Stochastic Bandits with Context Distributions

Johannes Kirschner, Andreas Krause

Department of Computer Science, ETH Zurich

Neurips 2019, Additional Material

Introduction: Stochastic Contextual Bandits

In the **contextual bandit** model, the learner interacts with an unknown environment.

In each iteration $t = 1, 2, \ldots, T$,

1) the environment provides a context c_t,
2) the learner observes the context and chooses an action x_t,
3) the environment reveals a stochastic reward $f(x_t, c_t) + \epsilon_t$.

The learner's objective is to maximize reward, and compete with the best-in-hindsight mapping from context to actions.

Examples: Bandits with Stochastic Context

Crop Variety Testing

Action: Crop Variety

Context: Weather Conditions

Movie Recommendation

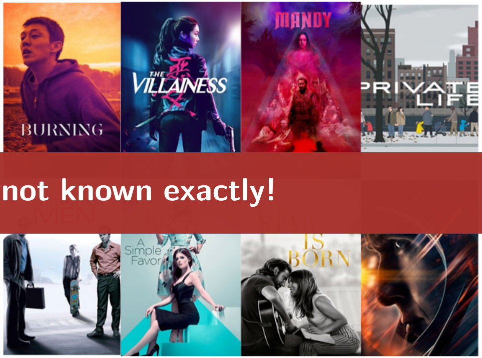
Action: Movie

Context: New User

Examples: Bandits with Stochastic Context

Crop Variety Testing

Movie Recommendation



Context is stochastic and not known exactly!

Action: Crop Variety

Context: Weather Conditions

Action: Movie

Context: New User

Our Setting: Bandits with Context Distributions

Notation: Action set $\mathcal{X}$, context set $\mathcal{C}$

$\phi_{x,c} \in \mathbb{R}^d$: Feature vectors, $\theta \in \mathbb{R}^d$: True parameter

For $t = 1, 2, \ldots, T$:

1: *Environment chooses distribution* $\mu_t \in \mathcal{P}(\mathcal{C})$,

2: *Learner observes* μ_t,

3: Learner chooses $x_t \in \mathcal{X}$,

4: *Environment secretly samples context* $c_t \sim \mu_t$,

5: Learner obtains reward $y_t = \phi_{x_t,c_t}^\top \theta + \epsilon_t$

Learner never observes context c_t!

Examples: Context Distributions

Crop Variety Testing

Context Distribution:

(Stochastic) Weather Prediction

Movie Recommendation

Context Distribution:

Based on user statistics

Bandits with Distributional Context: Regret

Impossible Baseline: $x_t^* = \arg\max_x \phi_{x,c_t}^\top \theta \quad (\rightarrow\ \Omega(T)$ regret)

Contextual Regret: Compare to best mapping $\pi^*(\mu_t) \rightarrow \mathcal{X}$.

$$x_t^* = \pi^*(\mu_t) = \arg\max_x \mathbb{E}_{c \sim \mu_t}[\phi_{x,c}^\top \theta]$$

Distributional Context Regret: $R_T := \sum_{t=1}^T \left(\phi_{x_t^*,c_t}^\top \theta - \phi_{x_t,c_t}^\top \theta\right)$

UCB for Bandits with Context Distributions

Key insight:

$$y_t = \phi_{x_t,c_t}^\top \theta + \epsilon_t = \mathbb{E}_{c\sim\mu_t}[\phi_{x_t,c}^\top \theta] + \underbrace{\phi_{x_t,c_t}^\top \theta - \mathbb{E}_{c\sim\mu_t}[\phi_{x_t,c}^\top \theta]}_{\textit{zero mean, acts like noise}} + \epsilon_t$$

$$= \mathbb{E}_{c\sim\mu_t}[\phi_{x_t,c}]^\top \theta + \xi_t$$

$$=: \bar{\phi}_{x,t}^\top \theta + \xi_t$$

We show: **UCB** on the expected features $\bar{\phi}_{x,t}$ achieves $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T})$.

Sample Based UCB for Context Distributions

Sample Averages: $\tilde{\phi}^l_{x,t} = \frac{1}{l}\sum_{i=1}^l \phi_{x,\tilde{c}_i}$ with samples $\tilde{c}_1, \dots \tilde{c}_l \sim \mu_t$.

$$y_t = \phi_{x_t,c_t}^\top \theta + \epsilon_t = \tilde{\phi}_{x_t,t}^\top \theta + \underbrace{\phi_{x_t,c_t}^\top \theta - \tilde{\phi}_{x_t,t}^\top \theta}_{=: \, b_t \text{ no longer zero mean!}} + \epsilon_t$$

Solution: Least-squares regression with *adversarial bias* b_t:

If $|b_t| \leq 1/\sqrt{t}$, *statistical rate* of least-squares estimator *unchanged!*

We show: **UCB** on sample-averaged features $\tilde{\phi}^{l=t}_{x,t}$ achieves $R_T \leq \tilde{\mathcal{O}}(d\sqrt{T})$

Summary of Contributions

Our setting: **Bandits with context distributions**

- UCB type-strategy with **regret bounds**
- **Sample-based algorithm** with guarantees
- **Kernelized setting** (Bayesian optimization)
- Variant with context observed after action choice.
- **Numerical experiments** on real-world data.

Arxiv: `https://arxiv.org/abs/1906.02685`