Distributionally-Aware Exploration for CVaR Bandits

Presenter: Hao Liang

The Chinese University of Hong Kong, Shenzhen, China

February 25, 2021

Mainly based on:
Tamkin, A., Keramati, R., Dann, C., & Brunskill, E. (2019). Distributionally-aware exploration for cvar bandits. In NeurIPS 2019 Workshop on Safety and Robustness on Decision Making.

Introduction

- Traditional multi-armed bandits aims at finding the optimal arm with maximal mean reward.
- However, risk sensitive objectives are often desirable in some high-stakes settings.
  - e.g. health-care, finance and machine control
- A popular risk-sensitive measure is the Conditional Value at Risk (CVaR).
- Consider MAB with CVaR as objective called CVaR bandit.

Notations

- Consider a stochastic K -armed bandit setting with rewards contained in $[0, U]$.
- $T_i(n)$ the number of times arm i has been pulled up to round n
- A_t the action taken during round t; $[m] := \{1, 2, ..., m\}$
- P_i the PDF of the distribution of rewards from the i-th arm
- $(X_{i,t})_{i \in [K], t \in [n]}$ denote a collection of independent random variables, 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 $X_{i,t}$ is $\hat{F}_{i,t}(x) = \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\left\{X_{i,s} \leq x\right\}$

Background

- Let X be a bounded random variable with CDF $F(x) = \mathbb{P}[X \leq x]$
- The CVaR at level α of a random variable X is then defined as

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

- Define the inverse CDF as $F^{-1}(u) = \inf\{x : F(x) \geq u\}$.
- When X has a continuous distribution, $\mathrm{CVaR}_\alpha(X) = \mathbb{E}_{X \sim F}\left[X \mid X \leq F^{-1}(\alpha) \right]$
- Write CVaR as a function of the $\mathrm{CDF} F, \mathrm{CVaR}_\alpha(F)$.

Background

- For continuous random variable X,

$$
\begin{aligned}
\mathrm{CVaR}_\alpha(X) &= \sup_\nu \left\{ \nu - \frac{1}{\alpha}\mathbb{E}\left[(\nu - X)^+\right] \right\} \\
&= \sup_\nu \left\{ \nu - \frac{1}{\alpha}\int_{-\infty}^{\nu} (\nu - x) f(x) dx \right\} \\
&= F^{-1}(\alpha) - \frac{1}{\alpha}\int_{-\infty}^{F^{-1}(\alpha)} (F^{-1}(\alpha) - x) f(x) dx \\
&= F^{-1}(\alpha) - \frac{F^{-1}(\alpha)}{\alpha} F(x)|_{-\infty}^{F^{-1}(\alpha)} + \frac{\int_{-\infty}^{F^{-1}(\alpha)} x f(x) dx}{\alpha} \\
&= \frac{\int_{-\infty}^{F^{-1}(\alpha)} x f(x) dx}{\int_{-\infty}^{F^{-1}(\alpha)} f(x) dx} = \mathbb{E}_{X\sim F}\left[X \mid X \leq F^{-1}(\alpha)\right]
\end{aligned}
$$

CVaR-regret

- Define the CVaR-regret at time n as

$$
\begin{aligned}
R_n^\alpha &:= \mathbb{E}\left[\sum_{t=1}^{n} \max_i \left(\mathrm{CVaR}_\alpha\left(F_i\right)\right) - \mathrm{CVaR}_\alpha\left(F_{A_t}\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}^\alpha\right] \\
&= \sum_{i=1}^{K} \mathbb{E}[T_i(n)]\Delta_i^\alpha,
\end{aligned}
$$

where the third line mimics the regret decomposition in risk-neutral MAB.

Motivation of algorithm

- CVaR-UCB computes an optimistic estimate of the CVaR of each arm, and then chooses the arm with the highest UCB in each turn.
- This optimistic estimate is based on the concentration of the empirical CDF via the DKW inequality: With probability at least $1-\delta$,

$$||\widehat{F}_{i,t}(\cdot) - F_i(\cdot)||_\infty \leq \sqrt{\frac{1}{2t}\ln(\frac{2}{\delta})}$$

- Specifically, the UCB of CVaR is constructed as follows
  - computes an optimistic estimate of the CDF via DKW inequality
  - the UCB of the CVaR value is set to be the CVaR of that optimistic CDF

Algorithm

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 $t = 1, \ldots, n$ **do**

  for $a = 1, \ldots, K$ **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;

Comparison with Direct Bonuses on the CVaR

- View the CDF as a set of samples.
- The optimistic CDF can be found very simply by shifting the lowest-reward samples to the maximum reward U

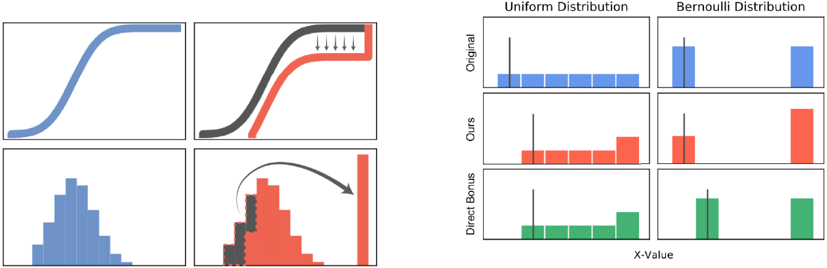


Figure: illustration of the method (left) and comparison with direct bonuses on the sample CVaR (right).

Comparison with Direct Bonuses on the CVaR

- A natural alternative to the proposal (Cassel et al.'18) 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, the DKW-based algorithm compute a lower bound on the empirical CDF, effectively shifting probability mass from the lower-reward tail to the max reward.
- The latter approach depends on the shape of the CDF itself while the former one is agnostic of the CDF structure, and relies only on the number of samples observed.

CVaR regret upper bound

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*

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

- The bounds differ on their dependence on the number of samples n and risk level α:
  - the problem-dependent bound is $O\left(U^2 \log n / \alpha^2\right)$
  - the problem-free bound grows as $O(U\sqrt{n}/\alpha)$
- For $\alpha = 1$, recover (in dominant terms) the well known UCB regret results

Proofs

Lemma 2 (An alternative representation of CVaR).

Let F be a CDF of a bounded non-negative random variable and $\nu \in \mathbb{R}$ be arbitrary. Then $\mathbb{E}_F\left[(\nu - X)^+\right] = \int_0^\nu F(y)dy$. Hence, one can write the CVaR of $X \sim F$ with $F(0) = 0$ as

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

Proof.

First

$$\begin{aligned}\mathbb{E}_F\left[(\nu - X)^+\right] &= \int_0^\nu (\nu - x)dF(x) = \nu \int_0^\nu dF(x) - \int_0^\nu x dF(x) \\ &= \nu F(x)|_0^\nu - (xF(x)|_0^\nu - \int_0^\nu F(x)dx) = \int_0^\nu F(x)dx\end{aligned}$$

$\square$

Proofs

Proof.

Plugging this identity into

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

□

Lemma 3 (Bounding difference of CVaR via distance between CDFs).

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:

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

Proofs

Proof.

Assume w.l.o.g. that $\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \geq 0$. A possible value of ν_F is $F^{-1}(\alpha)$.

$$\begin{aligned}
\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) &\leq \nu_F - \alpha^{-1}\mathbb{E}_F\left[(\nu_F - X)^+\right] - \left(\nu_F - \alpha^{-1}\mathbb{E}_G\left[(\nu_F - X)^+\right]\right) \\
&= \frac{1}{\alpha}\left(\mathbb{E}_G\left[(\nu_F - X)^+\right] - \mathbb{E}_F\left[(\nu_F - X)^+\right]\right) \\
&= \frac{1}{\alpha}\left(\int_0^{\nu_F} G(y)dy - \int_0^{\nu_F} F(y)dy\right) \\
&\leq \frac{1}{\alpha}\int_0^{\nu_F} |G(y) - F(y)|dy \leq \frac{\nu_F}{\alpha}\sup_y |F(y) - G(y)|
\end{aligned}$$

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

Proofs

Lemma 4 (Optimistic CDF results in optimistic estimate of CVaR).

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

Lemma 5 (Difference in CVaR).

Let $\hat{F}$ 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 F. Define $\tilde{F}(x) = [\hat{F}(x) - \epsilon 1\{x \in [0, U]\}]^+$. Then in event $\mathcal{G}$ the following inequality holds

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

Proofs

Proof.

By Lemma 3, the triangle-inequality and the definition of $\mathcal{G}$ and $\tilde{F}$

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

□

Lemma 6 (Down-shift is optimistic for CVaR).

In event $\mathcal{G}$ the following inequality holds

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

Proof of Theorem 1

- The proof closely follows the proof of UCB from [Lattimore'20]
- Let c_i^α denote the CVaR of arm i and $\hat{F}_{i,t}$ denote the empirical CDF of the i th arm before timestep t
- Define $\tilde{c}_i^\alpha(t)$ as $\tilde{c}_i^\alpha(t) = \mathrm{CVaR}_\alpha\left(\tilde{F}_{i,t}\right)$ where $\tilde{F}_{i,t}$ is defined as follows,

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

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

Proof of Theorem 1

- CVaR regret decomposes as $R_n^\alpha = \sum_{i=1}^{K} \Delta_i^\alpha \mathbb{E}\left[T_i(n)\right]$.
- Bound $\mathbb{E}\left[T_i(n)\right]$ for each suboptimal arm i.
- Assume arm 1 is the optimal arm
- Define the good event G_i as:

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

  where $u_i \in [n]$ will be chosen later.
- Show by contradiction that if G_i then $T_i(n) \leq u_i$
- $\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[T_i(n) \mathbb{I}\left\{G_i\right\}\right] + \mathbb{E}\left[T_i(n) \mathbb{I}\left\{G_i^c\right\}\right] \leq u_i + \mathbb{P}\left(G_i^c\right) n$

Proof of Theorem 1

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

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

- Hence $A_t = \arg\max_j \tilde{c}_j^{\alpha}(t-1) \neq i$, which is a contradiction.
- It is left to show $\mathbb{P}\left(G_i^c\right)$ is low.

Proof of Theorem 1

- $G_i^c = \left\{c_1^\alpha > \min_{t\in[n]} \tilde{c}_1^\alpha(t)\right\} \cup \left\{\tilde{c}_i^\alpha\left(u_i\right) > c_1^\alpha\right\}$
- Bound the probability of the first event

$$\begin{aligned}
\mathbb{P}\left(c_1^\alpha > \min_{t\in[n]} \tilde{c}_1^\alpha(t)\right) &= \mathbb{P}\left(\exists t \in [n] : c_1^\alpha > \tilde{c}_1^\alpha(t)\right) \\
&\le \mathbb{P}\left(\exists t \in [n] : \sup_x \left|\hat{F}_{1,t}(x) - F_1(x)\right| > \sqrt{\frac{\ln(2/\delta)}{2T_1(t)}}\right) \\
&\le n\delta
\end{aligned}$$

- Choose u_i such that $\Delta_i^\alpha \ge \epsilon_i\left(u_i\right)$, t_i the round at which arm i was chosen the u_i-th time

$$\begin{aligned}
\mathbb{P}\left(\tilde{c}_i^\alpha\left(u_i\right) > c_1^\alpha\right) &= \mathbb{P}\left(\tilde{c}_i^\alpha\left(u_i\right) - c_i^\alpha > \Delta_i^\alpha\right) \le \mathbb{P}\left(\tilde{c}_i^\alpha\left(u_i\right) - c_i^\alpha > \epsilon_i\left(u_i\right)\right) \\
&\le \mathbb{P}\left(\sup_x \left|\hat{F}_{i,t_i}(x) - F_i(x)\right| > \sqrt{\frac{\ln(2/\delta)}{2u_i}}\right) \le \delta
\end{aligned}$$

Proof of Theorem 1

- Substituting the two bound into

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

- Set $u_i = \left\lceil \frac{2\ln(2/\delta)U^2}{\alpha^2 \Delta_i^{\alpha 2}} \right\rceil$ so that $\Delta_i^\alpha \geq \epsilon_i\left(u_i\right)$ and choose $\delta = \frac{1}{n^2}$

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

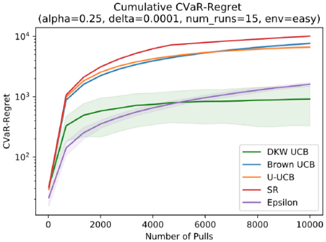
- Substituting this into CVaR-regret decomposition

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

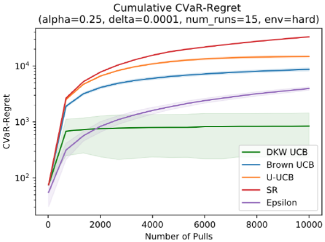
Proof of Theorem 1

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

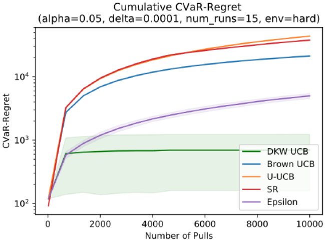
Truncated Normal Environments



(a) Easy Bandit with $\alpha = 0.25$



(b) Hard Bandit with $\alpha = 0.25$



(c) Hard Bandit with $\alpha = 0.05$

Figure: Compare CVaR-UCB with four others: 1) an ϵ-greedy algorithm with $\epsilon = 0{:}1$; 2) the CVaR best-arm identification algorithm from [Kolla'19]; 3) the U-UCB algorithm from [Cassel'18].; and 4) a variant of U-UCB called Brown-UCB. Means and 95% confidence intervals shown for fifteen runs, with $\delta = 10^{-4}$. Y-axis has log scale.

Comparison against a Tuned ϵ-Greedy Baseline

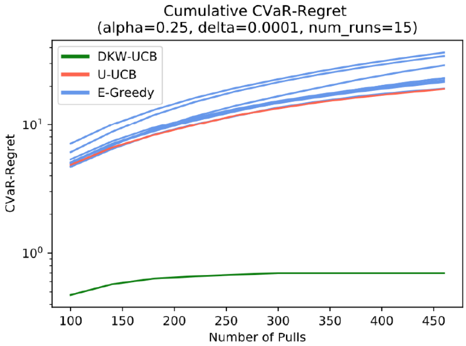


Figure: The ϵ-greedy algorithm was run with a wide range of starting epsilons and decay constants. It is important to verify that finding a successful decay schedule for ϵ-greedy is not easy. In the risk-neutral case, knowledge of the optimality gaps can be leveraged to create an decaying ϵ-greedy algorithm with logarithmic regret growth.

Dependence on Number of Arms

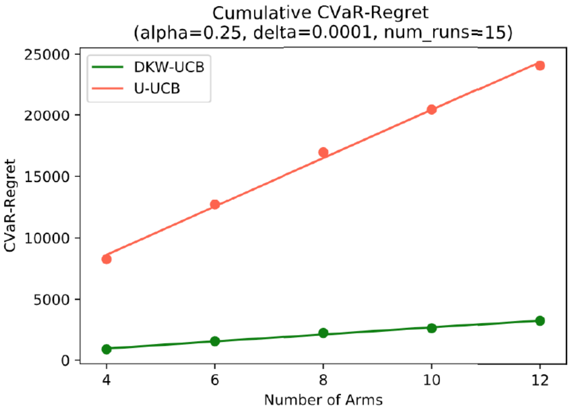


Figure: Cumulative CVaR-regret of our algorithm on the One Good Arm environment for different numbers of arms. Values were collected after 3500 pulls and averaged over 15 runs.

Proxy regret

- Cassel et al. introduced the notion of proxy regret as:

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

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

$$F_p = \sum_{i=1}^{K} p_i F_i$$

$$F_n^\pi = \frac{1}{n} \sum_{t=1}^{n} F_{\pi_t}$$

Proxy regret bounds for CvaR-UCB and U-UCB

Proposition 1.

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

It rules out the possibility that the algorithm's superior performance is due to the use of a different objective.