LC-Tsallis-INF: Generalized Best-of-Both-Worlds Linear Contextual Bandits

Masahiro Kato (Mizuho-DL Financial Technology Co., Ltd.) and Shinji Ito (The University of Tokyo and RIKEN AIP)

Mizuho-DL Financial Technology

Multi-Armed Bandit Problem

- K arms: $1, 2, \dots, K$. T rounds: $1,2, \dots, T$. Context $X_t \in \mathcal{X}$.
  - Arm a yields a loss $\ell_t(a, X_t)$ in round t.
- **Decision maker:** In each round $t = 1,2, \dots, T$.
  - Observes a context X_t.
  - Chooses arm $A_t \in [K]$ based on X_t and past observations.
  - Incurs a loss $\ell_t(A_t, X_t)$.
- **Goal**: minimize the cumulative loss $\sum_{t=1}^{T} \ell_t(A_t, X_t)$.
- Measure the performance by using the **regret**:

$$R_T := \max_{\rho \in \mathcal{R}} \mathbb{E}\left[\sum_{t=1}^{T} (\ell_t(A_t, X_t) - \ell_t(\rho(X_t), X_t))\right].$$

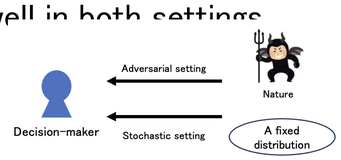
Best-of-Both-World Algorithms

- Stochastic bandits.
  - Distribution of the losses is fixed.
  - In many cases, $O(\log(T))$ regret is optimal.
- Adversarial bandits.
  - Losses are chosen to maximize the regret.
  - In many cases, $O(\sqrt{T})$ regret is optimal.

✓ Algorithms designed for one of the two settings tend to be suboptimal when applied to the other.

➢ **Best-of-Both-Worlds algorithms.**

- Perform well in both settings



Linear Contextual Bandits

- Loss $\ell_t(a, X_t)$ follows the linear models:

$$\ell_t(a, X_t) = \langle \phi(a, X_t), \theta_t \rangle + \varepsilon_t(a).$$

- $\phi(a, X_t) \in \mathbb{R}^d$ is a feature map.
- θ_t is a coefficient of a linear model.
- $\varepsilon_t(a)$ is a bounded error term with mean zero.

✓ Stochastic setting:

- θ_t is fixed and does not change across rounds: $\theta_1 = \theta_2 = \cdots = \theta_T$.

✓ Adversarial setting.

- θ_t is chosen to maximize the regret.

Our Contributions

- We propose the **LC-Tsallis-INF algorithm**.
- The regret is given as follows:
  - Stochastic setting: $O(\log(T))$ regret.
  - Stochastic setting with the margin condition: $O\left(\log(T)^{\frac{1+\beta}{2+\beta}} T^{\frac{1}{2+\beta}}\right)$ regret, where $\beta \in (0, +\infty]$.
  - Adversarial setting: $O(\sqrt{T})$ regret.
- Investigation of arm-dependent and arm-independent feature settings.

✓ Existing studies:

- $O(\log(T)^2)$ regret in stochastic bandits.
  - $O(\log(T))$ regret is possible if we incur a high computational cost.
- $\sqrt{T}$ regret in adversarial bandits.
- No margin conditions (only consider the case with $\beta = \infty$).

LC-Tsallis-INF Algorithms

- Estimator of the regression coefficient θ_t:

$$\hat{\theta}_t := \mathbb{E}[\phi(A_t, X_t)^\top \phi(A_t, X_t)]^{-1} \phi(A_t, X_t) \ell_t(A_t, X_t).$$

  - Computation of $\mathbb{E}[\phi(a, X_t)^\top \phi(a, X_t)]^{-1}$ depends on the assumption about the knowledge of the context distribution.
- Estimator of the loss: $\hat{\ell}_t(a, X_t) = \langle \phi(a, X_t), \hat{\theta}_t \rangle$
- Select arm a with the following probability:

$$\pi_t(a, X_t) = (1 - \gamma_t) q_t(a, X_t) + \gamma_t e^*(a, X_t).$$

  - $e^*(a, X_t)$ is an exploration policy, given by the decision maker in advance of the bandit trial.
  - γ_t is a parameter tuned in this algorithm.
- $q_t(a, X_t)$ is obtained by the Follow-The-Regularized Leader (FTRL) with the α-Tsallis entropy.

$$q_t(x) := \arg\min_{q \in \Delta_K} \left\{ \sum_{s=1}^{t} \langle \hat{\ell}_t(x), q \rangle + \frac{1}{\eta_{t-1}} \psi(q) \right\}.$$

  - $\hat{\ell}_t(x) = \left(\hat{\ell}_t(1, x), \hat{\ell}_t(2, x), \cdots, \hat{\ell}_t(K, x)\right)^\top$.
  - Δ_K is the $(K-1)$-dimensional probability simplex.
  - $\psi(q) := \frac{1}{\alpha}\left(1 - \sum_{a \in [K]} q(a)^\alpha\right)$ is the the α-Tsallis entropy. In our analysis, we use $\alpha = 1/2$.
  - η_t is a parameter tuned in this algorithm.

- We analyze the performances with various situations:
  - Adversarial regime with a self-bounding constraint
  - Margin condition with parameter $\beta \in (0, +\infty]$.
  - Arm-dependent and arm-independent feature maps.