Fast Pure Exploration via Frank-Wolfe

Po-An Wang[1], Ruo-Chun Tzeng[2], and Alexandre Proutiere[1]

Conference on Neural Information Processing Systems, 2021

[1]EECS, Division of Decision and Control System

[2]EECS, Division of Theoretical Computer Science
KTH Royal Institute of Technology

Pure exploration on structured bandits

Stochastic Multi-Armed Bandit (MAB)

K arms (K prob. distribution $\nu_1, \ldots, \nu_K$), the mean of ν_k is μ_k

ν_1

ν_2

ν_3

ν_4

ν_5

Stochastic Multi-Armed Bandit (MAB)

K arms (K prob. distribution $\nu_1, \ldots, \nu_K$), the mean of ν_k is μ_k

ν_1

ν_2

ν_3

ν_4

ν_5

In round t, an agent

1. pulls arm $A_t \in [K]$
2. receives the reward $X_{A_t}(t) \sim \nu_{A_t}$

Sequential sampling strategy: $A_t \in \mathcal{F}_t = \sigma[A_1, X_1, \ldots, A_{t-1}, X_{t-1}]$

Pure exploration with fixed confidence

Goal: Identify a certain answer $i^\star(\boldsymbol{\mu}) \in \mathcal{I}$

Example: Identify the best arm $i^\star(\boldsymbol{\mu}) = \mathrm{argmax}_{k\in[K]} \mu_k$

A strategy consists of

- a sampling rule A_t (arm to explore)
- a stopping rule τ (time to stop)
- a $\mathcal{F}_\tau$-measurable decision rule $\hat{\imath} \in \mathcal{I}$ (answer to return)

Pure exploration with fixed confidence

Goal: Identify a certain answer $i^\star(\boldsymbol{\mu}) \in \mathcal{I}$

Example: Identify the best arm $i^\star(\boldsymbol{\mu}) = \mathrm{argmax}_{k \in [K]} \mu_k$

A strategy consists of

- a sampling rule A_t (arm to explore)
- a stopping rule τ (time to stop)
- a $\mathcal{F}_\tau$-measurable decision rule $\hat{\imath} \in \mathcal{I}$ (answer to return)

We wish to minimize $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ subject to $\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq i^\star(\boldsymbol{\mu})] < \delta$

Structured bandits

"Side information" is encoded by the **structure**

Popular structures: Unstructured, Linear, Lipschitz, Dueling, Combinatorial, Unimodal, Monotone, Spectral and Cascading

Structured bandits

"Side information" is encoded by the **structure**

Popular structures: Unstructured, Linear, Lipschitz, Dueling, Combinatorial, Unimodal, Monotone, Spectral and Cascading

Question 1. What is the sample complex gain achievable when exploiting the structure?

Structured bandits

"Side information" is encoded by the **structure**

Popular structures: Unstructured, Linear, Lipschitz, Dueling, Combinatorial, Unimodal, Monotone, Spectral and Cascading

Question 1. What is the sample complex gain achievable when exploiting the structure?

Question 2. Can we devise a computational efficient algorithm achieving the promised gains for all structures?

Lower bound [GK16]

For any **good** strategy,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(\frac{1}{\delta})} \geq T^\star(\boldsymbol{\mu}),$$

where $T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \boldsymbol{\Sigma}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k)$

- Σ :$K-1$ simplex
- $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \boldsymbol{\Lambda} : i^\star(\boldsymbol{\lambda}) \neq i^\star(\boldsymbol{\mu})\}$
- $d(\mu_k, \lambda_k)$: KL-divergent of arm-k reward distribution under $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$

Lower bound [GK16]

For any **good** strategy,

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(\frac{1}{\delta})} \geq T^{\star}(\boldsymbol{\mu}),$$

where $T^{\star}(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega} \in \boldsymbol{\Sigma}} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k)$

- Σ : $K-1$ simplex
- $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} \in \boldsymbol{\Lambda} : i^{\star}(\boldsymbol{\lambda}) \neq i^{\star}(\boldsymbol{\mu})\}$
- $d(\mu_k, \lambda_k)$: KL-divergent of arm-k reward distribution under $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$

$\Rightarrow$ An optimal algorithm has a sampling strategy described by

$$\boldsymbol{\omega}^{\star}(\boldsymbol{\mu}) = \operatorname*{argmax}_{\boldsymbol{\omega} \in \boldsymbol{\Sigma}} F_{\boldsymbol{\mu}}(\boldsymbol{\omega}),$$

$$\text{where} \quad F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} \omega_k d(\mu_k, \lambda_k)$$

Frank-Wolfe based sampling (FWS)

Frank-Wolfe based sampling

- Devise a simple algorithm (FW-based) to track $\boldsymbol{x(t)} \xrightarrow{t\to\infty} \boldsymbol{\omega}^\star(\boldsymbol{\mu})$

Frank-Wolfe based sampling

- Devise a simple algorithm (FW-based) to track $\boldsymbol{x(t)} \xrightarrow{t\to\infty} \boldsymbol{\omega}^\star(\boldsymbol{\mu})$
- **Envelope theorem** shows that $F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \min_{j\in J} f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$, where J is a finite set and $f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$ is smooth $\forall j \in J$ ($F_{\boldsymbol{\mu}}$ **is non-smooth**)

Frank-Wolfe based sampling

- Devise a simple algorithm (FW-based) to track $\boldsymbol{x(t)} \xrightarrow{t\to\infty} \omega^\star(\mu)$
- **Envelope theorem** shows that $F_{\mu}(\omega) = \min_{j\in J} f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$, where J is a finite set and $f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$ is smooth $\forall j \in J$ (**F_{μ} is non-smooth**)
- To deal with non-smoothness, define

$$H_{F_\mu}(\omega, r) = \mathrm{cov}\left\{\nabla_\omega f_j(\omega, \mu) : j \in \mathcal{J}, f_j(\omega, \mu) < F_\mu(\omega) + r\right\}$$

Frank-Wolfe based sampling

- Devise a simple algorithm (FW-based) to track $\boldsymbol{x(t)} \xrightarrow{t\to\infty} \boldsymbol{\omega}^\star(\boldsymbol{\mu})$
- **Envelope theorem** shows that $F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \min_{j\in J} f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$, where J is a finite set and $f_j(\boldsymbol{\omega}, \boldsymbol{\mu})$ is smooth $\forall j \in J$ ($F_{\boldsymbol{\mu}}$ **is non-smooth**)
- To deal with non-smoothness, define

$$H_{F_{\boldsymbol{\mu}}}(\boldsymbol{\omega}, r) = \mathrm{cov}\left\{\nabla_{\boldsymbol{\omega}} f_j(\boldsymbol{\omega}, \boldsymbol{\mu}) : j \in \mathcal{J}, f_j(\boldsymbol{\omega}, \boldsymbol{\mu}) < F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) + r\right\}$$

- Update

$$\begin{cases} \boldsymbol{z}(t+1) \leftarrow \mathrm{argmax}_{\boldsymbol{z}\in\boldsymbol{\Sigma}} \min_{h\in H_{F_{\boldsymbol{\mu}}(\boldsymbol{x}(t), r_t)}} \langle \boldsymbol{z} - \boldsymbol{x}(t), h\rangle, \\ \boldsymbol{x}(t+1) \leftarrow \frac{t}{t+1}\boldsymbol{x}(t) + \frac{1}{t+1}\boldsymbol{z}(t+1) \end{cases}$$

FWS

Input: Confidence level δ, sequence $\{r_t\}_{t\geq 1}$

Initialization: Sample each arm once and update $\boldsymbol{\omega}(K), \boldsymbol{x}(K) = (\frac{1}{K}, \dots, \frac{1}{K})$, and $\hat{\boldsymbol{\mu}}(K)$
$t \leftarrow$K

While $tF_{\hat{\boldsymbol{\mu}}(t)}(\boldsymbol{\omega}(\boldsymbol{t}) < \beta(\delta, t)$ ←Stopping criteria or $\hat{\boldsymbol{\mu}}(t-1) \notin \Lambda$

IF $\sqrt{\lfloor t/K \rfloor} \in \mathbb{N}$ or $\hat{\boldsymbol{\mu}}(t-1) \notin \Lambda$, (Forced exploration) $\boldsymbol{z}(t) \leftarrow (\frac{1}{K}, \dots, \frac{1}{K})$

Else, (FW update)

$$\boldsymbol{z}(t) \leftarrow \operatorname*{argmax}_{\boldsymbol{z}\in\boldsymbol{\Sigma}} \min_{h \in H_{F_{\hat{\boldsymbol{\mu}}(t-1)}}(\boldsymbol{x}(t-1), r_t)} \langle \boldsymbol{z} - \boldsymbol{x}(t-1), h \rangle$$

Update $\boldsymbol{x}(t) \leftarrow \frac{t-1}{t}\boldsymbol{x}(t-1) + \frac{1}{t}\boldsymbol{z}(t)$

Sample $A_t \leftarrow \text{argmax}_k\, \boldsymbol{x}_k(t)/\boldsymbol{\omega_k}(t-1)$ (ties broken arbitrarily)
Update $\boldsymbol{\omega}(t)$ and $\hat{\boldsymbol{\mu}}(t)$

Output: $i^\star(\hat{\boldsymbol{\mu}}(t))$

Theoretical Results

Asymptotic optimality of FWS

Theorem

For most pure exploration problems in structured bandits, FWS satisfies:

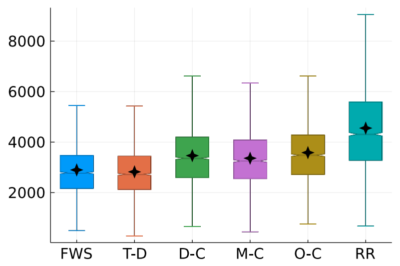
$\mathbb{P}_{\boldsymbol{\mu}}[\hat{\imath} \neq i^\star(\boldsymbol{\mu})] < \delta$ *and* $\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\log(\frac{1}{\delta})} \leq T^\star(\boldsymbol{\mu})$

With further assumptions, we can provide **non-asymptotic** upper bound for $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

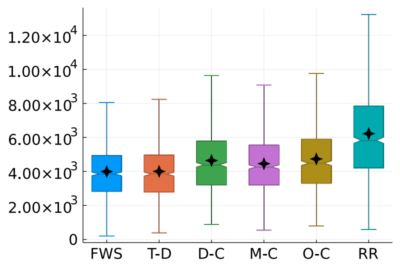
Numerical Results

Experiment (i) Unstructured bandits

Averaged sample complexity at $\delta = 0.01$



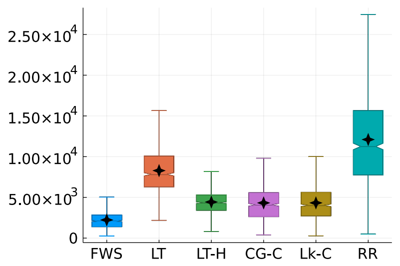
Bernoulli



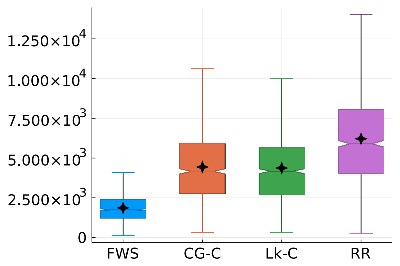
Gaussian

Experiment (ii) Linear bandits

Averaged sample complexity at $\delta = 0.01$



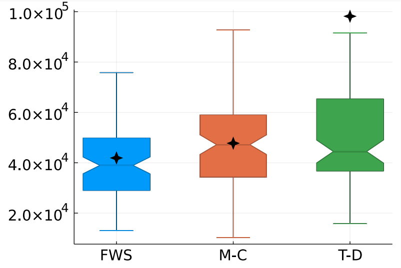
BAI



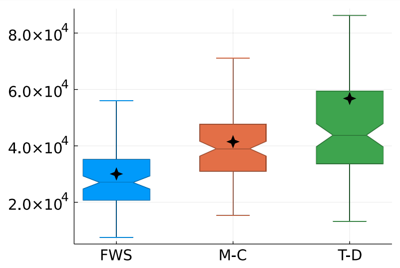
ThresholdingBandit

Experiment (iii) Lipschitz bandits

Averaged Sample complexity at $\delta = 0.01$



Experiment 1



Experiment 2

This is the first result for Lipschitz bandits in literatures

Related work and conclusion

Related works:

- LMA [Mén19]: Apply mirror ascent to update $\boldsymbol{x}(t)$
- Gamification [DMSV20, Sha21, JMKK21]: Use 2 player game to reach $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$

Unclear to extend the above approaches to general structures

Related work and conclusion

Related works:

- LMA [Mén19]: Apply mirror ascent to update $\boldsymbol{x}(t)$
- Gamification [DMSV20, Sha21, JMKK21]: Use 2 player game to reach $\boldsymbol{\omega}^\star(\boldsymbol{\mu})$

Unclear to extend the above approaches to general structures

Conclusion:

- FWS is computationally and statistically efficient for general pure exploration problems
- Theoretically, FWS matchs the instance-specific lower bounds
- Numerically, FWS outperforms all the other optimal algorithms in structured bandits

Reference

- Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko, *Gamification of pure exploration for linear bandits*, Proc. of ICML, 2020.
- Aurélien Garivier and Emilie Kaufmann, *Optimal best arm identification with fixed confidence*, Proc. of COLT, 2016.
- Marc Jourdan, Mojmír Mutnỳ, Johannes Kirschner, and Andreas Krause, *Efficient pure exploration for combinatorial bandits with semi-bandit feedback*, Proc. of ALT, 2021.
- Pierre Ménard, *Gradient ascent for active exploration in bandit problems*, arXiv (2019).
- Xuedong Shang, *Linbai: Gamification of pure exploration for linear bandits*, `https://github.com/xuedong/LinBAI.jl`, 2021, [Online; accessed 09-May-2021].