Best-of-Both-Worlds Algorithms for Partial Monitoring

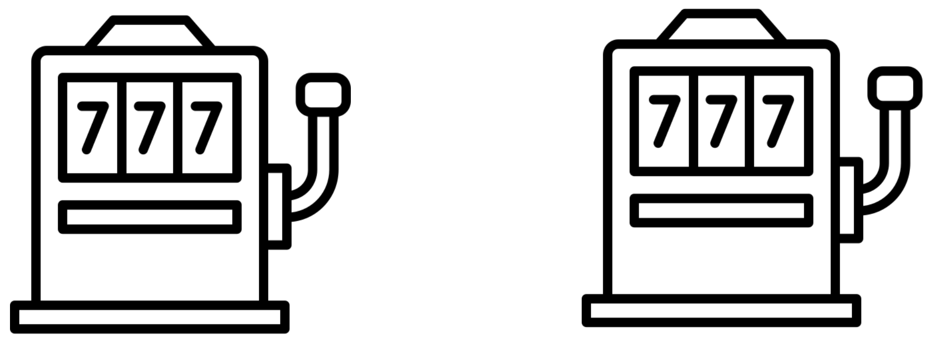
Taira Tsuchiya [1,2], Shinji Ito [3], Junya Honda [1,2]

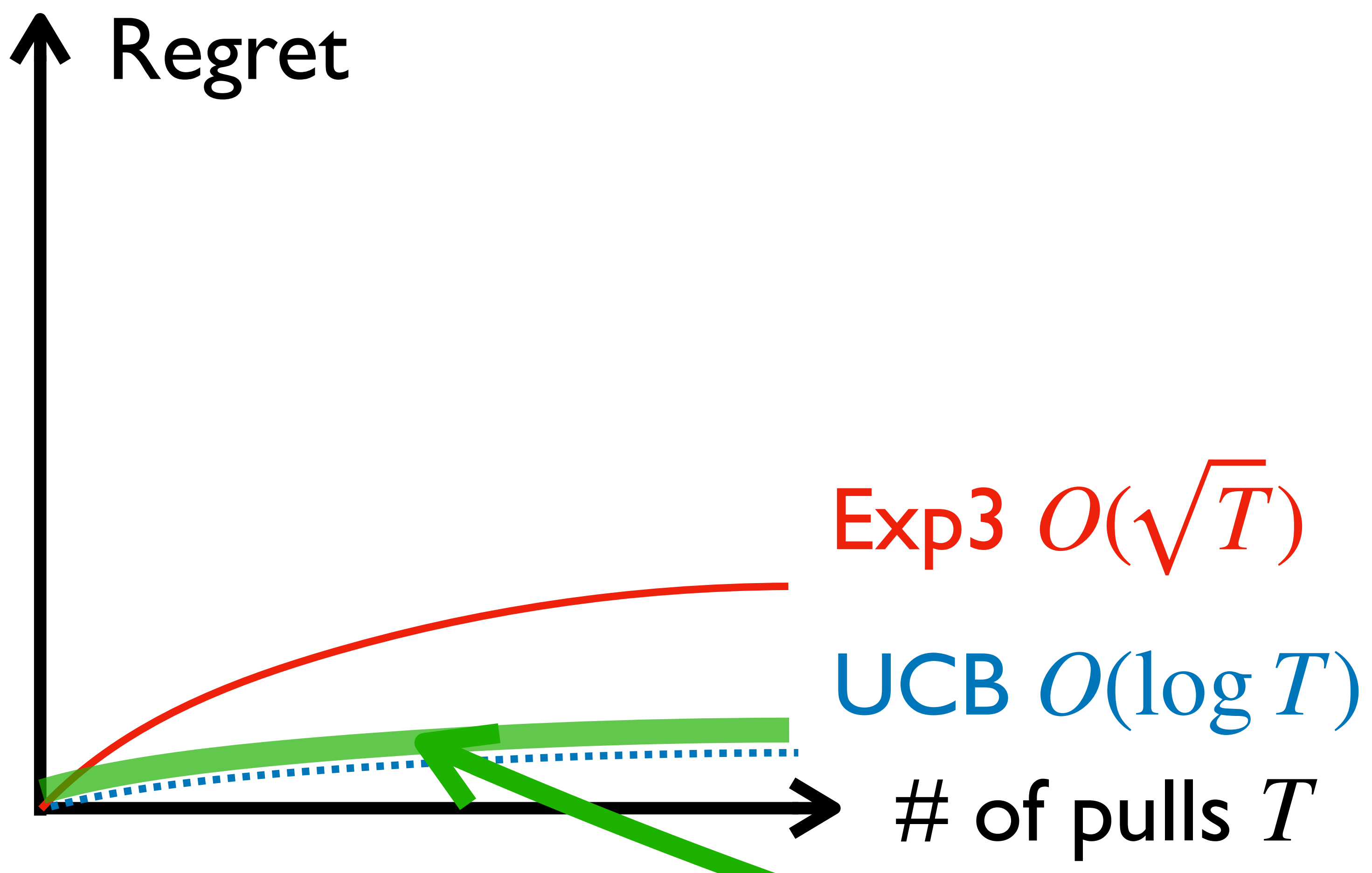
1. Kyoto University, 2. RIKEN AIP, 3. NEC Corporation

Feb 22, 2023, ALT2023, National University of Singapore

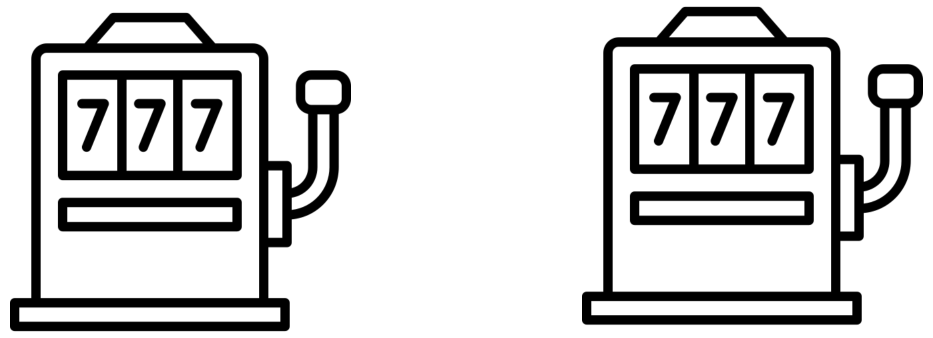
Introduction | Best-of-Both-Worlds in Bandits [Bubeck and Slivkins 2012]

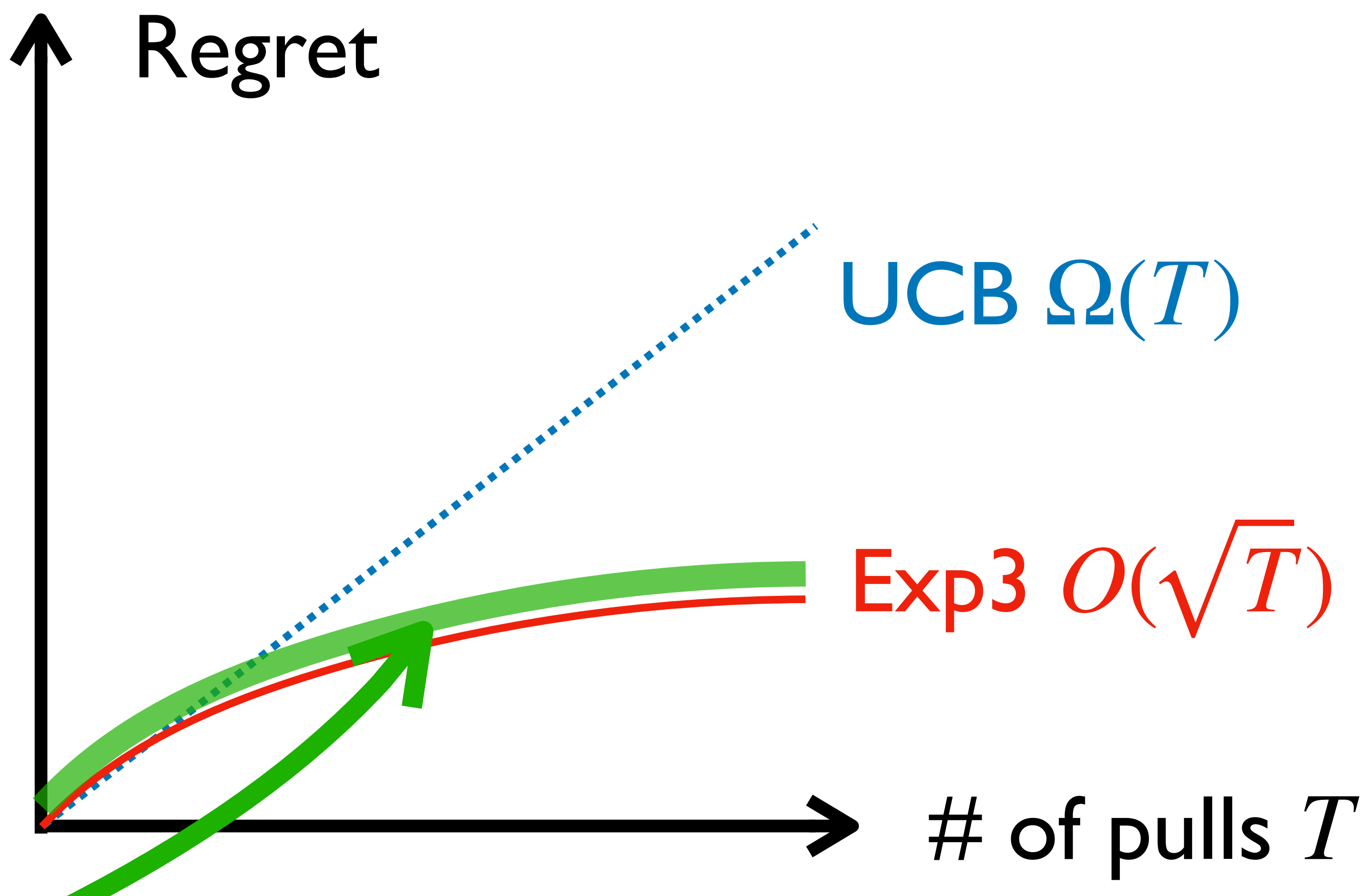
Stochastic regime





Adversarial regime





What We Hope :Achieving optimality for both stochastic and adversarial regimes w/o knowing the underlying regime = **Best-of-Both-Worlds (BOBW)**

Q. BOBW in more complex settings?

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In COLT 2012.

Research Question

Best-of-Both-Worlds is possible
in (relatively) **simple** settings

Many online decision-making problems

- full information
- multi-armed bandits
- online learning w/ feedback graphs
- dueling bandits
- dynamic pricing
- label efficient prediction
- ….

Special cases

Partial monitoring

Rewards are not directly observed

Q. Can we achieve **best-of-both-worlds** in **partial monitoring**?

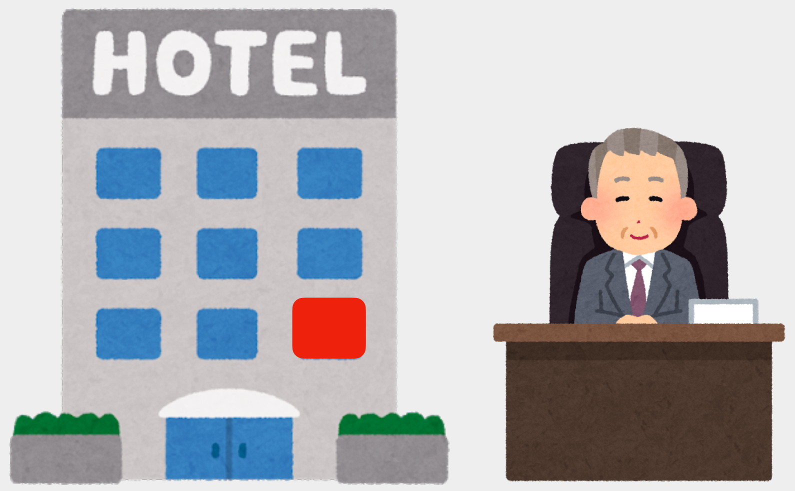
Outline

- Introduction: research question
- Preliminary: partial monitoring
- BOBW algorithm for locally observable games
- BOBW algorithm for globally observable games
- Summary

Partial Monitoring Example: Dynamic Pricing

Learner (= seller)

Hotel owner



decides **accommodation fee** of room from $\{\$1, \ldots, \$k\}$

$\underline{t = 1}$

accomm. fee $40

$\underline{t = 2}$

accomm. fee $80

$\underline{t = \cdots}$

Adversary

User's **evaluation price**

Use if accomm. fee $\leq \$90$

Use if accomm. fee $\leq \$50$

opportunity loss

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

$\$c$ (const.)

$(\because \$50 - \$80 < \$0)$

feedback

Buy

No-buy

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

Q. Possible to minimize the total loss only with limited feedbacks?

Formulation of Partial Monitoring

- Consider partial monitoring game $\mathbf{G} = (L, \Phi)$ with k-action and d-outcomes
- Loss matrix $L = (L_{ax}) \in [0,1]^{k\times d}$, feedback matrix $\Phi \in \Sigma^{k\times d}$ (Σ : set of feedback symbols)

observed to the player

Adversary selects outcomes $x_1, \ldots, x_T \in \{1,\ldots,d\}$
At each round $t = 1,\ldots,T$:
1. **Learner** selects action $A_t \in \{1,\ldots,k\}$
2. Learner incurs loss $L_{A_t x_t}$ and observes feedback $\Phi_{A_t x_t}$

- Goal: minimize regret R_T

$$R_T = \mathbb{E}\Big[\underbrace{\textstyle\sum_{t=1}^T L_{A_t x_t}}_{\text{cumulative losses of taken actions}} - \underbrace{\textstyle\sum_{t=1}^T L_{a^* x_t}}_{\text{cumulative losses of optimal action}}\Big], \quad a^* = \arg\min_{a\in[k]} \mathbb{E}\Big[\textstyle\sum_{t=1}^T L_{a x_t}\Big]$$

Example 1. Dynamic Pricing [Kleinberg & Leighton 2003]

k : discrete range of selling price
d : discrete range of evaluation price

(row: selling price, column: evaluation price)

selecting action
= determining the selling price

outcome
= evaluation price

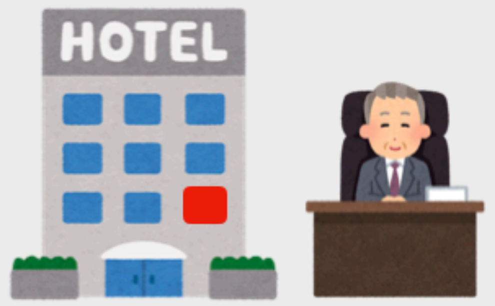
$\Sigma = \{\text{Buy}(\bigcirc), \text{No-Buy}(\times)\}$

Learner (= seller)

$t = 1$

Hotel owner

accomm. fee \$40



$t = 2$

decides accommodation fee of room from $\{\$1, \ldots, \$k\}$

accomm. fee \$80

$t = \cdots$

Adversary

User's evaluation price

Use if accomm. fee $\leq$ \$90

Use if accomm. fee $\leq$ \$50

opportunity loss

\$90 – \$40 = \$50

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

feedback

Buy

No-buy

loss matrix

$$L_{ax} = \begin{cases} x - a & \text{if } x \geq a \\ c & \text{otherwise} \end{cases}$$

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

$x \geq a$

$x < a$

feedback matrix

$$\Phi_{ax} = \begin{cases} \bigcirc & \text{if } x \geq a \\ \times & \text{otherwise} \end{cases}$$

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

$x \geq a$

$x < a$

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

Example 2. Apple Tasing, Matching Pennies

- Learner predicts label (positive or negative) of items in an online manner
- Three possible actions when labeling items:
  1. Label as positive (P)
  2. Label as negative (N)
  3. Ask a expert (A true label is revealed to the learner.)

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

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

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

Classification of Partial Monitoring Games

[Bartók, Pál & Szepesvári 2010, 2011]
[Lattimore & Szepesvári 2019]

- PM games fall into four classes based on their minimax regret $R_T(\mathbf{G}) = \inf_\pi \max_{x_1,\ldots,x_T} R_T(\pi, (x_t)_t, \mathbf{G})$

PM games

- Trivial: $R_T(G) = 0$
- Easy: $R_T(G) = \Theta(\sqrt{T})$
- Hard: $R_T(G) = \Theta(T^{2/3})$
- Hopeless: $R_T(G) = \Omega(T)$

Trivial and Easy: locally observable

Trivial, Easy and Hard: globally observable

(informal, stochastic)

Locally observable: Taking pair of actions $a, b \in [k]$ is enough to know $(L_a - L_b)^\top \nu^*$

Globally observable: Taking pair of actions $a, b \in [k]$ can be NOT enough to know $(L_a - L_b)^\top \nu^*$ but taking all actions is enough

G. Bartók, D. Pál, and Cs. Szepesvári. Toward a classification of finite partial-monitoring games. In ALT 2010.
G. Bartók, D. Pál, and Cs. Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In COLT 2011.
T. Lattimore and Cs. Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In ALT 2019.

Three Regimes in Partial Monitoring

- Stochastic regime: $x_t \overset{\text{i.i.d.}}{\sim} \nu^* \in \mathscr{P}_d$ (dist. over outcomes)
- Adversarial regime: x_t arbitrarily decided
- Stochastic regime w/ adversarial corruptions (for PM)

  (A MAB version was considered [Lykouris, Mirrokni & Leme 2018])

Adversarial regime: Pessimistic

Intermediate regime …?

Stochastic regime: Optimistic

<u>Outcomes sampled in i.i.d. manner</u>

$x'_1, \ldots, x'_T \sim \nu^*$

adversarial noise
at most C

$C = \mathbb{E}\left[\sum_{t=1}^{T} \|Le_{x_t} - Le_{x'_t}\|_\infty\right]$

<u>Outcomes with noise</u>

$x_1, \ldots, x_T$

$C = 0 \rightarrow$ stochastic regime
$C = T \rightarrow$ adversarial regime

Q. Can we achieve “best” in all regimes?

T. Lykouris, V. Mirrokni, and R.P. Leme. Stochastic bandits robust to adversarial corruptions. In STOC 2018.

Our Regret Bounds: Comparison with Existing Bounds

- Locally observable games

Corruption level: $C = \mathbb{E}\left[\sum_{t=1}^{T} \|Le_{x_t} - Le_{x_t'}\|_\infty\right], \; x_t' \sim \nu^*$

	Stochastic	Adversarial	Stochastic w/ Corruptions
[Tsuchiya+ 2020]	$O(\log T)$	NA	NA
[Lattimore+ 2020]	NA	$O(\sqrt{T})$	NA
Proposed	$O((\log T)^2)$	$O(\sqrt{T}\log T)$	$O((\log T)^2 + \sqrt{C}\log T)$

- Globally observable games

	Stochastic	Adversarial	Stochastic w/ Corruptions
[Lattimore+ 2020]	NA	$O(T^{2/3})$	NA
Proposed	$O((\log T)^2)$	$O((T\log T)^{2/3})$	$O((\log T)^2 + (C\log T)^{2/3})$

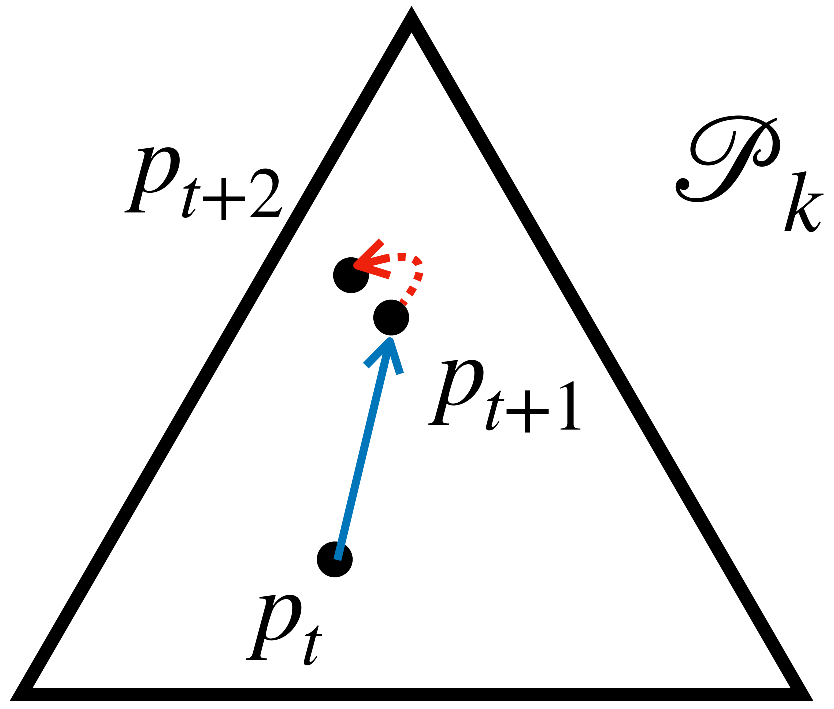
T. Tsuchiya, J. Honda, and M. Sugiyama. Analysis and design of Thompson sampling for stochastic partial monitoring. In NeurIPS 2020.
T. Lattimore and Cs. Szepesvári. Exploration by optimisation in partial monitoring. In COLT 2020.

Outline

- Introduction: research question
- Preliminary: partial monitoring
- BOBW algorithm for locally observable games
- BOBW algorithm for globally observable games
- Summary

Follow-the-Regularized-Leader in Bandits

- Follow-the-Regularized-Leader (FTRL):
  - One of the most common approaches for achieving BOBW [Wei & Luo 2018, Zimmert & Seldin, 2021, many!]
  - Determine action selection probability $p_t \in \mathscr{P}_k$ by minimizing "sum of estimated losses so far + convex regularizer":



sum of estimated losses convex regularization function

$$p_t \in \arg\min_{p \in \mathscr{P}_k} \underbrace{\langle \textstyle\sum_{s=1}^{t-1} \hat{y}_s, p \rangle}_{\text{sum of estimated losses}} + \underbrace{\psi_t(p)}_{\text{convex regularization function}}$$

$\hat{y}_s \in \mathbb{R}^k$: unbiased estimator of ℓ_s

- Common to transform the output of FTRL q_t to action selection probability $p_t \in \mathscr{P}_k$:

1. Compute $q_t \in \mathscr{P}_k$ by FTRL
2. Transform q_t to p_t : $p_t = \mathscr{T}_t(q_t)$

Important particularly in locally observable games

C.W. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In COLT 2018.
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. JMLR, 2021.

Exploration by Optimisation (ExpByOpt) [Lattimore & Szepesvári 2020]

- A technique to decide p_t from q_t and to favorably bound the *stability term* in PM
- We can bound the regret of FTRL w/ **negative Shannon entropy** $\psi_t(q) = -\eta_t^{-1} H(q)$ as

b-th dim of G: amount of information about action b when selecting action a and receive symbol Φ_{ax_t}

$$R_T \le \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathrm{penalty}(t) + \underbrace{(p_t - q_t)^\top L e_{x_t}}_{\text{transformation term}} + \underbrace{\frac{1}{\eta_t}\sum_{a=1}^{k} p_{ta}\Psi_{q_t}\left(\frac{\eta_t G(a,\Phi_{ax})}{p_a}\right)}_{\text{stability term } (\lesssim \text{ variance of loss estimator})}\right)\right] \qquad \Psi_q(z) = \langle q, \exp(-x) + x - 1\rangle$$

- ExpByOpt selects p_t from q_t by minimizing the sum of transformation and stability terms (for a worst-case outcome)

$$p_t = \mathcal{T}_t(q_t): \qquad \mathrm{opt}_q(\eta) := \mathrm{minimize}_{p\in\mathcal{P}_k} \max_{x\in[d]} \left[\frac{(p-q)^\top L e_x}{\eta} + \frac{1}{\eta^2}\sum_{a=1}^{k} p_a \Psi_q\left(\frac{\eta_t G(a,\Phi_{ax})}{p_a}\right)\right]$$

Theorem (informal). $\sup_{q\in\mathcal{P}_k} \mathrm{opt}_q(\eta) \le 3m^2k^3$ if $\eta \le 1/(2mk^2)$.

T. Lattimore and Cs. Szepesvári. Exploration by optimisation in partial monitoring. In COLT 2020.

Self-Bounding Technique : A technique to prove BOBW

[Zimmert & Seldin 2021]

- Use upper and *lower* bounds of regret depending on **FTRL outputs q_t**

Strategy. Suppose that using $Q = \mathbb{E}\left[\sum_{t=1}^{T}(1 - q_{ta^*})\right] \in [0,T]$ it holds that

$$R_T \lesssim \tilde{O}(\mathrm{polylog}(T)\sqrt{Q}) \quad \text{and} \quad R_T \geq \Delta_{\min} Q$$

<u>Adversarial regime:</u>

$$R_T \lesssim O(\mathrm{polylog}(T)\sqrt{Q}) \leq \tilde{O}(\sqrt{T})$$

<u>**Stochastic regime**</u>

$$R_T = 2R_T - R_T \lesssim \tilde{O}(\mathrm{polylog}(T)\sqrt{Q}) - \Delta_{\min} Q = O\left(\frac{\mathrm{polylog}(T)}{\Delta_{\min}}\right)$$

Require a "non-vacuous" lower bound

- ExpByOpt only considers the adversarial regime
  - Cannot derive a valid lower bound $R_T = \Omega(\Delta_{\min} Q)$ for applying the self-bounding technique (A naive use of EbO can lead to $p_a = 0$ and $q_a > 0$ for some $a \in [k]$)

J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. JMLR, 2021.

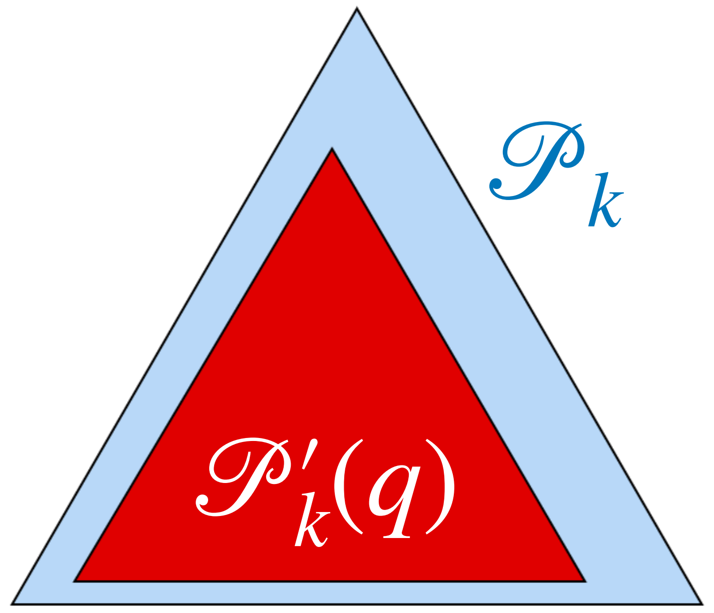
Solution: Restricting Feasible Set in Vanilla ExpByOpt

- Idea: Restrict a feasible set of the optimization problem to determine p_t from q_t

[Lattimore & Szepesvári 2020] $\mathrm{opt}_q(\eta) := \mathrm{minimize}_{p \in \mathscr{P}_k} \max_{x \in [d]} \left[\frac{(p-q)^\top L e_x}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \Psi_q \left(\frac{\eta_t G(a, \Phi_{ax})}{p_a} \right) \right]$

This work $\mathrm{opt}'_q(\eta) := \mathrm{minimize}_{p \in \mathscr{P}'_k(q)} \max_{x \in [d]} \left[\frac{(p-q)^\top L e_x}{\eta} + \frac{1}{\eta^2} \sum_{a=1}^{k} p_a \Psi_q \left(\frac{\eta_t G(a, \Phi_{ax})}{p_a} \right) \right]$

$\mathscr{P}'_k(q) = \{p \in \mathscr{P}_k : \underline{p_a \geq q_a/(2k)}$ for all $a \in [k]\} \subset \mathscr{P}_k$



This restriction leads to $R_T \geq \frac{1}{k} \Delta_{\min} Q$ and

Lemma (informal). $\sup_{q \in \mathscr{P}_k} \mathrm{opt}'_q(\eta) \leq 3m^2k^3$ if $\eta \leq 1/(2mk^2)$.

The component of regret is favorably bounded despite $\mathscr{P}'_k(q) \subset \mathscr{P}_k$.

Main Result for Locally Observable Games

- Combine the restricted EbO with the adaptive learning rate **with truncation**

$$\beta'_1 = c_1 \geq 1, \quad \beta'_{t+1} = \beta'_t + \frac{c_1}{\sqrt{1 + (\log k_\Pi)^{-1} \sum_{s=1}^{t} H(q_s)}}, \quad \beta_t = \max\left\{B, \beta'_t\right\}, \quad \text{and} \quad \eta_t = \frac{1}{\beta_t}$$

[Ito, Tsuchiya & Honda 2022]

Theorem. Consider non-degenerate locally observable games. Under some conditions,

Stochastic regime w/ adversarial corruptions

$$R_T = O\left(\frac{m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}} + \sqrt{\frac{C m^2 k^4 \log(T) \log(k_\Pi T)}{\Delta_{\min}}}\right)$$

Adversarial regime

$$R_T = O\left(m k^{3/2} \sqrt{T \log(T) \log k_\Pi}\right) + 2 m k^2 \log k_\Pi \qquad (k_\Pi \leq k)$$

- A first best-of-both-worlds algorithm for non-degenerate locally observable PM
- Adversarial: a factor of $\sqrt{\log T}$ worse than that by Lattimore & Szepesvári 2020

S. Ito, T. Tsuchiya, and J. Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In NeurIPS 2022.

Outline

- Introduction: research question
- Preliminary: partial monitoring
- BOBW algorithm for locally observable games
- BOBW algorithm for globally observable games
- Summary

Main Result for Globally Observable Games

- Shannon entropy regularizer + an adaptive learning rate leads to

Theorem. Consider globally observable games. Under some conditions,

Stochastic regime w/ adversarial corruptions

$$R_T = O\left(\frac{c_G^2 \log(T) \log(k_\Pi T)}{\Delta_{\min}^2} + \left(\frac{C^2 c_G^2 \log(T) \log(k_\Pi T)}{\Delta_{\min}^2} \right)^{1/3} \right)$$

Adversarial regime

$$R_T = O\left(\left(c_G^2 \log(T) \log(k_\Pi T) \right)^{1/3} T^{2/3} \right)$$

- Refining analysis replaces the hybrid regularizer with Shannon entropy

Tsallis entropy $T(q) = \sum_{a=1}^{k} \sqrt{q_i}$

[Zimmert, Luo & Wei 2019] $\psi_t(q) = -\eta_t^{-1}(T(q) + H(1-q))$

[Ito, Tsuchiya & Honda 2022] $\psi_t(q) = -\eta_t^{-1}(H(q) + H(1-q))$

Ours q_t becomes a closed-form

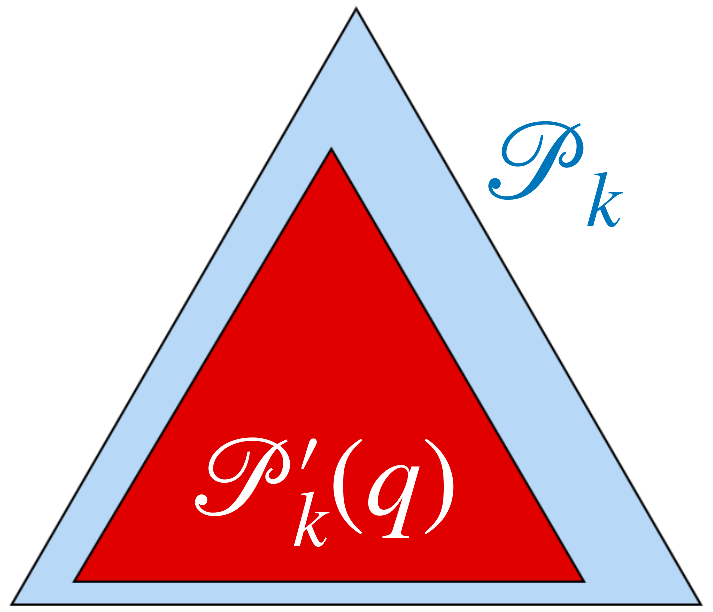
$$\psi_t(q) = -\frac{1}{\eta_t} H(q)$$

J. Zimmert, H. Luo, and C. Y. Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In AISTATS 2019.
S. Ito, T. Tsuchiya, and J. Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In NeurIPS 2022.

Summary | BOBW Algorithms for Partial Monitoring

Q. Is it possible to achieve BOBW in PM?
A. Yes, by FTRL, ExpByOpt, and adaptive learning rate!

<u>Locally observable games</u>
Extended exploration by optimisation for stochastic



$\sup_{q \in \mathcal{P}_k} \mathrm{opt}'_q(\eta) \leq 3m^2k^3$

↑ (transformation
+ stability terms) / learning rate

<u>Globally observable games</u>
A closed-form computation of q_t by refining analysis

Existing $\psi_t(q) = -\frac{1}{\eta_t}(H(q) + H(1-q))$

Ours $\psi_t(q) = -\frac{1}{\eta_t}H(q) \to q_{ta} \propto \exp\left(-\eta_t \sum_{s=1}^{t-1} \hat{y}_{sa}\right)$

- Future work
  - From $\mathrm{polylog} T$ to $\log T$
  - Remove the redundant $O(k)$ multiplicative factor in locally observable setting