Taming the monster every context: Unified framework for contextual bandits with offline regression oracles

Chicheng Zhang

Join work with Hao Qin (UArizona)

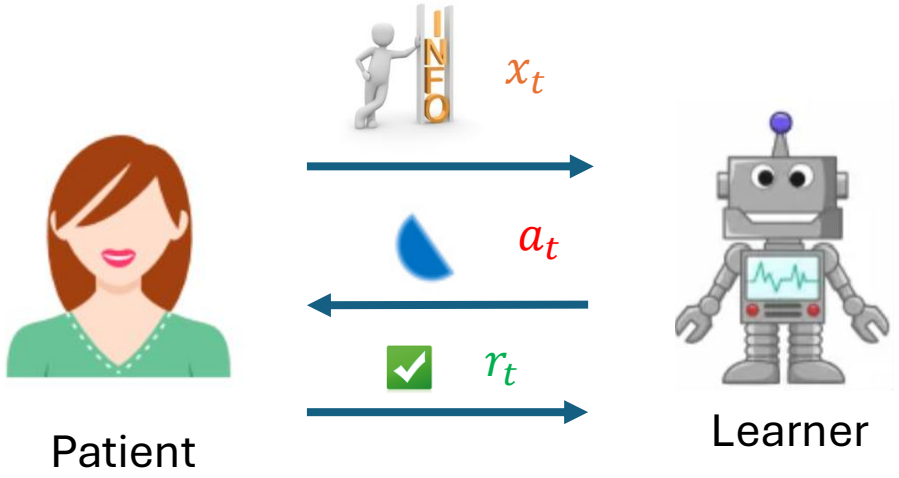
THE UNIVERSITY OF ARIZONA

Contextual bandits

- For time step $t = 1,2,..T$:
  - Receives context x_t from distribution D
  - Takes an action $a_t \in \mathcal{A}$
  - Receives reward $r_t = f^*(x_t, a_t) + \eta_t$

unknown true reward function

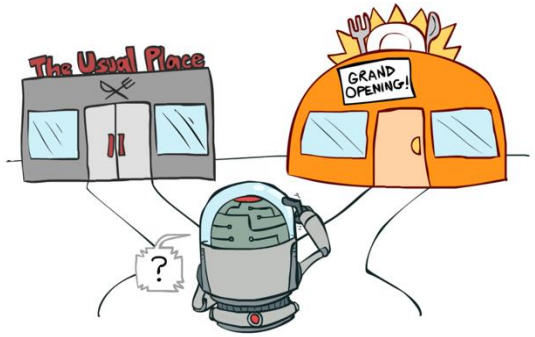
Zero-mean noise



- Learner's goal: minimize cumulative regret

$$\mathrm{Reg}(T) = \sum_{t=1}^{T} \max_{a \in \mathcal{A}} f^*(x_t, a) - f^*(x_t, a_t)$$

- Tradeoff: exploration vs. exploitation

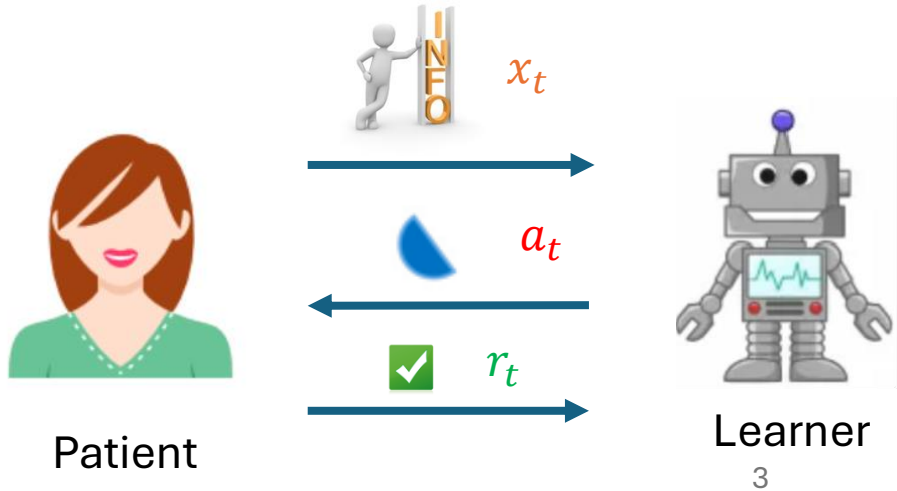


Regression-based contextual bandits

- **Assumption** (realizability): the learner has access to a reward class $\mathcal{F}$ that contains the true reward function $f^*(x, a) = \mathrm{E}[r_t \mid x_t = x, a_t = a]$

- Drives the development of many practical algorithms

  [e.g., Bietti et al, 2018; Foster et al, 2020]

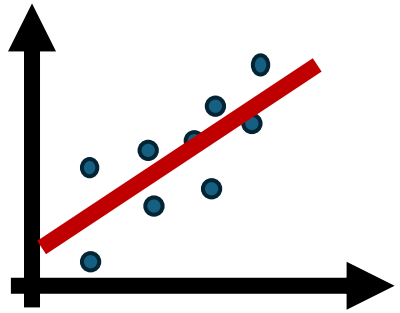


Regression oracles

- Computational primitives the learner has access to
- Offline regression oracle:

$(x_1, a_1, r_1), \ldots, (x_n, a_n, r_n)$ → Offline regression → $\hat{f} = \text{argmin}_{f \in \mathcal{F}} \sum_i (f(x_i, a_i) - r_i)^2$

More preferrable – implemented by standard ML libraries



- Online regression oracle:

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

x_t, a_t → Online regression → predicted reward $\hat{r}_t$

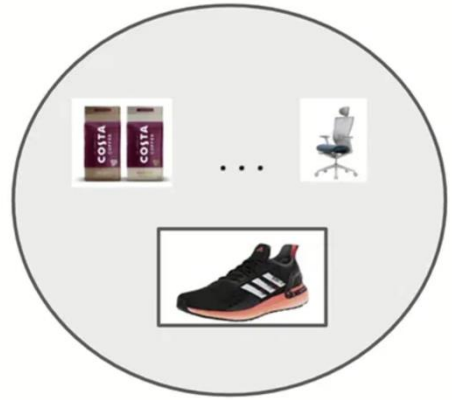
true reward r_t

Challenge: exploration with large action spaces

- Many contextual bandit applications have large action spaces:

Precision Medicine: dosage

Dynamic pricing: price



Product recommendation (Sen et al, 2021)

- Cannot afford to take every action even once
- Ideal: regret guarantee independent of $|\mathcal{A}|$

Combating large action spaces using structure

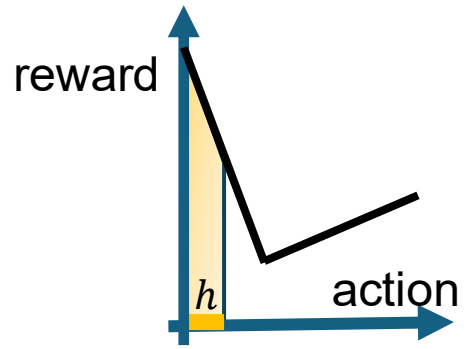
- Approach 1: structure in reward function class
  - E.g. Per-context linear reward (Demirer et al'19, Zhu et al,'22)

$$f^*(x,a) = \langle \theta^*(x), \phi(x,a) \rangle$$

Known feature extractor in R^d

  - Generalizes the linear bandit model (e.g. Dani et al, '08)

- Approach 2: regret against smoothed action distributions (Krishnamurthy et al'19, Zhu & Mineiro'22)
  - Allows utilizing Lipschitzness in reward functions



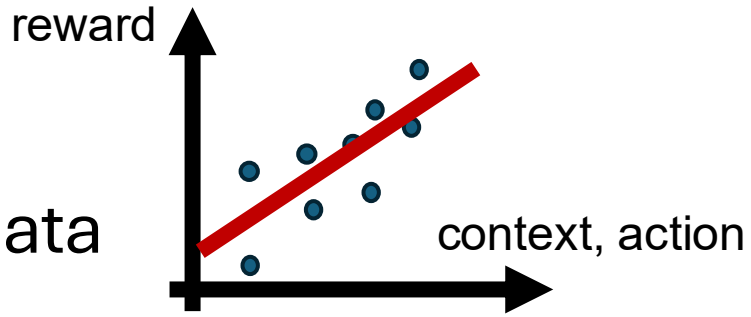
Large action space contextual bandits with offline regression oracles: state of the art

Algorithm with $\sqrt{T}$-regret guarantees	Total #oracle calls	Remark
Falcon [Simchi-Levi & Xu '20]	$\log T$	$\mathrm{poly}(\lvert\mathcal{A}\rvert)$ regret
Linear Falcon [Xu & Zeevi'20]	$\log T$	Per-context linear reward
UCCB [Xu & Zeevi'20]	T	General reward structure
E2D.Off [Foster et al'24]	T	General reward structure

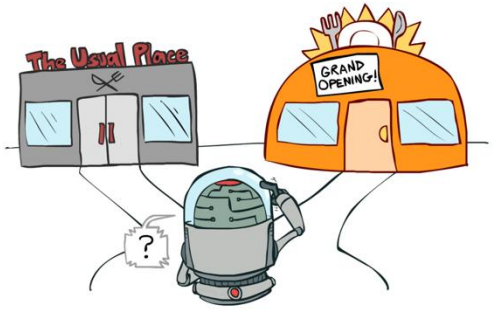
Can we design algorithms with $\log T$ offline oracle calls, utilizing General reward structure?

This work	$\log T$	General reward structure

Algorithm: Offline Estimation to Decision (OE2D)



- For each epoch m:
  - $\hat{f}_m \leftarrow$Offline regression on historical data
  - For time step t in epoch m:
    - Observe context x_t
    - Compute action distribution



$$\mathcal{F}_x = \{f(x,\cdot): f \in \mathcal{F}\}$$

$$p_t = \operatorname*{argmin}_{p} \max_{\lambda} \left(\underbrace{\mathrm{E}_{a\sim\lambda}\left[\hat{f}_m(x_t,a)\right] - \mathrm{E}_{a\sim p}\left[\hat{f}_m(x_t,a)\right]}_{\text{Exploit}} + \underbrace{\frac{1}{\gamma_m}\mathrm{Coverage}(p,\lambda;\mathcal{F}_x)}_{\text{Explore}} \right)$$

    - Sample action $a_t \sim p_t$, observe reward r_t

OE2D: remarks

- $\mathrm{Coverage}(p, \lambda; \mathcal{G})$:

  Source Target

  the utility of reward samples from p in evaluating λ's expected reward

- The solution p_t:
  - = Inverse Gap Weighting (Simchi-Levi & Xu, 2020, Foster & Rakhlin, 2020) with discrete action sets
  - = Exploration with log-determinant regularization (Xu & Zeevi'20, Foster et al'21) with linear structured reward functions
  - has guarantees mirroring the "Taming the Monster" exploration strategy (Agarwal et al'14)

- Efficient reduction from contextual bandits with large action spaces and general reward structure to offline estimation

Decision-Offline Estimation Coefficient (DOEC)

Definition The Decision-Offline Estimation Coefficient of reward class $\mathcal{G}$ is:

$$\mathrm{doec}_\gamma(\hat{g}, \mathcal{G}) = \min_p \max_\lambda \left(\underbrace{\mathrm{E}_{a\sim\lambda}[\hat{g}(a)] - \mathrm{E}_{a\sim p}[\hat{g}(a)]}_{\text{Decision}} + \underbrace{\frac{1}{\gamma}\mathrm{Coverage}(p, \lambda; \mathcal{G})}_{\text{Collecting Data for Offline Estimation}} \right)$$

and $\mathrm{doec}_\gamma(\mathcal{G}) = \max_{\hat{g}\in\mathcal{G}} \mathrm{doec}_\gamma(\hat{g}, \mathcal{G})$

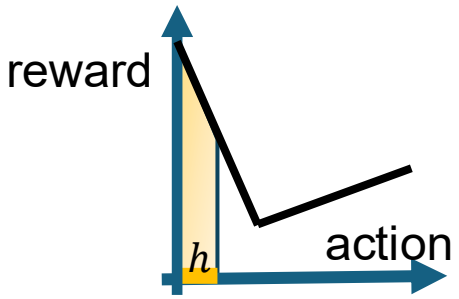
measures the difficulty of reducing contextual bandits to offline estimation

OE2D: performance guarantees

"effective #actions"

Theorem If $\max_x \mathrm{doec}_\gamma(\mathcal{F}_x) \lesssim \frac{D}{\gamma}$, then OE2D makes $\log T$ offline oracle calls and has regret $\lesssim \sqrt{DT \log|\mathcal{F}|}$.

Setting	Regret	Status
Discrete action space	$\sqrt{\lvert\mathcal{A}\rvert T \log\lvert\mathcal{F}\rvert}$	known
Per-context linear reward: $f^*(x,a) = \langle\theta^*(x), \phi(x,a)\rangle$, $\phi(x,a) \in \mathbb{R}^d$	$\sqrt{d\,T \log\lvert\mathcal{F}\rvert}$	known
Per-context structured reward: $\mathcal{F}_x$ has Eluder dimension d	$\sqrt{d\,T \log\lvert\mathcal{F}\rvert}$	new
Regret against h-smooth action distributions	$\sqrt{T/h \log\lvert\mathcal{F}\rvert}$	new

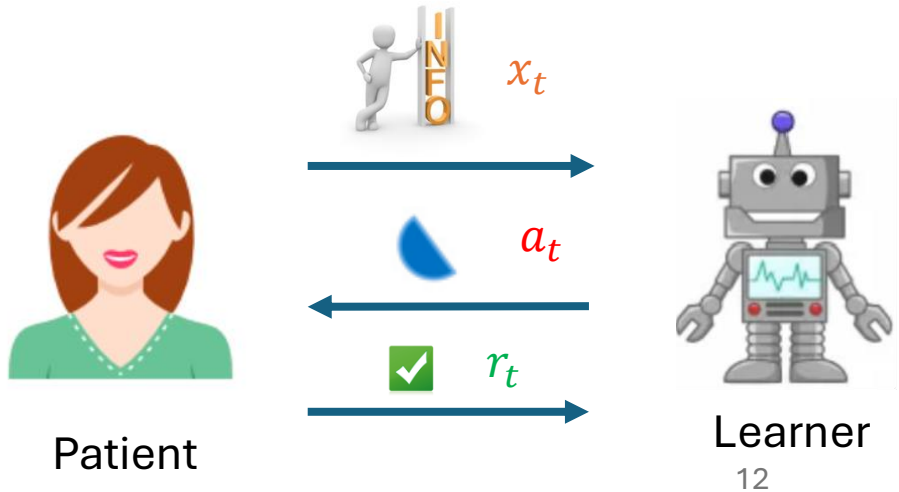


OE2D: extensions

- Extension 1: model misspecification, corruption, context distribution shifts
  - New results easily obtained from our modular regret guarantees

- Extension 2: cumulative regret guarantee *for every context* x
  - May be interesting for safety-critical applications

- Extension 3: $\mathrm{O}(\mathrm{loglog}\, T)$ calls to offline oracle if T known
  - Generalizes (Simchi-Levi & Xu '20)



12

DOEC vs. Decision-Estimation Coefficient (DEC)

DEC (Foster et al, 2021) enables a reduction from contextual bandits to *online regression*

$$\mathrm{dec}_\gamma(\hat{g}, \mathcal{G}) = \min_p \max_{\lambda,\, g^* \in \mathcal{G}} \left(\mathrm{E}_{a\sim\lambda}[g^*(a)] - \mathrm{E}_{a\sim p}[g^*(a)] - \gamma \mathrm{E}_{a\sim p}[\left(\hat{g}(a) - g^*(a)\right)^2] \right)$$

Decision

Online Estimation

Theorem (informal): Any exploration strategy p that certifies small DOEC also certifies small DEC.

Implication: exploration compatible with offline oracles are also compatible with online oracles!

Conclusion and open problems

- We propose the OE2D framework, establishing an efficient reduction from contextual bandits to offline estimation

- OE2D's regret is bounded by Decision-Offline Estimation Coefficient (DOEC), a new statistical complexity measure

- Open problems:
  - Computationally efficient implementation of OE2D?
  - Does DOEC characterize the fundamental statistical limit of contextual bandits?
  - Can we extend OE2D to approach other interactive decision making problems, e.g., partial monitoring /RL / RLHF?

Thank you!