Optimal Adversarial Policies in the Multiplicative Learning System With a Malicious Expert

S. Rasoul Etesami, Negar Kiyavash, Vincent Leon, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— We consider a learning system based on the conventional multiplicative weight (MW) rule that combines experts' advice to predict a sequence of true outcomes. It is assumed that one of the experts is malicious and aims to impose the maximum loss on the system. The system's loss is naturally defined to be the aggregate absolute difference between the sequence of predicted outcomes and the true outcomes. We consider this problem under both offline and online settings. In the offline setting where the malicious expert must choose its entire sequence of decisions a priori, we show somewhat surprisingly that a simple greedy policy of always reporting false prediction is asymptotically optimal with an approximation ratio of $1+O(\sqrt{\frac{\ln N}{N}})$, where N is the total number of prediction stages. In particular, we describe a policy that closely resembles the structure of the optimal offline policy. For the online setting where the malicious expert can adaptively make its decisions, we show that the optimal online policy can be efficiently computed by solving a dynamic program in $O(N^3)$. We also discuss a generalization of our model to multi-expert settings. Our results provide a new direction for vulnerability assessment of commonly-used learning algorithms to internal adversarial attacks.**

***Index Terms*— Adversarial learning, expert advice, Markov decision process, dynamic programming, approximation ratio.**

I. Introduction

THE focus of the vast literature on learning with expert advice is coming up with good prediction rules for the learning system even for the worst possible outcome sequence [1]–[6]. However, the proposed algorithms are not designed to be robust against malicious strategic experts. Given the prevalence of machine learning algorithms and, as a result, automated decision making in distributed settings in many real-world applications, the effect of malicious experts whose goal is to destroy the performance of the system by injecting false predictions cannot be ignored. In this paper, we address this issue by analyzing the performance of the *multiplicative weighted* (MW) learning algorithm [3], widely used in collaborative filtering, in the presence of malicious experts injecting false recommendations.

There are many motivating examples for considering the effect of malicious experts in real-world learning systems. To name a few, one can consider movie recommendation systems such as IMDB or Netflix, where the system relies on the users' feedback (experts) to rate the quality of the movies. However, the users do not always report the true ratings due to various reasons such as manipulating the outcome of the system toward their preferences [6], [7]. As another example, one can consider sensor fusion in networks where a malicious sensor can attack the system by injecting false signals and cause the central decision-maker to reach incorrect decisions [8]. Moreover, almost all cases of collaborative filtering or distributed decision making are vulnerable to internal attacks.

In this paper, we study the performance of the MW learning algorithm against adversarial attacks where the adversary's goal is to attack the system without having control over the system's prediction rule. The MW update rule is one of the most commonly used schemes for learning from expert advice [1], [9], [10], in which after each stage of prediction, when the true outcome is revealed, depending on whether the experts were correct or wrong on that stage, the system punishes or rewards the experts, respectively, by decreasing or increasing their relative weights by a multiplicative factor. Thus, learning with expert advice can be modeled in a multistage sequential decision-making framework where at each stage, the recommendation system combines the predictions of a set of experts about an unknown outcome with the aim of accurately predicting that outcome.

The problem that we consider here was originally proposed in [11] and subsequently studied in [8], where it was shown that in the case of *logarithmic* loss function, the optimal online policy for the malicious expert is a simple greedy policy. This, however, is not very surprising as the malicious expert's gain by reporting false predictions dominates his credibility loss due to the logarithmic nature of the loss function. As it was shown using numerical analysis in [8], [11], characterizing the optimal policy for the absolute loss function (which is the more interesting case and commonly used in MW learning systems) is much more challenging due to the strong coupling between the gain in reporting false prediction and the loss in credibility. In this work, we answer this question by showing that the same simple greedy algorithm is *asymptotically* optimal in the

Manuscript received December 30, 2019; revised September 17, 2020 and December 28, 2020; accepted January 6, 2021. Date of publication January 18, 2021; date of current version February 9, 2021. This work was supported by the U.S. National Science Foundation under Grant EPCN-1944403 and Grant CCF-1908308. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shantanu D. Rane. *(Corresponding author: S. Rasoul Etesami.)*

S. Rasoul Etesami and Vincent Leon are with the Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: etesami1@illinois.edu; jliang38@illinois.edu).

Negar Kiyavash is with the College of Management of Technology, EPFL, CH-1015 Lausanne, Switzerland (e-mail: negar.kiyavash@epfl.ch).

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: poor@princeton.edu).

Digital Object Identifier 10.1109/TIFS.2021.3052360

1556-6021 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

offline setting. Moreover, we show that although the optimal online policy can have a complicated structure, it can still be computed efficiently using a reduced-size dynamic program.

The problem that we study in this paper also belongs to the general family of many problems such as target tracking, distributed detection under the byzantine attacks, Sybil attack, and causative attack from the taxonomy of adversarial machine learning where the attacker can modify the data in training or during the operation in order to degrade the performance of a machine learning algorithm [12]–[15]. Our work is also related to [16]–[18] in which a learner plays against an adversary such that at each step, the learner has to choose an expert from a pool of experts to follow while the adversary adaptively sets the gains for the experts to maximize the overall regret incurred by the learner. The authors in [17] fully characterize the optimal online policies for the learner and the adversary in the case of 2 and 3 experts and provide some general insights into how to design an optimal algorithm for the learner and the adversary for an arbitrary number of experts. However, our work is different from those in the sense that the experts in our setting are themselves malicious and can act strategically. Moreover, the performance guarantee in our setting is in terms of the approximation factor rather than the notion of regret.

We consider the problem of learning with a malicious strategic expert under both *offline* and *online* settings. More specifically, we consider a system with two experts; one honest and the other malicious. The honest expert predicts the true outcome with some accuracy at each round, while the malicious expert strategically provides a prediction to maximize the loss incurred by the system. We assume that the adversary knows the true outcome and prediction rule of the learning system. For the offline setting, we assume that the adversary reports his entire sequence of predictions at the beginning of the horizon, while for the online setting, the adversary is allowed to look at the past information up to the current stage and then reports his next prediction. The problem that we address in this paper is two-fold: From the malicious expert's point of view, we are interested in knowing the optimal policy which imposes the maximum loss on the learning system, while from the system designer's point of view, we are interested in knowing how the widely-applied MW learning algorithm performs in the presence of a malicious expert.

As one of our main results, we show that for the case of the absolute loss function, the optimal offline policy can be approximated within a factor $1 + O(\sqrt{\frac{\ln N}{N}})$ of the one which reports false predictions at all the stages, where N is the total number of prediction stages. This can be viewed as a counterpart to the conventional regret minimization bounds obtained for the MW update rule. It is worth noting that obtaining such an approximation ratio is more challenging than obtaining regret bounds commonly used in expert advice settings. The reason is that the space of feasible policies is exponentially larger than the set of feasible actions. Therefore, for the offline setting, we approximate the optimal offline policy rather than the best action. We then extend our results to the online setting and characterize the optimal online policy using a dynamic program (DP). In particular, we show that the number of states of this dynamic program grows linearly in terms of the number of stages, which allows us to compute the optimal online policy in $O(N^3)$.

The paper is organized as follows: In Section II, we introduce the model formally and discuss some of its salient properties. In Section III, we provide our main results for the case of offline malicious expert and absolute loss function. In Section IV, we provide an efficient algorithm for computing the optimal online policy for the case of two experts, with an extension for the case of multiple honest experts. Simulation results for offline and online adversaries are provided in Section V. We conclude the paper by identifying some future directions of research in Section VI.

II. Problem Formulation

In this section, we first introduce the mathematical model formally as in [11] and then provide some of its salient properties, which will be used in our later analysis. In the remainder of this paper, we shall refer to the ill-intent expert as a malicious expert or an adversary, interchangeably.

Consider a learning system with two experts. At each round $k = 0, 1, 2, \ldots$, expert $i \in \{1, 2\}$ has a nonnegative weight denoted by $p_k^i \in [0, 1]$. We assume that both experts start with equal initial weight $p_0^1 = p_0^2 = 1$. We denote the prediction of the ith expert at stage k by $x_k^i \in \{0, 1\}$, and the true outcome by $y_k \in \{0, 1\}$. At stage k, the relative weight of expert $i \in \{1, 2\}$ is defined to be

$$\tilde{p}_k^i := \frac{p_k^i}{p_k^1 + p_k^2}. \tag{1}$$

In the kth stage, the learning system predicts the true outcome y_k using a weighted average rule given by

$$\hat{y}_k = \tilde{p}_k^1 x_k^1 + \tilde{p}_k^2 x_k^2, \tag{2}$$

and updates the experts' weights in the next time step depending on whether they were correct or wrong in the previous instance using the following multiplicative weight (MW) update rule:

$$p_{k+1}^i = \begin{cases} p_k^i \epsilon & \text{if } x_k^i \neq y_k, \\ p_k^i & \text{if } x_k^i = y_k. \end{cases} \tag{3}$$

Here $\epsilon \in (0, 1)$ is a fixed constant parameter set the learning system and reflects its aggressiveness on punishing/rewarding the experts. We note that the MW update rule (3) has been extensively used in the past literature [3], [10], [19], [20]. In particular, the MW learning system serves as an independent forecaster (executor). Unlike the adversary, the learning system is neither strategic nor has access to the information of the true outcomes: it merely takes the experts' advice and computes the prediction in each round using (2). After the true outcome y_k is revealed, the system incurs a loss $l(\hat{y}_k, y_k) = Q(|\hat{y}_k - y_k|)$, where $Q(\cdot) : [0, 1] \to \mathbb{R}^{\geq 0}$ can be some general nondecreasing function. In this paper, we shall only focus on the *absolute* loss function $Q(y) := y$, as it is the most common loss function used in the literature for the expert advice setting [3], [10].

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

We assume that expert 2 is the ***honest*** expert who makes a correct prediction with accuracy μ, i.e., the one that agrees with the true outcome with probability μ:

$$x_k^2 = \begin{cases} y_k & \text{w.p. } \mu, \\ 1 - y_k & \text{w.p. } 1-\mu. \end{cases}$$

Remark 1: For asymmetric accuracies $\{\mu_k\}_{k\in[N]}$, one can partition the horizon into epochs of a small constant length. As the honest expert predictions are independent, one may assume that the honest expert's expected accuracy within each epoch is close to its expected value denoted by μ. Therefore, our analysis can be viewed as a constant approximation of the heterogeneous model in the stationary regime.

Expert 1 is the ***malicious*** expert (adversary) who aims to impose the maximum loss on the system by taking the best adversarial action at each stage. We assume that expert 1 knows the true outcome y_k at time $k \in [N] := \{1, \ldots, N\}$, as well as the distribution of x_k^2, the prediction of expert 2.[1] One of our main objectives in this paper is to evaluate the robustness of the MW learning algorithm in the presence of a malicious expert. For that reason, we evaluate the system's performance against an adverse scenario where the adversary gets to know the true outcome one stage ahead of the honest expert (e.g., due to some side information from the outside). Another way to interpret such an adversarial model is to consider an arbitrary sequence of outcomes and a "conservative" adversary who attempts an attack only on the outcomes that it has full information about them. Therefore, upon the arrival of y_k, if the conservative adversary does not know that outcome, it will mimic the honest expert's response in order to keep its credibility for future outcomes. Otherwise, the adversary knows the true outcome y_k, in which case we are governed again by our adversarial model. We refer to Proposition 1 for a weaker adversary with no information about an arbitrary sequence of true outcomes.

*Definition 1: A malicious expert is called an **offline** adversary if he chooses his entire of sequence of predictions $\{x_k^1\}_{k=1}^N$ at the beginning of the horizon and then commits to it. A malicious expert is called an **online** adversary if the entire history of predictions and true outcomes $\{\tilde{p}_\ell^1, x_\ell^1, x_\ell^2, y_\ell\}_{\ell=1}^{k-1}$ are available to him, and then he decides x_k^1.*

Finally, the goal of the malicious expert (either offline or online) is to produce a sequence of predictions $\{x_k^1\}_{k=1}^N$ over a fixed finite horizon N in order to maximize the expected aggregate loss on the system given by:

$$\mathbb{E}_{x_1^2,\ldots,x_N^2}[\sum_{k=1}^N l(\hat{y}_k, y_k)] = \sum_{k=1}^N \mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)], \tag{4}$$

where the second expectation is taken over the past and current actions of the honest agent $x_1^2, \ldots, x_k^2$. In particular, an ***optimal policy*** for the offline/online malicious expert is a sequence of decisions which maximizes the objective function (4) with respect to its corresponding information set, i.e., a solution to the maximization problem $\max_{x_1^1,\ldots,x_N^1} \sum_{k=1}^N \mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)]$.

One of the major differences between the above model and the conventional expert advice problem is that in the latter, one assumes that all the experts are honest and report their true recommendations. In particular, the goal is to devise a learning scheme that intelligently combines the experts' recommendations to accurately predict the unknown outcomes, where it can be shown that the well-known MW learning rule achieves the minimum regret bound. However, the above adversarial model can be viewed as a dual to the expert advice problem where the MW rule is fixed as the underlying learning process, and the goal to evaluate how well this learning rule will perform in the presence of a malicious expert who strategically aims to maximize the loss of the system.

A. Preliminary Results

Here, we describe some of the important properties of the aforementioned adversarial model which will be used later to establish our main results. First we note that using the update rule (3) and the definition of relative weights (1), we have

$$\tilde{p}_{k+1}^1 = \begin{cases} \frac{1}{1+\left(\frac{1}{\tilde{p}_k^1}-1\right)\frac{1}{\epsilon}} & \text{if } x_k^1 = 1-y_k, x_k^2 = y_k, \\ \frac{1}{1+\left(\frac{1}{\tilde{p}_k^1}-1\right)\epsilon} & \text{if } x_k^1 = y_k, x_k^2 = 1-y_k, \\ \tilde{p}_k^1 & \text{if } x_k^1 = x_k^2, \end{cases} \tag{5}$$

In particular, from (5) one can easily see that the adversary's relative weight changes only when his prediction is at odds with the prediction of the honest agent (when both experts predict the same, the adversary's relative weight remains unchanged). As the update rule in (5) plays an important role in our analysis, we define a weight update function $g : (0,1] \to (0,1]$ and its inverse $g^{(-1)} : (0,1] \to (0,1]$ by

$$\begin{aligned} g(\rho) &:= \frac{1}{1+\left(\frac{1}{\rho}-1\right)\frac{1}{\epsilon}}, \\ g^{(-1)}(\rho) &:= \frac{1}{1+\left(\frac{1}{\rho}-1\right)\epsilon}. \end{aligned} \tag{6}$$

In fact, both $g(\rho)$ and its inverse $g^{(-1)}(\rho)$ are strictly increasing functions and we have $g(\rho) \le \rho \le g^{-1}(\rho), \forall \rho \in (0,1]$. An important feature of the functions $g(\rho)$ and $g^{(-1)}(\rho)$ is that for any integer $j \in \mathbb{Z}^+$, we have

$$\begin{aligned} g^{(j)}(\rho) &:= \underbrace{g(\ldots(g(\rho))}_{j \text{ times}} = \frac{1}{1+\left(\frac{1}{\rho}-1\right)\frac{1}{\epsilon^j}}, \\ g^{(-j)}(\rho) &:= \underbrace{g^{(-1)}(\ldots(g^{(-1)}(\rho))}_{j \text{ times}} = \frac{1}{1+\left(\frac{1}{\rho}-1\right)\epsilon^j}, \end{aligned} \tag{7}$$

where $g^{(j)}(\rho)$ and $g^{(-j)}(\rho)$ denote the composition of $g(\rho)$ and $g^{(-1)}(\rho)$ by themselves j times, respectively. In particular, we note that $g^{(0)}(\rho) \equiv \rho$.

[1] Note that the assumption that the adversary knows the prediction accuracy μ is not very restrictive as the adversary can always learn this distribution using the empirical history of observed actions taken by the honest expert.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

III. Optimal Offline Policy for the Absolute Loss

In this section, we analyze the optimal policy for the offline adversary and postpone our analysis for the case of the online adversary to Section IV. We recall that the offline adversary is the one who chooses his entire sequence of decisions (predictions) at the beginning of the horizon. More precisely, the offline adversary aims to maximize the expected loss of the learning system given by (4) over all the 2^N feasible sequences of the form $\{0,1\}^N$. Note that although the space of feasible solutions is exponentially large, we are only interested in obtaining polynomial-time computable policies. Thus, our goal is to approximate the optimal offline policy within only a negligible additive error term in the overall objective cost.

Toward this end, we first establish a sequence of lemmas to prove our main approximation result (Theorem 1). In fact, many of these lemmas do not make any use of the specific structure of the functions $g(\rho)$ and $Q(\cdot)$, and we state them in a more general form. Later, in order to provide more closed-form approximation results, we specialize these lemmas to the specific choice of $g(\rho)$ given in (6) and linear loss function $Q(y) = y$. It is worth noting that although we assumed that the learning algorithm starts with equal initial weight for both experts (i.e., the initial relative weight of the adversary is 0.5), however, we state our results for an offline adversary with generic initial relative weight ρ. The reason for this choice would become apparent subsequently. Next, we state the following lemma from [8, Lemma 1] whose proof is by induction on the horizon length N.

Lemma 1: For a loss function $l(\hat{y}, y) = Q(|\hat{y} - y|)$, with $Q : [0,1] \to \mathbb{R}^{\geq 0}$, the expected loss given in (4) *is fully determined by the initial relative weight of the adversary ρ, his policy $\Psi := (x_1^1, \ldots, x_N^1) \in \{0,1\}^N$, and the horizon length N.*

From Lemma 1 one can see that the adversary can take his optimal actions by only adjusting them relative to the honest expert's actions. Henceforth, the expected loss in (4) for a given policy $\Psi = (x_1^1, x_2^1, \ldots, x_n^1)$ of the offline adversary can be represented by $V_n^{\Psi}(\rho) := \sum_{k=1}^{n} \mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)]$, where ρ denotes the initial relative weight of the adversary.

Definition 2: Assume the adversary's initial weight is ρ and the number of stages is n. An adversary's policy is called a ***false policy*** *if he lies in all the stages, i.e., $x_k^1 = 1 - y_k, \forall k \in [n]$. It is called a* ***true policy*** *if the adversary tells the truth in all the stages, i.e., $x_k^1 = y_k, \forall k \in [n]$. We let $V_n^{\mathrm{f}}(\rho)$ and $V_n^{\mathrm{t}}(\rho)$ denote the expected loss of the system if the adversary follows the false policy and the true policy, respectively.*

Using the above definition, we can obtain closed-form relations for the expected loss of the false/true policies given in the following lemma. We will use these expressions as black-boxes in our approximation analysis.

Lemma 2: For a loss function $l(\hat{y}, y) := Q(|\hat{y} - y|)$, initial adversary's weight ρ, and n stages, we have

$$V_n^{\mathrm{f}}(\rho) = n(1-\mu)Q(1) + \sum_{j=0}^{n} \mathbb{P}(Z > j) Q(g^{(j)}(\rho)),$$

$$V_n^{\mathrm{t}}(\rho) = n\mu Q(0) + \sum_{j=0}^{n} \mathbb{P}(W > j) Q(1 - g^{(-j)}(\rho)),$$

where $Z \sim \mathrm{Bin}(n, \mu)$ and $W \sim \mathrm{Bin}(n, 1-\mu)$ are Binomial distributions with parameters μ and $1-\mu$, respectively.

Proof: Let us fix the adversary's policy to the false policy, and we look at all the possible sample paths which can be realized by predictions of the honest expert. Any sample-path in which the honest expert predicts correctly i times and makes a mistake $n-i$ times will occur with the same probability of $\mu^i(1-\mu)^{n-i}$. There are $\binom{n}{i}$ of such sample paths, and for any of them, independent of what positions the honest expert predicts correctly or wrongly, the incurred loss given the fixed adversary's false policy equals to $(n-i)Q(1) + \sum_{j=0}^{i-1} Q(g^{(j)}(\rho))$. The reason is that, for any of $n-i$ false predictions of the honest agent in this sample-path, the system incurs a loss of $Q(1)$ and by (5) the relative weight of the adversary does not change. Moreover, for the remaining i correct predictions and regardless of their order, the system incurs a loss of $\sum_{j=0}^{i-1} Q(g^{(j)}(\rho))$ (note that for $i = 0$ this term equals to 0). Therefore, by taking an expectation over all possible sample paths we have,

$$\begin{aligned}
V_n^{\mathrm{f}}(\rho) &= \sum_{i=0}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} (n-i) Q(1) \\
&\quad + \sum_{i=0}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} \sum_{j=0}^{i-1} Q(g^{(j)}(\rho)) \\
&= n(1-\mu)Q(1) + \sum_{i=0}^{n} \sum_{j=0}^{i-1} \binom{n}{i} \mu^i (1-\mu)^{n-i} Q(g^{(j)}(\rho)) \\
&= n(1-\mu)Q(1) + \sum_{j=0}^{n-1} \sum_{i=j+1}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} Q(g^{(j)}(\rho)) \\
&= n(1-\mu)Q(1) + \sum_{j=0}^{n} \mathbb{P}(Z > j) Q(g^{(j)}(\rho)),
\end{aligned}$$

where in the last equality we have used the fact that $Z \sim Bin(n, \mu)$ and $\mathbb{P}(Z > n) = 0$.

Similarly, to compute $V_n^{\mathrm{t}}(\rho)$ we can fix the adversary's policy to the true policy. Now for any sample-path realized by the honest expert with i correct predictions, the system incurs a loss of $i \cdot Q(0) + \sum_{j=0}^{n-i-1} Q(1 - g^{(-j)}(\rho))$. Finally, taking an expectation over all sample paths we get

$$\begin{aligned}
V_n^{\mathrm{t}}(\rho) &= \sum_{i=0}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} i \cdot Q(0) \\
&\quad + \sum_{i=0}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} \sum_{j=0}^{n-i-1} Q(1 - g^{(-j)}(\rho)) \\
&= n\mu Q(0) + \sum_{j=0}^{n-1} \sum_{i=0}^{n-j-1} \binom{n}{i} \mu^i (1-\mu)^{n-i} Q(1 - g^{(-j)}(\rho)) \\
&= n\mu Q(0) + \sum_{j=0}^{n} \mathbb{P}(W > j) Q(1 - g^{(-j)}(\rho)),
\end{aligned}$$

where the second equality is by switching the order of summations, and the last equality is by $W \sim \mathrm{Bin}(n, 1-\mu)$. □

Next, we note that every offline policy can be partitioned into several blocks such that within each block, the adversary follows either false or true policies. Thus we can characterize any offline policy by simply determining

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

the length of each of its sub-blocks. For this purpose, let $n_1, m_1, n_2, m_2, \ldots, n_k, m_k$, denote the partition of the entire horizon N into some sub-horizons of integer length for some positive integer k such that $N = \sum_{i=1}^{k}(n_i + m_i)$, and $m_i, n_i \in \mathbb{Z}^+$ (note that n_1 or m_k can also be zero). We assume that the adversary follows the false policy within each block of length n_i, and the true policy within each block of length m_i. Therefore, finding the optimal offline policy reduces to maximizing the expected loss (4) over all such partitions.

Lemma 3: Given an adversary's initial relative weight ρ, the relative weight of the adversary after lying n times and telling truth m times (in any arbitrary order) equals to $g^{(X-Y)}(\rho)$, where $X \sim \text{Bin}(n, \mu)$ and $Y \sim \text{Bin}(m, 1-\mu)$ are independent Binomial random variables.

Proof: Let $\bar{X}_i \sim \text{Ber}(\mu)$ and $\bar{Y}_i \sim \text{Ber}(1-\mu)$, $i = 1, 2, \ldots$ be independent Bernoulli random variables, and ρ be the initial weight of the adversary. Since at each stage the honest expert predicts independently from the earlier stages, a simple induction shows that if the adversary's weight at the beginning of the kth stage equals to $g^{(U)}(\rho)$ for some random variable U, then after the kth stage depending on whether he lies or tells the truth, his weight will change to $g^{(U+\bar{X}_k)}(\rho)$ or $g^{(U-\bar{Y}_k)}(\rho)$, respectively. Therefore, if we know that the adversary lies exactly n times and tells the truth m times, his relative weight at the end of this process will be equal to $g^{(X-Y)}(\rho)$, where X is the sum of n independent Bernoulli random variables of type $\bar{X}_i \sim \text{Ber}(\mu)$, and Y is the sum of m independent Bernoulli random variables of type $\bar{Y}_i \sim \text{Ber}(1-\mu)$. This implies that $X \sim \text{Bin}(n, \mu)$ and $Y \sim \text{Bin}(m, 1-\mu)$, and that X and Y are independent. □

Lemma 3 indicates that the distribution of the adversary's relative weight, induced by following an offline policy Ψ, only depends on the total number of times that the adversary lies or tells the truth, and not on the specific order of them. Note that this property only holds for the relative weight distribution, but not for the distribution of the accumulated loss at different stages. In fact, it can be shown that the distribution of loss depends critically on the order of the adversary's actions, and that is the main difficulty in the analysis of the optimal offline policy. We circumvent this issue in the following theorem by providing an approximation scheme that is asymptotically optimal as the number of stages approaches infinity.

Theorem 1: For any $\epsilon \in (0, 1)$, and the absolute loss function $l(\hat{y}, y) = |\hat{y} - y|$, we have $\frac{V_N^{\Psi^*}(0.5)}{V_N^{\text{f}}(0.5)} = 1 + O(\sqrt{\frac{\log_{1/\epsilon} N}{N}})$, where $V_N^{\Psi^*}(0.5)$ and $V_N^{\text{f}}(0.5)$ denote the expected loss by following the optimal policy and the false policy, respectively.

Proof: For simplicity and without any loss of generality we set $\epsilon = \frac{1}{e}$. For general $\epsilon \in (0, 1)$, the only difference in our analysis would be that the base of the natural logarithm will change to $\frac{1}{\epsilon}$. Using Lemma 2 specialized for the absolute loss function $Q(y) = y$, we can write

$$V_n^{\text{f}}(\rho) = (1-\mu)n + \sum_{j=0}^{n} \mathbb{P}(Z > j) g^{(j)}(\rho),$$
$$V_n^{\text{t}}(\rho) = (1-\mu)n - \sum_{j=0}^{n} \mathbb{P}(W > j) g^{(-j)}(\rho), \tag{8}$$

where $Z \sim \text{Bin}(n, \mu)$ and $W \sim \text{Bin}(n, 1-\mu)$. For any arbitrary but fixed $\rho \in (0, 1)$, let us define

$$f(r) := r - \ln(1 + ae^r),$$
$$\epsilon(r) := f(r+1) - f(r) - g^{(r)}(\rho),$$

where $a := \frac{1}{\rho} - 1$ and $r \in [0, \infty)$. Now we can write,

$$\begin{aligned}
V_n^{\text{f}}(\rho) &= (1-\mu)n + \sum_{j=0}^{n} \mathbb{P}(Z > j)[f(j+1) - f(j) - \epsilon(j)] \\
&= (1-\mu)n - \sum_{j=0}^{n} \mathbb{P}(Z > j)\epsilon(j) \\
&\quad + \sum_{j=0}^{n} [\mathbb{P}(Z > j-1) - \mathbb{P}(Z > j)] f(j) - f(0) \\
&= (1-\mu)n - \sum_{j=0}^{n} \mathbb{P}(Z > j)\epsilon(j) + \mathbb{E}[f(Z)] - f(0) \\
&\overset{(a)}{\leq} (1-\mu)n + \mathbb{E}[f(Z)] - f(0) \\
&\quad - \sum_{j=0}^{n} \mathbb{P}(Z > j) \left(\frac{1}{1+ae^{j+1}} - \frac{1}{1+ae^{j}} \right) \\
&\overset{(b)}{=} (1-\mu)n + \mathbb{E}[f(Z)] - f(0) + g^{(0)}(\rho) - \mathbb{E}[g^{(Z)}(\rho)] \\
&\overset{(c)}{=} n - \mathbb{E}[\ln(1 + (\frac{1}{\rho} - 1)e^Z)] - \ln(\rho) \\
&\quad + \rho - \mathbb{E}[\frac{1}{1 + (\frac{1}{\rho} - 1)e^Z}],
\end{aligned} \tag{9}$$

where (a) is due to Lemma 4 (Appendix A) which shows that $\epsilon(r) \geq \frac{1}{1+ae^{r+1}} - \frac{1}{1+ae^r}$, and (b) follows from (7) and the definition of expectation. Finally, (c) follows by substituting the expressions for $f(Z)$ and $g^{(Z)}(\rho)$. Similarly, to obtain an upper bound for $V_n^{\text{t}}(\rho)$, let us define

$$h(r) := \ln(a + e^r),$$
$$\delta(r) := h(r+1) - h(r) - g^{(-r)}(\rho).$$

Using identical steps as in the derivation of (9) and since by Lemma 4, $\delta(r) \leq \frac{1}{1+ae^{-(r+1)}} - \frac{1}{1+ae^{-r}}$, we get

$$\begin{aligned}
V_n^{\text{t}}(\rho) &\leq (1-\mu)n - \mathbb{E}[\ln((\frac{1}{\rho} - 1) + e^W)] - \ln(\rho) \\
&\quad - \rho + \mathbb{E}[\frac{1}{1 + (\frac{1}{\rho} - 1)e^{-W}}].
\end{aligned} \tag{10}$$

Next let us consider an arbitrary offline policy Ψ characterized by its false/true sub-block, i.e., $\Psi := n_1, m_1, \ldots, n_k, m_k$. Denote the expected loss under policy Ψ when the initial weight of the adversary was 0.5 by $V_N^{\Psi}(0.5)$. Moreover, for $\ell = 1, \ldots, k$, let $\bar{X}_\ell \sim \text{Bin}(n_\ell, \mu)$ and $\bar{Y}_\ell \sim \text{Bin}(m_\ell, 1-\mu)$ be pairwise independent Binomial distributions (i.e., for every i and every j, $\bar{X}_i$ and $\bar{Y}_j$ are independent) and define $X_\ell = \sum_{i=1}^{\ell} \bar{X}_i$, and $Y_\ell = \sum_{i=1}^{\ell} \bar{Y}_i$. Note that the pair X_ℓ and Y_ℓ are independent Binomial distributions. By linearity of expectation, and using Lemma 3 we can write,

$$\begin{aligned}
&V_N^{\Psi}(0.5) \\
&= V_{n_1}^{\text{f}}(0.5) + \mathbb{E}\left[V_{m_1}^{\text{t}}\left(g^{(X_1)}(0.5) \right) \right] \\
&\quad + \mathbb{E}\left[V_{n_2}^{\text{f}}\left(g^{(X_1 - Y_1)}(0.5) \right) \right] + \mathbb{E}\left[V_{m_2}^{\text{t}}\left(g^{(X_2 - Y_1)}(0.5) \right) \right] \\
&\quad + \ldots + \mathbb{E}\left[V_{m_k}^{\text{t}}\left(g^{(X_k - Y_{k-1})}(0.5) \right) \right].
\end{aligned} \tag{11}$$

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

Replacing $g^{(X_{\ell-1}-Y_{\ell-1})}(0.5)$ (or for brevity $g^{(X_{\ell-1}-Y_{\ell-1})}$) from (7) instead of ρ in (9), and taking expectation we have

$$\begin{aligned}
&\mathbb{E}\left[V^{\mathrm{f}}_{n_\ell}\left(g^{(X_{\ell-1}-Y_{\ell-1})}\right)\right] \\
&\quad\le \mathbb{E}\left[n_\ell - \mathbb{E}\Big[\ln\Big(1+\big(\frac{1}{g^{(X_{\ell-1}-Y_{\ell-1})}}-1\big)e^{Z}\Big)\Big]\right] \\
&\qquad -\mathbb{E}\left[\ln\left(g^{(X_{\ell-1}-Y_{\ell-1})}\right)\right] \\
&\qquad +\mathbb{E}\Big[g^{(X_{\ell-1}-Y_{\ell-1})} - \mathbb{E}\Big[\frac{1}{1+\big(\frac{1}{g^{(X_{\ell-1}-Y_{\ell-1})}}-1\big)e^{Z}}\Big]\Big] \\
&\quad= n_\ell - \mathbb{E}\left[\ln\left(\frac{1+e^{X_\ell-Y_{\ell-1}}}{1+e^{X_{\ell-1}-Y_{\ell-1}}}\right)\right] \\
&\qquad +\mathbb{E}\Big[\frac{1}{1+e^{X_{\ell-1}-Y_{\ell-1}}} - \frac{1}{1+e^{X_\ell-Y_{\ell-1}}}\Big], \qquad (12)
\end{aligned}$$

where the equality follows by simplifying the terms and noting that for the ℓ-the false block $Z := \bar{X}_\ell \sim \mathrm{Bin}(n_\ell, \mu)$, which is independent of $X_{\ell-1}$ and $Y_{\ell-1}$. Similarly, since for the ℓ-th true block $W := \bar{Y}_\ell \sim \mathrm{Bin}(m_\ell, 1-\mu)$, which is independent of X_ℓ and $Y_{\ell-1}$, by replacing $g^{X_\ell-Y_{\ell-1}}(0.5)$ instead of ρ into (10) and taking expectation we get

$$\begin{aligned}
&\mathbb{E}\left[V^{\mathrm{t}}_{m_\ell}\left(g^{(X_\ell-Y_{\ell-1})}(0.5)\right)\right] \\
&\quad\le (1-\mu)m_\ell - \mathbb{E}\Big[\ln\Big(\frac{e^{X_\ell-Y_{\ell-1}}+e^{\bar{Y}_\ell}}{1+e^{X_\ell-Y_{\ell-1}}}\Big)\Big] \\
&\qquad +\mathbb{E}\Big[\frac{1}{1+e^{X_\ell-Y_\ell}} - \frac{1}{1+e^{X_\ell-Y_{\ell-1}}}\Big]. \qquad (13)
\end{aligned}$$

Finally, substituting (12) and (13) into (11), we can write

$$\begin{aligned}
&V^{\Psi}_N(0.5) \\
&\le \sum_{\ell=1}^{k}(n_\ell+(1-\mu)m_\ell) \\
&\quad -\mathbb{E}\Big[\sum_{\ell=1}^{k}\Big(\ln\Big(\frac{e^{X_\ell-Y_{\ell-1}}+e^{\bar{Y}_\ell}}{1+e^{X_\ell-Y_{\ell-1}}}\Big)+\ln\Big(\frac{1+e^{X_\ell-Y_{\ell-1}}}{1+e^{X_{\ell-1}-Y_{\ell-1}}}\Big)\Big)\Big] \\
&\quad +2\sum_{\ell=1}^{k}\mathbb{E}\Big[\frac{1}{1+e^{X_\ell-Y_\ell}}-\frac{1}{1+e^{X_\ell-Y_{\ell-1}}}\Big] \\
&= \mu M + (1-\mu)N - \mathbb{E}\left[\ln\left(\prod_{\ell=1}^{k}\frac{e^{X_\ell-Y_{\ell-1}}+e^{\bar{Y}_\ell}}{1+e^{X_{\ell-1}-Y_{\ell-1}}}\right)\right] \\
&\quad +O(\sqrt{N\ln N}) \\
&= (1-\mu)N - \mathbb{E}\left[\ln\left(\frac{1+e^{Y_k-X_k}}{2}\right)\right] + O(\sqrt{N\ln N}) \\
&= (1-\mu)N + O(\sqrt{N\ln N}), \qquad (14)
\end{aligned}$$

where the first equality is due to Lemma 8 (Appendix A) and noting that $\sum_{\ell=1}^{k} n_\ell = M$ and $\sum_{\ell=1}^{k} m_\ell = N-M$. Moreover, the second equality holds because $\bar{Y}_\ell = Y_\ell - Y_{\ell-1}$ (which causes telescopic cancellation of the product terms inside of the natural logarithm) and noting that $X_0 = Y_0 = 0$, $\mathbb{E}[X_k] = \mu M$. Using (8), the expected loss of the false policy is at least,

$$V^{\mathrm{f}}_N(0.5) = (1-\mu)N + \sum_{j=0}^{N}\mathbb{P}(Z>j)g^{(j)}(0.5) \ge (1-\mu)N.$$

(Note that all the terms $\mathbb{P}(Z > j)g^{(j)}(0.5)$ are nonnegative.) This in view of (14) shows that $\frac{V^{\Psi^*}_N(0.5)}{V^{\mathrm{f}}_N(0.5)} = 1 + O(\sqrt{\frac{\ln N}{N}})$. □

Remark 2: It is important to distinguish the difference between the approximation ratio obtained in Theorem 1 and the sub-linear regret bounds commonly derived in regret minimization analysis. Here we allow the offline malicious expert to choose his policy *over the entire horizon (i.e., an arbitrary sequence of false/true predictions) and do not restrict him to his best* action *(i.e., only select one action and commit to it at all the stages). Interestingly, Theorem 1 shows that giving such an extra power to the malicious expert does not give him much advantage other than a negligible additive term.*

According to Theorem 1, the MW learning algorithm with an absolute loss function is not very robust against the malicious expert. The reason is that a naive malicious expert who follows a simple false policy can be nearly as harmful as any other highly strategic malicious expert. However, we should emphasize that such a conclusion is only valid under the adversarial model that we consider in this paper, i.e., when the adversary has access to the true outcome before making its decision at each stage. In fact, it would be interesting to evaluate the performance of the MW learning algorithm for less powerful adversarial models where the adversary has only partial information about the true outcomes.

A. Beyond Asymptotic Optimality for the Offline Policy

Theorem 1 shows that the false policy asymptotically achieves the same performance as the optimal offline policy. However, the exact structure of the optimal offline policy for a finite horizon N can be quite complex. Therefore, our goal in this section is to take one step further and provide a policy that closely resembles the structural patterns of the optimal offline policy. As it was shown in the proof of Theorem 1 one of the main reasons that there is a gap between the expected loss of the optimal offline policy and that of the false policy is the term:

$$B := \sum_{\ell=1}^{k}\mathbb{E}\left[\frac{1}{1+e^{X_\ell-Y_\ell}} - \frac{1}{1+e^{X_\ell-Y_{\ell-1}}}\right]$$

henceforth referred to as the ***bonus*** term. Here, $X_\ell \sim \mathrm{Bin}(N_\ell, \mu)$ and $Y_\ell \sim \mathrm{Bin}(M_\ell, 1-\mu)$ are independent Binomial distributions where $N_\ell = \sum_{i=1}^{\ell} n_i$ and $M_\ell = \sum_{i=1}^{\ell} m_i$. Therefore, it seems reasonable to expect the optimal offline policy (or a policy close to optimal), to maximize B in order to gain as much as possible from the bonus term. As such, we search the optimal offline policy Ψ^* among policies Ψ that satisfy the following two criteria:

- i) Ψ imposes at least as much loss as the false policy on the learning system, i.e., at least $(1-\mu)N - o(1)$,
- ii) Ψ maximizes the bonus gain B.

To maximize the bonus term, using Lemma 8 (Appendix A), it is enough to maximize

$$\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big], \qquad (15)$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and

$$\begin{aligned}
\mu_{2\ell} &= N_\ell\mu - M_\ell(1-\mu), & \sigma^2_{2\ell} &= \mu(1-\mu)(N_\ell+M_\ell), \\
\mu_{2\ell-1} &= N_\ell\mu - M_{\ell-1}(1-\mu), & \sigma^2_{2\ell-1} &= \mu(1-\mu)(N_\ell+M_{\ell-1}).
\end{aligned}$$

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

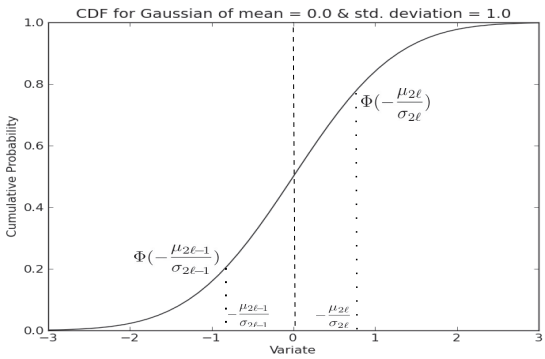

Fig. 1. Mean and standard deviation adjustment for maximizing the bonus term B using CDF of the standard normal distribution.

Now by adjusting the argument of $\Phi(\cdot)$ in (15) to periodically switch around 0 (see Figure 1), we obtain a positive gain from each of the summands in (15). This suggests that a policy in which $-\frac{\mu_{2\ell}}{\sigma_{2\ell}} = 1$, and $-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}} = -1$, would be a good candidate for maximizing the bonus term. Note that here the choice of 1 or -1 is not strict and it can be replaced by any two points close to zero such that the difference of the normal CDF evaluated at those points gives a sufficiently large gain. Solving $-\frac{\mu_{2\ell}}{\sigma_{2\ell}} = 1$ and $-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}} = -1$, by substituting the above expressions for $\mu_{2\ell}, \sigma_{2\ell}, \sigma_{2\ell}, \sigma_{2\ell-1}$, we obtain a ratio type policy in which the ratio of the false/true block lengths $\frac{n_\ell}{m_\ell}, \ell = 1, 2, \ldots$, is proportional to $\frac{1-\mu}{\mu}$. Therefore, to fulfill both criteria (i) and (ii), we introduce the following offline ***ratio*** policy:

Definition 3: Let a and b be the smallest positive integers such that $\frac{a}{b} = \frac{\mu}{1-\mu}$.[2] *We say that $\hat{\Psi}$ is a ratio policy if its block representation is of the form $(n_1, m_1, \ldots, m_{k-1}, n_k, m_k) = (b, a, b, a, b, \ldots, a, \frac{N}{2}, 0)$, where for $\ell = 1, 2, \ldots, k$, the adversary lies in n_ℓ-blocks and tells the truth in m_ℓ-blocks.*

To verify why the ratio policy is indeed a good offline policy, let us denote the number of lies and truths in $\hat{\Psi}$ by M and $N - M$, respectively. Due to Definition 3, $\hat{\Psi}$ has many more lies than the truth in its structure (note that $n_k = \frac{N}{2}$). Thus a similar analysis as in Lemma 2 for the false policy reveals that the expected loss in $\hat{\Psi}$ which is obtained from its last block $n_k = \frac{N}{2}$ is almost the same as the false policy minus a negligible constant which does not depend on N. As a result, the expected loss of $\hat{\Psi}$ which is obtained due to its heavy tail of false predictions, is at least $(1-\mu)N - o(1)$. On the other hand, the ratio policy $\hat{\Psi}$ gains a bonus due to its first $\frac{N}{2}$ stages. To evaluate the bonus term B for the ration policy $\hat{\Psi}$, we observe that due to Definition 3, $\mu_{2\ell} = 0$, and $\mu_{2\ell-1} = \mu b$, for all $\ell \in [k]$. Thus, for each $\ell \in [k]$, we have

$$\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}}) = \Phi(0) - \Phi(-\frac{\mu b}{\sigma_{2\ell-1}}) > 0.$$

In other words, each of these terms contributes positively to some constant amount to the bonus B. Since we chose a and b as the smallest positive integers such that $\frac{a}{b} = \frac{\mu}{1-\mu}$, this assures that the number of summands k in the bonus term B (which equals the number of switching between false/true blocks) is maximized. That is why we defined our ratio policy the way we did in Definition 3. This shows that the expected loss of the ratio policy is at least as high as that for the false policy (satisfying criterion (i)) with an additional bonus term B (satisfying criterion (ii)).

[2]Here for simplicity we have assumed that $\frac{\mu}{1-\mu}$ is a rational number, otherwise, we can always find positive integers such that $\frac{a}{b} \cong \frac{\mu}{1-\mu}$.

IV. Optimal Online Policy for the Absolute Loss

In this section, we consider the problem of finding the optimal online policy for the malicious expert, where we recall that the online adversary is the one who chooses his next action adaptively based on all the past revealed information up to the current stage. In order to be able to find the optimal online policy, we first cast it as a dynamic program and then show that it can be solved efficiently in $O(N^3)$.

For this purpose, let us assume again that the malicious expert is expert 1 and the other expert is the honest one who makes a correct prediction with probability μ. We assume that at stage k, expert 1 knows the true outcome y_k, the accuracy μ of the honest expert, and the entire history of predictions up to stage $k-1$, i.e., $\{\tilde{p}_\ell^1, \tilde{p}_\ell^2, x_\ell^1, x_\ell^2, y_\ell : \forall \ell \in [k-1]\}$. Given this information set, the goal of the online malicious expert is to produce a sequence of predictions $\{x_k^1\}_{k=1}^N$ over a fixed finite horizon N to maximize the expected accumulated loss of the system given by (4). Now let us define the *state* of the system at stage k to be the relative weight of the adversary at that stage, i.e., $\tilde{p}_k^1$. Note that as $\tilde{p}_k^1 + \tilde{p}_k^2 = 1, \forall k$, knowing $\tilde{p}_k^1$ is sufficient to determine the relative weight of the honest expert $\tilde{p}_k^2$.

Next let us define $c_{x_k^1}(\tilde{p}_k^1)$ to be the *current* loss that the online adversary can impose on the system at stage k by taking the action x_k^1, i.e.,

$$\begin{aligned} c_{x_k^1}(\tilde{p}_k^1) &= \mathbb{E}_{x_k^2}[|\hat{y}_k - y_k||x_1^1] \\ &= \begin{cases} 1-\mu+\mu\,\tilde{p}_k^1 & \text{if } x_k^1 = 1-y_k, \\ (1-\mu)(1-\tilde{p}_k^1) & \text{if } x_k^1 = y_k, \end{cases} \end{aligned} \tag{16}$$

where the second equality is by Lemma 1 specialized to the absolute loss function $Q(y) = y$. We can then cast the adversary's online optimal policy as a solution to an MDP in which the malicious expert's action at stage k imposes a current loss of $c_{x_k^1}(\tilde{p}_k^1)$ on the system and changes the state from $\tilde{p}_k^1$ to the next state $\tilde{p}_{k+1}^1$. In particular, the state transition of this MDP is given by the update rule (5), that is,

$$\tilde{p}_{k+1}^1 = \begin{cases} g(\tilde{p}_k^1) & \text{if } x_k^1 = 1-y_k, x_k^2 = y_k, \\ g^{(-1)}(\tilde{p}_k^1) & \text{if } x_k^1 = y_k, x_k^2 = 1-y_k, \\ \tilde{p}_k^1 & \text{if } x_k^1 = x_k^2. \end{cases} \tag{17}$$

Now the solution to this MDP can be obtained using dynamic programming, as shown in Algorithm 1. In this algorithm $V_{k+1}^*(\cdot)$ denotes the optimal value function, i.e., the optimally accumulated loss from time step $k+1$ onward. In particular, from Lemma 1, one can easily see that the optimal value function does not depend on the sequence of true outcomes and is only a function of the state and the number of remaining stages. Now by substituting the closed-form expressions of the current cost (16) and the state transition (17)

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 DP Algorithm

Initialize: $V_N^*(\cdot) = c_N(\cdot) = 0$.

For each step $k = N-1$ downto 0, find the optimal action

$$x_k^* := \arg\max_{x_k^1} \left\{ c_{x_k^1}(\tilde{p}_k^1) + \mathbb{E}[V_{k+1}^*(\tilde{p}_{k+1}^1)] \right\},$$

and the optimal value function,

$$V_k^*(\tilde{p}_k^1) = \max_{x_k^1} \left\{ c_{x_k^1}(\tilde{p}_k^1) + \mathbb{E}[V_{k+1}^*(\tilde{p}_{k+1}^1)] \right\}. \tag{19}$$

Output: sequence $x_{N-1}^*, V_{N-1}^*(\cdot), \ldots, x_0^*, V_0^*(\cdot)$.

into the DP Algorithm 1, and letting $\rho := \tilde{p}_k^1$ for brevity, we obtain the following closed-form recursion for computing the optimal value function:

$$V_k^*(\rho) = \max \Big\{ 1-\mu+\mu\rho+\mu V_{k+1}^*(g(\rho)) + (1-\mu)V_{k+1}^*(\rho), \\ (1-\mu)(1-\rho) + (1-\mu)V^*(g^{(-1)}(\rho)) + \mu V_{k+1}^*(\rho) \Big\}, \tag{18}$$

where the first term in the maximization (18) corresponds to the adversary's action at stage k being $x_k^1 = 1 - y_k$, and the second term corresponds to the adversary's action being $x_k^1 = y_k$. Unfortunately, due to the nonlinear structure of the transition functions $g(\rho)$ and $g^{(-1)}(\rho)$, as well as their joint convex/concave structure, solving the recursion (18) in a closed-form, seems to be a tedious task. Although at each stage of the above recursion, one needs to consider the maximum of two alternatives (so that the number of alternatives will grow exponentially in terms of the number of stages), however, in the following theorem, we show that most of these alternatives collapse on each other so that the optimal value function in (18) can be computed efficiently in polynomial time.

Theorem 2: The optimal policy for the online malicious expert can be found in $O(N^3)$, where N is the number of stages.

Proof: Let us consider a decision tree with a root node representing the initial relative weight of the adversary (i.e., $\rho = 0.5$) and such that the nodes in the k-th level of the tree that are at distance k from the root represent all the possible relative weights of the adversary after k stages (Figure 2). The key observation is that due to the property of $g(\cdot)$ and its inverse $g^{(-1)}(\cdot)$, the size of this decision tree does not grow exponentially such as a binary tree. In fact, a simple induction shows that the nodes in the k-th level of the tree can be grouped to form exactly $2k-1$ nodes representing all possible relative weights of the adversary up to stage k, given by $g^{(-k)}(0.5), g^{(-k+1)}(0.5), \ldots, g^{(k-1)}(0.5), g^{(k)}(0.5)$.[3] Therefore, the total number of tree nodes by such grouping (states in the DP) after N stages is at most $\sum_{k=1}^{N}(2k-1) = O(N^2)$. As a result, solving the dynamic recursion (18) backward by moving from the tree leaves toward the root, the number of computations to find the optimal online policy using the DP recursion (18) is at most $O(N \times N^2)$. □

[3]Using Lemma 3, one can even compute the distribution of the weights on the reduced nodes efficiently using Binomial distributions.

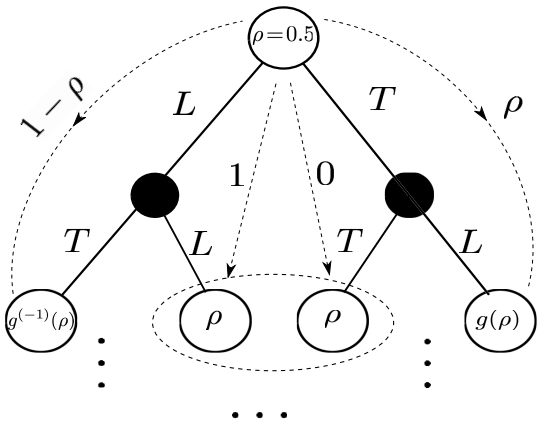

Fig. 2. Illustration of the first level (root) and the second level of the decision tree. The top actions connecting the root to the intermediate black circles correspond to the honest expert's decisions. The bottom actions connecting the black circles to the second level of the tree correspond to the decisions of the malicious expert. Although the second level originally has four nodes, two can be grouped and be reduced to only three states. The weights on the dashed paths denote the loss of the system by following that path.

In fact, in a recent work [21], the authors have analyzed the DP (18) in more detail by providing upper and lower bounds on the optimal value function using the approximated dynamic program's viscosity solution. More precisely, it was shown in [21] that for any online policy Ψ for the malicious expert:

$$\limsup_{N\to\infty} \frac{V_0^*(0.5)}{N} \le 1-\mu^2, \quad \liminf_{N\to\infty} \frac{V_0^*(0.5)}{N} > 1-\mu.$$

A. A Generalization to Multiple Experts

Here we provide a generalization of the problem to the case of many honest experts and one adversary. Without loss of generality, we again assume that the malicious expert is expert 1 and that all the other experts $i \in \{2, \ldots, K\}$ are honest who make a correct prediction with different probabilities μ_i (the accuracy of expert i). That is,

$$x_k^i = \begin{cases} y_k & \text{w.p.} \quad \mu_i, \\ 1-y_k & \text{w.p.} \quad 1-\mu_i. \end{cases}$$

We assume that at round k, expert 1 knows the true outcome y_k, the accuracy of the honest experts, and the whole history of predictions up to round $k-1$, i.e., $\{\tilde{p}_\ell^j, x_\ell^j, y_\ell : \forall \ell \in [k-1], j \in [K]\}$. Given this information set, the goal of the online malicious expert is to produce a sequence of predictions $\{x_k^1\}_{k=1}^N$ over a fixed finite horizon N in order to maximize the expected accumulated loss of the system,

$$\max_{x_1^1,\ldots,x_K^1} \mathbb{E}_{x_k^i, k\in[N], i\neq 1}\Big[\sum_{k=1}^{N} l(\hat{y}_k, y_k)\Big],$$

where the expectation is taken over the randomization of all the honest experts' predictions $\{x_k^i, k \in [N], i \neq 1\}$. We let $\vec{p}_k = (p_k^1, p_k^2, \ldots, p_k^K)$ be the *state* or *weight* vector of all experts at round k and $\vec{\tilde{p}}_k = (\tilde{p}_k^1, \ldots, \tilde{p}_k^K)$ be the corresponding normalized weight vector where $\tilde{p}_k^i = \frac{p_k^i}{\sum_{i\in[K]} p_k^i}$, $i \in [K]$

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

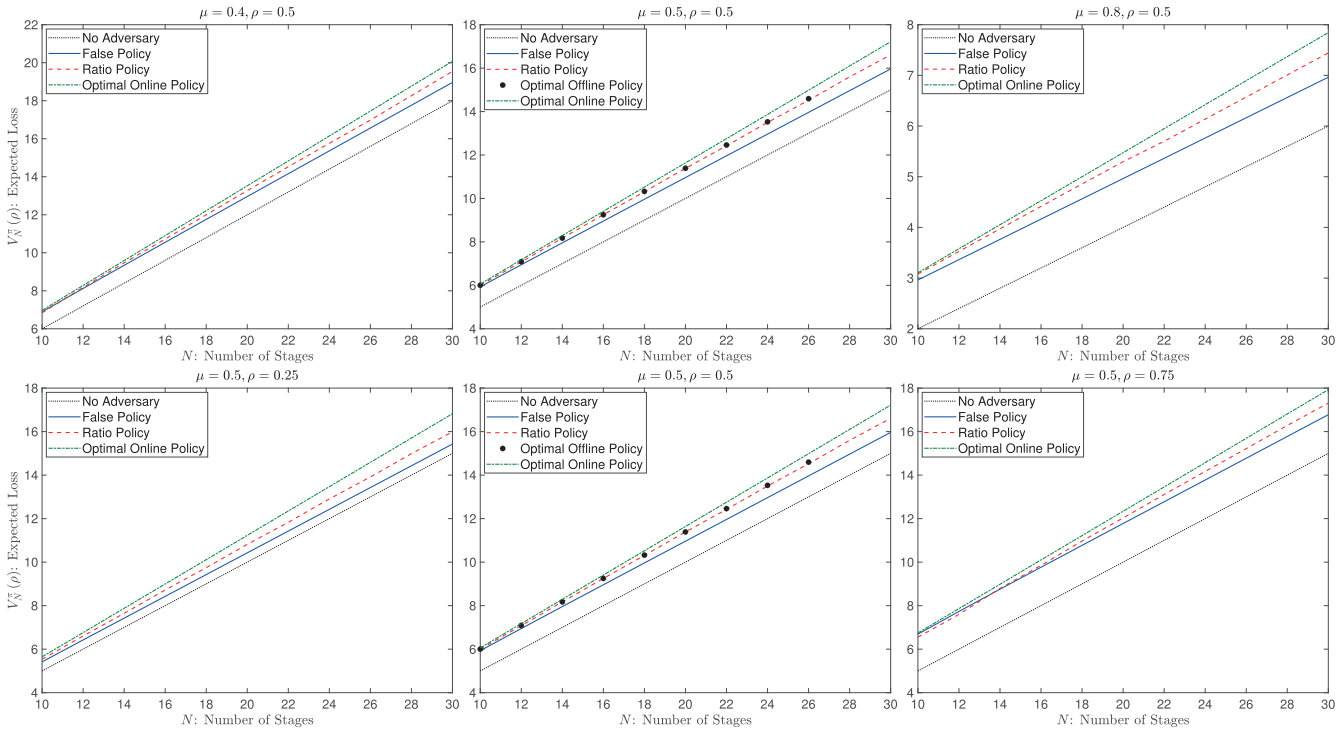

Fig. 3. Performance comparison of the false, ratio and optimal online policies for different accuracies and initial relative weights. In all figures we set $\epsilon = \frac{1}{e}$.

is the normalized weight of expert i. Note that we always have $\sum_{i\in[K]} \tilde{p}_k^i = 1, \forall k$. Moreover, we let $\phi_{x_k^1}(\vec{p}_k)$ be the state transition at stage k given that the online malicious expert takes the action $x_k^1 \in \{0, 1\}$, i.e.,

$$\phi_{x_k^1}(\vec{p}_k) = \left(p_{k+1}^1, \ldots, p_{k+1}^K\right),$$

where $p_{k+1}^i = p_k^i \epsilon^{l(x_k^i, y_k)}$. In addition, the learning algorithm predicts $\hat{y}_k$ at time k by

$$\hat{y}_k = \frac{\sum_{i\in[K]} p_k^i x_k^i}{\sum_{i\in[K]} p_k^i} = \sum_{i\in[K]} \tilde{p}_k^i x_k^i.$$

We note that for $K = 2$, this generalized setting coincides with that given in Section II. Although characterizing the optimal online policy structure in the generalized setting can be very complicated, in the next section, we provide some numerical experiments to study the behavior of the optimal policy with multiple honest experts.

V. Simulations

Performance of the false, ratio, and optimal online policies has been simulated numerically and compared in Figure 3. In addition to offline and online policies, the case of no adversary with two identical honest experts has also been simulated to show the effect of an adversary in the system.

As seen in all plots, an adversary simply adopting the false policy incurs extra loss compared to a system where a malicious expert is not present. The ratio policy imposes more loss than the false policy when the number of stages is large. The optimal online policy imposes a strictly greater loss than the optimal offline policy. Moreover, as the number of stages increases, the gap between the loss of the optimal online policy and offline policies also increases. Similarly, the gain of the bonus term B, which is the difference between the curves of the false policy and the ratio policy, increases as the number of stages increases. It can be seen that the ratio policy closely mimics the structure of the optimal offline policy. For instance, in the middle subfigures of Figure 3, the optimal offline expected loss for several values of $N = 10, 12, 14, 16, 18, 20, 22, 24, 26$ are plotted using black dots. As these values are very close to the expected loss of the ratio policy and even coincide in certain cases (e.g., $N = 10, 14, 16$), we believe that the optimal offline policy for finite N belongs to the class of ratio policies, given that one could properly round the block lengths using the CDF of normal distribution.

Finally, using the generalization to multiple experts in Section IV, a system with four honest experts and one adversary is simulated and compared to a system with one honest expert and one malicious expert. In the 5-expert model, all experts have identical initial weights. In the 2-expert model, the adversary's initial relative weight is $\rho = 0.2$, which is the same in the 5-expert model. The accuracy of the honest expert in the 2-expert model is the mean of the four honest experts' accuracy in the 5-expert model. Two cases are considered: the homogeneous case (accuracies of all honest experts are identical, $\mu_2 = \mu_3 = \mu_2 = \mu_5 = 0.5$) and the heterogeneous case (accuracies of honest experts are distinct, $\mu_2 = 0.3$, $\mu_3 = 0.4$, $\mu_4 = 0.6$, $\mu_5 = 0.7$). The mean of the accuracies of honest experts is the same for the two cases. The expected

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

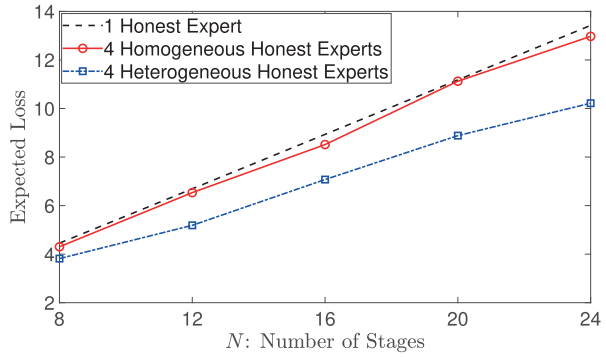

Fig. 4. Comparison of the online policies for 2-expert and 5-expert models. In the 2-expert model, $\mu = 0.5$, $\rho = 0.2$. In the homogeneous 5-expert model, $\mu_2 = \mu_3 = \mu_2 = \mu_5 = 0.5$, $\rho = 0.2$; in the heterogeneous 5-expert model, $\mu_2 = 0.3$, $\mu_3 = 0.4$, $\mu_4 = 0.6$, $\mu_5 = 0.7$, $\rho = 0.2$. $\epsilon = \frac{1}{e}$.

loss for the 5-expert model is estimated as follows. In each play, a sequence of actions $\{0, 1\}^N$ is randomly generated for each honest expert $i \in \{2, 3, 4, 5\}$ according to his accuracy μ_i, and the adversary chooses his optimal policy against the honest experts' strategies. This process is repeated 100 times, and the expected loss is approximated by the empirical mean of the losses for all the 100 plays.

Numerical results are shown in Figure 4. As the curve for the homogeneous 5-expert model is very close to that for the 2-expert model, it suggests that the 2-expert model can well approximate the system with multiple homogeneous honest experts by replacing all honest experts with a single one of combined relative weight and the same accuracy. The difference between the curves for the 2-expert model and the heterogeneous 5-expert model is greater, probably because the optimal online policy in the generalized heterogeneous setting is difficult to be approximated by only 2-experts.

VI. Conclusion and Future Directions

In this paper, we considered an adversarial learning system with two experts, of whom one is malicious. The malicious expert aims to impose the maximum loss on the system by strategically reporting false predictions. We analyzed the optimal policy for the malicious expert under both offline and online settings. In the offline setting, we showed that finding the adversary's optimal policy is a discrete optimization problem whose solution can be approximated within a negligible (sub-linear) additive term. In particular, we provided a more refined policy that closely mimics the optimal offline policy's behavior. We then considered the optimal online policy for the malicious expert and showed that it could be computed efficiently for two experts using a dynamic program. We also generalized the online setting to multiple experts.

This work opens many exciting directions for future research. It would be interesting to see whether the optimal policy structure for the online adversary can be characterized in a closed-form. One possible direction is to leverage the dynamic recursion (18) to show that the optimal value function possesses some nice properties such as convexity, which allows one to prove a class of optimal threshold policies for the online adversary [22]. Another interesting direction is to use learning schemes other than the MW algorithm (e.g., upper confidence bound algorithm) as the underlying learning scheme and study their robustness against adversarial attacks. When many experts in the system, can we use a mean-field approximation to approximate the optimal policies? Finally, it is interesting to study the game-theoretic version of this work in which the MW learning system can be strategic and not only penalizes the malicious expert but also detects and eliminates it from the system.

Appendix A
Auxiliary Lemmas

Proposition 1: The optimal online strategy for an adversary with no information about an arbitrary sequence of true outcomes $\{y_k\}$ *is to choose* $x_k^1 \in \{0, 1\}$ *with probability* $\frac{1}{2}$ *and independently for every* $k \in [N]$.

Proof: Let us fix an arbitrary stage k. The expected loss incurred in stage k is the conditional expectation given the entire history of outcomes and predictions up to stage $k-1$ taken over the past and current actions of the honest expert:

$$\mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)] \equiv \mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)|\{\tilde{p}_l^1, x_l^1, x_l^2, y_l\}_{l=1}^{k-1}].$$

As y_k can be chosen arbitrarily, and the honest expert's prediction at stage k is independent of the previous stages, the history of predictions up to stage $k-1$ cannot give any information to the adversary about y_k. As a result, we have

$$\mathbb{E}_{x_1^2,\ldots,x_k^2}[l(\hat{y}_k, y_k)] = \mathbb{E}_{x_k^2}[l(\hat{y}_k, y_k)]. \tag{20}$$

Therefore, in this case, the adversary becomes memoryless and treats every stage as a new restart. Now for the absolute loss function $l(\hat{y}, y) = |\hat{y} - y|$, one can compute the expected loss in a closed form for stage k using (20). Let us assume that the adversary chooses $x_k^1 = 0$ with probability q and chooses $x_k^1 = 1$ with probability $1-q$. Depending on the true outcome y_k, the expected loss equals one of the following terms:

$$\mathbb{E}_{x_k^2}[l(\hat{y}_k, y_k)|y_k = 0] = 1 - \mu + \mu\rho - q\rho,$$
$$\mathbb{E}_{x_k^2}[l(\hat{y}_k, y_k)|y_k = 1] = 1 - \mu + \mu\rho - (1-q)\rho,$$

where ρ denotes the relative weight of the adversary at the beginning of stage k. Since the adversary has no information about whether $y_k = 0$ or $y_k = 1$, it must choose q to maximize the minimum of the above two expressions. For $q = \frac{1}{2}$, the above equations coincide, which shows that predicting with probability $\frac{1}{2}$ at each stage is the optimal online strategy. □

Lemma 4: Let $f(r) := r - \ln(1 + ae^r)$, $h(r) := \ln(a + e^r)$,

$$\epsilon(r) := f(r+1) - f(r) - g^{(r)}(\rho),$$
$$\delta(r) := h(r+1) - h(r) - g^{(-r)}(\rho).$$

Then for any $r \geq 0$ *we have,*

$$0 \geq \epsilon(r) \geq \frac{1}{1 + ae^{r+1}} - \frac{1}{1 + ae^r},$$
$$0 \leq \delta(r) \leq \frac{1}{1 + ae^{-(r+1)}} - \frac{1}{1 + ae^{-r}},$$

where we recall that $a = \frac{1}{\rho} - 1$, *for some* $\rho \in (0, 1)$.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

Proof: Since $\frac{d}{dr}\epsilon(r) = \frac{ae^r(ae^r-e+2)}{(1+ae^r)^2(1+ae^{r+1})}$ has only one root given by $e^r = \frac{e-2}{a}$, by evaluating $\epsilon(r)$ in the root of its derivative as well as the boundary of its domain we get,

$$\epsilon(r) = \begin{cases} \frac{e-2}{e-1} - \ln(e-1) < 0 & \text{if } e^r = \frac{e-2}{a}, \\ 1 + \ln(\frac{1+a}{1+ae}) - \frac{1}{a+1} \le 0 & \text{if, } r = 0, \\ 0 & \text{if } \overset{a=0, \text{ or}}{\underset{a\to\infty}{r\to\infty,}}. \end{cases}$$

This shows that $\epsilon(r) \le 0,\ \forall r, a \in [0,\infty)$. On the other hand, for every $r > 0$, using the Mean-value Theorem we have $f(r+1) - f(r) = f'(\eta_r)$, for some $\eta_r \in [r, r+1]$. Since $f'(\eta_r) = \frac{1}{1+ae^{\eta_r}} > \frac{1}{1+ae^{r+1}}$, using (7) we can write,

$$\begin{aligned} \epsilon(r) &= f(r+1) - f(r) - g^{(r)}(\rho) \\ &\ge \frac{1}{1+ae^{r+1}} - g^{(r)}(\rho) \\ &= \frac{1}{1+ae^{r+1}} - \frac{1}{1+ae^r}. \end{aligned}$$

Similarly, $\frac{d}{dr}\delta(r) = -\frac{ae^r(e^r-(e-2)a)}{(a+e^r)^2(a+e^{r+1})}$, which has only one root at $e^r = (e-2)a$. Therefore, by evaluating $\delta(r)$ in the root of its derivative as well as the boundary points one can easily see that $\delta(r) \ge 0$. Again, using the Mean-value Theorem, there exists $\zeta_r \in [r, r+1]$ such that $h(r+1) - h(r) = h'(\zeta_r) = \frac{1}{1+ae^{-\zeta_r}} \le \frac{1}{1+ae^{-(r+1)}}$. This shows that,

$$\begin{aligned} \delta(r) &\le \frac{1}{1+ae^{-(r+1)}} - g^{(-r)}(\rho) \\ &= \frac{1}{1+ae^{-(r+1)}} - \frac{1}{1+ae^{-r}}. \end{aligned}$$

□

Lemma 5: Let $m_1, m_2, \ldots, m_k$ be positive integers and define $M_\ell := \sum_{j=1}^{\ell} m_j,\ \ell \in [k]$, (by convention we let $M_0 = 0$). Then $\sum_{\ell=1}^{k} \frac{m_\ell}{\sqrt{M_\ell}} = O(\sqrt{M_k \ln M_k})$.

Proof: Starting from the left-hand side and using Cauchy-Schwarz inequality, we have,

$$\sum_{\ell=1}^{k} \frac{m_\ell}{\sqrt{M_\ell}} = \sum_{\ell=1}^{k} \sqrt{m_\ell} \times \sqrt{\tfrac{m_\ell}{M_\ell}} \le \sqrt{M_k}\sqrt{\sum_{\ell=1}^{k} \tfrac{m_\ell}{M_\ell}}. \quad (21)$$

Next for every ℓ, we can write

$$\begin{aligned} \frac{m_\ell}{M_\ell} &= \underbrace{\frac{1}{m_\ell + M_{\ell-1}} + \frac{1}{m_\ell + M_{\ell-1}} + \ldots + \frac{1}{m_\ell + M_{\ell-1}}}_{m_\ell \text{ times}} \\ &\le \frac{1}{1+M_{\ell-1}} + \frac{1}{2+M_{\ell-1}} + \ldots + \frac{1}{m_\ell + M_{\ell-1}}. \end{aligned}$$

Summing the above relation for all $\ell = 1, \ldots, k$, we get

$$\begin{aligned} \sum_{\ell=1}^{k} \frac{m_\ell}{M_\ell} &\le \sum_{\ell=1}^{k} \frac{1}{1+M_{\ell-1}} + \frac{1}{2+M_{\ell-1}} + \ldots + \frac{1}{m_\ell + M_{\ell-1}} \\ &= \sum_{j=1}^{M_k} \frac{1}{j} \le 1 + \ln M_k. \end{aligned}$$

Using this relation into (21) we get the desired bound. □

Lemma 6 **Berry-Esseen Theorem [23]:** *Let V_i be independent random variables with mean a_i and variance s_i^2, and define $S_t := \sum_{i=1}^{t} V_i$. Then there exists an absolute constant c_0 such that for all t the CDF of S_t, denoted by $F_t(x)$, satisfies*

$$\sup_x \left| F_t(x) - \Phi\Big(\frac{x - \sum_{i=1}^{t} a_i}{\sqrt{\sum_{i=1}^{t} s_i^2}}\Big) \right| \le \frac{c_0 \max_i \{\frac{\mathbb{E}|X_i - a_i|^3}{s_i^2}\}}{\sqrt{\sum_{i=1}^{t} s_i^2}}.$$

Lemma 7: Let $X \sim \text{Bin}(n, \mu)$ and $Y \sim \text{Bin}(m, (1-\mu))$ be two independent Binomial distributions. Then, there exists a constant c such that $\left| \mathbb{E}\big[\frac{1}{1+e^{X-Y}}\big] - \Phi(-\frac{\nu}{\sigma}) \right| \le \frac{c}{\sigma}$, where $\nu = n\mu - (1-\mu)m$, $\sigma^2 = \mu(1-\mu)(n+m)$, and $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Proof: Let $p(\gamma)$ and $F(\gamma)$ denote the pmf and CDF of the random variable $X - Y$, respectively. Then,

$$\begin{aligned} \left| F(0) - \mathbb{E}[\tfrac{1}{1+e^{X-Y}}] \right| &= \left| \mathbb{P}(X - Y \le 0) - \mathbb{E}[\tfrac{1}{1+e^{X-Y}}] \right| \\ &= \Big| \sum_{i=-\infty}^{0} p(i) - \sum_{i=-\infty}^{\infty} \frac{p(i)}{1+e^i} \Big| \\ &= \Big| \sum_{i=0}^{\infty} \frac{p(-i)}{1+e^i} - \sum_{i=1}^{\infty} \frac{p(i)}{1+e^i} \Big| \\ &\le \sum_{i=0}^{\infty} \frac{p(-i)}{1+e^i} + \sum_{i=0}^{\infty} \frac{p(i)}{1+e^i} \\ &\le \sum_{i=0}^{\infty} p(-i)e^{-i} + \sum_{i=0}^{\infty} p(i)e^{-i} \\ &\le 2p_{\max} \sum_{i=0}^{\infty} e^{-i} = \frac{2e}{e-1} p_{\max}, \quad (22) \end{aligned}$$

where $p_{\max} = \max_i\{p(i)\}$. Next, using Berry-Esseen Theorem (Lemma 6), and noting that X and Y can be written as sum of n and m independent Bernoulli random variables $\text{Ber}(\mu)$ and $\text{Ber}(1-\mu)$, respectively, we get

$$\sup_\gamma \left| F(\gamma) - \Phi\Big(\frac{\gamma - \nu}{\sigma}\Big) \right| \le \frac{c_0(\mu^2 + (1-\mu)^2)}{\sigma}. \quad (23)$$

Now for every i we can write,

$$\begin{aligned} p(i) &= F(i) - F(i-1) \\ &\le \Phi(\tfrac{i-\nu}{\sigma}) - \Phi(\tfrac{i-1-\nu}{\sigma}) + \frac{2c_0(\mu^2 + (1-\mu)^2)}{\sigma} \\ &\le \Phi'(\eta) \times \left(\frac{i-\nu}{\sigma} - \frac{i-1-\nu}{\sigma} \right) + \frac{2c_0(\mu^2 + (1-\mu)^2)}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{\eta^2}{2}} \times \frac{1}{\sigma} + \frac{2c_0(\mu^2 + (1-\mu)^2)}{\sigma} \\ &\le \frac{1}{\sqrt{2\pi}\,\sigma} + \frac{2c_0(\mu^2 + (1-\mu)^2)}{\sigma} = \frac{c_1}{\sigma} \quad (24) \end{aligned}$$

where $c_1 := \frac{1}{\sqrt{2\pi}} + 2c_0(\mu^2 + (1-\mu)^2)$. The first inequality is due to (23) and the second inequality is by Mean-value Theorem for some $\eta \in [\frac{i-1-\nu}{\sigma}, \frac{i-\nu}{\sigma}]$. As a result, $p_{\max} \le \frac{c_1}{\sigma}$. Substituting (24) into (22), we have

$$\left| F(0) - \mathbb{E}[\frac{1}{1+e^{X-Y}}] \right| \le \frac{2ec_1}{(e-1)\sigma}. \quad (25)$$

Finally, adding (25) with (23) when $\gamma = 0$, and using the triangle inequality, we get

$$\left| \mathbb{E}[\tfrac{1}{1+e^{X-Y}}] - \Phi(-\tfrac{\nu}{\sigma}) \right| \le \frac{2ec_1}{(e-1)\sigma} + \frac{c_0(\mu^2 + (1-\mu)^2)}{\sigma} = \frac{c}{\sigma},$$

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.

where $c := \frac{2ec_1}{e-1} + c_0(\mu^2 + (1-\mu)^2)$ is a positive constant. □

Lemma 8: Let $X_\ell \sim \text{Bin}(N_\ell, \mu)$, $Y_\ell \sim \text{Bin}(M_\ell, 1-\mu)$, $\ell \in [k]$, be mutually independent (i.e., for every $i \neq j$, X_i and Y_j are independent). Moreover, assume $N_\ell = \sum_{i=1}^{\ell} n_i$ and $M_\ell = \sum_{i=1}^{\ell} m_i$, where $n_i, m_i \in \mathbb{Z}^+$, and $N_k + M_k = N$. Then $B := \sum_{\ell=1}^{k} \mathbb{E}\Big[\frac{1}{1+e^{X_\ell - Y_\ell}} - \frac{1}{1+e^{X_\ell - Y_{\ell-1}}}\Big] = O(\sqrt{N \ln N})$.

Proof: For any $\ell = 1, \ldots, k$, define

$$\begin{aligned} \mu_{2\ell} &:= N_\ell \mu - M_\ell(1-\mu), \qquad \sigma_{2\ell}^2 := \mu(1-\mu)(N_\ell + M_\ell), \\ \mu_{2\ell-1} &:= N_\ell \mu - M_{\ell-1}(1-\mu), \quad \sigma_{2\ell-1}^2 := \mu(1-\mu)(N_\ell + M_{\ell-1}), \end{aligned} \tag{26}$$

where $M_\ell = \sum_{i=1}^{\ell} m_i$ and $N_\ell = \sum_{i=1}^{\ell} n_i$ (recall that m_i and n_i denote, respectively, the lengths of the ith true and false blocks in an offline policy Ψ). Using Lemma 7 we can write

$$\begin{aligned} B &\overset{(a)}{\leq} 2\sum_{\ell=1}^{k}\left[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}}) + c\left(\frac{1}{\sigma_{2\ell}} + \frac{1}{\sigma_{2\ell-1}}\right)\right] \\ &\overset{(b)}{\leq} 2\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big] \\ &\quad + \frac{2c}{\sqrt{\mu(1-\mu)}}\sum_{\ell=1}^{k}\left(\frac{1}{\sqrt{2\ell}} + \frac{1}{\sqrt{2\ell-1}}\right) \\ &\leq 2\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big] + \frac{2c}{\sqrt{\mu(1-\mu)}}\int_0^{2k}\frac{1}{\sqrt{x}}dx \\ &\overset{(c)}{\leq} 2\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big] + 4c\sqrt{\frac{N}{\mu(1-\mu)}}, \end{aligned} \tag{27}$$

where (a) is due to Lemma 7, and (b) holds because $N_\ell + M_\ell = \sum_{i=1}^{\ell}(n_i + m_i)$ is the sum of 2ℓ positive integers, and thus $\sigma_{2\ell} \geq \sqrt{2\mu(1-\mu)\ell}$ (similarly $\sigma_{2\ell-1} \geq \sqrt{\mu(1-\mu)(2\ell-1)}$). Finally (c) holds because $2k \leq N$.

We proceed by showing a sub-linear upper bound on $\sum_{\ell=1}^{k}\big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\big]$. Let β be a constant defined by $\beta := \frac{1}{\sqrt{2\pi\,\mu(1-\mu)}}$. Using the Mean-Value Theorem and since $\Phi'(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \leq \frac{1}{\sqrt{2\pi}}, \forall x$, we can write,

$$\begin{aligned} &\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big] \\ &\quad\leq \sum_{\ell=1}^{k}\frac{1}{\sqrt{2\pi}}\left(\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}} - \frac{\mu_{2\ell}}{\sigma_{2\ell}}\right) \\ &\quad= \beta\sum_{\ell=1}^{k}\left(\frac{N_\ell\mu - M_{\ell-1}(1-\mu)}{\sqrt{N_\ell + M_{\ell-1}}} - \frac{N_\ell\mu - M_\ell(1-\mu)}{\sqrt{N_\ell + M_\ell}}\right) \\ &\quad\leq \beta\sum_{\ell=1}^{k}\left(\frac{-M_{\ell-1}}{\sqrt{N_\ell + M_{\ell-1}}} + \frac{M_\ell}{\sqrt{N_\ell + M_\ell}}\right) \\ &\quad\leq \beta\sum_{\ell=1}^{k}\frac{M_\ell - M_{\ell-1}}{\sqrt{N_\ell + M_\ell}} = \beta\sum_{\ell=1}^{k}\frac{m_\ell}{\sqrt{N_\ell + M_\ell}} \\ &\quad\leq \beta\sum_{\ell=1}^{k}\frac{m_\ell}{\sqrt{M_\ell}}. \end{aligned}$$

Finally, using Lemma 5 and noting that $\sum_{\ell=1}^{k} m_\ell \leq N$, we obtain

$$\sum_{\ell=1}^{k}\Big[\Phi(-\frac{\mu_{2\ell}}{\sigma_{2\ell}}) - \Phi(-\frac{\mu_{2\ell-1}}{\sigma_{2\ell-1}})\Big] = O(\sqrt{N \ln N}). \tag{28}$$

This together with (27) completes the proof. □

REFERENCES

[1] V. G. Vovk, "Aggregating strategies," in *Proc. 3rd Workshop Comput. Learn. Theory*, San Mateo, CA, USA: Morgan Kaufmann, 1990, pp. 371–383.

[2] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Mach. Learn.*, vol. 80, nos. 2–3, pp. 245–272, Sep. 2010.

[3] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice," *J. ACM*, vol. 44, no. 3, pp. 427–485, May 1997.

[4] D. Haussler, J. Kivinen, and M. K. Warmuth, "Tight worst-case loss bounds for predicting with expert advice," in *Computational Learning Theory*. Berlin, Germany: Springer, 1995, pp. 69–83.

[5] W. M. Koolen, "The pareto regret frontier," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 863–871.

[6] H. Yu, C. Shi, M. Kaminsky, P. B. Gibbons, and F. Xiao, "DSybil: Optimal sybil-resistance for recommendation systems," in *Proc. 30th IEEE Symp. Secur. Privacy*, May 2009, pp. 283–298.

[7] S. R. Etesami, S. Bolouki, A. Nedic, T. Basar, and H. V. Poor, "Influence of conformist and manipulative Behaviors on public opinion," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 1, pp. 202–214, Mar. 2019.

[8] A. Truong, S. R. Etesami, J. Etesami, and N. Kiyavash, "Optimal attack strategies against predictors–learning from expert advice," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 6–19, Jan. 2018.

[9] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Inf. Comput.*, vol. 108, no. 2, pp. 212–261, Feb. 1994.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[11] A. Truong and N. Kiyavash, "Optimal adversarial strategies in learning with expert advice," in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 7315–7320.

[12] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell.*, 2011, pp. 43–58.

[13] D. N. Tran, B. Min, J. Li, and L. Subramanian, "Sybil-resilient online content voting," in *Proc. NSDI*, vol. 9, no. 1, 2009, pp. 15–28.

[14] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting signature learning by training maliciously," in *Proc. Int. Workshop Recent Adv. Intrusion Detection*. Berlin, Germany: Springer, 2006, pp. 81–105.

[15] J. R. Douceur, "The sybil attack," in *Proc. Int. Workshop Peer Peer Syst.* Berlin, Germany: Springer, 2002, pp. 251–260.

[16] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Proc. 4th Prague Conf. Inf. Theory, Stat. Decesion Fundation, Random Processes*, 1965, pp. 263–272.

[17] N. Gravin, Y. Peres, and B. Sivan, "Towards optimal algorithms for prediction with expert advice," in *Proc. 27th Annu. ACM-SIAM Symp. Discrete Algorithms*, Jan. 2016, pp. 528–547.

[18] N. Gravin, Y. Peres, and B. Sivan, "Tight lower bounds for multiplicative weights algorithmic families," in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 80. Wadern, Germany: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[19] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, Jan. 2002.

[20] T. Lattimore and C. Szepesvri, *Bandit Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2020.

[21] E. Bayraktar, H. V. Poor, and X. Zhang, "Malicious experts versus the multiplicative weights algorithm in online prediction," 2020, *arXiv:2003.08457*. [Online]. Available: http://arxiv.org/abs/2003.08457

[22] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, no. 2. Belmont, MA, USA: Athena Scientific, 1995.

[23] A. C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variates," *Trans. Amer. Math. Soc.*, vol. 49, no. 1, pp. 122–136, Jan. 1941.

Authorized licensed use limited to: EPFL LAUSANNE. Downloaded on June 07,2021 at 09:04:19 UTC from IEEE Xplore. Restrictions apply.