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 $t = 1 \ldots T$

- 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 $\prod = \{\pi\}$ where $\pi : x \rightarrow a$
Goal: compete well with a set of policies \( \Pi = \{ \pi \} \)
where \( \pi : x \mapsto a \)

Previous best:

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

running time = \text{linear in } |\Pi|
Goal: compete well with a set of policies $\Pi = \{\pi\}$

where $\pi : x \rightarrow a$

Previous best:

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

running time $= \text{linear in } |\Pi|$ 

Our approach:

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

running time $= \text{polynomial in } \log |\Pi|$
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 |Π|* calls