אוניברסיטת תל אביב

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב

למידה בזמן אמת עם משוב יריבי מורעש

חיבור זה מוגש כחלק ממילוי הדרישות לקבלת התואר "מוסמך למדעים" (.M.Sc) בבית הספר למדעי המחשב באוניברסיטת תל-אביב

ע"י

אלון רסלר

העבודה הוכנה בהדרכתו של

פרופסור ישי מנצור

אוקטובר 2018

תקציר

אנו מציגים ולומדים מודלים של למידה בזמן אמת (online learning) כאשר המשוב (feedback) שנצפה ע"י האלגוריתם הלומד מורעש על ידי רעש אקראי וכאשר המשוב הוא משוב מלא (full feedback) או משוב בנדיט (bandit feedback).

אנו מתמקדים במודלים עם פונקצית הפסד בינארית כאשר היא מורעשת ע"י פעולת xor עם ביט אקראי שהוא משתנה מקרי ברנולי.

אנו מתמקדים בשני מודלים עיקריים: רעש קבוע – כאשר הביט האקראי מתפלג לפי התפלגות ברנולי עם פרמטר קבוע p, ורעש משתנה – כאשר הביט האקראי מתפלג לפי התפלגות ברנולי עם פרמטר משתנה p.

התוצאות העיקריות שלנו הן חסמים הדוקים על החרטה (regret) עבור למידה און-ליין עם רעש במודל היריבי.

TEL-AVIV UNIVERSITY

RAYMOND AND BEVERLY SACKLER

FACULTY OF EXACT SCIENCES

SCHOOL OF COMPUTER SCIENCE

Adversarial Online Learning with noise

Submitted as a partial fulfillment of the
requirements towards Master of Science degree
Tel Aviv University
School of Computer Science

by

Alon Resler

Under the supervision of
Prof. Yishay Mansour

October 2018

Abstract

We present and study models of adversarial online learning where the feedback observed by the learner is *noisy*, and the feedback is either *full information* feedback or *bandit* feedback. Specifically, we consider binary losses xored with the noise, which is a Bernoulli random variable. We consider both a constant noise rate and a variable noise rate. Our main results are tight regret bounds for learning with noise in the adversarial online learning model.

4

Acknowledgments

Contents

Chapter 1

Introduction

Online learning is a general framework for sequential decision-making under uncertainty. In each round, a learner chooses an action from a set of K available actions and suffers a loss associated with that action and observes "some" feedback about the losses. The losses in each round are arbitrary, possibly adversarial, and the goal of the learner is to minimize the cumulative loss over a fix time horizon T. We measure the performance of the learner using the *regret* which is the expected difference between the cumulative loss of the learner and that of the best fixed action.

Traditionally, there are two main types of feedback: *full-information* feedback and the *Bandit* feedback. In the *full-information* feedback, often referred to as *prediction with expert advice*, in each round the learner observes the losses of all actions. A typical example of the *full-information* feedback is a hypothetical stock investor who invests all of his money in one of K stocks on each day. At the end of the day, the investor incurs the outcome associated with the chosen stock and observes the outcomes of all the stocks. In the *Bandit* feedback the learner only observes the loss associated with the action played. The typical example of the *bandit* feedback is online advertising. Consider an Internet website that presents one of K ads to each user, and its goal is to maximize the number of clicked ads. Naturally, we know whether the user clicked on the presented ad, but we have no information about other ads (whether the user would have clicked on them, if they were presented).

Both models have been extensively studied and received significant practical and theoretical interest. The regret bound for the full information model is $\Theta(\sqrt{T \ln K})$, and for the bandit model is $\Theta(\sqrt{KT})$. (See, Cesa-Bianchi and Lugosi (2006); Bubeck and Cesa-Bianchi (2012).) However, both models assumes that the observed feedback is exact. In some real life scenarios, the feedback can be corrupted by noise, which is the focus of our work. For example, in the web-advertising example, we might observe an incorrect feedback since a click event might be missing (due to network connection problem or logging error), recorded incorrectly (due to a browser issue, such as privacy setting), or alternatively, the user might be misidentified (due to multiple users using the same computer).

Feedback type \ Noise model	Constant noise	Variable noise (Uniform)
Full information (known noise)	$\Theta(\frac{1}{\epsilon}\sqrt{T \ln K})$	$\Theta(T^{2/3} \ln^{1/3} K)$
Full Information (unknown noise)	$\Theta(\frac{1}{\epsilon}\sqrt{T \ln K})$	$\Theta(T)$
Bandit (known noise)	$\tilde{\Theta}(\frac{1}{\epsilon}\sqrt{TK})$	$\tilde{\Theta}(T^{2/3} K^{1/3})$
Bandit (unknown noise)	$\tilde{\Theta}(\frac{1}{\epsilon}\sqrt{TK})$	$\Theta(T)$

Figure 1.1: Results summery

In this paper, we present and study settings in which the feedback is corrupted by random noise. We assume that the losses are Boolean and that the noise is also Boolean, and the observation is the xor of the loss and the noise. For the noise we consider Bernoulli random variable with probability p, denoted by $B(p)$. We consider a few variations of the noise model:

For the *constant noise rate*, we assume that there is a fixed probability p for the noise (for all actions and rounds). For the *variable noise rate*, we assume that there exists a distribution D such that in each round t we draw a vector of probabilities, where is $p_{i,t}$ the noise of action i in round t. For both settings, we study both the case that the noise is known to the learner, and where it is unknown. Our main contribution is deriving tight regret bounds for those settings, both upper bounds (algorithms) and lower bounds (impossibility results). In the following we give a high level view of our results.

The *constant noise model* has a fixed parameter $\epsilon \in [0, 1]$ and for every round t the loss is xored with Bernoulli random variable with parameter $p = \frac{1-\epsilon}{2}$. For the full information model we have a tight regret bound of $\Theta(\frac{1}{\epsilon}\sqrt{T \ln K})$, both when the noise parameter is known and when it is unknown. For the bandit feedback model we have a tight regret bound of $\tilde{\Theta}(\frac{1}{\epsilon}\sqrt{TK})$, both when the noise parameter is known and when it is unknown.

The *variable noise model* has a distribution D over $[0,1]^K$ and at each round t, we draw from D a realized noise vector $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $p_{i,t} = (1 - \epsilon_{i,t})/2$ is the noise parameter for action i at round t. In the following we describe our results for the *uniform model*, where the marginal distribution of D of each action is uniform $[0,1]$. For the full information we have a contrast between the case where the realized noise is observed, where we have a tight regret bound of $\Theta(T^{2/3} \ln^{1/3} K)$, and the case where the realized noise is not observed, where we have a linear regret, i.e., $\Theta(T)$. For the bandit model we have a tight bound of $\tilde{\Theta}(T^{2/3} K^{1/3})$, when the realized noise is not observed, and linear regret, i.e., $\Theta(T)$, when the realized noise is not observed. We also discuss the case of a general distribution and derive regret bounds for other specific distributions. Our main results are summarized in Figure 1.

Related work

The work of Kocák et al. (2016) generalized a partial-feedback scheme proposed by (Mannor and Shamir, 2011; Alon et al., 2017), in which the learner observes losses associated with a subset of actions which depends on the selected action, and considered a zero mean noise added to the side observations. Their main result is an algorithm that guarantees a regret of $\tilde{O}(\sqrt{T})$, where the constant depends on a graph property.

The work of Wu et al. (2015) studies a stochastic model where the feedback of an action has the losses of each other action with an additive noise of a zero-mean Gaussian, where variance depends both on the action played and observed. For this model they derive problem-depend lower bounds and matching upper bounds.

Gajane et al. (2018) studied a stochastic bandit problem where the feedback is drawn from a different distribution than the rewards, but there exist a link function relating them. They provide lower and upper bound for this setting.

Binary sequence prediction with noise was studied by Weissman and Merhav (2000) and Weissman et al. (2001). They show upper bounds on the regret for binary sequence prediction with a constant noise rate (the binary sequence prediction model is implicitly a full feedback model). Their regret bound is similar to our regret bound (in the full information with constant noise).

There is a vast literature in statistics, operation research and machine learning regarding various noise models. In computational learning theory, popular noise models include random classification noise Angluin and Laird (1988) and malicious noise Valiant (1985); Kearns and Li (1993). The above noise models use the PAC model, and study the generalization error, while we consider an online setting and study the regret.

Chapter 2

Model and Preliminaries

1. The Model

We consider adversarial decision problem with finite *actions* (or *actions*) set $A = \{1, 2, \ldots, K\}$. On each round $t = 1, 2, \ldots, T$ the *environment* selects a loss vector $\vec{\ell}_t \in \{0, 1\}^K$ where $\ell_{i,t}$ is the loss associated with action i at round t. Then, the *learner* (or *algorithm*) chooses an *action* I_t and incurs a loss $\ell_{I_t, t}$.

The main difference between our models and the standard online model is that the learner observes a noisy feedback of the loss (to be specified separately in each setting). Before presenting our models we start with a general definition of a noisy feedback of a single loss.

Definition 1 *Let* $\ell \in \{0, 1\}$ *be a loss, and let* $\epsilon \in [0, 1]$ *be a parameter. We define the* ϵ-**noisy feedback** *to be the following the random variable*

$$c = \ell \oplus R_\epsilon$$

where R_ϵ *is Bernoulli random variable with parameter* $p = \frac{1-\epsilon}{2}$ *(i.e.,* $\Pr[R_\epsilon = 1] = p = \frac{1-\epsilon}{2}$ *).*

Using the above definition we present our four different settings, which are different in the feedback that the learner observes and the noise parameter selection. The settings are as follow:

1. **Full Information with Constant Noise:** In this setting, there exists a constant noise parameter $\epsilon \in [0, 1]$, such that for every round t the learner observes the ϵ-*noisy feedback*, $c_{i,t}$ for each action i, i.e., $c_{i,t} = \ell_{i,t} \oplus R_\epsilon$.

2. **Full Information with Variable Noise:** In this setting, there exists a distribution D over $[0, 1]^K$. At the beginning of each round t, we draw from D a realized noise $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $\epsilon_{i,t} \in [0, 1]$ is the noise parameter action i at round t. We assume that the noise vectors are drawn independently from D at each round t. The learner observes, for each action i, an $\epsilon_{i,t}$-*noisy feedback* $c_{i,t}$, i.e., $c_{i,t} = \ell_{i,t} \oplus R_{\epsilon_{i,t}}$.

3. **Bandit with Constant Noise:** In this setting, there exists a constant noise parameter $\epsilon \in [0, 1]$, such that for every round t the learner observes the ϵ-*noisy feedback* of the action he played, i.e., $c_{I_t,t} = \ell_{i,t} \oplus R_\epsilon$ where I_t is the action played in round t.

4. **Bandit with Variable Noise:** In this setting, there exists a distribution D over $[0, 1]^K$. At the beginning of each round t, we draw from D a realized noise $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $\epsilon_{i,t} \in [0, 1]$ is the noise parameter action i at round t. We assume that the noise vectors are drawn independently from D at each round t.

   The learner observes only the feedback for the action he played, i.e., $c_{I_t,t} = \ell_{I_t,t} \oplus R_{\epsilon_{I_t,t}}$, where I_t is the action played in time t.

Each of the models can have two variants: *known noise parameters*, where the learner observes the noise parameters or *unknown noise parameters*, where the learner doesn't observe the noise parameters. In the constant noise, the noise parameter is ϵ and in the variable noise, the noise parameters are the realized noise parameters at each round t, i.e., $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$. For our main results, we assume that the noise parameters are *known* to the learner. When we examine the setting where the learner does not know the noise parameters, we state it explicitly.

We measure the performance of the learner using the (expected) *regret* of the *true losses*, namely,

$$Regret(T) = \mathrm{E}\bigg[\sum_{t=1}^{T} \ell_{I_t,t}\bigg] - \min_{i \in A} E\bigg[\sum_{t=1}^{T} \ell_{i,t}\bigg]$$

where the losses are selected by an adversary and the expectation is taken over the randomness of the algorithm and the randomness of the noise.

The algorithms presented in this paper are variants of the *Exponential Weights Scheme* (see Algorithm 1). In the *Exponential Weights Scheme* (**EWS**) the algorithm maintains weight $w_{i,t}$ for each action i (initially $w_{i,t} = 1$). On round t the algorithm chooses an action proportional to the weights, based on a distribution q_t. After observing the feedback of round t, the algorithm updates the weights to $w_{i,t+1}$ using the previous weights $w_{i,t}$, the observations (i.e., $c_{i,t}$) and the noise parameter. Each noise setting determines how the feedback is constructed and observed (line 8 in the algorithm template). Each algorithm determines how to construct the loss estimate $\hat{\ell}_{i,t}$ (line 9 in the algorithm). We denote generically $\hat{\ell}_{i,t} = EST(i, c_{i,t}, \vec{q}_t, I_t)$, where EST is the loss estimate function that will be implemented differently in each setting and for each algorithm.

Notations: Let $\hat{L}_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\hat{\ell}_{i,t}$ and $\hat{L}_{k,T} = \sum_{t=1}^{T} \hat{\ell}_{k,t}$ the estimated loss of the online algorithm and of action k, respectively. We denote by $L_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t}$ and $L_{k,T} = \sum_{t=1}^{T} \ell_{k,t}$ the expected loss of the online algorithm and the loss of action k, respectively.

Algorithm 1 Exponential Weights Scheme

1: **Initialization:** $w_{i,1} = 1$ for all $i \in A$
2: **Parameters:** $\eta > 0$
3: $r \leftarrow a \bmod b$
4: **for** $t = 1, 2, ..., T$ **do**
5:      Construct the probability distribution q_t with

$$q_{i,t} = \frac{w_{i,t}}{W_t} \text{ where } W_t = \sum_{i=1}^{K} w_{i,t}$$

6:      Play a random action I_t according to q_t
7:      Incur loss $\ell_{I_t,t}$
8:      Observe feedback according to the specific settings
9:      Construct loss estimate $\hat{\ell}_{i,t} = EST(i, c_{i,t}, \vec{q}_t, I_t)$ for all $i \in A$
10:      Update weights for all $i \in A$:

$$w_{i,t+1} = w_{i,t} \exp(-\eta \hat{\ell}_{i,t})$$

We denote by $B(p)$ the *Bernoulli distribution* with parameter p and by $B(n, p)$ the *Binomial distribution* with n trials and parameter p.

2. Preliminaries

The following lemma establishes a well known property of EWS, and for completeness we give its proof.

Lemma 2 *Let $\eta > 0$ and a sequence of loss estimates $\hat{\ell}_1, \ldots, \hat{\ell}_T$ where t $\hat{\ell}_t : \{1, \ldots, K\} \to \mathbb{R}$ such that $-\eta \hat{\ell}_{i,t} \leq 1$ for all i and t, then the probability vectors $\vec{q}_1, \ldots, \vec{q}_T$ define in the Exponential Weights Scheme, for any action k, satisfies*

$$\sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} \hat{\ell}_{i,t} - \sum_{t=1}^{T} \hat{\ell}_{k,t} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} (\hat{\ell}_{i,t})^2$$

Proof The proof follows the standard analysis of exponential weighting schemes: let $W_t = \sum_{i=1}^{K} w_{i,t}$ using the algorithm update we can write

$$
\frac{W_{t+1}}{W_t} = \sum_{i=1}^{K} \frac{w_{i,t+1}}{W_t} = \sum_{i=1}^{K} \frac{w_{i,t} e^{-\eta \hat{\ell}_{i,t}}}{W_t} = \sum_{i=1}^{K} q_{i,t} e^{-\eta \hat{\ell}_{i,t}}
$$

$$
\leq \sum_{i=1}^{K} q_{i,t}(1 - \eta \hat{\ell}_{i,t} + \eta^2 (\hat{\ell}_{i,t})^2) \qquad \text{(using } e^x \leq 1 + x + x^2 \text{ for } x \leq 1)
$$

$$
= 1 - \eta \sum_{i=1}^{K} q_{i,t} \hat{\ell}_{i,t} + \eta^2 \sum_{i=1}^{K} q_{i,t}(\hat{\ell}_{i,t})^2
$$

Taking logs and using $ln(1-x) \leq -x$ for all x and summing for $t = 1, 2, ..., T$ yields

$$
ln\frac{W_{T+1}}{W_1} \leq -\eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} \hat{\ell}_{i,t} + \eta^2 \sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t}(\hat{\ell}_{i,t})^2
$$

Moreover, for any fixed action k we have $W_t \geq w_{k,t}$, thus:

$$
\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_{k,T+1}}{W_1} = -\eta \sum_{t=1}^{T} \hat{\ell}_{k,t} - \ln K
$$

Putting together and rearranging gives:

$$
\sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} \hat{\ell}_{i,t} - \sum_{t=1}^{T} \hat{\ell}_{k,t} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t}(\hat{\ell}_{i,t})^2
$$

∎

Chapter 3

Full Information Models

In this chapter, we consider the *Full Information* feedback models. In the first section we study the *constant noise* feedback setting and in the second section we study the *variable noise* feedback setting.

1. Constant Noise

In the first part, we derive an algorithm that uses the constant noise parameter ϵ and obtains regret bound of $O(\frac{1}{\epsilon}\sqrt{T \ln K})$. Then, we show how to obtain the same regret bound when the noise parameter ϵ is unknown. In the second part, we derive a lower bound, which shows that the regret of our algorithm is asymptotically optimal.

1.1 Algorithms

In this section we derive the algorithms that establish the upper bound on the regret. The idea is to construct an unbiased estimator for each loss. Let $\epsilon \in [0, 1]$ and let $p = \frac{1-\epsilon}{2}$ be the noise parameter. The unbiased estimator is

$$EST(i, c_{i,t}, \vec{q_t}, I_t) = \frac{c_{i,t} - p}{1 - 2p} = \hat{\ell}_{i,t} \ .$$

The estimator is unbiased since,

$$\mathrm{E}[\hat{\ell}_{i,t}] = \frac{p(1 - \ell_{i,t}) + (1 - p)\ell_{i,t} - p}{1 - 2p} = \ell_{i,t} \ .$$

The following theorem establishes the regret bound when we use the Exponential Weights Scheme with the above unbiased estimator.

Theorem 3 *Let $\epsilon \in [0, 1]$, denote $p = \frac{1-\epsilon}{2}$ and assume $T \geq \frac{1}{4}\ln K$. Then running Exponential Weights Scheme under the Full Information with Constant Noise setting with the following loss estimate*

$$EST(i, c_{i,t}, q_t, I_t) = \frac{c_{i,t} - p}{1 - 2p} = \hat{\ell}_{i,t}$$

and for $\eta = \epsilon \sqrt{\frac{\ln K}{T}}$ we have,

$$Regret(T) \leq \frac{2}{\epsilon}\sqrt{T \ln K}$$

Proof Since

$$\hat{\ell}_{i,t} = \frac{c_{i,t} - p}{1 - 2p} \in \{\frac{1 - p}{1 - 2p}, \frac{-p}{1 - 2p}\} = \{-\frac{1 - \epsilon}{2\epsilon}, \frac{1 + \epsilon}{2\epsilon}\},$$

we have that

$$-\eta\hat{\ell}_{i,t} \leq \epsilon\sqrt{\frac{\ln K}{T}}\frac{1 - \epsilon}{2\epsilon} \leq \frac{\sqrt{\frac{\ln K}{T}}}{2} \leq 1$$

where the last equation uses $T \geq \frac{1}{4}\ln K$. Thus, we can apply Lemma 2 and obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\hat{\ell}_{i,t} - \sum_{t=1}^{T} \hat{\ell}_{k,t} \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}(\hat{\ell}_{i,t})^2$$

Taking expectation on both sides and using that the estimator is unbiased (i.e., $E[\hat{\ell}_{i,t}] = \ell_{i,t}$) yields,

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \sum_{t=1}^{T} \ell_{k,t} \leq \frac{lnK}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}E[(\hat{\ell}_{i,t})^2]$$

Using the fact that $Regret(T) = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \min_{k \in A}\sum_{t=1}^{T} \ell_{k,t}$, we have,

$$Regret(T) \leq \frac{lnK}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}E[(\hat{\ell}_{i,t})^2]$$

We bound the second moments of the estimate as follows,

$$E[(\hat{\ell}_{i,t})^2] = p\frac{(\bar{\ell}_{i,t} - p)^2}{(1 - 2p)^2} + (1 - p)\frac{(\ell_{i,t} - p)^2}{(1 - 2p)^2} \leq \frac{1}{(1 - 2p)^2} = \frac{1}{\epsilon^2} \,,$$

where $\bar{\ell}_{i,t} = 1 - \ell_{i,t}$. Putting it back together and plugging $\eta = \epsilon\sqrt{\frac{\ln K}{T}}$ we obtain

$$Regret(T) \leq \frac{\ln K}{\eta} + \frac{\eta T}{\epsilon^2} \leq \frac{2}{\epsilon}\sqrt{T \ln K}$$

■

   In Theorem 3, the learner uses the noise parameter ϵ to derive an unbiased estimator. The following theorem shows that the same regret bound can be attained even when the leaner does not know the noise parameter ϵ.

Theorem 4 *Let $\epsilon \in [0, 1]$ and denote $p = \frac{1-\epsilon}{2}$. Running Exponential Weights Scheme under the Full Information with Constant Noise setting with the following loss estimate*

$$EST(i, c_{i,t}, \vec{q_t}, I_t) = c_{i,t} = \hat{\ell}_{i,t}$$

and for $\eta = \epsilon \sqrt{\frac{\ln K}{T}}$, we have,

$$Regret(T) \leq \frac{2}{\epsilon}\sqrt{T \ln K}$$

Proof By applying Lemma 2 and taking expectation on both sides we obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2]$$

Calculating the expectation of the estimator $\hat{\ell}_{i,t}$, and since $\ell_{i,t} \in \{0, 1\}$, we have,

$$\mathrm{E}[\hat{\ell}_{i,t}] = (1-p)\ell_{i,t} + p\bar{\ell}_{i,t} = (1-2p)\ell_{i,t} + p = |\ell_{i,t} - p|$$

For the second moment we have $(\hat{\ell}_{i,t})^2 = c_{i,t}^2 = c_{i,t} \leq 1$. Putting things together we have

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}|\ell_{i,t} - p| - \sum_{t=1}^{T}|\ell_{k,t} - p| \leq \frac{\ln K}{\eta} + \eta T \tag{3.1}$$

Using the notation of $\hat{L}_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\hat{\ell}_{i,t}$, $\hat{L}_{k,T} = \sum_{t=1}^{T}\hat{\ell}_{k,t}$, $L_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t}$, and $L_{k,T} = \sum_{t=1}^{T}\ell_{k,t}$, we can write inequality (3.1) as

$$\mathrm{E}[\hat{L}_{ON,T}] - \mathrm{E}[\hat{L}_{k,T}] \leq \frac{\ln K}{\eta} + \eta T$$

Denote by $G_{t,b} = \{i \in A \mid \ell_{i,t} = b\}$ the set of actions with loss $b \in \{0, 1\}$ in round t. Denote by $Q_t = \sum_{i \in G_{t,1}} q_{i,t}$ the distribution mass the learner gives actions in $G_{t,1}$. Using this notation we have $L_{ON,T} = \sum_{t=1}^{T} Q_t$. Now calculate the value of the estimated losses of the online algorithm,

$$\mathrm{E}[\hat{L}_{ON,T}] = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}|\ell_{i,t} - p| = \sum_{t=1}^{T}[p\sum_{i \in G_{t,0}} q_{i,t} + (1-p)\sum_{i \in G_{t,1}} q_{i,t}]$$

$$= \sum_{t=1}^{T}[p(1 - Q_t) + (1-p)Q_t] = \sum_{t=1}^{T}[p + (1-2p)Q_t]$$

$$= (1-2p)L_{ON,T} + pT$$

Similarly, for the term $\mathrm{E}[\hat{L}_{k,T}]$ we have,

$$\mathrm{E}[\hat{L}_{k,T}] = \sum_{t=1}^{T} |\ell_{k,t} - p| = \sum_{t|\ell_{t,k}=0} p + \sum_{t|\ell_{t,k}=1} (1 - p)$$

$$= p(T - L_{k,T}) + (1 - p)L_{k,T} = (1 - 2p)L_{k,T} + pT$$

Putting all together,

$$\mathrm{E}[\hat{L}_{ON,T}] - \mathrm{E}[\hat{L}_{k,T}] = (1-2p)L_{ON,T} + pT - [(1-2p)L_{k,T} + pT] = (1-2p)[L_{ON,T} - L_{k,T}]$$

Dividing by both sides of inequity by $(1 - 2p)$ and using $\eta = \sqrt{\frac{\ln K}{T}}$ we obtain that

$$Regret(T) = L_{ON,T} - \min_{k \in A} L_{k,T} \leq \frac{1}{1 - 2p}(\frac{\ln K}{\eta} + \eta T) = \frac{2}{\epsilon}\sqrt{T \ln K}$$

∎

1.2 Impossibility result

In this section we derive a lower bound on the regret for the Full Information with Constant Noise model. Our lower bound matches our upper bound, up to a constant factor. Specifically, the following theorem gives us a lower bound of $\Omega(\frac{1}{\epsilon}\sqrt{T \ln K})$ on the regret.

Theorem 5 *Consider the Full Information with Constant Noise setting with noise parameter $\epsilon \in (0, \frac{1}{2})$, $T \geq 2 \ln K$ and $K \geq e^{18}$. Then for any algorithm, there exists a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$Regret(T) = \Omega(\min\{\frac{1}{\epsilon}\sqrt{T \ln K}, T\})$$

To prove the theorem we first define the following adversarial loss assignment strategy:

- the adversary initially picks uniformly a *best action* $i^\star$ ($\forall i \; Pr[i^\star = i] = \frac{1}{K}$)

- at round t: the adversary draws losses for the actions from the following distributions:

  1. for $i^\star$: $\ell_{i^\star,t} \sim B(\frac{1}{2} - \delta)$
  2. for $i \neq i^\star$: $\ell_{i,t} \sim B(\frac{1}{2})$

where $\delta = \min\{\frac{1}{6\epsilon}\sqrt{\frac{\ln K}{T}}, \frac{1}{2}\}$. Now we calculate the distribution of the ϵ-**noisy feed-back** $c_{i,t}$. Starting with the best action we have

$$Pr[c_{i^\star,t} = 1] = Pr[\ell_{i^\star,t} = 1]Pr[R_\epsilon = 0] + Pr[\ell_{i^\star,t} = 0]Pr[R_\epsilon = 1]$$
$$= (\frac{1}{2} - \delta)\frac{1 + \epsilon}{2} + (\frac{1}{2} + \delta)\frac{1 - \epsilon}{2} = \frac{1}{2} - \epsilon\delta$$

For $i \neq i^\star$ we have

$$Pr[c_{i,t} = 1] = Pr[\ell_{i,t} = 1]Pr[R_\epsilon = 0] + Pr[\ell_{i,t} = 0]Pr[R_\epsilon = 1]$$
$$= \frac{1}{2}\frac{1 + \epsilon}{2} + \frac{1}{2}\frac{1 - \epsilon}{2} = \frac{1}{2}$$

Thus, we have: $c_{i^\star,t} \sim B(\frac{1}{2} - \epsilon\delta)$ and $c_{i,t} \sim B(\frac{1}{2})$ for $i \neq i^\star$.

The following is a standard claim regarding the minimum of i.i.d binomial random variables.

Lemma 6 *Let* $X_1, ..., X_{K-1}$ *be i.i.d random variables with distribute* $B(n, p)$ *such that* $p \in (\frac{1}{4}, \frac{1}{2})$, $n \geq 2\ln K$ *and* $K \geq e^{18}$. *Then with probability of at least* $\frac{1}{2}$ *we have*

$$\min\{X_1, ..., X_{K-1}\} \leq np - \sqrt{\frac{p}{9}n\ln K}$$

Proof Denote by $Y = \min\{X_1, ..., X_{K-1}\}$ then by interdependency we can write

$$Pr[Y \leq np - t] = 1 - Pr[\forall i \in \{1, 2, ..., K - 1\}X_i \geq np - t]$$
$$= 1 - (Pr[X_1 \geq np - t])^{K-1} \tag{3.2}$$

Now we want to bound $Pr[X_1 \geq np - t]$. Rearranging, and using *Lemma 5.2* of Klein and Young (1999), for $t \leq \frac{1}{2}pn$ we can bound

$$Pr[X_1 \geq np - t] = Pr[X_1 - np \geq -t] = 1 - Pr[X_1 - np \leq -t]$$
$$\leq 1 - \exp(-\frac{9t^2}{np}) = 1 - \frac{1}{K}$$

where in the last equation we take $t = \sqrt{\frac{p}{9}n\ln K} \leq \frac{1}{2}pn$. Plugging it back in (3.2) we obtain

$$Pr[Y \leq np - \sqrt{\frac{p}{9}n\ln K}] \geq 1 - (1 - \frac{1}{K})^{K-1} \geq \frac{1}{2}$$

∎

Denoting by $C_{i,T} = \sum_{t=1}^{T} c_{i,t}$, the sum of the noisy feedback of action i. Note that this is binomial random variable. In addition, for $i^\star$ we have $C_{i^\star,T} \sim B(T, \frac{1}{2} - \epsilon\delta)$ and for $i \neq i^\star$ we have $C_{i,T} \sim B(T, \frac{1}{2})$. By applying Lemma 6 on the noisy-feedbacks we show the following corollary.

Corollary 7 *With probability at least $\frac{1}{4}$ there exist action $j \neq i^\star$ such that $C_{j,T} < C_{i^\star,T}$.*

Proof Applying Lemma 6 on the $K-1$ actions with $c_{i,t} \sim B(\frac{1}{2})$ we obtain that with probability at least $\frac{1}{2}$ there exist action $j \neq i^\star$ such that

$$C_{j,T} \leq \frac{T}{2} - \sqrt{\frac{p}{9}T \ln K} < \frac{T}{2} - \frac{1}{6}\sqrt{T \ln K},$$

where the second inequality uses $p > 1/4$.
For the best action $i^\star$ we have $\mathrm{E}[C_{i^\star,T}] = \frac{T}{2} - \epsilon\delta T$.

- if $\delta = \frac{1}{6\epsilon}\sqrt{\frac{\ln K}{T}}$ we have

$$\mathrm{E}[C_{i^\star,T}] = \frac{T}{2} - \frac{1}{6}\sqrt{T \ln K}$$

  Using the fact that for binomial distribution, $B(n,q)$, the median is $\lfloor nq \rfloor$ or $\lceil nq \rceil$ we have that with probability at least $\frac{1}{2}$

$$C_{i^\star,T} \geq \frac{T}{2} - \frac{1}{6}\sqrt{T \ln K}$$

- if $\delta = \frac{1}{2}$ we have that the distribution for the ϵ-**noisy feedback** of the *best action*, $c_{i^\star,t}$ is $B(\frac{1-\epsilon}{2})$, therefore

$$\mathrm{E}[C_{i^\star,T}] = \frac{T}{2} - \frac{\epsilon}{2}T$$

$\delta = \frac{1}{2}$ implies $\epsilon \leq \frac{1}{3}\sqrt{\frac{\ln K}{T}}$ (as $\delta = \min\{\frac{1}{6\epsilon}\sqrt{\frac{\ln K}{T}}, \frac{1}{2}\}$) thus,

$$\frac{\epsilon}{2}T \leq \frac{T}{2}\frac{1}{3}\sqrt{\frac{\ln K}{T}} = \frac{1}{6}\sqrt{T \ln K}$$

  Therefore, we still have that with probability at least $\frac{1}{2}$

$$C_{i^\star,T} \geq \frac{T}{2} - \frac{1}{6}\sqrt{T \ln K}$$

Putting things together we obtain that with probability at least $\frac{1}{4}$ we have

$$C_{i^\star,T} > C_{j,T}$$

$\blacksquare$

The following lemma states that the action that has smaller observed noisy-loss has a higher probability to be the *best action*.

Lemma 8 *Let $C_{1,T}, \ldots, C_{K,T}$ be a realization of the noisy-feedbacks, such that $C_{j_1,T} < C_{j_2,T}$, where $j_1, j_2 \in A$. Then,*

$$\Pr[i^\star = j_1 \mid C_{1,T}, \ldots, C_{K,T}] > \Pr[i^\star = j_2 \mid C_{1,T}, \ldots, C_{K,T}]$$

Proof Using Bayes' theorem we have for action $j \in A$ that

$$
\begin{aligned}
\Pr[i^\star = j \mid C_{1,T}, \ldots, C_{K,T}] &= \frac{\Pr[C_{1,T}, \ldots, C_{K,T} \mid i^\star = j] \Pr[i^\star = j]}{\Pr[C_{1,T}, \ldots, C_{K,T}]} \\
&= \frac{\Pr[C_{j,T} \mid i^\star = j](\frac{1}{2}^T)^{K-1}\frac{1}{K}}{\Pr[C_{1,T}, \ldots, C_{K,T}]} = \frac{\Pr[C_{j,T} \mid i^\star = j]}{Z} \\
&= \frac{1}{Z}\left(\frac{1-\epsilon}{2}\right)^{C_{j,T}}\left(\frac{1+\epsilon}{2}\right)^{T-C_{j,T}}
\end{aligned}
$$

where $Z = \frac{(\frac{1}{2}^T)^{K-1}\frac{1}{K}}{\Pr[C_{1,T},\ldots,C_{K,T}]}$ is a constant (not depend on j). Therefore, if $C_{j_1,T} < C_{j_2,T}$ then

$$\Pr[i^\star = j_1 \mid C_{1,T}, \ldots, C_{K,T}] > \Pr[i^\star = j_2 \mid C_{1,T}, \ldots, C_{K,T}]$$

∎

Using the lemma we can show the following corollary.

Corollary 9 *consider an algorithm for "predicting the best action" problem: that is, the algorithm input is a realization $C_{1,T}, \ldots, C_{K,T}$, i.e., for one action $i^\star$ we have $C_{i^\star,T} \sim B(T, \frac{1}{2} - \epsilon\delta)$ and for $j \neq i^\star$ we have $C_{j,T} \sim B(T, \frac{1}{2})$ and the output is an action I_T - a prediction for which action is optimal. Then for any algorithm we have,*

$$\Pr[I_T \neq i^\star] \geq \frac{1}{4}$$

where the probability is taken over the randomness of the algorithm, the losses, the noise and the draw of $i^\star$.

Proof Lemma 8 implies that the optimal algorithm will predict

$$I_T = \arg\min_{j \in A}\{C_{1,T}, \ldots, C_{K,T}\}$$

From corollary 7 we have that for the optimal algorithm

$$\Pr[I_T \neq i^\star] \geq \frac{1}{4}$$

∎

Putting it all together we can now prove the theorem.

Proof of Theorem 5: For any round t we would think of the algorithm as algorithm for "predicting the best action" problem. Using this we can think of t as the the time horizon and by applying corollary 9 conclude that for every t we have

$$\Pr[I_t \neq i^\star] \geq \frac{1}{4}$$

Therefore the expectation of the regret, when the expectation is taken over the losses, the noise and the draw of $i^\star$ (note that the regret itself includes the randomness of the algorithm) satisfies,

$$\mathrm{E}[Regret(T)] = \sum_{t=1}^{T} \Pr[I_t \neq i^\star]\delta \geq \frac{1}{4}T\delta,$$

where $\delta = \min\{\frac{1}{6\epsilon}\sqrt{\frac{\ln K}{T}}, \frac{1}{2}\}$ concludes the proof. ∎

2. Variable Noise

In this section we investigate the *Full Information with Variable Noise* settings. Recall that in this setting we have a distribution D over $[0,1]^K$. At the beginning of each round t, we draw from D a realized noise $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $\epsilon_{i,t}$ is the noise parameter for action i at round t. We assume that the noise vectors are drawn independently from D at each round t (however, there can be correlations between the noise parameters $\epsilon_{i,t}$ of different actions at the same round t). The learner observes the realized noise $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$ and then picks an *action $I_t \in A$*. Then, the learner observes the $\epsilon_{i,t}$-*noisy feedback* $c_{i,t}$ for each action i. We denote by $p_{i,t} = \frac{1-\epsilon_{i,t}}{2}$.

The section is structured as follows. Initially, we investigate the case of a uniform distribution over $[0,1]$, that is, the marginal distribution of D for each action i is uniform over $[0,1]$, i.e., $\epsilon_{i,t} \sim U(0,1)$, where $U(0,1)$ is the uniform distribution on $[0,1]$. Following that, we generalize the regret bound for a general noise distribution D. We conclude with a few examples of specific distributions.

2.1 Uniform Noise Distribution

ALGORITHM

A simple potential approach to the problem is to try to use the *Exponential Weights Scheme* with the unbiased estimator

$$EST(i, c_{i,t}, \vec{q}_t, I_t) = \frac{c_{i,t} - p_{i,t}}{1 - 2p_{i,t}}$$

as in the constant noise settings. A close examination reveals that there is a problem when $p_{i,t}$ is close to $1/2$ (i.e., $\epsilon_{i,t}$ is close to 0). In such cases the estimator is unbounded

and can give a very high value. An intuitive idea is to avoid using feedbacks with high noise. This is implemented by the learner by having an additional parameter θ and ignoring feedbacks where $p_{i,t} > \frac{1-\theta}{2}$ (i.e., $\epsilon_{i,t} < \theta$). More formally, we use the *Exponential Weights Scheme* with the following estimator:

$$EST(i, c_{i,t}, \vec{q}_t, I_t) = \frac{c_{t,i} - p_{i,t}}{1 - 2p_{i,t}} \mathbb{1}_{\{p_{i,t} \leq \frac{1-\theta}{2}\}} = \hat{\ell}_{i,t}$$

The algorithm resulting from using the above estimator in the *Exponential Weights Scheme* is called **EW-Threshold**. We prove the following regret bound for **EW-Threshold**.

Theorem 10 *Let D be the noise distribution, such that for each action i the marginal distribution $\epsilon_{i,t}$ is distributed $U(0,1)$ (but not necessarily independent for different actions). The **EW-Threshold** algorithm with the parameters*

$$\eta = (\frac{\ln K}{T})^{2/3} \text{ and } \theta = (\frac{\ln K}{T})^{1/3}$$

has, in the Full Information with Variable Noise setting, a regret of at most,

$$Regret(T) \leq 3T^{2/3}(\ln K)^{1/3}$$

Proof By applying Lemma 2 and taking expectation on both sides we obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2] \tag{3.3}$$

Conditioning on $p_{i,t} \leq \frac{1-\theta}{2}$, the estimator $\hat{\ell}_{i,t}$ is biased, however, we can bound the deviation. Specifically,

$$\mathrm{E}[\hat{\ell}_{i,t}] = \theta * 0 + (1-\theta)\mathrm{E}[\hat{\ell}_{i,t} \mid p_{i,t} \leq \frac{1-\theta}{2}] = (1-\theta)\ell_{i,t}$$

This implies that

$$\ell_{i,t} - \theta \leq \mathrm{E}[\hat{\ell}_{i,t}] \leq \ell_{i,t}$$

To bound the second moment we have

$$\mathrm{E}[(\hat{\ell}_{i,t})^2] = \theta * 0 + (1-\theta)\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] \leq \mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}]$$

We bound the conditional expectation above as follows,

$$\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] = p_{i,t}\frac{(\bar{\ell}_{i,t} - p_{i,t})^2}{(1 - 2p_{i,t})^2} + (1-p_{i,t})\frac{(\ell_{i,t} - p_{i,t})^2}{(1 - 2p_{i,t})^2}$$

$$\leq \frac{1}{(1 - 2p_{i,t})^2} = \frac{1}{\epsilon_{i,t}^2}$$

Computing the expectation, given that the marginal is $U(0,1)$, we have,

$$\mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] \leq \mathrm{E}_{\epsilon \sim U(0,1)} [\frac{1}{\epsilon^2} \mathbb{1}_{\epsilon \geq \theta}]$$

$$= \int_\theta^1 \frac{1}{\epsilon^2} d\epsilon = -[\frac{1}{\epsilon}]_\theta^1 = \frac{1}{\theta} - 1 \leq \frac{1}{\theta}$$

Bounding the expressions in inequality (3.3) we obtain

$$\sum_{t=1}^T \sum_{i=1}^K q_{i,t} \mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^T \mathrm{E}[\hat{\ell}_{k,t}] \geq \sum_{t=1}^T \sum_{i=1}^K q_{i,t} \ell_{i,t} - \sum_{t=1}^T \ell_{k,t} - \theta T$$

$$\frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K q_{i,t} \mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K q_{i,t} \frac{1}{\theta} = \frac{\ln K}{\eta} + \frac{\eta T}{\theta}$$

Rearranging the terms gives us,

$$\sum_{t=1}^T \sum_{i=1}^K q_{i,t} \ell_{i,t} - \sum_{t=1}^T \ell_{k,t} \leq \frac{\ln K}{\eta} + \frac{\eta T}{\theta} + \theta T$$

Substituting $\eta = (\frac{\ln K}{T})^{2/3}$ and $\theta = (\frac{\ln K}{T})^{1/3}$ concludes the proof. ∎

IMPOSSIBILITY RESULT

In this section we derive a lower bound on the regret of $\Omega(T^{2/3}(\ln K)^{1/3})$. Together with the upper bound we obtain that for the *Full Information with Variable Noise* we have

$$Regret(T) = \Theta(T^{2/3}(\ln K)^{1/3})$$

For the lower bound we use a specific noise distribution D, denoted by D'. In D', all the noise of individual actions are identical, and uniformly distributed. Formally, we generate the noise parameters from D' as follows. We draw $\epsilon_t \sim U(0,1)$ and for every i we set $\epsilon_{i,t} = \epsilon_t$.

The idea behind the proof is to use adversarial strategy for loss assignment in the following way: when the noise is *low*, all the actions will have the same loss, but when the noise is *high*, one action, chosen randomly at the beginning, will be superior.

Theorem 11 *Any algorithm in the Full Information with Variable Noise setting with the noise distribution D', there exist a series of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$Regret(T) = \Omega(T^{2/3}(\ln K)^{1/3})$$

Proof Let $\theta = (\frac{\ln K}{T})^{1/3}$. Initially, the adversary choose an action $i^\star$ uniformly at random, and it will be the best action. Then, for each round t after observing ϵ_t, the adversary assigns losses as follow:

1. If $\epsilon_t \geq \theta$ then $\ell_{i,t} = 0$ for every action i.

2. Otherwise $(\epsilon_t < \theta)$ the adversary draw a loss for each action as follows: for action $i^\star$ the loss is drawn from $B(\frac{1}{2} - \frac{1}{6})$ and for any other action $j \neq i^\star$ it is drawn from $B(\frac{1}{2})$.

Denote by T' the number of *bad rounds*. Since $\mathrm{E}[T'] = \theta T$ and the fact that for Binomial distribution, $B(n,p)$, the median is $\lfloor np \rfloor$ or $\lceil np \rceil$ we conclude that with probability at least $1/2$ we have $T' \geq \theta T$. Condition on this event we assume that $T' = \theta T$ (if $T' > \theta T$ we take the first θT rounds to be T') we reduce the *bad rounds* to the constant noise setting in the following way:

In the *bad rounds* we have $\epsilon_t \sim U(0, \theta)$. If we assume that in the *bad rounds* we have $\epsilon_t = \theta$, namely a constant noise, then we only reduced the noise in the model. We call the model with $\epsilon_t = \theta$ and $T = T'$ the *reduced model*. Therefore, a lower bound for the regret in the *reduced model* is also a lower bound for a model where $\epsilon_t \sim U(0, \theta)$. Our *reduced model* is the **Full Information with Constant Noise** model with $T = T'$ and $\epsilon = \theta$. Denote by $Regret(T', \theta)$ the regret in the **Full Information with Constant Noise** model with horizon T' and noise parameter θ. Now, we can apply Theorem 5 on the *reduced model* and obtain that

$$Regret(T', \theta) \geq \gamma \frac{1}{\theta} \sqrt{T' \ln K}$$

where $\gamma > 0$ is a constant. Setting $T' = \theta T = T^{2/3} (\ln K)^{1/3}$ we obtain that

$$Regret(\theta T, \theta) \geq \frac{1}{\theta} \sqrt{\theta T \ln K} = \gamma T^{2/3} (\ln K)^{1/3}$$

Putting it back in the original model yields,

$$Regret(T) \geq \Pr[T' \geq \theta T] Regret(\theta T, \theta) \geq \frac{\gamma}{2} T^{2/3} (\ln K)^{1/3}$$

(We note that the number $\frac{1}{6}$ in the distribution $B(\frac{1}{2} - \frac{1}{6})$ the adversary uses, comes from the $\delta = \frac{1}{6\epsilon} \sqrt{\frac{\ln K}{T}}$ we use in the proof of theorem 5 with $\epsilon = \theta$ and $T = T'$). ■

2.2 General distributions

In this section we generalized the result of **EW-Threshold** to a general noise distribution D. We assume that the marginal distribution of each action i is the same and we denoting the CDF (Cumulative Distribution Function) of it by F. Then, we use our generalized bound to derive a sub-linear regret upper bound for distributions D that satisfies a given condition. We extend the proof of Theorem 5 and obtain the following general upper bound.

Theorem 12 *Let D be a distribution, such that the marginal distribution over $\epsilon_{i,t} \in (0,1)$ has a CDF F. Then, running* **EW-Threshold** *algorithm with parameters $\theta > 0$ and $\eta > 0$ satisfies,*

$$Regret(T) \leq \frac{\ln K}{\eta} + \eta T g(\theta) + F(\theta)T$$

where $g(\theta) = \mathrm{E}[\frac{1}{\epsilon^2}\mathbb{1}_{\epsilon \geq \theta}]$. Moreover, for $\eta = \sqrt{\frac{\ln K}{T g(\theta)}}$ we have

$$Regret(T) \leq 2\sqrt{g(\theta)T \ln K} + F(\theta)T$$

Proof We apply Lemma 2, and taking expectation on both sides, obtain,

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2] \qquad (3.4)$$

Conditioning on $p_{i,t} \leq \frac{1-\theta}{2}$, the estimator $\hat{\ell}_{i,t}$ is biased, and we have

$$\mathrm{E}[\hat{\ell}_{i,t}] = F(\theta) * 0 + (1 - F(\theta))\mathrm{E}[\hat{\ell}_{i,t} \mid p_{i,t} \leq \frac{1-\theta}{2}] = (1 - F(\theta))\ell_{i,t}$$

This implies that

$$\ell_{i,t} - F(\theta) \leq \mathrm{E}[\hat{\ell}_{i,t}] \leq \ell_{i,t}$$

To bound the second moment we have

$$\mathrm{E}[(\hat{\ell}_{i,t})^2] = F(\theta) * 0 + (1 - F(\theta))\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] \leq \mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}]$$

We bound the above conditional expectation as follows,

$$\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] = p_{i,t}\frac{(\bar{\ell}_{i,t} - p_{i,t})^2}{(1 - 2p_{i,t})^2} + (1 - p_{i,t})\frac{(\ell_{i,t} - p_{i,t})^2}{(1 - 2p_{i,t})^2}$$
$$\leq \frac{1}{(1 - 2p_{i,t})^2} = \frac{1}{\epsilon_{i,t}^2}$$

Bounding each side of inequality (3.4) we have

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \geq \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \sum_{t=1}^{T}\ell_{k,t} - F(\theta)T$$

$$\frac{\ln K}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\frac{1}{\epsilon^2}\mathbb{1}_{\epsilon \geq \theta}]$$

Rearranging it all yield

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \sum_{t=1}^{T}\ell_{k,t} \le \frac{\ln K}{\eta} + \eta T g(\theta) + F(\theta)T$$

∎

The following corollary gives a general upper bound that depends only on a property of the noise distribution D. We assume that all the marginal distributions of D are identical and with CDF F.

Corollary 13 *Let D be a noise distribution, where each marginal distribution has the same CDF F, and assume $F(\theta) \le \theta^\alpha$ for a given $\alpha > 0$. Then*

$$Regret(T) = O(T^{\frac{2+\alpha}{2+2\alpha}}(lnK)^{\frac{\alpha}{2(1+\alpha)}})$$

Proof Using Theorem 12 and the assumption we can write

$$Regret(T) \le 2\sqrt{g(\theta)T\ln K} + \theta^\alpha T$$

since $g(\theta) = \mathrm{E}[\frac{1}{\epsilon^2}\mathbb{1}_{\epsilon \ge \theta}] \le \frac{1}{\theta^2}$, we have,

$$Regret(T) \le \frac{2}{\theta}\sqrt{T\ln K} + \theta^\alpha T$$

taking $\theta = \left(\frac{2}{\alpha}\right)^{\frac{1}{1+\alpha}}\left(\frac{\ln K}{T}\right)^{\frac{1}{2(1+\alpha)}}$ gives

$$Regret = O(T^{\frac{2+\alpha}{2+2\alpha}}(\ln K)^{\frac{\alpha}{2(1+\alpha)}})$$

∎

To get an intuition for the bound of Corollary 13 we can consider a few intuitive settings of the parameter α. The uniform distribution has $\alpha = 1$, and the theorem yield $Regret(T) = \tilde{O}(T^{\frac{3}{4}})$, which is higher than the regret bound computed explicitly in Theorem 10, of $O(T^{2/3})$. When $\alpha \to \infty$ we have $Regret(T) \to \tilde{O}(T^{\frac{1}{2}})$, which is tight even without any noise. When $\alpha \to 0$, we have no restriction on the noise distribution, and indeed the theorem yields a linear regret bound.

To give an example for the bound of Theorem 12 we prove a regret bound for truncated exponential distribution.

Corollary 14 *Let D be a distribution, such that the marginal distribution over $\epsilon_{i,t} \in (0,1)$ has a PDF $f(x) = \frac{\lambda}{1-e^{-\lambda}}e^{-\lambda x}$ for $x \in (0,1)$ and $\lambda > 0$. Then, running **EW-Threshold** algorithm with parameters $\theta > 0$ and $\eta > 0$ satisfies,*

$$Regret(T) \le 3\lambda T^{2/3}(\ln K)^{1/3}$$

Proof Applying Theorem 12 gives

$$Regret(T) \leq 2\sqrt{g(\theta)T \ln K} + F(\theta)T \qquad\qquad (3.5)$$

To bound $g(\theta)$ we calculate

$$g(\theta) = \mathrm{E}[\frac{1}{\epsilon^2}\mathbb{1}_{\epsilon \geq \theta}] \int_\theta^1 \frac{1}{\epsilon^2} \frac{\lambda}{1 - e^{-\lambda}} e^{-\lambda\epsilon} d\epsilon \leq \frac{\lambda}{1 - e^{-\lambda}} \int_\theta^1 \frac{1}{\epsilon^2} d\epsilon$$

$$= \frac{\lambda}{1 - e^{-\lambda}}(\frac{1}{\theta} - 1) \leq \frac{\lambda}{1 - e^{-\lambda}}\frac{1}{\theta} \leq \frac{\lambda}{\theta}$$

Bounding the second term we use the inequality $1 - e^{-x} \leq x$ for $x > 0$ and obtain

$$F(\theta) = \frac{\lambda}{1 - e^{-\lambda}}(1 - e^{-\lambda\theta}) \leq \lambda^2\theta$$

Putting it back in (3.5) we have

$$Regret(T) \leq 2\sqrt{\frac{1}{\theta} \ln K} + \lambda^2\theta T$$

setting $\theta = \frac{1}{\lambda}(\frac{\ln K}{T})^{1/3}$ yields,

$$Regret(T) \leq 3\lambda T^{2/3}(\ln K)^{1/3}$$

∎

2.3 Importance of knowing the Noise

Until now we assume for the *Full Information with Variable Noise* that the learner observes the noise drawn for each action $p_{i,t}$, before picking an action. In the *Full Information with Constant Noise* we showed that this information is not critical and the same regret bound can be achieved without this information. The following theorem states that in the *Full Information with Variable Noise*, a learner cannot achieve sub-linear regret without observing the noise.

Theorem 15 *Fix an algorithm for the **Full Information with Variable Noise** model under the uniform marginal distribution and assume that in each round t the learner does not observes the noise parameters $(\epsilon_{1,t}, ..., \epsilon_{K,t})$. Then, there exist a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$Regret(T) = \Omega(T)$$

We prove the theorem for the case of $K = 2$. The idea behind the proof is to use stochastic adversarial strategy for loss assignment such that one action is significantly better than the other, but after applying noise on both, they look identical to the learner.

Proof Let the number of actions be $K = 2$. Assume that initially the adversary picks the best action uniformly (that is, with probability $\frac{1}{2}$ action 1 will be the best action and with probability $\frac{1}{2}$ action 2 will be the best action). Let $i^\star \in \{1, 2\}$ be a random variable denoting the best action and $j = 3 - i^\star$ denote the worse action. On round t, after observing the noise parameters $p_{1,t}$ and $p_{2,t}$, the adversary selects the losses as follow:

1. For the best action, $i^\star$, the loss is drawn at every round independently from a Bernoulli r.v. with parameter $1/4$, i.e., $\ell_{i^\star,t} \sim B(\frac{1}{4})$

2. For the worse action j: if $p_{j,t} < 1/4$ then the loss is $\ell_{j,t} = 0$, otherwise the loss is $\ell_{j,t} = 1$.

For the learner, observing the feedback $c_{i,t} = \ell_{i,t} \oplus r_{i,t}$, the loss of each action is a Bernoulli random variable. We will show that both actions will have the same probability of 1, namely $3/8$, and therefore indistinguishable by the learner.

Now we calculate the expected value of the observed feedback, $c_{i,t} = \ell_{i,t} \oplus r_{i,t}$, for each action in a single round. We note that this expectation is taken over the draw of $\epsilon_{i,t} \sim U(0,1)$, the draw $R_{i,t} \sim B(\frac{1-\epsilon_{i,t}}{2})$ and the draw of the losses $\ell_{i,t}$. We also note that if $\epsilon \sim U(0,1)$ then $p \sim U(0, \frac{1}{2})$.

The expected loss of best action, $\ell_{i^\star,t}$ is drawn independently from the noise parameter $\epsilon_{i^\star,t}$ and the Bernoulli noise $R_{i,t}$. Therefore, we have

$$\mathrm{E}[c_{i^\star,t}] = \mathrm{E}_p[\mathrm{E}_R[\mathrm{E}_\ell[\ell_{i^\star,t} \oplus R_{i^\star,t}] \mid p]] = \mathrm{E}_p[\mathrm{E}_R[\frac{1}{4}(1 \oplus R_{i^\star,t}) + \frac{3}{4}(0 \oplus R_{i^\star,t}) \mid p]]$$

$$= \frac{1}{4}\mathrm{E}_p[p_{i^\star,t} \cdot 0 + (1 - p_{i^\star,t}) \cdot 1] + \frac{3}{4}\mathrm{E}_p[p_{i^\star,t} \cdot 1 + (1 - p_{i^\star,t}) \cdot 0]$$

$$= \frac{1}{4} \cdot \frac{3}{4} + \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{8}$$

For the worse action, action j, we have

$$\mathrm{E}[c_{j,t}] = \mathrm{E}[\ell_{j,t} \oplus R_{j,t}] = \frac{1}{2}\mathrm{E}[0 \oplus R_{j,t} \mid p_{j,t} < 1/4] + \frac{1}{2}\mathrm{E}[1 \oplus R_{j,t} \mid \frac{1}{4} \le p_{j,t} < \frac{1}{2}]$$

$$= \frac{1}{2}\mathrm{E}[p_{j,t} \mid p_{j,t} < \frac{1}{4}] + \frac{1}{2}\mathrm{E}[1 - p_{j,t} \mid \frac{1}{4} \le p_{j,t} < \frac{1}{2}]$$

$$= \frac{1}{2} \cdot \frac{1}{8} + \frac{1}{2}(1 - \frac{3}{8}) = \frac{3}{8}$$

This implies that the feedback of both the best and worse action is a Bernoulli random variable with parameter $\frac{3}{8}$, i.e., $B(\frac{3}{8})$. This clearly implies that the learner cannot

distinguish between the two actions, and therefore, half the time it will select the worse action. The best action has an expected loss of $\frac{T}{4}$ while the worse action has a loss of $\frac{T}{2}$. This implies that the expected regret would be at least $\frac{T}{8}$.   ■

Chapter 4

Bandit Models

In this chapter we study bandit models, where the learner observes only the noisy feedback for the action selected. In our notation, the learner selects $I_t \sim q_t$ and observers only the feedback $c_{I_t,t}$.

1. Bandit with Constant Noise Model

1.1 Algorithm

Using our conclusion from the *Full Information with Constant Noise* setting, we present algorithm that do not use the noise parameter ϵ. Clearly, this establish upper bound for both settings: the *known noise* setting and the *unknown noise* setting.

Theorem 16 *Let $\epsilon \in [0,1]$ and denote $p = \frac{1-\epsilon}{2}$. Then, running Exponential Weights Scheme under the Bandit with Constant Noise setting with the following loss estimate*

$$EST(i, c_{i,t}, \vec{q}_t, I_t) = \frac{1}{q_{i,t}} c_{i,t} = \hat{\ell}_{i,t}$$

and $\eta = \epsilon \sqrt{\frac{\ln K}{TK}}$, guarantees

$$Regret(T) \leq \frac{2}{\epsilon} \sqrt{TK \ln K}$$

The proof of Theorem 16 is similar in spirit to the proof of Theorem 4. For completeness, we gives full proof.

Proof By applying Lemma 2 and taking expectation on both sides we obtain

$$\sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} \mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T} \mathrm{E}[\hat{\ell}_{k,t}] \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{K} q_{i,t} \mathrm{E}[(\hat{\ell}_{i,t})^2]$$

Calculating the expectation of the estimator $\hat{\ell}_{i,t}$, and since $\ell_{i,t} \in \{0, 1\}$, we have,

$$\mathrm{E}[\hat{\ell}_{i,t}] = q_{i,t} \frac{1}{q_{i,t}} \mathrm{E}[c_{i,t}] = \mathrm{E}[c_{i,t}] = (1-p)\ell_{i,t} + p\bar{\ell}_{i,t} = (1-2p)\ell_{i,t} + p = |\ell_{i,t} - p|$$

For the second moment, since $c_{i,t} \leq 1$ we have

$$E[(\ell_{i,t})^2] = q_{i,t}\frac{1}{q_{i,t}^2}E[c_{i,t}] \leq \frac{1}{q_{i,t}}$$

Putting things together we have

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}|\ell_{i,t} - p| - \sum_{t=1}^{T}|\ell_{k,t} - p| \leq \frac{\ln K}{\eta} + \eta TK \qquad (4.1)$$

Using the notation of $\hat{L}_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\hat{\ell}_{i,t}$, $\hat{L}_{k,T} = \sum_{t=1}^{T}\hat{\ell}_{k,t}$, $L_{ON,T} = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t}$, and $L_{k,T} = \sum_{t=1}^{T}\ell_{k,t}$, we can write inequality (4.1) as

$$E[\hat{L}_{ON,T}] - E[\hat{L}_{k,T}] \leq \frac{\ln K}{\eta} + \eta TK$$

Denote by $G_{t,b} = \{i \in A \mid \ell_{i,t} = b\}$ the set of actions with loss $b \in \{0,1\}$ in round t. Denote by $Q_t = \sum_{i \in G_{t,1}} q_{i,t}$ the distribution mass the learner gives actions in $G_{t,1}$. Using this notation we have $L_{ON,T} = \sum_{t=1}^{T} Q_t$. Now calculate the value of the estimated losses of the online algorithm,

$$E[\hat{L}_{ON,T}] = \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}|\ell_{i,t} - p| = \sum_{t=1}^{T}[p\sum_{i \in G_{t,0}} q_{i,t} + (1-p)\sum_{i \in G_{t,1}} q_{i,t}]$$

$$= \sum_{t=1}^{T}[p(1 - Q_t) + (1-p)Q_t] = \sum_{t=1}^{T}[p + (1 - 2p)Q_t]$$

$$= (1 - 2p)L_{ON,T} + pT$$

Similarly, for the term $E[\hat{L}_{k,T}]$ we have,

$$E[\hat{L}_{k,T}] = \sum_{t=1}^{T}|\ell_{k,t} - p| = \sum_{t|\ell_{t,k}=0} p + \sum_{t|\ell_{t,k}=1}(1-p)$$

$$= p(T - L_{k,T}) + (1-p)L_{k,T} = (1 - 2p)L_{k,T} + pT$$

Putting all together,

$$E[\hat{L}_{ON,T}] - E[\hat{L}_{k,T}] = (1-2p)L_{ON,T} + pT - [(1-2p)L_{k,T} + pT] = (1-2p)[L_{ON,T} - L_{k,T}]$$

Dividing by both sides of inequity by $(1 - 2p)$ and using $\eta = \sqrt{\frac{\ln K}{TK}}$ we obtain that

$$Regret(T) = L_{ON,T} - \min_{k \in A} L_{k,T} \leq \frac{1}{1 - 2p}(\frac{\ln K}{\eta} + \eta TK) = \frac{2}{\epsilon}\sqrt{TK \ln K}$$

$\blacksquare$

1.2 Impossibility result

In this section we present a lower bound that matches our upper bound, up to a constant factor.

Theorem 17 *Consider the Bandit with Constant Noise setting with noise parameter $\epsilon \in (0,1)$. Then, for any learner algorithm there exists a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$Regret(T) = \Omega(\min\{\frac{1}{\epsilon} \sqrt{TK}, T\})$$

The proof of the above theorem follows the methodology for lower bounds for *multi-arm bandit* problems, we follow here the methodology proposed in Slivkins (2017) and adapt it to our special setting.

The idea is to define a stochastic strategy for loss assignments such that the *learner* will have high expected regret, which implies that there exists a realization of a loss sequence such that on this loss sequence the *learner* has high regret.

We first define K different problem instances, one per action. Let $\beta \in (0,1)$ be a parameter. We denote by J_i the problem instance where action i loss is drawn from the distribution $B(\frac{1-\beta}{2})$ while the other actions loss is drawn from the distribution $B(\frac{1}{2})$. For problem instance J_i, we refer action i as the *best action*. The proof will show that in some sense those instances are indistinguishable for any algorithm.

For the proof, we will think of the online algorithm as a leaner making "prediction" for the best action at each round t. The main part of the proof is to show that if T is not large enough the algorithm has to have a constant mistake rate.

We denote by $\Pr[I_t = i | J_i]$ the probability that in instance J_i, at round t the algorithm selects action i (the best action in instance J_i). The following lemma shows that for many actions the algorithm will make a mistake.

Lemma 18 *Consider a deterministic algorithm for the Bandit with Constant Noise problem with noise $p = \frac{1-\epsilon}{2}$. There exist a constant γ such that if $t < \gamma \frac{K}{\epsilon^2 \beta^2}$ then there exist at least $\lceil \frac{K}{2} \rceil$ actions i such that*

$$\Pr[I_t = i | J_i] < \frac{3}{4}$$

Proof Consider the feedback distribution for each problem instance J_j and action i. First, if $\ell_{i,t} \sim B(\frac{1}{2})$ then $c_{i,t} \sim B(\frac{1}{2})$ (the noise does not have any influence). For the *best action*, i.e., j, we have $c_{j,t} \sim B(\frac{1-\epsilon\beta}{2})$ since

$$Pr[c_{j,t} = 1] = Pr[\ell_{j,t} = 1]Pr[R_\epsilon = 0] + Pr[\ell_{j,t} = 0]Pr[R_\epsilon = 1]$$

$$= (\frac{1-\beta}{2})(\frac{1+\epsilon}{2}) + (\frac{1+\beta}{2})(\frac{1-\epsilon}{2}) = \frac{1-\epsilon\beta}{2}$$

Applying *Lemma 2.10* of Slivkins (2017) on the feedbacks $c_{i,t}$ completes the proof. ∎

Corollary 19 *Choose the best action $i^\star$ uniformly from A and use instance $J_{i^\star}$. For any algorithm, for any round $t < \gamma \frac{K}{\epsilon^2 \beta^2}$, we have $\Pr[I_t \neq i^\star] \geq 1/8$.*

Proof For a deterministic algorithm the corollary follows since by Lemma 18 with probability at least $\frac{1}{2}$ the selected $i^\star$ is such that $\Pr[I_t \neq i^\star | J_{i^\star}] \geq \frac{1}{4}$. Since a randomized algorithm is a distribution over deterministic algorithms that claim hold also for randomized algorithms. ∎

Proof of Theorem 17: Let $\beta = \min\{\frac{\sqrt{\gamma}}{\epsilon}\sqrt{\frac{K}{T}}, 1\}$. By Corollary 19, we have that in each round t

$$\Pr[I_t \neq i^\star] \geq \frac{1}{8}$$

Denote by $\Delta_t = \mathrm{E}[\ell_{I_t,t}] - \mathrm{E}[\ell_{i^\star,t}]$ the regret of round t. Note that if $I_t \neq i^\star$ then $\Delta_t = \frac{1}{2} - \frac{1-\beta}{2} = \frac{\beta}{2}$. Therefore, the expected regret at round t is

$$\mathrm{E}[\Delta_t] = \Pr[I_t \neq i^\star]\frac{\beta}{2}$$

Summing over the rounds we have,

$$Regret(T) = \sum_{t=1}^{T} \mathrm{E}[\Delta_t] \geq \frac{1}{16}\beta T$$

Since $\beta = \min\{\frac{\sqrt{\gamma}}{\epsilon}\sqrt{\frac{K}{T}}, 1\}$, we have

$$Regret(T) \geq \min\{\frac{\sqrt{\gamma}}{16}\frac{1}{\epsilon}\sqrt{TK}, \frac{1}{16}T\}$$

∎

2. Bandit with Variable Noise Model

In this section we investigate the *Bandit with Variable Noise* settings. We concentrate on the case where the marginal distribution of D for each action i is the uniform distribution on $[0, 1]$.

2.1 Algorithm

We use the same idea as in the *Full Information settings* and ignore "too-noisy" rounds where the noise is close to $\frac{1}{2}$. More formally, we will run the *Exponential Weights Scheme* with the following estimator:

$$EST(i, c_{i,t}, \vec{q_t}, I_t) = \frac{1}{q_{i,t}}\frac{c_{i,t} - p}{1 - 2p}\mathbb{1}_{\{p_{i,t} \leq \frac{1-\theta}{2}\}}\mathbb{1}_{\{I_t=i\}} = \hat{\ell}_{i,t} ,$$

where θ is a parameter. We call the algorithm resulting from using the above estimator in the *Exponential Weights Scheme* as the **Exp3-Threshold**. The following theorem bounds the regret of the algorithm.

Theorem 20 *Let D be the noise distribution, such that for each action i the marginal distribution $\epsilon_{i,t}$ is distributed $U(0,1)$ (but not necessarily independent for different actions). The* **Exp3-Threshold** *algorithm with the parameters*

$$\eta = \frac{(\ln K)^{2/3}}{K^{1/3}T^{2/3}} \ and \ \theta = \frac{K^{1/3}(\ln K)^{1/3}}{T^{1/3}}$$

has, in the Bandit with Variable Noise, regret of at most

$$Regret(T) \leq 3T^{2/3}K^{1/3}(lnK)^{1/3}$$

Proof By applying Lemma 2 and taking expectation on both sides we obtain

$$\sum_{t=1}^{T}\sum_{i=1}^{K}q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K}q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2] \tag{4.2}$$

Conditioning on $p_{i,t} \leq \frac{1-\theta}{2}$, the estimator $\hat{\ell}_{i,t}$ is unbiased, since

$$\mathrm{E}[\hat{\ell}_{i,t} \mid p_t \leq \frac{1-\theta}{2}] = q_{i,t}[\frac{1}{q_{i,t}}\frac{p\bar{\ell}_{i,t} + (1-p)\ell_{i,t} - p}{1 - 2p}] = \ell_{i,t} .$$

However, overall the estimator is biased,

$$\mathrm{E}[\hat{\ell}_{i,t}] = \theta * 0 + (1-\theta)\mathrm{E}[\hat{\ell}_{i,t} \mid p_{i,t} \leq \frac{1-\theta}{2}] = (1-\theta)\ell_{i,t}$$

This implies that

$$\ell_{i,t} - \theta \leq \mathrm{E}[\hat{\ell}_{i,t}] \leq \ell_{i,t}$$

To bound the second moment we have

$$\mathrm{E}[(\hat{\ell}_{i,t})^2] = \theta * 0 + (1-\theta)\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] \leq \mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}]$$

The conditional expectation of the second moment is bounded as follows,

$$\mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_t \leq \frac{1-\delta}{2}] = \frac{1}{q_{i,t}}[p_t\frac{(\bar{\ell}_{i,t} - p_t)^2}{(1 - 2p_t)^2} + (1-p_t)\frac{(\ell_{i,t} - p_t)^2}{(1 - 2p_t)^2}] \leq \frac{1}{q_{i,t}}\frac{1}{(1 - 2p_t)^2} = \frac{1}{q_{i,t}}\frac{1}{\epsilon_t^2}$$

Since the marginal of the noise distribution D is uniform, we have,

$$\mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \mathrm{E}[(\hat{\ell}_{i,t})^2 \mid p_{i,t} \leq \frac{1-\theta}{2}] \leq \mathrm{E}_{\epsilon \sim U(0,1)}[\frac{1}{q_{i,t}}\frac{1}{\epsilon^2}\mathbb{1}_{\epsilon \geq \theta}]$$

$$= \frac{1}{q_{i,t}}\int_{\theta}^{1}\frac{1}{\epsilon^2}d\epsilon = -\frac{1}{q_{i,t}}[\frac{1}{\epsilon}]_{\theta}^{1} = \frac{1}{q_{i,t}}(\frac{1}{\theta} - 1) \leq \frac{1}{q_{i,t}}\frac{1}{\theta} \tag{4.3}$$

Bounding each side of inequality (4.2) we have

$$
\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[\hat{\ell}_{i,t}] - \sum_{t=1}^{T}\mathrm{E}[\hat{\ell}_{k,t}] \geq \sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \sum_{t=1}^{T}\ell_{k,t} - \theta T
$$

$$
\frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}[\frac{1}{q_{i,t}}\frac{1}{\theta}] = \frac{\ln K}{\eta} + \frac{\eta TK}{\theta}
$$

(4.4)

Rearranging it all yield

$$
\sum_{t=1}^{T}\sum_{i=1}^{K} q_{i,t}\ell_{i,t} - \sum_{t=1}^{T}\ell_{k,t} \leq \frac{\ln K}{\eta} + \frac{\eta TK}{\theta} + \theta T
$$

Substituting $\eta = \frac{(\ln K)^{2/3}}{K^{1/3}T^{2/3}}$ and $\theta = \frac{K^{1/3}(\ln K)^{1/3}}{T^{1/3}}$ concludes the proof. ∎

2.2 Impossibility result

We show a lower bound of $\Omega((TK)^{2/3})$. The proof is similar to the proof of Theorem 10 for the *Full Information* settings.

Theorem 21 *For any algorithm in the Bandit with Variable Noise setting with $D = U(0,1)$ as the noise parameters distribution, there exist a series of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$
Regret(T) = \Omega(T^{2/3}K^{1/3})
$$

Proof Let $\theta = (\frac{K}{T})^{1/3}$. Initially, the adversary choose an action $i^\star$ uniformly at random, and it will be the best action. Then, for each round t after observing ϵ_t, the adversary assigns losses as follow: fix $\beta = \frac{\sqrt{\gamma}}{\theta}\sqrt{\frac{K}{T}} = \sqrt{\gamma}(\frac{K}{T})^{1/6}$ and at round t do

1. if $\epsilon_t \geq \theta$ then $\ell_{i,t} = 0$ for every action i.

2. Otherwise ($\epsilon_t < \theta$) the adversary draw a loss for each action as follows: for action $i^\star$ the loss is drawn from $B(\frac{1}{2} - \beta)$ and for any other action $j \neq i^\star$ it is drawn from $B(\frac{1}{2})$.

Denote by T' the number of *bad rounds*. Since $\mathrm{E}[T'] = \theta T$ and the fact that for Binomial distribution, $B(n,p)$, the median is $\lfloor np \rfloor$ or $\lceil np \rceil$ we conclude that with probability at least $1/2$ we have $T' \geq \theta T$. Condition on this event we assume that $T' = \theta T$ (if $T' > \theta T$ we take the first θT rounds to be T') we reduce the *bad rounds* to the constant noise setting in the following way:

In the *bad rounds* we have $\epsilon_t \sim U(0, \theta)$. If we assume that in the *bad rounds* we have $\epsilon_t = \theta$, namely a constant noise, then we only reduced the noise in the model. We call the model with $\epsilon_t = \theta$ and $T = T'$ the *reduced model*. Therefore, a lower bound for

the regret in the *reduced model* is also a lower bound for a model where $\epsilon_t \sim U(0, \theta)$. Our *reduced model* is the **Bandit with Constant Noise** model with $T = T'$ and $\epsilon = \theta$. Denote by $Regret(T', \theta)$ the regret in the **Bandit with Constant Noise** model with horizon T' and noise parameter θ. Now, we can apply Theorem 15 on the *reduced model* and obtain that

$$Regret(T', \theta) \geq \gamma \frac{1}{\theta} \sqrt{T'K}$$

where $\gamma > 0$ is a constant. Setting $T' = \theta T = T^{2/3} K^{1/3}$ we obtain that

$$Regret(\theta T, \theta) \geq \frac{1}{\theta} \sqrt{\theta T K} = \gamma T^{2/3} K^{1/3}$$

Putting it back in the original model yields,

$$Regret(T) \geq \Pr[T' \geq \theta T] * Regret(\theta T, \theta) \geq \frac{\gamma}{2} T^{2/3} K^{1/3}$$

(We note here that the choice of β is according to the proof of Theorem 15 with $\epsilon = \theta$ and $T = T' = \theta T$). ∎

2.3 Importance of knowing the Noise

In Theorem 15 we showed that in the *Full Information with Variable Noise* setting, a learner cannot guarantee a sub-linear regret bound without observing the noise drawn for each action $p_{i,t}$ at each round t. Since in the *Bandit with Variable Noise* setting the feedback is a restriction of the feedback in the *Full Information with Variable Noise* setting, the same lower bound still holds, as stated in the following corollary.

Corollary 22 *Fix an algorithm for the **Bandit with Variable Noise** model under the uniform marginal distribution and assume that in each round t the learner does not observes the noise parameters $(\epsilon_{1,t}, ..., \epsilon_{K,t})$ before picking an action I_t, then there exist a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

$$Regret(T) = \Omega(T)$$

Chapter 5

Discussion

In this paper we investigated adversarial online learning problems where the feedback is corrupted by random noise. We presented and study different noise systems that apply to the *full information* feedback and the *bandit* feedback. We provided efficient algorithms, as well as upper and lower bounds on the regret.

This work can be extended in many ways. In our settings we apply the noise system on the classic *full information* and *bandit*. Similar noise system can be applied on intermediate models such as the one proposed by Mannor and Shamir (2011); Alon et al. (2017). A different corrupting settings can be consider too. For example, a settings in which an adversary is corrupting the feedbacks under some restrictions.

Bibliography

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM J. Comput.*, 46(6):1785–1826, 2017.

Dana Angluin and Philip Laird. Learning from noisy examples. *Mach. Learn.*, 2(4): 343–370, April 1988.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and non-stochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *ALT 2018-Algorithmic Learning Theory*, 2018.

Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM J. Comput.*, 22(4):807–837, 1993.

Philip Klein and Neal Young. On the number of iterations for dantzig-wolfe optimization and packing-covering approximation algorithms. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 320–327. Springer, 1999.

Tomás Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *AISTATS*, pages 1186–1194, 2016.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692, 2011.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2017.

L. G. Valiant. Learning disjunction of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'85, pages 560–566, 1985.

T Weissman and N Merhav. Universal prediction of binary individual sequences in the presence of noise. accepted to ieee trans. *Inform. Theory, September*, 2000.

Tsachy Weissman, Neri Merhav, and Anelia Somekh-Baruch. Twofold universal prediction schemes for achieving the finite-state predictability of a noisy individual binary sequence. *IEEE Transactions on Information Theory*, 47(5):1849–1866, 2001.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.