Sequential prediction under log-loss with side information

Alankrita Bhatt A2BHATT@UCSD.EDU
Department of Electrical and Computer Engineering
University of California San Diego
La Jolla, CA 92093, USA

Young-Han Kim YHK@UCSD.EDU
Department of Electrical and Computer Engineering
University of California San Diego
La Jolla, CA 92093, USA

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

The problem of online prediction with sequential side information under logarithmic loss is studied, and general upper and lower bounds on the minimax regret incurred by the predictor is established. The upper bounds on the minimax regret are obtained by constructing and analyzing a probability assignment based on mixture probability assignments in universal compression, and the lower bounds are obtained by way of a redundancy–capacity theorem. A tight characterization of the regret is provided in some special settings.

1. Introduction

We consider a variant of the problem of sequential prediction under log-loss with side information[1]. The particular variant under consideration was first studied by Fogel and Feder (2017). Let $X \in \mathcal{X}$ and $Y \in \{0,1\}$ denote two jointly distributed random variables. Let the marginal distribution of X be denoted by $P_X(x)$. A hypothesis f in the *hypothesis class* $\mathcal{F}$ determines the conditional distribution $P_f(y|x)$, or equivalently, the conditional probability mass function (pmf) $p_f(y|x)$, for $y \in \{0,1\}$ and $x \in \mathcal{X}$. Each hypothesis is characterized by a tuple $f = (g, \theta_0, \theta_1)$, where

1. $\theta_0, \theta_1 \in [0,1]$
2. $g \in \mathcal{G} \subset \{\mathcal{X} \to \{0,1\}\}$.

In other words, g belongs to a class $\mathcal{G}$ of binary functions. We assume that $\mathcal{G}$ has finite VC dimension, denoted by $\mathrm{VCdim}(\mathcal{G})$.

Given a chosen hypothesis $f = (g, \theta_0, \theta_1)$ we then have

$$Y|\{X = x\} \sim \mathrm{Bernoulli}(\theta_{g(x)}).$$

Thus, given the *side information* X, the random variable Y is distributed as either $\mathrm{Bernoulli}(\theta_0)$ or $\mathrm{Bernoulli}(\theta_1)$. Picking a hypothesis $f \in \mathcal{F}$, let $(X_i, Y_i)_{i=1}^n$ be drawn i.i.d. from the joint distribution of X and Y characterized by the hypothesis f as

$$P(x^n, y^n) = \prod_{i=1}^{n} P_X(x_i) P_f(y_i|x_i). \tag{1}$$

1. Extended abstract. Full version at (Bhatt and Kim, 2021)

The problem of sequential prediction under log-loss, also known as the sequential probability assignment problem, can be thought of as a game between the player and nature. First, nature picks a hypothesis $f \in \mathcal{F}$ unbeknownst to the player, and X^n, Y^n are then generated according to the law (1). At each time step $i \in [n]$, X_i is revealed to the player, who then assigns a pmf $q(\cdot|X^i, Y^{i-1})$ to Y_i conditioned on $X^i = (X_1, \ldots, X_i)$ and $Y^{i-1} = (Y_1, \ldots, Y_{i-1})$. Next, Y_i is revealed and the player incurs loss $-\log q(Y_i|X^i, Y^{i-1})$. Nature assigns the pmf $p_f(\cdot|X_i)$ at each time step i and incurs loss $-\log p_f(Y_i|X_i)$. The goal of the game is to minimize the expected value of cumulative loss relative to nature (known as the regret), without knowledge of f. Importantly, we wish to do this without knowing P_X either.

To make this notion precise, define the regret incurred by the probability assignment q when nature picked f and the distribution of X is P_X as

$$R_{n,P_X}(q, f) := \mathbb{E}\left[\sum_{i=1}^{n} \log \frac{1}{q(Y_i|X^i, Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)}\right]. \tag{2}$$

Then, the worst-case regret for the probability assignment q is

$$R_n(q) := \max_{P_X, f} R_{n,P_X}(q, f). \tag{3}$$

In this paper, we aim to calculate the min-max regret

$$R_n := \min_{q} R_n(q). \tag{4}$$

and discover a probability assignment q that is optimal or near-optimal in the sense of achieving $R_n(q)$ close to the optimal value (4).

The log-loss is of central importance in information theory as it connects two canonical problems in data science—compression and prediction; see the survey (Merhav and Feder, 1998). To motivate the use of the log-loss in the current problem, we view it as an extension of the problem of universal compression. Indeed, if there is no side information X present, then the problem is equivalent to universal compression of an i.i.d. Bernoulli source which has been well studied (Rissanen, 1983a,b, 1984; Xie and Barron, 1997, 2000). The minimax regret R_n then is significant operationally, representing the number of extra bits above the entropy one must pay as the price for compressing the source without knowing its distribution. Remarkably, one can show that $R_n = \frac{1}{2}\log n + o(\log n)$ in this setting. In a similar vein, Shkel et al. (2018) studied a closely related problem where a compressed version of the sequence Y^n is available as side information noncausally (i.e. not sequentially) and demonstrate its equivalence to lossy compression.

In the current setting, if the function g is known, then simple extensions of the techniques developed to tackle the problem of universal compression of an i.i.d. Bernoulli source can be used to show that $R_n \leq \log n + o(\log n)$, and we will elaborate on this important special case in the full paper. The problem becomes nontrivial when the function g is not known, and new techniques need to be developed to characterize R_n in this case.

In the standard study of classification in statistical learning theory, the loss function employed is the 0-1 loss or the indicator loss, and the notion of VC dimension plays a crucial role in characterizing the fundamental limits of binary classification (Shalev-Shwartz and Ben-David, 2014). In particular, $\text{VCdim}(\mathcal{G}) < \infty$ implies the PAC-learnability of the hypothesis class $\mathcal{G}$. Viewing the current setting as a log-loss variant of the standard classification problem studied in statistical

learning (which uses the indicator loss) motivates the choice of constraint $\mathrm{VCdim}(\mathcal{G}) < \infty$. A variant of the current problem with indicator loss instead of log-loss was studied in (Lazaric and Munos, 2012). We have considered a specific class of conditional distributions to compete against (recall that under hypothesis f we have $p_f(Y=0|X=x) = \mathrm{Bern}(\theta_{g(x)})$). As mentioned in the preceding paragraphs, our motivation stems from universal compression with side information, and to consider a log-loss variant of the standard binary classification problem. In both these cases, the choice of the considered class seems natural. However, in general, one could view this problem as an online conditional density estimation problem and correspondingly consider an arbitrary class $\mathcal{F}$ where any $f \in \mathcal{F}$ may characterize the conditional distribution $p_f(y|x)$ in a far more complex manner. It then makes sense to expect R_n in this case to depend on a measure of complexity of $\mathcal{F}$ akin to the VC dimension. Indeed, in (Rakhlin et al., 2015a) the authors develop a remarkable theory parallel to statistical learning theory when the data is non-i.i.d. They develop analogues of several combinatorial dimensions and the Rademacher complexity in the non-i.i.d. case. They then leverage this theory in (Rakhlin et al., 2015b) to study the minmax regret in several online learning problems (with adversarial data). This approach is employed to study sequential prediction with the log-loss in (Rakhlin and Sridharan, 2015) and (Bilodeau et al., 2020). However, it is important to note that the proofs in these works are nonconstructive—they proceed via using minmax duality and analyzing the dual game, which does not provide a strategy (i.e. a probability assignment) achieving the regret upper bound that is proven. Our method on the other hand involves construction of a sequential probability assignment. In the next subsection, we will mention and compare our results with the aforementioned two papers studying the log-loss.

1.1. Main Results

Our first main result is a probability assignment that yields an upper bound on R_n.

Theorem 1 *If $\mathcal{G}$ is such that VCdim$(\mathcal{G}) = d < \infty$, we have for an absolute constant $C \leq 250$, for a probability assignment q^* (which is specified in detail further on)*

$$R_n(q^*) \leq 125C\sqrt{dn}\log(2n) + d(\log n)^2 + 2. \tag{5}$$

Moreover, for any $P_X, f, \delta \in (0,1)$, with probability greater than $1-\delta$,

$$\sum_{i=1}^{n} \log \frac{1}{q^*(Y_i|X^i, Y^{i-1})} - \sum_{i=1}^{n} \log \frac{1}{p_f(Y_i|X_i)} \leq 25C\sqrt{dn}\log(2n)\left(C\sqrt{d} + \sqrt{2\log\frac{2\log n}{\delta}}\right) + d(\log n)^2 + 2 \tag{6}$$

In the full paper, we construct and analyze the probability assignment q^*. In (Fogel and Feder, 2017), the authors established that $R_n = O(d\sqrt{n}\log n)$, and $R_n \leq \left(2d + 1 + \log\frac{1}{\delta}\right)\sqrt{n}\log n$ with probability $\geq 1-\delta$. Our proof (and probability assignment) is different and achieves the same dependence on n, and a better dependence on δ in the high-probability version of the result.

We also establish a lower bound on R_n.

Theorem 2 *We have*

$$R_n \geq d + \log(n+1) - 2\sqrt{e}d^2e^{-3n/100d} - \log(\pi e).$$

The non-constructive approaches of the papers (Rakhlin and Sridharan, 2015) and (Bilodeau et al., 2020) mentioned earlier establish an $O(d \log n)$ upper bound for the $\mathcal{F}$ under consideration. In conjunction with Theorem 2 we see that the dependence of R_n on n is indeed $\Theta(\log n)$. This implies that the q^* employed to prove Theorem 1 is suboptimal, and raises the important question of constructing a better probability assignment achieving the tight upper bound.

Open Problem 1 *Construct a probability assignment q for the VC hypothesis class that achieves $O(d \log n)$ regret.*

As mentioned earlier, the problem of sequential probability assignment can be posed for any general (and possibly very complex) class $\mathcal{F}$ and viewed as an online conditional density estimation problem.

Open Problem 2 *Construct and analyze a probability assignment q for the case when $\mathcal{F}$ is a general hypothesis class.*

As a starting step towards Open Problem 1, we considered a few special cases of the function class $\mathcal{G}$ and distribution P_X and provide a sequential probability assignment achieving $O(d \log n)$ upper bound. These upper bounds constitute our third main result.

References

Alankrita Bhatt and Young-Han Kim. Sequential prediction under log-loss with side information. *arXiv preprint arXiv:2102.06855*, 2021.

Blair Bilodeau, Dylan Foster, and Daniel Roy. Tight bounds on minimax regret under logarithmic loss via self-concordance. In *International Conference on Machine Learning*, pages 919–929. PMLR, 2020.

Yaniv Fogel and Meir Feder. On the problem of on-line learning with log-loss. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2995–2999. IEEE, 2017.

Alessandro Lazaric and Rémi Munos. Learning with stochastic inputs and adversarial outputs. *Journal of Computer and System Sciences*, 78(5):1516–1537, 2012.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44 (6):2124–2147, 1998.

Alexander Rakhlin and Karthik Sridharan. Sequential probability assignment with binary alphabets and large classes of experts. *arXiv preprint arXiv:1501.07340*, 2015.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, 2015a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015b.

Jorma Rissanen. A universal data compression system. *IEEE Transactions on information theory*, 29(5):656–664, 1983a.

Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, pages 416–431, 1983b.

Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636, 1984.

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

Yanina Shkel, Maxim Raginsky, and Sergio Verdú. Sequential prediction with coded side information under logarithmic loss. In *Algorithmic Learning Theory*, pages 753–769, 2018.

Qun Xie and Andrew R Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

Qun Xie and Andrew R Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.