FY2025/2025年度

Sequential Decision Making Team

Shinji Ito

逐次的意思決定チーム　伊藤 伸志

RIKEN Center for Advanced Intelligence Project

Online Inverse Linear Optimization: Efficient Logarithmic-Regret Algorithm, Robustness to Suboptimality, and Lower Bound [Sakaue, Tsuchiya, Bao, Oki, NeurIPS'25]

Problem Setting

For $t = 1, \dots, T$:

Make prediction $\hat{c}_t \in \mathbb{B}^n$ of c^*.

Learner

Observe feasible region X_t and take action $x_t \in \operatorname{argmax}_{x \in X_t} \langle c^*, x \rangle$. Agent

Forward optimization

Observe (X_t, x_t) and update from $\hat{c}_t$ to $\hat{c}_{t+1}$.

$c^* \in \mathbb{B}^n$: linear objective vector of forward optimization.

$X_t \subseteq \mathbb{B}^n$: feasible region of forward optimization.

Regret (Opt. val. / Val. of learner's action)

$R_T^{c^*} := \sum_{t=1}^T \langle c^*, x_t - \hat{x}_t \rangle = \sum_{t=1}^T \langle c^*, x_t \rangle - \sum_{t=1}^T \langle c^*, \hat{x}_t \rangle$ for $\hat{x}_t \in \arg\max \{\langle \hat{c}_t, x \rangle \mid x \in X_t\}$.

Our results

- ONS-based method with $R_T^{c^*} = O(n \log T)$; the per-bound complexity is independent of T.
  - –The first method to have these two desirable properties!
- MetaGrad-based method to handle suboptimality of x_t.
- Lower bound of $R_T^{c^*} = \Omega(n)$.

Optimal Estimation of the Best Mean in Multi-Armed Bandits [Osogami, Honda, Komiyama, NeurIPS'25]

Goals

- Return an interval of width 2ε that contains the best mean $\mu^\star := \max_{i \in [K]} \mu_i$ with probability $\geq 1 - \delta$
- Minimize the sample complexity $\mathbb{E}[\tau]$

Applications: Estimating the worst performance of AI models across tasks, users, etc. with minimal test cases

Main results

Lower bound: For Gaussian rewards with variance R^2, the sample complexity $\mathbb{E}[\tau]$ of any correct algorithm must satisfy

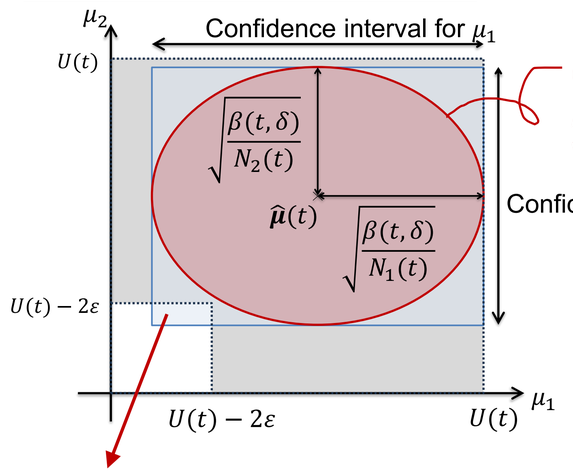
$$\liminf_{\delta \downarrow 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} \geq 2R^2 f(\boldsymbol{\mu})$$

Upper bound: For R-sub-Gaussian rewards, EllipsoidEst is correct, and its sample complexity satisfies

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}[\tau]}{\log(1/\delta)} = 2R^2 f(\boldsymbol{\mu})$$

Key ideas

Stopping rule based on confidence ellipsoid



Confidence ellipsoid: $\boldsymbol{\mu}$ is in the ellipsoid for all t with high probability

$$\hat{\mu}_i(t) := \frac{1}{N_i(t) + \lambda} \sum_{m=1}^{N_i(t)} x_{i,m}$$

$$\beta(t,\delta) := 2R^2 \log(1/\delta) + KR^2 \log t + K\lambda S^2$$

$\Longrightarrow$ Stop sampling when

$$\sum_{i \in [K]} N_i(t) (\hat{\mu}_i(t) - U(t) + 2\varepsilon)_+^2 \geq \beta(t,\delta)$$

for correctness

Geometric Resampling in Nearly Linear Time for Follow-the-Perturbed-Leader with Best-of-Both-Worlds Guarantee in Bandit Problems [Chen, Lee, Honda, ICML'25]

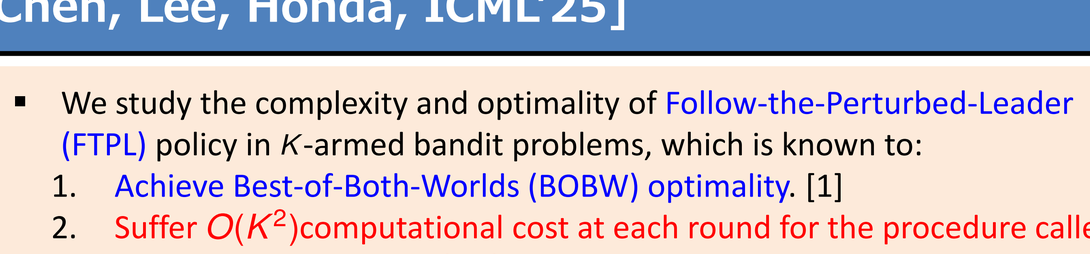
- We study the complexity and optimality of Follow-the-Perturbed-Leader (FTPL) policy in K-armed bandit problems, which is known to:
  1. Achieve Best-of-Both-Worlds (BOBW) optimality. [1]
  2. Suffer $O(K^2)$computational cost at each round for the procedure called Geometric Resampling (GR).

  → It's important to improve the computational efficiency.
- We propose a novel technique to replace GR.

  → Reduce the average complexity to $O(K \log K)$ with better regret.

Algorithms	Complexity	# of resampling $\mathbb{E}[M_t \mid \hat{L}_t, I_t]$	Worst-case M_t	Computational Requirement	Regret Bounds
GR [Neu+2016]	$O(K^2)$	Unbounded	Unbounded	None	BOBW
CGR I	$O(K \log K)$	Unbounded	Unbounded	None	BOBW
CGR II-unbiased	$O(K \log K)$	$O(K)$	Unbounded	Explicit $F(x)$ and $F^{-1}(x)$	BOBW
CGR II-biased	$O(K \log K)$	$O(K)$	$\lceil (K \vee 4) \log t \rceil$	Explicit $F(x)$ and $F^{-1}(x)$	BOBW

(Under Fréchet distribution with $\alpha = 2$)



Experimental results in **adversarial setting**; results in stochastic setting (shown in paper) is similar.

Regret of FTPL: CGR II $\leq$ CGR I $\leq$ Original GR, thanks to the smaller variance.

CGR I Samples More Than CGR II but Runs Faster:

- **Expected number of resampling:** CGR II < CGR I.
- **Cost of resampling once:** CGR I < CGR II (though both are $O(K)$).

→ **Runtime:** CGR I always runs faster than CGR II.

Corrupted Learning Dynamics in Games [Tsuchiya, Ito, Luo, COLT'25]

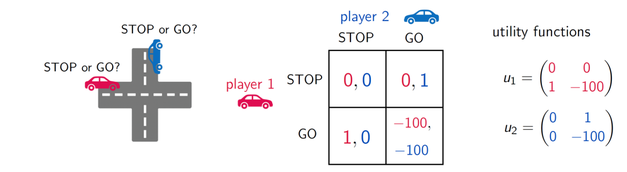
Learning in games

Multiple players interact in a shared environment,
each aiming to maximize their total rewards (= minimize their *regret*)
by iteratively adapting their strategies based on repeated interactions

Broader applications

- Minimax optimization (*e.g.*, $\min_x \max_y x^\top A y$)
- Multi-agent reinforcement learning
- Superhuman AI for poker, human-level AI for Stratego
- Alignment of LLMs
- …

Example 2. Game of chicken: two-player "STOP-or-GO" at intersection game



Nash eq: (pure) (STOP,GO), (GO,STOP),
(mixed) STOP with prob. 100/101 and GO with prob. 1/101

Research questions

- Can we adapt to deviations of the opponent from a given algorithm?
- Can we characterize regret and convergence rates to an equilibrium in such a corrupted game?

Our contributions

- Establish a framework of corrupted games, in which each player may deviate from a prescribed algorithm
- Give a nearly complete characterization of learning dynamics in corrupted games, by deriving regret upper and lower bounds in (normal-form) two-player zero-sum and multi-player general-sum games

Swap regret upper bounds of player i in multi-player general-sum games with n-players and m-actions after T rounds

$\widehat{C}_i \in [0, 2T]$: the cumulative amount of corruption in strategies for player i, $\widehat{S} = \sum_{i \in [n]} \widehat{C}_i$

References	Honest	Corrupted (if no corruption in observed utilities)
Chen and Peng (2020)	$\sqrt{n}(m \log m)^{3/4} T^{1/4}$	$\sqrt{mT \log m} + \widehat{C}_i$
Anagnostides et al. (2022)	$nm^{5/2} \log T$	$nm^{5/2} \log T + \sqrt{mT \log m} + \widehat{C}_i$
Ours	$nm^{5/2} \log T$	$nm^{5/2} \log T + \min\left\{\sqrt{\widehat{S}(nm^2 + m^{5/2}) \log T}, m\sqrt{T \log T}\right\} + \widehat{C}_i$