ON THE VALUE OF STOCHASTIC SIDE INFORMATION IN ONLINE LEARNING

Junzhang Jia, Xuetong Wu, Jamie Evans, Jingge Zhu

University of Melbourne
Department of Electrical and Electronic Engineering
Parkville, Victoria, Australia

ABSTRACT

We study the effectiveness of stochastic side information in deterministic online learning scenarios. We propose a forecaster to predict a deterministic sequence where its performance is evaluated against an expert class. We assume that certain stochastic side information is available to the forecaster but not the experts. We define the minimax expected regret for evaluating the forecaster's performance, for which we obtain both upper and lower bounds. Consequently, our results characterize the improvement in the regret due to the stochastic side information. Compared with the classical online learning problem with regret scales with $O(\sqrt{n})$, the regret can be negative when the stochastic side information is more powerful than the experts. To illustrate, we apply the proposed bounds to two concrete examples of different types of side information.

***Index Terms*—** Online learning, Expert advice, Minimax regret, Side information

1. INTRODUCTION

The online learning problem aims to make predictions for probabilistic/deterministic instances which arrive sequentially, and has become significantly popular in game theory and learning theory fields recently. Merhav and Feder [1] studied online learning problems for stochastic setup from an information-theoretic perspective, followed by [2]. In the deterministic setting, we will usually introduce a class of competitive predictors providing advice to the forecaster, namely the expert class [1, 3], and the learning performance is evaluated by the regret, i.e. the loss gap between the proposed forecaster and the best expert. To effectively leverage the experts, Littlestone and Warmuth [4] proposed a weighted majority algorithm, and the follow-up works such as [3, 5, 6] further proposed the randomized algorithms which produce logarithmic regret. In a more specific setup, Haussler et al. [7] considered binary and continuous instance spaces and provided an $\Omega(\sqrt{n\log N})$ worst-case regret, where n is the sample size and N is the number of experts. With respect to different loss functions, Cesa-Bianchi and Lugosi [8] and Vanli and Kozat [9] provided explicit upper and lower bounds on the regret for absolute loss and squared loss, respectively.

As a common situation in practice, the forecaster could access some additional resources which we call it *side information*, that may provide some useful knowledge on the sequence of interest. Cover and Ordentlich [10] first studied a portfolio investment problem where the sequence of interest is the stock vectors that may depend on some finite-valued states (as side information), and their proposed forecaster can achieve the same wealth as the best side information dependent investment strategy. Xie and Barron [11] studied the case when the sequence of interest is generated according to a pair-wise parametric distribution conditioning on the side information, and derived an logarithmic upper bound of the minimax regret. Cesa-Bianchi and Lugosi [8] analyzed the problem with an additional (deterministic) side sequence, then the learning performance depends on the occurrences of its agreed symbols compared to the sequence of interest. Recently, Bhatt and Kim [12] studied the probabilistic online learning problem where the side information is the auxiliary random symbols generated jointly with the data instance to be predicted, and analyzed the minimax regret under the logarithmic loss.

However, to the best of our knowledge there is no prior work discussing the formulation and effects of the stochastic side information under a deterministic online learning scenario. Inspired by the transfer learning problem [13] where people transfer the knowledge from one domain (source) to the domain of interest (target) with both the source and target data drawn from different but related distributions, we specify the formation of the side information that may depend on the target sequence with some stochasticity. In a similar spirit, we aim to explore the influence of a stochastic *sequential side information* (SSI) for predicting a *target sequence* of interest. In this paper, we propose a novel problem formulation where a forecaster tries to predicts a deterministic sequence with some stochastic side information, which is not known to the expert class. Then we develop an online learning framework with the expert class where we will additionally leverage the side information for prediction to minimise the regret with respect to the best expert. With the proposed algorithm, we provide both the lower and upper bounds on the minimax regret under the absolute loss, where the target sequence is selected adversarially to maximise the regret. From the results, we show that introducing SSI can improve the typical learning rate in [1, 7, 8] if the side information performs better than the best expert. On the other hand, the side information will not hurt our prediction if it fails to provide much useful information.

2. PROBLEM FORMULATION AND MAIN RESULTS

2.1. Prediction with Experts and Stochastic Side Information

We consider the online learning problem for a deterministic target sequence with the side information: we aim to design a forecaster that sequentially predict the outcome of an unknown target sequence $\mathcal{T}_n = (X_1^T, X_2^T, ..., X_n^T)$ where each instance X_t^T takes value in a set $\mathcal{X} \subseteq \mathbb{R}$. The prediction of the forecaster at time t, denoted by $\tilde{X}_t$, takes value in a space $\mathcal{D}$ which is a convex and nonempty subset of $\mathbb{R}$, and we also assume $\mathcal{X} \subset \mathcal{D}$. We will compare the forecaster with a class of experts. We use $\mathcal{F}_n^\theta$ to denote the prediction sequence made by the expert θ, and f_t^θ to denote the prediction at time t. Here we denote by θ the index of an expert, taking value in an index set $\Lambda = \{1, 2, \cdots, N\}$ and $N \in \mathbb{R}^+$ is the number of experts in the class. The performance of the predictions is evaluated by a non-negative loss function $\ell : \mathcal{D} \times \mathcal{X} \mapsto \mathbb{R}^+$. We assume that only the forecaster has access to the SSI which may provide extra information on target sequence, which is denoted by $\mathcal{S}_n = (X_1^S, X_2^S, ..., X_n^S)$ for each $X_t^S \in \mathcal{X}$. At each time t, the forecaster predicts the current target instance X_t^T with previous observations $(X_1^T, ..., X_{t-1}^T)$ up to time $t-1$ and the corresponding SSI $(X_1^S, ..., X_t^S)$ up to time t. In other words, the prediction $\tilde{X}_t$ can be regarded as a function of both SSI and target sequences $\tilde{X}_t(X_1^S, X_2^S, ..., X_t^S, X_1^T, X_2^T, ..., X_{t-1}^T)$. We also use $\tilde{\mathcal{T}}_n := (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n)$ to denote the sequence of the predictions.

Following the common assumption in the literature [8, 14] for deterministic online learning problems, we assume that the target sequence is an arbitrary sequence. It can even be viewed as adversarially chosen by the "environment" with the knowledge of the prediction rule of the forecaster. However, we assume the SSI is generated in a conditional independent stochastic fashion by $P^n(\mathcal{S}_n|\mathcal{T}_n) = \prod_t P(X_t^S|X_t^T)$ with some (known) conditional probability distribution $P(X|Y)$.

To evaluate the performance of the prediction sequences, we firstly define the cumulative loss L which takes two sequences $A_n := (a_1, a_2, ..., a_n)$ and $B_n := (b_1, b_2, ..., b_n)$ with length n as:

$$L(A_n, B_n) = \sum_{t=1}^{n} \ell(a_t, b_t) \tag{1}$$

We use the absolute loss $\ell(a, b) = |a - b|$ throughout this paper in order to derive the lower bounds [8]. We then define the *regret* for the deterministic online learning problems as the difference between the cumulative loss between our forecaster and the best expert:

$$L(\tilde{\mathcal{T}}_n, \mathcal{T}_n) - \min_\theta L(\mathcal{F}_n^\theta, \mathcal{T}_n). \tag{2}$$

2.2. Minimax Expected Regret

In this section, we consider a problem of minimising the expected regret for a worst-case target sequence $\mathcal{T}_n$, i.e. $\mathcal{T}_n$ that maximises the expectation (w.r.t the side sequence) of (2). To this end, we will study the *minimax expected regret* defined as follows. Unless specified, the expectation is always taken over the SSI conditional distribution $P^n(\mathcal{S}_n|\mathcal{T}_n)$.

$$R(n) := \inf_{\tilde{\mathcal{T}}_n} \sup_{\mathcal{T}_n} \mathop{\mathbb{E}}_{\mathcal{S}_n} \{L(\tilde{\mathcal{T}}_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n) - \min_\theta L(\mathcal{F}_n^\theta, \mathcal{T}_n)\}. \tag{3}$$

To evaluate the usefulness of the SSI, we introduce the maximum likelihood estimation of the target instances X_t^T given X_t^S:

$$\hat{X}_t^T(X_t^S) = \operatorname*{argmax}_{X_t^T} P(X_t^S|X_t^T). \tag{4}$$

Then we denote the maximum likelihood prediction sequence by $\hat{\mathcal{T}}_n(\mathcal{S}_n) := (\hat{X}_t^T(X_1^S), \hat{X}_t^T(X_2^S), ..., \hat{X}_t^T(X_n^S))$. Furthermore, we make the assumption that the expected cumulative loss induced by $\hat{\mathcal{T}}_n(\mathcal{S}_n)$ is upper bounded in the following.

Assumption 1. For any target sequence $\mathcal{T}_n$, it holds that

$$\mathop{\mathbb{E}}_{\mathcal{S}_n}[L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n)] \leq C_S(n), \tag{5}$$

where $C_S(n)$ is a finite value depending on n for $n < \infty$.

Assumption 1 does not significantly restrict the SSI as we only require that the total loss induced by $\hat{\mathcal{T}}_n(\mathcal{S}_n) \in \mathcal{X}$ is bounded and we do not specify the function $C_S(n)$ at this stage. Clearly, $C_S(n)$ depends the SSI through the conditional distribution $P(X_t^S|X_t^T)$, for which we will give two concrete examples in Section 3.

With definitions in place, we introduce Algorithm 1, which we call *Exp3 with SSI*. This algorithm is an extension of the classical Exponentially Weighted Average (Exp3) algorithm [8], which uses an exponentially updated mixture of the experts as the forecaster. Our algorithm further treats the maximum likelihood estimator $\hat{X}_t^T(X_t^S)$ as an additional expert, so that the prediction made by the forecaster will partially depend on the information provided by the SSI.

Algorithm 1 Exp3 with SSI

Initialize the weights for the SSI w_1^S and all experts w_1^θ to be 1;
for $t = 1$ to n **do**
  Receive X_t^S;
  $\tilde{X}_t = \frac{w_t^S \hat{X}_t^T(X_t^S) + \sum_{\theta=1}^N w_t^\theta f_t^\theta}{w_t^S + \sum_{\theta=1}^N w_t^\theta}$;
  Receive X_t^T;
  The experts suffer the loss $\ell(f_t^\theta, X_t^T)$;
  The SSI suffers the loss $\ell(\hat{X}_t^T(X_t^S), X_t^T)$;
  $w_{t+1}^S = w_t^S e^{-\eta \ell(\hat{\mathcal{T}}_n(\mathcal{S}_n), X_t^T)}$;
  for $\theta = 1$ to N **do**
    $w_{t+1}^\theta = w_t^\theta e^{-\eta \ell(f_t^\theta, X_t^T)}$
  end for
end for
return $\tilde{\mathcal{T}}_n = (\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n)$

Since the decision space $\mathcal{D}$ is convex and nonempty, the prediction $\tilde{X}_t$ in Algorithm 1 formed by a linear combination of the expert f_t^θ and $\hat{X}_t^T(X_t^S)$ is also guaranteed to lie in the decision space $\mathcal{D}$. In the following theorem, we give an upper bound on the minimax regret with the proposed algorithm.

Theorem 1. With Assumption 1, the minimax expected regret in (3) is upper bounded by:

$$R(n) \leq \sqrt{\frac{n}{2}\log(N+1)} + \min\{C_S(n) - L^*(n), 0\}, \quad (6)$$

where we define $L^*(n) = \inf_{\mathcal{T}_n} \min_{\theta} L(\mathcal{F}_n^{\theta}, \mathcal{T}_n)$.

Proof. To simplify the notations, we denote by L_θ the loss induced by the loss induced from the expert θ, i.e. $L_\theta = L(\mathcal{F}_n^{\theta}, \mathcal{T}_n)$. We also denote L_{min} as the minimum cumulative loss among all the experts and the SSI:

$$L_{min} = \min\{L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n), \min_{\theta} L_\theta\}. \quad (7)$$

Then by adding and subtracting the term L_{min} in R, we have

$$\begin{aligned} R(n) &= \inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L(\Xi_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n) - L_{min} \\ &\quad + L_{min} - \min_{\theta} L_\theta\} \quad (8) \\ &\leq \inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L(\Xi_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n) - L_{min}\} \\ &\quad + \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L_{min} - \min_{\theta} L_\theta\} \quad (9) \\ &\leq \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \underbrace{\{L(\Xi_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n)) - L_{min}\}}_{R_a} \\ &\quad + \underbrace{\sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L_{min} - \min_{\theta} L_\theta\}}_{R_b} \quad (10) \end{aligned}$$

Then we separately upper bound the quantity R_a and R_b. We regard the one realization of X^S as an expert, then by the Theorem 2.2 in [8], we have

$$R_a \leq \frac{\ln N + 1}{\eta} + \frac{n\eta}{2} = \sqrt{\frac{n}{2}\log(N+1)} \quad (11)$$

with the optimal selection of the learning factor that $\eta = \sqrt{\frac{8\ln(N+1)}{n}}$. Then we can remove the first supremum and expectation in (10) as R_a is upper bounded by a quantity only depends on N and n. Then for R_b, we have,

$$\begin{aligned} R_b &\leq \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L_{min} - \min_{\theta} L_\theta\} \quad (12) \\ &= \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{\min\{L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n), \min_{\theta} L_\theta\} - \min_{\theta} L_\theta\} \quad (13) \\ &\overset{(a)}{\leq} \sup_{\mathcal{T}_n} \{\min\{\mathbb{E}_{\mathcal{S}_n} (L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n)), \min_{\theta} L_\theta\} - \min_{\theta} L_\theta\} \quad (14) \\ &= \sup_{\mathcal{T}_n} \{\min\{\mathbb{E}_{\mathcal{S}_n} (L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n) - \min_{\theta} L_\theta), 0\}\} \quad (15) \\ &\leq \min\{C_S(n) - L^*(n), 0\}. \quad (16) \end{aligned}$$

The second term in equation (15) is straightforwardly taken from assumption 1. Inequality (a) holds according to the fact that L_θ does not depend on $\mathcal{S}_n$ and the inequality $\mathbb{E}(\min\{a,b\}) \leq \min\{\mathbb{E}(a), \mathbb{E}(b)\}$. The last inequality results from the Assumption 1 and taking the supremum separately. We then complete the proof by adding up eq (11) and eq (16). □

The learning rate in its current form is not determined since the rate of $C_S(n)$ and $L^*(n)$ may vary across different cases. In Section 3, we will provide two specific structures for $C_S(n)$ and $L^*(n)$ with two examples. Notably, Theorem 1 indicates that the effect of the SSI will be determined by the difference between $C_S(n)$ and $L^*(n)$. In particular, if the expected cumulative loss of the side information is smaller than the loss induced by the best expert, the SSI is indeed helpful for predicting the target sequence. In contrast, when the SSI induces a higher loss compared to the best expert, the regret is upper bounded by $\sqrt{\frac{n}{2}\log(N+1)}$, which is essentially the same as the learning bound without SSI [7, 8, 9] when N is large. As a result, the second term in theorem 1 indicates how much the SSI can improve on the regret. To examine the tightness of the proposed upper bound, we also develop a lower bound for a particular outcome space and a decision space in the following theorem.

Theorem 2. Consider $\mathcal{X} = \{0,1\}$ and $\mathcal{D} = [0,1]$, with the absolute loss $\ell(x,y) = |x-y|$, we have

$$R(n) \geq \sqrt{\frac{n}{2}\log(N+1)} + \left(\xi^* - \frac{1}{2}\right) n, \quad (17)$$

where $\xi^* = \inf_{\tilde{X}_1} \mathbb{E}_{Z, X_1^S} |\tilde{X}_1(X_1^S) - Z|$, in which Z and X_1^S are jointly distributed according to $P(Z, X_1^S)$, and Z is marginally Bernoulli distributed as $Z \sim \text{Ber}(\frac{1}{2})$, X_1^S is generated according to the conditional distribution $P(X_1^S | Z)$.

Proof. The proof is different compared with the previous work [7, 8] that the prediction p_t now depends on the target outcome X_t^T by referencing the advice from both experts and the SSI X_t^S. First of all, we lower bound the first term in (3) as follows.

$$\begin{aligned} &\inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{L(\Xi_n(\mathcal{S}_n, \mathcal{T}_n), \mathcal{T}_n) - \min_{\theta} L_\theta\} \\ &= \inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{\sum_t |p_t - X_t^T| - \min_{\theta} \sum_t |f_t^{\theta} - X_t^T|\} \quad (18) \\ &\overset{(a)}{\geq} \inf_{\Xi_n} \mathbb{E}_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \{\sum_t |p_t - X_t^T| - \min_{\theta} \sum_t |f_t^{\theta} - X_t^T|\} \quad (19) \\ &= \inf_{\Xi_n} \mathbb{E}_{\mathcal{T}_n} \mathbb{E}_{\mathcal{S}_n} \sum_t |p_t - X_t^T| - \mathbb{E}_{\mathcal{T}_n} \min_{\theta} \sum_t |f_t^{\theta} - X_t^T|, \quad (20) \end{aligned}$$

where inequality (a) holds since the worst-case target sequence will generate no lower regret than compared with any other stochastic target sequences. We now assume that the target instances and the SSI instances are generated according to a joint distribution $P(X_t^S, X_t^T) = P(X_t^T) P(X_t^S | X_t^T)$, here $P(X_t^T)$ is a Bernoulli distribution with probability $\frac{1}{2}$, i.e., $X_t^T \sim \text{Ber}(\frac{1}{2})$. Clearly, in this case the expected loss incurred by the expert cannot be smaller than $n/2$. Then the sequential prediction problem

becomes n repetitive one-instance prediction problem as follows:

$$\inf_{\tilde{\mathcal{T}}_n} \mathop{\mathbb{E}}_{\mathcal{T}_n,\mathcal{S}_n} \sum_t |\tilde{X}_t(X_t^S) - X_t^T| = n \mathop{\mathbb{E}}_{X_1^T,X_1^S} |\tilde{X}_1^*(X_1^S) - X_1^T|. \quad (21)$$

It is known that for the absolute loss, the optimal forecaster $\tilde{X}_1^*$ is determined by minimising $\sum_{X_1^T} |\tilde{X}_1(X_1^S) - X_1^T| P(X_1^T|X_1^S)$ where $P(X_1^T|X_1^S)$ is induced by the joint distribution $P(X_1^T, X_1^S)$. Then we denote by ξ^* the expected loss induced by $\tilde{X}_1^*$ in (21), and note that the optimality of $\tilde{X}_1^*$ is w.r.t. the individual loss $|\tilde{X}_1^*(X_1^S) - X_1^T|$. Following (20), we have,

$$\begin{aligned}
&\inf_{\Xi_n} \sup_{\mathcal{T}_n} \mathop{\mathbb{E}}_{\mathcal{S}_n} \{L(p, \mathcal{T}_n) - \min_\theta L_\theta\} \\
&\geq \left(n\xi^* - \frac{1}{2}n \right) + \left(\frac{1}{2}n - \mathop{\mathbb{E}}_{\mathcal{T}_n} \min_\theta \sum_t |f_t^\theta - X_t^T| \right) \quad (22) \\
&\geq \left(\xi^* - \frac{1}{2} \right) n + \sqrt{\frac{n}{2} \log N + 1}, \quad (23)
\end{aligned}$$

where the last step is derived with the same procedures from Theorem 3.7 in [8]. □

It can be easily checked that the first term in (17) is always negative since ξ is always smaller than $\frac{1}{2}$. So for large n, the lower bound is negative, showing that the loss produced by our forecaster could potentially be much smaller than the best expert. It can also be seen that if the term $C_s(n) - L^*(n)$ in the upper bound takes the form $-cn$ for some positive c, then the upper and lower bound are matched in terms of the scaling law. In the next section, we show two examples demonstrating this point.

3. EXAMPLES

In this section, we consider two concrete online learning problems and derive their corresponding upper and lower bounds to verify the effectiveness of the proposed bounds. To characterize the behavior of the expert class, we will further consider the expert class generating a cumulative loss that scales linearly in n. The following expert class with an example displays one of the possible case satisfying the linear loss expert.

Definition 1. (Constant Expert) We say an expert class is the constant expert class such that all experts in the class yield a fixed prediction for any target instances. Mathematically,

$$\text{For all } t \text{ from 1 to } n,\ f_t^\theta = c_\theta, \quad (24)$$

where c_θ is some constant in $\mathcal{D}$.

Since the constant expert class is independent to the target instances, we can directly calculate the amount $L^*(n)$ defined in Theorem 1 under a certain setup for the decision space $\mathcal{D}$ and output space $\mathcal{X}$. We give two examples as follows.

Example 1. Assume the decision space is $\mathcal{D} = [0,1]$, and we consider a constant expert class that each expert predict a fixed constant f_c^θ in $\mathcal{D}$]. Also, assume that there always exists two experts predicting 0.1 and 0.7 for any time t. We also consider the binary output space, e.g, $\mathcal{X} = \{0,1\}$. Then we have

$$L^*(n) = \inf_{\mathcal{T}_n} \min_\theta L(\mathcal{F}^\theta, \mathcal{T}_n) = 0.1n. \quad (25)$$

3.1. SSI via a binary symmetric channel

In this example, we consider a learning problem setup that $\mathcal{X} = \{0,1\}, \mathcal{D} = [0,1]$ under the absolute loss $\ell(x,y) = |x-y|$. We assume that the SSI is the output of a binary symmetric channel with the flipping probability δ with the target sequence being the input. That is,

$$P(X_t^T = X_t^S | X_t^S) = 1 - \delta, \quad (26)$$
$$P(X_t^T = 1 - X_t^S | X_t^S) = \delta. \quad (27)$$

It can be shown that the ML estimator $\hat{\mathcal{T}}_n(\mathcal{S}_n) = \mathcal{S}_n$ when $\delta < \frac{1}{2}$, and $\hat{\mathcal{T}}_n(\mathcal{S}_n) = \bar{\mathcal{S}}_n$ when $\delta > \frac{1}{2}$, where $\bar{\mathcal{S}}_n := (1 - X_1^S, 1 - X_2^S, \ldots, 1 - X_n^S)$ denotes the sequence consisting of the flipped SSI instances. For $\delta = \frac{1}{2}$, $\hat{\mathcal{T}}_n(\mathcal{S}_n)$ can be any value in $\mathcal{D}^n$. With the maximum likelihood estimator $\hat{\mathcal{T}}_n(\mathcal{S}_n)$, we can calculate the expected loss as:

$$\mathop{\mathbb{E}}_{\mathcal{S}_n} [L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n)] = n\big(\delta \wedge \bar{\delta}\big) \quad (28)$$

where $a \wedge b = \min\{a,b\}$ and $\bar{\delta} = 1 - \delta$, which satisfies Assumption 1 with $C_S(n) = n\big(\delta \wedge \bar{\delta}\big)$.

Corollary 1. Under the binary flipping channel setup, when $L^*(n)$ grows linearly in n, i.e. $L^*(n) = c_f n$, we have

$$R(n) \leq \sqrt{\frac{n}{2} \log(N+1)} + \min\big\{0, n(\delta \wedge \bar{\delta} - c_f)\big\}. \quad (29)$$

Proof. The proof is straightforwardly following Theorem 1 from equation (15)

$$\begin{aligned}
R(n) &= \sqrt{\frac{n}{2} \log(N+1)} \\
&\quad + \sup_{\mathcal{T}_n} \{\min\{\mathop{\mathbb{E}}_{\mathcal{S}_n} (L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n) - \min_\theta L_\theta), 0\}\} \quad (30) \\
&\quad (31) \\
&= \sqrt{\frac{n}{2} \log(N+1)} + \min\{0, n\big(\delta \wedge \bar{\delta} - c_f\big)\}. \quad (32)
\end{aligned}$$

□

Notice that if any expert in the expert class $\mathcal{F}_n^\theta = (f_1^\theta, f_2^\theta, \ldots, f_n^\theta)$ suffers a cumulative loss more than $\frac{1}{2}n$, one can construct a new expert class $(1 - f_1^\theta, 1 - f_2^\theta, \ldots, 1 - f_n^\theta)$ that suffers a loss smaller than $\frac{1}{2}n$. Hence we only consider the case that c_f is always smaller or equal to $\frac{1}{2}$. From Corollary 1, we notice that if $\delta \wedge \bar{\delta}$ is smaller than c_f, the regret is asymptotically negative and scale linearly with n. In the following, we give the corresponding lower bound.

Corollary 2. Under the binary symmetric channel setup, we have

$$R(n) \geq \sqrt{\frac{n}{2}\log(N+1)} - n\left(\frac{1}{2} - \delta \wedge \bar{\delta}\right). \tag{33}$$

We see that in the case when $\delta \wedge \bar{\delta} < c_f$, the upper and the lower bound is matched in terms of the scaling law of order $\Omega(-n)$ (although with a different constant).

3.2. SSI via a Zero-mean Gaussian Channel

Now we consider a different type of side information such that the side instance is the noisy version of the target instance pair-wise: $X_{S,t} = X_{T,t} + N_t$, where $N_t \sim \mathcal{N}(0, \sigma^2)$. Here we still assume the target instances are restricted in a binary outcome space $\{0,1\}$. Note that in this problem setup, the side instances are drawn from a distribution over the space R, which differs from the instance space $\mathcal{X}$.

We can easily determine the maximum likelihood estimator $\hat{X}_t^T(X_t^S)$ for this problem: $\hat{X}_t^T(X_t^S) = 1$ when $X_t^S \geq \frac{1}{2}$, and $\hat{X}_t^T(X_t^S) = 0$ when $X_t^S < \frac{1}{2}$. By introducing the cumulative density function of the standard normal distribution $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$, we have

$$\mathop{\mathbb{E}}_{\mathcal{S}_n}[L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n)] = n\Phi(-\frac{1}{2\sigma}). \tag{34}$$

Corollary 3. Under the zero-mean Gaussian channel setup, and when $L^*(n)$ is linear to n, i.e. $L^*(n) = c_f n$, we have

$$R(n) \leq \sqrt{\frac{n}{2}\log(N+1)} + \min\{0, n\left(\Phi(-\frac{1}{2\sigma}) - c_f\right)\} \tag{35}$$

Proof. Following the proof of Theorem 2, we need to specify the quantity ξ^*. We start by finding the optimal forecaster p_t for predicting the target instances X_t^T by minimising the absolute loss:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X_t^T} l(p_t(X_t^S), X_t^T) = \sum_{X_t^T} \sum_{X_t^S} |p_t(X_t^S) - X_t^T| P(X_t^S, X_t^T) \\
&= \sum_{X_t^S} \left(\sum_{X_t^T} |p_t(X_t^S) - X_t^T| P(X_t^T | X_t^S) \right) P(X_t^S).
\end{aligned} \tag{36}$$

Then minimising the expected loss w.r.t. $p_t(X_t^S)$ is equivalently minimising $\sum_{X_t^T} |p_t(X_t^S) - X_t^T| P(X_t^T | X_t^S)$ for any X_t^S. Given $X_t^S = 1$, we have

$$\begin{aligned}
&\sum_{X_t^T} |p_t(X_t^S = 1) - X_t^T| P(X_t^T | X_t^S = 1) &(37)\\
&= (1-\delta)|p_t(X_t^S = 1) - 1| + \delta |p_t(X_t^S = 1) - 0|. &(38)
\end{aligned}$$

Then we can obtain when $\delta < \frac{1}{2}$, $p_t^*(X_t^S) = X_t^S$, when $\delta > \frac{1}{2}$, $p_t^*(X_t^S) = X_t^S$, and when $\delta = \frac{1}{2}$, there are an infinite number of minimizers $p_t^*(X_t^S)$ between 0 and 1. One can verify that the optimal forecaster $p_t^*(X_t^S)$ is the maximum likelihood estimator $\hat{\mathcal{T}}_n(\mathcal{S}_n)$. Then similar to (28), we have

$$\begin{aligned}
\xi^* &= \mathop{\mathbb{E}}_{X_t^T} \mathop{\mathbb{E}}_{X_t^S} |p_t^* - X_t^T| = \frac{1}{n} \mathop{\mathbb{E}}_{\mathcal{S}_n} [L(\hat{\mathcal{T}}_n(\mathcal{S}_n), \mathcal{T}_n)] &(39)\\
&= \delta \wedge \bar{\delta}. &(40)
\end{aligned}$$

By substituting the ξ^* in Theorem 2 as (40), we completed the proof. □

It can be seen that the upper bound in this example behaves similarly to that in the binary symmetric channel case. When the quantity $\Phi(-\frac{1}{2\sigma})$ is smaller than c_f, the upper bound of the minimax regret becomes negative with a large n. Intuitively, when σ is large, the quantity $\Phi(-\frac{1}{2\sigma})$ will become larger, which decreases the effectiveness of the SSI.

Corollary 4. Under the zero-mean Gaussian channel setup, we have

$$R(n) \geq \sqrt{\frac{n}{2}\log(N+1)} + n\left(\Phi(-\frac{1}{2\sigma}) - \frac{1}{2}\right). \tag{41}$$

Similarly, as $\sigma > 0$, we have $\Phi(-\frac{1}{2\sigma}) < \Phi(0) = \frac{1}{2}$, the lower bound will become negative when n increases. Similar to the binary symmetric channel example, the upper and the lower bound is matched in terms of the scaling law if $c_f > \Phi(-\frac{1}{2\sigma})$.

4. CONCLUSION AND FUTURE WORKS

This work shows the upper and lower regret bounds on general deterministic online learning problems with two concrete examples, where an additional stochastic sequential side information sequence is revealed to the forecaster. The result infers the effectiveness of the the side information which may significantly improved the learning rate and shows the possibility of producing a negative regret. For future works, one may wish to find a tighter lower bound on the minimax regret based on more advanced algorithms, or more elementary proofs.

References

[1] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[2] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Online transfer learning: Negative transfer and effect of prior knowledge," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 1540–1545.

[3] V. Vovk, "A game of prediction with expert advice," *Journal of Computer and System Sciences*, vol. 56, no. 2, pp. 153–173, 1998.

[4] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and computation*, vol. 108, no. 2, pp. 212–261, 1994.

[5] V. G. Vovk, "Aggregating strategies," *Proc. of Computational Learning Theory, 1990*, 1990.

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

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

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

[9] N. D. Vanli and S. S. Kozat, "A unified approach to universal prediction: Generalized upper and lower bounds," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 646–651, 2014.

[10] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 348–363, 1996.

[11] Q. Xie and A. R. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 431–445, 2000.

[12] A. Bhatt and Y.-H. Kim, "Sequential prediction under log-loss with side information," in *Algorithmic Learning Theory*. PMLR, 2021, pp. 340–344.

[13] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[14] N. Cesa-Bianchi and F. Orabona, "Online learning algorithms," *Annual review of statistics and its application*, 2021.