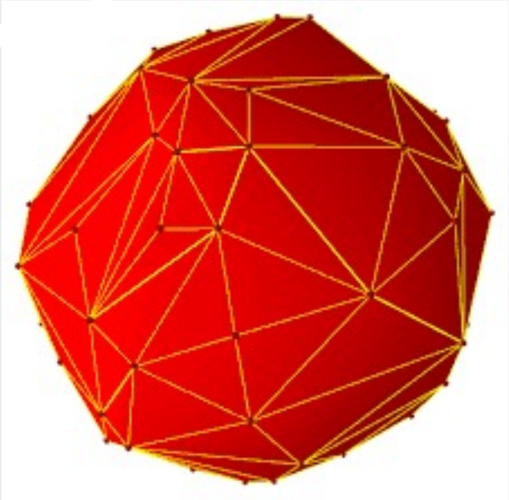
Optimal Learning for Structured Bandits

Negin Golrezaei (MIT)

Joint work with Bart Van Parys (MIT)



Structure of Constraints in Sequential Decision-Making,
Simons Institute for the Theory of Computing, Oct. 13, 2022.

Minor Revision at Management Science

Multi-armed Bandits

Online decision-making under uncertainty:

Fundamental trade-off:

- **Exploration**: collect information to discover the best arm
- **Exploitation**: exploit the collected information to play the arm that seems the best

Healthcare:	**Revenue management**:	**Online advertising:**
What drugs to prescribe?	What price to post?	Which ad to show to users?
(Arms: drugs)	(Arms: prices)	(Arms: ads)

Structured Multi-armed Bandits

Classical Multi-armed Bandits: Rewards of arms are independent of each other

In practice, they may NOT be independent

Healthcare:
Structure: Similar drugs have similar performance

Revenue management:
Structure: Demand goes down as price goes up

Online advertising:
Structure: Some ads are negatively correlated

Structural information makes arms correlated!

Structural Information

Why is structural information important?

Structural information allows for transfer learning

- Information obtained by one arm can be transferred to other arms

Thompson Sampling and UCB perform poorly for structured bandits!!

- They stop playing an arm as soon as they figure out they are suboptimal
- Playing suboptimal arms can help with transfer learning

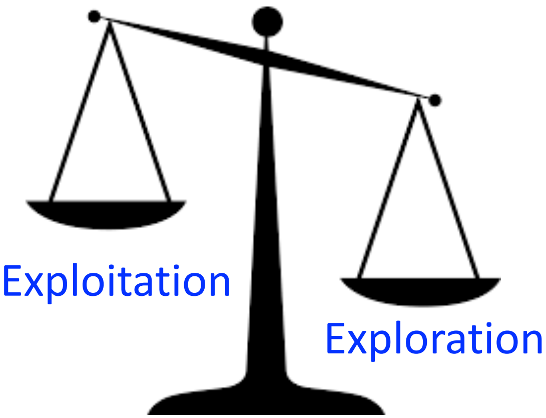
How to deal with structural information?

- Typical approach: Tailored algorithms for special structural problems: Lipschitz, linear, etc

Our approach: Unified framework that works for any convex structural information

Model

- Finite set of arms X with an unknown reward distribution
- A decision-maker needs to pull one of these arms per round over the course of T rounds
- Reward of arm $x \in X$ in round t is r with probability $P(r, x)$
  - P is unknown to the decision-maker
- There is an optimal arm $x^*(P)$ that has the highest average reward
- The decision-maker would like to identify the optimal arm with suffering a low regret



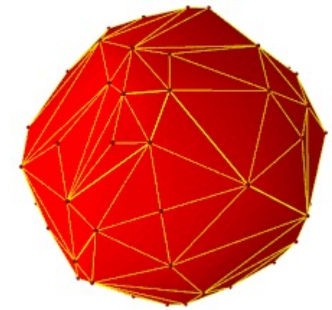
$$\textbf{Regret}_{\boldsymbol{\pi}}(T, P) = \text{best reward in hindsight} - \text{total obtained reward}$$

$$= \sum_x \mathrm{N_T(x)} \Delta(x, P)$$

$\mathrm{N_T(x)}$: Number of times we pull arm x in T rounds

$\Delta(x, P)$: The gap between expected reward of arm x and optimal arm $x^*(P)$

What About Structural Information?



Reward distribution P belongs to a convex set $\mathcal{P}$ (known)

Healthcare:
Structure: Similar drugs (d_1, d_2) have similar performance

$$\mathcal{P} = \left\{ Q : \left| \sum_r rQ(r, d_1) - \sum_r rQ(r, d_2) \right| \le \delta \right\}$$

Online advertising:
Structure: Some ads are negatively correlated

$$\mathcal{P} = \left\{ Q : \left| \sum_r rQ(r, x_D) + \sum_r rQ(r, x_R) \right| \le \delta \right\}$$

Using convex set $\mathcal{P}$, we can model existing structured bandit models: **Linear**, **convex**, **Lipschitz** bandits

- Existing structured bandit models only impose structures on the **mean reward of arms**
- We can impose structures on the entire reward distributions

Our Contributions and Main Results

- Design a unified learning algorithm for structured bandits
- Our **DU**al **S**tructure-based **A**lgorithm **(DUSA)** obtains optimal regret bound
- It mimics the dual counterpart of the regret lower bound to incorporate structural information
- It is computationally efficient
  - It solves a convex problem in only $O(\log(\mathrm{T}))$ periods
- DUSA is the first universally optimal algorithm for structured bandit that is computationally tractable

Related Work

- Learning under particular structural assumptions
  - **Linear structure** (Daniet al., 2008; Rusmevichientong and Tsitsiklis, 2010; Mersereau et al., 2009; Lattimore and Szepes-vari, 2017,…)
  - **Lipschitz structure** (Magureanu et al. 2014; Mao et al. 2018. Gupta et al. (2019),…)
  - **Structural information in contextual bandits** (Slivkins 2011, **Golrezaei et at 2020, …)**
  - **Structures in revenue management problems:** (Keskin et al 2014, Den Boer 2015, Agrawal etal 2017, Bubeck et al 2017, Ferreira et al 2018, **Golrezaei et al 2019**, Bastani et al 2021,...)
- Taking a unified approach:
  - Combes et al. (2017): Their algorithm mimics regret lower bound. But, it has to solve a semi-infinite optimization in every round
  - Russo and Van Roy (2018): balance reward gain with information gain. May not obtain the optimal regret bound

How to Design a Policy for ANY Structural Information?

Main idea: mimic something that directly encapsulates structural information!

How about mimicking the (information-theoretic) regret lower bound?

$$\lim_{T \to \infty} \mathrm{Regret}_{\pi}(\mathrm{T}, \mathrm{P}) \geq C(P) \log(T)$$

where

$$C(P) = \inf \sum_{x} \eta(x) \, \Delta(x, P)$$
$$s.t. \quad \text{sufficient exploration}$$

How to Design a Policy for ANY Structural Information?

Main idea: mimic something that directly encapsulates structural information!

How about mimicking the (information-theoretic) regret lower bound?

$$\lim_{T\to\infty} \mathrm{Regret}_{\pi}(\mathrm{T},\mathrm{P}) \geq C(P)\log(T)$$

where

$$C(P) = \inf \sum_{x} \eta(x)\, \Delta(x,P)$$

$$s.t. \quad \text{sufficinet exploration}$$

This condition encapsulates the structural information!

But How?

Regret Lower Bound: Sufficient Exploration Condition

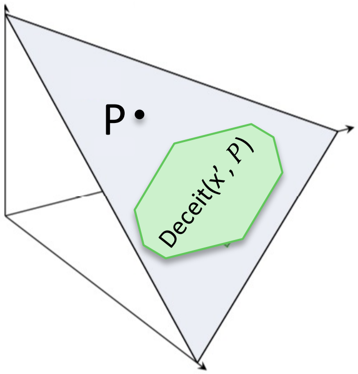
We have done enough exploration if we can distinguish the true distribution P from "deceitful" distributions!

Deceitful distributions (Deceit(x', P)):

1. Belong to convex set $\mathcal{P}$
2. They have the same distribution at $x^*(P)$
3. But, deceivingly have better arm (x') to play

We have done enough exploration if

$$\text{Distance}_{\eta}(\text{P, Deceit(x', P)}) \geq 1$$



This distance depends on **structural information** (convex set $\mathcal{P}$)

How to Design a Policy for ANY Structural Information?

Main idea: mimic something that directly encapsulates structural information!

How about mimicking the (information-theoretic) regret lower bound?

$$\lim_{T\to\infty} \mathrm{Regret}_{\pi}(\mathrm{T},\mathrm{P}) \geq C(P)\log(T)$$

where

$$C(P) = \inf \sum_{x} \eta(x)\, \Delta(x, P)$$

$$s.t.\quad \mathrm{Distance}_{\eta}(\mathrm{P}, \mathrm{Deceit}(\mathrm{x}', \mathrm{P})) \geq 1\ \forall x'$$

Lower bound contains/uses structural information

Mimicking Regret Lower Bound

The optimal solution to the lower bound problem: ($\eta(P)$)

- **Mimicking the Lower Bound:** Pull suboptimal arm x, $\eta(x, P)\log(T)$ times

A big issue: the regret lower bound is NOT available!

- The true reward distribution is NOT known

A high level idea: Compute the empirical reward distribution P_t and follow the empirical regret lower bound $C(P_t)$

If $P_t \to P$, the empirical regret lower bound $C(P_t) \to C(P)$

Mimicking the Regret Lower Bound Is not Easy!

- Solving regret lower bound is computationally expensive
- One does not want to solve the regret lower bound in each round
- If P_t does not converge to P, the idea of mimicking regret lower bound does not work!

First Challenge: Converting a Semi-infinite Lower Bound to Its Convex Counterpart

Regret lower bound (semi-infinite)

$$C(P) = \inf \sum_{x} \eta(x)\, \Delta(x, P)$$

$$s.t.\ \text{Distance}_\eta\ (\text{P, Deceit(x', P)}) \geq\ 1\ \ \forall x'$$

Dual counterpart (Convex)

$$C(P) = \inf_{\eta,\mu} \sum_{x} \eta(x)\, \Delta(x, P)$$

$$s.t\ \ \text{Dual}_\eta\ (\text{P, Deceit(x', P)};\ \mu) \geq\ 1\ \forall x'$$

$$\mu \text{ respects the structral information}$$

Weighted KL distance

$$\text{Distance}_\eta(\text{P, Deceit(x', P)}) =\ \min_{Q} \textstyle\sum_{r,x} \eta(x) P(r,x) \log(P(r,x)/Q(r,x))$$

$$\text{s.t.}\quad Q \in \text{Deceit(x', P)}$$

Let's dualize the distance function

Mimicking the Regret Lower Bound Is not Easy!

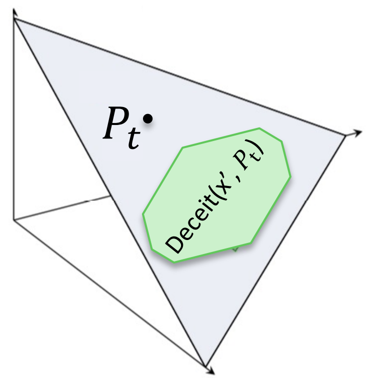
- ✓ Solving regret lower bound is computationally expensive
  - ✓ **Solve its dual instead**

- One does not want to solve the regret lower bound in each round

- If P_t does not converge to P, the idea of mimicking regret lower bound does not work!

Second Challenge: Avoid Solving the Regret Lower Bound in Each Round

- We don't need to resolve the regret lower bound if we have already obtained **enough information**
  - Don't resolve if we can distinguish P_t from Deceit(x', P_t)

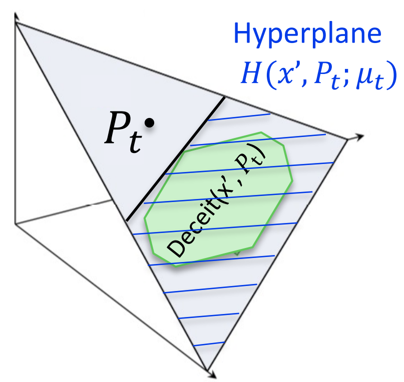


$$\text{Distance}_\eta\,(P_t, \text{Deceit}(x', P_t)) \geq 1$$

Testing this can be demanding!

- We design a simpler (one-dimensional) information test:

$$\text{Distance}_\eta\,(P_t, H(x', P_t; \mu_t)) \geq 1$$



- Can be tested by solving a 1-dimensional convex optimization problem

Mimicking the Regret Lower Bound Is not Easy!

- ✓ Solving regret lower bound is computationally expensive
  - ✓ **Solve its dual instead**
- ✓ One does not want to solve the regret lower bound in each round
  - ✓ **Design a simple information test**
- If P_t does not converge to P, the idea of mimicking regret lower bound does not work!

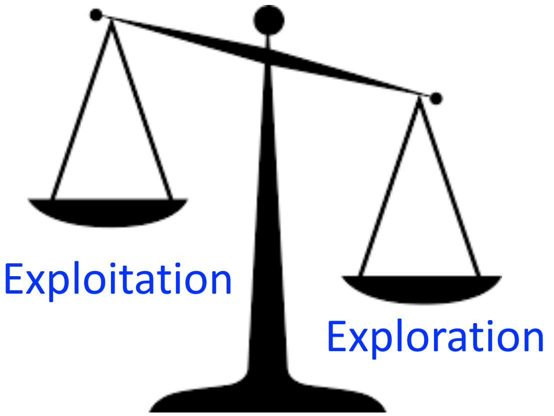
Third Challenge: Ensuring P_t Converges to P

We need to ensure that no arm is completely unexplored

(**Explore**) If $\min_x N_t(x) \leq \epsilon s_t$, pull the least played arm

$$x_t = \mathrm{argmin}_{\mathrm{x}} N_t(x)$$

s_t: number of exploration rounds during the first t rounds



Mimicking the Regret Lower Bound Is not Easy!

- ✓ Solving regret lower bound is computationally expensive
  - ✓ **Solve its dual instead**
- ✓ One does not want to solve the regret lower bound in each round
  - ✓ **Design a simple information test**
- ✓ If P_t does not converge to P, the idea of mimicking regret lower bound does not work!
  - ✓ **Do enough exploration**

Let's Put Everything Together: DUal Structure-based Algorithm (DUSA)

For every $t = |X|: T$

(**Exploit**) If you have collected enough information (i.e., $\text{Distance}_\eta\ (P_t, H(x', P_t; \mu_t)) \geq 1$), exploit by playing the best arm given P_t

(**Explore**) If $\text{Distance}_\eta\ (P_t, H(x', P_t; \mu_t)) < 1$

if $\min_x N_t(x) \leq \epsilon s_t$, pull the least played arm

$$x_t = \text{argmin}_x N_t(x)$$

Here, by following the (dual) regret lower bound

If not, solve the dual regret lower bound to obtain a target rate $(\eta(x, P_t))$ and pull the most behind arm:

$$x_t = \text{argmin}_x \frac{N_t(x)}{\eta(x, P_t)}$$

How did we use the structural information?

Main Theorem: Asymptotic Optimal Regret

Theorem (Regret bound for DUSA)

Under mild assumptions on the reward distribution $\mathrm{P} \in \mathcal{P}$, for any accuracy parameter $0 < \epsilon < \frac{1}{|X|}$, DUSA has the following two properties

- Optimal asymptotic regret:

Optimal regret bound

$$lim\ sup_{T\to\infty} \frac{\mathrm{Regret(T, P)}}{\log(T)} \leq (1+\epsilon)C(P) + O(\epsilon)$$

- Logarithmic number of exploration rounds:

$$\mathrm{E}[s_T] = O(\log(T))$$

Because of our information test, we only solve the dual convex problem in $O(\log(T))$ rounds

Proof Outline

Regret = Regret during **exploitation** + Regret during **exploration**

Exploitation: Obtain a finite regret because of information test.

- The probability that P_t is not close to P is small
- Regret is finite when P_t is close to P

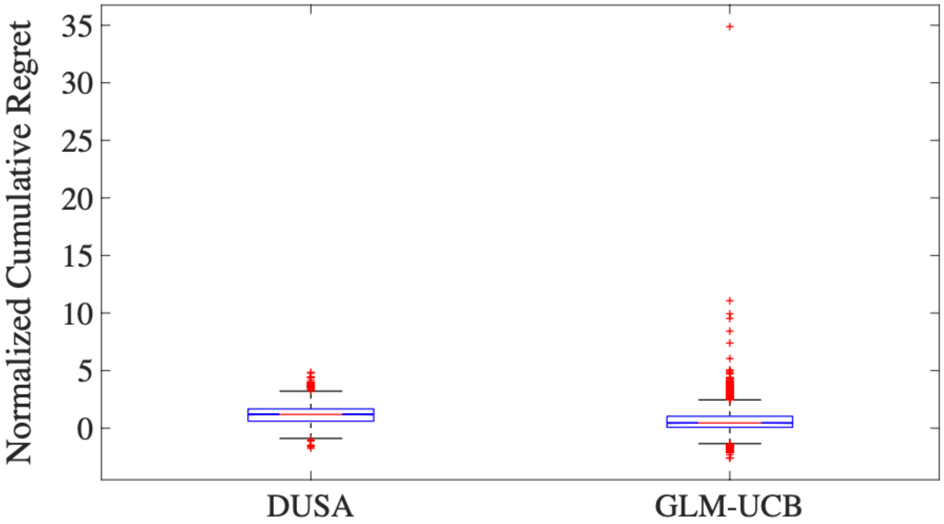
Exploration: Obtain $(1+\epsilon)C(P)\log(\mathrm{T})$ regret

- The probability that P_t is not close to P is small. Thus, our regret here is finite
- When P_t is close to $P, \eta(P_t)$ is close to $\eta(P)$. Thus,

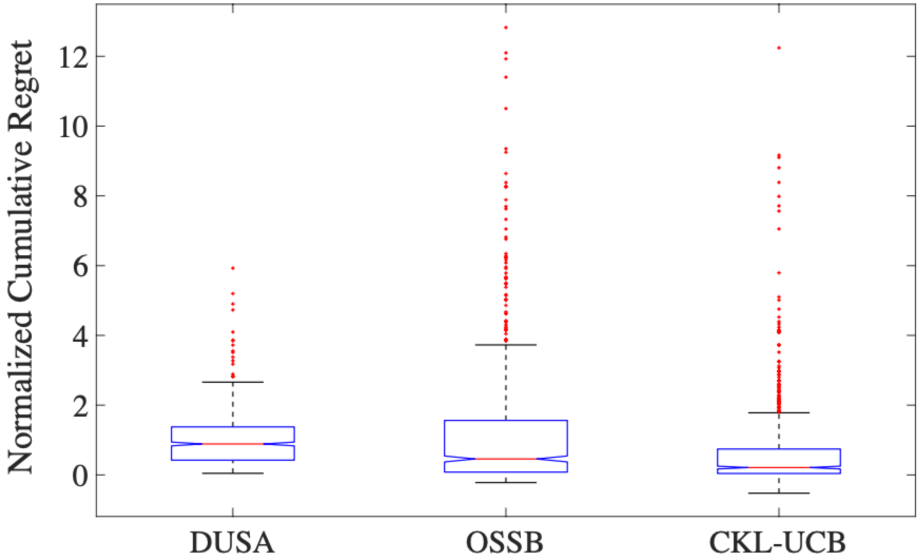
$$\text{Regret } \approx \sum_x \Delta(x,P)\big(\eta(x,P_t)\big)\log(T) \approx \sum_x \Delta(x,P)(\eta(x,P)(1+\epsilon))\log(T)$$

$$= C(P)(1+\epsilon)\log(T)$$

Numerical Studies for Well-known Structured Bandits



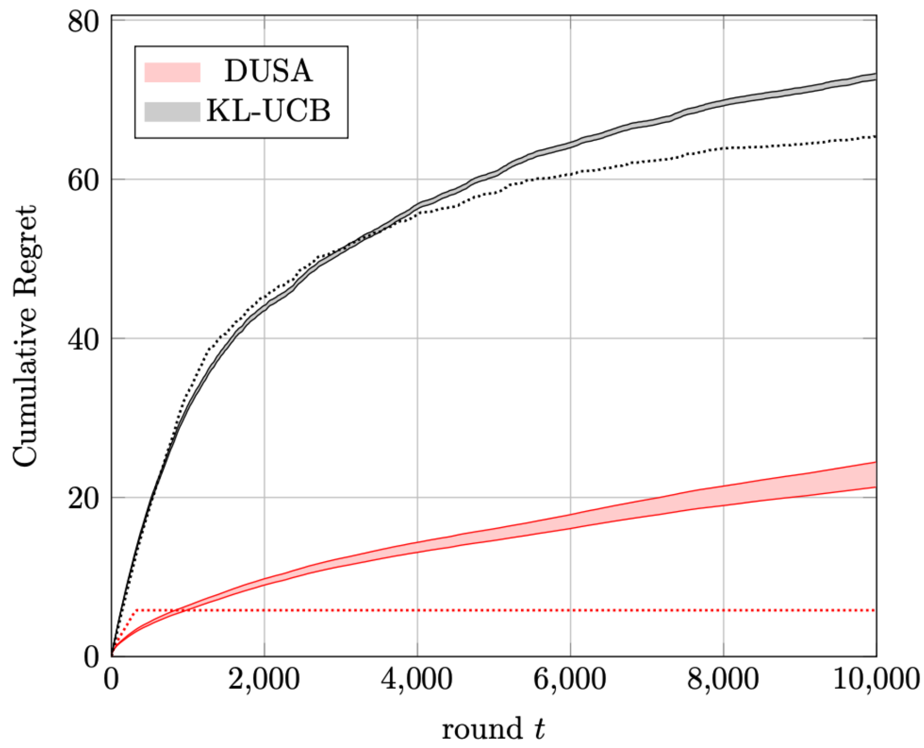
(a) Linear Bandits



(b) Lipschitz Bandits

- DUSA's regret is comparable to the regret of algorithms that are tailored for a specific structured bandits
- DUSA's regret is more concentrated around its median

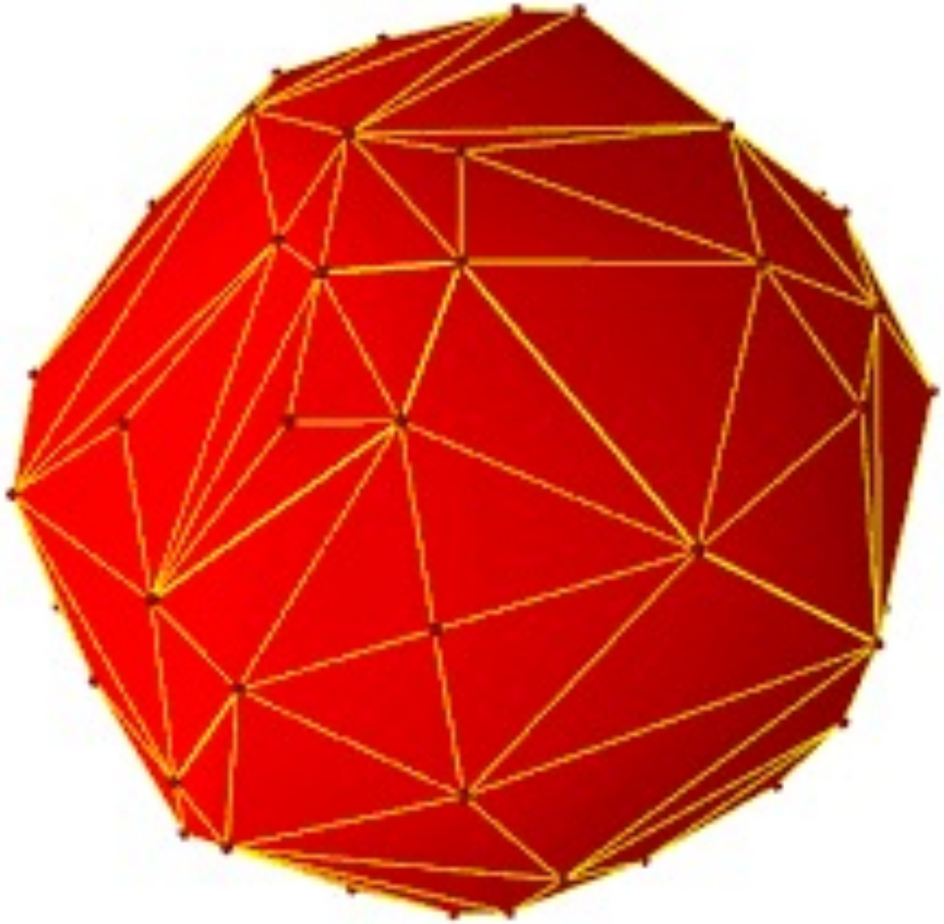
Numerical Studies for Novel Structured Bandits



- Divergence bandit: impose structures on the first and second moment of reward distributions

Takeaways

- Provide a unified framework to study structured bandits
- Present an algorithm called DUSA that obtains optimal regret bound for any convex structural information
- DUSA is the first universally optimal algorithm that is computationally tractable



Link to the paper: https://arxiv.org/abs/2007.07302

Email: golrezae@mit.edu