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 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

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 $\text{KL}(\nu, \nu')$ 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*

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

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

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.

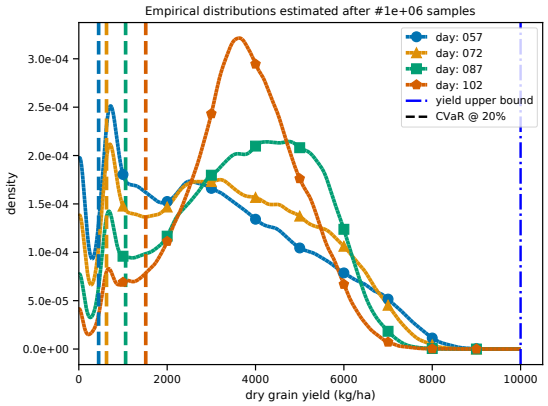

Figure 1: Empirical simulated yields and respective CVaRs at 20% estimated after 10^6 samples in DSSAT environment.

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_α			
	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.

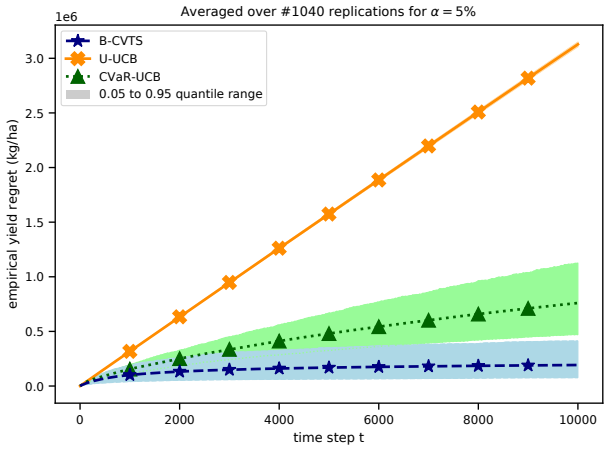

Figure 2: Regret comparison in `DSSAT` environment, averaged over 1040 experiment replications.

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.