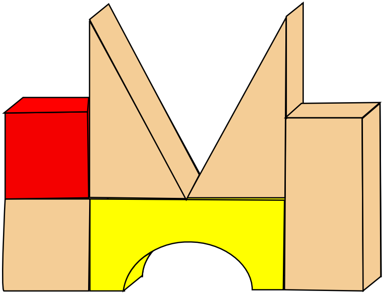
Sequential Reward Maximisation by Solving a Semi-infinite Covering LP



Wouter M. Koolen

CWI
Centrum Wiskunde & Informatica

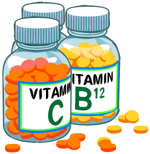
Team

Rémy Degenne

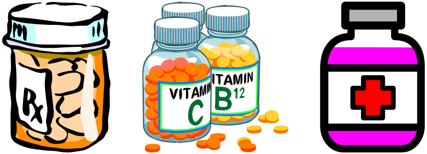
Han Shao (邵涵)

Wouter Koolen

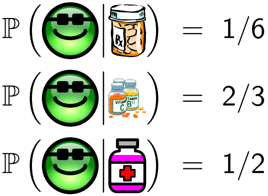
Stochastic Bandit



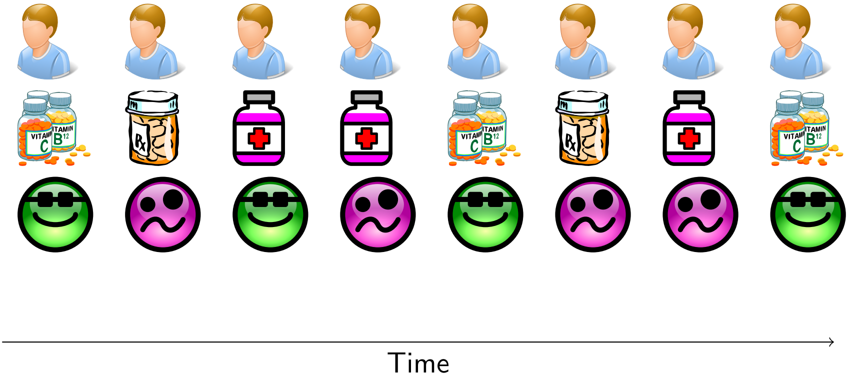
Stochastic Bandit



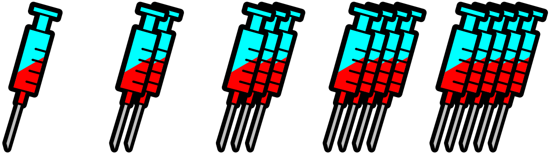
Model (Unknown)



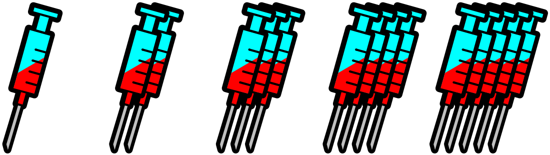
Stochastic Bandit interaction.



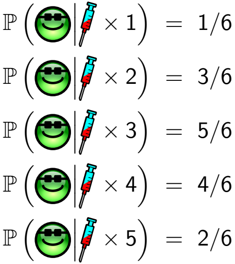
Structured Stochastic Bandit

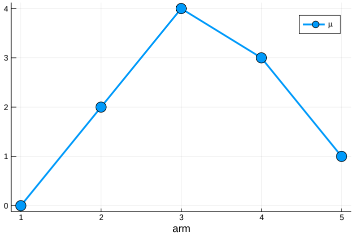


Structured Stochastic Bandit

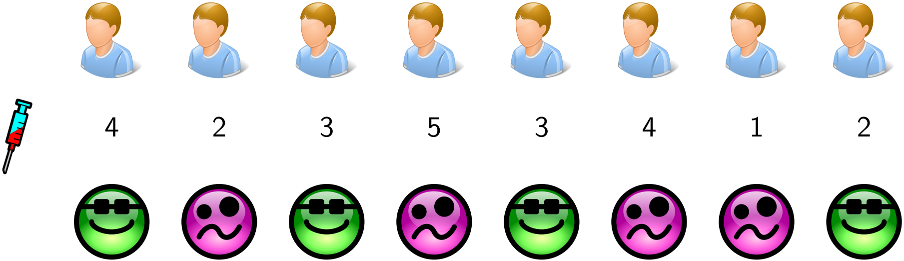


Model (Unknown)

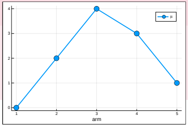


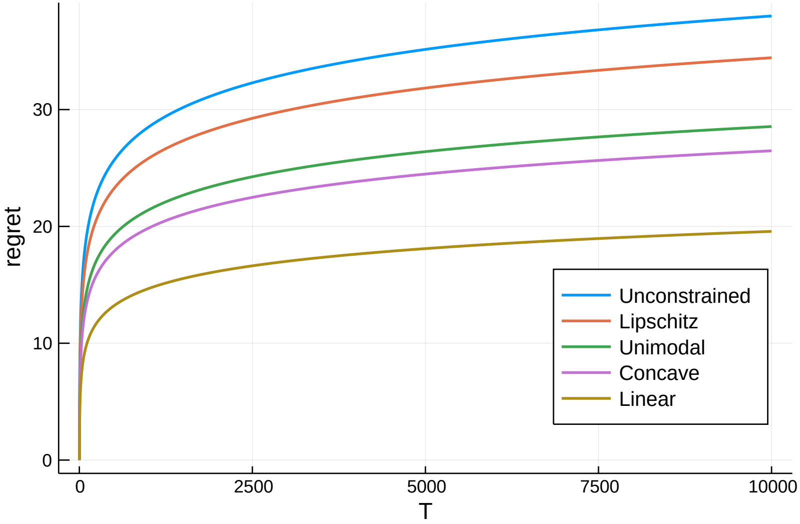


Structured Stochastic Bandit Interaction



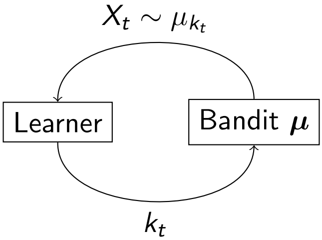
Desired behaviour





Outline

Setting



Structure $\mathcal{M} \subseteq R^K$.
MAB instance $\boldsymbol{\mu} \in \mathcal{M}$
Expfam $d(\mu, \lambda)$
Gaps $\Delta^k = \mu^* - \mu^k$
Regret

$$\sum_{t=1}^{T} \mathbb{E}[\Delta^{k_t}]$$

Goals

- Asymptotic Optimality
- Finite-time Regret Guarantees
- General Structure-Aware Methodology
- Computational Efficiency

Bandital Context

Regret

- Unimodal [Combes and Proutiere, 2014]
- Lipschitz [Magureanu, Combes, and Proutière, 2014]
- Rank-1 [Katariya, Kveton, Szepesvári, Vernade, and Wen, 2017]
- Linear [Lattimore and Szepesvári, 2017]
- OSSB [Combes, Magureanu, and Proutiere, 2017]

Pure Exploration

- Track-and-Stop (MAB) [Garivier and Kaufmann, 2016]
- Structure, Gaussian [Chen, Gupta, Li, Qiao, and Wang, 2017]
- Structure, ExpFam [Kaufmann and Koolen, 2018]
- Game core [Degenne, Koolen, and Ménard, 2019] **yesterday**

Outline

Argument [Graves and Lai, 1997]

Fix an **asymptotically consistent** algorithm for structure $\mathcal{M}$. Consider its behaviour on $\boldsymbol{\mu} \in \mathcal{M}$, and on any alternative bandit model $\boldsymbol{\lambda} \in \mathcal{M}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$:

$$\mathbb{E}_{\boldsymbol{\mu}}\left[N_T^{i^*(\boldsymbol{\mu})}\right]/T \to 1 \qquad \text{but} \qquad \mathbb{E}_{\boldsymbol{\lambda}}\left[N_T^{i^*(\boldsymbol{\mu})}\right]/T \to 0.$$

This stark **difference in behaviour** requires **discriminating information**! Specifically,

$$\mathrm{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}^T \,\middle\|\, \mathbb{P}_{\boldsymbol{\lambda}}^T\right) \;=\; \sum_k \mathbb{E}_{\boldsymbol{\mu}}[N_T^k] d(\mu^k, \lambda^k) \;\geq\; \ln T.$$

Instance-Dependent Regret Lower Bound

Any asymptotically consistent algorithm for structure $\mathcal{M}$ must incur on each $\boldsymbol{\mu} \in \mathcal{M}$ regret at least

$$V_T = \boxed{\min_{\boldsymbol{N} \geq \boldsymbol{0}} \sum_k N^k \Delta^k \quad \text{subject to} \quad \inf_{\boldsymbol{\lambda} \in \Lambda} \sum_k N^k d(\mu^k, \lambda^k) \geq \ln T}$$

where

$$\Lambda = \left\{ \boldsymbol{\lambda} \in \mathcal{M} \mid i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu}) \right\}$$

This is a (semi-infinite) **covering linear program**.

Operationalising the Lower Bound

Earlier work

At each time step

- compute **oracle sample counts** $\boldsymbol{N}^*(\hat{\boldsymbol{\mu}}_t)$ and advance $\boldsymbol{N}_t \to \boldsymbol{N}^*$, or
- **force exploration** to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

Operationalising the Lower Bound

Earlier work

At each time step

- compute **oracle sample counts** $\boldsymbol{N}^*(\hat{\boldsymbol{\mu}}_t)$ and advance $\boldsymbol{N}_t \to \boldsymbol{N}^*$, or
- **force exploration** to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

This talk

- Reformat lower bound as zero-sum "minigame".
- **Iteratively** solve minigame by full information online learning.
- Use iterates to advance $\boldsymbol{N}_t$.
- Add optimism to induce exploration.
- **Compose** regret bound from minigame regret + estimation regret

Minigame

We have $V_T = \frac{\ln T}{D^*}$ where

$$D^* = \max_{\boldsymbol{w} \in \triangle} \inf_{\boldsymbol{\lambda} \in \Lambda} \frac{\sum_k w^k d(\mu^k, \lambda^k)}{\sum_k w^k \Delta^k}$$

$w^k \propto N^k$ pulls

Minigame

We have $V_T = \frac{\ln T}{D^*}$ where

$$D^* = \max_{\boldsymbol{w} \in \triangle} \inf_{\boldsymbol{\lambda} \in \Lambda} \frac{\sum_k w^k d(\mu^k, \lambda^k)}{\sum_k w^k \Delta^k}$$

$$= \max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \Lambda} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}$$

$w^k \propto N^k$
pulls

$\tilde{w}^k \propto N^k \Delta^k$
regret

Minigame

We have $V_T = \frac{\ln T}{D^*}$ where

$$D^* = \max_{\boldsymbol{w} \in \triangle} \inf_{\boldsymbol{\lambda} \in \Lambda} \frac{\sum_k w^k d(\mu^k, \lambda^k)}{\sum_k w^k \Delta^k}$$

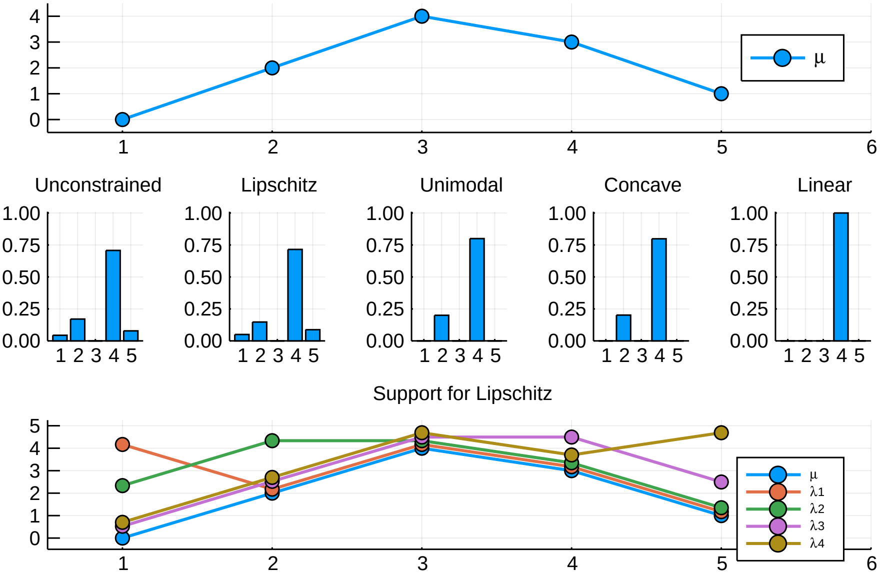
$w^k \propto N^k$
pulls

$$= \max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \Lambda} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}$$

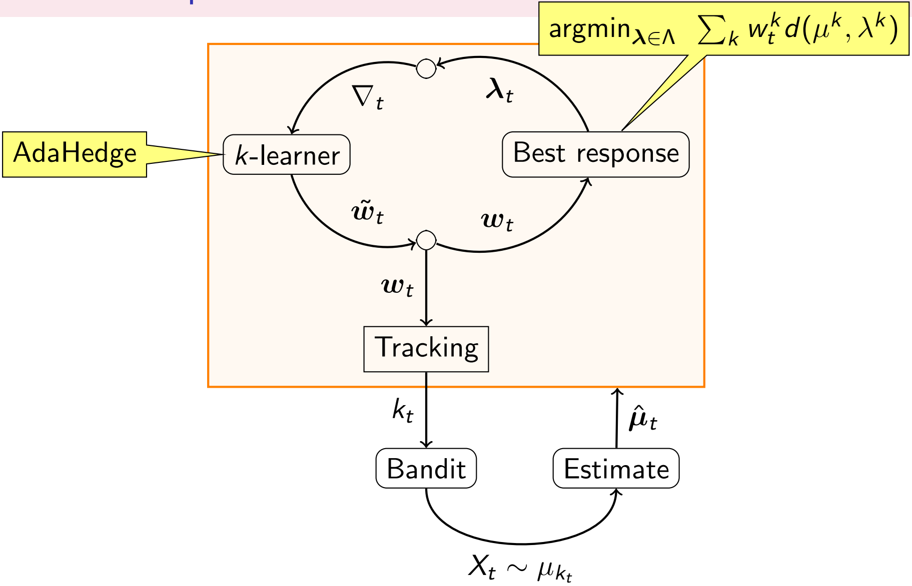
$\tilde{w}^k \propto N^k \Delta^k$
regret

$$= \inf_{\boldsymbol{q} \in \triangle(\Lambda)} \max_k \frac{\mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \left[d(\mu^k, \lambda^k) \right]}{\Delta^k}$$

Illustration



Overall Setup



Outline

Noise-free result

Let $\mathcal{B}_n^k$ be regret of full information online learning (AdaHedge) w. linear losses on the simplex.

Theorem

Consider running our algorithm until $\inf_{\boldsymbol{\lambda} \in \Lambda} \sum_{t=1}^n \sum_k w_t^k d(\mu^k, \lambda^k) \geq \ln T$. *The iterates* $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ *satisfy*

$$R_n = \sum_{t=1}^n \langle \boldsymbol{w}_t, \Delta \rangle \leq V_T + \frac{\mathcal{B}_n^k}{D^*}$$

Note

- Can get $k_1, \ldots, k_n$ using tracking (at cost $\Delta^{\max} \ln K$)
- Standard choice gives $n = O(\ln T)$ and $\mathcal{B}_n^k = O(\sqrt{n}) = O(\sqrt{\ln T}) = o(\ln T)$.

On Symmetry

Game-theoretic equilibrium is **symmetric** concept.

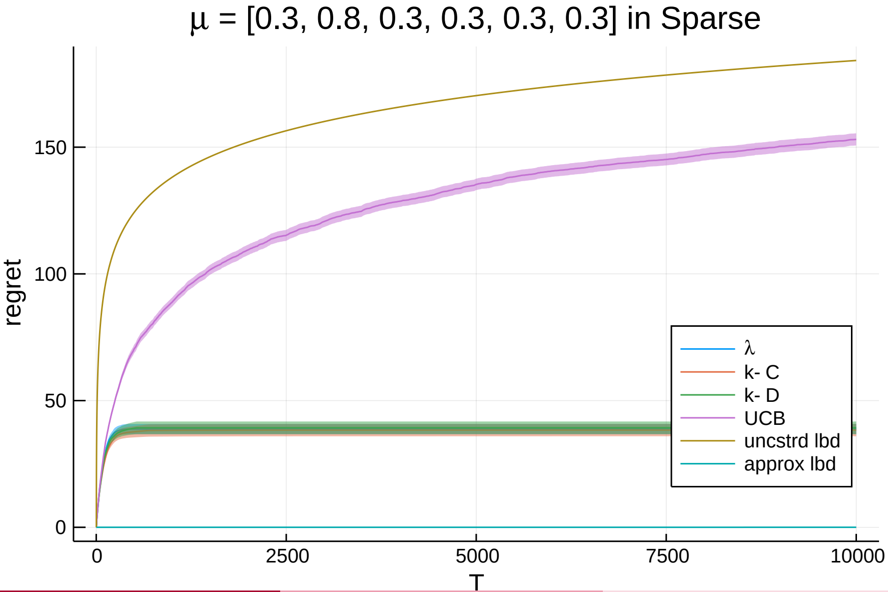
Can also focus on $\boldsymbol{\lambda}$-learner instead of k-learner. Interesting trade-offs

- More complex domain $\boldsymbol{\lambda} \in \Lambda$.
- No need for tracking, best response in k is “pure” arm.

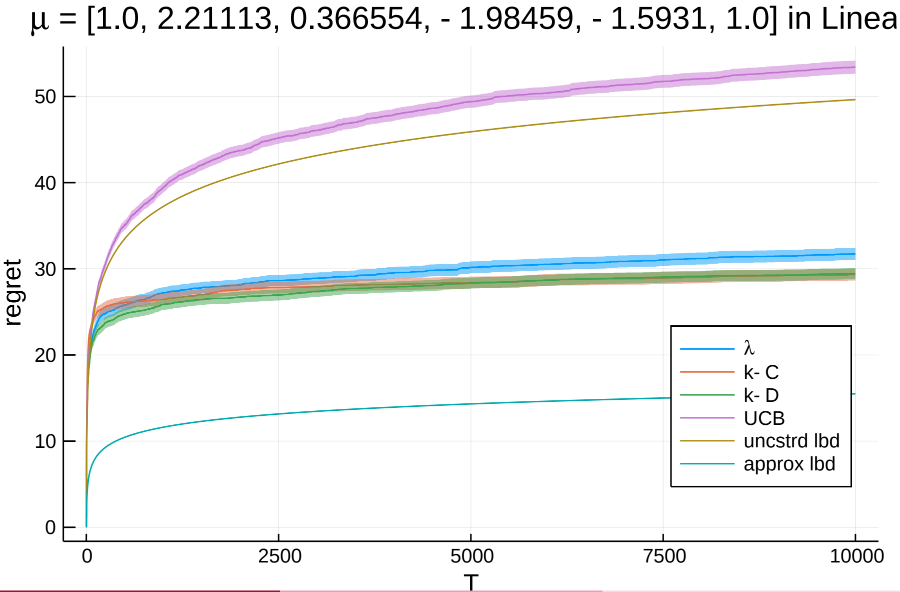
Will show both in experiments.

Outline

Experiment: Sparse



Experiment: Linear



Conclusion

Game equilibrium based technique for matching **instance dependent lower bounds** for structured stochastic bandits.
All you need is **Best Response oracle**.

- Fine tuning
- What about "lower-order" terms not scaling with $\ln T$?
- Is minigame interaction "easy data"? MetaGrad [Van Erven and Koolen, 2016]
- Minigames for other problems?

Conclusion

Game equilibrium based technique for matching **instance dependent lower bounds** for structured stochastic bandits.
All you need is **Best Response oracle**.

- Fine tuning
- What about "lower-order" terms not scaling with $\ln T$?
- Is minigame interaction "easy data"? MetaGrad [Van Erven and Koolen, 2016]
- Minigames for other problems?

Thank you!