Any questions during the talk are appreciated!

Towards Practical Algorithms for Online Decision-Making

Taira Tsuchiya

Kyoto University & RIKEN AIP

2022.07.08 @ INRIA Scool group seminar

Taira Tsuchiya

- 2nd year Ph.D. student at Kyoto University & RIKEN AIP
- Advised by Junya Honda

- Research interests
  - Wide range of statistical learning theory
  - Online-decision making problem, especially on bandits!

Today's Talk

1. Thompson sampling for stochastic partial monitoring (NeurIPS2020)
   - available at https://arxiv.org/abs/2006.09668

2. A best-of-both-worlds algorithm with variance-dependent regret bounds (COLT2022)
   - available at https://arxiv.org/abs/2206.06810

- Advertisement:
  - "Globally" optimal best arm identification for fixed-budget setting
  - available at https://arxiv.org/abs/2206.04646

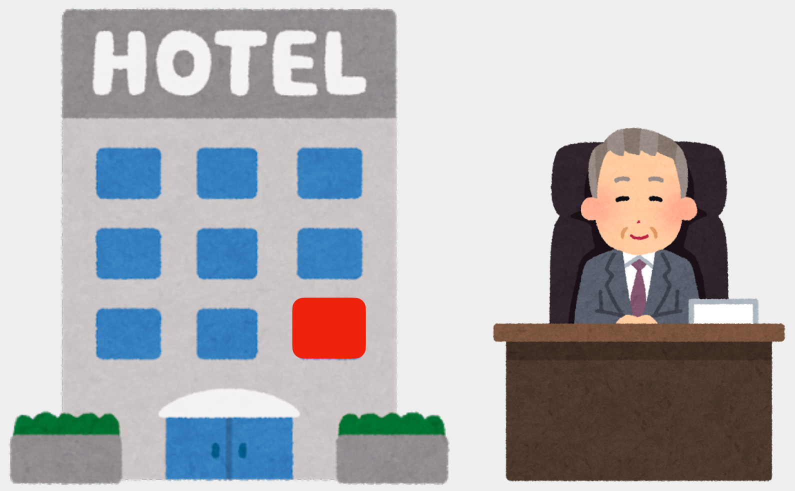
Analysis and Design of Thompson Sampling for Stochastic Partial Monitoring (NeurIPS2020)

Taira Tsuchiya [1,2], Junya Honda [2,3], Masashi Sugiyama [2,3]

1. The University of Tokyo, 3. RIKEN AIP

Partial Monitoring Example: Dynamic Pricing

Player (= seller)		**Environment** User's outcome $j(t)$ (= evaluation price)	opportunity loss	feedback
Hotel owner	$\underline{t = 1}$ selling price $\$40$	Use if selling price $\le \$90$	$\$90 - \$40 = \$50$	Buy
decides the price of room from $\{\$1, \ldots, \$N\}$	$\underline{t = 2}$ selling price $\$80$	Use if selling price $\le \$50$	$\$c$ (const.) ($\because \$50 - \$80 < \$0$)	No-buy



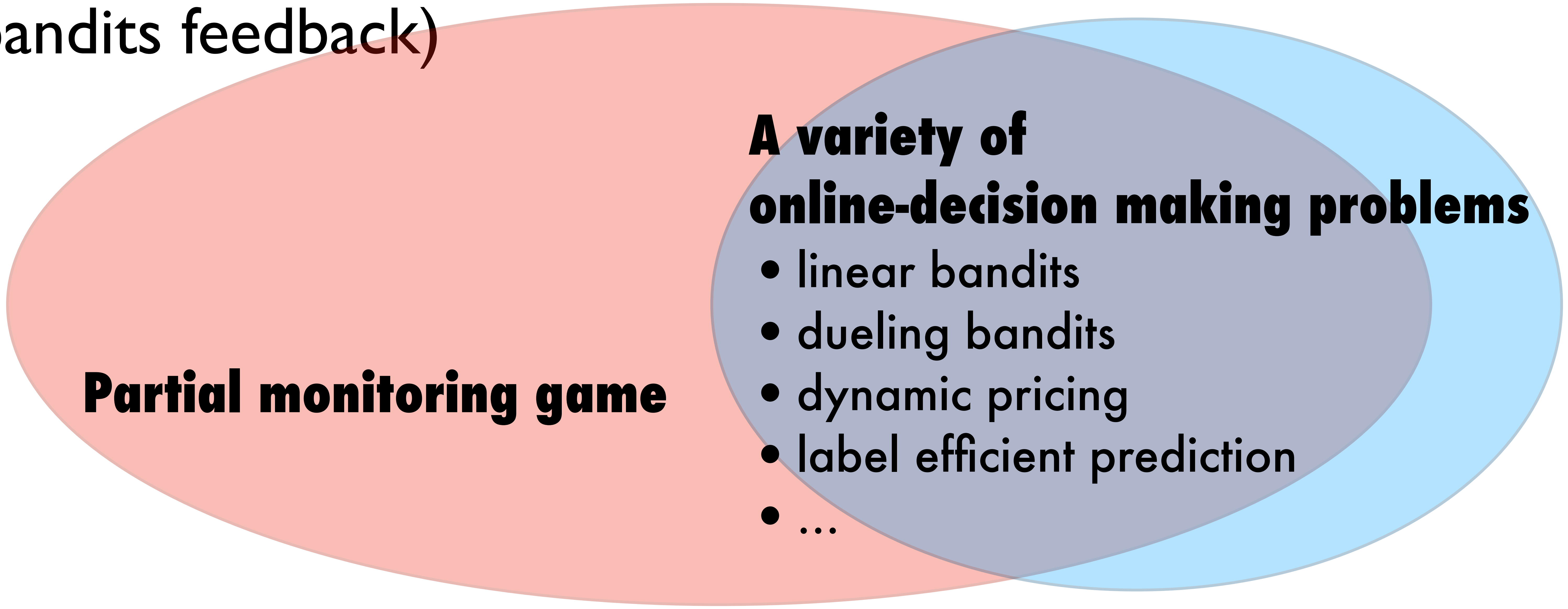
$\underline{t = \cdots}$

Only feedback (Buy or No-Buy) is observable to the seller!

Q. Is it possible to maximize the total reward (= minimize the total loss) only with limited feedback?

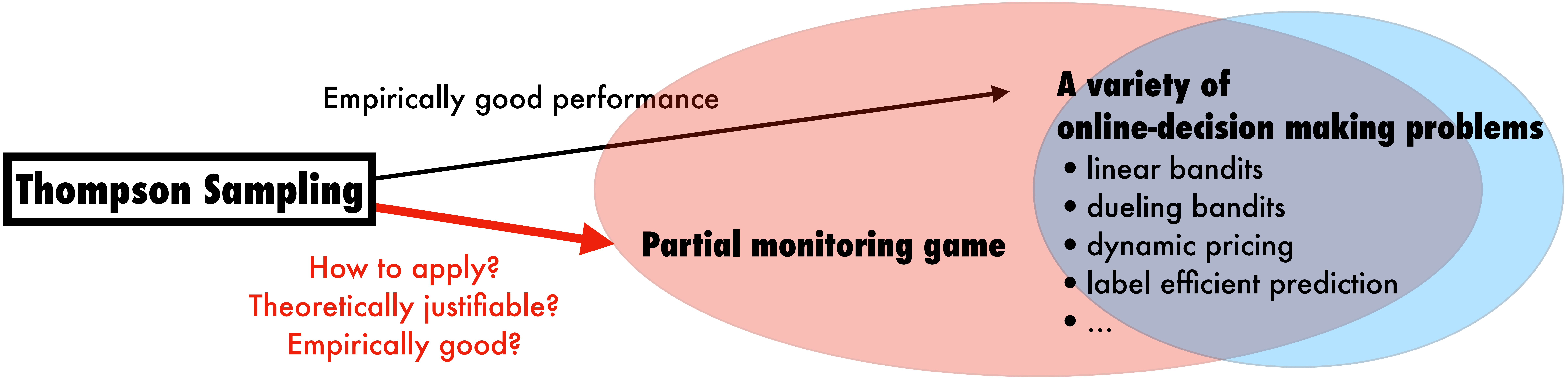
Partial Monitoring has Many Applications

- Online learning with full information
  (the loss is directly observed)
- Linear bandits
- Heteroscedastic bandits
- Dueling bandits
- Combinatorial bandits
  (both w/ (full-)bandits and semi-bandits feedback)
- Dynamic pricing
- Label efficient prediction
- ….



Research Question

- Partial Monitoring
  - General Framework for online-decision making problem with limited feedback
- Thompson Sampling
  - (Empirically) one of the most promising policies for various online decision-making problems
  - Handles the exploration/exploitation tradeoff by posterior sampling



Outline | Thompson Sampling for Partial Monitoring

- Introduction - partial monitoring and research question
- Background of partial monitoring
- Existing Thompson sampling based approach
- Proposed algorithms
- Regret upper bound
- Experiments
- Conclusion

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)
- PM game

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

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

- Goal: minimize pseudo-regret (= maximize total rewards)

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

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

L_i : i-th column of L

Example 1: Dynamic Pricing

N: the (discrete) range of selling price
M: the (discrete) range of evaluation price

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

action

outcome $\sim p^*$(strategy)

Player (= seller)

$t = 1$

Hotel owner

decides the price of room from $\{\$1, \ldots, \$N\}$

selling price $\$40$

$t = 2$

selling price $\$80$

$t = \cdots$

Environment

User's outcome $j(t)$ (= evaluation price)

Use if selling price $\leq \$90$

Use if selling price $\leq \$50$

opportunity loss

$\$90 - \$40 = \$50$

$\$c$ (const.) ($\because \$50 - \$80 < \$0$)

feedback

Buy

No-buy

$\Sigma = \{$Buy($\bigcirc$), No-Buy($\times$)$\}$

loss matrix (*)

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

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

$j \geq i$ $j < i$

feedback matrix (*)

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

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

$j \geq i$ $j < i$

(*row: selling price, column: outcome)

Example 2: Label Efficient Prediction [Cesa-Bianch+ 2005]

- Player predicts label (positive or negative) of the item in online manner
- There possible actions when labeling items:
  1. label as positive (P)
  2. label as negative (N)
  3. ask a expert (The true label is given.)

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

$c_{\mathsf{N}\to\mathsf{P}} > 0$: failure cost of N to P

$c_{\mathsf{P}\to\mathsf{N}} > 0$: failure cost of P to N

$q > 0$: cost of asking the expert

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

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. IEEE Transactions on Information Theory, 51(6):2152–2162, 2005.

Classification of Partial Monitoring Games [Bartók+ 2010, 2011]

Q. Can we achieve sub-liner regret for any PM game $G = (L, H)$?

A. No. We need some conditions.

e.g., $L = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, H = \begin{pmatrix} a & a \\ a & a \end{pmatrix}$

- PM games fall into four classes based on their minimax regret $R_T(G) = \inf_{\mathscr{A}} \sup_{p \in \mathscr{P}_M} \mathbb{E}_p[R_T(\mathscr{A}, p)]$

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 (trivial, easy)

globally observable games (trivial, easy, hard)

Taking pair of actions $i, j \in [N]$ is enough to know the value of $(L_i - L_j)^\top p$ (intuitively)

Taking pair of actions $i, j \in [N]$ can be NOT enough to know $(L_i - L_j)^\top p$

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. In *Algorithmic Learning Theory*, pages 224–238, 2010.
Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *the 24th Annual Conference on Learning Theory*, volume 19, pages 133–154, 2011.

Outline | Thompson sampling for Partial Monitoring

- Introduction - partial monitoring and research question
- Background of partial monitoring
- Existing Thompson sampling based approach
- Proposed algorithms
- Regret upper bound
- Experiments
- Conclusion

How to use Thompson Sampling in Partial Monitoring?

Target parameter: strategy $p \in \mathscr{P}_M$

A naive application of Thompson sampling:

1. calculating posterior distribution for parameters
   $f_t(p) := \pi(p \mid \text{observed data}(t)) \propto \pi(p) \prod_{i=1}^{N} \exp\left(-n_i \mathscr{D}_{\mathrm{KL}}\left(q_i^{(t)} \| S_i p\right)\right)$
2. sampling target parameters from posterior distribution
   sample $\tilde{p}_t \sim f_t(p)$

   😨 Complicated Posterior
3. deciding the best action (= arm) based on sampled parameters and take it
   take action $i(t) := \arg\min_{i \in [N]} \underbrace{L_i^\top \tilde{p}_t}_{\text{expected loss for action } i}$

n_i : the # of times action i was taken by time t
$q_i(t)$: empirical feedback dist. of action i at t
S_i : signal matrix of action i (Appendix)

Bayes-update Partial Monitoring (BPM-TS) [Vanchinathan+ 2014]

- Track strategy (p^*) estimate by Bayes-update with a Gaussian prior
- Assumption: the outcomes are generated from a Gaussian with covariance I_M and unknown mean (actually follows $\mathrm{Multi}(p^*)$)

😊 Fast computation

😊 One of the best experimental performances

😨 Discrepancy from the exact posterior $f_t(p)$

$$\mathcal{N}(\text{some params at } t) \underset{\text{discrepancy}}{\longleftrightarrow} \pi(p)\prod_{i=1}^{N} \exp\left(-n_i \mathcal{D}_{\mathrm{KL}}\left(q_i^{(t)} \| S_i p\right)\right)$$

😨 No theoretical analysis is given for TS setting

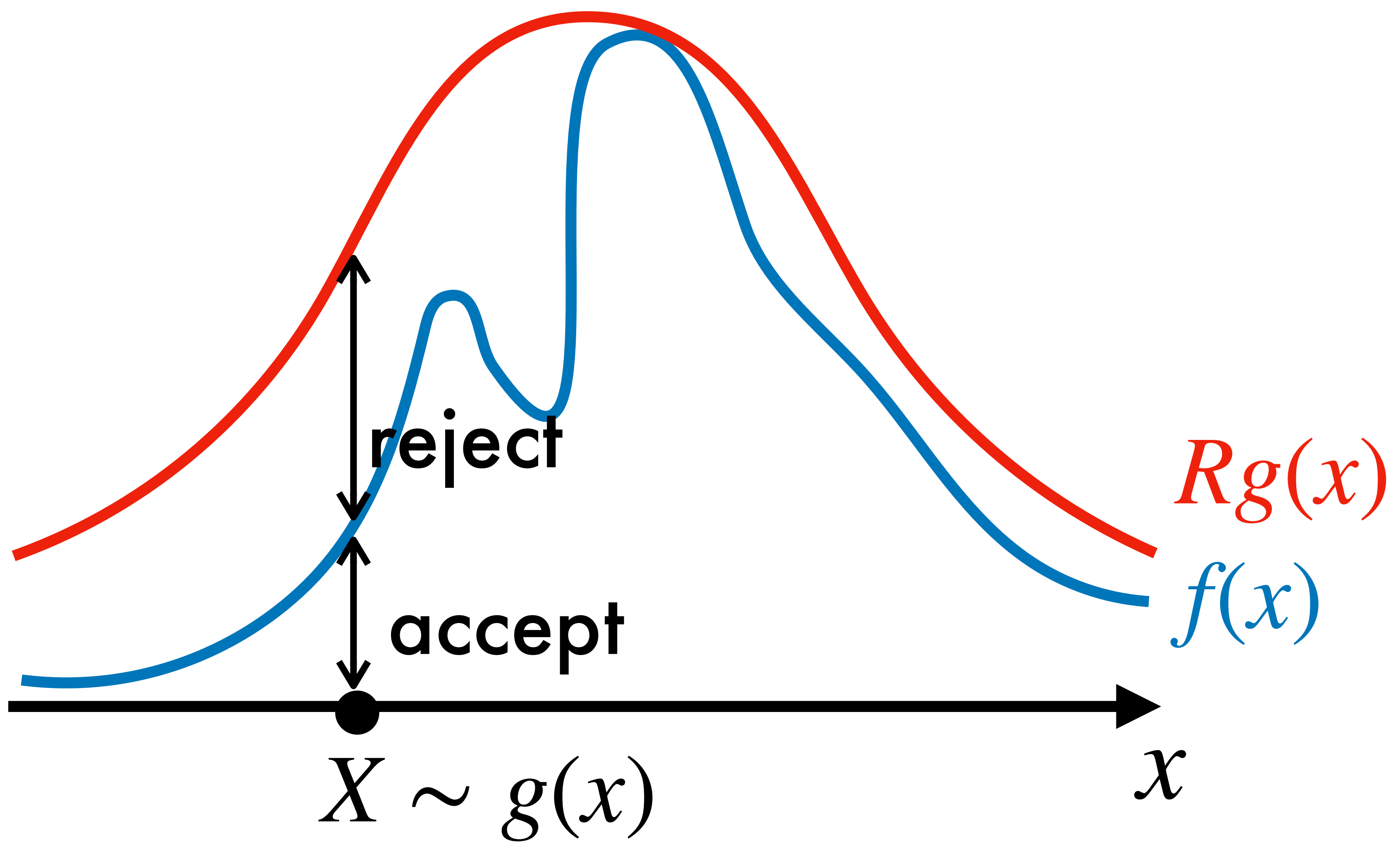
Hastagiri P Vanchinathan, Gábor Bartók, and Andreas Krause. Efficient partial monitoring with prior information. In Advances in Neural Information Processing Systems 27, pages 1691–1699, 2014.

Accept-Reject Sampling

- A method to obtain i.i.d. samples from a certain complex distribution $f(x)$
- Prepare a proposal distribution $g(x)$ and do the following:

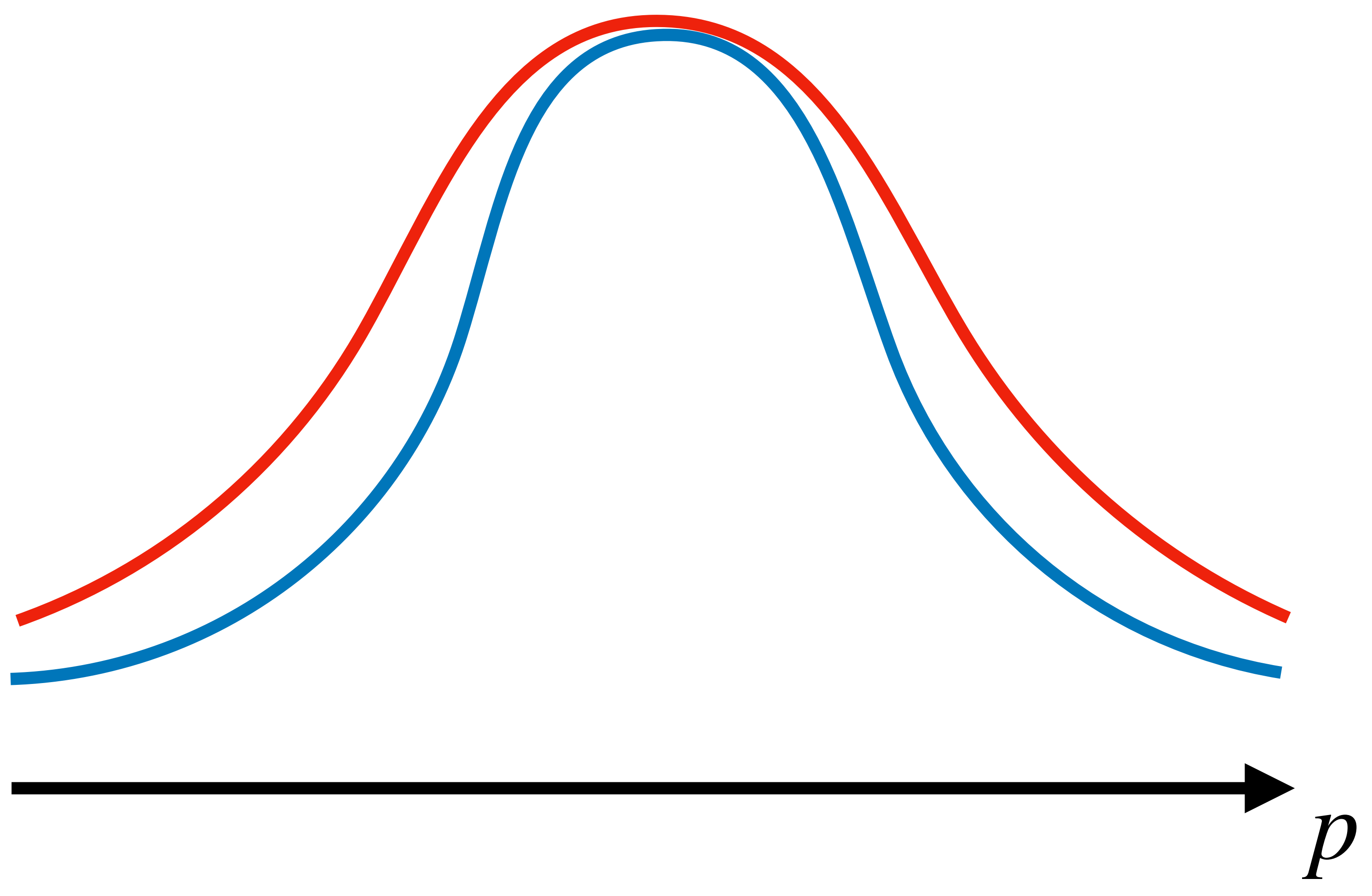
  Easy to obtain samples

  1. Generate sample $X \sim g(x)$
  2. Accept X with probability $\dfrac{f(X)}{Rg(X)}$, where $R = \sup_x \dfrac{f(x)}{g(x)}$
  3. Continue until getting accepted
- Need to prepare a *tight* proposal distribution



Proposed algorithm (TSPM) | Exact Posterior Sampling

1. Prepare a tight proposal distribution



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

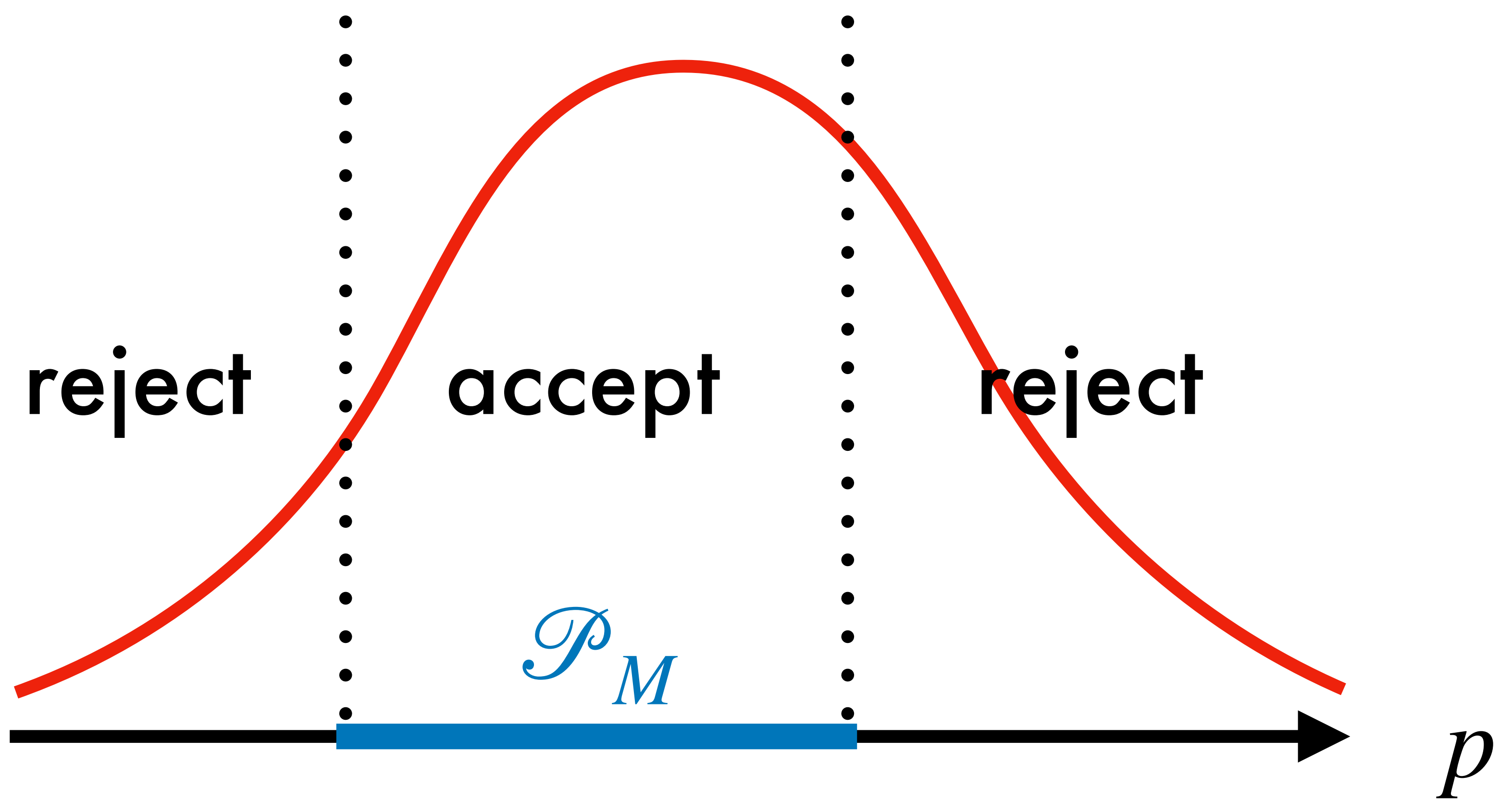
$$\geq \quad \text{Pinsker's inequality}$$

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

2. Sampling from the Gaussian distribution restricted to probability simplex $\mathscr{P}_M$

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

😨 restricted to $\mathscr{P}_M$



Summary of Proposed Method

Hard to directly sample target parameters from posterior distribution

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

Existing work [Vanchinathan+ 2014]

Approximate by Gaussian distribution

😊 Fast computation

😨 Discrepancy from the exact posterior

😨 No theoretical analysis is given for TS

Ours (**TSPM**)

Exact sampling by the tight proposal distribution

😊 Good empirical performance w/o much computational cost

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.

Outline | Thompson Sampling for Partial Monitoring

- Introduction - partial monitoring and research question
- Background of partial monitoring
- Existing Thompson sampling based approach
- Proposed algorithms
- Regret upper bound
- Experiments
- Conclusion

Logarithmic Regret Upper Bound

A, N : the # of feedback and action,
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 feedback)

Thm. (informal)

For any *linear* partial monitoring game with *local observability*, the expected pseudo-regret of **TSPM-Gaussian** is bounded by

$$O\left(\underbrace{\max\left\{ \frac{A \sum_{i \in [N]} \Delta_i}{\Lambda^2}, \frac{\sqrt{A} N^3 \max_{i \in [N]} \Delta_i}{\Lambda^2} \right\}}_{\text{some problem-dependent constants}} \underbrace{\log T}_{\text{dependence on time horizon } T} \right).$$

☑ The first logarithmic problem-dependent bound of TS for partial monitoring

☑ The first logarithmic bound of **Thompson sampling** for **Linear Bandits!**

What's New in Theoretical Analysis?

- Have to handle the effect of non-interested actions
- Bound the regret for each sub-optimal action $i \in [N]\backslash\{1\}$ (regret decomposition)

$$\mathrm{Reg}(T) = \textstyle\sum_{i\in[N]\backslash\{1\}} \Delta_i \underline{N_i(T+1)}$$

total # of times to pull sub-optimal action i

Multi-armed bandits

$\pi(\mu_j \mid \text{observed data}(t)) \longleftrightarrow \pi(\mu_k \mid \text{observed data}(t))$

independent for $j \neq k$

Partial monitoring

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

all actions except action i are of no-interest, but its statistic appear in posterior

$\boxed{\text{Approach: evaluate the worst-case effect of non-interested actions}}$

Lem. $\mathbb{E}[\text{worst-case statistics of non-interested actions}] = \mathrm{O}(\log T)$

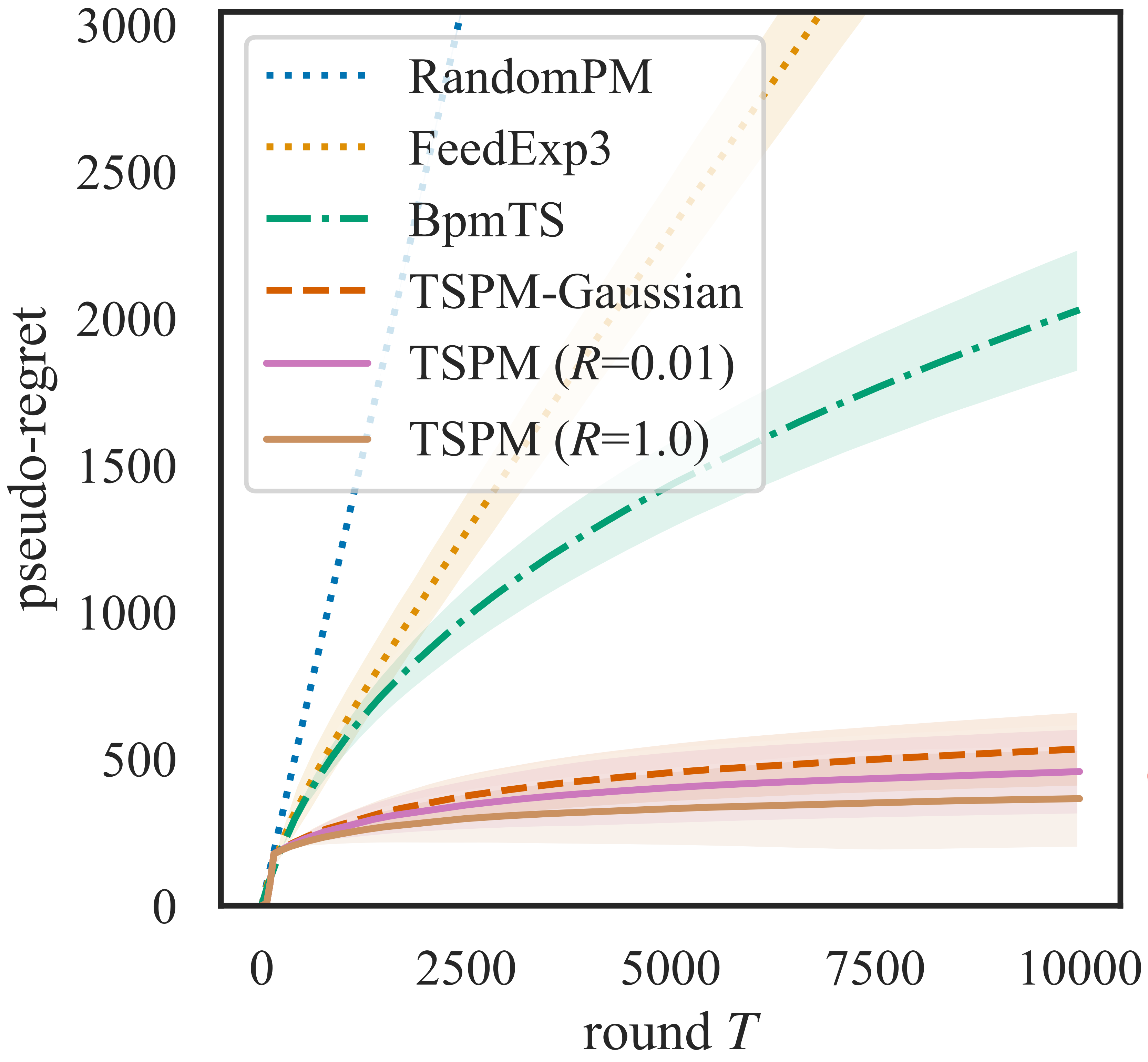
Outline | Thompson Sampling for Partial Monitoring

- Introduction - partial monitoring and research question
- Background of partial monitoring
- Existing Thompson sampling based approach
- Proposed algorithms
- Regret upper bound
- Experiments
- Conclusion

Performance Comparison on Dynamic Pricing

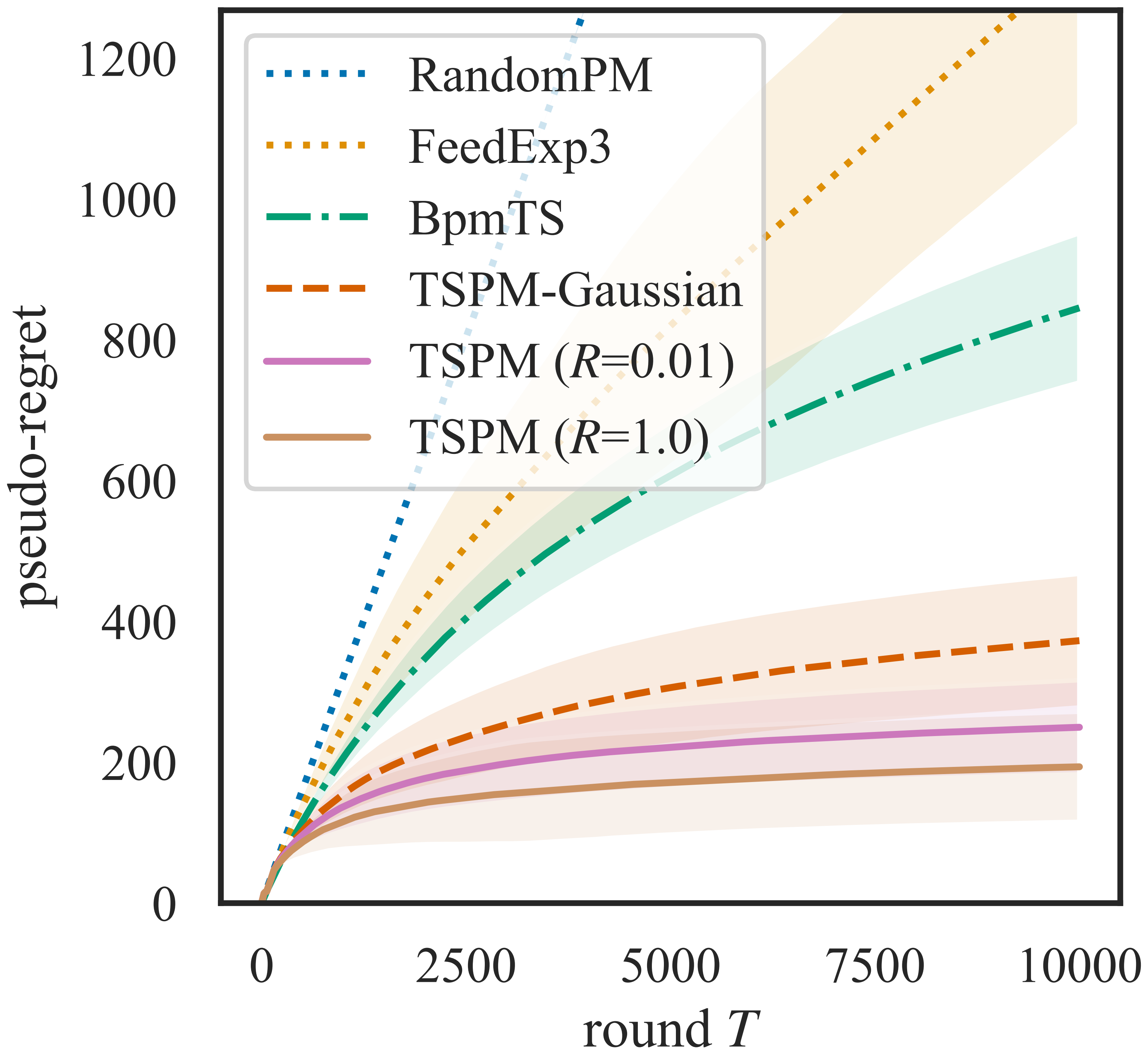
- Substantially better performance than existing methods

locally observable game



Ours

globally observable game



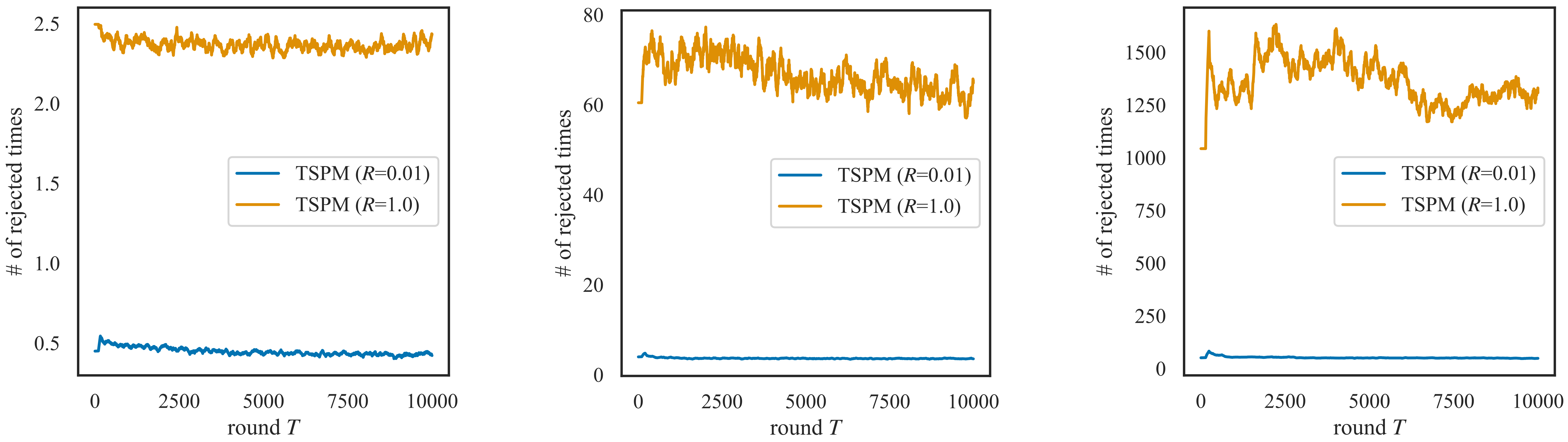
Ours

Frequency of Rejection in Accept-Reject Sampling

- Desirable Properties of Accept-Reject Sampling
  1. Frequency of rejection does not increase as round proceeds. OK
  2. Frequency of rejection does not increase as the support dimension $M-1$ increases.

OK (by setting R to be small value)

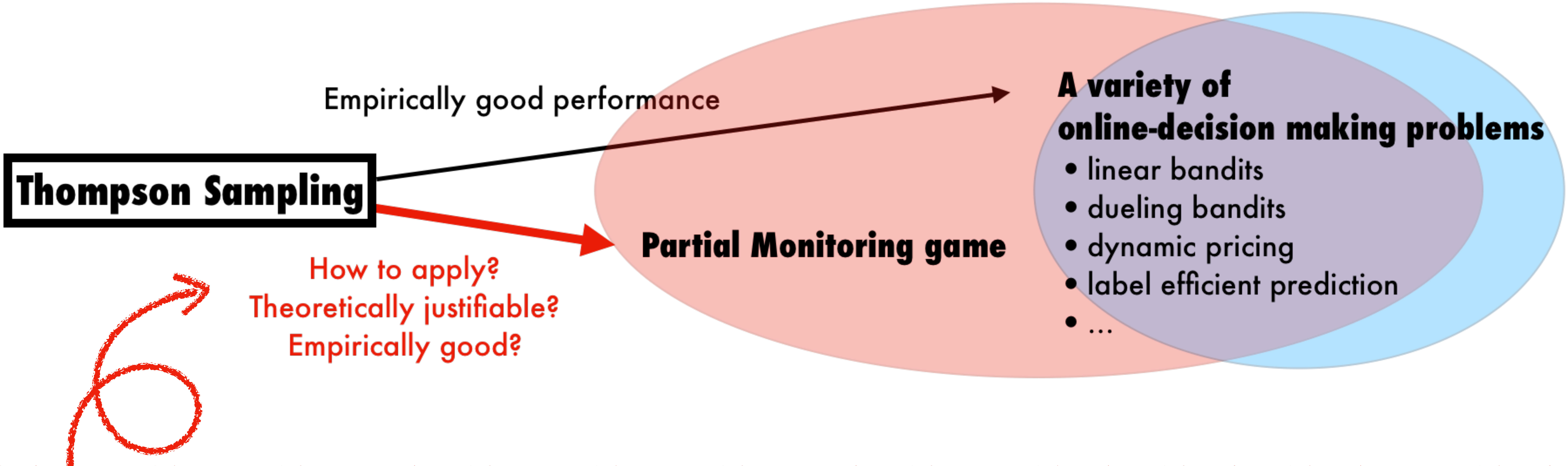
(locally observable game)



$N = M = 3$ $\quad$ $N = M = 5$ $\quad$ $N = M = 7$

$$\boxed{\text{Accept w.p. of } \mathrm{const} \cdot \frac{\text{posterior density}}{R \cdot \text{proposal dist. density}}}$$

Conclusion | Thompson Sampling is useful in PM!



Our contribution:

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

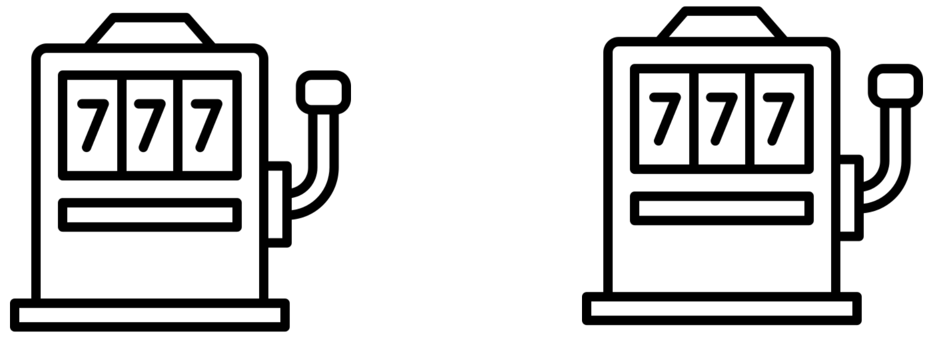
Adversarially Robust Multi-Armed Bandit Algorithm with Variance-Dependent Regret Bounds (COLT2022)

Shinji Ito [1,3], Taira Tsuchiya [2,3], Junya Honda [2,3]

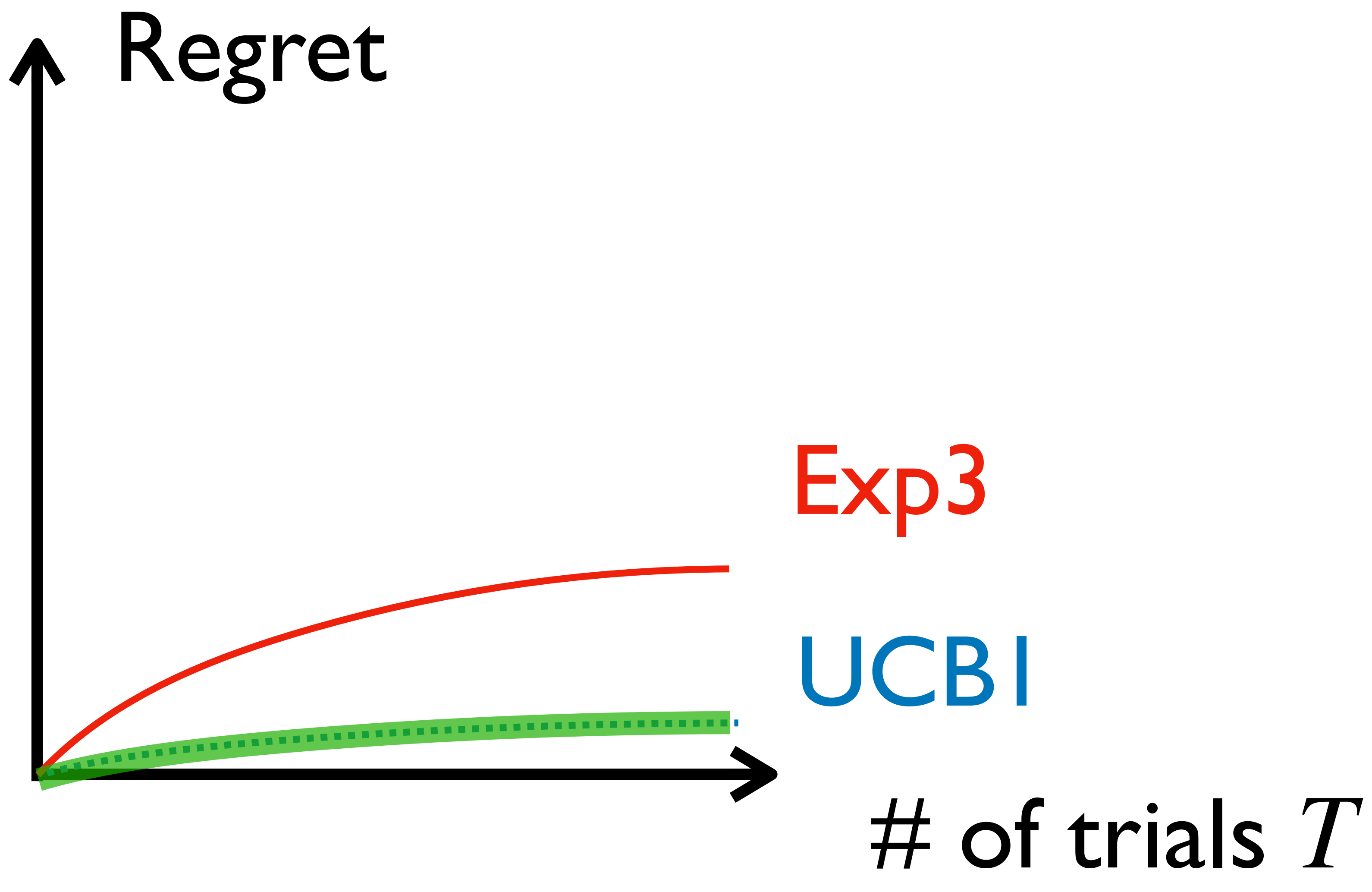
1. NEC Corporation, 2. Kyoto University, 3. RIKEN AIP

Any Policy Optimal Both for Stochastic and Adversarial?

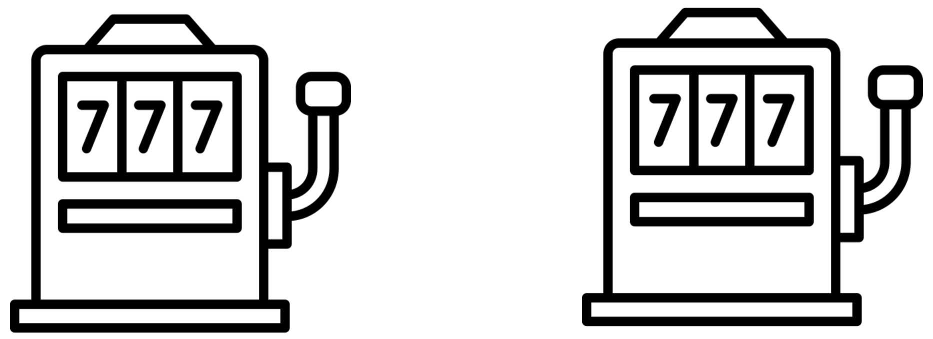
<u>Stochastic regime</u>



arm 1 | arm 2
Ber(0.9) | Ber(0.1)

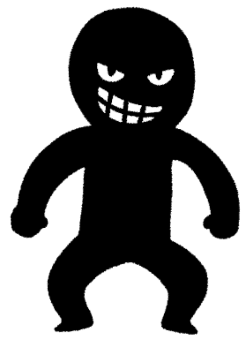


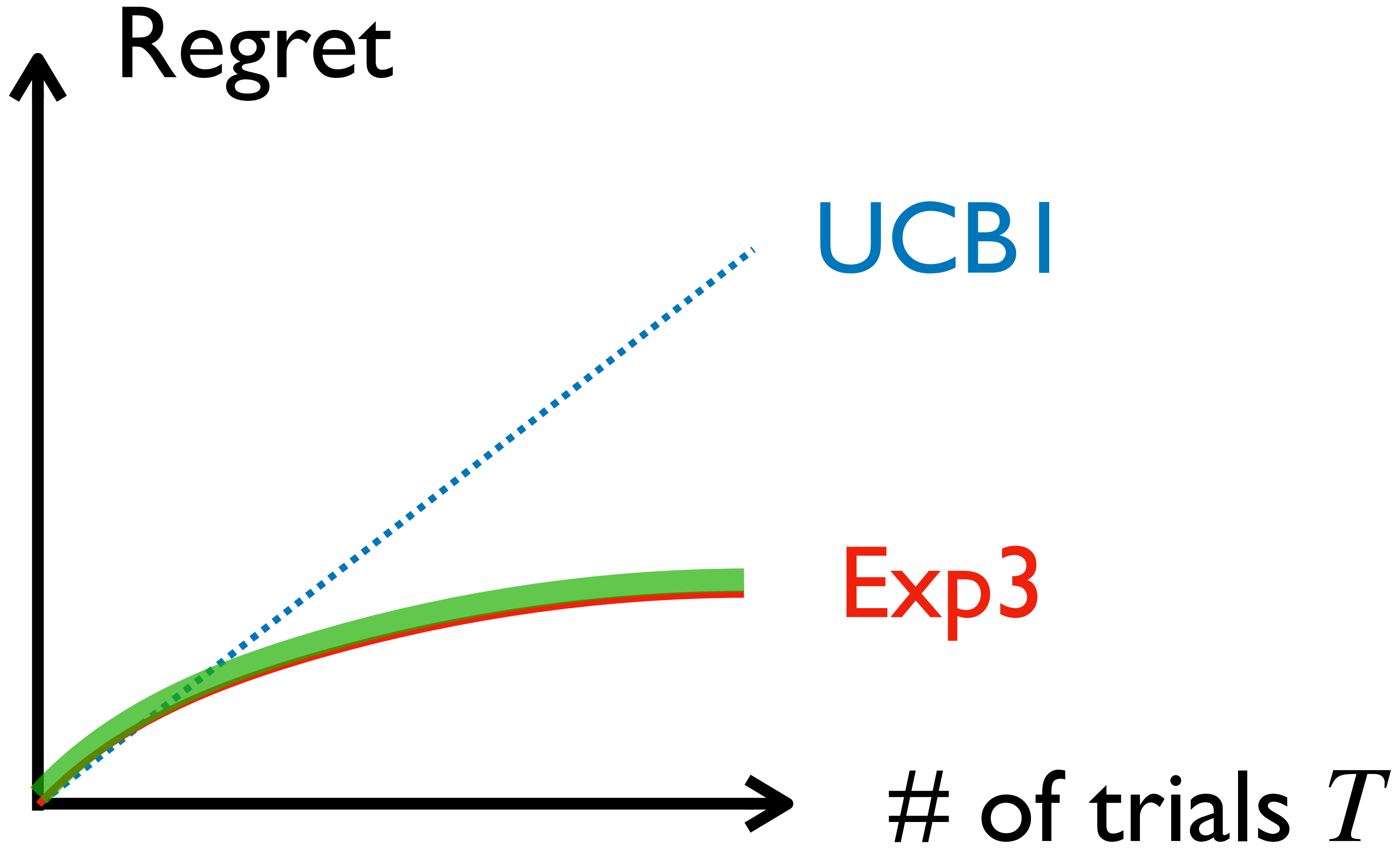
<u>Adversarial regime</u>



arm 1 | arm 2
Ber(0.9) | Ber(0.1)

A bad guy can exchange arms
at any time based on past history





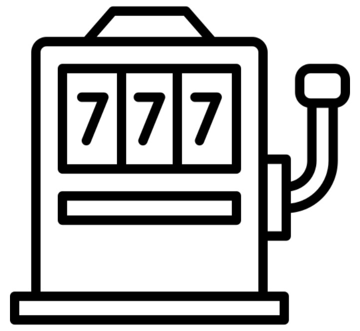
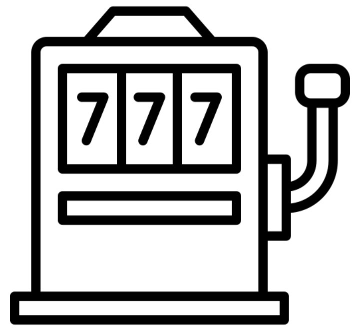
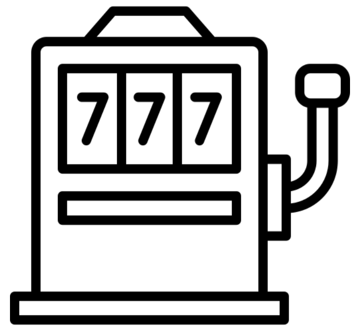
Q. Are there algorithm working well both for stochastic & adversarial **w/o knowing regime?**
If possible, can we make use of **distributional information?**

Outline | Variance-Dependent BOBW Algorithm

- Introduction to best-of-both-worlds algorithms
- Super quick introduction on vanilla multi-armed bandits
- Background (existing approaches & intermediate regime)
- Tsallis-INF algorithm
- Our work
  - Regret bounds on three regimes
  - Preliminary (Optimistic FTRL)
  - Proposed algorithm
  - Numerical experiments
- Conclusion

Stochastic Multi-armed Bandits

- Online decision making model with K unknown distributions (on $[0,1]$ (*)) $(P_i)_{i=1,\dots,K}$
  (= arm, action)



. . .


distribution / expected loss

Arm 1 P_1 / μ_1

Arm 2 P_2 / μ_2

Arm K P_K / μ_K

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

1. Player selects arm $I(t) \in \{1,\dots,K\}$
2. Observe stochastic loss of $I(t)$, $\ell_{t,I(t)} \sim P_{I(t)}$

Only the loss for selected arm $I(t)$ is observed

- Goal: minimize (pseudo-)regret: $\mathrm{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{I(t)} - \mu_{i^*})\right]$, $\quad i^* = \arg\min_i \mu_i$
- Need to handle the exploration & exploitation tradeoff

Pull the arm with large uncertainty

Pull the arm which looks optimal

*We consider losses instead of rewards in this talk

Algorithm for Stochastic Regime | UCB

- UCB algorithm : optimistically estimate the reward (= negative of the loss) of arms

[Auer+ 2002]

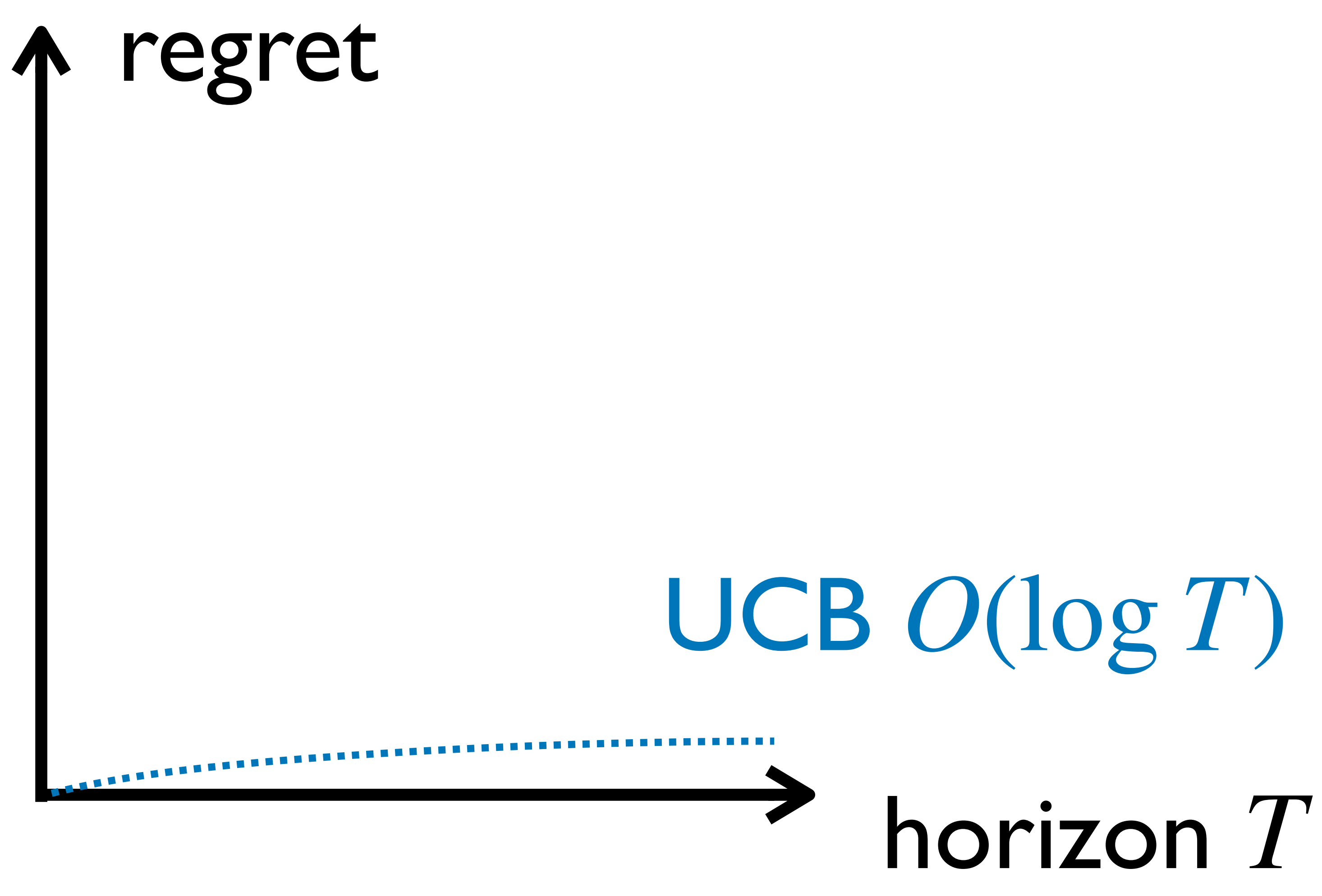
1. Pull each arm once in $[K] := \{1,\ldots,K\}$
2. Pull the arm with the highest <u>optimistically estimated rewards $\mathrm{UCB}_i(t-1)$</u>

$$\mathrm{UCB}_i(t-1) := \hat{\mu}_i(t-1) + \sqrt{\frac{\log T}{N_i(t-1)}}$$

reward mean

exploration term : The fewer the # of times an arm has been pulled so far, the more likely it is to be pulled.

$$\hat{\mu}_i(t-1) := \frac{1}{N_i(t-1)} \sum_{s=1}^{t-1} \tilde{\mu}_s \, 1[i(s) = i]$$



P. Auer, N. Cesa-Bianchi, & P. Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning, 2002.

Adversarial Multi-armed Bandits

Adversarial bandits: the losses for each step $\ell_t \in [0,1]^K$ are completely arbitrary

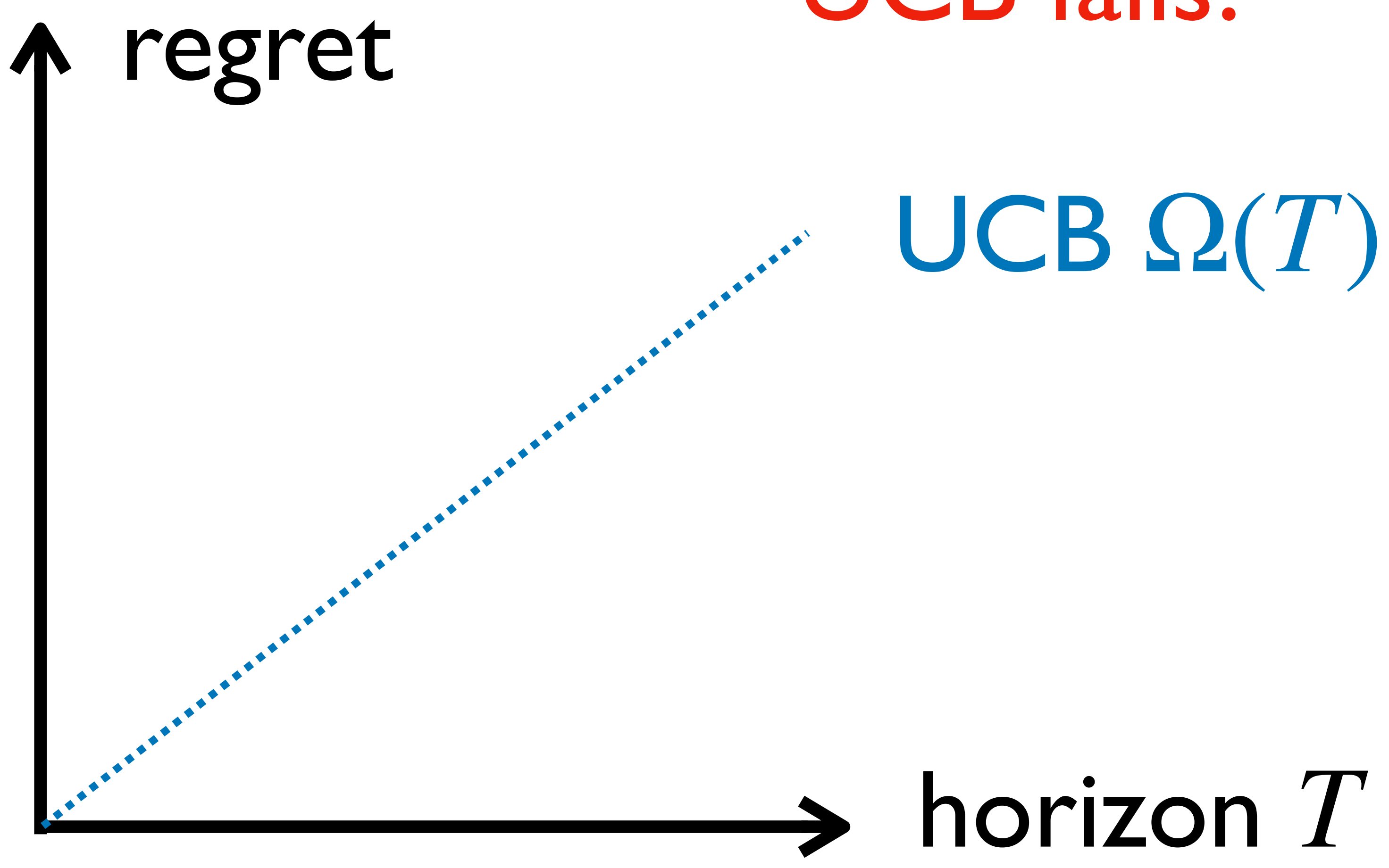
Stochastic bandits: the losses for each step follow the distribution $(P_i)_{i=1,\ldots,K}$ on $[0,1]$

Adversary selects losses $\ell_1, \ldots, \ell_T \subset [0,1]^K$

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

1. Player selects arm $I(t) \in \{1,\ldots,K\}$
2. Observe loss $\ell_{t,I(t)} \in [0,1]$ (adversarial)
   Observe stochastic loss of $I(t)$ from $P_{I(t)}$ (stochastic)

UCB fails!



- Goal: minimize the pseudo-regret

$$\mathrm{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I(t)}\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_{t,i}$$

Algorithm for Adversarial Regime | Exp3

- Exp3 (Exponential-weight algorithm for exploration & exploitation) algorithm [Auer+ 2002]

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

1. Draw arm $I(t) \in \{1,\ldots,K\}$ from distribution $(p_i(t))_i$,

$$p_i(t) \propto \exp\left(\sum_{s=1}^{t-1} \hat{\ell}_{s,i}\right)$$

sum of estimated losses so far

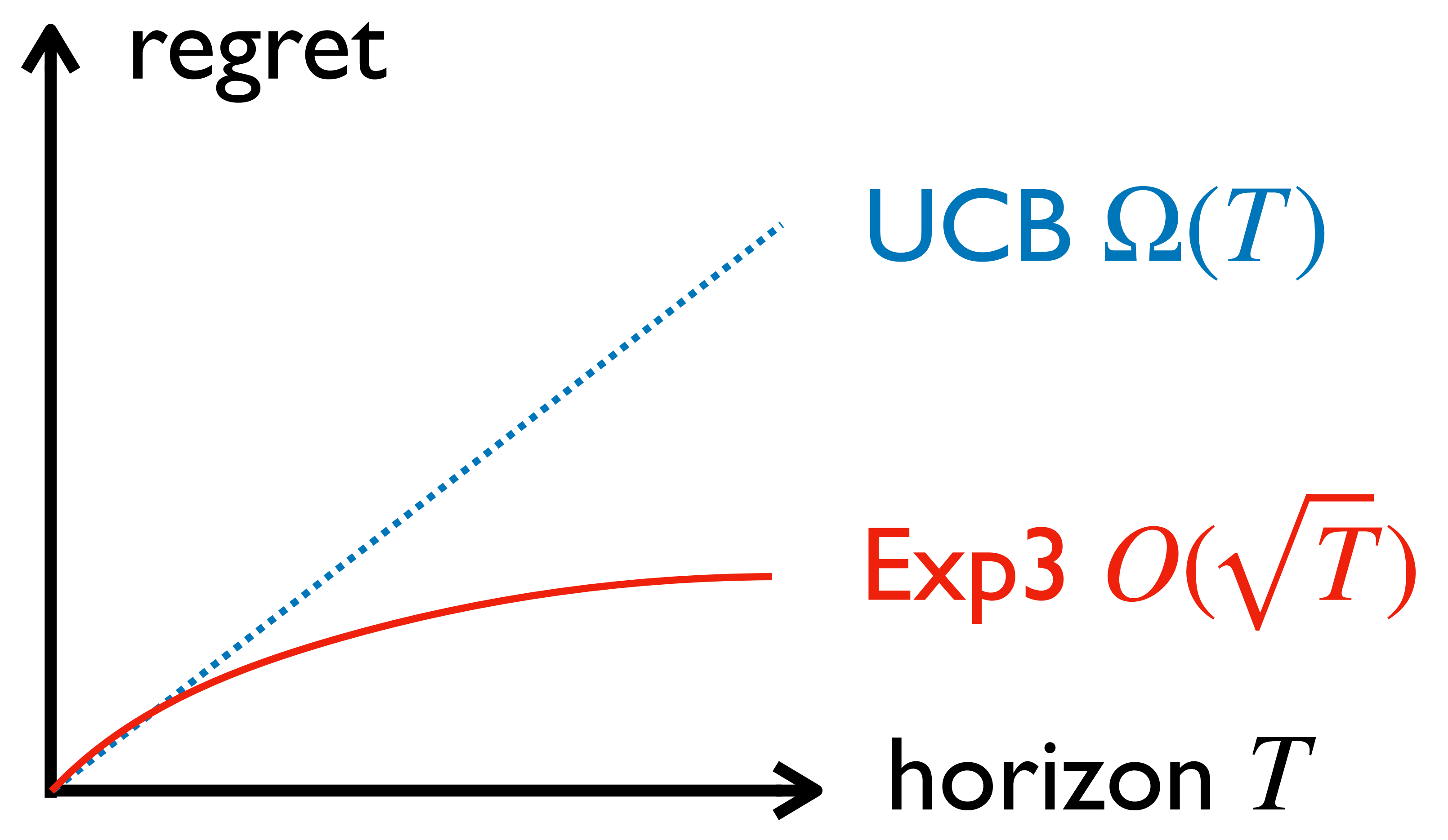
2. Observe loss $\ell_{t,I(t)} \in [0,1]$

3. Estimate the loss for each arm

$$\hat{\ell}_{t,i} = \begin{cases} \frac{\ell_{t,i}}{p_i(t)} & \text{if } i = I(t) \\ 0 & \text{otherwise} \end{cases}$$

Importance-weighted estimator

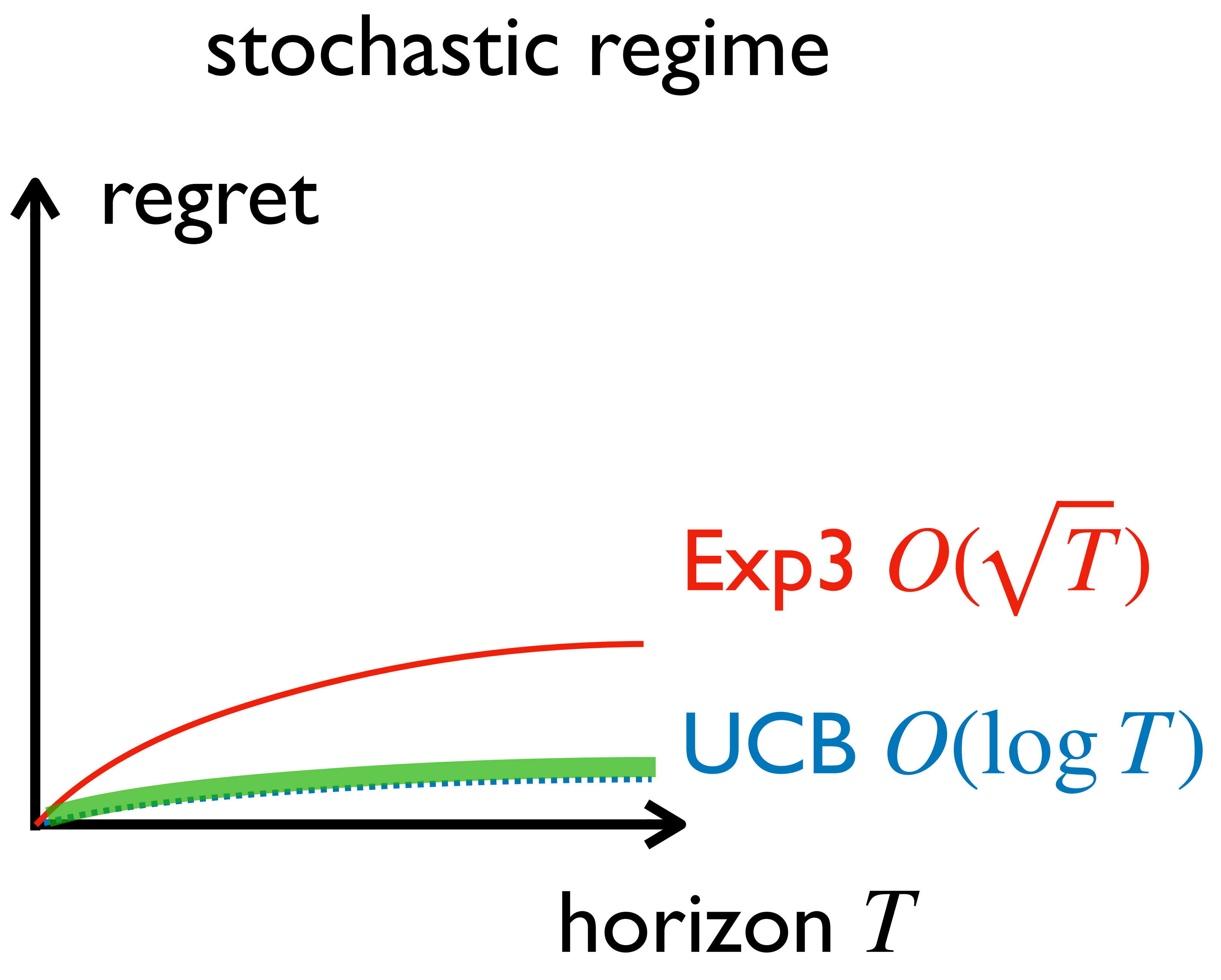
→ Unbiased estimator of ℓ_t : $\mathbb{E}_{I_t \sim p_i(t)}[\hat{\ell}_t] = \ell_t$

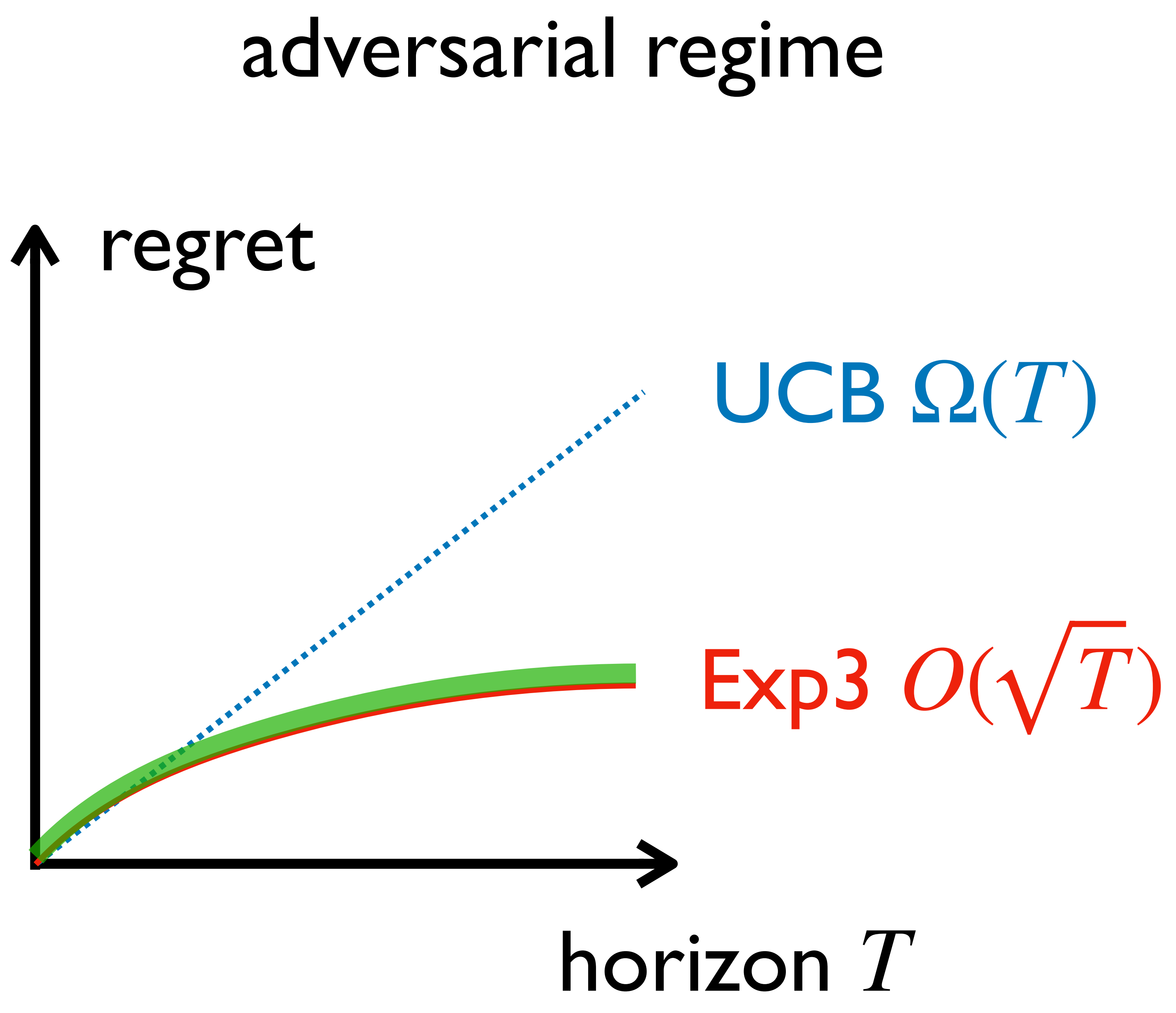


P. Auer, N. Cesa-Bianchi, Y. Freund, & R. E. Schapire. The non-stochastic multi-armed bandit problem. SIAM Journal on Computing, 2002.

All You Need is Exp3?

- No. Exp3 is not optimal for stochastic regimes





Q. Is it possible to achieve the optimality in both regimes
without knowing the underlying regime (= **best-of-both-worlds**)?

Outline | Variance-Dependent BOBW Algorithm

- Introduction to best-of-both-worlds algorithms
- Super quick introduction on vanilla multi-armed bandits
- Background (existing approaches & intermediate regime)
- Tsallis-INF algorithm
- Our work
  - Regret bounds on three regimes
  - Preliminary (Optimistic FTRL)
  - Proposed algorithm
  - Numerical experiments
- Conclusion

Existing Best-of-Both-Worlds Algorithms

Def. Optimal(*) in stochastic regime $\Leftrightarrow$ $\mathrm{Reg}_T = O(\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$

Optimal in adversarial regime $\Leftrightarrow$ $\mathrm{Reg}_T = O(\sqrt{KT})$

- Assume stochastic environment and check if the this assumption is satisfied
  - If it is determined that not satisfied, move to an algorithm for adversarial regime

[Bubeck & Slivkins 2012, Auer & Chiang 2016]

stochastic: optimal, adversarial: near-optimal

- Assume adversarial environment, and adopt to the stochastic environment if it's easy

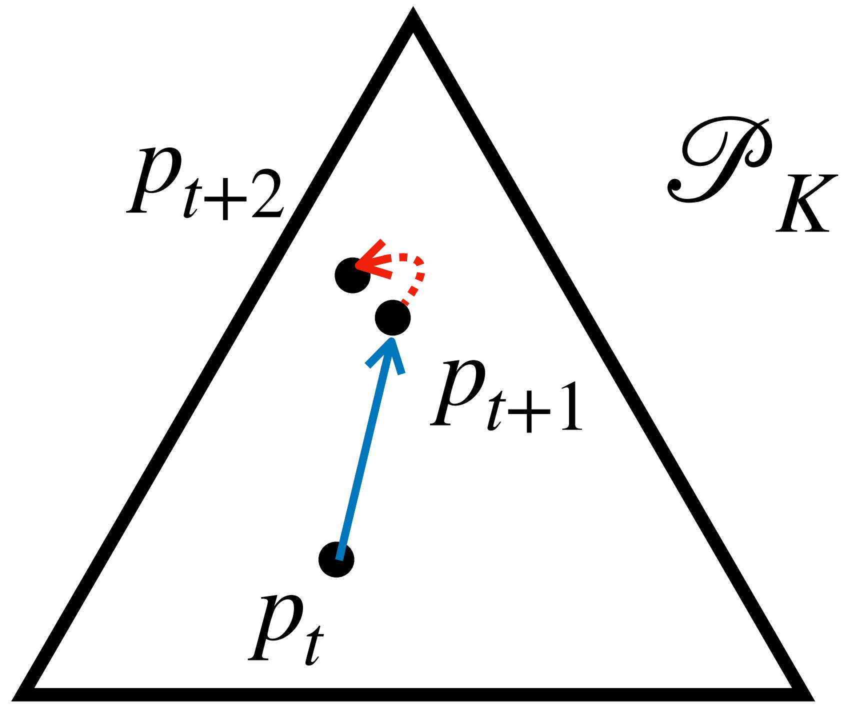
[Seldin & Slivkins 2014, Seldin & Lugosi 2017]

stochastic: near-optimal $O(\mathrm{polylog} T)$, adversarial: optimal

S´ebastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits.COLT, 2012.
Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. COLT, 2016.
Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. ICML, 2014.
Yevgeny Seldin and G´abor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. COLT, 2017.

FTRL with 1/2-Tsallis Entropy achieves Both Optimality!

[Zimmer & Seldin, 2021]



- Follow-the-Regularized-Leader (FTRL)
  - Select arm $I(t)$ based on distribution $p_t \in \mathscr{P}_K$ defined by:

sum of estimated losses   convex regularization function

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

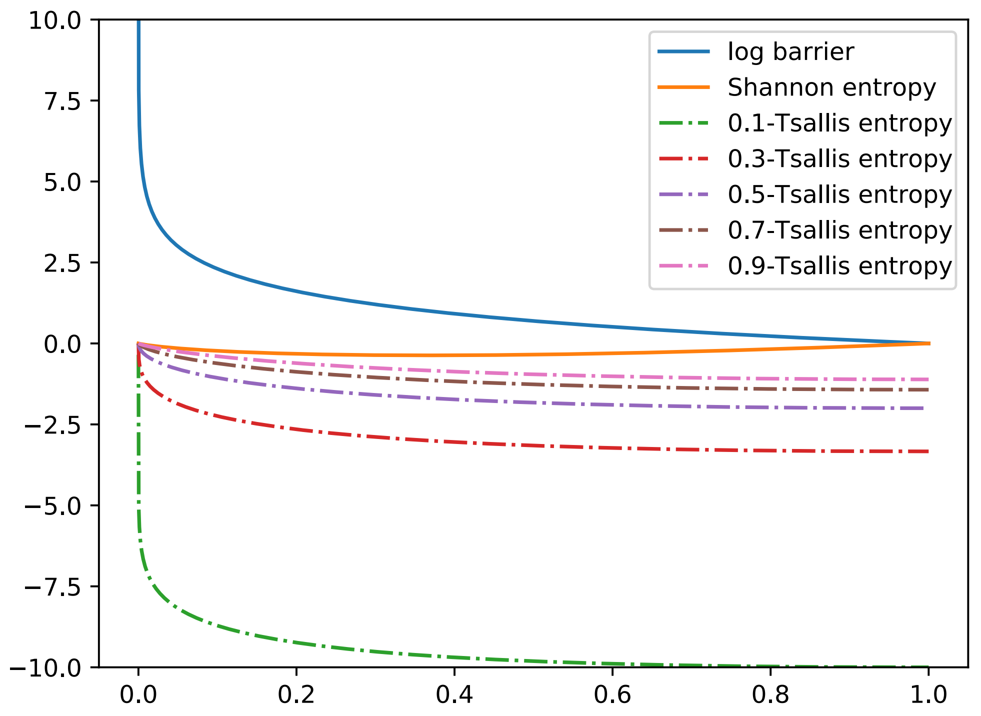
$\hat{\ell}_s \in \mathbb{R}^K$: unbiased estimator of ℓ_s

Convex regularizer determines the behavior of arm selection probability p_t especially around the border

- FTRL with 1/2-Tsallis entropy (with a certain learning rate) achieves the BOBW

α-Tsallis entropy $\xrightarrow{\alpha \to 0}$ log-barrier potential $\sum_{i=1}^{K} -\log(w_i)$

α-Tsallis entropy $\xrightarrow{\alpha \to 1}$ negative Shannon entropy $\sum_{i=1}^{K} w_i \log(w_i)$



J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 22(28):1–49, 2021.

Intermediate between Stochastic and Adversarial

- Adversarial bandits: Too pessimistic → Any practical regimes?
- Intermediate regime between stochastic & adversarial
  - ▸ Stochastically constrained adversarial regime: [Wei & Luo 2018]
    losses are drawn from distribution w/ fixed gaps & losses are allowed to change

$$\mathbb{E}[\ell_{t,i} - \ell_{t,j}] = \tilde{\Delta}_{i,j}$$

weather affects uniformly the will to buy

  - ▸ Stochastic regime with adversarial corruptions [Lykouris & Mirrokni 2018]

Losses generated from stochastic regime (*) $\bar{L}_T = (\bar{\ell}_1, \ldots, \bar{\ell}_T) \xrightarrow{\text{adversarial noise}} L_T = (\ell_1, \ldots, \ell_T)$

not observed

Adversarial reviews or clicks

★☆☆☆☆

  - ▸ (Adversarial regime with a self-bounding constraint) [Zimmer & Seldin, 2021]
    - General regimes including above stochastic, above two, & adversarial regimes

(*) More general model can be considered

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. COLT, 2018.
Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions.STOC, 2018.
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 22(28):1–49, 2021.

Regret Upper Bounds of Tsallis-INF Algorithm

- Follow-the-Regularized-Leader w/ 1/2-Tsallis entropy achieves the Best-of-Both-Worlds!

stochastic regime $(C = 0)$ ←UCB

$\cap$

stochastically constrained adversarial regime $(C = 0)$

$\cap$

stochastic regime w/ adversarial corruptions $(C = \sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty)$

$\cap$

adversarial regime ←Exp3

$O(\sqrt{KT})$

$\subset$ adversarial regime w/ a self-bounding constraint

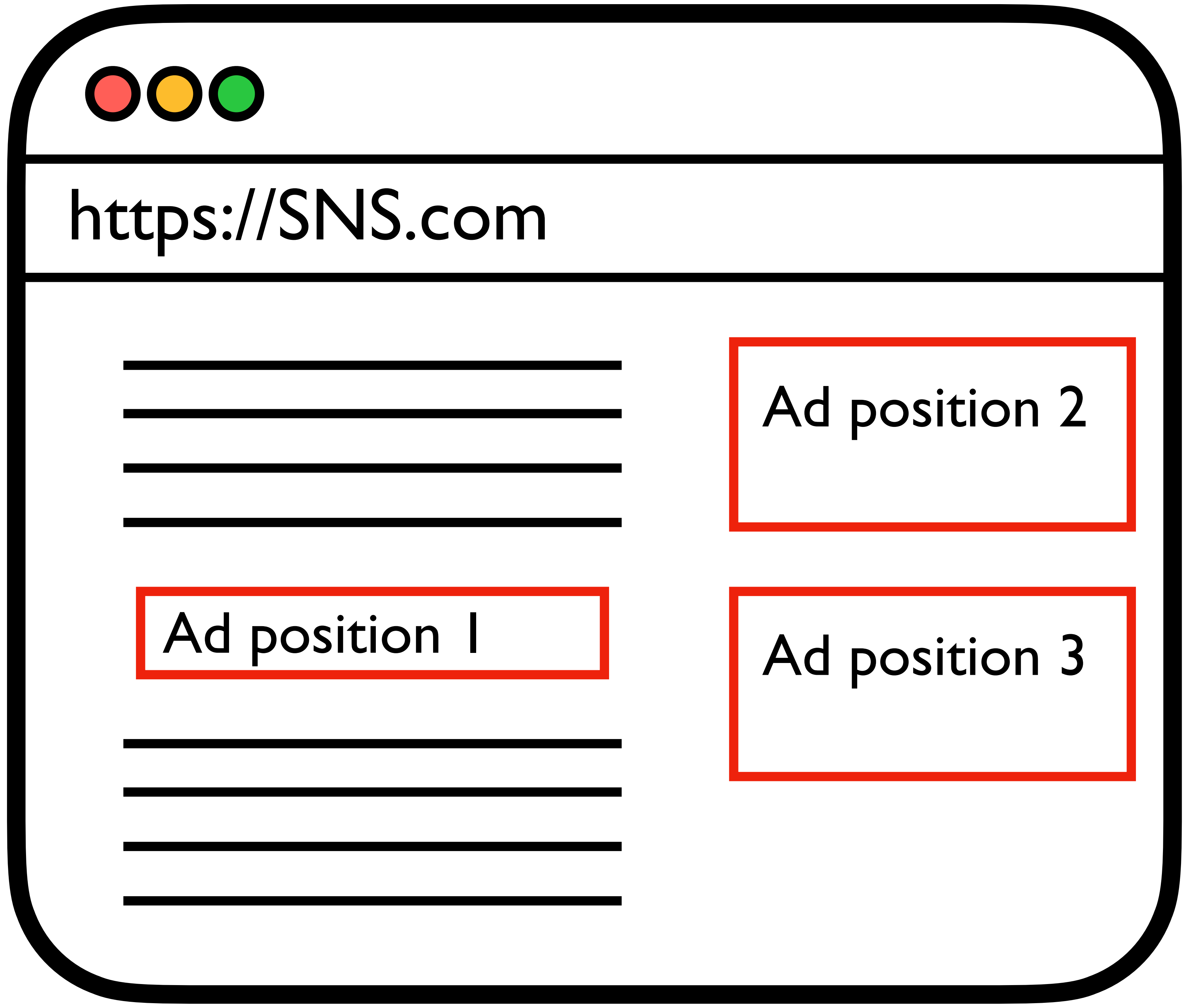
$$O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i} + C\right)$$

Q. Truly "optimal" bound for stochastic regime is e.g., $O\left(\sum_{i \neq i^*} \frac{\Delta_i}{d_{\inf}(i)} \log T\right)$.

Is it possible to use **distributional information** to obtain better bounds?

When is Distributional Information Useful?

- Maximizing the Click-Through Rate (CTR) using multi-armed bandits



- Running a website and try to let the users click ads
- Want to maximize the CTR

- CTR is around 1.0 to 10.0 % (Can be much more smaller!)
- Then, the variance of arm i is $\sigma_i^2 \simeq 0.01 \sim 0.1$
- If we could obtain the regret depending on σ_i^2, we would reduce the regret to 1% ~ 10%!

Outline | Variance-Dependent BOBW Algorithm

- Introduction to best-of-both-worlds algorithms
- Super quick introduction on vanilla multi-armed bandits
- Background (existing approaches & intermediate regime)
- Tsallis-INF algorithm
- Our work
  - Regret bounds on three regimes
  - Preliminary (Optimistic FTRL)
  - Proposed algorithm
  - Numerical experiments
- Conclusion

Regret Bounds: Existing Studies

Variance-dependent bound

Best-of-both-worlds

Robust against corruption

Data-dependent bound

	Stochastic	Adversarial	Stochastic with adversarial corruptions
UCB-V [Audiber+ 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert+ 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O\left(\sqrt{KT}\right)$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$
LB-INF [Ito, 2021]	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty,V_1\}\log T}\right)$	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$

- cumulative loss for the optimal arm: $L^* = \min_{i\in[K]}\sum_{t=1}^{T}\ell_i(t)\in[0,T]$
- empirical variation of loss vectors: $Q_\infty = \min_{\bar{\ell}\in\mathbb{R}^K}\sum_{t=1}^{T}\|\ell(t)-\bar{\ell}\|_\infty^2\in[0,T/4]$
- path-length of loss vectors: $V_1 = \sum_{t=1}^{T-1}\|\ell(t)-\ell(t+1)\|_1\in[0,T]$

Regret Bounds: This Study

	Stochastic	Adversarial	Stochastic with adversarial corruptions
UCB-V [Audiber+ 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	NA	NA
Tsallis-INF [Zimmert+ 2021]	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T\right)$	$O\left(\sqrt{KT}\right)$	$O\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T}\right)$
LB-INF [Ito, 2021]	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty,V_1\}\log T}\right)$	$O\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T+\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$
LB-INF-V (This work)	$O\left(\sum_{i:i\neq i^*}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty\}\log T}\right)$	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T+\sqrt{\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T}\right)$

- ☑ The first BOBW algorithm w/ variance-dependent bounds
- ☑ Proposed algorithm is corruption-robust & data-dependent

Regret Bounds: The Leading Constant Factor is Small!

	Stochastic	Gap from lower bound	Adversarial
UCB-V [Audiber+ 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	$\simeq 5$	NA
Tsallis-INF [Zimmert+ 2021]	$\simeq \sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T$	$\simeq 2$	$O\left(\sqrt{KT}\right)$
LB-INF [Ito, 2021]	$\simeq 36\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T$	$\simeq 72$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty,V_1\}\log T}\right)$
LB-INF-V (This work)	$\simeq \sum_{i:i\neq i^*}\max\left\{4\frac{\sigma_i^2}{\Delta_i},2\right\}\log T$	$\simeq 2$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty\}\log T}\right)$

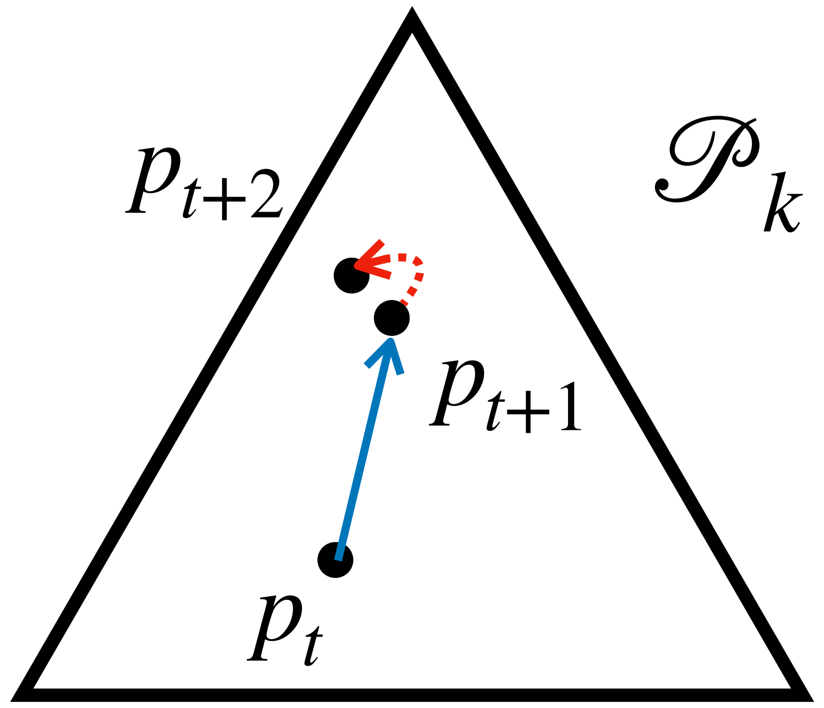
The leading constant of the regret upper bound is close to the lower bound (gap $\simeq 2$)

Regret Bounds: LB-INF-V with Path-Length Bound

	Stochastic	Gap from lower bound	Adversarial
UCB-V [Audiber+ 2009]	$O\left(\sum_{i:\Delta_i>0}\left(\frac{\sigma_i^2}{\Delta_i}+1\right)\log T\right)$	$\simeq 5$	NA
Tsallis-INF [Zimmert+ 2021]	$\simeq \sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log T$	$\simeq 2$	$O\left(\sqrt{KT}\right)$
LB-INF [Ito, 2021]	$\simeq 36\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T$	$\simeq 72$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty,V_1\}\log T}\right)$
LB-INF-V (This work)	$\simeq \sum_{i:i\neq i^*}\max\left\{4\frac{\sigma_i^2}{\Delta_i},2\right\}\log T$	$\simeq 2$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty\}\log T}\right)$
LB-INF-V' (This work)	$\simeq \sum_{i:i\neq i^*}\max\left\{8\frac{\sigma_i^2}{\Delta_i},4\right\}\log T$	$\simeq 4$	$O\left(\sqrt{K\min\{T,L^*,Q_\infty,V_1\}\log T}\right)$

- Modifications to the algorithm yield a path-length regret bound in exchange for a larger constant

Optimistic FTRL [Rakhlin & Sridharan, 2013]



- Follow-the-Regularized-Leader (FTRL)
  - Select arm $I(t)$ based on distribution $p_i(t) \in \mathscr{P}_k$ defined by:

sum of estimated rewards — convex regularization function

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

$\hat{\ell}_s \in \mathbb{R}^K$: unbiased estimator of ℓ_s

- Optimistic FTRL: optimistic prediction of $\ell(t)$ + FTRL
  - The arm selection probability is replaced with

$m(t) \in [0,1]^K$: optimistic prediction of $\ell(t)$

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

Useful when deriving data-dependent regret bound!

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In Conference on Learning Theory, pages 993–1019, 2013a.
Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In Advances in Neural Information Processing Systems, pages 3066–3074, 2013b.

Proposed Algorithm: LB-INF-V

- Optimistic FTRL: Reduce variances in unbiased estimator for loss vectors
  - Arm selection probability is replace with

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

convex regularization function

$m(t) \in [0,1]^K$: optimistic prediction of $\ell(t)$

  - Optimistic prediction $m(t) \in \mathbb{R}^K$: empirical mean of observed data

$$m_i(t) = \frac{\frac{1}{2} + \sum_{s=1}^{t-1} 1[I(s) = i]\, \ell_i(s)}{1 + \sum_{s=1}^{t-1} 1[I(s) = i]}$$

converges to μ_i

  - Unbiased estimator $\hat{\ell}_t \in \mathbb{R}^K$:

$$\hat{\ell}_i(t) = m_i(t) + \frac{1[I(t) = i]}{p_i(t)} (\ell_i(t) - m_i(t))$$

Reduce variances using $m_i(t)$

Proposed Algorithm: LB-INF-V

- Optimistic FTRL: Reduce variances in unbiased estimator for loss vectors
  - Arm selection probability is replace with

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

convex regularization function

$m(t) \in [0,1]^K$: optimistic prediction of $\ell(t)$

  - Regularization function is $\psi_t(p) = \sum_{i=1}^{K} \beta_i(t)\phi(p_i)$ with
    - $\phi(x) = x - 1 - \log(x) + \log(T) \cdot (x + (1-x)\log(1-x))$

Log-barrier regularization used in BROAD, LB-INF [Wei & Luo, 2018, Ito, 2021]

Entropy regularization function for $(1-x)$ used to handle the impact of the variance of the optimal arm

    - $\beta_i(t)$: adaptively chosen based on squared prediction error $(m_{I(t)}(t) - \ell_{I(t)}(t))^2$ of $m_{I(t)}(t)$

$\to \sigma_{I(s)}^2$ as $s \to \infty$

C.-Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In Conference on pages 1263–1291, 2018.
S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In Conference on Learning Theory, pages 2552–2583. PMLR, 2021b.

Regret Analysis: Stochastic Regime

- Definition of the regularization function ψ_t, and standard technique of OFTRL yields:

Lem. 1 For sufficiently large T, $R(T) \simeq O\left(\sum_{i \neq i^*} \sqrt{\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - m_i(t))^2 \log(T)} \right)$

- Definition of $m(t)$ yields:

Lem. 2 $\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - m_i(t))^2 \right] = O\left(\sigma_i^2\, \mathbb{E}[\sum_{t=1}^{T} p_i(t)] + \log(T) \right)$

- Combining the above two lemmas and Jensen's inequality we have

Prop. 1 For sufficiently large T, $R(T) = O\left(\sum_{i \neq i^*} \sqrt{\sigma_i^2\, \mathbb{E}[\sum_{t=1}^{T} p_i(t)] \log(T)} + K \log(T) \right)$

Regret Analysis: Stochastic Regime

Prop. 1 For sufficiently large T, $R(T) = O\left(\sum_{i\neq i^*} \sqrt{\sigma_i^2\, \mathbb{E}[\sum_{t=1}^T p_i(t)]\, \log(T)} + K\log(T) \right)$

- With a **self-bounding technique**, we have
  - From the definition of the regret, we have $\boxed{R(T) = \sum_{i\neq i^*} \Delta_i\, \mathbb{E}\left[\sum_{t=1}^T p_i(t)\right]}$
  - From Prop 1 & AM-GM, $$R(T) \le \sum_{i\neq i^*} \left(\frac{1}{2}\Delta_i \mathbb{E}\left[\sum_{t=1}^T p_i(t)\right] + O\left(\frac{\sigma_i^2}{\Delta_i}\log(T)\right)\right) + O(K\log(T))$$
  - From the above two bounds, we have

$$R(T) = 2R(T) - R(T)$$

$$= \sum_{i\neq i^*}\left(\Delta_i \mathbb{E}\left[\sum_{t=1}^T p_i(t)\right] + O\left(\frac{\sigma_i^2}{\Delta_i}\log(T)\right)\right) + O(K\log(T)) - \sum_{i\neq i^*}\Delta_i\,\mathbb{E}\left[\sum_{t=1}^T p_i(t)\right] = \boxed{O\left(\sum_{i\neq i^*}\left(\frac{\sigma_i^2}{\Delta_1} + 1\right)\log(T)\right)}$$

Regre Analysis: Adversarial & Data-Dependent Regime

- Definition of the regularization function ψ_t, and standard technique of OFTRL yields:

Lem. 1 For sufficiently large T, $R(T) \simeq O\left(\sum_{i \neq i^*} \sqrt{\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - m_i(t))^2 \log(T)}\right)$

- Definition of $m(t)$ yields

Lem. 3 It holds for any $\ell^* \in [0,1]^K$ that

$$\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - m_i(t))^2\right] = \mathbb{E}\left[\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - \ell_i^*)^2\right] + O(K \log(T))$$

Consequently,

$$\mathbb{E}\left[\sum_{t=1}^{T} 1[I(t) = i](\ell_i(t) - m_i(t))^2\right] = \min\{Q_\infty, L^* + R(T), T - L^* - R(T)\} + O(K \log(T))$$

- Combining the above two lemmas, $R(T) = O\left(\sqrt{K \min\{Q_\infty, L^*, T - L^*\} \log(T)} + K \log(T)\right)$

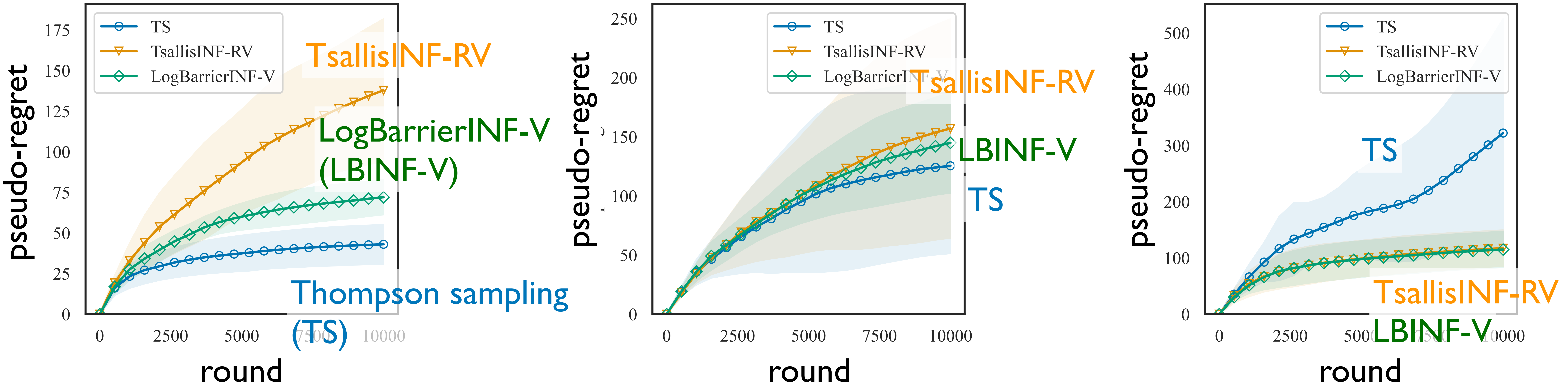
Numerical Comparison with Thompson Sampling (TS) & Tsallis-INF w/ RV-estimator

- Setting: Bernoulli distribution with $K = 5$

Experiment 1.
- Stochastic regime
- $\mu = (0.5, 0.9, \ldots, 0.9)$ $\to$ small σ_i^2

Experiment 2.
- Stochastic regime
- $\mu = (0.5, 0.55, \ldots, 0.55)$ $\to$ large σ_i^2

Experiment 3.
- Stochastically constrained adversarial regime
- $\Delta = 0.1$(same as Figure 3 in [Zimmert & Seldin 2021])



Conclusion & Future work

- OFTRL with adaptive learning rate achieves

stochastic regime $(C = 0)$

$\cap$

stochastic regime w/ adversarial corruptions $(C = \sum_{t=1}^{T} \|\bar{\ell}_t - \ell_t\|_\infty)$

$\cap$

adversarial regime

$O\left(\sqrt{K \min\{T, L^*, Q_\infty\} \log T}\right)$

$\boxed{L^*, Q_\infty : \text{ data-dependent measures}}$

$\subset$

adversarial regime w/ a self-bounding constraint

$\boxed{\sigma_i^2 : \text{ variance of arm } i}$

$$O\left(\sum_{i \neq i^*} (\frac{\sigma_i^2}{\Delta_i} + 1) \log T + \sqrt{C \sum_{i \neq i^*} (\frac{\sigma_i^2}{\Delta_i} + 1) \log T}\right)$$

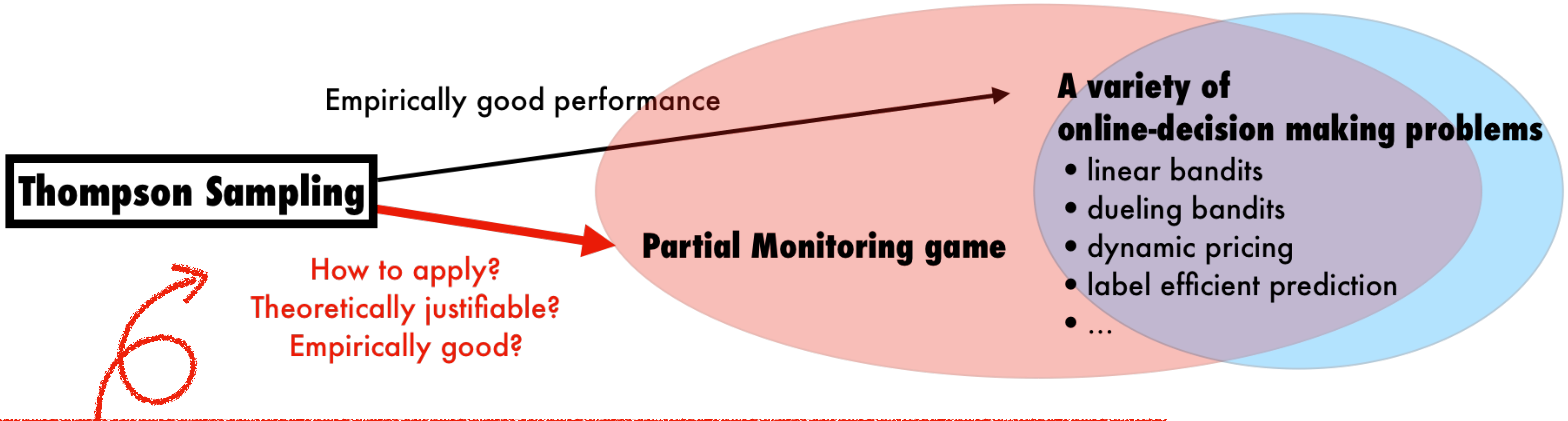
The leading constant of the regret upper bound is close to the lower bound (gap $\simeq 2$)

- Future work
  - Can we achieve a gap < 2 while preserving BOBW and/or corruption robustness?
  - Can we remove the assumption that the optimal arm is unique?

Summary of Today's Talk

1. Thompson sampling for stochastic partial monitoring (NeurIPS2020)
   - (possibly) practical since the algorithm can handle many problems and is empirically good
   - available at https://arxiv.org/abs/2006.09668

2. A best-of-both-worlds algorithm with variance-dependent regret bounds (COLT2022)
   - (possibly) practical since the algorithm can handle adversarial corruption and have "state-of-the-art" performance
   - available at https://arxiv.org/abs/2206.06810



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

Thank you for listening!