of Riou and Honda (2020) (NPTS). NPTS can be seen as an algorithm that computes for each arm a random average of the past observations. Our extension to CVAR-bandits required to interpret this operation as the computation of the *expectation* of a *random perturbation* of the empirical distribution, which can be replaced by the computation of the CVaR of this new distribution. Note that this idea generalizes beyond using the CVaR, that can be replaced with any criterion.

Algorithm 2 B-CVTS

Input: Level α, horizon T, K, upper bounds $B_1, \dots, B_K$
Init.: $t = 1$, $\forall k \in \{1, ..., K\}$, $\mathcal{X}_k = \{B_k\}$, $N_k = 1$
for $t \in \{2, \dots, T\}$ **do**
  for $k \in \{1, \dots, K\}$ **do**
    Draw $w_k \sim \mathcal{D}_{N_k}$
    Compute $c_{k,t} = C_\alpha(\mathcal{X}_k, w_k)$
  Pull arm $A_t = \text{argmax}_{k \in \{1,\dots,K\}} c_{k,t}$.
  Receive reward r_{t,A_t}.
  Update $\mathcal{X}_{A_t} = \mathcal{X}_{A_t} \cup \{r_{t,A_t}\}$, $N_{A_t} = N_{A_t} + 1$.

Remark 2. *Interestingly, B-CVTS also applies to multinomial distributions (that are bounded). The resulting strategy differs from M-CVTS due to the initialization step using the knowledge of the support in M-CVTS.*

3. Regret Analysis

In this section we prove, after defining this notion, that M-CVTS and B-CVTS are *asymptotically optimal* in terms of the CVaR regret for the distributions they cover.

3.1. Asymptotic Optimality in CVaR bandits

Lai and Robbins (1985) first gave an asymptotic lower bound on the regret for parameteric distribution, that was later extended by Burnetas and Katehakis (1996) to more general classes of distributions. We present below an intuitive generalization of this result for CVaR bandits.

Definition 1. *Let $\mathcal{C}$ be a class of probability distributions, $\alpha \in (0, 1]$, and $\mathrm{KL}(\nu, \nu')$ be the KL-divergence between $\nu \in \mathcal{C}$ and $\nu' \in \mathcal{C}$. For any $\nu \in \mathcal{C}$ and $c \in \mathbb{R}$, we define*

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{C}}(\nu, c) := \inf_{\nu' \in \mathcal{C}, \nu' \neq \nu} \{\mathrm{KL}(\nu, \nu') : \mathrm{CVaR}_\alpha(\nu') \geq c\} .$$

Theorem 1 (Regret Lower Bound in CVaR bandits)**.** *Let $\alpha \in (0, 1]$. Let $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_K$ be a set of bandit models $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K)$ where each ν_k belongs to the class of distribution $\mathcal{F}_k$. Let $\mathcal{A}$ be a strategy satisfying $\mathcal{R}_{\boldsymbol{\nu}}^\alpha(\mathcal{A}, T) = o(T^\beta)$ for any $\beta > 0$ and $\nu \in \mathcal{F}$. Then for any $\nu \in \mathcal{D}$, for any sub-optimal arm k, under the strategy $\mathcal{A}$ it holds that*

$$\lim_{T \to +\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[N_k(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{F}_k}(\nu_k, c^\star)},$$

where $c^\star = \max_{i \in [K]} \mathrm{CVaR}_\alpha(\nu_i)$.

Using (2), this result directly yields an asymptotic lower bound on the regret. The proof of Theorem 1 follows from a classical change-of-distribution argument, as that of any lower bound proof in the bandit literature. We detail it in Appendix D.1, following the proof of Theorem 1 in Garivier et al. (2019) originally stated for $\alpha = 1$. We discuss in Appendix D.2 how this lower bound yields a weaker regret bound expressed in terms of the CVaR gaps (by Pinsker).

In the next section we prove that M-CVTS matches the lower bound for the set of multinomial distribution when the support is known, and that B-CVTS matches the lower bound for the set of continuous bounded distribution with a known upper bound. Hence, under these hypotheses, the two algorithms are *asymptotically optimal*. Despite the recent development in CVaR bandits literature, to our knowledge no algorithm has been able to match this lower bound yet. These results are of particular interest because they show that this bound is attainable for CVaR bandit algorithms, at least for bounded distributions.

3.2. Regret Guarantees for M-CVTS and B-CVTS

Our main result is the following regret bound for M-CVTS, showing that it is matching the lower bound of Theorem 1 for multinomial distributions.

Theorem 2 (Asymptotic Optimality of M-CVTS)**.** *Let $\boldsymbol{\nu}$ be a bandit model with K arms, where the distribution of each arm $k \in \{1, \dots, K\}$ is multinomial with known support $\mathcal{X}_k \subset \mathbb{R}^{M_k}$ for some $M_k \in \mathbb{N}$. The regret of M-CVTS satisfies*

$$\mathcal{R}_{\boldsymbol{\nu}}(T) \leq \sum_{k : \Delta_k^\alpha > 0} \frac{\Delta_k^\alpha \log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{X}_k}(\nu_k, c_1^\alpha)} + o(\log T) .$$

We then provide a similar result for B-CVTS, for bounded and continuous distributions with a known upper bound.

Theorem 3 (Asymptotic Optimality of B-CVTS)**.** *Let $\boldsymbol{\nu}$ be a bandit model with K arms, where for each arm $k \in \{1, \dots, K\}$ its distribution ν_k belongs to $\mathcal{B}_k$, the set of continuous bounded distributions, and its supports $\mathcal{X}_k$ satisfies $\mathcal{X}_k \subset [0, B_k]$ for some known $B_k > 0$. Then the regret of B-CVTS on $\boldsymbol{\nu}$ satisfies*

$$\mathcal{R}_{\boldsymbol{\nu}}(T) \leq \sum_{k:\Delta_k^\alpha > 0} \frac{\Delta_k^\alpha \log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(\nu_k, c_1^\alpha)} + o(\log T) \ .$$

We postpone the detailed proofs of Theorem 2 and Theorem 3 respectively to Appendix B and Appendix C, and we highlight their main ingredients in this section. First, using Equation (2) it is sufficient to upper bound $\mathbb{E}[N_k(T)]$ for each sub-optimal arm k. To ease the notation we assume that arm 1 is optimal. Our analysis follows the general outline of that of Riou and Honda (2020), but requires several novel elements that are specific to CVaR bandits. First, the proof leverages some properties of the function $\mathcal{K}_{\inf}^\alpha$ for the sets of distributions we consider. Secondly, it requires novel boundary crossing bounds for Dirichlet distributions that we detail in Section 3.3.

The first step of the analysis is almost identical for the two algorithms and consists in upper bounding the number of selections of a sub-optimal arm by a *post-convergence* term (Post-CV) and a *pre-convergence* term (Pre-CV). The first term controls the probability that a sub-optimal arm *over-performs* when its empirical distribution is "close" to the true distribution of the arm, while the second term considers the alternative case. To measure how close two distributions are we use the L^∞ distance for multinomial distributions, while for general continuous arms we use the Levy distance (See Appendix A for definitions and details). We state the decomposition in Equation 3 below for a generic distance $d(F_{k,t}, F_k)$ between the empirical cdf of the arm at a time t and its true cdf. As in Section 2 we write $c_{k,t}^\alpha$ for the index assigned to arm k by the algorithm at time t. Then, for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ we define the events

$$\begin{aligned}\mathcal{C}_{t,k}^+ &= \{A_t = k, c_{k,t} \geq c_1^\alpha - \varepsilon_1, d(F_{k,t}, F_k) \leq \varepsilon_2\} \ , \\ \mathcal{C}_{t,k}^- &= \{A_t = k, c_{k,t} < c_1^\alpha - \varepsilon_1\} \\ &\quad \cup \{A_t = k, d(F_{k,t}, F_k) \geq \varepsilon_2\} \ .\end{aligned}$$

As $\{c_{k,t} \geq c_1^\alpha - \varepsilon_1, d(F_{k,t}, F_k) \leq \varepsilon_2\}$ is the complementary set of $\{c_{k,t} < c_1^\alpha - \varepsilon_1\} \cup \{d(F_{k,t}, F_k) > \varepsilon_2\}$ we obtain

$$\mathbb{E}[N_k(T)] \leq \underbrace{\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\mathcal{C}_{t,k}^+)\right]}_{\text{(Post-CV)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(\mathcal{C}_{t,k}^-)\right]}_{\text{(Pre-CV)}} \ . \quad (3)$$

For an arm k satisfying the hypothesis of Theorem 2, for all $\varepsilon > 0$ we show that the corresponding Post-Convergence term of M-CVTS satisfies

$$\text{(Post-CV)} \leq \frac{(1+\varepsilon)\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{X}_k}(\nu_k, c_1^\alpha)} + \mathcal{O}(1) \ , \quad (4)$$

while for an arm k satisfying the hypothesis of Theorem 3, for all $\varepsilon > 0$ the corresponding Post-Convergence term of B-CVTS satisfies

$$\text{(Post-CV)} \leq \frac{\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(\nu_k, c_1^\alpha) - \varepsilon} + \mathcal{O}(1) \ . \quad (5)$$

Finally, for both algorithms the Pre-Convergence term is asymptotically negligible for the families of distribution they cover, namely

$$\text{(Pre-CV)} = \mathcal{O}(1) \ . \quad (6)$$

We detail these results in Appendix B and Appendix C. In the next section we present some novel technical tools that we introduced in order to prove these results.

3.3. Technical challenges and tools

The proofs of (4), (5) and (6) follow the outline of Riou and Honda (2020), respectively for Multinomial Thompson Sampling and Non Parametric Thompson Sampling. However, replacing the linear expectation by the CVaR that is non-linear, causes several technical challenges that make the adaptation non-trivial. This is particularly true for the *boundary crossing probabilities* for Dirichlet random variables, that we define and analyze in this section. Our results aim at replacing the Lemma 13, 14, 15 and 17 of Riou and Honda (2020) in the proofs of Theorem 2 and Theorem 3.

Boundary crossing probabilities In this paragraph we highlight the construction of *boundary crossing* probabilities for Dirichlet random variables, which consists in providing upper and lower bounds of some terms of the form

$$\mathbb{P}_{w \sim \mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X}, w) \geq c\right),$$

for some known support $\mathcal{X} = (x_1, \dots, x_n)$, parameter $\beta \in \mathbb{R}_+^n$ of the Dirichlet distribution, and some real value c that will be defined in context. We introduce the set

$$\mathcal{S}_{\mathcal{X}}^\alpha(c) \quad = \quad \{p \in \mathcal{P}^n : C_\alpha(\mathcal{X}, p) \geq c\} \ ,$$

following the notations of Section 2 for $C_\alpha(\mathcal{X}, p)$. Thanks to the expression of the CVaR in Equation (1) we have

$$\mathcal{S}_{\mathcal{X}}^\alpha(c) = \cup_{m=1}^n \mathcal{S}_{m,\mathcal{X}}^\alpha(c) \ , \quad (7)$$

where we defined for all $m \in \{1, \dots, n\}$ the sets

$$\mathcal{S}^{\alpha}_{m,\mathcal{X}}(c) = \left\{ p \in \mathcal{P}^n, x_m - \frac{1}{\alpha} \sum_{i=1}^{n} p_i \left(x_m - x_i\right)^+ \geq c \right\}.$$

This set is closed and convex, hence $\mathcal{S}^{\alpha}_{\mathcal{X}}(c)$ is closed, and is the finite union of convex sets (but is not convex). These properties are crucial to prove the results of this section.

Bounded support size We first study the case when the size of the support is $|\mathcal{X}| = M$, for some known $M \in \mathbb{N}$ and when the considered distributions are the *frequency* of each observation in $\mathcal{X}$ out of $n \in \mathbb{N}$ many observations, which we represent by the set

$$\mathcal{Q}^M_n = \left\{ (\beta, p) \in \mathbb{N}^{*n} \times \mathcal{P}^M : p = \frac{\beta}{n} \right\} .$$

We then express bounds for boundary crossing probabilities on this set, in terms of n and M, where n should be considered much larger than M. Lemma 1 and 2 respectively provide an upper and lower bound on such probabilities.

Lemma 1 (Upper Bound). *For any $(\beta, p) \in \mathcal{Q}^M_n$, for any $c > C_\alpha(\mathcal{X}, p)$, it holds that*

$$\mathbb{P}_{w \sim Dir(\beta)}(w \in \mathcal{S}^{\alpha}_{\mathcal{X}}(c)) \leq C_1 M n^{M/2} \exp(-n \mathcal{K}^{\alpha,\mathcal{X}}_{\inf}(p, c)) ,$$

for some constant C_1.

Lemma 2 (Lower Bound). *For any $(M, n) \in N^2$ and $(\beta, p) \in \mathcal{Q}^M_n$, if n is large enough it holds that*

$$\mathbb{P}_{w \sim \mathrm{Dir}(\beta)}\left(w \in \mathcal{S}^{\alpha}_{\mathcal{X}}(c)\right) \geq C_2 \frac{\exp\left(-n \mathcal{K}^{\alpha,\mathcal{X}}_{\inf}(p, c)\right)}{n^{\frac{3M}{2}+1}} ,$$

for some constant $C_2 = \left(\frac{1}{\sqrt{2\pi}}\right)^M e^{-(M+1)/12}$.

The details of the proofs of these two results are to be found in Appendix E. Lemma 1 hinges on the Lemma 13 of Riou and Honda (2020) (see Appendix E), while the proof of Lemma 2 shares the core idea of the proof sketch of their Lemma 14. For both results we exploit the convexity of the sets $\mathcal{S}^{\alpha}_{m,\mathcal{X}}(c)$ (equation (7)). Lemma 2 is used in the proof of M-CVTS only. On the other hand, Lemma 1 is a core component of the proof of both M-CVTS and B-CVTS due to the quantization arguments used in the latter.

General support size We now detail some results that are specifically designed for the regret analysis of B-CVTS. For this reason, we consider a support $\mathcal{X} = (x_1, \dots, x_n)$ and the Dirichlet distribution $\mathcal{D}_n$ defined in Section 2. Here we focus on the Dirichlet sample, hence the support $\mathcal{X}$ is known. We further denote $u_{\mathcal{X}}$ the uniform distribution on $\mathcal{X}$, and $C_\alpha(\mathcal{X})$ its CVaR. We first establish an upper bound.

Lemma 3. *Let $\mathcal{X} = (x_0, \dots, x_n) \subset [0, B]^{n+1}$ for some known $B > 0$ and $n \in \mathbb{N}$, assuming that $x_0 = B$. For any $c > C_\alpha(\mathcal{X})$, and any $\eta > 0$ small enough it holds that*

$$\mathbb{P}_{w \sim \mathcal{D}_n}(C_\alpha(\mathcal{X}, w) \geq c) \leq \frac{B}{\eta} \exp^{-N(\mathcal{K}^{\alpha}_{\inf}(u_{\mathcal{X}}, c) - \eta C(B, \alpha, c))} ,$$

for some constant $C(B, \alpha, c)$.

We prove this result in Appendix E. It relies on deriving the dual form of the functional $\mathcal{K}^{\alpha}_{\inf}$ for discrete distributions, that is a result of independent interest.

Lemma 4. *If a discrete distribution F supported on $\mathcal{X}$ satisfies $\mathbb{E}_F\left[\frac{(y-c)\alpha}{(y-X)^+}\right] < 1$, then for any $c > CVaR_\alpha(F)$ it holds that*

$$\mathcal{K}^{\alpha}_{\inf}(F, c) = \inf_{y \in \mathcal{X}} \max_{\lambda \in \left[0, \frac{1}{\alpha(y-c)}\right)} g(y, \lambda, X) ,$$

with $g(y, \lambda, X) = \mathbb{E}_F\left[\log(1 - \lambda((y-c)\alpha) - (y-X)^+)\right]$.

If $\mathbb{E}_F\left[\frac{(y-c)\alpha}{(y-X)^+}\right] \geq 1$, then for any $c > CVaR_\alpha(F)$

$$\mathcal{K}^{\alpha}_{\inf}(F, c) = \inf_{y \in \mathcal{X}} \mathbb{E}_F\left(\frac{(y-X)^+}{(y-c)\alpha}\right) .$$

The detailed proof of this result is provided in Appendix D, where we also show that this expression matches the result of Honda and Takemura (2010) for $\alpha = 1$, and is similar to the one obtained by (Agrawal et al., 2020)[Theorem 6] for a more complex set of distributions (which is hence less explicit). Furthermore, Agrawal et al. (2020)[Lemma 4] prove the continuity of $\mathcal{K}^{\alpha,\mathcal{X}}_{\inf}$ under this condition, which is required in several part of our proofs. We propose a simplified proof of this result for the restriction to bounded distribution in Appendix D.

The last result we report in this section is a lower bound on the probability that a *noisy* CVaR in B-CVTS exceeds the CVaR of the empirical distribution.

Lemma 5. *Assume that $\mathcal{X} = (x_1, \dots, x_n)$ and $x_1 < \dots < x_n$, then $x_{\lceil n\alpha \rceil}$ is the empirical α quantile of the set and x_1 its minimum, and it holds that*

$$\mathbb{P}_{w \sim \mathcal{D}_n}\left(C_\alpha(\mathcal{X}, w) \geq C_\alpha(\mathcal{X})\right) \geq \frac{1}{25n^3}(x_{\lceil n\alpha \rceil} - x_1) .$$

This result is proved in Appendix E. Let us remark that in all the results presented in this section we consider a fixed support $\mathcal{X}$, while in B-CVTS the support is random and evolves with the time. This causes several challenges in the proof. In particular, the use of Lemma 5 in Appendix C.2.2 is not sufficient in itself to conclude and additional work is required to handle the random support.

Remark 3. *The results presented in this section contains most of the difficulty induced by the replacement of the expectation by the CVaR in the proofs. Extending these results to other criterion is an interesting future work and may help generalize the Non Parametric Thompson Sampling algorithms to broader settings.*

4. Experiments

In this section we report the results of experiments on the algorithms presented in the previous sections, first on synthetic examples, and then on a use-case study in agriculture based on the DSSAT agriculture simulator.

4.1. Preliminary Experiments

We first performed various experiments on synthetic data in order to check the good practical performance of M-CVTS and B-CVTS on settings that are simple to implement and are good illustrative examples of the performance of the algorithms. Due to space limitation, we report a complete description of the experiments and and an analysis of the results in Appendix G. We tested the TS algorithms on specified difficult instances and on randomly generated problems, against U-UCB and CVaR-UCB.

As an example of experiment with multinomial arms, we report in Table 1 the results of an experiment with 10^3 randomly generated problems with 5 arms drawn uniformly at random in $\mathcal{P}^{|\mathcal{X}|}$, where $\mathcal{X} = [0, 0.1, 0.2, \dots, 1]$, for $\alpha \in \{10\%, 50\%, 90\%\}$ and an horizon 10^4. These experiments confirm the benefits of TS over UCB approaches, as M-CVTS significantly outperforms its competitors for all levels of the parameter α. We also tested the algorithms with fixed instances (see Tables 5-8), with the same results, and further illustrated the asymptotic optimality of M-CVTS in Figures 7 and 8 by representing the lower bound presented in Section 3 along with the regret of the algorithm in logarithmic scale.

We also tested B-CVTS on different problems, using truncated gaussian mixtures (TGM). The results are presented in Tables 9-12, and again show the merits of the TS approach. We also performed an experiment with a small level $\alpha = 1\%$ (Table 13) and show that B-CVTS keeps the same level of performance in this case, while the other algorithm stay in the linear regime for the horizon we consider. Finally, we also experimented more arms ($K = 30$) and randomly generated TGM problems and report the results in Table 2. The means and variance of each arm satisfy $(\mu_k, \sigma_k) \sim \mathcal{U}([0.25, 1]^1 0 \times [0, 0.1]^1 0)$, and the probabilities of each mode are drawn uniformly, $p_k \sim \mathcal{D}_{19}$.

These very good results with synthetic data and its theoretical guarantees motivate using the B-CVTS algorithm in the real-world application we introduce in the next section.

Table 1: CVaR regret at time $T = 10^4$, averaged over 10^3 random instances with 5 multinomial arms supported on $\mathcal{X} = [0.1, 0.2, \dots, 1]$

α	U-UCB	CVaR-UCB	M-CVTS
10%	633.1	219.7	**38.8**
50%	368.8	187.9	**48.9**
90%	188.5	186.2	**42.7**

Table 2: Results for TGM arms with 10 modes, at $T = 10000$ averaged over 400 random instances with $K = 30$, $\alpha = 5\%$ (results: mean (std)).

T	U-UCB	CVaR-UCB	B-CVTS
10000	2149.9 (263)	2016.0 (265)	**210.9 (6.4)**
20000	4276.4 (538)	3781.3 (521)	**237.1 (15.4)**
40000	8493.4 (1085)	6894.1 (985)	**263.5 (17.9)**

4.2. Bandit application in Agriculture

Motivation Let us consider a farmer who must decide on a *planting date* (action) for a rainfed crop. Farmers have been reported to primarily seek advice that reduces uncertainty in highly uncertain decision making (McCown, 2002; Hochman and Carberry, 2011; Evans et al., 2017). Planting date is an example of such a decision as it will influence the probabilities of favorable meteorologic events during crop cultivation. These events are highly uncertain due to the length of crop growing cycles (e.g. 3 to 6 months for grain maize). For instance, because of the stochastic nature of the rainfalls and temperatures, a farmer will observe a range of different crop yields from year to year for the same planting date, all other technical choices being equal. Thus, assuming that the environment is stationary, each planting date corresponds to an underlying, unknown yield distribution, which can be modeled as an arm in a *bandit problem*. Depending on her profile, a farmer may be more or less risk averse, and the *Conditional Value at Risk* can be used to personalize her level of risk-aversion. For instance, a small-holder farmer looking for food security may seek to avoid very poor yields compromising auto-consumption (e.g $\alpha \leq 20\%$), while a market-oriented farmer may be more prone to risky choices in order to increase her profit but still not risk neutral (e.g $\alpha = 80\%$). Yield distributions are supposed to be *bounded*. Indeed, a finite yield potential can be defined under non-stressing conditions for a given crop and environment (Evans and Fischer, 1999; Tollenaar and Lee, 2002). Observed yields can be modeled as following Von Liebig's law of minimum (Paris, 1992): limiting factors will determine how much of the yield potential can be expressed.

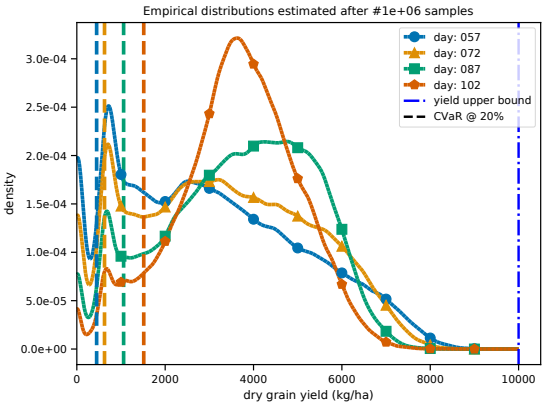


Figure 1: Empirical simulated yields and respective CVaRs at 20% estimated after 10^6 samples in DSSAT environment.

Setting Planting date decision-making support requires extensive testing prior to any real-life application, due the potential impact of wrong action-making, particularly in subsistence farming. For this reason, we consider the problem of facing many times the decision of a planting date in the DSSAT[3] simulator, to make an *in silico* decision. DSSAT, standing for *Decision Support System for Agrotechnology Transfer*, is a world-wide crop simulator, supporting 42 different crops, with more than 30 years of development (Hoogenboom et al., 2019). We specifically address maize planting date decision, as maize is a crucial crop for global food security (Shiferaw et al., 2011). Each simulation is assumed to be realistic, and starts from the same field initial conditions as ground measured. The simulator takes as input historical weather data, field soil measures, crop specific genetic parameters and a given crop management plan. Modeling is based on simulations of atmospheric, soil and plants compartments and their interactions. In the considered experiments, after a decision is made on planting date in the simulator, daily stochastic meteorologic features are generated according to historical data (Richardson and Wright, 1984) and injected in the complex crop model. At the end of crop cycle, a maize grain yield is measured to evaluate decision-making. We parameterized the crop-model under challenging rainfed conditions on shallow sandy soils, i.e. with poor water retention and fertility. Such experiment intends to be representative of realistic conditions faced by small-holder farmers under heavy environmental constraints, such as in Sub-Saharan Africa. Thus, this setting can help picturing how CVaR bandits may perform in real-world conditions. For the sake of the experiments, we built a bandit-oriented Python wrapper to DSSAT that we made available[4] to the bandit community for reproducibility.

[3] DSSAT is an Open-Source project maintained by the DSSAT Foundation, see https://dssat.net/.

[4] https://github.com/rgautron/DssatBanditEnv

Experiments We test bandit performances on the 4 armed DSSAT environment described in Table 3. To illustrate the non-parametric nature of these distributions, we report in Figure 1 estimations of their density obtained with Monte-Carlo simulations, as well as of their CVaRs. The resulting distributions are typically *multi-modal*, with one of their mode very close to zero (years of bad harvest), and with upper tails that cannot be properly characterized. However the practitioner can realistically assume that the distributions are upper-bounded, due to the physical constraints of crop-farming. The yield upper-bound is set to 10 t/ha thanks to expert knowledge for the considered conditions.

Table 3: Empirical yield distribution metrics in kg/ha estimated after 10^6 samples in DSSAT environment

day (action)	CVaR$_\alpha$			
	5%	20%	80%	100% (mean)
057	0	448	2238	3016
072	46	627	2570	3273
087	287	1059	3074	**3629**
102	**538**	**1515**	**3120**	3586

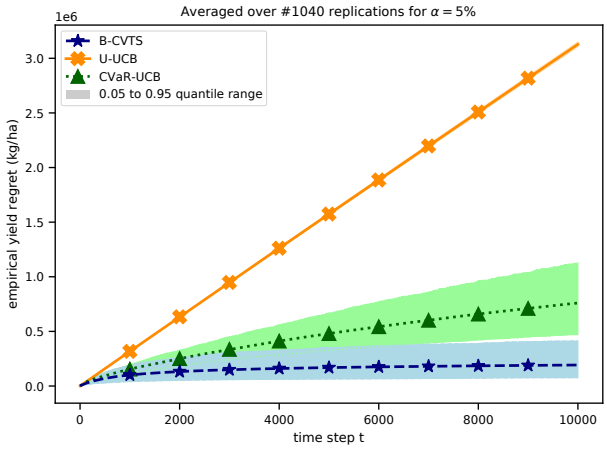
The presented DSSAT environment advocates for the use of algorithms specifically designed for CVaR bandits, as the optimal arm can change depending on the value of the parameter α. Our experiment consists in running 64 trajectories for three algorithms U-UCB, CVaR-UCB and B-CVTS defined in Section 2. Experiments are carried out with an horizon of 10^4 time steps, and we compare the results for each algorithm for $\alpha \in \{5\%, 20\%, 80\%\}$ to see how the parameter impacts their performance. Indeed we want a strategy to perform well on all α choices, allowing to freely model any farmer's risk aversion level. As shown in Figure 2 and Table 4, B-CVTS appears to be consistently better than its UCB counterparts in DSSAT environment for all tested α values, which is encouraging for real-life applications.

Table 4: Empirical yield regrets at horizon 10^4 in t/ha in DSSAT environment, for 1040 replications. Standard deviations in parenthesis.

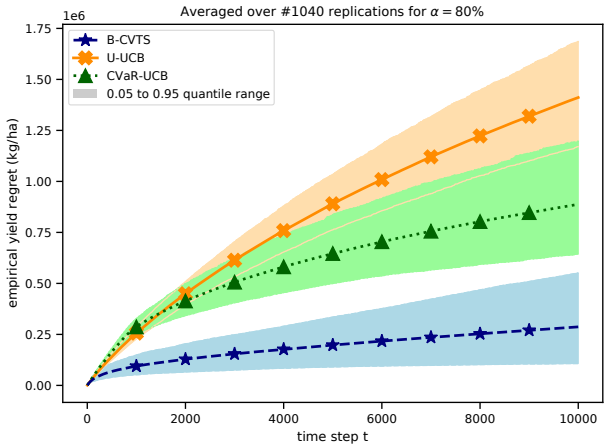
α	U-UCB	CVaR-UCB	B-CVTS
5%	3128 (3)	760 (14)	**192 (11)**
20%	4867 (11)	1024 (17)	**202 (10)**
80%	1411 (13)	888 (13)	**287 (12)**

Further experiments are reported in Appendix G. In particular we increase the number of arms, and empirically study the effect of over-estimating the support upper-bound: our results show that a "prudent" bound has little effect of the performance of the algorithms in the settings we consider. This property is of particular interest for the practitioner,

as a proper tuning of the support upper bound is the main limitation of the use of B-CVTS (and all bandit algorithms available for this problem). In most applications grounded on physical reality, the availability of such prudent upper-bound estimate is likely, and sufficient to ensure the practical performance of the B-CVTS algorithm.



$\alpha = 5\%$



$\alpha = 80\%$

Figure 2: Regret comparison in DSSAT environment, averaged over 1040 experiment replications.

Perspectives This first set of experiments using a challenging realistic crop simulator is promising, and motivates to further investigate the use of B-CVTS algorithm for crop-management support and other problems that can be modeled as CVaR bandits. B-CVTS enjoys appealing theoretical guarantees, and thanks to its simplicity and competitive empirical performances may be a good candidate for practitioners. In order to address real-world crop-management challenges, many questions remain to be considered, e.g. how to optimally generate mini-batches of recommendations to an ensemble of farmers in a semi-sequential procedure (in order to account for the long feedback time), how to incorporate distribution priors on crop-management options that could be pre-learnt *in silico* and refining them adaptively in the real world (thus, minimizing random exploration in the real world), how to include contextual information such as soil characteristics and local weather forecasts, or how handle non-stationarity, incorporating climate change progressive impact on an optimal planting date. Furthermore, the simplicity of the Non-Parametric Thompson Sampling algorithms make them appealing for generalization to other risk-aware settings, e.g risk-constrained (maximizing the mean under a condition on the CVaR) or with other risk metrics (mean-variance, entropic risk, etc). All of these open questions make interesting challenges for future works.

Acknowledgements

This work has been supported by the French Ministry of Higher Education and Research, Hauts-de-France region, Inria within the team-project Scool and the MEL. The authors acknowledge the funding of the French National Research Agency under projects BADASS (ANR-16-CE40-0002) and BOLD (ANR-19-CE23-0026-04) and the I-Site ULNE regarding project R-PILOTE-19-004-APPRENF.

The PhD of Dorian Baudry is funded by a CNRS80 grant. The PhD of Romain Gautron is co-funded by the French Agricultural Research Centre for International Development (CIRAD) and the Consortium of International Agricultural Research Centers (CGIAR) Big Data Platfrom.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

References

C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26:1487–1503, 2002.

S. Agrawal and N. Goyal. Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*, 2013.

S. Agrawal, W. M. Koolen, and S. Juneja. Optimal best-arm identification methods for tail-risk measures. *arXiv preprint arXiv:2008.07606*, 2020.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47, 2002.

C. Berge. *Topological Spaces: including a treatment of*

multi-valued functions, vector spaces, and convexity. 1997.

D. Brown. Large deviations bounds for estimating conditional value-at-risk. *Oper. Res. Lett.*, 35:722–730, 2007.

A. Burnetas and M. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2), 1996.

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*, 41, 2013.

A. Carpentier and M. Valko. Extreme bandits. In *Neural Information Processing Systems*, Montréal, Canada, Dec. 2014.

A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the 31st Annual Conference On Learning Theory*, 2018.

K. J. Evans, A. Terhorst, and B. H. Kang. From data to decisions: helping crop producers build their actionable knowledge. *Critical reviews in plant sciences*, 36(2): 71–88, 2017.

L. Evans and R. Fischer. Yield potential: its definition, measurement, and significance. *Crop science*, 39(6):1544–1551, 1999.

N. Galichet. *Contributions to multi-armed bandits: Risk-awareness and sub-sampling for linear contextual bandits*. PhD thesis, 2015.

N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, 2013.

A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Math. Oper. Res.*, 44:377–399, 2019.

Z. Hochman and P. Carberry. Emerging consensus on desirable characteristics of tools to support farmers' management of climate risk in australia. *Agricultural Systems*, 104(6):441–450, 2011.

M. J. Holland and E. M. Haress. Learning with cvar-based feedback under potentially heavy tails, 2020.

J. Honda and A. Takemura. An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

J. Honda and A. Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *Journal of Machine Learning Research*, 16:3721–3756, 2015.

G. Hoogenboom, C. Porter, K. Boote, V. Shelia, P. Wilkens, U. Singh, J. White, S. Asseng, J. Lizaso, L. Moreno, et al. The dssat crop modeling ecosystem. *Advances in crop modelling for a sustainable agriculture*, pages 173–216, 2019.

A. Kagrecha, J. Nair, and K. P. Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, 2019.

A. Kagrecha, J. Nair, and K. P. Jagannathan. Constrained regret minimization for multi-criterion multi-armed bandits. 2020.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012*, 2012.

N. Korda, E. Kaufmann, and R. Munos. Thompson Sampling for 1-dimensional Exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

T. Lattimore. A scale free algorithm for stochastic bandits with bounded kurtosis. 2017.

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

O. Maillard. Robust risk-averse stochastic multi-armed bandits. In *Algorithmic Learning Theory - 24th International Conference, ALT*, 2013.

B. B. Mandelbrot. The variation of certain speculative prices. In *Fractals and scaling in finance*. Springer, 1997.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Annals of Probability*, 18, 1990.

R. L. McCown. Changing systems for supporting farmers' decisions: problems, paradigms, and prospects. *Agricultural systems*, 74(1):179–220, 2002.

Q. Paris. The return of von liebig's "law of the minimum". *Agronomy Journal*, 84(6):1040–1046, 1992.

L. A. Prashanth, P. Krishna, Jagannathan, and R. K. Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, 2020.

A. Prashanth L, K. Jagannathan, and R. K. Kolla. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. *International Conference on Machine Learning*, 2019.

C. W. Richardson and D. A. Wright. Wgen: A model for generating daily weather variables. *ARS (USA)*, 1984.

C. Riou and J. Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In *Algorithmic Learning Theory - 31st International Conference, ALT 2020*, 2020.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

D. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning*, 11:1–96, 2018.

A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2012.

B. Shiferaw, B. M. Prasanna, J. Hellin, and M. Bänziger. Crops that feed the world 6. past successes and future challenges to the role played by maize in global food security. *Food security*, 3(3):307–327, 2011.

B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *International Conference on Machine Learning*, 2015.

A. Tamkin, R. Keramati, C. Dann, and E. Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM 2019*, 2020.

P. Thomas and E. Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, 2019.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.

M. Tollenaar and E. Lee. Yield potential, yield stability and stress tolerance in maize. *Field crops research*, 75(2-3): 161–169, 2002.

S. Vakili and Q. Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2015.

S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE J. Sel. Top. Signal Process.*, 10:1093–1111, 2016.

Q. Zhu and V. Tan. Thompson sampling algorithms for mean-variance bandits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A. Zimin, R. Ibsen-Jensen, and K. Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *CoRR*, abs/1405.0833, 2014.

A. Notations for the proofs

In this section, we introduce for convenience several notations that are used in the regret analysis.

A.1. General Notations

Model for multinomial arms In section B we denote by $(\mathcal{X}_1, \dots, \mathcal{X}_K)$ the supports of $(\nu_1, \dots, \nu_K)$, where $\mathcal{X}_k = (x_1^k, \dots, x_{M_k}^k)$ for some integer M_k. We also denote $\mathcal{P}^M = \{p \in \mathbb{R}^M : \forall i, p_i \geq 0, \sum_{j=1}^M p_j = 1\}$ the probability simplex in $\mathbb{R}^{M+1}$. Any multinomial distribution ν is characterized by its support and its probability vector p_ν: $\nu = (\mathcal{X}, p_\nu)$ with $p_\nu \in \mathcal{P}^{|\mathcal{X}|}$. Each arm ν_k is associated with a probability vector $p_k \in \mathcal{P}^M$.

Model for continuous bounded distributions In section C we assume that each arm k is supported in $[0, B_k]$ for some known value $B_k > 0$. We could assume supports of the form $[a_k, b_k]$ where only an upper bound on each b_k is known without loss of generality in the result, but we use this formulation for the sake of simplicity. We consider the set of continuous distributions in $[0, B_k]$, that we write $\mathcal{C}^{B_k}$. We still denote $\mathcal{P}^M = \{p \in \mathbb{R}^M : \forall i, p_i \geq 0, \sum_{j=1}^M p_j = 1\}$ the probability simplex in $\mathbb{R}^M$. The distribution of an arm k is ν_k, and its CVaR at a level α is denoted c_k^α.

Notations for the CVaR In the proofs in section of section B and section C we will encounter three different CVaR formulations for which we propose convenient notations,

- $\text{CVaR}_\alpha(F)$ for the CVaR of a distribution with a specified cumulative distribution function F.
- $C_\alpha(\mathcal{X}, w)$ for the CVaR of a discrete random variable of support $\mathcal{X} \subset [0, B]$ associated with a probability vector $w \in \mathcal{P}^{|\mathcal{X}|}$. According to our previous notation, $C_\alpha(\mathcal{X}, w) = \text{CVaR}_\alpha(F_w)$, where for all $y \in \mathbb{R}$,

$$F(y) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} w_x \mathbb{1}(y \leq x) \ .$$

- $C_\alpha(\mathcal{X})$ the CVaR of the uniform distribution on the discret support $\mathcal{X}$, which shortens the notations for the CVaR of the empirical distributions: $C_\alpha(\mathcal{X}) = C_\alpha(\mathcal{X}, 1_N)$ where $1_N = \underbrace{\left(\frac{1}{N}, \dots, \frac{1}{N}\right)}_{N terms}$.

We also introduce $c_k^\alpha = \text{CVaR}_\alpha(F_k)$ the CVaR of each arm's distribution, further assuming without loss of generality that

$$c_1^\alpha = \underset{k \in \{1, \dots, K\}}{\text{argmax}} \ c_k^\alpha.$$

Algorithm Both M-CVTS and B-CVTS can be formulated as an index policy, that is

$$A_t = \underset{k \in [K]}{\text{argmax}} \ c_{k,t}^\alpha,$$

where $c_{k,t}^\alpha$ is the index used in round t. We recall that $N_k(t) = \sum_{s=1}^t \mathbb{1}(A_s = k)$ denotes the number of selections of arms k. We define the σ-field $\mathcal{F}_t = \{A_\tau, r_{A_\tau} \text{ for } \tau = \{1, \dots, t\}\}$. In particular, knowing $\mathcal{F}_t$ allows to know $N_k(t)$ and the history of all arms available up to time t (i.e. all observations drawn before and including t).

Distances We define for multinomial distributions with same support $\mathcal{X}$ the distance

$$d : \mathcal{P}^{|\mathcal{X}|} \times \mathcal{P}^{|\mathcal{X}|} \to \mathbb{R}^+ : (p, q) \to ||p - q||_\infty = \sup_{m \in \{1, \dots, |\mathcal{X}|\}} |p_m - q_m| \ .$$

We also introduce the notation $D_L(F, G)$ for the Levy distance between two distributions with cdf F, and G. Namely, if two distributions are supported in $[0, B]$ for some $B > 0$

$$D_L(F,G) = \inf\left\{\varepsilon > 0 : \forall x \in [0,B], F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\right\} .$$

Furthermore, we recall the result from (Riou and Honda, 2020) stating that for two distributions of cdf F and G,

$$D_L(F,G) \leq ||F - G||_\infty$$

A.2. Specific notations for multinomial arms

We introduce for multinomial arms $N_k^m(t)$ as the number of times an element x_m has been observed during these pulls, so that $N_k(t) = \sum_{m=1}^{M_k} N_k^m(t)$. The Dirichlet posterior distribution given the observation after t rounds is then $\mathrm{Dir}(\beta_{k,t})$ where $\beta_{k,t} = (1 + N_k^1(t-1), 1 + N_k^1(t-1), \ldots, 1 + N_k^{M_k}(t-1))$. With this notation, the index is

$$c_{k,t}^\alpha = C_\alpha\left(\mathcal{X}_k, w_{k,t}\right) \quad \text{where} \quad w_{k,t} \sim \mathrm{Dir}(\beta_{k,t}).$$

We denote by $\beta_{k,t}$ the parameter of the Dirichlet distribution sampled at round t for arm k, and by $p_{k,t} \in \mathcal{P}^{M_k}$ the mean of this Dirichlet distribution: $p_{k,t} = \frac{1}{N_k(t-1)+M_k}\beta_{k,t}$. Observe that this probability vector can also be viewed as a biased version of the empirical probability distribution of arm k.

To ease the notation, we denote by $\mathcal{X} = (x_1, \ldots, x_M)$ the support of arm 1, while the support of any sub-optimal arm k is denoted by $\mathcal{X}_k = (x_1^k, \ldots, x_{M_k}^k)$.

A.3. Notations for continuous arms

For continuous arms we simply write $\mathcal{X}_n^k$ the history of observations available after n pulls of arm$k$, $X_0^k, X_1^k, X_2^k, \ldots, X_n^k$ in the order they have been collected, including as first term $X_0^k = B_k$, the upper bound of the support of k considered by the strategy. When considering (optimal) arm 1 we omit the exponent k to simplify notations. The Dirichlet distributions that we consider in this case are always of the form $\mathrm{Dir}((1,\ldots,1))$, where the size of the mean vector depends on the number of observation. We recall that this distribution is the uniform distribution on a simplex of fixed size N, and its average is the vector $\left(\frac{1}{N}, \ldots, \frac{1}{N}\right)$. We denote this distribution $\mathcal{D}_N$ to simplify the notations.

B. Proof of Theorem 2 : analysis of M-CVTS with multinomial arms

Thanks to Equation (3) presented in Section 3, the proof of Theorem 2 can be obtained by proving Equation (4) and Equation (6) for multinomial arms distributions. This consists in upper bounding the *pre-convergence* (Pre-CV) and *post-convergence* (Post-CV) terms presented in section 3. In this section we use the L^∞ distance presented in Section A.

B.1. Proof of Equation (4) : Upper Bound on the Post-Convergence term

We upper bound the term (Post-CV) $= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(A_t = k, c_{k,t}^\alpha \geq c_1^\alpha - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right]$, where $p_{k,t} = \frac{\beta_{k,t}}{N_k(t-1)}$ is the probability vector associated with the empirical distribution of an arm k, i.e the frequency of each item of the support $\mathcal{X}_k$ in the history of arm k. We recall that this former quantity is biased because of the initialization step which set $\beta_{k,0} = (1,\ldots,1)$ and $N_k^i(0) = 1$ (introducing the fictitious time $t = 0$ just before the algorithms starts). For any constant $n_0(T)$ we have

$$\begin{aligned}
\text{(Post-CV)} \leq & \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(A_t = k, c_{k,t}^\alpha \geq c_1^\alpha - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right] \\
\leq & \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}(A_t = k, N_k(t-1) \leq n_0(T), c_{k,t}^\alpha \geq c_1^\alpha - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right] \\
& + \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}(A_t = k, N_k(t-1) \geq n_0(T), c_{k,t}^\alpha \geq c_1^\alpha - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right] .
\end{aligned}$$

The first term is upper bounded by $n_0(T)$ as the event $(A_t = k, N_k(t-1) \leq n_0(T))$ can occur at most $n_0(T)$ times. So we have

$$\begin{aligned}
\text{(Post-CV)} \leq & n_0(T) + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(A_t = k, N_k(t-1) \geq n_0(T), c_{k,t}^{\alpha} \geq c_1^{\alpha} - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right] \\
\leq & n_0(T) + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(N_k(t-1) \geq n_0(T), c_{k,t}^{\alpha} \geq c_1^{\alpha} - \varepsilon_1, d(p_{k,t}, p_k) \leq \varepsilon_2)\right] \\
\leq & n_0(T) + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(N_k(t-1) \geq n_0(T), d(p_{k,t}, p_k) \leq \varepsilon_2\right) \times \mathbb{E}\left[\mathbb{1}\left(c_{k,t}^{\alpha} \geq c_1^{\alpha} - \varepsilon_1\right) | \mathcal{F}_{t-1}\right]\right] \\
= & n_0(T) + \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(N_k(t-1) \geq n_0(T), d(p_{k,t}, p_k) \leq \varepsilon_2\right) \times \mathbb{P}_{w \sim \mathrm{Dir}(\beta_{k,t})}\left(C_{\alpha}(\mathcal{X}_k, w) \geq c_1^{\alpha} - \varepsilon_1\right)\right] \ ,
\end{aligned}$$

where we upper bounded $\mathbb{1}(A_t = k)$ by 1, so that $\mathbb{1}\left(c_{k,t}^{\alpha} \geq c_1^{\alpha} - \varepsilon_1\right)$ is the only term that is not $\mathcal{F}_{t-1}$-measurable, and then used the law of total expectation. Now we can use Lemma 1 to control the probability term inside the expectation by

$$\mathbb{P}_{w \sim \mathrm{Dir}(\beta_k(t))}\left(C_{\alpha}(\mathcal{X}_k, w) \geq c_1^{\alpha} - \varepsilon_1\right) \leq C_1 M_k (N_k(t-1) + M_k)^{\frac{M_k}{2}} \exp\left(-(N_k(t-1) + M_k)\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_{k,t}, c_1^{\alpha} - \varepsilon_1)\right).$$

At this stage, our objective is to remove the randomness in this upper bound in order to bound uniformly the terms inside the expectation. To do this, we will use the fact that $d(p_{k,t}, p_k) \leq \varepsilon_2$ and $N_k(t-1) \geq n_0(T)$ together with the continuity of the function $\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}$ in its second argument, established in Lemma 6 defined in Appendix D.3.

Let $\varepsilon_3 \in (0, \mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_{k,t}, c_1^{\alpha}))$. There exists by continuity small enough values of ε_1 and ε_2 such that, if $d(p_{k,t}, p_k) \leq \varepsilon_2$,

$$\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_{k,t}, c_1^{\alpha} - \varepsilon_1) \geq \mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha} - \varepsilon_1) - \frac{\varepsilon_3}{4} \geq \mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \frac{\varepsilon_3}{2},$$

hence

$$\begin{aligned}
D := & (N_k(t-1) + M_k)^{\frac{M_k}{2}} \exp\left(-(N_k(t-1) + M_k)\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_{k,t}, c_1^{\alpha} - \varepsilon_1)\right) \\
\leq & (N_k(t-1) + M_k)^{\frac{M_k}{2}} \exp\left(-(N_k(t-1) + M_k)(\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3/2)\right).
\end{aligned}$$

Using the fact that for any $b > 0$ and $b > \varepsilon > 0$ there exists a constant C' such that $\forall t > 0$: $t \exp(-bt) \leq C' \exp(-t(b - \varepsilon))$, we further get

$$\begin{aligned}
D \leq & C' \exp\left(-(N_k(t-1) + M_k)(\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3)\right) \\
\leq & C' \exp\left(-(n_0(T) + M_k)(\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3)\right),
\end{aligned}$$

provided that $N_k(t-1) \geq n_0(T)$.

Putting things together, we proved that for every $\varepsilon_3 \in (0, \mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_{k,t}, c_1^{\alpha}))$, if ε_1 and ε_2 are small enough, then there exists a constant $C_1' > 0$ such that

$$\begin{aligned}
\text{(Post-CV)} \leq & n_0(T) + \sum_{t=1}^{T} C_1' \exp\left(-(n_0(T) + M_k)(\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3)\right) \\
\leq & n_0(T) + T C_1' \exp\left(-(n_0(T) + M_k)(\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3)\right).
\end{aligned}$$

Choosing $n_0(T) = \frac{\log T}{\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(\nu_k, c_1^{\alpha} - \varepsilon_1) - \varepsilon_3} - (M_k + 1)$ yields the upper bound

$$\text{(Post-CV)} \leq \frac{\log T}{\mathcal{K}_{\inf}^{\alpha, \mathcal{X}_k}(p_k, c_1^{\alpha}) - \varepsilon_3} + O(1).$$

Finally, we have shown that for any $\varepsilon_0 > 0$, if ε_1 and ε_2 are small enough, then

$$(\text{Post-CV}) \leq \frac{(1+\varepsilon_0)\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{X}_k}(p_k, c_1^\alpha)} + O(1),$$

which proves Equation (4).

B.2. Proof of Equation (6): Upper Bound on the Pre-Convergence term for multinomial distributions

In this section, we upper bound the term $(\text{Pre-CV}) = \mathbb{E}\left(\sum_{t=1}^T \mathbb{1}(A_t = k, \{c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1 \cup d(p_{k,t}, p_k) > \varepsilon_2\})\right)$.

We first decompose this term into

$$(\text{Pre-CV}) \leq \mathbb{E}\left(\sum_{t=1}^T \mathbb{1}(A_t = k, c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1)\right) + \mathbb{E}\left(\sum_{t=1}^T \mathbb{1}(A_t = k, d(p_{k,t}, p_k) > \varepsilon_2)\right).$$

Let us remark that the second term does not feature any CVaR, hence we can directly use the upper bound derived by (Riou and Honda, 2020) to get that, for any $\varepsilon_2 > 0$,

$$\mathbb{E}\left(\sum_{t=1}^T \mathbb{1}(A_t = k, d(p_{k,t}, p_k) > \varepsilon_2)\right) \leq KM\left(\frac{2M}{\varepsilon_2} + \frac{2}{\varepsilon_2^2}\right).$$

Hence, it remains to upper bound the term

$$A := \mathbb{E}\left(\sum_{t=1}^T \mathbb{1}(A_t = k, c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1)\right)$$

We write

$$\begin{aligned}
A \leq & \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(c_{A_t,t}^\alpha < c_1^\alpha - \varepsilon_1)\right] \\
\leq & \sum_{t=1}^T \sum_{n=1}^T \mathbb{E}\left[\mathbb{1}(c_{A_t,t}^\alpha < c_1^\alpha - \varepsilon_1, N_1(t) = n)\right] \\
\leq & \sum_{m=1}^T \sum_{n=1}^T \mathbb{E}\left[\mathbb{1}\left(\sum_{t=1}^T \mathbb{1}(c_{A_t,t}^\alpha < c_1^\alpha - \varepsilon_1, N_1(t) = n) \geq m\right)\right],
\end{aligned}$$

where we used as in (Riou and Honda, 2020) that for any series of events (E_t) it holds that

$$\sum_{t=1}^T \mathbb{1}(E_t) \leq \sum_{m=1}^T \mathbb{1}\left(\sum_{t=1}^T \mathbb{1}(E_t) \geq m\right).$$

We then introduce a random sequence $(\tau_i^n)_{i\in\mathbb{N}}$ where $\tau_i^n \in \mathbb{R} \cup \{+\infty\}$ is the i-th time at which the event $\{\max_{j>1} c_{j,t}^\alpha \leq c_1^\alpha - \varepsilon_1, N_1(t) = n\}$ holds. In order to ensure that this event occurs at least m times, then we need 1) that $\tau_i^n \leq +\infty$ for all $i \leq m$, and 2) $c_{1,\tau_i^n}^\alpha \leq c_1^\alpha - \varepsilon_1$ for all $i \leq m$, otherwise arm 1 would be drawn. Hence, we have the following inclusion of events

$$\left\{\sum_{t=1}^T \mathbb{1}(c_{A_t,t}^\alpha < c_1^\alpha - \varepsilon_1, N_1(t) = n) \geq m\right\} \subset \left\{\tau_i^n < +\infty, c_{1,\tau_i^n}^\alpha \leq c_1^\alpha - \varepsilon_1 \, \forall i \in \{1, \dots, m\}\right\}.$$

We then use the following arguments, for a fixed n: 1) since arm 1 has been drawn n times at all (finite) time steps τ_i^n, the random variables β_{1,τ_i^n} for i such that $\tau_i^n < \infty$ are all equal to some common value β_n, which is such that $\beta_n - 1$ follows a multinomial distribution $\mathrm{Mult}(n, p_1)$. 2) the $c_{1,\tau_i^n}^\alpha$ are independent conditionally to β_n and follow a $\mathrm{Dir}(\beta_n)$ distribution.

Therefore, we write

$$\begin{aligned}
A &\leq \sum_{n=1}^{T}\sum_{m=1}^{T}\mathbb{E}\left[\prod_{i=1}^{m}\mathbb{1}\left(\tau_i^n < +\infty, c^{\alpha}_{1,\tau_i^n} \leq c_1^{\alpha} - \varepsilon_1\right)\right] \\
&\leq \sum_{n=1}^{T}\sum_{m=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\prod_{i=1}^{m}\mathbb{P}_{w_i\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w_i)\leq c_1^{\alpha}-\varepsilon_1\right)\right] \\
&\leq \sum_{n=1}^{T}\sum_{m=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)^m\right] \\
&\leq \sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\sum_{m=1}^{T}\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)^m\right] \\
&\leq \sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\frac{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)}{1-\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)}\right].
\end{aligned}$$

Thank to this bound, we have transformed our problem into the study of properties of the Dirichlet distribution. Similarly, we may now upper bound the last term in the above inequality by considering different regions to which the mean of the Dirichlet distribution – that is, $\frac{\beta}{n+M}$ – belongs. However, in order to account for a general risk level α, the analysis is more intricate as we need to split the simplex into sub-spaces defined by different values of the CVaR, not the mean. This requires to establish new boundary crossing probabilities involving those sub-spaces, which we provide now.

We decompose the upper-bound on A into $\quad A \leq A_1 + A_2 + A_3, \quad$ where:

- $A_1 = \sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\frac{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)}\mathbb{1}\left(C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)\geq c_1^{\alpha}-\varepsilon_1/2\right)\right]$
- $A_2 = \sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\frac{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)}\mathbb{1}\left(c_1^{\alpha}-\varepsilon_1\leq C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)\leq c_1^{\alpha}-\varepsilon_1/2\right)\right]$
- $A_3 = \sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left[\frac{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)}\mathbb{1}\left(C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)\leq c_1^{\alpha}-\varepsilon_1\right)\right].$

We now upper bound each of these three terms by a constant, for any value of ε_1.

B.2.1. Upper bound on A_1

This term is the easiest to control. Indeed the set $\{p\in\mathcal{P}^M : C_\alpha(\mathcal{X},p)\leq c_1^{\alpha}-\varepsilon_1\}$ is closed and convex, hence we can apply the boundary crossing probability of Lemma 13 in (Riou and Honda, 2020) on this subset and write

$$\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)\leq C_1(n+M)^{\frac{M}{2}}\exp\left(-(n+M)\mathrm{KL}\left(\frac{\beta}{n+M},p^*_\beta\right)\right),$$

with $p^*_\beta = \underset{p:C_\alpha(\mathcal{X},p)\leq c_1^{\alpha}-\varepsilon_1}{\mathrm{argmin}}\ \mathrm{KL}(\frac{\beta}{n+M},p)$. Now the quantity

$$\delta = \inf_{\substack{q:C_\alpha(\mathcal{X},q)\geq c_1^{\alpha}-\varepsilon_1/2\\ p:C_\alpha(\mathcal{X},p)\leq c_1^{\alpha}-\varepsilon_1}}\mathrm{KL}(q,p)$$

satisfies $\delta>0$ due to the following argument: the infimum of the continuous function $(q,p)\mapsto\mathrm{KL}(q,p)$ on a compact set is necessarily achieved in a point (q^*,p^*). Assuming $\delta=0$ yields $q^*=p^*$ while $C_\alpha(\mathcal{X},q^*)\geq c_1^{\alpha}-\varepsilon_1/2$ and $C_\alpha(\mathcal{X},p^*)\leq c_1^{\alpha}-\varepsilon_1$ which is not possible as $\varepsilon_1>0$. Hence, if the event $\left\{C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)\geq c_1^{\alpha}-\varepsilon_1/2\right\}$ holds, one has

$$\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}\left(C_\alpha(\mathcal{X},w)\leq c_1^{\alpha}-\varepsilon_1\right)\leq C_1(n+M)^{\frac{M}{2}}\exp\left(-(n+M)\delta\right).$$

For $n \geq n_1$ large enough so that $C_1(n+M)^{M/2}\exp(-(n+M)\delta) < 1$ it follows that

$$A_1 \leq n_1 + \sum_{n=n_1+1}^{T} \mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(\mathbb{1}(C_\alpha(\mathcal{X},q) \geq c_1 - \varepsilon_1/2)\frac{C_1(n+M)^{\frac{M}{2}}\exp(-(n+M)\delta)}{1-C_1(n+M)^{\frac{M}{2}}\exp(-(n+M)\delta)}\right).$$

Furthermore, for any $\gamma > 1$, there exist some n_γ satisfying $\forall n \geq n_\gamma$, $\frac{1}{1-C_1(n+M)^{\frac{M}{2}}\exp(-(n+M)\delta)} \leq \gamma$, therefore:

$$\begin{aligned}
A_1 &\leq \max(n_1, n_\gamma) + \sum_{n=\max(n_1,n_\gamma)+1}^{T} \mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p_1)}\left(\gamma\mathbb{1}(C_\alpha(\mathcal{X},q) \geq c_1 - \varepsilon_1/2)C_1(n+M)^{\frac{M}{2}}\exp(-(n+M)\delta)\right)\\
&\leq \max(n_1, n_\gamma) + \sum_{n=\max(n_1,n_\gamma)+1}^{T} \gamma C_1(n+M)^{\frac{M}{2}}\exp(-(n+M)\delta),
\end{aligned}$$

and the right-hand side can be upper bounded by a constant.

B.2.2. Upper bound on A_2

For the term A_2 we ignore the numerator and hence study

$$A_2 \leq \sum_{n=1}^{T} \mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) \leq c_1^\alpha - \varepsilon_1/2\right)\frac{1}{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w) \geq c_1^\alpha - \varepsilon_1)}\right).$$

Using Lemma 2 we obtain

$$\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w) \geq c_1^\alpha - \varepsilon_1) \geq \frac{C_2}{(n+M)^{\frac{3M}{2}+1}}\exp\left(-(n+M)\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right)\right).$$

Now we note that, by definition, $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right) = 0$ if $c_1^\alpha - \varepsilon_1 \leq C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right)$, so the exponential term is equal to 1 in that case, leading to

$$\begin{aligned}
A_2 &\leq \sum_{n=1}^{T} \mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) \leq c_1^\alpha - \varepsilon_1/2\right)C_2^{-1}(n+M)^{3M/2+1}\right]\\
&\leq \sum_{n=1}^{T} C_2^{-1}(n+M)^{3M/2+1}\mathbb{P}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) \leq c_1^\alpha - \varepsilon_1/2\right).
\end{aligned}$$

To conclude, we make use of a concentration inequality for the empirical CVaR derived from Brown's inequality (Brown, 2007). However, to express the probability in the right-hand side in terms of the CVaR of an empirical distribution, we need to handle the bias of β induced by the initialization step. For this we use the fact that for any integers n_k, M:

$$\left|\frac{n_k}{n} - \frac{n_k+1}{n+M}\right| = \frac{1}{n+M}\left|\frac{n_k}{n}M - 1\right| \leq \frac{M}{n+M}.$$

As a direct consequence, when n is large enough the biased empirical distribution $\frac{\beta}{n+M}$ can be made as close to the empirical distribution $\frac{\beta-1}{n}$ as we want. Thanks to Lemma 7 we indeed get

$$\left|C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) - C_\alpha\left(\mathcal{X}, \frac{\beta-1}{n}\right)\right| \leq \frac{M^2 x_M}{\alpha(n+M)}.$$

So for $n \geq n'$ large enough it holds that

$$\mathbb{P}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) \leq c_1^\alpha - \varepsilon_1/2\right) \leq \mathbb{P}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(C_\alpha\left(\mathcal{X}, \frac{\beta-1}{n}\right) \leq c_1^\alpha - \varepsilon_1/3\right).$$

We are now ready to apply Lemma 9 (Brown inequality) in Appendix F, which yields

$$\mathbb{P}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(C_\alpha\left(\mathcal{X},\frac{\beta-1}{n}\right)\leq c_1^\alpha-\varepsilon_1/3\right)\leq\exp\left(-n\frac{2\alpha^2\,\varepsilon_1^2}{9}\right).$$

This entails that we can upper bound A_2 by a constant:

$$A_2\leq\sum_{n=1}^{T}C_2^{-1}(n+M)^{3M/2+1}\exp\left(-n\frac{2\alpha^2\,\varepsilon_1^2}{9}\right).$$

We remark that the assumption considered by Brown's inequality requires the random variables to be both *positive* and *bounded*. For instance, Prashanth L et al. (2019)[Theorem 3.3] proved that a similar result (i.e an exponential inequality with a scaling in $n\alpha^2\,\varepsilon^2$) hold for random variables that are non necessarily positive; their concentration bound is a little more complicated. Hence, we work with Brown's inequality in this proof for the sake of simplicity but our algorithm does not actually require the variables to be positive. Furthermore, note that Brown's inequality is not used here in order to control the first order terms of our regret bound, hence we do not focus on obtaining the tightest concentration bounds for the concentration of the empirical CVaR, as it only affects second order terms.

B.2.3. Upper bound on A_3

Similarly to what we presented in the previous section we write

$$A_3\leq\sum_{n=1}^{T}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left(\mathbb{1}\left(C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)<c_1^\alpha-\varepsilon_1\right)\frac{1}{\mathbb{P}_{w\sim\mathrm{Dir}(\beta)}(C_\alpha(\mathcal{X},w)\geq c_1^\alpha-\varepsilon_1)}\right)$$

and use Lemma 2 in order to lower bound the probability in the denominator:

$$\begin{aligned}A_3\leq\sum_{n=1}^{T}&C_2^{-1}(n+M)^{M/2+M}\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left[\mathbb{1}\left(C_\alpha\left(\mathcal{X},\frac{\beta}{n+M}\right)<c_1^\alpha-\varepsilon_1\right)\right.\\&\left.\times\exp\left((n+M)\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M},c_1^\alpha-\varepsilon_1\right)\right)\right].\end{aligned}$$

We now have to compute explicitly the term in expectation, using the fact that $\beta-1$ follows a multinomial distribution $\mathrm{Mult}(n,p)$. We upper bound this probability as follows

$$\begin{aligned}\mathbb{P}_{X\sim\mathrm{Mult}(n,p)}(X=\beta-1)&=\frac{n!}{(\beta_0-1)!\ldots(\beta_M-1)!}\prod_{i=1}^{M}p_i^{\beta_i-1}\\&=\frac{n!}{(\beta_0-1)!\ldots(\beta_M-1)!}\exp\left(\sum_{i=0}^{M}(\beta_i-1)\log(p_i)\right)\\&=\frac{n!}{(\beta_0-1)!\ldots(\beta_M-1)!}\exp\left(\sum_{i=1}^{M}(\beta_i-1)\log\left(\frac{\beta_i-1}{n}\right)-\sum_{i=1}^{M}(\beta_i-1)\log\left(\frac{\beta_i-1}{np_i}\right)\right)\\&=\frac{n!}{(\beta_0-1)!\ldots(\beta_M-1)!}\exp\left(\sum_{i=1}^{M}(\beta_i-1)\log\left(\frac{\beta_i-1}{n}\right)-n\mathrm{KL}\left(\frac{\beta-1}{n},p\right)\right)\\&=\mathbb{P}_{Y\sim\mathrm{Mult}\left(\frac{\beta-1}{n},p\right)}(X=\beta-1)\exp\left(-n\mathrm{KL}\left(\frac{\beta-1}{n},p\right)\right)\\&\leq\exp\left(-n\mathrm{KL}\left(\frac{\beta-1}{n},p\right)\right).\end{aligned}$$

Using again Lemma 7 and the continuity of the KL divergence, for any $\varepsilon'>0$ we can find $n_{\varepsilon'}$ large enough such that

$$n\mathrm{KL}\left(\frac{\beta-1}{n},p\right)\geq(n+M)\left(\mathrm{KL}\left(\frac{\beta}{n+M},p\right)-\varepsilon'\right).$$

Letting

$$\mathcal{C} = \left\{ \beta \in \{1, ..., n+1\}^M : C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) < c_1^\alpha - \varepsilon_1 \right\},$$

and combining the previous results one derives

$$\begin{aligned}
&\mathbb{E}_{\beta-1\sim\mathrm{Mult}(n,p)}\left[\mathbb{1}\left(C_\alpha\left(\mathcal{X}, \frac{\beta}{n+M}\right) < c_1^\alpha - \varepsilon_1\right)\exp\left((n+M)\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right)\right)\right] \\
&\leq \sum_{\beta\in\mathcal{C}} \mathbb{P}_{X-1\sim\mathrm{Mult}(n,p)}(X=\beta)\exp\left((n+M)\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right)\right) \\
&= \sum_{\beta\in\mathcal{C}} \exp\left(-(n+M)\left(\mathrm{KL}\left(\frac{\beta}{n+M}, p_1\right) - \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right) - \varepsilon'\right)\right) \\
&\leq \sum_{\beta\in\mathcal{C}} \exp\left(-(n+M)\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha\right) - \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n+M}, c_1^\alpha - \varepsilon_1\right) - \varepsilon'\right)\right) \\
&\leq \sum_{\beta\in\mathcal{C}} \exp\left(-(n+M)(\delta_1 - \varepsilon')\right),
\end{aligned}$$

where we introduced the quantity

$$\delta_1 = \inf_{p:\mathrm{CVaR}_\alpha^{\mathcal{X}}(p)\leq c_1^\alpha - \varepsilon_1} \left[\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p, c_1^\alpha) - \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p, c_1^\alpha - \varepsilon_1)\right].$$

Since $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}$ is continuous in its first argument and we take the infimum over a compact set, this infimum is reached for some distribution $p_{\inf}^\alpha$. That is, we can write

$$\delta_1 = \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p_{\inf}^\alpha, c_1^\alpha) - \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p_{\inf}^\alpha, c_1^\alpha - \varepsilon_1),$$

and $\mathrm{CVaR}_\alpha^{\mathcal{X}}(p_{\inf}^\alpha) \leq c_1^\alpha - \varepsilon_1$. Thanks to Lemma 6, we know that the mapping $c \mapsto \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p_{\inf}^\alpha, c)$ is strictly increasing for $c \geq c_1^\alpha - \varepsilon_1$, thus $\delta_1 > 0$. Finally, choosing $\varepsilon' = \frac{\delta_1}{2}$ and using the fact that $|\mathcal{C}| = (n+1)^M$ yields

$$A_3 \leq n_{(\delta_1/2)} + \sum_{n=n_{(\delta_1/2)}}^{T} C_2^{-1}(n+M)^{\frac{3M}{2}}(n+1)^M \exp\left(-(n+M)\frac{\delta_1}{2}\right).$$

This shows that A3 is upper-bounded by a constant.

C. Proof of Theorem 3 : analysis of B-CVTS for continuous bounded distributions

Similarly to the proof techniques used to analyze M-CVTS, we use Equation (3) presented in Section 3 in the proof of Theorem 3.

In particular, we first prove Equation (5) (Post-CV term) before proving Equation (6) (Pre-CV term), assuming that the arms are continuous, bounded, and that an upper bound on their support is known. In this section we use the Levy distance presented in Appendix A in order to compare the empirical cdf $F_{k,t}$ with F_k for each arm k.

C.1. Proof of Equation (5): Upper Bound on the Post-Convergence term

We upper bound the term (Post-CV) $= \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}(A_t = k, c_{k,t}^\alpha \geq c_1^\alpha - \varepsilon_1, D_L(F_{k,t}, F_k) \leq \varepsilon_2)\right]$. The change from using the L^∞ distance to using the Levy metric does not affect any argument in the beginning of the proof used to upper bound (Post-CV). Hence, following the same steps as in section B.1, for any $n_0(T)$ it holds that

$$\text{(Post-CV)} \leq n_0(T) + \sum_{t=1}^T \mathbb{E}\left[\mathbb{1}\left(N_k(t-1) \geq n_0(T), D_L(F_{k,t}, F_k) \leq \varepsilon_2\right) \times \mathbb{P}_{w\sim\mathcal{D}_{\mathrm{N_k(t)}}}\left(C_\alpha(\mathcal{X}_{k,t}, w) \geq c_1^\alpha - \varepsilon_1\right)\right].$$

We then use Lemma 3 in order to control the probability term inside the expectation as follows

$$\mathbb{P}_{w\sim\mathcal{D}_{\mathrm{N_k(t)}}}\left(C_\alpha(\mathcal{X}_{k,t}, w) \geq c_1^\alpha - \varepsilon_1\right) \leq \frac{1}{\eta}\exp\left(-N_k(t-1)\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1^\alpha - \varepsilon_1) - \eta C(\alpha, B_k, c_1^\alpha - \varepsilon_1)\right)\right),$$

where $\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1^\alpha - \varepsilon_1)$ is the functional defined in Section 3, applied on the set of continuous bounded distributions defined on $[0, B_k]$, $\mathcal{B}_k$.

We then proceed by removing the randomness in this upper bound by bounding uniformly the terms inside the expectation, using that $D_L(F_{k,t}, F_k) \leq \varepsilon_2$ and $N_k(t-1) \geq n_0(T)$. To do so, we now use the continuity of the mapping $\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}$, which is proved using Lemma 4 from (Agrawal et al., 2020). Note that this is not a trivial result, as for instance the Levy topology does not coincide with the one induced by the Kullback-Leibler divergence. Combining these elements it holds that for any ε_0, there exist $\eta > 0$ such that

$$\begin{aligned}\text{(Post-CV)} \leq & n_0(T) + \sum_{t=1}^{T} \frac{1}{\eta} \exp\left(-n_0(T)\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0\right)\right) \\ \leq & n_0(T) + \frac{1}{\eta} T \exp\left(-n_0(T)\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1) - \varepsilon_0\right)\right) .\end{aligned}$$

Choosing $n_0(T) = \frac{\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1^\alpha) - \varepsilon_0}$ we upper bound the post-convergence term as

$$\text{(Post-CV)} \leq \frac{\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{B}_k}(F_k, c_1^\alpha) - \varepsilon_0} + O(1).$$

C.2. Pre-Convergence term

In this section, we now focus on providing an upper bound on the remainder term (Pre-CV) $= \mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}(A_t = k, \{c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1 \cup D_L(F_{k,t}, F_k) > \varepsilon_2\})\right)$.

We first decompose this term into

$$\text{(Pre-CV)} \leq \mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}(A_t = k, c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1)\right) + \mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}(A_t = k, D_L(F_{k,t}, F_k) > \varepsilon_2)\right) .$$

Again, as the second term does not feature any CVaR, we can again use a result from Riou and Honda (2020) (section D.1) to get that, for any $\varepsilon_2 > 0$,

$$\mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}(A_t = k, D_L(F_{k,t}, F_k) > \varepsilon_2)\right) \leq K(M+1)\left(1 + \sum_{n=2}^{+\infty} 2(n+1)\exp\left(-2(n-1)\left(\varepsilon_2 - \frac{1}{n-1}\right)\right)\right).$$

Hence, if an arm is pulled a lot, its empirical distribution will be with high probability into a Levy ball of size ε_2 around its true distribution. Hence, it remains to upper bound the term

$$\bar{A} := \mathbb{E}\left(\sum_{t=1}^{T} \mathbb{1}(A_t = k, c_{k,t}^\alpha < c_1^\alpha - \varepsilon_1)\right) .$$

We follow the exact same steps as in section B.2 and obtain again an expression of the form

$$\bar{A} \leq \sum_{n=1}^{T} \mathbb{E}_{X_1,\dots,X_n}\left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n, w) \leq c_1^\alpha - \varepsilon_1\right)}{1 - \mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n, w) \leq c_1^\alpha - \varepsilon_1\right)}\right] ,$$

where we write here $\mathcal{X}_n$ as the support of the empirical distribution of arm 1 after receiving n observations. Let us recall that $\mathcal{D}_n$ is the uniform probability on the simplex of size n, namely the Dirichlet distribution with parameter $1_n = (1, \dots, 1)$. Thanks to the fact that $c_{A_t,t} < c_1^\alpha - \varepsilon_1 \Rightarrow c_{1,t} < c_1^\alpha - \varepsilon$ we can upper bound this term by the probability that the best arm under-performs.

As in section B.2, we split this expectation into different regions depending of the value of the CVaR of the empirical distribution (that includes the term $x_0 = B$ added at the beginning of the history of observations).

We split the upper bound on $\bar{A}$ into three terms

$$\bar{A} \leq \bar{A}_1 + \bar{A}_2 + \bar{A}_3 \ ,$$

where

- $\bar{A}_1 = \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n} \left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1\right)} \mathbb{1}\left(C_\alpha(\mathcal{X}_n) \geq c_1^\alpha - \varepsilon_1/2\right) \right] \ ,$
- $\bar{A}_2 = \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n} \left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1\right)} \mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \right],$
- $\bar{A}_3 = \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n} \left[\frac{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1)}{1-\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n,w)\leq c_1^\alpha-\varepsilon_1\right)} \mathbb{1}\left(C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1\right) \right].$

We now upper bound each of these three terms, for any value of ε_1.

C.2.1. Upper bound on $\bar{A}_1$

The first case is again easier than the two others, because in this case the CVaR of arm 1 is greater than $c_1^\alpha - \varepsilon_1/2$ and so we can upper bound the term $\bar{A}_1$ by upper bounding

$$\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n,w) \leq c_1^\alpha - \varepsilon_1\right) \ .$$

This term should be small as the empirical CVaR does not belong to the sub-space defined by this inequality. Furthermore, we can use a quantization of the random observations $X_1, \dots, X_n$, defining a number of bins M that we will specify later, and for any $i \in \{0, \dots, n\}$ the random variables $\widetilde{X}_i = \frac{\lfloor M X_i \rfloor}{M}$. Denoting the corresponding set of truncated observations $\widetilde{\mathcal{X}}_n$, since for all $i \in \{0, \dots, n\}, X_i - 1/M \leq \tilde{X}_i \leq X_i$ we then obtain that

$$C_\alpha(\mathcal{X}_n) - \frac{1}{M} \leq C_\alpha(\widetilde{\mathcal{X}}_n) \leq C_\alpha(\mathcal{X}_n) \ .$$

A similar control holds for $C_\alpha(\mathcal{X}_n, w)$ and $C_\alpha(\widetilde{\mathcal{X}}_n, w)$. Interestingly, these properties directly follow from the monotonicity of the CVaR, which is itself a property of any coherent risk measure (see Acerbi and Tasche (2002)).

Using these properties, we first have that

$$\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X}_n,w) \leq c_1^\alpha - \varepsilon_1\right) \leq \mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\widetilde{\mathcal{X}}_n,w) \leq c_1^\alpha - \varepsilon_1\right) \ ,$$

as well as

$$\mathbb{1}\left(C_\alpha(\mathcal{X}_n) \geq c_1^\alpha - \varepsilon_1/2\right) \leq \mathbb{1}\left(C_\alpha(\widetilde{\mathcal{X}}_n) \geq c_1^\alpha - \varepsilon_1/2 - \frac{1}{M}\right) \ .$$

Therefore, we have shown that we can upper bound the first term $\bar{A}_1$ by

$$\bar{A}_1 \leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n} \left(\mathbb{1}\left(C_\alpha\left(\widetilde{\mathcal{X}}_n,w\right) \geq c_1^\alpha - \frac{\varepsilon_1}{2} - \frac{1}{M}\right) \frac{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha\left(\widetilde{\mathcal{X}}_n,w\right) \leq c_1^\alpha - \varepsilon_1)}{1-\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha\left(\widetilde{\mathcal{X}}_n,w\right) \leq c_1^\alpha - \varepsilon_1)} \right) \ .$$

We then choose the discretization step $\frac{1}{M}$. First, we want this step to be small enough in order to preserve the order of the CVaRs, this in turns can be done by choosing ε_1 small enough. Secondly, we want that $c_1^\alpha - \frac{\varepsilon_1}{2} - \frac{1}{M} > c_1^\alpha - \varepsilon_1$. This condition requires $M > 2/\varepsilon_1$, so we choose (for instance) $M = \lceil 3/\varepsilon_1 \rceil$.

We can now resort to Lemma 13 of Riou and Honda (2020) in order to upper bound the probability involving the Dirichlet distribution, namely

$$\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha\left(\widetilde{\mathcal{X}}_n, w\right) \leq c_1^\alpha - \varepsilon_1) \leq C_1 n^{M/2} \exp\left(-n\mathcal{K}_{\inf}^{\widetilde{\mathcal{X}}_n}\left(\widetilde{F}_{k,n}, c_1^\alpha - \frac{\varepsilon_1}{2} - \frac{1}{M}\right)\right) ,$$

where $\widetilde{F}_{k,n}$ is the cdf of the empirical distribution corresponding to $\widetilde{\mathcal{X}}_n$. Since we could then transform our problem in order to consider multinomial random variables, we can use the same steps as in the corresponding part of the regret analysis of M-CVTS. Hence, following similar steps as in Appendix B.2.1, leads to the bound

$$\bar{A}_1 = O(1) .$$

C.2.2. Upper bound on $\bar{A}_2$

In order to control the term A_2 in Appendix B.2.2 we ignore the numerator and write

$$\bar{A}_2 \leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X}_n, w) \geq c_1^\alpha - \varepsilon_1)}\right] .$$

We then use Lemma 5 in order to upper bound the right hand term, which yields

$$\bar{A}_2 \leq \mathbb{E}_{x_1,\dots,x_n}\left(\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right) \frac{25n^3 \mathbb{1}(Y_1 < c_1^\alpha - \varepsilon_1)}{Y_{\lceil n\alpha\rceil} - Y_1}\right) .$$

Here, we have introduced $Y_1, \dots, Y_n$ to denote the ordered list of $(X_1, \dots, X_n)$ (i.e $Y_1 \leq Y_2 \leq \dots \leq Y_n$), where for any $j \in \mathbb{N}^*$ the variable X_j represents the j-th observation collected from arm 1. We also added the indicator $\mathbb{1}(Y_1 \leq c_1^\alpha - \varepsilon_1)$ because it is a necessary element of the next steps of the proof that aims at controlling Y_1. The inequality holds because if $Y_1 \geq c_1^\alpha - \varepsilon_1$ then $C_\alpha(\mathcal{X}_n, w) \geq c_1^\alpha - \varepsilon_1$ for any $w \in \mathcal{P}^n$. Then, under the events we consider it also holds that

$$Y_{\lceil n\alpha\rceil} \geq C_\alpha(\mathcal{X}_n) \geq c_1^\alpha - \varepsilon_1 .$$

Note that it is impossible to conclude at this step in general because the variable $Y_{\lceil n\alpha\rceil} - Y_1$ may be arbitrarily small in case all the n observations are very concentrated. However, if n is large and the distribution is *continuous* this event can only happen with a very low probability. This is a place in the proof where continuity is crucial. To do so, we upper bound the rest of the terms with a peeling argument on the values of Y_1. This is done using the closed-form formulas for the distribution of the minimum of n random variable that are independent and identically distributed. Indeed, if f_1 denotes the density of arms 1, and we write the cdf and pdf of the minimum of n independent observations of ν_1 respectively L_n and l_n, then it holds that $\forall x \in [0, B]$

$$L_n(x) = 1 - (1 - F_1(x))^n .$$

Now, since ν_1 is continuous it follows that in each point the density is $l_n(x) = nf_1(x)(1 - F_1(x))^{n-1}$. The next step consists in defining a strictly decreasing sequence $(a_k)_{k\geq 0}$, and to look at the intervals $[c_1^\alpha - a_k - \varepsilon_1, c_1^\alpha - a_{k+1} - \varepsilon_1]$. On each of these intervals we obtain by construction that $Y_{\lceil n\alpha\rceil} \geq c_1^\alpha - \varepsilon_1 \geq Y_1 + a_{k+1}$, and thus

$$\mathbb{E}_{\mathcal{X}_n}\left[\frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1}\mathbb{1}\left(Y_1 \in [c_1^\alpha - a_k - \varepsilon_1, c_1^\alpha - a_{k+1} - \varepsilon_1]\right)\right] \leq \frac{25n^3}{a_{k+1}} \times \mathbb{P}\left(Y_1 \in [c_1^\alpha - a_k - \varepsilon_1, c_1^\alpha - a_{k+1} - \varepsilon_1]\right) .$$

Using the properties of the density l_n it holds that

$$\begin{aligned}
\mathbb{P}\left(y_1 \in [c_1^\alpha - a_k - \varepsilon_1, c_1^\alpha - a_{k+1} - \varepsilon_1]\right) &= \int_{c_1^\alpha - \varepsilon_1 - a_k}^{c_1^\alpha - \varepsilon_1 - a_{k+1}} nf_1(x)(1 - F_1(x))^{n-1} dx \\
&\leq \sup_{x\in[0,B]} f_1(x) \int_{c_1^\alpha - \varepsilon_1 - a_k}^{c_1^\alpha - \varepsilon_1 - a_{k+1}} n(1 - F_1(x))^{n-1} dx \\
&\leq \sup_{x\in[0,B]} f_1(x)(a_k - a_{k+1}) n(1 - F_1(c_1^\alpha - \varepsilon_1 - a_k))^{n-1} .
\end{aligned}$$

With these results at hand, we can now aim at upper bounding $\bar{A}_2$. To this end, we first introduce

$$\begin{aligned}\bar{A}_2 &\leq \mathbb{E}_{x_1,\dots,x_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right)\frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1}\right] \\ &\leq \underbrace{\mathbb{E}_{x_1,\dots,x_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right)\frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1}\mathbb{1}(y_1 \leq c_1^\alpha - \varepsilon_1 - a_0)\right]}_{A_{21}} \\ &+ \underbrace{\mathbb{E}_{x_1,\dots,x_n}\left[\mathbb{1}\left(c_1^\alpha - \varepsilon_1 \leq C_\alpha(\mathcal{X}_n) \leq c_1^\alpha - \varepsilon_1/2\right)\frac{25n^3}{Y_{\lceil n\alpha\rceil} - Y_1}\mathbb{1}(y_1 \geq c_1^\alpha - \varepsilon_1 - a_0)\right]}_{A_{22}} .\end{aligned}$$

The left-hand side term can be handled thanks to Brown's inequality (Brown, 2007), that we restate in Lemma 9 for completeness, and discuss in Appendix B.2.2. Using that $Y_{\lceil n\alpha\rceil} - Y_1 \geq a_0$ on the considered interval, we obtain

$$A_{21} \leq \frac{25n^3}{a_0} e^{-2n\left(\frac{\alpha(a_0+\varepsilon_1)}{B_k}\right)^2} .$$

Regarding the second term A_{22} we have

$$A_{22} \leq \sup_{x\in[0,B]} nf_1(x) \times \sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}}(1 - F_1(c_1^\alpha - \varepsilon_1 - a_k))^{n-1} .$$

We first use that the cdf is increasing, which enables to upper bound $(1 - F_1(c_1^\alpha - \varepsilon_1 - a_k))^{n-1}$ by the quantity $(1 - F_1(c_1^\alpha - \varepsilon_1 - a_0))^{n-1}$. It remains to choose the sequence (a_k) in order to make the sum $\sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}}$ converge. We define recursively the sequence as $a_{k+1} = \frac{2^k}{2^k+1}a_k$, starting from $a_0 = \frac{c_1^\alpha - \varepsilon_1}{2}$. This way, $\sum_{k=0}^{+\infty} \frac{a_k - a_{k+1}}{a_{k+1}} = \sum_{k=0}^{+\infty} \frac{1}{2^k} = 2$. This shows that

$$A_{22} \leq 50n^4 \sup_{x\in[0,B]} f_1(x)\exp\left(-n\log(1 - F_1(c_1^\alpha - \varepsilon_1))\right) .$$

Hence, both terms A_{21} and A_{22} are asymptotically negligible, hence we can write that $\bar{A}_2 = O(1)$.

C.2.3. Upper bound on $\bar{A}_3$

We now turn to the last term $\bar{A}_2$, and first upperbound it as

$$\bar{A}_3 \leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n}\left[\mathbb{1}\left(C_\alpha(\mathcal{X}_n) < c_1^\alpha - \varepsilon_1\right)\frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X}_n, w) \geq c_1^\alpha - \varepsilon_1)}\right] .$$

We can use the same discretization arguments as in Appendix C.2.1 to handle this term. More precisely, we introduce a number of bins M' that is specified later in the proof, and for any $i \in \{0,\dots,n\}$ we again define $\widetilde{X}_i = \frac{\lfloor MX_i\rfloor}{M}$ and $\widetilde{\mathcal{X}}_n$ the corresponding set of truncated observations. Thanks to these definitions we can upper bound $\bar{A}_3$ as

$$\begin{aligned}\bar{A}_3 &\leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n}\left(\mathbb{1}\left(C_\alpha\left(\widetilde{\mathcal{X}}_n, w\right) < c_1^\alpha - \varepsilon_1\right)\frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\tilde{\mathcal{X}}_n, w) \geq c_1^\alpha - \varepsilon_1 - 1/M)}\right) \\ &\leq \sum_{n=1}^{T} \mathbb{E}_{\mathcal{X}_n}\left(\mathbb{1}\left(C_\alpha\left(\widetilde{\mathcal{X}}_n, w\right) < c_1^\alpha - \varepsilon_1 - \frac{1}{M}\right)\frac{1}{\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\tilde{\mathcal{X}}_n, w) \geq c_1^\alpha - \varepsilon_1 - 1/M)}\right) .\end{aligned}$$

Following Appendix C.2.1, we require ε_1 to be small enough in order to keep the same order for the CVaR of the discretized distributions. To proceed, we then choose M of the same order as $1/\varepsilon_1$, writing $\varepsilon_1 + 1/M = \varepsilon_1'$. Then we remark that this new formulation is exactly equivalent to the one we got in B.2.3. Hence, introducing

$$\delta_2 = \inf_{p \in \mathcal{P}^M : C_\alpha(\mathcal{X}, p) \leq c_1^\alpha - \varepsilon_1'} \left[\mathcal{K}_{\inf}^{\alpha, \tilde{\mathcal{X}}}(p, c_1^\alpha) - \mathcal{K}_{\inf}^{\alpha, \tilde{\mathcal{X}}}(p, c_1^\alpha - \varepsilon_1') \right] ,$$

where $\widetilde{\mathcal{X}}$ is $\widetilde{\mathcal{X}}_n$ where each item is only repeated once (the set built from $\widetilde{\mathcal{X}}_n$), the same steps allow us to finally obtain

$$\begin{aligned} \bar{A}_3 \leq & n_{(\delta_2/2)} + \sum_{n=n_{(\delta_1/2)}}^{T} C_2^{-1} (n+1)^{\frac{3M}{2}} (n+1)^M \exp\left(-(n+M) \frac{\delta_1}{2} \right) \\ \leq & n_{(\delta_2/2)} + \sum_{n=n_{(\delta_1/2)}}^{T} C_2^{-1} (n+1)^{\frac{5M}{2}} \exp\left(-(n+M) \frac{\delta_1}{2} \right) . \end{aligned}$$

Hence, this is again upper bounded by a constant. This final result concludes the proof of Equation (6) for continuous bounded distribution, which states that for B-CVTS

$$\text{(Pre-CV)} = O(1) .$$

D. Lower Bound and properties of $\mathcal{K}_{\inf}^{\alpha}$

In the classical bandit setting, asymptotic optimality is an important notion that has guided the design of algorithms, and we investigate in this section the optimal (problem-dependent) scaling of the CVaR-regret. We start by proving Theorem 1, and then investigate some properties of the obtained lower bound that permit to derive concentration inequalities for the Dirichlet distributions.

D.1. Proof of Theorem 1

In this section, we prove Theorem 1. We rely on the fundamental inequality (6) of (Garivier et al., 2019) which shows that, if ν and ν' are two bandit models in $\mathcal{D}$, for any $\mathcal{F}_T$-measurable random variable $Z \in [0, 1]$,

$$\sum_{k=1}^{K} \mathbb{E}_{\pi,\nu}[N_k(T)] \mathrm{KL}(\nu_k, \nu_k') \geq \mathrm{kl}(\mathbb{E}_{\pi,\nu}[Z], \mathbb{E}_{\pi,\nu'}[Z]) ,$$

where $\mathrm{kl}(x, y) = x \log\left(\frac{x}{y}\right) + (1-x) \log\left(\frac{1-x}{1-y}\right)$ denotes the binary relative entropy.

Fix $\nu = (\nu_1, \ldots, \nu_K) \in \mathcal{D}$ and let k be a sub-optimal arm in ν, that is $c_k^\alpha < c^\star$. Assume that there exists $\nu_k' \in \mathcal{D}_k$ such that $\mathrm{CVaR}_\alpha(\nu_k') > c^\star$ (if this does not hold, $\mathcal{K}_{\inf}^{\alpha, \mathcal{D}_k}(\nu_k, c^\star) = +\infty$ and the lower bound holds trivially). Then considering the alternative bandit model ν' in which $\nu_i' = \nu_i$ for all $i \neq k$ and ν_k' is defined above, we obtain

$$\mathbb{E}_{\pi,\nu}[N_k(T)] \mathrm{KL}(\nu_k, \nu_k') \geq \mathrm{kl}\left(\mathbb{E}_{\pi,\nu}\left[\frac{N_k(T)}{T}\right], \mathbb{E}_{\pi,\nu'}\left[\frac{N_k(T)}{T}\right] \right) .$$

Exploiting the fact that the strategy π has its CVaR-regret in $o(T^\beta)$ for any $\beta > 0$, one can prove that, for any β,

$$\mathbb{E}_{\pi,\nu}\left[N_k(T)\right] = o(T^\beta) \quad \text{and} \quad T - \mathbb{E}_{\pi,\nu'}\left[N_k(T)\right] = o(T^\beta)$$

since arm k is the (unique) optimal arm under ν'. Using the exact same arguments as (Garivier et al., 2019) enables to prove that

$$\liminf_{T \to \infty} \frac{\mathrm{kl}\left(\mathbb{E}_{\pi,\nu}\left[\frac{N_k(T)}{T}\right], \mathbb{E}_{\pi,\nu'}\left[\frac{N_k(T)}{T}\right] \right)}{\log(T)} \geq 1,$$

which yields

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\pi,\nu}[N_k(T)]}{\log(T)} \geq \frac{1}{\mathrm{KL}(\nu_k, \nu_k')} .$$

Taking the infimum over $\nu_k' \in \mathcal{D}_k$ such that $\mathrm{CVaR}_\alpha(\nu_k') > c^\star$ yields the result, by definition of $\mathcal{K}_{\inf}^{\alpha, \mathcal{D}_k}$.

D.2. Discussion on the scaling of the minimal regret

Using Lemma A.2 of Tamkin et al. (2020), we can show that for any distributions ν_F and ν_G with respective CDFs F and G that are supported in $[0,1]$,

$$|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)| \leq \frac{1}{\alpha}||F-G||_\infty \ .$$

It follows from Pinsker's inequality that $\mathrm{KL}(\nu_F,\nu_G) \geq \alpha^2 \left(\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)\right)^2/2$. Therefore, in a bandit model in which all ν_k are supported in $[0,1]$ (that is, all $\mathcal{D}_k$ are equal to $\mathcal{P}([0,1])$, the set of probability measures on $[0,1]$), it follows that

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(\nu_k, c^*) \geq (\alpha\Delta_k^\alpha)^2/2.$$

Combining this inequality together with the lower bound of Theorem 1, we obtain that the regret of an algorithm matching the lower bound is upper bounded by $\mathcal{O}\left(\sum_{k:c_k^\alpha<c^\star} \frac{\log(T)}{\alpha^2\Delta_k^\alpha}\right)$, which is precisely the scaling of the CVaR regret bounds obtained for the U-UCB (Cassel et al., 2018) and CVaR-UCB (Tamkin et al., 2020). Assuming the above inequalities are tight for some distributions (which may not be the case), one may qualify these algorithms as "order-optimal", as their CVaR regret makes appear the good scaling in the gaps (and in α), just like the UCB1 algorithm (Auer et al., 2002) for $\alpha = 1$. In this paper we go beyond order-optimality, and we strive to design algorithms that are asymptotically optimal.

D.3. Lemma 6: continuity of $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}$ for multinomial distributions

We state the following result on the continuity of the $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}$ functional for multinomial distributions.

Lemma 6. *The mapping $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}} : \mathcal{P}^M \times [x_1, x_M) \mapsto \mathbb{R}$ is continuous in its two arguments. Furthermore, for all $p \in \mathcal{P}^M$ the mapping $c \mapsto \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)$ is increasing on $(C_\alpha(\mathcal{X},p), x_M]$.*

Proof. We recall that a multinomial distribution ν is characterized by its finite support $\mathcal{X} = (x_1,\dots,x_M)$ and a probability vector $p \in \mathcal{P}^M = \left\{q \in \mathbb{R}^M : \forall i, q_i \geq 0, \sum_{j=1}^M q_j = 1\right\}$ such that $\mathbb{P}_{X\sim\nu}(X = x_k) = p_k$ for all $k \in \{1,\dots,M\}$. We assume that $x_1 \leq x_2 \leq \dots \leq x_M$.

Moreover, by a slight abuse of notation, we use $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)$ as a shorthand for $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}_\mathcal{X}}(\nu,c)$ where $\mathcal{D}_\mathcal{X}$ is the set of multinomial distribution supported on $\mathcal{X}$. That is, for all $p \in \mathcal{P}^M$ and $c \in [x_1, x_M]$,

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c) = \inf_{q\in\mathcal{P}^M} \{\mathrm{KL}(p,q) : C_\alpha(\mathcal{X},q) \geq c\} \ ,$$

where $\mathrm{KL}(p,q) = \sum_{i=1}^M p_i \log\left(\frac{p_i}{q_i}\right)$.

More precisely, we use the Berge's theorem (see, e.g. (Berge, 1997)) to prove the continuity of the mapping $\mathcal{K}_c : p \in \mathcal{P}^M \to \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)$ for any $c \in (0, x_M)$ and that of $\mathcal{K}_p : c \in [0, x_M) \to \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p,c)$ for any $p \in \mathcal{P}^M$. Those functions are of the form

$$\mathcal{K}_c(p) = \inf_{q\in\Gamma(p)} \mathrm{KL}(p,q) \ \text{ and } \ \mathcal{K}_p(c) = \inf_{q\in\Gamma(c)} \mathrm{KL}(p,q)\,,$$

where $\Gamma(c) = \Gamma(p) = \{q \in \mathcal{P}^M : C_\alpha(\mathcal{X},q) \geq c\}$. As $\mathrm{KL}(p,q)$ is continuous in both arguments and the feasible set is compact, it is sufficient to prove that the correspondences $c \mapsto \Gamma(c)$ and $p \mapsto \Gamma(p)$ are hemicontinuous, i.e that they are non-empty, lower hemicontinuous and upper hemicontinuous. For $\Gamma(p)$ the lower and upper hemicontinuity is trivial as the feasible set does not depend on p. Moreover, this set is non empty as the Dirac in x_M (represented by $q = (0,\dots,0,1)$) belongs to $\Gamma(p)$, for any $c \in [x_0, x_M]$. To conclude, it thus remains to prove that $c \mapsto \Gamma(c)$ is both a lower and upper hemicontinuous correspondence.

We first prove the lower hemicontinuity in every $c_0 \in [x_0, x_M)$. Consider an open set $V \in \mathcal{P}^{M+1}$ satisfying $V \cap \Gamma(c_0) \neq \emptyset$. We must prove that there exists $\varepsilon > 0$ such that for all $c \in (c_0-\varepsilon, c_0+\varepsilon)$, $V \cap \Gamma(c) \neq \emptyset$. Since $c \mapsto \Gamma(c)$ is nonincreasing in the sense of the set inclusion, it is sufficient to justify that $V \cap \Gamma(c_0+\varepsilon) \neq \emptyset$ for some $\varepsilon > 0$. Let $q_0 \in V \cap \Gamma(c_0)$. It holds that $C_\alpha(\mathcal{X},q_0) \geq c_0$. If the inequality is strict, there exists ε such that $q_0 \in V \cap \Gamma(c_0+\varepsilon)$. Assume that $C_\alpha(\mathcal{X},q_0) = c_0$. As V is open, there exists ε' such that any $B(q_0,\varepsilon') \subseteq V$. Now there must exist $q \in B(q_0,\varepsilon')$ such that $C_\alpha(\mathcal{X},q) > C_\alpha(\mathcal{X},q_0) = c_0$. Indeed, in order to construct such a probability vector, we can take out some mass from the

components of q corresponding to x_0 and assign it to the component corresponding to x_M. This increases the CVaR while staying in the ball provided that the change is small enough. Hence, there exists $\varepsilon > 0$ such that $q \in V \cap \Gamma(c_0 + \varepsilon)$.

To prove the upper hemicontinuity, we use the following sequential characterization: if c_n is a sequence taking values in $[x_0, x_M)$ that converges to c and q_n is a sequence taking values in $\mathcal{P}^{M+1}$ that converges to q, with $q_n \in \Gamma(c_n)$ for all n, one has to prove that $q \in \Gamma(c)$. This fact is a simple consequence of the continuity of $q \mapsto C_\alpha(\mathcal{X}, q)$ on $\mathcal{P}^{M+1}$, which is obvious as this function is the supremum of affine functions, as can be seen in Equation (8).

We know prove that the mapping $\mathcal{K}_p(c)$ is strictly increasing on $(C_\alpha(\mathcal{X}, p), x_M)$. The fact that this mapping is non-decreasing is a simple consequence of the fact that for $c < c'$, $\Gamma(c') \subseteq \Gamma(c)$. In order to prove the strict monotonicity it is sufficient to prove that the constraints are binding t the optimum. Assume that this is not the case, i.e. for some $c \in (C_\alpha(\mathcal{X}, p), x_M)$, $\exists p_c^* : \mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p, c) = \mathrm{KL}(p, p_c^*)$ such that $C_\alpha(\mathcal{X}, p_c^*) = c + \delta$ for some $\delta > 0$. By continuity of the CVaR, there exists some $\varepsilon > 0$ such that $\mathcal{B}(p_c^*, \varepsilon) \subset \Gamma(c + \delta/2) \subset \Gamma(c)$, where $\mathcal{B}(p_c^*, \varepsilon) = \{q \in \mathcal{P}^M : d(q, p_c^*) = ||q - p_c^*||_\infty \leq \varepsilon\}$. By definition of p_c^* we should have that for any distribution $q \in \mathcal{B}(p_c^*, \varepsilon)$, $\mathrm{KL}(p, q) \geq \mathrm{KL}(p, p_c^*)$. Consider a distribution $\widetilde{p}$ satisfying for some (i, j), $\widetilde{p}_i = p_{c,i}^* + \varepsilon$, $\widetilde{p}_j = p_{c,j}^* - \varepsilon$ and $\widetilde{p}_\ell = p_{c,\ell}^*$ for $\ell \neq i, j$. Then, it holds that:

$$\mathrm{KL}(p, p_c^*) - \mathrm{KL}(p, \widetilde{p}) = p_i \log\left(\frac{p_{c,i}^* + \varepsilon}{p_{c,i}^*}\right) + p_j \log\left(\frac{p_{c,j}^* - \varepsilon}{p_{c,j}^*}\right) .$$

By a simple Taylor expansion. we have $\mathrm{KL}(p, p_c^*) - \mathrm{KL}(p, \widetilde{p}) = \varepsilon\left(\frac{p_i}{p_{c,i}^*} - \frac{p_j}{p_{c,j}^*}\right) + o(\varepsilon^2)$. So, if ε is chosen small enough, the difference has the same sign as $\left(\frac{p_i}{p_{c,i}^*} - \frac{p_j}{p_{c,j}^*}\right)$. Since $p_c^* \neq p$ we are sure to find some coordinates satisfying $p_i > p_{c,i}^*$ and $p_j < p_{c,j}^*$, hence this term can be made positive for an appropriate choice of (i, j). This means that if the constraint is not binding at the optimum we can necessarily find a distribution in the feasible set with a lower KL-divergence with p, which is a contradiction. Hence, $\mathcal{K}_p$ is strictly increasing on $(C_\alpha^{\mathcal{X}}, x_M)$. □

D.4. Proof of Lemma 4: dual form of the $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}$ of multinomial distributions

In this section we derive the dual form of the function $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}$, where $\mathcal{X}$ is some finite support $\mathcal{X} = (x_1, \ldots, x_m) \in [0, 1]^M$. We let $\mathcal{P}^M$ denote he simplex of dimension M. We rewrite the optimization problem, defined for any $p \in \mathcal{P}^M$, $\alpha \in (0, 1]$ and $c \in [0, 1]$ as

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p, c) = \inf_{q \in \mathcal{P}^M} \{\mathrm{KL}(p, q) : C_\alpha(\mathcal{X}, q) \geq c\} .$$

First of all, we recall that $C_\alpha(\mathcal{X}, q) = \sup_{x \in \mathcal{D}} \left\{x - \frac{1}{\alpha}\mathbb{E}_{X \sim q}\left((x - X)^+\right)\right\}$. We then introduce the set

$$\begin{aligned}\mathcal{P}_{y,\alpha,c}^M &= \left\{q \in \mathcal{P}^M : y - \frac{1}{\alpha}\mathbb{E}_{X \sim q}((y - X)^+) \geq c\right\} \\ &= \left\{q \in \mathcal{P}^M : \mathbb{E}_{X \sim q}((y - X)^+) \leq (y - c)\alpha\right\} .\end{aligned}$$

Thanks to this definition we can rewrite the problem as

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p, c) = \min_{y \in \mathcal{D}}\left\{\inf_{q \in \mathcal{P}^M}\left\{\mathrm{KL}(p, q) : y - \frac{1}{\alpha}\mathbb{E}_{X \sim q}((y - X)^+) \geq c\right\}\right\} ,$$

where we used that $\{q : C_\alpha(\mathcal{X}, q) \geq c\} = \cup_{y \in \mathcal{D}}\left\{q \in \mathcal{P}^M : y - \frac{1}{\alpha}\mathbb{E}_{X \sim q}\left((X - y)^+\right) \geq c\right\}$.

Now, we can first solve the problem $\inf_{q \in \mathcal{P}_{y,\alpha,c}^M} \mathrm{KL}(p, q)$ for a fixed value of y, satisfying $y > c$ (else the feasible set is empty). We write the Lagrangian of this problem:

$$H(q, \lambda_1, \lambda_2) = \sum_{i=1}^M p_i \log\left(\frac{p_i}{q_i}\right) + \lambda_1\left(\sum_{i=1}^M q_i - 1\right) + \lambda_2\left(\sum_{i=1}^M q_i(y - x_i)^+ - \alpha(y - c)\right) ,$$

and want to solve $\max_{\lambda_1>0,\lambda_2>0} \min_q H(q,\lambda_1,\lambda_2)$. To this end, we write

$$\frac{\partial H}{\partial q_i} = -\frac{p_i}{q_i} + \lambda_1 + (y-x_i)^+ \ .$$

Setting the derivative to 0 yields

$$q_i = \frac{p_i}{\lambda_1 + \lambda_2 (y-x_i)^+} \ .$$

We can check that the inequality constraint is achieved. Moreover, exploiting the two constraints leads to $\lambda_1+\lambda_2\alpha(y-c)=1$. This finally gives

$$q_i = \frac{p_i}{1-\lambda_2((y-c)\alpha-(y-x_i)^+)} \ .$$

Note that this solution is only valid if $\lambda_2 \leq \frac{1}{\alpha(y-c)}$. We have two possibilities: 1) the maximum is achieved in $[0, \frac{1}{\alpha(y-c)})$, in this case we have

$$\begin{aligned}\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p,c) &= \inf_{y\in\mathcal{D}} \max_{\lambda\in\left[0,\frac{1}{\alpha(y-c)}\right)} \sum_{i=1}^{M} p_i \log\left(1-\lambda_2((y-c)\alpha-(y-x_i)^+)\right) \\ &= \inf_{y\in\mathcal{D}} \max_{\lambda\in\left[0,\frac{1}{\alpha(y-c)}\right)} \mathbb{E}_{X\sim p}[\log\left(1-\lambda_2((y-c)\alpha-(y-X)^+)\right)] \ .\end{aligned}$$

The other possibility is that the function is still increasing in $\lambda_2 = \frac{1}{\alpha(y-c)}$. For this case, we check the sign of $\frac{\partial \mathbb{E}_{X\sim F}[\log\left(1+\lambda_2((y-c)\alpha-(y-X)^+)\right)]}{\partial\lambda}$ at point $\lambda = \frac{1}{\alpha(y-c)}$, that is of $(y-c)\alpha\left(1-\mathbb{E}_F\left(\frac{(y-c)\alpha}{(y-X)^+}\right)\right)$. We see that the function can only be increasing if $\mathbb{E}_F\left(\frac{(y-c)\alpha}{(y-X)^+}\right) < 1$, and the solution is then $q_i = \frac{p_i(y-c)\alpha}{y-x_i}$, which provides $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(p,c) = \inf_y \mathbb{E}_F\left(\frac{(y-X)^+}{(y-c)\alpha}\right)$. This concludes the proof.

We remark that these results coincide with those of Honda and Takemura (2010) with $\alpha=1$ and $y=1$.

E. Auxiliary results

In this Section we provide some technical tools about the CVaR. In particular, we prove several results that were presented in Section 3, namely Lemma 1, Lemma 2, Lemma 3, and Lemma 5.

E.1. Some basic CVaR properties

In this section we develop some well-known properties of the CVaR. First, the definition of the CVaR as the solution of an optimization problem was first introduced by Rockafellar et al. (2000), to formalize previous heuristic definitions of the CVaR as an average over a certain part of the distribution. The definition (1) is indeed appealing as it applies to any distribution for which $\mathbb{E}[(x-X)^+]$ is defined, including both discrete and continuous distributions. To understand the CVaR it is particularly useful to look at its expression in these two particular cases. First, for any continuous distribution ν of CDF F it can be shown (see, e.g. Acerbi and Tasche (2002)) that

$$\mathrm{CVaR}_\alpha(\nu) = \mathbb{E}_{X\sim\nu}\left[X | X \leq F^{-1}(\alpha)\right] .$$

This expression provides a good intuition on what the CVaR represents, as the expectation of the distribution after excluding the best scenarios covering a fraction $(1-\alpha)$ of the total mass. A similar definition exists for real-valued distributions ν with discrete support $\mathcal{X} = (x_1, x_2, \dots)$ (either finite or infinite). Assuming that the sequence (x_i) is increasing and letting $p_i = \mathbb{P}_{X\sim\nu}(X=x_i)$, one has

$$\mathrm{CVaR}_\alpha(\nu) = \sup_{x_n\in\mathcal{X}} \left\{ x_n - \frac{1}{\alpha}\sum_{i=1}^{n-1} p_i(x_n - x_i) \right\} \ . \tag{8}$$

Indeed, the function to maximize in (1) is piece-wise linear, so the maximum is necessarily achieved in a point of discontinuity. In particular, we can easily prove that if n_α is the first index satisfying $\sum_{i=1}^{n_\alpha} p_i \geq \alpha$, then the supremum is achieved in n_α and

$$\begin{aligned}\mathrm{CVaR}_\alpha(\nu) &= x_{n_\alpha} - \frac{1}{\alpha}\sum_{i=1}^{n_\alpha-1} p_i(x_{n_\alpha} - x_i) \\ &= \frac{1}{\alpha}\left(\sum_{i=1}^{n_\alpha-1} p_i x_i + \left(\alpha - \sum_{i=1}^{n_\alpha-1} p_i\right) x_{n_\alpha}\right) .\end{aligned}$$

Hence in that case the CVaR can also be seen as an average when we consider the lower part of the distribution before reaching a total mass α.

From the general definition (1), one can also observe that for $\alpha = 1$, $\mathrm{CVaR}_\alpha(\nu) = \mathbb{E}_{X\sim\nu}(X)$. Moreover, the mapping $\alpha \mapsto \mathrm{CVaR}_\alpha(\nu)$ is continuous on $(0, 1]$. Thus, considering CVaR bandits allows to smoothly interpolate between classical bandits (that correspond to $\alpha = 1$) and risk-averse problems.

We also prove in this section a technical result needed in the proof of Theorem 2 that relates the CVaR of two distributions that are close in terms of the L^∞ distance defined in Appendix A.

Lemma 7 (CVaR of two discrete distributions in a L^∞ ball)**.** *Let p and q be the probability vectors of two discrete distribution of with shared support $\mathcal{X} = \{x_1, \dots, x_M\}$, then for any $\alpha \in (0, 1]$ and any $\varepsilon > 0$:*

$$C_\alpha(\mathcal{X}, p) - \frac{M||p-q||_\infty}{\alpha} x_M \leq C_\alpha(\mathcal{X}, q) \leq C_\alpha(\mathcal{X}, p) + \frac{M||p-q||_\infty}{\alpha} x_M .$$

Proof. For any $p \in \mathcal{P}^M$ and $q \in \mathcal{P}^M$ we write for simplicity $\varepsilon = \sup_{i\in\{0,1,\dots,M\}} |p_i - q_i| = ||p-q||_\infty$, so $\forall i$: $q_i - \varepsilon \leq p_i \leq q_i + \varepsilon$. Let's consider the optimisation problem used to compute the CVaR, $\forall x$:

$$\begin{aligned}x - \frac{1}{\alpha}\sum_{i=0}^{M} p_i\,(x - x_m)^+ &\leq x - \frac{1}{\alpha}\sum_{i=0}^{M}(q_i - \varepsilon)\,(x - x_m)^+ \\ &= x - \frac{1}{\alpha}\sum_{i=0}^{M} q_i\,(x - x_m)^+ + \frac{\varepsilon}{\alpha}\sum_{i=0}^{M}(x - x_m)^+ \\ &\leq x - \frac{1}{\alpha}\sum_{i=0}^{M} q_i\,(x - x_m)^+ + \frac{\varepsilon}{\alpha}(M+1)x_M .\end{aligned}$$

Taking the supremum on all possible values x on the left side of the inequality and then on the right side ensures the result. Then, replacing p by q proves the other inequality. □

E.2. Proof of Lemma 1

In this section we prove Lemma 1, introduced in Section 3.

Lemma 1 (Upper Bound)**.** *For any $(\beta, p) \in \mathcal{Q}_n^M$, for any $c > C_\alpha(\mathcal{X}, p)$, it holds that*

$$\mathbb{P}_{w\sim Dir(\beta)}(w \in \mathcal{S}_{\mathcal{X}}^\alpha(c)) \leq C_1 M n^{M/2} \exp(-n\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(p, c)) ,$$

for some constant C_1.

We recall that the set $\mathcal{Q}_n^M$ is defined as

$$\mathcal{Q}_n^M = \left\{(\beta, p) \in \mathbb{N}^{*n} \times \mathcal{P}^M : p = \frac{\beta}{n}\right\} .$$

The proof relies on Lemma 13 of (Riou and Honda, 2020) that we re-state below for completeness.

Lemma 8 (Lemma 13 in (Riou and Honda, 2020)). *Assume $w \sim \mathrm{Dir}(\beta)$ a Dirichlet distribution over the probability simplex $\mathcal{P}^M$. We assume that $\beta^T 1 = n$ and $\forall j \in \{1, \dots, M\}, \beta_j \geq 0$. We denote by $p = \frac{1}{n}\beta$ the mean of the Dirichlet distribution. Let $S \subset \mathcal{P}^{M+1}$ a closed convex set included in the probability simplex. The following bound holds:*

$$\mathbb{P}_{w\sim \mathrm{Dir}(\beta)}\left(w \in S\right) \leq C_1 n^{\frac{M}{2}} \exp\left(-n\mathrm{KL}\left(p, p^*\right)\right) ,$$

where $p^* = argmin_{x\in S}\mathrm{KL}(p, x)$.

Using the notation of Section 3, for any $w \in \mathcal{P}^M$, $\mathbb{P}(w \in \mathcal{S}^\alpha_{\mathcal{X}}(c)) \leq \sum_{m=1}^M \mathbb{P}(w \in \mathcal{S}^\alpha_{m,\mathcal{X}}(c))$. Then, using Lemma 8 for each subset $\mathcal{S}^\alpha_{m,\mathcal{X}}(c)$, which is closed and convex, we have

$$\mathbb{P}(w \in \mathcal{S}^\alpha_{\mathcal{X}}(c)) \leq C_1 n^{M/2} \sum_{m=1}^M \exp\left(-n\mathrm{KL}\left(\frac{\beta}{n}, p^*_m\right)\right) ,$$

where $p^*_m = \mathrm{argmin}_{x\in\mathcal{S}^\alpha_{m,\mathcal{X}}(c)}\mathrm{KL}\left(\frac{\beta}{n}, x\right)$.

We conclude by using that there exists some $i \in \{1, \dots, M\}$ such that $\mathcal{K}^{\alpha,\mathcal{X}}_{\inf}\left(\frac{\beta}{n}, c\right) = \mathrm{KL}\left(\frac{\beta}{n}, p^*_i\right)$, and for all $m \neq i$ $\mathrm{KL}(q, p^*_m) \geq \mathrm{KL}(q, p^*_i)$.

E.3. Proof of Lemma 2

We prove the Lemma 2 presented in Section 3.

Lemma 2 (Lower Bound). *For any $(M, n) \in N^2$ and $(\beta, p) \in \mathcal{Q}^M_n$, if n is large enough it holds that*

$$\mathbb{P}_{w\sim \mathrm{Dir}(\beta)}\left(w \in \mathcal{S}^\alpha_{\mathcal{X}}(c)\right) \geq C_2 \frac{\exp\left(-n\mathcal{K}^{\alpha,\mathcal{X}}_{\inf}(p, c)\right)}{n^{\frac{3M}{2}+1}} ,$$

for some constant $C_2 = \left(\frac{1}{\sqrt{2\pi}}\right)^M e^{-(M+1)/12}$.

We follow the sketch of the proof of Lemma 14 of (Riou and Honda, 2020) using Equation (8). We start by stating that there exists some $m \in \{1, \dots, M\}$ such that $C_\alpha\left(\mathcal{X}, \frac{\beta}{n}\right) = x_m - \frac{1}{\alpha}\sum_{i=1}^{m-1} q_i(x_m - x_i)$ and some p^* such that $\mathcal{K}^{\alpha,\mathcal{X}}_{\inf}\left(\frac{\beta}{n}, c\right) = \mathrm{KL}\left(\frac{\beta}{n}, p^*\right)$. The existence of p^* is ensured by the fact that the function $\mathcal{K}^{\alpha,\mathcal{X}}_{\inf}$ is the solution of the minimization of a continuous function on a compact set. We consider the set

$$\mathcal{S}_2 = \{w \in \mathcal{P}^{M+1} : w_i \in [0, p^*_i], \forall j \in \{m, \dots, M\} : w_j \geq p^*_j\}.$$

Let us remark that $\forall p \in \mathcal{S}_2, C_\alpha(\mathcal{X}, p) \geq C_\alpha(\mathcal{X}, p^*) \geq c$. Indeed, if we transfer some of the mass from some items of the support to larger items we can only increase the CVaR. It holds that

$$\begin{aligned}
\mathbb{P}_{w\sim Dir(\beta)}\left(C_\alpha(\mathcal{X}, w) \geq C_\alpha(\mathcal{X}, w)\right) &\geq \mathbb{P}_{w\sim Dir(\beta)}\left(w \in \mathcal{S}_2\right) \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \int_{x\in S_2} \prod_{i=1}^M x_i^{\beta_i - 1} dx \\
&\geq \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \prod_{i=m}^M (p^*_i)^{\beta_i - 1} \prod_{j=1}^{m-1} \int_{x_j=1}^{p^*_j} x_j^{\beta_j} dx_j \\
&= \frac{\Gamma(n)}{\prod_{i=1}^M \Gamma(\beta_i)} \prod_{i=m}^M (p^*_i)^{\beta_i - 1} \prod_{j=1}^{m-1} \frac{(p^*_j)^{\beta_j}}{\beta_j} .
\end{aligned}$$

We then use that the KL-divergence between two multinomial distributions has a simple form to compute:

$$\begin{aligned}\mathbb{P}_{w\sim Dir(\beta)}\left(w\in\mathcal{S}_2\right) &\geq \frac{\Gamma(n)}{n^m\prod_{i=0}^M\Gamma(\beta_i)}\prod_{i=m}^M\frac{\beta_i}{np_i^*}\prod_{j=0}^M\left(\frac{p_j^*n}{\beta_j}\right)^{\beta_j}\prod_{s=1}^M\left(\frac{\beta_s}{n}\right)^{\beta_s-1}\\ &\geq \frac{\Gamma(n)}{n^m}\prod_{i=m}^M\frac{\beta_i}{np_i^*}\exp\left(-n\mathrm{KL}\left(\frac{\beta}{n},p^*\right)\right)\prod_{s=1}^M\frac{(\beta_s/n)^{\beta_s-1}}{\Gamma(\beta_s)}\ .\end{aligned}$$

We then use the Lemma 12 of (Riou and Honda, 2020), which is an application of the Stirling formula, in order to lower bound the terms depending on the gamma function:

$$\begin{aligned}\mathbb{P}_{w\sim Dir(\beta)}\left(w\in\mathcal{S}_2\right) &\geq C_2n^{-\frac{M}{2}}\exp\left(-n\mathrm{KL}\left(\frac{\beta}{n},p^*\right)\right)\prod_{i=m}^M\frac{\beta_i}{np_i^*}\\ &\geq C_2n^{-\frac{M}{2}}\exp\left(-n\mathrm{KL}\left(\frac{\beta}{n},p^*\right)\right)\frac{1}{n^M}\\ &= C_2n^{-\frac{M}{2}}\exp\left(-n\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}\left(\frac{\beta}{n},c\right)\right)\frac{1}{n^M}\ ,\end{aligned}$$

where we used in the second line that $\frac{\beta_i}{np_i^*}\geq\frac{1}{n}$ and $C_2=\left(\frac{1}{\sqrt{2\pi}}\right)^M e^{-(M+1)/12)}$. This concludes the proof.

E.4. Proof of Lemma 3

We prove Lemma 3 that is introduced in Section 3.

Lemma 3. *Let $\mathcal{X}=(x_0,\dots,x_n)\subset[0,B]^{n+1}$ for some known $B>0$ and $n\in\mathbb{N}$, assuming that $x_0=B$. For any $c>C_\alpha(\mathcal{X})$, and any $\eta>0$ small enough it holds that*

$$\mathbb{P}_{w\sim\mathcal{D}_n}(C_\alpha(\mathcal{X},w)\geq c)\leq\frac{B}{\eta}\exp^{-N(\mathcal{K}_{\inf}^{\alpha}(u_{\mathcal{X}},c)-\eta C(B,\alpha,c))}\ ,$$

for some constant $C(B,\alpha,c)$.

Proof. We first use that $\{q:C_\alpha(\mathcal{X},q)\geq c\}=\cup_{y\in[c,B]}\left\{q\in\mathcal{P}^{n+1}:y-\frac{1}{\alpha}\mathbb{E}_{X\sim q}\left((X-y)^+\right)\geq c\right\}$ to write

$$\begin{aligned}\mathbb{P}_w\left(C_\alpha(\mathcal{X},w)\geq c\right)\leq&\mathbb{P}_w\left(\sup_{y\in[c,B]}\left\{y-\frac{1}{\alpha}\sum_{i=0}^n w_i(y-x_i)^+\right\}\geq c\right)\\ \leq&\int_c^B\mathbb{P}_w\left(y-\frac{1}{\alpha}\sum_{i=0}^n w_i(y-x_i)^+\geq c\right)\mathbb{P}_w\left(y=\mathrm{argsup}_{y\in[c,B]}\left\{y-\frac{1}{\alpha}\sum_{i=0}^n w_i(y-x_i)^+\right\}\right)dy\end{aligned}$$

The second term can have an arbitrarily complicated form, but we use the fact that the support is bounded, which enables to uniformly bound it by 1. We bound the first term by its supremum on $[0,B]$, hence

$$\begin{aligned}\mathbb{P}_w\left(C_\alpha(\mathcal{X},w)\geq c\right)\leq&B\sup_{y\in[c,B]}\mathbb{P}_w\left(y-\frac{1}{\alpha}\sum_{i=0}^n w_i(y-x_i)^+\geq c\right)\\ \leq&B\sup_{y\in[c,B]}\mathbb{P}_w\left(\alpha(y-c)-\sum_{i=0}^n w_i(y-x_i)^+\geq 0\right)\ .\end{aligned}$$

We then handle $\mathbb{P}\left(\alpha(y-c)-\sum_{i=0}^n w_i(y-x_i)^+\geq 0\right)$ for a fixed value of y. We here follow the path of Riou and Honda (2020), using that a Dirichlet random variable $w=(w_0,\dots,w_n)$ can be written in terms of $n+1$ independent random

variables $R_0, \dots, R_n$ following an exponential distribution, as $w_i = \frac{R_i}{\sum_{j=0}^n R_j}$. Using this property and multiplying by $\sum_{j=0}^n R_j$ we obtain

$$\begin{aligned}\mathbb{P}\left(\alpha(y-c) - \sum_{i=0}^n w_i (y-x_i)^+ \geq 0\right) \leq & \mathbb{P}\left(\sum_{i=0}^n R_i \left(\alpha(y-c) - (y-x_i)^+\right) \geq 0\right) \\ \leq & \mathbb{E}\left[\exp\left(t \sum_{i=0}^n R_i \left(\alpha(y-c) - (y-x_i)^+\right)\right)\right] ,\end{aligned}$$

where we used Markov's inequality for some $t \in \left[0, \frac{1}{(y-c)\alpha}\right)$. We then isolate the first term, writing

$$\begin{aligned}&\leq \prod_{i=0}^n \mathbb{E}\left[\exp\left(R_i t \left(\alpha(y-c) - (y-x_i)^+\right)\right)\right] \\ &\leq \exp\left(-\sum_{i=0}^n \log\left(1 - t\left(\alpha(y-c) - (y-x_i)^+\right)\right)\right) \\ &\leq \frac{1}{1-t\alpha(y-c)} \exp\left(-\sum_{i=1}^n \log\left(1 - t\left(\alpha(y-c) - (y-x_i)^+\right)\right)\right) \\ &\leq \frac{1}{1-t\alpha(y-c)} \left\{\exp\left(-N\mathbb{E}_{\widehat{F}}\left[\log\left(1 - t\left(\alpha(y-c) - (y-X)^+\right)\right)\right]\right)\right\} .\end{aligned}$$

Since the term $1 - t\alpha(y-c)$ can be arbitrarily small, we have to control the values of t in order to ensure that the constant before the exponential is not too large. Hence, we choose some constant $\eta > 0$, and write that for any $t \in [0, \frac{1-\eta}{\alpha(y-c)}]$ we have

$$\mathbb{P}_w\left(C_\alpha(\mathcal{X}, w) \geq c\right) \leq \frac{B}{\eta} \exp\left(-N\mathbb{E}_{\widehat{F}}\left[\log\left(1 - t\left(\alpha(y-c) - (y-X)^+\right)\right)\right]\right) ,$$

which leads to

$$\begin{aligned}\mathbb{P}_w\left(C_\alpha(\mathcal{X}, w) \geq c\right) \leq & \frac{B}{\eta} \inf_{t \in \left[0, \frac{1-\eta}{(y-c)\alpha}\right]} \exp\left(-N\mathbb{E}_{\widehat{F}}\left[\log\left(1 - t\left(\alpha(y-c) - (y-X)^+\right)\right)\right]\right) \\ \leq & \frac{B}{\eta} \exp\left(-N \sup_{t \in \left[0, \frac{1-\eta}{(y-c)\alpha}\right]} \mathbb{E}_{\widehat{F}}\left[\log\left(1 - t\left(\alpha(y-c) - (y-X)^+\right)\right)\right]\right) .\end{aligned}$$

At this step the dual form of the function $\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(\widehat{F})$ start to appear, however, we have to handle the interval on which the supremum is taken is $\left[0, \frac{1-\eta}{\alpha(y-c)}\right]$ instead of $\left[0, \frac{1}{\alpha(y-c)}\right]$. As in (Riou and Honda, 2020) we will use the concavity and the regularity of the function in the expectation in order to conclude. We write

$$\phi(t) = \frac{1}{n} \sum_{i=1}^n \log\left(1 - t(\alpha(y-c) - (y-x_i)^+)\right) .$$

As ϕ is concave it holds that for any $t \in \left[\frac{1-\eta}{(y-c)\alpha}, \frac{1}{(y-c)\alpha}\right)$ we have

$$\phi(t) \leq \phi\left(\frac{1-\eta}{\alpha(y-c)}\right) + \frac{\eta}{\alpha(y-c)} \phi'\left(\frac{1-\eta}{\alpha(y-c)}\right) .$$

At this step we only need to upper bound $\phi'\left(\frac{1-\eta}{\alpha(y-c)}\right)$ by a constant that would not depend on the values of $x_1, \dots, x_n$ and y. To do so, we use that all variables are bounded, and for any $t \in \left[\frac{1-\eta}{(y-c)\alpha}, \frac{1}{(y-c)\alpha}\right)$,

$$\begin{aligned}\phi'(t) &= -\mathbb{E}_{\widehat{F}}\left[\frac{(y-c)\alpha - (y-X)^+}{1-t\left[(y-c)\alpha-(y-X)^+\right]}\right] \\ &\leq -\mathbb{E}_{\widehat{F}}\left[\frac{(y-c)\alpha - y}{1-t\left[(y-c)\alpha-y\right]}\right] \\ &= \frac{(1-\alpha)y+\alpha c}{1-t(1-\alpha)y-t\alpha c}\ .\end{aligned}$$

We then replace t by $\frac{1-\eta}{(y-c)\alpha}$, which gives

$$\frac{\eta}{\alpha(y-c)}\phi'\left(\frac{1-\eta}{(y-c)\alpha}\right) \leq \eta\frac{(1-\alpha)y+\alpha c}{\eta\alpha(y-c)+(1-\eta)y}\ .$$

Then, we use that $(1-\alpha)y+\alpha c \leq B$, and that $\eta\alpha(y-c)+(1-\eta)y \geq (1-\eta)c \geq c$, so finally

$$\frac{\eta}{\alpha(y-c)}\phi'\left(\frac{1-\eta}{(y-c)\alpha}\right) \leq \eta\frac{(1-\alpha)B+\alpha c}{c}\ .$$

Summarizing these steps, we obtain

$$\sup_{t\in\left[0,\frac{1-\eta}{(y-c)\alpha}\right]}\phi(t) \leq \sup_{t\in\left[0,\frac{1}{(y-c)\alpha}\right]}\phi(t) + \eta\frac{(1-\alpha)B+\alpha c}{c}\ .$$

Hence, we finally conclude the proof using Lemma 4, to obtain

$$\mathbb{P}_w\left(C_\alpha(\mathcal{X},w)\geq c\right) \leq \frac{B}{\eta}\exp\left(-n\left(\mathcal{K}_{\inf}^{\alpha,\mathcal{X}}(\widehat{F},c) - \eta\frac{(1-\alpha)B+\alpha c}{c}\right)\right)\ .$$

□

E.5. Proof of Lemma 5

In this section we prove Lemma 5 introduced in Section 3.

Lemma 5. *Assume that $\mathcal{X} = (x_1, \dots, x_n)$ and $x_1 < \dots < x_n$, then $x_{\lceil n\alpha\rceil}$ is the empirical α quantile of the set and x_1 its minimum, and it holds that*

$$\mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X},w)\geq C_\alpha(\mathcal{X})\right) \geq \frac{1}{25n^3}(x_{\lceil n\alpha\rceil}-x_1)\ .$$

Proof. We assume that $\mathcal{X}$ is known and ordered, i.e $x_1 \leq x_2 \leq \dots \leq x_n$. We then write

$$A = \mathbb{P}_{w\sim\mathcal{D}_n}\left(C_\alpha(\mathcal{X},w)\geq C_\alpha(\mathcal{X})\right)\ .$$

Thanks to the definition of the CVaR provided by Equation (1) it holds that

$$A = \mathbb{P}_w\left(\sup_{y\in\mathcal{X}}\{y-\frac{1}{\alpha}\sum_{i=1}^n w_i(y-x_i)^+\} \geq \sup_{z\in\mathcal{X}}\{z-\frac{1}{\alpha n}\sum_{i=1}^n(z-x_i)^+\}\right)\ .$$

First, if we know $x_1, \dots, x_n$ then the second term is deterministic and the sup is actually achieved in $x_{\lceil n\alpha\rceil}$. Secondly, the inequality is true if at least one term in the left element satisfies it, so we can write

$$
\begin{aligned}
A =& \mathbb{P}\left(\sup_{z\in\mathcal{X}}\left\{z - \frac{1}{\alpha}\sum_{i=1}^{n} w_i(z-x_i)^+\right\} \geq x_{\lceil n\alpha\rceil} - \frac{1}{\alpha n}\sum_{i=1}^{n}(x_{\lceil n\alpha\rceil}-x_i)^+\right)\\
\geq& \mathbb{P}\left(x_{\lceil n\alpha\rceil} - \frac{1}{\alpha}\sum_{i=1}^{n} w_i(x_{\lceil n\alpha\rceil}-x_i)^+ \geq x_{\lceil n\alpha\rceil} - \frac{1}{\alpha n}\sum_{i=1}^{n}(x_{\lceil n\alpha\rceil}-x_i)^+\right)\\
=& \mathbb{P}\left(\sum_{i=1}^{n} w_i(x_{\lceil n\alpha\rceil}-x_i)^+ \leq \frac{1}{n}\sum_{i=1}^{n}(x_{\lceil n\alpha\rceil}-x_i)^+\right)\\
=& \mathbb{P}\left(\sum_{i=1}^{n} w_i\frac{B-(x_{\lceil n\alpha\rceil}-x_i)^+}{B} \geq \frac{1}{n}\sum_{i=1}^{n}\frac{B-(x_{\lceil n\alpha\rceil}-x_i)^+}{B}\right).
\end{aligned}
$$

As the variable $\frac{B-(x_{\lceil n\alpha\rceil}-x_i)^+}{B}$ belongs to $[0,1]$ we can apply the lemma 17 of Riou & Honda and get

$$
A \geq \frac{1}{25n^2B}\left(B - \frac{1}{n}\sum_{i=1}^{n}(B-(x_{\lceil n\alpha\rceil}-x_i)^+)\right) = \frac{1}{25n^3B}\sum_{i=1}^{n}(x_{\lceil n\alpha\rceil}-x_i)^+ .
$$

We conclude by simply omitting all the terms except $(x_{\lceil n\alpha\rceil} - x_1)$ in the sum. □

F. Brown-UCB a.k.a U-UCB

In this section, we present the instanciation of the U-UCB algorithm of (Cassel et al., 2018) for CVaR bandits, and discuss its links with the Brown-UCB idea proposed by (Tamkin et al., 2020), which propose to build a UCB strategy based on concentration inequalities proposed by (Brown, 2007).

F.1. Explicit form of U-UCB

The U-UCB bonus is written as $f\left(\frac{C\log t}{N_k(t-1)}\right)$ for some constant C and some function f defined as

$$
f(x) = \max\left\{2b\left(\frac{x}{a}\right)^{1/2}, 2b\left(\frac{x}{a}\right)^{q/2}\right\}.
$$

Following the Definition 3 in (Cassel et al., 2018) we can find the values of the constants a, b and q. The DKW inequality gives $a = 1$, while b and q are found as the smallest parameters satisfying the following inequality:

$$
|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)| \leq b(||F-G||_\infty + ||F-G||_\infty^q)
$$

If the distributions are upper bounded by some constant U then we have from (Tamkin et al., 2020) that it is sufficient to choose $b = \frac{U}{2\alpha}$ and $q = 1$. This yields the following explicit form for the U-UCB stratey:

$$
A_{t+1}^{\text{U-UCB}} = \operatorname*{argmax}_{k\in[K]}\left[\mathrm{CVaR}_\alpha(\widehat{\nu}_k) + \frac{U}{\alpha}\sqrt{\frac{C\log t}{2N_k(t)}}\right].
$$

In our experimental study, with use the constant $C = 2$ in the index of U-UCB (and $U = 1$ as we consider distributions that are bounded in $[0,1]$). This choice is motivated by the fact that (Cassel et al., 2018) show that for $C > 2$, U-UCB has a logarithmic *proxy regret*. As explained by (Tamkin et al., 2020), the proxy regret is an upper bound on the CVaR regret, hence U-UCB is guaranteed to have logarithmic CVaR regret in our setting.

Interestingly, by following an approach suggest by (Tamkin et al., 2020), we can recover the exact same algorithm as U-UCB, an propose a simple analysis of this algorithm directly in terms of CVaR regret.

F.2. Brown-UCB and its analysis

The authors of (Tamkin et al., 2020) propose to build on concentration inequalities for the empirical CVaR given by (Brown, 2007) to derive a UCB strategy in which the index of each arm adds a confidence bonus to the CVaR of its empirical distribution. However, their derivation of this Brown-UCB algorithm is not correct as they use the concentration inequalities originally given by (Brown, 2007) for the *loss* version of the CVaR. We propose a fix in this section, which consists in adapting the Brown inequalities to the *reward* version of the CVaR which we consider in this paper.

For bounded distributions there is a clear symmetry between the the two definitions of CVaR. In particular, for any distribution ν supported in $[0, B]$

$$\mathrm{CVaR}_\alpha(\nu) = B - \mathrm{CVaR}_\alpha^{\mathrm{loss}}(B - \nu) \ ,$$

where $1 - \nu$ denotes the distribution of $1 - X$ with $X \sim \nu$ and we write respectively CVaR and $\mathrm{CVaR}^{\mathrm{loss}}$ the reward and loss version of CVaR.

Proof.

$$\begin{aligned}
\mathrm{CVaR}_\alpha(\nu) &= \sup_{x\in[0,B]} \left\{ x - \frac{1}{\alpha}\mathbb{E}\left((x - X)^+\right) \right\} \\
&= \sup_{x\in[0,B]} \left\{ x - \frac{1}{\alpha}\mathbb{E}\left((B - X - (B - x))^+\right) \right\} \\
&= \sup_{y\in[0,1]} \left\{ B - y - \frac{1}{\alpha}\mathbb{E}\left((B - X - y)^+\right) \right\} \\
&= B - \inf_{y\in[0,1]} \left\{ y + \frac{1}{\alpha}\mathbb{E}\left((B - X - y)^+\right) \right\} \\
&= 1 - \mathrm{CVaR}_\alpha^{\mathrm{loss}}(B - \nu)
\end{aligned}$$

□

Remark 4. *Applying the same trick to $Y = -X$ provide that for a real random variable X then $\mathrm{CVaR}_\alpha(X) = -\mathrm{CVaR}_\alpha^{loss}(-X)$.*

This observation easily yield the following concentration inequalities, which are the counterpart of the Brown inequalities for the reward version of the CVaR.

Lemma 9 (Brown's inequalities for CVaR_α)**.** *If we write $\widehat{c}_n^\alpha$ the CVAR of an empirical distribution from n variables drawn from a distribution ν supported in $[0, B]$ and $\mathrm{CVAR}_\alpha(\nu) = c^\alpha$, we have:*

$$\mathbb{P}(\widehat{c}_n^\alpha \geq c^\alpha + \varepsilon) \leq 3\exp\left(-\frac{\alpha}{5}\left(\frac{\varepsilon}{B}\right)^2 n\right)$$
$$\mathbb{P}(\widehat{c}_n^\alpha \leq c^\alpha - \varepsilon) \leq \exp\left(-2\left(\frac{\alpha\,\varepsilon}{B}\right)^2 n\right)$$

We note that the upper and lower deviation have their probability bounded by a term whose scaling in α is different. By interverting the two inequalities, (Tamkin et al., 2020) proposed a "Brown-UCB" algorithm with an confidence bonus scaling in $1/\sqrt{\alpha}$ instead of $1/\alpha$ and obtained a regret bound that was actually contradicting the lower bound of Theorem 1.

Expression of Brown-UCB The inequalities in Lemma 9 permit to propose a UCB algorithm of the form

$$A_{t+1}^{\text{Brown-UCB}} = \operatorname*{argmax}_{k\in[K]} \mathrm{UCB}_k(N_k(t), t)$$

where $\mathrm{UCB}_k(n, t) = \widehat{c}_{k,n}^\alpha + \frac{U}{\alpha}\sqrt{\frac{f(t)}{2n}}$, where $\widehat{c}_{k,n}^\alpha$ is the CVaR of level α of the empirical distribution of the n first observations from arm k and $f(t)$ is an increasing function of t that will be specified later in the analysis. Indeed, one can easily check that

$$\mathbb{P}(\mathrm{UCB}_k(t, n) \leq c^\alpha) = \mathbb{P}\left(\widehat{c}_{k,n}^\alpha \leq c_k^\alpha - \frac{U}{\alpha}\sqrt{\frac{f(t)}{2n}}\right) \leq e^{-f(t)},$$

which justifies that fact that $\mathrm{UCB}_k(t,n)$ is an upper confidence bound on the CVaR of arm k.

Interestingly, we observe that for the choice $f(t) = C\log(t)$, Brown-UCB coincides with the U-UCB algorithm. We upper bound below the CVaR regret of Brown-UCB (or U-UCB) for $C > 2$, and recover that this regret is indeed logarithmic.

We analyze Brown-UCB for distributions supported in $[0, U]$ and a threshold function $f(t) = (2+\varepsilon)\log(t)$ for some $\varepsilon > 0$. For every sub-optimal arm k, we start with the classical decomposition

$$\begin{aligned}\mathbb{E}\left[N_k(T)\right] &= \sum_{t=0}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k)] \\ &= 1 + \sum_{t=K}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k, \mathrm{UCB}_1(N_1(t), t) \le c_1^\alpha)] + \sum_{t=K}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k, \mathrm{UCB}_k(N_k(t), t) \ge c_1^\alpha)] \ .\end{aligned}$$

We analyze separately these two terms. We use a union bound on the values of $N_1(t)$ and the second inequality in Lemma 9 to handle the first term:

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k, \mathrm{UCB}_1(t) \le c_1^\alpha)] &\le \sum_{t=K}^{T} \sum_{n=1}^{t} \mathbb{P}\left(N_1(t) = n, \mathrm{UCB}_1(n,t) \le c_1^\alpha\right) \\ &\le \sum_{t=1}^{T} \sum_{n=1}^{t} \mathbb{P}\left(\widehat{c}_{1,n}^\alpha \le c_1^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(t)}{2n}}\right) \\ &\le \sum_{t=1}^{T} t\exp(-f(t)) \ .\end{aligned}$$

With the choice $f(t) = (2+\varepsilon)\log(t)$, we get $\sum_{t=K}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k, \mathrm{UCB}_1(t) \le c_1^\alpha)] = O(1)$.

To handle the second term, we write the following:

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{E}[\mathbb{1}(A_{t+1} = k, \mathrm{UCB}_k(t) \ge c_1^\alpha)] &\le \sum_{t=K}^{T-1} \sum_{n=1}^{t} \mathbb{E}[\mathbb{1}(A_{t+1} = k, N_k(t) = n, \mathrm{UCB}_k(N_k(t), t) \ge c_1^\alpha)] \\ &\le \sum_{t=K}^{T-1} \sum_{n=1}^{t} \mathbb{E}\left[\mathbb{1}\left(A_{t+1} = k, N_k(t) = n, \widehat{c}_{k,n}^\alpha \ge c_1^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(t)}{2n}}\right)\right] \\ &\le \sum_{t=K}^{T-1} \sum_{n=1}^{t} \mathbb{E}\left[\mathbb{1}\left(A_{t+1} = k, N_k(t) = n, \widehat{c}_{k,n}^\alpha \ge c_1^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(T)}{2n}}\right)\right] \\ &\le \sum_{n=1}^{T} \mathbb{E}\left[\mathbb{1}\left(\widehat{c}_{k,n}^\alpha \ge c_1^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(T)}{2n}}\right) \sum_{t=K}^{T-1} \mathbb{1}\left(A_{t+1} = k, N_k(t) = n\right)\right] \\ &\le \sum_{n=1}^{T} \mathbb{E}\left[\mathbb{1}\left(\widehat{c}_{k,n}^\alpha \ge c_1^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(T)}{2n}}\right)\right] \\ &= \sum_{n=1}^{T} \mathbb{P}\left(\widehat{c}_{k,n}^\alpha \ge c_k^\alpha + \left(\Delta_k^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(T)}{2n}}\right)\right)\end{aligned}$$

Let $\beta > 0$ to be chosen later. Letting $n_0(T) = \left\lceil \frac{f(T)}{2\alpha^2(1-\beta)^2(\Delta_k^\alpha)^2} \right\rceil$, we have that for all $n \ge n_0(T)$,

$$\left(\Delta_k^\alpha - \frac{1}{\alpha}\sqrt{\frac{f(T)}{2n}}\right) \ge \beta\Delta_k^\alpha.$$

Therefore, using the first inequality in Lemma 9,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}(A_{t+1}=k,\mathrm{UCB}_k(t)\geq c_1^{\alpha})] &\leq n_0(T)+\sum_{n=n_0(T)+1}^{T}\mathbb{P}\left(\widehat{c}_{k,n}^{\alpha}\geq c_k^{\alpha}+\beta\Delta_k^{\alpha}\right)\\ &\leq n_0(T)+\sum_{n=1}^{T}3\exp\left(-\frac{\alpha}{5}\beta^2(\Delta_k^{\alpha})^2 n\right)\\ &= n_0(T)+O(1).\end{aligned}$$

Choosing for example $\beta = 1-\sqrt{1/2}$ yields that Brown-UCB with $f(t) = (2+\varepsilon)\log(T)$ satisfies

$$\mathbb{E}[N_k(T)] \leq \frac{(2+\varepsilon)\log(T)}{\alpha^2(\Delta_k^{\alpha})^2}+O_{\varepsilon}(1).$$

This permits to prove that Brown-UCB has a regret of the same order of magnitude as the CVaR-UCB algorithm proposed by (Tamkin et al., 2020).

G. Complementary Experiments

In this section we provide a more complete overview of the results of our experiments introduced but not detailed in Section 4. The first part of this section contains a comprehensive set of experiments on synthetic data in order to illustrate particular properties of the M-CVTS and B-CVTS algorithms. The second part provides additional experiments performed with the DSSAT crop-simulator, that complete the first set of experiments provided in Section 4.

G.1. Experiments on synthetic examples

Before testing the algorithms on a real-world use-case we performed experiments on simulated data in order to illustrate their empirical properties compared to existing UCB-like algorithms. For all the experiments we generally consider α ranging in $\{10\%, 50\%, 90\%\}$.

G.1.1. Experiments on Multinomial arms

We first introduce some experiments with multinomial arms in order to check the empirical performance of the M-CVTS algorithm. We ackowledge that M-CVTS has some advantage over its competitors as it is aware of the full support of the multinomial distribution while the UCBs only know an upper bound. For this reason we do not comment extensively on the performance gaps between the algorithms, but we are more interested in checking the *asymptotic optimality* of M-CVTS. Indeed, for multinomial distribution we implemented the lower bound described in Section 3 and illustrated it in Figures 7 and 8 for one of our experiments.

Multinomial Experiments 1 to 4: Choice of the distributions We run M-CVTS on different multinomial bandit problems in which all arms have the common support $(x_0, x_1, \ldots, x_{10}) = \{0, 0.1, \ldots, 0.9, 1\}$. In this setting we consider 5 multinomial distributions, that we write $(q_i)_{i\in\{1,\ldots,5\}}$, and visually represent in Figure 3. Those arms provide different distributions with interesting shapes and properties, using simple formulas to generate the probabilities.

As the order of arms' CVaR varies substantially depending on the value of α, a bandit algorithm aiming at minimizing the CVaR regret is necessary. For instance q_1 is the best arm for $\alpha \leq 20\%$, q_3 for $\alpha \in [30\%, 55\%]$, and q_5 for $\alpha \geq 55\%$. Furthermore, q_5 is typically a distribution that a *risk-averse* practitioner would like to avoid as its expectation is large at the cost of potential high losses, while q_1 is not satisfying for someone maximizing the expected reward due to the high concentration around 0.5. Interestingly, despite different shapes q_2 and q_4 are actually close in terms of CVaR, hence a bandit problem defined over these two distributions only is hard to solve. We illustrate the CVaRs of these different arms as a function of α in Figure 4.

We implemented four experiments with different subsets of these arms, for α in $\{10\%, 50\%, 90\%\}$:

- Experiment 1, $Q_1 = [q_1, q_2, q_3, q_4, q_5]$

- Experiment 2 (Hard for risk-averse learner), $Q_1 = [q_4, q_5]$
- Experiment 3 (Large gaps), $Q_1 = [q_1, q_2, q_3]$
- Experiment 4 (Small gaps), $Q_1 = [q_2, q_4]$

Experimental setup Each experiment consists in $N = 5000$ runs of each algorithm (namely U-UCB, CVaR-UCB and M-CVTS) up to a horizon $T = 10000$. We report the results in Tables 5, 6, 7 and 8.

Analysis of the results The results reveal that M-CVTS clearly outperforms the baselines for any level of α in all experiments. Experiment 4 is the only one for which the gap is not very large, because this CVaR bandit problem is a hard instance, and no algorithm reaches its asymptotic regime after 10^4 time steps. However, it is interesting to notice that M-CVTS can be better than the baselines even when it is not in this asymptotic regime.

For Experiment 1 and $\alpha \in \{10\%, 90\%\}$ we display the regret curves of both M-CVTS, CVaR-UCB and U-UCB in Figure 5 and Figure 6. We also add a $5\% - 95\%$ confidence bound around each curve. We see that the regret of U-UCB is linear for this time horizon with $\alpha = 10\%$, while the regrets of CVaR-UCB and M-CVTS have similar shapes and confidence intervals for the two values of α, hence they appear to be more robust to the parameter α than U-UCB in this setting. Nonetheless, in both cases M-CVTS largely outperforms CVaR-UCB. It is also interesting to remark that M-CVTS becomes clearly better than its competitors very early in the competition, which shows that M-CVTS can be a good choice for practitioners who would consider shorter horizons.

Optimality of M-CVTS Still on Experiment 1, we illustrate in Figures 7 and 8 the *asymptotic optimality* of M-CVTS by representing its regret (in logarithmic scale on the x axis) along with the asymptotic lower bound described in Section 3, again for $\alpha \in \{10\%, 90\%\}$. The fact that M-CVTS matches the asymptotic lower bound is verified in this experiment, as the regret of M-CVTS converges to a straight line which is parallel to the lower bound (still in logarithmic scale on the x axis). The small difference of slopes in Figure 7 might be due to the fact that we used a solver to solve the optimization problem involved in the lower bound, and we noticed that this solver was less precise for small values of α.

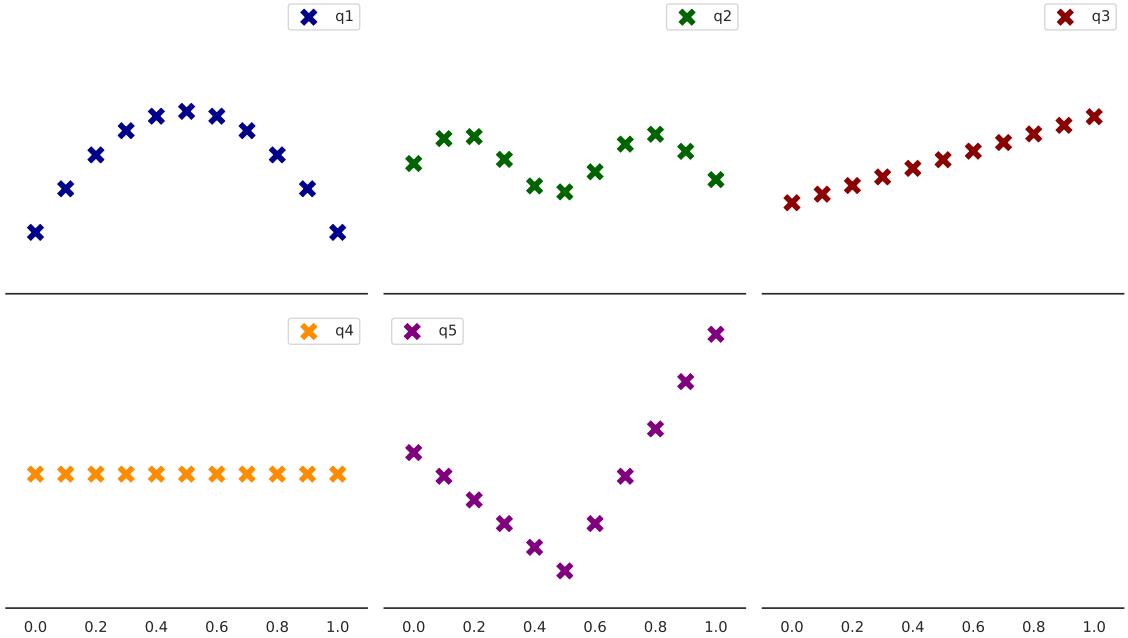


Figure 3: Visual representation of multinomial distributions q_1, q_2, q_3, q_4, q_5

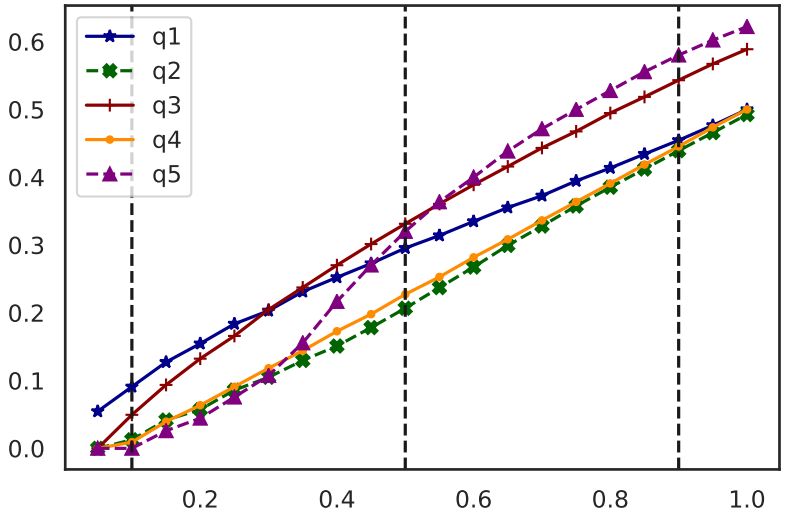


Figure 4: CVaR of the distributions $(q_i)_{i\in\{1,\dots,5\}}$ as a function of α vertical lines are the thresholds $\alpha = \{5\%, 50\%, 90\%\}$.

Table 5: Results for Multinomial Experiment 1 at $T = 10000$ for 5000 replications. Standard deviations in parenthesis. $Q = [q_1, q_2, q_3, q_4, q_5]$

α	U-UCB	CVaR-UCB	M-CVTS
10%	549.4 (3.3)	235.9 (19.6)	**35.1 (14.7)**
50%	283.6 (16.3)	181.5 (17.9)	**65.4 (30.6)**
90%	221.1 (23.7)	220.5 (23.7)	**43.7 (36.3)**

Table 6: Results for Multinomial Experiment 2 at $T = 10000$ for 5000 replications. for 5000 replications. $Q = [q_4, q_5]$

α	U-UCB	CVaR-UCB	M-CVTS
10%	44.5 (0.4)	26.7 (5.0)	**11.9 (7.8)**
50%	137.5 (18.9)	55.4 (13.6)	**17.7 (23.8)**
90%	53.3 (11.0)	54.3 (11.4)	**8.0 (5.6)**

Table 7: Results for Multinomial Experiment 3 at $T = 10000$ for 5000 replications. $Q = [q_1, q_2, q_3]$

α	U-UCB	CVaR-UCB	M-CVTS
10%	360.1 (3.9)	149 (16.8)	**23.0 (13.8)**
50%	217.2 (17)	117.6 (18.7)	**29.0 (25.6)**
90%	124.2 (16.5)	116.6 (16.0)	**17.3 (10.8)**

Table 8: Results for Multinomial Experiment 4 at $T = 10000$ for 5000 replications. $Q = [q_2, q_4]$

α	U-UCB	CVaR-UCB	M-CVTS
10%	17.7 (0.2)	16.4 (2.2)	**13.6 (8.6)**
50%	79 (7.8)	68.3 (14.4)	**27.8 (26.4)**
90%	27.1 (4.4)	26.0 (4.3)	**21.3 (15.6)**

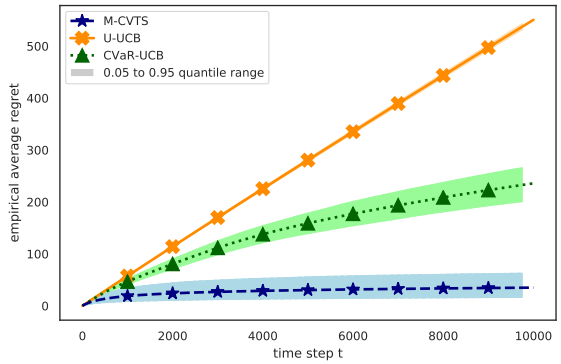


Figure 5: Experiment 1 with Multinomial arms, all algorithms, $\alpha = 10\%$

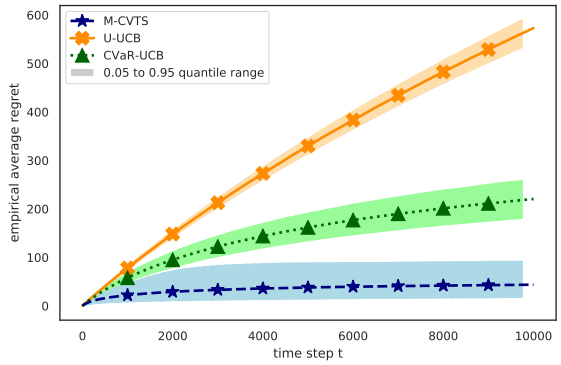


Figure 6: Experiment 1 with Multinomial arms, all algorithms, $\alpha = 90\%$

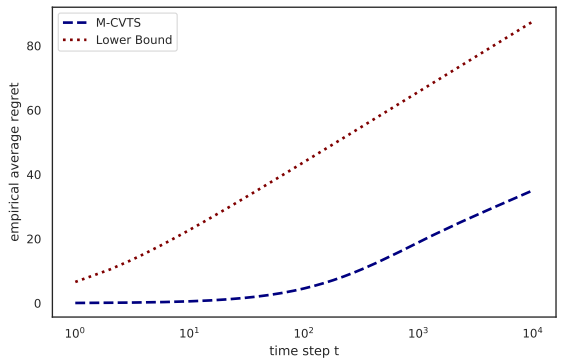


Figure 7: Experiment 1 with Multinomial arms, regret of M-CVTS and lower bound (abscissa log scale), $\alpha = 10\%$

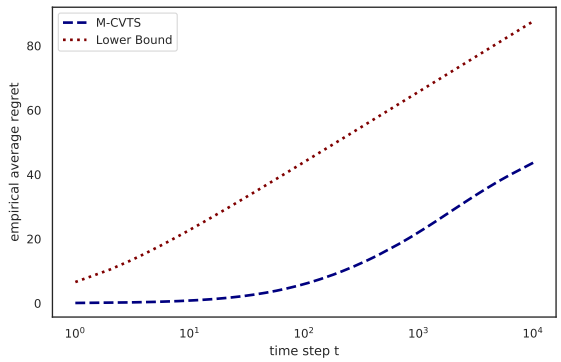


Figure 8: Experiment 1 with Multinomial arms, regret of M-CVTS and lower bound (abscissa log scale) $\alpha = 90\%$

G.1.2. Experiments on general Bounded distributions: Truncated Gaussian Mixtures

In this section we consider bounded multi-modal distributions, built by truncated Gaussian Mixture models in $[0, 1]$. We simply call these distributions Truncated Gaussian Mixtures (TGM for short). We first remark that these distributions are not continuous because they can have a positive mass in 0 and 1, but it is still a good illustrative example to check the performance of B-CVTS. Indeed, we also performed the same exact experiments making the distributions continuous (instead of truncating, we re-sampled observations until they lied in $[0, 1]$) and the results were deemed to be exactly the same.

Continuous Experiments 1 to 4: Bi-modal Gaussian mixtures We first consider experiments with two modes, each mode being equiprobable and having the same variance for simplicity ($\sigma = 0.1$) in all experiments. It is interesting to compare settings where some arms have modes that are both close to 0.5, and where other arms have a large mass of probability close to the two support bounds (one mode close to 1 and one close to 0).

We experiment 4 possible configurations of the modes, given by:

- $\mu_1 = (0.2, 0.5)$
- $\mu_2 = (0, 1)$

- $\mu_3 = (0.3, 0.6)$
- $\mu_4 = (0.1, 0.65)$

We indifferently call the arms by their means (saying arm μ_1 for the TGM arm with the parameter μ_1 along with the parameters we fixed).

As for the discrete setup in the previous section, we highlight some basic properties of their CVaRs: the distribution with parameter μ_2 has a larger mean than the one with μ_1, but the 50% CVaR of μ_1 is larger. We represented the CVaR for each parameter for different values of $\alpha \in (0, 1]$ in Figure 9, with the thresholds $\alpha \in \{0\%, 10\%, 90\%\}$ represented by the vertical lines. Interestingly, with these arms the most difficult problems are not necessarily those with smallest values of α. Indeed, for $\alpha = 80\%$ it may be particularly difficult to choose between μ_2 and μ_3, or between μ_1 and μ_4, while μ_3 is the clear winner for $\alpha = 10\%$ due to the distribution being very concentrated around 0.5. Furthermore, the distribution μ_2 is very concentrated around the bounds of the support but has a larger mean than the others, hence it becomes the best arm for values of α that are close to 1.

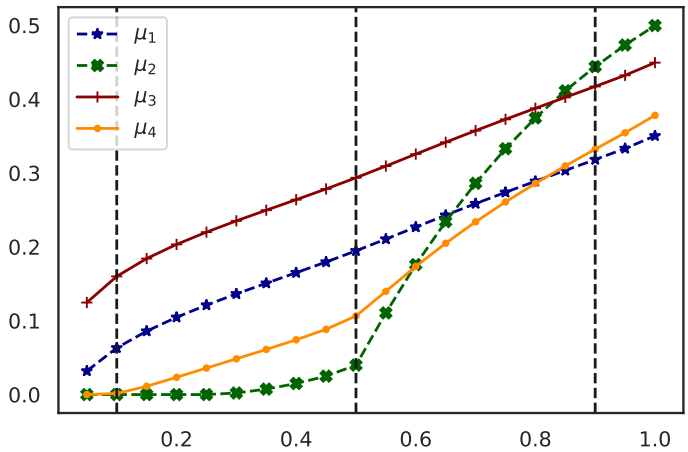


Figure 9: CVaR of each TGM distribution ν_i (with centers μ_i), $i = 1, \dots, 4$ for different values of α

We run the algorithms for $\alpha = 10\%, 50\%$ and 90% on four bandit problems with the following respective distributions:

- $\nu_1 = (\mu_1, \mu_2)$
- $\nu_2 = (\mu_1, \mu_3)$
- $\nu_3 = (\mu_1, \mu_4)$
- $\nu_4 = (\mu_i)_{i \in \{1,2,3,4\}}$

Results In Tables 9, 10, 11 and 12 we report the results for the four considered problems (mean regret and standard deviation at $T = 10000$). We also provide the regret curves, as for multinomial distributions, in order to check the logarithmic order of the regret of B-CVTS and its rate when T is large.

Again, the TS approach significantly outperforms the two UCB algorithms, which is a very interesting result: contrarily to the multinomial case, this time the three algorithms had the same level of information on arms' distributions. B-CVTS is consistently the best for all four problems we implemented and for all α levels.

Continuous Experiment 5: Robustness to small α We then check the robustness of B-CVTS to a smaller value of the parameter α by setting $\alpha = 1\%$, referred as Experiment 5. The bandit of Experiment 5 has six TGM arms with respective

Table 9: Results for TGM Experiment 1 at $T = 10000$ for 5000 replications. Standard deviations in parenthesis.

α	U-UCB	CVaR-UCB	B-CVTS
10%	274.9 (1.8)	5.3 (1.5)	**1.1 (0.5)**
50%	127.0 (19.3)	135.3 (41.1)	**29.8 (17.2)**
90%	80.5 (10.4)	53.5 (6.7)	**10.2 (17.9)**

Table 10: Results for TGM Experiment 2 at $T = 10000$ for 5000 replications.

α	U-UCB	CVaR-UCB	B-CVTS
10%	373.7 (4.1)	72.8 (9.6)	**4.1 (2.3)**
50%	135.8 (8.9)	37.9 (7.5)	**5.5 (2.7)**
90%	62.6 (7.1)	43.9 (5.1)	**5.0 (1.8)**

Table 11: Results for TGM Experiment 3 at $T = 10000$ for 5000 replications.

α	U-UCB	CVaR-UCB	B-CVTS
10%	269.4 (1.8)	23.2 (4.8)	**2.8 (1.5)**
50%	138.5 (12.4)	71.8 (19.0)	**14.7 (8.3)**
90%	53.1 (6.6)	34.5 (6.6)	**20.2 (22.4)**

Table 12: Results for TGM Experiment 4 at $T = 10000$ for 5000 replications.

α	U-UCB	CVaR-UCB	B-CVTS
10%	958.9 (4.8)	230.5 (25.3)	**10.4 (3.2)**
50%	318.4 (12.2)	147.7 (17.9)	**21.2 (6.4)**
90%	154.3 (11.9)	119.5 (11.7)	**25.1 (14.1)**

mean and variance parameters $\mu_{135} = (0.3, 0.6)$, $\mu_{246} = (0.25, 0.65)$, $\sigma_{12} = 0.05$, $\sigma_{34} = 0.06$, $\sigma_{56} = 0.07$. This experiment allows to additionally check if adding different variances to the arms affects the performance of the algorithms. However, we keep the probability of each mode to 0.5. This problem provides the following CVaR values for each arm at level 1%, respectively: $c_{1:6}^{0.01} = [0.18, 0.13, 0.15, 0.10, 0.13, 0.08]$. The results are reported in Table 13, in which we observe a very large performance gap between B-CVTS and UCB algorithms. This is particularly interesting because it shows that the UCB algorithms are not really able to learn for very small values of α (indeed $\alpha = 1\%$ is very small when drawing only a total number of 10^4 observations) before the horizon becomes extremely large. We already observed this behavior for CVaR-UCB in previous experiments, but this time we can see as well that its average regret is even higher than the one of U-UCB, and its variance spiked. On the other hand, B-CVTS seems to learn smoothly even for $\alpha = 1\%$, as its average regret only doubles between $T = 1000$ and $T = 5000$, and increases even less between $T = 5000$ and $T = 10000$.

Table 13: Results for TGM Experiment 5 ($\alpha = 1\%$) at $T = 10000$ for 5000 replications.

T	U-UCB	CVaR-UCB	B-CVTS
1000	49.1 (0.3)	53.2 (5.6)	**18 (37)**
5000	245 (1.1)	263.2 (24.7)	**35.5 (51)**
10000	489.1 (2.2)	518.4 (45.0)	**41 (66)**

Table 14: Results for TGM Experiment 6, at $T = 10000$ averaged over 400 random instances with $K = 30$ truncated Gaussian mixtures with 10 modes.

T	U-UCB	CVaR-UCB	B-CVTS
10000	2149.9 (263)	2016.0 (265)	**210.9 (6.4)**
20000	4276.4 (538)	3781.3 (521)	**237.1 (15.4)**
40000	8493.4 (1085)	6894.1 (985)	**263.5 (17.9)**

Continuous Experiment 6: Random Problems with more modes and more arms Finally, we further check the robustness of B-CVTS to more arms and more diverse distribution profiles by increasing the number of possible modes.

To do so, we implement an experiment with $K = 30$ arms, with TGM distributions with 10 modes exhibiting different means and variances, which covers a large variety of shapes of distributions. All of those parameters are drawn uniformly at random, and we summarize their distributions as $(\mu, \sigma) \sim \mathcal{U}([0.25, 1]^{10} \times [0, 0.1]^{10})$, and $p \sim \mathcal{D}_{10}$ (uniform distribution on the simplex, presented in Section 3). We name this setting TGM Experiment 6. The results of this experiment are reported in Table 14 for a parameter $\alpha = 0.05$ averaged over 400 random instances. Again, we choose a smaller value for α than in the previous extensive sets of experiments because problems with small α seem to be more challenging. The results highlight that best performances are obtained by B-CVTS.

Conclusions We preliminary evaluated the CVaR bandit algorithms on synthetic problems before testing them on realistic-world bandit environment in the next section. These experiments seem to highlight a greater robustness of B-CVTS to many different settings regarding different parameters: α level, the number of arms K and the different possible shapes of the distributions (symbolized by the number of modes in our synthetic experiments). In particular, B-CVTS is the only algorithm that has not shown to be affected by the value of α, as the two UCB algorithms had their respective performances degraded in some extent depending on α values.

G.2. Experiments with `DSSAT` crop-model

In this section we keep comparing B-CVTS with U-UCB and CVaR-UCB for $\alpha \in \{5\%, 10\%, 80\%\}$ as described in section 4. `DSSAT` is still parameterized with the same challenging conditions, but we generate two different problems thanks to the crop-simulator. For both presented experiments we consider $N = 1040$ runs of each algorithm up to a time horizon $T = 10000$. As explained in section 4, all `DSSAT` arms' distributions are empirically estimated from 10^6 samples in both experiments.

`DSSAT` Experiment 1: 7 armed planting date bandit We consider a bandit instance that consists of 7 arms, each arm corresponds to a planting date spaced of 15 days from the previous one. An illustration of the underlying distributions is given in Figure 10. In this case, the best arm is consistent with all values of α, as shown in Table 15. Nevertheless, arms exhibit different gaps when considering different values of α. This experiment intends to evaluate B-CVTS robustness for a greater number of real-world alike arms with a diversity of reward distribution shapes.

The results of this experiment are reported in Table 17. The regret curves for the three algorithms, with considered values of α parameter are illustrated in Figures 12, 14, and 16.

In this experiment, by exhibiting superior performances B-CVTS appears to be more robust than the UCB CVaR bandit algorithms relative to an increase in the number of arms. In practice for the planting-date problem, a global, few months planting-window is known but needs further refinements e.g. to identify the best two-week time slot for planting. That is to say, the number of arms is unlikely to be greater that what has been tested in this experiment, making B-CVTS a particularly fit-for-purpose candidate in this setup.

`DSSAT` Experiment 2: Impact of support upper bound over-estimation This configuration is the same than the one presented in Section 4, but here we largely over-estimate the yield upper-bound to 30 t/ha, when a close to reality yield upper bound is about 10 t/ha. From an agronomic point of view, this yield value is a very unlikely over-estimation in the given conditions. This experiment intends to empirically evaluate how a rough arms' upper-bound estimation affects algorithms' performances, when little expert knowledge is available. An illustration of the underlying distributions and how the upper-bound estimation is exaggerated is given in Figure 11 and corresponding metrics are reported in Table 16.

We provide the results of this experiment in Table 18, and display the regret curves in Figures 13, 15, and 17.

Experiment 2 addresses one possible concern for practitioners: the prerequisite of rewards' support upper bound. We empirically demonstrate that with realistic simulations, when a highly over-estimated, unrealistic support upper-bound is given – triple of expert's estimation –, B-CVTS keeps outperforming UCB-like CVaR bandit algorithms. We shown that this over-estimation did not affect B-CVTS performances compared to the situation of correct support upper-bound identification as presented in Section 4. In particular, it even slightly improved its performance for $\alpha = 80\%$. This result is counter-intuitive, but it can be explained by the fact that the extra exploration induced by the larger upper bound may have sped up learning in this particular case, improving overall performances. On the other hand, CVaR-UCB seems much more impacted by this over-estimation (regret is respectively increased by about 150%, 75% and 78% for $\alpha \in \{5\%, 20\%, 80\%\}$). Similarly U-UCB shown altered performances, despite its already unsatisfying results when considering the true upper

bound.

Conclusions B-CVTS appeared to be a satisfying candidate for real-world alike problems, as shown with the planting date bandits. We empirically showed the B-CVTS was best able to deal with a greater number of planting date arms than its UCB counterparts. We showed as well that B-CVTS remained the best performer despite considering a very unlikely support upper-bound estimation. We think that in many physical resource-based problems, this should be reassuring for practitioners, in particular when compared with UCB algorithms' sensibility to the input upper bound.

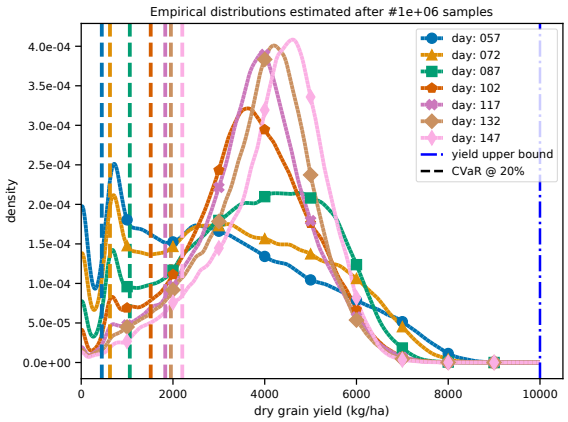


Figure 10: Experiment 1, 7 armed DSSAT environment empirical distributions ; 10^6 samples.

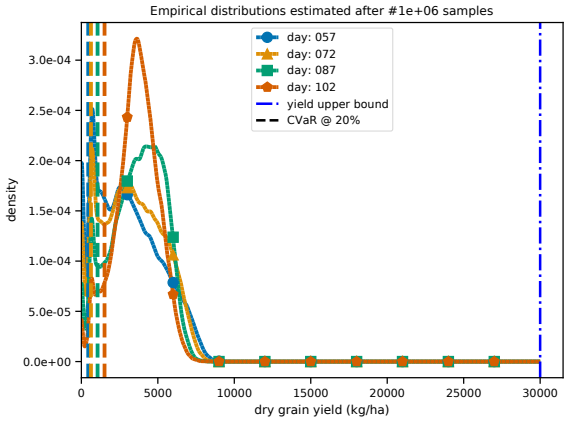


Figure 11: Experiment 2, 4 armed DSSAT environment empirical distributions with over-estimated support ; 10^6 samples.

Table 15: DSSAT Experiment 1 distribution metrics in kg/ha estimated from 10^6 samples..

day (action)	CVaR$_\alpha$			
	5%	20%	80%	100% (mean)
057	0	448	2238	3016
072	46	627	2570	3273
087	287	1059	3074	3629
102	538	1515	3120	3586
117	808	1832	3299	3716
132	929	1955	3464	3850
147	**1122**	**2203**	**3745**	**4112**

Table 16: DSSAT Experiment 2 distribution metrics in kg/ha estimated from 10^6 samples.

day (action)	CVaR$_\alpha$			
	5%	20%	80%	100% (mean)
057	0	448	2238	3016
072	46	627	2570	3273
087	287	1059	3074	**3629**
102	**538**	**1515**	**3120**	3586

Table 17: Results for DSSAT Experiment 1, empirical regret at $T = 10000$ in t/ha for 1040 replications. Standard deviations in parenthesis.

α	U-UCB	CVaR-UCB	B-CVTS
5%	5687 (5)	1891 (18)	**700 (22)**
20%	6445 (10)	1795 (19)	**489 (17)**
80%	3367 (14)	1580 (15)	**293 (8)**

Table 18: Results for DSSAT Experiment 2, empirical regret at $T = 10000$ in t/ha for 1040 replications.

α	U-UCB	CVaR-UCB	B-CVTS
5%	3179 (2)	759 (14)	**195 (11)**
20%	5644 (6)	1020 (17)	**202 (10)**
80%	2642 (10)	888 (13)	**284 (12)**

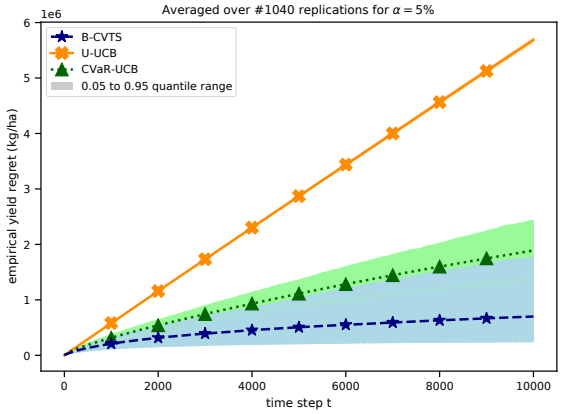


Figure 12: DSSAT Experiment 1, 7 armed bandit all algorithms, $\alpha = 5\%$; 1040 replications.

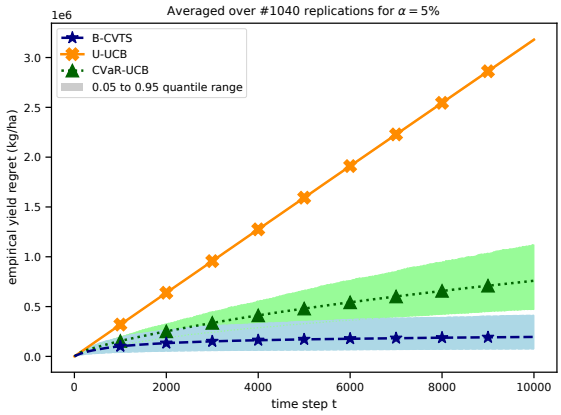


Figure 13: DSSAT Experiment 2, 4 armed over-estimated support upper bound, all algorithms, $\alpha = 5\%$; 1040 replications.

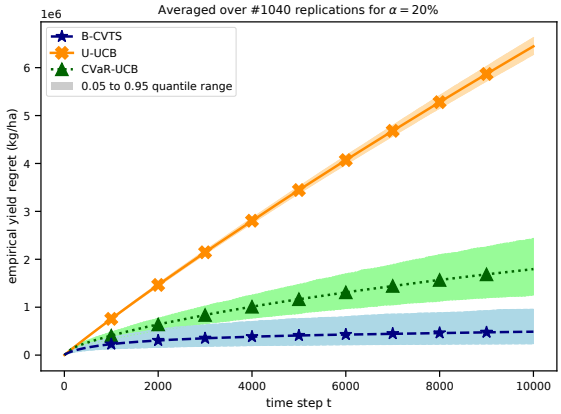


Figure 14: DSSAT Experiment 1, all algorithms, $\alpha = 20\%$

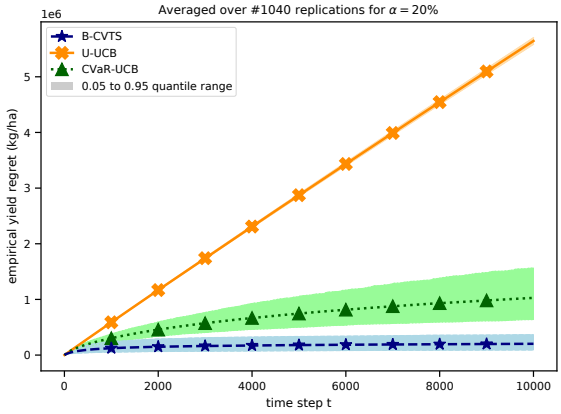


Figure 15: DSSAT Experiment 2, all algorithms, $\alpha = 20\%$

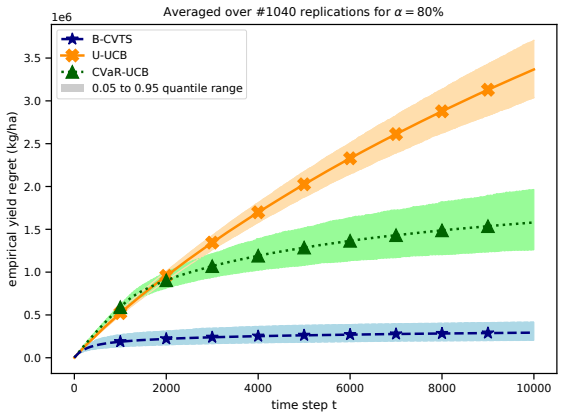


Figure 16: DSSAT Experiment 1, all algorithms, $\alpha = 80\%$

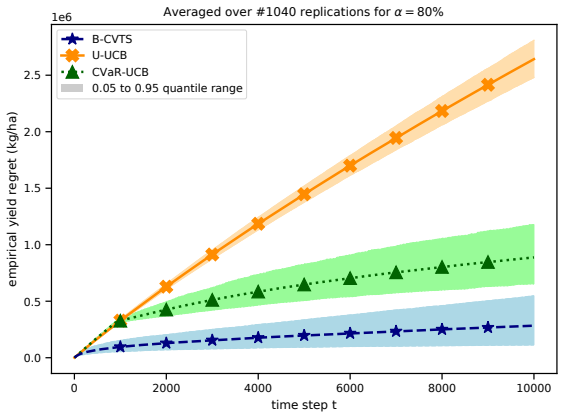


Figure 17: DSSAT Experiment 2, all algorithms, $\alpha = 80\%$