Further Adaptive Best-of-Both-Worlds Algorithm for Combinatorial Semi-Bandits

AISTATS2023

Taira Tsuchiya [1,2] • Shinji Ito [3] • Junya Honda [1,2] (1. Kyoto University, 2. RIKEN AIP, 3. NEC Corporation)

Bandit problems and Best-of-Both-Worlds Algorithms

Combinatorial Semi-bandits

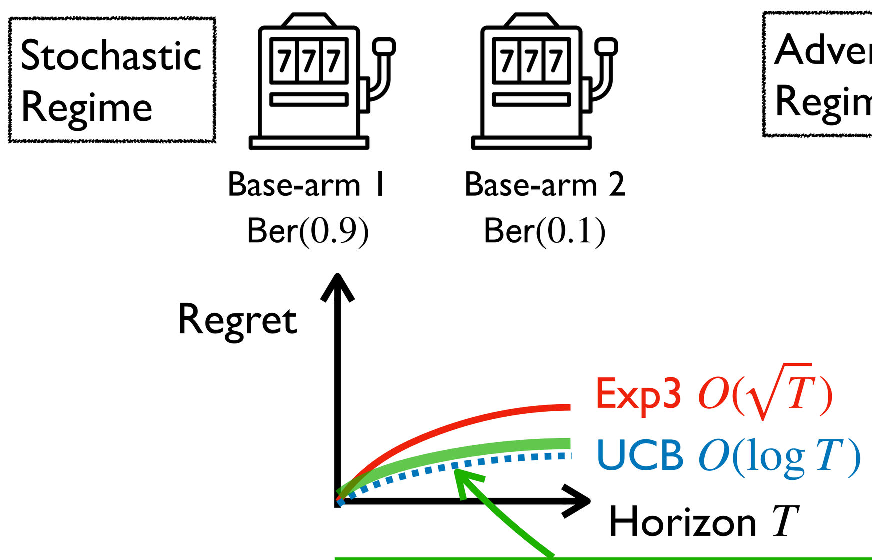
Given action set $\mathscr{A} \subset \{0,1\}^d$
Adversary selects loss vectors $\ell_1, \dots \ell_T \in [0,1]^d$
At each round $t = 1, \dots, T$:
1. Learner selects $a(t) \in \mathscr{A}$
2. Learner incurs a loss $\langle \ell(t), a(t) \rangle$ and observes $\ell_i(t)$ for $i \in [d]$ such that $a_i(t) = 1$

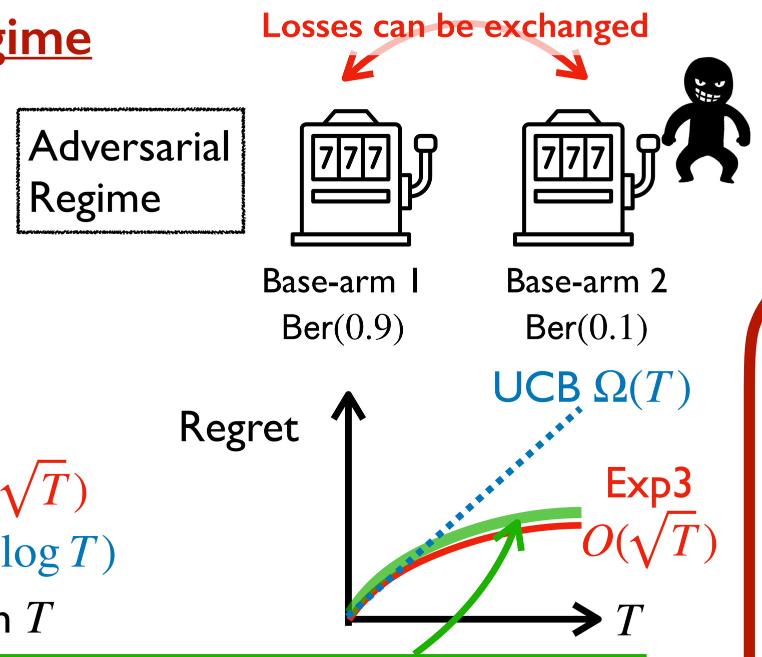
Goal: minimize regret R_T defined as

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \langle \ell(t), a(t)\rangle - \sum_{t=1}^T \langle \ell(t), a^* \rangle\right]$$

for $\quad a^* = \arg\min_{a \in \mathscr{A}} \sum_{t=1}^T \langle \ell(t), a \rangle$

Stochastic and Adversarial Regime





What we hope : Achieving optimality for both stoc. and adv. regimes w/o knowing underlying regime = Best-of-Both-Worlds (BOBW)

(Short) Research Background

There are many algorithms for BOBW algorithms mainly for multi-armed bandits [Bubeck & Slivkins 12, Zimmert & Seldin 21, etc.]

Research Question

Q. Can we improve existing BOBW algorithms for semi-bandits by exploiting problem structures more?

A. Yes, can show regret bounds with tight suboptimality gap and variance-dependent bounds!

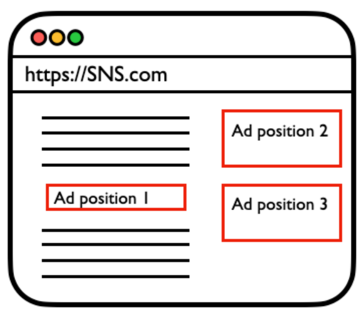
Background and Motivation

Why is distributional information useful?

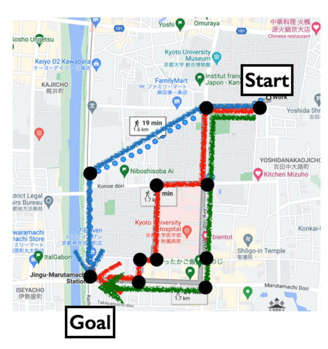
Practically, many problems have small variances nature

- Recommender system: very small CTRs
- Shortest path problem: The required time does not vary much

Variances of each base-arm is very small
→ Algorithms with variance-dependent bound should perform well



Ads allocation



Shortest-path problem

Q. Can we establish BOBW algorithms with variance-dependent bounds?

Theoretically, "Best" in BOBW literature is not the best; when $\ell_t \sim \mathrm{Ber}(\cdot)$

"Best" bounds in BOBW literatures $\quad$ The achievable bounds

$$R_T = O\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right) \gg R_T = \Omega\left(\sum_{i \neq i^*} \frac{\log T}{\mathrm{KL}(\mu_i, \mu_{i^*})}\right) \qquad \mu_i = \mathbb{E}[\ell_i], \ \Delta_i = \mu_i - \mu_{i^*}$$

[Lai & Robbins 85]

Q. Can we fill these gaps achieving BOBW simultaneously?

Better dependency on sub-optimality gap

Existing BOBW algorithms for combinatorial semi-bandits have a form of

$$R_T = O\left(\frac{dm \log T}{\Delta}\right)$$

for $\Delta = \min\{\langle \mu, a - a^* \rangle : a \in \mathscr{A} \backslash \{a^*\}\}$. [Zimmert, Luo & Wei 19, Ito 21]

Q. Can we achieve obtain high-resolution bounds with arm-wise suboptimality gaps $\Delta_{i,\min} = \min\{\langle \mu, a - a^* \rangle : a \in \mathscr{A} \backslash \{a^*\}, a_i = 1\}$ **?**

Proposed Algorithm

Optimistic Follow-the-Regularized-Leader (OFTRL)

- Let $\mathscr{X}$ be a convex full of $\mathscr{A}$
- OFTRL selects $x_t \in \mathscr{X}$ by minimizing "predicted losses for next round + observations so far + regularizer"

"optimistic" prediction of $\ell(t)$ $\quad$ estimated losses $\quad$ convex regularizer

$$x_t \in \arg\min_{x \in \mathscr{X}} \langle m(t) + \sum_{s=1}^{t-1} \hat{\ell}_s, x \rangle + \psi_t(p)$$

$\hat{\ell}_s \in \mathbb{R}^d$: unbiased estimator of ℓ_s

- OFTRL with $m(t) = 0$ is FTRL
- Most BOBW algorithms rely on (O)FTRL
- Choose a_t so that $\mathbb{E}[a_t \mid x_t] = x_t$

Algorithms and analysis are similar to [Ito, Tsuchiya & Honda 22]

Proposed Algorithms

We use OFTRL with following parameters:

Optimistic prediction of $\ell(t)$ $\quad$ $m(t)$ converge a.s. to μ

Method 1. least square (LS) estimation

$$m_i(t) = \frac{1}{1 + N_i(t-1)}\left(\frac{1}{2} + \sum_{s=1}^{t-1} a_i(s)\,\ell_i(s)\right)$$

→ smaller leading constant

Method 2. gradient descent (GD) estimation

$$m_i(1) = \frac{1}{2} \quad \text{and} \quad m_i(t+1) = \begin{cases} (1-\eta)m_i(t) + \eta \ell_i(t) & \text{if } i \in I(t) \\ m_i(t) & \text{otherwise} \end{cases}$$

Loss estimation: reduced-variance estimator

$$\hat{\ell}_i(t) = m_i(t) + \frac{a_i(t)}{x_i(t)}(\ell_i(t) - m_i(t))$$

$a_i(t)$: i-the element of $a(t)$

Regularizer: base-arm-wise log-barrier + Complement of Shannon entropy

Log-barrier $\quad$ Complement Shannon entropy

$$\psi_t(x) = \sum_{i=1}^d \beta_i(t)\phi(x_i) \quad \text{with} \quad \phi(z) = z - 1 - \log z + \log(T) \cdot (z + (1-z)\log(1-z))$$

with **learning rate** defined by $\quad$ $\to \sigma_i^2$ as $t \to \infty$

$$\beta_i(t) = \sqrt{(1+\epsilon)^2 + \frac{1}{\log T}\sum_{s=1}^{t-1} \alpha_i(s)}\ , \alpha_i(t) = a_i(t)(\ell_i(t) - m_i(t))^2 \min\left\{1, \frac{2(1 - x_i(t))}{x_i(t)^2 \log T}\right\}$$

→ variance-dependent bound

Regret Upper Bounds and Regret Analysis

Regret Upper Bounds

$\mathscr{R}$: bound for the stochastic regime of each algorithm

Reference	Stochastic	Adversarial	Stocastic w/ adv. corruptions
Audibert+ 14	–	$O(\sqrt{dmT})$	–
Kveton+ 15	$534 \sum_{i \in J^*} \frac{m \log T}{\Delta_{i,\min}}$	–	–
Zimmert+ 19	$O\left(\frac{dm \log T}{\Delta}\right)$	$O(\sqrt{dmT})$	$O(\mathscr{R} + \sqrt{C\mathscr{R}})$
Ito 21	$O\left(\frac{dm \log T}{\Delta}\right)$	$O(\sqrt{d \min\{L^*, Q_2, V_1\} \log T})$	$O(\mathscr{R} + \sqrt{C\mathscr{R}})$
Proposed (LS)	$\sum_{i \in J^*} \max\left\{\frac{4w\sigma_i^2}{\Delta_{i,\min}}, 2\right\} \log T$	$O(\sqrt{d \min\{L^*, Q_2\} \log T})$	$O(\mathscr{R} + \sqrt{C\mathscr{R}})$
Proposed (GD)	$\sum_{i \in J^*} \max\left\{\frac{8w\sigma_i^2}{\Delta_{i,\min}}, 4\right\} \log T$	$O(\sqrt{d \min\{L^*, Q_2, V_1\} \log T})$	$O(\mathscr{R} + \sqrt{C\mathscr{R}})$

Much better leading constant with tighter suboptimality gap $\Delta_{i,\min}$ and variance-dependency

Sketch of Proof

$m = \max_{a \in \mathscr{A}} \|a\|_1$
$I^* = \{i \in [d] : a_i^* = 1\}$
$J^* = \{i \in [d] : a_i^* = 0\}$
$v, w \le m$: action-set-dep const
$L^* = \min_{a \in \mathscr{A}} \mathbb{E}[\sum_{t=1}^T \langle \ell(t), a \rangle]$
$Q_2 = \mathbb{E}[\sum_{t=1}^T \|\ell(t) - \bar{\ell}\|^2]$
$(\bar{\ell} = T^{-1}\mathbb{E}[\sum_{t=1}^T \ell(t)])$
$V_1 = \mathbb{E}[\sum_{t=1}^{T-1} \|\ell(t) - \ell(t+1)\|_1]$
$C = \mathbb{E}[\sum_{t=1}^T \|\ell(t) - \ell'(t)\|_\infty]$

Lemma. In the stochastic regime w/ adversarial corruptions, it holds that

$$R_T \ge \mathbb{E}\left[\sum_{t=1}^T \left(\frac{1}{v(\mathscr{A})}\sum_{i \in I^*} \Delta'_{i,\min}(1 - a_i(t)) + \frac{1}{w(\mathscr{A})}\sum_{i \in J^*} \Delta_{i,\min} a_i(t)\right)\right] - 2Cm$$

The regret can be bounded by

$$R_T \le O\left(\sum_{i \in J^*} \sqrt{\beta_0^2 + \frac{\sigma_i^2 P_i}{\log T}} + \sum_{i \in I^*} \sqrt{\frac{Q_i}{(\log T)^{3/2}}}\right)$$

for $P_i = \mathbb{E}[\sum_{t=1}^T x_i(t)]$ and $Q_i = \mathbb{E}[\sum_{t=1}^T (1 - x_i(t))]$

Combining the upper and lower bounds, and taking the worst case w.r.t. (P_i) and (Q_i),

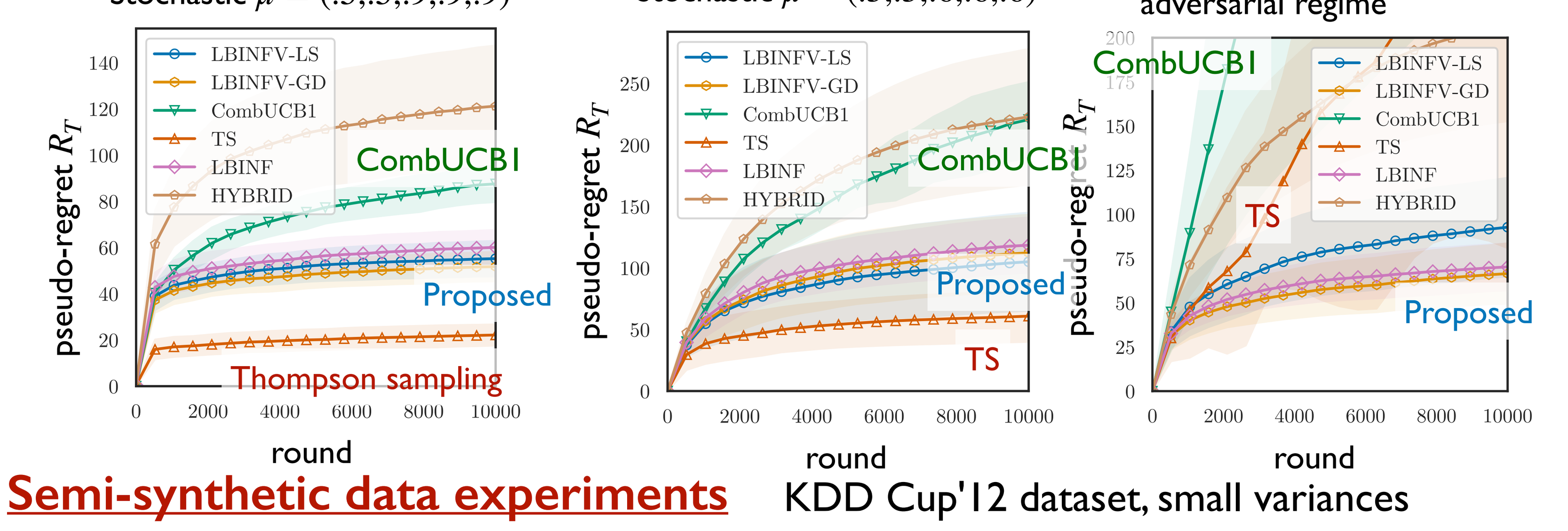
$$\frac{R_T}{\log T} \le 2\frac{R_T}{\log T} - \frac{R_T}{\log T} \le O\left(\sum_{i \in J^*} \frac{w(\mathscr{A})\sigma_i^2}{\Delta_{i,\min}} + |I^*| \frac{1}{\sqrt{\log T}} \frac{v(\mathscr{A})}{\Delta'_{i,\min}}\right)$$

Experiments

- m-set semi-bandits with with Bernoulli base-arms
- Compare proposed algorithms (LBINFV-LS and LBINFV-GD) with CombUCB1 [Kveton, Wen, Ashkan & Szepesvári 15] ,Thompson sampling [Wang & Chen 18] , HYBRID [Zimmert, Luo & Wei18] , LBINF [Ito 21]

Synthetic Data Experiments

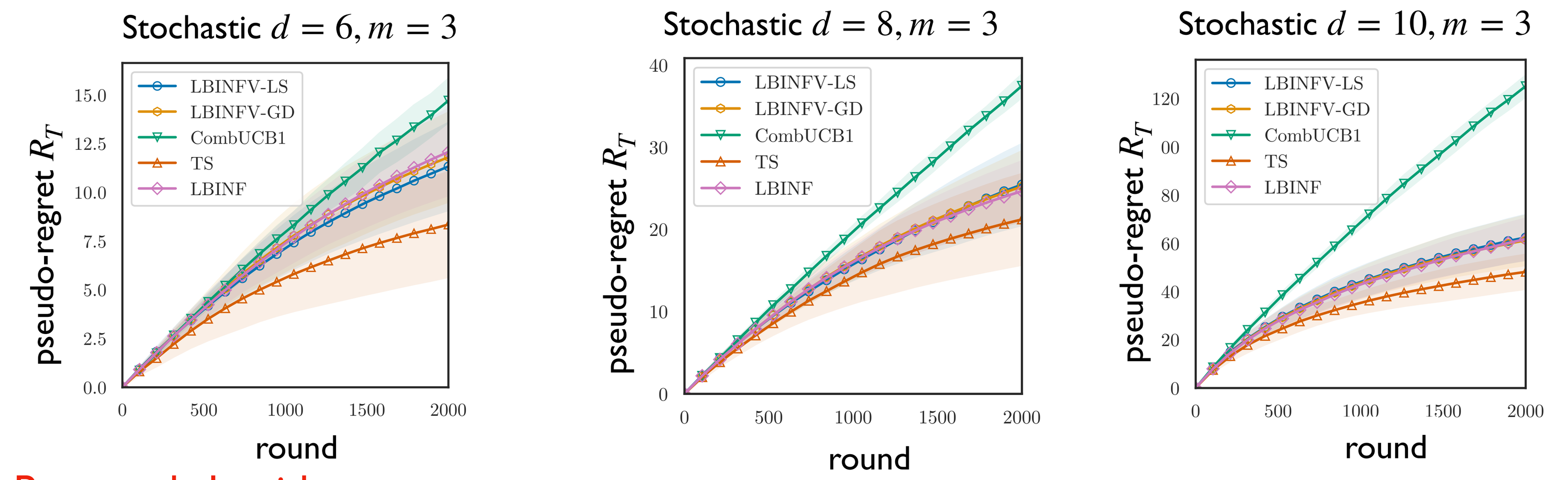
Stochastic $\mu = (.5,.5,.9,.9,.9)$ $\quad$ Stochastic $\mu = (.5,.5,.6,.6,.6)$ $\quad$ Stochastically constrained adversarial regime



Semi-synthetic data experiments

KDD Cup'12 dataset, small variances

Stochastic $d = 6, m = 3$ $\quad$ Stochastic $d = 8, m = 3$ $\quad$ Stochastic $d = 10, m = 3$



Proposed algorithms

- perform better than that of existing BOBW algorithms in small-variance stoc. regime
- perform best in slightly adversarial environments where Thompson Sampling fails

S. Bubeck & A. Slivkins. "The best of both worlds: Stochastic and adversarial bandits." COLT 2012.
J. Zimmert & Y. Seldin. "Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits." JMLR 2021.
S. Ito, T. Tsuchiya, & J. Honda" Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds." COLT 2022.
J.Y. Audibert, S. Bubeck, & G. Lugosi. "Regret in online combinatorial optimization." Mathematics of Operations Research 2014.
B. Kveton, Z. Wen, A. Ashkan, & Cs. Szepesvári. "Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits." AISTATS 2015.
J. Zimmert, H. Luo, & C.Y. Wei. "Beating stochastic and adversarial semi-bandits optimally and simultaneously." ICML 2019.
S. Ito." Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits." NeurIPS 2021.
S. Wang & W. Chen. "Thompson sampling for combinatorial semi-bandits." ICML 2018.