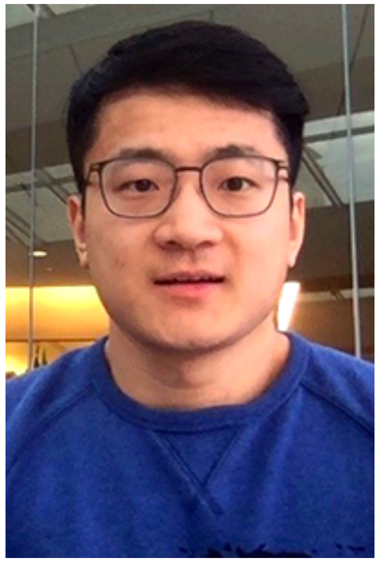
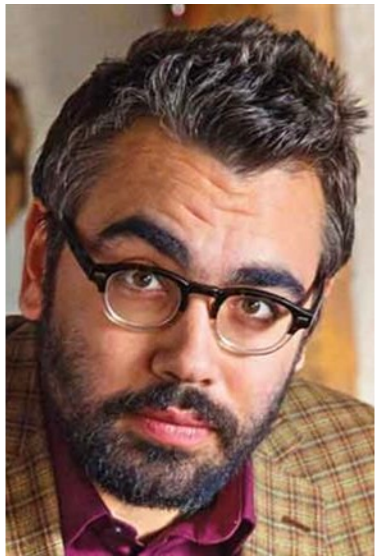
Contextual Bandits with Smooth Regret: Efficient Learning in Continuous Action Spaces

Yinglun Zhu[1] and 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 $a_t \in \mathscr{A}$.

Contextual bandits

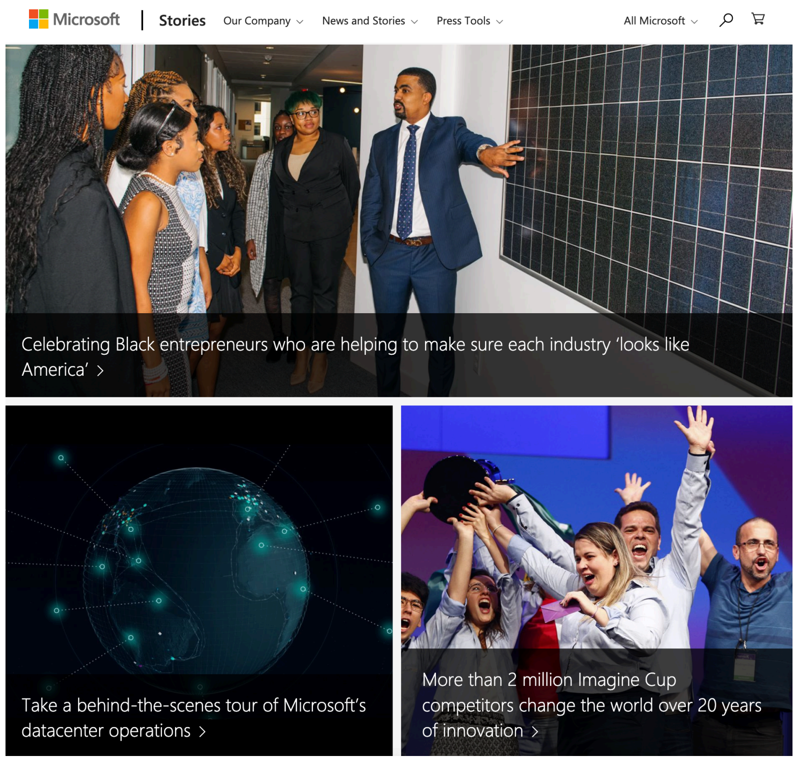
For each round $t = 1, \ldots, T$:

- Receive context x_t.
- Select action $a_t \in \mathcal{A}$.
- Observe loss $\ell_t(a_t) \in [0,1]$.

Contextual bandits

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

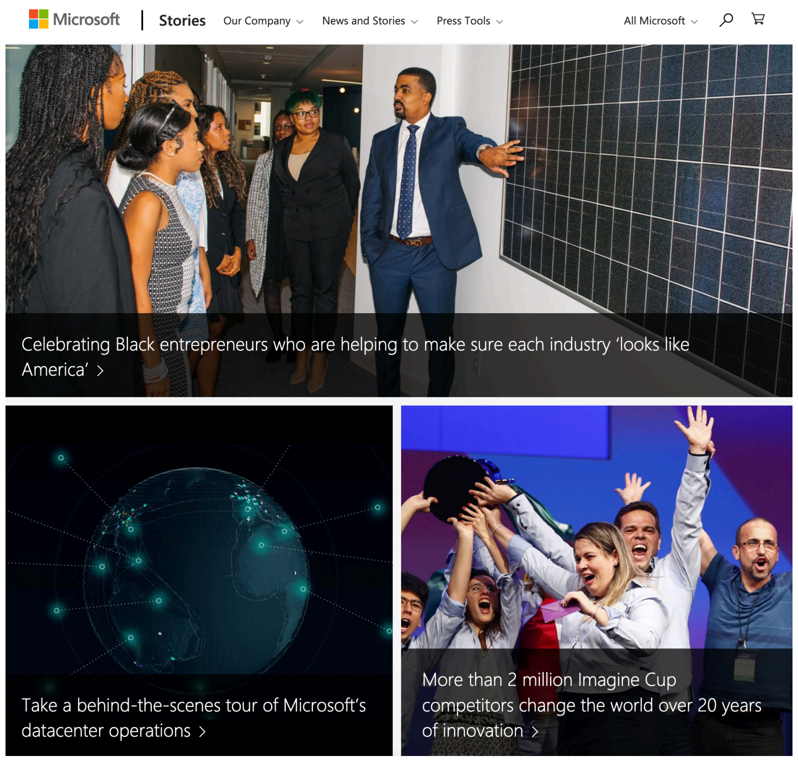
- Receive context x_t.
- Select action $a_t \in \mathscr{A}$.
- Observe loss $\ell_t(a_t) \in [0,1]$.



Contextual bandits

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

- Receive context x_t.
- Select action $a_t \in \mathcal{A}$.
- Observe loss $\ell_t(a_t) \in [0,1]$.



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

Existing guarantees

A standard realizability assumption

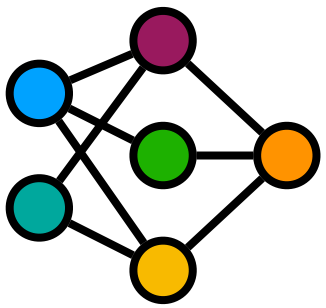
We assume $f^\star := \mathbb{E}[\ell_t \mid x_t] \in \mathcal{F}$ with a user-specified model class $\mathcal{F}$.

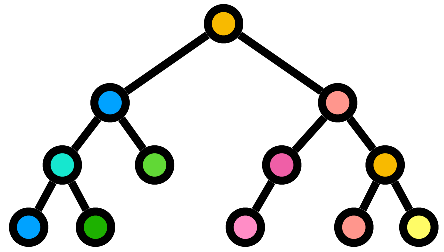
Existing guarantees

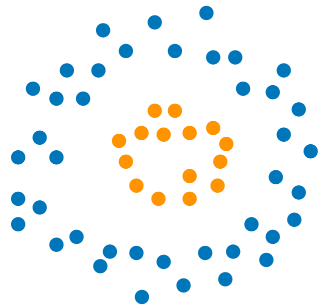
A standard realizability assumption

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

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





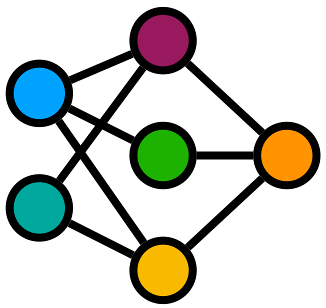


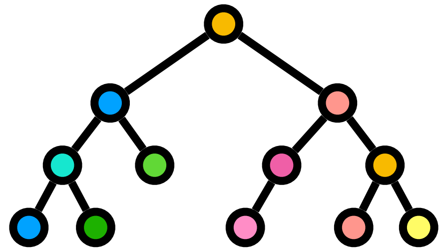
Existing guarantees

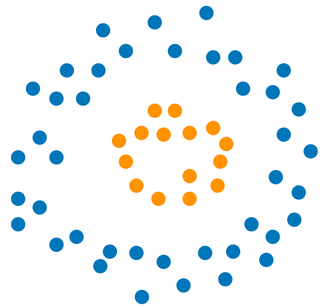
A standard realizability assumption

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

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





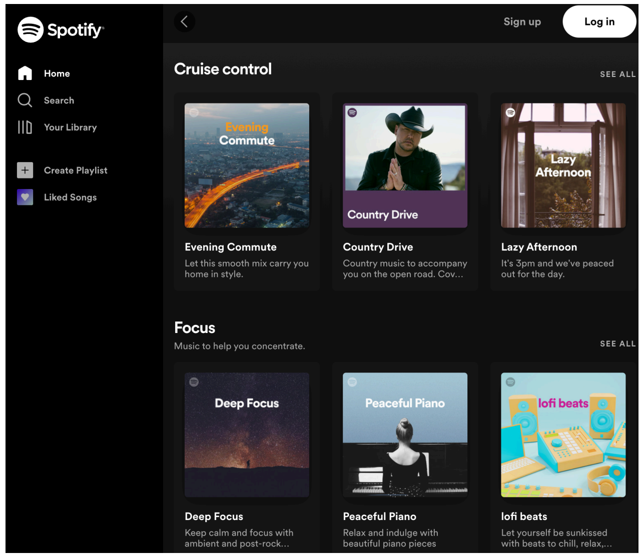


Theorem (Foster et al. 2020, Simchi-Levi et al. 2021)

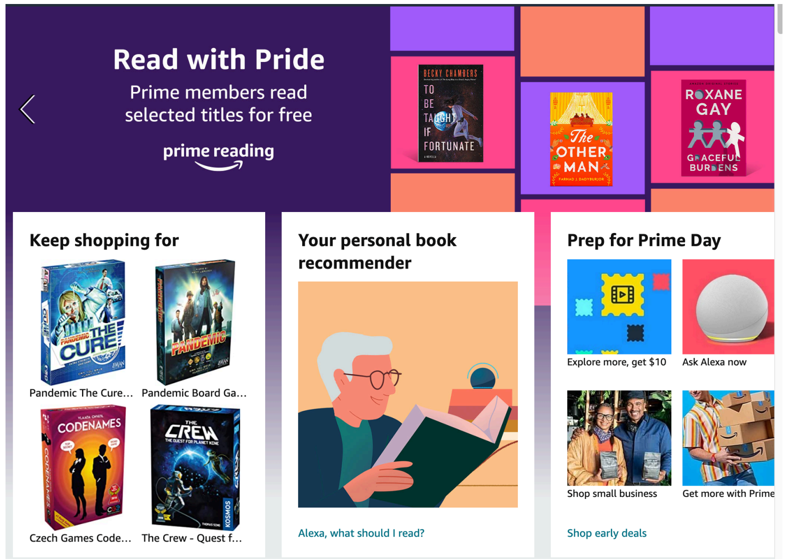
There exist efficient ALGs that achieve regret $O(\sqrt{|\mathcal{A}| T \log |\mathcal{F}|})$.

Large-scale recommendations

Large-scale recommendations

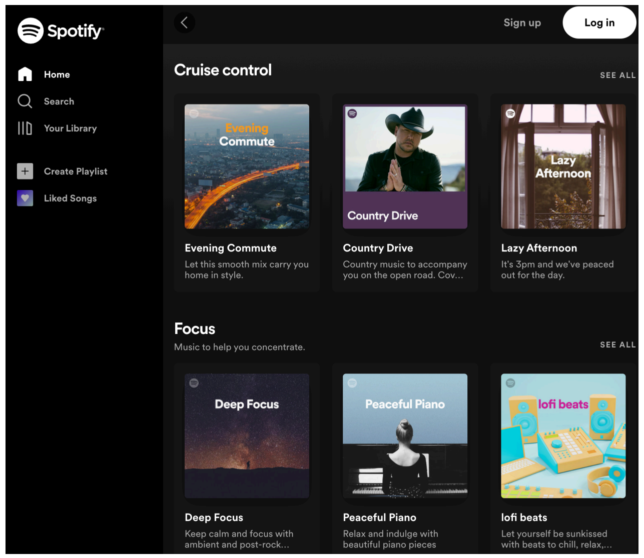


Spotify: 82 million songs

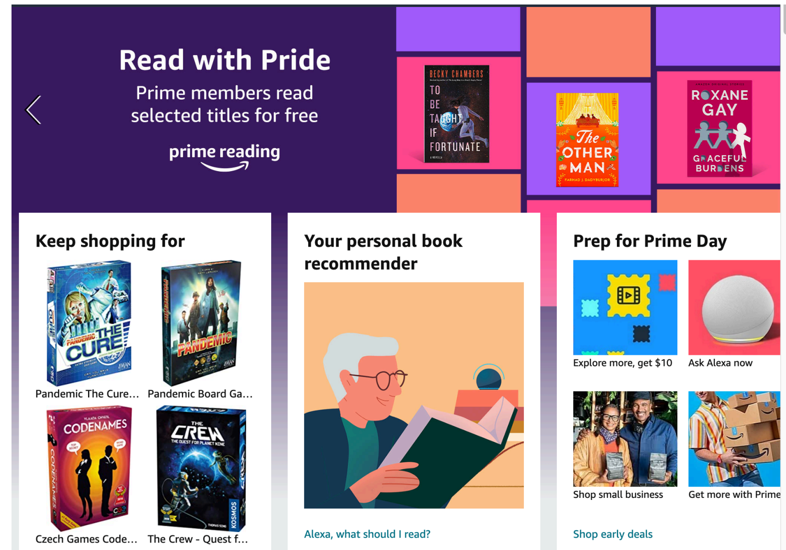


Amazon: 353 million commodities

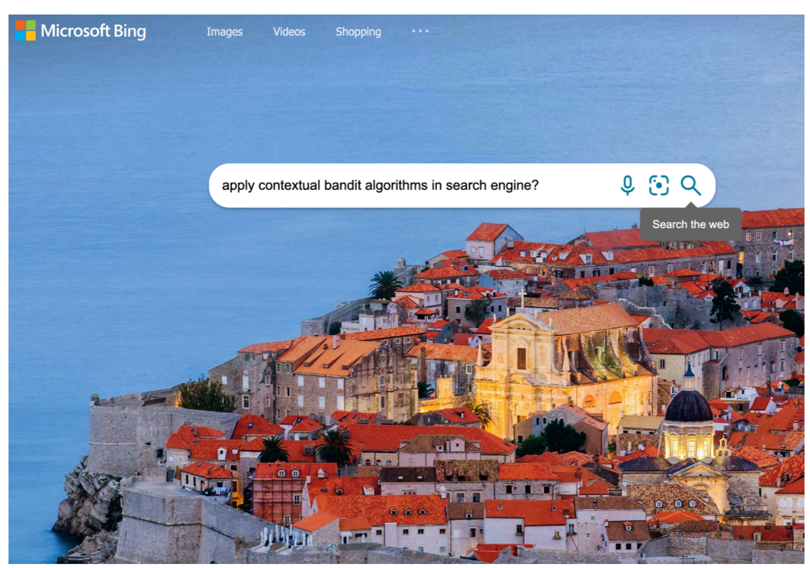
Large-scale recommendations



Spotify: 82 million songs

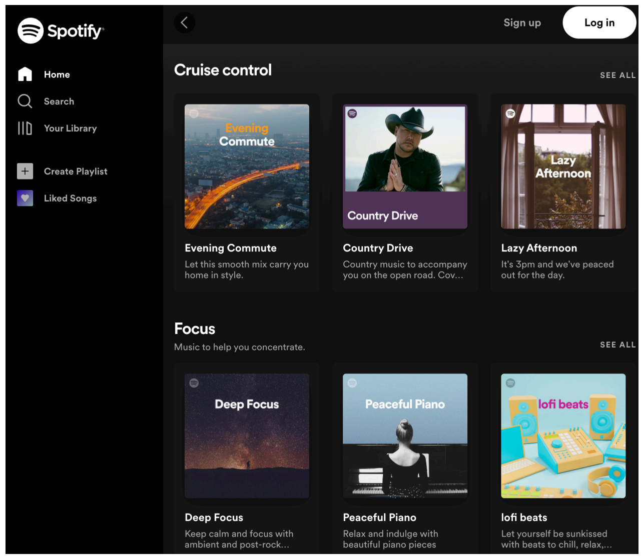


Amazon: 353 million commodities

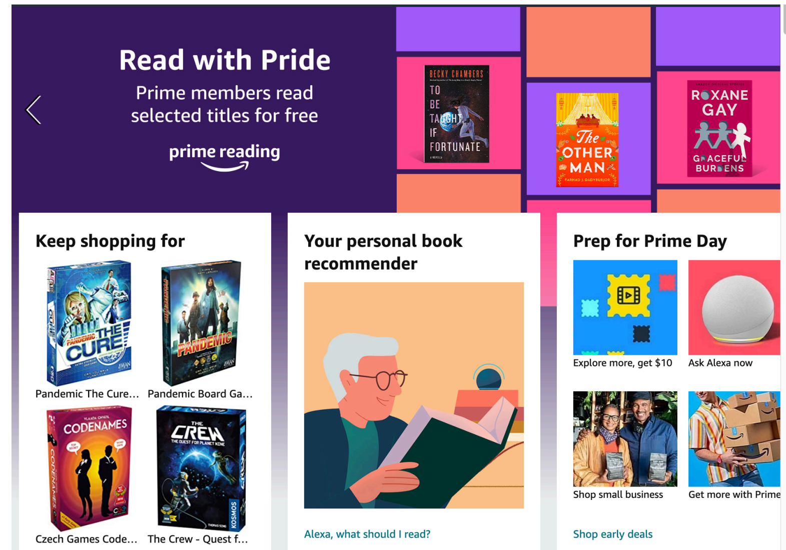


Search: dozens of billions of documents

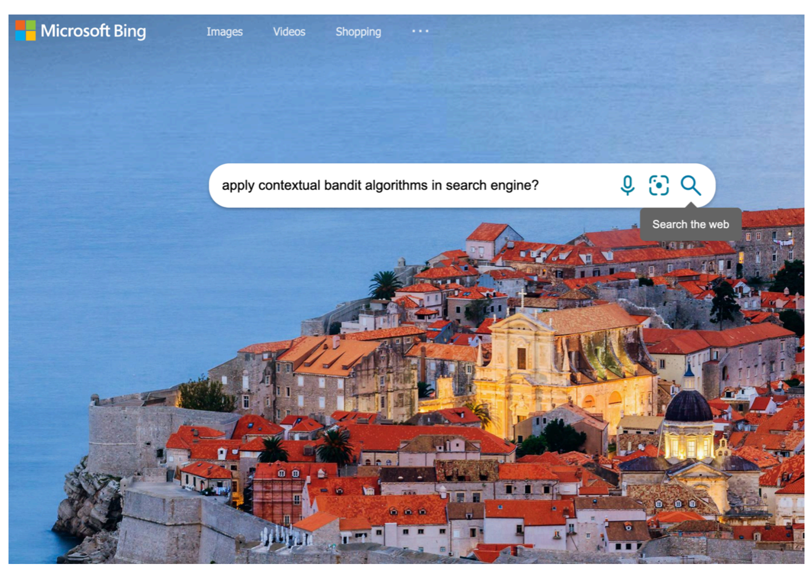
Large-scale recommendations



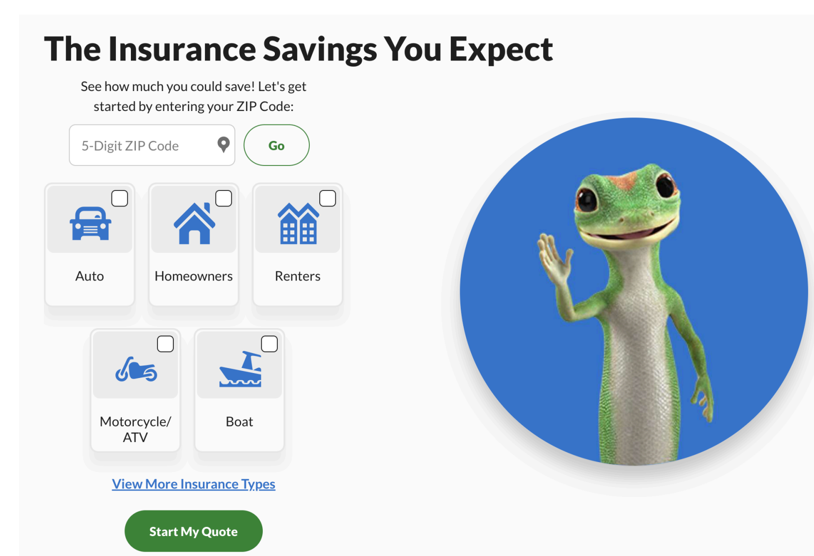
Spotify: 82 million songs



Amazon: 353 million commodities



Search: dozens of billions of documents



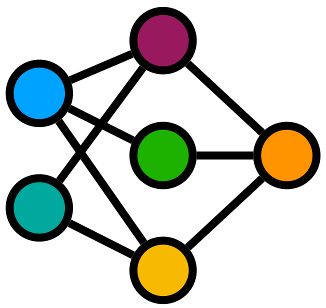
Personalized dynamic pricing: Continuous domain

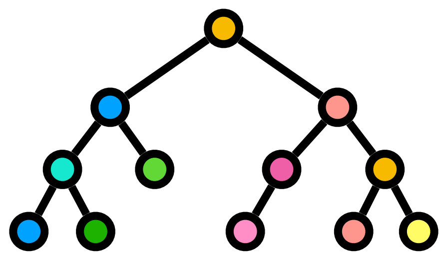
Existing guarantees

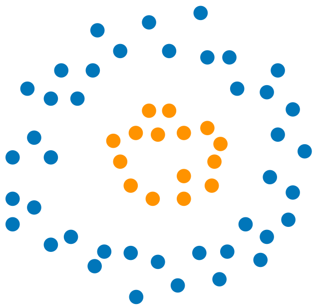
A standard realizability assumption

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

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







Theorem (Agarwal et al. 2012)

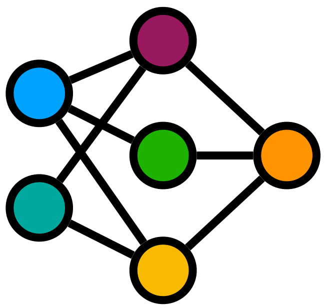
Any CB ALG must suffer worst-case regret $\Omega(\sqrt{|\mathcal{A}| T \log |\mathcal{F}|})$.

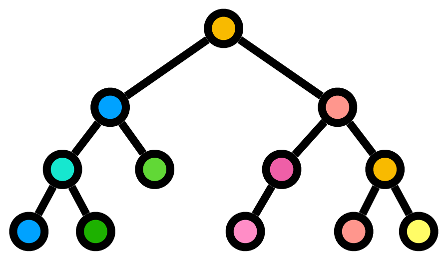
Existing guarantees

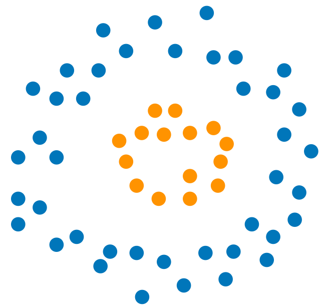
A standard realizability assumption

We assume $f^\star := \mathbb{E}[\ell_t \mid x_t] \in \mathscr{F}$ 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 $\Omega(\sqrt{|\mathscr{A}| T \log |\mathscr{F}|})$.

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

Adding additional structural assumptions

Adding additional structural assumptions

Linearity

f takes the form $f(x, a) := \langle \phi(x, a), \theta \rangle$ for an unknown $\theta \in \mathbb{R}^d$.

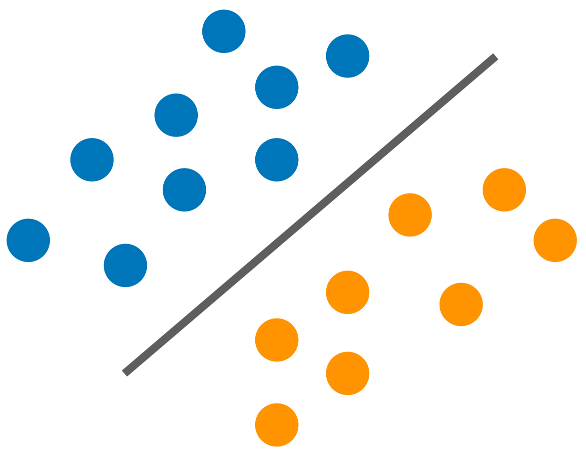
Studied in AL '99, Auer '02, CLRS '11, APS, '11, etc.

Adding additional structural assumptions

Linearity

f takes the form $f(x, a) := \langle \phi(x, a), \theta \rangle$ for an unknown $\theta \in \mathbb{R}^d$.

Studied in AL '99, Auer '02, CLRS '11, APS, '11, etc.



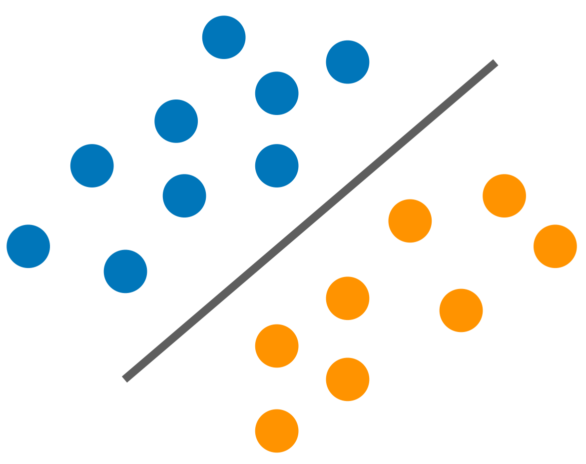
Leads to $d\sqrt{T}$ regret

Adding additional structural assumptions

Linearity

f takes the form $f(x, a) := \langle \phi(x, a), \theta \rangle$ for an unknown $\theta \in \mathbb{R}^d$.

Studied in AL '99, Auer '02, CLRS '11, APS, '11, etc.



Leads to $d\sqrt{T}$ regret

Lipschitzness

f is a 1-Lipschitz function.

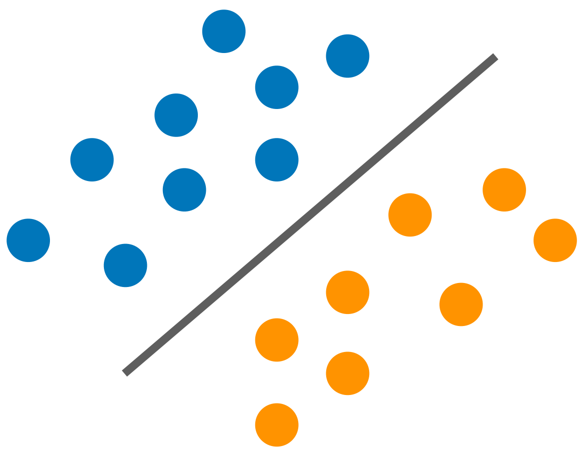
Studied in Agr '95, Kle '04, AOS '07, Sli, '14, etc.

Adding additional structural assumptions

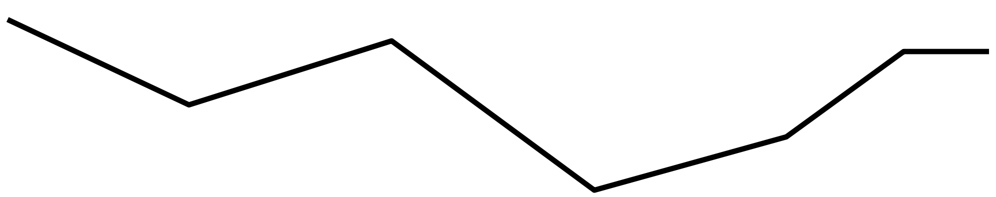
Linearity

f takes the form $f(x, a) := \langle \phi(x, a), \theta \rangle$ for an unknown $\theta \in \mathbb{R}^d$.

Studied in AL '99, Auer '02, CLRS '11, APS, '11, etc.



Leads to $d\sqrt{T}$ regret



Leads to $T^{2/3}$ regret

Lipschitzness

f is a 1-Lipschitz function.

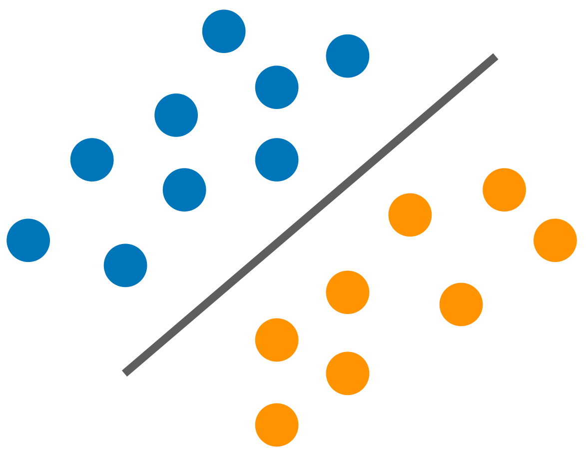
Studied in Agr '95, Kle '04, AOS '07, Sli, '14, etc.

Adding additional structural assumptions

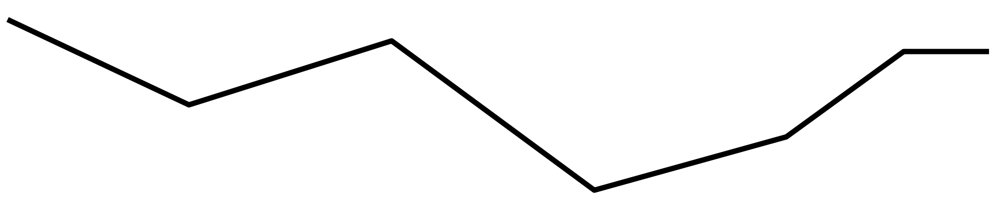
Linearity

f takes the form $f(x, a) := \langle \phi(x, a), \theta \rangle$ for an unknown $\theta \in \mathbb{R}^d$.

Studied in AL '99, Auer '02, CLRS '11, APS, '11, etc.



Leads to $d\sqrt{T}$ regret



Leads to $T^{2/3}$ regret

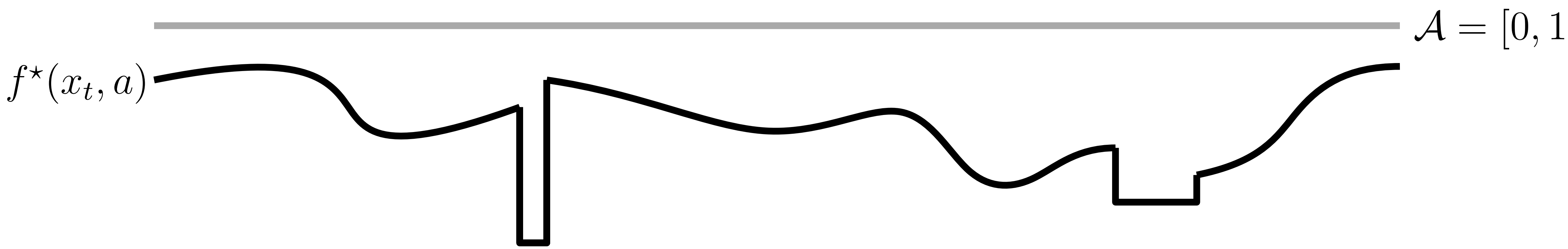
Lipschitzness

f is a 1-Lipschitz function.

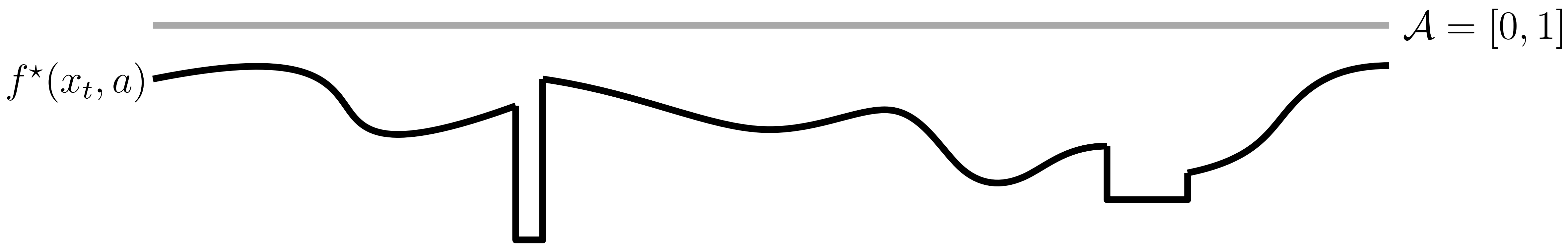
Studied in Agr '95, Kle '04, AOS '07, Sli, '14, etc.

Led to fruitful theoretical developments; but assumptions can be violated.

Beyond structural assumptions



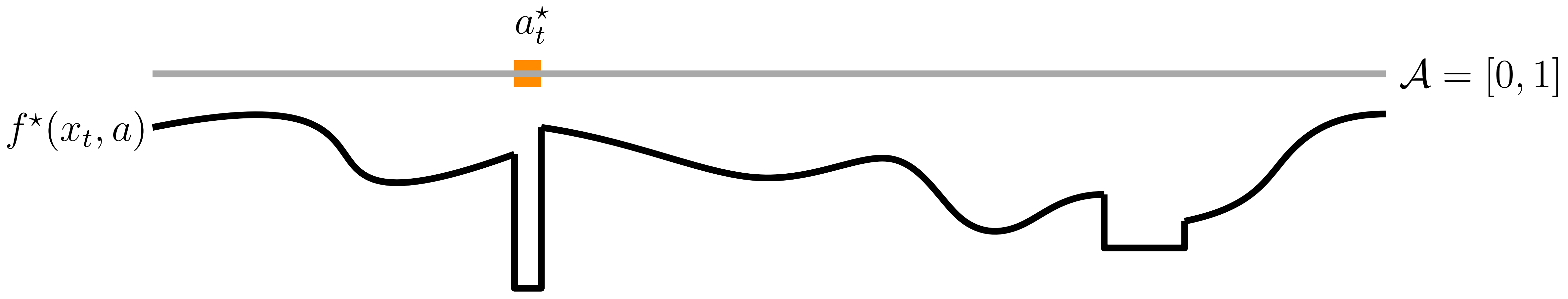
Beyond structural assumptions



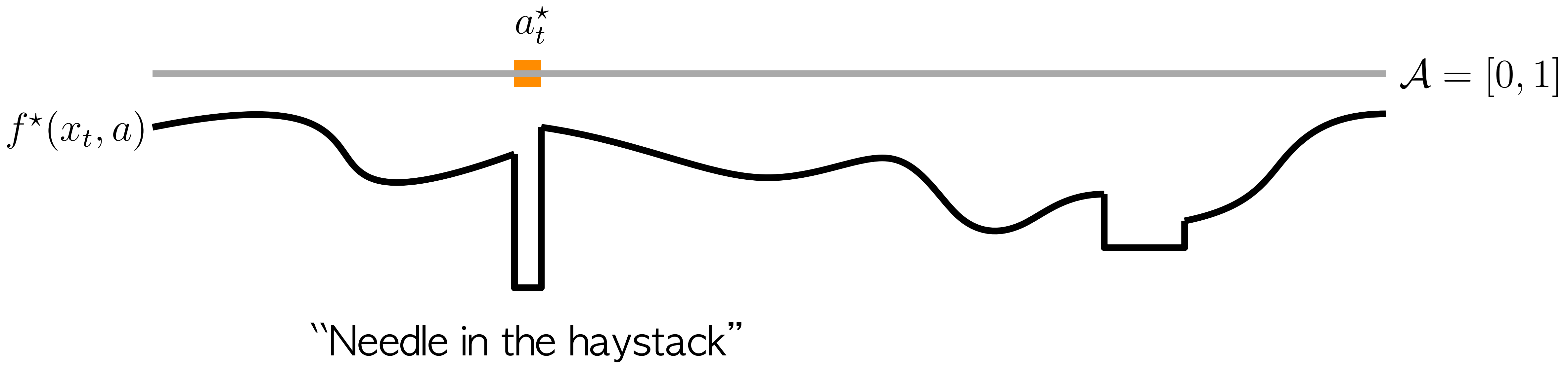
Difficulty: Need to handle general unstructured regression functions.

Competing against weaker benchmarks

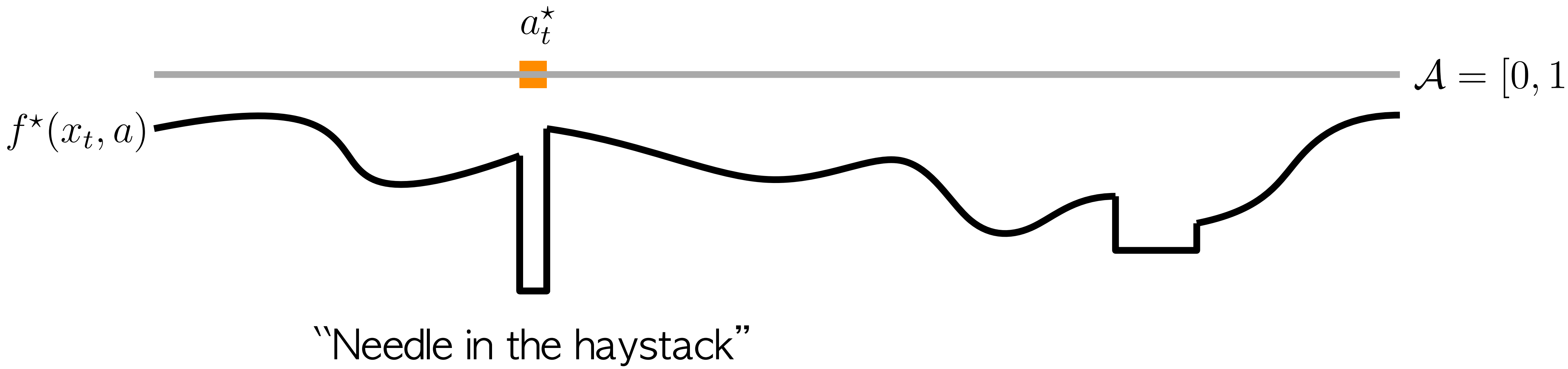
Competing against weaker benchmarks



Competing against weaker benchmarks



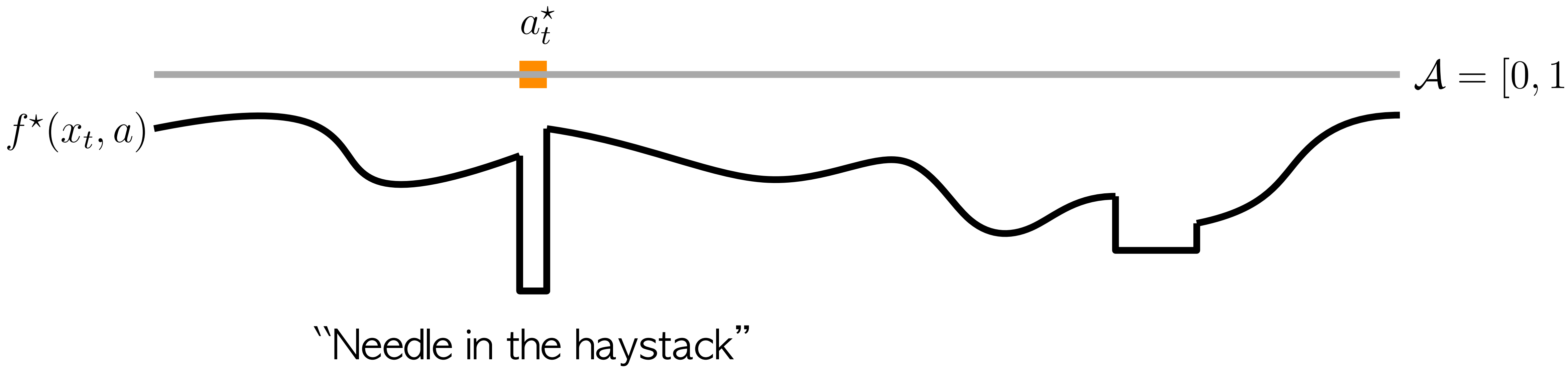
Competing against weaker benchmarks



Let μ be a base probability measure.

Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

Competing against weaker benchmarks



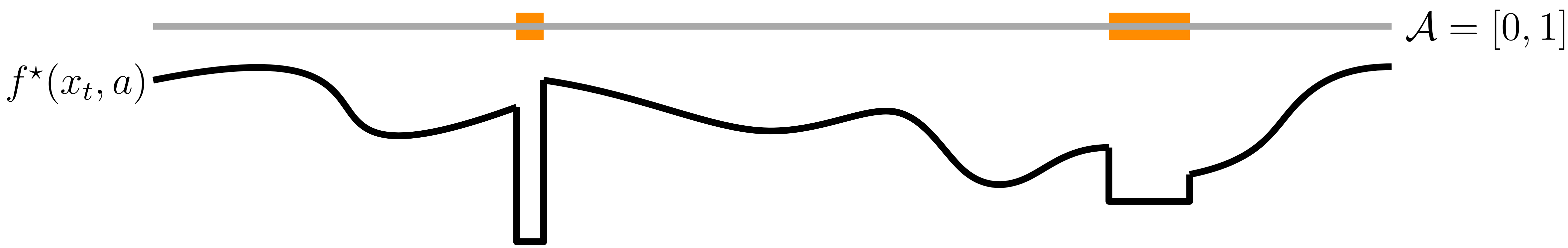
Let μ be a base probability measure.

Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

Compete against smoothed benchmark

$$\mathrm{Smooth}_h(x_t) := \inf_{Q \in \mathcal{Q}_h} \mathbb{E}_{a \sim Q}[f^\star(x_t, a)]$$

Competing against weaker benchmarks



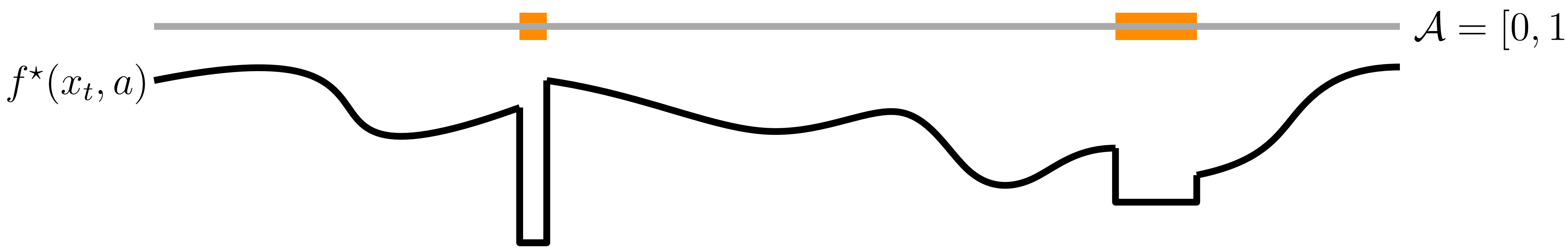
Let μ be a base probability measure.

Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

Compete against smoothed benchmark

$$\mathrm{Smooth}_h(x_t) := \inf_{Q \in \mathcal{Q}_h} \mathbb{E}_{a \sim Q}\left[f^\star(x_t, a)\right]$$

Competing against weaker benchmarks

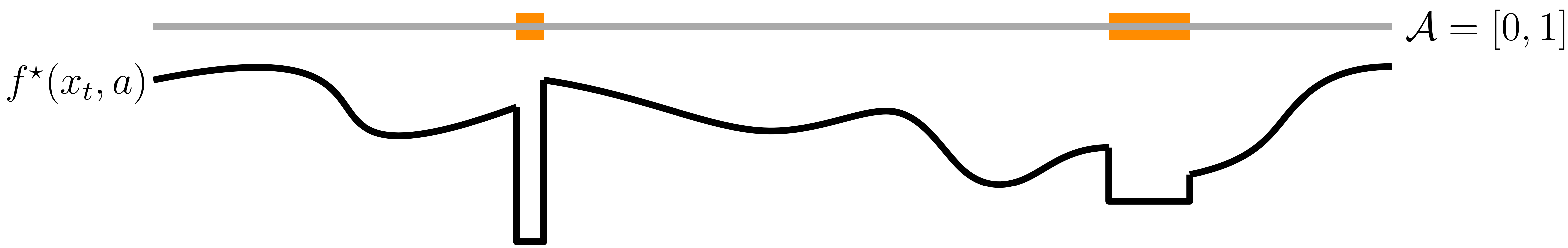


Let μ be a base probability measure.
Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

Compete against smoothed benchmark
$\text{Smooth}_h(x_t) := \inf_{Q \in \mathcal{Q}_h} \mathbb{E}_{a \sim Q}\left[f^\star(x_t, a)\right]$

Goal: Minimize smooth regret $\mathbf{Reg}_{\mathrm{CB},h}(T) := \sum_{t=1}^{T} f^\star(x_t, a_t) - \text{Smooth}_h(x_t)$

Competing against weaker benchmarks



Let μ be a base probability measure.
Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

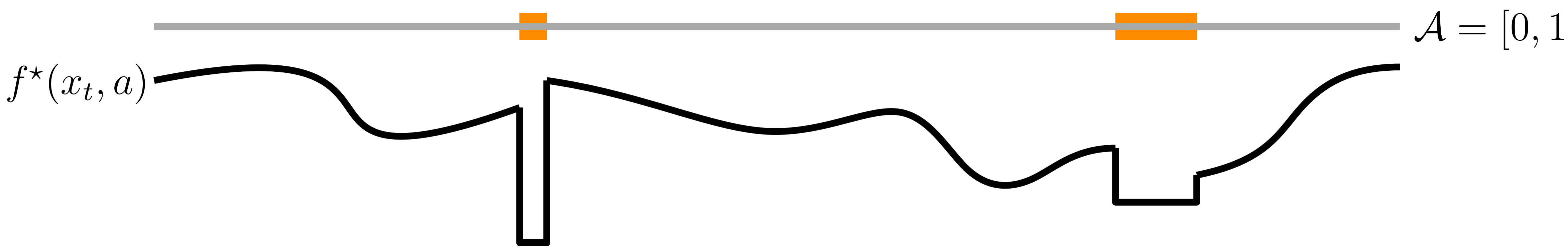
Compete against smoothed benchmark

$$\mathrm{Smooth}_h(x_t) := \inf_{Q \in \mathcal{Q}_h} \mathbb{E}_{a \sim Q}\left[f^\star(x_t, a)\right]$$

- Stronger than previous proposed smoothed benchmarks, e.g., Chaudhuri et al. 2018, Krishnamurthy et al. 2020.

Goal: Minimize smooth regret $\mathbf{Reg}_{\mathrm{CB},h}(T) := \sum_{t=1}^{T} f^\star(x_t, a_t) - \mathrm{Smooth}_h(x_t)$

Competing against weaker benchmarks



Let μ be a base probability measure.
Fix $h \in (0,1]$. Define $\mathcal{Q}_h := \{Q : dQ/d\mu \leq 1/h\}$

Compete against smoothed benchmark

$$\mathrm{Smooth}_h(x_t) := \inf_{Q \in \mathcal{Q}_h} \mathbb{E}_{a \sim Q}\left[f^\star(x_t, a)\right]$$

- Stronger than previous proposed smoothed benchmarks, e.g., Chaudhuri et al. 2018, Krishnamurthy et al. 2020.

Goal: Minimize smooth regret $\mathbf{Reg}_{\mathsf{CB},h}(T) := \sum_{t=1}^{T} f^\star(x_t, a_t) - \mathrm{Smooth}_h(x_t)$

- Recover minimax guarantees under standard regret and additional structural assumptions

Computational oracles

Computational oracles

Regression oracle

Online regression oracle such that

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

- $\mathbf{Reg}_{\mathsf{Sq}}(T) = O(\log|\mathscr{F}|)$ for general $\mathscr{F}$ using Vovk's aggregation algorithm (Vovk '98).
- Standard oracle studied in contextual bandits, e.g., FR '20, Zhang '21.

Computational oracles

Regression oracle

Online regression oracle such that

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

- $\mathsf{Reg}_{\mathsf{Sq}}(T)=O(\log|\mathcal{F}|)$ for general $\mathcal{F}$ using Vovk's aggregation algorithm (Vovk '98).
- Standard oracle studied in contextual bandits, e.g., FR '20, Zhang '21.

Sampling oracle

Sample action $a\sim\mu$ from the base probability measure μ.

- $O(\mathsf{H}(\mu))$ time to generate a random sample $a\sim\mu$ using DDG Tree (KY '76).

SquareCB for finite action

SquareCB (Foster et al. 2020)

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\widehat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\widehat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted prob. mass

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\widehat{a}_t := \arg\min_{a \in \mathscr{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted prob. mass

$$p_t(a) = \frac{1}{|\mathscr{A}| + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \widehat{a}_t)\right)}.$$

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\widehat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted prob. mass

$$p_t(a) = \frac{1}{|\mathcal{A}| + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \widehat{a}_t)\right)}.$$

- Sample action $a_t \sim p_t + (1 - p_t(\mathcal{A})) \cdot \mathbb{I}_{\widehat{a}_t}$.

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\widehat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted prob. mass

$$p_t(a) = \frac{1}{|\mathcal{A}| + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \widehat{a}_t)\right)}.$$

- Sample action $a_t \sim p_t + (1 - p_t(\mathcal{A})) \cdot \mathbb{I}_{\widehat{a}_t}$.
- Observe loss $\ell_t(a_t)$ and update regression oracle.

SquareCB for finite action

SquareCB (Foster et al. 2020)

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\hat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted RN derivative
$$\frac{dp_t}{d\mu}(a) = \frac{|\mathcal{A}|}{|\mathcal{A}| + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \hat{a}_t)\right)}.$$
- Sample action $a_t \sim p_t + (1 - p_t(\mathcal{A})) \cdot \mathbb{I}_{\hat{a}_t}$.
- Observe loss $\ell_t(a_t)$ and update regression oracle.

SmoothIGW for large action spaces

SmoothIGW

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\hat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted RN derivative

$$\frac{dp_t}{d\mu}(a) = \frac{1/h}{1/h + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \hat{a}_t)\right)}.$$

- Sample action $a_t \sim p_t + (1 - p_t(\mathcal{A})) \cdot \mathbb{I}_{\hat{a}_t}$.
- Observe loss $\ell_t(a_t)$ and update regression oracle.

SmoothIGW for large action spaces

SmoothIGW

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

- Obtain x_t from nature and $\hat{f}_t$ from regression oracle.
- Compute greedy action $\hat{a}_t := \arg\min_{a \in \mathcal{A}} \hat{f}_t(x_t, a)$.
- Construct inverse-gap-weighted RN derivative

$$\frac{dp_t}{d\mu}(a) = \frac{1/h}{1/h + \gamma \cdot \left(\hat{f}_t(x_t, a) - \hat{f}_t(x_t, \hat{a}_t)\right)}.$$

- Sample action $a_t \sim p_t + (1 - p_t(\mathcal{A})) \cdot \mathbb{I}_{\hat{a}_t}$.
- Observe loss $\ell_t(a_t)$ and update regression oracle.

Efficient rejection sampling

- Sample $\check{a}_t \sim \mu$ from base measure μ w/ sampling oracle.
- Sample Z from a Bernoulli dist. with mean $dp_t/d\mu(\check{a}_t)$.
- Play $\check{a}_t$ if $Z = 1$; play $\hat{a}_t$ otherwise.

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathscr{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log|\mathscr{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log|\mathcal{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
    - No structural assumptions on the model class.

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathcal{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathcal{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.
- Recover minimax guarantees under standard regret:

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathscr{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.

- Recover minimax guarantees under standard regret:
  - Discrete case w/ finite actions: Take $h = 1/|\mathscr{A}|$ leads to

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log|\mathcal{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.

- Recover minimax guarantees under standard regret:
  - Discrete case w/ finite actions: Take $h = 1/|\mathcal{A}|$ leads to

$$\mathrm{Reg}_{\mathrm{CB}}(T) = \Theta(\sqrt{|\mathcal{A}|T \log|\mathcal{F}|})$$

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathcal{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.

- Recover minimax guarantees under standard regret:
  - Discrete case w/ finite actions: Take $h = 1/|\mathcal{A}|$ leads to
  $$\mathsf{Reg}_{\mathsf{CB}}(T) = \Theta(\sqrt{|\mathcal{A}| T \log |\mathcal{F}|})$$
  - Continuous case under Hölder (Lipschitz) continuity w/ exponent α: Take $h = \tilde{O}(T^{-1/(2\alpha+1)})$ leads to

Theoretical guarantees

Theorem

Fix $h \in (0,1]$. SmoothIGW achieves $\sqrt{T/h \log |\mathscr{F}|}$ smooth regret, with per-round $O(1)$ calls to the regression/sampling oracles.

- An efficient ALG that works in large/continuous action spaces:
  - No structural assumptions on the model class.
  - $O(1/h)$ serves as the effective number of actions.

- Recover minimax guarantees under standard regret:
  - Discrete case w/ finite actions: Take $h = 1/|\mathscr{A}|$ leads to
  $$\mathrm{Reg}_{\mathrm{CB}}(T) = \Theta(\sqrt{|\mathscr{A}| T \log |\mathscr{F}|})$$
  - Continuous case under Hölder (Lipschitz) continuity w/ exponent α: Take $h = \tilde{O}(T^{-1/(2\alpha+1)})$ leads to
  $$\mathrm{Reg}_{\mathrm{CB}}(T) = \tilde{\Theta}(T^{(\alpha+1)/(2\alpha+1)})$$

An adaptive algorithm

An adaptive algorithm

Corral-SmoothIGW

- Initialize $O(\log T)$ base SmoothIGW, each with smoothness level $h_b = 2^{-b}$, for $b = 1, \ldots, O(\log(T))$.
- Apply the Corral (Agarwal et al. 2017) master ALG to balance over these base ALGs.

An adaptive algorithm

Corral-SmoothIGW

- Initialize $O(\log T)$ base SmoothIGW, each with smoothness level $h_b = 2^{-b}$, for $b = 1,\ldots,O(\log(T))$.
- Apply the Corral (Agarwal et al. 2017) master ALG to balance over these base ALGs.

- Inherit the computational efficiency of SmoothIGW up to $O(\log(T))$.
- Recover many known Pareto frontiers under standard regret:
  - bandits with unknown number of multiple best arms (ZN '20).
  - Hölder bandits with unknown smoothness parameter (Hadiji '19).

Empirical evaluations

Replicate the experiment setups from Majzoubi et al. 2020 on 5 OpenML regression datasets. CATS is the ALG proposed in Majzoubi et al. 2020.

Table 1. Average progressive loss, scaled by 1000, on continuous action contextual bandit datasets. 95% CIs reported.

	CATS	Ours (Linear)	Ours (RFF)
Cpu	$[55, 57]$	$[40.6, 40.7]$	$[\mathbf{38.6}, \mathbf{38.7}]$
Fri	$[183, 187]$	$[161, 163]$	$[\mathbf{156}, \mathbf{157}]$
Price	$[108, 110]$	$[70.2, 70.5]$	$[\mathbf{66.1}, \mathbf{66.3}]$
Wis	$[172, 174]$	$[138, 139]$	$[\mathbf{136.2}, \mathbf{136.6}]$
Zur	$[24, 26]$	$[24.3, 24.4]$	$[25.4, 25.5]$

Take-aways

Smooth regret is NOT a compromise

Facilitate the design of efficient ALGs