組合せ半バンディット問題における適応的 best-of-both-worlds 方策

（最近のバンディット問題における best-of-both-worlds 方策の進展）

土屋 平

京都大学 大学院情報学研究科

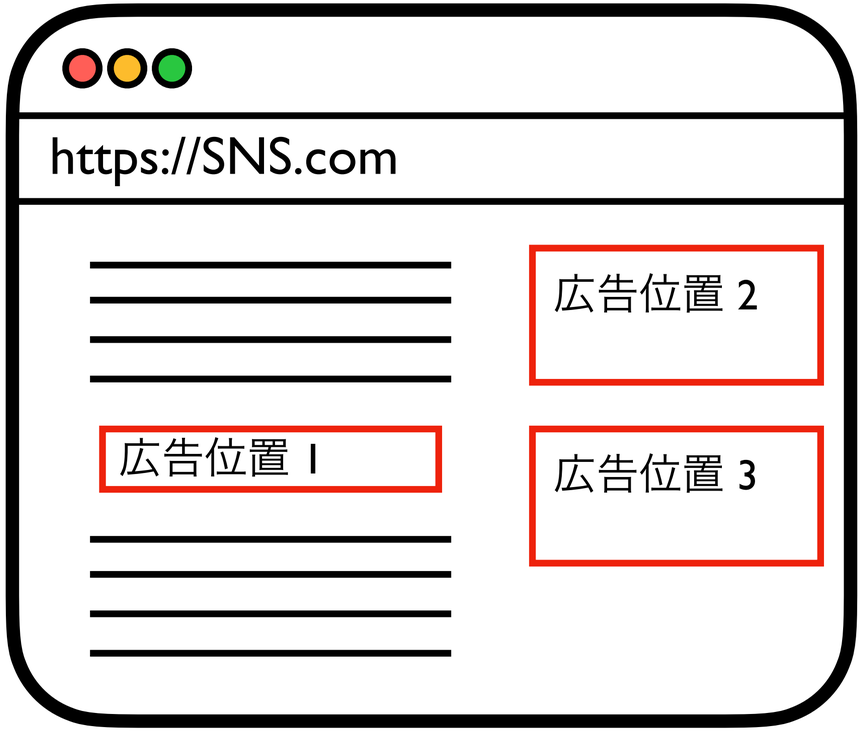
2023年9月7日 10:50-11:10, 大阪公立大学 中百舌鳥キャンパス

導入｜実世界には多くのオンライン意思決定問題が存在

オンライン広告配置

ウェブサイト上で

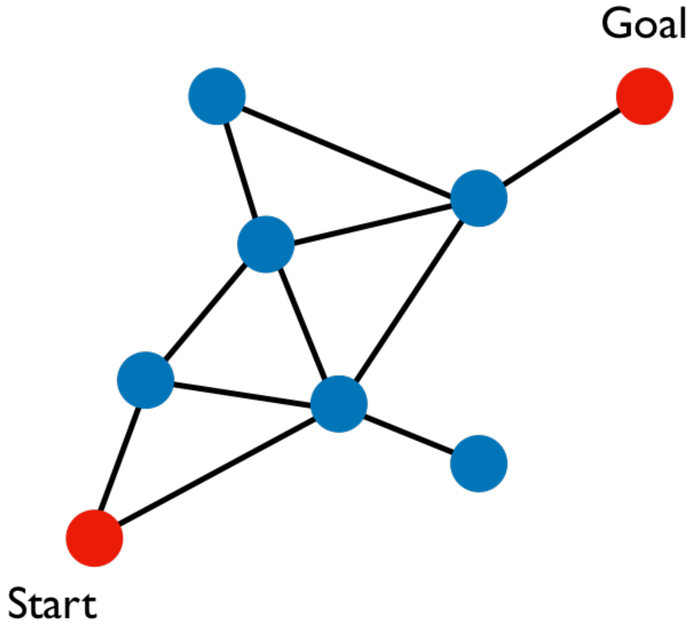
どの広告をどこに配置すると利益大？



オンライン経路制御

グラフ上のスタートからゴールへ

どの経路で {データ, 電力, … } を送ると累積コスト小？



クラウドソーシング

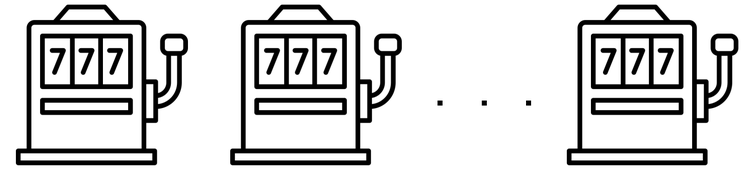
動的価格設定

etc …

バンディット問題として定式化可能

導入｜多腕バンディット問題

- k 個のスロットマシン（**アーム**，**行動**）が用意され，
  その中から1つ選び合計 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



敵対者が各時刻の**損失ベクトル** $\ell_1, \ldots, \ell_T \in [0,1]^k$ を決定
各ラウンド $t = 1, \ldots, T$ ：
1. **プレイヤー**がアーム $A_t \in [k] := \{1, \ldots, k\}$ を選択
2. プレイヤーがアーム A_t の損失 $\ell_{t,A_t} \in [0,1]$ を観測

← $\ell_{t,i} \in [0,1]$: 時刻 t でのアーム i の損失値

引いたアーム A_t についてのみ
損失が観測される

- プレイヤーの目標：累積損失の最小化 ＝ **リグレット** R_T の最小化

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} - \sum_{t=1}^{T} \ell_{t,i^*}\right], \quad i^* = \arg\min_{i \in [k]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$$

- 「探索」と「活用」のトレードオフに対処する必要がある

情報少で最適になりうるアームを選択

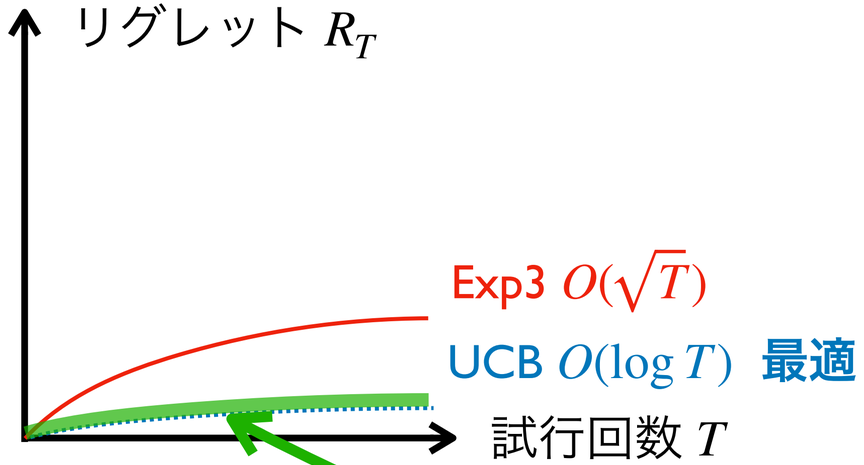
現時点で最適に見えるアームを選択

問題の難しさによらず最適な意思決定は可能か？

[Bubeck & Slivkins 2012, Zimmert & Seldin 2021]

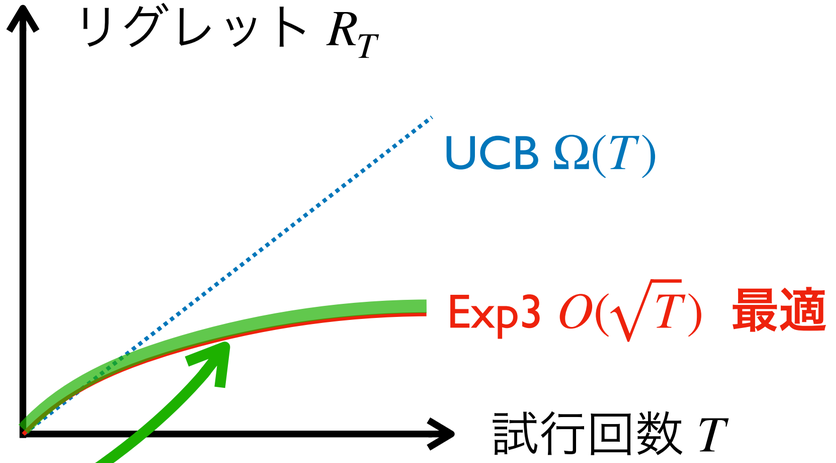
確率的環境

$\ell_{t,i} \sim \nu_i^*$ for all $i \in [k]$



敵対的環境

$\ell_1, \ldots, \ell_T \in [0,1]^k$ は任意に決定される



プレイヤーは背後の環境を知らない

望ましい方策：背後の環境を知らずに
確率的・敵対的環境の両方で高性能
= **Best-of-Both-Worlds (BOBW)**

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In COLT 2012.
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 2021.

比較的単純な設定では best-of-both-worlds が可能！

多腕バンディット問題などの

[Zimmert & Seldin 2021]

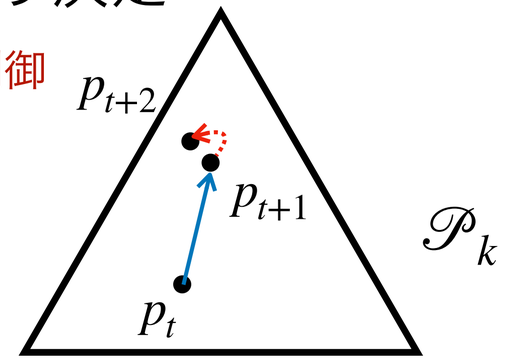
- Follow-the-Regularized-Leader：元々敵対的環境のためのアルゴリズム

1. アーム選択確率 $p_t \in \mathcal{P}_k$ を「これまでの観測 + 正則化」の最小化により決定：

これまでの推定損失の和　凸正則化関数：(主に) 境界への近づき方を制御

$$p_t \in \arg\min_{p \in \mathcal{P}_k} \langle \sum_{s=1}^{t-1} \hat{\ell}_s, p \rangle + \psi_t(p)$$

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



2. p_t をもとにアームを選択する：$A_t \sim p_t$

- Follow-the-regularized-leader に適切に正則化関数を設計すると，

多腕バンディット問題では best-of-both-worlds を達成可能

アーム数

確率的環境：$R_T = O\left(\frac{k \log T}{\Delta}\right)$　敵対的環境：$R_T = O(\sqrt{kT})$

損失期待値（一次の量）のみに依存した量

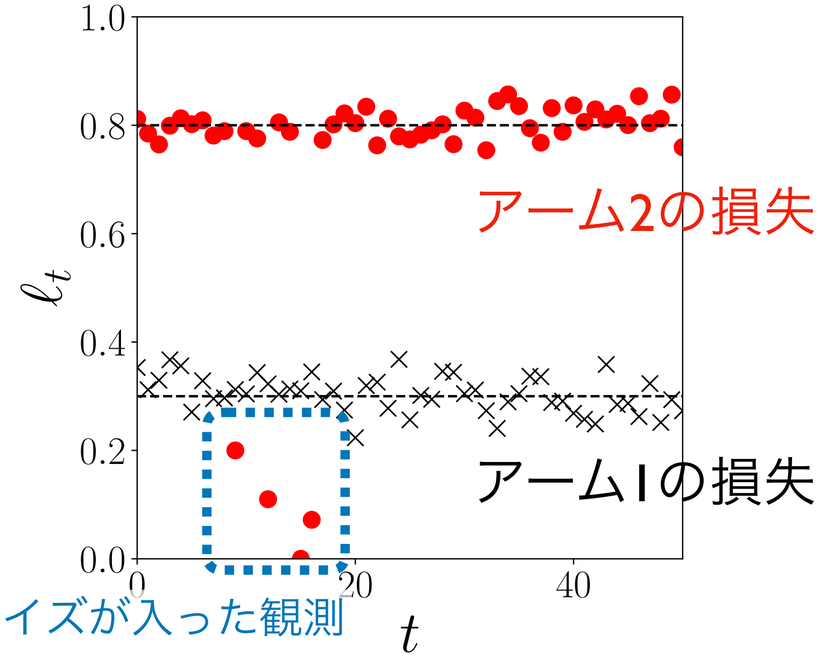
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 2021.

脱線：より現実的な環境

敵対的環境：非常に悲観的な設定

中間の環境？

確率的環境（有界損失）
楽観的な設定



敵対的汚染のある確率的環境 [Lykouris, Mirrokni & Leme 2018]

確率的環境から
生成された損失 (*)

$\ell'_1, \ldots, \ell'_T \sim \nu^*$

敵対的ノイズ

$C = \mathbb{E}\left[\sum_{t=1}^{T} \|\ell_t - \ell'_t\|_\infty\right]$

ノイズがのった損失

$\ell_1, \ldots, \ell_T$

$C = 0 \rightarrow$ 確率的環境

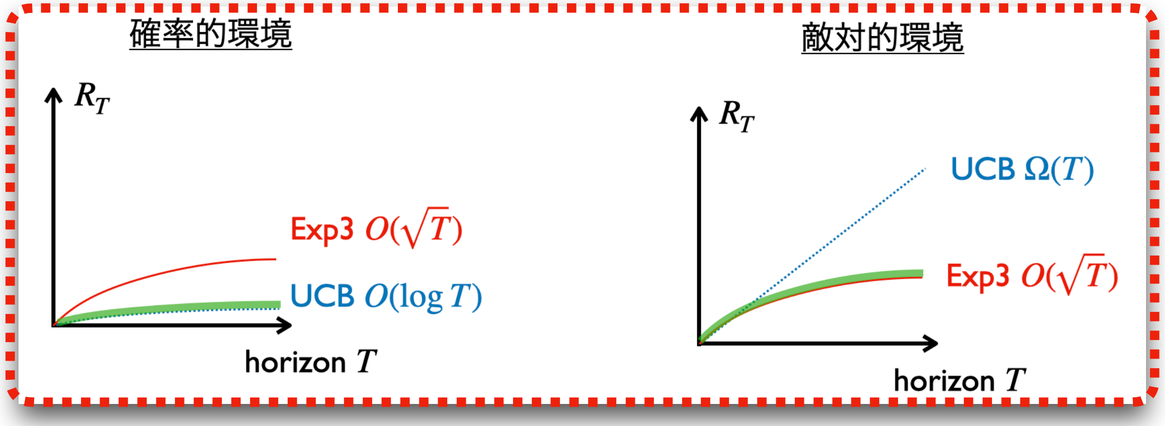
$C = 2T \rightarrow$ 敵対的環境

T. Lykouris, V. Mirrokni, and R.P. Leme. Stochastic bandits robust to adversarial corruptions. In STOC 2018.

研究課題

Best-of-Both-Worlds 方策と

敵対的汚染のある確率的環境



敵対的汚染のある確率的環境 with $C > 0$

FTRL に適当な正則化関数を用いると

多腕バンディット問題では

best-of-both-worlds を達成可能

確率的環境：$R_T = O\left(\frac{k \log T}{\Delta}\right)$　敵対的環境：$R_T = O(\sqrt{kT})$

課題1. 十分に適応的でない

環境の性質に適応的に：

Q1.1 損失の**スパース性**を活用可能？

Q1.2 損失の**分散**を活用可能？

課題2.

単純な設定のみ

現実的・複雑な問題でも：

Q2.1 組合せ構造を有する問題？

Q2.2 間接的な情報しか観測できない問題？

設定と主結果のみ紹介

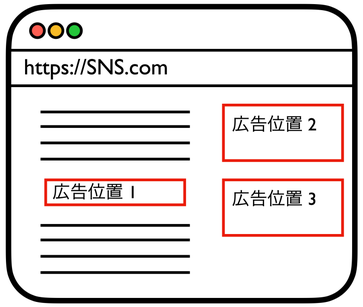
(Q1.1) スパース性の活用｜多腕バンディット問題

- **損失 $\ell_t \in [0,1]^k$ 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 [Kwon & Perchet 2016, Bubeck, Cohen & Li 2018]

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：

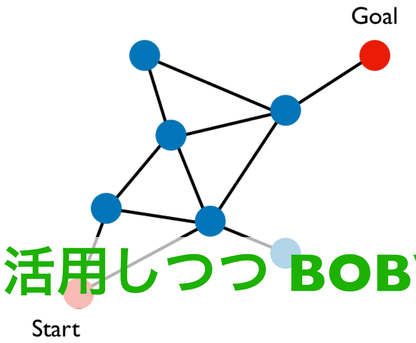
ほとんどの $a \in [k]$ で $r_{ta} := -\ell_{ta} = 0$



オンライン経路制御

ほとんどの経路でデータの損失は起きない：

ほとんどの $a \in [k]$ で $\ell_{ta} = 0$



Q. スパース性を活用しつつ BOBW を実現することは可能？

- 以下の保証を持つアルゴリズムを構築可能 [T, Ito, Honda 2023]

汚染のある確率的環境

$$R_T = \tilde{O}\left(\frac{s\log^2 T}{\Delta} + \sqrt{\frac{Cs\log^2 T}{\Delta}}\right)$$

敵対的環境

$$R_T = \tilde{O}\left(\sqrt{sT}\right)$$

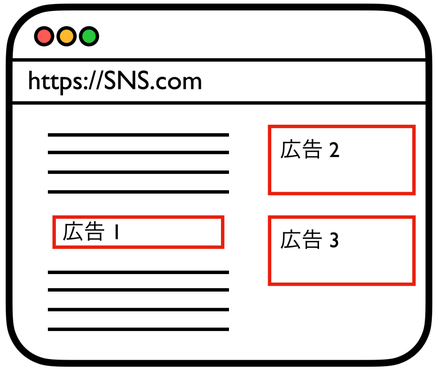
どちらもほぼ最適

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. JMLR, 2016.
S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In ALT 2018.
T. Tsuchiya, S. Ito, and J. Honda. Stability-penalty-adaptive Follow-the-regularized-leader: Sparsity, Game-dependency, and Best-of-both-worlds, 2023.

(Q1.2) 分布情報，分散の活用｜多腕バンディット問題

- **損失の分散**に注目：多くの実問題では分散が小さい

例1. オンライン広告配置問題



広告のクリック率は
非常に小さい

例2. オンライン最短経路問題

所要時間の変化は
小さい

[Chen et al. 2022]

分散依存のリグレット上界を持つ（各アームの分散を考慮する）方策は性能が良いはず

Q. 敵対的環境の性能を維持しつつ，確率的環境で分散依存のリグレット上界を達成可能？

- 以下の保証を持つアルゴリズムを構築可能 [Ito, T, Honda COLT 2022]

確率的環境 $R_T = O\left(\sum_{i:i \neq i^*} \left(\frac{\sigma_i^2}{\Delta_i} + 1 \right) \log T \right)$

敵対的環境 $R_T = \tilde{O}\left(\sqrt{kT}\right)$

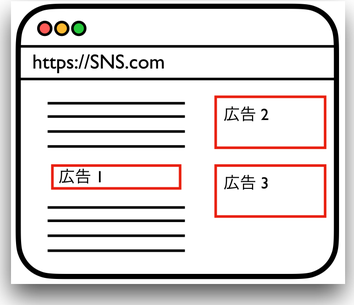
どちらもほぼ最適

X. Chen et al. Online eco-routing for electric vehicles using combinatorial multi-armed bandit with estimated covariance. Trans. Res. D: Transport and Environment, 2022.
S. Ito, T. Tsuchiya, and J. Honda. Adversarially Robust Multi-Armed Bandit Algorithm with Variance-Dependent Regret Bounds, In COLT2022.

(Q2.1) 組合せ半バンディット問題｜構造を伴う問題

- ここまで紹介した問題は実際には**組合せ構造**を持つ

例1.
オンライン広告配置問題



例2.
オンライン最短経路問題

Q. 組合せ構造がある場合にも分散を活用できる？

- 以下の保証を持つアルゴリズムを構築可能　[T, Ito, Honda AISTATS 2023]

確率的環境　$O\left(\sum_{i \in J^*}\left(\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}+c\right)\log T\right)$

汚染のある確率的環境　$O\left(\sum_{i \in J^*}\left(\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}+1\right)\log T+\sqrt{Cm\sum_{i \in J^*}\left(\frac{w(\mathcal{A})\,\sigma_i^2}{\Delta_{i,\min}}+c\right)\log T}\right)$

敵対的環境　$O\left(\sqrt{kmT\log T}\right)$　(+ データ依存上界)

T. Tsuchiya, S. Ito, and J. Honda. Further Adaptive Best-of-Both-Worlds Algorithm for Combinatorial Semi-Bandits, In AISTATS 2023.

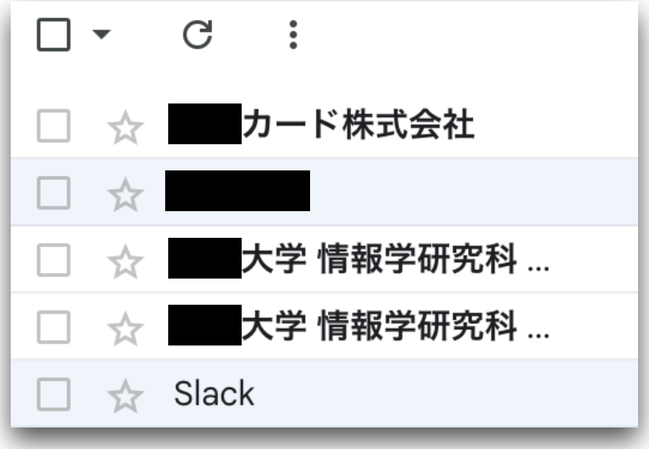
(Q2.2) 林檎の試食問題，メールのスパム判定

[Helmbold, Littlestone & Long 1992]

- メールボックスに来たメールが
  スパムかハム（not スパム）かを逐次的に判定
- 3通りの行動が存在:
  1. スパムとラベル付け (P)
  2. ハムとラベル付 (N)
  3. 人間に聞いて，正しいラベル （スパム or ハム）を教えてもらう
     → この場合のみ真のラベルを観測可能

スパム？ →

ハム？ →

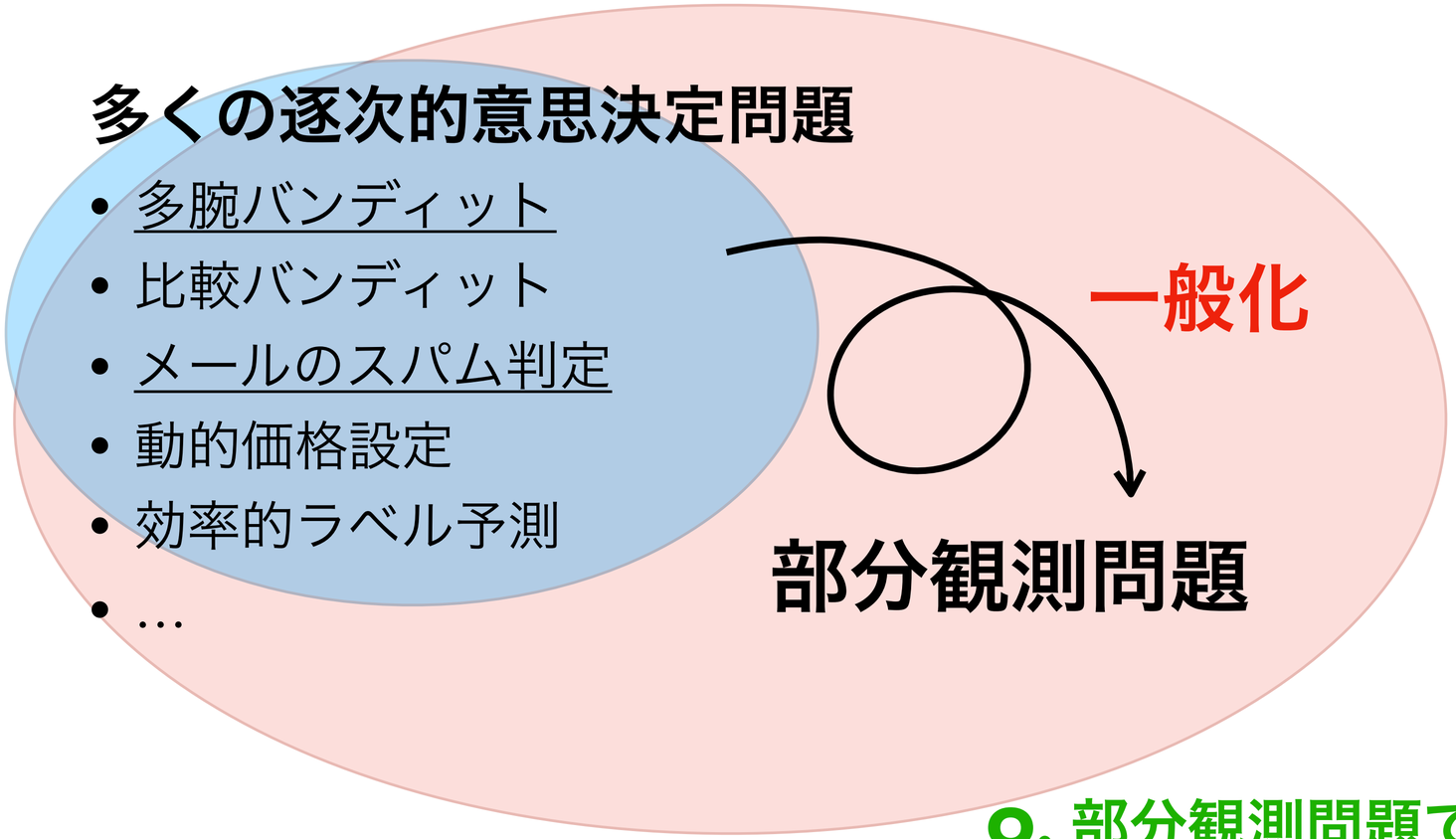


このような間接的な観測のみから意思決定を行う問題の枠組みは？

D. P. Helmbold, N. Littlestone, and P. M. Long, Apple tasting and nearly one-sided learning. In FOCS 1992.

(Q2.2) 部分観測問題｜間接的な観測から逐次的意思決定

- 逐次的意思決定問題の一般的な枠組み



Q. 部分観測問題で BOBW は可能？

(Q2.2) 部分観測問題における best-of-both-worlds 方策

[T, Ito, Honda ALT 2023]

- 局所的観測可能ゲーム

	確率的環境	敵対的環境	汚染のある確率的環境
[Tsuchiya+ 2020]	$O(\log T)$	NA	NA
[Lattimore+ 2020]	NA	$O(\sqrt{T})$	NA
提案法	$O((\log T)^2)$	$O(\sqrt{T}\log T)$	$O((\log T)^2 + \sqrt{C}\log T)$

- 大域的観測可能ゲーム

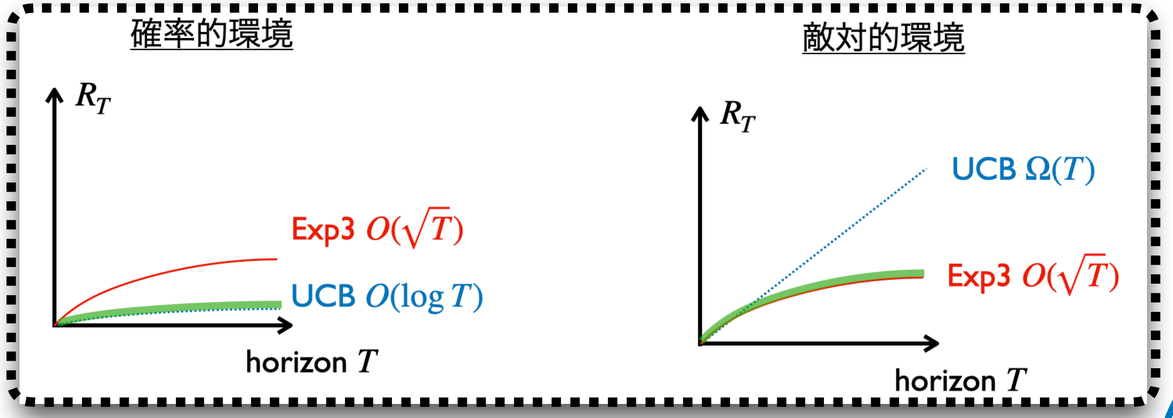
	確率的環境	敵対的環境	汚染のある確率的環境
[Lattimore+ 2020]	NA	$O(T^{2/3})$	NA
提案法	$O((\log T)^2)$	$O((T\log T)^{2/3})$	$O((\log T)^2 + (C\log T)^{2/3})$

T. Tsuchiya, J. Honda, and M. Sugiyama. Analysis and design of Thompson sampling for stochastic partial monitoring. In NeurIPS 2020.
T. Lattimore and Cs. Szepesvári. Exploration by optimisation in partial monitoring. In COLT 2020.
T. Tsuchiya, S. Ito, and J. Honda. Best-of-Both-Worlds Algorithms for Partial Monitoring, In ALT 2023.
T. Tsuchiya, S. Ito, and J. Honda. Stability-penalty-adaptive Follow-the-regularized-leader: Sparsity, Game-dependency, and Best-of-both-worlds, arXiv 2023.

まとめ｜Best-of-both-worlds 方策の進展

Best-of-Both-Worlds 方策と

敵対的汚染のある確率的環境



敵対的汚染のある確率的環境 with $C > 0$

FTRL に適当な正則化関数を用いると

多腕バンディット問題では

課題2.

単純な設定のみ

best-of-both-worlds を達成可能

確率的環境：$R_T = O\left(\frac{k \log T}{\Delta}\right)$　敵対的環境：$R_T = O(\sqrt{kT})$

課題1. 十分に適応的でない

Q1. より環境（損失の性質）に対して

適応的な方策を構築できるか？

A. 損失の**スパース性**や**分散**を活用可能

Q2. より**複雑な問題**においても

BOBW 方策を構築できるか？

A. 組合せバンディット問題や

部分観測問題で構築可能