Paper: https://arxiv.org/abs/2305.17301
Slide: https://tsuchhiii.github.io/pdf/slide/20230319.pdf



# Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds (NeurIPS 2023)

Taira Tsuchiya [1], Shinji Ito [2,3], Junya Honda [4,3]

1. The University of Tokyo, 2. NEC, 3. RIKEN, 4. Kyoto University

*Online Learning Seminar, March 19th 2024, The University of Tokyo*

# Introduction | Multi-armed bandits

- Select one of $k$ slot-machines for $T$ times to minimize the cumulative loss

The adversary determines loss vectors $\ell_1, \ldots, \ell_T \in [0,1]^k$

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

1. The learner selects arm $A_t \in \{1, \ldots, k\}$
2. The learner observes the loss of $A_t$ , $\ell_{t,A_t} \in [0,1]$

← $\ell_{t,i} \in [0,1]$ : loss of arm $i$ at time $t$

- Goal: minimize the cumulative loss $=$ minimize (pseudo-)regret $R_T$

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} - \sum_{t=1}^{T} \ell_{t,i^*}\right], \quad i^* = \arg\min_{i \in [k]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$$

# Environments in online learning and bandits

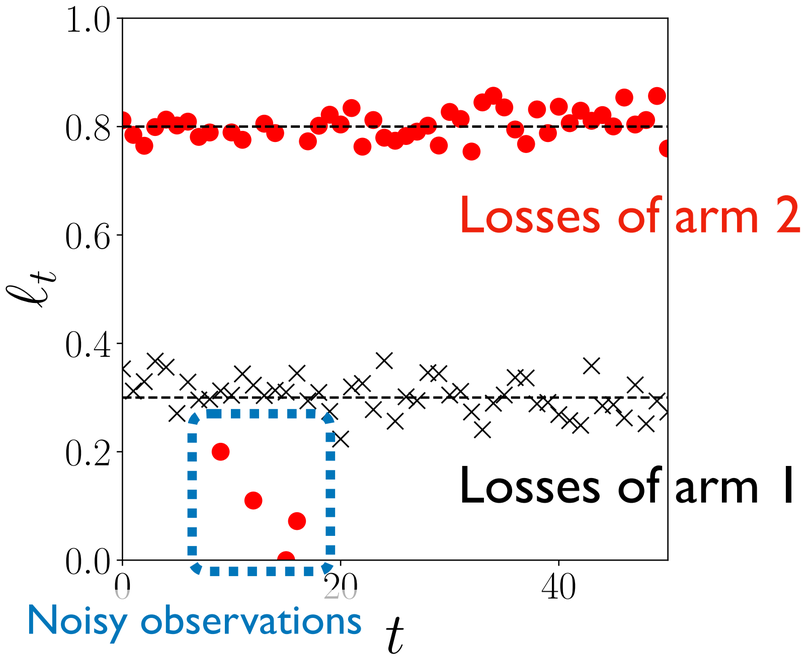
- In case of multi-armed bandits,
  - Adversarial environments: $\ell_1, \ldots, \ell_T \in [0,1]^k$, very pessimistic
  - Stochastic environments: $\ell_{t,i} \sim \nu_i^*$ for $i \in [k]$, somewhat optimistic
  - Stochastic environments with adversarial corruptions

# Stochastic environments with adversarial corruptions

**Adversarial env** : Too pessimistic

**Intermediate environment ...?**

**Stochastic env** (bounded losses)
Too optimistic




## Stochastic env with adversarial corruptions

[Lykouris, Mirrokni & Leme 2018]

Stochastically generated losses

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

Adversarial noise

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

Losses with noise

$\ell_1, \ldots, \ell_T$

$C = 0 \rightarrow$ Stochastic env

$C = 2T \rightarrow$ Adversarial env

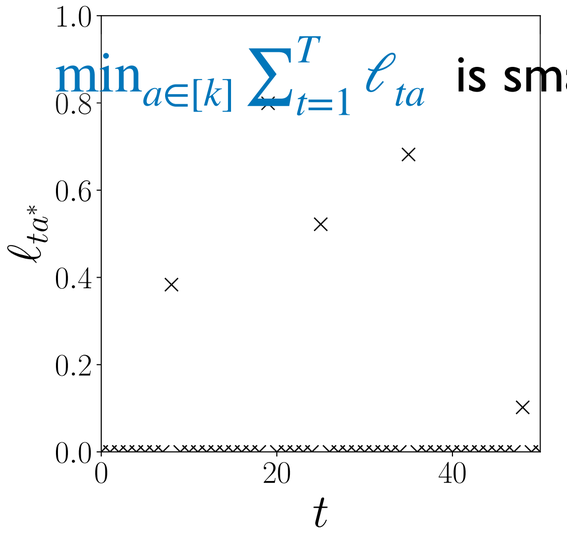
T. Lykouris, V. Mirrokni, and R.P. Leme. Stochastic bandits robust to adversarial corruptions. In STOC 2018.

# Adaptivity 1. Data-dependent bounds in adversarial env

- Adversarial env: $\ell_1, \ldots, \ell_T \in [0,1]^k$, too pessimistic
- Loss sequences in the real world usually have benign structures

first-order bound

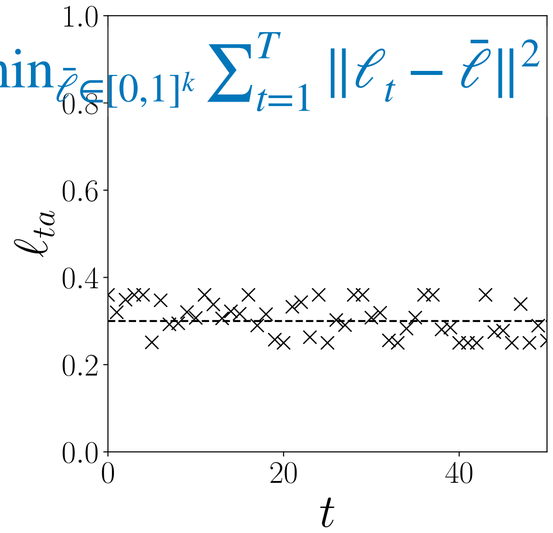
$\min_{a \in [k]} \sum_{t=1}^{T} \ell_{ta}$ is small




[Allenberg, Auer, Györfi & Ottucsák 2006]

second-order bound

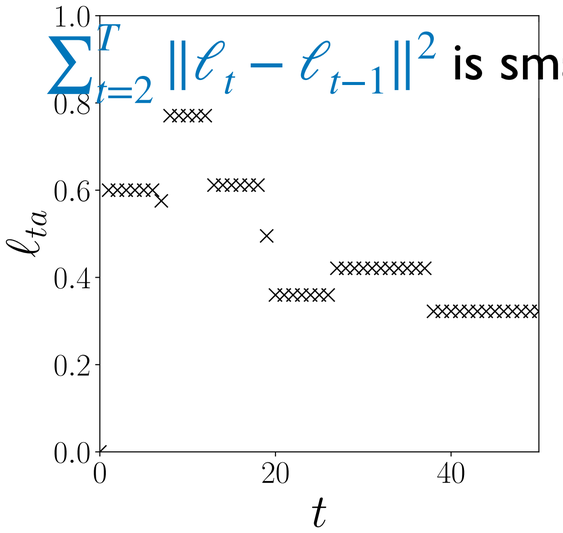
$\min_{\bar{\ell} \in [0,1]^k} \sum_{t=1}^{T} \|\ell_t - \bar{\ell}\|^2$ is small




[Hazan & Kale 2011]

path-length bound

$\sum_{t=2}^{T} \|\ell_t - \ell_{t-1}\|^2$ is small



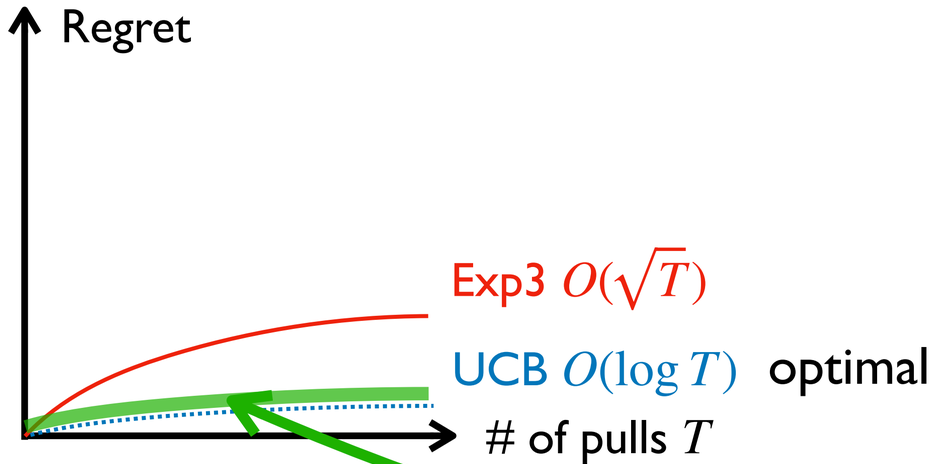

[Wei & Luo 2018]

**Data-dependent bounds**: Bounds that depend on the benign level of losses

# Adaptivity 2: Best-of-both-worlds [Bubeck & Slivkins 2012, Zimmert & Seldin 2021]

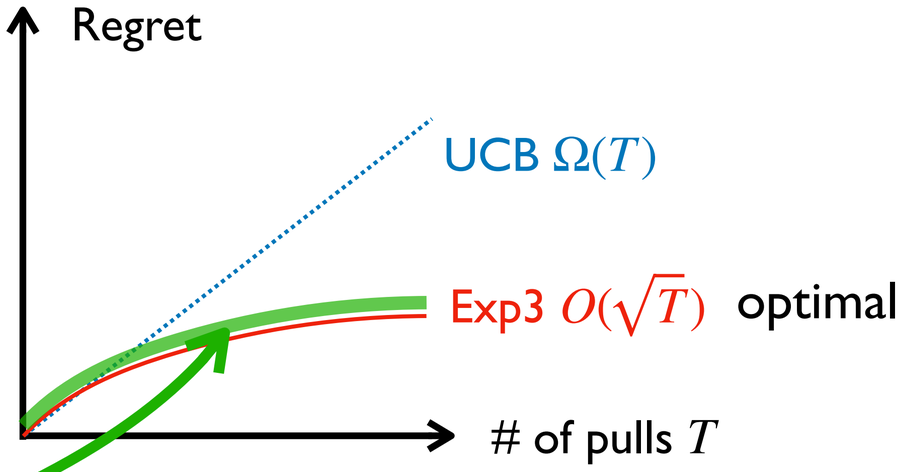
## Stochastic environment

$\ell_{t,i} \sim \nu_i^*$ for all $i \in [k]$




## Adversarial environment

$\ell_1, \ldots, \ell_T \in [0,1]^k$ are arbitrary decided




**What we want:** Achieving optimality for both stochastic and adversarial env without **knowing the underlying env** = **Best-of-Both-Worlds (BOBW)**

Better if it performs well also in stochastic env with adversarial corruption

S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In COLT 2012.
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 2021.

# Background and Research Question

- Many environment adaptivities can be realized by **Follow-the-Regularized-Leader (FTRL)** [Wei & Luo 2018, Zimmert & Seldin 2021, and many!]
- Need to design regularizers and learning rate in FTRL
- Only a few algorithms can achieve simultaneous environment adaptivities (e.g., data-dependent bounds & BOBW) (There are some for FTRL w/ log-barrier [Ito 2021, Ito-T-Honda 2022, T-Ito-Honda 2023])

**Research Question**

**Q.** Is it possible to establish an algorithm with
**a data-dependent bound and a BOBW guarantee simultaneously**?

**A.** Possible by **adapting learning rate of FTRL to multiple observations**!
→ Apply this to multi-armed bandits and partial monitoring

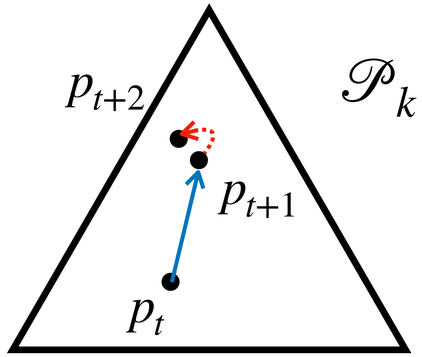
C.Y. Wei and H. Luo. More adaptive algorithms for adversarial bandits. In COLT 2018.
J. Zimmert and Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 2021.

# Outline

- Introduction
- Follow-the-Regularized-Leader and Stability-Penalty-Adaptive Learning Rate
- Case Study 1: Sparsity and Best-of-Both-Worlds in Multi-armed Bandits
- A Quick Introduction of Partial Monitoring
- Cases Study 2: Game-dependency and Best-of-Both-Worlds in Partial Monitoring
- Summary

# Follow-the-Regularized-Leader in Online Learning

- Follow-the-Regularized-Leader (FTRL):
  - Determine action selection probability $p_t \in \mathscr{P}_k$ by minimizing "sum of estimated losses so far + convex regularizer":




cumulative estimated losses  convex regularizer

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

$\hat{\ell}_t \in \mathbb{R}^k$ : some estimator of $\ell_t$

  - One of the most common approaches for achieving data-dependent bounds and BOBW

# Regret Bound: Penalty–Stability Decomposition

- FTRL can achieve these environment adaptivities
- For FTRL with the Shannon entropy regularizer with learning rate $(\eta_t)_{t=1}^T$,

$$R_T \lesssim \mathbb{E}\left[\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}\right] + \text{(insignificant term)} \quad \text{for} \quad \widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) h_{t+1} + \sum_{t=1}^T \eta_t z_t$$

penalty stability

  - **Examples** in multi-armed bandits:
    Shannon entropy regularizer and inverse-weighted estimator $\hat{\ell}_{ta} = \frac{\ell_{ta} 1[A_t = a]}{p_{ta}}$

$$h_t = H(p_t) = -\sum_{a=1}^k p_{ta} \log(p_{ta}) \le \log k, \quad z_t = \sum_{a=1}^k p_{ta} \hat{\ell}_{ta}^2 = \frac{\ell_{tA_t}^2}{p_{tA_t}}$$

# Can we make FTRL more adaptive?

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) h_{t+1} + \sum_{t=1}^{T} \eta_t z_t$$

penalty stability

- Existing adaptive learning rates $(\eta_t)_{t=1}^T$ depend on either the penalty or stability
  - With ***empirical*** stability $(z_s)_{s=1}^{t-1}$ and ***worst-case*** penalty terms $h_{\max} \geq \max_{t \in [T]} h_t$,
    we get **data-dependent bounds** [McMahan 2011 (AdaGrad); Lattimore & Szepesvári 2020, and so many!]

$$\text{e.g., In MAB, } \eta_t = \sqrt{\frac{\log k}{k + \sum_{s=1}^{t-1} \ell_{sA_s}^2 / p_{sA_s}}} \text{ corresponding to } z_t = \sum_{a=1}^{k} p_{ta} \hat{\ell}_{ta}^2 \text{ and } h_{\max} = \log k$$

  - With ***empirical*** penalty $(h_s)_{s=1}^{t-1}$ and ***worst-case*** stability $\bar{z} \geq \max_{t \in [T]} z_t$,
    we get **best-of-both-worlds** [Ito, T, & Honda 2022, T, Ito, & Honda 2023]

$$\beta_t = \frac{1}{\eta_t}, \quad \beta_1 > 0, \quad \beta_{t+1} = \beta_t + \frac{\text{const}}{\sqrt{\text{const} + \sum_{s=1}^{t-1} h_{s+1}}}$$

**Q.** Can we construct learning rates jointly dependent on the **empirical** stability and penalty?

# Stability-Penalty-Adaptive (SPA) Learning Rate

## Definition (informal)

A sequence of learning rates $(\eta_t)_{t=1}^T$ is ***stability-penalty-adaptive (SPA) learning rate*** if the update is written with a certain non-negative reals $((h_t, z_t, \bar{z}_t))_{t=1}^T$ as follows:

$$\beta_t = \frac{1}{\eta_t}, \quad \beta_1 > 0, \quad \beta_{t+1} = \beta_t + \frac{c_1 z_t}{\sqrt{c_2 + \bar{z} h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}}$$

update jointly dependent on stability $z_s$ & penalty $h_{s+1}$

## Theorem (informal)

Let $(\eta_t)_{t=1}^T$ be a SPA learning rate. Then under a certain condition on $((h_t, z_t, \bar{z}_t))_{t=1}^T$,

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \tilde{O}\left(\sqrt{c_2 + \bar{z}_t h_1 + \sum_{t=1}^{T} z_t h_{t+1}}\right)$$

regret bound jointly dependent on stability $z_s$ & penalty $h_{s+1}$

# SPA Learning Rate Generalizes Existing Learning Rates

- Letting $h_t \leftarrow h_{\max}$ for all $t \in [T]$ in SPA learning rate yields stability-dependent learning rate (AdaGrad-type learning rate) of

$$\eta_t = \frac{1}{\beta_t} \propto \sqrt{\frac{h_{\max}}{\text{const} + \sum_{s=1}^{t-1} z_s}}$$

- Letting $z_t \leftarrow \bar{z}$ for all $t \in [T]$ in SPA learning rate yields penalty-dependent learning rate of

$$\beta_t = \frac{1}{\eta_t}, \quad \beta_1 > 0, \quad \beta_{t+1} = \beta_t + \frac{\text{const}}{\sqrt{\text{const} + \sum_{s=1}^{t-1} h_{s+1}}}$$

**Q.** Can we simultaneously achieve BOBW and data-dependent bounds?
→ check in **multi-armed bandits** and **partial monitoring**

# Outline

- Introduction
- Follow-the-Regularized-Leader and Stability-Penalty-Adaptive Learning Rate
- Case Study 1: Sparsity and Best-of-Both-Worlds in Multi-armed Bandits
- A Quick Introduction of Partial Monitoring
- Cases Study 2: Game-dependency and Best-of-Both-Worlds in Partial Monitoring
- Summary

# Case Study 1. Sparsity in Multi-armed Bandits

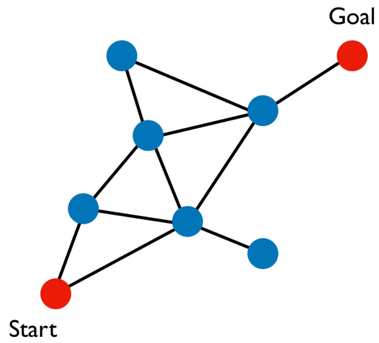
- Many problems involve sparse losses: $\ell_t \in [-1,1]^k$ with $s = \max_{t\in[T]} \|\ell_t\|_0 \ll k$

Online ads allocation

Most ads are not clicked on:
For most $a \in [k]$, $r_{ta} := -\ell_{ta} = 0$


Online shortest path

No data loss in most routes:
For most $a \in [k]$, $\ell_{ta} = 0$




- Sparsity-dependent bounds: Data-dependent bounds that depend on the sparsity level $s \ll k$
  - Lower bound: $\Omega(\sqrt{sT})$ [Kwon & Perchet, 2016]
  - Upper bound: $O(\sqrt{sT \log k})$ with **known** sparsity level $s$ [Kwon & Perchet, 2016, Bubeck, Cohen & Li, 2018]

J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. JMLR, 2016.
S. Bubeck, M. Cohen, and Y. Li. Sparsity, variance and curvature in multi-armed bandits. In ALT 2018.

# 1. Regret Upper Bounds | Sparsity and BOBW

**Theorem (informal)** There exists an algorithm based on SPA learning rate that achieves

Corrupted Stochastic Env. $R_T = O\left( \frac{s \log(T)\log(kT)}{\Delta_{\min}} + \sqrt{\frac{Cs \log(T)\log(kT)}{\Delta_{\min}}} \right)$ best-of-both-worlds

Adversarial Env. $R_T = O(\sqrt{sT \log(k)\ \log(T)})$ sparsity-dependent bound

- Technique: Evaluate the change of FTRL output. Let $h_t \simeq H(p_t)$. Then,

$$R_T \lesssim \mathbb{E}\left[\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}\right] \lesssim \tilde{O}\left(\sqrt{\textstyle\sum_{t=1}^T \mathbb{E}\left[z_t h_{t+1}\right]}\right) \lesssim \tilde{O}\left(\sqrt{\textstyle\sum_{t=1}^T \mathbb{E}\left[z_t h_t\right]}\right)$$

SPA learning rate

Lemma. $h_{t+1} \lesssim h_t + \epsilon$

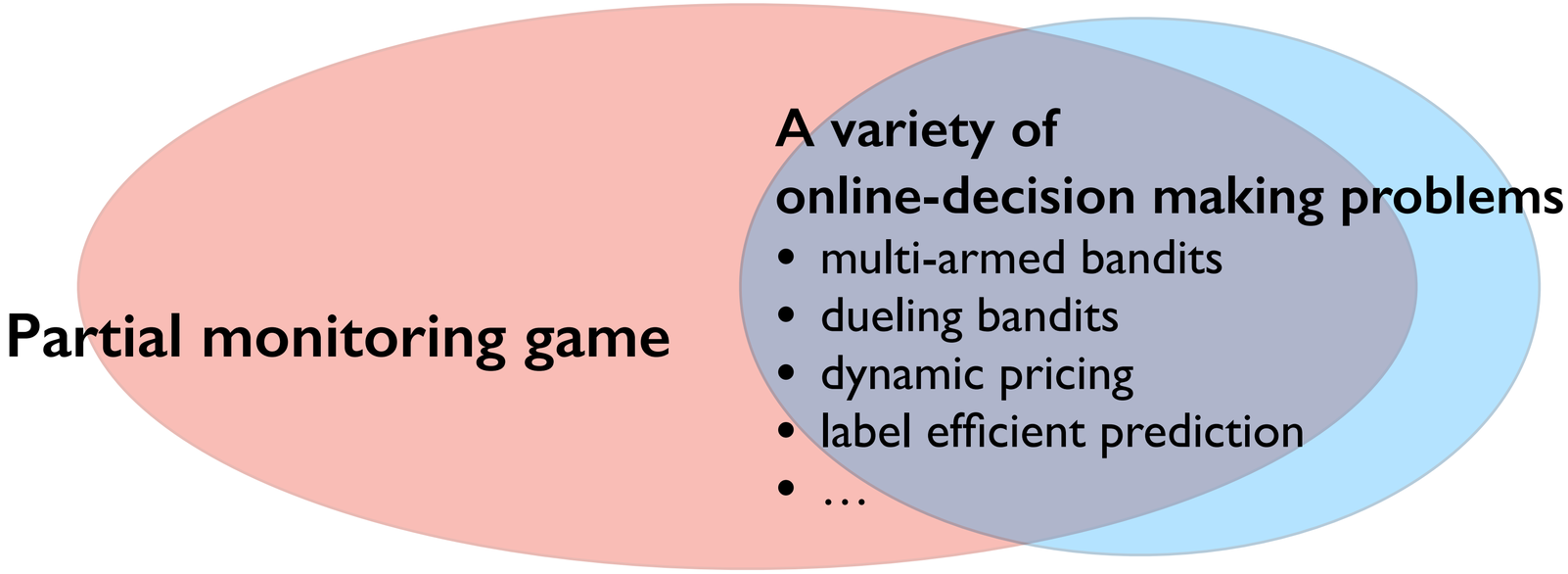
$$\longrightarrow \quad \mathbb{E}[z_t h_t] = \mathbb{E}[\mathbb{E}[z_t \,|\, p_t]\, h_t] \leq \begin{cases} s\,\mathbb{E}[h_t] \to \text{best-of-both-worlds} \\ \log(k)\,\mathbb{E}[z_t] \to \text{sparsity-dependent bound} \end{cases}$$

# Outline

- Introduction
- Follow-the-Regularized-Leader and Stability-Penalty-Adaptive Learning Rate
- Case Study 1: Sparsity and Best-of-Both-Worlds in Multi-armed Bandits
- A Quick Introduction of Partial Monitoring
- Cases Study 2: Game-dependency and Best-of-Both-Worlds in Partial Monitoring
- Summary

# Partial Monitoring:
# A general online decision-making problem

- Online learning with full information
- Multi-armed bandits
- Online learning with feedback graphs
- Dueling bandits
- Dynamic pricing
- Label efficient prediction
- ….



# Partial Monitoring Example: Dynamic Pricing

| **Learner** | **Adversary** | opportunity loss | feedback |
|---|---|---|---|
| $t = 1$ | User's **evaluation price** | | |
| Hotel owner<br>accomm. fee \$40 |  <br>Use if accomm. fee $\leq$ \$90 | $\$90 - \$40 = \$50$ | Buy |
| $t = 2$ | | | |
| decides **accommodation fee** of room from $\{\$1, \ldots, \$k\}$<br>accomm. fee \$80 |  <br>Use if accomm. fee $\leq$ \$50 | $\$c$ (const.)<br>( $\because \$50 - \$80 < \$0$) | No-buy |
| $t = \cdots$ | | | |

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

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

# Formulation of Partial Monitoring

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

**observed to the learner**

**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 $R_T$

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} L_{A_t x_t} - \sum_{t=1}^{T} L_{a^* x_t}\right], \quad a^* = \arg\min_{a \in [k]} \mathbb{E}\left[\sum_{t=1}^{T} L_{a x_t}\right]$$

cumulative losses of taken actions

cumulative losses of optimal action

# Example 1. Dynamic Pricing [Kleinberg & Leighton 2003]

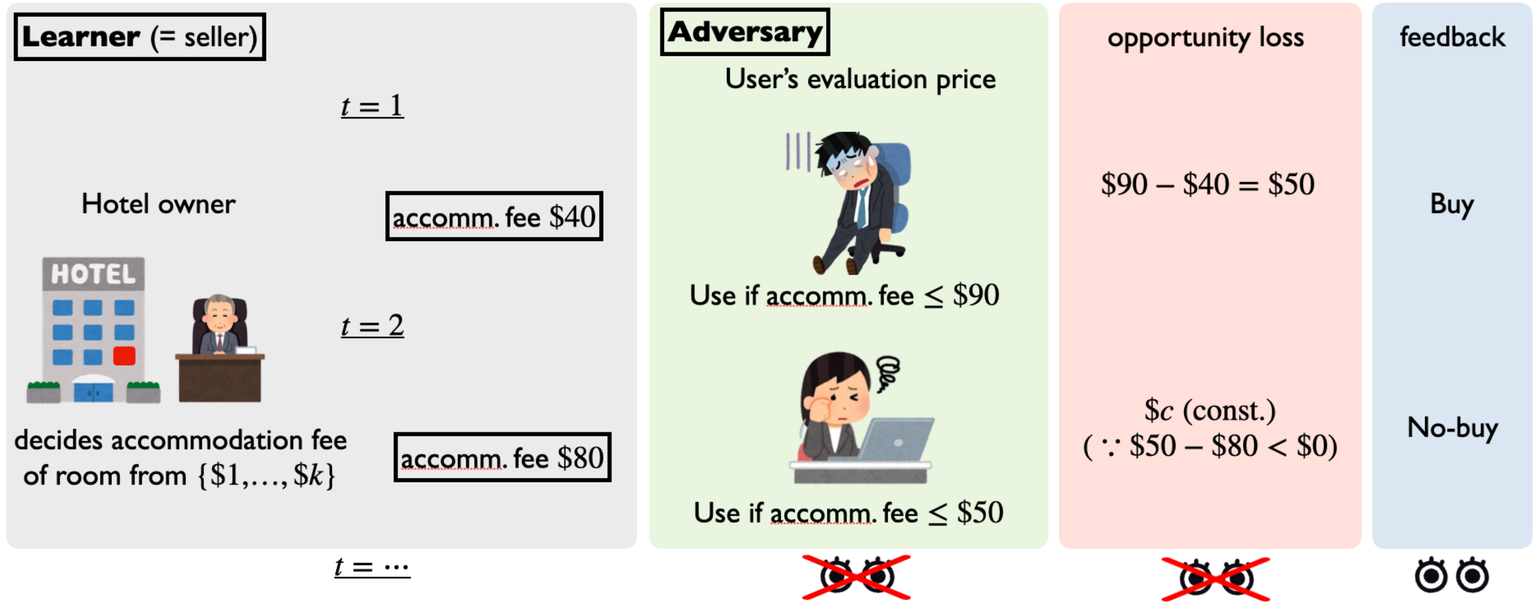
- Partial m… …ctions and $M$ outcomes
- loss ma… $\Sigma^{N\times M}$ ($\Sigma$ : set of feedback symbols)

$k$ : discrete range of accom. fee
$d$ : discrete range of evaluation price

selecting action
= determining the accom. fee

outcome
= evaluation price

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




(row: selling price, column: evaluation price)

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

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

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

($x \geq a$ / $x < a$)

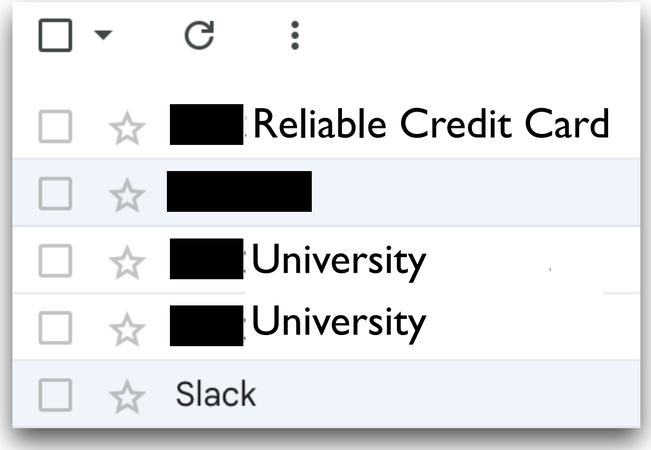
R. Kleinberg and T. Leighton, The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In FOCS 2003.

# Example 2. Apple Tasing, Matching Pennies [Helmbold, Littlestone & Long 1992]

- Sequentially determining whether emails received in the mailbox are spam or ham (not spam)
- Three possible actions when labeling emails:
  1. Label as spam (P)
  2. Label as ham (N)
  3. Consulting with humans to obtain the correct label
     (Oonly in this case, the true label can only be observed)

spam? →

ham? →




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

$c_{\mathrm{N}\to\mathrm{P}} > 0$ : failure cost of N to P
$c_{\mathrm{P}\to\mathrm{N}} > 0$ : failure cost of P to N
$q > 0$ : cost of asking the expert

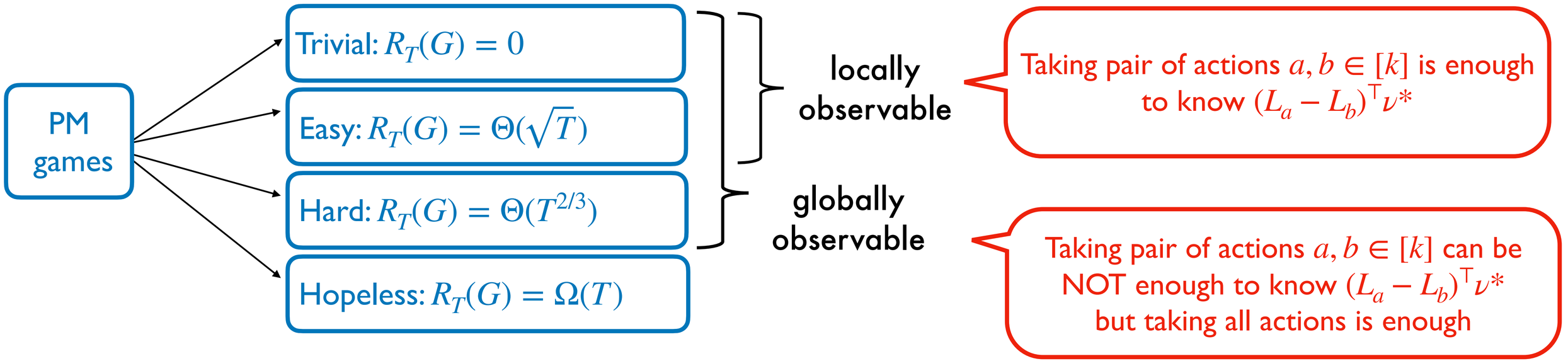
$$\Phi = \begin{pmatrix} \mathrm{None} & \mathrm{None} \\ \mathrm{None} & \mathrm{None} \\ \mathrm{P} & \mathrm{N} \end{pmatrix}$$

D. P. Helmbold, N. Littlestone, and P. M. Long, Apple tasting and nearly one-sided learning. In FOCS 1992.

# Classification of Partial Monitoring Games

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

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




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 Environments in Partial Monitoring

- Stochastic env: $x_t \overset{\text{i.i.d.}}{\sim} \nu^* \in \mathscr{P}_d$ (dist. over outcomes)
- Adversarial env: $x_t$ arbitrarily decided

**Adversarial env**: Pessimistic

Intermediate env …?

**Stochastic env**: Optimistic

- Stochastic env w/ adversarial corruptions (for PM)

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

Outcomes sampled in i.i.d. manner

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

adversarial noise at most $C$

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

Outcomes with noise

$x_1, \ldots, x_T$ 

$C = 0 \rightarrow$ 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.

# T-Ito-Honda (ALT2023): FTRL with Shannon entropy and "Exploration-by-Optimization"

- 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 |
|---|---|---|---|
| [T-Honda-Sugiyama 20] | $O(\log T)$ | NA | NA |
| [Lattimore-Szepesvári 20] | NA | $O(\sqrt{T})$ | NA |
| **[T-Ito-Honda 23]** | $O((\log T)^2)$ | $O(\sqrt{T}\log T)$ | $O((\log T)^2 + \sqrt{C}\log T)$ |

- Globally observable games

| | Stochastic | Adversarial | Stochastic w/ Corruptions |
|---|---|---|---|
| [Lattimore-Szepesvári 20] | NA | $O(T^{2/3})$ | NA |
| **[T-Ito-Honda 23]** | $O((\log T)^2)$ | $O((T\log T)^{2/3})$ | $O((\log T)^2 + (C\log T)^{2/3})$ |

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.
T. Tsuchiya, S. Ito, and J. Honda. Best-of-both-worlds algorithms for partial monitoring. In ALT 2023.

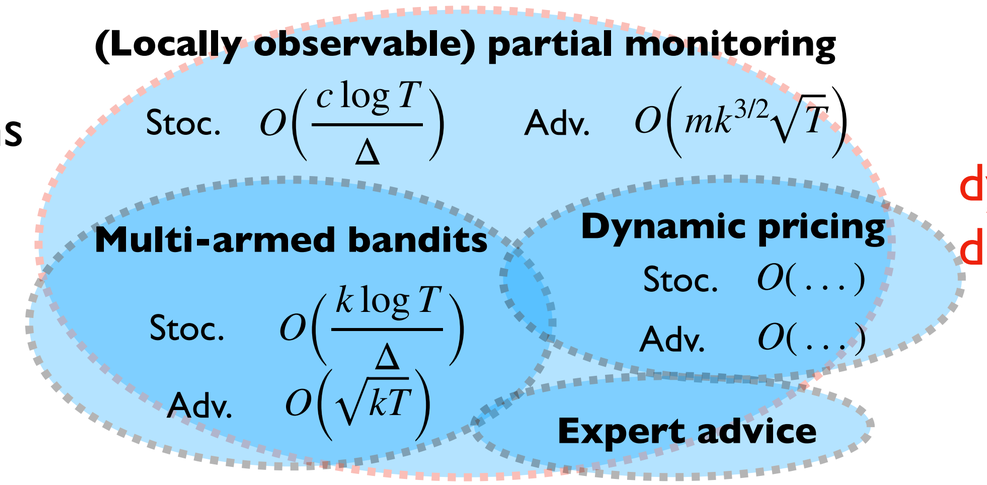
# Outline

- Introduction
- Follow-the-Regularized-Leader and Stability-Penalty-Adaptive Learning Rate
- Case Study 1: Sparsity and Best-of-Both-Worlds in Multi-armed Bandits
- A Quick Introduction of Partial Monitoring
- Cases Study 2: Game-dependency and Best-of-Both-Worlds in Partial Monitoring
- Summary

# Case Study 2. Game-dependency and BOBW in PM

- Partial monitoring: Very general framework for online decision-making under abstract feedback
- **Limitation**: Formulations and algorithms are conservative and (sometimes) not practical
- Desirable to automatically achieve regret that depends on the inherent difficulty of the problem being solved

Hierarchical structure of online decision-making problems




dynamically achieve the optimality defined by the structures of $L$ and $\Phi$
→ **game-dependent bounds**

[Lattimore & Szepesvári 2020]

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

# 2. Regret Upper Bounds

**Theorem (informal)** For locally observable partial monitoring games, by SPA learning rate,

Adversarial Env. $$R_T \leq \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} V_t' \log(k)\log(1+T)}\right] + o(\log T)$$

Corrupted Stochastic Env. $$R_T = O\left(\frac{r_{\mathscr{M}} \bar{V} \log(T)\log(kT)}{\Delta_{\min}} + \sqrt{\frac{C r_{\mathscr{M}} \bar{V} \log(T)\log(kT)}{\Delta_{\min}}}\right) + o(\log T)$$

$V_t'$, $\bar{V}$ : variables dependent on problem's inherent difficulty

Existing bounds: the value for ▢ is replaced with the worst-case scenario of the hardest problems.

↔ Our bounds: if the game is easier (possibly unknown), the value adjusts accordingly.

# Summary: Stability-Penalty-Adaptive FTRL

**The main term of regret upper bound of FTRL**

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \sum_{t=1}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) h_{t+1} + \sum_{t=1}^{T} \eta_t z_t$$

penalty stability

**Stability-penalty-adaptive learning rate**

$$\beta_{t+1} = \beta_t + \frac{c_1 z_t}{\sqrt{c_2 + \bar{z} h_1 + \sum_{s=1}^{t-1} z_s h_{s+1}}}$$

**Regret bound jointly dependent on stability and penalty**

$$\widehat{\mathrm{Reg}}_T^{\mathrm{SP}} = \tilde{O}\left( \sqrt{c_2 + \bar{z}_t h_1 + \sum_{t=1}^{T} z_t h_{t+1}} \right)$$

**1. Multi-armed bandits**

Sparsity-dependent bound
and best-of-both-worlds guarantee

**2. Partial monitoring**

Game-dependent bound
and best-of-both-worlds guarantee

Thank you!
paper: arxiv.org/abs/2305.17301