Course Notes on Inteactive Decision-making

Short Course given by Sasha Rakhlin
Princeton ML Theory Summer School 2024
Notes taken by Alexander Terenin

1 Refresher: Statistical Learning

1.1. Offline Regression. We have a dataset (x, y) with:

1. Classification: $x_i \in X$, $y_i \in [0, 1]$
2. IID data: $(x_t, y_t) \sim \mathcal{P}_{xy}$
3. Square loss: $\ell(f(x), y) = (f(x) - y)^2$
4. Function class: $\mathcal{F} = \{f : X \to [0, 1]\}$.

Denote the *conditional mean* by

Conditional mean

$$f^*(\cdot) = \mathbb{E}(y \mid x = (\cdot)) \tag{1}$$

Assume $f^* \in \mathcal{F}$. Then the *estimation error* of $\hat{f} : X \to [0, 1]$ is

Estimation error

$$\mathcal{E}(\hat{f}) = \mathbb{E}_x(\hat{f}(x) - f^*(x))^2. \tag{2}$$

Define the *empirical risk minimizer*

Empirical risk minimizer

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{T} \sum_{i=1}^{T} (f(x_i) - y_i)^2. \tag{3}$$

Lemma 1. *For finite $\mathcal{F}$, under the preceding assumptions, we have*

$$\mathbb{E}\, \mathcal{E}(\hat{f}) \lesssim \frac{\log |\mathcal{F}|}{T} \tag{4}$$

Proof. Homework: prove this via Bernstein's inequality. ■

The rate $\log |\mathcal{F}|$ is called the *fast rate*, in contrast with square-root rates.

Fast rate

1.2. Online Regression. For $t = 1, .., T$:

1. The learner chooses a potentially-random $\hat{f}_t : \mathcal{X} \to [0, 1]$.
2. Nature chooses and reveals $(x_t, y_t) \in X \times [0, 1]$.

We again assume that

$$\mathbb{E}(y_t \mid x_t = x) = f^*(x) \tag{5}$$

namely that the x-values are arbitrary, but y-values are drawn from a distribution $f^*(x)$. We also assume that $\hat{f}_t(\cdot) = \hat{f}_t(\cdot \mid \mathcal{H}_{t-1})$, namely the learner's strategy can only depend on the history, and in principle allow $\hat{f}_t \sim q_t \in \Delta(\mathcal{F})$. Define the *regret* Regret

$$\mathrm{Reg}_{\mathrm{sq}} = \sum_{t=1}^{T} (\hat{f}_t(x_t) - y_t)^2 - \min_{f \in \mathcal{F}} \sum_{t=1}^{T} (f(x_t) - y_t)^2 \tag{6}$$

and *estimation error* Estimation error

$$\mathrm{Est}_{\mathrm{sq}} = \sum_{t=1}^{T} (\hat{f}_t(x_t) - f^*(x_t))^2 \tag{7}$$

both of which are random variables. Note that all infima over $\mathcal{F}$ are achieved because it is assumed finite.

Lemma 2. *We have* Regret lower bound via estimation error

$$\mathbb{E}\,\mathrm{Est}_{\mathrm{sq}} \leq \mathbb{E}\,\mathrm{Reg}_{\mathrm{sq}} \tag{8}$$

Proof. Start from the previous expressions and add-subtract appropriate terms. ■

Lemma 3. *There exists a way of choosing $\hat{f}_t$ such that* Regret upper bound

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{sq}} \lesssim \frac{\log |\mathcal{F}|}{T}. \tag{9}$$

Specifically, the algorithm is averaged exponential weights: compute

$$q_t(f) \propto \exp\left(-\eta \sum_{s=1}^{t-1} (f(x_s) - y_s)^2 \right) \tag{10}$$

and choose $\hat{f}_t = \mathbb{E}_{f \sim q_t}(f)$ for an appropriate parameter η.

Proof. Exercise. ■

This is the same result as in statistical learning theory, but implies via regret a bound on the estimation error. Note that $\hat{f}_t$ above need not be in the

function class $\mathcal{F}$, particularly if it is finite. For more on this, one can read Cesa-Bianchi–Lugosi (2006), specifically the part about mixable losses.

The point is that we can do statistical estimation in settings even if the x-values are arbitrary. Later on, the x-values are going to come from some complicated distribution based on our own previous choices which is impossible to model, so our plan is going to be to treat them as arbitrary.

Aside on randomization. Let us remark why the learner does not need to randomize, in spite of the fact that they play first. Let $\ell : A \times B \to \mathbb{R}$ be convex, and note that

$$\inf_{p\in\Delta(A)} \sup_{q\in\Delta(B)} \mathop{\mathbb{E}}_{\substack{a\sim p\\ b\sim q}} \ell(a,b) = \inf_{p\in\Delta(A)} \sup_{b\in B} \mathop{\mathbb{E}}_{a\sim p} \ell(a,b) \leq \inf_{a\in A} \sup_{b\in B} \ell(a,b) \tag{11}$$

by taking the infimum over a smaller set. On the other hand, since optimizing over some set of numbers is equivalent to optimizing over distributions whose expectations map onto the same set of numbers, and since ℓ is convex, applying Jensen's Inequality gives

$$\inf_{a\in A} \sup_{b\in B} \ell(a,b) = \inf_{p\in\Delta(A)} \sup_{b\in B} \ell\left(\mathop{\mathbb{E}}_{a\sim p} a, b \right) \leq \inf_{p\in\Delta(A)} \sup_{b\in B} \mathop{\mathbb{E}}_{a\sim p} \ell(a,b). \tag{12}$$

This proves that if ℓ is convex, and a Nash equilibrium exists, then a deterministic Nash also exists.

2 Multi-armed Bandits

Here, we have a set of *decisions*, *actions*, or *arms* $\Pi = \{1, .., A\}$. The problem can be expressed as a repeated interaction. For $t = 1, .., T$:

1. The decision-maker selects $\pi_t \sim p_t \in \Delta(\Pi)$
2. The decision-maker observes a reward r_t.

We assume that $r_t \in [0, 1]$ and that the reward follow the optimal *model* Model

$$r_t \sim M^*(\cdot \mid \pi_t). \tag{13}$$

This choice of terminology will become clear later. Define the *optimal rewards* Optimal rewards

$$f^*(\pi) = \mathbb{E}^{M^*}(r \mid \pi) \tag{14}$$

where this notation means that the reward is drawn according to the distribution defined by M^* given the choice of an action π. The decision-maker's *regret* is Regret

$$\mathrm{Reg}_{\mathrm{DM}} = \sum_{t=1}^{T} \max_{\pi^*} f^*(\pi^*) - \mathop{\mathbb{E}}_{\pi_t\sim p_t} f^*(\pi_t). \tag{15}$$

In this setting, *exploration* is *necessary* in order to minimize regret: one cannot simply act according to the best actions seen in the past. Define

Exploration

$$\hat{f}_t(n) = \frac{1}{n_t(\pi)} \sum_{s=1}^{t-1} r_s \mathbb{1}_{\pi_s = \pi} \tag{16}$$

where $n_t(\pi)$ is the number of times an arm has been chosen up to time t. Define the *greedy algorithm* to be the one that chooses

Greedy algorithm

$$\pi_t = \arg\max_{\pi \in \Pi} \hat{f}_t(\pi) \tag{17}$$

and this algorithm fails on the following example: take $A = 2$, and let $M^*(\cdot \mid 1) = \delta_{1/2}$, and $M^*(\cdot \mid 2) = \mathrm{Ber}(3/4)$. Then under appropriate initialization, with probability $1/4$ we get zero from arm 2, and never choose it again, incurring linear regret. As consequence, we incur linear regret in expectation. One can instead consider *the ε-greedy algorithm*, which picks greedy $\hat{\pi}_t$ with probability $1 - \varepsilon$, and with probability ε picks $\pi_t \sim \mathrm{U}(\Pi)$, which we note makes sense because Π is finite.

The ε-greedy algorithm

Lemma 4. *Let $\varepsilon \sim \left(\frac{A \log \frac{AT}{\delta}}{T}\right)^{1/3}$. Then with probability $1 - \delta$, the ε-greedy algorithm has*

Regret bound: ε-greedy algorithm

$$\mathrm{Reg}_{\mathrm{DM}} \leq A^{1/3} T^{2/3} \left(\log \frac{AT}{\delta}\right)^{1/3}. \tag{18}$$

Proof sketch. Write

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{DM}} &= \sum_{t=1}^{T} f^*(\pi^*) - \mathbb{E}_{\pi_t \sim p_t} f^*(\pi_t) && (19)\\
&\overset{\text{(i)}}{\leq} (1-\varepsilon) \sum_{t=1}^{T} f^*(\pi^*) - f^*(\hat{\pi}_t) + \varepsilon T && (20)\\
&\overset{\text{(ii)}}{\leq} \sum_{t=1}^{T} \max_{\pi \in \{\pi_t, \pi^*\}} |f^*(\pi) - \hat{f}_t(\pi)| + \varepsilon T && (21)\\
&\leq \sum_{t=1}^{T} \sqrt{\frac{\mathbb{1}_{\pi_t = \pi}}{n_t(\pi)}} + \varepsilon T && (22)
\end{aligned}$$

where (i) conditions on whether exploration occurs or not, and if it does, just bounds this by the number of time steps, (ii) follows by a one-line exercise and by bounding $(1 - \varepsilon)$ by 1. We know that $n_t(\pi) \sim \frac{t\varepsilon}{A}$ by a back-of-the-envelope calculation, which gives

$$\mathrm{Reg}_{\mathrm{DM}} \leq \sum_{t=1}^{T} \sqrt{\frac{A}{\varepsilon t}} + \varepsilon T \tag{23}$$

which gives the morally correct answer which drops various log factors. ■

This is not the correct rate for this setting, but it is a start.

2.1. Optimism. A better approach is to apply the *upper confidence bound (UCB)* algorithm, an instance of a general technique called optimism. The idea is to construct

Upper confidence bound (UCB)

$$\overline{f}_t : \Pi \to \mathbb{R} \qquad \underline{f}_t : \Pi \to \mathbb{R} \tag{24}$$

such that with high probability, we have

$$\underline{f}_t \le f^* \le \overline{f}_t \tag{25}$$

pointwise. This goes back to Lai and Robbins (1985). Let

$$\pi_t = \arg\max_{\pi \in \Pi} \overline{f}_t(\pi). \tag{26}$$

Let's analyze this. Write the *regret* as

Regret

$$\begin{aligned}
\mathrm{Reg}_{\mathrm{DM}} &= \sum_{t=1}^{T} f^*(\pi^*) - f^*(\pi_t) && (27)\\
&\overset{\text{(i)}}{\le} \sum_{t=1}^{T} \overline{f}_t(\pi^*) - \underline{f}_t(\pi_t) && (28)\\
&\overset{\text{(ii)}}{\le} \sum_{t=1}^{T} \overline{f}_t(\pi_t) - \underline{f}_t(\pi_t) && (29)
\end{aligned}$$

where (i) holds by upper bounds, and (ii) holds by optimality of π_t. For an appropriate choice of *upper and lower confidence bounds*, letting $\hat{f}_t(\pi)$ be the mean, we have

Upper and lower confidence bounds

$$\overline{f}_t(\pi) = \hat{f}_t(\pi) + C\frac{\sqrt{\log(AT/\delta)}}{n_t(\pi)} \qquad \underline{f}_t(\pi) = \hat{f}_t(\pi) - C\frac{\sqrt{\log(AT/g)}}{n_t(\pi)}. \tag{30}$$

As consequence, we get

$$\mathrm{Reg}_{\mathrm{DM}} \overset{\text{(iii)}}{\le} \sum_{t=1}^{T} \frac{1}{\sqrt{n_t(\pi)}} \mathbb{1}_{\pi_t = \pi} \sqrt{\log(..)} \overset{\text{(iv)}}{\le} \sqrt{AT \log \frac{AT}{\delta}} \tag{31}$$

where (iii) follows by an appropriate calculation left as exercise, and (iv) follows by what is known as a potential lemma.

2.2. Contextual Bandits. Let's make the game more complicated by adding context. In this setting, we can also for instance study function approximation. For $t = 1, .., T$:

1. Observe $x_t \in X$.
2. Choose $\pi_t \sim p_t \in \Delta(\Pi)$.
3. Observe $r_t \sim M^*(\cdot \mid \pi_t, x_t)$.

Let

$$f^*(x, \pi) = \mathbb{E}^{M^*}(r \mid x, \pi) \tag{32}$$

We allow infinite X, potentially uncountable. We have some

$$\mathcal{F} = \{f : X \times \Pi \to [0, 1]\} \tag{33}$$

for which we assume the following.

Assumption 5. *We have $f^* \in \mathcal{F}$, which we call* REALIZABILITY. Realizability

We can consider settings where the x_t-values are generated randomly IID from some distribution $\mathcal{P}$, called the *stochastic setting*, or settings where they are arbitrary, called the *adversarial setting*. The best policy is Stochastic setting Adversarial setting

$$\pi^*(x_t) = \arg\max_{\pi \in \Pi} f^*(x_t, \pi). \tag{34}$$

If $\mathcal{X}$ is finite, we can apply ordinary multi-armed bandits separately for each x, which leads to $\sqrt{AT|X|}$ rates. Exercise: prove this.

We now ask: will optimism work? The general optimistic template is, at every t, construct $\mathcal{F}_t \in \mathcal{F}$, which is some version of a space of plausible models. Suppose that, with high probability, $f^* \in \mathcal{F}_t$ for all t. Define

$$\overline{f}_t(x, \pi) = \max_{f \in \mathcal{F}_t} f(x, \pi) \qquad \underline{f}_t(x, \pi) = \min_{f \in \mathcal{F}_t} f(x, \pi). \tag{35}$$

Then

$$\underline{f}_t \le f^* \le \overline{f}_t \tag{36}$$

which implies the *regret bound* Regret bound

$$\mathrm{Reg}_{\mathrm{DM}} \le \sum_{t=1}^{T} \overline{f}_t(x_t, \pi_t) - \underline{f}_t(x_t, \pi_t). \tag{37}$$

A natural definition of the *optimistic function class* is Optimistic function class

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \sum_{s=1}^{t-1} (r_s - f(x_t, \pi_s))^2 \le \sum_{s=1}^{t-1} (r_s - \hat{f}_t(x_s, \pi_s))^2 + \beta \right\} \tag{38}$$

where

$$\hat{f}_t \in \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t-1} (r_s - f(x_s, \pi_s))^2. \tag{39}$$

which picks the function class to be all functions which are close to the best-fit function in a regression sense up to some error β which we will choose later. This gives the appropriate analog of the preceding confidence interval. Note that $\mathcal{F}_t$ need not be nested.

Lemma 6. *If $\beta \sim \log \frac{|\mathcal{F}|}{\delta}$, then with probability $1-\delta$, we have:*

1. *$f^* \in \mathcal{F}_t$ for all t.*
2. *$\sum_{s=1}^{t-1} (f^*(x_s, \pi_s) - f(x_s, \pi_s))^2 \leq 4\beta$ for any $f \in \mathcal{F}_t$ and any $t = 1, .., T$.*

Proof. Exercise: Bernstein's inequality. ■

An aside: a more general form of this holds for arbitrary algorithms, as long as one randomizes π_s appropriately in lieu of UCB. Continuing, what we want to do next is show that

$$\overline{f}_t(x_t, \pi_t) - f^*(x_t, \pi_t) \stackrel{?}{\leq} \sum_{s=1}^{t-1} (f^*(x_s, \pi_s) - f(x^*, \pi_s))^2 \tag{40}$$

but this *regret bound fails in general*. In some sense, the only way this could work is if we can limit the number of times we are surprised by a new (x_t, π_t). If there are enough times that the new x_t is such that performance on it has nothing to do with performance in the past, we are in trouble.

Regret bound fails in general

A positive result is that this works for the *linear setting*, where

Linear setting

$$\mathcal{F} = \left\{ f(x, \pi) = \langle \theta, \phi(x, \pi) \rangle : \theta \in B_2^d(1), \|\theta\| < C \right\} \tag{41}$$

and B_2^d denotes a Euclidean ball. Note that ϕ here is fixed and not learned as in neural networks. In this case, $\mathcal{F}_t$ consists of functions f_θ such that

$$\|\theta - \hat{\theta}_t\|_\Sigma^2 \lesssim \beta \qquad\qquad \Sigma_t = \sum_{s=1}^{t-1} \phi\phi^T \tag{42}$$

which is very similar to what we had before. In this case, one can only be surprised on the order of d many times. Using an elliptic potential, we get

$$\mathrm{Reg}_{\mathrm{DM}} \lesssim \sum_{t=1}^{T} \|\phi(x_t, s_t)\|_{\Sigma_t^{-1}} \leq \sqrt{\beta d T \log(..)}. \tag{43}$$

The function class $\mathcal{F}$ can be approximated by $\log(\frac{1}{\varepsilon})^d$ elements. As consequence, we get the *regret bound for the linear setting*, namely

Regret bound for the linear setting

$$\log |\mathcal{F}| \propto d \log T \qquad\qquad \mathrm{Reg}_{\mathrm{DM}} \leq d\sqrt{T} \log(..). \tag{44}$$

2.3. Failure of any kind of optimism. Unfortunately, there also exist classes where optimism will fail, in a strong sense which suggests no such approach will work. Define

$$\mathcal{F} = \{f^*, f_1, .., f_N\} \qquad \Pi = \{\pi_g, \pi_b\} \qquad X = \{x_1, .., x_n\} \tag{45}$$

and

$$f^*(x_i, \pi_g) = 1 - \varepsilon \qquad f_j(x, \pi_g) = 1 - \varepsilon, \quad i \neq j \qquad f_i(x_i, \pi_g) = 0 \tag{46}$$
$$f^*(x_i, \pi_b) = 0 \qquad f_j(x, \pi_b) = 0, \quad i \neq j \qquad f_i(x_i, \pi_b) = 1. \tag{47}$$

The problem is that on new contexts, we don't share information. Let S_t be the set of x-values encountered so far, and take $x_t = x_i \notin S_t$ so that $\{f_k : x_k \notin S_t\} \subseteq \mathcal{F}_t$. For this choice of x-values, we get the *regret lower bound* Regret lower bound

$$\mathrm{Reg}_{\mathrm{DM}} \geq N(1 - \varepsilon). \tag{48}$$

where $N = |\mathcal{F}| - 1$ and $|X| = N$.

We ask: is this fundamental? Let's analyze *the ε-greedy strategy.* Recall that The ε-greedy strategy

$$p_t = \begin{cases} \hat{\pi}_t = \arg\max_{\pi \in \Pi} \hat{f}_t(x_t, \pi) & \text{w.p. } 1 - \varepsilon \\ \pi_t \sim \mathrm{U}(A) & \text{w.p. } \varepsilon. \end{cases} \tag{49}$$

Let's analyze this. We have the *regret upper bound* Regret upper bound

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{DM}} \leq (1 - \varepsilon) \sum_{t=1}^{T} f^*(x_t, \pi^*(x_t)) - f^*(x_t, \hat{\pi}_t) + \varepsilon T \tag{50}$$
$$\overset{\text{(i)}}{\leq} \sum_{t=1}^{T} \sum_{\pi \in \{\pi^*, \hat{\pi}_t\}} |f^*(x_t, \pi) - \hat{f}_t(x_t, \pi)| + \varepsilon T \tag{51}$$
$$= \sum_{t=1}^{T} \sum_{\pi \in \{\pi^*, \hat{\pi}_t\}} \frac{1}{\sqrt{p_t(x, \pi)}} \sqrt{p_t(x, \pi)} |f^*(x_t, \pi) - \hat{f}_t(x_t, \pi)| + \varepsilon T \tag{52}$$
$$\overset{\text{(ii)}}{\leq} \sum_{t=1}^{T} \left(\sum_{\pi \in \{\pi^*, \hat{\pi}_t\}} \frac{1}{p_t(x, \pi)} \right)^{1/2} \tag{53}$$
$$\times \left(\sum_{\pi \in \{\pi^*, \hat{\pi}_t\}} p_t(x, \pi)(f^*(x_t, \pi) - \hat{f}_t(x_t, \pi))^2 \right)^{1/2} + \varepsilon T \tag{54}$$
$$\leq \sum_{t=1}^{T} \sqrt{\frac{2A}{\varepsilon}} \left(\mathop{\mathbb{E}}_{\pi \sim p_t} (f^*(x_t, \pi) - \hat{f}_t(x_t, \pi))^2 \right)^{1/2} + \varepsilon T \tag{55}$$
$$\leq \sqrt{\frac{2A}{\varepsilon}} \sqrt{T \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p_t} (f^*(x_t, \pi) - \hat{f}_t(x_t, \pi))^2} + \varepsilon T \tag{56}$$

where (i) is by exercise, (ii) is by Cauchy–Schwarz. We therefore see the estimation error showing up: specifically, the type of online estimation error which appeared previously. We can show the inner sum is small via Bernstein's inequality to show

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} (f^*(x_t, \pi) - \hat{f}_t(x_t, \pi))^2 \le 2 \sum_{t=1}^{T} (f^*(x_t, \pi_t) - \hat{f}_t(x_t, \pi_t))^2 \tag{57}$$

$$+ \mathcal{O}\left(\log \frac{\mathcal{F}}{\delta} \right) \tag{58}$$

with probability $1 - \delta$, where we have replaced the expectation with respect to π by estimation error on the actual trajectory x_t, π_t that we have seen. One can show the *finite class estimation error bound*

Finite class estimation error bound

$$\mathrm{Est}_{\mathrm{sq}} \le \log |\mathcal{F}|. \tag{59}$$

Balancing terms, we get the *regret upper bound with rate*

Regret upper bound with rate

$$\mathrm{Reg}_{\mathrm{DM}} \lesssim \sqrt{\frac{A}{\varepsilon}} \sqrt{T \log \mathcal{F}} + \varepsilon T \propto T^{2/3} A^{1/3} (\log |\mathcal{F}|)^{1/3}. \tag{60}$$

We therefore see the *failure of optimism is not fundamental*: other algorithms can do better. In particular, one can show that the reason the ε-greedy algorithm fails to get a $\sqrt{T}$ rate because it explores uniformly: one can do better by tilting the exploratory distribution away from bad actions.

Failure of optimism is not fundamental

Let's see what a good distribution is: we will later see that this distribution comes from fundamental calculations, but it's better to start from the answer to get an idea where we are headed.

Definition 7. *Given a vector $\hat{f} = (\hat{f}(1), .., \hat{f}(n)) \in \mathbb{R}^A$, and $\gamma > 0$, let $\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{f}(\pi)$, and define the* INVERSE GAP WEIGHTING *distribution $p = \mathrm{IGW}_\gamma(\hat{f})$ by its probability mass function*

Inverse gap weighting

$$p(\pi) = \frac{1}{\lambda + 2\gamma(\hat{f}(\hat{\pi}) - \hat{f}(\pi))} \tag{61}$$

where $\lambda \in [1, A]$ is a uniquely-defined constant which ensures $\sum_{\pi \in \Pi} p(\pi) = 1$.

To ensure this is well-defined, note that

$$\frac{1}{\lambda} \le \sum_{\pi \in \Pi} p(\pi) \le \frac{A}{\lambda} \tag{62}$$

by various elementary inequalities on the sum and definition of $\hat{\pi}$.

Lemma 8. *For any $\hat{f} \in \mathbb{R}^A$, $\gamma > 0$, and $f^* \in \mathbb{R}^A$, if $p = \mathrm{IGW}_\gamma(\hat{f})$, then*

$$\mathop{\mathbb{E}}_{\pi \sim p} \max_{\pi^* \in \Pi} (f^*(\pi^*) - f^*(\pi)) \leq \frac{A}{\gamma} + \gamma \, \mathbb{E}_{\pi \sim p} (\hat{f}(\pi) - f^*(\pi))^2. \tag{63}$$

Proof. Exercise. ■

Call the algorithm which plays this distribution *SquareCB*. For $t = 1, .., T$, we: SquareCB

1. Observe x_t.
2. Compute $p_t = \mathrm{IGW}_\gamma(\hat{f}_t(x_t, 1), .., \hat{f}_t(x_t, A))$ where $\hat{f}_t$ is an online regression estimator with respect to $(x_1, \pi_1, r_1), .., x_{t-1}, \pi_{t-1}, r_{t-1}$.
3. Choose $\pi_t \sim p_t$ and observe r_t.

Theorem 9. *SquareCB satisfies* Regret bound for SquareCB

$$\mathbb{E}\,\mathrm{Reg}_{\mathrm{DM}} \leq \sqrt{AT\,\mathrm{Est}_{\mathrm{sq}}} \overset{\text{finite}}{\lesssim} \sqrt{AT \log |\mathcal{F}|}. \tag{64}$$

Proof. Write

$$\sum_{t=1}^{T} f^*(x_t, \pi^*) - \mathop{\mathbb{E}}_{\pi_t \sim p_t} f^*(x_t, \pi_t) \tag{65}$$

$$= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t \sim p_t} (f^*(x_t, \pi^*) - f^*(x_t, \pi_t) - \gamma(\hat{f}_t(x_t, \pi_t) - f^*(x_t, \pi_t))^2) \tag{66}$$

$$+ \gamma \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi_t \sim p_t} \gamma(\hat{f}_t(x_t, \pi_t) - f^*(x_t, \pi_t))^2 \tag{67}$$

$$\leq T\frac{A}{\gamma} + \gamma\,\mathrm{Est}_{\mathrm{sq}} \tag{68}$$

$$\leq \sqrt{AT\,\mathrm{Est}_{\mathrm{sq}}} \tag{69}$$

where $\mathrm{Est}_{\mathrm{sq}}$ is of order $\log |\mathcal{F}|$ in the finite setting. ■

SquareCB works for any $x_1, .., x_T$ as long as $\mathrm{Est}_{\mathrm{sq}}(\mathcal{F}, T, \delta) = o(T)$. There is nothing preventing us from generalizing this significantly. The magic is that the actual difficulty gets pushed into the online estimation oracle. Let us make some comments:

1. This result shows that if online regression is possible, then decision-making is also possible.
2. If $x_1, .., x_T$ are IID from $\mathcal{P}_x$, then SquareCB still works, even with ordinary offline regression, in the sense that $\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{s=1}^{t-1} (f(x_s, \pi_s) - r_s)^2$. This is a recent result due to Simchi-Levi and Xu (2021).

3. Inverse gap weighting was first introduced in a paper by Abe and Long (1999), in one of the first papers on linear contextual bandits. However, soon after, people instead started focusing on optimism, and their approach was mostly not followed-up on.

2.4. Structured Bandits. In contextual bandits, the class $\mathcal{F}$ was introduced mainly to capture structure in context sets X. We did not focus on structure in the action space Π. We now forget about X and let $\mathcal{F}$ capture structure in Π. For $t = 1, .., T$:

1. Choose $\pi_t \sim p_t \in \Delta(\Pi)$.
2. Observe $r_t \sim M^*(\cdot \mid \pi_t)$.

Now, we work with the setting where Π is large but structured: if we have to depend on $|\Pi|$, we are only willing to pay $\log |\Pi|$ at most.

Assumption 10. *Let $\mathcal{M}$ be the class of models, and assume $M^* \in \mathcal{M}$, which we call* REALIZABILITY. Realizability

Define the mean rewards under the model, greedy policy under the model, as well as the set of possible mean rewards to be

$$f^M(\pi) = \mathbb{E}^M(r \mid \pi) \quad \pi_M = \arg\max_{\pi \in \Pi} f^M(\pi) \quad \mathcal{F}_{\mathcal{M}} = \{f^M : M \in \mathcal{M}\}. \tag{70}$$

We write $\mathcal{F} = \mathcal{F}_{\mathcal{M}}$. The decision-making *regret* is Regret

$$\mathrm{Reg}_{\mathrm{DM}} = \sum_{t=1}^{T} f^{M^*}(\pi_{M^*}) - \mathop{\mathbb{E}}_{\pi_t \sim p_t} f^{M^*}(\pi_t). \tag{71}$$

We ask: is it possible to get

$$\mathrm{Reg}_{\mathrm{DM}} \overset{?}{\lesssim} \sqrt{T \log |\mathcal{F}|} \tag{72}$$

or not? The answer is *no*, in general. Consider the $\sqrt{T \log |\mathcal{F}|}$-*counterexample* $\sqrt{T \log |\mathcal{F}|}$-counterexample

$$\mathcal{F} = \{f_i : i = 1, .., A\} \qquad f_i(\pi) = \frac{1}{2} + \frac{1}{2}\mathbb{1}_{\pi = i} \tag{73}$$

and take the space of models to be Gaussians with means given by $\mathcal{F}$. Thus there is one optimal action for each f_i. Thus

$$\mathrm{Reg}_{\mathrm{DM}} \gtrsim A \tag{74}$$

but $|\mathcal{F}| = A$. Thus in general we cannot have $\log |\mathcal{F}|$. This tells us that *sharing information across contexts is easier than sharing information across actions.* We can therefore ask whether or not

$$\mathrm{Reg}_{\mathrm{DM}} \overset{?}{\lesssim} \sqrt{T \log |\mathcal{F}|} + \mathcal{C}(\mathcal{F}, \Pi) \tag{75}$$

where $\mathcal{C}(\mathcal{F}, \Pi)$ is some term which quantifies the complexity of $\mathcal{F}, \Pi$.

Define the *estimation-to-decisions (E2D) algorithm* as follows. For $t = 1, .., T$: Estimation-to-decisions (E2D) algorithm

1. Compute an online estimate $\hat{f}_t : \Pi \to \mathbb{R}$ on $(\pi_1, r_1), .., (\pi_{t-1}, r_{t-1})$.
2. Letting $\pi_f = \arg\max_{\pi \in \Pi} f(\pi)$, compute

$$p_t = \arg\min_{p \in \Delta(\Pi)} \max_{f \in \mathcal{F}_M} \mathbb{E}_{\pi \sim p} f(\pi_f) - f(\pi) - \gamma(f(\pi) - \hat{f}_t(\pi))^2. \tag{76}$$

Lemma 11. *For any f, any γ, and any f^*, the distribution $p = \mathrm{IGW}_\gamma(\hat{f})$ satisfies* Inequality for inverse gap weighting

$$\mathbb{E}_{\pi \sim p} f^*(\pi^*) - f^*(\pi) \leq \frac{A}{\gamma} + \gamma \mathbb{E}_{\pi \sim p}(f^*(\pi) - \hat{f}(\pi))^2. \tag{77}$$

Proof. Similar to preceding argument. ■

In particular, this obviously means that

$$\max_{\hat{f} \in \mathcal{F}} \min_{p \in \Delta(\Pi)} \max_{f^* \in \mathcal{F}} \mathbb{E}_{\pi \sim p} f^*(\pi^*) - f^*(\pi) - \gamma \mathbb{E}_{\pi \sim p}(f^*(\pi) - \hat{f}(\pi))^2 \leq \frac{A}{\gamma}. \tag{78}$$

It turns out that IGW_γ is an optimal solution to this problem when A is finite, which can be proved by considering the first-order optimality conditions. Define therefore the *decision-estimation coefficient* Decision-estimation coefficient

$$\mathrm{dec}^{\mathrm{sq}}_\gamma(\mathcal{F}, \hat{f}) = \min_{p \in \Delta(\Pi)} \max_{f^* \in \mathcal{F}} \mathbb{E}_{\pi \sim p} f^*(\pi^*) - f^*(\pi) - \gamma \mathbb{E}_{\pi \sim p}(f^*(\pi) - \hat{f}(\pi))^2. \tag{79}$$

Now, the question is: can we upper-bound regret by this quantity? We can, and this gives

$$\mathrm{Reg}_{\mathrm{DM}} \leq \min_{\gamma > 0} T \max_{\hat{f} \in \mathrm{co}(\mathcal{F})} \mathrm{dec}^{\mathrm{sq}}_\gamma(\mathcal{F}, \hat{f}) + \gamma \, \mathrm{Est}_{\mathrm{sq}} \tag{80}$$

where $\mathrm{co}(\mathcal{F})$ denotes the convex hull of $\mathcal{F}$. Recall that our model estimates $\hat{f}_t$ in online estimation did not necessarily live in $\mathcal{F}$: we needed to take mixtures, which is where the convex hull appears. Define

$$\mathrm{dec}^{\mathrm{sq}}_\gamma(\mathcal{F}) = \max_{\hat{f} \in \mathrm{co}(\mathcal{F})} \mathrm{dec}^{\mathrm{sq}}_\gamma(\mathcal{F}, \hat{f}). \tag{81}$$

We can think of this term as the *exploration complexity* of this problem class.

Theorem 12. *The minimax regret of a decision-maker can be lower-bounded by* Regret lower bound

$$\mathrm{Reg}_{\mathrm{DM}} \gtrsim \min_{\gamma} T \max_{\hat{f} \in \mathcal{F}} \mathrm{dec}^{\mathrm{sq}}_\gamma(\mathcal{F}, \hat{f}) + \gamma. \tag{82}$$

One can work with a slightly more refined notion called *constrained DEC*, defined in our particular setting with square loss as

Constrained DEC

$$\mathrm{dec}^{\mathrm{c,sq}}_{\varepsilon}(\mathcal{F}, \hat{f}) = \min_{p} \max_{f \in \mathcal{F}} \left[\mathop{\mathbb{E}}_{\pi \sim p} f(\pi_f) - f(\pi) : \mathbb{E}(f(\pi) - \hat{f}(\pi))^2 \le \varepsilon \right]. \tag{83}$$

Then, as long the mean is a sufficient statistic—which can be relaxed via the more refined notion given in the sequel—we have

$$\mathrm{dec}^{\mathrm{c,sq}}_{\underline{\varepsilon}}(\mathcal{F})T \lesssim \mathrm{Reg}_{\mathrm{DM}} \lesssim \mathrm{dec}^{\mathrm{c,sq}}_{\overline{\varepsilon}}(\mathcal{F})T \qquad \underline{\varepsilon} \sim \frac{1}{\sqrt{T}} \qquad \overline{\varepsilon} \sim \sqrt{\frac{\log |\mathcal{F}|}{T}}. \tag{84}$$

This removes the notion of a convex hull, and various other notions from the original work such as localization. Note that estimation error shows up in the upper but not lower bounds: understanding this is a key open problem.

2.5. Decision-making with structured observations. One can have situations like the preceding problem class but where information leaks between arms. For instance, one can construct settings where there is a Bernoulli arm which is optimal, and Gaussian arms which are suboptimal, where observing the Bernoulli leaks optimality. To handle this, one can define the more general *decision-making with structured observations* setting. For $t = 1, .., T$:

Decision-making with structured observations

1. Select $\pi_t \sim p_t \in \Delta(\pi)$.
2. Observe $(r_t, o_t) \sim M^*(\cdot \mid \pi_t)$.

Episodic reinforcement learning embeds into this problem class. In particular, this means that if we understand sample complexity of this problem class, we understand the sample complexity of reinforcement learning.

One can define the notion of a *decision-estimation coefficient* here too, with one modification: estimation error, at least for the discrete case, needs to be replaced with the squared Hellinger distance D^2_{H}. This gives

Decision-estimation coefficient

$$\mathrm{dec}^{\mathrm{H}}_{\gamma}(\mathcal{F}, \hat{f}) = \min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathop{\mathbb{E}}_{\pi \sim p} f^M(\pi^M) - f^M(\pi) - \gamma \mathop{\mathbb{E}}_{\pi \sim p} D^2_{\mathrm{H}}(M(\cdot \mid \pi), \widehat{M}(\cdot \mid \pi)). \tag{85}$$

The constrained version is defined similarly, but where we now have a Hellinger ball as our constraint. A good example to consider is the *cheating code* example, where information about other arms is hidden in the binary expansion of rewards in some arm. Thus, this notion of DEC tells us how to balance *regret*, which shows up in the terms $f^*(\pi^*) - f^*(\pi)$, with *information gain*, which shows up in the D_{H}-term.

There have been various refinements to this notion. One of the main open questions is whether and how to remove the analog of the estimation error term, as in the preceding case.