Online Learning with Nasty Experts

Nikolaos Papagiannis
Purdue University
W. Lafayette,USA
npapagia@purdue.edu

Wojciech Szpankowski
Purdue University
W. Lafayette, USA
szpan@purdue.edu

Changlong Wu
Purdue University
W. Lafayette, USA
wuchangl@hawaii.edu

***Abstract*—We study the problem of learning from expert advice in the presence of *perturbed* noisy losses. Unlike existing work that assumes stochastic perturbations, we consider the case where the perturbation occurs *arbitrarily*, subject to a budget C on each expert—an approach we term "nasty" experts. The learner's performance is evaluated on the underlying true losses while only observing the perturbed noisy losses. Assuming the existence of an expert that incurs zero true losses, we show that the minimax risk *equals* $\Theta(C \log K)$ for an expert class of size K. Remarkably, this risk *cannot* be achieved by the standard Exponentially Weighted Average (EWA) algorithm with constant learning rates, for which we establish a lower bound of $\Omega(C^2/\log K)$. We further demonstrate a nearly matching upper bound of $O(C^2 + C\log K)$ for the EWA algorithm. Our main proof technique is based on a novel *potential-based* analysis that is of independent interest. Finally, we demonstrate the effectiveness of our nasty expert model in the context of binary classification with Massart's noise, *without* knowing the noise upper bound.**

I. Introduction

Learning from expert advice [7] is a fundamental problem that arises naturally in many different domains, including information theory [8], online learning, and game theory [7]. Formally, the problem can be described as a repeated game between Nature and a learner. At each time step, the learner selects one expert from K available experts and follows their *advice* to make a prediction. After the prediction, Nature (or adversary) assigns a *loss* to each expert based on the prediction outcome. The learner's objective is to minimize its cumulative loss over a time horizon T.

While much of the literature [7], [3], [12], [13] assumes the *adversarial* selection of losses—i.e., arbitrarily chosen losses that may depend on the learner's history—the evaluation of learner's performance is typically based on the *observed* losses. However, many real-world applications involve more challenging scenarios: the observed losses can actually be *perturbed* versions of certain true (unobserved) losses. For instance, in the context of cybersecurity, one typically assumes that the adversary can *perturb* certain underlying *true* losses and revealing only the perturbed version of those losses—yet the learner should still perform well on the true losses.

Despite its foundational nature, the study of learning from generally *perturbed* losses while being evaluated on the (unobserved) *true* losses has primarily focused either on *stochastic* settings [14], [3], [15], [4] or on the so-called *partial monitoring* settings (e.g., bandits) [7, Section 6], where an unbiased estimation of the true losses is possible. An exception is the work in [1], which considers arbitrary corruption within a certain budget. However, [1] assumes that the true losses are sampled $i.i.d.$.

This paper investigates a novel prediction scenario where we assume the perturbation can occur *arbitrarily* (i.e., the true losses are *not* estimable), while still allowing adversarially selected true losses. We refer to this type of perturbation as "nasty" noise, borrowing from [6].

Formally, at each time step t, Nature selects an arbitrary *true* loss vector $\ell_t \in [0,1]^K$ (which assigns losses to the K experts) but reveals only a perturbed *noisy* version $\tilde{\ell}_t \in [0,1]^K$. The selection of ℓ_t and $\tilde{\ell}_t$ is completely arbitrary, except for the following constraints: (1) the discrepancy between the cumulative true and noisy losses for *any* expert is upper bounded by a parameter C, which may depend on the time horizon T; and (2) there exists one expert that incurs zero cumulative true loss (i.e., we consider the well-specified case).

Our goal is to minimize the learner's cumulative risk evaluated on the *true* losses ℓ_t while only observing the *noisy* losses $\tilde{\ell}_t$, under the *worst-case* selection of ℓ_t's and $\tilde{\ell}_t$'s that satisfies the constraints above.

a) Summary of Results: Our main contributions reveal several unexpected and notable results. We show that the expected minimax risk on the true losses equals $\Theta(C\log K)$, provided the parameter C is known to the learner. This result is achieved using a novel elimination-based algorithm (Algorithm 1). Notably, the best achievable risk on the *noisy* losses grows as $C + \log K$ using the standard Exponentially Weighted Average (EWA) algorithm [2, Theorem 2.1]. Our result demonstrates a striking distinction between the minimax risks on true and noisy losses.

We further show that the EWA algorithm (with constant learning rate), despite achieving the optimal risk on noisy losses, incurs a risk lower bound of $\Omega\left(\frac{C^2}{\log K}\right)$ for *true* losses. This is substantially worse than our optimal algorithm (Algorithm 1), which achieves a risk of $O(C\log K)$. If we take $C = K = \sqrt{T}$, the EWA incurs nearly linear risk $T/\log T$, whereas our algorithm achieves a risk growth of $\sqrt{T}\log T$.

Finally, for unknown C we show that the EWA algorithm (Algorithm 2) with learning rate $\eta = 1$ achieves an $O(C^2 + C\log K)$ upper bound on the *true* risk. This, together with the lower bound, establishes a nearly tight $\tilde{\Theta}(C^2)$ risk for EWA. Our proof leverages a non-trivial *potential*-based analysis, which we believe is of independent interest. Note that, the EWA algorithm has the advantage of not requiring knowledge of the parameter C. This also implies a prediction rule for binary classification under Massart's noise *without* knowledge of the noise upper bound, using the *pairwise testing* approach from [15].

b) Related Work: Our framework is related to the scenario of *prediction with limited feedback*, as discussed in [7, Section 6]. However, our "nasty expert" differs in that the noisy losses are selected *arbitrarily* (subject to certain constraints), rather than through a *time-independent* feedback function as in [7]. Prediction with *stochastic noise* was first investigated in [14], which studies sequential prediction with binary outcomes passing through a Binary Symmetric Channel (BSC). This setting was recently generalized in [5] (see also references therein) under the assumption that the true losses are estimable. A similarly stochastic setting for binary classification with Massart's noise (assume a known noise upper bound) and well-specified true labels was discussed in [3] and later extended in [15] to handle more general noise models. In [1], the authors considered a perturbation scheme similar to ours but imposed a different (more stringent) constraint on the perturbation budget C and assumed *i.i.d.* generated losses. The regret analysis of *observable* losses using the Exponentially Weighted Average (EWA) algorithm has been widely studied; see [7], [9], [11].

II. Problem Setup

Let K be the number of experts. We consider the following online framework, which operates in T rounds. At each time step $t = 1, 2, \dots, T$, the following events occur sequentially:

1. Nature selects the *true* loss vector $\ell_t \in [0,1]^K$ but keeps it secret from the learner.
2. The learner selects *distribution* $\hat{p}_t \in \Delta([K])$, samples $\hat{k}_t \sim \hat{p}_t$, and predicts $\hat{k}_t$.
3. Nature selects a *noisy* loss vector $\tilde{\ell}_t \in [0,1]^K$ and reveals it to the learner.

Let $C \leq T$ be an arbitrary parameter that controls the noise level, which may grow w.r.t. T. The goal of the learner is to find a prediction strategy $\hat{p}^T$ that minimizes:

$$\mathsf{risk}_{T,K,C} := \sup_{\ell^T, \tilde{\ell}^T} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t[\hat{k}_t]\right] = \sup_{\ell^T, \tilde{\ell}^T} \sum_{t=1}^{T} \langle \hat{p}_t, \ell_t \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is scalar product, and $\ell^T := \{\ell_1, \dots, \ell_T\}$, $\tilde{\ell}^T := \{\tilde{\ell}_1, \dots, \tilde{\ell}_T\}$ are subject to the following constraints:

1. There *exists* $k^* \in [K]$ such that
$$\sum_{t=1}^{T} \ell_t[k^*] = 0; \tag{2}$$
2. For *all* $k \in [K]$, we have
$$\sum_{t=1}^{T} |\ell_t[k] - \tilde{\ell}_t[k]| \leq C. \tag{3}$$

Here, the first constraint ensures that there exists an expert incurring 0 *true* cumulative loss, while the second constraint ensures that for any expert, the discrepancy between the true and noisy cumulative losses is upper bounded by C.

III. Main Results

Our first main result establishes *tight* upper and lower bounds for $\mathsf{risk}_{T,K,C}$ when the parameter C is known in advance.

Theorem 1. *Assume that the parameter C is known to the learner. Then, there exists a prediction strategy $\hat{p}^T$ (see Algorithm 1) such that:*

$$\mathsf{risk}_{T,K,C} \leq (2C+1) \cdot \log K + 2C.$$

Furthermore, for any *prediction rule (regardless if C is known or not) and any T, K, C with $T \geq C \cdot \log K$, we have:*

$$\mathsf{risk}_{T,K,C} \geq \frac{1}{2} C \cdot (\log K - 1).$$

Note that the upper bound holds for *all* C, K, and T; for example, it remains valid even when C and K grow with respect to T. To better understand Theorem 1, it is instructive to compare it with the classical setting of learning from expert advice. It can be shown [2, Theorem 2.1] that for the EWA algorithm (Algorithm 2) predicting $\hat{p}^T$ with $\eta = 1$ we have

$$\sum_{t=1}^{T} \langle \hat{p}_t, \tilde{\ell}_t \rangle \leq O(C + \log K). \tag{4}$$

Note that this risk bound is evaluated on the *noisy* losses $\tilde{\ell}_t$, not the *true* losses ℓ_t. To convert it to the risk in (1), one must also bound

$$\sum_{t=1}^{T} \langle \hat{p}_t, \hat{\ell}_t - \ell_t \rangle. \tag{5}$$

We emphasize that even though the discrepancy between ℓ_t and $\tilde{\ell}_t$ is controlled by (3), the quantity in (5) cannot be simply upper bounded by C.

In fact, as shown by the lower bound in Theorem 1, the asymptotic behavior of the risk on *true* losses grows as $C \cdot \log K$, rather than $C + \log K$. This means that the quantity in (5) can be the *dominant* contribution to the risk on true losses if one naively applies the predictor in (4). This is confirmed formally in our next main result:

Theorem 2. *Let $\hat{p}^T$ be predicted by the EWA algorithm (see Algorithm 2) with parameter $\eta = 1$. Then, for all C*

$$\mathsf{risk}_{T,K,C} \leq O(C^2 + C \log K).$$

Furthermore, for any given $\eta > 0$, $K \geq C$ and $T \leq C^2$, we have the risk incurred by Algorithm 2 satisfies

$$\mathsf{risk}_{T,K,C} \geq \Omega\left(\max\left\{\frac{\eta C^2}{\log K}, C \log K\right\}\right).$$

Note that Theorem 2 is remarkable, as it demonstrates that the EWA algorithm, despite being minimax optimal on the noisy losses, is *not* optimal for the true losses (c.f. Theorem 1). Although, the EWA algorithm has the advantage without requiring knowledge of the parameter C.

We stress that, even though the EWA algorithm may appear classical, the proof of Theorem 2 is non-trivial, as the potential analysis is applied to the unobservable *true* losses rather than the noisy losses. We defer the detailed proof to Section V.

We now present the following corollary, whose proof can be obtained by inspecting the proofs of Theorems 1 and 2.

Corollary 1. *Both Theorem 1 and Theorem 2 remain valid if we relax the constraints in (2) and (3) to the following:*

1. *There exists $k^* \in [K]$ such that $\sum_{t=1}^T \tilde{\ell}_t[k^*] \le C$;*
2. *For all $k \in [K]$, $\sum_{t=1}^T \tilde{\ell}_t[k] \ge \sum_{t=1}^T \ell_t[k] - C$.*

a) Implications: We provide an application to the online classification with *random* noises as investigated in [3], [16]. For clarity of exposition, we consider only the binary classification with the *Massart's* noise (defined below), although our results apply to more general noise models.

Let $\mathcal{H} := \{h_1, \cdots, h_K\} \subset \{0,1\}^{\mathcal{X}}$ be a hypothesis class of size K. For any noise upper bound $\gamma \in [0, \frac{1}{2})$, we consider the following online learning game. At the start of the game, Nature fixes $k^* \in [K]$ unknown to the learner. For each time step $t = 1, 2, \ldots, T$, the following events occur sequentially:

1. Nature selects feature $\mathbf{x}_t \in \mathcal{X}$ and reveals to learner;
2. Learner makes a (random) prediction $\hat{y}_t \in \{0,1\}$;
3. Nature then selects a (unknown) parameter $\gamma_t \le \gamma$ and reveal
$$\tilde{y}_t = \mathsf{Bernoulli}(\gamma_t) \oplus h_{k^*}(\mathbf{x}_t),$$
where $\oplus$ denotes binary addition.

The learner targets a learning rule $\hat{y}^T$ that minimizes:

$$\tilde{r}_T(\mathcal{H}, \gamma) := \sup_{\mathbf{x}^T, h_{k^*}, \gamma_1^T} \mathbb{E}\left[\sum_{t=1}^T 1\{h_{k^*}(\mathbf{x}_t) \neq \hat{y}_t\}\right],$$

where the expectation is over both $\tilde{y}^T$ and $\hat{y}^T$.

It was shown in [3] that

$$\tilde{r}_T(\mathcal{H}, \gamma) \le \frac{\log |\mathcal{H}|}{1 - 2\sqrt{\gamma(1-\gamma)}}. \quad (6)$$

However, a major limitation of [3] (as well as [16]) is that the upper bound γ is assumed *known* to the learner. This is undesirable in practice. We show in the following theorem that a constant risk holds even *without* the knowledge of γ.

Theorem 3. *For any finite class $\mathcal{H}$, there exists a prediction rule $\hat{y}^T$ such that for all (unknown) $\gamma \in [0, \frac{1}{2})$ we have*

$$\tilde{r}_T(\mathcal{H}, \gamma) \le O\left(\frac{\log^2 |\mathcal{H}|}{(1-2\gamma)^4}\right).$$

Sketch of Proof. We will leverage the *pairwise-testing* scheme introduced by [16]. For any $t \in [T]$ and $k \in [K]$, we define the cumulative *empirical loss* $c_t[k] := \sum_{r=1}^t 1\{h_k(\mathbf{x}_r) \neq \tilde{y}_r\}$. Moreover, we define the cumulative *pairwise testing* loss as

$$L_t[k] := \sup_{k' \neq k} \sum_{r=1}^t 1\{c_r[k] \le c_r[k'] \text{ and } h_k(\mathbf{x}_r) \neq h_{k'}(\mathbf{x}_r)\}.$$

We now define *true* loss $\ell_t[k] := 1\{h_k(\mathbf{x}_t) \neq h_{k^*}(\mathbf{x}_t)\}$; and

$$\text{noisy loss } \tilde{\ell}_t[k] := L_t[k] - L_{t-1}[k].$$

By a similar argument as [16, Example 4], with probability $\ge 1 - \delta$ over $\tilde{y}^T$, the constructed losses ℓ_t and $\tilde{\ell}_t$ satisfy the conditions in Corollary 1 with $C \le \frac{2\log(|\mathcal{H}|/\delta)}{(1-2\gamma)^2}$. The theorem follows by invoking Theorem 2 over the constructed ℓ_t and $\tilde{\ell}_t$ and taking expectation by integration over δ. □

To our knowledge, Theorem 3 is the first known *constant* risk for online binary classification under Massart's noise, *without* known the upper bound γ.

IV. Proof of Theorem 1

We first prove the upper bound:

Lemma 1. *The prediction rule in Algorithm 1 achieves*

$$\mathsf{risk}_{T,K,C} \le (2C+1) \cdot \log K + 2C.$$

Proof. The proof follows a similar argument to that in [16], with the key difference being that the losses are *real* instead of binary. For any time step t, we define the following *potential*:

$$E_t = \sum_{k \in S^t} \max\left\{0, 2C - \sum_{i=1}^{t-1} \ell_i[k]\right\},$$

where S^t is defined in Algorithm 1. Let $N_t = |S^t|$ and $D_t = \sum_{k \in S^t} \ell_t[k]$. The *true* expected loss is given by

$$\mathbb{E}[\ell_t[\hat{k}_t]] = \frac{D_t}{N_t},$$

where $\hat{k}_t$ is sampled as described in Step 5 of Algorithm 1.

We now observe the following key property:

$$D_t \le N_t - N_{t+1} + E_t - E_{t+1}.$$

To see this, note that for any $k \in S^t$, either k is removed from S^t, contributing at most $N_t - N_{t+1}$ to D_t; or its contribution to D_t is upper bounded by the difference $E_t - E_{t+1}$. Here, the second assertion uses the fact that if an expert is *not* removed from S^t, its *true* cumulative loss must be upper bounded by $2C$ due to the constraint (3).

We now observe that

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}[\ell_t[\hat{k}_t]] &\le \sum_{t=1}^T \frac{N_t - N_{t+1}}{N_t} + \sum_{t=1}^T \frac{E_t - E_{t+1}}{N_t} \\ &\le (2C+1)\sum_{t=1}^T \frac{N_t - N_{t+1}}{N_t} + 2C \\ &\le (2C+1)\log K + 2C,\end{aligned}$$

where the second inequality follows from the facts that $E_t \le 2C \cdot N_t$ and $N_t \ge N_{t+1}$, and the final inequality follows from a standard argument as in [10, Thm 2] or [16, Thm 3]. □

While the prediction rule of Algorithm 1 may appear simple, the following lemma shows that its risk is, up to a constant factor, the best we can hope for from any algorithm.

Lemma 2. *For any prediction rule $\hat{p}^T$ and any parameters T, K, C satisfying $T \ge C \cdot \log K$, we have*

$$\mathsf{risk}_{T,K,C} \ge \tfrac{1}{2}C \cdot (\log K - 1).$$

Proof. Our goal is to construct hard instances $\ell^T, \tilde{\ell}^T$ using a *probabilistic* argument. We partition the time horizon into

Algorithm 1 Elimination-based Algorithm

```
Input: Threshold $C > 0$, number of experts $K$, time horizon $T$
Initialize $S^0 \leftarrow [K]$ {Start with all experts}
for $t = 1, 2, \ldots, T$ do
    Sample $\hat{k}_t \sim \hat{p}_t$ where $\hat{p}_t[k] := \frac{1}{|S^t|}$ for $k \in S^t$
    Predict using expert $\hat{k}_t$
    Observe the noisy loss vector $\tilde{\ell}_t \in [0,1]^K$
    Update cumulative noisy loss for all experts:
        $$\tilde{L}_k^t \leftarrow \sum_{s=1}^{t} \tilde{\ell}_s[k] \quad \text{for } k \in S^t$$
    Update the set of surviving experts:
        $$S^{t+1} \leftarrow \{k \in S^t : \tilde{L}_k^t \leq C\}$$
end for
```

$\log K$ epochs, each of size C (with the other $T - C \log K$ positions padded with 0 losses). Here, we assume $\log K$ is an integer; otherwise, we use the largest power of 2, K', such that $\log K' \geq \log K - 1$. For each epoch $j \in [\log K]$, we maintain a *random* expert class $S^j \subset [K]$ with $|S^j| = \frac{K}{2^j}$. To construct this, we first select $S^0 := [K]$ and sample S^{j+1} as a random subset of S^j such that $|S^{j+1}| = |S^j|/2$ for all $j \geq 0$.

Having constructed the sets S^j, we define the true and noisy losses as follows. For any time step t in epoch j:

- $\ell_t[k] = 0$ if $k \in S^j$ and $\ell_t[k] = 1$ otherwise;
- $\tilde{\ell}_t[k] = 0$ if $k \in S^{j-1}$ and $\tilde{\ell}_t[k] = 1$ otherwise.

It is straightforward to verify that both constraints (2) and (3) are satisfied by this construction. This is because any expert accumulates discrepancies between the true and noisy losses only within a *single* epoch, and each epoch has size C.

Now, consider *any* prediction rule. At each epoch $j \leq \log K$, the predictor must choose $\hat{k}_t \in S^{j-1}$, otherwise the true loss is 1. Moreover, since the noisy losses are 0 for all $k \in S^{j-1}$, the predictor gains no information about S^j during epoch j. Therefore, the expected error at *every* step, taken over the randomness of both S^j and the predictor's internal randomness, is at least $\frac{1}{2}$ (since $|S^j| = |S^{j-1}|/2$).

This implies that the expected cumulative risk is at least $\frac{1}{2}C \cdot \log K$, since the predictor incurs an expected error of $1/2$ at *every* step, and by the linearity of expectation. The lemma then follows from the fact that there must *exist* realizations $\ell^T, \tilde{\ell}^T$ that achieve this expected cumulative risk bound. □

The proof of Theorem 1 then follows directly from Lemma 1 and Lemma 2.

V. Proof of Theorem 2

The upper bound follows from the standard Exponential Weighted Average (EWA) algorithm (Algorithm 2), but with a substantially more complicated analysis.

Lemma 3. *Taking $\eta = 1$ in Algorithm 2, the EWA predictor achieves*

$$\mathsf{risk}_{T,K,C} \leq O(C^2 + C \log K),$$

where O hides absolute constant independent of T, K, C.

Proof. The proof follows from a careful definition of a potential that controls the risk on the *true* losses. For any $k \in [K]$, we define

$$C_k^t = \max\left\{0, C - \sum_{i=1}^{t} |\tilde{\ell}_i[k] - \ell_i[k]|\right\},$$

as the remaining "budget" of expert k at time step t. We define the following *potential*:

$$E_t = \sum_{k=1}^{K} (C_k^t + 2) w_k^t,$$

where w_k^t is the weight for expert k at time step t as in Algorithm 2. Our goal is to relate the *change* of the potential with the expected error incurred by the predictor.

For any $k \in [K]$ and $t \in [T]$, we claim that

$$(C_k^{t+1} + 2) w_k^{t+1} \leq (C_k^t + 2) w_k^t - \ell_t[k] \cdot w_k^t.$$

To see this, if $\tilde{\ell}_t[k] \geq \ell_t[k]$, then $w_k^{t+1} \leq e^{-\ell_t[k]} w_k^t$. Therefore,

$$\begin{aligned}(C_k^{t+1} + 2) w_k^{t+1} &\leq (C_k^t + 2) e^{-\ell_t[k]} w_k^t \\ &= (C_k^t + 2) w_k^t - (C_k^t + 2)(1 - e^{-\ell_t[k]}) w_k^t \\ &\leq (C_k^t + 2) w_k^t - \ell_t[k] \cdot w_k^t,\end{aligned}$$

where the second inequality follows by that $C_k^t + 2 \geq 2$ and $1 - e^{-x} \geq \frac{x}{2}$ for $x \in [0,1]$. If $\tilde{\ell}_t[k] \leq \ell_t[k]$, we denote $e_k^t = \ell_t[k] - \tilde{\ell}_t[k]$ as the "budget" used by expert k at step t. We have

$$\begin{aligned}(C_k^{t+1} + 2) w_k^{t+1} &= (C_k^t + 2 - e_k^t) e^{-\ell_t[k] + e_k^t} w_k^t \\ &\overset{(a)}{\leq} \max_{x \in [0, \ell_t[k]]} \left\{(C_k^t + 2 - x) e^{-\ell_t[k] + x} w_k^t\right\} \\ &\overset{(b)}{\leq} (C_k^t + 2) w_k^t - \ell_t[k] w_k^t\end{aligned}$$

where (a) follows since the function $f(x) = (C_k^t + 1 - x) e^{-\ell_t[k] + x}$ has *critical point* $x = C + 1 \geq \ell_t[k]$, i.e., the maximum must be attained at boundary $x \in \{0, \ell_t[k]\}$. For $x = 0$, inequality (b) reduces to our previous case for $\tilde{\ell}_t[k] \geq \ell_t[k]$; for $x = \ell_t[k]$ the inequality (b) follows trivially.

Putting everything together, we have shown that

$$E_{t+1} \leq E_t - \sum_{k \in [K]} \ell_t[k] w_k^t.$$

Note that $\mathsf{err}_t := \frac{\sum_{k \in [K]} \ell_t[k] w_k^t}{\sum_{k \in [K]} w_k^t}$ is precisely the expected error at step t incurred by Algorithm 2. We have

$$\begin{aligned}E_{t+1} &\leq E_t - \mathsf{err}_t \cdot \sum_{k \in [K]} w_k^t \\ &\leq E_t - \frac{\mathsf{err}_t}{C + 2} E_t \\ &= \left(1 - \frac{\mathsf{err}_t}{C + 2}\right) E_t,\end{aligned}$$

where the second inequality follows by definition of E_t and the fact that $C_k^t \leq C$. This implies that

$$E_{T+1} \leq E_1 \prod_{t=1}^{T} \left(1 - \frac{\mathsf{err}_t}{C+2}\right) \leq E_1 e^{-\frac{1}{C+2}\sum_{t=1}^{T} \mathsf{err}_t}.$$

Finally, we note that $E_1 = (C+2)K$ and $E_{T+1} \geq e^{-C}$ (via constrains (2) and (3)). This implies

$$\sum_{t=1}^{T} \mathsf{err}_t \leq (C{+}2)\left(\log((C+2)K) + C\right) \leq O(C^2{+}C\log K).$$

This completes the proof. □

Algorithm 2 EWA algorithm with Noisy Losses

```
Input: Learning rate $\eta > 0$, number of experts $K$, time horizon $T$
Initialize weights $w_k^1 \leftarrow 1$ for all $k \in [K]$
for $t = 1, 2, \ldots, T$ do
    Compute probability distribution over experts:
        $\hat{p}_t[k] \leftarrow \frac{w_k^t}{\sum_{j=1}^{K} w_j^t}$   for $k \in [K]$
    Sample $\hat{k}_t$ from the distribution $\hat{p}_t$
    Predict using expert $\hat{k}_t$
    Observe the noisy loss vector $\tilde{\ell}_t \in [0,1]^K$
    Update weights for all experts:
        $w_k^{t+1} \leftarrow w_k^t \cdot \exp(-\eta \tilde{\ell}_t[k])$   for $k \in [K]$
end for
```

Quite surprisingly, we can show that the C^2 dependency is necessary for the EWA algorithm. This contrast substantially with the $O(C \log K)$ risk achieved by Algorithm 2.

Lemma 4. *For any given $\eta > 0$, $K \geq C$ and $T \leq C^2$, the risk incurred by Algorithm 2 satisfies*

$$\mathsf{risk}_{T,K,C} \geq \Omega\left(\max\left\{C \log K, \frac{\eta C^2}{\log K}\right\}\right).$$

Proof. The $\Omega(C \log K)$ lower bound follows from Lemma 2, as it holds for any algorithm. To prove the second lower bound, we construct specific hard true and noisy losses, ℓ^T and $\tilde{\ell}^T$, that attain the claimed lower bound. We partition the time horizon into C epochs, each of size C, and define the losses during each epoch j as follows:

- **First Epoch** ($j = 1$)**:**
  - **True Loss:** All experts incur a loss of 0, i.e., $\ell_t[k] = 0$ for all $k \in [K]$.
  - **Noisy Loss:** All experts except $k = 1$ incur a loss of 0, i.e., $\tilde{\ell}_t[k] = 0$ for $k \neq 1$, and $\tilde{\ell}_t[1] = 1$.
- **Subsequent Epochs** ($j \geq 2$)**:**
  - **Initial Time Steps** ($t \leq \frac{\log K}{\eta}$)**:**
    - **True Loss:** Expert $k = 1$ incurs no loss ($\ell_t[1] = 0$), while all other experts incur a loss of 1 ($\ell_t[k] = 1$ for $k \neq 1$).
    - **Noisy Loss:** Experts $k = 1$ and $k = j$ incur no loss ($\tilde{\ell}_t[k] = 0$ for $k \in \{1, j\}$), while all other experts incur a loss of 1 ($\tilde{\ell}_t[k] = 1$ for $k \notin \{1, j\}$).
  - **Remaining Time Steps** ($t > \frac{\log K}{\eta}$)**:**
    - **True Loss:** Expert $k = j$ incurs a loss of 1 ($\ell_t[j] = 1$), while all other experts incur a loss of 0 ($\ell_t[k] = 0$ for $k \neq j$).
    - **Noisy Loss:** All experts incur a loss of 0, i.e., $\tilde{\ell}_t[k] = 0$ for all $k \in [K]$.

It is easy to verify that expert $k = 1$ satisfies the constraint (2), and for any $k \neq 1$, the constraint (3) is satisfied since the discrepancy only occurs at epoch $j = k$ by construction, and the epoch length is C.

We now observe the following key properties of the construction:

1. The weight for expert $k = 1$ satisfies $w_1^t = e^{-\eta C}$ for all time steps $t \geq C$.
2. At the beginning of any epoch $j \geq 2$, we have $w_k^t = e^{-(j-1)\log K}$. After the initial $\frac{\log K}{\eta}$ steps, the weight becomes $w_k^t = e^{-j \log K}$ for all $k \notin \{1, j\}$, and for $k = j$, we have $w_j^t = e^{-(j-1)\log K}$.

For any epoch $j \leq \frac{\eta C}{\log K}$, during the initial $\frac{\log K}{\eta}$ time steps, we have

$$\sum_{k \geq 2} w_k^t \geq e^{-(j-2)\log K} \geq e^{-\eta C} = w_1^t.$$

This implies, by our construction of true losses and the definition of the EWA prediction rule, that the expected loss is $\geq \frac{1}{2}$. Moreover, at any time step $t > \frac{\log K}{\eta}$ during epoch j, we have

$$\sum_{k \neq j} w_k^t \leq K e^{-j \log K} = e^{-(j-1)\log K} = w_j^t.$$

Therefore, the expected loss incurred by EWA remains $\geq \frac{1}{2}$. Since each epoch has C time steps, the cumulative expected risk is lower bounded by

$$\frac{1}{2}\frac{\eta C^2}{\log K}$$

for the first $\frac{\eta C}{\log K}$ epochs. This completes the proof. □

Taking $\eta = 1$, we obtain the lower bound

$$\Omega\left(\max\left\{C \log K, \frac{C^2}{\log K}\right\}\right).$$

This matches the $O(C^2 + C \log K)$ upper bound upto only an $\log K$ factor. For instance, if we take $K, C = T^{\alpha}$ for some $\alpha \leq \frac{1}{2}$, the risk of EWA algorithm with $\eta = 1$ satisfies

$$\mathsf{risk}_{T,K,C} = \tilde{\Theta}(T^{2\alpha}),$$

where $\tilde{\Theta}$ hides only an $\log T$ factor. This differs substantially the $\tilde{\Theta}(T^{\alpha})$ risk achieved in Theorem 1.

Problem 1 (Open Problem)**.** *Is $\Omega(C^2)$ a lower bound for* any *algorithm* without *knowing C? Can we get better risk via EWA algorithm with smarter (adaptive) choice of η?*

References

[1] Idan Amir, Idan Attias, Tomer Koren, Yishay Mansour, and Roi Livni. Prediction with corrupted expert advice. *Advances in Neural Information Processing Systems*, 33:14315–14325, 2020.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[3] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

[4] Alankrita Bhatt and Young-Han Kim. Sequential prediction under log-loss with side information. In *Algorithmic Learning Theory*, pages 340–344. PMLR, 2021.

[5] Alankrita Bhatt and Victoria Kostina. Prediction with noisy expert advice. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 3546–3551. IEEE, 2024.

[6] Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. Pac learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[7] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

[8] L. D. Davisson. Universal noiseless coding. *IEEE Trans. Inf. Theory*, IT-19(6):783–795, Nov. 1973.

[9] Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196. PMLR, 2014.

[10] Sham Kakade and Adam T Kalai. From batch to transductive online learning. *Advances in Neural Information Processing Systems*, 18, 2005.

[11] Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20(83):1–28, 2019.

[12] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *NIPS*, 2010.

[13] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

[15] Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Robust online classification: From estimation to denoising. *arXiv preprint arXiv:2309.01698*, 2023.

[16] Changlong Wu, Ananth Grama, and Wojciech Szpankowski. Information-theoretic limits of online classification with noisy labels. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.