Universal Batch Learning with Log-Loss

Yaniv Fogel
School of Electrical Engineering
Tel Aviv University
Email: yanivfogel@mail.tau.ac.il

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

***Abstract*—In this paper we consider the problem of batch learning with log-loss, in a stochastic setting where given the data features, the outcome is generated by an unknown distribution from a class of models. Utilizing the minimax theorem and information-theoretical tools, we came up with the minimax universal learning solution, a redundancy capacity theorem and an upper bound on the performance of the optimal solution. The resulting universal learning solution is a mixture over the models in the considered class. Furthermore, we get a better bound on the generalization error that decays as $O(\log N/N)$, where N is the sample size, instead of $O(\sqrt{\log N/N})$ which is commonly attained in statistical learning theory for the empirical risk minimizer.**

I. Introduction

The recent years have seen a soaring interest in the field of inductive learning. Broadly speaking, inductive learning aims at utilizing past training examples of data features and outcomes into a good prediction of a new outcome, given its respective data feature. This is usually done for some loss function, that measures the accuracy of the prediction, and with respect to some models or hypotheses class, that represents possible connections between the data features and the outcomes.

A. Previous Results

Arguably the main theoretical framework for inductive learning is statistical learning theory, see e.g. [1],[2]. In this framework, the learner is given a batch of examples y^{N-1}, x^{N-1}, and some hypotheses class Θ, and is requested to form an hypothesis to predict a new outcome y_N, given the respective feature x_N, with a low expected loss.

A central concept in statistical learning theory is the Vapnik-Chervonenkis (VC) dimension, originally presented in [3]. Broadly speaking, the VC-dimension characterizes the richness of the hypotheses class from which the learner has to choose. If the VC-dimension is finite, there exist a uniform bound for the deviation between the empirical loss over the training data and the expected performance over a new data, for all possible hypotheses. Thus, one is guaranteed that the hypothesis with the best performances over the observed data will have a good performance with respect to new examples. Basically this is the reason why batch learning is mainly done by choosing the empirical risk minimizer.

Another field of interest is distribution estimation, where one is given a batch of outcomes y^{N-1}, and is requested to evaluate the probability for the next outcome y_N, see [4]. The measure of performance in this problem is often the regret, which is the difference between the loss associated with the estimated probability and that of the true probability. This is a simplified learning problem with no data features; however, its simplicity may provide a useful intuition for the more complicated learning problem.

In addition, we note that the task of learning can be thought of as a zero-sum game between the learner and nature, where the learner wants to minimize its expected loss, while nature may try to maximize it by choosing the data generating distribution in a possibly adversarial manner. Thus, the minimax theorems that provides sufficient conditions under which the min-max regret equals the max-min regret are useful. In this work we will mainly use the results introduced in [5].

Yet another relevant problem is universal coding or universal prediction with log-loss, where one has to assign a probability for the whole sequence y^N without knowing the true hypothesis according to which y^N is generated, see [6]. As in the distribution estimation problem, there are no data features. A fundamental result in this problem is the redundancy-capacity Theorem, see [7] and [8]. As we will show, our results in this paper may be viewed as an extension of the redundancy-capacity theorem for the single outcome prediction problem and furthermore for the single outcome batch learning problem where a data feature is given.

B. Main Contributions

In this paper we determine the minimax universal solution for the batch learning problem with log-loss, establish a redundancy-capacity theorem for it, and derive an upper bound for its expected regret.

We go on and show that in many cases the bound we derived guarantees a convergence rate of order $\frac{d\log N}{N}$, where d is the VC-dimension and N is the sampling size.

C. Paper Outline

Following the introduction, in section II we give the formal problem definitions for each of the problems to be discussed. Following that, in sections III and IV we discuss the problems of batch learning without and with data features, respectively. In section V we discuss two additional possible formulations of the batch learning problem and in section VI we compare our results to previously known results. We finally conclude and present possible directions for future research and mention briefly an extension of the results to the individual setting.

II. Formal Problem Definition

A. Learning without Data Features

Let us first formally define the simplified problem of learning without data features: First, both the learner and nature are given the outcomes space Y and a hypotheses class Θ, where each hypothesis defines a probability over the next outcome, $p_\theta(y)$. Next, the learner announces its "universal" probability assignment for a new outcome y_N given the training set y^{N-1}, which we denote $p_u(y_N|y^{N-1})$. Then, nature chooses some distribution over the hypotheses, $w(\theta)$. Note that this choice of $w(\theta)$ includes a deterministic choice of a particular θ. The regret will be the expected KL-divergence between $p_\theta(y_N)$ and $p_u(y_N|y^{N-1})$:

$$R(w(\theta),p_u)=\int_\theta w(\theta)\sum_{y^N}p_\theta(y^N)\log\left(\frac{p_\theta(y_N)}{p_u(y_N|y^{N-1})}\right) \quad (1)$$

where $p_\theta(y^N)=\prod_{t=1}^N p_\theta(y_t)$. We will be interested in the min-max regret:

$$R^*=\min_{p_u}\max_{w(\theta)}R(w(\theta),p_u) \quad (2)$$

To compare our results to results in statistical learning theory and the probably approximately correct (PAC) learning model, we will also discuss the probability over the training set that the expected regret will be greater than some ε.

B. Learning with Data Features

In the case of learning with data features we will consider a setting where the data features sequence $\tilde{x}^N$ is arbitrary and known both to the learner and nature. To avoid singular cases, the regret to be analysed will be the expected regret over all permutations of $\tilde{x}^N$, with uniform probability for each permutation:

$$\begin{aligned}R(w(\theta),p_u,\tilde{x}^N)&=E_{\theta,x^N,y^N}\left[\log\left(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})}\right)\right]\\&=\int_\theta\frac{w(\theta)}{N!}\sum_{x^N=perm(\tilde{x}^N)}\sum_{y^N}p_\theta(y^N|x^N)\log\left(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})}\right)\end{aligned} \quad (3)$$

where the notation $x^N=perm(\tilde{x}^N)$ is used to emphasize that x^N is some permutation of $\tilde{x}^N$. Note that $w(\theta)$ may depend on $\tilde{x}^N$, but it should be the same for each permutation. Note that under these permutations, each data feature $x_t\in\tilde{x}^N$ has an equal chance of being the one on which we will be tested.

As in the previous setting, here too we will be interested in the min-max regret:

$$R^*(\tilde{x}^N)=\min_{p_u}\max_{w(\theta)}R(w(\theta),p_u,\tilde{x}^N) \quad (4)$$

Before moving on, we note that one can come up with alternative formulations for this learning problem, with other assumptions regarding the feature sequence x^N. We will discuss two of those alternative formulations in section V.

III. Batch Learning Without Data Features

Our main result for the problem of batch learning without data features will be as follows:

Theorem 1. *The min-max regret of the problem defined in II-A satisfies:*

$$R^*=I(Y_N;\theta|Y^{N-1})\le\frac{1}{N}I(Y^N;\theta) \quad (5)$$

where $w(\theta)=\arg\max_{w(\theta)}I(Y_N;\theta|Y^{N-1})$.

Proof. Our first step in the analysis of the minimax regret would be to note that R is linear with respect to $w(\theta)$ for every p_u, and convex with respect to p_u for every $w(\theta)$, which is rather straight-forward.

Since the problem is indeed convex-concave and the minimax theorem can be applied, let us consider the optimal probability assignment p_u for a given $w(\theta)$. To this end, we will define the corresponding Lagrangian:

$$L=R+\sum_{y^{N-1}}\lambda_{y^{N-1}}\sum_{y_N}p_u(y_N|y^{N-1}) \quad (6)$$

where the Lagrange multipliers are due to the constraint $\sum_{y_N}p_u(y_N|y^{N-1})=1$. Next, we will find the optimal probability assignment by taking the derivative:

$$\begin{aligned}&\frac{\partial L}{\partial p_u}=-\int_\theta w(\theta)\frac{p_\theta(y^N)}{p_u(y_N|y^{N-1})}+\lambda_{y^{N-1}}=0\\&p_u(y_N|y^{N-1})=\frac{\int_\theta w(\theta)p_\theta(y^N)}{\lambda_{y^{N-1}}}=\frac{\int_\theta w(\theta)p_\theta(y^N)}{\int_\theta w(\theta)p_\theta(y^{N-1})}\end{aligned} \quad (7)$$

where we used the constraint $\sum_{y_N}p_u(y_N|y_{N-1})=1$ to get $\lambda_{y^{N-1}}$. Naturally, the induced solution can be obtained by a simple Bayes rule. This observation also allows us to interpret the minimax regret as the following mutual information:

$$\begin{aligned}R^*&=\max_{w(\theta)}\int_\theta w(\theta)\sum_{y^N}p_\theta(y^N)\log(\frac{p_\theta(y_N)}{p(y_N|y^{N-1})})\\&=\max_{w(\theta)}I(Y_N;\theta|Y^{N-1})\end{aligned} \quad (8)$$

The quantity in (8) may be hard to evaluate. Instead, we upper bound it as follows. First, we note that for all $w(\theta)$:

$$\begin{aligned}&I(Y_{t+1};\theta|Y^t)=H(Y_{t+1}|Y^t)-H(Y_{t+1}|Y^t,\theta)\\&\le H(Y_{t+1}|Y^{t-1})-H(Y_{t+1}|\theta)=H(Y_t|Y^{t-1})-H(Y_t|\theta)\\&=I(Y_t;\theta|Y^{t-1}).\end{aligned} \quad (9)$$

which immediately leads to the following bound:

$$R^*=I(Y_N;\theta|Y^{N-1})\le\frac{1}{N}\sum_{t=1}^N I(Y_t;\theta|Y^{t-1})=\frac{I(Y^N;\theta)}{N} \quad (10)$$

□

The mutual information $I(Y^N;\theta)$ has been investigated and evaluated in several interesting cases. For example, when $\theta\in$

$\mathbb{R}^k$ and $p_\theta(y)$ is twice continuously differentiable w.r.t θ, then $I(Y^N;\theta) = O(k\log(N))$, see [9], and the regret is bounded by $O\left(\frac{k\log(N)}{N}\right)$, where k is the number of parameters.

IV. Batch Learning with the minimax approach

A. Random Order over Data Features

Following the analysis section III, we will use a similar derivation for the problem with data features:

Theorem 2. *The min-max regret of the problem defined in II-B satisfies:*

$$R^*(\tilde{x}^N) = \frac{1}{N!}\sum_{x^N=perm(\tilde{x}^N)} I(Y_N;\theta|x^N,Y^{N-1}) \leq \frac{1}{N} I(Y^N;\theta|\tilde{x}^N). \tag{11}$$

$$where\ w(\theta) = \underset{w(\theta)}{arg\,max}\ \tfrac{1}{N!}\textstyle\sum_{x^N=perm(\tilde{x}^N)} I(Y_N;\theta|x^N,Y^{N-1}).$$

Proof. Again, using the minimax theorem and Lagrange multipliers, we get a Bayesian solution:

$$p_u(y_N|x^N,y^{N-1}) = \frac{\int_\theta w(\theta)p_\theta(y^N|x^N)}{\int_\theta w(\theta)p_\theta(y^{N-1}|x^{N-1})}. \tag{12}$$

Now, in this problem we get that the regret is simply the average mutual information between y_N and θ given x^N, y^{N-1}, where the average is over all permutations of $\tilde{x}^N$:

$$\begin{aligned} R &= \frac{1}{N!}\sum_{x^N=perm(\tilde{x}^N)} \int_\theta w(\theta)\sum_{y^N} p_\theta(y^N|x^N)\log\left(\frac{p_\theta(y_N|x_N)}{p(y_N|x^N,y^{N-1})}\right) \\ &= \frac{1}{N!}\sum_{x^N=perm(\tilde{x}^N)} I(Y_N;\theta|x^N,Y^{N-1}) \end{aligned} \tag{13}$$

Before moving on to bound the regret, we note that $p(y_N|x^N,y^{N-1})$ is unaffected by permutations of x^{N-1}. Thus, instead of averaging over all $N!$ permutations of $\tilde{x}^N$, one may simply average over the N possible choices of x_N.

Similarly to section III, we will show that the average mutual information $I(Y_{t+1};\theta|Y^t,x^N)$ is decreasing with t:

$$\begin{aligned} &\sum_{x^N=perm(\tilde{x}^N)} \left[I(Y_{t+1};\theta|Y^t,x^N) - (Y_t;\theta|Y^{t-1},x^N)\right] \\ &= \sum_{x^N=perm(\tilde{x}^N)} \left[H(Y_{t+1}|Y^{t-1},x^N) - H(Y_{t+1}|\theta,Y^t,x^N)\right] \\ &- \sum_{x^N=perm(\tilde{x}^N)} \left[H(Y_t|Y^t,x^N) + H(Y_t|Y^{t-1},\theta,x^N)\right] \\ &= \sum_{x^N=perm(\tilde{x}^N)} \left[H(Y_{t+1}|Y^t,x^N) - H(Y_t|Y^{t-1},x^N)\right] \\ &+ \sum_{x^N=perm(\tilde{x}^N)} \left[H(Yt|\theta,x_t) - H(Y_{t+1}|\theta,x_{t+1})\right] \\ &\overset{(b)}{=} \sum_{x^N=perm(\tilde{x}^N)} \left[H(Y_{t+1}|Y^t,x^N) - H(Y_t|Y^{t-1},x^N)\right] \\ &\leq \sum_{x^N=perm(\tilde{x}^N)} \left[H(Y_{t+1}|Y^{t-1},x^N) - H(Y_t|Y^{t-1},x^N)\right] \overset{(c)}{=} 0 \end{aligned} \tag{14}$$

where both (b) and (c) are due to the fact that for each $x \in \tilde{x}^N$, the probability that $x_t = x$ equals to the probability that $x_{t+1} = x$; thus, summing up over the permutations cancels out both $H(Y_{t+1}|\theta,x_{t+1}) - H(Y_t|\theta,x_t)$ and $H(Y_{t+1}|Y^{t-1},x^N) - H(Y_t|Y^{t-1},x^N)$.

This leads to the following bound:

$$\begin{aligned} R^* &= \frac{1}{N!}\sum_{x^N=perm(\tilde{x}^N)} I(Y_N;\theta|x^N,Y^{N-1}) \\ &\leq \frac{1}{N*N!}\sum_{x^N=perm(\tilde{x}^N)}\sum_{t=1}^{N} I(Y_t;\theta|x^N,Y^{t-1}) \\ &= \frac{1}{N*N!}\sum_{x^N=perm(\tilde{x}^N)} I(Y^N;\theta|x^N) \end{aligned} \tag{15}$$

Since we assumed that $p_\theta(y^N|x^N) = \prod_{t=1}^N p_\theta(y_t|x_t)$, we get that the mutual information $I(Y^N;\theta|x^N)$ is not effected by the permutations, and thus:

$$R^* \leq \frac{1}{N} I(Y^N;\theta|\tilde{x}^N). \tag{16}$$

□

This result states that the expected loss may be bounded by the mutual information between y^N and θ given the whole sequence x^N. Let us consider this mutual information for hypotheses classes of the form considered in [10], where each data feature is assigned to one of two groups, and for each group there is a different probability over y: In this case, assuming that the VC dimension of the hypotheses class is a finite d, we get:

$$R \leq \frac{\max_{x^N} I(Y^N;\theta|x^N)}{N} \leq \frac{d\log(eN) + 2\log(N)}{N} \tag{17}$$

where the $d\log(eN)$ is due to the number of possible partitions, which is bounded by $(\frac{eN}{d})^d$ by Sauer's lemma, and the $2\log(N)$ is due to the uncertainty in the probability assignment for each group.

B. PAC-results

We have bounded the expected regret over both the training set and the test sample. This is in contrast with the PAC model, that introduces a bound over the probability that the expected regret will be larger then some ε, where the probability is with respect to the training set and the expectation is with respect to the test sample.

Nevertheless, one may achieve PAC-like results using the method described above. To this end, note that the regret can also be written as follows:

$$R = E_{\theta,x^N,y^{N-1}}\left[D\left(p_\theta(y_N|x_N)||p_u(y_N|x_N,y^{N-1},x^{N-1})\right)\right] \tag{18}$$

where $D(p_\theta||p_u)$ is the KL-divergence, and the sequences x^{N-1} and y^{N-1} are the training set. Since the KL-divergence is non-negative, by Markov's inequality:

$$Pr\{D(p_\theta||p_u) \geq \varepsilon\} \leq \frac{R}{\varepsilon}. \quad (19)$$

Returning to the problems mentioned above, where each hypothesis assigns the data feature into a group and assigns some probability function for each group, we get:

$$Pr\{D(p_\theta||p_u) \geq \varepsilon\} \leq \frac{(d+2)\log(N)}{N*\varepsilon}. \quad (20)$$

Alternatively, we get that with probability of at least $1-\delta$ the following holds:

$$D(p_\theta||p_u) \leq \frac{(d+2)\log(N)}{N*\delta}. \quad (21)$$

V. Alternative Formulations

As mentioned, one may consider several other formulations regarding the data feature sequence x^N for the batch learning problem.

A. i.i.d Data Features

Similarly to standard assumptions in batch learning, one may assume that the data features are generated by an unknown i.i.d process with distribution $p(x)$. In this setting, the regret satisfies:

$$R_{i.i.d}(w(\theta), p_u, p(x)) = E_{\theta,x^N,y^N}\left[\log\left(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})}\right)\right]$$
$$= \int_{\theta,x^N} w(\theta)\prod_{t=1}^{N} p(x_t)\sum_{y^N} p_\theta(y^N|x^N)\log\left(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})}\right) \quad (22)$$

Intuitively, this should be an easier task then the one we have considered above: On one hand, if the x^N are generated by an i.i.d process then their order is obviously random. On the other hand, in our previous result we allowed the prior $w(\theta)$ to be dependent on the specific values $\tilde{x}^N$, where here it can only depend on $p(x^N)$. So, for this case, we prove the following theorem:

Theorem 3. *For every $p(x)$, the min-max regret of (22) is bounded by:*

$$R^*_{i.i.d}(p(x)) \leq \frac{1}{N} I(Y^N;\theta|X^N) \quad (23)$$

where $w(\theta) = \underset{w(\theta)}{arg\,max}\ \frac{1}{N!}\sum_{x^N=perm(\tilde{x}^N)} I(Y_N;\theta|x^N,Y^{N-1})$.

Proof. Let us use the results obtained in Theorem 2:

$$R^*_{i.i.d}(p(x)) =$$
$$\min_{p_u}\max_{w(\theta)} \int_{\theta,x^N} w(\theta)p(x^N)\sum_{y^N} p_\theta(y^N|x^N)\log(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})})$$
$$\overset{(d)}{=} \min_{p_u}\max_{w(\theta)} \int_{\theta,\tilde{x}^N} \frac{w(\theta)p(\tilde{x}^N)}{N!}\sum_{x^N} E_{Y^N}\log(\frac{p_\theta(Y_N|x_N)}{p_u(Y_N|x^N,Y^{N-1})})$$
$$\overset{(e)}{\leq} \min_{p_u}\int_{\tilde{x}^N} p(\tilde{x}^N)\max_{w(\theta)}\int_\theta \frac{w(\theta)}{N!}\sum_{x^N} E_{Y^N}\log(\frac{p_\theta(Y_N|x_N)}{p_u(Y_N|x^N,Y^{N-1})})$$
$$\overset{(f)}{\leq} \int p(\tilde{x}^N)I(Y^N;\theta|\tilde{x}^N) = I(Y^N;\theta|X^N) \quad (24)$$

where (d) is due to the fact that since the $x^N \overset{i.i.d}{\sim} p(x)$, then for every pair $x^N,\tilde{x}^N$ that are the same up to a permutation, $p(x^N) = p(\tilde{x}^N)$[1]. Inequality (e) holds since we now allow $w(\theta)$ to depend on the specific $\tilde{x}^N$, and (f) holds since we can use the optimal probability assignment for the random permutation problem. □

Clearly, a PAC performance bound can also be obtained in a similar way to that used in the random permutation formulation.

B. Individual Data Features Sequence

One may wonder if the permutations are really necessary, and try to derive similar results for a setting where the learning is done with respect to a specific data features sequence:

$$R^*_{ind} = \min_{p_u}\max_{w(\theta)}\int_\theta w(\theta)\sum_{y^N} p_\theta(y^N|x^N)\log(\frac{p_\theta(y_N|x_N)}{p_u(y_N|x^N,y^{N-1})}). \quad (25)$$

Note that one may also apply the minimax theorem and derive a Bayesian probability assignment in this setting. The main difference with respect to the setting in section IV is that now we cannot apply (14), and thus cannot get a bound which is based on the mutual information $I(Y^N;\theta|x^N)$.

Our next example will show that this is not a mere technicality, and that in many cases this problem in not learnable: Consider the hypotheses class of probabilistic one-dimensional barrier threshold, where each θ consists of a barrier parameter b, and the two probability assignments for each of the two groups. For binary outcome space $Y = \{0,1\}$, this can be written as follows:

$$p_{\theta=(b,p_1,p_2)}(y_t = 1|x_t) = \begin{cases} p_1 & \text{if } x_t \leq b \\ p_2 & \text{if } x_t > b \end{cases} \quad (26)$$

Now, assume that the data features sequence is some x^N, and denote by x_l,x_r the samples that are closest to x_N from

[1] This indicates that any exchangeable distribution on x^N may be assumed

the left and right side, accordingly. Consider an adversary that chooses $w(\theta)$ as follows:

$$w(b=\frac{xl+x_N}{2}, p_1=0, p_2=1)=\frac{1}{2} \quad (27)$$

$$w(b=\frac{xr+x_N}{2}, p_1=0, p_2=1)=\frac{1}{2} \quad (28)$$

Notice that the outcomes sequence y^{N-1} will always be the same: $y_t = I(x_t > x_N)$. Now, even if the learner knows $w(\theta)$, we get that $p(y_N=1|x^N,y^{N-1}) = Pr\{x_N > b\} = \frac{1}{2}$. Since this is the Bayesian probability assignment, which is optimal, we get that for this problem:

$$\begin{aligned} &R^*_{ind} \geq R_{ind}(w,q^*) = I(Y_N;\theta|Y^{N-1},x^N) \\ &= H(Y_N|x^N,Y^{N-1}) - H(Y_N|\theta,x^N,Y^{N-1}) = 1-0=1. \end{aligned} \quad (29)$$

Clearly, the regret does not converge to zero as $N \to \infty$.

VI. Comparison to Previous Results

A. Empirical Risk minimization

A classical, basic result in learning theory considers a learner based on the empirical risk minimizer and bound its regret by finding a uniform bound on the difference between the empirical loss and the expected loss. Note that for log-loss one should choose a version of the empirical risk minimizer which is bounded away from the edges to avoid unbounded loss. Using this uniform bound (usually obtained by Hoeffding inequality), an error bound of order $\sqrt{\frac{d\log N+\log\frac{1}{\delta}}{N}}$ can be obtained following, e.g., [1],[2].

Comparing this error bound to the error bound of order $\frac{d\log N}{N\delta}$ found in this paper, it is clear that our result achieves a better convergence rates for any fixed δ. Furthermore, the expected generalization error we attain decays as $O(\log N/N)$, instead of $O(\sqrt{\log N/N})$ attained by the uniform bound on the performance of the empirical risk minimizer. Nevertheless, if one wishes in the PAC formulation to have $\delta \to 0$ as $N \to \infty$, at high enough rate, our result may yield a slower convergence rate. In particular for $\delta \sim \frac{1}{N}$, the bound (21) does not lead to a vanishing error, while the empirical risk minimizer converges. We believe, however, that since we have used the weak Markov's inequality our PAC-like bound can probably be improved .

B. Add-β Universal Batch Predictor

In the problem of probability estimation where there are no data features and the hypotheses class Θ is the probability simplex over Y, a classical predictor is the add-β predictor where a constant β is added to the empirical counts in the training. This estimator follows the well-known Bayes reasoning with a Dirichlet prior, see, e.g., [11]. Now, it was shown in [12] that the estimator's expected regret is of order $\frac{k-1}{N}$, where k is the alphabet size. Yet, our results provide a bound of $\frac{(k-1)\log N}{2N}$. The reason, of course, is that we loosely upper bounded $\sum_{t=1}^{N} I(Y_t;\theta|Y^{t-1})$ by $I(Y^N;\theta)$. This suggests that our results may be improved for the general learning problem.

VII. Conclusions and Future Research Directions

In this paper we discussed the problem of batch learning with respect to the logarithmic loss function. A min-max solution was introduced. This solution has a Bayesian flavor, and yields an upper bound over the expected regret which, for model classes with finite VC dimension d, is of order $\frac{d\log(N)}{N}$, both when the data features are generated by an i.i.d process and when the data features are arbitrary but introduced to the learner in a random order.

There are quite a few topics for further research following the approach suggested in this paper. For example, on the technical side - improving the regret bound. Another question is what can be the performance if the learner (as in the empirical risk minimizer) must choose one of the hypotheses. A more fundamental question is how to choose the model class: on one hand we want that the model class can approximate well the true behavior, but on the other hand it should be restricted enough to be learned. This brings the question: can the presented approach be extended to the case where the true behavior is not part of the hypotheses class?

The last point may hint to an even more ambitious goal - batch learning in the individual setting. Incidentally, this is the subject of our latest work [13].

Finally, this work and these additional challenges are part of the quest to establish information theory approach in the theoretical framework of statistical learning.

References

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

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

[3] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.

[4] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Conference on Learning Theory*, 2015, pp. 1066–1100.

[5] K. Fan, "Minimax theorems," *Proceedings of the National Academy of Sciences*, vol. 39, no. 1, pp. 42–47, 1953.

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

[7] 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.

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

[9] B. S. Clarke and A. R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," *Journal of Statistical planning and Inference*, vol. 41, no. 1, pp. 37–60, 1994.

[10] Y. Fogel and M. Feder, "On the problem of on-line learning with log-loss," in *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 2995–2999.

[11] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[12] R. E. Krichevskiy, "Laplace's law of succession and universal encoding," *IEEE Transactions on information theory*, vol. 44, no. 1, pp. 296–303, 1998.

[13] Y. Fogel and M. Feder, "Batch learning in the individual setting," *Draft*, 2018.