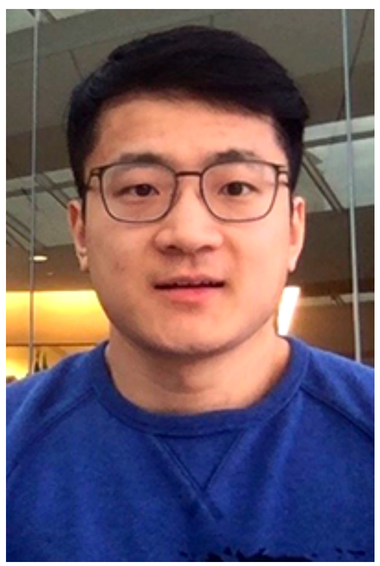
Contextual Bandits with Large Action Spaces: Made Practical




Yinglun Zhu[1], Dylan Foster[2], John Langford[2], Paul Mineiro[2]

[1]University of Wisconsin-Madison
[2]Microsoft Research NYC

Contextual bandits

Contextual bandits

For each round $t = 1, \ldots, T$:

Contextual bandits

For each round $t = 1, \ldots, T$:

- Receive context x_t.

Contextual bandits

For each round $t = 1, \ldots, T$:

- Receive context x_t.
- Select action $a_t \in \mathcal{A} := [1, \ldots, A]$.

Contextual bandits

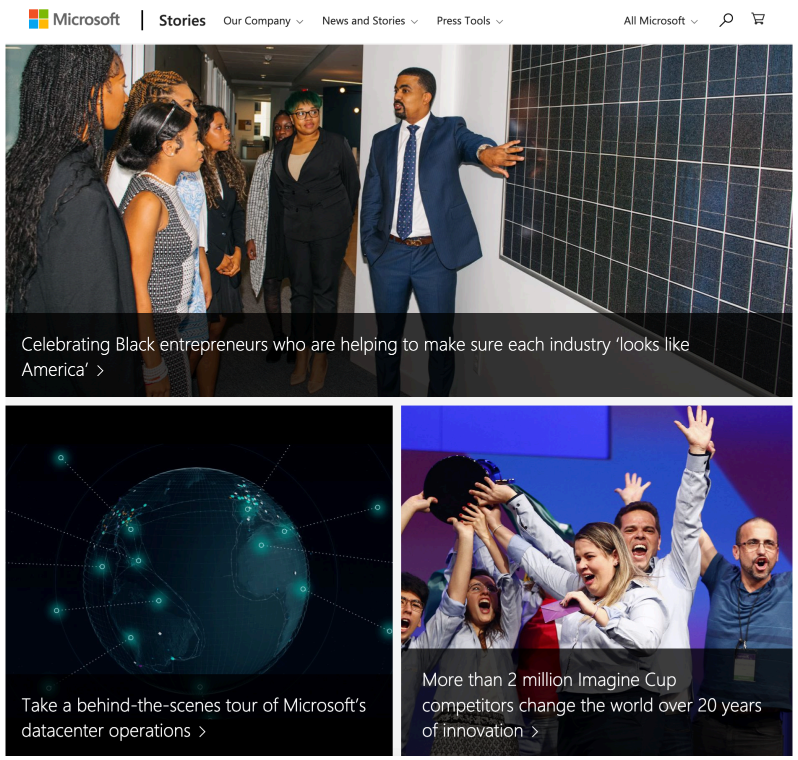
For each round $t = 1, \ldots, T$:

- Receive context x_t.
- Select action $a_t \in \mathscr{A} := [1, \ldots, A]$.
- Observe reward $r_t(a_t) \in [-1, 1]$.

Contextual bandits

For each round $t = 1, \ldots, T$:

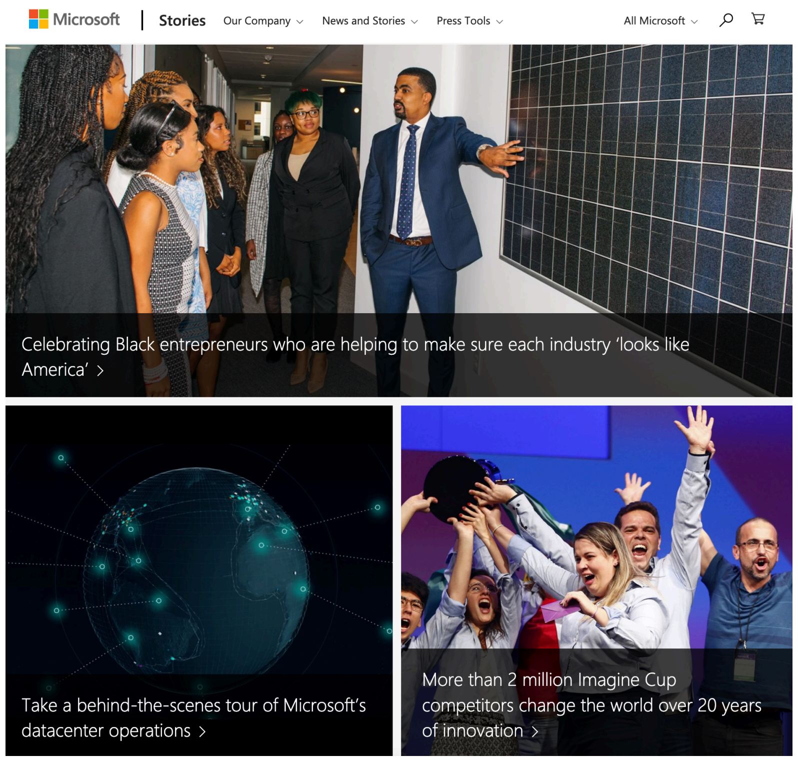
- Receive context x_t.
- Select action $a_t \in \mathscr{A} := [1, \ldots, A]$.
- Observe reward $r_t(a_t) \in [-1, 1]$.



Contextual bandits

For each round $t = 1, \ldots, T$:

- Receive context x_t.
- Select action $a_t \in \mathcal{A} := [1, \ldots, A]$.
- Observe reward $r_t(a_t) \in [-1, 1]$.



Goal: Minimize regret $\mathrm{Reg}_{\mathrm{CB}}(T) := \sum_{t=1}^{T} r_t(\pi^\star(x_t)) - r_t(a_t)$.

Existing guarantees

A standard realizability assumption

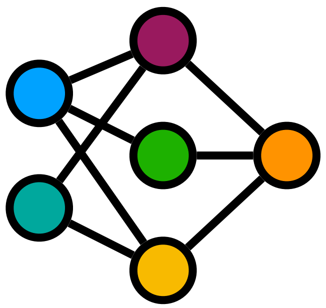
We assume $f^\star := \mathbb{E}[r_t \mid x_t] \in \mathscr{F}$ with a user-specified model class $\mathscr{F}$.

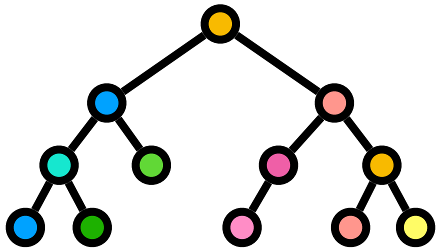
Existing guarantees

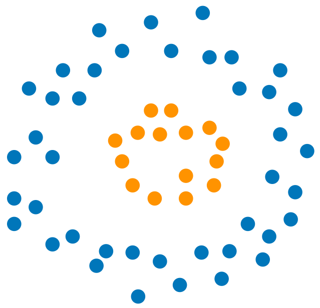
A standard realizability assumption

We assume $f^\star := \mathbb{E}[r_t \mid x_t] \in \mathscr{F}$ with a user-specified model class $\mathscr{F}$.

Rich function approximation for $\mathscr{F}$: Neural nets, decision trees, kernels, etc.





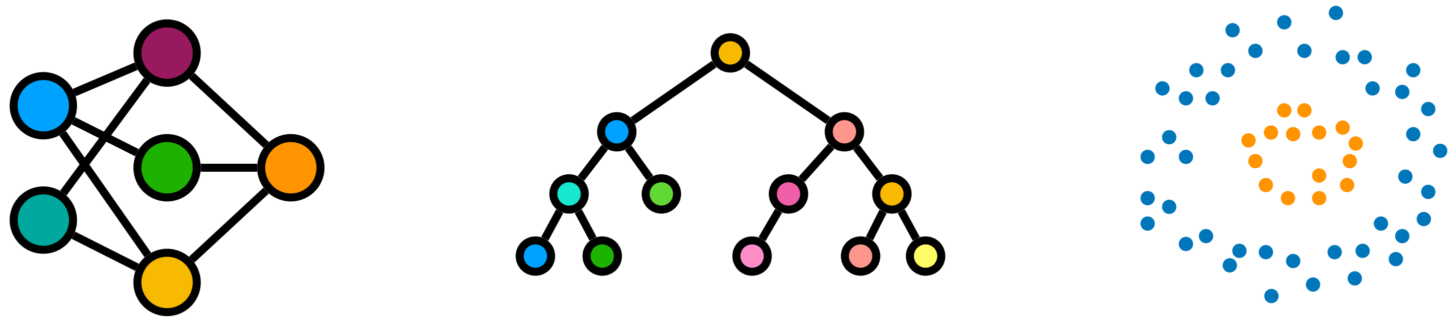


Existing guarantees

A standard realizability assumption

We assume $f^\star := \mathbb{E}[r_t \mid x_t] \in \mathscr{F}$ with a user-specified model class $\mathscr{F}$.

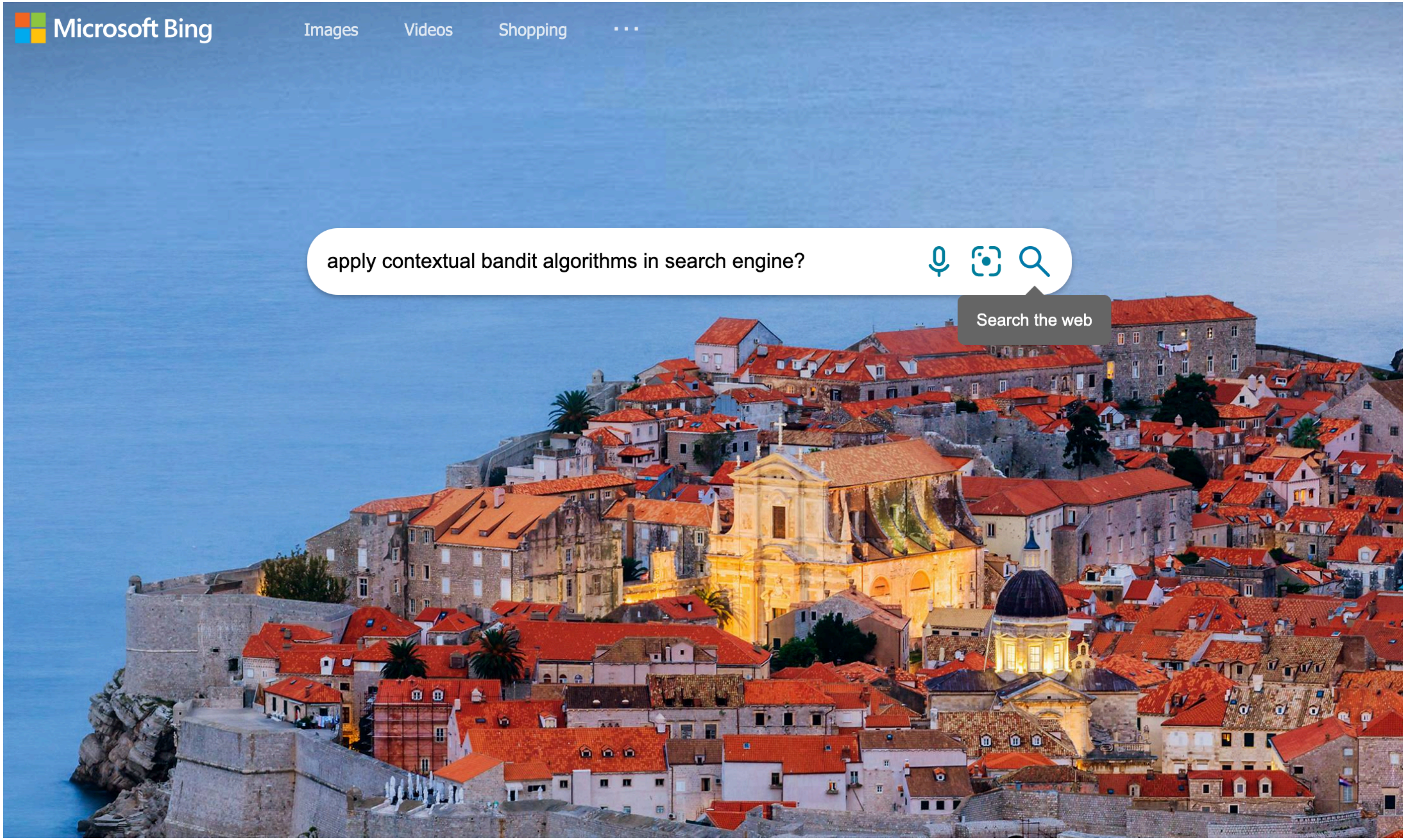
Rich function approximation for $\mathscr{F}$: Neural nets, decision trees, kernels, etc.



Theorem (Foster et al. 2020, Simchi-Levi et al. 2021)

There exist efficient ALGs that achieve regret $O(\sqrt{AT \log |\mathscr{F}|})$.

Large-scale recommendations

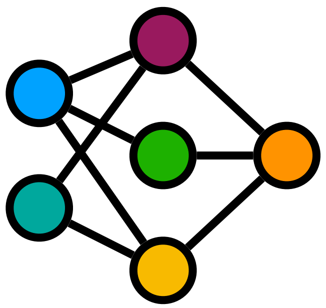


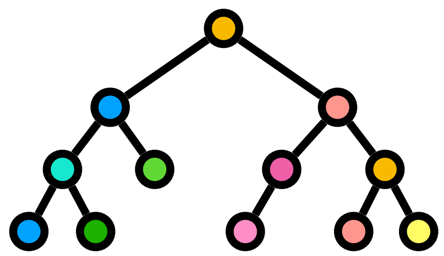
Existing guarantees

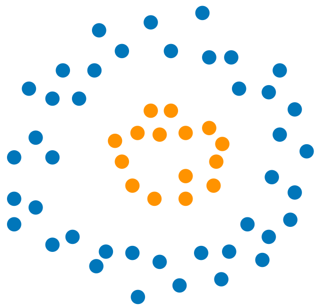
A standard realizability assumption

We assume $f^\star := \mathbb{E}[r_t \mid x_t] \in \mathscr{F}$ with a user-specified model class $\mathscr{F}$.

Rich function approximation for $\mathscr{F}$: Neural nets, decision trees, kernels, etc.







Theorem (Agarwal et al. 2012)

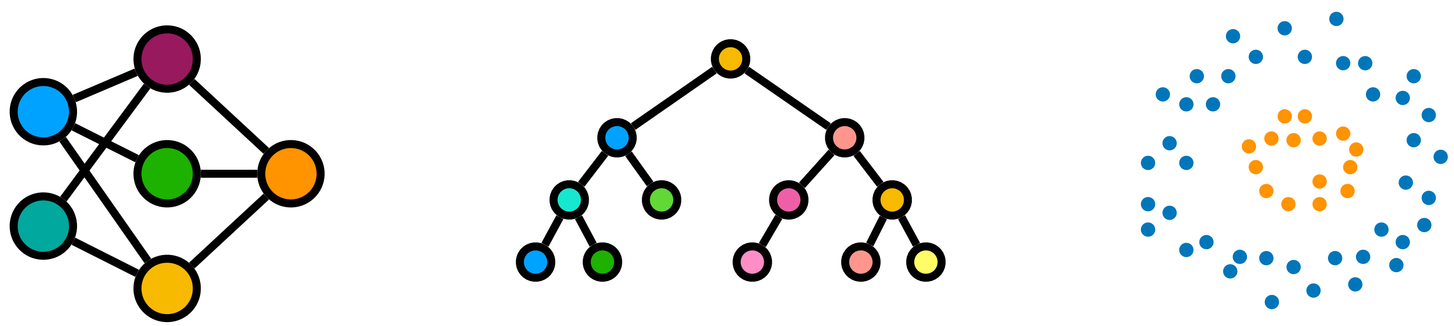
Any CB ALG must suffer worst-case regret $\Omega(\sqrt{AT \log |\mathscr{F}|})$.

Existing guarantees

A standard realizability assumption

We assume $f^\star := \mathbb{E}[r_t \mid x_t] \in \mathscr{F}$ with a user-specified model class $\mathscr{F}$.

Rich function approximation for $\mathscr{F}$: Neural nets, decision trees, kernels, etc.



Theorem (Agarwal et al. 2012)

Any CB ALG must suffer worst-case regret $\Omega(\sqrt{AT \log |\mathscr{F}|})$.

Question: Can we develop efficient ALGs to handle large action space problems?

A modeling assumption

Function approximation

We consider the following model class

$$\mathcal{F} := \{f_g(x, a) = \langle \phi(x, a), g(x) \rangle : g \in \mathcal{G}\},$$

where $\phi(x, a) \in \mathbb{R}^d$ is known feature embedding, and $\mathcal{G} : \mathcal{X} \to \mathbb{R}^d$ models the unknown context embedding.

A modeling assumption

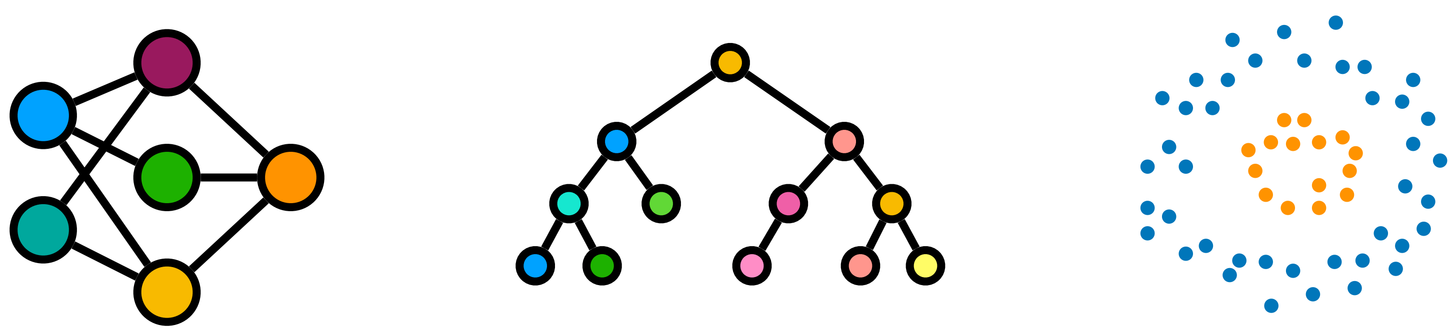
Function approximation

We consider the following model class

$$\mathscr{F} := \{f_g(x, a) = \langle \phi(x, a), g(x) \rangle : g \in \mathscr{G}\},$$

where $\phi(x, a) \in \mathbb{R}^d$ is known feature embedding, and $\mathscr{G} : \mathscr{X} \to \mathbb{R}^d$ models the unknown context embedding.

- Recover the finite action case when $\phi(x, a)$ is one-hot encoding and linear contextual bandits when $g(x) = \theta$ is constant.
- Allow general models for $\mathscr{G}$: Neural nets, decision trees, kernels, etc.



A modeling assumption

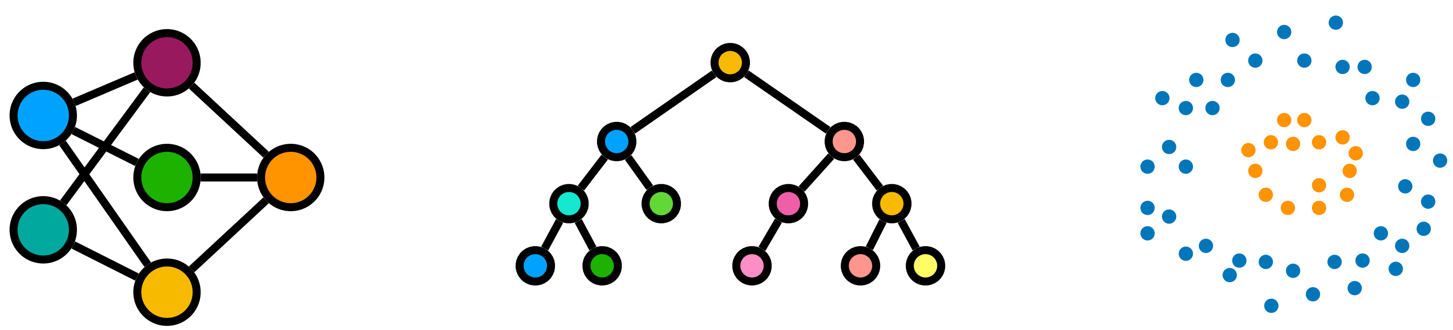
Function approximation

We consider the following model class

$$\mathscr{F} := \{f_g(x,a) = \langle \phi(x,a), g(x) \rangle : g \in \mathscr{G}\},$$

where $\phi(x,a) \in \mathbb{R}^d$ is known feature embedding, and $\mathscr{G} : \mathscr{X} \to \mathbb{R}^d$ models the unknown context embedding.

- Recover the finite action case when $\phi(x,a)$ is one-hot encoding and linear contextual bandits when $g(x) = \theta$ is constant.
- Allow general models for $\mathscr{G}$: Neural nets, decision trees, kernels, etc.



Linearly-structured actions with general function approximation

Computational oracles

Regression oracle

Online regression oracle such that

$$\sum_{t=1}^{T}\left(\hat{f}_t(x_t, a_t) - r_t(a_t)\right)^2 - \inf_{f\in\mathcal{F}}\sum_{t=1}^{T}\left(f(x_t, a_t) - r_t(a_t)\right)^2 \leq \mathsf{Reg}_{\mathsf{Sq}}(T)\,.$$

- $\mathsf{Reg}_{\mathsf{Sq}}(T) = O(\log|\mathcal{F}|)$ for general $\mathcal{F}$, and $\mathsf{Reg}_{\mathsf{Sq}}(T) = O(d)$ for linear models.
- Standard oracle studied in contextual bandits, e.g., FR '20, Zhang '21.

Computational oracles

Regression oracle

Online regression oracle such that

$$\sum_{t=1}^{T} \left(\hat{f}_t(x_t, a_t) - r_t(a_t)\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(f(x_t, a_t) - r_t(a_t)\right)^2 \leq \mathbf{Reg}_{\mathsf{Sq}}(T)\,.$$

- $\mathbf{Reg}_{\mathsf{Sq}}(T) = O(\log|\mathcal{F}|)$ for general $\mathcal{F}$, and $\mathbf{Reg}_{\mathsf{Sq}}(T) = O(d)$ for linear models.
- Standard oracle studied in contextual bandits, e.g., FR '20, Zhang '21.

Action optimization oracle

For any $\theta \in \mathbb{R}^d$ and $x \in \mathcal{X}$, returns $a^\star := \arg\max_{a \in \mathcal{A}} \langle \phi(x, a), \theta \rangle\,.$

Computational oracles

Regression oracle

Online regression oracle such that

$$\sum_{t=1}^{T} \left(\hat{f}_t(x_t, a_t) - r_t(a_t)\right)^2 - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \left(f(x_t, a_t) - r_t(a_t)\right)^2 \leq \mathbf{Reg}_{\mathsf{Sq}}(T)\,.$$

- $\mathbf{Reg}_{\mathsf{Sq}}(T) = O(\log|\mathcal{F}|)$ for general $\mathcal{F}$, and $\mathbf{Reg}_{\mathsf{Sq}}(T) = O(d)$ for linear models.
- Standard oracle studied in contextual bandits, e.g., FR '20, Zhang '21.

Action optimization oracle

For any $\theta \in \mathbb{R}^d$ and $x \in \mathcal{X}$, returns $a^\star := \arg\max_{a \in \mathcal{A}} \langle \phi(x, a), \theta \rangle$.

- Poly-time algorithms for combinatorial problems; hashing-based MIPS in general.
- Previously studied in linear bandit/pure exploration, e.g., DHK '08, CGLQW '17.

Algorithms and guarantees

Algorithmic framework

At each round $t = 1, \ldots, T$:

- Obtain $\hat{f}_t$ from the regression oracle.
- Efficiently compute optimal design wrt a $\hat{f}_t$-reweighted embedding.
- Sample an action from mixture of optimal design/greedy action.
- Update regression oracle.

Algorithms and guarantees

Algorithmic framework

At each round $t = 1, \ldots, T$:

- Obtain $\hat{f}_t$ from the regression oracle.
- Efficiently compute optimal design wrt a $\hat{f}_t$-reweighted embedding.
- Sample an action from mixture of optimal design/greedy action.
- Update regression oracle.

Key idea: Explore optimal design → generalize across actions.

- Novel $\hat{f}_t$-reweighted embedding to balance exploration/exploitation.
- Efficient computation of optimal design using action opt. oracle.

Algorithms and guarantees

Algorithmic framework

At each round $t = 1, \ldots, T$:

- Obtain $\hat{f}_t$ from the regression oracle.
- Efficiently compute optimal design wrt a $\hat{f}_t$-reweighted embedding.
- Sample an action from mixture of optimal design/greedy action.
- Update regression oracle.

Key idea: Explore optimal design → generalize across actions.

- Novel $\hat{f}_t$-reweighted embedding to balance exploration/exploitation.
- Efficient computation of optimal design using action opt. oracle.

Theorem

Our ALG achieves $\sqrt{\text{poly}(d) \cdot T}$-regret, with per-round $O(1)$ calls to the regression oracle and $\tilde{O}(d^3)$ calls to the linear optimization oracle.

- no explicit dependence on # actions both statistically and computationally

A large-scale exhibition

Amazon 3m dataset

A large-scale dataset that aims at predicting commodity identity based on text descriptions.

- Contexts: Text description of a commodity.
- Actions: Around 3 million different commodities.
- Rewards: $r_t(a_t) = \mathbb{I}(x_t$ describes commodity $a_t)$.

Table 1: Comparison with the previous state-of-the-art

Algs.	Averaged rewards
Sen et al. 2021	0.19
Ours	**0.43**