Adversarial Online Learning with noise

Alon Resler [1] **Yishay Mansour** [1 2]

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.

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. In the *Bandit* feedback the learner only observes the loss associated with the action played.

Both models have been extensively studied and received significant practical and theoretical interest. The regret bound for the full information model is $\Theta(\sqrt{T \ln K})$ (see (Littlestone & Warmuth, 1994; Freund & Schapire, 1997; Kalai & Vempala, 2005)), and for the bandit model is $\Theta(\sqrt{KT})$ (see (Auer et al., 2002; Audibert & Bubeck, 2009; Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012)).

Both models assume that the observed feedback is exact, namely that we always get the correct ground-truth. In some real life scenarios, the feedback (ground-truth) might be corrupted by noise, which is the focus of our work.

For a motivating example, consider an Internet website that presents one out of K ads to each user, and its goal is to maximize the number of clicked ads. In real world scenarios, we might observe an incorrect feedback regarding the realization of the click. This can happen for many reasons: (1) network connection (from the user to the advertising server, which is different from the content server), (2) browser issues, including privacy setting, (3) misidentification of the user (due to multiple users using the same IP address, or blocking cookies). All those can be modeled as a Bernoulli noise.

Another important aspect for this web advertising example to consider are invalid clicks, which can be either malicious clicks (done by robots, etc) or unintentional clicks. We would like to maximize the "valid clicks", and ignore invalid ones. (In fact, web search companies developed tools to address such issues.) The invalid clicks, out of the entire click stream, can be viewed as noise. Some "users" are more prone to generate invalid clicks, which can be captured by variable Bernoulli noise across time steps.

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 parameter 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 paragraphs

[*]Equal contribution [1]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel [2]Google Research, Israel. Correspondence to: Alon Resler <alonress@gmail.com>, Yishay Mansour <mansour.yishay@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Figure 1. Results summary

Feedback \Noise	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	$\tilde{\Theta}(\frac{1}{\epsilon}\sqrt{TK})$	$\Theta(T)$

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 to the learner 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 parameters are observed by the learner, where we have a tight regret bound of $\Theta(T^{2/3}\ln^{1/3} K)$, and the case where the realized noise parameters are 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 parameters are observed, and linear regret, i.e., $\Theta(T)$, when the realized noise parameters are 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 & 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 & 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 & Laird, 1988) and malicious noise (Valiant, 1985; Kearns & 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.

Paper Organization: In the next section we formalize our model. In section 3 we study the *full information with constant noise* settings, providing algorithms and matching lower bounds. In section 4 we study *full information with variable noise* settings and derives algorithm, analyzes their regret, and proves a matching lower bound for specific noise distribution. In section we 5 study the *bandit feedback* settings both for the constant noise and variable noise model. We include in the paper proofs sketches for part of the bounds shown. Full proofs for all of the bounds are given in the supplementary material.

2. Model

We consider adversarial decision problem with finite *actions* set $A = \{1,2,\ldots,K\}$. On each round $t = 1,2,\ldots,T$ the *environment* (or *adversary*) 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. The *learner* (or *algorithm*) chooses an *action* I_t, without observing $\vec{\ell_t}$. Then, the learner 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 parameters $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $\epsilon_{i,t} \in [0,1]$ is the noise parameter of action i at round t. We assume that the noise parameters 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 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 parameters $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$, where $\epsilon_{i,t} \in [0,1]$ is the noise parameter of action i at round t. We assume that the noise parameters 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 round 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 does not observe the noise parameters. The *known noise parameters* settings can be divided farther into two variant: in the *informed* model, the learner observes the noise parameter at the beginning of round t, before drawing an action I_t. In the *uninformed* model, the player observes the noise parameter at the end of round t, after drawing I_t and observing the losses. In this paper we show upper bounds (algorithms) in the *uninformed* model and tight lower bounds in the *informed* model, concluding that the *informed* and *uninformed* models are equivalent with respect to asymptotic regret bounds.

Algorithm 1 Exponential Weights Scheme

Initialization: $w_{i,1} = 1$ for all $i \in A$
Parameters: $\eta > 0$
for $t = 1$ **to** T **do**
  Construct the probability distribution q_t with

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

  Play a random action I_t according to q_t
  Incur loss $\ell_{I_t,t}$
  Observe feedback according to the specific settings

$$observed\_feedback = Model\_feedback(\ell_t, I_t, \epsilon_t)$$

  Construct loss estimate

$$\hat{\ell}_{i,t} = EST(i, \vec{q}_t, I_t, observed\_feedback)$$

  for all $i \in A$
  Update weights for all $i \in A$:

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

end for

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., $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$.

The adversaries we consider are nonoblivious. Namely, the losses at time t can be arbitrary functions of the past player's actions and the realise noise drawn in each round.

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

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

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 observa-

tions (i.e., $c_{i,t}$) and the noise parameter (if known). Each noise setting determines how the feedback is constructed and observed depending on the losses ℓ_t, the selected action I_t and the noise ϵ_t (line 7 in the algorithm template). Each algorithm determines how to construct the loss estimate $\hat{\ell}_{i,t}$ (line 8 in the algorithm). We denote generically $\hat{\ell}_{i,t} = EST(i, \vec{q}_t, I_t, observed\_feedback)$, 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 $\hat{L}_{k,T} = \sum_{t=1}^{T}\hat{\ell}_{k,t}$ be 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.

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.

3. Full Information with Constant Noise model

In this section, we consider the *Full Information with Constant Noise* feedback model. 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.

3.1. Algorithms

We start with 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

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

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 2 *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*

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

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

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

The following lemma establishes a well known property of EWS.

Lemma 3 *Let $\eta > 0$ and a sequence of loss estimates $\hat{\ell}_1, \ldots, \hat{\ell}_T$ where $\hat{\ell}_t : \{1,\ldots,K\} \to \mathbb{R}$ 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*

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

Full proof of Theorem 2 follows by using Lemma 3, the fact that the estimator is unbiased, and bounding the second moment of the estimator by $\mathrm{E}[(\hat{\ell}_{i,t})^2] \leq \frac{1}{\epsilon^2}$.

In Theorem 2, 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*

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

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

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

3.2. 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 $\epsilon \in (0, \frac{1}{2})$ and $T \geq 2\ln K$. Then for any algorithm, there exists a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

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

The proof idea is to define a stochastic strategy for loss assignment, which is a distribution over problem instances. Then, by showing that any algorithm suffers high expected regret, where the expectation is over the problem instances defined by the strategy, conclude that there exists a problem instance with high regret.

Proof sketch for $K \geq 27$ 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 feedback** $c_{i,t}$. A simple calculation yields that for $i^\star$ we have $Pr[c_{i^\star,t} = 1] = \frac{1}{2} - \epsilon\delta$ and for $i \neq i^\star$ we have $Pr[c_{i,t} = 1] = \frac{1}{2}$, where the probability is taken over the draw of $R_{i,t} \sim B(\frac{1-\epsilon}{2})$ and the draw of the losses $\ell_{i,t}$.

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

Denote by $C_{i,T} = \sum_{t=1}^{T} c_{i,t}$, the sum of the noisy feedback of action i, and note that it 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 $C_{i,T} \sim B(T, \frac{1}{2})$.
Using a standard claim regarding the minimum of i.i.d. binomial random variables we show that with probability at least $\frac{1}{4}$ there exist action $j \neq i^\star$ such that $C_{j,T} < C_{i^\star,T}$.
Standard Bayesian argument gives us

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

Applying this argument for each round t, implies that the optimal algorithm satisfies,

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

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 depends on 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. □

4. Full Information with Variable Noise model

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 parameters 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 picks an *action* $I_t \in A$. Then the learner observes the realized noise $(\epsilon_{1,t}, \ldots, \epsilon_{K,t})$ and 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.

4.1. Uniform Noise Distribution

4.1.1. Algorithm

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

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

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:

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

The algorithm resulting from using the above estimator in the *Exponential Weights Scheme* is called **EW-Threshold**.

Theorem 6 *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,

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

4.1.2. 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*, with uniform marginals, 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 of the following theorem 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 7 *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*

$$Regret(T) \geq \frac{1}{48} T^{2/3}(\ln K)^{1/3}$$

Proof sketch Let $\theta = c$. 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 rounds where $\epsilon_t \leq \theta$ (*bad rounds*). Since $\mathrm{E}[T'] = \theta T$ we have that with probability at least $\frac{1}{2}$ 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

$$Regret(T', \theta) \geq \frac{1}{24\theta}\sqrt{T' \ln K} = \frac{1}{24} T^{2/3}(\ln K)^{1/3}$$

Where in the last equality we use $T' = \theta T,\ \theta = (\frac{\ln K}{T})^{1/3}$
Putting it back in the original model yields,

$$\begin{aligned} Regret(T) &\geq \Pr[T' \geq \theta T]\ Regret(\theta T, \theta) \\ &\geq \frac{1}{2}\ \frac{1}{24} T^{2/3}(\ln K)^{1/3} \end{aligned}$$

(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 $T = T'$). □

4.2. 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 6 and obtain the following general upper bound.

Theorem 8 *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

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

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 9 *Let D be a noise distribution, where each marginal distribution has the same CDF F, and assume $F(\theta) \leq \theta^\alpha$ for a given $\alpha > 0$. Then, running **EW-Threshold** algorithm with $\eta = \sqrt{\frac{\ln K}{T g(\theta)}}$, where $g(\theta) = \mathrm{E}[\frac{1}{\epsilon^2}\mathbb{1}_{\{\epsilon \geq \theta\}}]$ satisfies,*

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

To get an intuition for the bound of Corollary 9 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 6, of $O(T^{2/3})$ (this is because we use the bound $g(\theta) \leq \frac{1}{\theta^2}$ which is not tight). 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 8 we prove a regret bound for truncated exponential distribution.

Corollary 10 *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,*

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

4.3. 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, $p_{i,t}$, with the feedback. 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 11 *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*

$$Regret(T) \geq \frac{1}{8} T$$

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 sketch We prove the theorem for the case of $K = 2$. Assume that initially the adversary picks the best action uniformly. 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 $1/4$.
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. A simple calculation of the expected value of the observed feedback, where the 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 $\ell_{i,t}$, yields that both actions will have the same probability of 1, namely $3/8$, and therefore indistinguishable by the learner.

This clearly implies that the learner cannot distinguish between the two actions, and therefore, in expectation, 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 an expected loss of $\frac{T}{2}$. This implies that the expected regret would be at least $\frac{T}{8}$. □

5. Bandit Models

In this section 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 $c_{I_t,t}$.

5.1. Bandit with Constant Noise Model

5.1.1. 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 12 *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*

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

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

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

The proof of Theorem 12 is similar in spirit to the proof of Theorem 4.

5.1.2. Impossibility result

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

Theorem 13 *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 $\vec{\ell}_1, ..., \vec{\ell}_T$ and a constant $\gamma > 0$ such that*

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

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.

5.2. 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 $[0,1]$.

5.2.1. 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:

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

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 14 *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*

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

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

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

5.2.2. Impossibility result

We show a lower bound of $\Omega((TK)^{2/3})$. The proof is similar to the proof of Theorem 7 for the *Full Information* settings. For the lower bound we use the distribution D' we used in theorem 7: we first draw $\epsilon_t \sim U(0,1)$ and for every i we set $\epsilon_{i,t} = \epsilon_t$.

Theorem 15 *For any algorithm in the Bandit with Variable Noise setting with D' as the noise parameters distribution, there exist a series of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ and a constant $\gamma > 0$ such that*

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

5.3. Importance of knowing the Noise

In Theorem 11 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 16 *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 $(\epsilon_{1,t}, ..., \epsilon_{K,t})$, then there exist a sequence of loss vectors $\vec{\ell}_1, ..., \vec{\ell}_T$ such that*

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

6. 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 & 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.

Acknowledgements

This work was supported in part by a grant from the Israel Science Foundation (ISF) and by the Tel Aviv University Yandex Initiative in Machine Learning. AR would like to thank Dor Elboim for fruitful discussions on probability.

References

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM J. Comput.*, 46(6):1785–1826, 2017.

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

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Freund, Y. and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Gajane, P., Urvoy, T., and Kaufmann, E. Corrupt bandits for preserving local privacy. In *ALT 2018-Algorithmic Learning Theory*, 2018.

Kalai, A. and Vempala, S. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

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

Kocák, T., Neu, G., and Valko, M. Online learning with noisy side observations. In *AISTATS*, pp. 1186–1194, 2016.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

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

Slivkins, A. Introduction to multi-armed bandits, 2017. URL http://slivkins.com/work/MAB-book.pdf.

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

Weissman, T. and Merhav, N. Universal prediction of binary individual sequences in the presence of noise. *IEEE Trans. Inform. Theory, September*, 2000.

Weissman, T., Merhav, N., and Somekh-Baruch, A. 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.

Wu, Y., György, A., and Szepesvári, C. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pp. 1360–1368, 2015.