2019 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 13–16, 2019, PITTSBURGH, PA, USA

MINIMAX ACTIVE LEARNING VIA MINIMAL MODEL CAPACITY

Shachar Shayovitz and Meir Feder

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

ABSTRACT

Active learning is a form of machine learning which combines supervised learning and feedback to minimize the training set size, subject to low generalization errors. Since direct optimization of the generalization error is difficult, many heuristics have been developed which lack a firm theoretical foundation. In this paper, a new information theoretic criterion is proposed based on a minimax log-loss regret formulation of the active learning problem. In the first part of this paper, a Redundancy Capacity theorem for active learning is derived along with an optimal learner. Building on this, a new active learning criterion is proposed which naturally induces an exploration - exploitation trade-off in feature selection. In the second part, the linear separator hypotheses class with additive label noise is considered and a low complexity algorithm is proposed which optimizes the active learning criterion from the first part. This greedy algorithm is based on the Posterior Matching scheme for communication with feedback and is shown that for BSC and BEC label noise, the proposed information theoretic criterion decays at an exponential rate.

Index Terms— Active Learning, Linear Separator, Posterior Matching

1. INTRODUCTION

In classical supervised learning, a training set (features and labels) is provided and the learning algorithm, for example, optimizes its model parameters to minimize the empirical error with the hope that it will also minimize the generalization error. In this passive learning setting, the training set is randomly selected from a pool of available examples. In order to avoid large generalization errors, the training set is usually very large and redundant, consequently generalization bounds for passive learning decay slowly with the sample size.

In active learning, the learner has access to an unlabeled data-set and sequentially chooses features to label based on past observed examples. The problem is how to adaptively choose these features? A lot of the work in this field has dealt with proposing a heuristic for feature selection, analyzing its performance and comparing to different lower bounds [1]. One well studied approach is based on the disagreement region introduced by Hanneke in [2]. This region contains all the features for which at least two candidate learners do not agree on. Thus, querying the label of such a feature may be helpful to reduce the candidate pool. The general algorithmic framework of disagreement based active learning in the presence of noise was introduced with the A^2 algorithm by Balcan in [3] and other related work in [4] and [5]. However, this approach has high computational complexity and is not aggressive enough in feature selection, hence is far from attaining the optimal sample complexity.

Another approach is margin based active learning which has better sample and computational complexity than the disagreement based approaches. The idea is to sample features at carefully selected regions inside the disagreement region, specifically near the edges of this region. This approach was introduced by Balcan in [6] and followed up in [7] and [8]. While this approach has much better sample and computational complexity than the disagreement based methods, it is not generic enough and only achieves the optimal sample complexity for linear separators in a noiseless setting.

Several approaches consider information-theoretic criteria for selecting the features such as Mutual Information [9], Fisher Information [10] and Entropy [11]. These criteria are typically ad-hoc and the aim is to provide the learning algorithm with some "uncertainty" measure for feature selection. The most common method is Maximum Uncertainty (MU), where the feature which maximizes the label entropy is selected under the assumption that it provides the most informative example. However, this scheme may be too aggressive and lead to large generalization errors since high label entropy may be due to noise and corrupt the learner.

In the first part of this work an information theoretic active learning approach is proposed which unlike those in [9], [10] and [11] optimizes the expected log-loss regret of a test sample. The resulting scheme generalizes the notion of maximum uncertainty with adding the goal of effectiveness in making the prediction. In the second part, the special case of the linear separator model is considered and a greedy low complexity scheme is shown to attain the desired criterion at an optimal exponential rate.

978-1-7281-0824-7/19/$31.00

2. MINIMAX ACTIVE AND PASSIVE LEARNING PROBLEM FORMULATION

In this section a minimax criterion for learning is presented which applies to both active and passive learning. We concentrate on the stochastic setting which assumes a parametric model between the features and the labels unknown to the learner. Specifically, we assume that there is a parameteric family of hypotheses $p\left(y|x,\theta\right)$ and the true distribution of the data corresponds to a specific θ in the family. Our analysis is using log-loss and probabilistic learners which assign a probability to each possible label with a labeling budget of N queries. The objective is to sequentially select features based on past examples and construct a learner, $q\left(y|x,x^N,y^N\right)$, which will perform as well as the best model in the hypotheses class: $p(y|x,\theta)$, i.e. the oracle. A related analysis for passive learning was provided in [12] but assumes i.i.d training samples.

Since the learner has no access to θ, we wish to minimize the maximal expected log-loss regret of this learner. In this sense, the optimal learner would perform as close as possible to the oracle in the worst case. As a first step we write the minimax log-loss regret after observing N samples:

$$\hat{R}=\min_{q}\max_{\theta}\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q\left(y|x,x^N,y^N\right)}\right)\right\} \tag{1}$$

where x,y,x^N,y^N and θ are the test feature and label, observed features and labels and model parameter respectively.

We can equivalently optimize (1) with the distribution $\pi(\theta)$:

$$\hat{R}=\min_{q}\max_{\pi(\theta)}\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q\left(y|x,x^N,y^N\right)}\right)\right\} \tag{2}$$

This is because $\pi(\theta)$ which puts all probability mass on the worst case θ is a least favorable prior. The expectation is performed over the joint probability $p\left(y,x,\theta,x^N,y^N\right)$.

In active learning, the feature to be labeled, x_t, is selected via some selection policy, $\phi(x_t|x^{t-1},y^{t-1})$ which is based on the previous examples. This selection may be stochastic which means that after observing the past examples there is a probability for choosing a specific feature to be labeled.

The joint distribution $p\left(y,x,\theta,x^N,y^N\right)$ is expressed in the following manner using Bayes formula where given θ,x, y is independent of x^N,y^N.

$$\begin{aligned}p\left(y,x,\theta,x^N,y^N\right)=&p\left(y|\theta,x\right)p\left(x^N,y^N|\theta,x\right)\cdot\\&\cdot\pi(\theta|x)p(x)\end{aligned} \tag{3}$$

We can write the following conditional,

$$p\left(x^N,y^N|\theta,x\right)=\Pi_{t=1}^{N}p\left(y_t|x_t,\theta\right)\phi\left(x_t|x^{t-1},y^{t-1}\right) \tag{4}$$

Since there is no dependence on x in the RHS of (4), we get,

$$p\left(x^N,y^N|\theta,x\right)=p\left(x^N,y^N|\theta\right) \tag{5}$$

plugging (5) in to (3)

$$\begin{aligned}p\left(y,x,\theta,x^N,y^N\right)=&p\left(y|\theta,x\right)\Pi_{t=1}^{N}p\left(y_t|x_t,\theta\right)\cdot\\&\cdot\phi\left(x_t|x^{t-1},y^{t-1}\right)\pi(\theta|x)p(x)\end{aligned} \tag{6}$$

Applying (6) to (2),

$$\begin{aligned}\hat{R}=\min_{q}\sum_{x^N,y^N,x}&\phi\left(x_t|x^{t-1},y^{t-1}\right)p(x)\cdot\\&\cdot\max_{\pi(\theta|x)}\sum_{\theta}\Pi_{t=1}^{N}p\left(y_t|x_t,\theta\right)\pi(\theta|x)\cdot\\&\cdot D_{KL}\left(p\left(y|x,\theta\right)||q\left(y|x,x^N,y^N\right)\right)\end{aligned} \tag{7}$$

In passive learning, the $\left\{\phi(x_t|x^{t-1},y^{t-1})\right\}_{t=1}^{N}$ chooses x_t uniformly at random over the available features. However, in active learning we wish to optimize the examples taken at each step t, so we would like to minimize (7) over $\left\{\phi(x_t|x^{t-1},y^{t-1})\right\}_{t=1}^{N}$. Therefore the final active learning problem formulation can be stated as,

$$R=\min_{\{\phi_t\}_{t=1}^{N}}\mathbf{E}_{\mathbf{x}}\left\{\min_{q}\max_{\pi(\theta|x)}\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q\left(y|x,x^N,y^N\right)}\right)\right\}|x\right\} \tag{8}$$

3. OPTIMAL ACTIVE LEARNER

In this section, a capacity redundancy theorem is derived and the optimal learner, based on (x^N,y^N), is shown to be a mixture over the hypotheses class. This capacity can then be used as a criterion for active learning.

Theorem 3.1 (Redundancy-Capacity)**.** *The minimax active learning problem defined in (8) is equivalent the conditional model capacity,*

$$R=\min_{\{\phi(x_t|x^{t-1},y^{t-1})\}_{t=1}^{N}}\max_{\pi(\theta)}I\left(Y;\theta|X,Y^N,X^N\right) \tag{9}$$

The optimal learner is:

$$q^*\left(y|x,x^N,y^N\right)=\sum_{\theta}p\left(\theta|y^N,x^N\right)p\left(y|\theta,x\right) \tag{10}$$

where $\pi(\theta|x)$ maximizes $I\left(Y;\theta|X=x,Y^N,X^N\right)$ (capacity achieving distribution) for each x.

Note that unlike commonly used heuristic methods for active learning such as maximum uncertainty, the conditional model capacity, $\max_{\pi(\theta)}I\left(Y;\theta|X,Y^N,X^N\right)$, inherently optimizes an exploration-exploitation trade-off due to the fact that it maximizes the uncertainty in choosing the training sample, while minimizing the uncertainty in predicting the test sample.

Proof. First we find the best learner q, which optimizes (8) for a given $\left\{\phi(x_t|x^{t-1},y^{t-1})\right\}_{t=1}^{N}$. Note that q is optimal for both passive and active learning. Using (5),

$$\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q\left(y|x,x^N,y^N\right)}\right)|x\right\}=\mathbf{E}_{\mathbf{x^N},\mathbf{y^N}|\mathbf{x}}\left(f(\pi(\theta|x),q)\right) \tag{11}$$

where

$$\begin{aligned}&f(\pi(\theta|x),q)=\\&=\sum_{\theta}p\left(\theta|y^N,x^N,x\right)D_{KL}\left(p\left(y|x,\theta\right)||q\left(y|x,x^N,y^N\right)\right)\end{aligned} \tag{12}$$

Since $f(\pi(\theta|x),q)$ is convex in $q\left(y|x,x^N,y^N\right)$ and concave (linear) in $\pi(\theta|x)$ and the set of distributions is the probability simplex which is compact and convex, then we can apply the minimax theorem [13]. Plugging (11) in to (8) and using the minimax theorem,

$$\begin{aligned}R=\min_{\{\phi_t\}_{t=1}^N}\mathbf{E}_{\mathbf{x},\mathbf{x^N},\mathbf{y^N}}\max_{\pi(\theta|x)}\min_{q}\sum_{\theta}p\left(\theta|y^N,x^N,x\right)\cdot\\\cdot D_{KL}\left(p\left(y|x,\theta\right)||q\left(y|x,x^N,y^N\right)\right)\end{aligned} \tag{13}$$

The optimal q which minimizes the KL divergence is:

$$q^*\left(y|x,x^N,y^N\right)=\sum_{\theta}p\left(y,\theta|x,y^N,x^N\right) \tag{14}$$

Using (14),

$$\begin{aligned}&\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q^*\left(y|x,x^N,y^N\right)}\right)|x\right\}=\\&=\mathbf{E}_{\mathbf{x^N},\mathbf{y^N}|\mathbf{x}}\sum_{y,\theta}p\left(y|\theta,x,x^N,y^N\right)p\left(\theta|y^N,x^N,x\right)\cdot\\&\cdot\log\left(\frac{p\left(y|x,\theta,x^N,y^N\right)}{p\left(y|x,y^N,x^N\right)}\right)\end{aligned} \tag{15}$$

We can average over x and then the expected regret of the optimal predictor given a fixed selection strategy and N examples is the conditional mutual information between the test label and model parameters:

$$\mathbf{E}\left\{\log\left(\frac{p\left(y|x,\theta\right)}{q^*\left(y|x,x^N,y^N\right)}\right)\right\}=I\left(Y;\theta|X,Y^N,X^N\right) \tag{16}$$

with,

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

and $\pi(\theta|x)$ maximizes the mutual information (capacity achieving distribution) for each x. □

However, finding the optimal $\{\phi_t\}_{t=1}^N$ which minimize (9) is dependent on the hypotheses class and can be difficult to solve directly. In the following section we concentrate on the linear separator model class and not try to solve (9) directly but propose a novel algorithm which provides near optimal performance for the proposed criterion. We will show that for the linear separator, using a variation of MU can provide close to optimal sample complexity.

4. ONE DIMENSIONAL LINEAR SEPARATOR WITH ADDITIVE NOISE

In this section we discuss the one-dimensional linear separator model with additive label noise. This hypotheses class is defined by two parameters, θ_0 and p. The relation between a feature x and a noisy label y is defined by first defining an intermediate binary random variable v:

$$p(v|x,\theta)=\begin{cases}1 & \text{if } x>\theta_0\\0 & \text{otherwise}\end{cases} \tag{18}$$

Then y is the output of a binary channel whose input is v. We address two such channels, binary symmetric (BSC) and erasure (BEC) channels. In BSC, the channel models a case where the true label v is flipped with some probability p, while BEC models a case where the output cannot be labeled (and outputs erasure/error) with some probability.

We prove that for this hypotheses class, the maximum uncertainty policy is equivalent to the Posterior Matching (PM) scheme presented in [14] (with appropriate input channel distribution) and show that the conditional entropy of the test point decays exponentially with the number of examples. Moreover, we provide the exponent for this decay which is equivalent to the mutual information defined by the noisy channel (W) and the input distribution to the noisy channel (Q) - $I(Q,W)$.

The maximum uncertainty is a greedy algorithm for sequentially selecting training examples. The algorithm selects features based on greedy optimization of the instantaneous conditional entropy of the classifier. This algorithm is described in the following pseudo-code Algorithm 1.

Algorithm 1 Maximum Uncertainty

procedure MU
  $X^0, Y^0 =$ empty sets
  $i = 1$
  while $i \leq n$ **do**
    $x_i \leftarrow \arg\min_\eta H(Y_i|X_i=\eta, X^{i-1}, Y^{i-1})$.
    $y_i \leftarrow Label(x_i)$.
    $X^i \leftarrow [X^{i-1}, x_i], Y^i \leftarrow [Y^{i-1}, y_i]$.
    $i = i+1$

Lemma 4.1. *The feature selected by maximum uncertainty for one dimensional linear separators with BSC or BEC label noise is the median of $p(\theta|X^{i-1}, Y^{i-1})$. That is,*

$$\eta^* = \arg\min_{\eta} H(Y_i|X_i = \eta, X^{i-1}, Y^{i-1}) \tag{19}$$

where $\eta^* = F^{-1}_{\theta|X^{i-1},Y^{i-1}}\left(\frac{1}{2}\right)$ and $F_{\theta|X^{i-1},Y^{i-1}}(\theta)$ is a cumulative distribution function.

Proof. For the BSC channel this result is very simple and due to limited space we do not include it. For the BEC channel, the entropy can be computed as,

$$\begin{aligned} &H(Y_i|X_i = \eta, x^{i-1}, y^{i-1}) = \\ &= H_B(p) + (1-p)H_B(F_{\theta|x^{i-1},y^{i-1}}(\eta)) \end{aligned} \tag{20}$$

where $H_B(\cdot)$ denotes the binary entropy function. This means that maximizing (20) is equivalent taking the posterior's median. □

In the following theorem the convergence of MU for one dimensional linear separator with BSC or BEC is analyzed. The conditional mutual information, $I(\theta; Y|X, X^n, Y^n)$, reduces to the conditional entropy $H(Y|X, x^n, y^n)$, since the noise is additive, which means that $H(Y|X, \theta, X^n, Y^n) = H_B(p)$.

Theorem 4.2. *For the one dimensional linear separator hypotheses class with BSC or BEC label noise and uniform $p(x)$, the MU algorithm produces a selection policy for which the conditional entropy $H(Y|X, x^n, y^n)$ converges to $H_B(p)$ at an exponential rate $2^{-nI(Q;W)}$.*

Proof. We define an auxiliary Bernoulli random variable $Q \sim Ber\left(\frac{1}{2}\right)$ and write the one dimensional linear separator's output for the selected feature η as,

$$v_i = F_Q^{-1}\left(F_{\theta|X^{i-1},Y^{i-1}}(\theta_0))\right) \tag{21}$$

Using Lemma (4.1), our selection policy selects a point η that once labeled is equivalent to performing PM. Since we deal with an additive channel, we can use Lemma 2 in [14],

$$\lim_{n\to\infty} p(\theta|x^n, y^n) = 2^{nI(Q;W)} \tag{22}$$

for points θ which lie $2^{-nI(Q;W)\pm\epsilon}$ from θ_0.

Note that the value on the RHS of (22) is independent of θ_0 and for any threshold this value will be the same and is only dependent on the level of noise in the channel.

In order to analyze the asymptotic performance, we will compute the conditional entropy,

$$\begin{aligned} &H(Y|X, X^n, Y^n) = \\ &= \int H_B\left(\int P(Y=1|x,\theta)p(\theta|x^n, y^n)d\theta\right)\cdot \\ &\cdot p(x^n, y^n)p(x)dxdx^n dy^n \end{aligned} \tag{23}$$

For BSC, the binary entropy conditioned on a specific x, can be written as,

$$\begin{aligned} &H(Y|x, x^n, y^n) = \\ &= H_B\left(\left(p(1 - F_{\theta|x^n,y^n}(x)) + (1-p)F_{\theta|x^n,y^n}(x)\right)\right) \end{aligned} \tag{24}$$

Therefore,

$$\begin{aligned} &H(Y|X, x^n, y^n) = \int_0^{\theta_0 - 2^{-\frac{nI(Q;W)}{2}}} H_B(p)\, p(x)dx + \\ &+ \int_{\theta_0+2^{-\frac{nI(Q;W)}{2}}}^{1} H_B(1-p)\, p(x)dx + \int_{\theta_0-2^{-\frac{nI(Q;W)}{2}}}^{\theta_0+2^{-\frac{nI(Q;W)}{2}}} \\ &H_B\left(\left(\left(p(1 - F_{\theta|x^n,y^n}(x)) + (1-p)F_{\theta|x^n,y^n}(x)\right)\right)p(x)dx \end{aligned} \tag{25}$$

which can be upper bounded by,

$$\begin{aligned} H_B(p) \le H(Y|X, x^n, y^n) \le 2^{-nI(Q;W)} + \\ + (1 - 2^{-nI(Q;W)})H_B(p) \end{aligned} \tag{26}$$

For the BEC case,

$$H(Y|X = x, x^n, y^n) = H_B(p) + (1-p)H_B(F_{\theta|x^n,y^n}(x)) \tag{27}$$

Therefore,

$$H(Y|X, x^n, y^n) \le H_B(p) + 2^{-nI(Q;W)}(1-p) \tag{28}$$

This concludes the proof that active learning via PM converges to $H_B(p)$ at rate $2^{-nI(Q;W)}$. □

5. MULTIDIMENSIONAL LINEAR SEPARATOR MODELS WITH ADDITIVE NOISE

In this section we will extend Theorem 4.2 to features embedded in $\mathbb{R}^d$. We will constrain our discussion to features $\underline{x} \in \mathbb{R}^d$ which satisfy $\|\underline{x}\| \le R$ with uniform $p(\underline{x})$. The hypotheses class contains all possible homogeneous linear separators with normal vector $\underline{w}$, partitioning the d-sphere to two disjoint sets. The relation between feature $\underline{x}$ and noisy label y is defined by introducing the auxiliary binary random variable v,

$$p(v|\underline{x}, \underline{w}) = \begin{cases} 1 & \text{if } \underline{w}^T\underline{x} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

and y is the output of a binary channel (BSC or BEC) whose input is v.

For convenience, we define the spherical coordinates of a unit vector $\underline{x} \in \mathbb{R}^d$ are defined as:

$$\underline{x} = \left[cos(\theta_1), sin(\theta_1)cos(\theta_2), ..., \Pi_{i=1}^{d-1} sin(\theta_i)\right] \tag{30}$$

5.1. Successive Posterior Matching

In this section, we propose a label efficient low complexity algorithm for learning a linear separator with noisy additive labels for the criterion in (9). The basic idea is to successively localize the spherical coordinates of the normal vector $\underline{w}$ using the PM scheme and then use a mixture learner over all the normal vectors. We will prove that this algorithm, which we denote as Successive Posterior Matching (SPM), can achieve an exponential improvement over passive learning in sample complexity.

The SPM is detailed in Algorithm 2 below where the estimations of the spherical coordinates of $\underline{w}$ are denoted as: $\hat{\underline{\theta}}$. In the initialization stage, each entry in $\hat{\underline{\theta}}$ is set to $\frac{\pi}{2}$ and its respective posterior is uniform. Next, at iteration i, SPM localizes the boundary, T_i, between two hyper-spaces by querying points $\underline{x}$ with spherical coordinates fixed to the current estimation $\hat{\underline{\theta}}$ but coordinate θ_i is changed using PM to localize this boundary. After n label queries, the median of $p(T_i|\underline{x}^n, y^n)$ is selected and $\frac{\pi}{2}$ is added to account for the fact that we need the spherical coordinate, θ_i, of the normal vector. This process is repeated for the next angle θ_{i-1}. Note that the number of labeling operations is $n_T = nd$ and the computational complexity is polynomial.

Algorithm 2 Active Learning via Successive Posterior Matching

procedure SPM
  Init: $\hat{\underline{\theta}} = [\frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}, ..., \frac{\pi}{2}]$,
  Init: $\forall i \in [1 : d-1], p(\theta_i|\underline{x}^0, y^0) = Unif[0, \pi]$
  for $i \leftarrow d-1$ to 1 **do**
    for $k \leftarrow 1$ to n **do**
      $\hat{\theta}_i = F^{-1}_{T_i|\underline{x}^{k-1}, y^{k-1}}\left(\frac{1}{2}\right)$
      $\underline{x}_k = [\Pi_{l=1}^{d-1} \sin(\hat{\theta}_l), \cos(\hat{\theta}_{d-1})\Pi_{l=1}^{d-2} \sin(\hat{\theta}_l)$
        $, ..., \cos(\hat{\theta}_i)\Pi_{l=1}^{i-1} \sin(\hat{\theta}_l), ..., \cos(\hat{\theta}_1)]$
      $y_k = Label(\underline{x}_k)$
      Update $p(T_i|\underline{x}^k, y^k)$
    $\hat{\theta}_i = \hat{\theta}_i + \frac{\pi}{2}$

Theorem 5.1. *For the d dimensional homogeneous linear separator with BSC or BEC label noise and uniform $p(\underline{x})$, the SPM algorithm produces a selection policy for which the conditional entropy $H(Y|X, \underline{x}^{n_T}, y^{n_T})$ converges to $H_B(p)$ at the exponential rate $2^{-\frac{n_T}{d} I(Q;W)}$.*

Proof. We denote the set of point $\{\underline{x}(i)\}_{k=1}^n$ as the points selected at iteration i:

$$\underline{x}(i)_k = [\Pi_{j=i+1}^{d-1} sin(\hat{\theta}_j) sin(\theta_i), \\ cos(\hat{\theta}_{d-1})\Pi_{j=i+1}^{d-2} sin(\hat{\theta}_j) sin(\theta_i), ..., cos(\theta_i), 0, 0, ..., 0] \tag{31}$$

These points belong to a plane defined by the previously estimated coordinates $\underline{\theta}_{i+1}^{d-1}$ and PM find the boundary of T_i in the interval $[0, \pi]$. Therefore, after labeling n points, $\{\underline{x}(i)\}_{k=1}^n$, the posterior $p(T_i|\underline{x}(i)^n, y^n) \approx 2^{nI(Q,W)}$, for points θ which lie in an interval of size $2^{-nI(Q;W)}$.

Since we wish to estimate the normal vector $\underline{w}$, we use the fact that $\underline{w}$ must satisfy $\underline{w}^T \underline{x} = 0$ for all point $\underline{x}$ which lie near this boundary. Since all the points on the inspected plane with high probability can lie on the boundary, then:

$$\underline{w}^T \underline{x}_k = \Pi_{j=1}^{i-1} sin(\phi_j) sin(\theta_i(\underline{x}(i))) sin(\theta_i(\underline{w}))\gamma + \\ + \Pi_{j=1}^{i-1} sin(\phi_j) cos(\theta_i(\underline{x}(i))) cos(\theta_i(\underline{w})) = 0 \tag{32}$$

where γ is a recursive function based on the angles $\underline{\theta}_{i+1}^{d-1}(\underline{w})$ and $\hat{\underline{\theta}}_{i+1}^{d-1}$.

Equation (32) is a two-dimensional correlation between two vectors where $\theta_i(\underline{x}(i))$ has a range of $2^{-nI(Q;W)}$. For large enough n, γ is arbitrarily close to 1, thus $p(\theta_i(\underline{w})|\underline{\theta}_{i+1}^{d-1}(\underline{w}), \underline{x}^{n_T}, y^{n_T}) \approx 2^{-nI(Q;W)}$.

Using the chain rule for probabilities,

$$p(\underline{\theta}(\underline{w})|\underline{x}^{n_T}, y^{n_T}) = \Pi_{i=1}^d p(\theta_i(\underline{w})|\underline{\theta}_{i+1}^{d-1}(\underline{w}), \underline{x}^{n_T}, y^{n_T}) \tag{33}$$

For the BSC case (BEC is very similar),

$$P(Y = 1|\underline{x}, \underline{x}^{n_T}, y^{n_T}) = \\ = p \int 1\left(\underline{x}^T \underline{w} \le 0\right) \Pi_{i=1}^d p(\theta_i|\underline{\theta}^{i-1}, \underline{x}_i^n, y_i^n) d\underline{\theta} + \\ + (1-p) \int 1\left(\underline{x}^T \underline{w} \ge 0\right) \Pi_{i=1}^d p(\theta_i|\underline{\theta}^{i-1}, \underline{x}_i^n, y_i^n) d\underline{\theta} \tag{34}$$

Since each angle in the vector $\underline{\theta}(\underline{w})$ is contained in an interval which is equal to $2^{-nI(Q,W)}$, then whenever the angle between $\underline{x}$ and all the unit vectors $\underline{w}$ defined by $\underline{\theta}(\underline{w})$ is greater than $sin(2^{-nI(Q,W)}) \approx 2^{-nI(Q,W)}$, then $H(Y|X, \underline{x}^{n_T}, y^{n_T}) = H_B(p)$.

Therefore we can bound the conditional entropy,

$$H(Y|X, x^{n_T}, y^{n_T}) \le \int_{\frac{|\langle \underline{x}, \underline{\hat{w}} \rangle|}{|\underline{x}|} > 2^{-nI(Q,W)}} H_B(p)\, p(\underline{x}) d\underline{x} + \\ + \int_{\frac{|\langle \underline{x}, \underline{\hat{w}} \rangle|}{|\underline{x}|} \le 2^{-nI(Q,W)}} p(\underline{x}) d\underline{x} \tag{35}$$

where $\underline{\hat{w}}$ is the unit vector with spherical coordinates corresponding to the median of each uncertainty interval.

Thus,

$$H(Y|X, x_T^n, y_T^n) \le 2^{-\frac{n_T}{d} I(Q;W)} + \\ + (1 - 2^{-\frac{n_T}{d} I(Q;W)}) H_B(p) \tag{36}$$

□

SPM was simulated for BEC and BSC with $d = 3$ for two different p values. In Fig.1, it is shown that the decay in error probability is exponential compared to the passive bound. Moreover, it can be seen that the decay rate is close to the capacities of the BEC and BSC channels.

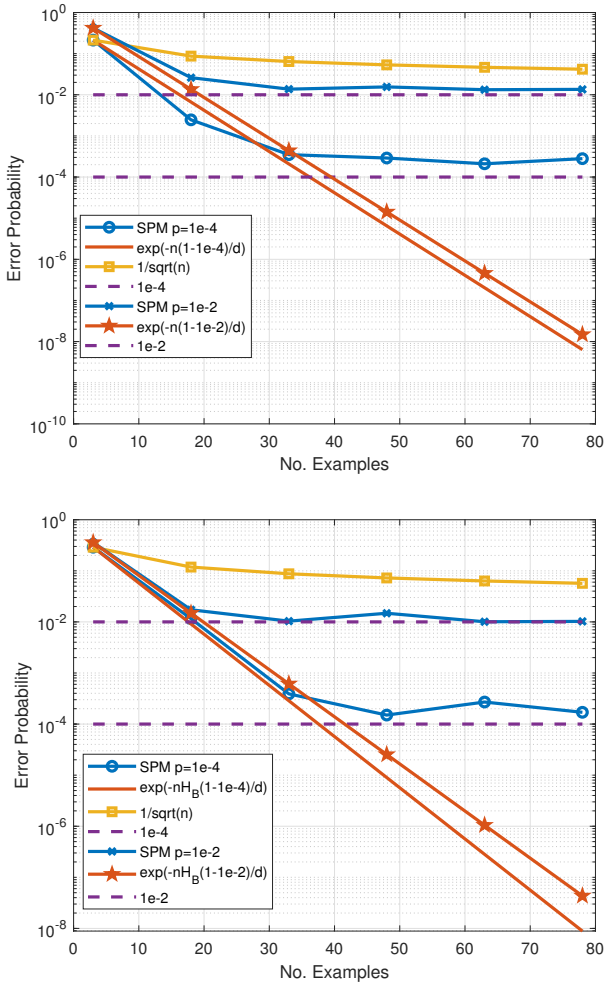

Fig. 1. SPM error probability for BEC (up) and BSC (down)

6. DISCUSSION

In this work a new information theoretic criterion for active learning has been proposed. It relies on a Redundancy-Capacity equivalence between the minimax log-loss regret and the channel capacity between the model parameters and the test label given the test feature and the training set. Moreover, an active learning algorithm has been proposed based on the Posterior Matching scheme, which was shown to achieve exponential improvement over passive learning for the family of homogeneous linear separator over $\mathbb{R}^d$ with BSC or BEC label noise, with decay factor equal to $\frac{I(Q,W)}{d}$.

7. REFERENCES

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

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

[3] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *Journal of Computer and System Sciences*, vol. 75, no. 1, pp. 78–89, 2009.

[4] A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang, "Agnostic active learning without constraints," in *Advances in Neural Information Processing Systems*, 2010, pp. 199–207.

[5] S. Dasgupta, D. J. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Advances in neural information processing systems*, 2008, pp. 353–360.

[6] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 35–50.

[7] M.-F. Balcan and P. Long, "Active and passive learning of linear separators under log-concave distributions," in *Conference on Learning Theory*, 2013, pp. 288–316.

[8] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM, 2014, pp. 449–458.

[9] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," *arXiv preprint arXiv:1112.5745*, 2011.

[10] J. Sourati, M. Akcakaya, T. K. Leen, D. Erdogmus, and J. G. Dy, "Asymptotic analysis of objectives based on fisher information in active learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1123–1163, 2017.

[11] Y. Guo and R. Greiner, "Optimistic active-learning using mutual information." in *IJCAI*, vol. 7, 2007, pp. 823–829.

[12] 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.

[13] J. v. Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische annalen*, vol. 100, no. 1, pp. 295–320, 1928.

[14] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *2007 IEEE International Symposium on Information Theory*. IEEE, 2007, pp. 391–395.