Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring

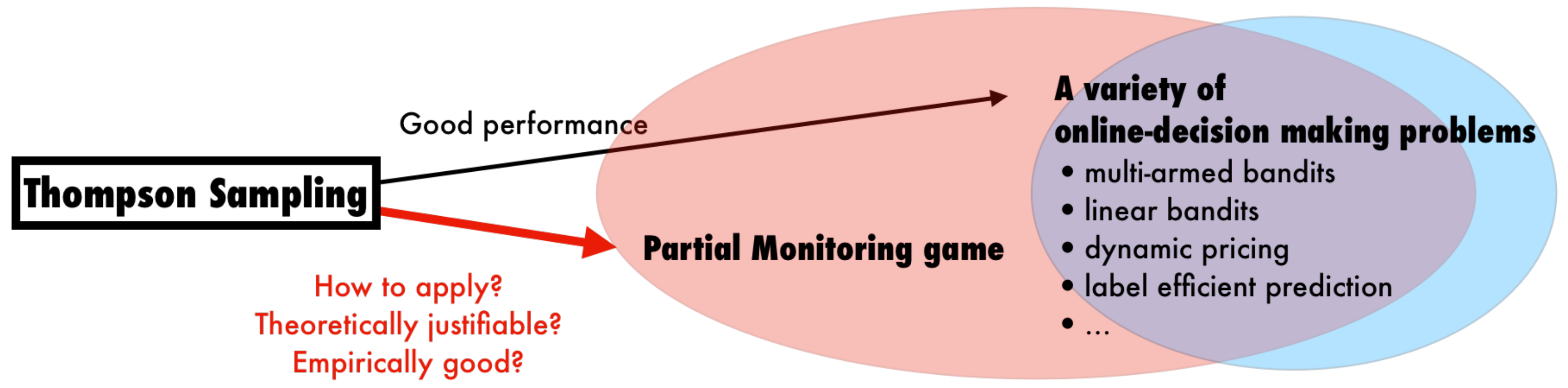
Taira Tsuchiya [1,2]
Junya Honda [1,2]
Masashi Sugiyama [2,1]

1. 東京大学 THE UNIVERSITY OF TOKYO
2. RIKEN

NEURAL INFORMATION PROCESSING SYSTEMS

Research Question

- Partial Monitoring (PM)
  - General framework for online-decision making with limited feedback
- Thompson Sampling (TS)
  - One of the most promising policies, especially for bandit problems
  - Handles the exploration/exploitation tradeoff by posterior sampling



Our Contribution

1. A novel TS-based algorithm based on a tight proposal distribution
2. First logarithmic regret upper bound both for PM and linear bandits

Background of Partial Monitoring

Formulation

- Partial monitoring game $G = (L, H)$ with N actions and M outcomes
- loss matrix $L = (\ell_{i,j}) \in \mathbb{R}^{N \times M}$, feedback matrix $H = (h_{i,j}) \in \Sigma^{N \times M}$ (Σ: set of feedback symbols)

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

1. **Player** selects action $i(t) \in \{1, \dots, N\}$ and play the action
2. **Opponent** selects outcome $j(t) \overset{\text{i.i.d.}}{\sim} \mathrm{Multi}(\underset{\text{strategy}}{\underline{p^*}})$ $(p^* \in \underset{\text{prob. simplex}}{\underline{\mathscr{P}_M}})$
3. Player suffers a loss $\ell_{i(t),j(t)}$ and observe feedback $y(t) = h_{i(t),j(t)}$

- Goal: minimize pseudo-regret

$$\mathrm{Reg}(T, p^*) = \sum_{t=1}^{T} \Big(\underbrace{L_{i(t)}^\top p^*}_{\text{expected loss for taken action}} - \underbrace{L_1^\top p^*}_{\text{expected loss for best action 1}} \Big)$$

w.l.o.g. action 1 is optimal

L_i : i-th column of L

- Seller (= player) sells an item for a specific price $i(t) \in [N]$
- Buyer (= opponent) comes with an evaluation price $j(t) \in [M]$

Example: Dynamic Pricing (dp-hard)

$$\ell_{i,j} = \begin{cases} j - i & (j \ge i) \\ c & (\text{otherwise}) \end{cases} \qquad h_{i,j} = \begin{cases} \text{buy} & (j \ge i) \\ \text{no-buy} & (\text{otherwise}) \end{cases}$$

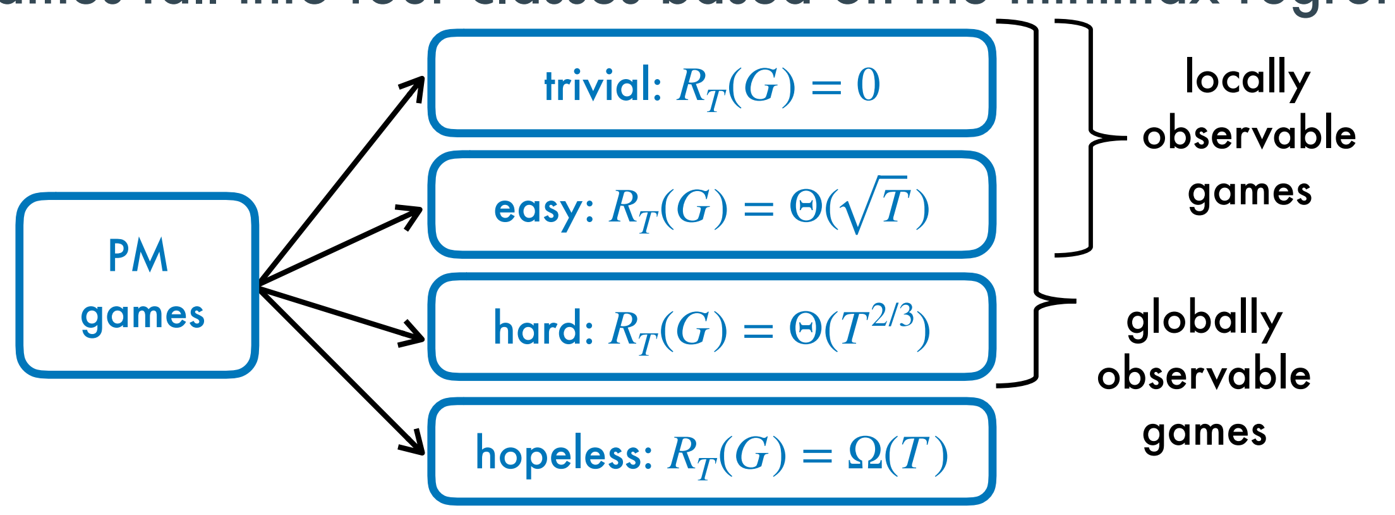
$$L = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ c & 0 & 1 & 2 & 3 \\ c & c & 0 & 1 & 2 \\ c & c & c & 0 & 1 \\ c & c & c & c & 0 \end{pmatrix} \quad H = \begin{pmatrix} \text{buy} & \text{buy} & \text{buy} & \text{buy} & \text{buy} \\ \text{no-buy} & \text{buy} & \text{buy} & \text{buy} & \text{buy} \\ \text{no-buy} & \text{no-buy} & \text{buy} & \text{buy} & \text{buy} \\ \text{no-buy} & \text{no-buy} & \text{no-buy} & \text{buy} & \text{buy} \\ \text{no-buy} & \text{no-buy} & \text{no-buy} & \text{no-buy} & \text{buy} \end{pmatrix}$$

$j \ge i$

$j < i$

Classification of Partial Monitoring Games [Bartók+ 2011]

PM games fall into four classes based on the minimax regret $R_T(G)$


PM games
trivial: $R_T(G) = 0$
easy: $R_T(G) = \Theta(\sqrt{T})$
hard: $R_T(G) = \Theta(T^{2/3})$
hopeless: $R_T(G) = \Omega(T)$
locally observable games
globally observable games

e.g., The dp-hard game belongs to the hard class.

Using Thompson Sampling for PM

1. Calculate a posterior dist. for the target parameter (strategy p^*)
   - $f_t(p) := \pi(p \mid \{i(s), y(s)\}_{s=1}^{t}) \propto \pi(p) \prod_{i=1}^{N} \exp\left(-n_i \mathscr{D}_{\mathrm{KL}}\left(q_i^{(t)} \| S_i p\right)\right)$
2. Sample the target parameter from the posterior distribution
   - sample $\tilde{p}_t \sim f_t(p)$
3. Decide the best action based on the sampled parameter and take it
   - take action $i(t) := \arg\min_{i \in [N]} L_i^\top \tilde{p}_t$

n_i : the # of times action i was taken by t
$q_i(t)$: emp fb dist of action i at t
S_i : signal matrix for action i

😨 Complicated posterior

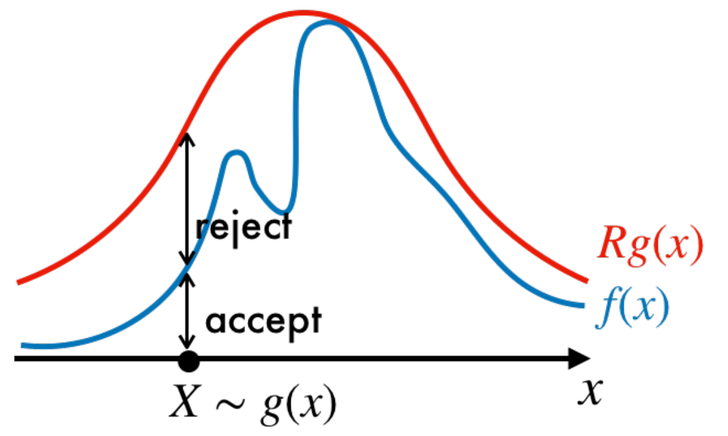
Existing Approach: BPM-TS [Vanchinathan+ 2014]

- Track strategy param. by Bayes-update with a Gaussian conjugate prior
- Assumption: the outcomes are generated from a Gaussian with covariance I_M and unknown mean (actually follows $\mathrm{Multi}(p^*)$)
- pros: fast computation
- cons: discrepancy from the exact posterior $f_t(p)$ & no theory

Proposed Algorithm (TSPM)

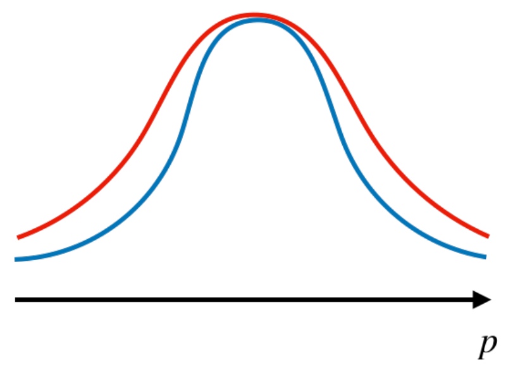
Accept-Reject Sampling

- A method to obtain i.i.d. samples from a complex distribution $f(x)$
- Prepare a *tight* proposal distribution $g(x)$ and do the following:
  1. Generate sample $X \sim g(x)$
  2. Accept X w.p. $f(X)/Rg(X)$, where $R = \sup_x f(x)/g(x)$
  3. Continue until getting accepted



TSPM (**TS**-based algorithm for **PM**)

Prepare a tight proposal distribution



Gaussian distribution

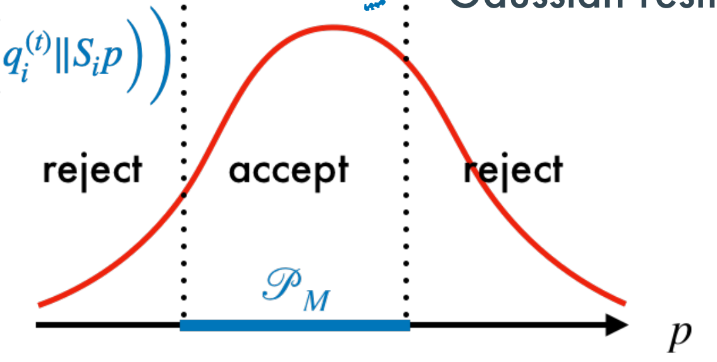
$$\overbrace{R\pi(p) \prod_{i=1}^{N} \exp\left(-n_i \|q_i^{(t)} - S_i p\|^2\right)}^{\text{Gaussian distribution}}$$

$\mathbin{\rotatebox{90}{$\geq$}}$ Pinsker's inequality

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

😨 restricted to $\mathscr{P}_M$

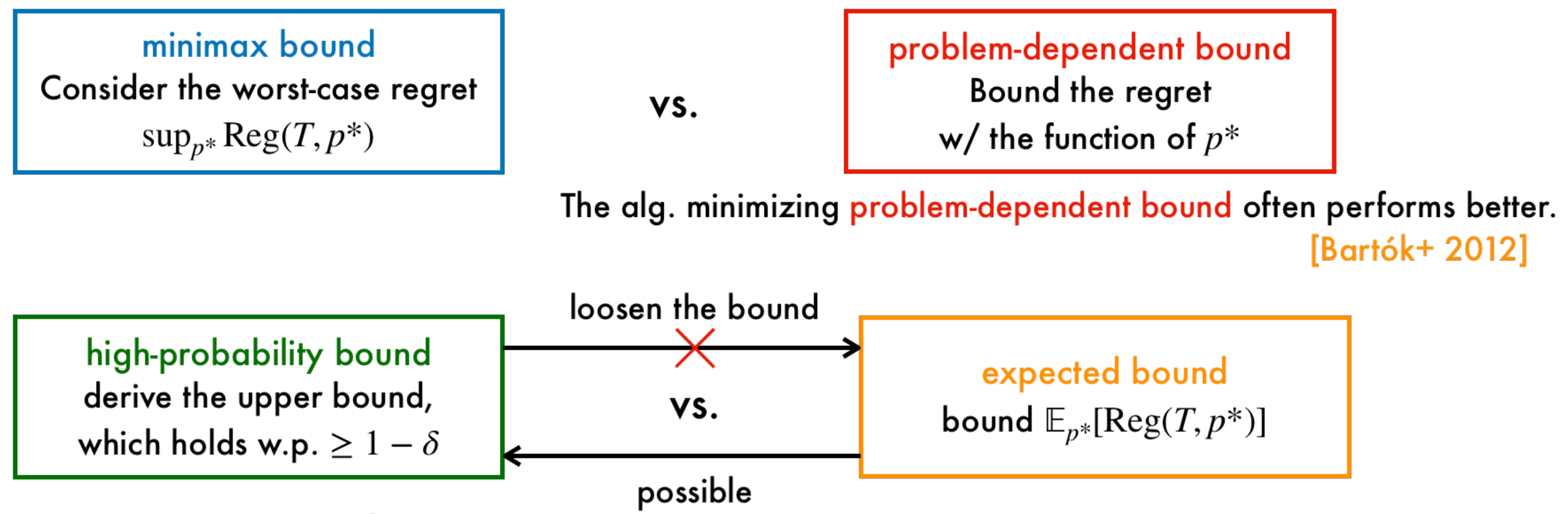
Sampling parameters from Gaussian restricted to $\mathscr{P}_M$



Theoretical Analysis

Types of Regret Upper Bounds

focus on the problem-dependent expected bound



minimax bound
Consider the worst-case regret
$\sup_{p^*} \mathrm{Reg}(T, p^*)$

vs.

problem-dependent bound
Bound the regret
w/ the function of p^*

The alg. minimizing problem-dependent bound often performs better. [Bartók+ 2012]

high-probability bound
derive the upper bound, which holds w.p. $\ge 1 - \delta$

vs.

expected bound
bound $\mathbb{E}_{p^*}[\mathrm{Reg}(T, p^*)]$

Regret Upper Bound

Theorem (informal). For any PM game w/ (strong) local observability, the expected pseudo-regret of **TSPM-Gaussian** is bounded by

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

Δ_i : sub-optimality gap for action i
$\Lambda = \min_{j \neq k} \Delta_{j,k} / \|z_{j,k}\|$
($\Delta_{j,k}$: loss gap between action j and k,
$z_{j,k} \in \mathbb{R}^{2A}$: vector relating loss and fb)

- The first logarithmic problem-dependent bound of TS for PM
- The first logarithmic bound of TS for *linear bandits*

What's Difficult in Theoretical Analysis?

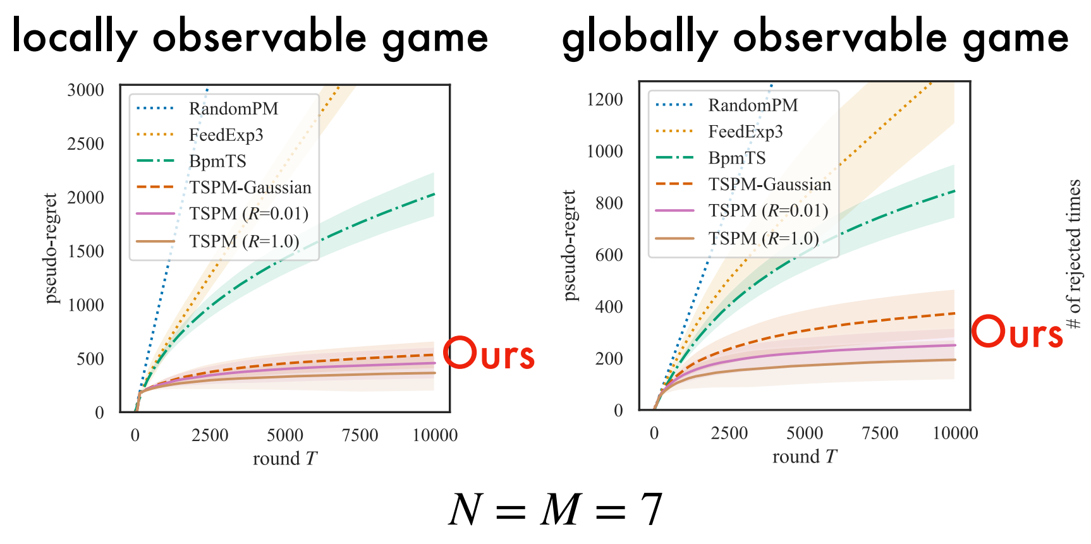
- Have to handle the effect of non-interested actions

$$\pi(p) \prod_{j=1}^{N} \exp\left(-n_j \mathscr{D}_{\mathrm{KL}}\left(q_j^{(t)} \| S_j p\right)\right)$$

  - Approach: evaluate the worst-case effect of non-interested actions
  - **Lemma**. $\mathbb{E}$[worst-case statistics of non-interested actions] $= \mathrm{O}(\log T)$
- Bound the probability that the optimal action is taken from below
  - Approach: use an argument of super-martingale

Experiments on Dynamic Pricing

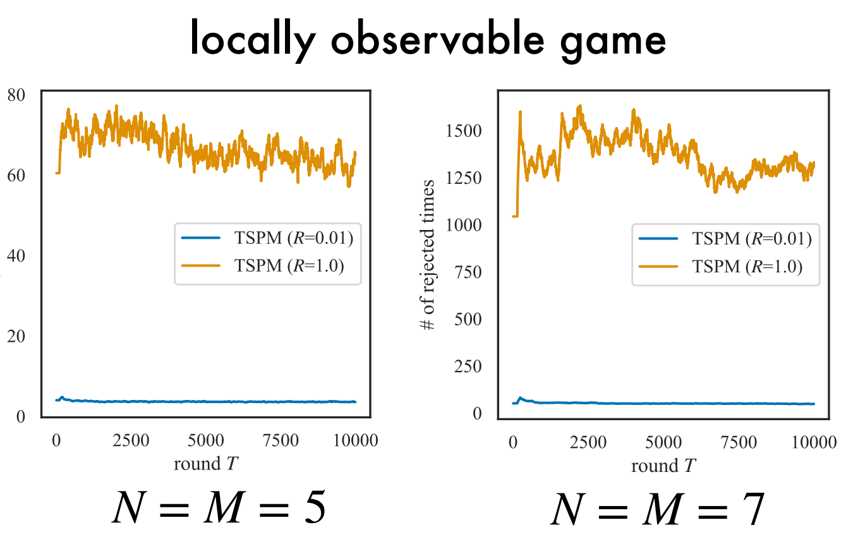
Regret Comparison



$N = M = 7$

substantially better performance than existing methods

Frequency of Rejection



$N = M = 5$ $N = M = 7$

freq of rejection does not increase as round proceeds & M becomes large

References

G. Bartók et al. (2011). "Minimax regret of finite partial-monitoring games in stochastic environments." In ICML'11.

G. Bartók et al. (2012). "An adaptive algorithm for finite stochastic partial monitoring." In ICML'12.

H. Vanchinathan et al. (2014). "Efficient partial monitoring with prior information." In NeurIPS'14.