部分観測問題における トンプソン抽出アルゴリズムの設計と解析

土屋 平[1,2] 本多 淳也[1,2] 杉山 将[2,1]

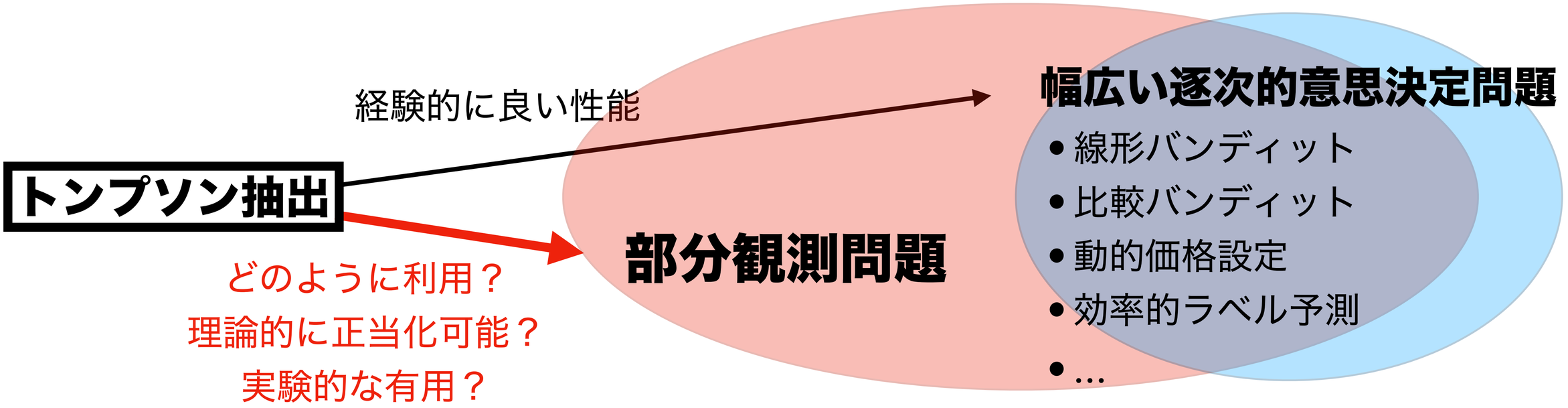
概要: トンプソン抽出の部分観測問題における理論的性質はこれまで未知であり，また，既存法は事後分布のヒューリスティックな近似に基づいている．そこで，近似を用いないトンプソン抽出に基づいたアルゴリズムを考案した．さらに，提案法がラウンド数 T に対して，$O(\log T)$ のリグレット上界を達成可能であることを示した．この上界は，部分観測問題におけるトンプソン抽出の最初の上界であり，さらに線形バンディットにおける最初の対数オーダーの上界でもある．

1. 東京大学 THE UNIVERSITY OF TOKYO

2. RIKEN

研究課題

- 部分観測問題
  - 限られたフィードバックをもとに逐次的意思決定を行う一般的な枠組み
- トンプソン抽出
  - 幅広い逐次的意思決定問題に対して，経験的に最も有用な方策の1つ
  - 「探索」と「活用」のトレードオフを事後分布からのサンプリングによって扱う



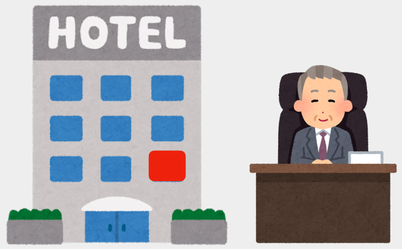
部分観測問題の例：動的価格設定

プレイヤー（= 販売者）

$t = 1$日目

ホテルのオーナー
（販売者）

宿泊料 4,000円



$t = 2$日目

ホテル1泊の宿泊料を決定
$\{1000円, \ldots, N円\}$

宿泊料 8,000円

$t = \cdots$

敵対者　利用者の
内部状態 $j(t)$
（= 評価価格）

宿泊料 $\leq$ 9,000円 なら利用

宿泊料 $\leq$ 5,000円 なら利用

（機会）**損失**

$9{,}000 - 4{,}000$
$= 5{,}000$円

c 円（定数）
$(\because 5000 - 8000 < 0)$

フィードバック

宿泊する

宿泊
しない

販売者は，フィードバック（宿泊する or 宿泊しない）のみ観測可能

Q. 限られた観測のみから，全体の損失を最小化する（= 全体の報酬を最大化する）ことは可能か？

部分観測問題の定式化

- N 個の行動と M 個の内部状態からなる部分観測問題 $G = (L, H)$
- 損失行列 $L = (\ell_{i,j}) \in \mathbb{R}^{N \times M}$，フィードバック行列 $H = (h_{i,j}) \in \Sigma^{N \times M}$

（Σ：フィードバック記号の集合）

For round $t = 1, \ldots, T$:

1. **プレイヤー**が行動 $i(t) \in \{1, \ldots, N\}$ を選択する
2. **敵対者**が内部状態を選択する $j(t) \overset{\text{i.i.d.}}{\sim} \mathrm{Multi}(p^*)$ $(p^* \in \mathscr{P}_M)$
   - p^*：戦略
   - $\mathscr{P}_M$：確率単体
3. プレイヤーが損失 $\ell_{i(t),j(t)}$ を被り，フィードバック $h_{i(t),j(t)}$ を観測する

- 目標: 擬リグレットの最小化（= 全体の損失の最小化）

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \left(L_{i(t)}^{\top} p^* - L_1^{\top} p^* \right)$$

実際に取った行動の期待損失　最適な行動の期待損失

行動1が最適であるとする

$L_i \in \mathbb{R}^M$: L の i 列目

部分観測問題におけるトンプソン抽出の適用

未知パラメータ: 敵対者の戦略 p^*

トンプソン抽出の素朴な一般化:

1. 目的パラメータの事後分布を計算する

   $\pi(p \mid \text{時刻 } t \text{ までの観測データ}) \propto \pi(p) \prod_{i=1}^{N} \exp\left(-n_i \mathscr{D}_{\mathrm{KL}}\left(q_i^{(t)} \| S_i p\right)\right)$

2. 事後分布から目的パラメータをサンプリングする

   $\tilde{p}_t \sim \pi(p \mid \text{時刻 } t \text{ までの観測データ})$

3. サンプルしたパラメータをもとに最適な行動を選択する

   行動 $i(t) := \arg\min_{i \in [N]} \underbrace{L_i^\top \tilde{p}_t}_{\text{行動 } i \text{ の期待損失}}$ を取る

😨 複雑な事後分布

n_i : 時刻 t までに行動 i を取った回数

$q_i(t)$: 時刻 t における行動 i を取ったときの経験フィードバック分布

S_i : 行動 i の信号行列

Bayes-update PM for TS (BPM-TS) [Vanchinathan+ 2014]

- 戦略 p^* の推定値をガウス分布の事前分布を用意してベイス的更新
- 仮定: 内部状態は共分散行列 I_M と平均未知のガウス分布に従う

  （実際には $\mathrm{Multi}(p^*)$ に従う）

😊 高速な計算

😊 経験的に最も性能の良いアルゴリズムの一つ

😨 真の事後分布との乖離

$$\mathcal{N}(\text{時刻 } t \text{ におけるパラメータ}) \xleftrightarrow[\text{乖離}]{} \pi(p)\prod_{i=1}^{N}\exp\left(-n_i\mathcal{D}_{\mathrm{KL}}\left(q_i^{(t)}\|S_i p\right)\right)$$

😨 理論解析が与えられていない

Hastagiri P Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In Advances in Neural Information Processing Systems 27, pages 1691–1699, 2014.

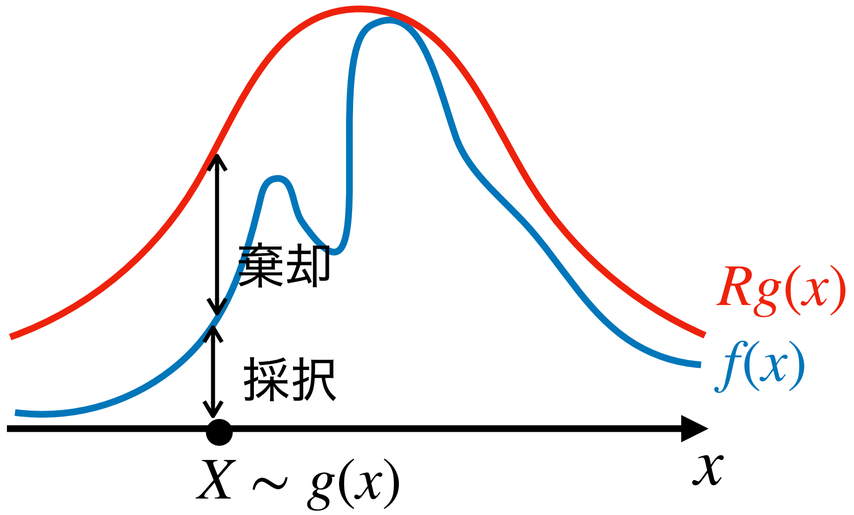
棄却サンプリング

- 複雑な分布 $f(x)$ から独立同分布に従う標本を得る手法
- 提案分布 $g(x)$ を用意し，以下を行う:

容易にサンプルを得られる分布

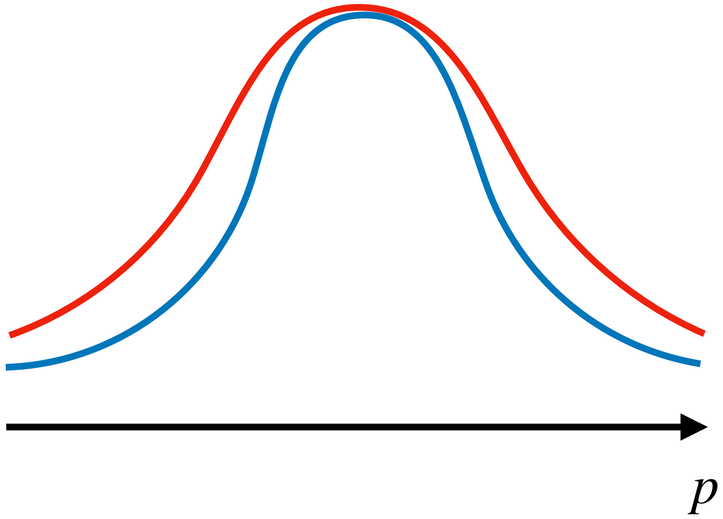
1. サンプル $X \sim g(x)$ を得る
2. 確率 $f(X)/Rg(X)$ で X を採択する $R = \sup_x f(x)/g(x)$
3. 採択されるまで繰り返す

- タイトな提案分布を用意する必要がある



提案法 (TSPM)

1. タイトな提案分布を用意する



$$\overbrace{R\pi(p)\prod_{i=1}^{N}\exp\left(-n_i\|q_i^{(t)}-S_ip\|^2\right)}^{\text{ガウス分布}}$$ 提案分布

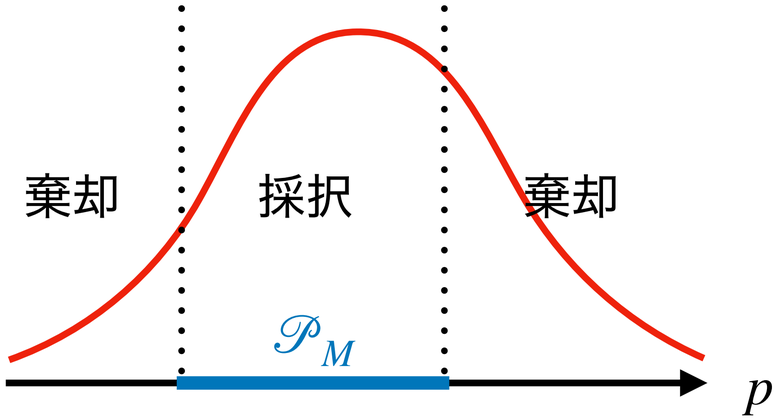
VI　ピンスカーの不等式

$$\pi(p)\prod_{i=1}^{N}\exp\left(-n_i\mathcal{D}_{\mathrm{KL}}\left(q_i^{(t)}\|S_ip\right)\right)$$ 事後分布

2. 確率単体 $\mathcal{P}_M$ 上に制限されたガウス分布からサンプリング

$$\overbrace{\pi(p)\prod_{i=1}^{N}\exp\left(-n_i\|q_i^{(t)}-S_ip\|^2\right)}^{\text{ガウス分布}}$$

😨 確率単体 $\mathcal{P}_M$ に制限



リグレット上界

A, N：フィードバックと行動の選択肢数

Δ_i：行動 i の最適行動との期待損失の差

$\Lambda = \min_{j \neq k} \Delta_{j,k} / \|z_{j,k}\|$（$\Delta_{j,k}$：行動 j と k の期待損失の差，

$z_{j,k} \in \mathbb{R}^{2A}$：損失とフィードバックを関連させるベクトル）

定理（簡略版）.

任意の局所的観測可能な線形部分観測問題に対して，

TSPM-Gaussianの期待リグレットが以下で抑えられる:

$$O\left(\max\left\{ \frac{A \sum_{i \in [N]} \Delta_i}{\Lambda^2}, \frac{\sqrt{A} N^3 \max_{i \in [N]} \Delta_i}{\Lambda^2} \right\} \log T \right).$$

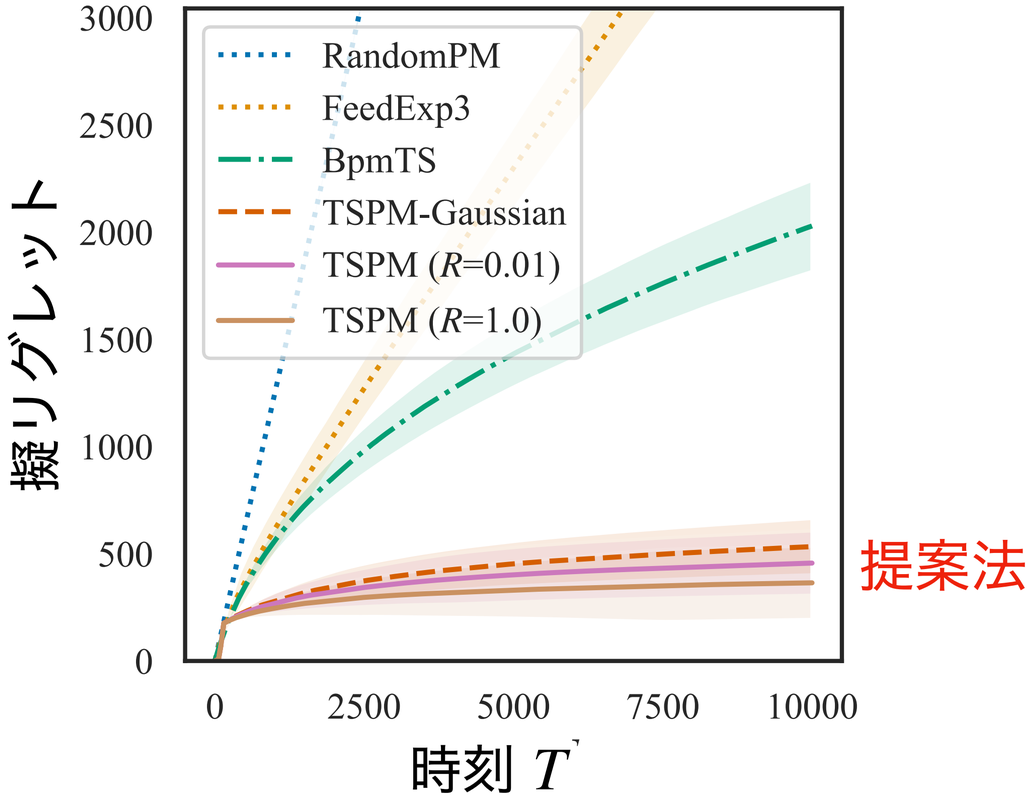
問題依存の定数　　ラウンド数 T への対数依存

- トンプソン抽出の部分観測問題に対する，初めての問題依存対数リグレット
- トンプソン抽出の線形バンディット問題に対する，初めての対数リグレット

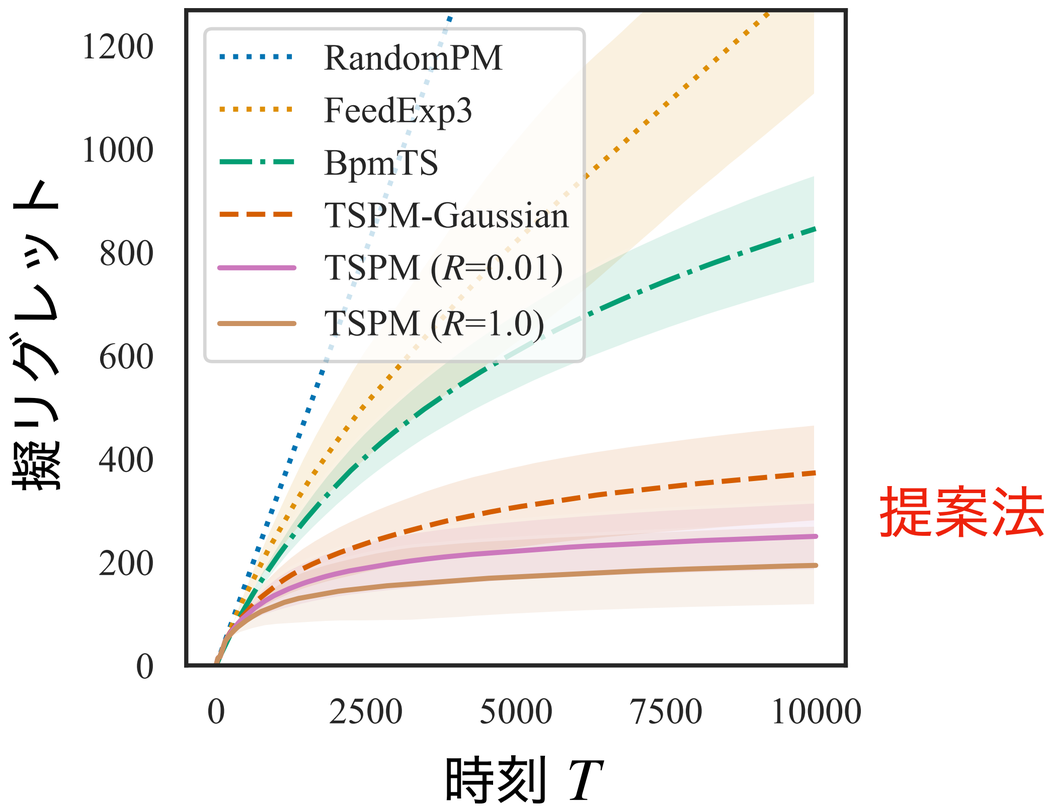
動的価格設定におけるリグレット比較実験

- 既存手法の性能を大きく改善

局所的観測可能ゲーム



大域的観測可能ゲーム

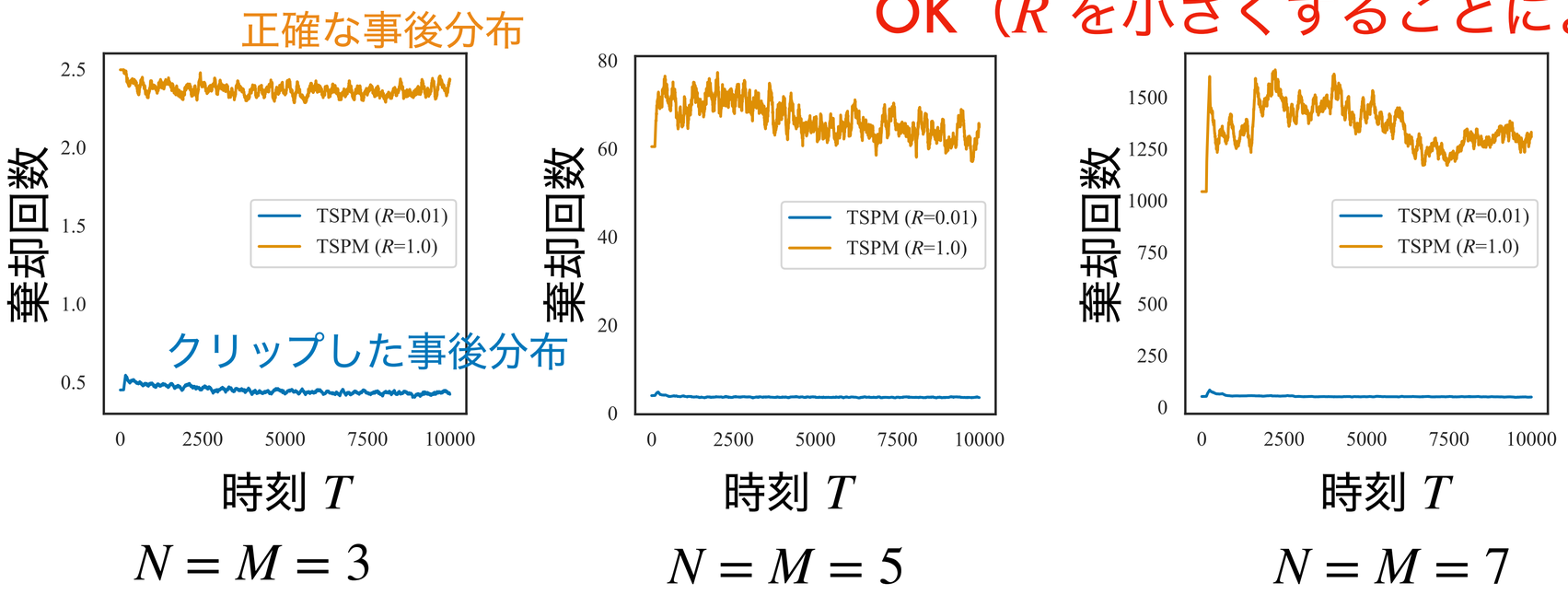


棄却サンプリングにおける棄却頻度

- 棄却サンプリングにおいて望ましい性質
  1. ラウンド数が進むにつれ，棄却頻度が増えない　OK
  2. 分布のサポートの次元 $M-1$ が増えるにつれ，棄却頻度が増えない

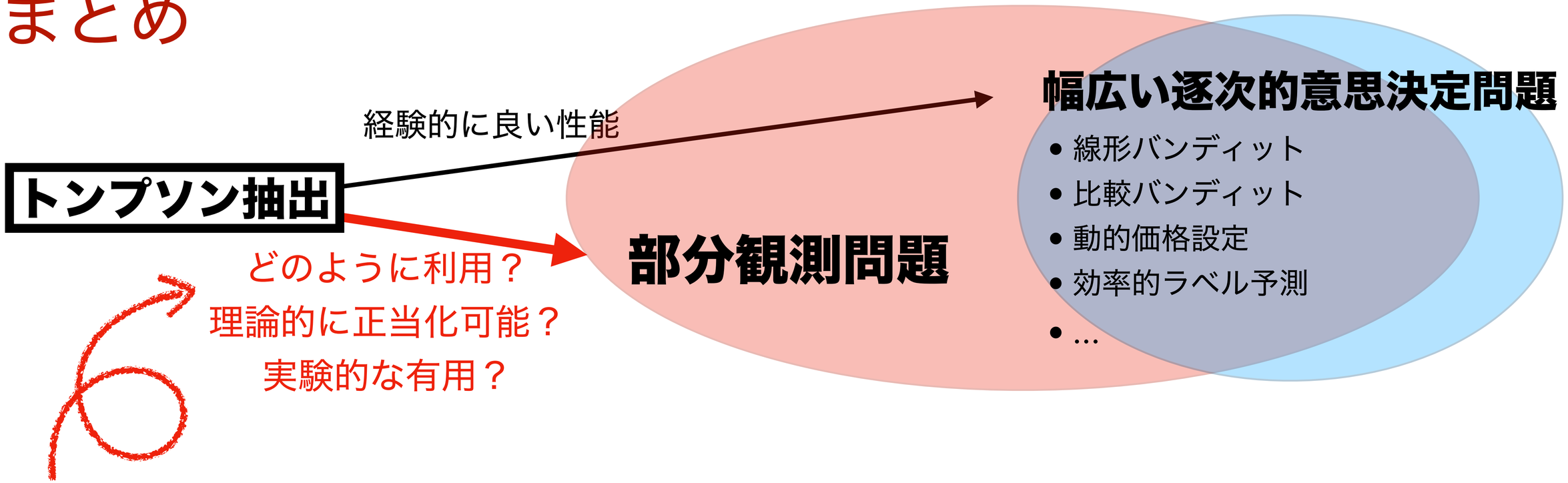
OK（R を小さくすることにより）

- 局所的観測可能ゲーム
- 動的価格設定



定数 $\cdot \dfrac{\text{事後分布の密度関数}}{R \cdot \text{提案分布の密度関数}}$ の確率で採択

まとめ



1. タイトな提案分布からのサンプリングによる新しいトンプソン抽出アルゴリズム
2. 部分観測問題と線形バンディットに対しての新しいリグレット上界

本論文の詳細は以下の論文で参照可能

T. Tsuchiya, J. Honda, M. Sugiyama,
Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring, In NeurIPS 2020 (to appear).