Universal Active Learning via Conditional Mutual Information Minimization

Shachar Shayovitz, *Member, IEEE*, Meir Feder, *Life Fellow, IEEE*

***Abstract*—Modern machine learning systems require massive amounts of labeled training data in order to achieve high accuracy rates which is very expensive in terms of time and cost. Active learning is an approach which uses feedback to only label the most informative data points and significantly reduce 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, a new information theoretic criterion is proposed based on a minimax log-loss regret formulation of the active learning problem. First, a Redundancy Capacity theorem for active learning is derived along with an optimal learner. This leads to a new active learning criterion which naturally induces an exploration - exploitation trade-off in feature selection and generalizes previously proposed heuristic criteria. The new criterion is compared analytically and via empirical simulation to other commonly used information theoretic active learning criteria. Next, the linear hyper-plane hypotheses class with possibly asymmetric label noise is considered. The achievable performance for the proposed criterion is analyzed using a new low complexity greedy algorithm based on the Posterior Matching scheme for communication with feedback. It is shown that for general label noise and bounded feature distribution, the new information theoretic criterion decays exponentially fast to zero.**

***Index Terms*—Minimax learning, active learning, posterior matching, feedback.**

I. Introduction

IN SUPERVISED learning, a training set (features and labels) is provided to the learner which optimizes its model parameters to minimize the empirical error on that training set with the hope of low generalization error. In this passive learning setting, the training set is randomly drawn from some pool of available examples and an expert labels them prior to training. 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 impossible to do so. The cost associated with labeling is high (time and money wise) especially when very large training sets are needed. Consequently, only a small random sub-set is labeled which may be un-representative of the true underlying model between features and labels, thus large generalization errors might occur. To avoid this, the training set is redundant and usually larger than required. Consequently, generalization bounds for passive learning error probability do not decay exponentially fast.

In active learning, the learner has access to a large set of unlabeled examples and can interact with an 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 the next data point to be labeled?

In the last decade there has been significant progress in active learning research. Most rigorous results and bounds are, however, for binary linear classification or regression problems. Most papers 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: [1], [4], [5], [6], [7], [8], [9] and [10].

One well studied approach is based on the disagreement region introduced by Hanneke in [3]. 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 *et al.* in [5] and other related work in [9], [11] and [12].

Another approach which has proven effective is margin based active learning which has better label and computational complexity than disagreement based approaches. The idea is not to sample features in all the disagreement region but at carefully selected regions inside, specifically near the edges of this region. This approach was introduced in [6] and continued in [13] and [14]. While this approach has much better computational complexity than the disagreement based approach, it is not robust to noise. Also, since this algorithm samples points based on some known prior distribution on the features, the exponential decay will only work for log-concave distributions.

In addition, several approaches consider information-theoretic criteria for selecting features [15], [16] and [17]. The most common method is uncertainty sampling or Maximum Uncertainty (MU), where the feature with the highest label entropy given the training is selected. In some sense this approach is very similar to the margin based approach in [13]. However, this aggressive, essentially greedy, scheme may lead to large generalization errors since noise might produce very

Manuscript received October 15, 2020; revised March 17, 2021; accepted April 12, 2021. Date of publication April 23, 2021; date of current version June 21, 2021. *(Corresponding author: Shachar Shayovitz.)*

The authors are with the Iby and Aladar Fleischman Faculty of Engineering, Tel Aviv University, Tel Aviv 69978, Israel (e-mail: shachar.shay@gmail.com).

Digital Object Identifier 10.1109/JSAIT.2021.3073842

high entropy and corrupt the training set. Suppose a very noisy feature is presented to the learner, then the probability assigned to all the labels will be very low (essentially uniform), causing the label entropy to be very high. The learner will thus learn the noise modalities instead of useful information.

In [17] an information theoretic criterion is proposed which is based on maximizing the mutual information between the model and the selected features and provides good performance. The criterion is based on reducing the number of possible hypotheses maximally fast, i.e., to minimize the uncertainty about parameters using Shannon's entropy. This criterion also appears as an upper bound on information based complexity of stochastic optimization in [18] and also for experimental design of experiments in [19] and [20]. This criterion represents the average reduction in uncertainty on the model θ after observing the label Y_t of feature X_t based on the available training. Since this maximization is generally very difficult, a greedy algorithm is proposed, which seeks the data point X_t that maximizes the decrease in expected posterior entropy. This approach was empirically investigated in [21], where a Bayesian method to perform deep learning was proposed and several heuristic active learning acquisition functions were explored within this framework. It was shown that the performance of this criterion, denoted as BALD , was the best. However, this criterion does not to take into account the test distribution $p(x)$ and thus may select examples which are not informative for the test case at hand.

All the aforementioned papers either proposed heuristic criteria for performing active learning or gave rigorous guarantees only for binary linear classification and regression problems. In this contribution, active learning is addressed from an information theoretic point of view. First, passive and active learning are formulated as a minimax log-loss regret problem and a capacity-redundancy theorem is developed. Next, a novel active learning criterion which implicitly optimizes an exploration-exploitation trade off in feature selection is proposed. This criterion, denoted as UAL (Universal Active Learning), is compared to BALD and MU analytically for linear regression and using an empirical binary classification test. Finally, this criterion is analyzed for the linear separator hypothesis class. In order to analyze UAL's achievable performance in this class, a low complexity, noise robust and label efficient algorithm is proposed for bounded prior distributions on the test feature, x.

This paper is organized as follows. In Section II a new active learning criterion (UAL) is derived based on a mini-max regret formulation of the learning problem. This criterion is analyzed and compared with other active learning criteria. In Section III, UAL and BLAD are analyzed for the linear regression hypothesis class and the relation to optimal design of experiments is described. In Section IV, active learning with linear separators is addressed and a low complexity, noise robust algorithm, denoted as SPM, is presented. It is shown via simulations that SPM achieves very good performance in terms of error probability. Moreover, it is proved that SPM generates an active learning selection policy for which UAL decays exponentially to zero. This fact links the two contributions in this paper: a new information theoretic active learning criterion and an achievable upper bound for noisy linear separators.

II. Minimax Active Learning in the Stochastic Setting

In this section, the stochastic active learning problem is formulated for the log-loss cost function in a setting similar to the one described as stochastic universal prediction in [22]. In this setting, the probability distribution of a label, y, given a feature, x, is given as $p(y|x,\theta)$ with a parameter $\theta \in \Theta$, where Θ is a set containing all the parameters of a hypothesis class.

In the active learning setting, the objective is to sequentially select features and collect N training examples (features $x^N = \{x\}_{i=1}^N$ and labels $y^N = \{y\}_{i=1}^N$) which derive a probabilistic learner for a test label y, given a test feature x: $q(y|x, x^N, y^N)$, such that it will perform as close as possible to the best learner in the hypotheses class: $p(y|x,\theta)$, i.e., the Oracle. A related analysis for passive learning was provided in [23] but assumes i.i.d training samples.

Since the learner has no access to θ, we wish to minimize the maximal (worst θ) expected (with respect to the true distribution) log-loss regret of this learner to the Oracle. In this sense, we wish to minimize the regret of the learner to the Oracle in worst case.

The minimax log loss regret, R_ϕ, after learning N examples for a specific feature selection policy ϕ, is:

$$R_\phi = \min_q \max_\theta \mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\big(y|x,x^N,y^N\big)}\right)\right\} \tag{1}$$

where the expectation in (1) is performed over the joint probability:

$$p\big(y,x,x^N,y^N|\theta\big) = p(y|\theta,x)\Pi_{t=1}^N p(y_t|x_t,\theta) \phi\Big(x_t|x^{t-1},y^{t-1}\Big)p(x) \tag{2}$$

and $\phi(x_t|x^{t-1},y^{t-1})$ is the sequential selection policy which gives a probability distribution for each training feature, x_t, based *only* on the past observed training data x^{t-1}, y^{t-1}. Another assumption we make is that $p(x|\theta) = p(x)$ since the feature prior should be independent of the model.

Remark 1: Note that the selection may be stochastic, which means that after observing the past examples there may be some randomness in choosing the next feature. For example, in passive learning, the distributions $\{\phi(x_t|x^{t-1},y^{t-1})\}_{t=1}^N$ are uniform, since the examples are drawn uniformly from the training pool.

In active learning we wish to optimize the examples taken, or in our case, the selected policy, ϕ. Therefore we would like to minimize (1) over $\{\phi(x_t|x^{t-1},y^{t-1})\}_{t=1}^N$. The final active learning problem formulation can be stated as finding the policy ϕ which minimizes R_ϕ, i.e.,

$$R = \min_{\{\phi_t\}_{t=1}^N} \min_q \max_\theta \mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\big(y|x,x^N,y^N\big)}\right)\right\} \tag{3}$$

We derive the following theorem which is the basis for the active learning criterion.

Theorem 1 (Redundancy-Capacity): The minimax active learning problem defined in (3) is equivalent to the conditional model capacity,

$$R = \min_{\{\phi(x_t|x^{t-1},y^{t-1})\}_{t=1}^{N}} \max_{\pi(\theta)} I\left(Y;\theta|X,Y^N,X^N\right) \tag{4}$$

The optimal learner is:

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

where $\pi(\theta)$ is a capacity achieving distribution for the channel $\theta \to Y$ given the test X and training X^N, Y^N.

A proof for Theorem 1. was provided in [24] but a shorter proof in also provided in Appendix A. for completeness. Note that for any prior distribution $\pi(\theta)$ with $\theta \in \Theta$, any policy ϕ and a given model class $p(y|x,\theta)$, the mutual information $I(\theta; Y|X, X^N, Y^N)$ is well defined. The active learning designer finds ϕ and π which solve the minimax problem in (25). Once $\pi(\theta)$ is known, the optimal learner q^* is given by (5) for any realization of x^N, y^N, x.

Theorem 1, is denoted as *Redundancy-Capacity* since it is very similar to the classical result in universal prediction, with the same name, proposed in [25]. In universal prediction, a stream of samples is given sequentially to a predictor and the objective is to predict the next sample based on the constraint that the samples originate from a source belonging to some predefined set of distributions. The *Redundancy-Capacity* links the minimax prediction problem with channel capacity.

Theorem 1, proposes a new criterion for optimal selection policy in active learning. The objective is to find a selection policy which will *minimize* the conditional capacity between the model parameters and test label given the test feature and training data. This is different than active learning strategies used today which do not take into account the test feature prior, $p(x)$, and instead maximize the mutual information between the training and model ignoring the test set if available. In practical applications, if the test set is available, then there will be a dedicated pre-processing stage to prune the training set from data points which seem irrelevant to the test scenario. This step is implicitly preformed by the proposed criterion. This has the potential to significantly improve performance in active learning for priors which are multi-modal and help avoid learning sub-spaces of features which are non-informative for the test scenario. There is of course an issue on how to find $p(x)$, and if such a probability even exists. However, since the main bottleneck in machine learning is the labeling process and not the amount of training features, then we can assume we can estimate the feature probability in some way and come up with an approximation of $p(x)$ or the relative occurrences of features in real life.

The following theorem states that the optimal $\phi(x_t|x^{t-1},y^{t-1})$ places all the probability mass on a specific feature, and is essentially deterministic given the history.

Theorem 2 (Optimal Selection Policies): The selection policies which optimize (25) are deterministic:

$$\phi\left(x_t|x^{t-1},y^{t-1}\right) = \delta\left(x_t - f\left(x^{t-1},y^{t-1}\right)\right)$$

where $\delta(\cdot)$ is the Dirac or Kronecker delta function for continuous or discrete x_t respectively and $f(x^{t-1},y^{t-1})$ is a deterministic function from the history x^{t-1}, y^{t-1} sequence to a feature x_t.

The proof in provided in Appendix B.

The optimization of (25) is unfortunately intractable for many hypotheses classes. The reason is that the number of candidate policies grows exponentially fast and thus infeasible to search for the best possible policy. Moreover, the objective function is not sub-modular or adaptively sub-modular [26] and thus greedy algorithms are not guaranteed to converge in the general case. In future work, we plan to explore different methods to find approximately optimal solutions for this problem.

A. Interpretation of the Proposed Criterion

The proposed criterion, which is denoted as UAL, can be decomposed in the following manner using the chain rule:

$$\begin{aligned} I\left(\theta; Y|X,Y^N,X^N\right) &= I(\theta; Y|X) + I\left(\theta; Y^N|X^N,Y,X\right) \\ &\quad - I\left(\theta; Y^N|X^N\right) \end{aligned} \tag{6}$$

$I(\theta; Y|X)$ does not depend on the selection policy and the optimization is only on the difference between two other mutual information terms. We denote $I(\theta; Y^N|X^N, Y, X)$ and $I(\theta; Y^N|X^N)$ as the exploitation and exploration respectively.

Exploitation means that if the test feature and label, (X, Y), were known in advance, then we would like to select the training examples which will be as correlative to the test as possible. If we select training features X^N such that for a given X, Y, the training labels Y^N will be highly indicative of the test, then these labels will be independent of the model parameter θ and thus $I(\theta; Y^N|X^N, Y, X)$ will be minimized. Moreover, this criterion takes into account the prior probability $p(x)$ and tries to find the best examples averaged across this prior.

Exploration is identical to the criterion used in [17] which basically means that one wants to find the most uncertain example in the pool. Therefore, UAL balances between exploration and exploitation and finds the most informative example given the specific test set at hand.

B. Relation to Other Information Theoretic Active Learning Criteria

In this section the relation UAL and other criteria such as BALD [17] and Maximum Uncertainty (MU) [16] is analyzed. First, a brief review of these criteria is provided.

The MU criterion [16] selects the feature based on:

$$x_t^* = \operatorname*{argmax}_{x_t} H\left(Y_t|X_t = x_t, x^{t-1}, y^{t-1}\right) \tag{7}$$

MU basically selects a feature in the training set whose conditional label entropy based on the current training set is the highest. Since the current model cannot label this feature well, then this example can improve the learning process best. However, this example may be noisy and produce high entropy, thus the learner will now add noise to the training set and this is of course not helpful to the learning task and the labeling budget.

The BALD criterion is defined as:

$$x_t^* = \underset{x_t}{\operatorname{argmax}}\, I\Big(\theta;\, Y_t|X_t = x_t, x^{t-1}, y^{t-1}\Big) \tag{8}$$

According to [17], the objective is to find a feature x_t that maximises the decrease in expected posterior entropy and that will reduce the hypotheses class as fast as possible. It is obvious by the definition of mutual information, MU is an upper bound on BALD.

Both of these criteria make sense in an intuitive manner but lack a firm mathematical foundation rooted in a clear learning problem formulation. There is no clear concept what is the prior of the model θ and no use of the prior on the test features, $p(x)$. Nevertheless, BALD is used to produce very good results for active learning using deep neural network [27] and [21].

In order to relate BALD to UAL, for discrete valued labels Y, (6), becomes:

$$I\big(\theta;\, Y|X, Y^N, X^N\big) \le I(\theta;\, Y|X) + \sum_{i=1}^{N} H(Y_i|X_i, Y, X) - I\big(\theta;\, Y^N|X^N\big) \tag{9}$$

which can be further simplified using the chain rule:

$$I\big(\theta;\, Y|X, Y^N, X^N\big) \le I(\theta;\, Y|X) + N\log_2\left(|\mathcal{A}|\right) - \sum_{i=1}^{N} I\Big(\theta;\, Y_i|X_i, Y^{i-1}X^{i-1}\Big) \tag{10}$$

where $Y \in \mathcal{A}$ and $|\mathcal{A}|$ is the size of the alphabet of the random label Y.

The first two terms in (10) are constant and thus the minimization of the R.H.S is only performed on the third term, which turns into a maximization of $\sum_{i=1}^{N} I(\theta;\, Y_i|X_i, Y^{i-1}X^{i-1})$. Therefore, BALD is basically a greedy algorithm which tries to minimize an upper bound on UAL for certain hypotheses classes. BALD does not take into account the minimization of the term $I(\theta;\, Y^N|X^N, Y, X)$ which, as described before, is related to refining the model for the specific test distribution.

C. Empirical Comparison Between Criteria

In this section, UAL is compared to BALD, MU and passive learning in an empirical test using Gaussian Process Classification (GPC) over a synthetic data set. GP's are a powerful and popular non-parametric tool for regression and classification and a detailed introduction to them can be found in [28].

In [17], BALD was analyzed using GPC and compared to other active learning algorithms including MU. In this section, the same mathematical model and approximations as in [17] are used and repeated again for clarity.

The probabilistic model underlying GPC is as follows:

$$\begin{aligned} f &\sim GP(\mu(\cdot), k(\cdot,\cdot)) \\ y|\underline{x}, f &\sim \text{Bernoulli}\big(\Phi\big(f(\underline{x})\big)\big) \end{aligned} \tag{11}$$

where the parameter f, is a function of a feature point $\underline{x}$ 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 $\Phi(f(\underline{x}))$, where Φ is the Gaussian CDF.

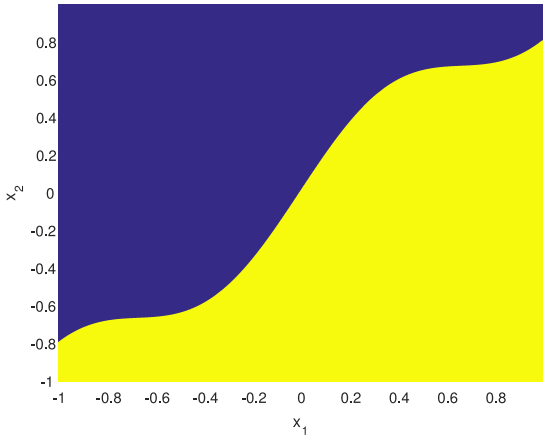

Fig. 1. Training Set, different colors indicate the label of each feature.

Inference in GPC is intractable, since given a training set, the posterior over f (per feature $\underline{x}$) becomes non-Gaussian and complicated. In the following test, Expectation Propagation (EP) [29] was used for approximating this posterior.

UAL requires the computation of $I(f;\, y|\underline{x}, \underline{x}^n, y^n)$, which can be written using (6) as:

$$I\big(f;\, y|\underline{x}, \underline{x}^n, y^n\big) = Const + I\big(f;\, y^n|\underline{x}, y, \underline{x}^n\big) - I\big(f;\, y^n|\underline{x}^n\big) \tag{12}$$

Defining $\mathcal{D} = \{\underline{x}^{n-1}, y^{n-1}\}$ and using the approximations from [17]:

$$I\big(f;\, y_n|\underline{x}_n, \mathcal{D}\big) \approx H\left(\Phi\left(\frac{\mu_{\underline{x}_n,\mathcal{D}}}{\sqrt{\sigma^2_{\underline{x}_n,\mathcal{D}} + 1}}\right)\right) - \frac{C}{\sqrt{\sigma^2_{\underline{x}_n,\mathcal{D}} + C^2}} e^{\left(\frac{-\mu^2_{\underline{x}_n,\mathcal{D}}}{2\left(\sigma^2_{\underline{x}_n,\mathcal{D}} + C^2\right)}\right)} \tag{13}$$

where $C = \frac{\phi \ln 2}{2}$ and $\mu_{\underline{x}_n,\mathcal{D}}, \sigma^2_{\underline{x}_n,\mathcal{D}}$ are the mean and variance of the Gaussian approximation (using EP) for the posterior $p(f|\underline{x}_n, \mathcal{D})$.

Defining $\tilde{\mathcal{D}} = \{\underline{x}^{n-1}, y^{n-1}, x, y\}$ and using (12) to approximate $I(f;\, y_n|\underline{x}_n, \tilde{\mathcal{D}})$:

$$I\Big(f;\, y_n|\underline{x}_n, \tilde{\mathcal{D}}\Big) \approx \mathbb{E}_{\underline{x}}\left(\sum_{y=-1}^{1} I\Big(f;\, y_n|\underline{x}_n, \tilde{\mathcal{D}}\Big) p\big(y|\underline{x}, \mathcal{D}\big)\right) \tag{14}$$

where $p(y|\underline{x}, \mathcal{D}) \approx \int \Phi(f)\mathcal{N}(f;\, \mu_{\mathcal{D}}, \sigma^2_{\mathcal{D}})df$

The synthetic data set consists of two dimensional feature vectors with binary labels as shown in Figure 1, where the yellow color indicates '−1' label and blue is '+1'. In Figure 2, the test set is shown and is basically a smaller sub-set of the feature space. This simulates a scenario where the test is concerned with a particular region of the feature space and there is no real need to learn the whole labeling function which may be very complex and require many data points. In practice, there may be a pre-processing stage which prunes the training set from data point which are irrelevant to the test, but this requires domain knowledge regarding the similarity between data points. On the other hand, UAL, implicitly, takes

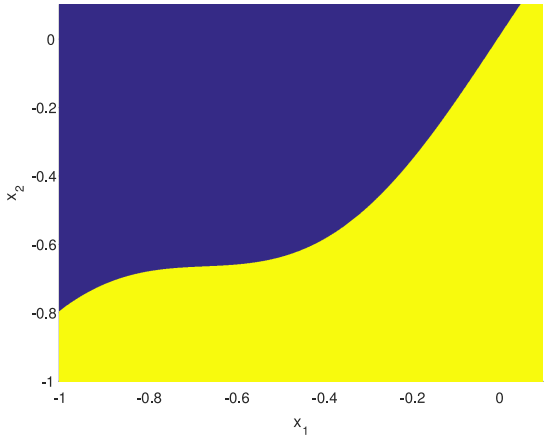

Fig. 2. Test Set, different colors indicate the label of each feature.

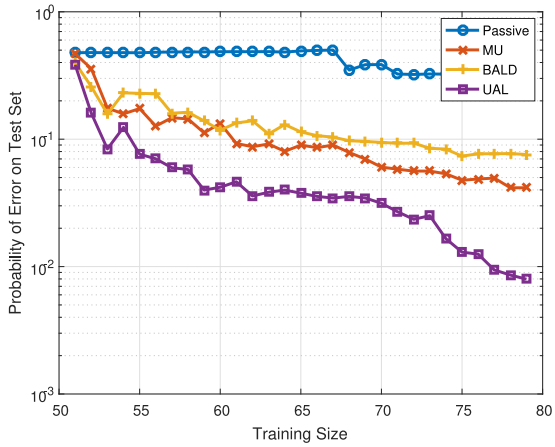

Fig. 3. Error Probability as computed on the test set.

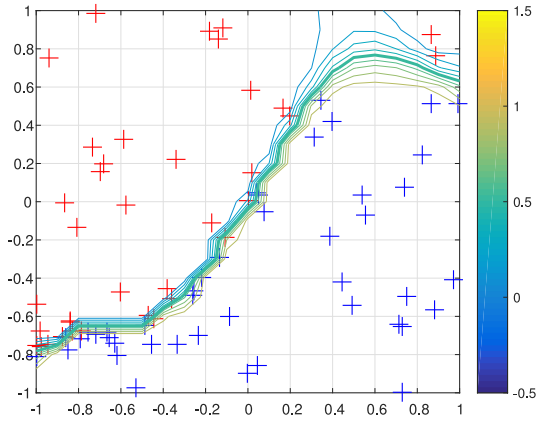

Fig. 4. GPC predictive probability contour lines with data points acquired using UAL.

into account the unlabelled test data to improve the resulting classifier.

The unlabelled test set is given to the learner along with an initial labelled training set and the active selection of training data starts. The active learning process 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 3. It can be observed that passive has the worst performance in terms of sample complexity given error probability. BALD and MU have comparable performance since they do not utilize the test set features and simply sample the boundary curve at multiple locations. UAL does take into account the test set and in Figure 4 one can see a large concentration of training point in the test set region. In Figure 4, the contours of the predictive probability for each test point is plotted. Also, the labelled training data consisting of: 50 random initial training data points and 30 data points selected by UAL is plotted. We can see good fit to the test set as depicted in Figure 2.

III. Linear Regression

In this section, UAL is applied for the linear regression hypothesis class. It is shown that UAL aligns with commonly used criteria for this setting and provides an alternative derivation for these criteria.

The underlying model for the hypothesis class, is defined as:

$$y = \underline{x}^T \underline{\theta} + z \tag{15}$$

where $y, \underline{x}, \underline{\theta}$ and n are the observed response/ label, feature vector, model parameters vector and additive white Gaussian noise with zero mean and unit variance respectively. We also add a power constraint $\mathbb{E}(\frac{1}{d}\|\underline{\theta}\|^2) \leq \sigma_\theta^2$, where d is the dimension of the vector $\underline{\theta}$.

The goal of active learning in this setting is to pick a small number of feature vectors, $\underline{x}^N$, from the space of possible features so that the underlying model, which relates input variables to output responses, is estimated accurately. The optimal solution to the linear regression problem is called the Ordinary Least Squares (OLS) solution. The linear regression model has the property that the error covariance matrix depends on neither the true parameter vector $\underline{\theta}$ nor the observed response y. This suggests that we can "optimize" the covariance of the estimator a-priori, even before taking any measurements, transforming the problem from interactive querying an oracle to subset selection of feature vectors. Active linear regression has been studied extensively under the mathematical field of "Optimal Experiment Design" and a summary of this field can be found in [30].

In linear experimental design, multiple feature vectors are sampled and a linear model is derived:

$$\underline{y} = X\underline{\theta} + \underline{z}$$

where $X \in \mathbb{R}^{nxp}$ is a matrix of n feature vectors, $y \in \mathbb{R}^n$ is the vector of observable responses and $\underline{z} \in \mathbb{R}^n$ is an i.i.d Gaussian noise vector with zero mean and finite variance σ_Z^2.

The classical experimental design is defined as selecting a small subset $S \subset \{1, \ldots, n\}$ r rows, X_S, from X so that estimating $\underline{\theta}$ is optimized on the selected design X_S. Using the selected training set, one can derive the OLS solution for the parameter vector $\underline{\theta}$ and since we are looking for S such that X_S is most statistically efficient, the optimal design problem reduces to minimizing the covariance matrix $\Sigma^{-1} = (X_S^T X_S)^{-1}$. In [30] several optimality criteria have been developed for measuring how well Σ^{-1} is minimized on a selected design X_S. In [31],

performance guarantees for the greedy solution of experimental design problems are provided. In particular, it focuses on A and E optimal designs, for which typical guarantees do not apply since the mean-square error and the maximum eigenvalue of the estimation error covariance matrix are not sub-modular.

A. Comparison Between BALD and UAL

Applying UAL and assuming the noise is Gaussian with the response model (15):

$$I\left(\underline{\theta}; y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) = h\left(y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) - h\left(y_{test}|\underline{\theta}, \underline{x}_{test}, \underline{x}^n, y^n\right) \quad (16)$$

Since the noise, z is independent from the label y given the feature vector $\underline{x}$, then

$$I\left(\underline{\theta}; y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) = h\left(y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) - h(z) \quad (17)$$

Using the expression for Gaussian entropy,

$$I\left(\underline{\theta}; y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) = h\left(y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) - \frac{1}{2}\log\left(2\pi e\sigma_Z^2\right) \quad (18)$$

UAL first finds the prior $\pi(\underline{\theta})$ which maximizes the mutual information in (18). Since there is a power constraint on $\underline{\theta}$ then $\underline{y}$ will also be power limited due to the linear model.

The distribution which will maximize the differential entropy for $\underline{y}|X, \underline{\theta}$ under the power constraint is an i.i.d Gaussian distribution. This distribution can be achieved if the prior $\underline{\theta} \sim \mathcal{N}(0, \sigma_\theta^2 I_d)$ is used. Therefore, in the case of the linear regression hypothesis class, the capacity achieving prior can be computed analytically.

Using [32],

$$I\left(\underline{\theta}; y_{test}|\underline{x}_{test}, \underline{x}^n, y^n\right) = \mathbb{E}_{\underline{x}_{test}}\left(\log\left(1 + \underline{x}_{test}^T Q \underline{x}_{test}\right)\right) \quad (19)$$

where $Q = (X^T X + \frac{1}{\sigma_\theta^2} I_d)^{-1}$ is the inverse covariance matrix of $p(\underline{\theta}|\underline{x}^n, y^n)$ which is also Gaussian and thus easy to compute using Kalman filtering. The expectation is performed on the distribution of the test features $\underline{x}_{test}$.

Upper bounding (19) we get,

$$\mathbb{E}_{\underline{x}_{test}}\left(\log\left(1 + \underline{x}_{test}^T Q \underline{x}_{test}\right)\right) \leq \mathbb{E}_{\underline{x}_{test}}\left(\underline{x}_{test}^T Q \underline{x}_{test}\right) \quad (20)$$

where the bound is tight when $\underline{x}_{test}^T Q \underline{x}_{test} << 1$, which corresponds to high SNR scenarios.

Therefore,

$$\min_{\underline{x}^n} I\left(\underline{\theta}; y|\underline{x}_{test}, \underline{x}^n, y^n\right) \leq \min_{\underline{x}^n} Tr\left(\mathbb{E}\left(\underline{x}_{test}\underline{x}_{test}^T\right) Q\left(\underline{x}^n\right)\right) \quad (21)$$

This criterion is closely related to the A and V optimal design criteria [30]. Note that the matrix Q is a function of the training features $\underline{x}^n$ only and have no dependence on their respective labels y^n. Therefore, there is no real need for online feedback in the active linear regression problem and the training set problem can be cast as a subset selection problem performed offline. This problem is NP hard and thus approximate solutions are needed.

Another observation is that (21) is exactly the transductive experimental design proposed heuristically in [33] and UAL has provided the mathematical reasoning for this criterion.

On the other hand, BALD will try to sequentially maximize the conditional mutual information $I(\theta; y^n|\underline{x}^n)$. It is not clear which prior $\pi(\underline{\theta})$ should be used for BALD since this was not addressed in [17]. Thus, we will use the same prior used in UAL and the same entropy calculation to get the respective BALD criterion:

$$\min_{\underline{x}^n} I\left(\theta; y^n|\underline{x}^n\right) = \min_{\underline{x}^n} \log\det\left(Q\left(\underline{x}^n\right)\right) \quad (22)$$

BALD converges to D-optimal design [30]. Note that D-optimal design is a sub-modular objective and thus greedy optimization as BALD suggests will provide a close to optimal solution.

Therefore, UAL and BALD converge to two different experimental design criteria which are suited for different applications as described in [30]. Note that MU in this case will be identical to BALD since $h(y_t|\underline{x}_t, \underline{\theta}, \underline{x}^{t-1}, y^{t-1}) = h(z)$.

IV. Linear Separators With Label Noise

In this section, UAL is analyzed for learning half-spaces in $\mathbb{R}^n$. This learning problem is probably the most well studied for active learning with well established bounds and algorithms for different label noise models and feature priors. In [13], the first algorithm to achieve near optimal sample complexity for a noiseless Oracle, using margin based active learning was proposed. This algorithm performs well under low noise conditions and log-concave feature distributions. In [34], an efficient Perceptron-based algorithm for active learning homogeneous half-spaces under the uniform distribution over the unit sphere was proposed. This algorithm performs well also under the bounded noise condition [35], where each label is flipped with probability at most $\eta \leq 0.5$. In [36], a margin based algorithm is presented which handles bounded noise using a polynomial regression approach for shrinking the disagreement region. However, all these algorithms achieve good sample complexity only under log concave feature priors and symmetric binary noise models, i.e.,:

$$P(y = 1|x \geq \theta) = P(y = 0|x \leq \theta)$$

In this contribution, we would like to address a general case of noisy Oracles for learning hyper-planes. We would like to analyze the achievable performance of UAL and verify that it behaves as other active learning criteria behave for this hypothesis class.

The model for the noisy Oracle is based on an hypotheses class composed of a one dimensional linear separator with threshold θ_0, followed by a BAC (Binary Asymmetric Channel) with parameters (p, q), as described in Fig. (5). The higher dimensional linear separator is generalized accordingly. The parameters p, q are assumed in this work to be known a-priori and in future work we will address the joint estimation of θ, q and p using active learning.

The algorithm which will be developed in this section can handle any Discrete Memory-less Channel (DMC) noise which can be asymmetric as shown in Fig. (5). Also, we will not

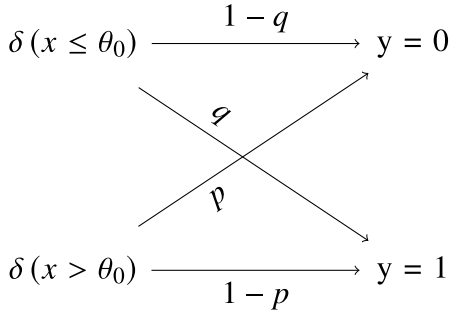

Fig. 5. Noisy Asymmetric One Dimensional Linear Separator.

use the assumption of log-concave feature priors since our algorithm will not randomly draw features from the pool and thus eliminate the need for log-concavity of this prior. The algorithm proposed will have a polynomial computational complexity making it usable for real-world usage. Finally, the achievable performance for UAL in the linear separator hypothesis class is examined with the proposed algorithm.

A. Communication With Noiseless Feedback and the Posterior Matching Scheme

The problem of active learning a classifier with a noisy oracle is closely related to communication over a noisy binary channel with noiseless feedback. In this section, we will discuss this relationship and provide a short overview of Posterior Matching, [37], which is a capacity achieving transmission scheme which utilizes noiseless feedback.

In the world of communications theory the problem of achieving capacity in noisy binary channels is well studied, where the underlying objective is to develop coding and decoding schemes which approach zero error probability as the block length grows. In order to achieve transmission at capacity approaching rates, one needs to develop complex channel codes and employ computationally intensive decoding algorithms. Feedback cannot increase the capacity of memoryless channels as proved by *Shannon*, but utilization of noiseless feedback can boost reliability, allow rate adaptation to cope with unknown channels and significantly simplify transmission schemes. In [38], Horstein presented a simple feedback utilizing scheme for the Binary Symmetric Channel (BSC). In that work, information is represented by a uniformly distributed message point over the unit interval, its binary expansion representing an infinite random binary sequence. The message point is then conveyed to the receiver in an increasing resolution by always indicating whether it lies to the left or to the right of its posterior distribution's median, which is also available to the transmitter via feedback. This, in analogy to active learning, is to transmit the point which answers the most informative binary question that can be posed by the receiver based on its received information. Bits from the binary representation of the message point are decoded by the receiver whenever their respective intervals accumulate a sufficient posterior probability mass.

In [37], Shayevitz and Feder showed that Horstein's method is a specific instance of a more general approach which they called Posterior Matching (PM). This scheme utilizes the noiseless feedback to achieve capacity for any Discrete Memory less Channel (DMC). The flow of PM is as follows: At each time instance, the transmitter computes the posterior distribution of the message point given the receiver's observations. According to the posterior, it "shapes" the message point into a random variable that is independent of the receiver's observations and has the desired input distribution, and transmits it over the channel. Intuitively, this random variable captures the information still missing at the receiver, described in a way that best matches the channel input. In the special cases of a BSC with uniform input distribution, PM is reduced to Horstein's scheme.

The PM scheme is defined for a channel input and output X and Y respectively with known prior and channel transition probability law: $P(x)$ and $P(Y|X)$ respectively. As with active learning, the channel output Y_{t-1} is passed to the transmitter via noiseless feedback and helps the PM scheme to generate a new channel input X_t. The receiver can then use all the received signals Y^t to generate an estimate of the message θ_0.

The next channel input is given by:

$$X_{t+1} = F_X^{-1}\left(F_{\theta_0|Y^t}\left(\theta_0|Y^t\right)\right) \tag{23}$$

where F_X, $F_{\theta_0|Y^t}$ and θ_0 are c.d.f's and the message respectively.

B. One Dimensional Noisy Linear Separator

In Fig. 6 the basic flow diagram of the learning problem is shown. The feature x_t is selected by a selection policy ϕ based on the past training. This feature is passed through a one dimensional linear separator, generating a single bit, representing the true label associated with this feature. This label is passed through a noisy channel and this is basically the mechanism generating the training features and labels.

The learning flow in Fig. 6 can also be viewed as a communications problem, as observed by [1] and others. The oracle and learner can be viewed as: a transmitter, channel and feedback as detailed in the dashed boxes. The transmitter's output is the "clean" label bit generated by some feedback driven coding scheme. In order to have as few oracle labeling operations as possible, the objective will be that the Oracle "transmit" as few bits over the noisy channel as possible and still have enough samples so the classifier will have low error probability. The input to the noisy Oracle can be viewed as a coding function on a message θ and then transmission through the noisy channel. This is exactly the same as designing a transmission scheme which achieves capacity over this channel.

This scheme generates $X_t \sim P_X$ which are independent of Y^{t-1} and this is basically a two step procedure of zooming in on the interesting region in the posterior and matching to the channel input distributing.

In the next theorem, it is shown that PM based active learning (with appropriate input channel distribution) produces a selection policy such that the active learning criterion for the one dimensional threshold decays exponentially fast to zero. Moreover, this result provides an exponent for the decay of (4),

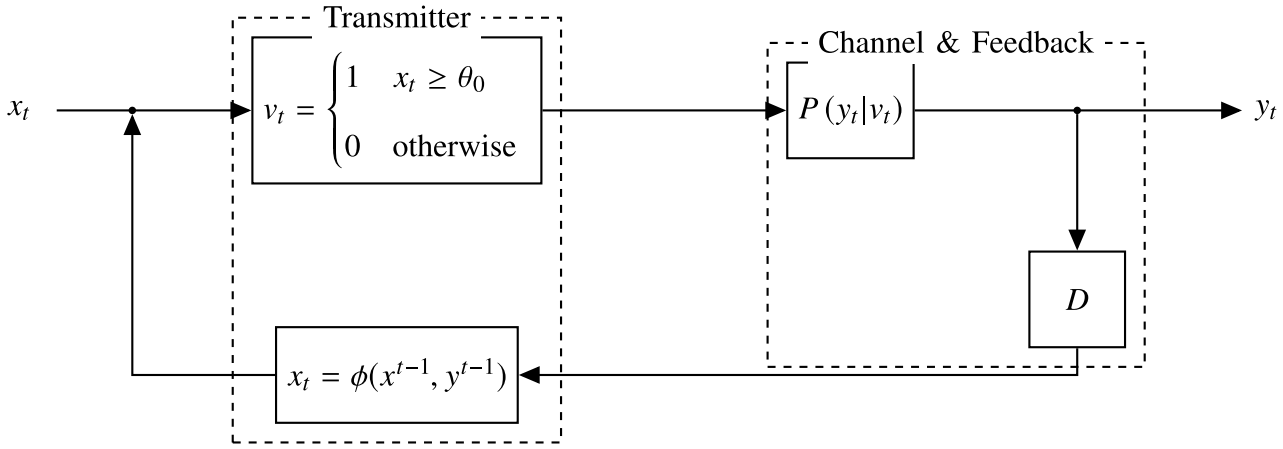

Fig. 6. 1-Dimensional Noisy Linear Separator Block Diagram.

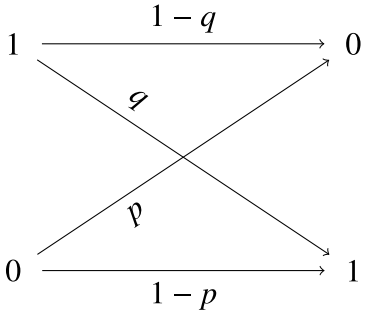

Fig. 7. Binary Asymmetric Channel.

which is equivalent to the *Shannon* capacity of the noisy channel (W) - C_W.

Theorem 3: The 1-dimensional barrier hypotheses class is defined as:

$$p(v|x,\theta) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

where the input is x, output is v and the threshold is θ. The output, v, is the input to a Binary Asymmetric Channel (BAC) with output, y, as defined in Figure 7 and $\forall x \in X$, $p(x) \leq \alpha$.

PM induced active learning produces a selection policy such that:

$$\lim_{n\to\infty} I\big(\theta; Y|X, x^n, y^n\big) = O\Big(2^{-nC_W}\Big)$$

where C_W is the Shannon capacity of the BAC channel W and $\pi(\theta)$ is a uniform distribution on the appropriate interval.

The proof is detailed in Appendix C.

Remark 2: What happens if p and q are unknown? For the Binary Symmetric Channel (BSC), if there exists an upper bound on p, then one can transmit, in principle, at any rate below the capacity derived from this upper bound. More generally, this is proved in [39, Th. 8] under the discussion on *Mismatch Achievability*. In that theorem, Shayevitz and Feder prove that when the true channel is $p(Y^*|X^*)$ and induces some stationary input distribution $p(X^*)$. Then a scheme designed for a pair of an input distribution $p(X)$ and a noisy channel $p(Y|X)$ will have a penalty in the rate (relative to $I(X^*;Y^*)$) given by: $D(p(Y^*|X^*)||p(Y|X)|p(X^*)) - D(p(Y^*)||p(Y))$, where D is the Kullback-Leibler divergence. Therefore, one can use PM with a mismatched prior and channel model and achieve a rate which is lower than the actual capacity of the channel.

In Theorem 3, the prior $\pi(\theta)$ is chosen to be uniform since the convergence of PM to the correct message θ is guaranteed for a uniform prior on the messages. However, the mutual information maximizing prior $\pi^*(\theta)$ of (4) may not be uniform. In the next theorem, it is proven that when using the capacity achieving prior and the training set selected by PM (using uniform prior), UAL still decays to zero at the same exponential rate.

Theorem 4: Given a training set (x^n, y^n) selected by PM using a uniform prior $\pi_u(\theta)$ and

$$\pi^*(\theta) = \underset{\pi(\theta)}{\operatorname{argmax}}\, I\big(Y; \theta|X, Y^n, X^n\big)$$

Then,

$$\lim_{n\to\infty} I\big(\theta; Y|X, x^n, y^n\big) = O\Big(2^{-nC_W}\Big)$$

where the conditional mutual information above is computed using the prior $\pi^*(\theta)$

This theorem basically means that the uniform prior is as good as the capacity achieving prior.

Proof: Since (x^n, y^n) were selected using PM, then based on the results from [39], the posterior satisfies:

$$\lim_{n\to\infty} \sup_{\theta_1} \int_{\theta_1}^{\theta_1+2^{-nC_W}} p\big(\theta|x^n, y^n\big)d\theta = 1 \tag{25}$$

Using Bayes,

$$\lim_{n\to\infty} \frac{1}{Z} \sup_{\theta_1} \int_{\theta_1}^{\theta_1+2^{-nC_W}} p\big(y^n|x^n, \theta\big)\pi_u(\theta)d\theta = 1 \tag{26}$$

where $Z = \int p(y^n|x^n,\theta)\pi_u(\theta)d\theta$.

We define the interval $A = [\theta_1, \theta_1 + 2^{-nC_W}]$ and thus:

$$\lim_{n\to\infty} \int_{\theta\in A^c} p\big(y^n|x^n, \theta\big)\pi_u(\theta)d\theta = 0 \tag{27}$$

where the set A^c is the complementary set to A.

For the capacity achieving prior $\pi^*(\theta)$, a given training set size n and using Hölder's inequality:

$$0 \leq \int_{\theta\in A^c} p\big(y^n|x^n,\theta\big)\frac{\pi_u(\theta)\pi^*(\theta)}{\pi_u(\theta)}d\theta \leq \int_{\theta\in A^c} p\big(y^n|x^n,\theta\big)\pi_u(\theta)d\theta \int_{\theta\in A^c}\frac{\pi^*(\theta)}{\pi_u(\theta)}d\theta \tag{28}$$

Based on the multiplication law for limits and since the two limits exist, then:

$$0 \leq \lim_{n\to\infty}\int_{\theta\in A^c} p\big(y^n|x^n,\theta\big)\pi^*(\theta)d\theta \leq \lim_{n\to\infty}\int_{\theta\in A^c} p\big(y^n|x^n,\theta\big)\pi_u(\theta)d\theta \lim_{n\to\infty}\int_{\theta\in A^c}\frac{\pi^*(\theta)}{\pi_u(\theta)}d\theta \tag{29}$$

Using (27):

$$0 \leq \lim_{n\to\infty}\int_{\theta\in A^c} p(y^n|x^n,\theta)\pi^*(\theta)d\theta \leq 0\cdot\lim_{n\to\infty}\int_{\theta\in A^c}\frac{\pi^*(\theta)}{\pi_u(\theta)}d\theta \tag{30}$$

Since $\lim_{n\to\infty}\int_{\theta\in A^c}\frac{\pi^*(\theta)}{\pi_u(\theta)}d\theta$ exists and is finite for any probability distribution $\pi^*(\theta)$, then

$$\lim_{n\to\infty}\int_{\theta\in A^c} p\big(y^n|x^n,\theta\big)\pi_u(\theta)d\theta = 0.$$

■

Theorem 3 confirms that UAL behaves similarly to other criteria in the one dimensional linear separator hypothesis class. Moreover, the decay factor for this convergence is provided, which is the *Shannon* capacity of the noisy channel. In the next section, higher dimensional linear separators will be addressed and the exponential decay of UAL will be demonstrated using a novel active learning algorithm.

C. Active Learning Hyper-Planes via Successive Posterior Matching

In this section, a label efficient, low complexity algorithm for active learning high dimension linear separators with noisy labels under bounded prior distributions is proposed. The basic idea is to successively localize the spherical coordinates of the normal vector $\underline{w}$, representing the linear separator, using PM. This algorithm, which is denoted as Successive Posterior Matching (SPM) achieves an exponential improvement over passive learning in label complexity with the label noise channel capacity divided by the dimension as the exponent's decay coefficient.

In this setup, the features $\underline{x}\in\mathbb{R}^d$ are assumed to have a bounded feature distribution, $p(\underline{x})\leq\alpha$, for all $\underline{x}$. The hypotheses class contains all possible homogeneous hyper-planes with normal vector $\underline{w}$. The relation between feature $\underline{x}$ and label v is defined as follows,

$$p\big(v|\underline{x},\underline{w}\big) = \begin{cases} 1 & \text{if } \underline{w}^T\underline{x} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

However, labeling may be a noisy process and the oracle may make errors. The noisy label y, outputted by the oracle is modeled as the output of a binary asymmetric channel detailed graphically in Fig. 7. It is important to note here that the proposed algorithm SPM can also work for a noisy channel with $K\geq 2$ possible output labels and the binary channel is used here for simplicity purposes.

Algorithm 1 Active Learning via Successive Posterior Matching

1: **procedure** SPM
2: Init: $\underline{\hat{\theta}} = [\frac{\pi}{2},\frac{\pi}{2},\frac{\pi}{2},\ldots,\frac{\pi}{2}]$,
3: Init: $\forall i\in[1{:}d-1], p(\theta_i) = Unif[0,\pi]$
4: **for** $i\leftarrow d-1$ to 1 **do**
5: **for** $k\leftarrow 1$ to n **do**
6: $\hat{\theta}_i = F^{-1}_{\theta_i|\underline{x}^i_{1:k-1},y^i_{1:k-1}}\left(\frac{p-0.5}{p+q-1}\right)$
7: $\underline{x}^i_k = [\Pi_{l=1}^{d-1}\sin(\hat{\theta}_l), \cos(\hat{\theta}_{d-1})\Pi_{l=1}^{d-2}\sin(\hat{\theta}_l), \ldots, \cos(\hat{\theta}_i)\Pi_{l=1}^{i-1}\sin(\hat{\theta}_l),\ldots,\cos(\hat{\theta}_1)]$
8: $y^i_k = Label(\underline{x}^i_k)$
9: Update $p(\theta_i|\underline{x}^i_{1:k}, y^i_{1:k})$
10: $\hat{\theta}_i = \hat{\theta}_i + \frac{\pi}{2}$

It is assumed that the parameters of the noisy channel p, q are known a-priori and can be different.

SPM is detailed in Algorithm 1, where the estimations of the spherical coordinates of $\underline{w}$ are denoted by $\underline{\hat{\theta}}$. In the initialization stage, each entry in $\underline{\hat{\theta}}$, is set to $\frac{\pi}{2}$ and its respective posterior is uniform.

Next, in iteration i, SPM localizes the boundary between two hyper planes by querying points $\underline{x}$ with spherical coordinates fixed to $\underline{\hat{\theta}}$ and sweeping over θ_i. After acquiring n training points using PM, the median of $p(\theta_i|\underline{x}^n, y^n)$ is computed. In order to generate $\hat{\theta}_i$, $\frac{\pi}{2}$ is added to the computed median to account for the fact that the normal vector needs to be estimated. This process repeats for the next angle θ_{i-1}. Note that the number of labeling operations is dn where $d+1$ and n are the dimension of the vector space and the number of labeling operations for each iteration, respectively.

In order to analyze the performance of SPM for UAL, the capacity achieving prior $\pi(\underline{\theta})$ needs to be computed. This is quite difficult and a clear analytical solution is hard to find. Therefore, a uniform prior is used and achieves close to optimal performance based on the reasoning from Theorem 4.

The convergence of SPM is detailed in the following theorem:

Theorem 5: Suppose $\underline{x}\in\mathbb{R}^{d+1}$ with a bounded p.d.f on the test feature $\forall\underline{x}$, $p(\underline{x})\leq\alpha$. Also, assume the Oracle is some member of a d dimensional homogeneous hyper-plane hypotheses class followed by a BAC.

Then, SPM algorithm produces a selection policy which satisfies:

$$\lim_{n\to\infty} I\big(\theta; Y|X,\underline{x}^n,y^n\big) = O\Big(2^{-\frac{n}{d}C_W}\Big)$$

where n is the total number of Oracle queries and C_W is the Shannon capacity of the BAC with transition probability W.

The proof is provided in Appendix D.

Note that the update function in step 9 refers to a Bayesian computation of the posterior of the threshold point,

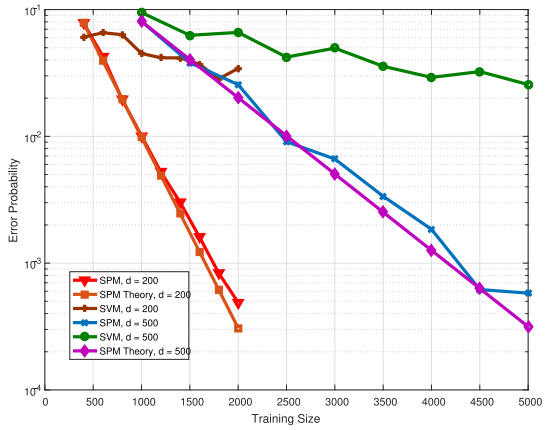

Fig. 8. Error probability for linear separator in $\mathbb{R}^{200}$ and $\mathbb{R}^{500}$ under BAC label noise.

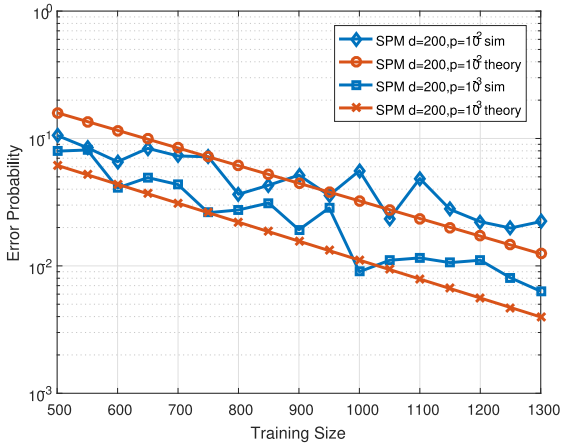

Fig. 9. Error probability for linear separator in $\mathbb{R}^{200}$ with different noise levels.

based on all the observed training examples. The posterior $p(\theta_i|\underline{x}^i_{1:n}, y^i_{1:n})$ is updated at each iteration and the threshold point needs to be localized with very high accuracy. The Naïve approach would be to quantize the interval $[0, \pi]$ and compute the posterior for each quantiztion level. However, this approach is computationally expensive and also limited in accuracy. Since the hypothesis class is a linear separator followed by a noisy binary channel, then the posterior of the intersection angle is a multiplication of different step functions. This enables SPM to only maintain a list of the step points and update the value of the posterior between these points. Since the number of points is exactly the number of training examples, then the complexity of the calculation is proportional to n, and so the whole computational complexity of the algorithm is linear with n and mathematically exact.

D. Simulation Results

In this section, SPM is compared to a widely used passive learning algorithm for learning hyper planes - Support Vector Machine (SVM) which is known to perform very well even in noisy conditions. The comparison will be for feature spaces with $d = 200$ and $d = 500$ and using a BAC with $q = 10^{-3}$ and $p = 10^{-4}$. A Monte Carlo simulation was implemented to estimate the error probability for an active learner based on SPM and a passive learner based on SVM. In Figure 8, the error probabilities as a function of the total number of labels performed are presented for different space dimensions. Each test for $d = 200$ or $d = 500$ has the SVM and SPM error probabilities and also the trend line as predicted by Theorem 5 for SPM. It can be seen that the error probability decay is exponential with the decay factor related to the channel capacity divided by the degrees of freedom, which is in agreement with the theory. In Figure 9, the error probabilities for $d = 200$ with different noise levels: $p = 10^{-2}$ and $p = 10^{-3}$ are plotted and it can be seen that the theory holds in these cases too.

V. Conclusion

In this work, a new criterion for active learning motivated by a Minimax redundancy view of the learning problem was introduced. The relation between this criterion and commonly used active learning criteria was analyzed. The proposed criterion, UAL, intrinsically balances an exploration-exploitation trade-off and thus has the potential to outperform commonly used uncertainty maximization criteria.

Later, the linear separator problem with asymmetric noise was considered and a low complexity, noise robust algorithm for active learning has been presented. It was proven that this algorithm achieves exponential decay of the proposed criterion and empirically shown that the error probability decays exponentially with the same rate to reach the Oracle's error probability.

In future work, the Minimax redundancy objective will be modified, so it will be able to address the fact that the model generating the labels and the inference model may not be in the same hypothesis class.

Appendix A
Proof of Theorem 1

Proof: The proof is very similar to the one in [25] but with small technical modifications. First, we induce a probability measure $\pi(\theta)$ over the parameter θ:

$$R = \min_{\{\phi_t\}_{t=1}^N} \min_q \max_{\pi(\theta)} \mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\left(y|x,x^N,y^N\right)}\right)\right\} \tag{32}$$

where the worst θ is with probability one.

Then, observe that,

$$\mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\left(y|x,x^N,y^N\right)}\right)\right\} = \sum_\theta \pi(\theta) \sum_{x^N,y^N,x} p\left(y^N,x^N,x|\theta\right) D_{KL}\left(p(y|x,\theta)||q\left(y|x,x^N,y^N\right)\right) \tag{33}$$

Since (33) is a non-negative weighted sum of convex functions (for each (x, x^N, y^N), the KL divergence is convex in $q(y|x, x^N, y^N)$) and concave (linear) in $\pi(\theta)$, and the set of distributions is the probability simplex which is compact and convex, then we can apply the Minimax theorem [40].

Plugging (33) in to (32) and using the Minimax theorem,

$$R = \min_{\{\phi_t\}_{t=1}^N} \max_{\pi(\theta)} \min_q \mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\left(y|x,x^N,y^N\right)}\right)\right\} \tag{34}$$

Now we can find the learner q (for each x, x^N, y^N) which optimizes (34) for a given $\{\phi(x_t|x^{t-1}, y^{t-1})\}_{t=1}^N$ and $\pi(\theta)$.

Note that:

$$p\big(\theta|y^N, x^N, x\big) = \frac{p\big(y^N, x^N, x, \theta\big)}{p\big(y^N, x^N, x\big)}$$

Then,

$$\mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q\big(y|x, x^N, y^N\big)}\right)\right\} = \mathbf{E}_{\mathbf{x^N},\mathbf{y^N},\mathbf{x}} \sum_\theta p\big(\theta|y^N, x^N, x\big) D_{KL}\big(p(y|x,\theta)||q\big(y|x, x^N, y^N\big)\big) \tag{35}$$

Then, the optimal q which minimizes the KL divergence is:

$$q^*\big(y|x, x^N, y^N\big) = \sum_\theta p\big(\theta|y^N, x^N\big) p(y|\theta, x) \tag{36}$$

Note that q is optimal regardless of the selection policy and thus optimal for both passive and active learning. The predictor q is a function of the training set and test feature but also loosely dependent (for large N) on $\pi(\theta)$. The optimal prior $\pi(\theta)$ is different for a given selection policy though.

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:

$$\mathbf{E}\left\{\log\left(\frac{p(y|x,\theta)}{q^*\big(y|x, x^N, y^N\big)}\right)\right\} = I\big(Y; \theta|X, Y^N, X^N\big) \tag{37}$$

and $\pi(\theta)$ maximizes the mutual information (capacity achieving distribution) for a given policy. ■

APPENDIX B
PROOF OF THEOREM 2.

Proof: We wish to analyze the conditional mutual information $I(\theta; Y|X = x, Y^n = y^n, X^n = x^n)$. First, we analyze the posterior:

$$p\big(y|x, y^n, x^n\big) = \sum_\theta p(\theta|x, y^n, x^n) p\big(y|x, y^n, x^n, \theta\big) \tag{38}$$

Using the fact that given θ and x, y is independent of x^n, y^n:

$$p\big(y|x, y^n, x^n\big) = \sum_\theta p(\theta|y^n, x^n, x) p(y|x, \theta) \tag{39}$$

Using Bayes,

$$p(\theta|y^n, x^n, x) = \frac{p(y^n, x^n, x|\theta)\pi(\theta)}{p(y^n, x^n, x)} \tag{40}$$

Therefore,

$$p\big(\theta|y^n, x^n, x\big) = \frac{\Pi_{t=1}^n p(y_t|x_t,\theta)\phi\big(x_t|x^{t-1}, y^{t-1}\big)p(x|\theta)\pi(\theta)}{\sum_\theta p(x|\theta)\pi(\theta)\Pi_{t=1}^n p(y_t|x_t,\theta)\phi\big(x_t|x^{t-1}, y^{t-1}\big)} \tag{41}$$

Eliminating $\phi(x_t|x^{t-1}, y^{t-1})$

$$p\big(\theta|y^n, x^n, x\big) = \frac{\Pi_{t=1}^n p(y_t|x_t,\theta) p(x|\theta)\pi(\theta)}{\sum_\theta p(x|\theta)\pi(\theta)\Pi_{t=1}^n p(y_t|x_t,\theta)} \tag{42}$$

Therefore,

$$p\big(y|x, y^n, x^n\big) = \sum_\theta p(y|x,\theta) \frac{\Pi_{t=1}^n p(y_t|x_t,\theta) p(x|\theta)\pi(\theta)}{\sum_\theta p(x|\theta)\pi(\theta)\Pi_{t=1}^n p(y_t|x_t,\theta)} \tag{43}$$

and thus, for a given $\pi(\theta)$, the value of the posterior $p(y|x, y^n, x^n)$ does not depend on the value of the selection policy.

We can write the conditional mutual information explicitly,

$$I\big(\theta; Y|X, Y^n, X^n\big) = \sum_{x,y^n,x^n} I\big(\theta; Y|X = x, Y^n = y^n, X^n = x^n\big) \cdot p\big(x, y^n, x^n\big) \tag{44}$$

Then,

$$I\big(\theta; Y|X, Y^n, X^n\big) = \sum_{x,y^n,x^n} I\big(\theta; Y|X = x, Y^n = y^n, X^n = x^n\big) \cdot \left(\sum_\theta p(x|\theta)\pi(\theta)\Pi_{t=1}^n p(y_t|x_t,\theta) \times \phi\Big(x_t|x^{t-1}, y^{t-1}\Big)\right) \tag{45}$$

which can be written as,

$$I\big(\theta; Y|X, Y^n, X^n\big) = \sum_{y^n,x^n} \phi\Big(x_t|x^{t-1}, y^{t-1}\Big) \cdot I\big(\theta; Y|X, Y^n = y^n, X^n = x^n\big) \cdot \left(\sum_\theta \pi(\theta)\Pi_{t=1}^n p(y_t|x_t,\theta)\right) \tag{46}$$

Since the weighted average of positive values (mutual information is always larger or equal to 0) is always bigger than the minimum of the set, we come to the conclusion that the optimal selection strategy is a delta function for each step which correspond to the trajectory x^n, y^n which minimizes the conditional mutual information $I(\theta; Y|X, X^n = x^n, Y^n = y^n)$. ■

APPENDIX C
PROOF OF THEOREM 3

Proof: In [37], it is proved that PM achieves capacity on the BAC. Achieving capacity essentially means that the maximum amount of bits are transmitted and decoded without error with the minimal amount of channel uses. This is analogous to high accuracy on θ_0 (low generalization error) using as few Oracle calls as possible. This is exactly the target of active learning and we will now show that PM on BAC is equivalent to a specific active learning policy for the hypotheses class discussed here.

The proposed selection scheme selects the training feature, x_t, based on previously observed labels y^{t-1} (x^{t-1} are a deterministic function of y^{t-1}):

$$x_t = F^{-1}_{\theta|y^{t-1}}\left(\frac{p-0.5}{p+q-1}\right) \tag{47}$$

Therefore, the input to the BAC, v_t, is computed according to:

$$v_t = \begin{cases} 0 & x_t \le \theta_0 \\ 1 & x_t > \theta_0 \end{cases} \tag{48}$$

Now, we would like to show that this selection of x_t generates v_t which achieves capacity for the BAC.

Define an auxiliary Bernoulli random variable $Q \sim Ber(\frac{p-0.5}{p+q-1})$ and use the fact that a Cumulative Distribution Function (CDF) is always increasing, then v_t can also be described as:

$$v_t = F_Q^{-1}\left(F_{\theta|Y^{t-1}}(\theta_0)\right) \tag{49}$$

which is exactly the PM scheme for a BAC channel with p, q!

Therefore, the error probability on the message θ approaches zero as the number of channel uses, n, goes to infinity:

$$\lim_{n\to\infty} \sup_{\theta_1} \int_{\theta_1}^{\theta_1+2^{-nC_W}} p\left(\theta|y^n, x^n\right)d\theta = 1 \tag{50}$$

This means that most of the probability mass is centred in an interval of length 2^{-nC_W} where the true barrier, θ_0, resides, where Q is the input distribution to the BAC and W is the BAC transition probability.

Now we can analyze the active learning criterion for the PM selection with training x^n, y^n. We will compute the desired mutual information using the difference of the two conditional entropies:

$$H\left(Y|X, X^n = x^n, Y^n = y^n\right) = \int H_B\left(\int P(Y=1|x,\theta)p\left(\theta|x^n, y^n\right)d\theta\right)p(x)dx \tag{51}$$

and the conditional entropy with θ:

$$H\left(Y|X, \theta, X^n = x^n, Y^n = y^n\right) = \int H_B(P(Y=1|x,\theta))p\left(\theta|x^n, y^n, x\right)d\theta p(x)dx \tag{52}$$

For BAC, the binary entropy conditioned on a specific x and θ, can be written as,

$$H\left(Y|x, x^n, y^n\right) = H_B\left(\left(q\left(1 - F_{\theta|x^n,y^n}(x)\right) + (1-p)F_{\theta|x^n,y^n}(x)\right)\right) \tag{53}$$

$$H\left(Y|x, \theta, x^n, y^n\right) = H_B(q\delta(x \le \theta) + (1-p)\delta(x > \theta)) \tag{54}$$

Therefore,

$$\lim_{n\to\infty} H\left(Y|X, x^n, y^n\right) = \int_0^{\theta_1} H_B(q)p(x)dx + \int_{\theta_1+2^{-nC_W}}^{1} H_B(1-p)p(x)dx + \int_{\theta_1}^{\theta_1+2^{-nC_W}} H_B(q(1-F_\theta(x)) + (1-p)F_\theta(x))p(x)dx \tag{55}$$

where θ_1 is estimated using (50) for a division of the interval $(0, 1)$ to bins on length 2^{-nC_W}.

Similarly,

$$\lim_{n\to\infty} H\left(Y|X, \theta, x^n, y^n\right) \ge \int_0^{\theta_1} H_B(q)p(x)dx + \int_{\theta_1+2^{-nC_W}}^{1} H_B(1-p)p(x)dx \tag{56}$$

Therefore the desired mutual information can be upper bounded by,

$$0 \le \lim_{n\to\infty} I\left(\theta; Y|X, X^n, Y^n\right) \le \alpha 2^{-nC_W} \tag{57}$$

This concludes the proof that active learning via PM achieves exponential decay and the important takeaway here is that the decay factor is dependent on the channel and the input distribution which achieved the capacity. ■

APPENDIX D
PROOF OF THEOREM 5

Proof: Assume there is a homogeneous hyper-plane separating two complementary volumes in $\mathbb{R}^d$. This hyper-plane is defined by a unit length normal vector $\underline{w}$ which can be described by its spherical coordinates $\underline{\theta}$.

The idea of SPM is to successively estimate the spherical coordinates of $\underline{w}$ using PM, one coordinate at a time. In the first iteration, the spherical coordinate, θ_{d-1} is estimated and used for the estimation of the next spherical coordinate, θ_{d-2}. This process repeats until all the coordinates are estimated.

A. SPM Flow

In the first step of Algorithm 1, which corresponds to θ_{d-1}, SPM searches for the intersection point between the hyper plane defined by $\underline{w}$ and an arc, $\underline{r}(\phi)$ defined by the following description:

$$\underline{r}(\phi) = [sin(\phi), cos(\phi), 0, 0, \ldots, 0]$$

for $\phi \in (0, \pi)$.

This problem is a 1-dimensional noisy barrier model on the interval $(0, \pi)$, thus PM will query points in this interval and provide an estimate of the intersection point. The estimated intersection point, $\underline{x}_n^{d-1}$, after n training points can be described as:

$$\underline{x}_n^{d-1} = [sin(\phi_n), cos(\phi_n), 0, 0, \ldots, 0] \tag{58}$$

where ϕ_n is final queried point (angle) in the interval $(0, \pi)$.

The relation between ϕ_n and the estimate $\hat{\theta}_{d-1}$ of the spherical coordinate θ_{d-1} (of $\underline{w}$), is given by:

$$\hat{\theta}_{d-1} = \phi_n + \frac{\pi}{2} \tag{59}$$

Using (50), the following holds:

$$\lim_{n\to\infty} p\left(\theta_{d-1}|\underline{x}^n, y^n\right) = 2^{nC_W} \tag{60}$$

for any $\theta_{d-1} \in [\hat{\theta}_{d-1} - 2^{-nC_W-1}, \hat{\theta}_{d-1} + 2^{-nC_W-1}]$

In the next iteration of step 4 in Algorithm 1, the intersection between the hyper-plane and the arc, $\underline{r}(\phi)$:

$$\underline{r}(\phi) = \Big[sin\Big(\hat{\theta}_{d-1}\Big)sin(\phi), cos\Big(\hat{\theta}_{d-1}\Big)sin(\phi), cos(\phi), 0, 0, \ldots, 0\Big]$$

for $\phi \in (0, \pi)$

The estimated intersection point after n training points:

$$\underline{x}_n^{d-2} = \Big[sin\Big(\hat{\theta}_{d-1}\Big)sin(\phi_n), cos\Big(\hat{\theta}_{d-1}\Big)sin(\phi_n), cos(\phi_n), 0, 0, \ldots, 0\Big]$$

Again, the estimated spherical coordinate is:

$$\hat{\theta}_{d-2} = \phi_n + \frac{\pi}{2} \tag{61}$$

This process goes on for all the spherical coordinates.

B. Proof Idea

Now that we have detailed the mechanism generating the estimates for the spherical coordinates, we can show how the active learning criterion decays for this training set selection policy. The main idea is to show that most of the probability mass of the joint posterior for the spherical coordinates reside inside a narrow enough cone in space, such that the active learning criterion decays exponentially fast to zero.

The active learning criterion, which is the conditional mutual information, is a difference of the conditional entropy of the test label Y given the training and test feature X:

$$H\Big(Y|\underline{X}, \underline{X}^{dn} = \underline{x}^{dn}, Y^{dn} = y^{dn}\Big) \int H_B\Big(\int P\big(Y = 1|\underline{x}, \underline{\theta}\big)p\Big(\underline{\theta}|\underline{x}^{dn}, y^{dn}\Big)d\underline{\theta}\Big)p\big(\underline{x}\big)d\underline{x} \tag{62}$$

and the conditional entropy of the test label Y given the training, test feature X and model parameter θ:

$$H\Big(Y|\underline{X}, \underline{\theta}, \underline{X}^{dn} = \underline{x}^{dn}, Y^{dn} = y^{dn}\Big) = \int H_B\big(P\big(Y = 1|\underline{x}, \underline{\theta}\big)\big)p\Big(\underline{\theta}|\underline{x}^{dn}, y^{dn}\Big)d\underline{\theta}p(\underline{x})d\underline{x} \tag{63}$$

The spherical coordinates posterior can be decomposed using the chain rule for probabilities,

$$p\big(\underline{\theta}|\underline{x}^{n_T}, y^{n_T}\big) = \Pi_{i=d-1}^{1} p\Big(\theta_i|\underline{\theta}_{i+1}^{d-1}, \underline{x}^{n_T}, y^{n_T}\Big) \tag{64}$$

where $n_T = dn$.

We will now concentrate on the individual posteriors and show that they concentrate to the correct spherical coordinates fast.

C. Posterior for θ_{d-2}

For simplicity, we will first compute the posterior $p(\theta_{d-2}|\theta_{d-1}, \underline{x}^{n_T}, y^{n_T})$. After running the PM scheme for θ_{d-2}, all normal vectors, $\underline{w}$, which are possible candidates for the true normal vector, must satisfy the following equality with the estimated threshold point $\underline{x}_n^{d-2}$:

$$\lim_{n\to\infty} \Pr\Big(\Big|\underline{w}^T\underline{x}_n^{d-2}\Big| \le 2^{-nI}|\theta_{d-1}, \underline{x}^n, y^n\Big) = 1$$

This equality basically creates a constraint on the possible values θ_{d-2}, can take and we can explicitly write this as:

$$\Big|\underline{w}^T\underline{x}_n^{d-2}\Big| = \Big|\sin\Big(\hat{\theta}_{d-1}\Big)sin(\phi_n)sin(\theta_{d-1})sin(\theta_{d-2}) + cos\Big(\hat{\theta}_{d-1}\Big)sin(\phi_n)cos(\theta_{d-1})sin(\theta_{d-2}) + cos(\phi_n)cos(\theta_{d-2})\Big| \le 2^{-nI}$$

which can be written as:

$$|sin(\phi_n)sin(\theta_{d-2})\gamma_{d-1} + cos(\phi_n)cos(\theta_{d-2})| \le 2^{-nI} \tag{65}$$

where,

$$\gamma_{d-1} = sin(\theta_{d-1})sin\Big(\hat{\theta}_{d-1}\Big) + cos\Big(\hat{\theta}_{d-1}\Big)cos(\theta_{d-1}) \tag{66}$$

We note that γ_{d-1} is an inner product between two unit length vectors and thus:

$$\gamma_{d-1} = \cos\Big(\theta_{d-1} - \hat{\theta}_{d-1}\Big)$$

and according to (60), with probability approaching to 1 as n goes to infinity, $\gamma_{d-1} \le \cos(2^{-nC_W})$. We also note that since 2^{-nC_W} is small then we can approximate γ_{d-1} using its Taylor expansion:

$$\gamma_{d-1} \approx 1 - \frac{2^{-2nC_W}}{2} \tag{67}$$

Therefore we can approximate (65) as,

$$\Big|sin(\phi_n)sin(\theta_{d-2})\Big(1 - \frac{2^{-2nI}}{2}\Big) + cos(\phi_n)cos(\theta_{d-2})\Big| \le 2^{-nI} \tag{68}$$

This is equivalent to:

$$\Big|\cos(\phi_n - \theta_{d-2}) - \tfrac{2^{-2nI}}{2}\sin(\phi_n)sin(\theta_{d-2})\Big| \le 2^{-nI} \tag{69}$$

We will use the reverse triangle inequality and get:

$$\Big||\cos(\phi_n - \theta_{d-2})| - \Big|\frac{2^{-2nI}}{2}\sin(\phi_n)sin(\theta_{d-2})\Big|\Big| \le 2^{-nI} \tag{70}$$

Therefore,

$$|\cos(\phi_n - \theta_{d-2})| \le 2^{-nI} + \frac{2^{-2nI}}{2}$$

For large enough n, we can expand cosine around $\frac{\pi}{2}$ and get that the angles θ_{d-2} satisfy:

$$\Big|\hat{\theta}_{d-2} - \theta_{d-2}\Big| \le 2^{-nI} + \frac{2^{-2nI}}{2} \tag{71}$$

Which basically means that:

$$\lim_{n\to\infty} \Pr\Big(\Big|\hat{\theta}_{d-2} - \theta_{d-2}\Big| \le 2^{-nI} + \frac{2^{-2nI}}{2}|\theta_{d-1}, \underline{x}^n, y^n\Big) = 1 \tag{72}$$

which basically means for large enough n (d is fixed):

$$\lim_{n\to\infty} \Pr\Big(\Big|\hat{\theta}_{d-2} - \theta_{d-2}\Big| \le 2^{-nI}|\theta_{d-1}, \underline{x}^n, y^n\Big) \approx 1 \tag{73}$$

Therefore, we approximately get the same condition as in (60).

D. Posterior for θ_i

We can now move to the general case of θ_i. We will show using recursion, that the posterior concentrates to the correct value appropriately. The final threshold point after n labeling operations for the i'th spherical coordinate is defined as:

$$\underline{x}_n^i = \Big[\sin(\phi_n)\Pi_{l=1,l\neq i}^{d-1}\sin\Big(\hat{\theta}_l\Big), \cos\Big(\hat{\theta}_{d-1}\Big)\sin(\phi_n)\Pi_{l=1,l\neq i}^{d-2}\sin\Big(\hat{\theta}_l\Big), \ldots, \cos(\phi_n)\Pi_{l=1}^{i-1}\sin\Big(\hat{\theta}_l\Big), \ldots, \cos\Big(\hat{\theta}_1\Big)\Big]$$

Again, due to PM, the following holds:

$$\lim_{n\to\infty}\Pr\Big(\big|\underline{w}^T\underline{x}_n^i\big| \leq 2^{-nI}|\underline{\theta}_{i+1}^{d-1}, \underline{x}^{nT}, y^{nT}\Big) = 1$$

We define the following recursion rule:

$$\gamma_i = sin(\theta_i)sin\Big(\hat{\theta}_i\Big)\gamma_{i+1} + cos\Big(\hat{\theta}_i\Big)cos(\theta_i) \tag{74}$$

with (66) as the initial condition.

The inner product ,$\underline{w}^T\underline{x}_n^i$, can be written as:

$$\underline{w}^T\underline{x}_n^i = \sin(\phi_n)sin(\theta_i)\gamma_{i+1} + cos(\phi_n)cos(\theta_i) \tag{75}$$

If we knew that $\gamma_{i+1} \approx 1 - \frac{2^{-2nC_W}}{2}$ we could use the same arguments from the previous section to bound the posterior. Using (73), $\gamma_{d-2} \approx 1 - \frac{2^{-2nC_W}}{2}$ and applying (74) in recursion, we get:

$$\lim_{n\to\infty}\Pr\Big(\Big|\hat{\theta}_i - \theta_i\Big| \leq 2^{-nI}|\theta_{i+1}^{d-1}, \underline{x}^{dn}, y^{dn}\Big) \approx 1. \tag{76}$$

E. Asymptotic Decay of Mutual Information

Finally, we will use the posteriors computed in the previous sections to give an upper bound on the conditional mutual information. The multiplication of the posteriors, $p(\theta_i|\underline{\theta}_{i+1}^{d-1}, \underline{x}_i^{n_T}, y_i^{n_T})$, form a cone with probability approaching 1 in $\underline{x} \in \mathbb{R}^d$. The unit vector $\hat{w}$ is a vector in the center of this cone. Using the results on $p(\theta_i|\underline{\theta}_{i+1}^{d-1}, \underline{x}_i^{n_T}, y_i^{n_T})$, (53) and (54), we can compute upper bounds on the conditional mutual information.

For the BAC,

$$\begin{aligned} P\big(Y=1|\underline{x}, \underline{x}^{n_T}, y^{n_T}\big) &= q\int 1\big(\underline{x}^T\underline{w}\leq 0\big)\Pi_{i=1}^d p(\theta_i|\underline{\theta}^{i-1}, \underline{x}_i^{n_T}, y_i^{n_T})d\underline{\theta} \\ &+ (1-p)\int 1\big(\underline{x}^T\underline{w}\geq 0\big)\Pi_{i=1}^d p\Big(\theta_i|\underline{\theta}^{i-1}, \underline{x}_i^{n_T}, y_i^{n_T}\Big)d\underline{\theta} \end{aligned} \tag{77}$$

Therefore,

$$\begin{aligned} \lim_{n\to\infty} H\big(Y|\underline{X}, \underline{x}^n, y^n\big) &= \int_{\frac{\langle \underline{x},\hat{w}\rangle}{|\underline{x}|}\leq -2^{-nC_W}} H_B(q)p(\underline{x})d\underline{x} \\ &+ \int_{\frac{\langle \underline{x},\hat{w}\rangle}{|\underline{x}|}>2^{-nC_W}} H_B(1-p)p(\underline{x})d\underline{x} + \int_{\frac{|\langle \underline{x},\hat{w}\rangle|}{|\underline{x}|}\leq 2^{-nC_W}} \\ H_B\big(q\big(1-F_{\underline{\theta}|\underline{x}^n,y^n}(\underline{x})\big) &+ (1-p)F_{\underline{\theta}|\underline{x}^n,y^n}(\underline{x})\big)p(\underline{x})d\underline{x} \end{aligned} \tag{78}$$

Similarly,

$$\begin{aligned} \lim_{n\to\infty} H\big(Y|\underline{X}, \theta, \underline{x}^n, y^n\big) &\geq \int_{\frac{\langle \underline{x},\hat{w}\rangle}{|\underline{x}|}\leq -2^{-nC_W}} H_B(q)p(\underline{x})d\underline{x} \\ &+ \int_{\frac{\langle \underline{x},\hat{w}\rangle}{|\underline{x}|}>2^{-nC_W}} H_B(1-p)p(\underline{x})d\underline{x} \end{aligned} \tag{79}$$

Therefore the desired mutual information can be upper bounded by,

$$0 \leq \lim_{n\to\infty} I\big(\theta; Y|\underline{X}, \underline{X}^n, Y^n\big) \leq \alpha 2^{-\frac{n_T}{d}C_W}. \tag{80}$$

■

REFERENCES

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

[2] M. Raginsky and A. Rakhlin, "Lower bounds for passive and active learning," in *Proc. NIPS*, 2011, pp. 1026–1034.

[3] S. Hanneke, "A bound on the label complexity of agnostic active learning," in *Proc. ICML*, 2007, pp. 353–360.

[4] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Mach. Learn.*, vol. 28, nos. 2–3, pp. 133–168, 1997.

[5] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *J. Comput. Syst. Sci.*, vol. 75, no. 1, pp. 78–89, 2009.

[6] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," in *Proc. COLT*, 2007, pp. 35–50.

[7] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[8] S. Dasgupta, "Coarse sample complexity bounds for active learning," in *Proc. NIPS*, 2006, pp. 235–242.

[9] S. Dasgupta, D. J. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Proc. ISAIM*, 2008, pp. 353–360.

[10] S. Hanneke, "Theoretical foundations of active learning," Mach. Learn. Dept., Carnegie–Mellon Univ., Pittsburgh, PA, USA, Rep. CICS-P-160, 2009.

[11] A. Beygelzimer, D. J. Hsu, J. Langford, and T. Zhang, "Agnostic active learning without constraints," in *Proc. NIPS*, 2010, pp. 199–207.

[12] V. Koltchinskii, "Rademacher complexities and bounding the excess risk in active learning," *J. Mach. Learn. Res.*, vol. 11, pp. 2457–2485, Sep. 2010.

[13] M.-F. Balcan and P. Long, "Active and passive learning of linear separators under log-concave distributions," in *Proc. COLT*, 2013, pp. 288–316.

[14] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," in *Proc. STOC*, 2014, pp. 449–458.

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

[16] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. 1648, 2009.

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

[18] M. Raginsky and A. Rakhlin, "Information-based complexity, feedback and dynamics in convex programming," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 7036–7056, Oct. 2011.

[19] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, 1992.

[20] V. V. Fedorov, *Theory of Optimal Experiments*. Amsterdam, The Netherlands: Elsevier, 2013.

[21] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. ICML*, 2017, pp. 1183–1192.

[22] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[23] Y. Fogel and M. Feder, "Universal batch learning with log-loss," in *Proc. ISIT*, 2018, pp. 21–25.

[24] S. Shayovitz and F. Meir, "Minimax active learning via minimal model capacity," in *Proc. Mach. Learn. Signal Process. Workshop (MLSP)*, 2019, pp. 1–6.

[25] R. G. Gallager, "Source coding with side information and universal coding," in *Proc. ISIT*, 1979, pp. 1093–1097.
[26] D. Golovin and A. Krause, "Adaptive submodularity: Theory and applications in active learning and stochastic optimization," *J. Artif. Intell. Res.*, vol. 42, pp. 427–486, Nov. 2011.
[27] A. Kirsch, J. van Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning," in *Proc. NeurIPS*, 2019, pp. 7024–7035.
[28] C. E. Rasmussen, "Gaussian processes in machine learning," in *Proc. Adv. Lectures Mach. Learn.*, 2003, pp. 63–71.
[29] T. P. Minka, "Expectation propagation for approximate Bayesian inference," 2013. [Online]. Available: arXiv:1301.2294.
[30] F. Pukelsheim, *Optimal Design of Experiments*, SIAM, Philadelphia, PA, USA, 2006.
[31] L. Chamon and A. Ribeiro, "Approximate supermodularity bounds for experimental design," in *Proc. NIPS*, 2017, pp. 5403–5412.
[32] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, 1999.
[33] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proc. ICML*, 2006, pp. 1081–1088.
[34] S. Yan and C. Zhang, "Revisiting perceptron: Efficient and label-optimal learning of halfspaces," in *Proc. NIPS*, 2017, pp. 1056–1066.
[35] P. Massart and É. Nédélec, "Risk bounds for statistical learning," *Ann. Stat.*, vol. 34, no. 5, pp. 2326–2366, 2006.
[36] P. Awasthi, M.-F. Balcan, N. Haghtalab, and H. Zhang, "Learning and 1-bit compressed sensing under asymmetric noise," in *Proc. COLT*, 2016, pp. 152–192.
[37] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. ISIT*, 2007, pp. 391–395.
[38] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inf. Theory*, vol. IT-9, no. 3, pp. 136–143, Jul. 1963.
[39] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1186–1222, Mar. 2011.
[40] J. V. Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, vol. 100, no. 1, pp. 295–320, 1928.

Shachar Shayovitz (Member, IEEE) was born in Israel, in 1985. He received the B.Sc. degree in electrical and computer engineering from Technion (Israel Institute of Technology), Haifa, Israel, in 2007, and the M.Sc. degree (*cum laude*) in electrical engineering from Tel Aviv University, Tel Aviv, Israel, in 2013. He is currently pursuing the Ph.D. degree in the field of information theory with Tel Aviv University. Since 2013, he has been developing signal processing algorithms for Intel, Vayyar Imaging, and General Motors Research and Apple. His research interests include machine learning, statistical signal processing, and information theory.

Meir Feder (Life Fellow, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University in 1980 and 1984, respectively, and the Sc.D. degree in electrical engineering and ocean engineering from the Massachusetts Institute of Technology (MIT) and the Woods Hole Oceanographic Institution in 1987. After being a Research Associate and a Lecturer with MIT, he joined the School of Electrical Engineering, Tel-Aviv University, where he is currently a Chaired Professor and the Head of the newly established Tel-Aviv University Center for Artificial Intelligence and Data Science (TAD). He is also a Visiting Professor with the Department of EECS, MIT. Parallel to his academic career, he is closely involved with the high-tech industry. He founded five companies, among them are Peach Networks that developed an interactive TV solution (Acq: MSFT) and Amimon that provided the highest quality, robust and no delay wireless high-definition A/V connectivity (Acq:LON.VTC). Recently, with his renewed interest in machine learning and AI, he co-founded Run:ai, a virtualization, orchestration, and acceleration platform for AI infrastructure. He is also an Active Angel Investor and serves on the board/advisory board of several U.S. and Israeli companies. He received several academic and professional awards, including the IEEE Information Theory Society Best Paper Award for his work on universal prediction, the Creative Thinking Award of the Israeli Defense Forces, and the Research Prize of the Israeli Electronic Industry, awarded by the President of Israel. For the development of Amimon's chip-set, that uses a unique MIMO implementation of joint source-channel coding for wireless video transmission, he received the 2020 Scientific and Engineering Award of the Academy of Motion Picture Arts and Sciences.