Combining Batch and Online Prediction

Yaniv Fogel
School of Electrical Engineering
Tel-Aviv University
Email: Yaniv.fogel8@gmail.com

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

***Abstract*—We study a variation of the stochastic, realizable batch learning problem where there is a training set of N symbols and the prediction is then tested over L symbols. We prove an equivalent of the Redundancy-Capacity Theorem, find the leading term of the regret for the multinomial case and also discuss, informally, a general parametric hypothesis class. We implement a variant of the Arimoto-Blahut algorithm to calculate the optimal minimax redundancy and show, for the binary case, the resulting regret and the approximated capacity-achieving prior.**

I. Introduction

Universal prediction is a well-established and researched problem; see, e.g., [1] for a survey paper. In the information theory approach, prediction is done by providing a probability distribution for the predicted outcome, and the incurring loss is the information loss or the log loss. There are several variations for this problem, which are now outlined.

First, there is a hypothesis class, which in the information-theoretic approach is a class of models or probability distributions $\{P_\theta(\cdot), \theta \in \Theta\}$ used to explain the data. The various settings differ on the assumption of how the data is generated. In the stochastic setting, it is assumed that the data comes from some unknown distribution. In the stochastic realizable case, the unknown distribution is one of the models in the hypothesis class, whereas in the mis-matched, or unrealizable case, the data distribution is not necessarily in the class but may belong to a larger set of distributions. In the individual setting, on the other hand, nothing is assumed of the data, and it is an arbitrary individual sequence. In the stochastic setting, the goal of the universal predictor is to attain the same performance as a predictor that knows the true distribution. In the individual setting, the goal is to attain the performance of a predictor that knows the sequence but must use a model from the class.

Another variation comes from the distinction between batch and online. The classical work on universal prediction assumes online prediction, which, in the case of prediction with log-loss, is equivalent to assigning a probability and accumulating the loss over the entire sequence. On the other hand, in batch prediction, there is a training sequence, and prediction is performed and tested on a single new symbol. Batch prediction is the common situation in machine learning, but it did not receive much attention in information-theoretic universal prediction until recently, [2]. The goal of this paper is to consider a case between online and batch: there is a training sequence of size N, and then the predictor continues online to predict L symbols, or equivalently in prediction with log-loss, to allocate a probability for L test symbols.

This combination of training and test, between batch and online, can be analyzed in both the stochastic and individual settings. Furthermore, universal prediction can be extended to universal supervised learning where the sequence is composed not only of "labels" to be predicted, denoted $\underline{y}$, but there is also a feature sequence $\underline{x}$ from which the labels are learned in the training and the test. In this paper we consider the simplest case of a realizable stochastic setting with no feature data, only label symbols to predict. Clearly, the additional variations can be defined and analyzed in future work.

The paper is organized as follows: Following a brief review of previous work, we present the formal problem statement. Then, we show a min-max theorem for our problem, present the asymptotic min-max regret for the multinomial case, and discuss a general parametric hypothesis class. The min-max regret is given as a supremum over a prior on the hypothesis class of a conditional mutual information. Thus, we then present an Arimoto-Blahut algorithm that evaluates that "capacity-achieving" prior and shows some of its results for the binary case. Last, we conclude and present directions for future work.

II. Previous Works

A. Online Prediction

Perhaps the more established case to this date is the one where there is no training set. In this case, the probability assignment for a whole sequence of size L, $y^L = y_1, \ldots, y_L$ can be decomposed into a sequential probability assignment:

$$q(y^L) = \prod_{t=1}^{L} q(y_t | y^{t-1}).$$

One can define the following min-max problem:

$$R^*_{online} = \min_{q(\cdot)} \max_{\theta} \frac{1}{L} \sum_{y^L} p_\theta(y^L) \log \left(\frac{p_\theta(y^L)}{q(y^L)} \right). \quad (1)$$

Arguably, the most fundamental result in this setting is the celebrated Capacity-Redundancy theorem due to [3], [4] and [5]. According to this Theorem, the min-max optimal regret is equal to the capacity of a channel from Θ to y^L:

$$R^*_{online} = \max_{w(\theta)} \frac{1}{L} I(Y^L; \Theta). \quad (2)$$

where the maximization is over all valid distributions over Θ, which defines the hypothesis class $\left\{P_\theta(y^L), \theta \in \Theta\right\}$. We also get that the optimal probability assignment is a mixture $q(y^L) = \int_\theta w^*_{online}(\theta) p_\theta(y^L) d\theta$ of the models in the class, where the mixture weight $w^*_{online}(\theta) = \arg\max_{w(\theta)} I(Y^L;\Theta)$ is the capacity-achieving prior.

Following this result, it was shown in [6] that under some mild regularity conditions, if $\theta \in \mathbf{R}^k$ then $R^*_{online} = \frac{k\log(L)}{2L} + o\left(\frac{\log(L)}{L}\right)$. Interestingly, this leading term is achieved for all smooth prior distributions over θ, and thus, the capacity achieving prior usually maximizes the second term.

There are many results regarding the capacity-achieving prior. First, if we only consider "smooth" priors, it was shown in [7] - again, under some rather mild regularity conditions - that Jefferys' prior achieves the maximal mutual information. However, for any finite L, it was shown in [8] that the capacity achieving prior is, in fact, discrete and takes non-zero values at a finite number of points $\theta_0, \theta_1, ..., \theta_K$.

The problem of theoretically identifying the capacity-achieving prior might prove to be challenging for many channels. Nevertheless, there are numerical methods of evaluating this prior and the resulting capacity. Notably, [9] and [10] proposed a method for computing the capacity achieving prior for discrete channels using alternating maximization. This concept has been further developed in many papers, including [11], where a variation of the Arimoto-Blahut algorithm that achieves faster convergence rates using a one-step iteration has been proposed.

The Arimoto-Blahut algorithm and its variations have also been used specifically to study the capacity-achieving prior of $I(Y^L;\Theta)$, see, for example, [12] and [13].

B. Batch Learning

The problem of batch learning, common in machine learning, is perhaps less studied in the information-theoretic approach than online learning. In batch learning, there are N training symbols y^N, and the goal is to predict an unknown test symbol, y_{N+1}. In the realizable, stochastic setting of batch prediction, the problem can be formulated as follows:

$$R^*_{batch} = \min_{q(\cdot|\cdot)} \max_\theta \sum_{y^N} p_\theta(y^N) \sum_{y_{N+1}} p_\theta(y_{N+1}|y^N) \log\left(\frac{p_\theta(y_{N+1}|y^N)}{q(y_{N+1}|y^N)}\right)$$
$$= \min_{q(\cdot|\cdot)} \max_\theta \sum_{y^{N+1}} p_\theta(y^{N+1}) \log\left(\frac{p_\theta(y_{N+1}|y^N)}{q(y_{N+1}|y^N)}\right).$$

Let us first consider the multinomial case where $y \in \{0, ..., m-1\}$ and Θ is the simplex over an alphabet of size m, i.e. $\Theta = \{\theta \in \mathbf{R}^m : \sum_{j=0}^{m-1} \theta[j] = 1, 0 \le \theta[j] \le 1\}$. Perhaps the earliest work in this case is due to [14], who considered "add-β" rules of the form $q(y_{N+1} = j|y^N) = \frac{n_j+\beta}{N+m\cdot\beta}$, where n_j is the number of appearances of the j-th letter in the training set y^N. In this case, [14] concluded that the min-max optimal β leads to a regret of order $\frac{0.509(m-1)}{N} + o\left(\frac{1}{N}\right)$, and showed a lower bound of $\frac{m-1}{2N} + o\left(\frac{1}{N}\right)$ using a uniform prior over Θ. The gap between the lower and the upper bound was closed in [15], who deviated from the "add-β" rule and used a predictor that is essentially a "add-$\frac{3}{4}$" probability assignment but with some modifications at the edges when $n_j \in \{0, 1\}$. Interestingly, in [16] it was shown that when the edges are excluded from Θ, an "add-$1+\sqrt{\frac{1}{6}}$" rule is both min-max and max-min optimal, leading to $\frac{1}{2N} - \frac{2.06...}{N^2} + o\left(\frac{1}{N^2}\right)$ regret. A numerical analysis of the capacity-achieving prior for the batch learning case, for a variety of test set sizes, has been conducted in [17], with a relatively small number of values of θ allowed. A more elaborated numerical analysis, including a variation of the Arimoto-Blahut algorithm, has recently been conducted in [18]. In addition, a min-max theorem for this batch learning case, as well as cases where there is also a feature sequence $\underline{x}$, is presented in [2]. Recently, [19] have considered the case of batch learning with more than one test sample, analyzing "add-β" learners for the multinomial case and for first-order Markov chains.

III. Formal Problem Statement

Consider the stochastic batch learning problem where instead of a single test, the test set contains L samples:

$$R\left(\theta, q(\cdot|\cdot)\right) = \frac{1}{L} \sum_{y^{N+L}} p_\theta(y^{N+L}) \log \frac{p_\theta(y^L|y^N)}{q(y^L|y^N)}. \quad (3)$$

Naturally, we will be interested in the min-max optimal regret and the probability assignment that achieves it:

$$R^* = \min_{q(\cdot|\cdot)} \max_\theta R\left(\theta, q(\cdot|\cdot)\right). \quad (4)$$

In this paper, we will deal with hypothesis classes where given the true hypothesis, the past does not contribute any more information regarding future results, i.e., $\forall L, N : p_\theta(y^L|y^N) = p_\theta(y^L)$. We will limit ourselves to the case where the outcomes space Y is finite. Following previous works in batch learning, we will use the natural logarithm.

IV. Theoretical Results

Our first result, unsurprisingly, is a minimax theorem for the stochastic batch-learning problem with varying test set sizes:

Theorem 1: Assume that Θ is bounded. The min-max optimal regret is equal to the max-min regret where the maximum is over all priors over θ, and the regret (redundancy) equals the capacity of a channel between Y^L and Θ given Y^N:

$$R^* = \max_{w(\theta)} \min_{q(\cdot|\cdot)} \int_\theta w(\theta) R\left(\theta, q(\cdot|\cdot)\right) d\theta$$
$$= \max_{w(\theta)} \frac{1}{L} I(Y^L;\Theta|Y^N).$$

where the maximum is over all valid distributions over Θ.

Furthermore, if we denote by $w^*(\theta)$ the prior that achieves the max-min optimal regret, then the max-min and min-max optimal probability assignment is given by a Bayesian mixture using $w^*(\theta)$:

$$q^*(y^L|y^N) = \frac{\int_\theta w^*(\theta)p_\theta(y^{N+L})}{\int_\theta w^*(\theta)p_\theta(y^N)}$$

Proof of Theorem 1: First, define the regret with respect to $w(\theta)$

$$R\left(w(\theta), q(\cdot|\cdot)\right) = \int_\theta w(\theta) R\left(\theta.q(\cdot|\cdot)\right) d\theta \tag{5}$$

It is easy to verify that:

$$R^* = \min_{q(\cdot|\cdot)} \max_{w(\theta)} R\left(w(\theta), q(\cdot|\cdot)\right).$$

This is due to the fact that one can always center all mass of $w(\theta)$ at any specific θ, and on the other hand for every $w(\theta)$, the average regret over $\theta \in \Theta$ cannot be larger than the maximal value.

Now, we know that $R\left(w(\theta), q(\cdot|\cdot)\right)$ is convex and lower semi-continuous with respect to $q(\cdot|\cdot)$ and linear with respect to $w(\theta)$. Since the set of all possible conditional probabilities $q(\cdot|\cdot)$ is a compact Hausdorff space, we can invoke Theorem 2 in [20] and get that:

$$\min_{q(\cdot|\cdot)} \max_{w(\theta)} R\left(w(\theta), q(\cdot|\cdot)\right) = \max_{w(\theta)} \min_{q(\cdot|\cdot)} R\left(w(\theta), q(\cdot|\cdot)\right).$$

Next, incorporating the constraint that for all training y^N, $\sum_{y^L} q(y^L|y^N) = 1$, and solving the corresponding constrained minimization over $q(\cdot|\cdot)$ we get the following optimal probability assignment for every $w(\theta)$:

$$q(y^L|y^N) = \frac{\int_\theta w(\theta)p_\theta(y^{N+L})}{\int_\theta w(\theta)p_\theta(y^N)},$$

and so we get:

$$R^* = \max_{w(\theta)} I(Y^L; \Theta|Y^N).$$

Now, if we denote by $w^*(\theta)$ the capacity-achieving prior, than we know from KKT conditions that for every θ, $q^*(\cdot|\cdot)$ achieves a divergence of at most $\max_{w(\theta)} I(Y^L; \Theta|Y^N)$ at every point θ, and thus $q(\cdot|\cdot)$ achieves the min-max optimal regret R^*.

■

V. Multinomial Distribution

Consider now the multinomial case, where $y \in \{0, ..., m-1\}$ and Θ is the simplex over an alphabet of size m. Note that for this case, when $N = 0$ we have $R^* = \frac{(m-1)\log(L)}{2L} + o(\frac{1}{L})$, while for $L = 1$ we have $R^* = \frac{(m-1)}{2N} + o\left(\frac{1}{N}\right)$. The following result is a generalization that allows us to interpolate between the two scenarios:

Theorem 2: For the multinomial case, the min-max optimal value of:

$$R^* = \min_{q(\cdot|\cdot)} \max_{\theta\in\Theta} \frac{1}{L} \sum_{y^{N+L}} p_\theta(y^{N+L}) \log\left(\frac{p_\theta(y^L)}{q(y^L|y^N)}\right). \tag{6}$$

is equal to

$$\frac{1}{L}\left[\sum_{k=0}^{L-1} \frac{m-1}{2(N+k)} + o\left(\frac{1}{N+k}\right)\right]. \tag{7}$$

Proof of Theorem 2: First, we will show that this regret is achievable. To this end, consider a learner who, given every training set y^N and partial set of results $y^{t-1}, t \in \{1, ..., L\}$, uses the probability assignment suggested in [15] for the next y_t. We will denote this probability assignment by $q_{SB}(y_t|y^N, y^{t-1})$. Since $p_\theta(y^L) = \prod_{t=1}^L p_\theta(y_t)$, we get:

$$R\left(\theta, q_{SB}(\cdot|\cdot)\right) = \frac{1}{L}\sum_{t=1}^{L} \sum_{y^N, y^t} p_\theta(y^{N+t}) \log\left(\frac{p_\theta(y_t)}{q_{SB}(y_t|y^N, y^{t-1})}\right).$$

Now, from [15] we know that:

$$\max_\theta R\left(\theta, q_{SB}(\cdot|\cdot)\right) = \frac{1}{L}\sum_{t=0}^{L-1}\left[\frac{m-1}{2(N+t)} + o\left(\frac{1}{N+t}\right)\right].$$

To show that one cannot get a better regret, we will consider an adversary that chooses some prior distribution $w(\theta)$ over the models. It is quite clear that for all $w(\theta)$:

$$R^* \geq \max_{q(\cdot|\cdot)} \frac{1}{L}\int_{\theta)} w(\theta)\left[\sum_{y^{N+l}} p_\theta(y^{N+L}) \log\left(\frac{p_\theta(y^L)}{q(y^L|y^N)}\right)\right] \tag{8}$$

where now the learner knows $w(\cdot)$, but not the specific θ according to which the data is generated. Using Lagrange multipliers, we get that the optimal probability assignment is a simple Bayesian mixture:

$$q^*_{w(\theta)}(y^L|y^N) = \frac{\int_\theta w(\theta)p_\theta(y^{N+L})}{\int_\theta w(\theta)p_\theta(y^N)} \tag{9}$$

which can be written as a product of Bayesian sequential probability assignments:

$$\begin{aligned} q^*_{w(\theta)}(y^L|y^N) &= \prod_{t=1}^{L} \frac{\int_\theta w(\theta)p_\theta(y^{N+t})}{\int_\theta w(\theta)p_\theta(y^{N+t-1})} \\ &= \prod_{t=1}^{L} q^*_{w(\theta)}(y_t|y^{N+t-1}). \end{aligned}$$

Now, by [14], we know that by choosing a uniform prior, $w_1(\theta)$, for the multinomial case, we have for every N, t:

$$\begin{aligned} &\int_\theta w_1(\theta)p_\theta(y^{N+t+1}) \log\left(\frac{p_\theta(y_{N+t+1})}{q^*_{w_1}(y_{N+t+1}|y^{N+t})}\right) \\ &= \frac{m-1}{2(N+t)} + o\left(\frac{1}{N+t}\right). \end{aligned} \tag{10}$$

and thus:

$$R \geq \frac{1}{L}\sum_{t=0}^{L-1}\left[\frac{m-1}{2(N+t)}+o\left(\frac{1}{N+t}\right)\right] \tag{11}$$

■

Now, when $L, N \gg 1$ the regret can be approximated by $R^* \simeq (m-1)\frac{\log(N+L)-\log(N)}{2L}$, which can be further simplified for three different cases:

- When $L \ll N$ we get:

$$R^* \simeq (m-1)\frac{\log(1+\frac{L}{N})}{2L} \simeq \frac{m-1}{2N} \tag{12}$$

  which is exactly what we get when $L = 1$. This is not surprising since when $L \ll N$, the information derived from the additional test samples is negligible.
- when $L \gg N$ we get:

$$\begin{aligned} R^* &\simeq (m-1)\frac{\log(L)-\log(1+\frac{N}{L})-\log(N)}{2L} \\ &\simeq \frac{(m-1)}{2L}\log(L)+O\left(\frac{\log(N)}{L}\right). \end{aligned} \tag{13}$$

  The leading term here coincides with what we get in online prediction, which makes sense because the data that the training set adds is negligible.
- when $\frac{L}{N} = \alpha$ we get:

$$R^* \simeq (m-1)\frac{\log(1+\alpha)}{2L} = (m-1)\frac{\log(1+\alpha)}{2N\cdot\alpha} \tag{14}$$

  we get an expression that is more similar to the $L = 1$ case, with a multiplicative factor of $\frac{\log(1+\alpha)}{\alpha}$.

We note that for the batch learning case, where $L = 1$ and the edges are excluded, the second term of the regret is proportional to $\frac{1}{N^2}$, a result due to [16]. Thus, we conjecture that the second term of the regret, in our case, will be proportional to $\sum_{l=0}^{L-1}\frac{1}{(N+l)^2}$.

VI. Laplace's Approximation Approach

In this section, we will present an informal derivation that might lead to a more general result than the one presented in the multinomial case if formalized correctly. We will do so by utilizing Laplace's method of integration; see [21] for a translation of the original work. Assume some continuous, bounded away from zero prior $w(\theta)$ and an hypothesis classes $\Theta \in \mathbf{R}^d$ such that Laplace's approximation holds:

$$\begin{aligned} p(y^N) &= \int_{\theta\in\Theta} w(\theta)p_\theta(y^N)d\theta \\ &\simeq \left(\frac{2\pi}{NJ_{\hat{\theta}(y^N)}}\right)^{\frac{d}{2}} w\left(\hat{\theta}(y^N)\right)p_{\hat{\theta}(y^N)}(y^N) \end{aligned}$$

where $\hat{\theta}(y^N) = \arg\max_{\theta\in\Theta} p_\theta(y^N)$ and J_θ is the Fisher's information evaluated at θ.

Now, note that $I(Y^{N+L};\Theta|Y^N) = H(Y^{N+L}) - H(Y^N) - H(Y^L|\theta)$. Using Laplace's approximation above, we get that when $N \gg 1, d$

$$H(Y^N) = \frac{k\log(N)}{2} - \mathbf{E}\left[\log\left(p_{\hat{\theta}(y^N}(y^N)\right)\right] + O(d)$$

and thus:

$$\begin{aligned} I(Y^{N+L};\Theta|Y^N) &= \frac{k}{2}\left(\log(N+L)-\log(N)\right) \\ &- \mathbf{E}\left[\log\left(\frac{p_{\hat{\theta}(y^{N+L})}(y^{N+L})}{p_{\hat{\theta}(y^N)}(y^N)p_\theta(y^L)}\right)\right] + O(d) \end{aligned}$$

In other words, we get that the normalized mutual information behaves as $\frac{k\log(1+\frac{L}{N})}{2L} = \frac{k}{2N}\frac{\log(1+\alpha)}{\alpha}$ if we can show that $\mathbf{E}\left[\log\left(\frac{p_{\hat{\theta}(y^{N+L})}(y^{N+L})}{p_{\hat{\theta}(y^N)}(y^N)p_\theta(y^L)}\right)\right]$ is negligible.

This expression, derived for the general parametric case, coincides with the result we got for the multinomial case where we have $m-1$ parameters. Nevertheless, note that for the multinomial case, we also dealt with the edges where Fisher's information is undefined. Thus, this derivation cannot be used directly for the "close" multinomial case (including the edges).

VII. Capacity Achieving Prior

A. Arimoto-Blahut Algorithm for Batch Learning

In order to verify our theoretical results regarding the leading term of the regret and to understand the behavior of the capacity achieving prior and the second term of the regret, we have implemented a variation of the Arimoto-Blahut algorithm, [9] [10], for the batch-learning problem with varying test set sizes. Our variation is based on the accelerated Arimoto-Blahut algorithm proposed in [11]. For simplicity, we will present results for the binary case, where $y \in \{0, 1\}$.

The most simplified form of the algorithm is presented as Algorithm 1: Naturally, the algorithm receives as input both N and L. In addition, The algorithm receives a discrete set of possible θ, which we denote by $\vec{\theta} = [\theta_0, \theta_1, ..., \theta_M]$, and an initial guess for the prior, $w_0(\vec{\theta})$.

Note that for the classical case, where $N = 0$, the actual capacity-achieving prior is indeed discrete, see [8]. Here, however, this is not guaranteed, and indeed, we will show that for some values of N, L, we get an optimal $w(\vec{\theta})$ which is non-zero for most of the values θ_j even when $M \gg N, L$.

The basic step of the algorithm is as follows: Given some $w_i(\theta)$, we can calculate for each θ_j the divergence between $p_{\theta_j}(y^L)$ and $p(Y^L|y^N)$:

$$D_{\theta_j}\left(p_\theta(y^L)||p(y^L|y^N)\right) = \sum_{y^{N+L}} p_{\theta_j}(y^{N+L})\log\left(\frac{p_{\theta_j}(y^L)}{p(y^L|y^N)}\right)$$

where the conditional probability is based upon $w_i(\vec{\theta})$:

$$p(y^L|y^N) = \frac{\sum_j w(\theta_j)p_{\theta_j}(y^{N+L})}{\sum_j w(\theta_j)p_{\theta_j}(y^N)} \tag{15}$$

Then, given a design parameter λ, each $w(\theta_j)$ is multiplied by $\exp\left(\lambda D_{\theta_j}\left(p_\theta(y^L)||p(y^L|y^N)\right)\right)$. A normalization of the new $w_{j+1}(\theta)$ is performed to get a valid distribution function.

It should be noted that for every $\vec{\theta}$ and every guess $w(\vec{\theta})$ for $w(\theta)$, we can get a bound over the min-max optimal regret. This observation was made in the past for similar cases, for example, [22], but we will give the proof here as well:

Corollary 1: For every $\vec{\theta}, w(\vec{\theta})$, the following holds:

$$\frac{1}{L}I(Y^L;\Theta|Y^N) \leq R^* \leq \max_{\theta\in\Theta}\frac{1}{L}D_\theta\left(p_\theta(y^L)||p(y^L|y^N)\right). \tag{16}$$

Proof: Since $R^* = \max_{w(\theta)}\frac{1}{L}I(Y^L;\Theta|Y^N)$, any specific $\vec{\theta}, w(\vec{\theta})$ is a trivial lower bound over the min-max optimal regret. As for the second inequality, note that every $\vec{\theta}, w(\vec{\theta})$ defines a probability assignment $q_{w(\vec{\theta})}(\cdot|\cdot)$, and trivially:

$$R^* = \min_{q(\cdot|\cdot)}\max_\theta R\left(\theta, q(\cdot|\cdot)\right) \leq \max_\theta R\left(\theta, q_{w(\vec{\theta})}(\cdot|\cdot)\right) \tag{17}$$

■

This observation allows us to derive both lower and upper bounds over R^* from every $w(\vec{\theta})$ we get from the algorithm.

Algorithm 1 Arimoto-Blahut algorithm for Batch Learning

Input: $N, L, \vec{\theta} = [\theta_0, \theta_1, ..., \theta_M], \lambda, w_0(\vec{\theta})$
Output: $w(\vec{\theta})$
LOOP Process
for $i = 0$ to $I-2$ **do**

$$\tilde{w}_{i+1}(\theta_j) = w_i(\theta_j)\cdot\exp\left(\lambda D_{\theta_j}\left(p_\theta(y^L)||p(y^L|y^N)\right)\right)$$

$$w_{i+1}(\theta_j) = \frac{\tilde{w}_{i+1}(\theta_j)}{\sum_{j'=0}^{M}\tilde{w}_{i+1}(\theta_{j'})}$$

end for
return $w(\vec{\theta})$

VIII. Numerical Results

Our first results are presented in Figure 1, where we compare the theoretical leading term with the lower and upper bounds derived from the Arimoto-Blahut algorithm as a function of the total number of observations $L+N$ and for various values of $\alpha = \frac{L}{N}$.

In most cases, the difference between the lower and upper bound in Figure 1 is so small that it can barely be seen, indicating that the numerical calculations coincide with the theoretical ones. Also note that, as expected, the regret for large test sets with fixed training is larger than that for large training sets with fixed test set size.

The full version of this paper presents additional results, including the resulting prior for various combinations of L, N.

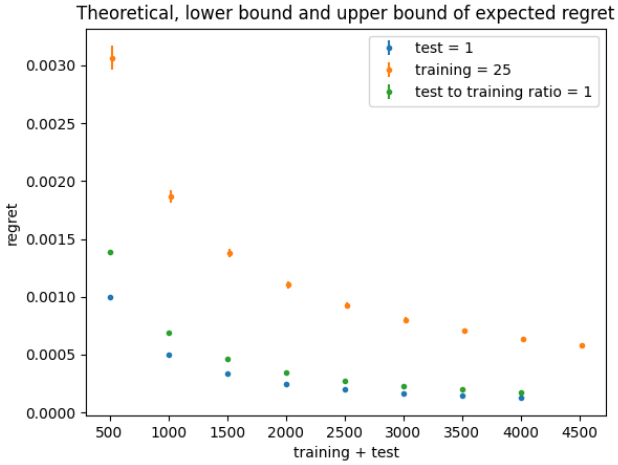

Fig. 1. Theoretical Leading Term with Numerical Upper and Lower bounds

IX. Conclusions and Future Work

In this paper, we have studied the generalization of the batch learning problem where the size of the test set is not necessarily a single sample. We have proven a minimax theorem for a general case and provided a characterization of the min-max optimal regret for the multinomial case. Utilizing the Arimoto-Blahut algorithm, we have also implemented a numerical calculation of the capacity-achieving prior for the binary case and showed both the min-max optimal regret and the capacity-achieving prior for various regimes of the test size L and the training set size N.

There are several directions for future work. First, it will be interesting to see if we can theoretically calculate the capacity achieving prior or at least some properties of it for various values of L and N, and see if it interpolates between the known results for the special cases of $N = 0$ (online) and $L = 1$ (batch with single test).

In addition, it would also be interesting to consider other hypothesis classes, for example, Markov models or hypothesis classes that utilize data features $\underline{x}$. Also, it seems plausible to extend the results to the mismatched, unrealizable, stochastic case.

Another interesting line of work will be to improve the Arimoto-Blahut algorithm we have implemented. One possible direction might be to update $\vec{\theta}$ as in [23] for example.

Finally, it would be interesting to extend the results for individual batch learning, such as [24], to the case of arbitrary test set sizes. This might provide the link for the individual setting from the online case, which is well-defined and understood, to the batch case, which seems to be hard to define in the individual case.

REFERENCES

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

[2] Y. Fogel and M. Feder, "Universal batch learning with log-loss," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 21–25.

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

[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] 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.

[6] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information theory*, vol. 30, no. 4, pp. 629–636, 1984.

[7] 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.

[8] H. Witsenhausen, "Some aspects of convexity useful in information theory," *IEEE Transactions on Information Theory*, vol. 26, no. 3, pp. 265–271, 1980.

[9] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[10] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[11] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated blahut-arimoto-type algorithms," in *Information theory workshop*. IEEE, 2004, pp. 66–70.

[12] H. H. Mattingly, M. K. Transtrum, M. C. Abbott, and B. B. Machta, "Maximizing the information learned from finite data selects a simple model," *Proceedings of the National Academy of Sciences*, vol. 115, no. 8, pp. 1760–1765, 2018.

[13] M. C. Abbott and B. B. Machta, "A scaling law from discrete to continuous solutions of channel capacity problems in the low-noise limit," *Journal of Statistical Physics*, vol. 176, pp. 214–227, 2019.

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

[15] D. Braess and T. Sauer, "Bernstein polynomials and learning theory," *Journal of Approximation Theory*, vol. 128, no. 2, pp. 187–206, 2004.

[16] F. Komaki, "Asymptotically minimax Bayesian predictive densities for multinomial models," *Electronic Journal of Statistics*, vol. 6, pp. 934–957, 2012.

[17] ——, "Bayesian predictive densities based on latent information priors," *Journal of Statistical Planning and Inference*, vol. 141, no. 12, pp. 3705–3715, 2011.

[18] E. Goldstein, "Numerical calculations for universal batch learning," 2023.

[19] M. Bondaschi and M. Gastpar, "Batch universal prediction," *arXiv preprint arXiv:2402.03901*, 2024.

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

[21] S. M. Stigler, "Laplace's 1774 memoir on inverse probability," *Statistical Science*, vol. 1, no. 3, pp. 359–363, 1986.

[22] I. Naiss and H. H. Permuter, "Extension of the blahut–arimoto algorithm for maximizing directed information," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 204–222, 2013.

[23] C.-I. Chang and L. Davisson, "On calculating the capacity of an infinite-input finite (infinite)-output channel," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1004–1010, 1988.

[24] Y. Fogel and M. Feder, "Permutation invariant individual batch learning," in *2023 IEEE Information Theory Workshop (ITW)*. IEEE, 2023, pp. 142–146.