Best-of-Both-Worlds Algorithms for Linear Contextual Bandits

Yuko Kuroki   Alberto Rumi   Taira Tsuchiya

Fabio Vitale   Nicolò Cesa-Bianchi

CENTAI

UNIVERSITÀ DEGLI STUDI DI MILANO

POLITECNICO MILANO

東京大学 THE UNIVERSITY OF TOKYO

RIKEN AIP

AISTATS 2024

Multi-Armed Bandits

- K-arms (actions)
- **Environment** determines the losses to arms $\ell_t = (\ell_t(1), \ell_t(2), \ldots, \ell_t(K)) \in \mathbb{R}^K$ at each time step $t = 1, 2, \ldots, T$ hidden to the learner

At each time step $t = 1, 2, \ldots, T$

- **Learner** selects an action $A_t \in [K]$ and incurs a loss $\ell_t(A_t)$
- **Learner** observes a feedback: Only the loss for chosen arm $\ell_t(A_t)$ is revealed

Multi-Armed Bandits

- K-arms (actions)
- **Environment** determines the losses to arms $\ell_t = (\ell_t(1), \ell_t(2), \ldots, \ell_t(K)) \in \mathbb{R}^K$ at each time step $t = 1, 2, \ldots, T$ hidden to the learner

At each time step $t = 1, 2, \ldots, T$

- **Learner** selects an action $A_t \in [K]$ and incurs a loss $\ell_t(A_t)$
- **Learner** observes a feedback: Only the loss for chosen arm $\ell_t(A_t)$ is revealed

Goal is to minimize the expected **regret** against the best action in hindsight

$$R_T := \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \ell_t(A_t)}_{\text{cumulative losses of the learner}} - \underbrace{\sum_{t=1}^{T} \ell_t(a^*)}_{\text{cumulative losses of the best action}}\right], \quad a^* := \arg\min_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a)\right]$$

Contextual Information in Real Worlds

We often have access to **contextual information** in various domains such as online advertising, medical diagnosis, and finance.

Example: Recommendation Systems

- Context: User's profile or past purchase history
- Goal: Providing personalized product recommendation

Linear Contextual Bandits

At each time step $t = 1, 2, \ldots, T$

- Environment determines a loss vector $\boldsymbol{\theta}_{t,a} \in \mathbb{R}^d$ for each $a \in [K]$
- Environment draws the **context vector** $X_t \sim \mathcal{D}$
- Learner observes current context X_t and chooses action $A_t \in [K]$
- Learner incurs and observes $\ell_t(X_t, A_t)$

Linear Contextual Bandits

At each time step $t = 1, 2, \ldots, T$

- Environment determines a loss vector $\boldsymbol{\theta}_{t,a} \in \mathbb{R}^d$ for each $a \in [K]$
- Environment draws the **context vector** $X_t \sim \mathcal{D}$
- Learner observes current context X_t and chooses action $A_t \in [K]$
- Learner incurs and observes $\ell_t(X_t, A_t)$

Goal is to minimize the expected regret against the optimal policy π^*:

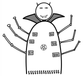
$$R_T := \max_{\pi^* \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \Big(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^*(X_t))\Big)\right],$$

where $\Pi = \{\pi : \mathcal{X} \to [K]\}$ is the set of all deterministic policies and $\mathcal{X} \subseteq \mathbb{R}^d$ is the context space

Adversarial and Stochastic Regimes

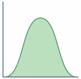
Adversarial Regime

$\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_{t,a} \rangle$, where $\theta_{t,a}$ **for** $a \in [K]$ **is chosen by an adversary**



Stochastic Regime

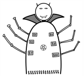
$\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_a \rangle + \varepsilon_t(X_t, a)$, where θ_a **for** $a \in [K]$ **is fixed and unknown**;
$\varepsilon_t(X_t, a)$ is bounded zero-mean noise



Adversarial and Stochastic Regimes

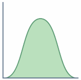
Adversarial Regime

$\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_{t,a} \rangle$, where **$\theta_{t,a}$ for $a \in [K]$ is chosen by an adversary**



Stochastic Regime

$\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_a \rangle + \varepsilon_t(X_t, a)$, where **$\theta_a$ for $a \in [K]$ is fixed and unknown**;
$\varepsilon_t(X_t, a)$ is bounded zero-mean noise



(Corrupted Stochastic Regime)

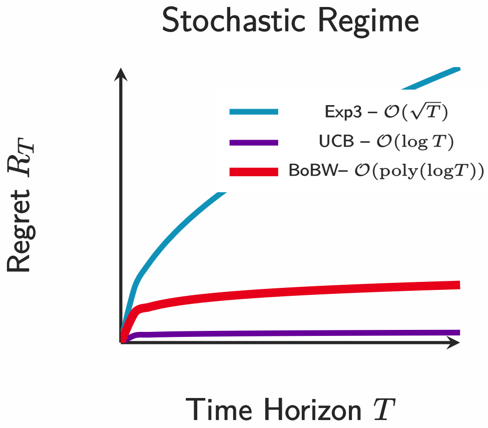
Intermediate regime between adversarial and stochastic one

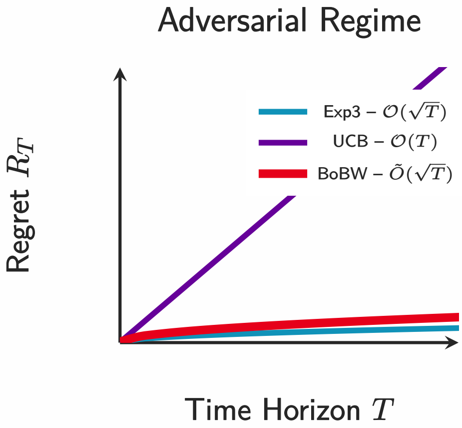
$\ell_t(X_t, a) := \langle X_t, \boldsymbol{\theta}_{t,a} \rangle + \varepsilon_t(X_t, a)$, where $\boldsymbol{\theta}_{t,a}$ satisfies $\sum_{t=1}^{T} \max_{a \in [K]} \|\boldsymbol{\theta}_{t,a} - \boldsymbol{\theta}_a\|_2 \leq C$
for fixed and unknown $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ and unknown **corruption level $C > 0$**

Best-of-Both-Worlds Algorithms

Research Question

Can we establish an algorithm achieving optimal rates in both **stochastic** and **adversarial** regimes **without any prior knowledge of the environment**?





First BoBW Results for Linear Contextual Bandits

Main Contributions (Informal)

	Stochastic	Adversarial
Worst-case	$\mathcal{O}\left(dK \operatorname{poly}\log(T)\right)$	$\tilde{\mathcal{O}}\left(\sqrt{dK\,T}\right)$
Data-dependent	$\mathcal{O}\left((dK)^2 \operatorname{poly}\log(T)\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\Lambda^*}\right)$

Λ^*: data-dependent quantity (cumulative second moment for the losses incurred by the algorithm)

Follow-the-Regularized-Leader (FTRL)

At each round t:

$$p_t(\cdot|X_t) := \arg\min_{r\in\Delta([K])} \left\{ \sum_{s=1}^{t-1} \langle r, \tilde{\ell}_s(X_t)\rangle + \psi_t(r) \right\}$$

estimated cumulative losses
up to previous rounds

$\tilde{\ell}_s(X_t) := (\langle X_t, \tilde{\theta}_{s,1}\rangle, \ldots, \langle X_t, \tilde{\theta}_{s,K}\rangle)$, $\tilde{\theta}_{s,a}$ is the (biased) estimate for $\theta_{s,a}$.

Shannon entropy regularizer: $\psi_t(r) = -\eta_t^{-1} \sum_{a\in[K]} r_a \ln r_a$

Loss Estimation

The estimator of $\theta_{t,a}$ is $\tilde{\theta}_{t,a} := \hat{\Sigma}^{+}_{t,a} X_t \ell_t(X_t, A_t) \mathbb{1}\left[A_t = a\right]$

where $\hat{\Sigma}^{+}_{t,a}$ is the biased estimate of $\Sigma^{-1}_{t,a} := \mathbb{E}[\mathbb{1}\left[A_t = a\right] X_t X_t^\top \mid \mathcal{F}_{t-1}]^{-1}$

Entropy-dependent Learning Rate

Update Rule for Learning Rate (Informal)

$$\eta_{t+1}^{-1} \leftarrow \eta_t^{-1} + \frac{c}{\sqrt{1 + (\log K)^{-1} \Sigma_{s=1}^{t} H(p_s(\cdot|X_s))}}$$

so that we control

adversarial regime: η_t^{-1} would become $\mathcal{O}(\sqrt{t})$

stochastic regime: η_t^{-1} would become $\mathcal{O}(t)$

H: Shannon entropy

FTRL Analysis for i.i.d. Sample of Context $X_0 \sim \mathcal{D}$

(Expected regret for a fixed X_0)

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\eta_{t+1}^{-1} - \eta_t^{-1}\right) H(p_{t+1}(\cdot|X_0))\right] + \mathbb{E}\left[\sum_{t=1}^{T} \eta_t \cdot \text{(variance of loss estimates)}\right]$$

(+prob. dependent constant)

Main Result

Theorem

FTRL with Shannon entropy achieves:

$$R_T^{\text{adv}} = \mathcal{O}\left(\sqrt{T\left(d + \frac{\log T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right) K \log(K) \log(T)}\right) \quad \text{for the adversarial regime}$$

$$R_T^{\text{sto}} = \mathcal{O}\left(\frac{K}{\Delta_{\min}}\left(d + \frac{\log T}{\lambda_{\min}(\boldsymbol{\Sigma})}\right) \log(KT) \log T\right) \quad \text{for the stochastic regime}$$

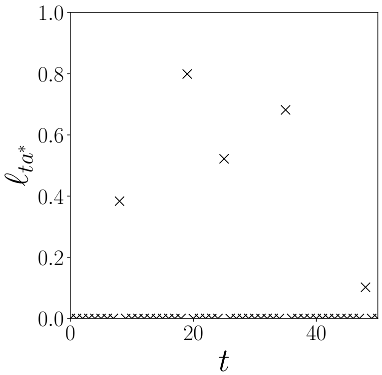
$$R_T^{\text{cor}} = \mathcal{O}\left(R_T^{\text{sto}} + \sqrt{C R_T^{\text{sto}}}\right) \quad \text{for the corrupted stochastic regime}$$

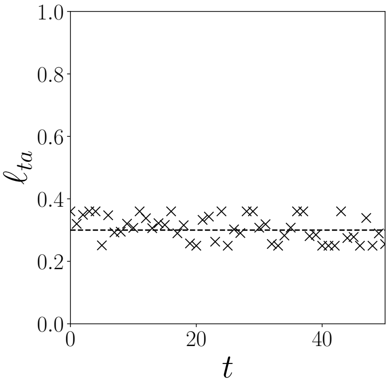
$\Delta_{\min}$: minimum suboptimality gap over the context space

$\lambda_{\min}(\boldsymbol{\Sigma}) :=$ minimum eigenvalue of $\mathbb{E}[XX^\top]$ $\quad C$: corruption level

Our bound recovers the best-known result in the adversarial regime of Neu and Olkhovskaya (2020) and Zierahn et al. (2023) up to log-factors

Benefits of Data-dependent Regret Bounds





$L^* := \mathbb{E}\left[\Sigma_{t=1}^{T} \ell_t(X_t, \pi^*(X_t))\right]$ $(\leq T)$

Cumulative loss of the optimal policy

$\bar{\Lambda} := \mathbb{E}\left[\Sigma_{t=1}^{T} (\ell_t(X_t, A_t) - \langle X_t, \bar{\theta} \rangle)^2\right]$ $(\leq T)$
with average vector $\bar{\theta}$

Cumulative variance of a policy

Overview

Additional Assumptions

- The learner has access to $\Sigma_{t,a}^{-1}$ **to get unbiased estimators**.
- $\mathcal{D}$ is a log-concave distribution **to make the unbiased estimators stable**.

Techniques

Optimistic FTRL
Continuous Exponential Weights

$\longrightarrow$

Black-Box Reduction
Dann, Wei, and Zimmert (2023)

Data-dependent
BoBW

Main Results on Deta-Dependent BoBW

Theorem

	Stochastic	Adversarial	$\sqrt{C}$
Main Theorem	$\mathcal{O}\left(\frac{(dK)^2 \mathrm{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\Lambda^*}\right)$	✓
Corollary	$\mathcal{O}\left(\frac{(dK)^2 \mathrm{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\min\{L^*, \bar{\Lambda}\}}\right)$	✓

Λ^*: cumulative variance of a policy w.r.t. a predictable loss sequence $m_{t,a}$ for $a \in [K]$
L^*: cumulative loss of the best policy
$\bar{\Lambda}$: cumulative second moment for the losses incurred by the algorithm

- Our result has extra $\sqrt{d}$ in the adversarial regime (Olkhovskaya et al. (2023)).
- For a choice of $m_{t,a}$, we use the online optimization method as in Ito et al. (2020).
- This allows a single algorithm to achieve first/second-order bounds simultaneously.

Summary

First BoBW Bounds for Linear Contextual Bandits

	Stochastic	Adversarial	$\sqrt{C}$
Worst-case	$\mathcal{O}\left(\frac{dK \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\mathcal{O}\left(\sqrt{TK\left(d+\log T\right)\log(T)\log(K)}\right)$	✓
Data-dependent	$\mathcal{O}\left(\frac{(dK)^2 \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\Lambda^*}\right)$	✓
First/second order	$\mathcal{O}\left(\frac{(dK)^2 \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\min\{L^*, \bar{\Lambda}\}}\right)$	✓

L^*: cumulative loss of the best action

$\Lambda^*(\bar{\Lambda})$: cumulative second moment for the losses incurred by the algorithm

Summary

First BoBW Bounds for Linear Contextual Bandits

	Stochastic	Adversarial	$\sqrt{C}$
Worst-case	$\mathcal{O}\left(\frac{dK \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\mathcal{O}\left(\sqrt{TK\left(d+\log T\right)\log(T)\log(K)}\right)$	✓
Data-dependent	$\mathcal{O}\left(\frac{(dK)^2 \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\Lambda^*}\right)$	✓
First/second order	$\mathcal{O}\left(\frac{(dK)^2 \operatorname{poly}\log(dKT)}{\Delta_{\min}}\right)$	$\tilde{\mathcal{O}}\left(dK\sqrt{\min\{L^*, \bar{\Lambda}\}}\right)$	✓

L^*: cumulative loss of the best action

$\Lambda^*(\bar{\Lambda})$: cumulative second moment for the losses incurred by the algorithm

Thank you!

Appendix

Loss Estimation

Loss Estimation

The estimator of $\theta_{t,a}$ is $\tilde{\theta}_{t,a} := \hat{\Sigma}^{+}_{t,a} X_t \ell_t(X_t, A_t) \mathbb{1}\left[A_t = a\right], \quad \forall a \in [K]$,

where $\hat{\Sigma}^{+}_{t,a}$ is the biased estimate of $\Sigma^{-1}_{t,a} := \mathbb{E}_t[\mathbb{1}\left[A_t = a\right] X_t X_t^{\top}]^{-1}$.

Estimate $\Sigma^{-1}_{t,a}$

Use **simulator to generate i.i.d. contexts** from distribution $\mathcal{D}$
(Matrix Geometric Resampling with Adaptive Iteration Numer M_t)

Unique Challenges

- We need to deal with a biased estimate of the loss vector
- We require redesigning adaptive learning rates, exploration rates, and iteration numbers of MGR. ($\gamma_t = \alpha_t \cdot \eta_t$, $M_t = \left\lceil \frac{4K}{\gamma_t \lambda_{\min}(\boldsymbol{\Sigma})} \log(t) \right\rceil$ and $\alpha_t = \frac{4K \log(t)}{\lambda_{\min}(\boldsymbol{\Sigma})}$).

Continuous MWU Method

OFTRL: learner has access to a loss predictor $m_{t,a} \in \mathbb{R}^d$ for each action a at round t.

MWU

The learner computes the density $p_t(\cdot|X_t)$ supported on $\Delta([K])$ and based on the continuous exponential weights $w_t(\cdot|X_t)$:

$$w_t(r|X_t) := \exp\left(-\eta_t\left(\sum_{s=1}^{t-1}\langle r, \widehat{\ell}_s(X_t)\rangle + \langle r, \boldsymbol{m}_t(X_t)\rangle\right)\right),$$

$$p_t(r|X_t) := \frac{w_t(r|X_t)}{\int_{\Delta([K])} w_t(y|X_t)\ dy}, \quad \forall r \in \Delta([K]).$$

Conputational Time

The continuous exponential weights incur a high (yet polynomial) sampling cost, resulting in $\mathcal{O}\big((K^5 + \log T)g_{\boldsymbol{\Sigma}_t}\big)$ per round running time, where $g_{\boldsymbol{\Sigma}_t}$ is the time to construct the covariance matrix for each round

Data-Dependent Importance Weighting Stability

Data-Dependent Importance Weighting Stability

Given an adaptive sequence of weights $q_1, q_2, \ldots \in (0, 1]$, the learner observes the feedback in round t with probability q_t. Let upd_t be 1 if observation occurs and 0 otherwise. Then, for any $\tau \in [T]$ and $a^* \in [K]$,
$R_\tau(a^*) = \mathbb{E}\left[\Sigma_{t=1}^{\tau} \ell_t(X_t, A_t) - \ell_t(X_t, a^*)\right]$ is bounded by

$$\mathcal{O}\left(\sqrt{\kappa_1(d, K, T)}\left(\sqrt{\mathbb{E}\left[\Sigma_{t=1}^{\tau} \frac{\mathsf{upd}_t \cdot (\ell_t(X_t, A_t) - \langle X_t, m_{t,A_t}\rangle)^2}{q_t^2}\right]} + \mathbb{E}\left[\frac{\sqrt{50dK}}{\min_{j \le \tau} q_j}\right]\right)\right).$$