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

Hao Qin

Chicheng Zhang

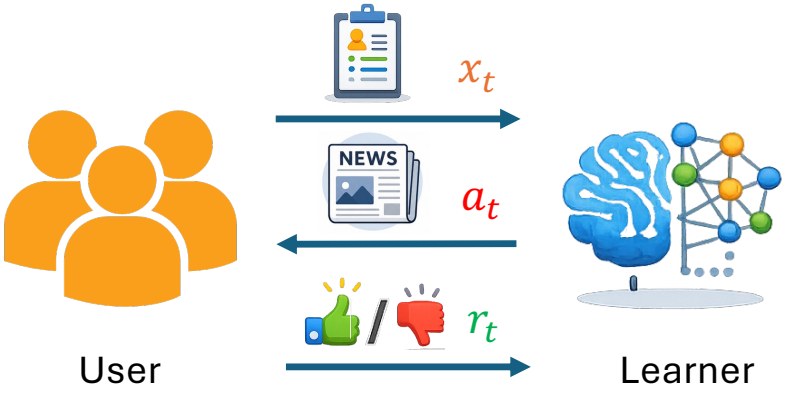
THE UNIVERSITY OF ARIZONA

COLT 2026
San Diego

Contextual bandits (CB)

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

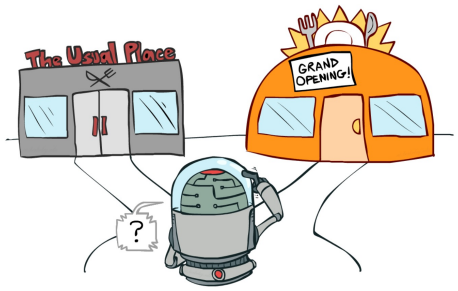
unknown reward function zero-mean noise



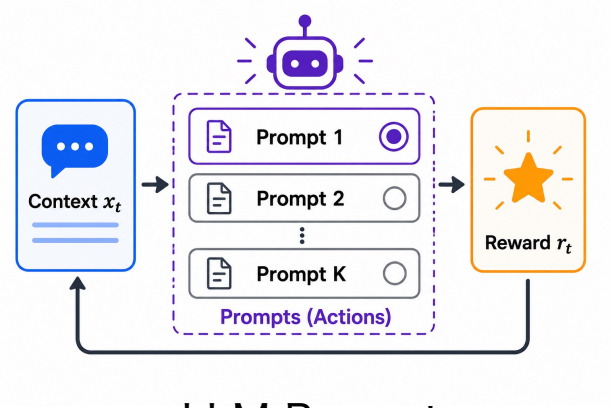
- Goal: minimize regret

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

- Tradeoff: exploration vs. exploitation



Challenge: exploration in large action space



LLM Prompt Optimization
[Hong et al. 2026]

Dynamic pricing: price

Product recommendation
[Sen et al, 2021]

- Take every action once
- Ideal: regret guarantee independent of $|\mathcal{A}|$

Hong et al., 2026 MASPOB: Bandit-Based Prompt Optimization for Multi-Agent Systems with Graph Neural Networks

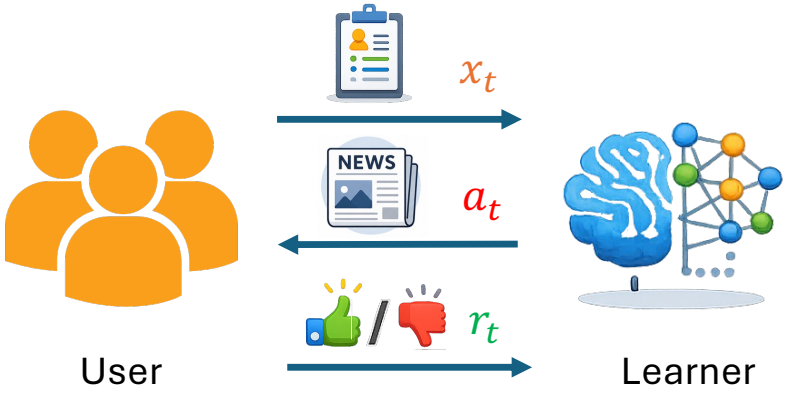
Regression-based CB

- **Assumption (realizability):**

  learner's model class contains $f^*(x, a) = \mathrm{E}[r_t \mid x_t = x, a_t = a]$

- Drives the development of many practical algorithms

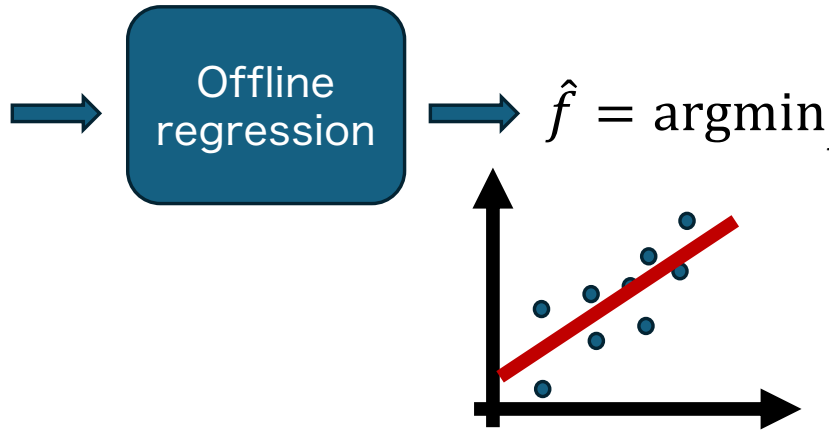
  [e.g., Bietti et al, 2018; Foster et al, 2020]



Regression oracles

learner has access to computational primitives: Offline regression oracle

$(x_1, a_1, r_1),$
$\dots,$
(x_n, a_n, r_n)



$\hat{f} = \text{argmin}_{f \in \mathcal{F}} \sum_i (f(x_i, a_i) - r_i)^2$

More preferrable – implemented
by standard ML libraries

CB with Offline Regression Oracle (SOTA)

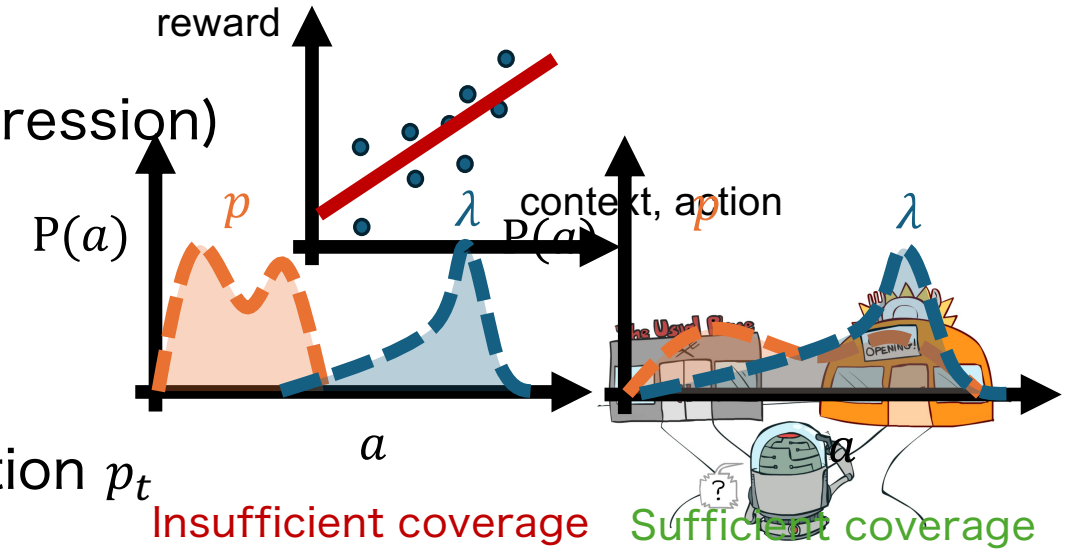
Algorithm with $\sqrt{T}$-regret		Total # oracle calls	Remark
Falcon	[Simchi-Levi & Xu '20]	$\log T$	$\mathrm{poly}(\lvert\mathcal{A}\rvert)$ regret
Linear Falcon	[Xu & Zeevi'20]	$\log T$	Per-context linear reward
UCCB	[Xu & Zeevi'20]	T	General reward structure
E2D.Off	[Foster et al'24]	T	General reward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

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

***OE2D* [This work]**		$\log T$	General reward structure

Algorithm: Offline Estimation to Decision (OE2D)

- For each epoch m:
  - $\hat{f}_m \leftarrow$ historical data (offline regression)
  - For time step t in epoch m:
    - Observe context x_t
    - Compute an action distribution p_t
    - Sample action $a_t \sim p_t$
    - observe reward r_t



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

Exploit

Explore

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

Decision-Offline Estimation Coefficient (DOEC)

$$\mathrm{doec}_{\gamma}(\hat{g}, \mathcal{G}) = \min_{p} \max_{\lambda} \left(\mathrm{E}_{a\sim\lambda}[\hat{g}(a)] - \mathrm{E}_{a\sim p}[\hat{g}(a)] + \frac{1}{\gamma} \mathrm{Coverage}(p, \lambda; \mathcal{G}) \right)$$

Decision

Collecting Data for Off-policy Estimation

DOEC measures the difficulty of reducing contextual bandits to offline estimation

Algorithm: Offline Estimation to Decision (OE2D)

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

Exploit Explore

Setting	Status
Discrete action space	Falcon [Simchi-Levi & Xu '20]
Per-context GLM reward $f^*(x,a) = \sigma(\langle \theta^*(x), \phi(x,a) \rangle)$, $\phi(x,a) \in \mathbb{R}^d$	New: Generalizes Linear Falcon [Xu & Zeevi, 2020]
Regret against h-smooth action distributions [Zhu & Mineiro, 2022]	New

OE2D: regret guarantees

Theorem "effective #actions"

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

Setting	D	Regret
Discrete action space	$\lvert\mathcal{A}\rvert$	$\sqrt{\lvert\mathcal{A}\rvert T\log\lvert\mathcal{F}\rvert}$
Per-context GLM reward $f^*(x,a) = \sigma(\langle\theta^*(x), \phi(x,a)\rangle)$, $\phi(x,a) \in \mathbb{R}^d$	d	$\sqrt{d\,T\log\lvert\mathcal{F}\rvert}$
Regret against h-smooth action distributions [Zhu & Mineiro, 2022]	$1/h$	$\sqrt{T/h\log\lvert\mathcal{F}\rvert}$

DOEC vs. Decision-Estimation Coefficient (DEC)

DEC [Foster et al, 2021] enables a reduction from contextual bandits to online regression

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

1. Propose OE2D framework:

   Efficient reduction from contextual bandits to offline estimation

2. Statistical complexity measure:

   Decision-Offline Estimation Coefficient (DOEC)

Thank you!

Open problems:

Can we extend OE2D to approach more interactive decision-making problems: partial monitoring / RL / RLHF / Semi-bandit ?

Relaxed coverage

Additional assumptions on the reward function class saves the computational cost

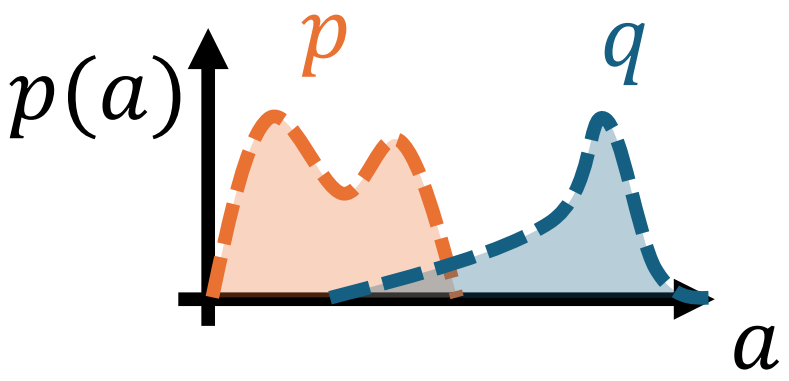
Setting	Relaxed Coverage	p_t
Discrete action space	$\sum_a \frac{q(a)}{p(a)}$	$\frac{1}{v - \gamma g(a)}$
Per-context GLM reward $f^*(x,a) = \sigma(\langle \theta^*(x), \phi(x,a) \rangle)$, $\phi(x,a) \in \mathbb{R}^d$	$\mathrm{trace}(\Sigma_p^{-1}\Sigma_q)$	Convex Optimization
Regret against h-smooth action distributions [Zhu & Mineiro, 2022]	$\frac{1}{h}\mathrm{E}_{a\sim\mu}\left[\frac{q(a)}{p(a)}\right]$	$\frac{1}{h \vee (v - \gamma g(a))}$

Coverage

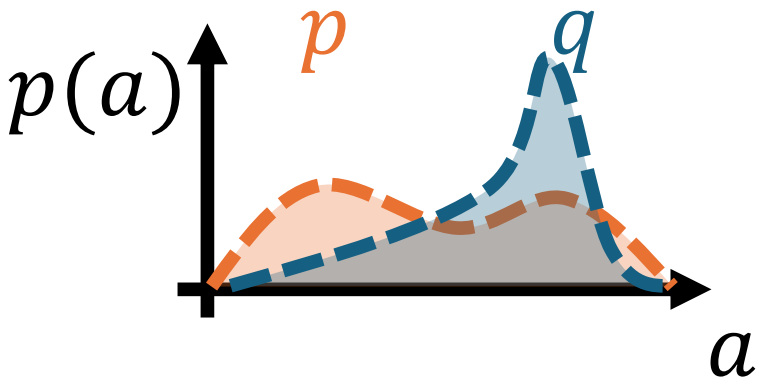
Evaluation error under target policy q

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

Estimation error under behavior policy p



Insufficient coverage (high value)



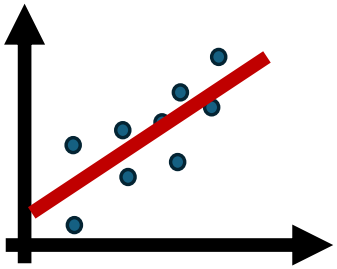
sufficient coverage (low value)

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 → $\hat{f} = \mathrm{argmin}_{f \in \mathcal{F}} \sum_i (f(x_i, a_i) - r_i)^2$

More preferrable
– implemented by standard ML libraries



PyTorch

scikit learn

DMLC XGBoost

LightGBM

  - Online regression oracle:

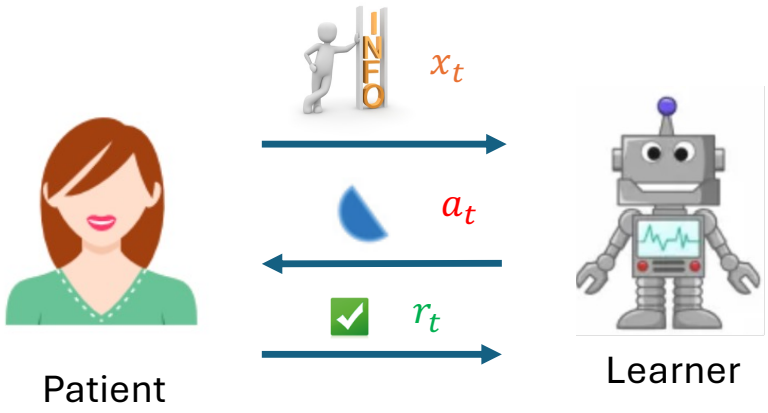
For $t = 1, \ldots, T$:

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

true reward r_t

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: $O(\log\log T)$ calls to offline oracle if T known
  - Generalizes (Simchi-Levi & Xu '20)



DOEC vs. Decision-Estimation Coefficient (DEC)

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

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

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!

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