Foundations of Reinforcement Learning and Interactive Decision Making – Exercises Solutions

Ludovic Schwartz

Contents

Chapter 1

Introduction

Exercise 1.1 Proposition 1, Part 2.

Consider the setting of Proposition 1 where $(x^1, y^1), \ldots, (x^T, y^T)$ are i.i.d., $\mathcal{F} = \{f : \mathcal{X} \to [0,1]\}$ is finite, the true regression function satisfies $f^* \in \mathcal{F}$, and $Y_i \in [0,1]$ almost surely. Prove that empirical risk minimizer $\widehat{f}$ with respect to square loss satisfies the following bound on excess risk. With probability at least $1 - \delta$:

$$\mathcal{E}(\widehat{f}) \preceq \frac{\log(|\mathcal{F}|/\delta)}{T}.$$

1. For a fixed function $f \in \mathcal{F}$, consider the random variable

$$Z_i(f) = (f(x^i) - y^i)^2 - (f^*(x^i) - y^i)^2$$

for $i = 1, \ldots, T$. Show that

$$\mathbb{E}\left[Z_i(f)\right] = \mathbb{E}\left[(f(x^i) - f^*(x^i))^2\right] = \mathcal{E}(f).$$

2. Show that for any fixed $f \in \mathcal{F}$, the variance $\mathbb{V}\left[Zi(f)\right]$ is bounded as

$$\mathbb{V}\left[Z_i(f)\right] \leq 4\mathbb{E}\left[(f(x^i) - f^*(x^i))^2\right].$$

3. Apply Bernstein's inequality (Lemma 5) to show that with for any $f \in \mathcal{F}$, with probability at least $1 - \delta$,

$$\mathcal{E}(f) \leq 2(\widehat{L}(f) - \widehat{L}(f^*)) + \frac{C\log(1/\delta)}{T},$$

for an absolute constant C, where $\widehat{L}(f) = \frac{1}{T}\sum_{t=1}^{T}(f(x^t) - y^t)^2$.

4. Extend this probabilistic inequality to simultaneously hold for all $f \in \mathcal{F}$ by taking the union bound over $f \in \mathcal{F}$. Conclude as a consequence that the bound holds for $\widehat{f}$, the empirical minimizer, implying (1.38)

Solution : 1. Developping the squares and rearranging the terms, we have

$$Z_i(f) = (f(x^i) - f^*(x^i))^2 + 2(f^*(x^i) - y_i)(f^*(x^i) - f(x^i)).$$

Now remark that $\mathbb{E}\left[f^*(x_i) - y_i\right] = \mathbb{E}\left[\mathbb{E}\left[f^*(x_i) - y_i\right| x_i\right]] = 0$ where we have used the tower rule for the first equality and the fact that $f^*(x_i) = \mathbb{E}\left[y_i\right| x = x_i\left]\right.$. Thus

$$\begin{aligned}\mathbb{E}\left[Z_i(f)\right] &= \mathbb{E}\left[(f(x^i) - f^*(x^i))^2\right] + \mathbb{E}\left[2(f^*(x^i) - y_i)(f^*(x^i) - f(x^i))\right] \text{ (by linearity of } \mathcal{E}.) \\ &= \mathbb{E}\left[(f(x^i) - f^*(x^i))^2\right]\end{aligned}$$

Finally the fact that $\mathbb{E}\left[(f(x^i) - f^*(x^i))^2\right] = \mathcal{E}(f)$ comes from Lemma 1.

2. Remark that $Z_i(f) = (f(x_i) - f^*(x_i))(f(x_i) + f^*(x_i) - 2y_i)$ therefore

$$\begin{aligned}Z_i(f)^2 &= (f(x_i) - f^*(x_i))^2(f(x_i) + f^*(x_i) - 2y_i)^2 \\ &\leq 4(f(x_i) - f^*(x_i))^2\end{aligned}$$

where we have used the fact that $f(x_i) + f^*(x_i) - 2y_i \in [-2, 2]$. Finally

$$\mathbb{V}\left[Z_i(f)\right] \leq \mathbb{E}\left[Z_i(f)^2\right] \leq 4\mathbb{E}\left[(f(x_i) - f^*(x_i))^2\right] = 4\mathcal{E}(f)$$

which concludes.

3. First we recall Bernstein's inequality:

Lemma 1. *Let $Z_1, \ldots, Z_T$ be i.i.d with variance $\mathbb{V}\left[Z_i(f)\right] = \sigma^2$, and range $Z - \mathbb{E}\left[Z\right] \leq B$ almost surely. Then with probability at least $1 - \delta$,*

$$\frac{1}{T}\sum_{i=1}^{T} Z_i - \mathbb{E}\left[Z\right] \leq \sigma\sqrt{\frac{2\log(1/\delta)}{T}} + \frac{B\log(1/\delta)}{3T}.$$

Applying Lemma 1 for $Z_i = -Z_i(f)$ where $B = 2$ and $\sigma = 2\sqrt{\mathcal{E}(f)}$ yields

$$\widehat{L}(f^*) - \widehat{L}(f) + \mathcal{E}(f) \leq \sqrt{\frac{8\mathcal{E}(f)\log(1/\delta)}{T}} + \frac{2\log(1/\delta)}{3T} \tag{1.1}$$

where we have used the fact that $\widehat{L}(f) - \widehat{L}(f^*) = \frac{1}{T}\sum_{t=1}^{T} Z_i(f)$. Now in order to get the bound expected in the question we will need to use the following lemma which is a rewriting of the AMGM inequality.

Lemma 2. *For any $a, b \geq 0$ and $\eta > 0$ we have,*

$$\sqrt{ab} \leq \frac{a}{4\eta} + \eta b.$$

Proof. Starting from AMGM we have $\sqrt{xy} \leq \frac{x+y}{2}$. Since the inequality holds for any $x, y \geq 0$ we can pick $x = \frac{a}{2\eta}$ and $y = 2\eta b$. to conclude the proof. □

Let's apply Lemma 2 with $a = 8\mathcal{E}(f)$ and $b = \frac{\log(1/\delta)}{T}$, then for $\eta > 0$ we have

$$\sqrt{\frac{8\mathcal{E}(f)\log(1/\delta)}{T}} \leq \frac{2\mathcal{E}(f)}{\eta} + \eta\frac{\log(1/\delta)}{T}.$$

Taking $\eta = \frac{1}{2}$, substituing the above inequality in 1.1 and rearranging terms gives

$$\mathcal{E}(f) \leq 2(\widehat{L}(f^*) - \widehat{L}(f)) + \frac{7\log(1/\delta)}{6T}.$$

4. For sake of concreteness we detail the union bound argument. Let $\delta' > 0$, from the previous question we have

$$\begin{aligned}
\mathbb{P}\left[\exists f \in \mathcal{F} : \mathcal{E}(f) \geq 2(\widehat{L}(f^*) - \widehat{L}(f)) + \frac{7\log(1/\delta')}{6T}\right] &= \mathbb{P}\left[\bigcup_{f\in\mathcal{F}} \mathcal{E}(f) \geq 2(\widehat{L}(f^*) - \widehat{L}(f)) + \frac{7\log(1/\delta')}{6T}\right] \\
&\leq \sum_{f\in\mathcal{F}} \mathbb{P}\left[\mathcal{E}(f) \geq 2(\widehat{L}(f^*) - \widehat{L}(f)) + \frac{7\log(1/\delta')}{6T}\right] \\
&\leq \delta'|\mathcal{F}|
\end{aligned}$$

Now taking $\delta' = \frac{\delta}{|\mathcal{F}|}$ we have with probability $1-\delta$ that for any f

$$\mathcal{E}(f) \leq 2(\widehat{L}(f^*) - \widehat{L}(f)) + \frac{7\log(|\mathcal{F}|/\delta)}{6T}$$

and in particular this holds for $\widehat{f}$. □

Exercise 1.2 ERM in Online Learning

Consider the problem of Online Supervised Learning with indicator loss $\ell(f(x), y) = \mathbb{I}\{f(x) \neq y\}$, $\mathcal{Y} = \mathcal{Y}' = \{0, 1\}$, and a finite class $\mathcal{F}$.

1. Exhibit a class $\mathcal{F}$ for which ERM cannot ensure sublinear growth of regret for all sequences, i.e. there exists a sequence $(x^1, y^1), \ldots, (x^T, y^T)$ such that

$$\sum_{t=1}^{T} \ell(\hat{f}^t(x^t), y^t) - \min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x^t), y^t) = \Omega(T),$$

   where $\hat{f}^t$ is the empirical minimizer for the indicator loss on $(x^1, y^1), \ldots, (x^{t-1}, y^{t-1})$. Note : The construction must have $|\mathcal{F}| \leq C$, where C is an absolute constant that does not depend on T.

2. Show that if data are i.i.d., then in expectation over the data, ERM attains a sublinear bound $O(\sqrt{T\log|\mathcal{F}|})$ on regret for any finite class $\mathcal{F}$.

Solution : 1. We will do a constrution with $|\mathcal{F}| = 2$, we consider the two constants functions $f_0 = 0$ and $f_1 = 1$. We look at the following $(y_t)_{t\geq 1}$:

$$1, 0, 0, 1, 1, 0, 0, \dots$$

Everytime the algorithm sees more 1 than 0, it will predict 1 and make a mistake and inversely it it sees more 0 than 1 (If there is the same number, ties are broken arbitrarilies and the algorithm will be wrong half the time). That means that after T round, our algorithm makes at least $\frac{3T}{4} - 3$ mistakes. Conversely, either f_1 or f_0 will make less than $\frac{T}{2}$ mistakes. This means that ERM suffers linear regret on that sequence.

2. We operate under the more general assumption that ℓ is bounded in $[0, 1]$. We introduce some notation :

- For a fixed $f \in \mathcal{F}$, $\hat{\ell}_t(f) = \sum_{s=1}^{t} \ell(f(x_t), y_t)$, $\hat{f}^{t+1} = \arg\min_{f\in\mathcal{F}} \hat{\ell}_t(f)$.
- $\mu \in \Delta(\mathcal{X} \times \mathcal{Y})$ the distribution under which the data is drawn.
- For a fixed $f \in F$, $\bar{\ell}(f) = \mathbb{E}_{x,y\sim\mu}[\ell(f(x), y)]$, $f^* = \arg\min_{f\in\mathcal{F}} \bar{\ell}(f)$.

We start by a decomposition of the regret :

$$\begin{aligned}
\mathbb{E}[R_n] &= \mathbb{E}\left[\sum_{t=1}^{T} \ell(\hat{f}^t(x_t), y_t) - \min_{f\in\mathcal{F}} \sum_{t=1}^{T} \ell(f(x_t), y_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left[\ell(\hat{f}^t(x_t, y_t)) - \bar{\ell}(f^*)\right]\right] + T\mathbb{E}\left[\bar{\ell}(f^*) - \hat{\ell}_T(\hat{f}^{T+1})\right]
\end{aligned}$$

Our main technical tool is Hoeffding's Lemma. Since ℓ is bounded, we have that for a fixed f, for any $\delta > 0$

$$\mathbb{P}\left[|\bar{\ell}(f) - \hat{\ell}_t(f)| \geq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2t}}\right] \leq \delta.$$

In particular, taking a union bound over all functions in $\mathcal{F}$, we have that :

$$\mathbb{P}\left[\exists f \in \mathcal{F}, |\bar{\ell}(f) - \hat{\ell}_t(f)| \geq \sqrt{\frac{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{2t}}\right] \leq \delta.$$

We also state the following Lemma for integrating tail bounds :

Lemma 3. *Let A,B be positive reals and X a real valued random variable such that* $\mathbb{P}\left[X \geq \sqrt{A \log\left(\frac{B}{\delta}\right)}\right] \leq \delta$, *then*

$$\mathbb{E}[X] \leq 2\sqrt{A \log B}.$$

Proof. We have by a change of variable $\mathbb{P}\left[X - \sqrt{A \log(B)} \geq \epsilon^2\right] \leq \exp\left(\frac{-\epsilon^2}{A}\right)$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \sqrt{A \log B} + \mathbb{E}\left[X - \sqrt{A \log B}\right] \\
&\leq \sqrt{A \log B} + \int_0^\infty \mathbb{P}\left[X - \sqrt{A \log B} \geq \epsilon\right] d\epsilon \\
&= \sqrt{A \log B} + \int_0^\infty \exp -\frac{\epsilon^2}{A}\, d\epsilon \\
&= \sqrt{A \log B} + \frac{\sqrt{A\pi}}{2} \\
&\leq 2\sqrt{A \log B}.
\end{aligned}$$

□

Now, with probability at least $1 - \delta$, we have

$$|\bar{\ell}(f^*) - \hat{\ell}_T(f^{T+1})| \leq \sqrt{\frac{\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{2T}}.$$

Integrating this bound with Lemma 3, we get

$$T\mathbb{E}\left[\bar{\ell}(f^*) - \hat{\ell_T}(f^{\hat{T}+1})\right] \leq \sqrt{2T \log(2|\mathcal{F}|)}.$$

It remains to deal with the term $\sum_{t=1}^T \ell(\hat{f}^t(x_t, y_t)) - \bar{\ell}(f^*)$. Here, since (x_t, y_t) is independent from $\hat{f}^t$, we have that for any $t \geq 2$,

$$\mathbb{E}\left[\ell(\hat{f}^t(x_t), y_t) \,|(x_1, y_1, \ldots, x_{t-1}, y_{t-1})\right] = \bar{\ell}(\hat{f}^t)$$

Now we bound with probability at least $1 - \delta$

$$\begin{aligned}
\bar{\ell}(\hat{f}^t) &= \hat{\ell}_{t-1}(\hat{f}^t) + (\bar{\ell}(\hat{f}^t) - \hat{\ell}_{t-1}(\hat{f}^t)) \\
&\leq \hat{\ell}_{t-1}(f^*) + (\bar{\ell}(\hat{f}^t) - \hat{\ell}_{t-1}(\hat{f}^t)) \\
&= \bar{\ell}(f^*) + (\hat{\ell}_{t-1}(f^*) - \bar{\ell}(f^*)) + (\bar{\ell}(\hat{f}^t) - \hat{\ell}_{t-1}(\hat{f}^t)) \\
&\leq \bar{\ell}(f^*) + \sqrt{\frac{2\log\left(\frac{2|\mathcal{F}|}{\delta}\right)}{t-1}}.
\end{aligned}$$

By the tower rule of expectation, we have that

$$\begin{aligned}\mathbb{E}\left[\ell(\hat{f}^t(x_t), y_t) - \bar{\ell}(f^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\ell(\hat{f}^t(x_t), y_t)\,|(x_1, y_1, \ldots, x_{t-1}, y_{t-1})\right] - \bar{\ell}(f^*)\right] \\ &= \mathbb{E}\left[\bar{\ell}(\hat{f}^t) - \bar{\ell}(f^*)\right] \\ &\leq \sqrt{\frac{8\log(2|\mathcal{F}|)}{t-1}}.\end{aligned}$$

Where the last line is obtained integrating the tail bound via Lemma 3. Now it only remains to put everything together, we have

$$\begin{aligned}\mathbb{E}[R_n] &= \mathbb{E}\left[\sum_{t=1}^{T}\left[\ell(\hat{f}^t(x_t, y_t)) - \bar{\ell}(f^*)\right]\right] + T\mathbb{E}\left[\bar{\ell}(f^*) - \hat{\ell}_T(\hat{f}^{T+1})\right] \\ &\leq 1 + \sum_{t=2}^{T}\sqrt{\frac{8\log(2|\mathcal{F}|)}{t-1}} + \sqrt{2T\log(2|\mathcal{F}|)}. \\ &\leq \sqrt{32T\log(2|\mathcal{F}|)} + \sqrt{2T\log(2|F|)} \\ &= O(\sqrt{T\log|\mathcal{F}|}),\end{aligned}$$

where the second inequality comes from the elementary inequality $\sum_{t=2}^{T}\frac{1}{\sqrt{t-1}} \leq 2(\sqrt{T}-1)$. □

Exercise 1.3 Low Noise

1. For a nonnegative random variable X, prove that for any $\eta > 0$,

$$\log\mathbb{E}\left[\exp\left(-\eta(X - \mathbb{E}[X])\right)\right] \leq \frac{\eta^2}{2}\mathbb{E}\left[X^2\right].$$

Hint : use the fact that $\log x \leq x - 1$ and $\exp(-x) \leq 1 - x + x^2/2$for $x \geq 0$.

2. Consider the setting of Proposition 3, Part 1(Generic Loss). Prove that the randomized variant of the Exponential Weights Algorithm satisfies, for any $f^* \in \mathcal{F}$,

$$\sum_{t=1}^{T}\mathbb{E}_{\hat{f}^t\sim q^t}\left[\ell(\hat{f}^t(x^t), y^t)\right] - \ell(f^*(x^t), y^t) \leq \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_{\hat{f}^t\sim q^t}\left[\ell(\hat{f}^t(x^t), y^t)^2\right] + \frac{\log|\mathcal{F}|}{\eta}.$$

for any sequence of data and nonnegative losses. Hint : replace Hoeffding's Lemma by the result of the first question.

3. Suppose $\ell(f(x), y) \in [0, 1]$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $f \in \mathcal{F}$. Suppose that there is a "perfect expert $f^* \in \mathcal{F}$ such that $\ell(f^*(x^t)y^t) = 0$ for all $t \in [T]$. Conclude that the above algorithm, with an appropriate choice of η, enjoys a bound of $O(\log|\mathcal{F}|)$ on the cumulative loss of the algorithm

(equivalently, the fast rate $\frac{\log|F|}{T}$ for the average regret). This setting is called "zero noise."

4.Consider the binary classification problem with indicator loss, and suppose $\mathcal{F}$ contains a perfect expert, as above. The *Halving Algorithm* maintains a version space $\mathcal{F}^t = \{f \in \mathcal{F} : f(x^s) = y^s, s < t\}$ and given x^t, follows the majority vote of remaining experts in $\mathcal{F}^t$. Show that this algorithm incurs cumulative loss at most $O(\log|\mathcal{F}|)$. Hence, the Exponential Weights Algorithm can be viewed as an extension of the Halving algorithm to settings where the optimal loss is non-zero.

Solution : 1. Using the hint, we have

$$\begin{aligned}
\log \mathbb{E}\left[\exp\left(-\eta(X - \mathbb{E}\left[X\right])\right)\right] &= \eta\mathbb{E}\left[X\right] + \log \mathbb{E}\left[\exp -\eta X\right] \\
&\leq \eta\mathbb{E}\left[X\right] + \mathbb{E}\left[\exp(-\eta X) - 1\right] \\
&\leq \eta\mathbb{E}\left[X\right] + \mathbb{E}\left[1 - \eta X + \frac{\eta^2 X^2}{2} - 1\right] \\
&= \eta\mathbb{E}\left[X\right] - \eta\mathbb{E}\left[X\right] + \frac{\eta^2}{2}\mathbb{E}\left[X^2\right] \\
&= \frac{\eta^2}{2}\mathbb{E}\left[X^2\right].
\end{aligned}$$

Where the first inequality is valid because the exponential is always positive and the second one because $\eta X \geq 0$.

2. We take the same notation as in the proof of Proposition 3. We have the potential $\Phi_\eta^t = -\log \sum_{f\in\mathcal{F}} \exp\left(-\eta \sum_{i=1}^t \ell(f(x^i), y^i)\right)$. Rewriting the inequality of question 1 as $\eta\mathbb{E}\left[X\right] \leq -\log \mathbb{E}\left[\exp X\right] + \frac{\eta^2}{2}\mathbb{E}\left[X^2\right]$ and applying it to $X = \ell(\hat{f}^t(x^t), y^t) \geq 0$ where $\hat{f}^t \sim q_t$, we get

$$\eta\mathbb{E}_{f\sim q_t}\left[\ell(f(x^t), y^t)\right] \leq -\log \mathbb{E}_{f\sim q_t}\left[\exp\left(\ell(f(x^t), y^t)\right)\right] + \frac{\eta^2}{2}\mathbb{E}_{f\sim q_t}\left[\ell(f(x^t), y^t)^2\right].$$

As in the proof of Proposition 3, we remark that

$$-\log \mathbb{E}_{f\sim q_t}\left[\exp\left(\ell(f(x^t), y^t)\right)\right] = \Phi_\eta^t - \Phi_\eta^{t-1},$$

and summing over t we get

$$\eta\sum_{t=1}^T \mathbb{E}_{f\sim q_t}\left[\ell(f(x^t), y^t)\right] \leq \Phi_h^T - \Phi_n^0 + \frac{\eta^2}{2}\sum_{t=1}^T \mathbb{E}_{f\sim q_t}\left[\ell(f(x^t), y^t)^2\right].$$

As in the proof of Proposition 3, we can upper bound for any $f^* \in \mathcal{F}$.

$$\Phi_\eta^T \leq \eta\sum_{t=1}^T \ell(f^*(x^t), y^t),$$

while $\Phi_\eta^0 = -\log|\mathcal{F}|$. Hence we have that for any $f^* \in \mathcal{F}$

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{f}^t \sim q^t}\left[\ell(\hat{f}^t(x^t), y^t)\right] - \ell(f^*(x^t), y^t) \leq \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}_{\hat{f}^t \sim q^t}\left[\ell(\hat{f}^t(x^t), y^t)^2\right] + \frac{\log|\mathcal{F}|}{\eta}.$$

3. We choose $\eta = 1$, using the fact that $\ell(\cdot,\cdot) \in [0,1]$, we have that $\ell^2 \leq \ell$. Hence with the result of the previous question and the choice of f^* as the perfect expert, we get

$$\sum_{t=1}^{T} \mathbb{E}_{f \sim q_t}\left[\ell(f(x^t), y^t)\right] \leq \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}_{f \sim q_t}\left[\ell(f(x^t), y^t)\right] + \log|\mathcal{F}|.$$

And

$$\sum_{t=1}^{T} \mathbb{E}_{f \sim q_t}\left[\ell(f(x^t), y^t)\right] - \ell(f^*(x_t), y_t) \leq 2\log|\mathcal{F}|.$$

4. It is easy to see that for this algorithm at a given round, either no mistakes is made or at least half of the experts are discarded. Since this can happen only $\lfloor \log_2|\mathcal{F}| \rfloor$ times, the regret is of order $O(\log|\mathcal{F}|)$.

□

Chapter 2

Multi-Armed Bandits

Exercise 2.4 Adversarial Bandits

In this exercise, we will prove a regret bound for adversarial bandits (Section 2.5), where the sequence of rewards (losses) is non-stochastic. To make a direct connection to the Exponential Weights Algorithm, we switch from rewards to losses, mapping $r_t \circ 1 - r_t$, a transformation that does not change the problem itself. To simplify the presentation, suppose that a collection of losses

$$\left\{\boldsymbol{\ell}^t(\pi) \in [0,1] : \pi \in [A], t \in [T]\right\}$$

for each action π and time step t is arbitrary and chosen before round $t = 1$; this is referred to as an oblivious adversary. We denote by $\boldsymbol{\ell}^t = (\boldsymbol{\ell}^t(1), \ldots, \boldsymbol{\ell}^t(A))$ the vector of losses at time t. The protocol for the problem of adversarial multi-armed bandits (with losses) is as follows

- for $t = 1, \ldots, T$ do
  - Select decision $\pi^t \in \Pi := \{1, \ldots, A\}$ by sampling $\pi^t \sim p^t$
  - Observe loss $\boldsymbol{\ell}^t(\pi^t)$

Let p^t be the randomization distribution of the decision-maker on round t. Expected regret can be written as

$$\mathbb{E}\left[\mathbf{Reg}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p^t, \boldsymbol{\ell}^t \right\rangle\right] - \min_{\pi \in [A]} \sum_{t=1}^{T} \left\langle \boldsymbol{e}_\pi, \boldsymbol{\ell}^t \right\rangle.$$

Since only the loss of the chosen action $\pi_t \sim p_t$ is observed, we cannot directly appeal to the Exponential Weights Algorithm. The solution is to build an unbiased estimate of the vector $\boldsymbol{\ell}^t$ from the single real-valued observation $\boldsymbol{\ell}^t(\pi^t)$.

1. Prove that the vector $\widetilde{\boldsymbol{\ell}}^t(\cdot|\pi^t)$ defined by

$$\widetilde{\boldsymbol{\ell}}^t(\pi|\pi^t) = \frac{\boldsymbol{\ell}^t(\pi)}{p^t(\pi)}\mathbf{1}_{\{\pi^t=\pi\}}$$

is an *unbiased estimate* for $\boldsymbol{\ell}^t(\pi)$ for all $\pi \in [A]$. In vector notation, this means

$$\mathbb{E}_{\pi^t\sim p^t}\left[\widetilde{\boldsymbol{\ell}}^t(\cdot|\pi^t)\right] = \boldsymbol{\ell}^t.$$

Conclude that

$$\mathbb{E}\left[\mathbf{Reg}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\left\langle p^t, \widetilde{\boldsymbol{\ell}}^t\right\rangle\right]\right] - \min_{\pi\in[A]}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\left\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t\right\rangle\right]\right].$$

Above, we use the shorthand $\widetilde{\boldsymbol{\ell}}^t = \tilde{\boldsymbol{\ell}}^t(\cdot|\pi^t)$.

2. Show that given π',

$$\mathbb{E}_{\pi\sim p'}\left[\widetilde{\boldsymbol{\ell}}^t(\pi|\pi')^2\right] = \frac{\boldsymbol{\ell}^t(\pi')^2}{p^t(\pi^t)}, \text{ so that } \mathbb{E}_{\pi^t\sim p^t}\mathbb{E}_{\pi\sim p^t}\left[\widetilde{\boldsymbol{\ell}}^t(\pi|\pi^t)^2\right] \le A.$$

3. Define

$$p^t(\pi) \propto \exp\left\{-\eta\sum_{s=1}^{t-1}\left\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^s(\cdot|\pi^s)\right\rangle\right\},$$

which corresponds to the exponential weights algorithm on the estimated losses $\widetilde{\boldsymbol{\ell}}^s$. Apply (1.41) to the estimated losses to show that for any $\pi \in [A]$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\left\langle p^t, \widetilde{\boldsymbol{\ell}}^t\right\rangle\right]\right] - \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\left\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t\right\rangle\right]\right] \lesssim \sqrt{AT\log A}$$

Hence, the price of bandit feedback in the adversarial model, as compared to full-information online learning, is only A.

Solution : 1. From a direct calculation

$$\begin{aligned}\mathbb{E}_{\pi^t\sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi^t)\right] &= \sum_{\pi^t\in[A]} p^t(\pi_t)\frac{\boldsymbol{\ell}^t(\pi)}{p^t(\pi)}\mathbb{I}_{\{\pi^t=\pi\}}\\ &= p^t(\pi)\frac{\boldsymbol{\ell}^t(\pi)}{p^t(\pi)}\\ &= \boldsymbol{\ell}^t(\pi).\end{aligned}$$

In particular, by linearity of the expectation, we have that for any vector $x \in \mathbb{R}^{[}A]$, $\langle x, \boldsymbol{\ell}^t\rangle = \mathbb{E}_{\pi^t\sim p^t}\left[\langle x, \tilde{\boldsymbol{\ell}}^t\rangle\right]$. This directly implies

$$\mathbb{E}\left[\mathbf{Reg}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[\langle p^t, \widetilde{\boldsymbol{\ell}}^t \rangle\right]\right] - \min_{\pi \in [A]} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t \rangle\right]\right].$$

2. By a direct computation

$$\begin{aligned}\mathbb{E}_{\pi \sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi')^2\right] &= \sum_{\pi \in [A]} p^t(\pi) \frac{\boldsymbol{\ell}^t(\pi)^2}{p^t(\pi)^2} \mathbb{I}_{\{\pi' = \pi\}} \\ &= \frac{\boldsymbol{\ell}^t(\pi')^2}{p^t(\pi')}.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}_{\pi^t \sim p^t}\left[\mathbb{E}_{\pi \sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi^t)\right]\right] &\leq \mathbb{E}_{\pi^t \sim p^t}\left[\frac{\boldsymbol{\ell}^t(\pi^t)^2}{p^t(\pi^t)}\right] \\ &\leq \sum_{\pi^t \in [A]} p^t(\pi^t) \cdot \frac{\boldsymbol{\ell}^t(\pi^t)^2}{p^t(\pi^t)} \\ &\leq \sum_{\pi^t \in [A]} \boldsymbol{\ell}^t(\pi^t)^2 \\ &\leq A.\end{aligned}$$

3. In that setting, applying (1.41) to the sequence of losses $\tilde{\boldsymbol{\ell}}^t$ yields for any $\pi \in [A]$

$$\sum_{t=1}^{T} \langle p^t, \tilde{\boldsymbol{\ell}}^t \rangle - \sum_{t=1}^{T} \langle \boldsymbol{e}_\pi, \tilde{\boldsymbol{\ell}}^t \rangle \leq \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi^t)^2\right] + \frac{\log A}{\eta}.$$

Now using the result of the first question, we can bound the regret

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Reg}\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\langle p^t, \widetilde{\boldsymbol{\ell}}^t\rangle\right]\right] - \min_{\pi\in[A]}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t\rangle\right]\right] \\
&= \max_{\pi\in[A]}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\langle p^t, \widetilde{\boldsymbol{\ell}}^t\rangle\right] - \sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p^t}\left[\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t\rangle\right]\right] \\
&= \max_{\pi\in[A]}\mathbb{E}\left[\sum_{t=1}^{T}\langle p^t, \widetilde{\boldsymbol{\ell}}^t\rangle - \sum_{t=1}^{T}\langle \boldsymbol{e}_\pi, \widetilde{\boldsymbol{\ell}}^t\rangle\right] \\
&\leq \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_{\pi\sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi^t)^2\right] + \frac{\log A}{\eta}\right] \\
&= \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}_{\pi^t\sim p_t}\left[\mathbb{E}_{\pi\sim p^t}\left[\tilde{\boldsymbol{\ell}}^t(\pi|\pi^t)^2\right]\right] + \frac{\log A}{\eta}\right] \\
&\leq \mathbb{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}A + \frac{\log A}{\eta}\right] \\
&= \frac{\eta AT}{2} + \frac{\log A}{\eta} \\
&= \sqrt{2AT\log A},
\end{aligned}$$

where the 3rd and 4th equality come from the tower rule of expectation, the first inequality from (1.41), the second inequality from the previous question and the last equality from the choice $\eta = \sqrt{\frac{2\log[A]}{AT}}$. □

Chapter 3

Contextual Bandits

Exercise 3.5 Unstructured Contextual Bandits

Consider a contextual bandit problem with a finite set $\mathcal{X}$ of possible contexts, and a finite set of actions $\mathcal{A}$. Show that running UCB independently for each context yields a regret bound of the order $\widetilde{\mathcal{O}}(\sqrt{|\mathcal{X}||\mathcal{A}|T})$ in expectation, ignoring logarithmic factors. In the setting where $\mathcal{F} = \mathcal{X} \times \mathcal{A} \to [0,1]$ is unstructured, and consists of all possible functions, this is essentially optimal.

Exercise 3.6 ϵ-Greedy with Offline Oracles

In Proposition 8, we analyzed the ϵ-Greedy contextual bandit algorithm assuming access to an online regression oracle. Because we appeal to online learning, this algorithm was able to handle adversarial contexts $x^1, \dots, x^T$. In the present problem, we will modify the ϵ-greedy algorithm and proof to show that if contexts are stochastic (that is $x^t \sim \mathcal{D}\ \forall t$, where $\mathcal{D}$ is a fixed distribution), ϵ-greedy works even if we use an *offline oracle*. We consider the following variant of ϵ-greedy. The algorithm proceeds in epochs $m = 0, 1, \dots$ of doubling size

$$\{2\}, \{3,4\}, \{5 \dots 8\}, \dots \{2^m + 1, 2^{m+1}\}, \dots, \{T/2 + 1, T\};$$

we assume without loss of generality that T is a power of 2, and that an arbitrary decision is made on round $t = 1$. At the end of each epoch $m - 1$, the offline oracle is invoked with the data from the epoch, producing an estimated model $\widehat{f}^m$. This model is used for the greedy step in the next epoch m. In other words, for any round $t \in [2^m + 1, 2^{m+1}]$ of epoch m, the algorithm observes $x^t \sim \mathcal{D}$, chooses an action $\pi^t \sim \text{unif}[A]$ with probability ϵ and chooses the greedy action

$$\pi^t = \arg\max_{\pi \in [A]} \widehat{f}^m(x^t, \pi)$$

with probability $1-\epsilon$. Subsequently, the reward r^t is observed.

1. Prove that for any $T \in \mathbb{N}$ and $\delta > 0$, by setting ϵ appropriately, this method ensures that with probability at least $1-\delta$,

$$\mathbf{Reg} \lesssim A^{1/3}T^{1/3}\left(\sum_{m=1}^{\log_2 T} 2^{m/2}\mathbf{Est}_{\mathrm{Sq}}^{\mathrm{off}}(\mathcal{F}, 2^{m-1}, \delta/m^2)^{1/2}\right)^{2/3}$$

2. Recall that for a finite class, ERM achieves $\mathbf{Est}_{\mathrm{Sq}}^{\mathrm{off}}(\mathcal{F}, T, \delta) \lesssim \log(|\mathcal{F}|/\delta)$. Show that with this choice, the above upper bound matches that in Proposition 8, up to logarithmic in T factors.

Exercise 3.7 Model Misspecification in Contextual Bandits

In Proposition 10, we showed that for contextual bandits with a general class $\mathcal{F}$, SquareCB attains regret

$$\mathbf{Reg} \lesssim \sqrt{AT\mathbf{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta)}.$$

To do so, we assumed that $f^* \in \mathcal{F}$, where $f^*(x,a) := \mathbb{E}_{r\sim M^*(\cdot|x,a)}[r]$; that is, we have a well- specified model. In practice, it may be unreasonable to assume that we have $f^* \in \mathcal{F}$. Instead, a weaker assumption is that there exists some function $\bar{f} \in \mathcal{F}$ such that

$$\max_{x\in\mathcal{X}, a\in\mathcal{A}} |\bar{f}(x,a) - f^*(x,a)| \le \epsilon$$

for some $\epsilon > 0$; that is, the model is ϵ-*misspecified*. In this problem, we will generalize the regret bound for SquareCB to handle misspecification. Recall that in the lecture notes, we assumed that the regression oracle satisfies

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t}\left[(\widehat{f}^t(x^t,\pi^t) - f^*(x^t,\pi^t))^2\right] \le \mathbf{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta).$$

In the misspecified setting, this is too much to ask for. Instead, we will assume that the oracle satisfies the following guarantee for every sequence:

$$\sum_{t=1}^{T}(\widehat{f}^t(x^t,\pi^t) - r^t)^2 - \min_{f\in\mathcal{F}}\sum_{t=1}^{T}(f(x^t,\pi^t) - r^t)^2 \le \mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T).$$

Whenever $f^* \in \mathcal{F}$, we have $\mathbf{Est}_{\mathrm{Sq}}(\mathcal{F}, T, \delta) \lesssim \mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T) + \log(1/\delta)$ with probability at least $1-\delta$. However, it is possible to keep $\mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T)$ small even when $f^* \notin \mathcal{F}$. For example, the averaged exponential weights algorithm satisfies this guarantee with $\mathbf{Reg}_{\mathrm{Sq}} \lesssim \log|\mathcal{F}|$, regardless of whether $f^* \in \mathcal{F}$.

We will show that for every $\delta > 0$, with an appropriate choice of γ, SquareCB (that is, the algorithm that chooses $p^t = \mathbf{IGW}_\gamma(\widehat{f}^t(x^t, \cdot))$ ensures that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{AT(\mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T) + \log(1/\delta))} + \epsilon A^{1/2} T.$$

1. Show that for any sequence of estimators $\widehat{f}^1, \ldots, \widehat{f}^t$, by choosing $p_t = \mathbf{IGW}_\gamma(\widehat{f}^t(x^t, \cdot))$, we have that

$$\mathbf{Reg} = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[f^*(x^t, \pi^*(x^t)) \right] \lesssim \frac{AT}{\gamma} + \gamma \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \right] + \epsilon T.$$

2. Show that the following inequality holds for every sequence

$$\sum_{t=1}^{T} (\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \leq \mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T) + 2 \sum_{t=1}^{T} (r^t - \bar{f}(x^t, \pi^t))(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t)).$$

3. Using Freedman's inequality (Lemma 35), show that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \right] \leq 2 \sum_{t=1}^{T} (\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 + \mathcal{O}(\log(1/delta)).$$

4. Using Freedman's inequality once more, show that with probability at least $1 - \delta$;

$$2 \sum_{t=1}^{T} (r^t - \bar{f}(x^t, \pi^t)(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t)) \leq \frac{1}{4} \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \right] + \mathcal{O}(\epsilon^2 T + \log(1/\delta)).$$

Conclude that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[(\widehat{f}^t(x^t, \pi^t) - \bar{f}(x^t, \pi^t))^2 \right] \lesssim \mathbf{Reg}_{\mathrm{Sq}(\mathcal{F}, T)} + \epsilon^2 T + \log(1/\delta).$$

5. Combining the previous results, show that for any $\delta > 0$, by choosing $\gamma > 0$ appropriately, we have that with probability at least $1 - \delta$,

$$\mathbf{Reg} \lesssim \sqrt{AT(\mathbf{Reg}_{\mathrm{Sq}}(\mathcal{F}, T) + \log(1/\delta))} + \epsilon A^{1/2} T.$$

Chapter 4

Structured Bandits

Chapter 6

General Decision Making