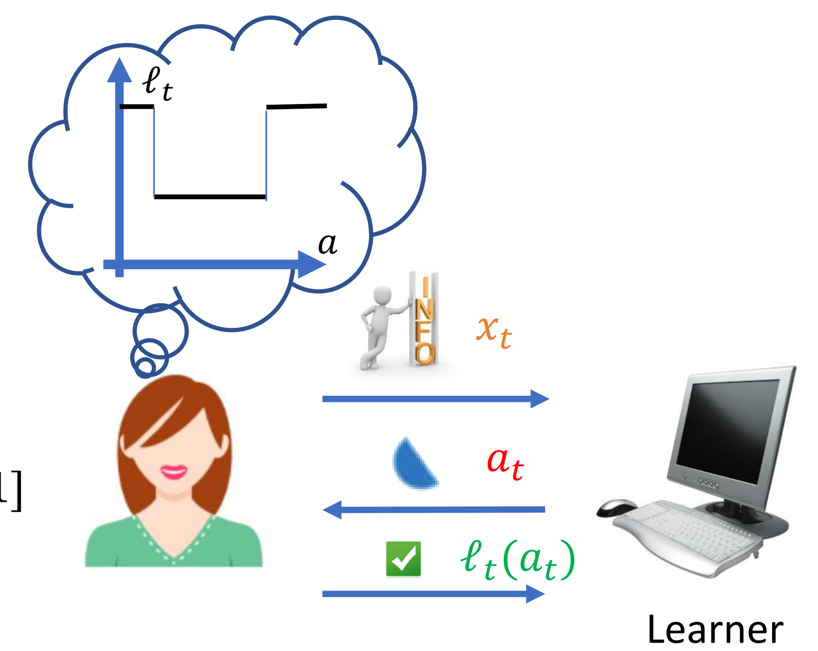
Efficient Contextual Bandits with Continuous Actions

Maryam Majzoubi[1], Chicheng Zhang[2], Rajan Chari[3], Akshay Krishnamurthy[3], John Langford[3], Alex Slivkins[3]
[1] *New York University*
[2] *University of Arizona*
[3] *Microsoft Research*

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Contextual Bandits (CB):



For time step $t = 1, 2, \dots, T$:

- Receives context x_t
- Takes an action $a_t \in A = [0,1]$
- Receives loss $\ell_t(a_t) \in [0,1]$

Learner's goal: minimize cumulative loss $\sum_{t=1}^{T} \ell_t(a_t)$

In many practical settings the **action** chosen is ***continuous-valued***.

Precision Medicine: dosage

Dynamic pricing: price

Networking: packet sending rate

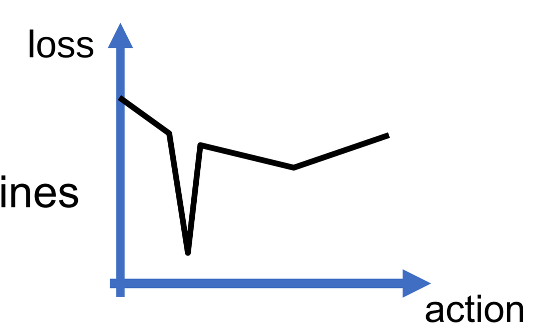
Challenges with continuous actions:

Discrete action spaces:

- Can afford trying all possible actions through "exploration"

Continuous action spaces:

- Need additional geometrical assumptions to guarantee competitiveness with "usual" baselines

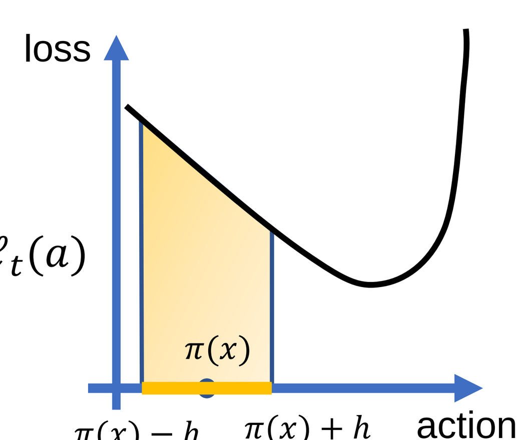


Smoothed regret for continuous-action CB

Smoothed regret [KLSZ19]:

$$\mathrm{SReg}(T, \Pi, h) = \sum_{t=1}^{T} \ell_t(a_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} \mathrm{E}_{a \sim \pi_h(\cdot|x)} \ell_t(a)$$

where $\pi_h(\cdot\,|x)$ = uniform($[\pi(x) - h, \pi(x) + h]$)



- Admits **assumption-free** nontrivial guarantees
- Recovers many existing results in contextual bandits with smooth loss assumptions, e.g. Lipschitz losses

➢ **Goal**: develop efficient algorithms with sublinear smoothed regret

CATS: Continuous Action Trees with Smoothing

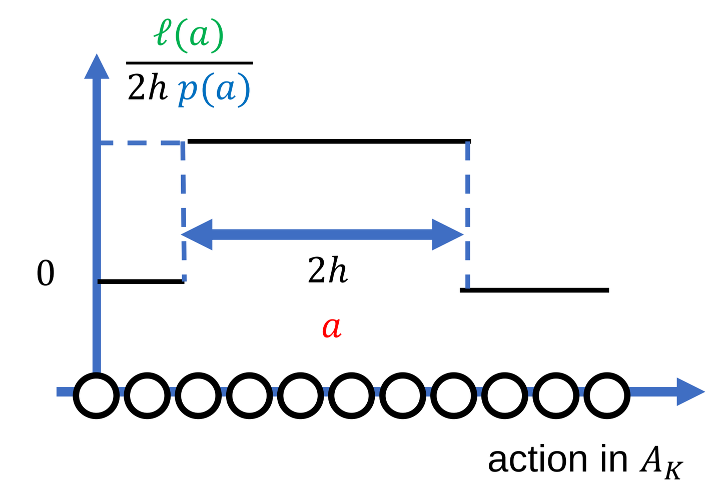
Key idea 1: reduce CB learning to importance-weighted (IW) multiclass learning

Input: interaction log $S = \{(x, a, \ell(a), p(a))\}$

1. Consider policy class Π taking actions in $A_K = \left\{0, \frac{1}{K}, \dots, \frac{K-1}{K}\right\}$.
2. For every input, generate cost-sensitive label using IPW loss estimate:

$$\hat{L}(\pi_h) = \frac{1}{|S|} \sum_S \frac{\pi_h(a|x)}{p(a)} \ell(a) = \frac{1}{|S|} \sum_S \tilde{c}\,(\pi(x))$$

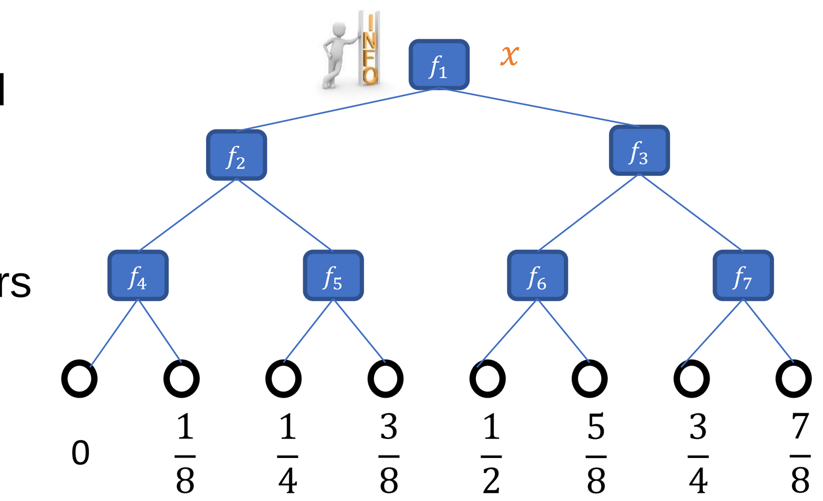
where cost vector $\tilde{c}$ is:



Key idea 2: Using tree policies to reduce IW multiclass learning to binary classification

Tree policy: special form of decision tree with leaves associated with fixed action labels in A_K

- Internal nodes are binary classifiers
- Execution time: $O(\log K)$



Training tree policies: we use the filter tree algorithm [BLR09], and show:

1. it can be implemented with $\boldsymbol{O}(\log \boldsymbol{K})$ time per example (with $\tilde{c}$ constructed above)
2. it achieves statistical consistency under realizability

Online contextual bandit learning guarantees

Theorem: *CATS* with input tree policy class $\Pi_{K,F}$:

- (computationally) has time cost $\boldsymbol{O}(\log \boldsymbol{K})$ per example,
- (statistically) has a smoothed regret guarantee of

$$\mathrm{SReg}(T, \Pi_{K,F}, h) \leq O\left(\left(\frac{K^2\, T^2 \ln|F|}{h}\right)^{1/3}\right)$$

under certain realizability assumptions

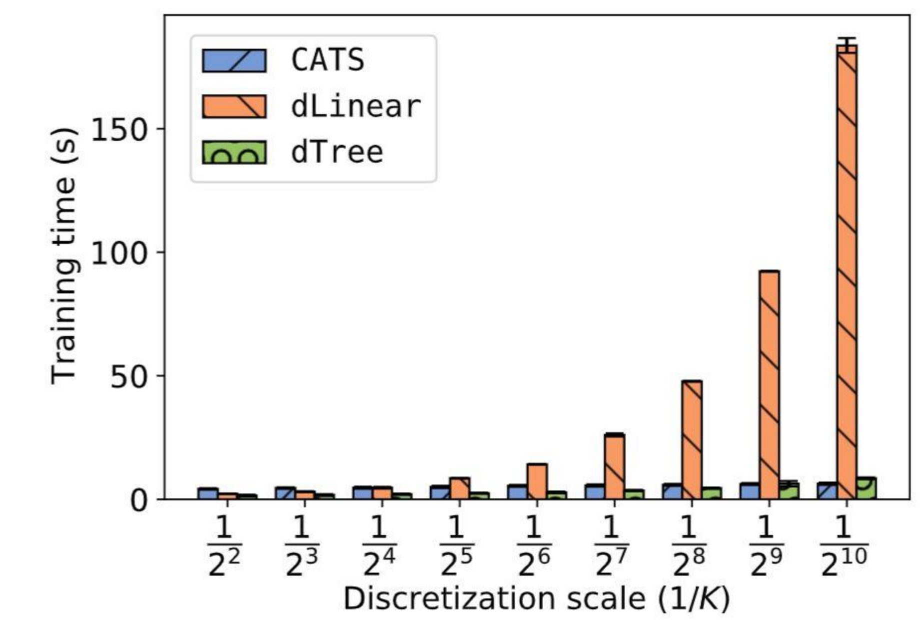
Experiments

We evaluate our learning algorithm on regression-based contextual bandit simulation environments, and compare with two baselines *dTree and dLinear*, that perform naïve discretization with epsilon-greedy exploration strategy.

Online contextual bandit learning:

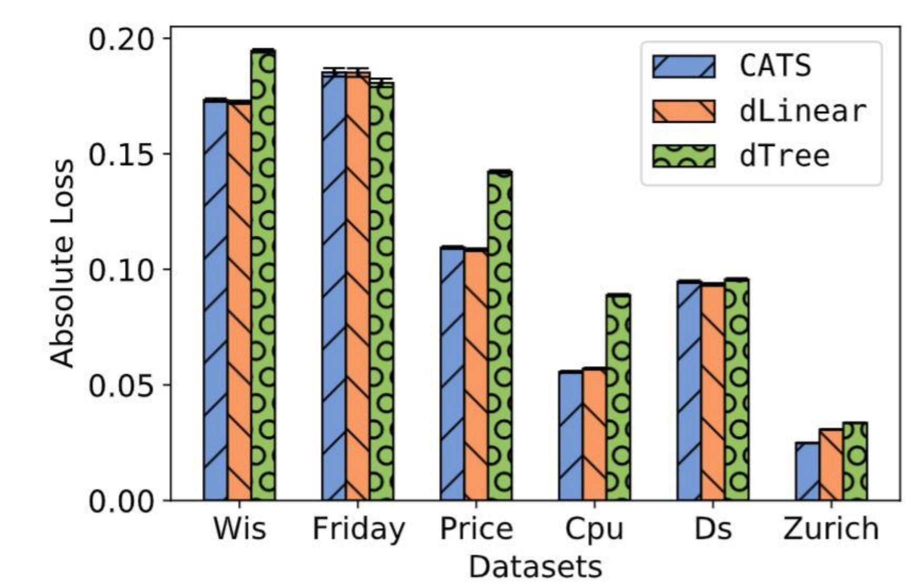
Time cost comparison:

CATS and *dTree* have much better scalability with respect to K compared to *dLinear*.



Online loss comparison:

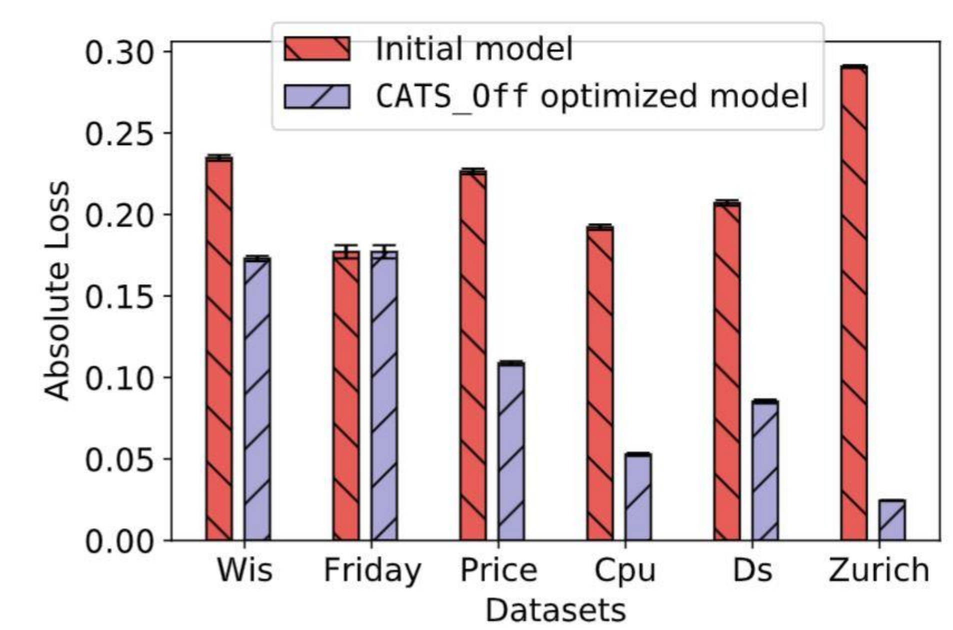
CATS and *dLinear* have lower average losses than *dTree*.



Off-policy optimization:

Key advantage over naïve discretization: it can use interaction log collected by one policy to do off-policy optimization over smoothing parameter h and discretization level K.

It produces tree policies that have significantly smaller test losses than the original policies.



References

[KLSZ19] Akshay Krishnamurthy, John Langford, Aleksandrs Slivkins, and Chicheng Zhang. Contextual bandits with continuous actions: smoothing, zooming, and adapting. COLT 2019.
[BLR09] Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. ALT 2009.