オンライン意思決定ユニット　本多 淳也
Online Decision Making Unit　Junya Honda

背景：意思決定を繰り返し行う問題では，対象についての知識が事前になく試行錯誤を通じてしか情報が得られない場面が多く現れる（例：広告配信・新薬の割り当て）
目標：これまでの観測結果に基づいて次の行動を適応的に選択することで得られる利益/知識を最大化

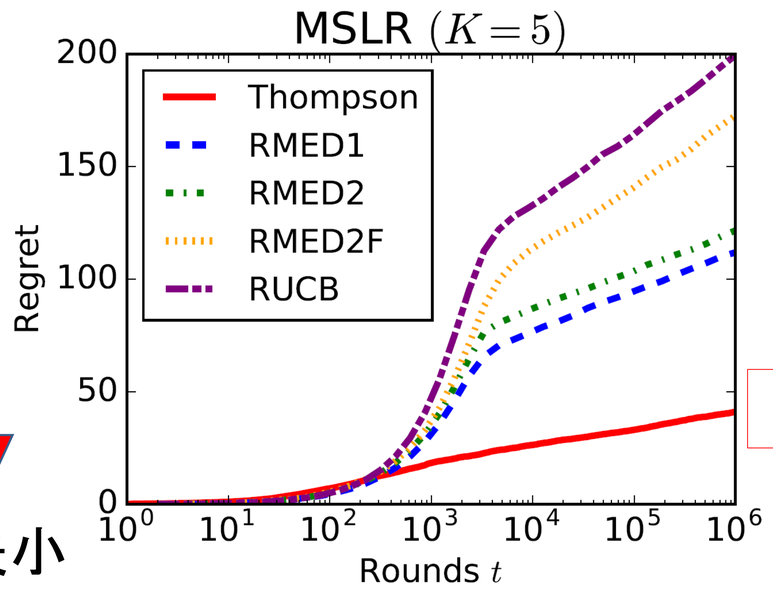
定式化：多腕バンディット問題
- 時刻ごとにスロットマシンK台のうち1台を選択
- 選択した台iに対応付けられた確率分布P_iにしたがう報酬X_iを観測
- 目標：累積報酬の最大化（リグレット最小化）/ 優れた台の発見（純粋探索問題）

Dueling Bandits with Qualitative Feedback

X. Liyuan, J. Honda, M. Sugiyama (AAAI'17)

- 観測が量的でなく質的なバンディット問題を考えたい
  例) 新薬の薬効（完治/効果あり/効果なし），
  　　レストランの評価（おいしい/まあまあ/微妙）
  - 「累積報酬」等がそもそも素朴には定義されない
- 従来手法：分位点バンディット
  - 報酬分布の分位点により累積報酬を定義
  - 明確な優劣関係を捉えられない場合がある
- 関連問題：比較バンディット問題
  - 各時刻にスロットマシンのペアを選択
  - マシン間の（雑音付き）優劣関係のみを観測
  - コンドルセ勝者・ボルダ勝者等の経済学由来の規準で報酬を定義
- 本研究：
  - 質的観測の設定が比較バンディット問題に帰着できることに基づき，比較バンディット由来の自然な報酬を用いることを提案
  - 比較バンディットへの帰着を用いる戦略に比べて実験的にも理論的にも高性能な戦略を構成

損失小

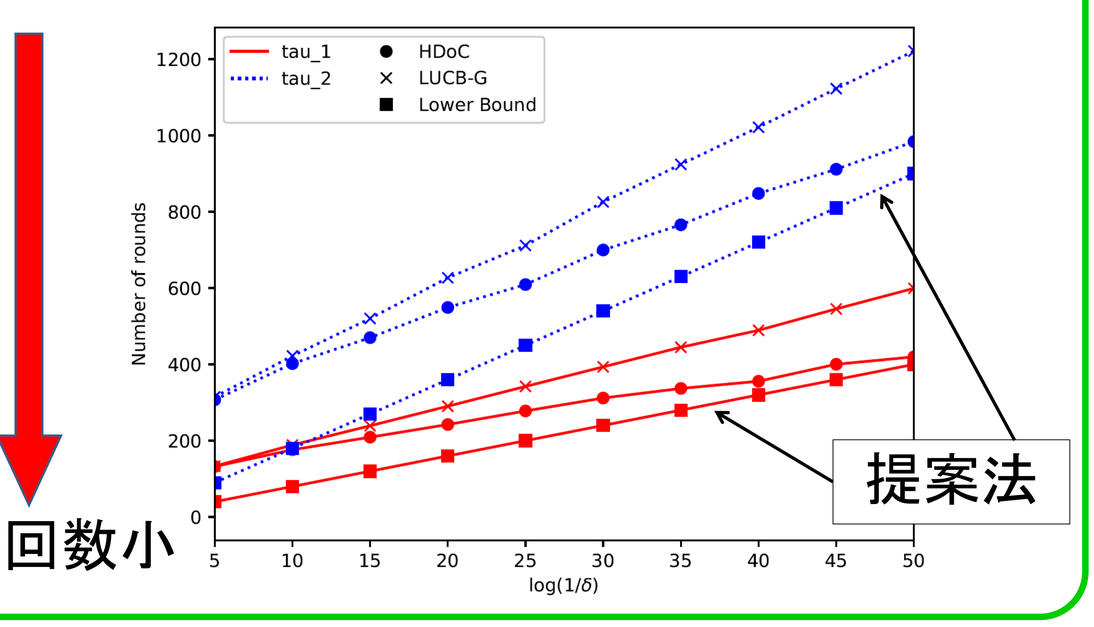


Good Arm Identification via Bandit Feedback

H. Kano, J. Honda, K. Sakamaki, K. Matsuura, A. Nakamura, M. Sugiyama (MLJ)

- 期待報酬が優れたスロットマシンをなるべく少ない総試行回数で発見したい
  - が，厳密な真に最適な台を探すのは原理的に膨大な試行回数が必要
- 従来手法：しきい値付きバンディット
  - 期待報酬がしきい値以上の台とそれ以外に分類
  - ある台の期待報酬がしきい値以下であることを確認することに多くの試行が費やされてしまう
- 本研究の定式化：期待報酬がしきい値以上の台(good arm)が存在する場合に，それらのうち1個をなるべく早く出力
  - 期待報酬が大きい台ほど，それがgood armであることを確認するのに要する標本数が少ない
  - 一方，どの台の累積報酬が大きそうかは少ない試行回数では分からない
- 本研究では累積報酬最大化との類似性を利用した戦略を提案し，その漸近最適性を証明

試行回数小



その他の研究

- 報酬が期待値と分散が未知の正規分布にしたがう設定におけるバンディットアルゴリズムの理論解析
  - 漸近最適と考えられてきたKL-UCBとよばれる戦略が理論限界を達成しないことを証明，かつ微細な修正により漸近最適にできることを証明 (JMLR)
- 人種や年齢等のセンシティブな説明変数がある場合の公平性を考慮した回帰問題を定式化
  - 公平性と精度のトレードオフに関する非凸最適化問題の多項式時間厳密解法を構成 (ICML'18，連続最適化チームとの共同研究)