Feedback Coding for Active Learning

Gregory Canal   Matthieu Bloch   Christopher Rozell

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA
{gregory.canal, matthieu, crozell}@gatech.edu

Abstract

The iterative selection of examples for labeling in active machine learning is conceptually similar to feedback channel coding in information theory: in both tasks, the objective is to seek a minimal sequence of actions to encode information in the presence of noise. While this high-level overlap has been previously noted, there remain open questions on how to best formulate active learning as a communications system to leverage existing analysis and algorithms in feedback coding. In this work, we formally identify and leverage the structural commonalities between the two problems, including the characterization of encoder and noisy channel components, to design a new algorithm. Specifically, we develop an optimal transport-based feedback coding scheme called *Approximate Posterior Matching* (APM) for the task of active example selection and explore its application to Bayesian logistic regression, a popular model in active learning. We evaluate APM on a variety of datasets and demonstrate learning performance comparable to existing active learning methods, at a reduced computational cost. These results demonstrate the potential of directly deploying concepts from feedback channel coding to design efficient active learning strategies.

1 INTRODUCTION

Active learning is an area of modern machine learning that studies how data points can be sequentially selected for labeling to train a model with as few labeled

examples as possible (Settles, 2009). Minimizing the number of labeled examples is critical in any learning scenario where labels are expensive to obtain, such as in healthcare applications where a medical expert must hand-label each training example (Liu, 2004), or where only a limited number of examples can be evaluated, such as in drug discovery (Warmuth et al., 2003).

The active selection of data points shares many technical parallels with channel coding with feedback, where a message is encoded into a sequence of symbols transmitted across a noisy channel and each symbol is selected based on the message and past channel outputs. In active learning, the optimal classifier parameters play the role of the "message" while the sequence of examples with noisy labels plays the role of "channel outputs" available through feedback to select the next example for labeling. Both feedback channel coding and active learning seek to minimize the number of encoder actions, leverage a history of noisy observations to select the next most informative action, must account for observation noise, and should operate in a computationally efficient manner. Although there exists a large literature studying the intersection of information theory with machine learning (Xu and Raginsky, 2017) and specifically active learning (Naghshvar et al., 2015), there remain open questions about the best ways to directly leverage techniques in channel coding for active example selection.

The main contribution of this work is a formulation of general active learning problems in terms of a feedback coding system, and a demonstration of this approach through the application and analysis of active learning in logistic regression. To motivate this approach, we first examine active learning through the lens of feedback channel coding by identifying communications system components, including a deterministic encoder, noisy channel, channel input constraints, and capacity-achieving distribution. With these components identified, we show how typical structural constraints in active learning problems prevent the direct application of existing feedback coding approaches such as *posterior*

matching (Ma and Coleman, 2011). We address this challenge by proposing *Approximate Posterior Matching* (APM), an optimal transport-based active learning scheme that extends posterior matching to account for the type of encoder constraints found in active learning problems.

To demonstrate the power of this approach, we apply APM to Bayesian logistic regression, a popular model in active learning. We identify the communication system components in logistic regression, derive a corresponding APM selection scheme (APM-LR), provide analytical results concerning each selected example's information content, and empirically demonstrate on several datasets how APM-LR attains a sample complexity comparable to other active logistic regression methods at a reduced computational cost. While this example scenario highlights the capabilities of APM as a specific data selection method, the feedback communications framework we develop provides a unified approach for designing and analyzing active learning systems in general.

1.1 Related Work

Modern active learning methods vary considerably in their approach to example selection, ranging from coreset construction (Pinsler et al., 2019; Sener and Savarese, 2018) and adversarial learning of informative examples (Sinha et al., 2019) to ensemble measures of example utility (Beluch et al., 2018) and Bayesian information acquisition methods (Gal et al., 2017; Kirsch et al., 2019). Bayesian active learning methods are intimately related to concepts in information and coding theory, and the intersection between these topics has a long history rooted in the study of sequential design of experiments (Lindley, 1956; Chernoff, 1959) and active hypothesis testing (Burnashev and Zigangirov, 1974). Since this early work, direct estimation and maximization of information gain has emerged as a popular active learning method (MacKay, 1992), and has been approximated for computational tractability (Houlsby et al., 2011). More recently, Naghshvar et al. (2015) have studied the direct application of an information-theoretic active hypothesis testing method to active learning problems. This method is limited to discriminating between a finite number of hypotheses (as opposed to estimating arbitrary model parameters) and to our knowledge has not been applied to popular machine learning models such as logistic regression. Other works have described at a high-level the similarities between active learning and coding with feedback over a noisy channel but do not exploit this observation to leverage existing coding schemes for example selection (Chen et al., 2015; Arias-Castro et al., 2013).

Posterior matching (Shayevitz and Feder, 2011; Ma

and Coleman, 2011) is a general feedback coding scheme that has been applied to tasks beyond telecommunications such as brain-computer interfacing (Omar et al., 2010; Tantiongloc et al., 2017) and aircraft path planning (Akce et al., 2010), but has limited application to example selection in active learning. Castro and Nowak (2008) study an active learning algorithm related to posterior matching that learns decision boundaries in discretized spaces, but does not directly maximize information about hyperplane parameters in a continuous space as we do here. More generally, to our knowledge existing work has not framed the task of active learning as a feedback communications system for the purpose of identifying an equivalent capacity-achieving distribution and selecting examples whose channel input distribution most closely approximates it, as we do here.

Logistic regression is a popular setting for the study of active learning, and has served as a testbed for the evaluation of competing example selection techniques. Yang and Loog (2018) surveyed modern active learning methods for logistic regression and evaluated them on many datasets. They generally found that uncertainty sampling and random sampling match or exceed the performance of more sophisticated (and computationally intensive) example selection methods. Uncertainty sampling, where examples closest to the estimated decision boundary are selected for labeling, is arguably the most popular active learning method for linear classification (Tong and Koller, 2001). Other active learning methods for linear classifiers are discussed in the literature related to learning halfspaces under bounded noise (Zhang et al., 2020).

2 ACTIVE LEARNING AS A COMMUNICATIONS MODEL

Let $\mathcal{U} \subseteq \mathbb{R}^d$ denote a pool of unlabeled examples from which at each training iteration $n \in \mathbb{N}$ an example $x_n \in \mathcal{U}$ is selected for labeling by an expert, who assigns label $Y_n \in \{1, 2, \ldots K\}$ according to a probabilistic model (all random variables are capitalized in this work). We consider a Bayesian framework in which we assume the existence of ground-truth model parameters $\theta \in \Theta$ distributed according to a prior p_θ that parameterizes a distribution $p(Y \mid x, \theta)$ governing the expert's labeling behavior. As is common in active learning, we assume that the labels $\{Y_n\}$ are independent when conditioned on θ. At each iteration n, a learning algorithm A is trained on a labeled dataset $\mathcal{L}_n = \{(x_i, y_i)\}_{i=1}^n$ (using lowercase to denote previously observed labels), resulting in a trained model with parameters $\widehat{\theta}_n \in \Theta$. The task of active learning is to design a policy π_n that, at each iteration, uses

the label history $\mathcal{L}_{n-1}$ to select example x_n from the remaining unlabeled examples $\mathcal{U}_n := \mathcal{U} \setminus \{x_i\}_{i=1}^{n-1}$, such that the classifier trains a generalizable model with as few labeled examples as possible.

In active logistic regression, θ encodes the weights of a linear separator, with $\Theta = \mathbb{R}^d$ (we consider only homogeneous logistic regression in this work). We assume a Gaussian prior $p_\theta \sim \mathcal{N}(0, \frac{1}{\lambda}I)$ with hyperparameter $\lambda > 0$. The label $Y \in \{-1, 1\}$ for data example x is assumed to be distributed according to

$$p(Y = 1 \mid x, \theta) = \frac{1}{1 + e^{-x^T\theta}}. \qquad (1)$$

Given a labeled dataset $\mathcal{L}$, we consider a maximum a posteriori (MAP) learning algorithm given by the convex program

$$\mathrm{A}(\mathcal{L}) = \arg\max_{\theta \in \mathbb{R}^d} \ \ln p_\theta \prod_{(x,y) \in \mathcal{L}} p(y \mid x, \theta)$$

$$= \arg\min_{\theta \in \mathbb{R}^d} \ \frac{\lambda}{2} \|\theta\|_2^2 + \sum_{(x,y) \in \mathcal{L}} \ln(1 + e^{-yx^T\theta}). \qquad (2)$$

Our key insight in this work is to define an intermediate variable $L = h_\theta(x)$, where $h_\theta(x) := x^T\theta$, and decompose the labeling distribution in (1) into $p(Y = 1 \mid x, \theta) = p(Y = 1 \mid L) = \frac{1}{1+e^{-L}}$. This decomposition of the labeling distribution into a deterministic function $h_\theta(x)$ and conditional distribution $p(Y \mid L)$ can be found in many machine learning models. For instance, in Bayesian neural networks (Gal et al., 2017), $h_\theta(x)$ is typically given by the composition of several nonlinear layers with $L = h_\theta(x)$ encoding the final layer feature vector, and $p(Y \mid L)$ is given by the softmax function. Figure 1a depicts this decomposition for logistic regression, and Figure 1b illustrates the full active learning decomposition in the general case.

By decomposing active learning in this manner, we are able to draw direct connections to feedback channel coding, in which a message θ is encoded into a sequence of symbols $\{L_n\}$, transmitted across a channel with transition probability $p(Y \mid L)$ yielding noisy output symbols $\{Y_n\}$, and subsequently decoded into an estimated message $\widehat{\theta}_n$. The availability of noiseless feedback from the channel output to the encoder provides the encoder with the history of received symbols, and allows it to adaptively select an informative channel input (Figure 1c). By comparing Figures 1b and 1c, we can see the direct correspondence between active learning and channel coding with feedback: model parameters θ serve as the message, which is encoded by function h (parameterized by x_n) into channel input $L_n = h_\theta(x_n)$. Label distribution $p(Y \mid L)$ can be interpreted as a noisy channel, with label Y_n as the channel output. Algorithm A decodes labeled data
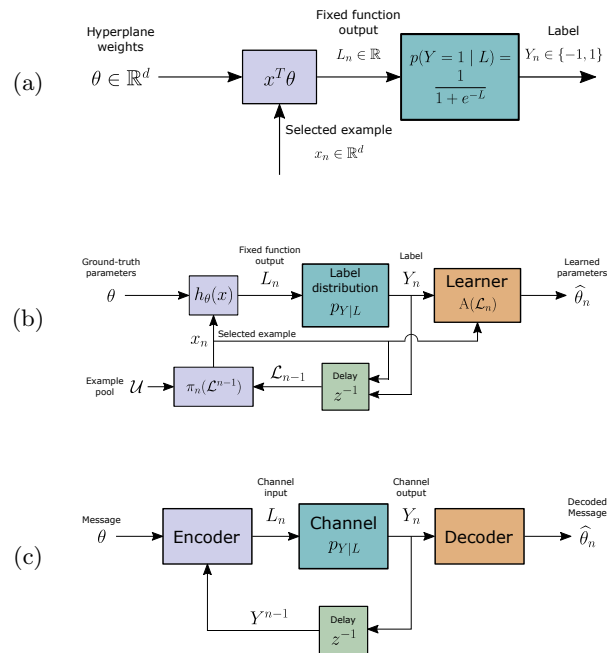

Figure 1: (a) Decomposition of logistic regression into an inner product between hyperplane θ and example x_n, and a logistic label distribution that depends only on this product. (b) Active learning decomposed into a deterministic function h, label distribution $p(Y \mid L)$, and feedback of the labeling history $\mathcal{L}_{n-1}$ to example selection policy π_n. (c) Coding with feedback, where a message is transmitted across a noisy channel as a sequence of symbols and subsequently decoded. By comparing (b) and (c), one can draw direct connections between active learning and coding with feedback.

$\mathcal{L}_n$ into a decoded message $\widehat{\theta}_n$, and $\mathcal{L}_n$ is passed as noiseless feedback to the encoder. This formulation of active learning as a feedback communications system allows one to leverage existing tools in channel coding for the design of an example selection scheme π_n. While similar decompositions have been observed in prior work (Naghshvar et al., 2015; Chen et al., 2015), we believe our work is the first to use this approach to analyze active learning in a real-world setting such as logistic regression.

2.1 Optimal Feedback Coding

In devising a feedback coding scheme for selecting a sequence of channel inputs $\{L_n\}$, there are several quantities that characterize optimal performance. We denote the *mutual information* $I(L;Y)$ between random variables L and Y as a function of marginal distribution p_L and conditional distribution $p_{Y|L}$ (using the notation p_L and $p_{Y|L}$ interchangeably with $p(L)$ and

$p(Y \mid L))$ given by $I(p_L, p_{Y|L})$:

$$I(p_L, p_{Y|L}) := \int_{L,Y} p_L \, p_{Y|L} \log_2 \frac{p_{Y|L}}{p_Y},$$

where p_Y denotes the output distribution of channel $p_{Y|L}$ with input distribution p_L. Letting $y^i := \{y_1, \ldots y_i\}$ denote the history of observed channel outputs, at iteration n we seek to maximize the *information gain* $I(\theta; Y_n \mid y^{n-1})$, which measures the one-step decrease in uncertainty about the message upon receiving each channel output. For deterministic encoders, information gain is equal to $I(L_n; Y_n \mid y^{n-1}) = I(p_{L_n|y^{n-1}}, p_{Y|L})$ (Cover and Thomas, 2006). Note that for a fixed channel $p_{Y|L}$, information gain is only a function of the channel input distribution $p_{L_n|y^{n-1}}$, conditioned on the history of channel outputs.

A key quantity in channel coding is the *channel capacity C*, defined as the maximum mutual information across the channel for any channel input distribution p_L within some class $\mathcal{C}$:

$$p_L^*(\mathcal{C}) := \underset{p_L \in \mathcal{C}}{\arg\max} \, I(p_L, p_{Y|L}) \quad C := I(p_L^*, p_{Y|L}).$$

The *capacity-achieving distribution $p_L^*(\mathcal{C})$* is the input distribution in $\mathcal{C}$ that maximizes information across the channel. Through achievability and converse arguments, a central result in information theory is that optimal coding schemes, when marginalized over the message set, should induce the capacity-achieving distribution on the channel input (Shannon, 1948). In working towards applying existing feedback coding schemes to active example selection, we first characterize the capacity-achieving distribution for logistic regression, which is a core contribution of our work and forms the basis of our novel active logistic regression scheme in Section 3.

Channel Capacity in Logistic Regression. Letting $f(\ell) := \frac{1}{1+e^{-\ell}}$, we observe from Figure 1a that logistic regression has a binary output channel with transition probability $p(Y = 1 \mid L) = f(L)$. Without constraints on the channel input, the information gain can be maximized by placing masses of equal weight at $\pm\infty$. However, logistic regression imposes the structural constraint $L = x^T\theta$, so that such a distribution would require data points of infinite energy for finite model weights. Therefore, to characterize logistic regression capacity in practice, we consider the capacity-achieving distribution within the class of power-constrained distributions given by $\mathcal{C}_P := \{p_L : \mathbb{E}[L^2] \leq P\}$; we discuss the selection of P in Section 3. With this class defined, we have our first result.

Proposition 2.1 (Capacity of Logistic Regression). *For $p(Y = 1 \mid L) = f(L)$, we have $p_L^*(\mathcal{C}_P) = B_{\sqrt{P}}$,*

where B_t is defined as $B_t(\ell) := \frac{1}{2}\delta(\ell - t) + \frac{1}{2}\delta(\ell + t)$ and δ denotes the Dirac delta function. Furthermore, we have $C = I(B_{\sqrt{P}}, f) = 1 - h_b(f(\sqrt{P}))$, where h_b denotes the binary entropy function.

The proof follows closely to that of Singh et al. (2009) for the one-bit quantized Gaussian channel; the proofs of Proposition 2.1 and all subsequent results are presented in the Supplement.

2.2 Posterior Matching

By characterizing the channel capacity and capacity-achieving distribution of active learning models, we enable the use of existing feedback coding schemes that achieve capacity. Recently, a capacity-achieving feedback coding scheme known as *posterior matching* has been developed to select a sequence of channel inputs $\{L_n\}$ to maximize the information gain across a given channel $p_{Y|L}$. The central concept is to construct an encoder that by definition induces $p_{L_n|y^{n-1}} = p_L^*$ for every n, which in essence hands the decoder the information that it is still "missing" (Ma and Coleman, 2011). This involves the construction of an encoder mapping $S_{y^{n-1}} : \theta \to L$ parameterized by y^{n-1} such that $S_{y^{n-1}}(\theta) \sim p_L^*$ for every n.

While posterior matching is an attractive feedback coding scheme, there are challenges in applying it to active learning: given the structural constraints of any particular active learning problem as depicted in Figure 1b, it may not always be the case that a mapping from $p_{\theta|\mathcal{L}_{n-1}}$ to p_L^* exists, since the encoder is constrained to the set of mappings given by $\{h_\theta(x) : x \in \mathcal{U}_n\}$.[1] For example, in active logistic regression under mild assumptions, there exists no x such that $h_\theta(x) \sim p_L^*$, as shown in the following proposition.

Proposition 2.2. *Under a log-concave prior distribution p_θ, in Bayesian logistic regression for any n there exists no x_n that induces $p_{L_n|\mathcal{L}_{n-1}} \sim p_L^*$.*

Since we assume a Gaussian prior p_θ (which is log-concave), Proposition 2.2 applies and therefore there exists no active logistic regression scheme π_n corresponding to a posterior matching mapping from θ to p_L^*. We suspect that the infeasibility of p_L^* holds generally in other real-world machine learning models (e.g., Bayesian neural networks) due to similar structural constraints imposed by $h_\theta(x)$, preventing the direct application of posterior matching for example selection. In the next section, we extend concepts from

---

[1] The analogous distribution to $p_{L_n|y^{n-1}}$ in active learning is $p_{L_n|\mathcal{L}_{n-1}}$. When considering only deterministic example selection schemes, $p_{L_n|\mathcal{L}_{n-1}}$ is induced directly from $p_{\theta|\mathcal{L}_{n-1}}$, through $h_\theta(x)$.

posterior matching to a novel active learning scheme compatible with this constrained encoder structure.

2.3 Approximate Posterior Matching

To address the impossibility of finding $x \in \mathcal{U}$ that induces p_L^* on L, we introduce a scheme that instead selects an example x_n such that $p_{L|\mathcal{L}_{n-1}}$ is distributed "as close as possible" to p_L^*, as measured by a distance between distributions. Specifically, we use the 2-Wasserstein distance because of its convenient geometric properties and compatibility with non-overlapping distribution supports (Arjovsky et al., 2017). The p-Wasserstein distance between distributions μ and ν is given by

$$ W_p(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu, \nu)} \int_u \int_v |u - v|^p \gamma(u, v) \right)^{\frac{1}{p}}, $$

where $\Pi(\mu, \nu)$ is the set of couplings with marginal distributions μ and ν (Villani, 2008). Our selection scheme, called *Approximate Posterior Matching* (APM), is then given by

$$ x_n = \pi_n(\mathcal{L}_{n-1}) := \arg\min_{x \in \mathcal{U}_n} W_2(p_{L_n|\mathcal{L}_{n-1}}, p_L^*). \quad (3) $$

While APM is intuitively appealing because it steers the induced channel distribution as close as possible to p_L^*, we justify this strategy in the next section for the case of logistic regression by showing that information gain does in fact approach its maximum possible value as $W_2(p_{L|\mathcal{L}_{n-1}}, p_L^*)$ is minimized.

3 APM IN LOGISTIC REGRESSION

Under the power constraint $\mathbb{E}[L^2] \leq P$, Proposition 2.1 establishes that the capacity-achieving distribution in the logistic regression system is given by $B_{\sqrt{P}}$. We now show an information continuity result for this capacity-achieving distribution, which provides a mathematical justification for the APM Wasserstein distance minimization in (3).

Theorem 3.1. *Let $\widetilde{C}_n = \max_{x \in \mathcal{U}_n} I(p_{L_n|\mathcal{L}_{n-1}}, f)$ denote the maximum information gain from any example selected at iteration n, and suppose $P > 0$ is selected such that $p_{L_n|\mathcal{L}_{n-1}} \in \mathcal{C}_P$ for any $x \in \mathcal{U}_n$. Then for any $x \in \mathcal{U}_n$,*

$$ \widetilde{C}_n - I(p_{L_n|\mathcal{L}_{n-1}}, f) \leq K_P W_2(p_{L_n|\mathcal{L}_{n-1}}, B_{\sqrt{P}}), $$

where $K_P > 0$ is a constant that only depends on P.

For decreasing $W_2(p_L, B_{\sqrt{P}})$, this result bounds $I(p_{L_n|\mathcal{L}_{n-1}}, f)$ towards its maximum possible information gain $\widetilde{C}_n$. In other words, minimizing the distance $\widetilde{C}_n$.

to the *known* capacity-achieving distribution (even if not achievable in practice) ensures that the information gain approaches its maximum value within the set of possible input distributions — a value which is *unknown* a priori. As we shall see in the results and experiments that follow, targeting the known capacity-achieving distribution affords geometric simplifications and computational benefits over the strategy of directly selecting the example that achieves $\widetilde{C}_n$. Unlike APM, the latter method does not benefit from analytical knowledge of the information structure of the channel and constraint set, and so it must instead conduct an expensive brute-force maximization of information gain.

3.1 Closed-form Results

For logistic regression, the calculation of $W_2(p_L, B_t)$ takes a convenient closed-form expression, which simplifies the example selection in (3):

Proposition 3.2. *For $t > 0$, with $\text{med}_{p_L}(L)$ denoting the median of L according to distribution p_L,*

$$ W_2^2(p_L, B_t) = \mathbb{E}_{p_L}[L^2] - 2t \, \mathbb{E}_{p_L}[|L - \text{med}_{p_L}(L)|] + t^2 $$

We can simplify this expression even further when p_L is normally distributed.

Corollary 3.2.1. *For $L \sim \mathcal{N}(\mu, \sigma^2)$,*

$$ W_2^2(p_L, B_t) = \mu^2 + \left(\sigma - \sqrt{\frac{2}{\pi}} t \right)^2 + \left(1 - \frac{2}{\pi} \right) t^2. $$

At iteration n, suppose that $p_{\theta|\mathcal{L}_{n-1}}$ is approximated by $\mathcal{N}(\mu_n, \Sigma_n)$, resulting in channel input $L_n = \theta^T x_n$ being distributed as $\mathcal{N}(\mu_n^T x_n, x_n^T \Sigma_n x_n)$. Although $p_{\theta|\mathcal{L}_{n-1}}$ is not normally distributed in logistic regression, it is common to make this approximation in practice (Bishop, 2006). By applying Corollary 3.2.1 and omitting constant terms, we derive our APM selection policy for logistic regression with power constraint P.

Definition 3.1. Approximate Posterior Matching for Logistic Regression (APM-LR):

$$ \pi_n(\mathcal{L}_{n-1}) = \arg\min_{x \in \mathcal{U}_n} (\mu_n^T x)^2 + \left(\sqrt{x^T \Sigma_n x} - \sqrt{\frac{2}{\pi} P} \right)^2. \quad (4) $$

This objective is a combination of two terms: the first term corresponds to minimizing the distance between example x and the posterior mean hyperplane. If μ_n is taken as an estimate of θ, this term corresponds to the well-known *uncertainty sampling* active learning method, which samples points close to the current hyperplane estimate (Tong and Koller, 2001). The second term prefers examples that align with the direction of maximum posterior covariance. Specifically, for
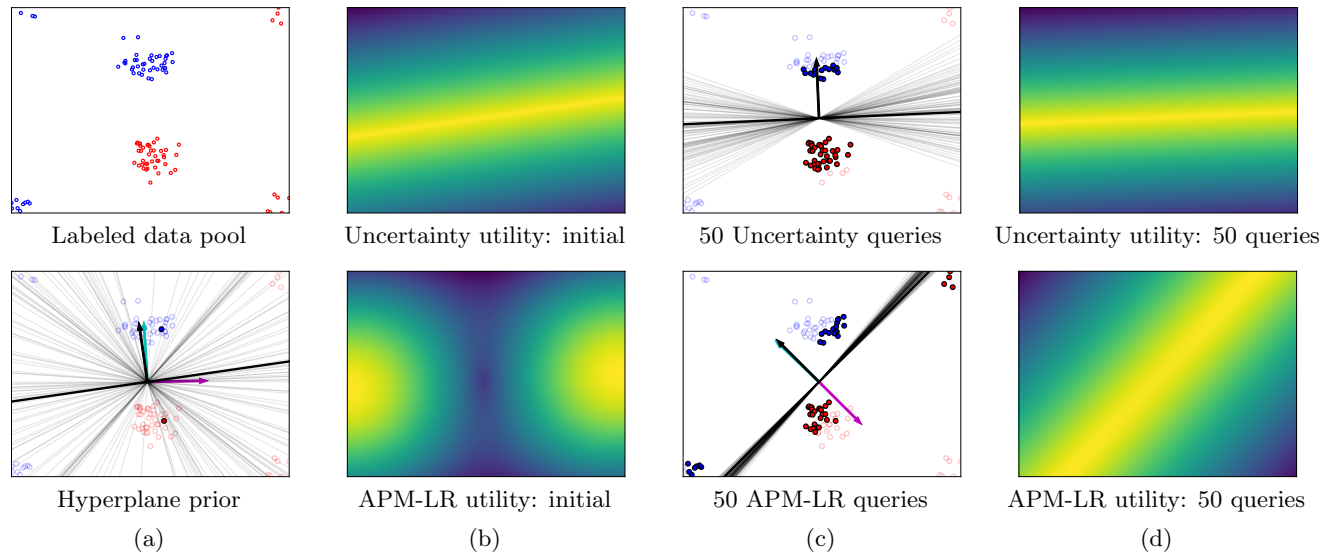
Figure 2: (a) Top: linearly separable dataset (optimal hyperplane is diagonal) demonstrating the failure of uncertainty sampling (dataset adapted from Huang et al. (2010)). Bottom: samples from hyperplane posterior, given two seed labels. The black hyperplane and corresponding normal vector depict the initial logistic regression solution, the cyan arrow indicates the normal vector to the posterior mean hyperplane, and the purple arrow indicates the maximal eigenvector of the posterior. (b) Utility function heatmap for uncertainty sampling (top) and APM-LR (bottom) — the unlabeled example with the highest utility is selected for labeling. Uncertainty sampling selects examples close to the current hyperplane, while APM-LR selects examples that are both close to the posterior mean hyperplane and align with the direction of largest posterior variance. (c-d) After 50 queries, uncertainty sampling (c-top) has not selected samples in the dataset corners, leading to sampling bias and continued sampling of the center clusters (d-top). Meanwhile, APM-LR (c-bottom) has sufficiently explored the dataset, while continuing to sample examples in only the most ambiguous regions (d-bottom).

$x^T \Sigma_n x < \frac{2}{\pi} P$, the second term is a decreasing function of $x^T \Sigma_n x$, encouraging x to align with posterior covariance eigenvectors with large eigenvalues.

These two terms together can be interpreted as encouraging "exploitation" and "exploration," respectively: the first term encourages the selection of examples that are close to the current estimate of θ, exploiting this estimate to only query examples whose labels are ambiguous. The second term balances this exploitation by probing in directions of the hyperplane posterior that have not yet been sufficiently explored, reducing uncertainty about the hyperplane itself. Figure 2 visualizes this tradeoff in comparison to uncertainty sampling, which only queries examples close to the current hyperplane estimate and does not account for the fact that there may be directions of the hyperplane posterior that have not been sufficiently explored. This myopic behavior is an instance of *sampling bias*, a well-known phenomenon in active learning where a policy continually selects examples that reinforce the learner's belief in an incorrect hypothesis (Dasgupta, 2011; Beygelzimer et al., 2009; Farquhar et al., 2021). The balance of exploitation and exploration terms in APM-LR helps prevent this type of sampling bias, in

a spirit similar to other active learning methods that balance uncertainty reduction with diverse example selection (Dasgupta and Hsu, 2008; Huang et al., 2010).

An attractive computational feature of (4) is that the posterior mean and covariance can be estimated *once* at each selection iteration and then simply projected onto each candidate example, resulting in a computational cost of only $O(d^2)$ per example evaluation. Note that these computational advantages along with the natural balance between exploration and exploitation in APM-LR emerged naturally from first-principles of feedback coding, demonstrating the potential of identifying the capacity-achieving distribution and applying APM as a universal means of designing geometrically intuitive, computationally efficient active selection schemes.

4   EXPERIMENTAL RESULTS

We evaluate the performance of APM-LR against baseline example selection methods for logistic regression on a variety of datasets from different tasks, as measured by holdout test accuracy and selection com-

pute time.[2] For each method, we follow Yang and Loog (2018) and set the regularization parameter in (2) to $\lambda = 0.01$, which we solve with the LIBLINEAR solver (Fan et al., 2008). After each example is labeled, we approximate $p_{\theta | \mathcal{L}_{n-1}}$ with a normal distribution by applying the variational approximation described in Jaakkola and Jordan (2000), which is solved in only a few iterations of an expectation-maximization procedure (referred to here as "VariationalEM"). The final component needed to apply APM-LR is the selection of power constraint P in (4).

Selecting Power Constraint Although our approach is rooted in feedback coding theory, regarding the power constraint there are two key differences between our model and traditional communications systems. First, unlike telecommunications systems that have physical restrictions such as limited battery levels, in our framework there is no external prescription of the power budget P and therefore we can select any valid upper bound on the channel input power induced by the unlabeled examples. Secondly, unlike coding schemes which globally maximize information gain over the entire trajectory of channel inputs, we seek to myopically maximize the one-step information gain at every channel input. Since the goal at each iteration is to separately solve a local information maximization problem, there is no need for the power constraint P to be constant across iterations, and therefore we set a separate power constraint P_n for each iteration.

Since the selection of P_n parameterizes the target distribution in APM-LR, it is important for P_n to be set as tight as possible so that the target capacity-achieving distribution is well-matched to the set of feasible channel input distributions. This is because at each iteration the capacity-achieving distribution serves as a proxy for the optimal input distribution induced by a real example, and a setting of P_n that is too loose will result in APM targeting a proxy that is not well-matched to the feasible input distributions. To select a satisfactory setting of P_n, we derive an upper bound on the channel input power to use as an implicit constraint.

Suppose for a given dataset that there exists a known $B > 0$ such that $\|x\|_2 < B$ (this is a reasonable assumption in many real-world settings). Let $\lambda_1(M)$ denote the largest magnitude eigenvalue of matrix M. We then have (with expectations taken with respect to $p_{L_n | \mathcal{L}_{n-1}}$)

$$\mathbb{E}[L_n^2] = x^T(\mu_n \mu_n^T + \Sigma_n)x \leq B^2 \lambda_1(\mu_n \mu_n^T + \Sigma_n).$$

For each n we can therefore set $P_n = B^2 \lambda_1(\mu_n \mu_n^T +$

Algorithm 1: Approximate Posterior Matching for Logistic Regression (APM-LR)

Input: data pool $\mathcal{X}$, hyperparameter $\lambda > 0$, horizon N, initial training set $\mathcal{L}$
1: $\mu \leftarrow 0$, $\Sigma \leftarrow \frac{1}{\lambda}I$
2: $B \leftarrow \max_{x \in \mathcal{U}} \|x\|_2$
3: $\mathcal{U} \leftarrow \mathcal{X}$
4: **for** $n = 1$ **to** N **do**
5: $\quad P \leftarrow B^2 \lambda_1(\Sigma)$
6: $\quad x^* \leftarrow \arg\min_{x \in \mathcal{U}}(\mu^T x)^2 + \left(\sqrt{x^T \Sigma x} - \sqrt{\frac{2}{\pi}P}\right)^2$
7: $\quad y^* \leftarrow \text{ExpertLabel}(x^*)$
8: $\quad \mathcal{U} \leftarrow \mathcal{U} \setminus \{x^*\}$, $\mathcal{L} \leftarrow \mathcal{L} \cup (x^*, y^*)$
9: $\quad \mu, \Sigma \leftarrow \text{VariationalEM}(\mathcal{L})$
10: $\quad \theta^* \leftarrow \text{A}(\mathcal{L})$ i.e., eq. (2)
11: **end for**
Output: hyperplane θ^*

Σ_n). In our experiments we select a slightly modified parameter $P_n = B^2 \lambda_1(\Sigma_n)$, which we justify as a more practical heuristic in the Supplement. We summarize APM-LR in full in Algorithm 1, including power constraint calculation and variational posterior updating.

Datasets We follow previous work in active learning for logistic regression (Huang et al., 2010; Yang and Loog, 2018) and test each method on several UCI datasets (Dua and Graff, 2017) including *vehicle*, *letter*, *austra*, and *wdbc*. We also evaluate performance on several synthetic datasets including the dataset depicted in Figure 2 (adapted from Huang et al. (2010)), which we refer to as *cross* (see Supplement for details on all datasets). For each simulation trial, we first randomly divide the dataset into an equally-sized data pool ($\mathcal{U}$) and held-out test set. We normalize $\mathcal{U}$ to zero-mean and coordinate-wise unit-variance, and apply the same transformation to the test set. Before evaluating each example selection method, the training dataset ($\mathcal{L}$) is seeded to consist of one randomly selected labeled example from each class.[3]

Baseline Methods We evaluate the following baseline methods, each described with their computational cost per candidate example evaluation (see Supplement for details):

- *Uncertainty*: select closest example to current hyperplane estimate (i.e. $\arg\min_{x \in \mathcal{U}_n} x^T \widehat{\theta}_{n-1}$) at cost $O(d)$. The action of Uncertainty sampling is comparable to that of the first term in (4).

- *Random*: each example is selected uniformly at ran-

---

[2]Code at https://github.com/siplab-gt/APM-LR

[3]Our experiments are synchronized across data selection methods: each trial uses the same training/test split and seed examples for each tested method.
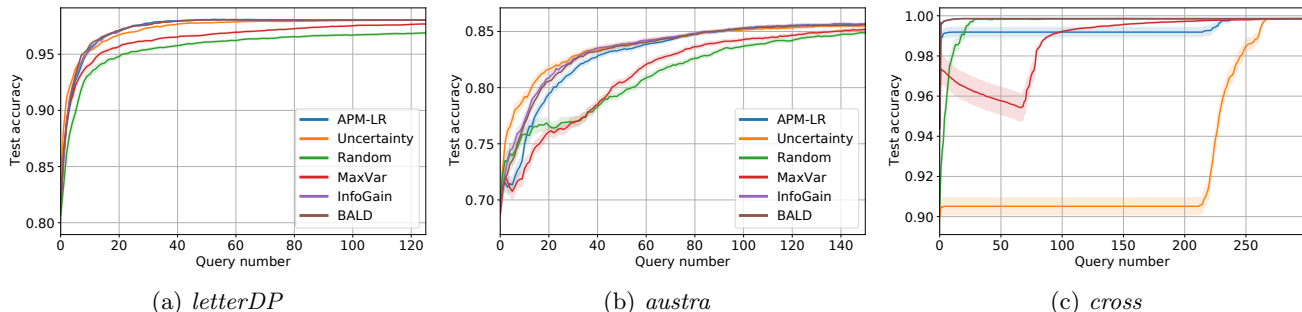
(a) *letterDP* (b) *austra* (c) *cross*

Figure 3: Average test classification accuracy plotted against number of labeled examples (error bars show ±1 standard error) across select UCI datasets (a-b) and the synthetic *cross* dataset (c, with legend shared with a-b and omitted for visual clarity). Overall, APM-LR performs comparably to other methods seeking to approximately maximize information gain. While uncertainty sampling performs well on some datasets (a-b), it can fail in cases where it suffers from sampling bias (c). Most of the tested active learning methods (except the control, MaxVar) outperform random sampling. For visual clarity we show different numbers of queried examples for each dataset.

dom from $\mathcal{U}_n$, at $O(1)$ cost.

- *MaxVar*: to isolate the effect of the second term in (4), we evaluate a control strategy that selects the example that induces the largest channel input variance (i.e. $\arg\max_{x \in \mathcal{U}_n} x^T \Sigma_n x$), at cost $O(d^2)$.

- *InfoGain*: selects the example with the largest information gain $I(\theta; Y_n \mid \mathcal{L}_{n-1})$, estimated by sampling s times from the normally approximated hyperplane posterior (here we set $s = 100$) and for each candidate example evaluating a Monte Carlo approximation of information gain, at $O(ds)$ cost.

- *BALD*: we approximate the logistic function $f(\ell)$ with a probit function and apply the probit regression active learning method of Houlsby et al. (2011), at cost $O(d^2)$. Like APM-LR, BALD approximates the action of InfoGain and only requires the mean and covariance of the normally approximated hyperplane posterior.

InfoGain is the most computationally intensive selection method, since it requires a brute-force Monte Carlo approximation of information gain for each candidate example. BALD and APM-LR have the next least expensive cost per example at $O(d^2)$, followed by Uncertainty and Random sampling.

Performance Comparison In Figure 3, we compare the learning performance of each data selection method by plotting holdout test accuracy against number of queried examples (excluding the seed set) across select datasets (see Supplement for full results). We generally find that the tested active data selection methods outperform random sampling. The exception is MaxVar, which performs comparably to random selection and worse than APM-LR. Although simple Un-

certainty sampling matches the performance of other active methods on several datasets (Figure 3a-b) as previously observed by Yang and Loog, in additional tests on synthetic datasets we find that APM-LR outperforms uncertainty sampling. This is the case for the *cross* dataset (Figure 3c), demonstrating how Uncertainty sampling can be susceptible to sampling bias that leads to insufficient exploration (see Supplement for additional failure mode analysis). These tests together lend evidence to the mixture of terms in (4) having combined benefits over pure exploration of directions with large posterior variance or pure exploitation of ambiguous examples with respect to the current hyperplane estimate. Finally, APM-LR generally performs similarly to InfoGain and BALD, both of which directly approximate the action of information gain maximization, in contrast to APM's geometric, indirect approach.

Table 1 depicts the computational cost for each method across select datasets (see Supplement for full results and expanded timing evaluations). Similar to the analysis in Yang and Loog (2018), for each method we evaluate the cumulative compute time to select the first 40 examples (excluding seed examples and time for model retraining), and compute the median time over all trials. We see that InfoGain is the most expensive of all methods, since it directly approximates information gain with Monte Carlo sampling. BALD has the next highest cost, followed by APM-LR — the two latter methods only require a single computation of posterior mean and variance, which can be projected onto each candidate example. Uncertainty sampling and random sampling have the lowest computational cost.

Although BALD can also be computed using only the

	letterDP	*austra*	*cross*
APM-LR	0.336	0.150	0.125
Uncertainty	0.149	0.063	0.053
BALD	4.230	1.770	1.521
InfoGain	12.755	5.089	2.722
Random	0.005	0.003	0.002
MaxVar	0.118	0.050	0.040

Table 1: Comparison of median cumulative time (s) for each method to select the first 40 examples (excluding seed points and time for model retraining). Generally, APM-LR has a cost an order of magnitude lower than InfoGain and BALD (which directly approximate the action of information maximization), while Uncertainty, MaxVar, and Random sampling have the cheapest cost.

posterior mean and covariance, it is unclear how the approximation in BALD can be applied beyond probit regression. In contrast, the APM formulation in (3) can be applied generally to *any* active learning problem that can be decomposed into a deterministic encoder and noisy channel, along with a known capacity-achieving distribution. The combined results of Figure 3 and Table 1 suggest that the universal APM approach of leveraging this analytical knowledge of the capacity-achieving distribution affords a geometric active selection approach that performs well in terms of both sample and computational complexity.

5 CONCLUSION

To our knowledge, our work is the first effort to both reframe active learning as a feedback communications system and utilize analytical knowledge of the corresponding capacity-achieving distribution to derive an active learning scheme. The analytical and empirical results in this work for the special case of logistic regression demonstrate the potential of this coding-based active learning approach: information continuity results show how examples selected with APM-LR have information gain approaching their maximum possible value, APM-LR has a convenient geometrical formulation resulting from analytical knowledge of the capacity-achieving distribution for logistic regression (characterized here for the first time) that can lead to computationally efficient example selection, and when tested on multiple datasets APM-LR performs comparably to baseline active learning methods including brute-force information maximization. APM-LR's attractive balance between exploration and exploitation emerged naturally from first-principles of channel coding, extending beyond the common approach of uncertainty sampling.

More generally, a fundamental feature of Approximate

Posterior Matching is that analytical knowledge of the capacity-achieving distribution converts the usually unwieldy *information maximization* problem in active learning to a *geometric* problem. In logistic regression, this geometry led to computational advantages over direct information maximization, and we conjecture that similar benefits may emerge in more complex settings. Additionally, the general formulation of APM in (3) presents several opportunities to leverage existing computational algorithms to aid example selection, including estimating p_L^* when it is analytically unknown (Blahut, 1972; Arimoto, 1972) and optimizing Wasserstein distances with state-of-the-art methods (Peyré and Cuturi, 2019). Overall, we believe that our coding-theoretic approach opens several new directions for future work in active learning.

References

Akce, A., Johnson, M., and Bretl, T. (2010). Remote teleoperation of an unmanned aircraft with a brain-machine interface: Theory and preliminary results. In *2010 IEEE International Conference on Robotics and Automation*, pages 5322–5327.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. (2013). On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481.

Arimoto, S. (1972). An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Beygelzimer, A., Dasgupta, S., and Langford, J. (2009). Importance weighted active learning. In

Bottou, L. and Littman, M., editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 49–56, Montreal. Omnipress.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Information science and statistics. Springer, New York, NY. Softcover published in 2016.

Blahut, R. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473.

Burnashev, M. V. and Zigangirov, K. (1974). An interval estimation problem for controlled observations. *Problems of Information Transmission*, 10(3):223—-231.

Castro, R. M. and Nowak, R. D. (2008). Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353.

Chen, Y., Hassani, S. H., Karbasi, A., and Krause, A. (2015). Sequential information maximization: When is greedy near-optimal? volume 40 of *Proceedings of Machine Learning Research*, pages 338–363, Paris, France. PMLR.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, USA.

Dasgupta, S. (2011). Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781. Algorithmic Learning Theory (ALT 2009).

Dasgupta, S. and Hsu, D. (2008). Hierarchical sampling for active learning. pages 208–215.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Farquhar, S., Gal, Y., and Rainforth, T. (2021). On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*.

Gal, Y., Islam, R., and Ghahramani, Z. (2017). Deep Bayesian active learning with image data. volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192, International Convention Centre, Sydney, Australia. PMLR.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (2011). Bayesian active learning for classification and preference learning.

Huang, S., Jin, R., and Zhou, Z.-H. (2010). Active learning by querying informative and representative examples. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 892–900. Curran Associates, Inc.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Kirsch, A., van Amersfoort, J., and Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

Liu, Y. (2004). Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Computer Sciences*, 44(6):1936–1941. PMID: 15554662.

Ma, R. and Coleman, T. P. (2011). Generalizing the posterior matching scheme to higher dimensions via optimal transportation. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 96–102.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604.

Naghshvar, M., Javidi, T., and Chaudhuri, K. (2015). Bayesian active learning with non-persistent noise. *IEEE Transactions on Information Theory*, 61(7):4080–4098.

Omar, C., Akce, A., Johnson, M., Bretl, T., Ma, R., Maclin, E., McCormick, M., and Coleman, T. P. (2010). A feedback information-theoretic approach to the design of brain–computer interfaces. *International Journal of Human–Computer Interaction*, 27(1):5–23.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pinsler, R., Gordon, J., Nalisnick, E., and Hernández-Lobato, J. M. (2019). Bayesian batch active learning as sparse subset approximation. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sener, O. and Savarese, S. (2018). Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shayevitz, O. and Feder, M. (2011). Optimal feedback communication via posterior matching. *IEEE Transactions on Information Theory*, 57(3):1186–1222.

Singh, J., Dabeer, O., and Madhow, U. (2009). On the limits of communication with low-precision analog-to-digital conversion at the receiver. *IEEE Transactions on Communications*, 57(12):3629–3639.

Sinha, S., Ebrahimi, S., and Darrell, T. (2019). Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tantiongloc, J., Mesa, D. A., Ma, R., Kim, S., Alzate, C. H., Camacho, J. J., Manian, V., and Coleman, T. P. (2017). An information and control framework for optimizing user-compliant human–computer interfaces. *Proceedings of the IEEE*, 105(2):273–285.

Tong, S. and Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., and Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, 43(2):667–673. PMID: 12653536.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 2524–2533. Curran Associates, Inc.

Yang, Y. and Loog, M. (2018). A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401 – 415.

Zhang, C., Shen, J., and Awasthi, P. (2020). Efficient active learning of sparse halfspaces with arbitrary bounded noise.