A Simple and Adaptive Learning Rate for FTRL in Online Learning with Minimax Regret of $\Theta(T^{2/3})$ and its Application to Best-of-Both-Worlds

https://arxiv.org/abs/2405.20028

Taira Tsuchiya and Shinji Ito

The University of Tokyo & RIKEN

December 12, 2024

General Online Learning Framework

Given a finite action set $\mathcal{A} = [k] := \{1, \ldots, k\}$ and an observation set $\mathcal{O}$

for $t = 1, 2, \ldots, T$ **do**

- Environment determines a loss function $\ell_t \colon \mathcal{A} \to [0, 1]$
- Learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing ℓ_t
- Learner then suffers a loss $\ell_t(A_t)$ and observes a feedback $o_t \in \mathcal{O}$

General Online Learning Framework

Given a finite action set $\mathcal{A} = [k] := \{1, \dots, k\}$ and an observation set $\mathcal{O}$

for $t = 1, 2, \dots, T$ **do**

  Environment determines a loss function $\ell_t \colon \mathcal{A} \to [0, 1]$

  Learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing ℓ_t

  Learner then suffers a loss $\ell_t(A_t)$ and observes a feedback $o_t \in \mathcal{O}$

Learner's Goal: Minimize the **(pseudo-)regret** R_T

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \sum_{t=1}^{T} \ell_t(a^*)\right] \quad \text{for} \quad a^* \in \operatorname*{arg\,min}_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a)\right]$$

General Online Learning Framework

Given a finite action set $\mathcal{A} = [k] := \{1, \ldots, k\}$ and an observation set $\mathcal{O}$

for $t = 1, 2, \ldots, T$ **do**
- Environment determines a loss function $\ell_t \colon \mathcal{A} \to [0, 1]$
- Learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing ℓ_t
- Learner then suffers a loss $\ell_t(A_t)$ and observes a feedback $o_t \in \mathcal{O}$

Learner's Goal: Minimize the **(pseudo-)regret** R_T

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t) - \sum_{t=1}^{T} \ell_t(a^*)\right] \quad \text{for} \quad a^* \in \arg\min_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a)\right]$$

Examples of this framework

- expert problem: observe entire loss vectors $o_t = \ell_t \in [0, 1]^k$
- multi-armed bandits: observe a loss of chosen arm $o_t = \ell_t(A_t)$

Follow-the-Regularized-Leader (FTRL)

A highly powerful framework for such online learning problems

Select an action selection probability vector q_t over $\mathcal{A}$
by minimizing the sum of cumulative (estimated) loss $\sum_{s=1}^{t-1} \widehat{\ell}_s(q)$ so far plus convex regularizer ψ:

$$q_t \in \operatorname*{arg\,min}_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\}, \quad A_t \sim q_t$$

- $\mathcal{P}_k$: the set of probability distributions over $\mathcal{A} = [k]$
- $\beta_t > 0$: (a reciprocal of) learning rate at round t

Follow-the-Regularized-Leader (FTRL)

A highly powerful framework for such online learning problems

Select an action selection probability vector q_t over $\mathcal{A}$
by minimizing the sum of cumulative (estimated) loss $\sum_{s=1}^{t-1} \widehat{\ell}_s(q)$ so far plus convex regularizer ψ:

$$q_t \in \arg\min_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\}, \quad A_t \sim q_t$$

- $\mathcal{P}_k$: the set of probability distributions over $\mathcal{A} = [k]$
- $\beta_t > 0$: (a reciprocal of) learning rate at round t

FTRL can perform adaptively to various properties of underlying loss functions
by designing its regularizer ψ and learning rate $(\beta_t)_t$!
$\to$ Q. How to tune the learning rate?

Stability–Penalty Decomposition

The regret of FTRL is roughly bounded as

$$R_T \lesssim \underbrace{\sum_{t=1}^{T} \frac{z_t}{\beta_t}}_{\text{stability term}} + \underbrace{\beta_1 h_1 + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) h_t}_{\text{penalty term}} .$$

- **stability** term: large when the difference in FTRL outputs, q_t and q_{t+1}, is large
- **penalty** term: due to the strength of the regularizer

There is a tradeoff between these two terms.

Examples of z_t and h_t

When using FTRL with the negative Shannon entropy regularizer $-H(\cdot)$ (Exp3) in MAB [Aue+02],
penalty is $h_t = H(q_t)$ or $h_t = \log k$, stability is $z_t = \mathbb{E}\left[\|\widehat{\ell}_t\|^2_{(\nabla^2 \psi(q_t))^{-1}}\right]$.

Adaptive Learning Rate in the Literature

Adaptive learning rates allow us to achieve various adaptive bounds

e.g., data-dependent bounds (first-order/second-order/path-length bounds), best-of-both-worlds bounds

- Use **empirical** stability $(z_s)_{s=1}^{t-1}$ and **worst-case** penalty terms $h_{\max} \geq \max_t h_t$

  e.g., AdaGrad [MS10; DHS11], first-order algorithms [AHR12], and many!

$$1/\beta_t = \sqrt{\frac{\mathrm{const}}{\mathrm{const} + \sum_{s=1}^{t-1} z_s}}$$

Adaptive Learning Rate in the Literature

Adaptive learning rates allow us to achieve various adaptive bounds

e.g., data-dependent bounds (first-order/second-order/path-length bounds), best-of-both-worlds bounds

- Use **empirical** stability $(z_s)_{s=1}^{t-1}$ and **worst-case** penalty terms $h_{\max} \geq \max_t h_t$

  e.g., AdaGrad [MS10; DHS11], first-order algorithms [AHR12], and many!

$$1/\beta_t = \sqrt{\frac{\mathrm{const}}{\mathrm{const} + \sum_{s=1}^{t-1} z_s}}$$

- Use **empirical** penalty $(h_s)_{s=1}^{t-1}$ and **worst-case** stability terms $z_{\max} \geq \max_t z_t$

  for best-of-both-worlds bounds *e.g.*, [ITH22; TIH23a]

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

Adaptive Learning Rate in the Literature

Adaptive learning rates allow us to achieve various adaptive bounds

e.g., data-dependent bounds (first-order/second-order/path-length bounds), best-of-both-worlds bounds

- Use **empirical** stability $(z_s)_{s=1}^{t-1}$ and **worst-case** penalty terms $h_{\max} \geq \max_t h_t$

  e.g., AdaGrad [MS10; DHS11], first-order algorithms [AHR12], and many!

$$1/\beta_t = \sqrt{\frac{\mathrm{const}}{\mathrm{const} + \sum_{s=1}^{t-1} z_s}}$$

- Use **empirical** penalty $(h_s)_{s=1}^{t-1}$ and **worst-case** stability terms $z_{\max} \geq \max_t z_t$

  for best-of-both-worlds bounds *e.g.*, [ITH22; TIH23a]

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

- Use both empirical stability and penalty [TIH23b; JLL23; ITH24]

  for simultaneous data-dependent bounds and best-of-both-worlds bounds or Tsallis entropy regularizer

Adaptive Learning Rate in the Literature

Adaptive learning rates allow us to achieve various adaptive bounds

e.g., data-dependent bounds (first-order/second-order/path-length bounds), best-of-both-worlds bounds

- Use **empirical** stability $(z_s)_{s=1}^{t-1}$ and **worst-case** penalty terms $h_{\max} \geq \max_t h_t$

  e.g., AdaGrad [MS10; DHS11], first-order algorithms [AHR12], and many!

$$1/\beta_t = \sqrt{\frac{\text{const}}{\text{const} + \sum_{s=1}^{t-1} z_s}}$$

- Use **empirical** penalty $(h_s)_{s=1}^{t-1}$ and **worst-case** stability terms $z_{\max} \geq \max_t z_t$

  for best-of-both-worlds bounds *e.g.*, [ITH22; TIH23a]

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

- Use both empirical stability and penalty [TIH23b; JLL23; ITH24]

  for simultaneous data-dependent bounds and best-of-both-worlds bounds or Tsallis entropy regularizer

Almost all adaptive learning rates are for problems with a minimax regret of $\Theta(\sqrt{T})$

$\leftrightarrow$ Limited investigation into problems with a minimax regret of $\Theta(T^{2/3})$

Research Questions

There are many important online learning problems with a minimax regret of $\Theta(T^{2/3})$:

- partial monitoring with global observability [BPS11; LS19]
- graph bandits with weak observability [Alo+15]
- bandits with paid observations [Sel+14]
- dueling bandits [SKM21]
- online ranking [CT17]
- bandits with switching costs [Dek+14]

Research Questions

There are many important online learning problems with a minimax regret of $\Theta(T^{2/3})$:

- partial monitoring with global observability [BPS11; LS19]
- graph bandits with weak observability [Alo+15]
- bandits with paid observations [Sel+14]
- dueling bandits [SKM21]
- online ranking [CT17]
- bandits with switching costs [Dek+14]

Research Question

Can we provide a unified adaptive learning rate framework for online learning with a minimax regret of $\Theta(T^{2/3})$, which allows us to achieve a certain adaptivity?

Objective Function that Adaptive Learning aims to Minimize

In online learning with the minimax regret of $\Theta(T^{2/3})$, it is common to use forced exploration for FTRL:

$$q_t \in \arg\min_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\}, \quad A_t \sim p_t = (1 - \gamma_t) q_t + \gamma_t u \quad \text{for } u \in \mathcal{P}_k$$

Objective Function that Adaptive Learning aims to Minimize

In online learning with the minimax regret of $\Theta(T^{2/3})$, it is common to use forced exploration for FTRL:

$$q_t \in \arg\min_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\}, \quad A_t \sim p_t = (1-\gamma_t) q_t + \gamma_t u \quad \text{for } u \in \mathcal{P}_k$$

The regret of FTRL with a somewhat large exploration rate γ_t is known to be bounded as

Stability–penalty–bias decomposition

$$\mathsf{R}_T \lesssim \underbrace{\sum_{t=1}^{T} \frac{z_t}{\beta_t \gamma_t}}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} (\beta_t - \beta_{t-1}) h_t}_{\text{penalty term}} + \underbrace{\sum_{t=1}^{T} \gamma_t}_{\text{bias term}} \qquad (1)$$

Objective Function that Adaptive Learning aims to Minimize

In online learning with the minimax regret of $\Theta(T^{2/3})$, it is common to use forced exploration for FTRL:

$$q_t \in \arg\min_{q \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \widehat{\ell}_s(q) + \beta_t \psi(q) \right\}, \quad A_t \sim p_t = (1 - \gamma_t) q_t + \gamma_t u \quad \text{for } u \in \mathcal{P}_k$$

The regret of FTRL with a somewhat large exploration rate γ_t is known to be bounded as

Stability–penalty–bias decomposition

$$\mathsf{R}_T \lesssim \underbrace{\sum_{t=1}^{T} \frac{z_t}{\beta_t \gamma_t}}_{\text{stability term}} + \underbrace{\sum_{t=1}^{T} (\beta_t - \beta_{t-1}) h_t}_{\text{penalty term}} + \underbrace{\sum_{t=1}^{T} \gamma_t}_{\text{bias term}} \qquad (1)$$

Goal: construct adaptive learning rate that minimizes (1) under the constraints that $(\beta_t)_t$ is non-decreasing and β_t depends on $(z_{1:t}, h_{1:t})$ or $(z_{1:t-1}, h_{1:t})$.

Step 1: Choose Exploration Rate γ_t

A naive way: choose $\gamma_t = \sqrt{z_t/\beta_t}$ so that the stability term and the bias term match.
$\rightarrow$ this choice does not work well because to obtain a regret bound of (1), a lower bound of $\gamma_t \geq u_t/\beta_t$ for some $u_t > 0$ is needed.
(This lower bound is used to control the magnitude of the loss estimator $\widehat{\ell}_t$.)

Step 1: Choose Exploration Rate γ_t

A naive way: choose $\gamma_t = \sqrt{z_t/\beta_t}$ so that the stability term and the bias term match.
$\rightarrow$ this choice does not work well because to obtain a regret bound of (1), a lower bound of $\gamma_t \geq u_t/\beta_t$ for some $u_t > 0$ is needed.
(This lower bound is used to control the magnitude of the loss estimator $\widehat{\ell}_t$.)

Alternative solution: consider the exploration rate of

$$\gamma_t = \gamma'_t + u_t/\beta_t \quad \text{for} \quad u_t > 0$$

With these choices, setting $\gamma'_t = \sqrt{z_t/\beta_t}$ yields

$$\begin{aligned} \text{Eq.(1)} &\leq \sum_{t=1}^{T} \left(\frac{z_t}{\beta_t \gamma'_t} + (\beta_t - \beta_{t-1}) h_t + \left(\gamma'_t + \frac{u_t}{\beta_t} \right) \right) \\ &= \sum_{t=1}^{T} \Bigg(\underbrace{2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t}}_{\text{stability + bias}} + \underbrace{(\beta_t - \beta_{t-1}) h_t}_{\text{penalty}} \Bigg) =: F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \,. \end{aligned}$$

Step 2: Choose Learning Rate β_t

Idea: choose β_t so that stability + bias terms and penalty term match! (inspired by [ITH24])

$$2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} = (\beta_t - \beta_{t-1}) h_t \tag{2}$$

Inspired by the above matching, consider

Stability–Penalty–Bias Matching (SPB-Matching, Rule 2 in the paper)

$$\beta_t = \beta_{t-1} + \frac{1}{\widehat{h}_t}\left(2\sqrt{\frac{z_{t-1}}{\beta_{t-1}}} + \frac{u_{t-1}}{\beta_{t-1}}\right) \quad \text{and} \quad \gamma_t = \sqrt{z_t/\beta_t} + u_t/\beta_t$$

Assume that when choosing β_t, we have an access to $\widehat{h}_t \geq h_t$.

Designed by following the simple principle of matching the stability, penalty, and bias elements!

Main Result (1): SPB-matching

Theorem

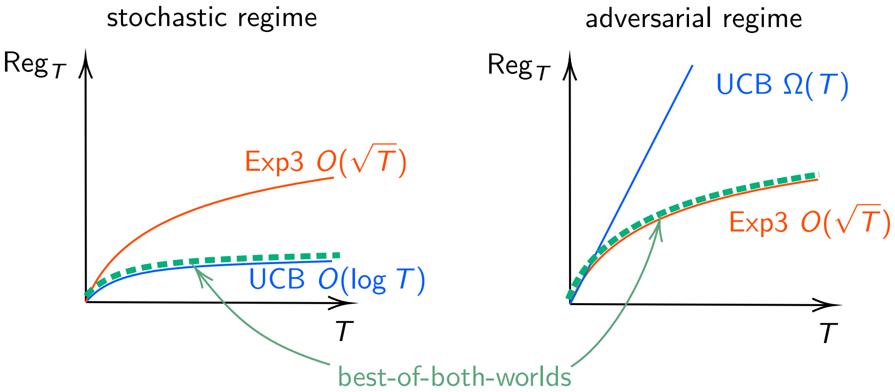
If learning rate β_t is given by SPB-matching, then for all $\epsilon \geq 1/T$,

$$
\begin{aligned}
&F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \\
&\lesssim \min\left\{ \left(\sum_{t=1}^{T} \sqrt{z_t \widehat{h}_{t+1} \log(\epsilon T)} \right)^{\frac{2}{3}} + \left(\sqrt{z_{\max} \widehat{h}_{\max}} \Big/ \epsilon \right)^{\frac{2}{3}}, \left(\sum_{t=1}^{T} \sqrt{z_t \widehat{h}_{\max}} \right)^{\frac{2}{3}} \right\} \\
&\quad + \min\left\{ \sqrt{\sum_{t=1}^{T} u_t \widehat{h}_{t+1} \log(\epsilon T)} + \sqrt{u_{\max} \widehat{h}_{\max}/\epsilon}\,, \sqrt{\sum_{t=1}^{T} u_t \widehat{h}_{\max}} \right\} .
\end{aligned}
$$

- Depending on the stability component z_t and the penalty component h_t simultaneously
- Different from the existing stability–penalty adaptive type bounds $O\left(\sqrt{\sum_{t=1}^{T} z_t \widehat{h}_{t+1} \log T}\right)$ in [TIH23b; JLL23; ITH24]

Application: Best-of-Both-Worlds Algorithms

Best-of-Both-Worlds (BOBW) algorithm:
achieve a near-optimal regret for stochastic and adversarial envs **simultaneously**



FTRL is known to be useful for constructing BOBW algorithms.

Main Result (2): BOBW for Problems with a Minimax Regret of $\Theta(T^{2/3})$

FTRL with α-Tsallis entropy $H_\alpha(p) = \frac{1}{\alpha}\sum_{i=1}^{k}(p_i^\alpha - p_i)$:

$$q_t = \arg\min_{p \in \mathcal{P}_k} \left\{ \sum_{s=1}^{t-1} \langle \widehat{\ell}_t, p \rangle + \beta_t(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p)) \right\}, \quad \alpha \in (0,1)\,,\ \bar{\alpha} = 1 - \alpha\,,$$

Theorem (informal)

The FTRL with **SPB-matching** β_t *for* z_t *and* h_t *satisfying* **a condition** *achieves*

$$\mathsf{R}_T \lesssim \begin{cases} (z_{\max} h_1)^{1/3} T^{2/3} + \sqrt{u_{\max} h_1 T} & \textit{adversarial} \\ \frac{\rho}{\Delta^2}\log(T\Delta^2) + \left(\frac{C^2\rho}{\Delta^2}\log\left(\frac{T\Delta}{C}\right)\right)^{1/3} & \textit{corrupted stochastic} \\ \frac{\rho}{\Delta^2}\log(T) & \textit{stochastic} \end{cases}$$

for a problem-dependent constant $\rho > 0$*. (*Δ*: minimum suboptimality gap)*

The condition can be satisfied in several problems with a minimax regret of $\Theta(T^{2/3})$ ↓

Case Study (1): Partial Monitoring with Global Observability

Partial monitoring: a general sequential decision-making problem with **limited feedback**

Consider PM game $\mathbf{G} = (\mathcal{L}, \Phi)$ with k-actions and d-outcomes
for loss matrix $\mathcal{L} \in [0,1]^{k \times d}$, feedback matrix $\Phi \in \Sigma^{k \times d}$ (Σ: the set of feedback symbols)

Learner observes $\mathcal{L}$ and Φ
for $t = 1, 2, \ldots, T$ **do**
- Environment determines an outcome $x_t \in \{1, \ldots, d\}$
- Learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing x_t
- Learner then suffers an **unobserved** loss $\mathcal{L}_{A_t, x_t}$ and observes a symbol $\Phi_{A_t, x_t} \in \Sigma$

Goal: Minimize the regret

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^T \mathcal{L}_{A_t, x_t} - \sum_{t=1}^T \mathcal{L}_{a^*, x_t}\right] \quad \text{for} \quad a^* = \arg\min_{a \in \{1,\ldots,k\}} \mathbb{E}\left[\sum_{t=1}^T \mathcal{L}_{a, x_t}\right]$$

There exists a class called **globally observable** games with minimax regret of $\Theta(T^{2/3})$,
which is characterized by the relationship between $\mathcal{L}$ and Φ.

Regret bounds for globally observable partial monitoring.
T: the number of rounds, k: the number of actions, Δ: minimum suboptimality gap,
$c_{\mathcal{G}}$: a game-dependent constant, MS-type: an improved bound by [MS21]

References	Stochastic	Adversarial	Corrupted
[KHN15]	$D \log T$	–	–
[LS20]	–	$(c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}$	–
[TIH23a]	$\frac{c_{\mathcal{G}}^2 \log T \log(kT)}{\Delta^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log T \log(kT))^{1/3}$	✓
[TIH24]	$\frac{c_{\mathcal{G}}^2 k \log T}{\Delta^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log T)^{1/3}$	✓
Ours	$\frac{c_{\mathcal{G}}^2 \log k \log T}{\Delta^2}$	$(c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}$	✓ (MS-type)

Case Study (2): Graph Bandits with Weak Observability

Graph bandits: interpolation and extrapolation of expert problems and multi-armed bandits

Learner observes a directed graph $G = (V, E)$ for $V = \{1, \ldots, k\}$
for $t = 1, 2, \ldots, T$ **do**
  Environment determines a loss vector $\ell_t \colon V \to \mathbb{R}$
  Learner selects an action $A_t \in \mathcal{A}$ based on past observations without knowing ℓ_t
  Learner then suffers a loss $\ell(A_t)$ and observes a set of losses $\{\ell_t(a) \colon (A_t, a) \in E\}$

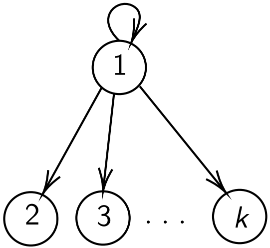


Figure: a weakly observable graph

Goal: Minimize the regret R_T

There exists a class called **weakly observable** graphs with minimax regret of $\Theta(T^{2/3})$, characterized by the structure of feedback graph G.

Regret bounds for weakly observable graph bandits with no self-loops.
T: the number of rounds, k: the number of actions, Δ: minimum suboptimality gap,
δ: domination number (satisfying $\delta^* \leq \delta$), δ^*: fractional domination number (satisfying $\delta^* \leq \delta$)

References	Stochastic	Adversarial	Corrupted
[Alo+15]	–	$(\delta \log k)^{1/3} T^{2/3}$	–
[Che+21]	–	$(\delta^* \log k)^{1/3} T^{2/3}$	–
[ITH22]	$\frac{\delta \log T \log(kT)}{\Delta^2}$	$(\delta \log T \log(kT))^{1/3} T^{2/3}$	✓
[DWZ23][a]	$\frac{\delta \log k \log T}{\Delta^2}$	$(\delta \log k)^{1/3} T^{2/3}$	✓
Ours	$\frac{\delta^* \log k \log T}{\Delta^2}$	$(\delta^* \log k)^{1/3} T^{2/3}$	✓(MS-type)

[a] A hierarchical reduction-based approach, rather than a direct FTRL method, discarding past observations as doubling-trick. The variable δ can be replaced with δ^*.

Case Study (3): MAB with Paid Observations

MAB with paid observations: a variant of the multi-armed bandits (MAB) problem

for $t = 1, 2, \ldots, T$ **do**

- Environment determines a loss vector $\ell_t \colon [k] \to \mathbb{R}$
- Learner observes cost vector $c_t \in \mathbb{R}^k_{\geq 0}$
- Learner selects an action $A_t \in [k]$ and chooses a set of actions $S_t \subseteq [k]$, for which we can observe losses.
- Learner then suffers a loss $\ell(A_t) + \sum_{i \in S_t} c_{ti}$ and observes a set of losses $\{\ell_{ti} \colon i \in S_t\}$.

Goal: Minimize the sum of the standard regret and the observation costs R_T given by

$$\mathrm{R}_T^{\mathrm{cost}} = \mathrm{R}_T + \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in S_t} c_{ti}\right].$$

The minimax regret of this setting is $\Theta(T^{2/3})$.

Upper bounds on $\mathrm{R}_T^{\mathrm{cost}}$ for MAB with paid observations.
T: the number of rounds, k: the number of actions, Δ: minimum suboptimality gap,
c: paid cost for observing a loss of actions

References	Stochastic	Adversarial	Corrupted
[Sel+14]	–	$(ck \log k)^{1/3} T^{2/3} + \sqrt{T \log k}$	–
Ours	$\frac{\max\{c, 1\} k \log k \log T}{\Delta^2}$	$(ck \log k)^{1/3} T^{2/3} + \sqrt{T \log k}$	✓ (MS-type)

Summary

- Investigated online learning with a minimax regret of $\Theta(T^{2/3})$
- Established a simple and adaptive learning rate framework called **stability–penalty–bias matching (SPB-matching)**
- FTRL with SPB-matching and Tsallis entropy regularization improves the existing BOBW regret bounds based on FTRL
  for partial monitoring with global observability, graph bandits with weak observability, and MAB with paid observations
- Future work: investigate if we can apply SPB-matching to other problems with a minimax regret of $\Theta(T^{2/3})$, such as bandits with switching costs [Dek+14] and dueling bandits with Borda winner [SKM21]

[AHR12] Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. "Interior-Point Methods for Full-Information and Bandit Online Learning". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4164–4175.

[Alo+15] Noga Alon et al. "Online Learning with Feedback Graphs: Beyond Bandits". In: *Proceedings of The 28th Conference on Learning Theory*. Vol. 40. 2015, pp. 23–35.

[Aue+02] Peter Auer et al. "The Nonstochastic Multiarmed Bandit Problem". In: *SIAM Journal on Computing* 32.1 (2002), pp. 48–77.

[BPS11] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. "Minimax Regret of Finite Partial-Monitoring Games in Stochastic Environments". In: *Proceedings of the 24th Annual Conference on Learning Theory*. Vol. 19. 2011, pp. 133–154.

[Che+21] Houshuang Chen et al. "Understanding Bandits with Graph Feedback". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 24659–24669.

[CT17] Sougata Chaudhuri and Ambuj Tewari. "Online Learning to Rank with Top-k Feedback". In: *Journal of Machine Learning Research* 18.103 (2017), pp. 1–50.

[Dek+14] Ofer Dekel et al. "Bandits with switching costs: $T^{2/3}$ regret". In: *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*. 2014, pp. 459–467.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Journal of Machine Learning Research* 12.61 (2011), pp. 2121–2159.

[DWZ23] Chris Dann, Chen-Yu Wei, and Julian Zimmert. "A Blackbox Approach to Best of Both Worlds in Bandits and Beyond". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. 2023, pp. 5503–5570.

[ITH22] Shinji Ito, Taira Tsuchiya, and Junya Honda. "Nearly Optimal Best-of-Both-Worlds Algorithms for Online Learning with Feedback Graphs". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 28631–28643.

[ITH24] Shinji Ito, Taira Tsuchiya, and Junya Honda. "Adaptive Learning Rate for Follow-the-Regularized-Leader: Competitive Analysis and Best-of-Both-Worlds". In: *arXiv preprint arXiv:2403.00715* (2024).

[JLL23] Tiancheng Jin, Junyan Liu, and Haipeng Luo. "Improved Best-of-Both-Worlds Guarantees for Multi-Armed Bandits: FTRL with General Regularizers and Multiple Optimal Arms". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 30918–30978.

[KHN15] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. "Regret Lower Bound and Optimal Algorithm in Finite Stochastic Partial Monitoring". In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015, pp. 1792–1800.

[LS19] Tor Lattimore and Csaba Szepesvári. "Cleaning up the neighborhood: A full classification for adversarial partial monitoring". In: *Proceedings of the 30th International Conference on Algorithmic Learning Theory*. Vol. 98. 2019, pp. 529–556.

[LS20] Tor Lattimore and Csaba Szepesvári. "Exploration by Optimisation in Partial Monitoring". In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. 2020, pp. 2488–2515.

[MS10] H. Brendan McMahan and Matthew J. Streeter. "Adaptive Bound Optimization for Online Convex Optimization". In: *The 23rd Conference on Learning Theory*. 2010, pp. 244–256.

[MS21] Saeed Masoudian and Yevgeny Seldin. "Improved Analysis of the Tsallis-INF Algorithm in Stochastically Constrained Adversarial Bandits and Stochastic Bandits with Adversarial Corruptions". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. 2021, pp. 3330–3350.

[Sel+14] Yevgeny Seldin et al. "Prediction with Limited Advice and Multiarmed Bandits with Paid Observations". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. 1. 2014, pp. 280–287.

[SKM21] Aadirupa Saha, Tomer Koren, and Yishay Mansour. "Adversarial Dueling Bandits". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. 2021, pp. 9235–9244.

[TIH23a] Taira Tsuchiya, Shinji Ito, and Junya Honda. "Best-of-Both-Worlds Algorithms for Partial Monitoring". In: *Proceedings of The 34th International Conference on Algorithmic Learning Theory*. 2023, pp. 1484–1515.

[TIH23b] Taira Tsuchiya, Shinji Ito, and Junya Honda. "Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.

[TIH24] Taira Tsuchiya, Shinji Ito, and Junya Honda. "Exploration by Optimization with Hybrid Regularizers: Logarithmic Regret with Adversarial Robustness in Partial Monitoring". In: *arXiv preprint arXiv:2402.08321* (2024).