FY2023/2023年度

Sequential Decision-Making Team
Shinji Ito
逐次的意思決定チーム　伊藤 伸志

Center for Advanced Intelligence Project

逐次的意思決定チームでは，予測の不確実性や環境の変動の中で，逐次的に合理的な判断を下すためのアルゴリズムや理論の開発に取り組んでいます．本ポスターでは，さまざまな環境に自動的に適応し，環境の特性に応じた性能向上を達成する**環境適応的アルゴリズム**および**両環境最適アルゴリズム**についての研究成果を紹介します．

Stability-penalty-adaptive Follow-the-regularized-leader: Sparsity, Game-dependency, and Best-of-both-worlds

NeurIPS2023

*投稿時所属（括弧内現所属）

Taira Tsuchiya[1,2 (4)], Shinji Ito[3 (3,2)], Junya Honda[1,2] (1. Kyoto University, 2. RIKEN AIP, 3. NEC, 4. The University of Tokyo)

オンライン意思決定問題における Follow-the-Regularized-Leader の環境適応性 ~ 多腕バンディット問題を例に ~

多腕バンディット問題 (MAB)

k 個のスロットマシンが用意され合計 T 回プレイし，累積報酬を最大化（= 累積損失を最小化）する問題

敵対者が損失ベクトル $\ell_1, \dots, \ell_T \in [0,1]^k$ を決定

各ラウンド $t = 1, \dots, T$:

1. プレイヤーがアーム $A_t \in \{1, \dots, k\}$ を選択
2. アーム A_t の損失 $\ell_{t,A_t} \in [0,1]$ を観測

目標：累積損失の最小化 = **リグレット** R_T の最小化

$$a^* = \arg\min_{a \in \{1,\dots,k\}} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right]$$

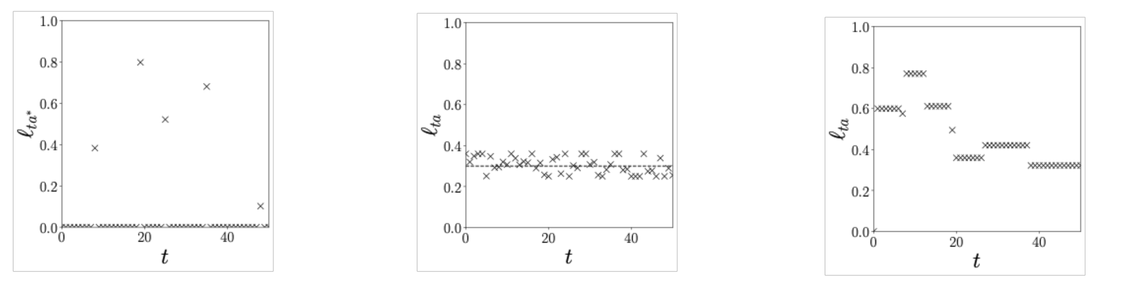
$$R_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t} - \sum_{t=1}^T \ell_{t,a^*}\right]$$

バンディット問題における環境

確率的環境	$\ell_{t,a} \sim \nu_a^*$ for all $a \in [k]$
汚染のある確率的環境	中間の環境
敵対的環境	$\ell_1, \dots, \ell_T \in [0,1]^k$ は任意

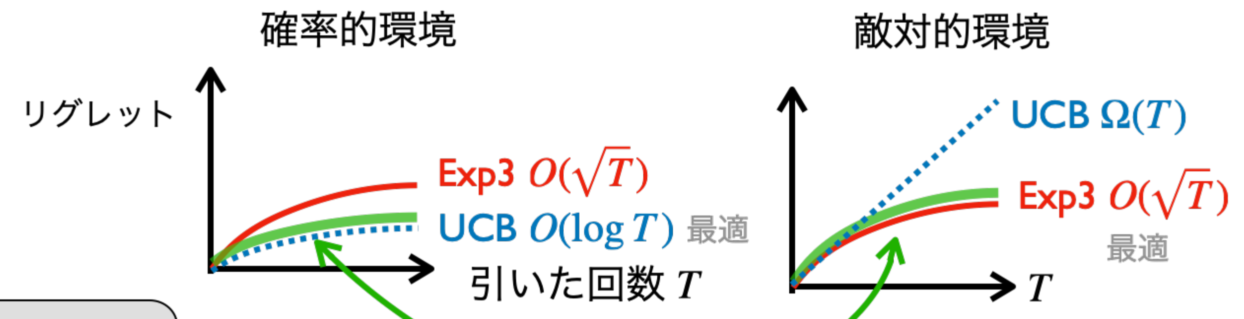
環境適応性：損失系列の性質に適応的意思決定による性能改善

データ依存上界：敵対的環境における損失の良性度合いに依存した上界



$\min_{a \in [k]} \sum_{t=1}^T \ell_{ta}$ が小　$\min_{\bar{\ell} \in [0,1]^k} \sum_{t=1}^T \|\ell_t - \bar{\ell}\|^2$ が小　$\sum_{t=2}^T \|\ell_t - \ell_{t-1}\|$ が小

両環境最適性：確率的環境と敵対的環境に対する同時最適性



プレイヤーは背後の環境を知らない

望ましい方策：確率的・敵対的環境両方で高性能 = Best-of-Both-Worlds (BOBW, 両環境最適性)

研究背景

- これらの環境適応性の多くは Follow-the-Regularized-Leader (FTRL) で実現可能 [Wei & Luo 18, Zimmert & Seldin 21, etc]
- それらは FTRL の正則化関数，**学習率**を設計することで実現
- 複数の環境適応性を同時に達成することができるアルゴリズムは僅か

研究課題

Q. **データ依存上界とBOBW を同時に達成**するような アルゴリズムを設計可能か？

A. **FTRL の学習率を複数の観測量に同時適応的に**することで実現可能！

→ **MAB** と**部分観測問題**で応用

Follow-the-Regularized-Leader と適応的学習率

Follow-the-Regularized-Leader (FTRL)

- FTRL：アーム選択確率 $p_t \in \mathscr{P}_k$ を "累積推定損失 + 正則化" の最小化で決定

累積推定損失　強凸正則化関数

$$p_t = \arg\min_{p \in \mathscr{P}_k} \left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p \right\rangle + \frac{1}{\eta_t} \phi_t(p)$$

$\hat{\ell}_s \in \mathbb{R}^k : \ell_s$ の推定量

$\mathscr{P}_k$: $(k-1)$-次元確率単体　学習率

- 多くのデータ依存上界や BOBW 保証は FTRL に基づく方策がほとんど
- 学習率 η_t をこれまでの観測系列に基づいて決定することで実現
  → **適応的学習率**と呼ばれる

適応的学習率：エントロピー正則化 $\phi_t(p) = \sum_{a=1}^k p_a \log p_a$ の場合

- 学習率 $(\eta_t)_{t=1}^T$ を用いたFTRL のリグレット上界の主要なパートは 以下の $\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}$ の期待値からなる

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \underbrace{h_{t+1}}_{\text{penalty}} + \sum_{t=1}^T \eta_t \underbrace{z_t}_{\text{stability}}$$

penalty: 正則化強で大

stability: p_t と p_{t+1} の差が大のとき大

- 既存の FTRL の適応的学習率 $(\eta_t)_{t=1}^T$ は，penalty か stability のどちらかのみに依存
  - η_t を**経験的な** stability $(z_s)_{s=1}^{t-1}$ と**最悪ケース**の penalty $h_{\max}(\geq \max_{t \in [T]} h_t)$ で決定
    → データ依存上界を誘導 [McMahan 2011; Lattimore & Szepesvári 2020, and many!]
  - η_t を**経験的な** penalty $(h_s)_{s=1}^{t-1}$ と**最悪ケース**の stability $z_{\max}(\geq \max_{t \in [T]} z_t)$ で決定
    → Best-of-both-worlds を誘導 [Ito, Tsuchiya & Honda, 2022, Tsuchiya, Ito & Honda 2023]

Q. Penalty と Stability の経験的な値に同時に適応的な学習率を構成できるか？

Stability と Penalty に同時適応的な学習率（SPA 学習率）

定義. (informal)

学習率 $(\eta_t)_{t=1}^T$ が *stability-penalty-adaptive (SPA)* 学習率 であるとは，
ある非負実数 $((h_t, z_t, \bar{z}_t))_{t=1}^T$ が適当な条件を満たし，以下の形で書けることをいう：

$$\beta_t = \frac{1}{\eta_t}, \quad \beta_1 > 0, \quad \beta_{t+1} = \beta_t + \frac{c_1 z_t}{\sqrt{c_2 + \bar{z} h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}}$$

stability z_s と penalty h_{s+1} に同時に依存した設計

定理. (informal)

$(\eta_t)_{t=1}^T$ を SPA 学習率とする．これを構成する $((h_t, z_t, \bar{z}_t))_{t=1}^T$ が

Stability 条件：$\dfrac{\sqrt{c_2 + \bar{z}_t h_1}}{c_1}(\beta_1 + \beta_t) \geq \epsilon + z_t$ for all $t \in [T]$ for some $\epsilon > 0$

を満たすならば，

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \tilde{O}\left(\sqrt{c_2 + \bar{z}_t h_1 + \sum_{t=1}^T z_t h_{t+1}}\right)$$

stability z_s と penalty h_{s+1} に同時に依存したリグレット上界

Q. 実際に BOBW とデータ依存上界を同時に達成可能か？ → MAB, PM で検証

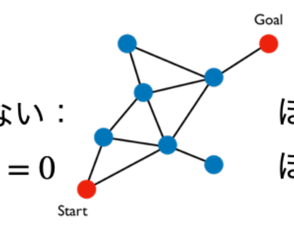
Case Study 1: 多腕バンディット問題におけるスパース性

スパース性とスパース性依存上界 (∈ データ依存上界)

- 損失 $\ell_t \in [-1,1]^k$ のスパース性 $s = \max_{t \in [T]} \|\ell_t\|_0 \ll k$ はあらゆる問題に登場

オンライン広告配置

ほとんどの広告はクリックされない：ほとんどの $a \in [k]$ で $r_{ta} := -\ell_{ta} = 0$



オンライン経路制御

ほとんどの経路でデータの損失なし　ほとんどの $a \in [k]$ で $\ell_{ta} = 0$

- スパース性依存上界：損失のスパース性 $s \ll k$ に依存したデータ依存上界

下界 $\Omega(\sqrt{sT})$ [Kwon & Perchet 2016]，上界 $O(\sqrt{sT \log k})$ [Kwon & Perchet 2016, Bubeck, Cohen & Li 2018]

スパース性依存上界と best-of-both-worlds の同時実現

定理. (informal) SPA学習率に基づくアルゴリズムを構成できて，

確率的環境：

$$R_T = O\left(\frac{s \log(T) \log(kT)}{\Delta_{\min}}\right)$$

敵対的環境：

$$R_T = O(\sqrt{sT \log(k) \log(T)})$$

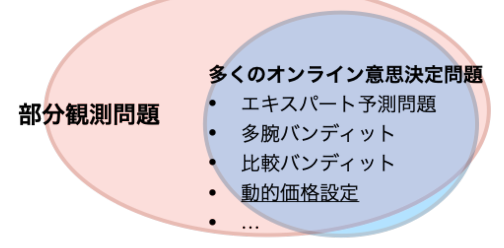
追加のテクニック：**1.** スパース性の推定, **2.** 負損失の対処，**3.** FTRL 出力の変化解析 ↓

$$R_T \lesssim \mathbb{E}\left[\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}\right] \underset{\text{SPA}}{\lesssim} \tilde{O}\left(\sqrt{\sum_{t=1}^T \mathbb{E}[z_t h_{t+1}]}\right) \underset{\text{補題. } h_{t+1} \lesssim h_t + \epsilon}{\lesssim} \tilde{O}\left(\sqrt{\sum_{t=1}^T \mathbb{E}[z_t h_t]}\right)$$

Case Study 2: 部分観測問題における経験的なゲーム依存性

部分観測問題とその例

抽象的なフィードバックの下で オンライン意思決定を行う**非常に一般的**な枠組み

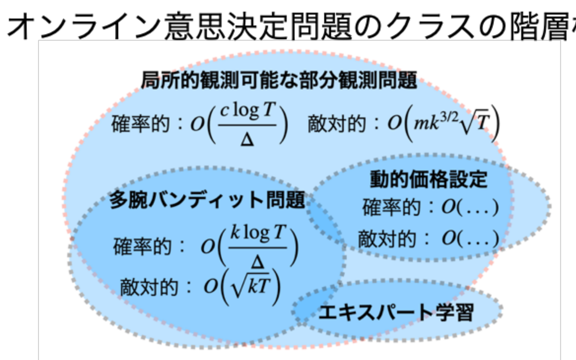


部分観測問題の課題とゲーム依存性の上界

課題. 部分観測問題は**アルゴリズム・定式化が保守的**で，実用的ではない

例. 二腕Bernoulliバンディット問題は $k \times 2^k$ 個のパラメータが必要

オンライン意思決定問題のクラスの階層構造



解いている問題の本来の難しさに依存した リグレットを自動で達成してほしい

= ゲーム依存性上界 [Lattimore & Szepesvári 2020]

経験的なゲーム依存性と best-of-both-worlds の同時実現

$$V_t' \simeq \min_{p \in \mathscr{P}_k'} \max_{x \in [d]} \left[\frac{(p - q_t)^\top L e_x}{\eta_t} + \underbrace{\frac{1}{\eta_t^2} \sum_{a=1}^k p_a \Psi_{q_t}\left(\frac{\eta_t G(a, \Phi_{ax})}{p_a}\right)}_{\text{stability 項}}\right] \leq \begin{cases} 1/2 & \text{if 完全情報設定} \\ k/2 & \text{if 多腕バンディット} \\ 3m^2k^3 & \text{if 局所観測部分観測問題} \end{cases} =: \bar{V}$$

$V_t', \bar{V}$: 問題に依存した変数

定理. (informal) 局所的観測可能なとき，SPA学習率に基づくアルゴリズムは，

確率的環境：

$$R_T = O\left(\frac{r_{\mathscr{M}} \bar{V} \log(T) \log(kT)}{\Delta_{\min}}\right)$$

敵対的環境：

$$R_T = O\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^T V_t' \log(k) \log(T)}\right]\right)$$

既存のアルゴリズムでは の部分が，最も難しい問題の最悪ケースの値に定まっていたが，仮にゲームが簡単なゲーム（未知かもしれない）に属するなら，それに対応した値になる

参考文献

C.Y. Wei & H. Luo. More adaptive algorithms for adversarial bandits. In COLT 2018.

J. Zimmert & Y. Seldin. "Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits." JMLR 2021.

J. Kwon & V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. JMLR 2016.

S. Bubeck, M. Cohen, & Y. Li. Sparsity, variance and curvature in multi-armed bandits. In ALT 2018.

T. Lattimore & Csaba Szepesvári. "Exploration by optimisation in partial monitoring." In COLT 2020.