

Efficient Optimal Learning for Contextual Bandits

M. Dudík*

D. Hsu[†]

S. Kale*

N. Karampatziakis[‡]

J. Langford*

L. Reyzin[§]

T. Zhang[#]

*Yahoo!; [†]Microsoft; [‡]Cornell; [§]Georgia Tech; [#]Rutgers

Contextual Bandits

For $t = 1 \dots T$

observe x

take action a

observe reward r

IID assumption:

x sampled i.i.d.

$\mathbb{P}(r | x, a)$ identical (but unknown) in each step

Goal: maximize reward

Goal: compete well

with a set of policies $\Pi = \{\pi\}$

where $\pi : \mathcal{X} \mapsto \mathcal{A}$

Goal: compete well
with a set of policies $\Pi = \{\pi\}$

where $\pi : \mathcal{X} \mapsto \mathcal{A}$

Previous best:

$$\text{regret} = O\left(\sqrt{TA \log |\Pi|}\right)$$

running time = *linear* in $|\Pi|$

Goal: compete well
with a set of policies $\Pi = \{\pi\}$
where $\pi : \mathcal{X} \mapsto \mathcal{A}$

Previous best:

$$\text{regret} = O\left(\sqrt{TA \log |\Pi|}\right)$$

running time = *linear* in $|\Pi|$

Our approach:

$$\text{regret} = O\left(\sqrt{TA \log |\Pi|}\right)$$

running time = *polynomial* in $\log |\Pi|$

How is that possible?

How is that possible?

Thought experiment:

rewards *for all actions* observed

collect data

optimize empirical risk

How is that possible?

Thought experiment:

rewards *for all actions* observed

collect data

optimize empirical risk

cost-sensitive
classification



How is that possible?

Our approach:

transform *partial feedback* into *full feedback*
call cost-sensitive learner

only *polylog* $|\Pi|$ calls

