Taking a hint: how to leverage loss predictors in contextual bandits

Chen-Yu Wei (USC) Haipeng Luo (USC) Alekh Agarwal (MSR)

Contextual Bandit Setup

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$,

Contextual Bandit Setup

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$,
- learner receives x_t

Contextual Bandit Setup

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$,
- learner receives x_t
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Contextual Bandit Setup

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$,
- learner receives x_t
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Multi-armed bandit is a special case of contextual bandit (MAB = CB without contexts)

Regret Bounds

A policy π is a mapping: $\mathcal{X}$ (contexts) $\longrightarrow [K]$ (action)

Regret Bounds

A policy π is a mapping: $\mathcal{X}$ (contexts) $\longrightarrow [K]$ (action)

Suppose that the learner is given a fixed policy set Π. The goal of the learner is to be competitive w.r.t. the best policy in Π.

$$\mathrm{Reg} = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi(x_t))\right]$$

Regret Bounds

A policy π is a mapping: $\mathcal{X}$ (contexts) $\longrightarrow [K]$ (action)

Suppose that the learner is given a fixed policy set Π. The goal of the learner is to be competitive w.r.t. the best policy in Π.

$$\mathrm{Reg} = \max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi(x_t))\right]$$

Minimax regret is $\mathcal{O}(\sqrt{KT \ln |\Pi|})$ (we simplify it as $O(\sqrt{T})$)

- Exp4, ILOVETOCONBANDITS (ACFS'02, AHKLLS'14)

Regret Bounds

A policy π is a mapping: $\mathcal{X}$ (contexts) $\longrightarrow [K]$ (action)

Suppose that the learner is given a fixed policy set Π. The goal of the learner is to be competitive w.r.t. the best policy in Π.

$$\text{Reg} = \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t) - \sum_{t=1}^{T} \ell_t(\pi(x_t))\right]$$

Minimax regret is $\mathcal{O}(\sqrt{KT\ln|\Pi|})$ (we simplify it as $O(\sqrt{T})$)

- Exp4, ILOVETOCONBANDITS (ACFS'02, AHKLLS'14)

Question: Can we do better when the losses are *predictable*?

Contextual Bandit with Loss Predictors

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$,
- learner receives x_t
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Contextual Bandit with Loss Predictors

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and a loss predictor $m_t \in [0,1]^K$
- learner receives x_t and m_t
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Contextual Bandit with Loss Predictors

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and a loss predictor $m_t \in [0,1]^K$
- learner receives x_t and m_t
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Examples of m_t:

- $m_t(a) = f_\theta(x_t, a)$ (some learned or fixed loss regressor f_θ)
- $m_t(a) = \textbf{avg}(\widehat{\ell}_{t-\tau}(a), \ldots, \widehat{\ell}_{t-1}(a))$ (for slowly changing MAB)

Contextual Bandit with Loss Predictors

Key Q: if $\mathcal{E} = \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$ is small, can we improve over $\mathcal{O}(\sqrt{T})$?

(note: $\mathcal{E} \leq T$ always holds)

Contextual Bandit with Loss Predictors

Key Q: if $\mathcal{E} = \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$ is small, can we improve over $\mathcal{O}(\sqrt{T})$?
(note: $\mathcal{E} \le T$ always holds)

Previous work studied the full-information setting and the MAB setting (RS'13, WL'18), and get $\mathrm{Reg} = \mathcal{O}(\sqrt{\mathcal{E}})$.

Contextual Bandit with Loss Predictors

Key Q: if $\mathcal{E} = \sum_{t=1}^{T} \|\ell_t - m_t\|_\infty^2$ is small, can we improve over $\mathcal{O}(\sqrt{T})$?
(note: $\mathcal{E} \le T$ always holds)

Previous work studied the full-information setting and the MAB setting (RS'13, WL'18), and get $\text{Reg} = \mathcal{O}(\sqrt{\mathcal{E}})$.

Prior Works on Contextual Bandits:
Closely related to **doubly-robust** methods that use **loss estimators** m_t to reduce the variance of off-policy evaluation.
Theoretical benefits in the online exploration scenario? (no prior work)

A More General Setting

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and M loss predictors $m_t^1, \ldots, m_t^M \in [0,1]^K$
- learner receives x_t and $m_t^1, \ldots, m_t^M$
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

A More General Setting

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and M loss predictors $m_t^1, \ldots, m_t^M \in [0,1]^K$
- learner receives x_t and $m_t^1, \ldots, m_t^M$
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Key Q: if $\mathcal{E}^* = \min_i \sum_{t=1}^{T} \left\|\ell_t - m_t^i\right\|_\infty^2$ is small, can we improve over $\mathcal{O}(\sqrt{T})$?

A More General Setting

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

- environment chooses a context $x_t \in \mathcal{X}$, a loss vector $\ell_t \in [0,1]^K$, and M loss predictors $m_t^1, \ldots, m_t^M \in [0,1]^K$
- learner receives x_t and $m_t^1, \ldots, m_t^M$
- learner chooses action $a_t \in [K]$ and observes its loss $\ell_t(a_t)$

Key Q: if $\mathcal{E}^* = \min_i \sum_{t=1}^T \left\|\ell_t - m_t^i\right\|_\infty^2$ is small, can we improve over $\mathcal{O}(\sqrt{T})$?

MAB or full-information setting: $\mathrm{Reg} = \mathcal{O}(\sqrt{\mathcal{E}^* + \ln M})$ (RS'13)

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \le \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \ge \sqrt{T} \end{cases} \quad = \quad \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \leq \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \geq \sqrt{T} \end{cases} \quad = \quad \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

cf. MAB or full-info: $\sqrt{\mathcal{E}}$

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}} T^{\frac{1}{4}} & \text{when } \mathcal{E} \le \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \ge \sqrt{T} \end{cases} = \min\left\{\sqrt{\mathcal{E}} T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

cf. MAB or full-info: $\sqrt{\mathcal{E}}$

- The tight bound is unachievable if the learner does not know $\mathcal{E}$: we show $\omega(\sqrt{\mathcal{E}} T^{\frac{1}{4}})$ and $O(\sqrt{\mathcal{E}} T^{\frac{1}{3}})$ for unknown $\mathcal{E}$.

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \le \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \ge \sqrt{T} \end{cases} \quad = \quad \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

cf. MAB or full-info: $\sqrt{\mathcal{E}}$

- The tight bound is unachievable if the learner does not know $\mathcal{E}$: we show $\omega(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ and $O(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ for unknown $\mathcal{E}$.

cf. MAB or full-info: tight bound is achievable without knowing $\mathcal{E}$

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \le \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \ge \sqrt{T} \end{cases} = \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

  cf. MAB or full-info: $\sqrt{\mathcal{E}}$

- The tight bound is unachievable if the learner does not know $\mathcal{E}$: we show $\omega(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ and $O(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ for unknown $\mathcal{E}$.

  cf. MAB or full-info: tight bound is achievable without knowing $\mathcal{E}$

- For $M > 1$ (multiple predictor case): we show $\Omega(\sqrt{\mathcal{E}^*}T^{\frac{1}{4}} + M)$ and $O(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}})$

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \le \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \ge \sqrt{T} \end{cases} = \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

  cf. MAB or full-info: $\sqrt{\mathcal{E}}$

- The tight bound is unachievable if the learner does not know $\mathcal{E}$: we show $\omega(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ and $O(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ for unknown $\mathcal{E}$.

  cf. MAB or full-info: tight bound is achievable without knowing $\mathcal{E}$

- For $M > 1$ (multiple predictor case): we show $\Omega(\sqrt{\mathcal{E}^*}T^{\frac{1}{4}} + M)$ and $O(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}})$

  cf. MAB or full-info: $O(\ln M)$ overhead

Our results

- Regret tight bound (in $\Theta(\cdot)$) when $M = 1$:

$$\begin{cases} \sqrt{\mathcal{E}}T^{\frac{1}{4}} & \text{when } \mathcal{E} \leq \sqrt{T} \\ \sqrt{T} & \text{when } \mathcal{E} \geq \sqrt{T} \end{cases} \quad = \quad \min\left\{\sqrt{\mathcal{E}}T^{\frac{1}{4}}, \sqrt{T}\right\}.$$

  cf. MAB or full-info: $\sqrt{\mathcal{E}}$

- The tight bound is unachievable if the learner does not know $\mathcal{E}$: we show $\omega(\sqrt{\mathcal{E}}T^{\frac{1}{4}})$ and $O(\sqrt{\mathcal{E}}T^{\frac{1}{3}})$ for unknown $\mathcal{E}$.

  cf. MAB or full-info: tight bound is achievable without knowing $\mathcal{E}$

- For $M > 1$ (multiple predictor case): we show $\Omega(\sqrt{\mathcal{E}^*}T^{\frac{1}{4}} + M)$ and $O(\sqrt{M\mathcal{E}^*}T^{\frac{1}{4}})$

  cf. MAB or full-info: $O(\ln M)$ overhead

For all upper bound results, we give 1) algorithms for general adversarial sequences, and 2) ERM oracle-efficient algorithms for i.i.d. sequences.

Adversarial Setting + Single Predictor + Known $\mathcal{E}$

Algorithm

EXP4

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

- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q'_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a)-m_t(a))\mathbf{1}[a_t=a]}{p_t(a)} + m_t(a)$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a) - m_t(a))\mathbf{1}[a_t = a]}{p_t(a)} + m_t(a)$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a) - m_t(a))\mathbf{1}[a_t = a]}{p_t(a)} + m_t(a)$

$$\mathrm{Reg} \leq \frac{\ln|\Pi|}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^{T} p_t(a_t)(\widehat{\ell}_t(a_t) - m_t(a_t))^2\right]$$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a)-m_t(a))\mathbf{1}[a_t=a]}{p_t(a)} + m_t(a)$

$$\mathrm{Reg} \le \frac{\ln|\Pi|}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^{T} p_t(a_t)(\widehat{\ell}_t(a_t) - m_t(a_t))^2\right] = \mathcal{O}(\sqrt{\mathcal{E}}) \quad ??$$

Algorithm

Optimistic EXP4

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

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a) - m_t(a))\mathbf{1}[a_t = a]}{p_t(a)} + m_t(a)$

$$\mathrm{Reg} \le \frac{\ln |\Pi|}{\eta} + 2\eta \mathbb{E}\left[\sum_{t=1}^{T} p_t(a_t)(\widehat{\ell}_t(a_t) - m_t(a_t))^2\right] = \mathcal{O}(\sqrt{\mathcal{E}}) \quad ??$$

issue: requires $\widehat{\ell}_t(a_t) - m_t(a_t) \ge -\frac{1}{\eta}$

Algorithm

Optimistic EXP4

For $t = 1, \dots, T$:

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = \sum_{\pi:\pi(x_t)=a} Q_t(\pi)$ and sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

loss estimator: $\widehat{\ell}_t(a) = \frac{(\ell_t(a)-m_t(a))\mathbf{1}[a_t=a]}{p_t(a)} + m_t(a)$

$$\text{Reg} \leq \frac{\ln|\Pi|}{\eta} + 2\eta\mathbb{E}\left[\sum_{t=1}^{T} p_t(a_t)(\widehat{\ell}_t(a_t) - m_t(a_t))^2\right] = \mathcal{O}(\sqrt{\mathcal{E}}) \quad ??$$

issue: requires $\widehat{\ell}_t(a_t) - m_t(a_t) \geq -\frac{1}{\eta}$

naive fix: ensure $p_t(a) \geq \eta$ via uniform exploration $\Rightarrow \Omega(\sqrt{T})$ regret :(

Key Technique: Action Remapping

For $t = 1, \dots, T$:

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = (1-\mu) \sum_{\pi:\pi(x_t)=a} Q_t(\pi) + \frac{\mu}{K}$
- sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Key Technique: Action Remapping

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

- define **"baseline"** $a_t^* = \operatorname{argmin}_a m_t(a)$,
- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = (1-\mu) \sum_{\pi:\pi(x_t)=a} Q_t(\pi) + \frac{\mu}{K}$
- sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Key Technique: Action Remapping

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

- define **"baseline"** $a_t^* = \operatorname{argmin}_a m_t(a)$,

$$\mathcal{A}_t = \{a : m_t(a) \leq m_t(a_t^*) + \sigma\},$$

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = (1-\mu) \sum_{\pi:\pi(x_t)=a} Q_t(\pi) + \frac{\mu}{K}$
- sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Key Technique: Action Remapping

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

- define **"baseline"** $a_t^* = \mathrm{argmin}_a\, m_t(a)$,

$$\mathcal{A}_t = \{a : m_t(a) \le m_t(a_t^*) + \sigma\}, \quad \phi_t(a) = \begin{cases} a, & \text{if } a \in \mathcal{A}_t \\ a_t^*, & \text{else} \end{cases}$$

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\pi(x_t)))$
- compute $p_t(a) = (1-\mu) \sum_{\pi:\pi(x_t)=a} Q_t(\pi) + \frac{\mu}{K}$
- sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\pi(x_t)))$

Key Technique: Action Remapping

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

- define **"baseline"** $a_t^* = \operatorname{argmin}_a m_t(a)$,

$$\mathcal{A}_t = \{a : m_t(a) \le m_t(a_t^*) + \sigma\}, \quad \phi_t(a) = \begin{cases} a, & \text{if } a \in \mathcal{A}_t \\ a_t^*, & \text{else} \end{cases}$$

- compute $Q_t(\pi) \propto Q'_t(\pi) \exp(-\eta m_t(\phi_t(\pi(x_t))))$
- compute $p_t(a) = (1-\mu) \sum_{\pi : \phi_t(\pi(x_t)) = a} Q_t(\pi) + \frac{\mu}{|\mathcal{A}_t|} \mathbf{1}[a \in \mathcal{A}_t]$
- sample $a_t \sim p_t$
- compute $Q'_{t+1}(\pi) \propto Q'_t(\pi) \exp(-\eta \widehat{\ell}_t(\phi_t(\pi(x_t))))$

Intuition

1. $\mathcal{A}_t$ excludes actions whose $m_t(a)$ are bad (i.e., large).

Intuition

1. $\mathcal{A}_t$ excludes actions whose $m_t(a)$ are bad (i.e., large).
2. Because $\mathcal{E} = \sum_t \|\ell_t - m_t\|_\infty^2$ is small, $\ell_t(a)$ for $a \notin \mathcal{A}_t$ are also generally bad.

Intuition

1. $\mathcal{A}_t$ excludes actions whose $m_t(a)$ are bad (i.e., large).
2. Because $\mathcal{E} = \sum_t \|\ell_t - m_t\|_\infty^2$ is small, $\ell_t(a)$ for $a \notin \mathcal{A}_t$ are also generally bad.
3. Because of 2., it suffices to explore the actions in $\mathcal{A}_t$ (this reduces the regret overhead due to exploration compared to standard uniform exploration).

i.i.d. Setting + Single Predictor + Known $\mathcal{E}$

$$(x_t, \ell_t, m_t) \sim \mathcal{D}$$

+ Oracle efficient

ϵ-Greedy and Oracle-efficiency

ERM-oracle: $\operatorname{argmin}_{\pi \in \Pi} \sum_{s=1}^{t} \widehat{\ell}_s(x_s, \pi(x_s))$

ϵ-Greedy and Oracle-efficiency

ERM-oracle: $\operatorname{argmin}_{\pi \in \Pi} \sum_{s=1}^{t} \widehat{\ell}_s(x_s, \pi(x_s))$

The simplest oracle-efficient CB algorithm: ϵ-Greedy

ϵ-Greedy and Oracle-efficiency

ERM-oracle: $\operatorname{argmin}_{\pi \in \Pi} \sum_{s=1}^{t} \widehat{\ell}_s(x_s, \pi(x_s))$

The simplest oracle-efficient CB algorithm: ϵ-Greedy

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

- find $\pi_t = \mathrm{ERM}(\{x_s, \widehat{\ell}_s\}_{s<t})$ (1 ERM-Oracle call)
- compute $p_t(a) = (1-\mu)\mathbf{1}[a = \pi_t(x_t)] + \frac{\mu}{K}$
- sample $a_t \sim p_t$ and construct loss estimator $\widehat{\ell}_t$

ϵ-Greedy and Oracle-efficiency

ERM-oracle: $\operatorname{argmin}_{\pi\in\Pi} \sum_{s=1}^{t} \widehat{\ell}_s(x_s, \pi(x_s))$

The simplest oracle-efficient CB algorithm: ϵ-Greedy

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

- find $\pi_t = \mathrm{ERM}(\{x_s, \widehat{\ell}_s\}_{s<t})$ (1 ERM-Oracle call)
- compute $p_t(a) = (1-\mu)\mathbf{1}[a = \pi_t(x_t)] + \frac{\mu}{K}$
- sample $a_t \sim p_t$ and construct loss estimator $\widehat{\ell}_t$

ϵ-Greedy with Action Remapping and Catoni's Estimator

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

- find $\pi_t = \operatorname{argmin}_{\pi} \mathrm{Catoni}(\{x_s, \widehat{\ell}_s\}_{s<t})$
- compute $p_t(a) = (1-\mu)\mathbf{1}[a = \phi_t(\pi_t(x_t))] + \frac{\mu}{|\mathcal{A}_t|}\mathbf{1}[a \in \mathcal{A}_t]$
- sample $a_t \sim p_t$ and construct loss estimator $\widehat{\ell}_t$

ϵ-Greedy and Oracle-efficiency

ERM-oracle: $\mathrm{argmin}_{\pi \in \Pi} \sum_{s=1}^{t} \widehat{\ell}_s(x_s, \pi(x_s))$

The simplest oracle-efficient CB algorithm: ϵ-Greedy

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

- find $\pi_t = \mathrm{ERM}(\{x_s, \widehat{\ell}_s\}_{s<t})$ (1 ERM-Oracle call)
- compute $p_t(a) = (1-\mu)\mathbf{1}[a = \pi_t(x_t)] + \frac{\mu}{K}$
- sample $a_t \sim p_t$ and construct loss estimator $\widehat{\ell}_t$

ϵ-Greedy with Action Remapping and Catoni's Estimator

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

- find $\pi_t = \mathrm{argmin}_{\pi} \mathrm{Catoni}(\{x_s, \widehat{\ell}_s\}_{s<t})$ ($\ln T$ ERM-Oracle call)
- compute $p_t(a) = (1-\mu)\mathbf{1}[a = \phi_t(\pi_t(x_t))] + \frac{\mu}{|\mathcal{A}_t|}\mathbf{1}[a \in \mathcal{A}_t]$
- sample $a_t \sim p_t$ and construct loss estimator $\widehat{\ell}_t$

Summary

- We initiate the study of online contextual bandits with loss predictors.

Summary

- We initiate the study of online contextual bandits with loss predictors.
- For the single predictor $(M = 1)$ case, we give complete answers (matching regret lower and upper bound).

Summary

- We initiate the study of online contextual bandits with loss predictors.
- For the single predictor $(M = 1)$ case, we give complete answers (matching regret lower and upper bound).
- There are sharp contrasts between CB and MAB
  - regret dependence on $\mathcal{E}$
  - regret dependence on M
  - whether the prior knowledge of $\mathcal{E}$ matters

Summary

- We initiate the study of online contextual bandits with loss predictors.
- For the single predictor $(M = 1)$ case, we give complete answers (matching regret lower and upper bound).
- There are sharp contrasts between CB and MAB
  - regret dependence on $\mathcal{E}$
  - regret dependence on M
  - whether the prior knowledge of $\mathcal{E}$ matters
- Future work: empirical evaluation of our algorithms