Distributionally-Aware Exploration for CVaR Bandits

Alex Tamkin
Department of Computer Science
Stanford University
atamkin@stanford.edu

Ramtin Keramati
Department of Computer Science
Stanford University
keramati@cs.stanford.edu

Christoph Dann
Machine Learning Department
Carnegie Mellon University
cdann@cdann.net

Emma Brunskill
Department of Computer Science
Stanford University
ebrun@cs.stanford.edu

Abstract

Risk sensitive objectives are often desirable in settings like healthcare or finance, where agents are more sensitive to worst-case than average outcomes. However, current risk-sensitive algorithms for multi-armed bandits utilize exploration bonuses which do not adapt to the distributional nature of these objectives. In this paper, we consider multi-armed bandits with a popular risk-sensitive objective called the Conditional Value at Risk (CVaR). We present an novel optimism-based algorithm for this setting: instead of adding bonuses to the CVaR estimate of each arm, we apply optimism *at the sample level*, generating an optimistic set of samples for each arm and then computing CVaR estimates from them instead. We present regret bounds for our algorithm, along with experiments showing order-of-magnitude improvements over baselines and prior work. Further experiments demonstrate how sample-level optimism enables our algorithm to adapt to the shape of each arm distribution in ways that exploration bonuses do not.

1 Introduction

In a K-armed bandit problem, an agent samples from one of K distributions at each turn. Informally, the goal of the agent is to choose from "good" arms as often as possible. This requires the agent to balance exploration of all possible arms with exploitation of the knowledge it has gained from previous samples. This theoretical framework has applications ranging from clinical trial design [1] to financial portfolio optimization [2] to optimizing websites [3].

The multi-armed bandits literature has traditionally used the expected value of an arm's reward distribution as the proxy for the "goodness" of that arm. However, in many interesting cases, it is important to consider the full distributions over the potential rewards, and the desired objective may be a risk-sensitive measure of this distribution. For example, a patient undergoing a surgery for a knee replacement will (hopefully) only experience that procedure once or twice, and may well be interested in the distribution of potential results for a single procedure, rather than what may happen on average if he or she were to undertake that procedure hundreds of time. Finance and (machine) control are other cases where interest in risk-sensitive outcomes are common.

A popular risk-sensitive measure of a distribution of outcomes is the Conditional Value at Risk (CVaR) [4]. Intuitively, CVaR is the expected reward in the worst α-fraction of outcomes, and has seen extensive use in financial portfolio optimization [5, 6], often under the name "expected shortfall." In this paper, we concern ourselves with quickly finding a policy with low CVaR-regret, defined as

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the cumulative difference across timesteps between the CVaR of the optimal arm and the average CVaR of arms chosen by our algorithm.

One sample-efficient principle for exploration is optimism in the face of uncertainty (OFU) where additional bonus terms are added to empirical estimates of the quantity of interest to promote exploration. This principle can perform very well empirically but OFU approaches are known to not adapt to a particular problem structure unless the bonus terms are explicitly designed to account for such structure [7–9].

While existing OFU approaches for risk-sensitive bandits derive bonus terms as upper-confidence bounds on the CVaR of the rewards [10], one might wonder whether tighter bonuses could be obtained by taking the shape of the arm distribution into account. This may be especially relevant for an objective like CVaR which is more sensitive to one side of the distribution than the other; for example, if the distribution of observed samples is skewed towards the minimum observed reward, one might be more confident in the corresponding CVaR estimate than if the samples were skewed in the other direction.

Working towards this goal, we present a method that applies optimism directly at the sample level, generating a new set of "optimistic samples" for each arm *before* computing the sample CVaR from them. This process enables our algorithm to take the shape of the distribution into account, as opposed to count-based reward bonuses added to the sample CVaR, which are agnostic to it. To construct these optimistic samples, we use the Dvoretzky–Kiefer–Wolfowitz (DKW) concentration inequality [11], which provides bounds on the true cumulative distribution function (CDF) given a set of sampled outcomes. These bounds take the form of confidence bands around the empirical distribution function (EDF), and we show how to modify our samples to match the optimistic band.

We will shortly illustrate how DKW allows us to both preserve strong theoretical guarantees and enable better empirical performance by being more sensitive to the underlying CDF distribution, compared to bounds that are agnostic to this shape. Empirically, we observe that our DKW approach of bounding the CDF achieves an order of magnitude lower CVaR-regret than Cassel et al. [10]'s approach which uses direct bonuses on the sample CVaR. These improvements on a variety of simulated environments showcase the importance of distributionally-aware exploration for quickly learning risk-sensitive policies.

2 Related Work

Much work on risk-aware multi-armed bandits [12–15] considers mean-variance objectives. We here consider CVaR due to its advantage as a coherent risk measure [4]. Galichet et al. [16] consider a CVaR-optimizing framework, but only analyze the case where $\alpha \to 0$, which corresponds to finding the arm distribution with the greatest essential infimum. In a pure exploration setting, Kolla et al. [17] consider the task of finding the arm with the optimal CVaR with a successive rejects algorithm.

Outside the bandits framework, Brown [18] and Thomas and Learned-Miller [19] consider the problem of quantifying the uncertainty of a CVaR estimate from a set of samples. Both present concentration inequalities for CVaR, with the latter showing that the DKW inequality can provide much tighter bounds in practice.

There has been very recent related work [10] for bandits with more general risk measures. However, for learning CVaR bandit policies, their work relies on a more generic upper confidence bound that works in the exploration bonus framework, whereas we apply optimism at the sample level.

3 Background and Notation

Let X be a bounded random variable with cumulative distribution function $F(x) = \mathbb{P}[X \leq x]$. The *conditional value at risk (CVaR)* at level $\alpha \in (0, 1)$ of a random variable X is then defined as [20]:

$$\mathrm{CVaR}_\alpha(X) := \sup_\nu \left\{ \nu - \frac{1}{\alpha}\mathbb{E}[(\nu - X)^+] \right\}$$

We define the inverse CDF as $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. It is well known that when X has a continuous distribution, $\mathrm{CVaR}_\alpha(X) = \mathbb{E}_{X\sim F}[X | X \leq F^{-1}(\alpha)]$ [21]. For ease of notation we sometimes write CVaR as a function of the CDF F, $\mathrm{CVaR}_\alpha(F)$.

We consider a stochastic K-armed bandit setting with rewards contained in $[0, U]$. $T_i(n)$ is the number of times arm i has been pulled up to round n; A_t is the action taken during round t; $[m]$ denotes the set $\{1, \dots m\}$ for any m; and P_i is the PDF of the distribution of rewards from the ith arm. Let $(X_{i,t})_{i\in[k],t\in[n]}$ denote a collection of independent random variables (the samples of our arms), with the pdf of X_{it} equal to P_i. $X_t = X_{A_t,T_{A_t}(t)}$ is the reward in round t. The empirical distribution function of $X_{i,s}$ is $\hat{F}_{i,t}(x) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}\{X_{i,s} \le x\}$.

Here, we define the *CVaR-regret* at time n as

$$R_n^\alpha = n \max_i(\mathrm{CVaR}_\alpha(F_i)) - \mathbb{E}[\sum_{t=1}^{n} \mathrm{CVaR}_\alpha(F_{A_t})] = \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}^\alpha\right]$$

where F_a is the CDF of the distribution of rewards from the ath arm, A_t is the action taken at time t and $\Delta_a^\alpha = \max_j(\mathrm{CVaR}_\alpha(F_j)) - \mathrm{CVaR}_\alpha(F_a)$ is the suboptimality of arm a with respect to CVaR. We also consider an alternate notion of regret for the CVaR setting from Cassel et al. [10] in Section 5 and Appendix A.2.

4 Algorithm

We present our algorithm, CVaR-UCB, in Algorithm 1. CVaR-UCB computes an optimistic estimate of the CVaR of each arm from available samples, and then chooses the arm with the highest upper-confidence bound in each turn. This optimistic estimate is based on the concentration of the empirical cumulative distribution function (CDF) via the Dvoretzky-Kiefer-Wolfowitz [DKW; 11] inequality. DKW provides a deviation bound on the maximum difference between the empirical CDF and the true CDF as a function of the number of samples observed. We can use the DKW bound to compute an optimistic estimate of the CVaR value of an arm by computing the CVaR of the optimistic confidence band around the CDF. This optimistic CVaR can be found very simply by shifting the lowest-reward samples to the maximum reward U and then taking the empirical CVaR of the resulting "optimistic samples". Figure 1a illustrates this process, both in terms of a confidence band around CDF and in terms of generating an optimistic set of samples.

Algorithm 1: CVaR-UCB

Input: Risk level α, reward range U, horizon n

Choose each arm once;
Set $\hat{F}_a$ as the CDFs of each arm a on $[0, U]$ for all $a \in [K]$;
Set $T_a \leftarrow 1$;
for $t = 1, \dots, n$ **do**
  for $a = 1, \dots, K$ **do**
    $\epsilon_a \leftarrow \sqrt{\frac{\ln(2n^2)}{2T_a}}$;
    $\tilde{F}_a(x) \leftarrow \left(\hat{F}_a(x) - \epsilon_i \mathbf{1}\{x \in [0, U)\}\right)^+$;
    $\mathrm{UCB}_a^{\mathrm{DKW}}(t) \leftarrow \mathrm{CVaR}_\alpha(\tilde{F}_a)$;
  Play action $A_t = \mathrm{argmax}_i\, \mathrm{UCB}_i^{\mathrm{DKW}}(t)$;
  $T_{A_t} \leftarrow T_{A_t} + 1$;
  Update empirical CDF $\hat{F}_{A_t}$ of arm A_t;

We now show that this simple method has strong regret bounds. To our knowledge this is the first time that DKW-based optimism has been used in risk-sensitive RL or bandits. For proofs see Appendix A.

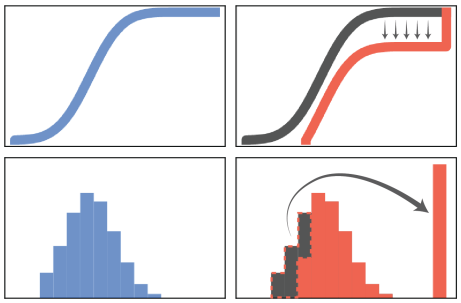
Theorem 1. *Consider CVaR-UCB on a stochastic K-armed bandit problem with rewards bounded in* $[0, U]$. *For any given horizon* n *the expected CVaR-regret after this horizon is bounded as*

$$R_n^\alpha \le \sum_{i\in[K]:\Delta_i^\alpha>0} \frac{4U^2\ln(\sqrt{2}n)}{\alpha^2\Delta_i^\alpha} + 3\sum_{i=1}^{K}\Delta_i^\alpha; \qquad R_n^\alpha \le \frac{4U}{\alpha}\sqrt{nK\ln(\sqrt{2}n)} + 3KU$$

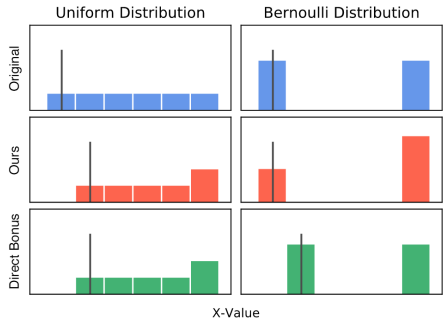
Note that the bounds differ on their dependence on the number of samples n and risk level α: the problem-dependent bound is $O\left(U^2\log n/\alpha^2\right)$, while the problem-independent bound grows as $O\left(U\sqrt{n}/\alpha\right)$. Observe that for $\alpha = 1$, we recover (in dominant terms) the well known upper confidence bound regret results for comparing to the arm with the best expected reward [22].

5 Comparison with Direct Bonuses on the CVaR

In our proposed approach we compute an optimistic estimate of the CDF and then extract a CVaR from this optimistic estimate. In contrast, in standard bandit methods optimizing for expected outcomes,



(a) Calculation of $\text{UCB}_i^{\text{DKW}}$. **Top-left**: The empirical CDF of arm i. **Top-right**: The lower DKW confidence band (a shifted-down version of the empirical CDF). **Bottom-left**: A histogram of samples from arm i. **Bottom-right**: A histogram of samples optimistically perturbed to be close to the lower DKW band. This is done by replacing the ϵ_i-fraction smallest samples with the maximum reward. The empirical CVaR_α of these new samples is $\text{UCB}_i^{\text{DKW}}$.



(b) While our method shifts the lowest-reward samples to the maximum value, direct bonuses on the sample CVaR effectively shift all samples to the right equally. For the uniform distribution (left), both have the same effect, leading to an equivalent CVaR estimate (vertical black line). However, for a Bernoulli distribution, our method can leave the empirical CVaR estimate unchanged while direct bonuses always result in a looser estimate.

Figure 1: Illustration of our method (1a) and comparison with direct bonuses on the sample CVaR (1b).

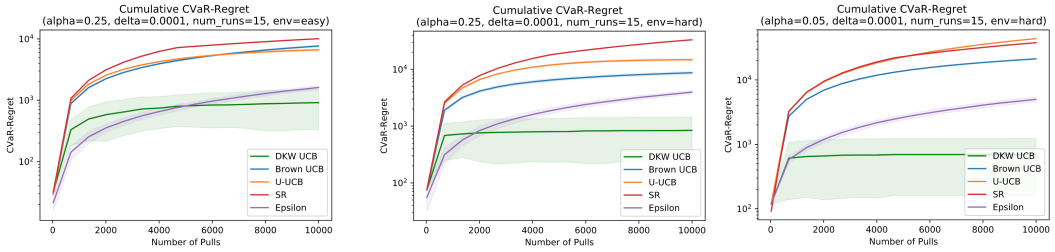
one simply adds a direct bonus to the mean to compute the upper-confidence bound. Therefore, a natural alternative to our proposal, as introduced by Cassel et al. [10], is to directly compute the empirical CDF, extract the empirical CVaR and then add a bonus based on the number of samples. Procedurally this is equivalent to right-shifting each observed sample, in contrast to our algorithm in which we use DKW to compute a lower bound on the empirical CDF, effectively shifting probability mass from the lower-reward tail to the max reward.

Interestingly we will shortly observe that empirically there is a significant difference between these two styles of approaches. In particular, our approach of first computing a bound on the empirical CDF will yield a CVaR transformation that depends on the shape of the CDF itself. In contrast, adding a bonus directly to the empirical CVaR is agnostic of the CDF structure, and relies only on the number of samples observed. To gain some geometric intuition for the superiority of our method, consider a distribution where the probability mass in the low-reward tail is clustered around a single point. In this scenario, one would hope to estimate the CVaR quite quickly as the tail variance [23] is small, and indeed the lowest α fraction of optimistic samples will be almost unchanged by our optimistic perturbation. Figure 1b illustrates this by looking at the optimistic CVaR estimates generated for two different distributions (for an empirical simulation, see Appendix B.4).

Cassel et al. [10] show that their U-UCB approach achieves a problem-dependent "proxy regret" bound of order $O\left(U^2 \log n/\alpha^2\right)$ which matches our bound above.[1] Note however that their proxy regret is potentially a slightly less standard notion of regret. While we show in Appendix A.2 that it is an upper bound on our CVaR-regret, our CVaR-regret is amenable to a simpler analysis and is still a good measure of the algorithm's performance; e.g., to achieve sub-linear CVaR-regret the algorithm needs to play the optimal arm more and more frequently. However, to rule out the possibility that our algorithm's superior performance is due to the use of a different objective, Proposition 2 shows that CVaR-UCB achieves the same dependency on α and n as U-UCB, even when evaluating on proxy regret.

We further illustrate the benefits of our approach over direct bonuses by devising a variant of U-UCB with even better dependence on risk level α. This algorithm, Brown-UCB, leverages a CVaR concentration inequality from Brown [18] to compute a bonus. We show in the Appendix A.1 that its problem-dependent CVaR-regret grows at $O\left(U^2 \log n/\alpha\right)$. Similar to Thomas and Learned-

[1] To be exact, the proxy regret as presented in Cassel et al. [10] is not a cumulative notion, and thus the bound is presented as $O\left(U^2 \log n/\alpha^2 n\right)$, with an additional n present in the denominator. For clarity's sake, we deal with the cumulative version here.



(a) Easy Bandit with $\alpha = 0.25$ (b) Hard Bandit with $\alpha = 0.25$ (c) Hard Bandit with $\alpha = 0.05$

Figure 2: Cumulative CVaR-regret of CVaR-UCB (green; our algorithm), ϵ-greedy (purple), Cassel et al. [10]'s U-UCB (orange), Brown-UCB Brown [18] (blue), and Kolla et al. [17]'s successive rejects algorithm (red) for different bandit setups. While the arm bonuses in rightshifting algorithms are only dependent on the number of samples gathered, the those in DKW-UCB depend further on the particular values of those samples, leading to larger variance than the other algorithms. Means and 95% confidence intervals shown for fifteen runs, with $\delta = 10^{-4}$. Y-axis has log scale.

Miller [19]'s observation in the concentration inequalities setting, we will soon see that while our direct bonus approach has a slightly improved dependency on the risk level compared to the CVaR-UCB above, the latter's empirical advantages can be significant, highlighting the practical significance of leveraging the specific CDF structure when computing a bonus.

Proposition 2. *Consider a stochastic K-armed bandit problem with rewards bounded in* $[0, U]$. *For any given horizon* n *and risk level* α, *both* CVaR*-UCB and U-UCB incur proxy regret with* $O(\frac{\log n}{n})$ *and* $O(1/\alpha^2)$ *dependency on the horizon and risk level, respectively.*

6 Experiments

We present a series of experiments demonstrating the superior performance of our algorithm across several types of distributions and numbers of arms. Details, when not stated, are deferred to Appendix B.

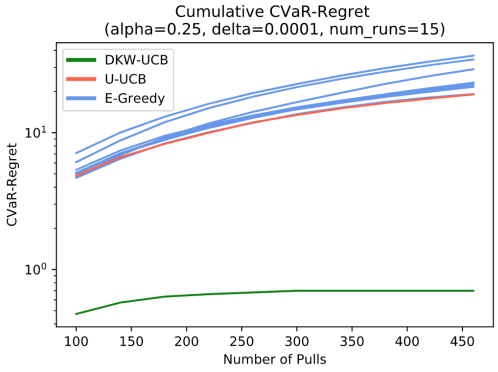
6.1 Truncated Normal Environments

First, we present results on 3-armed truncated-normal bandits with varying risk levels and CVaR-regret gaps. We compare our CVaR-UCB with four others: 1) an ϵ-greedy algorithm, which chooses a random arm with probability $\epsilon = 0.1$, and otherwise the arm with highest empirical CVaR; 2) the CVaR best-arm identification algorithm from Kolla et al. [17]; 3) the U-UCB algorithm from Cassel et al. [10]; and 4) a variant of U-UCB called Brown-UCB presented in Appendix A.1.

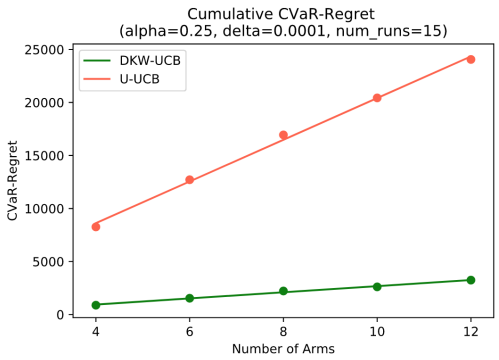
The results in Figure 2 show that the CVaR-regret of our algorithm is an order of magnitude lower than the reward bonus-based algorithms we compared against (note the log-scale). Furthermore, as expected, the cumulative-CVaR regret of our algorithm grows logarithmically, as expected. This is in contrast to the linear growth of ϵ-greedy and Kolla et al. [17], which was not designed for regret minimization.

6.2 Comparison against a Tuned ϵ-Greedy Baseline

In practice, the parameters used for ϵ-greedy can have a substantial impact on its performance. For example, in the risk-neutral case, knowledge of the optimality gaps can be leveraged to create an decaying ϵ-greedy algorithm with logarithmic regret growth [24]. Thus, to demonstrate the practical relevance of our algorithm, it is important to check that finding a successful decay schedule for ϵ-greedy is not easy. Thus, we introduce the Bernoulli Bandit environment, which is designed to penalize algorithms which explore either too little or too much. In addition to comparing against our algorithm, we also compare against the U-UCB algorithm from Cassel et al. [10]. The results of our experiments are shown in Figure 3a, and show that all algorithms incur at least an order of magnitude higher cumulative CVaR-regret than ours.



(a) Cumulative CVaR-regret of our algorithm (green), ϵ-greedy (blue), and Cassel et al. [10]'s U-UCB (red) on the Bernoulli Bandit environment. The ϵ-greedy algorithm was run with a wide range of starting epsilons and decay constants. Results averaged over 15 runs. Y-axis has log scale.



(b) Cumulative CVaR-regret of our algorithm on the One Good Arm environment for different numbers of arms. Values were collected after 3500 pulls and averaged over 15 runs.

Figure 3: Cumulative CVaR-regret comparison on two additional environments.

6.3 Dependence on Number of Arms

We also performed an empirical evaluation of our algorithm's dependence on number of arms K in an environment called One Good Arm.

In Figure 3b, we plot the cumulative CVaR-regret of our algorithm and Cassel et al. [10]'s U-UCB after 3500 pulls for various values of k. The empirical results show what our bound predicts: the CVaR-regret grows linearly with number of arms K. Moreover, our algorithm continues to significantly outperform U-UCB as the number of arms increases.

7 Conclusion

We present *distributionally-aware optimism* as a simple and effective way to reap the benefits of optimism in risk-sensitive multi-armed bandits. We provide theoretical results matching state-of-the-art and empirical results showing that our algorithm achieves an order-of-magnitude lower CVaR-regret than exploration bonus-based alternatives and other baselines. In the future, we aim to expand our analysis to contextual linear bandits, which capture the high-dimensionality and individualized treatment desired in many applications, including healthcare.

References

[1] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[2] Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[3] John White. *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.", 2012.

[4] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[5] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. 2002.

[6] Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

[7] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org, 2017.

[8] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

[9] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*, 2018.

[10] Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *COLT*, 2018.

[11] Aryeh Dvoretzky, Jack Kiefer, Jacob Wolfowitz, et al. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[12] Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[13] Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

[14] Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

[15] Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

[16] Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

[17] Ravi Kumar Kolla, Krishna Jagannathan, et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

[18] David B Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

[19] Philip Thomas and Erik Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, pages 6225–6233, 2019.

[20] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[21] Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

[22] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[23] Emiliano A Valdez. On tail conditional variance and tail covariances. *UNSW Actuarial Studies, Sydney*, 2004.

[24] Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

[25] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Appendices

A Proofs for multi-armed bandit setting

We consider a stochastic K-armed bandit setting with rewards contained in $[0, U]$. $T_i(n)$ is the number of times arm i has been pulled up to round n; A_t is the action taken during round t; $[m]$ denotes the set $\{1, \dots m\}$ for any m; and F_i is the CDF of the distribution of rewards from the ith arm. Let $(X_{i,t})_{i\in[k],t\in[n]}$ denote a collection of independent random variables (the samples of our arms) drawn i.i.d. from $X_{i,t} \sim F_i$. We denote by $X_t = X_{A_t,T_{A_t}(t)}$ the reward in round t. The empirical distribution function of $X_{i,T_i(t)}$ is $\hat{F}_{i,t}(x) = \frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} \mathbb{I}\{X_{i,s} \leq x\}$. We use $\Delta_i^\alpha = \max_j(\mathrm{CVaR}_\alpha(F_j)) - \mathrm{CVaR}_\alpha(F_i)$ for the suboptimality of arm i with respect to CVaR risk level α.

Lemma A.1. *Let F be a CDF of a bounded non-negative random variable and $\nu \in \mathbb{R}$ be arbitrary. Then $\mathbb{E}_F[(\nu - X)^+] = \int_0^\nu F(y)dy$. Hence, one can write the conditional value at risk of a variable $X \sim F$ for any CDF F with $F(0) = 0$ as*

$$\mathrm{CVaR}_\alpha(F) = \sup_\nu \left\{ \frac{1}{\alpha} \int_0^\nu (\alpha - F(y))dy \right\}.$$

Proof. We rewrite $\mathbb{E}_F[(\nu - X)^+]$ as follows

$$\begin{aligned}
\mathbb{E}_F[(\nu - X)^+] =& \mathbb{E}_F[(\nu - X)\mathbf{1}\{X \leq \nu\}] = \nu F(\nu) - \mathbb{E}_F[X\mathbf{1}\{X \leq \nu\}] \\
\overset{①}{=}& \nu F(\nu) - \mathbb{E}_F\left[\mathbf{1}\{X \leq \nu\} \int_0^\infty \mathbf{1}\{X > y\}dy\right] \\
\overset{②}{=}& \nu F(\nu) - \int_0^\infty \mathbb{P}_F\left[y < X \leq \nu\right] dy \\
=& \nu F(\nu) - \int_0^\nu (F(\nu) - F(y))dy = \int_0^\nu F(y)dy
\end{aligned}$$

where ① follows from $a = \int_0^a dx = \int_0^\infty \mathbf{1}\{a > x\}dx$ which holds for any $a \geq 0$ and ② uses Tonelli's theorem to exchange the two integrals. Plugging this identity into

$$\nu - \frac{1}{\alpha}\mathbb{E}_F[(\nu - X)^+] = \frac{1}{\alpha}\left(\nu\alpha - \int_0^\nu F(y)dy\right) = \frac{1}{\alpha}\int_0^\nu (\alpha - F(y))dy$$

and taking the sup over ν gives the desired result. □

Lemma A.2. *Let F and G be the CDFs of two non-negative random variables and let ν_F, ν_G be a maximizing value of ν in the definition of $\mathrm{CVaR}_\alpha(F)$ and $\mathrm{CVaR}_\alpha(G)$ respectively. Then:*

$$\begin{aligned}
|\,\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)| \leq& \frac{1}{\alpha}\int_0^{\max\{F^{-1}(\alpha), G^{-1}(\alpha)\}} |G(y) - F(y)|dy \\
\leq& \frac{\max\{F^{-1}(\alpha), G^{-1}(\alpha)\}}{\alpha} \sup_x |F(x) - G(x)|
\end{aligned}$$

Proof. Assume w.l.o.g. that $\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \geq 0$. Denote by ν_F any maximizing value of ν in the definition of $\mathrm{CVaR}_\alpha(F)$. By Proposition 4.2 and Equation (4.9) in Acerbi and Tasche [21], a possible value of ν_F is $F^{-1}(\alpha)$. Then we can write the differences in CVaR as

$$\begin{aligned}
\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \leq& \nu_F - \alpha^{-1}\mathbb{E}_F[(\nu_F - X)^+] - (\nu_F - \alpha^{-1}\mathbb{E}_G[(\nu_F - X)^+]) \\
=& \frac{1}{\alpha}(\mathbb{E}_G[(\nu_F - X)^+] - \mathbb{E}_F[(\nu_F - X)^+]). \qquad (1)
\end{aligned}$$

Using Lemma A.1 in Equation (1) gives

$$\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \leq \frac{1}{\alpha}\left(\int_0^{\nu_F} G(y)dy - \int_0^{\nu_F} F(y)dy\right)$$

$$\leq \frac{1}{\alpha}\int_0^{\nu_F} |G(y) - F(y)|dy \leq \frac{\nu_F}{\alpha}\sup_y |F(y) - G(y)|.$$

We can in full analogy upper-bound $\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F)$ and arrive at the statement. □

Lemma A.3. *Let G and F be CDFs of non-negative random variables so that $\forall x \geq 0: \quad F(x) \geq G(x)$. Then for any $\alpha \in [0,1]$, we have $\mathrm{CVaR}_\alpha(F) \leq \mathrm{CVaR}_\alpha(G)$.*

Proof. Consider now the following difference

$$\frac{1}{\alpha}\int_0^{\nu} (\alpha - G(y))dy - \frac{1}{\alpha}\int_0^{\nu} (\alpha - F(y))dy = \frac{1}{\alpha}\int_0^{\nu} (F(y) - G(y))dy \geq 0.$$

By Lemma A.1, we have that

$$\begin{aligned}&\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F)\\ &= \sup_\nu \left\{\frac{1}{\alpha}\int_0^{\nu} (\alpha - G(y))dy\right\} - \sup_\nu \left\{\frac{1}{\alpha}\int_0^{\nu} (\alpha - F(y))dy\right\}.\end{aligned}$$

Let ν_F denote a value of ν that achieves the supremum in $\frac{1}{\alpha}\int_0^{\nu}(\alpha - F(y))dy$ (which exists by Proposition 4.2 and Equation (4.9) in Acerbi and Tasche [21]). Then

$$\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F) \geq \frac{1}{\alpha}\int_0^{\nu_F} (\alpha - G(y))dy - \frac{1}{\alpha}\int_0^{\nu_F} (\alpha - F(y))dy \geq 0.$$

□

Lemma A.4 (Difference in CVaR)**.** *Let F be the CDF of a random variable bounded by $[0, U]$ and $\hat{F}$ be the empirical CDF obtained by n, $i.i.d$ samples drawn from F. Let $\epsilon > 0$ and $\mathcal{G} = \left\{\sup_x |F(x) - \hat{F}(x)| \leq \epsilon\right\}$ be the event that the empirical CDF is uniformly ϵ-close to the true CDF F. Define $\tilde{F}(x) = 0 \vee (\hat{F}(x) - \epsilon \mathbf{1}\{x \in [0, U]\})$. Then in event $\mathcal{G}$ the following inequality holds*

$$|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(\tilde{F})| \leq \frac{2\tilde{F}^{-1}(\alpha)\epsilon}{\alpha}.$$

Proof. By Lemma A.2, the triangle-inequality and the definition of $\mathcal{G}$ and $\tilde{F}$

$$\begin{aligned}|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(\tilde{F})| &\leq \frac{\tilde{F}^{-1}(\alpha)}{\alpha}\sup_x |F(x) - \tilde{F}(x)|\\ &\leq \frac{\tilde{F}^{-1}(\alpha)}{\alpha}\sup_x |F(x) - \hat{F}(x)| + \frac{\tilde{F}^{-1}(\alpha)}{\alpha}\sup_x |\hat{F}(x) - \tilde{F}(x)|\\ &\leq \frac{2\tilde{F}^{-1}(\alpha)\epsilon}{\alpha}.\end{aligned}$$

□

Lemma A.5 (Down-shift is optimistic for CVaR)**.** *Let F be the CDF of a random variable bounded by $[0, U]$ and $\hat{F}$ be the empirical CDF obtained by n, $i.i.d$ samples drawn from F. Let $\epsilon > 0$ and $\mathcal{G} = \left\{\sup_x |F(x) - \hat{F}(x)| \leq \epsilon\right\}$ be the event that the empirical CDF is uniformly ϵ-close to the true CDF F. Define $\tilde{F}(x) = 0 \vee (\hat{F}(x) - \epsilon \mathbf{1}\{x \in [0, U]\})$. Then in event $\mathcal{G}$ the following inequality holds*

$$\mathrm{CVaR}_\alpha(F) \leq \mathrm{CVaR}_\alpha(\tilde{F})$$

Proof. By construction of $\tilde{F}$, we have $\tilde{F}(x) \leq F(x)$ for all x on $\mathcal{G}$ and hence the statement follows by Lemma A.3. □

Theorem A.6 (DKW-UCB regret bound)**.** *Consider DKW-UCB on a stochastic k-armed bandit problem with bounded rewards in range* $[0, U]$*. For any horizon* n*, if* $\delta = 1/n^2$ *then the expected CVaR-regret after the* n*th timestep is bounded by*

$$R_n^\alpha \leq \sum_{i=1}^{K} \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha} + 3\sum_{i=1}^{K} \Delta_i^\alpha$$

Additionally the CVaR-regret after the n*th timsestep is also bounded by,*

$$R_n^\alpha \leq 4\sqrt{nk\ln(\sqrt{2}n)}\frac{U}{\alpha} + 3\sum_{i}^{k} \Delta_i^\alpha$$

Proof. Our proof closely follows the proof of UCB from [25]. Let c_i^α denote the CVaR of arm i and $\hat{F}_{i,t}$ denote the empirical CDF of the ith arm before timestep t. Define $\tilde{c}_i^\alpha(t)$ as

$$\tilde{c}_i^\alpha(t) = \mathrm{CVaR}_\alpha(\tilde{F}_{i,t})$$

Where $\tilde{F}_{i,t}$ is defined as follows,

$$\tilde{F}_{i,t}(x) = \left(\hat{F}_{i,t} - \sqrt{\frac{\ln(2/\delta)}{2T_i(t)}}\mathbf{1}\{x \in [0, U]\}\right)^+$$

$$\epsilon_i(t) = \frac{U}{\alpha}\sqrt{\frac{2\ln(2/\delta)}{T_i(t)}}$$

First observe that CVaR decomposes as $R_n^\alpha = \sum_{i=1}^K \Delta_i^\alpha \mathbb{E}[T_i(n)]$. We want to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. Without loss of generality we assume arm 1 is the optimal arm. Define the "good event" G_i as:

$$G_i = \{c_1^\alpha \leq \min_{t\in[n]} \tilde{c}_1^\alpha(t)\} \cap \bigcup_{i\in[K]} \{\tilde{c}_i^\alpha(u_i) \leq c_1^\alpha\}$$

We chose $u_i \in [n]$ later. Following Lattimore and Szepesvári [25] we can show by contradiction that if G_i then $T_i(n) \leq u_i$. First, since $T_i(n) \leq n$ we can write:

$$\mathbb{E}[T_i(n)] = \mathbb{E}[T_i(n)\mathbb{I}\{G_i\}] + \mathbb{E}[T_i(n)\mathbb{I}\{G_i^c\}] \leq u_i + \mathbb{P}(G_i^c)n \tag{2}$$

Suppose $T_i(n) > u_i$ on event G_i, that means arm i was played more than u_i times over n rounds and so there must be a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$.

$$\begin{aligned}
\tilde{c}_i^\alpha &= \mathrm{CVaR}_\alpha\left(\hat{F}_{i,t-1} - \sqrt{\frac{\ln(2/\delta)}{2T_i(t-1)}}\right) \\
&= \mathrm{CVaR}_\alpha\left(\hat{F}_{i,u_i} - \sqrt{\frac{\ln(2/\delta)}{2u_i}}\right) \\
&< c_1^\alpha \\
&< \tilde{c}_1^\alpha(t-1)
\end{aligned}$$

Where the second line follows by $T_i(t-1) = u_i$ and the third and the fourth follows by the definition of event G_i. Hence $A_t = \arg\max_j \tilde{c}_j^\alpha \neq i$, which is a contradiction, so when G_i occurs $T_i(n) \leq u_i$. It is left to show the probably of the complement of the good event is low. Consider G_i^c

$$G_i^c = \{c_1^\alpha > \min_{t\in[n]} \tilde{c}_1^\alpha(t)\} \cup \{\tilde{c}_i^\alpha(u_i) > c_1^\alpha\} \tag{3}$$

and let us first consider the probability of the first part

$$\mathbb{P}\left(c_1^\alpha > \min_{t\in[n]} \tilde{c}_1^\alpha(t)\right) = \mathbb{P}\left(\exists t \in [n] \ : \ c_1^\alpha > \tilde{c}_1^\alpha(t)\right)$$

and using optimism as shown in Lemma A.5 with $\epsilon = \sqrt{\frac{\ln(2/\delta)}{2T_1(t)}}$ we can upper bound this probability as

$$\leq \mathbb{P}\left(\exists t \in [n] \ : \ \sup_x |\hat{F}_{1,t}(x) - F_1(x)| > \sqrt{\frac{\ln(2/\delta)}{2T_1(t)}}\right)$$

and combining a union bound over the first $n \geq T_1(t)$ samples of arm 1 with the Dvoretzky-Kiefer-Wolfowitz inequality, we further bound this as

$$\leq n\delta.$$

For the second term of the failure event in Equation (3), recall $\Delta_i^\alpha = c_1^\alpha - c_i^\alpha$, and we chose u_i such that $\Delta_i^\alpha \geq \epsilon_i(u_i)$

$$\mathbb{P}(\tilde{c}_i^\alpha(u_i) > c_1^\alpha) = \mathbb{P}(\tilde{c}_i^\alpha(u_i) - c_i^\alpha > \Delta_i^\alpha) \leq \mathbb{P}(\tilde{c}_i^\alpha(u_i) - c_i^\alpha > \epsilon_i(u_i))$$

applying Lemma A.4 and let t_i be the round at which arm i was observed the u_ith time

$$\leq \mathbb{P}\left(\sup_x |\hat{F}_{i,t_i}(x) - F_i(x)| > \sqrt{\frac{\ln(2/\delta)}{2u_i}}\right) \leq \delta$$

where the final bound follows from the Dvoretzky-Kiefer-Wolfowitz inequality. Hence, the probability of the failure event is bounded as $\mathbb{P}(G_i^c) \leq (n+1)\delta$. Substituting this bound into (2):

$$\mathbb{E}[T_i(n)] \leq u_i + n(n+1)\delta \tag{4}$$

It remains to determine u_i which can be chosen as the first integer that satisfies $\Delta_i^\alpha \geq \epsilon_i(u_i)$:

$$u_i = \left\lceil \frac{2\ln(2/\delta)U^2}{\alpha^2 {\Delta_i^\alpha}^2} \right\rceil$$

Substituting into (4), and choosing $\delta = \frac{1}{n^2}$:

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{2\log(2n^2))U^2}{\alpha^2 {\Delta_i^\alpha}^2} \right\rceil + 2 \leq 3 + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 {\Delta_i^\alpha}^2}$$

Substituting this into CVaR-regret decomposition, we get the desired bound

$$R_n^\alpha = \sum_{i=1}^k \Delta_i^\alpha \mathbb{E}[T_i(n)] \leq \sum_{i=1}^K \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha} + 3\sum_{i=1}^K \Delta_i^\alpha$$

One can also prove a sublinear regret bound that does not depend on the reciprocal of the gaps.

$$\begin{aligned}
R_n^\alpha = \sum_{i=1}^k \Delta_i^\alpha \mathbb{E}[T_i(n)] &= \sum_{i:\Delta_i^\alpha < \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i^\alpha \geq \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i^\alpha \geq \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i^\alpha \geq \Delta} \left(3\Delta_i^\alpha + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha}\right) \\
&\leq n\Delta + \frac{4k\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta} + \sum_{i=1}^k 3\Delta_i^\alpha \\
&\leq 4\sqrt{nk\ln(\sqrt{2}n)}\frac{U}{\alpha} + 3\sum_{i=1}^k \Delta_i^\alpha \\
&\leq 4\frac{U}{\alpha}\sqrt{nk\ln(\sqrt{2}n)} + 3kU
\end{aligned}$$

Where the first inequality follows by $\sum_{i:\Delta_i^\alpha < \Delta} T_i(n) \leq n$, and the last line follows by choosing $\Delta = \frac{U}{\alpha}\sqrt{\frac{4k\ln(\sqrt{2}n)}{n}}$. □

Algorithm 2: Brown-UCB for MABs

Input: Risk level α, reward range U, max # of pulls n
Output: Series of actions $A_1, A_2, \dots A_n$.

Choose each arm once.
Initialize $t = 1$, Set $\delta = 1/n^2$.
for $t = 1, \dots, n$ **do**
  for $i = 1, \dots, k$ **do**
    $\text{UCB}_i^{\text{Brown}}(t) = \text{CVaR}_\alpha(\hat{F}_{i,t}) + U\sqrt{\frac{5\log(3/\delta)}{\alpha T_i(t)}}$;
  Play action $A_t = \text{argmax}_i \, \text{UCB}_i^\alpha(t)$;
  Update empirical CDF of arm A_i;

A.1 Brown-UCB

The Brown-UCB algorithm presented in section 6 uses the upper confidence bound presented in Brown [18]. Similar to Cassel et al. [10] we compute the empirical CVaR of each arm and add an optimism bonus to build an upper confidence bound.

$$UCB_i^{\text{Brown}}(t) = \text{CVaR}_\alpha(\hat{F}_{i,t}) + U\sqrt{\frac{5\log(3/\delta)}{\alpha T_i(t)}}$$

we use $\delta = 1/n^2$. Algorithm 2 describes the algorithm.

Theorem A.7 (Brown-UCB regret bound)**.** *Consider Brown-UCB on a stochastic k-armed bandit problem with bounded rewards in range $[0, U]$. For any horizon n, if $\delta = 1/n^2$ then the expected CVaR-regret after the nth timestep is bounded by*

$$R_n^\alpha \leq \sum_{i=1}^{K} \frac{40\ln(\sqrt{3}n)U^2}{\alpha\Delta_i^\alpha} + 3\sum_{i=1}^{K}\Delta_i^\alpha$$

Additionally the CVaR-regret after the nth timsestep is also bounded by,

$$R_n^\alpha \leq 4\sqrt{\frac{10kn\ln(\sqrt{3}n)}{\alpha}}U + 3KU$$

Proof. Our proof style mimics theorem A.6. Let c_i^α denote the CVaR of arm i, and $\hat{c}_i^\alpha(t)$ be the empirical CVaR of the ith arm before timestep t. First, we observe that the CVaR-regret decomposes as $R_n^\alpha = \sum_{i=1}^k \Delta_i^\alpha \mathbb{E}[T_i(n)]$. The general strategy we use is to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. To do this, we define the "good" event

$$G_i = \{c_1^\alpha < \min_{t\in[n]} \text{UCB}_1^\alpha(t)\} \cap \{\text{UCB}_i^\alpha(u_i) < c_1^\alpha\}$$

where $u_i \in [n]$ is a constant we will choose later. We need to show two things, first if G_i occurs, then $T_i(n) \leq u_i$. Second, The complement event G_i^c occurs with low probability (governed by our future choice of u_i). Because $T_i(n) \leq n$, we will have

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\}T_i(n)] \leq u_i + \mathbb{P}(G_i^c)n. \tag{5}$$

For the case where G_i is true, we can show by contradiction that $T_i(n) \leq u_i$, similar to Lattimore and Szepesvári [25]. The next step is to upper bound $\mathbb{P}(G_i^c)$. By its definition,

$$G_i^c = \{c_1^\alpha \geq \min_{t\in[n]} \text{UCB}_1^\alpha(t)\} \cup \{\text{UCB}_i^\alpha(u_i) \geq c_1^\alpha\}$$

The first set can be decomposed into a union of inequalities $\bigcup_{t=1}^{n}\{c_1^\alpha \geq UCB_1^\alpha(t)\}$. We apply the concentration inequality from Brown [18, Theorem 4.2]. Combining all probability bounds of individual events with a union bound, we bound the probability of $\{c_1^\alpha \geq \min_{t\in[n]} \text{UCB}_1^\alpha(t)\}$ as $n\delta$.

$$\begin{aligned}\mathbb{P}\left(c_1^\alpha \geq \min_{t\in[n]} UCB_1^\alpha(t)\right) &\leq \mathbb{P}\left(\bigcup_{t=1}^{n}\{c_1^\alpha \geq UCB_1^\alpha(t)\}\right) \\ &\leq \sum_{t=1}^{n}\left(\mathbb{P}(c_1^\alpha \geq \hat{c}_1^\alpha + U\sqrt{\frac{5\ln(3/\delta)}{\alpha t}}\right) \leq n\delta\end{aligned}$$

Since the second event is contained in $\{UCB_i^\alpha(u_i) \geq c_1^\alpha\}$ we can simply apply the Brown [18, Corollary 3.1] concentration inequality to the second set as well. Assume that u_i is chosen large enough that $\Delta_i^\alpha - U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}} \geq c\Delta_i^\alpha$:

$$\begin{aligned}\mathbb{P}\left(\hat{c}_i^\alpha + U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}} \geq c_1^\alpha\right) &= \mathbb{P}\left(\hat{c}_i^\alpha - c_i^\alpha \geq \Delta_i^\alpha - U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}}\right) \\ &\leq \mathbb{P}\left(\hat{c}_i^\alpha - c_i^\alpha \geq c\Delta_i^\alpha\right) \\ &\leq \exp\left(-2u_i\frac{\alpha^2(c\Delta_i^\alpha)^2}{U^2}\right)\end{aligned}$$

Substituting into (5), we obtain

$$\mathbb{E}[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(-2u_i\frac{\alpha^2(c\Delta_i^\alpha)^2}{U^2}\right)\right)$$

Choosing $u_i = \left\lceil \frac{5U^2\ln(3/\delta)}{\alpha(1-c)^2(\Delta_i^\alpha)^2}\right\rceil$ and $c = \frac{1}{2}$ then yields

$$\mathbb{E}[T_i(n)] \leq \frac{40U^2\ln(\sqrt{3}n)}{\alpha(\Delta_i^\alpha)^2} + 3$$

when combined with the CVaR-regret decomposition results in

$$R_n^\alpha = \sum_{i=1}^{k}\Delta_i^\alpha\mathbb{E}[T_i(n)] \leq \sum_{i=1}^{K}\frac{40\ln(\sqrt{3}n)U^2}{\alpha\Delta_i^\alpha} + 3\sum_{i=1}^{K}\Delta_i^\alpha$$

To get a final result not dependent on each arm's optimality gap, we decompose the CVaR-regret further as

$$\begin{aligned}R_n^\alpha &= \sum_{i=1}^{k}\Delta_i^\alpha\mathbb{E}[T_i(n)] \\ &= \sum_{i:\Delta_i^\alpha<\Delta}\Delta_i^\alpha\mathbb{E}[T_i(n)] + \sum_{i:\Delta_i^\alpha\geq\Delta}\Delta_i^\alpha\mathbb{E}[T_i(n)] \\ &\leq n\Delta + \sum_{i:\Delta_i^\alpha\geq\Delta}\left(3\Delta_i^\alpha + \frac{40U^2\ln(\sqrt{3}n)}{\alpha\Delta_i^\alpha}\right) \\ &\leq 4\sqrt{\frac{10kn\ln(\sqrt{3}n)}{\alpha}}U + 3\sum_{i=1}^{k}\Delta_i^\alpha \\ &\leq 4\sqrt{\frac{10kn\ln(\sqrt{3}n)}{\alpha}}U + 3KU\end{aligned}$$

where the inequality follows because $\sum_{i:\Delta_i^\alpha<\Delta} T_i(n) \leq n$. Choosing $\Delta = \sqrt{\frac{40kU^2\log(\sqrt{3}n)}{n\alpha}}$ and simplifying produces the desired problem-independent bound. □

A.2 Proxy Regret

Cassel et al. [10] introduced the notion of proxy regret for risk aware multi-arm bandits as:

$$\bar{R}_\pi(n) = \mathrm{CVaR}_\alpha(F_{p^\star}) - \mathbb{E}[\mathrm{CVaR}_\alpha(F_n^\pi)]$$

where $p^\star = \mathrm{argmax}_{p\in\Delta_{K-1}} \mathrm{CVaR}_\alpha(F_p)$ where Δ_{K-1} is the $K-1$ dimensional simplex:

$$\Delta_{K-1} = \left\{ p = (p_1, \ldots, p_K) \in \mathbb{R}^K | \sum_{i=1}^K p_i = 1, p_i \geq 0 \right\}$$

and

$$F_p = \sum_{i=1}^K p_i F_i$$
$$F_n^\pi = \frac{1}{n}\sum_{t=1}^n F_{\pi_t}$$

Where $F^{(i)}$ is the distribution of arm i and π_t is the policy at step t. Here we establish a formal relation between this notion and CVaR-regret, defined in section 3.

Proposition A.8. CVaR *is a convex function of the CDF. Concretely, if* $\sum_{\alpha_i} = 1$ *and* $\alpha_i \geq 0$*:*

$$\mathrm{CVaR}_\alpha\left(\sum_i \alpha_i F_i(x)\right) \leq \sum_i \alpha_i \mathrm{CVaR}_\alpha(F_i(x))$$

Proof. Define the mixture distribution $\hat{F}(x) = \sum_i \alpha_i F_i(x)$:

$$\begin{aligned}
\mathrm{CVaR}_\alpha(\sum_i \alpha_i F_i(x)) &= \frac{1}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} (\alpha - \sum_i \alpha_i F_i(x))dx = \frac{1}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} \sum_i \alpha_i(\alpha - F_i(x))dx \\
&= \sum_i \frac{\alpha_i}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} (\alpha - F_i(x))dx \\
&\leq \sum_i \frac{\alpha_i}{\alpha}\int_0^{{F_i}^{-1}(\alpha)} (\alpha - F_i(x))dx = \sum_i \alpha_i \mathrm{CVaR}_\alpha(F_i(x))
\end{aligned}$$

Where the last inequality followed by the fact that $\int_0^y (\alpha - F(x))dx$ attains its maximum at $F^{-1}(y)$ □

Proposition A.9. *Consider a notion of proxy regret* $\bar{R}_\pi(n)$ *defined in [10] as:*

$$\bar{R}_\pi(n) = \mathbb{E}[\mathrm{CVaR}_\alpha(F_{p^\star}) - \mathrm{CVaR}_\alpha(F_n^\pi)]$$

The notion of regret R_n^α *defined in equation 3 satisfies the following inequality.*

$$\bar{R}_\pi(n) \geq \frac{1}{n} R_\alpha^n$$

Proof. First note that By the convexity of CVaR shown in proposition A.8 we have $\mathrm{CVaR}_\alpha(F_{p^\star}) = \mathrm{CVaR}_\alpha(F_{i^\star})$ where $i^\star = \mathrm{argmax}_i \mathrm{CVaR}_\alpha(F_i)$.

$$\begin{aligned}
\mathrm{CVaR}_\alpha(F_{p^\star}) &= \mathrm{CVaR}_\alpha\left(\sum_{i=1}^K p_i^\star F_i\right) \\
&\leq \sum_{i=1}^K p_i^\star \mathrm{CVaR}_\alpha(F_i) \\
&\leq \max_i \mathrm{CVaR}_\alpha(F_i)
\end{aligned}$$

The bound is tight by setting $p^*_{i^\star} = 1$ and $p^*_i = 0 : i \neq i^\star$. Then by using the linearity of expectation and the convexity of CVaR we have:

$$\begin{aligned}
\bar{R}_\pi(n) &= \mathrm{CVaR}_\alpha(F_{p^\star}) - \mathbb{E}[\mathrm{CVaR}_\alpha(F^\pi_n)] \\
&= \mathrm{CVaR}_\alpha\left(F_{i^\star}\right) - \mathbb{E}[\mathrm{CVaR}_\alpha(F^\pi_n)] \\
&= \mathrm{CVaR}_\alpha\left(F_{i^\star}\right) - \mathbb{E}[\mathrm{CVaR}_\alpha\left(\frac{1}{n}\sum_{t=1}^{n} F_{\pi(t)}\right)] \\
&\geq \mathrm{CVaR}_\alpha\left(F_{i^\star}\right) - \frac{1}{n}\mathbb{E}[\sum_{t=1}^{n}\mathrm{CVaR}_\alpha\left(F_{\pi_i}\right)] \\
&= \frac{1}{n}\left(n\,\mathrm{CVaR}_\alpha\left(F_{i^\star}\right) - \mathbb{E}[\sum_{t=1}^{n}\mathrm{CVaR}_\alpha\, F_{\pi_i})]\right) = \frac{1}{n}R^n_\alpha
\end{aligned}$$

This completes the proof. □

Here we reiterate and prove the proposition 2.

Proposition A.10 (Proposition 2)**.** *Consider a stochastic K-armed bandit problem with rewards bounded in $[0, U]$. For any given horizon n and risk level α, both* CVaR*-UCB and U-UCB incur proxy regret with $O(\frac{\log n}{n})$ and $O(1/\alpha^2)$ dependency on the horizon and risk level, respectively.*

Proof. First note that proxy regret decomposes as (Equation 15 in [10]):

$$\bar{R}_\pi \leq \frac{L}{n}\sum_{i\neq i^\star}\mathbb{E}[T_i(n)]||F_{i^\star} - F_i||$$

Where $i^\star = \arg\max_i \mathrm{CVaR}_\alpha(F_i)$ is the optimal arm, and:

$$L = b\left(1 + \max_{i,j\in\{1,\dots,K\}}||F_i - F_j||^{q-1}\right)$$

$$||F|| = \max\{||F||_\infty, \int_{-\infty}^{0} x dF\}$$

For CVaR_α, $q = 2$ and $b = \frac{1}{\alpha}(1 + \frac{3}{\min\{\alpha, 1-\alpha\}})$ (proposition 4 in Appendix E.2. [10]). Following results from Theorem A.6, follows that for CVaR-UCB:

$$\bar{R}_\pi \leq \frac{L}{n}\sum_{i\neq i^\star}\left(3 + \frac{4\log(\sqrt{2}n)}{\alpha^2\Delta_i^2}U_i\right)||F_{i^\star} - F_i||$$

The right hand side has $O(1/\alpha^2)$ and $O(\log(n)/n)$ dependency on the risk level and horizon, respectively.

Similarly for U-UCB we have (Theorem 2 in [10])

$$\bar{R}_{U-UCB} \leq \frac{L}{n}\sum_{i\neq i^\star}\left(\frac{\alpha'\log n}{\phi(\Delta_i/2)} + \frac{\alpha'+6}{\alpha'-2}\right)||F_{i^\star} - F_i||$$

For $\alpha' \geq 2$. Where

$$\phi(y) = \min\{a(\frac{y}{2b})^2, a(\frac{y}{2b})^{2/q}\}$$

For CVaR_α we have $b \leq \frac{1}{\alpha}$ and $q = 2$, which yields $\phi(\Delta_i/2) \leq a\alpha^2\Delta_i^2$. Hence the RHS has $O(1/\alpha^2)$ and $O(\log(n)/n)$ dependency on the risk level and horizon. □

B Experimental Details

B.1 Three Arm Bandit Environments

In each distribution, the arms correspond to truncated normal distributions with different means and variances. The parameters of these distributions, along with their CVaRs, are shown in Table 1. The

Environment	Arm 1	Arm 2	Arm 3
Easy, $\alpha = 0.25$	3.18	**9.32**	-1.02
Hard, $\alpha = 0.25$	6.59	**9.32**	6.91
Hard, $\alpha = 0.05$	4.91	**6.52**	2.43

(a) CVaR values of different arms in our simulations. In all cases, Arm 2 (bolded) has the highest CVaR.

Parameter	Arm 1	Arm 2	Arm 3
Easy & Hard μ	10	15	16
Easy σ	6	5	15
Hard σ	3	5	8

(b) Means and standard deviations for each arm for the Easy and Hard environments. All distributions were truncated at $\pm 1\sigma$.

Table 1: Setup for the Three-Arm Multi-Arm Bandit Environments.

difference between the Easy and Hard environments is that the arms in the Hard environment feature a smaller difference between their CVaR values compared to those in the Easy environment. We also assess our algorithm's performance on two runs of the Hard environment with different levels of α. Note that we began each experiment by pulling each arm $\lceil 1/\alpha \rceil$ times to ensure there are enough samples so that the empirical CVaR is properly defined.

B.2 Bernoulli Environment

The environment consists of two arms: a deterministic arm that always returns 0.1 reward, and a stochastic that returns reward 1 with probability 0.8 and returns 0 otherwise. The $\text{CVaR}_{0.25}$ of the stochastic arm is 0.2, compared to 0.1 for the deterministic arm. This environment highlights an algorithm's shortcomings in handling exploration: an suboptimal algorithm will either not explore enough and settle for the inferior deterministic arm, or explore for too long and incur large CVaR-regret.

We ran a decaying ϵ-greedy algorithm with a range of exponential decay constants spanning from 10^{-1} to 10^{-6} and initial epsilons ranging from 1 to 10^{-3}. Results shown in Figure 3a.

Note that the successive rejects algorithm requires upper and lower bounds of the CVaR-regret gaps between the optimal arm and all other arms. To ensure as favorable a comparison as possible, we provide the tightest possible bounds based on the true CVaRs of the arms.

Dependence on Number of Arms The One Good Arm environment consists of one normal distribution with mean 0 and variance 1 ($\text{CVaR}_{0.25} \approx -0.703$), and the rest of the arms with mean 0.1 and standard deviation 10 ($\text{CVaR}_{0.25} \approx -6.928$). Distributions were truncated at $\pm 1\sigma$.

B.3 Proxy Regret

In Figure 4, we show a comparison of DKW-UCB and Cassel et al. [10]'s U-UCB on the Bernoulli Bandit environment, using Cassel et al. [10]'s notion of Proxy Regret. DKW-UCB achieves much smaller proxy regret than U-UCB does. Note that unlike the CVaR-regret, Proxy Regret is a non-cumulative notion.

B.4 Tightness of upper confidence bound

As described in section 5 DKW-UCB provides a problem dependent upper confidence bound that empirically can be much tighter than the upper confidence bounds provided in Cassel et al. [10] and Brown [18]. In this section we consider two different reward distributions as in the experiments:

- Bernoulli, with mean $p = 1/2$
- Truncated normal, mean $\mu = 0$, standard deviation $\sigma = 1$, min value -2 and max value 2

Figure 5 shows the tightness of each bound based on the number of samples. The dashed line is the true $\text{CVaR}_{0.25}$ of each distribution, and the y-axis is the computed upper confidence bound of CVaR with each different method, where DKW-UCB is our proposed method (shifting the empirical CDF downward and compute the CVaR), Brown-UCB computes the empirical CDF and add a value based on Brown [18] and similarly U-UCB is the bound presented in Cassel et al. [10]. DKW-UCB shows a much tighter bound in both experiments. Additionally, Figure 5 (c) shows a zoomed version of the

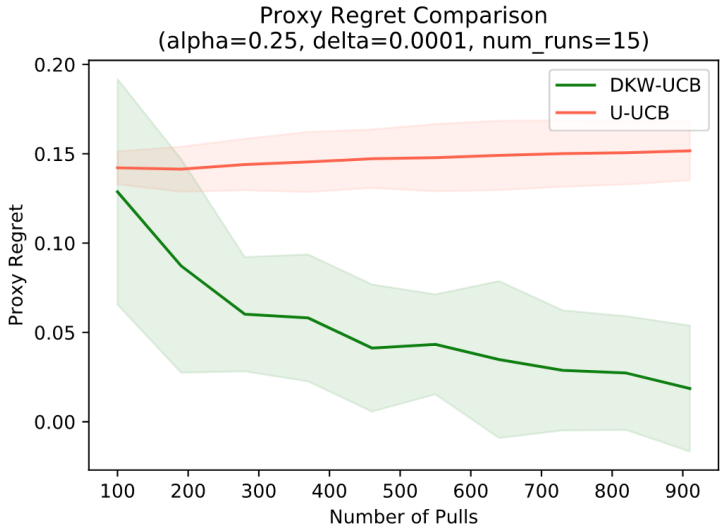


Figure 4: Proxy-Regret comparison between DKW-UCB and U-UCB on the Bernoulli Bandit environment. Means and 95% confidence intervals plotted for 15 runs.

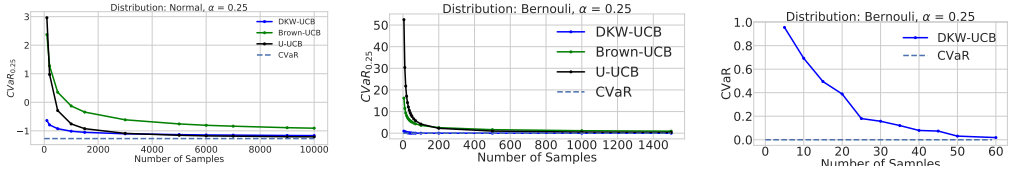


Figure 5: (a) Truncated normal distribution, (b) Bernoulli distribution, (c) Bernoulli distribution, a closer look

Bernoulli bound where the DKW-UCB upper confidence bound exactly matches the CVaR value after around 60 samples.