Structured Bandits

Baiyu Wang

University of Southampton

Multi-Armed Bandit Protocol

for $t = 1, ..., T$ **do**
Select decision $\pi^t \in \Pi := \{1, ..., A\}$.
Observe reward $r^t \in \mathcal{R}$

Contextual Bandit Protocol

for $t = 1, ..., T$ **do**
Observe context $x^t \in \mathcal{X}$
Select decision $\pi^t \in \Pi := \{1, ..., A\}$.
Observe reward $r^t \in \mathcal{R}$

Measure performance for Bandit problem:

$$\textbf{Reg} := \sum_{t=1}^{T} f^\star(\pi^\star) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(\pi^t)]$$

Practical Issue

- In medicine, the treatment may be a continuous variable, such as a dosage. The treatment could even by a high-dimensional vector (such as dosages for many different medications).
- In pricing applications, a seller might aim to select a continuous price or vector in order to maximize their returns.
- In routing applications, the decision space may be finite, but combinatorially large. For example, the decision might be a path or flow in a graph.

$\Rightarrow$ Decision space Π is not finite set.

Regret Bound

Multi-Armed Bandit:

- ε-Greedy algorithm: $\mathbf{Reg} \lesssim A^{1/3} T^{2/3} \cdot \log^{1/3}(AT/\delta)$.
- UCB algorithm: $\mathbf{Reg} \lesssim \sqrt{AT \log(AT/\delta)}$.
- Posterior Sampling Algorithm: $\mathbf{Reg} \lesssim \sqrt{AT \log(A)}$ / $\sqrt{AT \log |\mathcal{F}|}$
- Exp3 Algorithm: $\mathbf{Reg} \lesssim \sqrt{AT \log A}$

Contextual Bandit:

- ε-Greedy algorithm: $\mathbf{Reg} \lesssim A^{1/3} T^{2/3} \cdot \mathbf{Est}_{\mathsf{Sq}}(\mathcal{F}, T, \delta)^{1/3}$.
- LinUCB algorithm: $\mathbf{Reg} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta) \log(1 + T/d)}$

Necessity of structural assumptions

Let $\Pi = [A]$, and let $\mathcal{F} = \{f_i\}_{i \in [A]}$, where

$$f_i(\pi) := \frac{1}{2} + \frac{1}{2}\mathbb{I}\{\pi = i\}.$$

- **Reg** $\gtrsim A$ and $\log|\mathcal{F}| = \log(A)$ for this setting.
- **Reg** $\lesssim \sqrt{AT\log|\mathcal{F}|}$ is impossible if $A \gg T$.

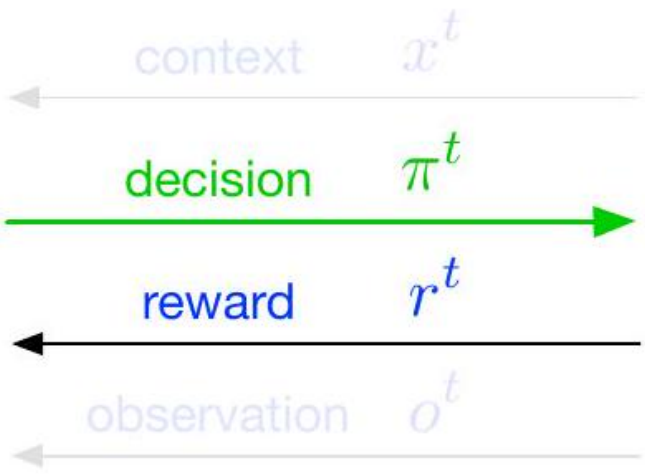
Structured Bandit

Structured Bandit Protocol

for $t = 1, ..., T$ **do**

Select decision $\pi^t \in \Pi$. Π *is large and potentially continuous.*

Observe reward $r^t \in \mathbb{R}$.



Assumption and Regret

- Decision: $\pi^t \in \Pi$, for $t = 1, ..., T$.
- Stochastic rewards: Rewards are generated independently via $r^t \sim M^\star(\cdot|\pi^t)$.
- Mean reward function: $f^\star(\pi) := \mathbb{E}[r|\pi]$.
- Regret function: $\mathbf{Reg} := \sum_{t=1}^{T} f^\star(\pi^\star) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^\star(\pi^t)]$, where $\pi^\star := \arg\max_{\pi \in \Pi} f^\star(\pi)$.
- Data history: $\mathcal{H}^t = (\pi^1, r^1), \ldots, (r^t, \pi^t)$.
- The decision-maker has access to a class $\mathcal{F} \subset \{f : \Pi \to \mathbb{R}\}$ such that $f^\star \in \mathcal{F}$.

Upper Confidence Bound

Assumption and definition:

- Assume $\mathcal{F} = \{f : \Pi \to [0,1]\}$ and $r^t \in [0,1]$ almost surely.
- $\widehat{f}^t = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{t-1} \left(f\left(\pi^i\right) - r^i\right)^2$.
- Confidence sets $\mathcal{F}^1 = \mathcal{F}$ and

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} : \sum_{i=1}^{t-1} \left(f\left(\pi^i\right) - r^i\right)^2 \leq \sum_{i=1}^{t-1} \left(\widehat{f}^t\left(\pi^i\right) - r^i\right)^2 + \beta \right\},$$

  where $\beta := 8 \log(|\mathcal{F}|/\delta)$.
- $\bar{f}^t(\pi) := \max_{f \in \mathcal{F}^t} f(\pi)$ is the upper confidence bound.

Generalized UCB algorithm:

$$\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$$

Upper Confidence Width

Lemma 10

Let $\pi^1, ..., \pi^T$ be chosen by an arbitrary (and possibly randomized) decision-making algorithm. With probability at least $1-\delta$, $f^\star \in \mathcal{F}^t$ for all $t \in [T]$. Moreover, with probability at least $1-\delta$, for all $\tau \leq T$, all $f \in \mathcal{F}^\tau$ with $\beta = 8\log(|\mathcal{F}/\delta|)$ satisfy

$$\sum_{t=1}^{\tau} \mathbb{E}_{\pi^t \sim p^t}\left[(f(x^t,\pi^t) - f^\star(x^t,\pi^t))^2\right] \leq 4\beta.$$

From Lemma 10, $f^\star \in \mathcal{F}$ with high probability. The regret is bounded by upper confidence width

$$\mathbf{Reg} \leq \sum_{t=1}^{T} \left(\bar{f}^t(\pi^t) - f^\star(\pi^t)\right).$$

When will the confidence shrink?

LinearUCB Algorithm

We hope Π and $\mathcal{F}$ have nice structure to get a better regret bound.

Recall linearUCB algorithm:

$$\mathcal{F} = \left\{\pi \mapsto \langle \theta, \phi(\pi)\rangle \mid \theta \in \Theta \subset \mathrm{B}_2^d(1)\right\}$$

From Proposition 7 we know $\mathbf{Reg} \lesssim \sqrt{dT \log |\mathcal{F}|}$.

Is there a general version when we move beyond linear model?

Definition

Eulder Dimension

$\mathcal{F} \subset (\Pi \to \mathbb{R})$ and $f^\star : \Pi \to \mathbb{R}$ be given, and define $\underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \varepsilon)$ as the length of the longest sequence of decisions $\pi^1, \ldots, \pi^d \in \Pi$ such that for all $t \in [d]$, there exists $f^t \in \mathcal{F}$ such that

$$\left|f^t\left(\pi^t\right) - f^\star\left(\pi^t\right)\right| > \varepsilon, \quad \text{and} \quad \sum_{i<t}\left(f^t\left(\pi^i\right) - f^\star\left(\pi^i\right)\right)^2 \le \varepsilon^2$$

Eluder dimension: $\mathrm{Edim}_{f^\star}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \ge \varepsilon} \underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \varepsilon') \vee 1$,
$\mathrm{Edim}(\mathcal{F}, \varepsilon) = \max_{f^\star \in \mathcal{F}} \mathrm{Edim}_{f^\star}(\mathcal{F}, \varepsilon)$.

- When $\beta = \varepsilon^2$ in confidence set, then

$$\sum_{t=1}^{T} \mathbb{I}\left\{\bar{f}^t(\pi^t) - f^\star(\pi^t) > \varepsilon\right\} \le \mathrm{Edim}_{f^\star}(\mathcal{F}, \varepsilon).$$

- Monotonicity with respect to ε.

Regret Bound for UCB

Regret Bound

For a finite set of functions $\mathcal{F} \subset (\Pi \to [0,1])$, using $\beta = 8\log(|\mathcal{F}|/\delta)$, the generalized UCB algorithm guarantees that with probability at least $1-\delta$,

$$\begin{aligned}\mathbf{Reg} &\lesssim \min_{\varepsilon > 0}\{\sqrt{\mathrm{Edim}(\mathcal{F}, \varepsilon) \cdot T\log(|\mathcal{F}|/\delta)} + \varepsilon T\} \\ &\lesssim \sqrt{\mathrm{Edim}\left(\mathcal{F}, T^{-1/2}\right) \cdot T\log(|\mathcal{F}|/\delta)}\end{aligned}$$

Proof of Regret Bound

Define

$$\overline{\mathcal{F}}^t = \left\{ f \in \mathcal{F} \mid \sum_{i<t} \left(f\left(\pi^i\right) - f^\star\left(\pi^i\right) \right)^2 \leq 4\beta \right\}.$$

By Lemma 10, $f^\star \in \mathcal{F}^t$ and $\mathcal{F}^t \subseteq \overline{\mathcal{F}}^t$ for $t = 1, ..., T$ with probability at least $1 - \delta$.

Define

$$w^t(\pi) = \sup_{f \in \overline{\mathcal{F}}^t} \left[f(\pi) - f^\star(\pi) \right].$$

The regret can be bounded by

$$\textbf{\textit{Reg}} \leq \sum_{t=1}^{T} \bar{f}^t\left(\pi^t\right) - f^\star\left(\pi^t\right) \leq \sum_{t=1}^{T} w^t\left(\pi^t\right)$$

Fix $\varepsilon > 0$,

$$\sum_{t=1}^{T} w^t\left(\pi^t\right) \leq \sum_{t=1}^{T} w^t\left(\pi^t\right) \mathbb{I}\left\{w^t\left(\pi^t\right) > \varepsilon\right\} + \varepsilon T.$$

Lemma 1

Fix a function class $\mathcal{F}$, function $f^\star \in \mathcal{F}$, and parameter $\beta > 0$. For any sequence $\pi^1, \ldots, \pi^T$, if we define

$$w^t(\pi) = \sup_{f \in \mathcal{F}} \left\{ f(\pi) - f^\star(\pi) : \sum_{i<t} \left(f\left(\pi^i\right) - f^\star\left(\pi^i\right)\right)^2 \leq \beta \right\}$$

then for all $\alpha > 0$,

$$\sum_{t=1}^{T} \mathbb{I}\left\{w^t\left(\pi^t\right) > \alpha\right\} \leq \left(\frac{\beta}{\alpha^2} + 1\right) \cdot \underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \alpha)$$

Order the indices $\{1, \ldots, T\}$ as $\{i_1, \ldots, i_T\}$, so that

$$w^{i_1}\left(\pi^{i_1}\right) \geq w^{i_2}\left(\pi^{i_2}\right) \geq \ldots \geq w^{i_T}\left(\pi^{i_T}\right).$$

Consider any index τ for which $w^{i_\tau}\left(\pi^{i_\tau}\right) > \varepsilon$. For any $\alpha > \varepsilon$, if we have $w^{i_\tau}\left(\pi^{i_\tau}\right) > \alpha$, Lemma 1 implies that

$$\tau \leq \sum_{t=1}^{T} \mathbb{I}\left\{w^t\left(\pi^t\right) > \alpha\right\} \leq \left(\frac{4\beta}{\alpha^2} + 1\right) \underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \alpha) \leq \frac{5\beta}{\alpha^2} \underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \alpha).$$

Hence,

$$w^{i_\tau}\left(\pi^{i_\tau}\right) \leq \sqrt{\frac{5\beta\, \mathrm{Edim}(\mathcal{F}, \varepsilon)}{\tau}}$$

With this,

$$\sum_{t=1}^{T} w^t\left(\pi^t\right) \mathbb{I}\left\{w^t\left(\pi^t\right) > \varepsilon\right\} \lesssim \sum_{t=1}^{T} \sqrt{\frac{\beta \operatorname{Edim}(\mathcal{F}, \varepsilon)}{t}} \lesssim \sqrt{\beta \operatorname{Edim}(\mathcal{F}, \varepsilon) T}$$

The regret is bounded by

$$\begin{aligned} \mathbf{Reg} &\leq \sum_{t=1}^{T} w^t\left(\pi^t\right) \mathbb{I}\left\{w^t\left(\pi^t\right) > \varepsilon\right\} + \varepsilon T \\ &\lesssim \sqrt{\beta \operatorname{Edim}(\mathcal{F}, \varepsilon) T} + \varepsilon T, \end{aligned}$$

where $\varepsilon > 0$ is arbitrary.

Proof of Lemma 1

The proof will be finished by following steps:

1. Define α-dependent, which is similar to the Eulder dimension.
2. For $w^t(\pi^t) > \alpha$ in a sequence $\pi^1, ..., \pi^t$, we find at most $\frac{\beta}{\alpha^2}$ of disjoint subsequences that are α-dependent of π^t.
3. For any sequence, we find that there exists one π in the sequence α-dependent with at least $\lfloor \tau/d \rfloor$ disjoint subsequence.
4. Establish relationship between Step 2 and Step 3.

Step 1:

Let $d = \underline{\mathrm{Edim}}_{f^\star}(\mathcal{F}, \alpha)$.

α-independent:

> π is α-independent of $\pi^1, \ldots, \pi^t$ if $\exists\ f \in \mathcal{F}$
> such that $|f(\pi) - f^\star(\pi)| > \alpha$ and $\sum_{i=1}^t \left(f\left(\pi^i\right) - f^\star\left(\pi^i\right)\right)^2 \leq \alpha^2$.

α-dependent:

> π is α-dependent on $\pi^1, \ldots, \pi^t$ if
> $\forall\ f \in \mathcal{F}$ with $\sum_{i=1}^t \left(f\left(\pi^i\right) - f^\star\left(\pi^i\right)\right)^2 \leq \alpha^2, |f(\pi) - f^\star(\pi)| \leq \alpha$.

Step 2:

For any t, if $w^t(\pi^t) > \alpha$, then π_t is α-dependent on at most β/α^2 disjoint subsequences of $\pi^1, \ldots, \pi^{t-1}$.

Let f be $|f(\pi^t) - f^\star(\pi^t)| > \alpha$. If π^t is α-dependent on $\pi^{i_1}, \ldots, \pi^{i_k}$ but $w^t(\pi^t) > \alpha$, we must have

$$\sum_{j=1}^{k} \left(f\left(\pi^{i_j}\right) - f^\star\left(\pi^{i_j}\right)\right)^2 \geq \alpha^2$$

If there are M such disjoint sequences, we have

$$M\alpha^2 \leq \sum_{i<t} \left(f\left(\pi^i\right) - f^\star\left(\pi^i\right)\right)^2 \leq \beta$$

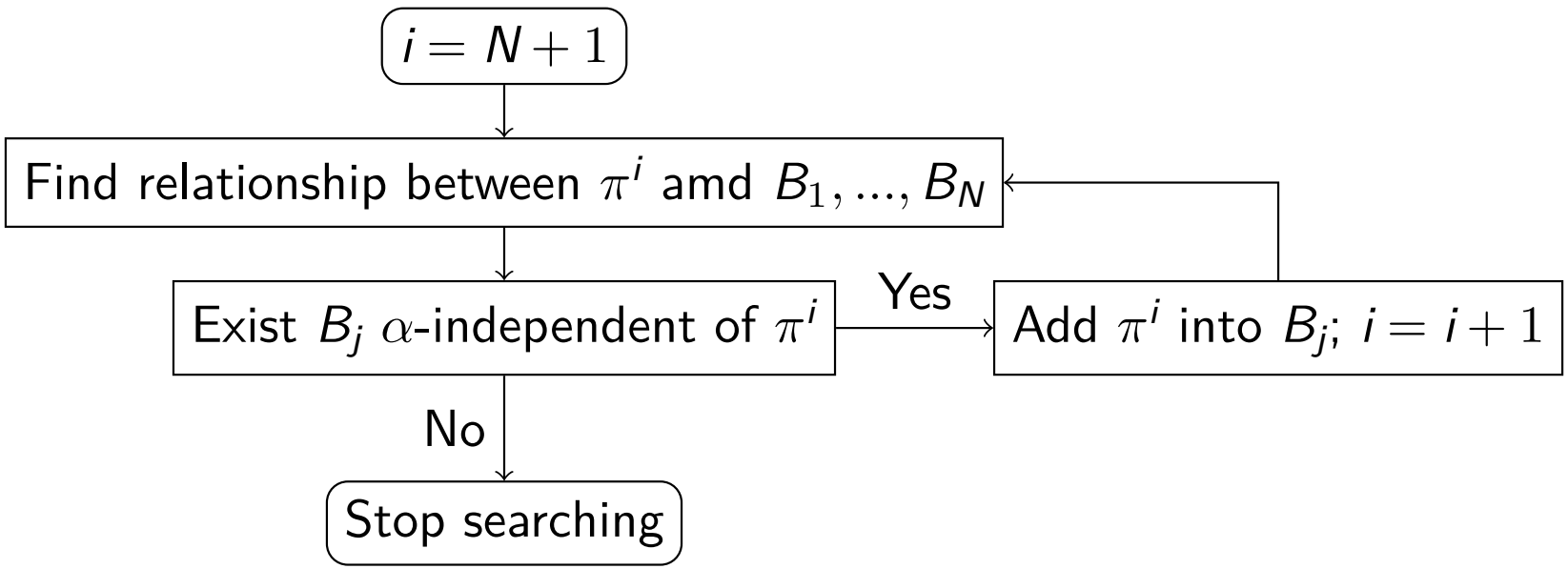
so $M \leq \frac{\beta}{\alpha^2}$.

Step 3:

For τ and any sequence $\left(\pi^1, \ldots, \pi^\tau\right)$, there is some π^j is α-dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences of $\pi^1, \ldots, \pi^{j-1}$.

Let $N = \lfloor \tau/d \rfloor$, and let $B_1, \ldots, B_N$ be subsequences of $\pi^1, \ldots, \pi^\tau$, and $B_i = \left(\pi^i\right)$.

Do the following step



If stop at $i = \tau$, we have $\sum_{i=1}^{N} |B_i| \geq \tau \geq dN$. Moreover, $|B_i| \leq d$, so in this case π^τ is α-dependent on all B_i.

Step 4:

Let $(\pi^{t_1}, \ldots, \pi^{t_\tau})$ be the subsequence $\pi^1, \ldots, \pi^T$ consisting of all elements for which $w^{t_i}(\pi^{t_i}) > \alpha$.

From **Step 2**, each π^{t_i} is dependent on at most β/α^2 disjoint subsequences of $(\pi^{t_1}, \ldots, \pi^{t_\tau})$.

From **Step 3**, one element is dependent on at least $\lfloor \tau/d \rfloor$ disjoint subsequences.

We must have $\lfloor \tau/d \rfloor \leq \beta/\alpha^2$, and which implies that $\tau \leq \left(\beta/\alpha^2 + 1\right) d$.

When will this bound be vacuous?

Consider generalised linear models

$$\mathcal{F} = \left\{\pi \mapsto \sigma(\langle\theta, \phi(\pi)\rangle) \mid \theta \in \Theta \subset \mathrm{B}_2^d(1)\right\}$$

Let $\sigma(z) = +\operatorname{relu}(z)$ or $\sigma(z) = -\operatorname{relu}(z)$, where $\operatorname{relu}(z) := \max\{z, 0\}$ is the ReLU function.

The lower bound of Eulder dimension is

$$\operatorname{Edim}(\mathcal{F}, \varepsilon) \gtrsim e^d$$

for constant ε.

Suboptimality of Optimism

The UCB algorithm is useful for some special cases, it does not attain optimal regret for any structured bandit problem. We see an example by adding "cheating" action into decision space.

Example: Let $A \in \mathbb{N}$ be a power of 2 and consider the following $\mathcal{F}$.

- The decision space is $\Pi = [A] \cup \mathcal{C}$, where $\mathcal{C} = \left\{c_1, \ldots, c_{\log_2(A)}\right\}$ is a set of "cheating" actions.
- For all actions $\pi \in [A], f(\pi) \in [0,1]$ for all $f \in \mathcal{F}$.
- For each $f \in \mathcal{F}$, rewards for actions in $\mathcal{C}$ take the following form. Let $b(f) = \left(b_1(f), \ldots, b_{\log_2(A)}(f)\right) \in \{0,1\}^{\log_2(A)}$ be a binary encoding for the index of $\pi_f \in [A]$. For each action $c_i \in \mathcal{C}$, we set

$$f(c_i) = -b_i(f).$$

Simple analyses

- Suppose that rewards are Gaussian with $r \sim \mathcal{N}\left(f^\star(\pi), 1\right)$ under π.
- If we do this for each $c_i \in \mathcal{C}$, which will incur $\widetilde{O}(\log_2(A))$ regret.
- We can stop exploring, and commit to playing $\pi_{f^\star}$ for remaining rounds.

$\Rightarrow$ with probability at least $1-\delta$,

$$\textbf{Reg} \lesssim \log_2^2(A/\delta).$$

UCB algorithm

- $c_i \in \mathcal{C}$ have $f(c_i) \leq 0$ for all $f \in \mathcal{F}$, we have $\bar{f}^t(c_i) \leq 0$.
- UCB algorithm only pull actions in $[A]$, ignoring the cheating actions.

$\Rightarrow$

$$\textbf{Reg} \gtrsim \sqrt{AT}.$$

The "cheating" actions are guaranteed to have low reward, UCB avoids them even though they are very informative.

We conclude that:

- Balance the tradeoff between optimizing reward and acquiring information to get optimal regret bound.
- The Inverse Gap Weighting algorithm attained optimal sample complexity for any choice of class $\mathcal{F}$, and all that needed to change was how to perform estimation.

Online Regression Oracle

Online Regression Oracle

At each time $t \in [T]$, an online regression oracle returns, given

$$\left(\pi^1, r^1\right), \ldots, \left(\pi^{t-1}, r^{t-1}\right)$$

with $\mathbb{E}\left[r^i \mid \pi^i\right] = f^\star\left(\pi^i\right)$ and $\pi^i \sim p^i$, a function $\widehat{f}^t : \Pi \to \mathbb{R}$ such that

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left(\widehat{f}^t\left(\pi^t\right) - f^\star\left(\pi^t\right)\right)^2 \leq \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)$$

with probability at least $1-\delta$. Here, $p^i\left(\cdot \mid \mathcal{H}^{i-1}\right)$ is the randomization distribution for the decision-maker.

E2D Algorithm

Estimation-to-Decision (E2D) for Structured Bandits

Input: Exploration parameter $\gamma > 0$.

for $t = 1, \ldots, T$ do

Obtain $\widehat{f}^t$ from online regression oracle with $\left(\pi^1, r^1\right), \ldots, \left(\pi^{t-1}, r^{t-1}\right)$.

Compute

$$p^t = \arg\min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot \left(f(\pi) - \widehat{f}^t(\pi)\right)^2\right].$$

Select action $\pi^t \sim p^t$.

Decision-to-Estimation Coefficient

Decision-Estimation Coefficient is a complexity measure for $\mathcal{F}$, and it is defined as:

$$\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f}) = \min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}[\underbrace{f(\pi_f) - f(\pi)}_{\text{regret of decision}} - \gamma \cdot \underbrace{(f(\pi) - \widehat{f}(\pi))^2}_{\text{information gain for obs.}}]$$

$$\mathrm{dec}_\gamma(\mathcal{F}) = \sup_{\widehat{f} \in \mathrm{co}(\mathcal{F})} \mathrm{dec}_\gamma(\mathcal{F}, \widehat{f})$$

Regret Bound for E2D

Rgeret Bound

The E2D algorithm with exploration parameter $\gamma > 0$ guarantees that with probability at least $1 - \delta$,

$$\textbf{Reg} \leq \mathrm{dec}_{\gamma}(\mathcal{F}) \cdot T + \gamma \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)$$

We can optimize over the parameter γ in the result above, which yields

$$\begin{aligned}\textbf{Reg} &\leq \inf_{\gamma > 0} \left\{ \mathrm{dec}_{\gamma}(\mathcal{F}) \cdot T + \gamma \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta) \right\} \\ &\leq 2 \cdot \inf_{\gamma > 0} \max \left\{ \mathrm{dec}_{\gamma}(\mathcal{F}) \cdot T, \gamma \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta) \right\}\end{aligned}$$

For finite classes, the exponential weights method give $\mathrm{Est}_{\mathrm{sq}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$, and this bound specializes to

$$\textbf{Reg} \lesssim \inf_{\gamma > 0} \max \left\{ \mathrm{dec}_{\gamma}(\mathcal{F}) \cdot T, \gamma \cdot \log(|\mathcal{F}|/\delta) \right\}$$

Proof of Regret Bound

We write

$$\begin{aligned}\mathbf{Reg} &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^\star\left(\pi^\star\right) - f^\star\left(\pi^t\right)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^\star\left(\pi^\star\right) - f^\star\left(\pi^t\right)\right] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[\left(f^\star\left(\pi^t\right) - \widehat{f}^t\left(\pi^t\right)\right)^2\right] \\ &\quad + \gamma \cdot \mathbf{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)\end{aligned}$$

For each t, since $f^\star \in \mathcal{F}$, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^t \sim p^t}\left[f^\star(\pi^\star) - f^\star(\pi^t)\right] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[\left(f^\star(\pi^t) - \widehat{f}^t(\pi^t)\right)^2\right] \\
&\leq \sup_{f \in \mathcal{F}}\left\{\mathbb{E}_{\pi^t \sim p^t}\left[f(\pi_f) - f(\pi^t)\right] - \gamma \cdot \mathbb{E}_{\pi^t \sim p^t}\left[\left(f(\pi^t) - \widehat{f}^t(\pi^t)\right)^2\right]\right\} \\
&= \inf_{p \in \Delta(\Pi)} \sup_{f \in \mathcal{F}} \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot \left(f(\pi^t) - \widehat{f}^t(\pi^t)\right)^2\right] \\
&= \mathrm{dec}_\gamma\left(\mathcal{F}, \widehat{f}^t\right)
\end{aligned}$$

We conclude that

$$\textbf{Reg} \leq \sup_{\widehat{f}} \mathrm{dec}_\gamma(\mathcal{F}, \widehat{f}) \cdot T + \gamma \cdot \textbf{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)$$

Example: Multi-Armed Bandit

Decision-to-Estimation Bound (IGW minimizes the DEC)

For the Multi-Armed Bandit setting, where $\Pi = [A]$ and $\mathcal{F} = \mathbb{R}^A$

- the Inverse Gap Weighting distribution $p = \mathrm{IGW}_{4\gamma}(\widehat{f})$ is the exact minimizer for $\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f})$.
- $\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f}) = \frac{A+1}{4\gamma}$.

Inverse Gap Weighting

Give a vector $\widehat{f} = (\widehat{f}(1), ..., \widehat{f}(A)) \in \mathbb{R}^A$, the Inverse Gap Weighting distribution $p = IGW_\gamma(\widehat{f}(1), ..., \widehat{f}(A))$ with parameter $\gamma \geq 0$ is defined as

$$p(\pi) = \frac{1}{\lambda + 2\gamma(\widehat{f}(\hat{\pi}) - \widehat{f}(\pi))},$$

where $\hat{\pi} = \arg\max_\pi \widehat{f}(\pi)$, and $\lambda \in [1, A]$ satisfy $\sum_\pi p(\pi) = 1$

Proposition 9

Consider a finite decision space $\Pi = \{1, ..., A\}$. For any vector $\widehat{f} \in \mathbb{R}^A$ and $\gamma > 0$, define $p = IGW_\gamma(\widehat{f}(1), ..., \widehat{f}(A))$. This strategy guarantees that for all $f^\star \in \mathbb{R}^A$,

$$\mathbb{E}[f^\star(\pi^\star) - f^\star(\pi)] \leq \frac{A}{\gamma} + \gamma \cdot \mathbb{E}_{\pi \sim p}[(\widehat{f}(\pi) - f^\star(\pi))^2].$$

- By rewriting Proposition 9, we can deduce that the DEC is bounded by $\frac{A+1}{4\gamma}$.
- Proposition 14 shows that IGW is actually the best possible distribution for this minimax problem.

Proof of DEC Bound

Rewrite the definition of DEC:

$$\begin{aligned}
&\min_{p\in\Delta([A])} \max_{f\in\mathbb{R}^A} \mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\right] \\
=\ &\min_{p\in\Delta([A])} \max_{f\in\mathbb{R}^A} \max_{\pi^\star\in[A]} \mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\right] \\
=\ &\min_{p\in\Delta([A])} \max_{\pi^\star\in[A]} \max_{f\in\mathbb{R}^A} \mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\right].
\end{aligned}$$

Proof step:

- IGW achieve lower bound of

  $$\textbf{Inner Equaiton} = \max_{\pi^\star\in[A]} \max_{f\in\mathbb{R}^A} \mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\right]$$

- IGW achieve the condition of

  $$\min_{p\in\Delta([A])} \textbf{Inner Equaiton}$$

.

Step 1

Firstly, we find the lower bound.

For any fixed p and $\pi^\star$, first-order conditions for optimality imply that

$$f(\pi) = \widehat{f}(\pi) - \frac{1}{2\gamma} + \frac{1}{2\gamma p(\pi^\star)} \mathbb{I}\{\pi = \pi^\star\}$$

It gives that

$$\mathbb{E}_{\pi\sim p}\left[f(\pi^\star) - f(\pi)\right] = \mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1 - p(\pi^\star)}{2\gamma p(\pi^\star)}$$

and

$$\gamma\mathbb{E}_{\pi\sim p}\left[\left(f(\pi) - \widehat{f}(\pi)\right)^2\right] = \frac{1 - p(\pi^\star)}{4\gamma} + \frac{\left(1 - p(\pi^\star)\right)^2}{4\gamma p(\pi^\star)} = \frac{1}{4\gamma p(\pi^\star)} - \frac{1}{4\gamma}$$

For any $p \in \Delta(\Pi)$,

$$\begin{aligned}
&\max_{\pi^\star \in [A]} \max_{f \in \mathbb{R}^A} \mathbb{E}_{\pi \sim p}\left[f(\pi^\star) - f(\pi) - \gamma(f(\pi) - \widehat{f}(\pi))^2\right] \\
&= \max_{\pi^\star \in [A]} \left\{\mathbb{E}_{\pi \sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}\right\} - \frac{1}{4\gamma} \\
&\geq \mathbb{E}_{\pi^\star \sim p}\left[\mathbb{E}_{\pi \sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}\right] - \frac{1}{4\gamma} \\
&= \frac{A}{4\gamma} - \frac{1}{4\gamma}
\end{aligned}$$

Next, we prove that IGW can achieve this lower bound.

Let $p = \mathrm{IGW}_{4\gamma}(\widehat{f})$, for all $\pi^\star$,

$$\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}$$

$$= \mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{\lambda}{4\gamma} + \widehat{f}(\widehat{\pi}) - \widehat{f}(\pi^\star) = \mathbb{E}_{\pi\sim p}[\widehat{f}(\widehat{\pi}) - \widehat{f}(\pi)] + \frac{\lambda}{4\gamma}$$

Hence,

$$\max_{\pi^\star\in[A]}\left\{\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}\right\}$$

$$= \min_{\pi^\star\in[A]}\left\{\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}\right\}$$

$$= \mathbb{E}_{\pi^\star\sim p}\left\{\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi^\star) - \widehat{f}(\pi)\right] + \frac{1}{4\gamma p(\pi^\star)}\right\} = \frac{A}{4\gamma}$$

$p = \mathrm{IGW}_{4\gamma}(\widehat{f})$ achieves the optimal value.

Step 2

Relaxing to $p \in \mathbb{R}_+^A$. Define

$$g_{\pi^\star}(p) = \widehat{f}(\pi^\star) + \frac{1}{4\gamma p(\pi^\star)}.$$

Let $\nu \in \mathbb{R}$ be a Lagrange multiplier, and Lagrangian is

$$\mathcal{L}(p, \nu) = g_{\pi^\star}(p) - \sum_\pi p(\pi)\widehat{f}(\pi) + \nu \left(\sum_\pi p(\pi) - 1 \right)$$

Based on **Step 1**, we aim to show that $p \in \Delta(\Pi)$ is optimal for

$$\min_{p \in \Delta(\Pi)} \max_{\pi^\star \in [A]} \left\{ \mathbb{E}_{\pi \sim p} \left[\widehat{f}(\pi^\star) - \widehat{f}(\pi) \right] + \frac{1}{4\gamma p(\pi^\star)} \right\} - \frac{1}{4\gamma}$$

It is sufficient to find ν such that

$$0 \in \partial_p \mathcal{L}(p, \nu)$$

where ∂_p denotes the subgradient with respect to p.

For a convex function $h(x) = \max_y g(x, y)$, we have

$$\partial_x h(x) = \mathrm{co}\left(\left\{\nabla g(x, y) \mid g(x, y) = \max_{y'} g(x, y')\right\}\right).$$

Hence,

$$\partial_p \mathcal{L}(p, \nu) = \nu \mathbf{1} - \widehat{f} + \mathrm{co}\left(\left\{\nabla_p g_{\pi^\star}(p) \mid g_{\pi^\star}(p) = \max_{\pi'} g_{\pi'}(p)\right\}\right)$$

Now, let $p = \mathrm{IGW}_{4\gamma}(\widehat{f})$. From **Step 1**, we know that $g_\pi(p) = g_{\pi'}(p)$ for all π, π', so

$$\partial_p \mathcal{L}(p, \nu) = \nu \mathbf{1} - \widehat{f} + \mathrm{co}\left(\{\nabla_p g_{\pi^\star}(p)\}_{\pi^\star \in \Pi}\right).$$

Since $\nabla_p g_{\pi^\star}(p) = -\frac{1}{4\gamma p^2(\pi^\star)} e_{\pi^\star}$,

$$\boldsymbol{\delta} := \sum_\pi p(\pi) \nabla_p g_\pi(p) = \left\{-\frac{1}{4\gamma p(\pi)}\right\}_{\pi \in \Pi} = \left\{-\frac{\lambda}{4\gamma} - \widehat{f}(\widehat{\pi}) + \widehat{f}(\pi)\right\}_{\pi \in \Pi}.$$

Therefore, $\boldsymbol{\delta} \in \mathrm{co}\left(\{\nabla_p g_{\pi^\star}(p)\}_{\pi^\star \in \Pi}\right)$, by choosing $\nu = \frac{\lambda}{4\gamma} + \widehat{f}(\widehat{\pi})$, we have

$$\nu \mathbf{1} - \widehat{f} + \boldsymbol{\delta} = 0,$$

which finishes the proof.

Conclusion

- Introduce Sructured Bandit, and generalise the decision space Π into large and potentially continuous space.
- Applying UCB algorithm with Eulder dimension to solve Structured Bandit problem, where

$$\textbf{Reg} \lesssim \min_{\varepsilon>0}\{\sqrt{\mathrm{Edim}(\mathcal{F},\varepsilon)\cdot T\log(|\mathcal{F}|/\delta)}+\varepsilon T\}.$$

- Using Estimation-to-Decision (E2D) for Structured Bandit, which provide the optimal rate. The regret bound is

$$\textbf{Reg} \leq \mathrm{dec}_{\gamma}(\mathcal{F})\cdot T+\gamma\cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F},T,\delta)$$