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

Taira Tsuchiya [1,2] Shinji Ito [1,2]

[1]The University of Tokyo [2]RIKEN

NEURAL INFORMATION PROCESSING SYSTEMS

General Online Learning Framework

Given a finite action set $\mathcal{A} = [k] \coloneqq \{1, \dots, k\}$ 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 $o_t \in \mathcal{O}$

Goal: Minimize the **regret** $\mathsf{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_t(A_t) - \sum_{t=1}^T \ell_t(a^*)\right]$ for $a^* \in \arg\min_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^T \ell_t(a)\right]$

- 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)

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 $\mathcal{A} = [k]$, $\beta_t > 0$: learning rate at round t)

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

Stability–Penalty Decomposition

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

- **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.

For example when using FTRL with the negative Shannon entropy $-H(\cdot)$ (Exp3) in MAB (Auer et al., 2002b), $h_t = H(q_t)$ or $h_t = \log k$ and $z_t = \mathbb{E}\left[\|\widehat{\ell}_t\|^2_{(\nabla^2 \psi(q_t))^{-1}}\right]$.

Adaptive Learning Rate in the Literature

- Use **empirical** stability $(z_s)_{s=1}^{t-1}$ and **worst-case** penalty terms $h_{\max} \geq \max_t h_t$
  e.g., AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011), first-order algorithms (Abernethy et al., 2012)
  $$1/\beta_t = \sqrt{\text{const}/(\text{const} + \textstyle\sum_{s=1}^{t-1} z_s)}$$
- 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.*, (Ito et al., 2022a; Tsuchiya et al., 2023a)
  $$\beta_1 > 0\,, \quad \beta_{t+1} = \beta_t + \text{const}/\sqrt{\text{const} + \textstyle\sum_{s=1}^{t-1} h_{s+1}}$$
- **Use both empirical stability and penalty (Tsuchiya et al., 2023c; Jin et al., 2023; Ito et al., 2024) for simultaneous data-dependent bounds and best-of-both-worlds bounds or for Tsallis-entropy regularizer**

Almost all adaptive learning rates are for problems with a minimax regret of $\Theta(\sqrt{T})$
↔ 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 $\Theta(T^{2/3})$:
e.g., partial monitoring with global observability (Bartók et al., 2011; Lattimore and Szepesvári, 2019a), graph bandits with weak observability (Alon et al., 2015), bandits with paid observations (Seldin et al., 2014), dueling bandits (Saha et al., 2021), online ranking (Chaudhuri and Tewari, 2017)

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?

Stability–Penalty–Bias Decomposition

Common to use forced exploration for FTRL in online learning with the minimax regret of $\Theta(T^{2/3})$:

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

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

Stability–penalty–bias decomposition

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

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

Stability–Penalty–Bias Matching Learning Rate

Step 1: Choose Exploration Rate γ_t

A naive approach: choose $\gamma_t = \sqrt{z_t/\beta_t}$ so that the stability term and the bias term match.
→ this choice does not work well because to obtain a regret bound of $(\star)$, 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 $\gamma_t = \gamma'_t + u_t/\beta_t$ for $u_t > 0$

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

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

Step 2: Choose Learning Rate β_t

Idea: choose β_t so that stability + bias terms and penalty term match! (inspired by Ito et al. (2024))

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

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): Regret Bound by SPB-matching

Theorem

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

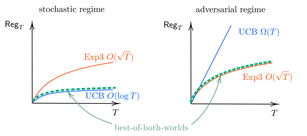
$$\begin{aligned}F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) \lesssim \min &\left\{ \left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{t+1} \log(\epsilon T)} \right)^{\frac{2}{3}} + \left(\sqrt{z_{\max} \widehat{h}_{\max}} \Big/ \epsilon \right)^{\frac{2}{3}}, \left(\sum_{t=1}^T \sqrt{z_t \widehat{h}_{\max}} \right)^{\frac{2}{3}}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

- 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 Tsuchiya et al. (2023c); Jin et al. (2023); Ito et al. (2024)

Application: Best-of-Both-Worlds Algorithms

Best-of-Both-Worlds (BOBW) algorithm achieves a near-optimal regret for stochastic and adversarial environments **simultaneously**

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

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

Theorem

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

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

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 Studies for Problems with a Minimax Regret of $\Theta(T^{2/3})$

Partial monitoring with global observability: a general sequential decision-making problem with feedback symbols ($c_{\mathcal{G}}$: a game-dependent constant)

(Ours) $\mathsf{R}_T \lesssim c_{\mathcal{G}}^2 \log k \log T/\Delta^2$ stochastic env $\mathsf{R}_T \lesssim (c_{\mathcal{G}} T)^{2/3} (\log k)^{1/3}$ adversarial env

Graph bandits with weak observability: interpolation and extrapolation of expert problems and multi-armed bandits ($\delta^* \leq k$: fractional domination number)

(Ours) $\mathsf{R}_T \lesssim \delta^* \log k \log T/\Delta^2$ stochastic env $\mathsf{R}_T \lesssim (\delta^* \log k)^{1/3} T^{2/3}$ adversarial env

MAB with paid observations (new BOBW result): you need to pay a cost of c to observe a loss (c: paid cost for observations)

(Ours) $\mathsf{R}_T \lesssim \max\{c, 1\} k \log k \log T/\Delta^2$ stochastic $\mathsf{R}_T \lesssim (ck \log k)^{1/3} T^{2/3} + \sqrt{T \log k}$ adversarial