# Instance-Optimal PAC Algorithms for Contextual Bandits

Zhaoqi Li\*, Lillian Ratliff\*, Houssam Nassif<sup>†</sup>, Kevin Jamieson\*, Lalit Jain\*

\*University of Washington

†Amazon

# **Contextual Bandit Setting**

- At each time  $t = 1, 2, \cdots$ :
  - Context  $c_t \in C$  arrives,  $c_t \sim \nu \in \Delta_C$
  - Choose action  $a_t \in A$
  - Receive reward  $r_t$ ,  $\mathbb{E}[r_t | c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$
- Policy class  $\Pi$ , each  $\pi \in \Pi, \pi : \mathbb{C} \to \mathbb{A}$
- Average reward:  $V(\pi) := \mathbb{E}_{c \sim \nu}[r(c, \pi(c))]$
- Optimal policy:  $\pi_{\star} := \arg \max V(\pi)$  $\pi \in \Pi$

### $(\epsilon, \delta)$ – PAC Guarantee

Return  $\hat{\pi}$  satisfying,  $V(\hat{\pi}) \geq V(\pi_*) - \varepsilon$  with probability greater than  $1 - \delta$  in a minimum number of samples.

### **Contributions:**

- Show the first instance-dependent lower bound for PAC contextual bandits
- Design sampling procedure that achieves this lower bound
- Design a computationally efficient algorithm - allowing context space C and policy space  $\Pi$  to be **infinite**!





# **Regret Minimization Not Enough**

- Regret heavily studied:
  - computationally efficient
  - also see [Zanette et al. 2021]

### Two Problems

- Minimax Result! Does not adapt to hardness of instance. a)

• ILOVETOCONBANDITS [Agarwal et al. 2014] achieves  $R_T = O(\sqrt{|A|} T \log(\Pi))$ ,

• Modification gives ( $\epsilon, \delta$ )- PAC algorithm w/ sample complexity  $O(|A|\log(\Pi/\delta)/\epsilon^2)$ ,

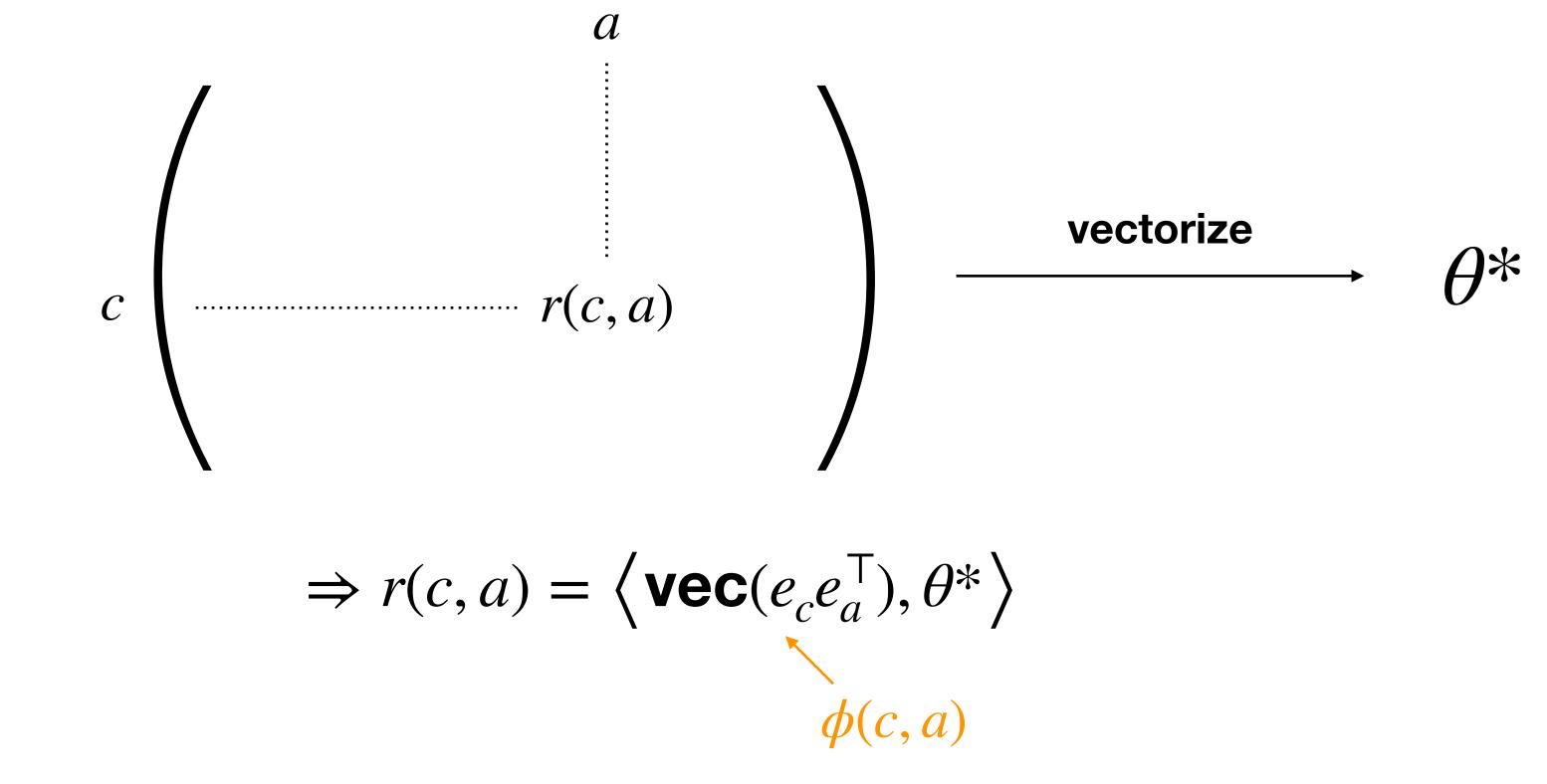
**True for any policy class! Not** capturing difficulty for learning  $\pi_*$ 

b) Can construct an example, where any optimal regret algorithm won't be instance optimal!



## **Agnostic Setting Reduces to Linear**

- Let  $\theta^* \in \mathbb{R}^{|C| \times |A|}$  where  $[\theta^*]_{c,a} = r(c,a)$



Lower bound motivated by best-arm identification in linear bandits [Fiez et al. 2019]

## **Contribution 1: A Lower Bound** $p) = \sum \nu_c \sum p_{c,a} \phi(c,a) \phi(c,a)^{\mathsf{T}},$ c a

Let 
$$\phi_{\pi} := \mathbb{E}_{c \sim \nu}[\phi(c, \pi(c))]$$
 and  $A(p)$ 

$$\rho_{\Pi,0} = \min_{\substack{p_c \in \triangle_A, \forall c \in C \\ p_c \in A}} p_c \in P_c \in C$$

2021]

**Theorem [Li et al. 2022]** Let  $\tau$  be the stopping time of the algorithm. Any  $(0,\delta)$ -PAC algorithm satisfies  $\tau \ge \rho_{\Pi,0} \log(1/2.4\delta)$  with high probability where  $\max_{\pi \in \Pi \setminus \pi_*} \frac{\|\phi_{\pi_*} - \phi_{\pi}\|_{A(p)^{-1}}^2}{\Delta(\pi)^2}.$ 

• This bound is better than the sample complexity bound based on disagreement coefficients [Foster et al. 2020] and decision-estimation coefficients [Foster et al.

## **Contribution 2: An Instance-Optimal Algorithm**

• In each round, Choose  $p_c \in \Delta_A, \forall$ 

 $\min_{p_c \in \Delta_A, \forall c \in C} \max_{\pi \in \Pi} \left( -\Delta(\pi) + \sqrt{\frac{\|\phi_{\pi} - \phi_{\pi}\|}{\|\phi_{\pi} - \phi_{\pi}\|}} \right)$ 

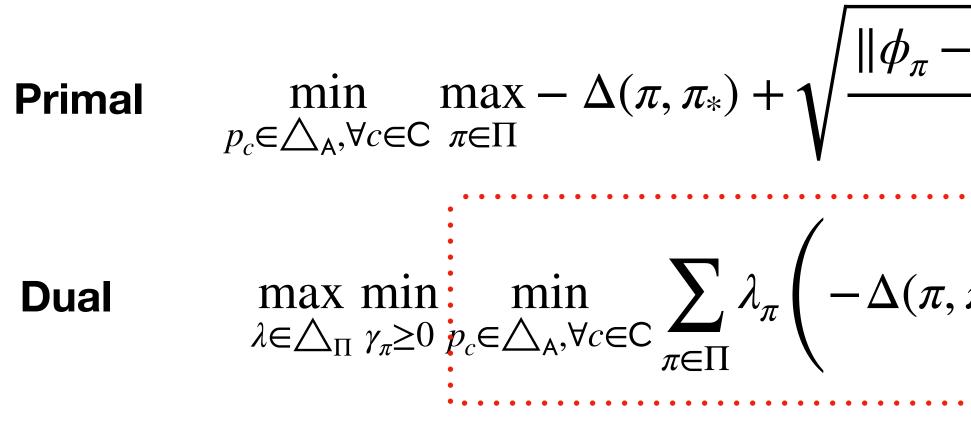
$$\frac{|c| \in \mathbb{C} \text{ and } n \text{ such that}}{|-\phi_{\pi_*}||_{A(p)^{-1}}^2 \log(1/\delta)} \le 2^{-l}$$

$$n_l$$

**Theorem [Li et al. 2022]** The algorithm returns an  $(\epsilon, \delta)$ -PAC policy with at most  $O(\rho_{\Pi,\epsilon} \log(|\Pi|/\delta) \log_2(1/\epsilon))$  samples.

# **Contribution 3: An Efficient Algorithm**

• Consider the dual formulation of the design of the previous algorithm:



- The dual objective is concave in  $\lambda$  and locally strongly convex in  $\gamma$ , so the saddle point problem can be solved
- Frank-Wolfe subroutine gives us a sparse yet good enough solution  $\lambda$

$$\frac{-\phi_{\pi_*}\|_{A(p)^{-1}}^2 \log(1/\delta)}{n}$$

$$\pi_*) + \gamma_{\pi} \|\phi_{\pi} - \phi_{\pi_*}\|_{A(p)^{-1}}^2 + \frac{\log(1/\delta)}{2\gamma_{\pi}n} \right).$$

analytical solution  $\Rightarrow$  implicitly maintain  $p_c$  for all  $c \in C$  simultaneously!



# Thank you!