Achieving Near-optimal Regrets in Confounded Contextual Bandits

Extended Abstract

Xueping GONG
HKUST
Hong Kong, China
xgongah@connect.ust.hk

Jiheng Zhang
HKUST
Hong Kong, China
jiheng@ust.hk

ABSTRACT

The contextual bandit problem is a theoretically justified framework with wide applications in various fields. While the previous study on this problem usually requires independence between noise and contexts, our work considers a more sensible setting where the noise becomes a latent confounder that affects both contexts and rewards. Such a confounded setting is more realistic and could expand to a broader range of applications. However, the unresolved confounder will cause a bias in reward function estimation and thus lead to a large regret. To deal with the challenges brought by the confounder, we apply the dual instrumental variable regression, which can correctly identify the true reward function. We prove the convergence rate of this method is near-optimal in two types of widely used reproducing kernel Hilbert spaces. Therefore, we can design a computationally efficient and regret-optimal algorithm based on the theoretical guarantees for confounded bandit problems.

KEYWORDS

Correlated Contexts; Instrumental Variable; Near-optimal Regret

ACM Reference Format:
Xueping GONG and Jiheng Zhang. 2023. Achieving Near-optimal Regrets in Confounded Contextual Bandits: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 PROBLEM FORMULATION

We consider a contextual bandit problem with K arms. A learner interacts with the environment in several rounds $t = 1, 2, \cdots, T$, where T is the time horizon. At each round t, the environment generates a context c_t in a compact set C. The learner is given an action set $\mathcal{A}$ with cardinality K. The context and action spaces can be discrete or included in $\mathbb{R}^d$. The learner will obtain a reward y_t after it chooses the action $a_t \in \mathcal{A}$. We model the reward function for the context-action pair as $y_t = f(c_t, a_t) + e_t$, where e_t is a bounded σ^2-subgaussian noise term. For notation brevity, we let $x_t := (c_t, a_t)$ and $\mathcal{X} := C \times \mathcal{A}$. Furthermore, we consider a correlated bandit problem and allow for noise e_t that is potentially correlated with the context c_t, namely, $\mathbb{E}[e_t|c_t] \neq 0$. We use a causal model in Figure 1 to describe the correlation. As the structural causal graph shows, the noise term E serves as an unobserved confounder, making the causal identification between the context-action pair X and the reward Y challenging.

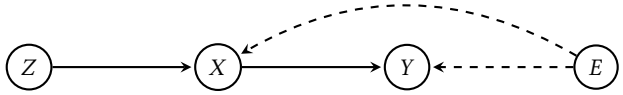

Figure 1: A causal model with unobserved confounders and instrumental variables.

To address the challenge, we assume that the learner can observe an extra *instrumental variables* z_t at each round, and that the tuple (c_t, z_t) is i.i.d. generated according to Figure 1. The properties of instrumental variables: *relevance, exclusion, unconfoundedness*, allows us to estimate an unbiased $\hat{f}(x)$ that captures the structural relationship between X and Y [2, 5]. The natural filtration $\mathcal{F}_t$ is defined w.r.t. the sequence of contexts, actions, instrumental variables and the collected rewards up to t. A policy π of the learner is a non-anticipatory decision sequence of actions in $\mathcal{A}$, i.e., $\pi_t : \mathcal{F}_{t-1} \to \mathcal{A}$. The expected regret of an algorithm is defined to be $Reg(T) = \sum_{t=1}^{T} \mathbb{E}_\pi[f(c_t, a_t^*) - f(c_t, a_t)]$, where the optimal action is $a_t^* := \arg\max_{a \in \mathcal{A}} f(c_t, a)$, and $a_t = \pi(\mathcal{F}_{t-1})$ is the arm pulled according to the policy π at step t. The expectation is taken w.r.t. the randomness in algorithms. Our goal is to find algorithms to minimize the expected regret.

2 ALGORITHM DESIGN

Dual Formulation. We first provide a dual formulation of IV regressions. To avoid ambiguity, we denote f^* as the truth and f as a general functional variable. Since samples are stochastic and noisy, we propose the following minimization problem

$$\min_{f \in \mathcal{H}} R(f) := \frac{1}{2}\mathbb{E}_{YZ}[(Y - \mathbb{E}_{X|Z}[f(X)])^2]. \tag{1}$$

The true structural function f^* can be identified by the optimum of the above minimization problem if f^* is in the function space $\mathcal{H}$. The conditional expectation operator $\mathbb{E}_{X|Z}[\cdot]$ is difficult to approximate by samples, because of the limited sample size and the possibly high dimensions of X and Z. Since (1) is a convex problem with respect to f, we can solve its dual problem to obtain a solution:

$$R(f) = \max_{u \in \mathcal{U}} \Psi(f, u), \tag{2}$$

where $\Psi(f, u) = \mathbb{E}_{XYZ}[f(X)u(Y, Z) - Yu(Y, Z) - \frac{1}{2}u(Y, Z)^2]$. Proved in [1, 4], the optimal solution $u^*(y, z)$ to the above problem takes the form $u^*(y, z) = \mathbb{E}_{X|z}[f(X)] - y$.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

RKHS. The function spaces $\mathcal{H}$ and $\mathcal{U}$ are chosen to be reproducing kernel Hilbert spaces associated with positive definite and continuous kernels $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $l : (\mathcal{Y}\times\mathcal{Z})\times(\mathcal{Y}\times\mathcal{Z}) \to \mathbb{R}$, respectively. Let $\phi(x) := k(x,\cdot)$ and $\varphi(y,z) := l((y,z),\cdot)$ be the canonical feature maps of $\mathcal{H}$ and $\mathcal{U}$, respectively. Due to the properties of RKHS, $\Psi(f,u)$ can be rewritten as

$$\Psi(f,u) = \langle \mathfrak{C}_{YZX} f - r, u\rangle_{\mathcal{U}} - \frac{1}{2}\langle u, \mathfrak{C}_{YZ} u\rangle_{\mathcal{U}},$$

where $r = \mathbb{E}_{YZ}[Y\varphi(Y,Z)]$, $\mathfrak{C}_{YZ} = \mathbb{E}_{YZ}[\varphi(Y,Z)\otimes\varphi(Y,Z)]$, $\mathfrak{C}_{YZX} = \mathbb{E}_{XYZ}[\varphi(Y,Z)\otimes\phi(X)]$. For a deep insight into the covariance operator $\mathfrak{C}_{YZ}$ and the cross-covariance operator $\mathfrak{C}_{YZX}$ (its adjoint is denoted as $\mathfrak{C}_{XYZ}$), we refer the readers to [1].

In empirical versions, corresponding operators might not be invertible, so a regularizer $\lambda = (\lambda_1, \lambda_2)$ can be added to the solution. The modified empirical version is

$$\begin{aligned}\hat{\Psi}_\lambda(f,u) \quad = \quad & \tfrac{1}{n}\textstyle\sum_{i=1}^n (f(x_i) - y_i)u(y_i,z_i) + \tfrac{\lambda_2}{2}\|f\|^2_{\mathcal{H}} \\ & -\tfrac{1}{2n}\textstyle\sum_{i=1}^n u(y_i,z_i)^2 - \tfrac{\lambda_1}{2}\|u\|^2_{\mathcal{U}}.\end{aligned}$$

The solution obtained from the following empirical minimization-maximization problem $\min_{f\in\mathcal{H}}\max_{u\in\mathcal{U}}\hat{\Psi}_\lambda(f,u)$ is denoted by dualIV$(k,l,\lambda_1,\lambda_2)$. We summarize the assumptions for function spaces and kernels in dualIV as followings.

Assumption 2.1.

- *(Realizability) The RKHSs $\mathcal{H}$ and $\mathcal{U}$ are correctly specified, i.e., $f^* \in \mathcal{H}$ and $u^* \in \mathcal{U}$.*
- *(Invertibility) The covariance operators $\mathfrak{C}_{YZ}, \mathfrak{C}_{XYZ}\mathfrak{C}_{YZ}^{-1}\mathfrak{C}_{YZX}$ are invertible.*
- *(Continuity) The referred kernels of $\mathcal{H}$ and $\mathcal{U}$ are continuous on compact sets.*

Concentration Inequalities. The convergence rate of dual methods will provide a useful guide to the algorithm design. We prove the following theorem for $\tilde{d}$-dimensional spaces, which matches the minimax lower bound in Theorem 2.2.

Theorem 2.1. *Let the $\tilde{d}$-dimensional RKHSs $\mathcal{H}$ and $\mathcal{U}$ associated with kernels k and l satisfying Assumption 2.1. Consider a dataset $(X_i,Y_i,Z_i)_{i=1}^n$ i.i.d. sampled according to Figure 1, and $\hat{f}_n$ is obtained from dualIV with regularization parameters $\lambda_i \simeq \sqrt{\tilde{d}\tau/n}$ for $i = 1,2$. Then, there exists a constant M which depends on the true structural function f^* and spaces $\mathcal{H},\mathcal{U}$, such that for all $\tau,\delta > 0$, the convergence rate of $\hat{f}_n$ satisfies $\|\hat{f}_n - f^*\|_{L^2(\mathbb{P}_X)} \le \sqrt{\frac{M\tilde{d}(\tau+\delta)}{n}}$ with probability at least $1 - 2e^{-\tau} - e^{-\delta}$.*

Theorem 2.2. *Consider data $(X_i,Y_i)_{i=1}^n$ following the relationship $Y_i = f(X_i) + E_i$, where E_i is a Gaussian or truncated Gaussian noise, and f is in a $\tilde{d}$-dimensional function space $\mathcal{H}$. For any estimation algorithm π, there exists a function $f \in \mathcal{H}$ such that $\|f - \hat{f}_n^\pi\|_{L^2(\mathbb{P}_X)} = \Omega\left(\sqrt{\tilde{d}/n}\right)$ for the estimator $\hat{f}_n^\pi$ obtained by π from the data.*

The Epoch Learning Strategy. This epoch strategy (e.g., [3, 6]) is purely due to technical reasons (Theorem 2.1 requires i.i.d. data), and we want to avoid a more complicated construction of martingales. Instead of feeding all the previous data into dualIV, we only feed the data in the previous epoch. Motivated by greedy algorithms, we design an action sampling policy based on the inverse gap weighting technique. The sampling policy keeps fixed during an epoch though changes over epochs. Hence, we can obtain an i.i.d. sequence within each epoch, and elegantly balance exploration and exploitation as we move along the epochs. We further prove that the regret of this algorithm is rate-optimal for properly selected tuning parameters.

Algorithm 1 DualIV Regression with an Epoch Learning Strategy

Input: epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$, confidence parameter δ, kernel functions k, l, tuning parameters η, η_1, η_2

Determine $\tilde{d}$ from kernels k and l
for epoch $m = 1, 2, \cdots$, **do**
  Collect (only) the data in epoch $m-1$ in $\mathcal{D}_{m-1}$
  Let parameters $\lambda_i = \eta_i\sqrt{\tilde{d}/|\mathcal{D}_{m-1}|}$ for $i = 1,2$
  Implement dualIV with input λ_1, λ_2, k, l and $\mathcal{D}_{m-1}$, and then obtain $\hat{f}_m$ (for epoch 1, $\hat{f}_1 = 0$)
  Compute $\gamma_m = \sqrt{\frac{\eta K|\mathcal{D}_{m-1}|}{\tilde{d}\log(2m^2/\delta)}}$ (for the first epoch, $\gamma_1 = 1$)
  for round $t = \tau_{m-1}+1, \cdots, \tau_m$ **do**
    Observe the context c_t and the instrumental variable z_t
    Compute $\hat{f}_m(c_t,a)$ for each action $a \in \mathcal{A}$ and the following probabilities

$$p_t(a) = \begin{cases} \frac{1}{K+\gamma_m(\hat{f}_m(c_t,\hat{a}_t)-\hat{f}_m(c_t,a))}, \text{ for all } a \neq \hat{a}_t \\ 1 - \sum_{a\neq\hat{a}_t} p_t(a), \text{ for } a = \hat{a}_t. \end{cases}$$

    where $\hat{a}_t = \max_{a\in\mathcal{A}}\hat{f}_m(c_t,a)$.
    Sample $a_t \sim p_t(\cdot)$ and take the action a_t
    Observe a reward y_t

Theorem 2.3. *Suppose that Assumption 2.1 hold in kernelized contextual bandit settings. Moreover, the epoch schedule is set to be $\tau_m = 2^m$, and the tuning parameters are properly selected to match the constant in Theorem 2.1. Then, with probability at least $1-\delta$, the expected regret $Reg(T)$ of Algorithm 1 is at most*

$$O\left(2\eta^{-1/2}\sqrt{KT\tilde{d}\log(2\log^2(T)/\delta)} + \|f^*\|_{\mathcal{H}}\sqrt{8T\log(2/\delta)}\right).$$

If we set $\delta = \tilde{d}/T$ and take expectation w.r.t. the randomness in contexts, then the expected regret can be further reduced to

$$O\left(\sqrt{K\tilde{d}T\log(2T\log^2(T)/\tilde{d})} + \|f^*\|_{\mathcal{H}}\sqrt{T\log(2T/\tilde{d})} + \tilde{d}\|f^*\|_{\mathcal{H}}\right)$$

via the property of conditional expectation. As a simple corollary of Theorem 2.3 and Theorem 2.4, the regret of this order is rate-optimal if we ignore the $O(\log\log T)$ term and constants.

Theorem 2.4. *Assume that $\mathcal{H}$ is $\tilde{d}$-dimensional. Moreover, $K \le 2^{\tilde{d}/2}$ and $T \ge \tilde{d}(\log K)^{1+\epsilon}$ for any small constant $\epsilon > 0$. For any algorithm π, there exists a bandit instance with a reward function $f \in \mathcal{H}$ such that $Reg(T) \ge \Omega\left(\sqrt{\tilde{d}T\log K\log(T/\tilde{d})}\right)$.*

ACKNOWLEDGMENTS

This work was supported in part by Hong Kong Research Grant Council (HKRGC) Grants 16208120 and 16214121.

REFERENCES

[1] Xueping Gong and Jiheng Zhang. 2022. Dual Instrumental Method for Confounded Kernelized Bandits. *arXiv preprint arXiv:2209.03224* (2022).

[2] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. 2017. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*. PMLR, 1414–1423.

[3] Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. 2019. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on learning theory*. PMLR, 2294–2340.

[4] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. 2020. Dual instrumental variable regression. *Advances in Neural Information Processing Systems* 33 (2020), 2710–2721.

[5] Stephen Powell. 2018. The Book of Why. *Journal of MultiDisciplinary Evaluation* 14, 31 (2018), 47–54.

[6] David Simchi-Levi and Yunzong Xu. 2021. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research* (2021).