Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds

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

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

NEURAL INFORMATION PROCESSING SYSTEMS

Environment Adaptivity of Follow-the-Regularized-Leader in Online Decision-Making Problems: Multi-armed Bandits Case

Multi-armed bandits (MAB)

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

The adversary determines loss vectors $\ell_1, \dots, \ell_T \in [0,1]^k$
For $t = 1, \dots, T$:
1. The learner selects arm $A_t \in \{1, \dots, k\}$
2. The learner observes the loss of A_t, $\ell_{t,A_t} \in [0,1]$

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

$$a^* = \arg\min_{a \in \{1,\dots,k\}} \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right]$$

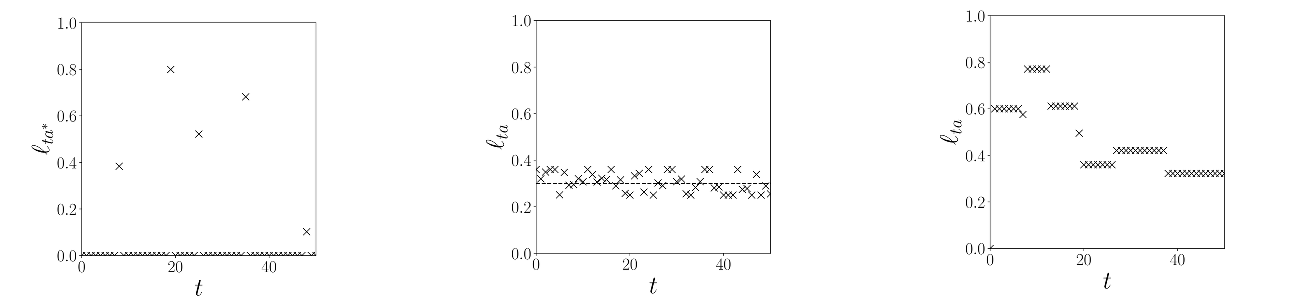
$$R_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t} - \sum_{t=1}^T \ell_{t,a^*}\right]$$

Environments in bandit problems

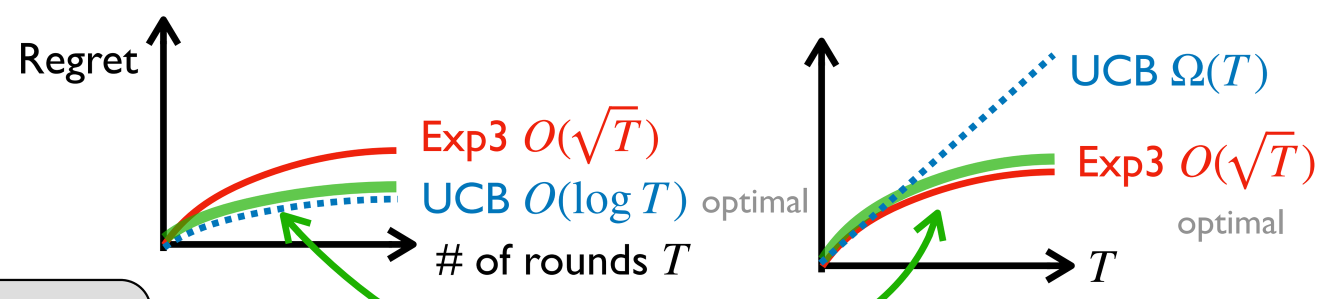
Stochastic Environment	$\ell_{t,a} \sim \nu_a^*$ for all $a \in [k]$
Corrupted	Intermediate regime
Adversarial Environment	$\ell_1, \dots, \ell_T \in [0,1]^k$ are arbitrarily determined

Environment adaptivity

Data-dependent bounds:
Bounds that depend on the benign level of losses in adversarial env.



Best-of-both-worlds: simultaneous optimality in stoc. & adv. env.



The learner has no knowledge on environment

What we desire: simultaneous optimality
= Best-of-Both-Worlds (BOBW)

Background

- Many environment adaptivities can be realized by Follow-the-Regularized-Leader (FTRL) [Wei & Luo 18, Zimmert & Seldin 21, etc]
- Need to design regularizers and learning rate in FTRL
- Only a few algorithms can achieve simultaneous environment adaptivities (e.g., data-dependent bounds & BOBW)

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 simultaneously!
→ apply this to MAB and partial monitoring

Follow-the-Regularized-Leader and Adaptive Learning Rate

Follow-the-Regularized-Leader (FTRL)

- Decide arm selection probability $p_t \in \mathscr{P}_k$ by minimizing "cumulative losses + regularizer"

cumulative estimated loss (strongly-)convex regularizer

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

learning rate

$\hat\ell_s \in \mathbb{R}^k$: estimator of ℓ_s

$\mathscr{P}_k$: $(k-1)$-dim. prob. simplex

- Most of data-dependent bounds and BOBW bounds are obtained by FTRL (or OMD)
- Achieved by adaptively determining the learning rate η_t based on previous observations
→ called **adaptive learning rate**

Adaptive learning rate w/ entropic regularizer $\phi_t(p) = \sum_{a=1}^k p_a \log p_a$

- Main part of the regret of FTRL with learning rate $(\eta_t)_{t=1}^T$ is the expectation of the following $\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}$:

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

penalty / stability

penalty: large when the regularization is strong
stability: large when p_t and p_{t+1} are far apart

- Existing adaptive learning rate $(\eta_t)_{t=1}^T$ depends only on the penalty or stability
  - η_t with *empirical* stability $(z_s)_{s=1}^{t-1}$ & *worst-case* penalty $h_{\max}(\geq \max_{t\in[T]} h_t)$
    → induces data-dependent bounds [McMahan 2011; Lattimore & Szepesvári 2020, and many!]
  - η_t with *empirical* penalty $(h_s)_{s=1}^{t-1}$ & *worst-case* stability $z_{\max}(\geq \max_{t\in[T]} z_t)$
    → induces best-of-both-worlds bounds [Ito, Tsuchiya & Honda, 2022, Tsuchiya, Ito & Honda, 2023]

Q. Can we construct adaptive learning rate simultaneously dependent on the empirical penalty and stability?

Stability-penalty-adaptive Learning Rate (SPA learning rate)

Definition. (informal)

Learning rate $(\eta_t)_{t=1}^T$ is *stability-penalty-adaptive (SPA) learning rate* if there exist non-negative reals $((h_t, z_t, \bar z_t))_{t=1}^T$ satisfying a certain condition and η_t is

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

design that jointly depends on stability z_s and penalty h_{s+1}

Theorem. (informal)

Let $(\eta_t)_{t=1}^T$ be a SPA learning rate. If $((h_t, z_t, \bar z_t))_{t=1}^T$ in the SPA learning rate satisfies

Stability condition: $\frac{\sqrt{c_2 + \bar z_t h_1}}{c_1}(\beta_1 + \beta_t) \geq \epsilon + z_t$ for all $t \in [T]$ for some $\epsilon > 0$

then

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

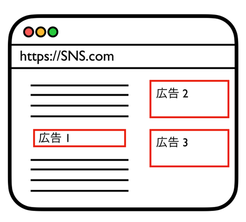
bound that jointly depends on stability z_s and penalty h_{s+1}

Q. Possible to achieve BOBW and data-dependent bounds simultaneously?
→ Verifying cases of multi-armed bandits and partial monitoring

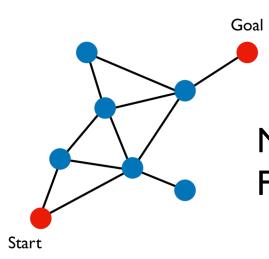
Case Study 1: Sparsity and BOBW in Multi-armed Bandits

Sparsity-dependent bounds (∈ data-dependent bounds)

- Many problems have 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 path control
No data loss in most routes:
For most $a \in [k]$, $\ell_{ta} = 0$

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

Simultaneously achieving sparsity-dependent and BOBW bounds

Theorem. (informal) There exists an algorithm based on the SPA learning rate achieving

Stochastic Env. $R_T = O\left(\frac{s\log(T)\log(kT)}{\Delta_{\min}}\right)$ Adversarial Env. $R_T = O(\sqrt{sT\log(k)\log(T)})$

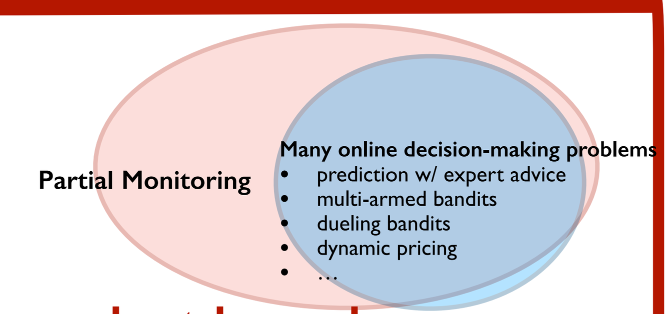
techniques: 1. sparsity estimation, 2. handle negative losses, 3. evaluate change of FTRL output

$$R_T \lesssim \mathbb{E}\left[\widehat{\mathrm{Reg}}_T^{\mathrm{SP}}\right] \underset{\text{SPA learning rate}}{\lesssim} \tilde O\left(\sqrt{\sum_{t=1}^T \mathbb{E}[z_t h_{t+1}]}\right) \underset{\text{Lemma. } h_{t+1} \lesssim h_t + \epsilon}{\lesssim} \tilde O\left(\sqrt{\sum_{t=1}^T \mathbb{E}[z_t h_t]}\right)$$

Case Study 2: Game-dependency and BOBW in Partial Monitoring

Partial monitoring and examples

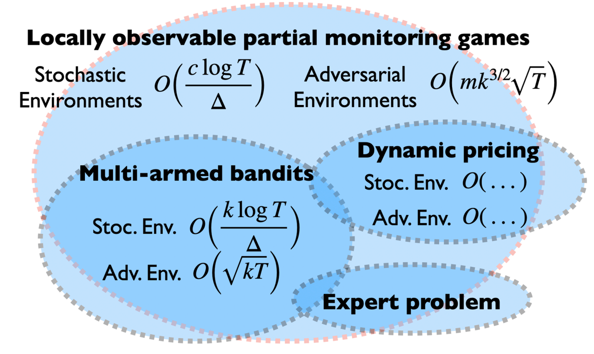
Very general framework for online decision-making under abstract feedback



Limitation of partial monitoring and game-dependent bounds

Formulations and algorithms are conservative and thus (sometimes) not practical

Hierarchical structure of online decision-making problems



Regret bounds that automatically depends on the inherent difficulty of the problem being solved
= game-dependent bounds [Lattimore & Szepesvári 2020]

Simultaneously achieving game-dependent and BOBW bounds

stability term

$$V_t' \simeq \min_{p\in\mathscr{P}_k'} \max_{x\in[d]} \left[\frac{(p-q_t)^\top L e_x}{\eta_t} + \frac{1}{\eta_t^2}\sum_{a=1}^k p_a \Psi_{q_t}\left(\frac{\eta_t G(a,\Phi_{ax})}{p_a}\right)\right] \leq \begin{cases} 1/2 & \text{if expert problems} \\ k/2 & \text{if MAB} \\ 3m^2k^3 & \text{if Locally observable PM games} \end{cases} =: \bar V$$

$V_t', \bar V$: game-dependent variables

Theorem. (informal) For locally observable PM games, an alg. w/ SPA learning rate can

Stochastic Env. $R_T = O\left(\frac{r_{\mathcal{M}} \bar V \log(T)\log(kT)}{\Delta_{\min}}\right)$ Adversarial Env. $R_T = O\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^T V_t' \log(k)\log(T)}\right]\right)$

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.

References
C.Y. Wei & H. Luo. More adaptive algorithms for adversarial bandits. In COLT 2018.
J. Zimmert & Y. Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. JMLR 2021.
J. Kwon & V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. JMLR 2016.
S. Bubeck, M. Cohen, & Y. Li. Sparsity, variance and curvature in multi-armed bandits. In ALT 2018.
T. Lattimore & Csaba Szepesvári. "Exploration by optimisation in partial monitoring." In COLT 2020.