Active Learning for Individual Data via Minimal Stochastic Complexity

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@tau.ac.il

***Abstract*—Modern machine learning systems require massive amounts of labeled training data in order to achieve high accuracy rates. Active learning uses feedback to label the most informative data points and significantly reduces the labeling effort. Many heuristics for selecting data points have been developed in recent years which are usually tailored to a specific task and a general unified framework is lacking.**

In this work, the individual data setting is considered and an active learning criterion is proposed. In this setting the features and labels, both in the training and the test, are specific individual, deterministic quantities. Motivated by connections between source coding and minimax learning, the proposed criterion attempts to find data points which minimize the average Predictive Normalized Maximum Likelihood (pNML) on the unlabeled test set. It is shown using a real data set that the proposed criterion performs better than other active learning criteria.

I. INTRODUCTION

In supervised learning, a training set is provided to the learner which optimizes its model parameters to minimize the empirical error on that training set with the hope that this will imply a low test set error. The training set is randomly drawn from a pool of available examples and an expert labels them prior to training. This process is considered passive learning since the learner is passive in the data collection and labelling tasks. Many machine learning applications today rely heavily on the assumption that humans can annotate all the available data for training. However, the massive amounts of data available today make it an impossible task. The time and financial cost associated with labeling is high, especially when very large training sets are needed. Consequently, only a small random sub-set is labeled which may not represent the true underlying model between features and labels. To avoid this, the training set is redundant and usually much larger than required. Consequently, passive learning requires much more data than needed to attain a vanishing test error probability.

In active learning, the learner has access to a large set of unlabeled examples and can interact with a labelling expert. The learner sequentially chooses which data point he wishes the expert to label based on previously observed examples. This feedback loop has the potential to significantly reduce the number of examples needed to achieve a given accuracy level. The fundamental problem is how to choose which data points to be labeled?

In the last decade there has been significant progress in active learning research. Most contributions deal with proposing a heuristic for feature selection, analyzing its performance and comparing to different lower bounds [1], [2] and [3]. Some of the algorithms and heuristics which have been proposed for active learning include: [4], [5], [6], [1], [7], [8], [9] and [10].

Several approaches consider information-theoretic active learning criteria [11], [12], [13] and [14]. The most common method is Maximum Uncertainty (MU) sampling, where the feature with the highest label predictive entropy given the training is selected. In some sense this approach is very similar to the margin based approach in [15]. However, this aggressive, essentially greedy, scheme may lead to large generalization errors since noise might produce high predictive entropy and corrupt the training set.

In [13], a criterion called Bayesian Active Learning by Disagreement (BALD), is proposed which maximizes the mutual information between the model parameters θ and the data point to be selected, $\hat{x}_t$, given the available training: $\hat{x}_t = \text{argmax}_{x_t} I(\theta; y_t|x^t, y^{t-1})$. The idea is to reduce the number of possible hypotheses maximally fast, i.e. to minimize the uncertainty about model parameters using Shannon's entropy. This criterion also appears as an upper bound on information based complexity of stochastic optimization in [16] and also for experimental design in [17] and [18]. This approach was empirically investigated in [19], where a Bayesian method to perform deep learning was proposed and several heuristic active learning acquisition functions were explored within this framework.

Another approach to optimize the training set is by taking into account the un-labelled test set. Since the trained model will be tested using the test set, one should select training points which have the most relevance to the test set. Essentially, there is no real need to learn the labeling function over the whole feature space which may be very complex and requires many data points. In practice, a preprocessing stage prunes the training set from data points which are irrelevant to the test, but this requires domain knowledge regarding the similarity between training and test sets. Criteria such as BALD and MU do not to take into account the unlabelled test set and select data points based solely on the training pool. In [14], a criterion denoted as Universal Active Learning (UAL) was derived based on universal source coding

and minimax regret minimization. UAL utilizes the unlabeled test set in order to learn data points which are most relevant to the test set. It was shown in [14] that UAL is related to BLAD and MU and is basically a generalization of the two. In [20], UAL is also proposed using heuristic reasoning and denoted as Expected Predictive Information Gain (EPIG).

However, UAL assumes that the data is generated according to a distribution which belongs to a given hypothesis class. This assumption cannot be verified on real world data thus limiting the application of UAL. In this work, the problem of active learning in the individual data setting is considered. This is the most general data setting, in which the data is not assumed to originate from any distribution but an arbitrary / individual set of features and labels. In the next sections we will introduce the criterion and analyze it for binary classification.

Throughout this paper, the following notation for a sequence of samples will be used $x^t = (x_1, x_2, ..., x_t)$. The variables $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ will represent the features and labels respectively with $\mathbb{X}$ and $\mathbb{Y}$ being the sets containing the alphabets of features and labels respectively.

II. Individual Data Setting

In supervised machine learning, a training set consisting of N pairs of examples is provided to the learner:

$$z^N = \{(x_n, y_n)\}_{n=1}^N \tag{1}$$

where x_n is the n-th data instance and y_n is its corresponding label. The goal of a learner is to predict an unknown test label y given its test data, x, by assigning a probability distribution $q\left(\cdot|x, z^N\right)$ for each training z^N.

We consider the individual setting proposed in [21] and [22]. In this setting we assume that there is no conditional distribution relating a feature x to a label y, but the sequence $z^t = \{x^t, y^t\}$ is an individual sequence. Unlike the stochastic setting [22] in which the data follows a distribution $f(y|x)$, which is assumed to be part of some parametric family of hypotheses.

In order to have a well posed problem, the learning objective is to compete with a reference learner, a genie from a known hypothesis class, P_Θ and perform as well as it does on the test set, using the same training data, z^n.

Denote Θ as a general index set, this class is a set of conditional probability distributions

$$P_\Theta = \{p(y|x,\theta)\,|\theta \in \Theta\} \tag{2}$$

In addition, it is assumed that the reference learner knows the test label value y but is restricted to use a model from the given hypothesis set P_Θ. This reference learner then chooses a model, $\hat{\theta}$, that attains the minimum loss over the training set and the test sample:

$$\hat{\theta} = \arg\max_{\theta\in\Theta}\left[\log p(y|x,\theta) + \sum_{n=1}^{N}\log p(y_n|x_n,\theta) + \log(w(\theta))\right] \tag{3}$$

where performance is evaluated using the logloss function, i.e. $-\log\left(q\left(\cdot|x, z^N\right)\right)$.

Note that in this work we extended the individual setting of [21] and allowed the usage of some prior $w(\theta)$ over the parameter space which may be useful for regularization purposes.

The learning problem is defined as the log-loss difference between a learner $\mathbf{q}$ and the reference learner (genie):

$$R_n(q, y; x) = \log\left(\frac{p\left(y|x,\hat{\theta}\right)}{q\left(y|x, z^n\right)}\right) \tag{4}$$

An important result for this setting is provided in [23] and provides a closed form expression for the minimax regret along with the optimal learner, q_{pNML}:

Theorem 1 (Fogel and Feder (2018))**.** *The universal learner, denoted as the pNML, minimizes the worst case regret:*

$$R_n = \min_q \max_{y\in\mathbb{Y}} \log\left(\frac{p\left(y|x,\hat{\theta}\right)}{q\left(y|x, z^n\right)}\right)$$

The pNML probability assignment and regret are:

$$q_{pNML}(y|x, z^N) = \frac{p\left(y|x,\hat{\theta}\right)}{\sum_y p\left(y|x,\hat{\theta}\right)}$$

$$R_n = \log\sum_{y\in\mathbb{Y}} p\left(y|x,\hat{\theta}\right)$$

The pNML regret is associated with the *stochastic complexity* of an hypothesis class as discussed in [24] and [25].

III. Individual Data Active Learning

In this section, active learning in the individual setting is presented. In active learning, the learner sequentially selects data instances, x_n, based on some criterion and produces N training examples $\{x^N, y^N\}$. The objective is to select a subset of the training set and derive a probabilistic learner $q\left(y|x, x^N, y^N\right)$ which will attain the minimal prediction error among all training sets of the same size. Most selection criteria are based on uncertainty quantification of data instances in order to quantify their informativeness. However, in the individual setting, there is no natural uncertainty measure since there is no distribution $f(y|x)$ governing the data. Therefore, we have to resort to a different approach to perform active learning in this setting.

We propose to use the minimax regret R_n as defined in Theorem 1 as an active learning criterion which essentially quantifies the uncertainty of the whole training set z^n for a given un-labelled test feature x. Since R_n is a pointwise quantity, we propose to accumulate it across all the features in the test set.

Therefore, we propose the following criterion for selecting x_n:

$$C_n^A = \min_{x^n\in\mathbb{X}^n}\max_{y^n\in\mathbb{Y}^n}\sum_x\left(\sum_{y\in\mathbb{Y}} p\left(y|x,\hat{\theta}\left(x, y, z^n\right)\right)\right) \tag{5}$$

This problems is difficult to solve for a general hypothesis class in batch form, so we define a greedy form which we denote as Individual Active Learning (IAL):

$$C^A_{n|n-1} = \min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \sum_x \left(\sum_{y \in \mathbb{Y}} p\left(y|x, \hat{\theta}\left(x, y, x_n, y_n, z^{n-1}\right)\right)\right) \quad (6)$$

where z^{n-1} is given.

In the next sections, we will analyze the performance of IAL for binary classification. First, we will prove that IAL coincides with binary search for binary linearly separable data. This result is very important since it provides a sanity check for the use of IAL as an active learning criterion. Finally, we will derive the IAL for Gaussian Process Classification (GPC) and analyze its performance on real data.

A. Binary Classification with Separable Data

In this section we discuss the one dimensional barrier where the data is separable. The idea is to show that in this simple case, the proposed criterion reduces to simple binary search which is known to be optimal.

The 1-dimensional barrier hypotheses class is defined as:

$$p(y=1|x,\theta) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where the input is $x \in [0,1]$, output is $y \in \{0,1\}$ and the unknown threshold is $\theta \in [0,1]$.

Theorem 2. *For one dimensional separable data, the criterion in Eq. (6) induces a selection policy which coincides with binary search.*

Proof. The greedy criterion defined in Eq. (6) can be written as:

$$C^A_{n|n-1} = \min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \sum_{v \in \mathbb{V}} \int_{u \in \mathbb{U}} p\left(v|u, \hat{\theta}^n\right) du \quad (8)$$

where $\hat{\theta}^n$ is the maximum likelihood estimation based on training and test data:

$$\hat{\theta}^n = \arg\max_{\theta \in \Theta} p\left(y^n, v|x^n, u, \theta\right) \quad (9)$$

Since z^{n-1} is assumed known in the greedy case and the likelihood function is basically a multiplication of indicator functions resulting in a rectangle window function around the correct barrier. There are multiple solutions for the maximum likelihood estimator and we select an arbitrary point in the posterior's support.

We can write the likelihood for z^{n-1} as:

$$p\left(y^{n-1}|x^{n-1}, \theta\right) \propto 1\left(\theta \geq \theta^{n-1}_{min}\right) 1\left(\theta < \theta^{n-1}_{max}\right) \quad (10)$$

where θ^{n-1}_{min} and θ^{n-1}_{max} represent the support of the posterior on θ given x^{n-1}, y^{n-1}.

Once we select a new feature point x_n (any point in the support of $p\left(y^{n-1}|x^{n-1}, \theta\right)$), then based on its label y_n which can be arbitrary (we train for the worst), the likelihood window function gets split again.

For $y_n = 0$:

$$p\left(y^n|x^n, \theta\right) \propto 1\left(\theta \geq x_n\right) 1\left(\theta < \theta^{n-1}_{max}\right) \quad (11)$$

For $y_n = 1$:

$$p\left(y^n|x^n, \theta\right) \propto 1\left(\theta \geq \theta^{n-1}_{min}\right) 1\left(\theta < x_n\right) \quad (12)$$

Now we look at the expression $p\left(v|u, \hat{\theta}^n\right)$:

$$\begin{aligned} &\min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \sum_{v \in \mathbb{V}} \int_{u \in \mathbb{U}} p\left(v|u, \hat{\theta}^n\right) du \\ &= \min_{x_n \in \mathbb{X}} \max\{l_0, l_1\} \end{aligned} \quad (13)$$

where

$$l_0 = |1 - \theta^{n-1}_{max}| + 2|x_n - \theta^{n-1}_{max}| + |x_n| = 1 + |x_n - \theta^{n-1}_{max}|$$

and

$$l_1 = |\theta^{n-1}_{min}| + 2|\theta^{n-1}_{min} - x_n| + |1 - x_n| = 1 + |\theta^{n-1}_{min} - x_n|$$

Therefore, the point x_n which minimizes the maximal length is the mid point of the interval $\left[\theta^{n-1}_{min}, \theta^{n-1}_{max}\right]$

□

This theorem indicates that IAL can be a valid active learning criterion. The more interesting and relevant scenario is general binary classification on possibly non separable data. In the next section we will look at IAL for GPC and examine its performance.

IV. Gaussian Process Classification

In this section we will analyze IAL for Gaussian Process Classification (GPC). GPC is a powerful, non-parametric kernel-based model that poses a challenging problem for information-theoretic active learning since the parameter space is infinite dimensional and the posterior distribution is analytically intractable. A detailed introduction to GPC can be found in [26].

In [13], BALD was analyzed for GPC and compared to other active learning algorithms including MU. In [14], UAL was analyzed for GPC too and was shown to perform well when given access to the un-labelled test set. In this section we use the mathematical model of [13] which is repeated here for clarity.

The probabilistic model underlying GPC is as follows:

$$\begin{aligned} f &\sim GP(\mu(\cdot), k(\cdot,\cdot)) \\ y|x, f &\sim Bernoulli\left(\Phi\left(f_x\right)\right) \end{aligned} \quad (14)$$

where the parameter f is a function of a feature point x and is assigned a Gaussian process prior with mean $\mu(\cdot)$ and covariance function $k(\cdot,\cdot)$. The label y is Bernoulli distributed with probability $\Phi(f_x)$, where Φ is the Gaussian CDF.

Without any prior, pNML will give over confident scores for models with very high degrees of freedom such as GPC.

Therefore, we need to add some regularization [25]. A regularization prior will limit the possible solutions of the hypothesis class and avoid over-fitting.

For GPC, IAL can be written as:

$$C^A_{n|n-1} = \min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \sum_{v \in \mathbb{V}} \int_{u \in \mathbb{U}} p\left(v|\hat{f}_u\right) du \tag{15}$$

The maximum likelihood estimate, $\hat{f}_u$ (for test point u) is based on training and test data. Note that the test label v is not known to the learner and thus all options will be covered by IAL.

The MAP estimation for the model parameter vector, $\underline{f}$:

$$\hat{\underline{f}} = \arg\max_{\underline{f}} p\left(y^n, v|x^n, u, \underline{f}\right) p(\underline{f}) \tag{16}$$

where $p(\underline{f})$ is the regularization prior over the latent vector $\underline{f}$.

IAL for GPC in the binary case:

$$C^A_{n|n-1} = \min_{x_n \in \mathbb{X}} \max_{y_n \in \mathbb{Y}} \int_{u \in \mathbb{U}} \left(\Phi\left(\hat{f}_u^{v=1}\right) + \left(\Phi\left(-\hat{f}_u^{v=-1}\right)\right)\right) du \tag{17}$$

where $\hat{f}_u^{v=1}$ and $\hat{f}_u^{v=-1}$ are the maximum likelihood estimates of the latent parameter f_u for test point u with corresponding label v.

Inference in GPC is intractable, since given a training set, the posterior over $\underline{f}$ becomes non-Gaussian and complicated. In order to compute IAL in this case, we need to use approximate inference to model the posterior distribution on the latent model $\underline{f}$. We note that (17) is only dependent on $\hat{f}_{x_n}$ and $\hat{f}_u$, so we can write the expression for the MAP estimation as:

$$\hat{f}_{x_n}^{y_n}, \hat{f}_u^v = \arg\max_{f_{x_n}, f_u} p\left(v|f_u\right) p\left(y_n|f_{x_n}\right) q\left(f_{x_n}, f_u|y^{n-1}, x^{n-1}\right) \tag{18}$$

where $q\left(f_{x_n}, f_u|y^{n-1}, x^{n-1}\right)$ is a multivariate Gaussian distribution approximating the posterior on f_{x_n}, f_u based on $D^{n-1} = \{x^{n-1}, y^{n-1}\}$.

The resulting IAL is summarized in Algorithm 1. First, the algorithm uses an approximate inference method to compute a Gaussian approximation for the posterior using the available training set. Next, for each training point all possible labels are attached along with a sweep on the test set with all possible labels. We run MAP estimation for all the different configurations of training and test and recover the MAP estimate for the test points. We accumulate the probability of the test label given these estimations (pNML regrets). Finally, we find the training point, for which the worst case regret is minimal over the sum of the test points. For all subsequent tests, Expectation Propagation (EP) [27] was used for approximating this posterior using the Matlab GPML toolbox [28].

A. Simulation Results

In this section, IAL is compared to UAL, BALD, MU and passive learning in an empirical analysis using Gaussian Process Classification (GPC) over a real data set. The set is the USPS hand-written digits data set [29]. There are total of 9298 handwritten single digits between 0 and 9, each of which consists of 16 × 16 pixel image. Half of 9298 digits are designated as training and the other half are as test. Pixel values are normalized to be in the range of [-1, 1]. The objective is to classify the digit 0 versus $2, 4, 7$ and 8. In order to reduce the dimension of the data, PCA is applied using the un-labeled training data. Finally, after centering and PCA, the 5 largest Eigenvalues of the PCA are used as the feature space for classification.

A small random subset of the unlabeled test set is given to the learner (15 random samples) along with an initial labelled training set (3 random examples). Active learning is performed by adding a new data point each iteration based on the different criteria. For each iteration, the error probability on the test set is computed and this is shown in Figure 1. It can be observed that random selection has the worst performance in terms of sample complexity given error probability. UAL and MU have comparable performance and slightly better than BALD. IAL has the best performance since it takes into account the test set and is not constrained by the assumption that the data is generated by some class of distributions unlike UAL.

Algorithm 1 Individual Active Learning

procedure IAL - GPC
  Input: Training Data $\{D^{n-1}\}$, Training and Test samples $\{x\}^N$ and $\{u\}^K$.
  Output: Next data point for labelling - x_n
  Run approximate inference algorithm using $\{D^{n-1}\}$ to get posterior Gaussian density $q(\underline{f}|D^{n-1})$
  $\underline{s} = 0$
  for $i \leftarrow 1$ to N **do**
    for $j \in \{-1, 1\}$ **do**
      Set label j for feature x_i
      for $k \leftarrow 1$ to K **do**
        for $l \in \{-1, 1\}$ **do**
          Set label l for feature u_k
          Compute

$$\hat{f}^l_{u_k}, \hat{f}^j_{x_i} = \underset{f_{u_k}, f_{x_i}}{\operatorname{argmax}} \Phi\left(l \cdot f_{u_k}\right) \Phi\left(j \cdot f_{x_i}\right) q\left(f_{x_i}, f_{u_k}|D^{n-1}\right)$$

$$s_{i,j} = s_{i,j} + \Phi\left(l \cdot \hat{f}^l_{u_k}\right)$$

  $\hat{i} = \operatorname{argmin}_i \max_j \underline{s}$
  $x_n = x_{\hat{i}}$

V. Conclusions

In this work, a new active learning criterion for the individual data setting was proposed. It was shown that minimizing the proposed criterion will decrease the minimax regret for any arbitrary data sequence. It has been shown that for binary classification, this criterion coincides with binary search for separable data and is optimal for this scenario. Finally, an empirical test was conducted where the criterion was analyzed in comparison with several other active learning criteria and proved to be superior in terms of sample complexity.

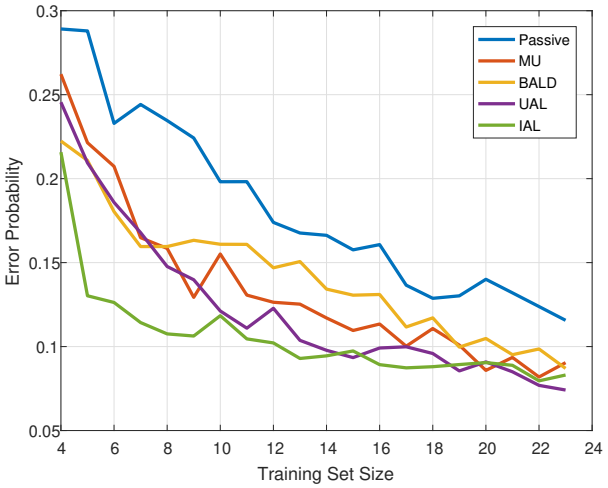

Fig. 1: Error Probability: Hand-written digits data set

REFERENCES

[1] Rui M Castro and Robert D Nowak, "Minimax bounds for active learning," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2339–2353, 2008.

[2] Maxim Raginsky and Alexander Rakhlin, "Lower bounds for passive and active learning," pp. 1026–1034, 2011.

[3] Steve Hanneke, *A bound on the label complexity of agnostic active learning*, Citeseer, 2007.

[4] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby, "Selective sampling using the query by committee algorithm," *Machine learning*, vol. 28, no. 2-3, pp. 133–168, 1997.

[5] Maria-Florina Balcan, Alina Beygelzimer, and John Langford, "Agnostic active learning," *Journal of Computer and System Sciences*, vol. 75, no. 1, pp. 78–89, 2009.

[6] Maria-Florina Balcan, Andrei Broder, and Tong Zhang, "Margin based active learning," pp. 35–50, 2007.

[7] David Cohn, Les Atlas, and Richard Ladner, "Improving generalization with active learning," *Machine learning*, vol. 15, no. 2, pp. 201–221, 1994.

[8] Sanjoy Dasgupta, "Coarse sample complexity bounds for active learning," pp. 235–242, 2006.

[9] Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni, "A general agnostic active learning algorithm," pp. 353–360, 2008.

[10] Steve Hanneke, "Theoretical foundations of active learning," Tech. Rep., CARNEGIE-MELLON UNIV PITTSBURGH PA MACHINE LEARNING DEPT, 2009.

[11] Yuhong Guo and Russell Greiner, "Optimistic active-learning using mutual information.," vol. 7, pp. 823–829, 2007.

[12] Burr Settles, "Active learning literature survey," Tech. Rep., University of Wisconsin-Madison Department of Computer Sciences, 2009.

[13] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.

[14] Shachar Shayovitz and Meir Feder, "Universal active learning via conditional mutual information minimization," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 720–734, 2021.

[15] Maria-Florina Balcan and Phil Long, "Active and passive learning of linear separators under log-concave distributions," pp. 288–316, 2013.

[16] Maxim Raginsky and Alexander Rakhlin, "Information-based complexity, feedback and dynamics in convex programming," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 7036–7056, 2011.

[17] David JC MacKay, "Information-based objective functions for active data selection," *Neural computation*, vol. 4, no. 4, pp. 590–604, 1992.

[18] Valerii Vadimovich Fedorov, *Theory of optimal experiments*, Elsevier, 2013.

[19] Yarin Gal, Riashat Islam, and Zoubin Ghahramani, "Deep bayesian active learning with image data," pp. 1183–1192, 2017.

[20] Andreas Kirsch, Tom Rainforth, and Yarin Gal, "Test distribution-aware active learning: A principled approach against distribution shift and outliers," *arXiv preprint arXiv:2106.11719*, 2021.

[21] Koby Bibas and Meir Feder, "Distribution free uncertainty for the minimum norm solution of over-parameterized linear regression," *arXiv preprint arXiv:2102.07181*, 2021.

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

[23] Yaniv Fogel and Meir Feder, "Universal batch learning with log-loss," pp. 21–25, 2018.

[24] Fernando E Rosas, Pedro AM Mediano, and Michael Gastpar, "Learning, compression, and leakage: Minimising classification error via meta-universal compression principles," in *2020 IEEE Information Theory Workshop (ITW)*. IEEE, 2021, pp. 1–5.

[25] Aurick Zhou and Sergey Levine, "Amortized conditional normalized maximum likelihood: Reliable out of distribution uncertainty estimation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12803–12812.

[26] Carl Edward Rasmussen, "Gaussian processes in machine learning," pp. 63–71, 2003.

[27] Thomas P Minka, "Expectation propagation for approximate bayesian inference," *arXiv preprint arXiv:1301.2294*, 2013.

[28] Carl Edward Rasmussen and Hannes Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

[29] "Usps hand written data set," http://www.gaussianprocess.org/gpml/data/.