Redundancy Capacity Theorem for On-Line Learning Under a Certain Form of Hypotheses Class

Shachar Shayovitz
School of Electrical Engineering
Tel Aviv University
Email: shachar.shay@gmail.com

Meir Feder
School of Electrical Engineering
Tel Aviv University
Email: meir@eng.tau.ac.il

***Abstract*—In this paper we consider the problem of on-line learning in the stochastic setting under a certain form of hypotheses class. We prove an equivalence between the minimax redundancy and capacity of the channel between the class parameters and the labels conditioned on the data features (side information). Our proof extends Gallager's Redundancy Capacity theorem for universal prediction to on-line learning with the considered form of hypotheses class. Moreover, this result confirms the optimality of previous ad-hoc universal learners, or universal predictors with side information, but more importantly, extends these previous results to more general hypotheses classes.**

I. Introduction

Universal prediction is the task of predicting the next sample in a sequence, based only on previous samples with almost no *a-priori* knowledge on the sequence. There are primarily two settings for this problem. In the first, the sequence is individual/arbitrary, i.e, it is not generated by any underlying statistical model, but we fit to that sequence a model P_θ from a certain hypotheses class. In the second, denoted stochastic, the sequence is generated by some source P_θ from the hypotheses class. Both settings have been studied thoroughly and the reader is referred to an overview of this field in [1]. In this paper we will concentrate on the stochastic setting but in the on-line learning problem and on universal on-line learners which do not know the value of θ.

As in any prediction problem, we wish to examine the predictor's accuracy based on some cost function. We will concentrate on the log-loss which is a commonly used cost function in many applications such as classification, data compression and more. For a causal predictor $b(\cdot|y^{t-1}) \geq 0, \sum_{y_t} b(y_t|y^{t-1}) = 1$ of y_t, where $y^{t-1} = \{y_1, y_2, ..., y_{t-1}\}$ are the past labels, the log-loss for a specific y_t is defined as $-\log b(y_t|y^{t-1})$. In the stochastic case, it was shown in [1], that the predictor that minimizes the expected log-loss is $P_\theta(y_t|y^{t-1})$ for every θ. However, we are interested in universal predictors which do not have access to the parameter θ. Therefore, the log-loss of the universal predictor is examined in comparison to the log-loss of the optimal predictor which knows θ. Basically, the objective now becomes the difference between the cumulative log-loss of the universal predictor and the cumulative log-loss of the optimal predictor and this difference is called the *regret*.

Moreover, a common extension to the stochastic case is where there is an unknown distribution $\pi(\theta)$ over the parameter space. The on-line predictor in this case wishes to minimize the average behavior of the *regret*, without knowing $\pi(\theta)$. We denote the average regret, which averages over the data sequence and class parameters as the *redundancy*. One interesting task is to find the universal predictor that minimizes the *redundancy* for the worst possible prior $\pi(\theta)$ and compute the value of the resulting *minimax redundancy*. If the *minimax redundancy* grows sub-linearly with the data length n as it goes to infinity, the *redundancy* rate goes to zero and effectively the universal predictor performs asymptotically as well as the optimal predictor, had θ (or more generally $\pi(\theta)$) been known in hindsight.

An important result in universal prediction is the Redundancy Capacity Theorem. First introduced in [2] and then independently in [3] and in [4], the theorem states that the *minimax redundancy* is equal to the channel capacity between θ and the data sequence measurements. Interestingly, it shows the strong underlying connections between universal sequential prediction and channel coding. This theorem also shows how to construct the sequential universal predictor which achieves the minimax redundancy. Moreover, in [5] it is shown that the channel capacity is essentially a lower bound also in a stronger sense, that is, for "most" sources in the class. This result extends Rissanen's lower bound for parametric families.

Consider now the on-line learning problem with log-loss as discussed in [6], [7], [8], [9], [10] and [11]. In on-line learning, the goal is to predict the label of a feature based on past features and associated labels. The features may be considered as side information of the labels that are causally available to the predictor. *Can the Capacity Redundancy theorem be extended to on-line learning?* The answer for arbitrary hypotheses classes is unknown and there is no general derivation of a universal minimax predictor. Nevertheless, some special cases were considered, e.g., in [12], Sec. IX, where the hypotheses class was modeled as a multiplication of several i.i.d sources, where each source is determined by a different value of the side information sample. In that work, a very specific universal predictor was proposed and it was shown that it achieves the asymptotic minimax regret. In [13], a closely related problem of universal portfolio with side information was considered. In that work, a universal portfolio was proposed, which attains

the optimal growth rate compared to the best state constant rebalanced portfolios, where the side information determines the state. This optimal portfolio is again a multiplication of mixtures of portfolios. The universal predictors proposed in [12] and [13] attain the asymptotic minimax regret but it is unclear if these predictors are optimal in the non asymptotic *minimax redundancy* sense and if a Capacity Redundancy equivalence exists in these scenarios.

In this paper a Redundancy-Capacity theorem for the on-line learning problem is proven that holds for a certain form of hypotheses class. The optimal predictor that achieves the *minimax redundancy* is also derived. As a special case, under the hypotheses classes used in [12] and [13] our theorem holds and the optimal universal predictor is a multiplication of mixtures which coincides with the universal predictors proposed in these works.

The hypotheses class defined in the next section basically partitions the labels sequence to several sub-sequences based on the output of some arbitrary finite state machine. Each sub-sequence is controlled by a distribution which may be or not different than the distributions of the other sub-sequences. An example for this class can be a sequence of horse races in which the labels indicate the winning horse in each race and the side information indicates the weather (sunny or rainy) in each race. It is reasonable to assume that the winning probability for each horse can change conditioned on the weather.

Even though the proposed hypotheses class is large and can describe interesting problems, there are many important hypotheses classes which cannot be described in this form. These include any partitioning of the labels sequence which depends on unknown parameters. For example, when the feature is any real value in the interval $\mathbb{R}^d$ and is the input to a Perceptron with unknown parameters which define the hyper-planes that separate $\mathbb{R}^d$. The Perceptron's output partitions the labels sequence to sub-sequences with different distributions. However, this partitioning is dependent on the unknown parameters which are the same for each sub-sequence. This prohibits the redundancy from factoring as in the model proposed in this paper, and the theorem proven later does not hold for that case. A very interesting research question would be to find out whether a Redundancy Capacity theorem for such hypotheses class exist.

II. Problem Definition and Main Result

We consider a labels sequence, $y^n = \{y_1, y_2, ..., y_n\}$, and a side information or features sequence, $x^n = \{x_1, x_2, ..., x_n\}$, where here for all t, $x_t \in \{1, 2, ..., K\}$ and $y_t \in \{1, 2, ..., M\}$. We are interested in on-line probabilistic prediction of the label y_t for the feature x_t based on all past features and associated labels, x^{t-1} and y^{t-1}.

A. Hypotheses Class Definition

We define the hypotheses class, i.e, the family of conditional distributions $p^{\underline{\theta}}(y^t|x^t)$, parameterized by the vector $\underline{\theta}$, which generates the labels sequence, y^t, for the features, x^t. We assume the conditional distribution of the labels given the features, is a multiplication of up to K different distributions, which are determined by the features: x^t.

Specifically, the distributions which make up the hypotheses class for this problem are of the following form,

$$p^{\underline{\theta}}\left(y^t|x^t\right) = \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^t\right) \tag{1}$$

where $\underline{\theta} = [\theta_0, \theta_1, \theta_2, ..., \theta_{K-1}]$, $\theta_i \in \Theta$ and $\underline{y}_j^t = \{y_i, 0 \leq i \leq t | x_i = j\}$.

Basically, the features partition the labels sequence to K different sub-sequences. We define Θ as the parameterization space of the different distributions. For example, with multivariate Gaussian distributions, each element in Θ would contain the mean vector and covariance matrix of a respective distribution. This model defines up to K different distributions, each is parametrized by the corresponding θ_j. The hypotheses class defined in [12] satisfies (1) and thus our novel result can also be applied to their scenario.

We will now find the conditional distribution of y_t, $p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right)$ using the model in (1), Bayes law and total probability,

$$p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) = \frac{p^{\underline{\theta}}\left(y^t|x^t\right)}{\sum_{y_t} p^{\underline{\theta}}\left(y^t|x^t\right)} \tag{2}$$

Plugging (1) in (2)

$$p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) = \frac{\Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^t\right)}{\sum_{y_t} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^t\right)} \tag{3}$$

Eliminating common factors,

$$p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) = \frac{p^{\theta_{x_t}}\left(\underline{y}_{x_t}^t\right)}{\sum_{y_t} p^{\theta_{x_t}}\left(\underline{y}_{x_t}^t\right)} \tag{4}$$

Thus the conditional probability of y_t based on all the history is determined by its associated feature x_t and the past labels with the same feature value.

$$p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) = p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) \tag{5}$$

B. Minimax Redundancy Definition

As discussed in the introduction, we are interested in the regret which is defined as follows,

$$Reg(x^n) = \sum_{t=1}^{n} \left(\ln p^{\underline{\theta}}\left(y_t|y^{t-1}, x^t\right) - \ln q\left(y_t|y^{t-1}, x^t\right)\right) \tag{6}$$

where $q(y_t|y^{t-1}, x^t)$ is an arbitrary (universal) conditional probability.

The regret for a sequence y^n with features x^n is the difference between the log-loss of the universal predictor and the log-loss of the true distribution. We can now define the redundancy for a specific x^n, as the expectation of the regret

over y^n and $\underline{\theta}$ conditioned on x^n. Therefore the redundancy can be defined as follows,

$$R(x^n) = E_{y^n,\underline{\theta}|x^n}\left\{ Reg(x^n) \right\} \tag{7}$$

Plugging (1) and (5) into (6)

$$\begin{aligned} R(x^n) = &\sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^t\right) \\ &\sum_{t=1}^{n} \left(\ln p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) - \ln q\left(y_t|y^{t-1},x^t\right)\right) \end{aligned} \tag{8}$$

The objective of on-line learning is to find a sequential universal predictor which solves the following minimax problem,

$$R_{minimax}(x^n) = \min_{q_1,q_2,\ldots,q_n} \max_{\pi(\underline{\theta}|x^n)} R(x^n) \tag{9}$$

where $q_t = q(y_t|y^{t-1},x^t)$ is a universal predictor for y_t.

Note that the maximization is for $\pi(\underline{\theta}|x^n)$, i.e, there could be a different distribution for the parameters vector for each features sequence.

C. Main Result

Theorem 1: Capacity Redundancy Theorem for On-Line Learning Under a Certain Hypotheses Class:

1) The minimax redundancy problem which was defined in (9) for the hypotheses class of (1) is solved by the following on-line learner

$$q_t(y_t|y^{t-1},x^t) = \sum_{\theta_{x_t}} w(\theta_{x_t}) p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right)$$

where the weight $w(\theta_{x_t})$ depends on the current feature x_t and takes into account only previous labels with the same feature x_t,

$$w(\theta_{x_t}) = \frac{\pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)}{\sum_{\theta_{x_t}} \pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)}$$

2) There is an equivalence between the minimax redundancy and the sum of channel capacities.

$$R_{minimax}(x^n) = \sum_{j=1}^{K} C_j(x^n)$$

where,

$$C_j(x^n) = \max_{\pi(\theta_j|x^n)} I\left(\theta_j; \underline{y}_j^n|x^n\right)$$

which is the capacity of the channel between θ_j and $\underline{y}_j^n$ given x^n.

D. Observations

The resulting predictor is a multiplication of K independent mixture predictors for each sub sequence defined by the feature. One might think that this result is quite straightforward, since the conditional distribution factorizes based on the features. Therefore, there should not be any information gained from observing the labels associated with other features.

However, the values θ_j are unknown a-priori and might be correlated or in the extreme case, equal! Thus, in principle there might be something to be gained by observing the sub sequences associated with other features too. If two subsequences originate from sources with identical distributions, then jointly processing them might improve the estimation error of their properties.

Nevertheless, the theorem above states that in the minimax redundancy sense, the optimal predictor processes each sub sequence independently. This issue is the main focus of the proof in the next section.

Finally, it is important to note that the following theorem applies to any discrete valued function, $f(x^t)$ of the features x^t which partition the labels sequence.

III. Proof of the Capacity Redundancy Theorem with Side Information

The Capacity - Redundancy theorem in [2], shows the equivalence between *minimax redundancy* and the *capacity* (maximum mutual information) of the channel between $\underline{\theta}$ and the measurements y^n. We wish to find a similar equivalence for the case where the features are known sequentially. The proof technique in [2] cannot be used in this case since it will result in a different prior distribution for each time index. Therefore, the proof will be as follows. First, we will find the *maximin redundancy* solution. Then, we will propose a predictor whose redundancy will upper bound the *minimax redundancy*. Finally, we will show that this upper bound is equal to the *maximin redundancy* and thus we will be able to conclude that the proposed predictor achieves the *minimax redundancy* and that the *minimax redundancy* equals the sum of capacities.

A. Maximin Redundancy Solution

We can write the maximin problem,

$$\begin{aligned} R_{maximin}(x^n) = &\max_{\pi(\underline{\theta}|x^n)} \min_{q_1,q_2,q_3,\ldots,q_n} \\ &\sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) \\ &\sum_{t=1}^{n} \left(\ln p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) - \ln q\left(y_t|y^{t-1},x^t\right)\right) \end{aligned} \tag{10}$$

Since we minimize for each q_t separately, then we can write,

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta}|x^n)} \sum_{t=1}^{n} \min_{q_t} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) \left(\ln p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) - \ln q\left(y_t|y^{t-1},x^t\right)\right) \tag{11}$$

The conditional probability $q_t(y_t|y^{t-1},x^t)$ is a different function for each y^{t-1} and x^t. Therefore we can move the summation on y^n outside the $\min$, except for the summation on y_t and find $q_t(y_t|y^{t-1},x^t)$ for a specific sequence y^{t-1}.

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta}|x^n)} \sum_{t=1}^{n} \sum_{y^{t-1}} \min_{q_t} \sum_{y_t} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \Pi_{\{1\le j\le K|j\neq x_t\}} p^{\theta_j}\left(\underline{y}_j^t\right) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right) p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) \left(\ln p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) - \ln q\left(y_t|y^{t-1},x^t\right)\right) \tag{12}$$

We can write the inner minimization problem using the Kullback Leibler divergence, $D_{KL}(\cdot||\cdot)$,

$$q_t^* = \arg\min_{q_t} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^{t-1}\right) D_{KL}\left(p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right)||q\left(y_t|y^{t-1},x^t\right)\right) \tag{13}$$

The optimal q_t^* is thus,

$$q_t^*(y_t|y^{t-1},x^t) = \sum_{\underline{\theta}} \alpha_t(\underline{\theta}) p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) \tag{14}$$

where $\alpha_t(\underline{\theta})$ is defined as,

$$\alpha_t(\underline{\theta}) = \frac{\pi(\underline{\theta}|x^n)\Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^{t-1}\right) p^{\theta_j}\left(y_t|\underline{y}_j^{t-1}\right)}{\sum_{y_t}\sum_{\underline{\theta}} \pi(\underline{\theta}|x^n)\Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^{t-1}\right) p^{\theta_j}\left(y_t|\underline{y}_j^{t-1}\right)} \tag{15}$$

which is equivalent to,

$$q_t^*(y_t|y^{t-1},x^t) = \frac{\sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) p^{\underline{\theta}}\left(y^t|x^t\right)}{\sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) p^{\underline{\theta}}\left(y^{t-1}|x^{t-1}\right)} \tag{16}$$

Therefore, the universal predictor is basically a mixture of all the distributions in the model,

$$q^*(y^n|x^n) = \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) p^{\underline{\theta}}\left(y^n|x^n\right) \tag{17}$$

Now, using (17), we can sum up the logarithms in (12),

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta})} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) D_{KL}\left(p^{\underline{\theta}}\left(y^n|x^n\right)||q^*\left(y^n|x^n\right)\right) \tag{18}$$

Substituting (17) in (18),

$$R_{maximin}(x^n) = \max_{\pi(\underline{\theta}|X^n=x^n)} I\left(\underline{\theta}; Y^n|X^n=x^n\right) \tag{19}$$

Note that given x^n, we basically have K independent channels. Therefore, the distribution $\pi(\underline{\theta}|x^n)$ which maximizes (19) induces independence,

$$\pi(\underline{\theta}|x^n) = \Pi_{j=1}^{K} \pi(\theta_j|x^n) \tag{20}$$

We can now substitute (20) into (17) and get,

$$q^*(y^n|x^n) = \sum_{\underline{\theta}} \Pi_{j=1}^{K} \pi(\theta_j|x^n) p^{\theta_j}\left(\underline{y}_j^n\right) \tag{21}$$

Due to independence,

$$q^*(y^n|x^n) = \Pi_{j=1}^{K} \sum_{\theta_j} \pi(\theta_j|x^n) p^{\theta_j}\left(\underline{y}_j^n\right) \tag{22}$$

Then the maximin optimal universal predictor is in fact a multiplication of mixtures.

Also, plugging (20) into (16) gives us the maximin on-line predictor:

$$q_t^*(y_t|y^{t-1},x^t) = \sum_{\theta_{x_t}} w(\theta_{x_t}) p^{\theta_{x_t}}\left(y_t|\underline{y}_{x_t}^{t-1}\right) \tag{23}$$

where the weights $w(\theta_{x_t})$

$$w(\theta_{x_t}) = \frac{\pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)}{\sum_{\theta_{x_t}} \pi(\theta_{x_t}|x^n) p^{\theta_{x_t}}\left(\underline{y}_{x_t}^{t-1}\right)} \tag{24}$$

If we also insert (20) into (19) we get,

$$R_{maximin}(x^n) = \sum_{j=1}^{K} \max_{\pi(\theta_j|x^n)} I\left(\theta_j; \underline{y}_j^n|x^n\right) \tag{25}$$

B. Upper Bound on Minimax Redundancy

In this section we will find an upper bound for the minimax solution which will turn out to be equivalent to the maximin solution. We will then use the inequality $Maximin \le Minimax$ to prove that this solution is minimax optimal.

The minimax problem is defined in (9) and we propose to minimize over a smaller set of universal predictors, each of the following form,

$$q(y^n|x^n) = \Pi_{j=1}^{K} q_j\left(\underline{y}_j^n\right) \tag{26}$$

where $\underline{y}_j^n = \{y_i, 0 \le i \le n | x_i = j\}$.

We define the set Q as the set of all predictors which satisfy (26). Plugging (26) in to (9) we get,

$$\tilde{R}_{minimax}(x^n) = \min_{q(y^n|x^n)\in Q} \max_{\pi(\underline{\theta}|x^n)} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) \left(\ln \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) - \ln \Pi_{j=1}^{K} q_j\left(\underline{y_j}^n\right)\right) \tag{27}$$

where $R_{minimax}(x^n) \le \tilde{R}_{minimax}(x^n)$

Since the distributions in the logarithms can be factored,

$$\tilde{R}_{minimax}(x^n) = \min_{q(y^n|x^n)\in Q} \max_{\pi(\underline{\theta}|x^n)} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) \ln\left(\Pi_{j=1}^{K} \frac{p^{\theta_j}\left(\underline{y}_j^n\right)}{q_j\left(\underline{y}_j^n\right)}\right) \tag{28}$$

Using the fact that the summation of logarithms equals the logarithm of multiplication,

$$\tilde{R}_{minimax}(x^n) = \min_{q(y^n|x^n)\in Q} \max_{\pi(\underline{\theta}|x^n)} \sum_{\underline{\theta}} \pi(\underline{\theta}|x^n) \sum_{y^n} \Pi_{j=1}^{K} p^{\theta_j}\left(\underline{y}_j^n\right) \left(\sum_{j=1}^{K} \ln \frac{p^{\theta_j}\left(\underline{y}_j^n\right)}{q_j\left(\underline{y}_j^n\right)}\right) \tag{29}$$

Each logarithmic term in the summation depends only on $\underline{y}_j^n$ and θ_j, thus

$$\tilde{R}_{minimax}(x^n) = \min_{q(y^n|x^n)\in Q} \max_{\pi(\underline{\theta}|x^n)} \sum_{j=1}^{K} \sum_{\theta_j} \pi(\theta_j|x^n) \sum_{\underline{y}_j^n} p^{\theta_j}\left(\underline{y}_j^n\right) \ln \frac{p^{\theta_j}\left(\underline{y}_j^n\right)}{q_j\left(\underline{y}_j^n\right)} \tag{30}$$

We insert the maximization inside the summation and since Q is a family of factored distributions, we can minimize for each factor separately,

$$\tilde{R}_{minimax}(x^n) = \sum_{j=1}^{K} \min_{q(\underline{y}_j^n|\underline{x}_j^n)} \max_{\pi(\theta_j|x^n)} \sum_{\theta_j} \pi(\theta_j|x^n) D_{KL}\left(p^{\theta_j}\left(\underline{y}_j^n\right)||q_j\left(\underline{y}_j^n\right)\right) \tag{31}$$

The inner function $\sum_{\theta_j} \pi(\theta_j|x^n) D_{KL}\left(p^{\theta_j}\left(\underline{y}_j^n\right)||p\left(\underline{y}_j^n\right)\right)$ is convex-concave and thus the maximin problem is equal to the minimax problem for this specific term.

Therefore,

$$\tilde{R}_{minimax}(x^n) = \sum_{j=1}^{K} \max_{\pi(\theta_j|x^n)} \min_{q(\underline{y}_j^n|\underline{x}_j^n)} \sum_{\theta_j} \pi(\theta_j|x^n) D_{KL}\left(p^{\theta_j}\left(\underline{y}_j^n\right)||q\left(\underline{y}_j^n\right)\right) \tag{32}$$

The universal predictor which minimizes the convex combination of divergences is,

$$q\left(\underline{y}_j^n\right) = \sum_{\theta_j} \pi(\theta_j|x^n) p^{\theta_j}\left(\underline{y}_j^n\right) \tag{33}$$

Inserting (26) and (33)

$$q(y^n|x^n) = \Pi_{j=1}^{K} \sum_{\theta_j} \pi(\theta_j|x^n) p^{\theta_j}\left(\underline{y}_j^n\right) \tag{34}$$

Therefore (34) is equivalent to (22). Once we insert (33) into (31) we get,

$$\tilde{R}_{minimax}(x^n) = \sum_{j=1}^{K} \max_{\pi(\theta_j|x^n)} I\left(\theta_j; \underline{y}_j^n|x^n\right) \tag{35}$$

However, (35) is equal to (25) and the upper bound on the minimax is equal to the maximin solution, thus the maximin solution is equal to the minimax solution.

IV. Conclusion

In this paper, a Capacity Redundancy theorem for on-line learning was proven, under a hypotheses class which partitions the labels sequence based on the features which are the output of an arbitrary finite state machine. We have also developed an on-line universal predictor that can achieve the *minimax redundancy* for an appropriate choice of prior distribution. It was shown that even if the sub sequences originated from identical distributions, there is nothing to be gained from jointly processing them and the optimal predictor (in the minimax redundancy sense) is composed of independent predictors for each feature value.

Moreover, we have shown that the universal predictors proposed in [12] and [13] are in-fact special cases of the predictor in this paper, which according to our proof achieves the *minimax redundancy* even non asymptotically.

References

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

[2] R. G. Gallager, "Source coding with side information and universal coding, unpublished manuscript," 1979.

[3] L. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Transactions on Information Theory*, vol. 26, no. 2, pp. 166–174, 1980.

[4] B. Y. Ryabko, "Coding of a source with unknown but ordered probabilities," *Problems of Information Transmission*, vol. 15, no. 2, pp. 134–138, 1979.

[5] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 714–722, 1995.

[6] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[7] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced lectures on machine learning*. Springer, 2004, pp. 169–207.

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

[9] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[10] S. Ben-David, D. Pál, and S. Shalev-Shwartz, "Agnostic online learning," 2009.

[11] A. Rakhlin, K. Sridharan, and A. Tewari, "Online learning: Stochastic, constrained, and smoothed adversaries," in *Advances in neural information processing systems*, 2011, pp. 1764–1772.

[12] 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.

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