部分観測問題におけるThompson抽出

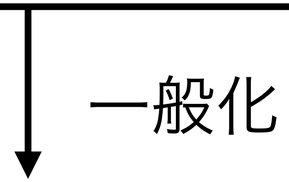
土屋 平 / Taira Tsuchiya

＊本発表は本多淳也先生，杉山将先生との共同研究に基づきます.

Toshiba Symposium 2020, 2020年12月22日

講演の概要

前半: バンディット問題とThompson抽出



後半: 部分観測問題におけるThompson抽出の有効性に関する研究の紹介

[Tsuchiya, Honda & Sugiyama NeurIPS 2020]

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

多腕バンディット問題

- 未知の分布の集合 $(P_i)_{i=1,\dots,N}$ が備わった，逐次的意思決定モデル
  (= アーム，行動)

	アーム 1	アーム 2	・・・	アーム N
分布/期待損失	P_1 / μ_1	P_2 / μ_2		P_N / μ_N

For round $t = 1,\dots,T$:
1. プレイヤーがアーム $i(t) \in \{1,\dots,N\}$ を選択する
2. アーム $i(t)$ の確率的な損失($\sim P_{i(t)}$)を観測する

引いたアーム $i(t)$ についてのみ損失が観測される

目標: ラウンド全体の期待損失を最小化する

- 「探索」と「活用」のトレードオフに対処する必要がある

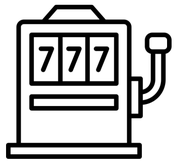
情報の少なく最適になり得るアームを引く

現時点で最適に見えるアームを引く

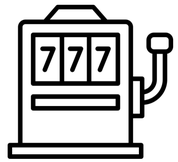
Thompson抽出 [Thompson 1933]

- あらゆる逐次的意思決定問題において経験的に最も有用な方策の1つ

例. 2腕Bernoulliバンディット



アーム1の損失
$\sim \mathrm{Ber}(\mu_1)$



アーム2の損失
$\sim \mathrm{Ber}(\mu_2)$

$$\mathrm{Ber}(\mu) = \begin{cases} 1 & (\text{確率 } \mu) \\ 0 & (\text{確率 } 1-\mu) \end{cases}$$

- 各ラウンド $t = 1, \ldots, T$ で以下を行う

1. 目的パラメータの事後分布を計算する

   $\pi(\mu_1 \mid \text{時刻 } t \text{ までの観測データ}), \pi(\mu_2 \mid \text{時刻 } t \text{ までの観測データ})$　(目的パラメータ: μ_1, μ_2)
2. 事後分布から目的パラメータをサンプリングする

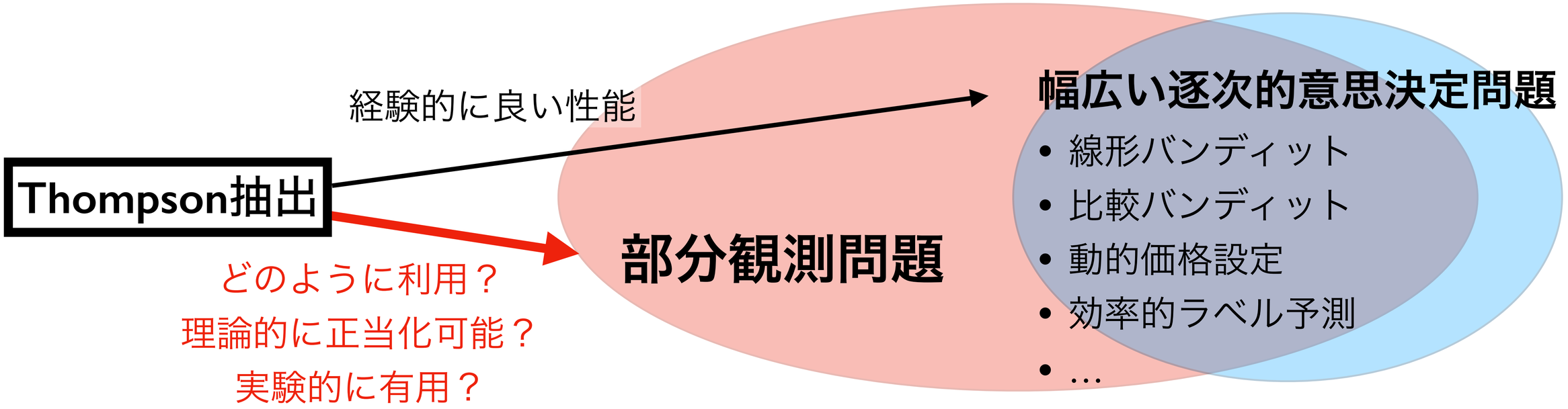
   $\tilde{\mu}_1 \sim \pi(\mu_1 \mid \text{時刻 } t \text{ までの観測データ}), \tilde{\mu}_2 \sim \pi(\mu_2 \mid \text{時刻 } t \text{ までの観測データ})$
3. サンプルしたパラメータをもとに最適な行動（= 腕）を選択する

   $\tilde{\mu}_1 \leq \tilde{\mu}_2$ ならばアーム 1 を引き，それ以外ならばアーム 2 を引く

W. R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933.

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問題の一般的な枠組み
- Thompson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



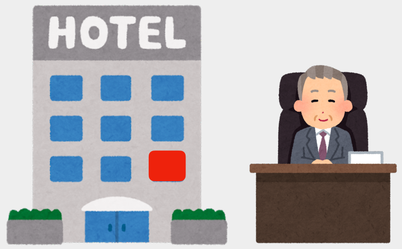
部分観測問題の例: 動的価格設定

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

フィードバックのみから，全体の損失を最小化する（= 全体の報酬を最大化する）ことを目指す

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

例1. 動的価格設定 [Kleinberg & Leighton 2003]

N: （離散化した）販売価格
M: （離散化した）評価額

(*) 行: 販売価格, 列: 内部状態

行動

内部状態 ~ p^*(戦略)

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 ≤ 9,000円 なら利用

宿泊料 ≤ 5,000円 なら利用

（機会）**損失**

9,000 − 4,000
= 5,000円

c 円（定数）
(∵ 5000 − 8000 < 0)

フィードバック

宿泊する

宿泊しない

損失行列 (*)

$$\ell_{i,j} = \begin{cases} j - i & (j \geq i) \\ c & (\text{otherwise}) \end{cases}$$

$$L = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ c & 0 & 1 & 2 & 3 \\ c & c & 0 & 1 & 2 \\ c & c & c & 0 & 1 \\ c & c & c & c & 0 \end{pmatrix}$$

$j \geq i$

$j < i$

フィードバック行列 (*)

$$h_{i,j} = \begin{cases} \bigcirc & (j \geq i) \\ \times & (\text{otherwise}) \end{cases}$$

$$H = \begin{pmatrix} \bigcirc & \bigcirc & \bigcirc & \bigcirc & \bigcirc \\ \times & \bigcirc & \bigcirc & \bigcirc & \bigcirc \\ \times & \times & \bigcirc & \bigcirc & \bigcirc \\ \times & \times & \times & \bigcirc & \bigcirc \\ \times & \times & \times & \times & \bigcirc \end{pmatrix}$$

$j \geq i$

$j < i$

$\Sigma = \{$宿泊する($\bigcirc$), 宿泊しない($\times$)$\}$

R. Kleinberg, & T. Leighton, The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In FOCS 2003.

例2. 効率的ラベル予測 [Cesa-Bianch+ 2005]

- プレイヤーが商品のラベル（正か負か）をオンラインに決定する
- 3通りの行動が存在:
  1. 正とラベル付け (P)
  2. 負とラベル付 (N)
  3. 専門家に聞いて，正しいラベルを教えてもらう

$$L = \begin{pmatrix} 0 & c_{\mathrm{N}\to\mathrm{P}} \\ c_{\mathrm{P}\to\mathrm{N}} & 0 \\ q & q \end{pmatrix}$$

$c_{\mathrm{N}\to\mathrm{P}} > 0$: N を P と間違えるコスト

$c_{\mathrm{P}\to\mathrm{N}} > 0$: P を N と間違えるコスト

$q > 0$: 専門家に聞くコスト

$$H = \begin{pmatrix} \mathrm{None} & \mathrm{None} \\ \mathrm{None} & \mathrm{None} \\ \mathrm{P} & \mathrm{N} \end{pmatrix}$$

N. Cesa-Bianchi, G. Lugosi, & G. Stoltz. Minimizing regret with label efficient prediction. IEEE Transactions on Information Theory, 2005.

部分観測問題におけるThompson抽出の適用

未知パラメータ: 敵対者の戦略 p^*

Thompson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 $i(t) := \arg\min_{i \in [N]} L_i^{\top} \tilde{p}_t$ を取る

   $L_i^{\top} \tilde{p}_t$: 行動 i の期待損失

複雑な事後分布

→サンプリングが困難 😨

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

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

😨 理論解析が与えられていない

H. P Vanchinathan, G. Bartók, & A. Krause. Efficient partial monitoring with prior information. In NeurIPS 2014.

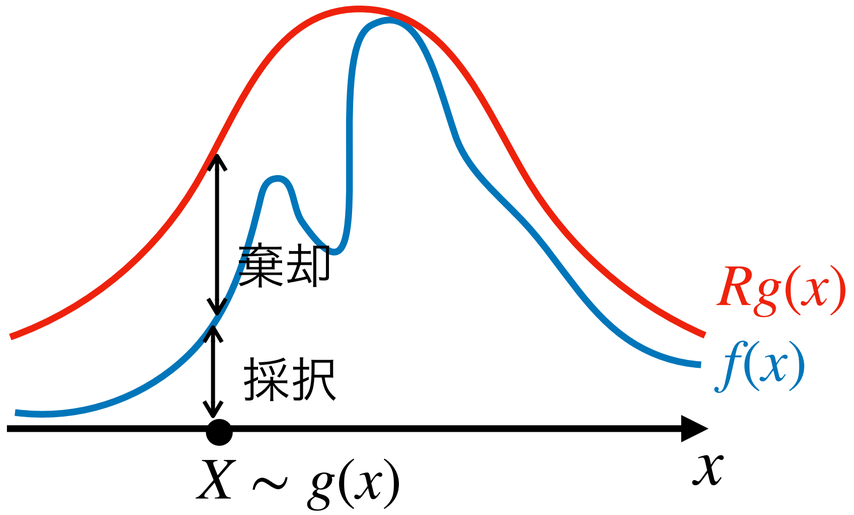
棄却サンプリング

- 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

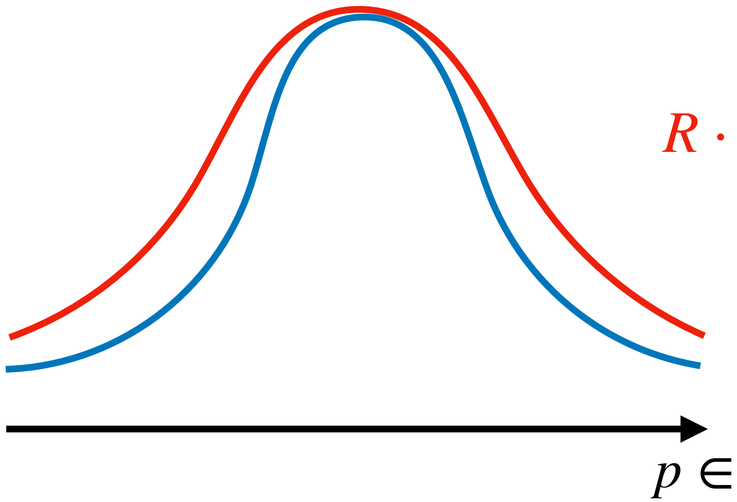
1. サンプル $X \sim g(x)$ を得る
2. 確率 $f(X)/Rg(X)$ で X を採択する $R = \sup_x f(x)/g(x)$
3. 採択されるまで繰り返す

- タイトな提案分布を用意する必要がある



提案法 (TSPM)

Step 1. タイトかつサンプリングが容易な提案分布を用意する



$$R \cdot \overbrace{\pi(p)}^{\text{事前分布}} \cdot \overbrace{\prod_{i=1}^{N} \exp\left(-n_i \|q_i^{(t)} - S_i p\|^2\right)}^{M\text{次元ガウス分布}}$$

提案分布

VI　ピンスカーの不等式

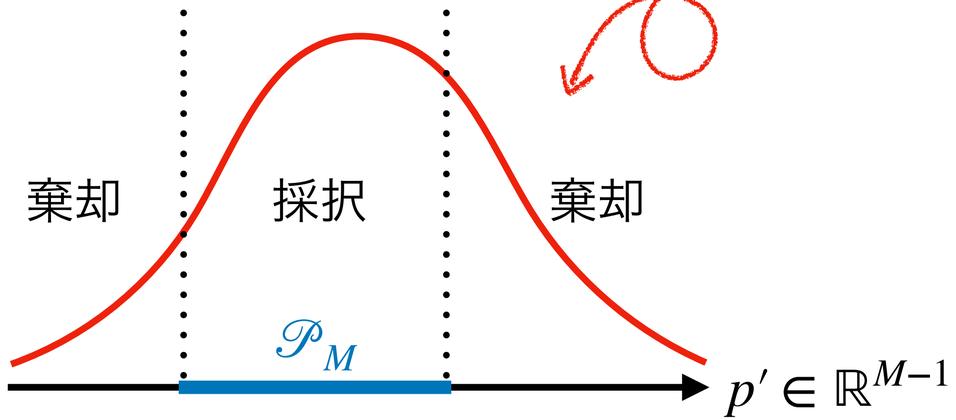
$$\pi(p) \cdot \prod_{i=1}^{N} \exp\left(-n_i \mathscr{D}_{\mathrm{KL}}\left(q_i^{(t)} \| S_i p\right)\right)$$

事後分布

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

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

😨 確率単体 $\mathscr{P}_M$ 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:

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

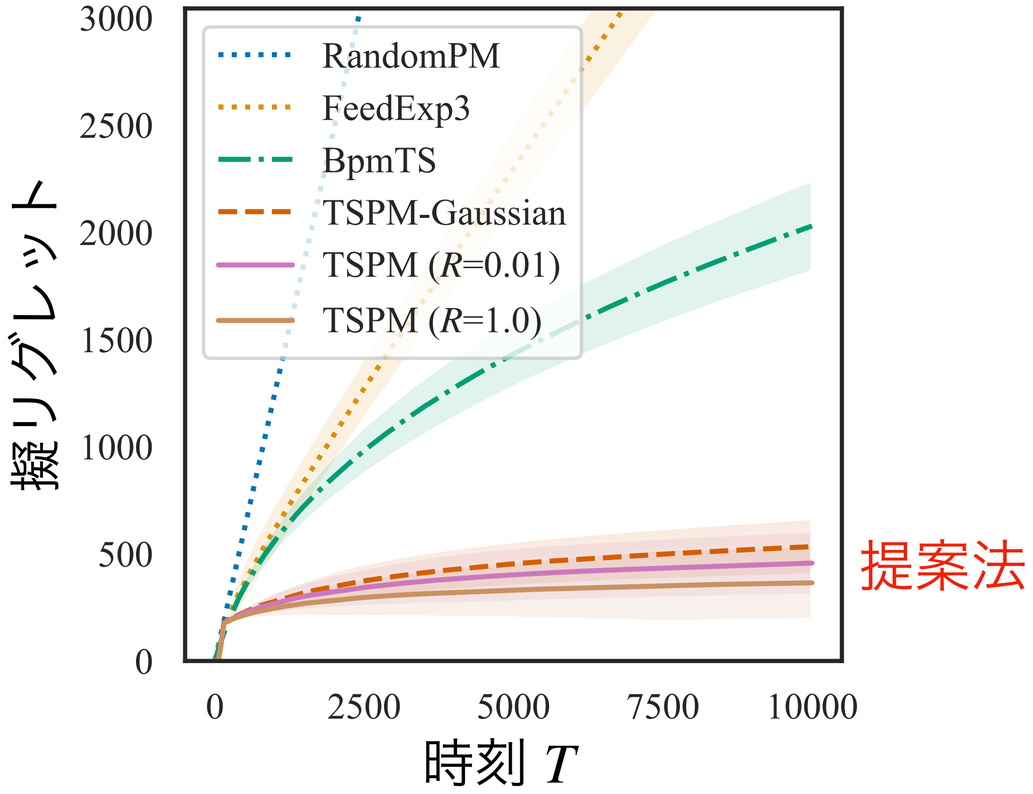
問題依存の定数　　ラウンド数 T への対数依存

- 部分観測問題に対する，Thompson抽出の初めての問題依存対数リグレット
- 線形バンディット問題に対する，Thompson抽出の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

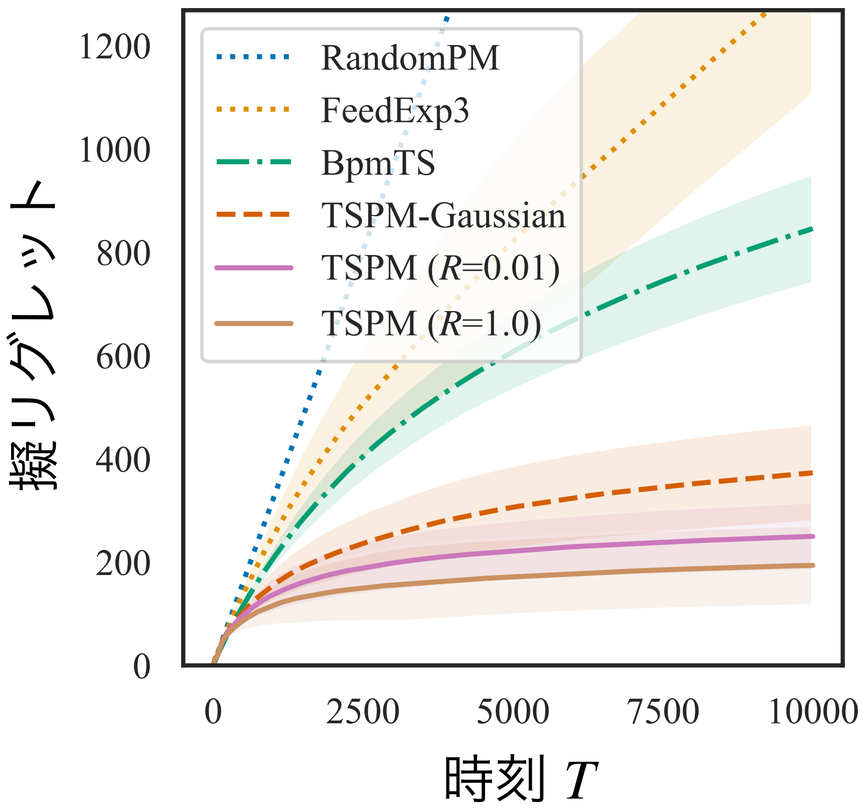
- 既存手法の性能を大きく改善

局所的観測可能ゲーム



提案法

大域的観測可能ゲーム



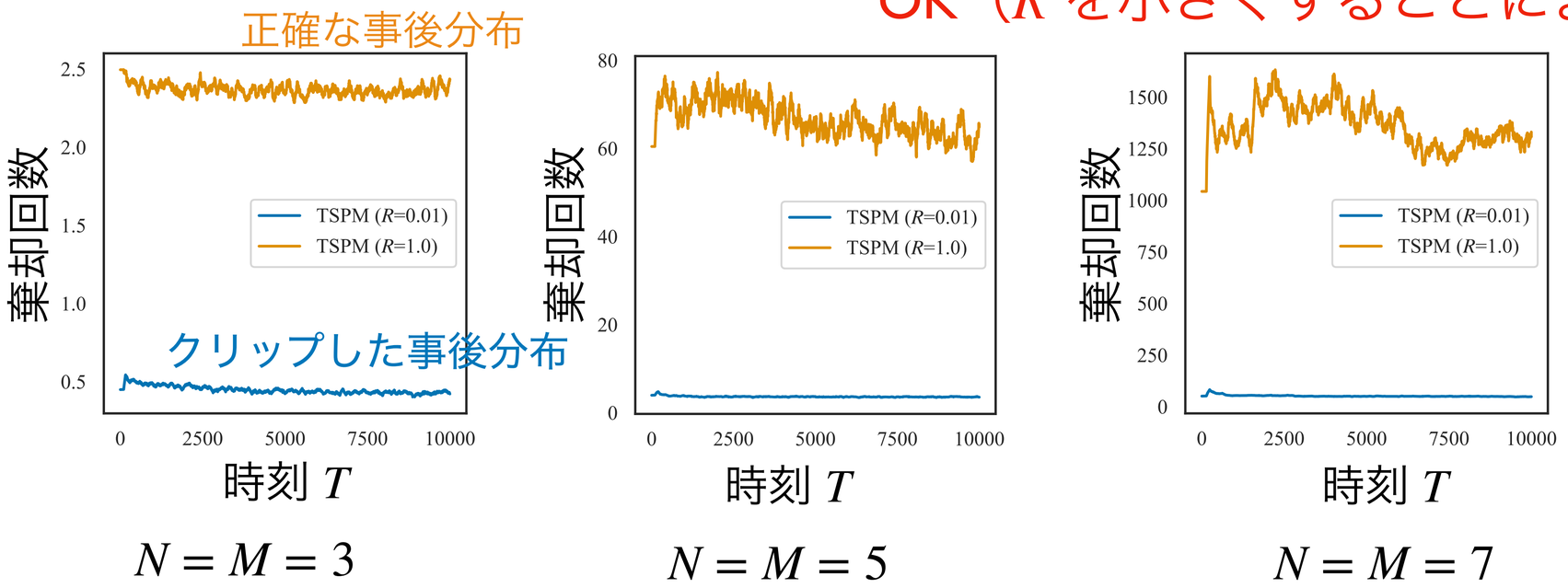
提案法

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

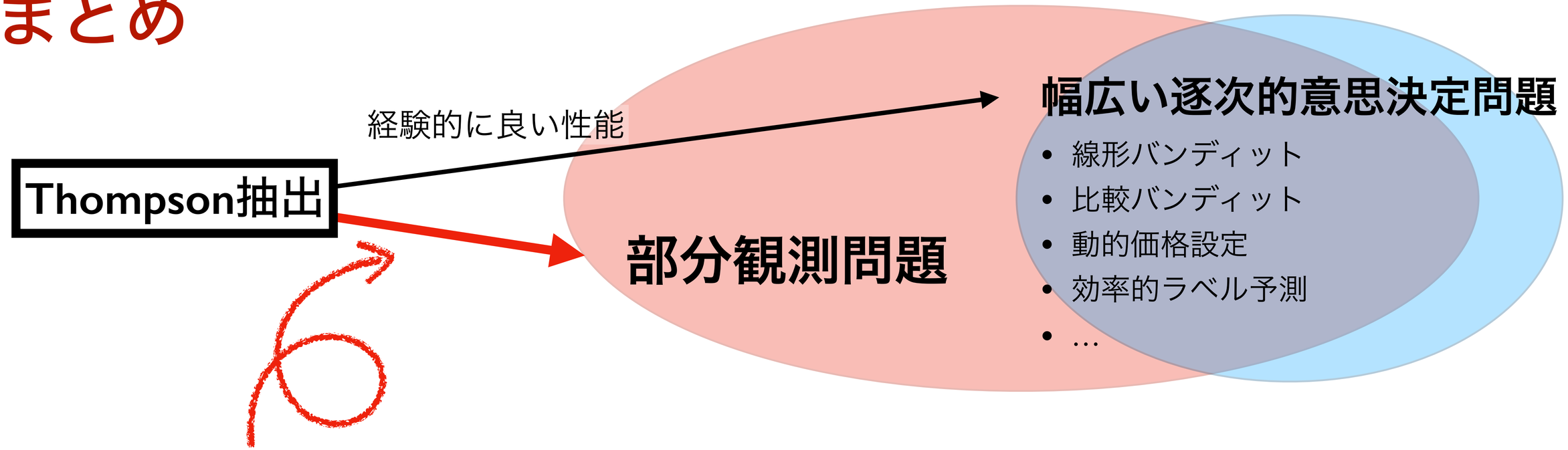
OK（R を小さくすることにより）

- 局所的観測可能ゲーム
- 動的価格設定



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

まとめ



1. どのように利用？

   タイトな提案分布からのサンプリングによる新しいThompson抽出アルゴリズム

2. 理論的に正当化可能？

   大域的観測可能性を満たす連続緩和した問題に対して可能

3. 実験的に有用？

   Yes! 理論的性質を示していない大域的観測可能ゲームに対しても有用