Risk-Sensitive Bandits: Arm Mixtures Optimality and Regret-Efficient Algorithms

Arpan Mukherjee

Department of Electrical, Computer, and Systems Engineering

CNI, IISc Bangalore

September 30, 2024

Collaborators

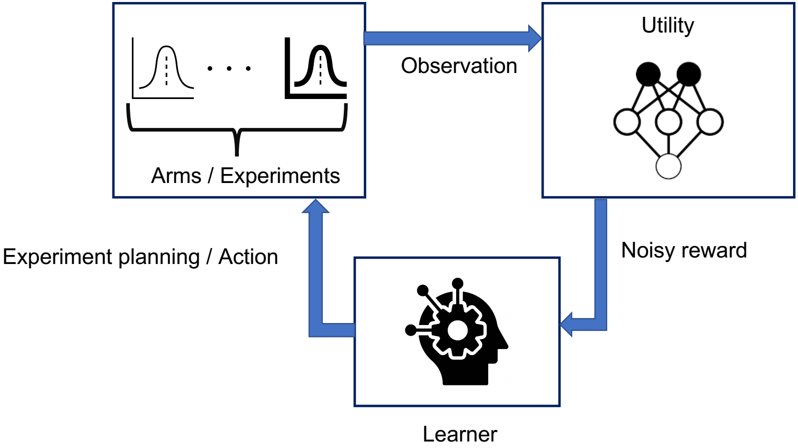
Meltem Tatlı (RPI)

Karthikeyan Shanmugam (Deepmind India)

Prashanth LA (IIT M)

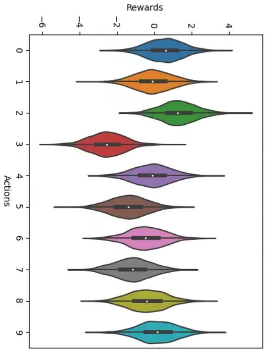
Ali Tajer (RPI)

Multi-Armed Bandits: A Sequential Experimental Design Framework



Multi-armed Bandits: Objectives

Means $\boldsymbol{\mu} \triangleq [\mu_1, \cdots, \mu_K]$ unknown



Regret minimization (Exploration-Exploitation trade-off)

Minimize cumulative regret:

$$R_T \triangleq T\mu_{a^\star} - \sum_{s=1}^{T} \mathbb{E}[X_{A_s}]$$

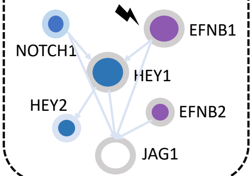
Best arm identification (Pure Exploration)

Identify the arm with the largest mean

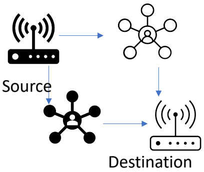
$$a^\star \triangleq \arg\max_{i \in [K]} \mu_i$$

Bandit Settings and Applications

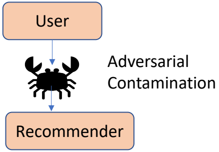
Causal (Gene Network)



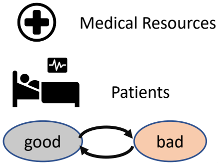
Combinatorial (Routing)



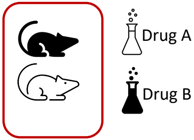
Contaminated (Recommender)



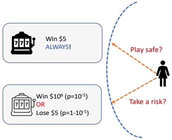
Restless (Healthcare)



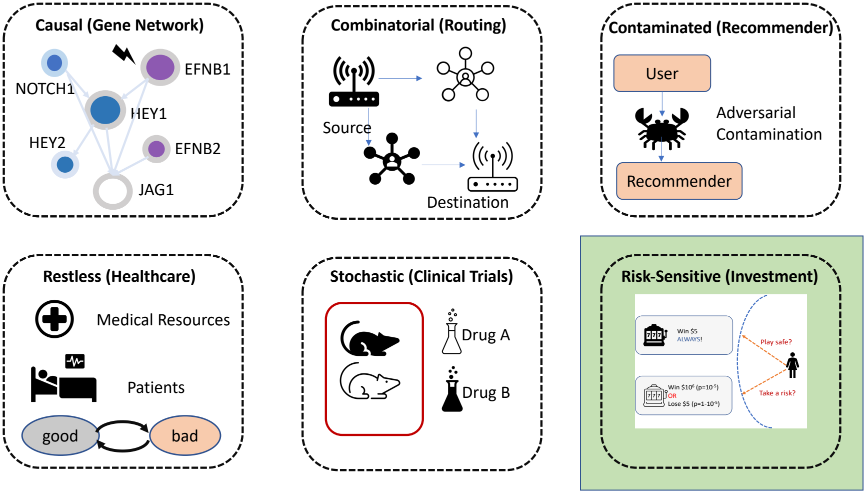
Stochastic (Clinical Trials)



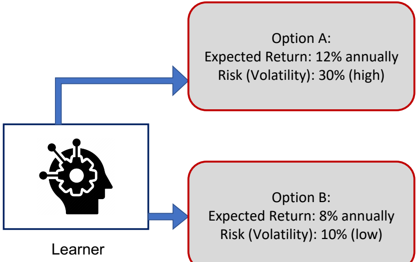
Risk-Sensitive (Investment)

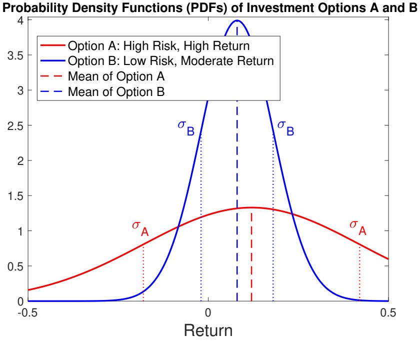


Bandit Settings and Applications – Focus



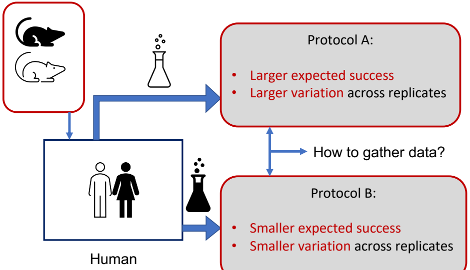
Risk-Sensitive Decision Making





- **Option A:** Larger average reward (mean), larger risk (variance)!
- **Option B:** Smaller average reward (mean), smaller risk (variance)!

Linking Risk-Sensitivity and Experimental Design



- Human-in-the-loop decision making is sensitive to decision risks
- Example bandit applications: clinical trials / investment portfolios
- Average reward is risk-neutral – **not suitable**
- **Question:** How to sequentially control risk?
- Use **Risk-Sensitive Utilities:** Functions of arm distributions (not just the first moment)
- **Examples:** Variance, CVaR, Gini deviation, Sharpe ratio, many others

Risk Nomenclature

Distortion Riskmetric (Wang et. al. 2022)

$$U(\mathbb{P}) := \int_0^\infty h\Big(\mathbb{P}(X \geq x)\Big)\mathrm{d}x$$

- $h : [0,1] \mapsto [0, B]$
- Distortion function, $h(0) = 0$

Risk measures

- Distortion function is **monotone**
- Distortion function is **translation invariant**
- Examples: VaR, CVaR, expected shortfall, quantile-based measures

Deviation measures

- Distortion function is **concave / convex**
- Distortion function **may not be monotone**
- Examples: Gini deviation, mean-median deviation, inter-quantile range

Rensselaer

Risk-Sensitive Bandits: Existing Literature

Sporadic investigations on *specific risk measures*:

- **Quantile-based measures**:
  - Szorenyi et. al. [2015] (regret minimization)
  - David et. al. [2018] (best arm identification)
  - Zhang et. al. [2021] (best arm identification)
- **CVaR**:
  - Baudry et. al. [2018] (regret minimization)
  - Agrawal et. al. [2021] (best arm identification)

Focus: Towards a unifying approach...

- Gopalan et. al. [2017] (regret minimization for distortion risk measures)
- Cassel et. al. [2018] (and [2023]) (empirical distribution performance measures (EDPMs))
- Chang and Tan [2022] (regret minimization for EDPMs)
- Prashanth and Bhat [2022] (regret minimization for EDPMs)

Rensselaer

Objective: Minimization versus Maximization

- Majority of investigations focus on **minimizing** risk
- Few investigations maximize risk measures
  - maximizing risk $\Leftrightarrow$ looking at **gains** instead of losses
  - Examples: Baudry et. al. [2018] and Cassel et. al. [2018/2023] maximize CVaR
  - Khurshid et. al. [2024] maximizes variance to eliminate high volatile arms
- Goal of this work: unconstrained maximization of distortion riskmetrics
  - **Application:** high-volatile trading, traders seek **riskiest** policies for maximizing returns
  - Maximizing **entropy-based deviation measures** well-known in finance

A Unified Framework for Risk Measures (Cassel et. al. 2018 / 2023)

- Let $a^\star$ denote the risk-maximizing arm, i.e.,

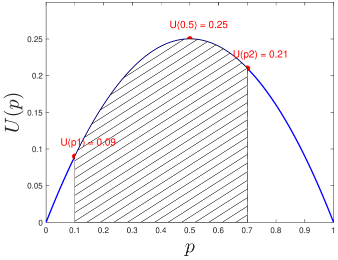
$$a^\star \triangleq \arg\max_{i\in[K]} U(\mathbb{F}_i)$$

- Goal: Minimize the average regret

$$\mathfrak{R}_{\nu}^{\pi}(T) \triangleq U(\mathbb{F}_{a^\star}) - \mathbb{E}_{\nu}^{\pi}\left[U\left(\sum_{i\in[K]} \frac{\tau_T^{\pi}(i)}{T}\mathbb{F}_i\right)\right]$$

- **Assumptions:**
  - The utility is convex $\implies$ solitary best arm
  - The utility is *stable* in an **abstract** semi-normed space – CDF estimates admit exponential convergence to the ground truth
  - Utility is Lipschitz

Assumptions and Gaps in the Literature...



- Convexity **does not hold** for various riskmetrics!
- Concave + non-monotone distortion function $\Longrightarrow$ optimal mixtures!
- Counter-example: **Gini deviation**, $K = 2$ arms

$$U(\alpha p_1 + (1-\alpha)p_2) > \max\{U(p_1), U(p_2)\}$$

Question: Can we construct regret-efficient algorithms for riskmetrics which have optimal mixtures?

Key Challenge: Estimation problem instead of detection problem – how to **track mixtures**?

Revised Objective: Regret w.r.t. Infinite Horizon Oracle Policy

- *Mixtures* may be optimal as opposed to solitary arms
- Oracle Policy: Policy that attains the *maximum* utility over an *infinite horizon*, i.e.,

$$\boldsymbol{\alpha}_{\boldsymbol{\nu}}^{\star} \in \arg \sup_{\boldsymbol{\alpha} \in \Delta^{K-1}} U\Big(\sum_{i \in [K]} \alpha(i)\, \mathbb{F}_i\Big)$$

- Goal: Define regret w.r.t. the oracle policy

$$\mathfrak{R}_{\boldsymbol{\nu}}^{\pi}(T) \triangleq U\left(\sum_{i \in [K]} \alpha_{\boldsymbol{\nu}}^{\star}(i)\mathbb{F}_i\right) - \mathbb{E}_{\boldsymbol{\nu}}^{\pi}\left[U\left(\sum_{i \in [K]} \frac{\tau_T^{\pi}(i)}{T}\mathbb{F}_i\right)\right]$$

- **Assumption:** Hölder continuous utility, Hölder exponent q

Rensselaer

Algorithm Design – Challenges

Attributes of canonical algorithms	Identify sub-optimal arms by sampling them at most $O(\log T)$ times	Risk-neutral algorithms are generally parametric in nature	Risk-sensitive algorithms (Cassel et. al.) have strong assumptions
Challenges in the risk-sensitive setting	➢ Doesn't work for mixtures ➢ **No "sub-optimal" arm**	➢ Risk-sensitive setting is **non-parametric**	➢ **"stable" utilities:** exponential convergence of CDF estimates in an *abstract semi-normed space* **may not hold!** ➢ **convex utilities:** may not be true

Algorithm Design Components

Estimate mixing coefficients in a regret-efficient way

Track the estimated mixtures in a regret-efficient way

- ETC-based mechanism
- UCB-based mechanism

- Under-sampling as an efficient method for tracking mixtures

Risk-Sensitive Explore Then Commit for Mixtures (RS-ETC-M)

Component 1: Estimating mixtures...

- **Step 1** (Explore): Estimate CDFs, draw each arm $\lceil N(\varepsilon)/K \rceil$ times ($N(\varepsilon)$ is instance-dependent)
- **Step 2** (Estimate): Using CDF estimates $\mathbb{F}^{\mathrm{E}}_{t,i}$ of each arm, estimate mixing coefficients through discretization

$$\boldsymbol{\alpha}_{N(\varepsilon)} \in \underset{\boldsymbol{\alpha} \in \Delta_{\varepsilon}^{K-1}}{\operatorname{argmax}} \, U\Big(\sum_{i \in [K]} \alpha(i) \mathbb{F}^{\mathrm{E}}_{t,i} \Big)$$

- Why discretize?
  1. Computational tractability – always computable provided we have plug-in estimates
  2. **Transforms** the problem into a finite-armed bandit instance in terms of discrete mixing coefficients

Rensselaer

Risk-Sensitive Explore Then Commit for Mixtures (RS-ETC-M)

Component 2: Tracking the estimated mixtures...

- **Step 2** (Commit): Sample arms in a way that **best matches** the allocation fractions to the estimated mixing coefficient
- Define $\mathcal{S} \triangleq [K-1]$ as the first $K-1$ arms

$$\tau_T^{\mathrm{E}}(i) \triangleq \begin{cases} \max\left\{\left\lceil \frac{N(\varepsilon)}{K} \right\rceil, \lfloor T\widehat{\alpha}_{N(\varepsilon)}(i) \rfloor\right\}, & \text{if } \; i \in \mathcal{S} \\ T - \sum_{i \in \mathcal{S}} \tau_T^{\mathrm{E}}(i), & \text{otherwise} \end{cases}$$

Drawback

Assumes **knowledge of instannce-dependent parameters** (through $N(\varepsilon)$)

Rensselaer

Risk-Sensitive Upper Confidence Bound for Mixture (RS-UCB-M)

Component 1: Estimating mixtures...

- **Step 1** (Forced exploration): Form reliable estimates of arm CDFs, draw each arm ζT times
  - Forced exploration is **absent** in canonical UCB
  - **Reason:** sub-optimal arms **should not** be sampled over $O(\log T)$ times
  - In our setting, mixtures may **necessitate** a linear order of exploration for every arm!

Open question

Can we design a regret-efficient algorithm that **implicitly** explores arms in a linear order?

Risk-Sensitive Upper Confidence Bound for Mixture (RS-UCB-M)

- **Step 2** (Estimating optimal mixtures): Using CDF estimates $\mathbb{F}^{\mathrm{U}}_{t,i}$ of each arm:
  - **Optimistic estimate:** For any mixture $\boldsymbol{\alpha} \in \Delta^{K-1}$, define the upper confidence bound (UCB):

$$\mathrm{UCB}_t(\boldsymbol{\alpha}) \triangleq \underbrace{U\Big(\sum_{i\in[K]} \alpha(i)\mathbb{F}^{\mathrm{U}}_{t,i}\Big)}_{\text{estimated utility}} + \underbrace{\mathcal{L}\sum_{i\in[K]} \alpha(i)\cdot \mathrm{diam}^q(i)\left(\frac{\log T + 0.15}{\tau^{\mathrm{U}}_t(i)}\right)^{\frac{q}{2}}}_{\text{upper confidence bound}}$$

  - Estimate mixture through *discretization*:

$$\boldsymbol{\alpha}_t \in \operatorname*{argmax}_{\boldsymbol{\alpha}\in\Delta^{K-1}_{\varepsilon}} \mathrm{UCB}_t(\boldsymbol{\alpha})$$

Risk-Sensitive Upper Confidence Bound for Mixture (RS-UCB-M)

Component 2: Tracking the estimated mixtures...

- **Step 3** (Tracking): *Undersample* according to the estimated mixing coefficients, i.e., for all $t > KT\zeta$,

$$A_{t+1} \triangleq \operatorname*{argmax}_{i \in [K]} \left\{ T\alpha_t(i) - \tau_t^{\mathrm{U}}(i) \right\}$$

- No instance dependence

- **Empirically** performs **better than randomly sampling** according to the estimated mixtures

Rensselaer

Regret Analysis

Regret Decomposition

Regret = discretization error + CDF estimation error + sampling estimation error

1. **Discretization error:** Error due to discretization
2. **CDF estimation error:** Error in estimating arm CDFs from rewards
3. **Sampling estimation error:** Error in tracking estimated mixing coefficients

For analyzing the errors, we consider the space of distributions endowed with the 1-**Wasserstein metric**.

1. Exponential convergence of CDF estimates directly follows from DKW (bounded support)
2. Easily extensible to unbounded sub-Gaussian distributions (Prashanth and Bhatt [2022])

Sublinear Regret $\Leftrightarrow$ Convergence in Sampling Fractions

Key finding: UCB + under-sampling is a **regret-efficient** way of tracking mixtures. How?

Lemma (Convergence in mixing coefficient estimates)

After a finite time instant $T(\varepsilon)$, at any time $t > T(\varepsilon)$, the probability that the RS-UCB-M algorithm selects a sub-optimal discrete mixing coefficient is upper-bounded as

$$\mathbb{P}\Big(\exists\, t \in [T(\varepsilon), T] : \boldsymbol{\alpha}_t \neq \bar{\boldsymbol{\alpha}}^\star\Big) \;\leq\; T\left(\left(\frac{1}{T^2}+1\right)^K - 1\right)$$

After a finite time instant, UCB always picks the correct discrete optimal coefficient $\bar{\boldsymbol{\alpha}}^\star$ with a high probability.

Sublinear Regret $\Leftrightarrow$ Convergence in Sampling Fractions

Lemma (Tracking using under-sampling incurs sub-linear regret)

With high probability, we have

$$\left| \frac{\tau_t(i)}{t} - \bar{\alpha}^\star(i) \right| < \frac{K}{T} \quad \text{for all} \quad t > T(\varepsilon)$$

- $T(\varepsilon)$ **inversely proportional** to $\varepsilon^{2/q}$
- Larger the discretization level, faster the convergence to the discrete optimal solution, larger the discretization error

Bounds on Error Components

- CDF estimation error: $O\left(T^{-q/2}(\log T)^{q/2}\right)$ – does not depend on the discretization level

- Sampling Error: $O\left(T\left(\left(\frac{1}{T^2}+1\right)^K-1\right)+\left(\frac{K}{T}\right)^q\right)$ – valid for $T > T(\varepsilon)$ (a finite time instant)

- **Final step:** *Optimize* the discretization level (best possible ε)

Performance Guarantees

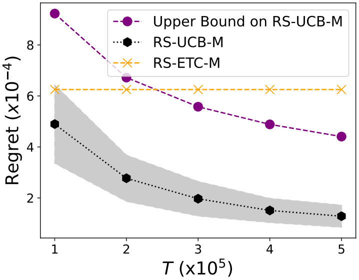
Table: Regret bounds of ETC-type ($\mathfrak{R}_\nu^{\mathrm{E}}(T)$) and UCB-type ($\mathfrak{R}_\nu^{\mathrm{U}}(T)$) algorithms.

Risk-sensitive Utilities [a]	$\mathfrak{R}_\nu^{\mathrm{U}}(T)$	$\mathfrak{R}_\nu^{\mathrm{E}}(T)$
Risk-neutral Mean Value	$O(\sqrt{\log T/T})$	$O(\log T/T)$
Dual Power	$O(\sqrt{\log T/T})$	$O(\log T/T)$
Quadratic	$O(\sqrt{\log T/T})$	$O(\log T/T)$
CVaR	$O(\sqrt{\log T/T})$	$O(\log T/T)$
PHT ($s = 1/2$)	$O((\log T/T)^{1/4})$	$O(\sqrt{\log T/T})$
Wang's Right-Tail Deviation	$O((\log T/T)^{1/4})$	$O(T^{-1/3}(\log T)^{1/4})$
Gini Deviation	$O(\sqrt{\log T/T})$	$O(T^{-1/3}\sqrt{\log T})$

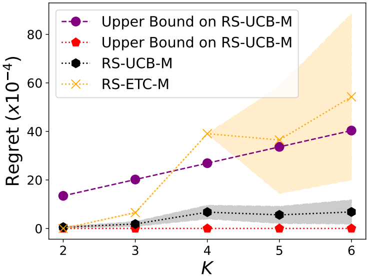
[a] In the first five rows, solitary arms are optimal. In the last two rows, mixtures of arms are optimal.

- RS-ETC-M has **better** regret guarantees for solitary arms (known gap information)
- For mixtures, RS-UCB-M **better** for Gini deviation
- For canonical bandits, ETC and UCB have similar performance guarantees!

Empirical Evaluations



Regret versus time horizon T



Regret versus number of arms K

Figure. Regret of the algorithms for different parameters

- **Utility:** Gini deviation
- $K = 2$, $\boldsymbol{\nu} = [0.4, 0.9]^\top$, $\zeta = 0.1$
- $\boldsymbol{\alpha}^\star_{\boldsymbol{\nu}} = [0.8, 0.2]^\top$

- Regret decomposition in canonical bandits:

$$\mathfrak{R}(T) = \mathbb{E}\left[\sum_{i \neq a^\star} \underbrace{\left(\mu_{a^\star} - \mu_i\right)}_{\text{gap}} \times \underbrace{\tau_t(i)}_{\#\text{times chosen}}\right]$$

- Create principal & alternate bandit instances
- Principal instance **same** as alternate instance **except one sub-optimal arm** of the principal instance
- Use *change of measures* to argue that no policy can have a “small” regret for both instances

Issue

Canonical regret decomposition does not work – **no sub-optimal arms!**

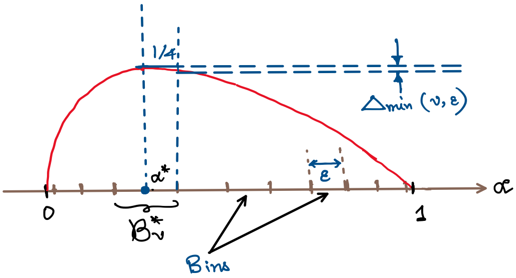
Minimax Lower Bound (K=2)

How to decompose regret?

- Say, the utility is **Gini deviation**
- Pick a discretization level ε
- The discretization scheme is such that $\alpha^\star$ lies at the center of one of the discrete bins
- We have the following rgret decomposition:

$$\mathfrak{R}^\pi_{\boldsymbol{\nu}}(T) \geq \underbrace{\Delta_{\text{min}}(\boldsymbol{\nu}, \varepsilon)}_{\text{minimum gap}} \times \underbrace{\mathbb{P}^\pi_{\boldsymbol{\nu}}\left(\widehat{\alpha}^\pi_T \notin \mathcal{B}^\star_{\boldsymbol{\nu}}\right)}_{\text{Probability of error}}$$

Wednesday, September 25, 2024 7:34 PM



Minimax Lower Bound (K=2)

- Construct the following bandit instances:
  1. Principal instance $\boldsymbol{\nu}$: $\Big(\mathrm{Bern}(p)\ ,\ \mathrm{Bern}(1-p)\Big)$
  2. Alternate instance $\boldsymbol{\nu}_1$: $\Big(\mathrm{Bern}(p+\delta)\ ,\ \mathrm{Bern}(1-p)\Big)$
  3. Alternate instance $\boldsymbol{\nu}_2$: $\Big(\mathrm{Bern}(p)\ ,\ \mathrm{Bern}(1-p-\delta)\Big)$
- For any $k \in \{1, 2\}$, the minimax regret is lower-bounded by:

$$\begin{aligned}\mathfrak{R}^\star(T) &\geq \frac{1}{2}\left(\mathfrak{R}^\pi_{\boldsymbol{\nu}}(T) + \mathfrak{R}^\pi_{\boldsymbol{\nu}_k}(T)\right) \\ &\geq \frac{1}{2}\min\left\{\Delta_{\mathrm{min}}(\boldsymbol{\nu}, \varepsilon), \Delta_{\mathrm{min}}(\boldsymbol{\nu}_k, \varepsilon)\right\} \times \left(\mathbb{P}^\pi_{\boldsymbol{\nu}}(\widehat{\alpha}^\pi_T \notin \mathcal{B}^\star_{\boldsymbol{\nu}}) + \mathbb{P}^\pi_{\boldsymbol{\nu}_k}(\widehat{\alpha}^\pi_T \in \mathcal{B}^\star_{\boldsymbol{\nu}})\right)\end{aligned}$$

Minimax Lower Bound (K=2)

💡 Looks familiar! Lower bound using total variation + Brutagnolle-Huber inequality?

$$\mathfrak{R}^\star(T) \;\geq\; \frac{1}{2}\min\left\{\Delta_{\min}(\boldsymbol{\nu},\varepsilon), \Delta_{\min}(\boldsymbol{\nu}_k,\varepsilon)\right\} \times \exp\left(-\sum_{i\in[K]} \mathbb{E}^\pi_{\boldsymbol{\nu}}[\tau^\pi_T(i)] D_{\mathrm{KL}}(\boldsymbol{\nu}(i)\|\boldsymbol{\nu}_k(i))\right)$$

Yes! However, the principal and the alternate bandit instances should satisfy the following properties.

(P1) Principal and alternate instances should have different optimal bins

(P2) The alternate instances should not be "too different" from the principal instance. Specifically,

$$\frac{1}{D_{\mathrm{KL}}(\boldsymbol{\nu}(1)\|\boldsymbol{\nu}_1(1))} + \frac{1}{D_{\mathrm{KL}}(\boldsymbol{\nu}(2)\|\boldsymbol{\nu}_2(2))} \;\geq\; T$$

Minimax Lower Bound (Theorem)

Q. How to set p and δ in he bandit instances, such that (P1) and (P2) are satisfied?

A. Set $p = 0.5 + \eta$, $\varepsilon = \delta/4$ for (P1), and $\delta = 1/\sqrt{T}$ for (P2).

Final step: Find a lower bound on the minimum utility gap $\Delta_{\min}(\boldsymbol{\nu}, \varepsilon)$. For Gini deviation, we have

$$\Delta_{\min}(\boldsymbol{\nu}, \varepsilon) \geq \frac{1}{4}\varepsilon^2\eta^2 .$$

Theorem (Minimax Lower Bound)

For Gini deviation, for a bandit instance with $K = 2$ arms, the minimax lower bound on the regret is of the order $\Omega(1/T)$.

Rensselaer

Risk-Sensitive Decision Making: Takeaways...

- Risk-sensitivity is an important aspect for human-in-the-loop decision-making
- Existing algorithms works only when **solitary arms** are optimal
- **Key observation:** Various risk measures exhibit optimal mixtures
- RS-UCB-M and RS-ETC-M algorithms proposed for safe decision making, **regret-efficient**, works for **mixtures**
- **Key idea:** Optimistic estimate for mixtures, undersampling for tracking mixtures

Open Questions

Closing the regret gap

- Can we close the gap between the regret upper bound and the minimax lower bound? Current gap of the order $O(1/\sqrt{T})$.
- Can we incorporate the dependence on the number of arms K in the minimax lower bound?

Instance-dependent lower bound

Can we devise instance-dependent lower bounds for risk-sensitive bandits with optimal mixtures?

Structred bandits

How do we extend risk-sensitive decision-making for the larger class of distortion riskmetrics to structured bandits, such as linear bandits, causal bandits, and restless bandits?

Heavy-tailed bandits

Can we derive exponential convergence in CDF estimates for heavy-tailed bandits? What are the performance guarantees for risk-sensitive decision making for heavy-tailed bandits?

Discussion