Taming the Monster Every Context:

Complexity Measure and Unified Framework for Offline-Oracle Efficient Contextual Bandits

Hao Qin, Chicheng Zhang *The University of Arizona*

UNIVERSITY OF ARIZONA College of Science

Abstract

1. We introduce ***Offline Estimation to Decisions (OE2D)***, an algorithmic framework that reduces contextual bandit learning with general reward approximation to offline regression.
2. *OE2D* achieves near-optimal regret for contextual bandits and is computational efficient with appropriate relaxation.
3. We propose ***Decision-Offline Estimation Coefficient (DOEC)***, a new complexity measure capturing the difficulty of reducing online contextual bandits to offline regression.
4. We relate the *DOEC* to *Decision Estimation Coefficient* (*DEC*) [1], bridging offline- and online-oracle-efficient contextual bandit algorithm design.

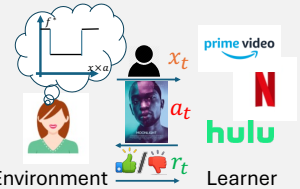
Context Bandit

For time step $t = 1, 2, \dots, T$:

Receives context x_t

Takes an action $a_t \in A$

Receives reward $r_t = f^*(x_t, a_t) + \varepsilon_t$



Λ-Regret:

$$\mathrm{Regret}_\Lambda(p \mid x) = \max_{q \in \Lambda} \mathrm{E}_{a' \sim q}[f^*(x, a')] - \mathrm{E}_{a \sim p}[f^*(x, a)]$$

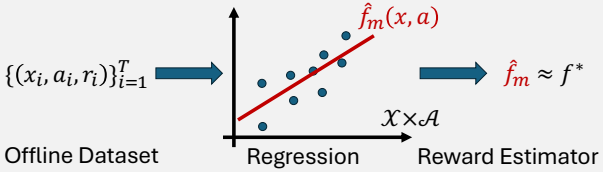
benchmark set Λ

Cumulative Λ-Regret

$$\mathrm{Regret}_\Lambda(\mathrm{Alg}, \mathrm{T}) = \sum_{t=1}^{T} \mathrm{E}_x[\mathrm{Regret}_\Lambda(p_t \mid x_t)]$$

Goal: minimize $\mathrm{Regret}_\Lambda(\mathrm{Alg}, \mathrm{T})$ efficiently, with the help of an offline regression oracle

Offline Regression Oracle

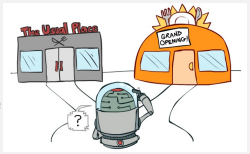


Algorithm: *Offline Estimation to Decision (OE2D)*

For epoch m = 1, 2, …,

#Offline Estimation

Compute $\hat{f}_m$ via offline regression oracle



For t in epoch m:

Observe context x_t

#Decision-making

Compute p_t as the solution of ***exploitative F-design*** *w.r.t.* $\hat{f}_m$

$$\underset{p \in \mathrm{co}(\Lambda)}{\operatorname{argmin}} \max_{q \in \Lambda} \left(\left(\mathrm{E}_{a \sim q}[\hat{f}_m(x_t, a)] - \mathrm{E}_{a \sim p}[\hat{f}_m(x_t, a)]\right) + \frac{1}{\gamma_m} \mathrm{Coverage}_\varepsilon(p, q; \mathcal{F}_{x_t}) \right)$$

Exploitation Exploration

Sample $a_t \sim p_t$

Decision-Offline Estimation Coefficient (DOEC)

$\mathcal{G}$ is a reward function class, g is current believed reward function

$$\mathrm{doec}_{\gamma,\varepsilon}(g, \mathcal{G}, \Lambda) := \min_{p \in \mathrm{co}(\Lambda)} \max_{q \in \Lambda} \left(\mathrm{E}_{a \sim q}[g(a)] - \mathrm{E}_{a \sim p}[g(a)] + \frac{1}{\gamma} \mathrm{Coverage}_\varepsilon(p, q; \mathcal{G}) \right)$$

doec being small means that we can find exploration strategies that are simultaneously near-greedy and collects useful data for future learning .

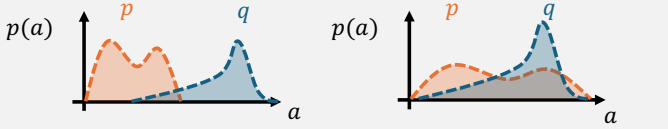
F-design [5]: $\gamma \to 0$ => pure exploration

Coverage: a measure of the ability of using data from behavior policy p to evaluate target policy q

Evaluation error under target policy q

$$\mathrm{Coverage}_\varepsilon(p, q; \mathcal{G}) = \sup_{g, g' \in \mathcal{G}} \frac{\left(\mathrm{E}_{a \sim q}[g(a) - g'(a)]\right)^2}{\varepsilon + \mathrm{E}_{a \sim p}[\left(g(a) - g'(a)\right)^2]}$$

Estimation error under behavior policy p



Insufficient coverage (high value) sufficient coverage (low value)

Computational Efficiency

1. OE2D makes $\log(T)$ number of oracle calls:

Algorithm	Oracle	# Oracle Calls	Assumptions on Reward
E2D [1]	Online	T	General
Falcon [2]	Offline	$\log(T)$	Discrete Action
Lin.Falcon [3]	Offline	$\log(T)$	Per-context Linear
UCCB [3]	Offline	T	General
E2D.Off [4]	Offline	T	General
OE2D	Offline	$\log(T)$	General

2. OE2D computes exploitative F-design efficiently via **relaxation**

Assumptions	Relaxed Coverage	Solution
Discrete Action	$\sum_a \frac{q(a)}{p(a)}$	$p(a) = \frac{1}{v - \gamma g(a)}$
h-Smooth regret	$\frac{1}{h}\mathrm{E}_{a \sim \mu}\left[\frac{q(a)}{p(a)}\right]$	$p(a) = \frac{1}{h \vee (v - \gamma g(a))}$
Per context Generalized Linear Model (GLM)	$\mathrm{trace}(\Sigma_p^{-1}\Sigma_q)$	*Convex Optimization*

Regret Guarantee

If $\mathrm{doec}_{\gamma,\varepsilon}(f, \mathcal{F}_x, \Lambda) \lesssim \tilde{\mathrm{O}}\left(\frac{D}{\gamma}\right)$:

$$\mathrm{Regret}(\mathrm{OE2D}, T) \lesssim \tilde{\mathrm{O}}\left(\sqrt{TD \cdot \log(|\mathcal{F}|)}\right)$$

Examples:

Assumptions	D
Discrete Action	$\lvert A \rvert$
h-Smooth regret	$1/h$
Generalized GLM	d
per-context Eluder dimension	d_ε

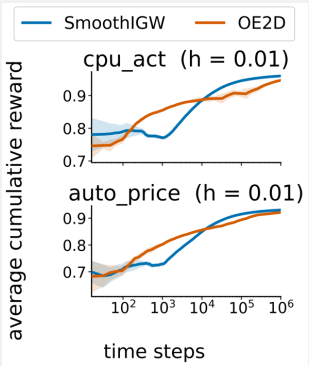
Connection between DEC [1] and DOEC

$$\mathrm{DEC}_\gamma \lesssim \mathrm{DOEC}_{\gamma,\varepsilon}$$

Experiments

h-Smooth regret setting

Compared with *SmoothIGW* [6]



(up) OpenML cpu_act
(down) OpenML autoPrice

References:

[1] The statistical complexity of interactive decision making. D.J. Foster et al.

[2] Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. D.Simchi-Levi and Y. Xu.

[3] Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. Y. Xu and A. Zeevi.

[4] Online estimation via offline estimation: An information-theoretic framework. D.J. Foster, et al.

[5] The Non-linear F -Design and Applications to Interactive Learning. A. Agarwal et al.

[6] Contextual Bandits with Smooth Regret: Efficient Learning in Continuous Action Spaces. Zhu and Mineiro