Structured Bandits (part 2)

Chapter 4.3-4.4

Reinforcement Learning and Bandit Algorithms Joint Reading Group, Spring 2024

Review of Part 1

Revisit Structured Bandits

Multi-Armed Bandit:

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

Motivation: Decision space Π is large and potentially continuous. (not finite set). $\rightarrow$ Replace A with some intrinsic measure of complexity.

Failure of UCB

Regret Bound with Eluder dimension

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

$$\mathbf{Reg} \lesssim \sqrt{\mathrm{Edim}\left(\mathcal{F}, T^{-1/2}\right) \cdot T \log(|\mathcal{F}|/\delta)}$$

The UCB algorithm is useful for some special cases, it does not attain optimal regret for any structured bandit problem.

- **relu class models**: $\mathrm{Edim}(\mathcal{F}, \varepsilon) \gtrsim e^d \rightarrow$ Eulder dimension is still large (overly pessimistic)
- **Cheating Code**: we can find simple algorithms that give

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

  while with UCB we have $\mathbf{Reg} \gtrsim \sqrt{AT}$.

E2D and $\mathrm{dec}_\gamma(\mathcal{F})$

Estimation-to-Decision (E2D) Algorithm

Input: Exploration parameter $\gamma > 0$.

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

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

- Select action $\pi^t \sim p^t$, where

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

Decision-Estimation Coefficient is a complexity measure for $\mathcal{F}$:

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

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

Regret Bound for E2D

Proposition 13. The E2D algorithm with exploration parameter $\gamma > 0$ guarantees with probability at least $1 - \delta$,

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

where $\mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)$ is the estimation error from online oracle and scales as $\log(|\mathcal{F}|/\delta)$ for finite $\mathcal{F}$.

Therefore, for regret bound we just need to bound DEC.

Actually, any specific choice of $p \in \Delta(\Pi)$ gives an upper bound of DEC.

Proposition 14. 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 $\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f})$.
- $\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f}) = \frac{A+1}{4\gamma}$.

4.3 Decision-Estimation Coefficient: Examples

Example 1: Background of Cheating Code

Cheating Code: Settings

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

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

- Determine each $b_i(f^*)$, which will incur $\widetilde{O}(\log_2(A))$ regret.
- Then stop exploring, and commit to playing $\pi_{f^\star}$ for remaining rounds.

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

- UCB algorithm only pull actions in $[A]$, ignoring the cheating actions.

$$\Rightarrow \text{Classic bound:} \quad \textbf{Reg} \gtrsim \sqrt{AT}.$$

New Regret bound with DEC for Cheating Code

Proposition 15 (DEC for Cheating Code)

Consider the cheating code. For this class $\mathcal{F}$, we have

$$\mathrm{dec}_{\gamma}(\mathcal{F}) \lesssim \frac{\log_2(A)}{\gamma}$$

Remark:

- this result implies $\mathrm{Reg} \lesssim \sqrt{\log_2(A) T \log |\mathcal{F}|}$.
- the strategy p that certifies the bound on the DEC is not necessarily the exact DEC minimizer (the distributions $p^1, \ldots, p^T$ played by E2D may be different.).
- Using a slightly more refined version of the E2D algorithm (Foster, Golowich and Han, 2023), one can improve the bound to match the $\log(A)$ given earlier.

Proof of Proposition 15

For simplicity, we work on $\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f})$ for $\widehat{f} \in \mathcal{F}$, not for $\widehat{f} \in \mathrm{co}(\mathcal{F})$. Define

$$p = (1-\varepsilon)\pi_{\widehat{f}} + \varepsilon \cdot \mathrm{unif}(\mathcal{C}).$$

We want to show with $\varepsilon = 2\frac{\log_2(A)}{\gamma}$, it yields

$$\mathrm{dec}_\gamma(\mathcal{F}, \widehat{f}) \lesssim \frac{\log_2(A)}{\gamma}$$

For minimax problem of

$$\mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\right],$$

Let's consider two cases:

First , if $\pi_f = \pi_{\widehat{f}}$, then

$$\mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\right] \leq \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi)\right]$$
$$= \mathbb{E}_{\pi \sim p}\left[f(\pi_{\widehat{f}}) - f(\pi)\right] \leq 2\varepsilon.$$

Second, suppose that $\pi_f \neq \pi_{\widehat{f}}$. Note

$$\mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}(\pi))^2\right] \leq 2 - \gamma\cdot\mathbb{E}_{\pi\sim p}\left[(f(\pi) - \widehat{f}(\pi))^2\right]$$

Observe that since $\pi_f \neq \pi_{\widehat{f}}$, if we let $b_1, \ldots, b_{\log_2(A)}$ and $b'_1, \ldots, b'_{\log_2(A)}$ denote the binary representations for π_f and $\pi_{\widehat{f}}$, there must exist i such that $b_i \neq b'_i$. Hence

$$\mathbb{E}_{\pi\sim p}\left[(f(\pi) - \widehat{f}(\pi))^2\right] \geq \frac{\varepsilon}{\log_2(A)}\left(f(c_i) - \widehat{f}(c_i)\right)^2 = \frac{\varepsilon}{\log_2(A)}$$

We conclude that in the second case,

$$\mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma\cdot(f(\pi) - \widehat{f}(\pi))^2\right] \leq 2 - \gamma\frac{\varepsilon}{\log_2(A)}$$

Putting the cases together, we have

$$\mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot (f(\pi) - \widehat{f}(\pi))^2\right] \leq \max\left\{2\varepsilon, 2 - \gamma \frac{\varepsilon}{\log_2(A)}\right\}$$

To balance these terms, set

$$\varepsilon = 2\frac{\log_2(A)}{\gamma}$$

which leads to the result. □

Example 2: Background of Linear Bandit

Linear Bandit: Settings

- Decision space: arbitrary Π. Define $\mathcal{F} = \{\pi \mapsto \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$, where $\Theta \subseteq \mathrm{B}_2^d(1)$ and $\phi : \Pi \to \mathrm{B}_2^d(1)$ is a fixed feature map (known).
- Special case of the linear contextual bandit problem

G-optimal Design

Definition: G-optimal Design

For any compact set $\mathcal{Z} \subseteq \mathbb{R}^d$ with $\dim \operatorname{span}(\mathcal{Z}) = d$, there exists a distribution $p \in \Delta(\mathcal{Z})$, called the *G*-optimal design, which has

$$\sup_{z \in \mathcal{Z}} \left\langle \Sigma_p^{-1} z, z \right\rangle \leq d \tag{4.23}$$

where $\Sigma_p := \mathbb{E}_{z \sim p}\left[zz^\top\right]$.

The G-optimal design ensures coverage in every direction of the decision space. Special cases include:

- When $\mathcal{Z} = \Delta([A])$, $p = \operatorname{unif}(e_1, \ldots, e_A)$ is an optimal design
- When $\mathcal{Z} = \mathrm{B}_2^d(1)$, $p = \operatorname{unif}(e_1, \ldots, e_A)$ is an optimal design.
- For any positive definite matrix $A \succ 0$, letting $\lambda_1, \ldots, \lambda_d$ and $v_1, \ldots, v_d$ denote the eigenvalues and eigenvectors for A, respectively, $p = \operatorname{unif}\left(\lambda_1^{-1/2} v_1, \ldots, \lambda_d^{-1/2} v_d\right)$ is an optimal design.

Regret (DEC) bound

- Generalised ε-greedy algorithm gives **Reg** $\lesssim d^{1/3} T^{2/3} \log |\mathcal{F}|$.
- We can obtain a d/γ bound on the DEC, which leads to **Reg**$\lesssim \sqrt{dT}$.

Algorithm: D2E+IGW with G-Optimal Design

- Define $\bar{\phi}(\pi) = \phi(\pi) / \sqrt{1 + \frac{\gamma}{d}\left(\hat{f}(\pi_{\hat{f}}) - \hat{f}(\pi)\right)}$, where $\pi_{\hat{f}} = \arg\max_{\pi \in \Pi} \hat{f}(\pi)$.
- Let $\bar{q} \in \Delta(\Pi)$ be the G-optimal design, and define $q = \frac{1}{2}\bar{q} + \frac{1}{2}\mathbb{I}_{\pi_{\hat{f}}}$.
- For each $\pi \in \Pi$, set

$$p(\pi) = \frac{q(\pi)}{\lambda + \frac{\gamma}{d}\left(\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi)\right)}$$

Proposition 17: This strategy certifies that

$$\mathrm{dec}_\gamma(\mathcal{F}) \lesssim \frac{d}{\gamma}$$

Proof of Proposition 17

Fix f, denote $\eta = \gamma/d$. Minimax problem in DEC,

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

Handle the regret term: decomposition (same as Proposition 9)

$$\begin{aligned}&\mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi)\right] \\ =&\underbrace{\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi_{\widehat{f}}) - \widehat{f}(\pi)\right]}_{\text{(I) exploration bias}} + \underbrace{\mathbb{E}_{\pi\sim p}\left[\widehat{f}(\pi) - f(\pi)\right]}_{\text{(II) est error on policy}} + \underbrace{f(\pi_f) - \widehat{f}(\pi_{\widehat{f}})}_{\text{(III) est error at opt}}\end{aligned}$$

(I) and (II)

For (I)

$$\mathbb{E}_{\pi\sim p}\left[\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi)\right]=\sum_{\pi}\frac{q(\pi)\left(\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi)\right)}{\lambda+\eta\left(\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi)\right)}\leq\sum_{\pi}\frac{q(\pi)}{\eta}\leq\frac{1}{\eta}$$

For (II)

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

(III): Est error at opt

Decomposition:

$$(\mathrm{III}) = f(\pi_f) - \widehat{f}(\pi_f) - \left(\widehat{f}\left(\pi_{\widehat{f}}\right) - \widehat{f}(\pi_f)\right) = \left\langle \theta - \widehat{\theta}, \phi\left(\pi_f\right) \right\rangle - \left(\widehat{f}\left(\pi_{\widehat{f}}\right) - \widehat{f}(\pi_f)\right),$$

where $f(\pi) = \langle \theta, \phi(\pi) \rangle$ and $\widehat{f}(\pi) = \langle \widehat{\theta}, \phi(\pi) \rangle$.

Define $\Sigma_p = \mathbb{E}_{\pi \sim p}\left[\phi(\pi)\phi(\pi)^\top\right]$, we have

$$\begin{aligned}
\left\langle \theta - \widehat{\theta}, \phi\left(\pi_f\right) \right\rangle &= \left\langle \Sigma_p^{1/2}(\theta - \widehat{\theta}), \Sigma_p^{-1/2}\phi\left(\pi_f\right) \right\rangle \\
&\le \left\| \Sigma_p^{1/2}(\theta - \widehat{\theta}) \right\|_2 \left\| \Sigma_p^{-1/2}\phi\left(\pi_f\right) \right\|_2 \\
&\le \frac{\gamma}{2} \left\| \Sigma_p^{1/2}(\theta - \widehat{\theta}) \right\|_2^2 + \frac{1}{2\gamma} \left\| \Sigma_p^{-1/2}\phi\left(\pi_f\right) \right\|_2^2 \\
&= \frac{\gamma}{2} \mathbb{E}_{\pi \sim p}\left[\left(\widehat{f}(\pi) - f(\pi)\right)^2 \right] + \frac{1}{2\gamma} \left\langle \phi\left(\pi_f\right), \Sigma_p^{-1}\phi\left(\pi_f\right) \right\rangle
\end{aligned}$$

(III): Est error at opt

Observe that $\Sigma_p \succeq \frac{1}{2}\bar{\Sigma}_{\bar{q}}$, hence

$$\begin{aligned}\left\langle \phi\left(\pi_f\right), \Sigma_p^{-1}\phi\left(\pi_f\right)\right\rangle &\le 2\left\langle \phi\left(\pi_f\right), \bar{\Sigma}_{\bar{q}}^{-1}\phi\left(\pi_f\right)\right\rangle \\ &= 2\left(1+\eta\left(\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi_f)\right)\right)\left\langle \bar{\phi}\left(\pi_f\right), \bar{\Sigma}_{\bar{q}}^{-1}\bar{\phi}\left(\pi_f\right)\right\rangle \\ &\le 2d\left(1+\eta\left(\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi_f)\right)\right),\end{aligned}$$

where we defined $\bar{\phi}(\pi) = \phi(\pi)/\sqrt{1+\frac{\gamma}{d}\left(\hat{f}\left(\pi_{\hat{f}}\right)-\hat{f}(\pi)\right)}$ and $\bar{q}$ is the G-optimal design for $\{\bar{\phi}(\pi)\}_{\pi\in\Pi}$.

$$\Sigma_p \succeq \frac{1}{2}\sum_{\pi}\frac{\bar{q}(\pi)}{\lambda+\eta\left(\widehat{f}\left(\pi_{\widehat{f}}\right)-\widehat{f}(\pi)\right)}\phi(\pi)\phi(\pi)^\top \succeq \frac{1}{2}\sum_{\pi}\bar{q}(\pi)\bar{\phi}(\pi)\bar{\phi}(\pi)^\top =: \frac{1}{2}\bar{\Sigma}_{\bar{q}}$$

(III): Est error at opt

Therefore:

$$(\mathrm{III}) \leq \frac{\gamma}{2}\mathbb{E}_{\pi\sim p}\left[\left(\widehat{f}(\pi) - f(\pi)\right)^2\right] + \underbrace{\frac{1}{2\gamma}\left\langle \phi\left(\pi_f\right), \Sigma_p^{-1}\phi\left(\pi_f\right)\right\rangle - \left(\widehat{f}\left(\pi_{\widehat{f}}\right) - \widehat{f}\left(\pi_f\right)\right)}_{(\mathrm{IV})},$$

where

$$(\mathrm{IV}) \leq \frac{2d}{2\gamma} + \frac{2d\eta}{2\gamma}\left(\widehat{f}\left(\pi_{\widehat{f}}\right) - \widehat{f}\left(\pi_f\right)\right) - \left(\widehat{f}\left(\pi_{\widehat{f}}\right) - \widehat{f}\left(\pi_f\right)\right) \leq \frac{d}{\gamma},$$

which completes the proof. □

Remarks on Regret Bound

- One can show $\text{dec}_\gamma(\mathcal{F}) \gtrsim \frac{d}{\gamma}$
- Combining this result with Proposition 13 and using the averaged exponential weights algorithm gives $\text{Reg} \lesssim \sqrt{dT \log(|\mathcal{F}|/\delta)}$.
- So far, we have shown

$$\text{dec}_\gamma(\mathcal{F}) \lesssim \frac{\text{eff-dim}(\mathcal{F}, \Pi)}{\gamma}$$

  where $\text{eff-dim}(\mathcal{F}, \Pi)$ is some quantity that (informally) reflects the amount of exploration required.
- In general, DEC can have slower decay rate than $\gamma^{-1} \Rightarrow$ optimal rate worse than $\sqrt{T}$.

Example 3: Nonparametric Bandits

Consider the Lipschitz bandits in metric spaces:
Let Π to be a metric space equipped with metric ρ, and define

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is } 1\text{-Lipschitz w.r.t } \rho\}$$

Objective: give bound on the DEC which depends on the $\mathcal{N}_\rho(\Pi, \varepsilon)$.

Define $\Pi' \subseteq \Pi$ as an ε-cover with respect to ρ if

$$\forall \pi \in \Pi \quad \exists \pi' \in \Pi' \quad \text{s.t.} \quad \rho(\pi, \pi') \leq \varepsilon$$

Suppose $\mathcal{N}_\rho(\Pi, \varepsilon) \leq \varepsilon^{-d}$ for all $\varepsilon > 0$. Let $\widehat{f} : \Pi \to [0,1]$ and $\gamma \geq 1$, consider:

- Let $\Pi' \subseteq \Pi$ witness the covering number $\mathcal{N}_\rho(\Pi, \varepsilon)$.
- Let p be IGW distribution, restricted to the (finite) decision space Π'

DEC bound for Lipschitz Bandits

Proposition 18: DEC bound for Lipschitz Bandits

By setting $\varepsilon \propto \gamma^{-\frac{1}{d+1}}$, this strategy certifies that

$$\operatorname{dec}_\gamma(\mathcal{F}, \widehat{f}) \lesssim \gamma^{-\frac{1}{d+1}}$$

This leads to **Reg** $\lesssim T^{\frac{d+1}{d+2}}$, which cannot be improved.

Proof: Since f is 1-Lipschitz and Π' is the ε-cover for Π, there exists $\iota(\pi) \in \Pi'$ such that $\rho(\pi, \iota(\pi)) \leq \varepsilon$. Consequently,

$$\begin{aligned} \mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi)\right] &\leq \mathbb{E}_{\pi\sim p}\left[f(\iota(\pi_f)) - f(\pi)\right] + \left|f(\pi_f) - f(\iota(\pi_f))\right| \\ &\leq \mathbb{E}_{\pi\sim p}\left[f(\iota(\pi_f)) - f(\pi)\right] + \varepsilon \end{aligned}$$

Proof of Proposition 18

since $\iota(\pi_f) \in \Pi'$, Proposition 9 ensures for p from inverse gap weighting over Π', we have

$$\mathbb{E}_{\pi \sim p}\left[f(\iota(\pi_f)) - f(\pi)\right] \leq \frac{|\Pi'|}{\gamma} + \gamma \cdot \mathbb{E}_{\pi \sim p}\left[\left(f(\pi) - \widehat{f}(\pi)\right)^2\right]$$

As we assume $\mathcal{N}_\rho(\Pi, \varepsilon), |\Pi'| \leq \varepsilon^{-d}$,

$$\mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma \cdot \left(f(\pi) - \widehat{f}(\pi)\right)^2\right] \leq \varepsilon + \frac{\varepsilon^{-d}}{\gamma}$$

Choosing $\varepsilon \propto \gamma^{-\frac{1}{d+1}}$ leads to the result. □

Example 4: DEC subsumes Edim

Consider any class $\mathcal{F}$ with values in $[0, 1]$. For all $\gamma \geq e$, we have

$$\mathrm{dec}_{\gamma}(\mathcal{F}) \lesssim \inf_{\varepsilon > 0} \left\{ \varepsilon + \frac{\mathrm{Edim}(\mathcal{F} - \mathcal{F}, \varepsilon) \log^2(\gamma)}{\gamma} \right\} + \gamma^{-1}$$

As a special case, this implies that E2D enjoys a regret bound for generalized linear bandits similar to that of UCB.

Example 5: Bandits with Concave Rewards

Take $\Pi \subseteq \mathrm{B}_2^d(1)$ and define

$$\mathcal{F} = \{f : \Pi \to [0,1] \mid f \text{ is concave and } 1\text{-Lipschitz w.r.t } \ell_2\}$$

For this setting, Lattimore (2020) shows

$$\mathrm{dec}_\gamma(\mathcal{F}) \lesssim \frac{d^4}{\gamma} \cdot \mathrm{polylog}(d, \gamma)$$

For the relu function class

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

above bound leads to $\sqrt{\mathrm{poly}(d)\,T}$ regret bound.

⇒Eluder dimension is overly pessimistic, as it grows exponentially for this class.

4.4 Relationship to Optimism and Posterior Sampling

Combine E2D with Confidence Sets

Algorithm: E2D with Confidence Set

Input: $\gamma > 0$, confidence radius $\beta > 0$.
For $t = 1, \ldots, T$ do
Obtain $\widehat{f}^t$ from online regression oracle with $(\pi^1, r^1), \ldots, (\pi^{t-1}, r^{t-1})$.
Set

$$\mathcal{F}^t = \left\{ f \in \mathcal{F} \mid \sum_{i<t} \mathbb{E}_{\pi^i \sim p^i} \left[\left(\widehat{f}^i(\pi^i) - f^\star(\pi^i) \right)^2 \right] \leq \beta \right\}$$

Select action $\pi^t \sim p^t$, with

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

Same as E2D, except that at each step, we compute a confidence set $\mathcal{F}^t$.
If $\beta = \mathrm{Est}_{\mathrm{sq}}(\mathcal{F}, T, \delta)$, then it ensures that with probability at least $1 - \delta$,

$$\textit{\textbf{Reg}} \leq \sum_{t=1}^{T} \mathrm{dec}_\gamma \left(\mathcal{F}^t \right) + \gamma \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)$$

Relation to usual UCB

Proposition 20

The UCB strategy $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\operatorname{dec}_0\left(\mathcal{F}^t\right) \leq \bar{f}^t\left(\pi^t\right) - \underline{f}\left(\pi^t\right) \tag{4.27}$$

the confidence width might be large for a given round t, but by the pigeonhole argument

$$\sum_{t=1}^{T} \operatorname{dec}_0\left(\mathcal{F}^t\right) \leq \sum_{t=1}^{T} \bar{f}^t\left(\pi^t\right) - \underline{f}^t\left(\pi^t\right) \leq \widetilde{O}(\sqrt{AT})$$

Meaningful only if $\mathcal{F}^1, \ldots, \mathcal{F}^T$ are shrinking (fast).

Proposition 21

For any $\gamma > 0$, the UCB strategy $\pi^t = \arg\max_{\pi \in \Pi} \bar{f}^t(\pi)$ certifies that

$$\operatorname{dec}_\gamma\left(\mathcal{F}^t, \widehat{f}^t\right) \leq \bar{f}^t\left(\pi^t\right) - \widehat{f}^t\left(\pi^t\right) + \frac{1}{4\gamma}$$

Proof of Proposition 21

By choosing $\pi^t = \arg\max_{\pi\in\Pi} \bar{f}^t(\pi)$, we have

$$\begin{aligned}
\operatorname{dec}_\gamma\left(\mathcal{F}, \widehat{f}^t\right) &= \min_{p\in\Delta(\Pi)} \max_{f\in\mathcal{F}_t} \mathbb{E}_{\pi\sim p}\left[\max_{\pi^\star} f(\pi^\star) - f(\pi) - \gamma\cdot\left(\widehat{f}^t(\pi) - f(\pi)\right)^2\right] \\
&\le \max_{f\in\mathcal{F}_t}\left[\max_{\pi^\star} f(\pi^\star) - f\left(\pi^t\right) - \gamma\cdot\left(\widehat{f}^t\left(\pi^t\right) - f\left(\pi^t\right)\right)^2\right] \\
&\le \max_{f\in\mathcal{F}_t}\left[\bar{f}^t\left(\pi^t\right) - f\left(\pi^t\right) - \gamma\cdot\left(\widehat{f}^t\left(\pi^t\right) - f\left(\pi^t\right)\right)^2\right] \\
&= \underbrace{\max_{f\in\mathcal{F}_t}\left[\widehat{f}^t\left(\pi^t\right) - f\left(\pi^t\right) - \gamma\cdot\left(\widehat{f}^t\left(\pi^t\right) - f\left(\pi^t\right)\right)^2\right]}_{\le \frac{1}{4\gamma}} \\
&\quad + \bar{f}^t\left(\pi^t\right) - \widehat{f}^t\left(\pi^t\right).
\end{aligned}$$

Connection to Posterior Sampling

Define a natural dual (max-min) analogue of the DEC

$$\underline{\mathrm{dec}}_{\gamma}(\mathcal{F},\widehat{f}) = \sup_{\mu\in\Delta(\mathcal{F})} \inf_{p\in\Delta(\Pi)} \mathbb{E}_{f\sim\mu}\mathbb{E}_{\pi\sim p}\left[f(\pi_f) - f(\pi) - \gamma\cdot(f(\pi)-\widehat{f}(\pi))^2\right]$$

The adversary selects a prior distribution μ over models in $\mathcal{M}$, and the learner (with knowledge of the prior) finds a decision distribution p that balances the average tradeoff between regret and information acquisition when the underlying model is drawn from μ.

Equivalence of Primal and Dual

Under mild regularity conditions, we have

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

Remarks:

- Any bound on the dual DEC immediately yields a bound on the primal DEC. We bring existing tools for Bayesian bandits and reinforcement learning to bear on the primal DEC.
- the dual DEC is always bounded by a Bayesian complexity measure known as the *information ratio*, which is used throughout the literature on Bayesian bandits and reinforcement learning.

Incorporating with Contextual Bandits

Algorithm: E2D for Contextual Structured Bandits

Input: Exploration parameter $\gamma > 0$.
for $t = 1, \ldots, T$ do
- Observe $x^t \in \mathcal{X}$.
- Obtain $\widehat{f}^t$ from online regression oracle with $(x^1, \pi^1, r^1), \ldots, (x^{t-1}, \pi^{t-1}, r^{t-1})$.
- Compute

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

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

Regret Bound of Contextual E2D

The E2D algorithm with exploration parameter $\gamma > 0$ guarantees that

$$\mathrm{Reg} \le \sup_{x \in \mathcal{X}} \mathrm{dec}_\gamma(\mathcal{F}(x, \cdot)) \cdot T + \gamma \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta),$$

where $\mathcal{F}(x, \cdot) = \{f(x, \cdot) \mid f \in \mathcal{F}\}$. (Proof is identical to Proposition 13.)

- For finite decisions, if $\mathcal{F} = \mathbb{R}^A$, SquaredCB is precisely the special case of Contextual E2D (IGW distribution is the exact DEC minimiser).
- Going beyond the finite-action setting: e.g.,

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

  Applying Proposition 17 gives $\sup_{x \in \mathcal{X}} \mathrm{dec}_\gamma(\mathcal{F}(x, \cdot)) \lesssim \frac{d}{\gamma}$, so that Proposition 23 gives $\mathrm{Reg} \lesssim \sqrt{dT \cdot \mathrm{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)}$.

Conclusion

- In this Chapter, we introduced Structured Bandit, which generalises the decision space Π into large and potentially continuous space, where UCB could fail.
- Using Estimation-to-Decision (E2D) framework (combined with other schemes, e.g., IGW), which provides a better (optimal) regret rate:

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

- Seen some examples on how to bound $\mathrm{dec}_{\gamma}(\mathcal{F})$