Closing the Computational-Statistical Gap in BAI for Combinatorial Semi-bandits

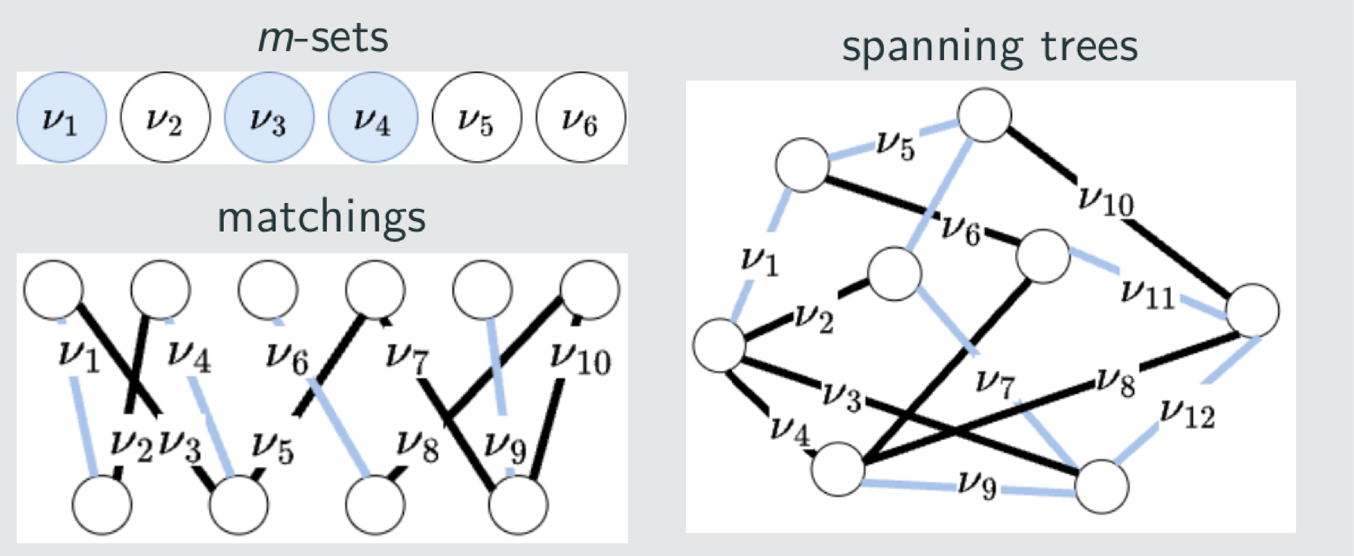
Ruo-Chun Tzeng[1], Po-An Wang[1], Alexandre Proutiere[1]   Chi-Jen Lu[2]

[1]EECS, KTH Royal Institute of Technology, Sweden
[2]Institute of Information Science, Academia Sinica, Taiwan

Combinatorial BAI with semi-bandit feedback

Input: K arms $(\nu_k)_{k\in[K]}$ with mean $\boldsymbol{\mu}\in\mathbb{R}^K$ and $\mathcal{X}\subseteq\{0,1\}^K$.

Example: Gaussian reward $\nu_k = \mathcal{N}(\mu_k, 1), \forall k\in[K]$



Rule: At each round $t\in\mathbb{N}$, the learner pulls an action $\boldsymbol{x}(t)\in\mathcal{X}$ and observe $y_k(t)\sim\nu_k$ iff $x_k(t)=1$, and returns her estimated best action $\hat{\imath}\in\mathcal{X}$ when she decides to terminate at round τ.

Goal: Design a δ-PAC learning strategy s.t. the best action $\boldsymbol{i}^\star(\boldsymbol{\mu})\in\operatorname{argmax}\langle\boldsymbol{\mu},\boldsymbol{x}\rangle$ is identified w.p. $\geq 1-\delta$ and $\mathbb{P}_{\boldsymbol{\mu}}[\tau<\infty]=1$ while minimizing $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Prior works: a computational-statistical gap

Any δ-PAC algorithm satisfies $\mathbb{E}_{\boldsymbol{\mu}}[\tau]\geq T^\star(\boldsymbol{\mu})\mathrm{kl}(\delta,1-\delta)$, where

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{\boldsymbol{\omega}\in\Sigma} F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) \text{ with } F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \inf_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\mu})}\sum_{k=1}^K \frac{\omega_k(\mu_k-\lambda_k)^2}{2}.$$

Track-and-Stop [6] is statistically optimal but requires to repeatedly solve $T^\star(\hat{\boldsymbol{\mu}}(t-1))^{-1}\Rightarrow$ computationally inefficient.

FWS [8] at the FW-update round has to solve a potentially $\mathcal{O}(2^K)$ many convex programs $\Rightarrow$ computationally inefficient

CombGame [7] is MCP-oracle efficient and statistically optimal $\Rightarrow$ left open the design of an efficient MCP-oracle

Designing efficient MCP based on a structural observation

Let $f_{\boldsymbol{x}}(\boldsymbol{\omega},\boldsymbol{\mu}) = \inf_{\boldsymbol{\lambda}\in\mathbb{R}:\langle \boldsymbol{i}^\star(\boldsymbol{\mu})-\boldsymbol{x},\boldsymbol{\lambda}\rangle<0}\sum_{k=1}^K\frac{\omega_k(\mu_k-\lambda_k)^2}{2}$ s.t. $F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \min_{\boldsymbol{x}\neq\boldsymbol{i}^\star(\boldsymbol{\mu})} f_{\boldsymbol{x}}(\boldsymbol{\omega},\boldsymbol{\mu})$.

Property of $f_{\boldsymbol{x}}$ and its Lagrangian dual $g_{\boldsymbol{\omega},\boldsymbol{\mu}}$:

$f_{\boldsymbol{x}}(\boldsymbol{\omega},\boldsymbol{\mu}) = \max_{\alpha\geq 0} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x},\alpha)$ (known by [2])

$g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x},\alpha)$ is linear in $\boldsymbol{x}$ and concave in α (our observation)

These properties $\Rightarrow F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \min_{\boldsymbol{x}\neq\boldsymbol{i}^\star(\boldsymbol{\mu})}\max_{\alpha\geq 0} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x},\alpha)$ as a two-player zero-sum game.

We not only want to estimate $F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$ but also the *equilibrium* action $\boldsymbol{x}_e$ s.t. $F_{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \max_{\alpha\geq 0} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x}_e,\alpha)$. This rules out many existing results from applying.

- $\boldsymbol{x}_e$ is required to solve $\max_{\boldsymbol{\omega}\in\Sigma} F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$ by the first-order methods
- Last-iterate convergence [1, 3] are mostly for saddle-point problems

Algorithm 1: (ϵ,θ)-MCP$(\boldsymbol{\omega},\boldsymbol{\mu})$

for $n=1,2,\cdots$ **do**

(Follow-the-Perturbed-Leader) $\boldsymbol{\mathcal{Z}}_n\sim\exp(1)^K$ and $\eta_n = \frac{c_0}{\sqrt{n}}$

$$\boldsymbol{x}^{(n)}\in\operatorname*{argmin}_{\boldsymbol{x}\neq\boldsymbol{i}^\star(\boldsymbol{\mu})}\left(\sum_{m=1}^{n-1} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x},\alpha^{(m)}) + \frac{\langle\boldsymbol{\mathcal{Z}}_n,\boldsymbol{x}\rangle}{\eta_n}\right)$$

(Best-Response) $\alpha^{(n)}\in\operatorname*{argmax}_{\alpha\geq 0} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x}^{(n)},\alpha)$

if $\sqrt{n} > \frac{c_\theta(1+\epsilon)}{\epsilon\hat{F}}$, *where* $\begin{cases}\hat{F} = g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x}^{(n_\star)},\alpha^{(n_\star)})\\ n_\star\in\operatorname{argmin}_{m\leq n} g_{\boldsymbol{\omega},\boldsymbol{\mu}}(\boldsymbol{x}^{(m)},\alpha^{(m)})\end{cases}$

then return $(\hat{F},\boldsymbol{x}^{(n_\star)})$;

end

(Theorem 1) Let $\epsilon,\theta\in(0,1)$ and $(\boldsymbol{\omega},\boldsymbol{\mu})\in\Sigma_+\times\Lambda$.

- $\mathbb{P}_{\boldsymbol{\mu}}\left[F_{\boldsymbol{\mu}}(\boldsymbol{\omega})\leq\hat{F}\leq(1+\epsilon)F_{\boldsymbol{\mu}}(\boldsymbol{\omega})\right]\geq 1-\theta$
- the number of $\boldsymbol{i}^\star$-oracle calls: $\mathcal{O}\left(\frac{K^3D^5\ln K\ln\theta^{-1}\|\boldsymbol{\mu}\|_\infty^4\left\|\boldsymbol{\omega}^{-1}\right\|_\infty^2}{\epsilon^2F_{\boldsymbol{\mu}}(\boldsymbol{\omega})^2}\right)$

Our Perturbed Frank-Wolfe Sampling (P-FWS)

We use stochastic smoothing [5, 4] to overcome the nonsmoothness of $F_{\boldsymbol{\mu}}$ as: all we need is $\boldsymbol{i}^\star$-oracle and its required graident can be evaluated by envelope theorem [8].

The smoothed objective $\bar{F}_{\boldsymbol{\mu},\eta}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\mathcal{Z}}\sim\mathrm{Uniform}(B_2)}[F_{\boldsymbol{\mu}}(\boldsymbol{\omega}+\eta\boldsymbol{\mathcal{Z}})]$ satisfies:

- $\nabla\bar{F}_{\boldsymbol{\mu},\eta}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\mathcal{Z}}\sim\mathrm{Uniform}(B_2)}[\nabla F_{\boldsymbol{\mu}}(\boldsymbol{\omega}+\eta\boldsymbol{\mathcal{Z}})]$
- $\bar{F}_{\boldsymbol{\mu},\eta}$ is $\frac{\ell K}{\eta}$-smooth and $\bar{F}_{\boldsymbol{\mu},\eta}(\boldsymbol{\omega})\xrightarrow{\eta\downarrow 0}F_{\boldsymbol{\mu}}(\boldsymbol{\omega})$

High-level design of P-FWS

Let $\mathcal{X}_0$ be a set s.t. $\forall k\in[K]$, there exists $\boldsymbol{x}\in\mathcal{X}_0$ s.t. $x_k=1$.

P-FWS alternate between two phases:

$\begin{cases}\text{pull each } \boldsymbol{x}\in\mathcal{X}_0 \text{ once} & \text{(to avoid high cost and boundary cases)}\\ \text{pull } \boldsymbol{x}(t)\in\operatorname{argmax}_{\boldsymbol{x}\in\mathcal{X}}\langle\nabla\bar{F}_{\hat{\boldsymbol{\mu}}(t-1),\eta_t}(\hat{\boldsymbol{\omega}}(t-1)),\boldsymbol{x}\rangle & \text{(ideal FW update)}\end{cases}$

(Theorem 2) Let $\boldsymbol{\mu}\in\Lambda$ and $\delta\in(0,1)$. P-FWS is δ-PAC, finishes in finite time, $\mathbb{P}_{\boldsymbol{\mu}}\left[\limsup_{\delta\to 0}\frac{\tau}{\ln\delta^{-1}}\leq T^\star(\boldsymbol{\mu})\right]=1$, $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ is bounded by Poly(K) in moderate-confidence regime and achieves the minimal in high-confidence regime, and the total number of $\boldsymbol{i}^\star$-oracle calls is bounded by Poly(K).

References

[1] K. Abe, K. Ariu, M. Sakamoto, K. Toyoshima, and A. Iwasaki. Last-iterate convergence with full-and noisy-information feedback in two-player zero-sum games. In *Proc. of AISTATS*, 2023.

[2] L. Chen, A. Gupta, and J. Li. Pure exploration of multi-armed bandit under matroid constraints. In *Proc. of COLT*, 2016.

[3] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *Proc. of ITCS*, 2019.

[4] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 2012.

[5] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. of SODA*, 2005.

[6] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Proc. of COLT*, 2016.

[7] M. Jourdan, M. Mutnỳ, J. Kirschner, and A. Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Proc. of ALT*, 2021.

[8] P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast pure exploration via frank-wolfe. In *Proc. of NeurIPS*, 2021.